

On key modulators of higher-order chromatin structure



André J. Faure

EMBL – European Bioinformatics Institute

Magdalene College

A dissertation submitted to the University of Cambridge for the degree of

Doctor of Philosophy

September, 2013

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

No part of this work has been submitted or is currently being submitted for any other qualification.

This document does not exceed the word limit of 60,000 words as defined by the Biology Degree Committee.

September 2, 2013

André J. Faure

To Jean, Mary, Somerset and Guido

Acknowledgements

First and foremost, thanks are due to my supervisor Paul Flicek. His guidance, patience, trust and encouragement towards independence were crucial ingredients that made this work possible.

All of the research presented in this thesis is the result of fruitful collaborations with excellent and enthusiastic experimental collaborators in the groups of Duncan Odom and Matthias Merkschlager, the heads of which also provided valuable mentorship. It has been a stimulating experience and great pleasure working with Dominic Schmidt, Hegias Mira-Bontenbal and Vlad Seitan.

I was fortunate to have a friendly and helpful thesis advisory committee comprising Nicholas Luscombe, Eileen Furlong and Florian Markowetz. I would also like to thank my former supervisors, Nicola Mulder and Cathal Seoighe, who fostered my initial interest in computational biology, leading me to take up the challenge presented by this work.

Thanks are also due to the EMBL PhD programme for funding, as well as the EMBL-EBI and Sanger Institute PhD communities for well-organised, supportive and dynamic student environments. I am greatly indebted to all colleagues at the EMBL-EBI for helpful advice, discussions and those who proofread my thesis, including Petra Schwalie, Albert Vilella, John Marioni, Graham Ritchie, David Thibert, Mikhail Spivakov, Camille Berthelot, Ângela Gonçalves, Steven

Wilder, Anika Oellrich and Valentina Lotchkova. Regular lively research meetings with members of the Odom and Flicek groups created a stimulating and creative environment for generating new ideas and potential research directions.

In particular, I would like to thank the wonderful folks with which I shared my office over past few years. These include current and former members of Paul's research group (Petra, David, Albert, Graham, Tom, Camille, Emily) and members of Ewan Birney's research group (Daniel, Alison, Mikhail, Markus, Dace, Sander, Nils, Sandro). Day-to-day interactions, the friendly and helpful atmosphere, as well as the weekly "Fliceker's" meetings made my experience at the EMBL-EBI a fun and memorable one. The absurd comments, cultural peculiarities, ridiculous memes and constant supply of sweet treats will not be easily forgotten.

It has been a great privilege to be a student at Cambridge and to call Audley Cottage, Magdalene College my home. The unique experiences and special relationships forged with housemates, friends and colleagues will stay with me forever.

A very special thanks to my two grandfathers, Guido and Somerset whose love of science and the natural world inspired me from a young age to pursue my passion. Lastly to my dear parents, Mary and Jean, and siblings, Mario, Jeanita, Erin and Andrew, thank you so much for your support, encouragement and generosity. Your unfailing love and belief in me has always sustained me.

Abstract

The billions of DNA bases that comprise mammalian genomes encode the instructions to assemble complex organisms consisting of hundreds of different cell types. Apart from the challenge of packaging this genetic material into a nuclear space that is six orders of magnitude smaller than its linear sequence, the molecules need to be organised in such a way as to ensure the coordinated activity of their twenty-odd thousand genes. Regulatory elements recruit specific factors to direct tissue-specific expression programmes, but the contribution of three-dimensional chromatin structure to this process is still unclear.

Cohesin and CTCF are two ubiquitous and highly conserved nuclear factors that have emerged as key modulators of higher-order genome organisation. In this thesis I integrate genome-wide binding, expression and chromatin conformation data in order to better characterise the diverse functions of these proteins. The main conclusions from this analysis are three-fold: (i) CTCF-independent cohesin binding is associated with cell-type-specific gene regulation and the stabilisation of highly occupied *cis*-regulatory modules, (ii) cohesin-based interactions within pre-existing architectural compartments enable discrete gene expression states, and (iii) the CTCF paralog CTCFL/BORIS perturbs gene regulation and displaces cohesin from promoter-proximal sites when present in somatic cells.

Results from these three studies provide further insight into the roles of cohesin and CTCF in the context of tissue-specific expression, and strengthen the link between these proteins and the dynamic control of genome architecture.

Contents

Contents	vi
List of Figures	x
List of Abbreviations	xiii
1 Introduction	1
1.1 Genomes, gene regulation and chromatin structure	1
1.1.1 Transcriptional regulation	3
1.1.2 Chromatin structure	5
1.1.2.1 DNA sequence and first-order structure	6
1.1.2.2 Nucleosomes, histone variants and modifications	9
1.1.2.3 Higher-order chromatin and looping structures	13
1.1.2.4 Chromatin compartmentalisation	17
1.1.2.5 Chromosome territories, nuclear bodies and sub-nuclear position	20
1.2 CCCTC-binding factor, CTCF	21
1.3 The cohesin protein complex	24
1.4 Brother of Regulator of Imprinted Sites, CTCFL	25
1.5 High-throughput approaches in functional genomics	27
1.5.1 High-throughput sequencing technologies	28
1.5.2 Chromatin state assays	29
1.5.2.1 ChIP-seq data analysis	33
1.5.3 Transcriptomic assays	35
1.5.3.1 Differential expression analysis with RNA-seq	36

1.5.4	Chromatin conformation assays	37
1.5.4.1	Hi-C data analysis	39
2	Cohesin regulates tissue-specific expression by stabilising highly occupied <i>cis</i>-regulatory modules	42
2.1	Summary	42
2.2	Introduction	43
2.3	Results	44
2.3.1	CTCF-independent cohesin binding is associated with master regulators and enhancers	46
2.3.2	CNC sites occur preferentially at multiply-bound <i>cis</i> -regulatory modules (CRMs)	52
2.3.3	CNC presence is associated with liver-specific gene expression	55
2.3.4	Maximally occupied CRMs show similar properties to HOT regions	57
2.3.5	Cohesin intensity explains disparities between motif score and ChIP signal	58
2.3.6	ONECUT1 ChIP signal is reduced at weak motifs in heterozygous <i>Rad21</i> ^{+/-} mouse liver cells	60
2.3.7	Mirrored binding of CTCF near transcription start sites and cohesin-bound enhancers is associated with elevated expression levels	64
2.4	Discussion	66
2.5	Methods	69
2.5.1	Experimental methods	69
2.5.1.1	ChIP sequencing	69
2.5.2	Computational methods	70
2.5.2.1	Read mapping and peak calling	70
2.5.2.2	Cohesin-non-CTCF site definition and peak clustering	70
2.5.2.3	Motif analysis and selection	71
2.5.2.4	Mouse ES cell data analysis	71
2.5.2.5	CRM clustering and analysis	71

2.5.2.6	Expression analysis	74
2.5.2.7	Motif presence prediction	75
2.5.2.8	Wild type versus heterozygous <i>Rad21</i> ^{+/-} differential binding analysis	75
3	Cohesin-based chromatin interactions enable regulated gene expression within pre-existing architectural compartments	76
3.1	Summary	76
3.2	Introduction	77
3.3	Results	79
3.3.1	Cohesin binding predicts perturbed long-range interactions in cohesin-deficient thymocytes	79
3.3.2	Cohesin depletion perturbs gene expression in open compartments	84
3.3.3	Genes that are sensitive to cohesin dosage are bound by cohesin, CTCF and NIPBL	85
3.3.4	Predictive features of genes that show cohesin-dependent expression	85
3.3.5	Cohesin depletion perturbs long-range interactions within architectural compartments and compresses the dynamic range of gene expression	91
3.4	Discussion	96
3.5	Methods	99
3.5.1	Experimental methods	99
3.5.2	Computational methods	100
3.5.2.1	ChIP-seq read mapping and peak calling	100
3.5.2.2	RNA-seq data analysis	100
3.5.2.3	Hi-C data analysis	100
3.5.2.4	Multinomial logistic regression model	102

4	The CTCF paralog CTCFL/BORIS regulates gene expression and displaces cohesin from promoter-proximal sites when expressed in somatic cells	103
4.1	Summary	103
4.2	Introduction	104
4.3	Results	105
4.3.1	CTCFL associates with the majority of active promoters in human K562 leukaemia cells and mouse ES cells	105
4.3.2	In contrast to CTCF, CTCFL binds clusters of GC-rich motifs and low complexity repeats	108
4.3.3	CTCFL-regulated genes are enriched for active, CGI promoters, imprinting and relevance to cancer	113
4.3.4	The relationship between CTCFL and cohesin binding	116
4.4	Discussion	117
4.5	Methods	122
4.5.1	Experimental methods	122
4.5.1.1	ChIP sequencing	122
4.5.1.2	Microarray experiments	122
4.5.2	Computational methods	123
4.5.2.1	ChIP-seq read mapping and peak calling	123
4.5.2.2	Motif analysis	123
4.5.2.3	Microarray analysis	124
5	Conclusions and future work	125
	Publications	131
	References	132

List of Figures

1.1	Regulatory interactions involving transcription by RNAP2.	4
1.2	Multiple levels of chromatin structure.	7
1.3	The relationship between genomic sequence, chromatin state and transcriptional activity.	11
1.4	Long-range regulation by chromatin looping at selected developmental loci.	15
1.5	Higher-order organisation of chromatin into compartments and TADs.	18
1.6	Models for cohesin function in sister chromatid cohesion and loop formation, with and without CTCF.	23
1.7	Schematic representation of the Illumina/Solexa sequencing process.	30
1.8	Diagram showing wet-lab workflow of a typical ChIP-seq experiment.	32
1.9	3C-based methods to assay chromatin conformation.	38
2.1	Genome-wide distribution of CRMs in primary mouse liver as measured by ChIP-seq.	45
2.2	Cohesin binding and RNAP2 pausing.	47
2.3	Cohesin peak shift with respect to sequence-specific factors. . . .	48
2.4	Within-CRM binding correlations reveal distinct modes of cohesin binding in mouse liver cells.	50
2.5	Within-CRM binding correlations reveal distinct modes of cohesin binding in mouse ES cells.	51
2.6	Cohesin-non-CTCF (CNC) binding occurs preferentially at multiply-bound CRMs.	53
2.7	<i>AutoClass</i> CRM clustering results.	54

LIST OF FIGURES

2.8	CNC sites are associated with liver-specific gene expression. . . .	56
2.9	Cohesin ChIP signal is significantly associated with TF motif score.	59
2.10	Performance results of all motif classifiers.	61
2.11	ONECUT1 ChIP-seq in <i>Rad21</i> ^{+/-} cells reveals preferential loss of binding events at sites without motifs.	63
2.12	Simultaneous CTCF binding within promoters and nearby en- hancers is associated with elevated expression levels.	65
2.13	Motifs and motif statistics.	72
2.14	Choice of <i>K</i> for K-means CRM clustering analysis.	73
3.1	Chromosomal compartments are resilient to the depletion of the cohesin subunit RAD21 from non-cycling thymocytes in vivo. . .	80
3.2	Cohesin depletion perturbs long-range interactions in cohesin-deficient thymocytes.	82
3.3	Cohesin binding predicts perturbed long-range interactions in cohesin- deficient thymocytes.	83
3.4	Differentially expressed genes in cohesin-deficient thymocytes as assayed by RNA-seq.	84
3.5	Characteristics of cohesin-sensitive genes.	86
3.6	Univariate analysis of predictors of differential gene expression in cohesin-deficient thymocytes.	88
3.7	Multivariate analysis of predictors of differential gene expression in cohesin-deficient thymocytes.	89
3.8	Investigating top predictors of differential gene expression in cohesin- deficient thymocytes.	90
3.9	Impact of cohesin-deficiency on homotypic cohesin-based and al- ternative interactions.	92
3.10	Impact of cohesin-deficiency on all pair-wise feature-based interac- tions.	93
3.11	Selectivity of increased interactions and the effect of compartment size on all pair-wise feature-based interactions.	94
3.12	Length scale of lost and gained cohesin-dependent feature-based interactions.	95

LIST OF FIGURES

3.13	Cohesin depletion compresses the dynamic range of gene expression.	97
4.1	Schematic alignment of mouse CTCFL and CTCF protein sequences.	104
4.2	Venn diagrams showing the overlap between CTCF and CTCFL binding events in K562 and mouse ES cells.	106
4.3	Heatmap representation of chromatin ChIP-seq fragment profiles near CTCFL and CTCF peaks.	107
4.4	Validation of RNAi-mediated knockdown of CTCFL in K562 cells and characterisation of FLAG-CTCFL ES cells.	108
4.5	CTCFL associates with the majority of active promoters in human K562 leukaemia cells and mouse ES cells.	109
4.6	<i>De novo</i> motif characterisation and distribution within CTCF and CTCFL-bound regions.	110
4.7	CTCFL binding preferences are stronger than CTCF for a diverse set of GC-rich motif words.	112
4.8	CTCFL binding is enriched in regions with high GC content and low DNA sequence complexity.	113
4.9	Impact of CTCFL on gene expression in K562 cells and ES cells. .	114
4.10	Impact of CTCFL on gene class expression in K562 cells and ES cells.	115
4.11	Preferential RAD21 depletion at promoter-proximal CTCFL sites in FLAG-CTCFL mouse ES cells.	118
4.12	RAD21 depletion at promoter-distal CTCFL sites in FLAG-CTCFL mouse ES cells.	119
4.13	Validation of reduced RAD21 binding at CTCFL sites in FLAG-CTCFL mouse ES cells.	120

List of Abbreviations

3C chromosome conformation capture

4C circular 3C

5C 3C-carbon copy

ATP adenosine triphosphate

BORIS Brother of Regulator of Imprinted Sites

cDNA complementary DNA

CGI CpG island

ChIA-PET chromatin interaction analysis by paired-end tag sequencing

ChIP chromatin IP

CNC cohesin-non-CTCF

CpG 5-CG-3

CRM *cis*-regulatory module

CT chromosome territory

CTCF CCCTC binding factor

CTCFL CTCF-like

DHS DNase hypersensitive site

DNase deoxyribonuclease 1

DNA deoxyribonucleic acid

eRNA enhancer RNA

ENCODE Encyclopedia Of DNA Elements

ER estrogen-receptor

ES embryonic stem

FAIRE Formaldehyde-Assisted Isolation of Regulatory Elements

FDR false discovery rate

FISH fluorescent *in situ* hybridisation

GO Gene Ontology

GTF general transcription factor

HCP high CpG content promoter

HOT high occupancy target

ICE iterative correction and eigenvector decomposition

ICR imprinting control region

IP immunoprecipitation

lncRNA long ncRNA

LAD lamina-associated domain

LCP low CpG content promoter

LCR locus control region

mRNA messenger RNA

modENCODE Model Organism ENCODE

MNase micrococcal nuclease

ncRNA non-coding RNA

NFR nucleosome-free region

NGS	next-generation sequencing
PCA	principle components analysis
PCR	polymerase chain reaction
PIC	pre-initiation complex
PRC	polycomb repressive complex
PWM	position weight matrix
qRT-PCR	quantitative reverse transcription PCR
rRNA	ribosomal RNA
RNA	ribonucleic acid
RNAi	RNA interference
RNAP2	RNA polymerase II
RRBS	reduced representation bisulphite sequencing
siRNA	small interfering RNA
SIMA	structured interaction matrix analysis
SINE	short-autonomous interspersed element
SMC	structural maintenance of chromosomes
SNP	single-nucleotide polymorphism
tRNA	transfer RNA
TAD	topologically associating domain
Tcra	T-cell receptor alpha
TF	transcription factor
TSS	transcription start site
UCNE	ultraconserved non-coding elements
WT	wild type

Chapter 1

Introduction

1.1 Genomes, gene regulation and chromatin structure

Mouse formatting conventions are used for gene and protein symbols throughout this thesis as nearly all of the results are based on data from mouse liver, embryonic stem cells and thymocytes.

The head-spinning pace of advancements in molecular biology beginning in the 20th century has been characterised by the fusion and synergy between previously separate scientific fields, catalysed by the availability of enabling technologies. Less than a century of research has provided a highly detailed understanding of the basic mechanisms and molecules central to the functioning and evolution of all living things: nucleic acids and proteins.

The convergence of two fields in particular: biochemistry and genetics, traces a path from discoveries such as the residing of genes – the basic units of inheritance – on chromosomes, to the molecular structure of DNA and culminating with the reference sequence of the human genome at the turn of the 21st century^[1]. Together with the availability of reference sequences of several model organisms including baker’s yeast^[2], *Arabidopsis thaliana*^[3], worm^[4] and fruit fly^[5], this signalled the start of an era of genome-wide analyses. Thanks to this revolution, scientists are now able to probe the complex interplay between nucleic acids and proteins on a whole-genome scale, which

has dramatically accelerated biomedical research^[6].

In hindsight, new discoveries have birthed entirely new fields of research, with qualitative studies giving way to more data-intensive approaches. The deluge of data generated by high-throughput experiments has increasingly called for more sophisticated statistical and computational approaches, both in terms of data collection and processing, as well as in downstream analysis and interpretation. Global “omics” fields of research, spearheaded by genomics, proteomics and metabolomics, are now codependent on their data-centric partners which include bioinformatics, computational biology and systems biology. The explicit aim of the latter is the integration of diverse sources of data in order to gain insight into emergent properties of biological systems assumed to be intractable by classical reductionist approaches^[7]. To achieve this goal and keep up with the pace of developments in molecular biology, it is clear that inter-disciplinary approaches will continue to be important.

The complexity of living things is a result of processes which occur at the cellular level, starting with the control and organisation of information encoded in their DNA. However, with the availability of sequenced genomes covering the breadth of the tree of life, it has become clear that neither genome size nor total gene count correlate well with organismal complexity^[8]. This suggests that the regulation of gene expression is fundamental to complex processes like normal development, cellular differentiation and homeostasis. Alterations in spatial and/or temporal gene expression regulation can have dramatic phenotypic effects and some studies suggest that this may be a particularly powerful, if not the dominant, mode of evolution^[9]. Insight into mechanisms of gene regulation and mis-regulation is essential for the understanding of cancer and other diseased states, and is expected to be important for the future development of medicines and other tailor-made therapeutics.

In my thesis I use computational approaches to study mechanisms of gene regulation through the lens of chromatin structure and the key DNA-binding proteins which modulate it. The aim of my research has been to better understand the complex interplay between genome sequence, chromatin structure and transcription, particularly in the context of cell-type-specific gene expression. In the following sections I provide a brief overview of eukaryotic transcriptional regulation, followed by a review of the multiple levels of chromatin structure and the ways in which these are known to relate

to changes in the expression of associated genes.

1.1.1 Transcriptional regulation

Gene expression results in functional gene products from the relatively stable form of genetic information stored in the genome. The process consists of multiple steps, with the first being transcription from DNA into RNA. The regulation of this process, which ultimately affects the concentration of the resulting gene product, is the topic of this section. Downstream events, each providing opportunities for additional regulation and fine-tuning, include RNA processing, nuclear transport, and in the case of proteins, translation and post-translational modification.

In higher eukaryotes, the role of transcription is shared by three separate RNA polymerases (I, II and III), where RNA polymerase II (RNAP2) transcribes all protein-coding genes as well as some small untranslated RNAs with regulatory functions^[10]. RNA polymerase I (RNAP1) and III (RNAP3) each have specialised roles in the transcription of large ribosomal RNA (rRNA) genes, 5S RNA and transfer RNA (tRNA) genes. Transcription by the large RNAP2 molecular machine is itself a complex multistep process beginning with the assembly of the pre-initiation complex (PIC) within the core promoter region near the transcription start site (TSS; Figure 1.1). The PIC consists of RNAP2 in complex with general transcription factors (GTFs), which despite their high number (12), are all necessary for accurate initiation of transcription^[11].

Although these components of the basal transcriptional machinery are also sufficient for transcription initiation from core promoter elements *in vitro*, this is not generally the case within normal cellular conditions. The combination of various proteins and RNAs which together help to package DNA into chromatin (see Section 1.1.2), present a barrier to the intrinsic transcriptional ability of the PIC^[11]. Therefore, in addition to RNAP2 and the GTFs, the action of DNA-binding transcription factors (TFs) in association with their corresponding coactivators, is necessary to overcome these repressive conditions and result in a net “activated” transcriptional state. Promoter-proximal DNA sequence elements as well as more distant enhancers – in general referred to as *cis*-regulatory sequences – in part define the TF binding neighbourhood of specific target genes^[12] (Figure 1.1). These elements tend to be clustered in discrete regions

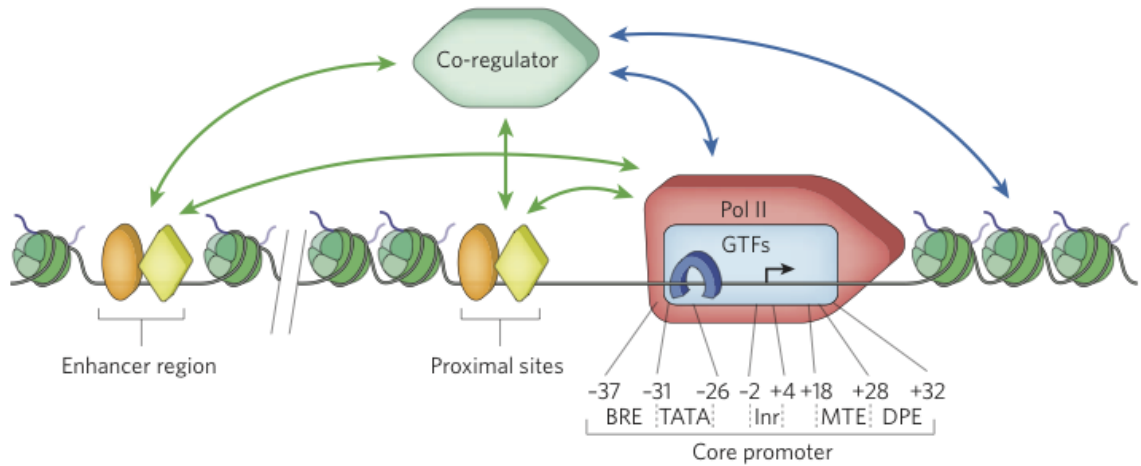


Figure 1.1: Regulatory interactions involving transcription by RNAP2. GTFs as well as RNAP2 are shown localised within the core promoter. TFs (orange and yellow shapes) bound to promoter-proximal as well as distal regulatory elements (e.g. enhancers) act primarily via co-regulators to affect the basal transcriptional machinery and chromatin status. Figure adapted from Fuda et al.^[12]

termed *cis*-regulatory modules (CRMs), which facilitates the coordinated binding of multiple TFs^[13]. Also, unlike components of the basal transcriptional machinery, the expression of many TFs is tightly restricted to particular developmental time-points, tissues and/or cell-cycle stages^[14]. This enables the appropriate control of complex expression programmes in eukaryotes. In view of their pivotal cellular role, the large ENCODE^[15] and modENCODE^[16] projects have mapped the binding positions of a collection of TFs across diverse tissues to help catalog the complete set of regulatory elements in the genome. However, understanding precisely how these networks of regulatory elements act alone and in concert to perform their functions remains a major goal in the field of functional genomics.

TFs can act directly on the transcriptional machinery at the core promoter, but generally do so in conjunction with co-regulators, which are often multi-protein complexes. Co-regulator mechanisms include the modulation of PIC assembly, as well as the downstream control of transcriptional initiation and elongation^[11]. The multiple discrete steps involved in the production of a complete RNA provide many points of

regulation. By measuring the accumulation of RNAP2 and its various covalently modified forms along the length of the transcribed region, genes can be classified according to their transcriptional state^[17]. This can help determine important rate-limiting steps in the process as well as the impact of specific co-regulators. Effects on transcriptional output are positive for activator TFs and their positive co-regulators (coactivators), and negative for repressor TFs and their corresponding negative co-regulators (corepressors). In the case of activators, an additional layer of regulation is provided in the form of the aptly named mediator complex, which is required for interactions with GTFs and to increase their activity^[18]. Co-regulators can also act more indirectly by modifying chromatin structure in a myriad of ways, but broadly their effect is to either create a more permissive local environment or to reduce the accessibility of regulatory regions to TFs and the core transcriptional machinery^[11].

The availability of TFs and their diverse binding preferences combined with the multiple ways in which co-regulators can modulate transcription and the surrounding chromatin architecture, illustrates the extraordinary combinatorial nature of eukaryotic transcriptional regulatory control. Indeed, recent results showing the surprisingly rapid rate of TF binding divergence in closely related species suggests that transcriptional regulatory landscapes are fertile ground for evolutionary innovation^[19,20].

1.1.2 Chromatin structure

The DNA contained in one human cell is about two meters long, but is organised in such a way that it fits within a nucleus that is six orders of magnitude smaller than its linear sequence. Far from being static, the three-dimensional (3D) structure of eukaryotic genomes is highly dynamic, most obviously reflecting functional requirements when compared between different cell cycle stages. In mitotic cells, the genome is coerced into the iconic shape of condensed chromosomes, which facilitates faithful transmission of genetic information to daughter cells. However, during interphase, the genome requires a more complex organisation that allows for the regulated transcription of genes, as well as DNA replication and repair.

Eukaryotic genomes are packaged in chromatin, which consists of DNA in complex with various proteins and structural RNAs (Figure 1.2A). There are a number of factors

that influence the structure and composition of chromatin at different resolutions, from the covalent modification of individual bases, to the position of entire chromosomes within the nucleus. In the following sections I briefly review these different levels of chromatin structure and describe the relevance of each in the context of transcriptional regulation.

1.1.2.1 DNA sequence and first-order structure

Certain aspects of chromatin structure, for example its bending properties, are directly related to nucleotide composition. For example, the structure of polyA (or polyT) runs are inherently stiff^[23], whereas AT-rich dinucleotides are more flexible. These aspects, to a large extent, determine the placement of nucleosomes along the underlying DNA sequence and this in turn can influence the accessibility of *cis*-regulatory elements to TFs^[24]. In the case of the regulatory region immediately upstream of the TSS – a major position of functional TF binding sites – sequence properties which directly influence nucleosome occupancy have been shown to intrinsically encode gene expression variability^[25].

The dinucleotide CG (or CpG) is a particularly important regulatory sequence in vertebrates and occurs in three known forms: unmethylated, methylated (5mC) and hydroxymethylated (5hmC). Methylated cytosine residues in this sequence, catalysed by the action of DNA methyltransferases, are generally associated with a repressed or silent chromatin state. Although the hydroxymethylated form and the TET enzymes which facilitate its conversion from 5mC have only very recently begun to be studied on a genome-wide scale, there are clues that this mark may be particularly important as a demethylation intermediate during developmental reprogramming events^[26] (see below). With a typical genome G+C (or GC) content of 40%, the observed CpG frequency of one every ≈ 100 base pairs (bp) represents an approximate four-fold depletion. In view of the fact that the great majority of CpGs are methylated (80%), this is thought to be due, in part, to the hypermutability of 5mC to thymine (T). This would also account for the preferential evolutionary retention of CpG islands (CGIs) – approximately 1000 bp regions of locally increased CpG density near most human TSSs – which tend to lack methylation^[27].

The status of cytosine residues represents an impediment to the binding of methylation-

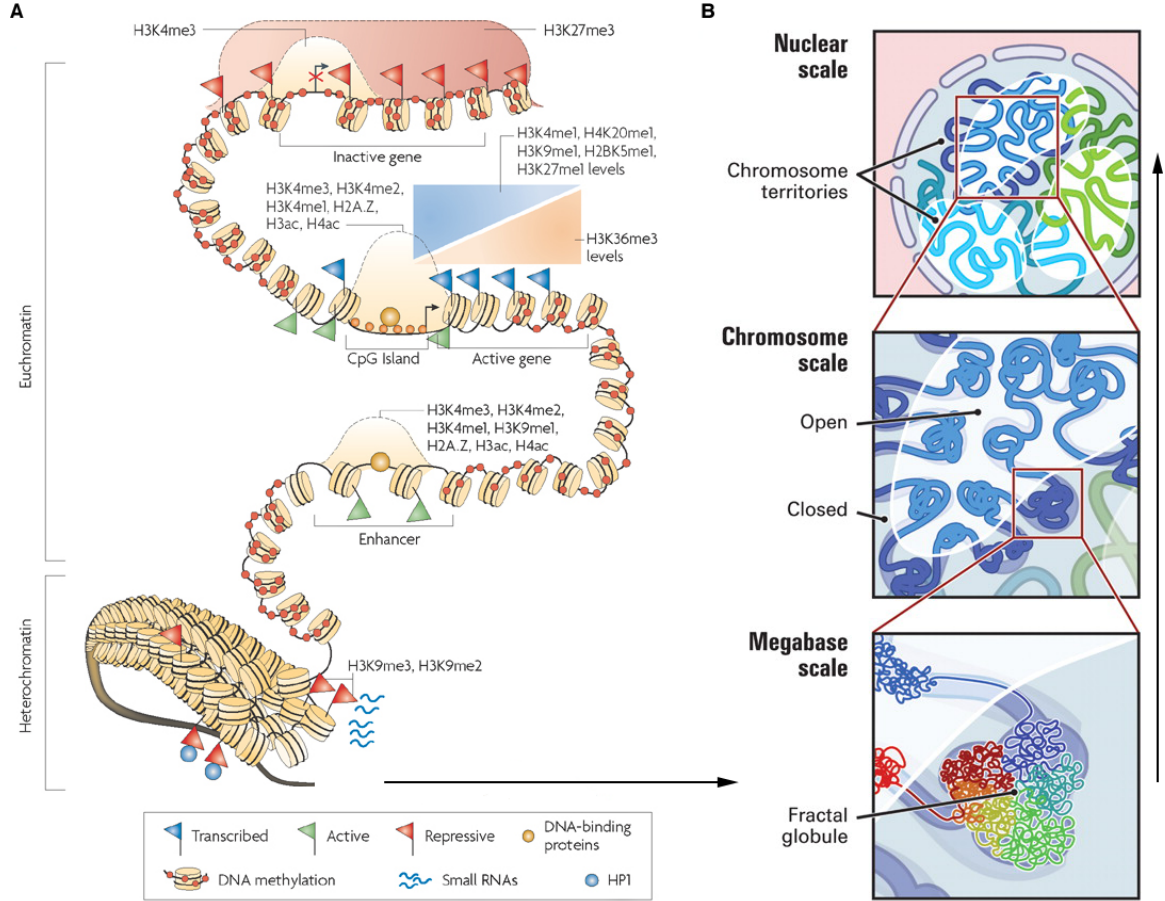


Figure 1.2: Multiple levels of chromatin structure. (A) Features such as DNA methylation (red spheres), nucleosome occupancy and compaction (cylinders), histone variants and modifications (flags), DNA-binding proteins (orange spheres), co-regulators (e.g. HP1, blue spheres) and small RNAs (blue) contribute to chromatin state and are associated with regional differences in transcriptional activity. Through-space interactions, represented by the spatial proximity of the enhancer with the juxtaposed promoter, enable regulation of target genes by distal *cis*-regulatory elements. (B) Evidence from microscopy and 3C-based assays suggest that chromatin is locally folded and hierarchically organised into TADs (bottom) and “open” or “closed” compartments (middle; see Section 1.1.2.4). Subnuclear positioning and the territorial organisation of chromatin are also associated with differences in transcriptional states (see Section 1.1.2.5). Figure adapted from Schones et al.^[21] and Lieberman-Aiden et al.^[22]

sensitive TFs, including “CXXC” motif-containing DNA-binding proteins such as CFP1 and MLL1, which bind unmethylated CGIs exclusively. Importantly, the direction of a causal relationship linking DNA hypomethylation, CFP1 binding and a transcriptionally permissive chromatin state has been established by experiments using an artificial promoterless CGI^[28]. DNA methylation can also exclude CTCF and other factors^[29] from their binding sites thereby influencing gene expression in more complex ways (see Section 1.2). Conversely, “readers” of methylation marks include MBD1-4, KAISO and MECP2, which bind methylated DNA exclusively and facilitate transcriptional repression. The latter acts through its relationship with a histone deacetylase (HDAC)-containing co-repressor complex^[30].

DNA methylation status, and its associated effect on gene expression, can be faithfully transmitted to daughter cells through cell division^[31]. In this context, DNA methylation is an example of an “epigenetic” trait, defined as a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence^{[?] 1}. Within a multi-cellular organism, the control and narrowing of possible expression states by epigenetic mechanisms, such as DNA methylation, enables successive generations of cells to proceed from a pluripotent state towards one of a number of possible terminally differentiated states. Thus epigenetic effects are central to the process of normal development^[30]. However epigenetic traits also have the potential to be transmitted across generational boundaries and a subset of pre-existing chromatin marks need to be erased and reset genome-wide to ensure totipotency at the start of each new generation. Correspondingly, there are two waves of genome-wide DNA demethylation that are essential for the establishment of new methylation landscapes during normal mammalian development: the first occurs during male and female germ cell development and the second wave occurs between fertilisation and implantation^[32]. However, there are a number of germline differentially methylated regions (gDMRs) that retain specific maternal or paternal methylation patterns throughout development. This phenomenon, referred to as “imprinting”, results in allele-specific gene expression in a parent-of-origin fashion i.e. one parental allele is predominantly expressed in the progeny^[32]. DNA methylation is therefore important not only for the transmission of expression states between cells of the same developing organism, but also regulates the trans-generational inheritance of these states.

Epigenetic changes in animals can also occur in response to environmental signals

such as diet and stochastically with ageing, where it is thought to contribute to late-onset diseases. For example, the mis-regulation of the epigenome is a hallmark of oncogenesis, with widespread promoter hypermethylation and associated silencing of important tumour-suppressor genes in cancer cells^[33].

1.1.2.2 Nucleosomes, histone variants and modifications

Rather than existing naked in the nucleus, eukaryotic genomes are normally clothed in histone proteins, which serve to spool approximately 147 bp stretches of DNA into repeating units called nucleosomes (Figure 1.2A). At the core of this fundamental particle is an octamer consisting of two copies of each histone (H2A, H2B, H3 and H4), whose amino acid tails protrude from the resulting DNA-protein complex. The exchange of these histones and the post-translational modification of their tails enables signals to be overlaid on the “raw” DNA sequence, which include the biological analogs of bookmarking, highlighting, annotation and strikeout functionality in traditional text. Similarly to CpG methylation signals, these histone signals (or marks) are read and interpreted by chromatin modifying enzymes and elements of the transcriptional machinery. However, in contrast to DNA methylation, the myriad histone modifications and variants are less directly linked to the nucleotide sequence, can be combinatorially deployed and seem to show a high level of dynamism and redundancy in their use in the cell. The resulting associations are diverse including the usual epigenetic suspects such as the reinforcement of established patterns of expression and the stable repression of genes whose activation is inappropriate or detrimental to cell state, but also encompass many other aspects of transcriptional regulatory fine control. The extension of corresponding locus-specific techniques to genome-wide assays and subsequent integration with other data such as gene expression, TF binding and genomic sequence annotation are helping to provide a more unified view of this primary structure of chromatin.

High resolution whole genome nucleosome occupancy maps, as assayed by MNase-seq (see Section 1.5), have shown a surprising amount of non-random nucleosome positioning when aggregated over cell populations. Most nucleosome positions in the human genome are more stably positioned than would be expected by random chance and a substantial proportion (8.7%) have highly consistent placement^[34]. The presence of nucleosomes can occlude TF binding sites as evidenced by the typically low proportion of possible binding locations that are occupied in vivo. The result is that TF

localisation tends to be biased towards nucleosome-free regions (NFRs), which often occur immediately upstream of regulated genes^[35] (Figure 1.2A). Sequence characteristics in part determine these patterns of nucleosome occupancy, but the binding of TFs and the action of ATP-dependent chromatin remodelling enzymes such as those of the SWI/SNF family^[36] can act to override these preferences. In particular, the binding of CTCF^[37] and NRSF/REST^[38] have been shown to result in highly regular spacing of surrounding nucleosomes and similar patterns of phasing have been observed near TSSs^[39]. The association between chromatin “openness” and regulatory activity is the rationale behind genome-wide approaches that include DNase-seq^[40], FAIRE-seq^[41] and Sono-seq^[42], which directly assay chromatin accessibility (see Section 1.5). Importantly, these methods do not distinguish between the many possible reasons for this accessibility, but instead provide a window into the functionally engaged portion of the genome. Indeed high density mapping of DNase 1 cleavages can help to identify the specific transcription factors bound from the analysis of accessible sequence motifs^[43,44].

There is a huge and growing catalog of identified histone marks comprising at least eleven types of post-translational modification, including methylation, acetylation, ubiquitylation and formylation, which occur at over 60 different amino acid residues^[45]. Although there is correlative evidence linking many of these marks with particular genomic features and activities, their specific functions in the context of transcriptional regulation remain largely uncharacterised. Nevertheless, maps of chromatin modifications have been successfully used as a proxy for transcriptional state, as well as in the hunt for novel genes and regulatory elements^[46,47]. For instance, peaks of histone H3 lysine 4 trimethylation (H3K4me3) are associated with the TSSs of actively transcribed genes and this mark tends to occur at CGI-containing promoters, or high CpG content promoters (HCPs), by default^[48]. The link between H3K4me3 and “open” chromatin is supported by corresponding hypersensitivity to DNase 1 digestion and elevated levels of histone acetylation, both of which are generally associated with increased chromatin accessibility. HCPs are also characterised by RNAP2 occupancy at the TSS, although the absence of transcription in many cases indicates that these types of promoters are subject to additional layers of transcriptional regulation beyond initiation.

Although the relationship between HCPs and H3K4me3 could be explained by the binding of CFP1 and action of associated H3K4 methyltransferase complexes at unmethylated CGIs (as mentioned above), there is evidence that the act of transcription

initiation itself may also contribute to the open chromatin status and deposition of this mark^[49]. Similarly to this positive feedback loop between transcription and chromatin state, the interdependence between TFs and chromatin state has been suggested to be important in the maintenance of accessible (euchromatic) regions of the genome and the spreading of repressive (heterochromatic) marks^[50]. Coupled with cross-talk between the large number of different histone modifications^[50], this model of chromatin as a platform for the integration of signals from many TFs, may provide robustness to established expression states and help to facilitate complex gene regulation^[51] (Figure 1.3).

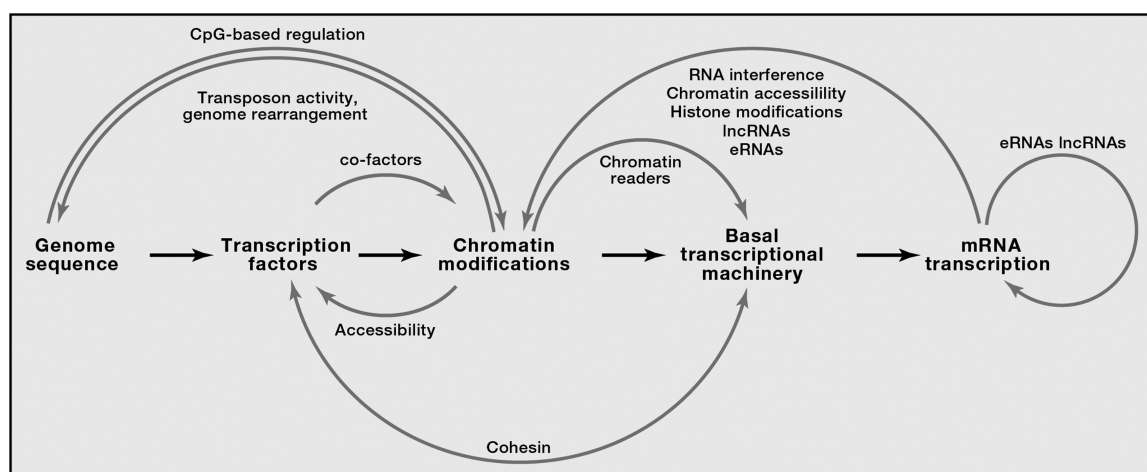


Figure 1.3: The relationship between genomic sequence, chromatin state and transcriptional activity. Rather than a simple direct relationship between genome sequence and gene expression (linear arrows), multiple factors influencing transcription provide additional layers of complexity (curved arrows). Chromatin state represents a “hub” or platform for the integration of up- and downstream regulatory signals adding robustness and responsiveness to the system. See Section 1.3 for the role of cohesin in transcriptional regulation. Figure adapted from Merckenschlager et al.^[52]

HCP genes requiring tissue-specific regulation, for example those encoding developmental transcription factors, can be inactivated by Polycomb repressive complexes (PRC1 and PRC2) whose presence is characterised by the deposition of H3K27me3^[53]. The simultaneous (bivalent) marking of HCPs by H3K4me3 and H3K27me3 is thought

to represent a poised transcriptional state, as the corresponding genes are inactive in pluripotent cells but become either rapidly activated or stably repressed depending on the subsequent developmental stage^[54]. More stable gene repression, or silencing, is associated with large regions of H3K9me2 and H3K9me3, which recruit heterochromatic protein 1 (HP1)^[55].

In contrast to default-on HCP genes, which are enriched for house-keeping functions, genes without promoter CGIs (low CpG content promoters, LCPs) require the binding of specialised TFs for RNAP2 recruitment, full activation and associated H3K4me3 marking. This suggests that substantial LCP regulation occurs at the level of transcription initiation^[53]. Furthermore, these genes tend to have tissue-specific functions relevant to terminally differentiated cells, and their poised status in undifferentiated cells is accompanied by H3K4me2 but not H3K4me3 at the promoter. Although stable inactivation of LCP genes is not associated with particular histone modifications, they may be DNA methylated^[56].

Other chromatin features associated with active promoters are H3K4me1 and the histone variant H2A.Z, which conveys protection from DNA methylation^[57]. Actively transcribed gene bodies possess chromatin modifications distinct from those at active promoters including H3K36me3 and H3K79me2, which correspond to the abundance of elongating RNAP2 and accompanying histone methyltransferases^[49]. Within transcripts, expressed exons are enriched for H3K36me3, which may be at least partly explained by increased exonic nucleosome occupancy. Nevertheless, these findings suggest a link between chromatin structure and co-transcriptional splicing, where spliceosome activity and recruitment is either aided by nucleosome barrier-induced RNAP2 slowing or the level of H3K36me3 itself^[53].

The task of comprehensively identifying enhancers genome-wide is complicated by their cell-type-specific activity and the diversity of TFs which bind them. However, by exploiting similarities in their chromatin context, promoter distal loci occupied by the active enhancer-associated protein EP300 were used to determine a more general histone modification signature: H3K4me1 in the absence of H3K4me3^[58]. Other histone marks that occur at enhancers include H2BK5me1, H3K4me2, H3K9me1, H3K27me1, H3K36me1^[46] and H3K27ac. The latter can be deposited by EP300 and CREBBP and also occurs frequently at active mammalian promoters^[59]. This and other similarities

in the modifications that occur at enhancers and promoters may be a result of their proximity in 3D space and “sharing” of physically associated chromatin modifying enzymes at either genomic region^[53].

Insulators are regulatory elements that interfere with interactions between distinct genomic regions. A classical example is the blocking of enhancers from acting on nearby promoters, but there is evidence that insulator function extends to more general roles in the formation of barriers to maintain separate chromatin and transcriptional states^[60]. Insulators as defined by the presence of CTCF (see Section 1.2) are not consistently associated with a particular chromatin signature apart from a modest enrichment of the histone variant H2A.Z^[46]. In addition to these punctate functional elements, distinct histone modifications are also associated with megabase sized replication time zones^[61] and large homogenous regions of active or repressed chromatin thought to be involved in the formation of subnuclear structures such as Polycomb bodies^[62], transcription factories^[63] and lamina associated domains (see Section 1.1.2.5).

Although a “histone code” analogous to that of the genetic code remains elusive^[64], unsupervised machine learning approaches have enabled the vast amounts of epigenomic data generated by projects such as ENCODE to be summarised in an easily-interpretable manner^[65]. This type of integrative analysis and annotation of chromatin state can provide a starting point for the characterisation of non-coding DNA and the interpretation of individual genome sequences.

1.1.2.3 Higher-order chromatin and looping structures

Beyond insight from assays of nucleosome occupancy, chromatin accessibility and modifications, which are agnostic to the 3D structure of the genome, comparatively little is understood about higher-order chromatin organisation. The “beads on a string” configuration, consisting of linear arrays of nucleosomes, is subject to further compaction by linker histone H1 binding to spacer DNA between successive nucleosomes. This is thought to form condensed solenoidal fibres or other repetitive structures with a diameter of ≈ 30 nm which are associated with heterochromatin (Figure 1.2A). However, neither the detailed organisation of these structures nor the cell types or stages where they exist in vivo are clear^[66].

Euchromatin, which is associated with transcriptional activity, needs to be organised in such a way as to ensure regulated gene expression. Batteries of regulatory elements – spread in and around transcriptional units – are also often located at great distances from their target genes. This may be a particular feature of mammalian genomes which have a relatively low gene density and therefore a high proportion of non-coding (potentially) regulatory DNA per gene^[67]. Locus-specific studies of genes encoding developmental TFs suggest the importance of such wide-ranging regulatory landscapes comprising potent, but distant, tissue-specific enhancers^[68]. In chromatin looping models, these elements exert their positive effects on target promoters by forming direct, through-space interactions with components of the transcriptional machinery. Hypothesised mechanisms that could initiate such interactions include linking by large protein complexes, tracking or scanning of enhancer-bound factors (e.g. RNAP2), as well as more passive free diffusion scenarios^[69]. Alternative models of enhancer function include “placeholder” of pioneer factors for TFs expressed at a later developmental time-point, spreading of associated chromatin states and the indirect action of enhancer-associated non-coding transcripts (eRNAs). However, evidence from fluorescent *in situ* hybridisation (FISH) and, more recently, chromosome conformation capture (3C) techniques, which involve formaldehyde mediated chromatin fixation (cross-linking) and sequencing to identify pairs of regions in close proximity (see Section 1.5.4), provide support for enhancer-promoter looping as a widespread phenomenon^[69].

Chromatin looping in the context of enhancer-promoter interactions was first studied at the *β-Globin* gene cluster. Initial 3C experiments showed that the locus control region (LCR) forms long-range (≈ 50 kb) contacts with downstream promoters of either embryonic or adult globin genes, which are expressed according to their corresponding developmental stages in erythrocytes (Figure 1.4A). In addition to these dynamic multi-gene interactions, the LCR also forms more stable contacts with up- and downstream DNase 1 hypersensitive sites (DHSs) that have been proposed to contribute to the creation of a chromatin micro-environment, or hub, conducive to globin gene transcription^[70]. Another well-studied gene subject to long-range control is *Sonic hedgehog* (*Shh*), which encodes a signalling protein with a role in developmental patterning. The complex spatial expression patterns of *Shh* in various embryonic structures is regulated by an array of enhancers extending beyond an upstream gene desert. For example, transient promoter contact of the ZRS element, separated by a linear distance of almost one megabase and spanning an intervening gene, is required for *Shh* expression in

the developing limb bud^[71] (Figure 1.4B). In addition to these one-to-many (β -Globin) and many-to-one (*Shh*) enhancer looping configurations, the collinearly expressed *HoxD* gene cluster is regulated by long-range interactions with multiple activating elements in a partially redundant manner^[72] (many-to-many; Figure 1.4C).

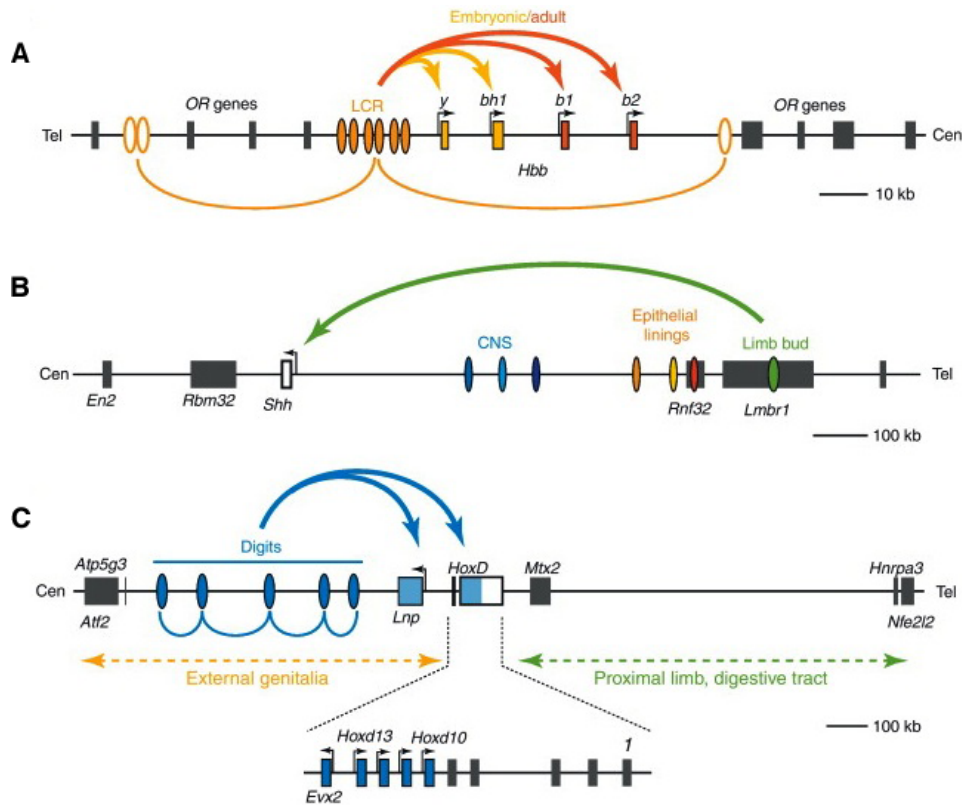


Figure 1.4: Long-range regulation by chromatin looping at selected developmental loci. (A) Schematic of the β -Globin gene cluster showing developmental stage-specific contacts (curved arrows) involving the LCR (orange ovals) and promoters of either embryonic or adult globin gene promoters. (B) Long-range looping interaction active in cells of the developing limb bud involving the *Shh* gene and the distal ZRS enhancer element (green oval) located in an intron of the *Lmbr1* gene. Other tissue-specific elements (blue and orange ovals) are indicated. (C) Multiple enhancer elements (blue ovals) interact with each other to coordinate the expression of the *HoxD* gene cluster. Figure adapted from Montavon et al.^[73]

The action of specific proteins bound at loop endpoints is likely needed to boost interaction frequencies above that expected under situations of random diffusion^[74].

Many different types of factors have been suggested to mediate loop formation in the context of both positive and negative regulation, including activator TFs, Polycomb complexes, chromatin remodellers, RNAP2, as well as insulator and architectural proteins^[75]. However, in the case of LCR contacts with β -*Globin* genes, it has been shown that the role of the RNAP2 transcriptional machinery itself is not required to maintain long-range interactions^[76]. Furthermore, although ZRS deletion interferes with *Shh* transcription in budding limbs, the resulting 3D conformation of the locus is unchanged^[77]. Similarly, long-range interactions with enhancer elements controlling digit expression in the *HoxD* cluster are also present in the developing forebrain despite the total absence of transcription^[72]. These findings suggest that enhancers, and other regulatory elements, may often take advantage of pre-existing chromatin structures rather than relying on the act of transcription or recruited tissue-specific factors to create these long-distance contacts *de novo*.

Chromatin contacts have also been detected between insulator sites, where the directly bound factor CTCF (see Section 1.2) and associated cohesin complex (see Section 1.3) have been shown to co-depend on each other to facilitate loop formation. CTCF-based loops deployed in a constitutive or condition-specific manner, as they are near the well-studied mono-allelically expressed *Igf2* gene, influence transcription by modulating local chromatin topology. There are many different conceivable and demonstrated mechanisms whereby such long-range interactions can enable complex gene regulation. Indeed the thousands of binding sites where CTCF and cohesin co-occur and potentially form chromatin loops have led to suggestions that these proteins may help to set up the global 3D structure of the genome^[78,79].

On a more local level, chromatin loops linking the 5' and 3' ends of individual genes seem to favour messenger RNA (mRNA) synthesis by RNAP2 over corresponding non-coding RNAs (ncRNAs) at bidirectionally transcribed promoters. These gene loops are also thought to enable efficient recycling of components of the transcriptional machinery and associated factors, thereby contributing to fully productive gene expression^[80].

Apart from contacting themselves in this way, genes have been shown to reach out and touch other co-regulated genes in the nucleus, as has been demonstrated for globin genes^[81], estrogen receptor alpha (ER) induced genes^[82] and the promoters of genes actively transcribed by RNAP2 in general^[83]. Furthermore, instead of being restricted to

long-distance communication with target promoters, there is evidence that regulatory elements can also interact with each other to coordinate gene activation^[84]. The full complement of possible connections involving genes and regulatory elements is therefore represented in the nucleus. This suggests that gene regulation in higher eukaryotes involves complex 3D networks of chromatin interactions^[85], as is supported by a recent targeted high-resolution study of interactions involving human gene promoters^[86].

1.1.2.4 Chromatin compartmentalisation

From the imaging of live cells, it is clear that chromosome positions, although broadly consistent over the lifetime of a particular cell, are highly variable between cells^[85]. Yet, in the face of this large-scale structural heterogeneity, individual cells in the population still manage to maintain their expression states, which are dependent on 3D regulatory interactions, as discussed above. This apparent paradox may be explained by a domain-like organisation of chromatin, where sub-regions of the genome are hierarchically compartmentalised^[85]. In this model, chromatin is reproducibly structured on a local, sub-domain level, whereas the relative distribution of these domains within the nucleus is more variable between cells. Indeed, there is currently support for two such genome-wide levels of organisation: chromatin compartments and topologically associating domains (TADs; Figure 1.5).

The first computational analysis of results from Hi-C – a high-throughput genome-wide version of 3C (see Section 1.5.4) – in human cell lines revealed that interphase chromatin can be classified into megabase-sized architectural compartments, denoted as A and B (Figure 1.5). A compartments are associated with marks of open chromatin, gene-dense regions and active transcription, whereas B compartments tend to be “closed”, gene-poor and less transcriptionally active^[22]. The fact that loci within the same compartment show a preference for similar interacting partners, probably reflects their spatial clustering in the nucleus. Active genes within compartment A are thought to co-associate at transcription hotspots (or factories) in the nuclear interior, whereas repressed genes congregate at distinct regions such as the peripherally located nuclear lamina^[85] (see Section 1.1.2.5). Consistent with ties to transcriptional activity, compartments exhibit cell-type-specific changes, where large regions containing genes expressed in a lineage-specific fashion switch compartment assignment during differentiation^[22,88,89].

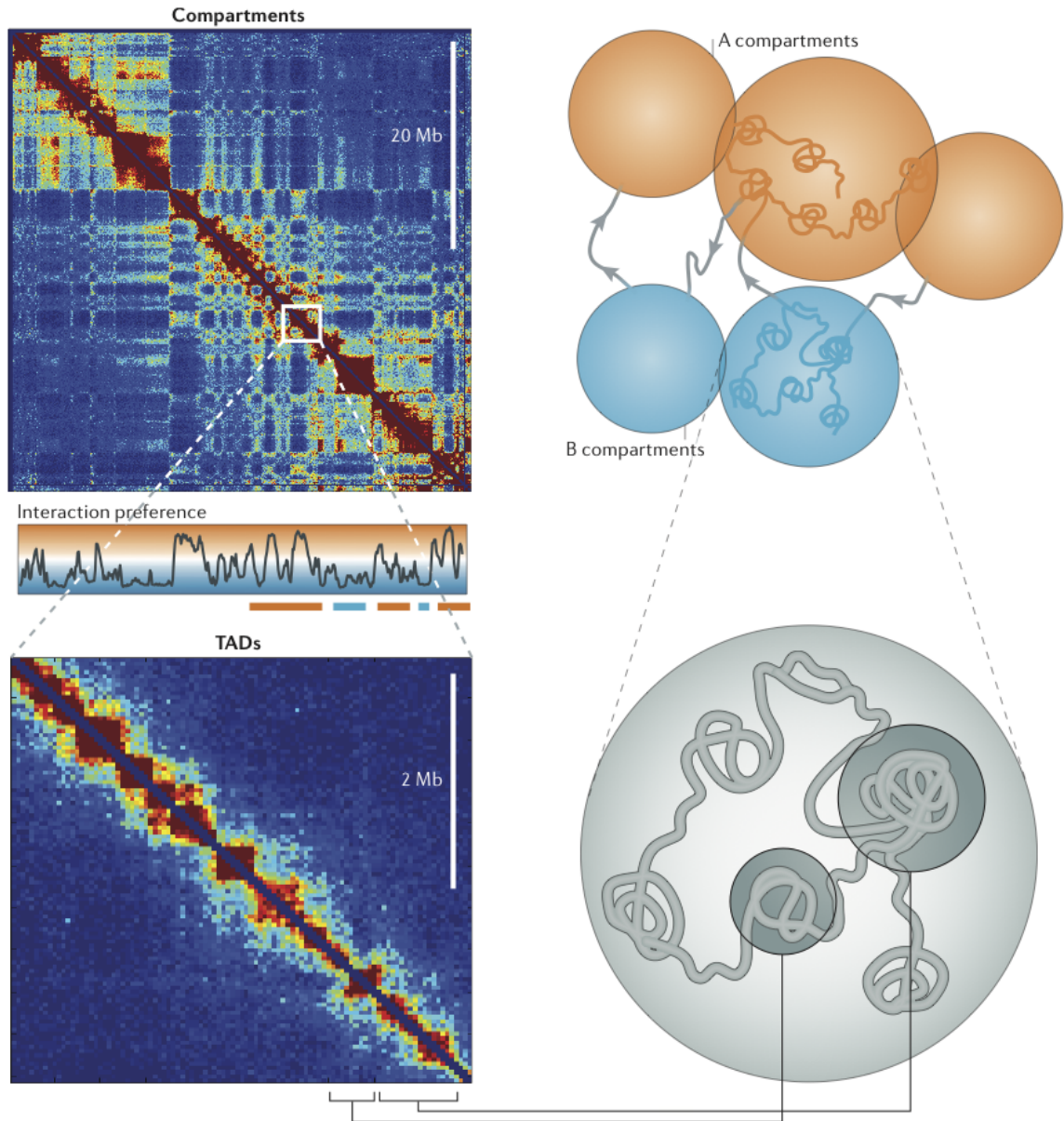


Figure 1.5: Higher-order organisation of chromatin into compartments and TADs. A and B compartments with distinct interaction profiles are indicated above and comprise TADs, which exist at a lower level in the hierarchy of chromatin structure (below). Corresponding cartoon models of the 3D chromatin structure represented by the Hi-C maps (left) are shown. Figure adapted from Dekker et al. [87]

Recent genome-wide (Hi-C) data in *Drosophila*^[90], mouse and human cells^[91], as well as high-resolution 5C (see Section 1.5.4) interaction maps of the mouse X-inactivation centre^[92], have identified discrete locally-constrained blocks of chromatin, or TADs. Loci within the same TAD interact more frequently with themselves than with other loci (Figure 1.5). Although TADs vary in size, both visual inspection and computational analysis show that they tend to comprise hundreds of kilobases of DNA (*median* = 880 kb in mice) and cover $\approx 90\%$ of the genome. In contrast to compartments, TADs seem to be stable across cell types and highly conserved between different species, suggesting that they are fundamental building blocks of higher-order chromatin structure in mammals^[91]. Furthermore, TADs contain coordinately expressed genes^[92] and enhancer-promoter interactions occur more frequently within the same TAD than across domain boundaries^[91]. These observations point towards a role for TADs in defining local gene regulatory neighbourhoods and constraining the universe of possible chromatin interactions that a particular locus can participate in. Chromatin in live cell nuclei is locally dynamic with substantial diffusional mobility up to distances of $0.25\mu\text{m}$ ^[93,94], which corresponds to a megabase or so of DNA. Therefore, within TADs, it is expected that regulatory elements would have ample opportunity to encounter their targets by chance alone.

The unbiased all-versus-all nature of the genome-wide Hi-C assay provides the opportunity to study fundamental biophysical properties of chromatin. Using results from simulations of different polymer models, the “fractal globule” was found to be the most consistent with empirical data at the megabase scale^[22]. In this model, the extremely long and flexible polymer is condensed and locally compact – rather than being spread – enabling knot-free packaging and easy unfolding (or refolding) of chromosomal sub-regions. 3C and associated techniques to assay chromatin conformation also allow for the bottom-up construction of 3D models representing the structure of large chromosomal regions^[95] or even whole genomes^[96,97,98]. By modelling the global structure of the genome and capturing cell-to-cell variability with conformational ensembles, it is expected that these physics-based approaches will enable the discovery of novel organisational principles^[99].

1.1.2.5 Chromosome territories, nuclear bodies and subnuclear position

It is becoming increasingly clear that position within the nucleus is an important aspect of gene regulation. Imaging studies, and more recently Hi-C, have confirmed that individual chromosomes tend to occupy specific domains in the nucleus, termed chromosome territories (CT)^[100]. Apart from their largely mutually exclusive volumes, the radial arrangement of chromosomes also correlates with gene density. For example, chromosomes with high gene density (e.g. human chr19) consistently occupy interior positions, whereas gene poor chromosomes (e.g. human chr18) occur more frequently near the nuclear periphery^[101]. These observations, based on chromosome positions, point towards structural and spatial mechanisms of gene expression regulation that can operate at the chromosome-wide level. A particularly dramatic example of this is the global compaction and almost complete silencing of one X chromosome in female mammal cells. The action of the Xist long non-coding RNA (lncRNA), which coats the inactivated X chromosome, is central to the formation of this discrete and repressive CT known as the Barr body^[102].

Genome-wide mapping of nuclear lamina-associated domains (LADs) in human fibroblasts, revealed that these regions correlate with low levels of gene expression and repressive chromatin marks^[103]. Although not always sufficient, artificial recruitment to the nuclear lamina can facilitate transcriptional repression^[104] and dynamic (dis-)associations with this compartment have been shown to accompany differentiation events^[105]. Recently, specific lamina-associating sequences (LASs) enriched for GAGA motifs have been implicated in the targeting of genomic loci to the nuclear periphery^[106]. In contrast to the repressive effects at the lamina, contact with components of nuclear pore complexes (NPCs) is associated with euchromatin and active transcription^[107,108]. However, many of these nucleoporin-associated regions (NARs) are bound by diffusible NPC proteins in the nucleoplasm and specific binding to NPC-tethered forms occurs only for a small set of inactive genes in *Drosophila*^[109]. Therefore, it remains to be determined whether the nuclear pore indeed represents a chromatin compartment that is functionally distinct from the nuclear lamina^[110].

Transcription factories are nuclear foci where actively transcribed genes come together at sites of locally increased regulatory protein and RNAP2 concentration. Al-

though the dynamic colocalisation of transcribed genes with clusters of active RNAP2 molecules has been observed in vivo^[111], a model where the latter are immobilised to a nuclear substructure is still controversial^[63]. High levels of expression – facilitated by multiple simultaneously transcribing RNAP2 molecules – and processes such as bidirectional transcription are examples of some of the challenges faced by the idea that genes are recruited to, and “reeled” through, fixed transcription factories. Other issues such as their variable number between cells, compositional details and mechanism of formation still need to be resolved^[63]. On the other hand, more is known about transcription by RNAP1 and RNAP3 at the nucleolus, which is formed around clustered arrays of ribosomal genes from multiple chromosomes. By genome-wide mapping of nucleolar-associated domains (NADs), it was found that silenced RNAP2-dependent genes also occur at nucleoli^[112]. Therefore, like LADs, these regions can serve as a scaffold for the sequestration of heterochromatin.

1.2 CCCTC-binding factor, CTCF

CCCTC binding factor (CTCF) is an essential^[113,114,115] and widely expressed nuclear protein with a DNA-binding domain that is highly conserved from fly to human^[116]. CTCF uses combinations of its 11 zinc fingers to recognise an information-rich core motif common to most binding events, a newly discovered but less frequent additional motif^[117] and possibly other protein binding partners^[118]. Originally identified as a transcriptional regulator of the *c-Myc* oncogene^[119,120,121], its known repertoire has expanded to include a diverse array of well-characterised gene regulatory functions from transcriptional activation to repression and silencing^[122,123,124].

CTCF’s apparently unique ability to directly bind vertebrate insulators and role in ensuring the correct architecture and expression at imprinted genes, in particular *Igf2/H19*, have been extensively studied^[125,126,127,128]. Allelic differences in the expression of the mammalian *Igf2* and *H19* genes are the result of a nearby gDMR, the imprinting control region (ICR), harbouring multiple clustered CTCF binding sites. 3C experiments indicate that DNA methylation-free maternal ICR recruitment of CTCF blocks the looping of a distant enhancer to the *Igf2* promoter and its subsequent expression. However, paternal ICR methylation prevents CTCF binding and associated insulation effects, thereby allowing these enhancer-promoter interactions to occur, re-

sulting in activated expression of the *Igf2* gene^[129,130,131,132].

Despite demonstrated roles in gene regulation and development at isolated loci, recent comparisons of genome-wide CTCF occupancy across multiple tissues and species have revealed high levels of cell-type invariance^[133,134] and conservation^[20,117] in its patterns of binding. Differential CTCF binding in cancer cell lines correlates with widespread DNA methylation differences and therefore may be a particular feature of tumour cells^[135]. Other features of CTCF binding include a relatively low level of enrichment at promoters compared to other transcription factors and, consistent with insulation, its frequent presence at the boundaries between active and inactive chromatin domains^[134]. Furthermore, CTCF binding sites demarcate the borders of regions localised to the nuclear periphery (LADs)^[103] and are enriched at TAD boundaries^[91], both of which represent important features of higher-order chromatin structure. These findings, together with results showing that CTCF binding events correlate with intra- and inter-chromosomal interactions in the human genome as measured by Hi-C^[22,136], point towards a global genome-wide organisational role^[78]. However, rather than acting in isolation, CTCF has been shown to depend on cohesin for its looping function (Figure 1.6C, see Section 1.3).

Given CTCF’s diverse roles, it is not surprising that the protein is linked to a number of diseases including Beckwith-Wiedemann syndrome, characterised by loss of CTCF binding sites in the ICR^[138], neurological disorders^[139], as well as cancer, leading to its classification as a tumour suppressor^[140]. Host CTCF has also been shown to bind and regulate gene expression in a number of viral genomes, including Epstein-barr virus^[141] and Kaposi sarcoma-associated herpes virus^[142].

ChIP-seq analyses have highlighted the role of transposable elements in the evolution of high numbers of CTCF binding sites closely matching the long motif consensus in mammalian genomes. Specifically, B2 repeats – members of the short-autonomous interspersed elements (SINE) family – possessing embedded CTCF binding sites have undergone recent massive lineage-specific expansions in rodents^[117]. As CTCF binding could conceivably prevent methylation-dependent silencing of transposons^[135,143], such “passenger” CTCF motifs may have played a role in their retention as active elements. These results add weight to previous evidence suggesting the importance of retroelement-driven remodelling of regulatory networks in general^[144].

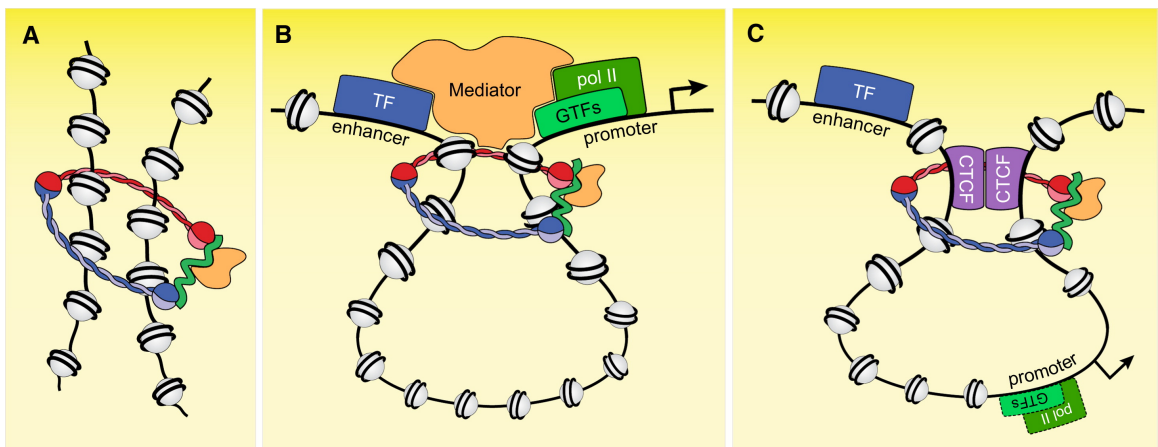


Figure 1.6: Models for cohesin function in sister chromatid cohesion and loop formation, with and without CTCF. (A) The embrace model whereby cohesin is thought to hold sister chromatids together. (B) Model of enhancer-based activation of target genes where cohesin stabilises loop formation and facilitates bridging of the mediator complex between distal TFs and promoter-proximal RNAP2. (C) Model of CTCF-based insulator activity wherein cohesin facilitates looping between distinct CTCF binding events thereby disrupting enhancer-promoter interactions. Figure adapted from Cuylen et al.^[137]

1.3 The cohesin protein complex

The evolutionarily conserved cohesin protein complex plays an essential role in chromosome cohesion during mitosis and meiosis^[145]. The core of the complex is a heterodimer of structural maintenance of chromosomes (SMC) subunits (SMC1A and SMC3) connected by a third subunit RAD21 (MCD1/SCC1 in budding yeast), forming an unusual tripartite ring-like structure^[146,147]. RAD21 is bound to a fourth member (either STAG1, STAG2 or STAG3) and it has been proposed that the complex mediates cohesion by embracing sister chromatids^[148] (Figure 1.6A). Several other proteins are associated with cohesin including NIPBL (Nipped-B in fly, Scc2 in budding yeast), which is required for loading of cohesin onto chromatin^[149].

Although essential for sister chromatid cohesion, Nipped-B was first identified in *Drosophila* as a result of its function in gene regulation, where it was suggested to facilitate enhancer-promoter interactions^[150]. Similarly, mutations in core components of the cohesin complex can affect gene expression, and have been linked to developmental defects (cohesinopathies) in a number of different species^[151,152,153,154,155,156,157]. Beyond its presence on sister chromatids during cell division, cohesin is also expressed in post-mitotic cells and is loaded onto unreplicated chromosomes in telophase^[155,158,159]. Together, these findings point towards an important noncanonical role of cohesin in regulating gene expression.

More recently, genome-wide maps of cohesin binding in mammalian cells reveal that the complex functionally associates with the majority of CTCF sites. Cohesin, which has no known DNA-binding domain, is recruited to chromatin via CTCF C-terminal tail interactions with the STAG2 subunit where it assists in performing its function as an enhancer-blocking insulator^[142,159,160,161,162] (Figure 1.6C). Direct evidence from cohesin knockdown experiments implicates the complex in facilitating long-range interactions between CTCF sites at the *H19/Igf2* locus, as well as at others including the IFNG, apolipoprotein and hemoglobin, beta genes^[163,164,165,166].

ChIP-seq experiments in MCF-7 and HepG2 human cancer cells show that cohesin also binds to thousands of sites in a CTCF-independent manner. In stark contrast to

relatively invariant CTCF sites, these cohesin-non-CTCF (CNC) binding events differ dramatically between cell-types. CNC sites colocalise with tissue-specific TFs such as ER in MCF-7 cells, and contribute to global gene expression. Cohesin is also highly enriched at ER-bound regions that participate in chromatin looping interactions as assayed by ChIA-PET^[167,168] (see Section 1.5.4). A study in mouse embryonic stem (ES) cells, which highlighted subunits of both the cohesin and mediator complexes as key contributors to ES cell state, found analogous patterns of CNC binding and co-occupancy with pluripotency regulators, including POU5F1 (also known as OCT4), at interacting promoter and enhancer regions^[169] (Figure 1.6B). Finally, results from 3C experiments show that cohesin is required for similar promoter-enhancer interactions within the T-cell receptor alpha (Tcra) locus^[170]. Taken together with previous findings, this firmly establishes a role for the complex in the widespread mediation of long-range transcriptional control.

Therefore, consistent with the theme of modularity and reuse in biology, it seems likely that evolution has re-purposed a protein complex originally involved in a universal eukaryotic function in *trans* (mitosis), for a related topological function in *cis* (intra-chromosomal looping) that is central to cell-type-specific regulation in multi-cellular organisms^[171]. However, rather than duplication and diversification as is the case in the neofunctionalisation of paralogs^[172], the same cohesin complex fulfils these distinct roles, which predominate at different stages of the cell cycle.

1.4 Brother of Regulator of Imprinted Sites, CTCFL

Initially discovered in a screen of rodent testes extracts, CTCF-like (CTCFL) or Brother of Regulator of Imprinted Sites (BORIS) was named according to its structural similarity and inferred paralogy to the well-characterised CTCF protein^[173].

The *Ctcf* gene is thought to have arisen via a duplication event during vertebrate evolution at least 210 mya^[174] and although the 11 zinc finger DNA-binding domains in the respective proteins are highly similar, their C- and N-terminal domains show no significant similarity in mouse or human^[173], pointing towards differences in their interactions with cofactors. Results from expression profiling in amniotes suggest that

a functional change occurred in CTCFL early in the evolution of therian mammals that resulted in its germ-line-restricted expression pattern^[174].

Although there are conflicting reports as to the specific cell type and time-point of localisation in the testis, the most recent studies show transient presence of CTCFL in nuclei of spermatogonia and preleptotene spermatocytes prior to the onset of meiosis^[175]. Initial observations that the mutually exclusive expression of CTCFL and CTCF in testis is correlated with the re-setting of genome-wide DNA methylation led to a hypothesised role for CTCFL in the developmental reprogramming of the epigenome^[173], but the precise details of this putative function remain elusive. Nevertheless, CTCFL's role as a male germ cell gene regulator is strengthened by research showing that its deficiency affects the expression of a number of testis-specific genes resulting in spermatogenesis defects^[175].

Abnormally high levels of CTCFL have also been detected in a number of human tumour samples and cancer cell lines. Here its expression has been linked to promoter demethylation and de-repression of genes normally exclusively expressed in testis^[176]. Consequently CTCFL is classified as a cancer-testis antigen (CTA) and the protein is currently the focus of anti-tumour vaccine development efforts^[177]. On the other hand, CTCF is generally considered a tumour suppressor gene^[140]. This, together with CTCF and CTCFL's normally mutually exclusive expression patterns and opposite effects on their targets, namely the *Igf2/H19* ICR epigenetic status^[178], *Bag1*^[179] and *hTERT*^[180] genes, as well as the *Ctcf* gene itself^[181], suggests that the paralogs may be antagonistic regulators of their common binding loci.

In contrast to CTCF which exists as a single isoform, the expression of CTCFL seems to provide more opportunities for regulation, involving three alternative promoters capable of producing 23 different splice isoforms and corresponding to a total of 17 distinct protein products in human cells^[177]. CTCFL's cellular position may also be subject to additional layers of control, with reports of its accumulation in either the cytoplasm or localisation near specific nuclear structures, including the centrosome and nucleolus, whereas CTCF is more generally distributed in the nucleoplasm^[176,182].

Whether under normal physiological conditions in the male germline or in cancer cells, the study of CTCFL is complicated by its transient presence, structural com-

plexity and expression variability. However, close structural and functional links to CTCF and cohesin as well as its recent evolution, tissue-restricted expression pattern and prevalence in cancer cells, make it an intriguing subject of mammalian chromatin research.

1.5 High-throughput approaches in functional genomics

A number of key technologies have accelerated the study of genome function. Scaling up from successful attempts to determine viral and organelle DNA sequences starting in the late 70's, machines implementing automated Sanger sequencing – the dominant method for almost two decades – enabled the first cellular genomes to be sequenced, including the finished human genome reference^[183]. Despite providing a fundamental resource for locus-specific analyses and biomedical research in general, whole-genome DNA sequences also facilitated the first assays of genome functioning on a global scale. Polymerase chain reaction (PCR)-based techniques, for example, that quantify the expression of individual genes (qRT-PCR) gave way to high-throughput genome-wide alternatives such as microarrays, which allow the full complement of an organism's known genes to be interrogated simultaneously.

DNA microarrays consist of hundreds of thousands of short synthesised oligonucleotide probes, complementary to known transcript sequences, immobilised on a solid support surface in a two dimensional array. Gene expression values are determined by measuring the intensity (and therefore quantity) of fluorescently labelled sample complementary DNA (cDNA) that hybridises to these probes^[184]. Microarrays with overlapping probes that “tile” the genome (tiling arrays) have also been designed to assay chromatin state and find regulatory elements i.e. features with no clear *a priori* genomic distribution (see Section 1.5.2), but there remain a number of limitations of array-based technologies. Microarrays measure sample abundance indirectly, are subject to saturation effects (limited dynamic range) and issues related to cross-hybridisation to non-exactly matching probe sequences^[185]. Although tiling arrays can be used to discover new, short and low-abundance RNAs, conventional sequencing is usually needed to determine precise transcript structures^[186]. Single-nucleotide poly-

morphism (SNP)-chips can be used to assay population variation at known sites and for diagnostic purposes^[187], but microarrays are otherwise blind to the departure of individual genomes from the reference sequence.

The recent development of high-throughput, or next-generation, sequencing (NGS) technologies has enabled the rapid and relatively cheap sequencing of massive amounts of DNA. In addition to revolutionising whole-genome (re-)sequencing, this has led to a wave of NGS-based techniques, adapting originally locus-specific functional assays to their genome-wide alternatives. Apart from continually falling sequencing costs, other benefits of these “digital” assays (e.g. DNase-seq, ChIP-seq, 4C-seq) over their microarray-based forerunners (DNase-chip, ChIP-chip, 4C-chip) include greater coverage, dynamic range and resolution^[188].

Shankar Balasubramanian and David Klenerman of the University of Cambridge originally developed the particularly successful DNA sequencing technology that was commercialised in the joint founded Solexa Inc.^[189] and subsequently acquired by Illumina Inc. In the next section, I provide a brief description of NGS technologies, focussing on this Illumina/Solexa platform, which was used to generate the data presented in this thesis. The following sections discuss the techniques and downstream data analysis corresponding to NGS-based chromatin state assays, transcriptomic assays and chromatin conformation assays. Common steps in the analysis of NGS data involve quality assessment of raw sequences and alignment to the reference genome or transcriptome to determine their identity. There is a vast and growing list of stand-alone, command-line and web-based computational tools for downstream data processing, where the method of choice depends on the application and biological question being addressed. The R statistical environment and Bioconductor genomics toolbox^[190] also provide a number of useful packages for data manipulation and integration, which were used extensively throughout this thesis.

1.5.1 High-throughput sequencing technologies

Current commercially-available NGS technologies all employ massively parallel approaches which take place on a solid support surface, but differ in a number of respects including template preparation, sequencing and imaging, as well as downstream anal-

ysis^[191]. In the case of Illumina/Solexa, double-stranded sample DNA is sheared into 200-300 bp fragments followed by end-repair and adapter ligation (Figure 1.7A). The single-stranded template DNAs are then attached to random positions on the dense “lawn” of primers covering the surface of a flow cell, which contains eight lanes allowing parallel sequencing of multiple samples (Figure 1.7B). Solid-phase amplification is accomplished by repeated cycles of “bridging” of the immobilised fragments to nearby primers generating double-stranded DNA (Figure 1.7C), followed by denaturing, resulting in millions of molecular clusters each containing identical copies of single-stranded template DNA (Figure 1.7D). Sequencing-by-synthesis then proceeds by a four-colour cyclic reversible termination (CRT) method, where individual labelled nucleotides are incorporated into the growing oligonucleotide chains by polymerases, which are then laser imaged (Figure 1.7D). Fluorescent dye and terminator removal allows the process to be repeated, with each cycle determining the identity of a single template nucleotide per cluster (Figure 1.7E).

Other technologies include Roche/454 and Life/ABI’s SOLiD, which use emulsion PCR followed by pyrosequencing and sequencing-by-ligation respectively, as well as machines by Pacific Biosciences that involve immobilisation of polymerase molecules (instead of templates/primers)^[191]. The latter technology and other single molecule nanopore-based approaches on the horizon promise much longer read lengths, allowing for comprehensive sequencing of repetitive DNA and improvements in the assembly of genomes/transcriptomes.

1.5.2 Chromatin state assays

NGS-based assays of linear chromatin state can be broadly grouped into four categories: (i) DNA methylation, (ii) chromatin accessibility, (iii) nucleosome positioning and (iv) protein-DNA interactions. Bisulphite-based methods to assay DNA methylation status involve the chemical modification of all unmethylated cytosines in the genome into uracils, which appear as thymines after PCR amplification. By sequencing the resulting DNA fragments and comparing them to the reference, the methylation status of individual cytosines can be inferred^[193]. Reduced representation bisulphite sequencing (RRBS-seq) lowers the cost of whole-genome BS-seq by enriching for restriction fragments within a specific size range thereby limiting the assay to a small fraction of the

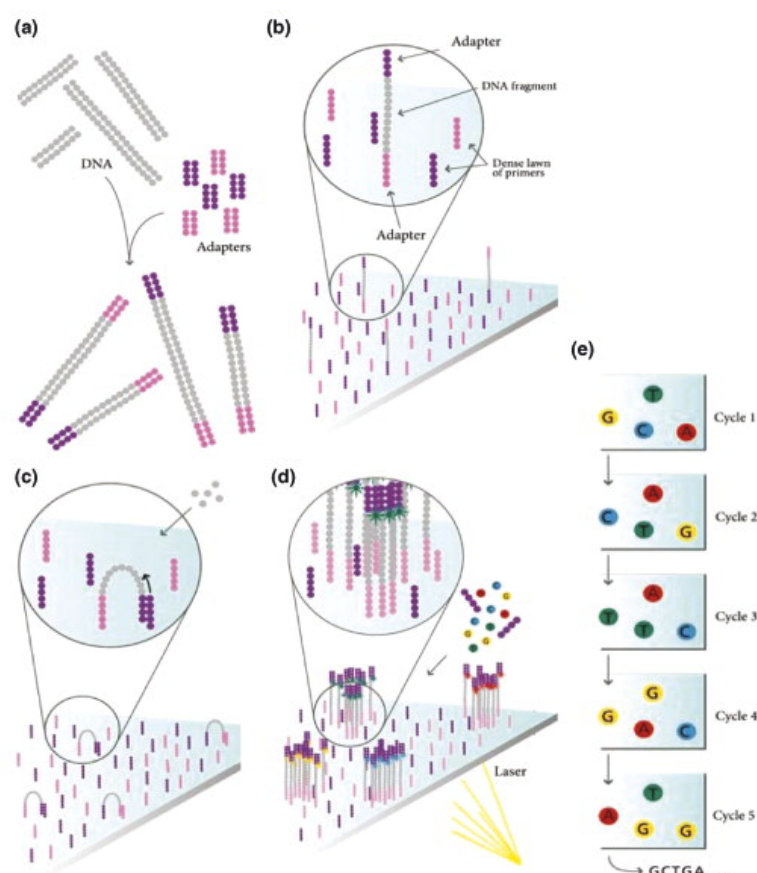


Figure 1.7: Schematic representation of the Illumina/Solexa sequencing process. Figure adapted from Strausberg et al.^[192]

genome^[194]. Although both methods allow for interrogation of cytosine methylation at single bp resolution, they do not distinguish between 5mC and the recently discovered 5hmC. On the other hand, MethylCap-seq and MeDIP-seq use methyl-binding domain proteins and specific antibodies respectively to enrich sonicated DNA for either 5mC or 5hmC fractions before sequencing^[195].

This general approach of enriching or manipulating sample DNA to obtain fragments of interest, followed by sequencing, read alignment to the corresponding reference genome and computational identification of genomic regions with increased “signal” is a common theme among the remaining assays discussed. In the case of DNase-seq, rather than sonication, sample nuclei are treated with the DNase 1 restriction enzyme which preferentially cleaves at sites of accessible chromatin, for example nucleosome-depleted TSSs and other TF-engaged regulatory regions^[196]. The genomic density of mapped read start sites provides quantitative cleavage site information, where sites of locally increased signal, or peaks, correspond to increased chromatin accessibility (DHSs). Other methods used to map “open” regions include FAIRE-seq^[41] and the related Sono-seq^[42], both of which involve cross-linking of chromatin using formaldehyde before sonication. Conversely, DNA fragments associated with individual mononucleosomes can be isolated using micrococcal nuclease-digestion (MNase-seq), which preferentially cleaves linker regions between adjacent nucleosomes^[197].

Finally, genomic regions bound by – or in close proximity to – a protein of interest, be it a TF or a modified histone, are determined using chromatin immunoprecipitation followed by sequencing (ChIP-seq). Briefly, this involves sonication of cross-linked chromatin, isolation of protein-associated fragments using a corresponding antibody, reversal of cross-links and sequencing of the recovered DNA to determine the fragments originally bound by the specific protein^[198] (Figure 1.8). Importantly, a control sample prepared using the same procedure, but excluding the immunoprecipitation step, is needed to control for regional and cell-type-specific biases in sonication and platform-specific sequencing efficiency^[199]. This “Input” (Sono-seq) data is used in downstream computational analysis as described in the following section. Other experimental considerations affecting the ChIP enrichment, or signal-to-noise ratio, of target protein associated sites include antibody efficiency, sequencing depth and library complexity.

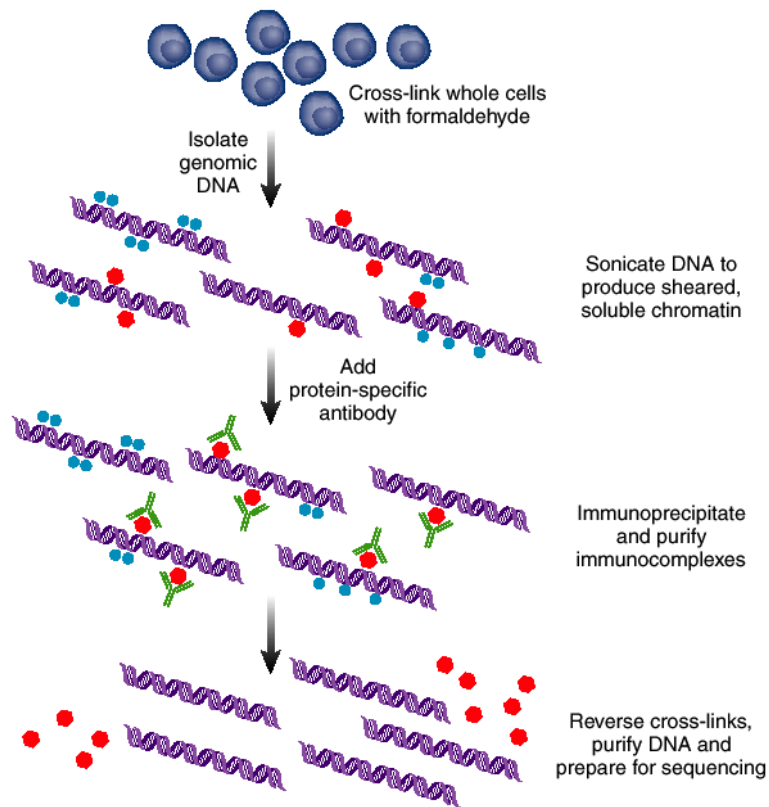


Figure 1.8: Diagram showing wet-lab workflow of a typical ChIP-seq experiment. Steps involve cross-linking of proteins to DNA, fragmentation by sonication, isolation of fragments of interest by immunoprecipitation, reversal of cross-links, DNA purification and sequencing. Red circles represent the protein of interest. Figure adapted from Mardis et al.^[200]

1.5.2.1 ChIP-seq data analysis

The data analysis of a ChIP-seq experiment starts with the quality assessment of raw FASTQ output files which contain short read sequences and associated per-base quality values indicating the estimated probability that the corresponding base call is incorrect. The Babraham Bioinformatics FastQC application^[201] and Bioconductor R package ShortRead^[202] both provide web-browsable reports with aggregate quality metrics for individual raw read files, including quality score distributions, per-base sequence content, sequence duplication levels and overrepresented sequences. These checks can help identify and potentially remedy sub-standard datasets: high levels of duplicates indicates low sequence diversity in the original library, potentially the result of PCR over-amplification; adapter contamination can be identified by overrepresentation of corresponding adapter sequences. The latter can be remedied by trimming 5' or 3' read ends to maximise mappable sequences.

Popular tools available to map reads to the reference genome such as ELAND^[189], Maq^[203], BWA^[204] and Bowtie^[205], differ in indexing strategies, speed and sensitivity. Base quality values are used to inform the scoring of alignments (all except ELAND) and indexing with the Burrows-Wheeler transform (used in the latter two tools) increases speed and memory-efficiency. After alignment, a typical ChIP-seq pipeline filters out identical duplicates (possible PCR artefacts), low quality mappings or reads with high mapping uncertainty, those corresponding to scaffolds or non-canonical chromosomes, followed by visualisation of read pile-ups, or coverage vectors, in a genome browser like Ensembl^[206] or IGV^[207].

Good quality ChIP-seq datasets produce obvious peaks of read density that are readily identified by visual inspection and a common next step is peak calling, which produces a set of discrete genomic intervals of significant enrichment likely representing bona fide sites of protein localisation. Although this corresponds to a somewhat arbitrary thresholding of the genome into “bound” and “unbound” regions, ignoring the complexity of the underlying biology, strength/frequency of binding and cell-to-cell differences etc., it is a computationally convenient first step in the analysis. Similarly to short-read mapping software, there are a large number of available tools for peak calling, which use properties like peak shape, strand shift (offset due to sequencing of bound fragment ends), read depth and control reads to distinguish between IP-related signal and noise^[208]. The popular MACS peak caller uses a window-based

approach to determine significantly enriched regions and an empirical false discovery rate (FDR) calculated by randomly shuffling IP and Input reads^[209]. SWEmbl – a TF peak caller developed at the EMBL-EBI and used in this thesis – employs a strand-specific Markov model-based approach where a tuneable score function is incremented and decremented according to IP and Input read distributions respectively. Many tools, including SWEmbl, provide an estimate of the peak “summit”, or position of highest enrichment, which is useful for binding site motif-related analyses (see below). Another consideration is the localisation pattern of the assayed protein: point-source factors such as TFs tend to occupy punctate, sharply defined regions; broad-source factors are associated with large genomic domains. The CCAT peak caller provides specific configurations for both types of factor^[210].

The output from peak calling is typically tens of thousands of regions, but this number can be variable across replicates and depends on experimental factors, the protein of interest, the peak calling algorithm used, as well as sequencing depth, with more reads leading to increased sensitivity to detect higher numbers of peaks^[199]. Nevertheless, most methods consistently detect the highest-scoring peaks, so these are a justifiable focus set for DNA motif-finding analyses with tools including MEME^[211] and Nested-MICA^[212]. Previously identified position weight matrices (PWMs) corresponding to TF binding preferences are stored in databases like TRANSFAC^[213] and these can also be used to identify overrepresented motifs by scanning peak sequences. In this thesis I combined these two approaches to find motifs from both above-mentioned *de novo* motif discovery algorithms that best distinguished “bound” from randomly sampled “unbound” genomic regions, which were retained for downstream analysis.

The distribution and functional significance of ChIP-seq peaks can be investigated by direct overlap or proximity to nearby annotated genomic features. For example, testing TF-bound genes for enrichment of functional terms – as provided by the Gene Ontology^[214] (GO) – may help formulate hypotheses regarding the biological purpose of binding. The construction of the background set, or null distribution, is an important consideration in these enrichment tests, and other statistical hypothesis tests based on genome information. In the absence of 3D structural information, putative target genes of individual TFs or clusters thereof (CRMs) can be assigned naïvely based on minimal distance or according to more sophisticated strategies as those used by the GREAT tool^[215].

Constructing aggregate read/fragment profiles centred on peak summit positions, motifs or other annotated features, allows the study of more subtle biological effects, including peak shape, shift or enrichment changes in different conditions/contexts, that are likely to be missed by peak overlap analyses. Recent approaches, originally developed to identify differentially expressed transcripts in RNA-seq datasets (see Section 1.5.3), have been co-opted to detect significant changes in the ChIP signal intensity of individual peaks across conditions. These are included in packages such as EdgeR^[216] and DESeq^[217] – and wrapped by DiffBind^[218] – which model the distribution of mapped reads in the genome with the negative binomial distribution and use tests that capture read count variability across biological replicates. These methods are more robust and sensitive than those based on peak presence/absence, which are unable to detect the magnitude and statistical significance of ChIP signal intensity changes. Lastly, as such analyses of genome-scale data often involve thousands of comparisons, this needs to be taken into account by correcting for multiple testing.

1.5.3 Transcriptomic assays

Two major goals of genome-wide analyses at the RNA level, transcriptome annotation and quantification, have traditionally been tackled using separate technologies: Sanger sequencing of cDNA or expressed sequence tag (EST) libraries, and microarray expression analysis of known transcripts respectively^[219]. The advent of NGS has enabled the replacement of both with faster and cheaper approaches based on high-throughput RNA sequencing (RNA-seq). Briefly, the experimental protocol involves isolation of sample RNA, optional fractionation (e.g. poly(A)+ for mRNA), fragmentation and reverse transcription to cDNA, followed by sequencing^[185]. Despite their limitations, gene expression microarrays are still commonly used, with mature computational methods available for associated pre-processing, normalisation^[220] and differential expression analysis^[221,222].

Apart from already-mentioned benefits of NGS-based approaches, one particular to RNA-seq is non-reliance on existing genomic sequence, and tools such as Velvet^[223] and Trans-ABYSS^[224] have been applied to the task of *de novo* transcriptome assembly. There are also programs for the easier task of genome-guided transcriptome reconstruc-

tion, allowing for the discovery of novel transcripts, splice junctions and RNA sequence variations (SNPs)^[185]. In this case it is necessary to use spliced read aligners (allowing gaps) and, optionally, paired-end reads – where both ends of the isolated fragments are sequenced – to improve alignment results. An example computational workflow for transcriptome reference-based quantification and differential expression analysis is outlined in the following subsection, which is relevant if comprehensive transcript sequences are already available.

1.5.3.1 Differential expression analysis with RNA-seq

Following FASTQ quality assessment (see Section 1.5.2.1), RNA-seq reads are aligned to transcriptome cDNA sequences, either independently obtained (e.g. Ensembl) or assembled from the data itself. The result is a list of read counts per transcript, but two sources of unwanted variation need to be addressed before comparisons can be made between genes and across experiments: (i) longer transcripts, when sonicated, generate more RNA/cDNA fragments (and therefore reads) than shorter transcripts, even when expressed at identical levels and (ii) read counts are directly related to the total number of sequences per lane/sample. Normalisation using the fragments per kilobase (kb) of transcript per million mapped reads/read-pairs (FPKM) measure accounts for both of these factors^[219]. Reads mapping equally well to multiple transcripts, or multi-mapping reads, represent a particular challenge to RNA-seq data analysis due to the identity, or similarity, of transcripts originating from paralogs or different isoforms of the same gene. Initial methods either discarded these entirely or assigned reads mapping to multiple locations proportionally according to uniquely mapping reads. However, more sophisticated probabilistic approaches, like those provided by MMSEQ^[225], RSEM^[226] and Cufflinks^[227], are needed to resolve cases of overlapping features i.e. transcript isoforms sharing exonic sequence. Count equivalents of the normalised and deconvoluted expression estimates can then be compared across replicate experiments using EdgeR^[216] or DESeq^[217] at the transcript or gene (aggregate) level to determine differentially expressed transcripts or genes respectively (see Section 1.5.2.1).

1.5.4 Chromatin conformation assays

Initial studies of nuclear architecture were carried out using DNA or RNA FISH, which is a low-throughput and low resolution approach based on fluorescence microscopy, only allowing the visualisation of a handful of sites in individual cells^[228]. However, recently developed molecular techniques based on 3C enable quantitative measurements of chromatin contacts present in a population of cells, providing average maps of chromosome folding at up to kilobase resolution^[87]. The classical 3C assay starts by fixing cells with formaldehyde to covalently link interacting loci (Figure 1.9A). This is followed by restriction enzyme digestion and ligation of fragment ends, favouring those within the same cross-linked complex which are in close spatial proximity. Single hybrid DNA molecules, representing interactions of interest, are then detected using locus-specific PCR^[229] (Figure 1.9B).

All other 3C-based assays use the same basic principle but differ mainly in throughput. Chromosome conformation capture-on-chip^[230,231] (4C), or circular 3C^[232], uses inverse PCR to selectively amplify fragments interacting with a single locus (Figure 1.9B), which are then identified by microarray analysis or NGS^[233] (4C-seq). The result is a high resolution genome-wide profile of contacts involving the selected “bait” region (one-to-all). 3C-carbon copy^[234] (5C) uses multiplexed ligation-mediated amplification (LMA) to quantify many pairs of interactions in parallel (many-to-many, Figure 1.9B). However, the number of primers that would be needed for simultaneous interrogation of all restriction fragments in the genome prohibits a global 5C experiment.

Hi-C was developed to address the need for an unbiased genome-wide chromatin conformation assay^[22] (all-to-all). In this technique, biotin labelling of restriction fragment ends before ligation enables the selection and subsequent paired-end sequencing of sheared DNA possessing a ligation junction i.e. hybrid molecules (Figure 1.9B). Other genome-wide adaptations of the 3C approach include chromatin interaction analysis by paired-end tag sequencing^[168] (ChIA-PET), which uses an IP step to enrich for interactions involving a particular protein of interest.

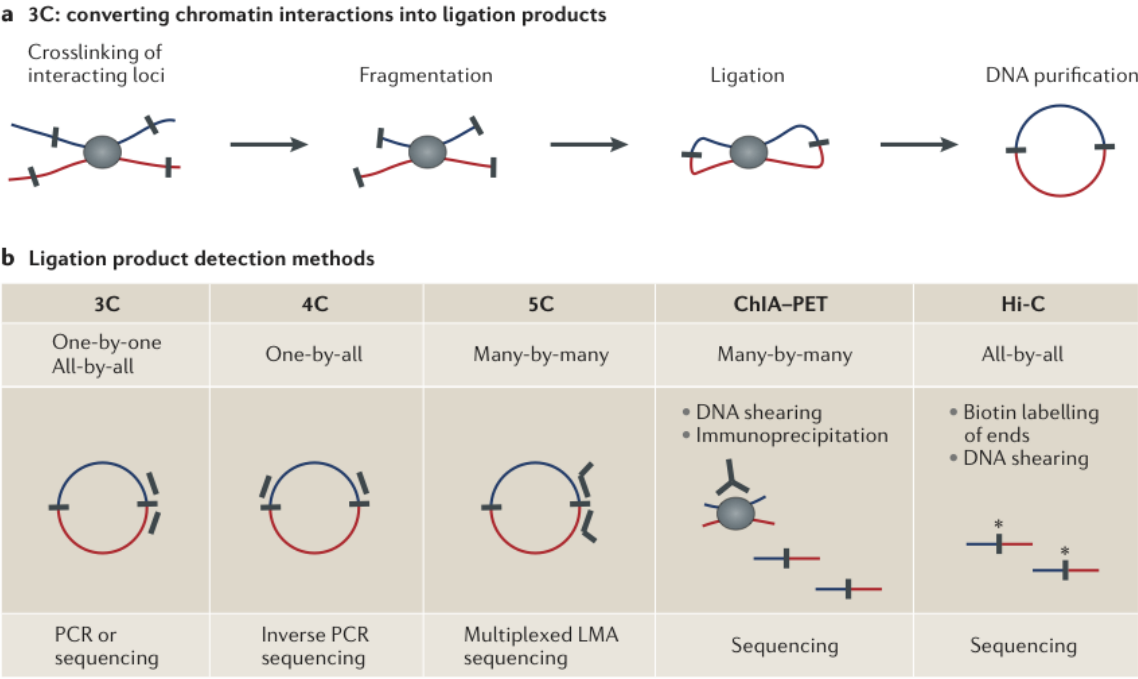


Figure 1.9: 3C-based methods to assay chromatin conformation. (A) The general scheme common to all methods consists of ligating cross-linked chromatin fragments thereby capturing information about originally proximal genomic regions. (B) Various 3C-based methods arranged according to throughput (from left to right). Figure adapted from Dekker et al. ^[87]

1.5.4.1 Hi-C data analysis

The result of a Hi-C experiment is on the order of hundreds of millions of paired-end reads each potentially representing a single ligation event. The computational analysis and visualisation of Hi-C data poses a number of unique challenges related to the sheer volume and dimensionality of data generated due to the all-to-all nature of the assay. Steps common to all workflows include raw read pre-processing, mapping and filtering to obtain valid pairs representing putative chromatin interactions. However, there is currently no consensus regarding subsequent normalisation and analysis techniques, and the choice of which likely depends on the particular biological question at hand.

As each read in a pair originates from a distinct genomic locus, they are aligned to the reference genome independently using a standard single-end short-read mapping tool, like Bowtie or BWA, and merged for downstream processing. Depending on the level of sonication and the length of the reads, some sequences may contain the ligation junction itself (5'-AAGCTAGCTT-3' in the case of HindIII), which can be removed by corresponding read truncation to improve mapping efficiency. This is the approach taken by HiCUP and HOMER^[89,235], both publicly available software packages for Hi-C data pre-processing and analysis. An alternative iterative mapping technique is provided by the hiclib pipeline^[236], which takes a “trial and error” approach where reads are repeatedly trimmed by 5 bp until a unique alignment is obtained. After mapping, reads are filtered to remove invalid pairs such as those representing self-ligations (self circles), unligated fragments (dangling ends) or inferred fragment lengths outside the selected size range.

By dividing the genome into bins – typically sized between 100 kb and 1 Mb – the filtered Hi-C data can be visualised as a two-dimensional symmetrical matrix, or heatmap, where the value/intensity in each cell corresponds to the number of read pairs linking loci represented by row and column indices (Figure 1.5). Such heatmaps of “raw” Hi-C data reveal the striking dependence of contact frequency on interaction distance i.e. the vast majority of read pairs represent proximal interactions close to the matrix diagonal. When averaged over all loci in the genome, this characteristic length relationship shows 100-fold differences in intra-chromosomal contact frequencies, with power-law scaling between distances of 0.7 Mb and 10 Mb^[22,87]. Another important source of variation is locus-specific read depth related to sequencing and other technical biases. Taking both of these factors into account, HOMER defines the expected

number of Hi-C reads in each cell as a function of estimated region-specific totals and interaction distance^[89,235]. An approximate solution to the set of nonlinear equations is then obtained using an iterative hill climbing optimisation technique, resulting in a matrix of expected reads, termed the background model, with row and column totals very close to that of the observed data. Normalised matrices are obtained by dividing the observed Hi-C matrices by their corresponding background models.

An alternative normalisation approach specifically models technical biases affecting contacts at the restriction fragment level (mappability, fragment length and GC content) and combines the estimated parameters into a single correction factor per fragment pair^[237]. A third related, but less computationally expensive, method is used in iterative correction and eigenvector decomposition (ICE), which operates on binned data yet produces comparable normalised Hi-C maps^[236]. Importantly, these two latter methods leave the distance-dependent relationship intrinsic to Hi-C data intact.

Principle components analysis (PCA) is a mathematical technique useful for transforming high-dimensional datasets into a set of orthogonal “meta-variables” (principle components) ranked by explained variance (highest to lowest). Performing PCA on a normalised Hi-C matrix and considering the first principle component (PC1) is a way to summarise the complex and difficult to interpret contact matrix into a one-dimensional vector where each value corresponds to the “connectivity” of each position (or bin) in the genome i.e. loci with similar PC1 values show broadly similar genome-wide contact profiles. The method of thresholding on PC1, termed eigenvector analysis of chromosomal organisation^[22,236], is used to classify the genome into one of two compartments, which tend to show distinct patterns of transcriptional activity (see Section 1.1.2.4).

Apart from exploring global properties of chromatin architecture, another goal in Hi-C data analysis is to discover significant looping interactions between genomic regions. HOMER uses the binomial distribution to model the number of Hi-C reads linking two loci (successes) out of those available (combined region-specific totals). Using the background model to calculate the probability of success, significant departure from the expected number of reads can be determined using a binomial test^[89,235].

Finally, although probing chromatin interactions between individual genomic features such as TF binding events is intractable with current Hi-C methods, an approach

called structured interaction matrix analysis (SIMA) has been developed, which pools information across many such sites thereby boosting the effective resolution^[89]. Briefly, SIMA counts the number of Hi-C reads connecting features of interest within a pair of predefined genomic domains, normalises by the expected count provided by the background model and determines significance based on a randomisation test i.e. shuffling feature positions many times. In this way, it is possible to determine whether particular features (e.g. CTCF sites), or feature pairs (e.g. promoters and enhancers), are enriched for chromatin interactions, possibly suggesting a role in facilitating particular 3D topologies^[89] (see Section 3.3.5).

Chapter 2

Cohesin regulates tissue-specific expression by stabilising highly occupied *cis*-regulatory modules

2.1 Summary

The cohesin protein complex contributes to transcriptional regulation in a CTCF-independent manner by colocalising with master regulators at tissue-specific loci. The regulation of transcription involves the concerted action of multiple transcription factors (TFs) and cohesin's role in this context of combinatorial TF binding has not been previously explored. To investigate cohesin-non-CTCF (CNC) binding events in vivo we mapped cohesin and CTCF, as well as a collection of tissue-specific and ubiquitous transcriptional regulators, using ChIP-seq in primary mouse liver. We observe a positive correlation between the number of distinct TFs bound and the presence of CNC sites. In contrast to regions of the genome where cohesin and CTCF colocalise, CNC sites coincide with the binding of master regulators and enhancer-markers and are significantly associated with liver-specific expressed genes. We also show that cohesin presence partially explains the commonly observed discrepancy between TF motif score and ChIP signal. Evidence from these statistical analyses in wild type cells, and comparisons to maps of TF binding in *Rad21*-cohesin haploinsufficient mouse liver, suggests that cohesin helps to stabilise large protein-DNA complexes. Finally, we observe that the presence of mirrored CTCF binding events at promoters and their nearby

cohesin-bound enhancers is associated with elevated expression levels.

This study is the result of a collaboration between Dr. Duncan Odom’s laboratory at the Cancer Research UK Cambridge Institute and Dr. Paul Flicek’s research group at the EMBL European Bioinformatics Institute. Dr. Dominic Schmidt performed most of the experiments for this project and I carried out the computational analysis, except where otherwise specified. This chapter is based on a paper that has recently been published in *Genome Research*^[238] and is adapted here with the publisher’s permission.

2.2 Introduction

There is increasing evidence to suggest that changes in higher-order genome structure and subnuclear chromatin localisation are crucial for lineage specification and temporal/tissue-specific transcriptional regulation^[239]. In view of CTCF’s involvement in mediating chromatin loops at specific developmentally regulated genomic loci, it has been suggested that the primary role of CTCF may be the genome-wide organisation of chromatin architecture^[78]. However, given the similarities in CTCF occupancy across different cell types^[134], it is not clear how CTCF alone could configure the three-dimensional structure of the genome in a dynamic way. Considering that cohesin and CTCF colocalise genome-wide, it is an attractive hypothesis that cohesin contributes to this global organisational function. Indeed, mutations responsible for human cohesinopathies, such as Cornelia de Lange Syndrome, severely disrupt the subnuclear organisation of chromatin and cause aberrant nucleolar morphology when induced in budding yeast^[240]. Furthermore, studies at specific loci show that cohesin dynamically controls the spatial conformation of chromatin required for normal development and differentiation, in a cell-division independent way^[163,170].

Recent research showing that cohesin occurs at *cis*-regulatory elements apart from CTCF and contributes to global gene expression in human cancer cell lines and mouse ES cells, implicates the widely expressed cohesin complex in tissue-specific functions. In this context, causal evidence at the *Tcra* locus, as well as genome-wide correlative evidence involving ER and pluripotency factors, also links cohesin to the stabilisation of long-range regulatory interactions (see Section 1.3). In order to investigate *in vivo* patterns of cohesin binding in depth – particularly independent of CTCF – we mapped

both factors together with a collection of ten TFs using ChIP-seq in primary mouse liver. We collected additional data from the same tissue for several histone modifications and other functional DNA-protein interactions, providing a comprehensive map of cohesin’s role in tissue-specific transcriptional regulation. We show that this role is likely to be functionally similar across multiple tissues by demonstrating that cohesin’s presence at binding events of liver-specific TFs parallels its localisation with ES cell-specific factors. Results from the computational analysis and integration of these datasets with expression and sequence-level information provide a number of novel insights, including evidence of a potential role for cohesin in the stabilisation of highly occupied *cis*-regulatory modules.

2.3 Results

We performed ChIP-seq experiments in primary mouse liver with antibodies targeted to CTCF, three cohesin subunits (RAD21, STAG1 and STAG2), ten TFs (CEBPA, HNF4A, FOXA1, FOXA2, ONECUT1, HNF1A, PKNOX1, REST, GABPA, E2F4), two co-activators (EP300, CREBBP), five histone-modifications (H3K4me1, H3K4me3, H3K36me3, H3K79me2, H2AK5ac) and RNA polymerase II (RNAP2) (see Section 2.5.1).

The TFs for our analysis were chosen to include both ubiquitously expressed factors and liver-specific regulators, two of which have well-characterised evolutionary dynamics^[19]. We additionally profiled chromatin marks associated with active TSSs, enhancers and transcribed genes, providing a comprehensive picture of the genome function and the transcriptional regulatory network active in mouse liver cells. Figure 2.1A displays a number of key regulatory features of the data in the vicinity of the predominantly liver-specific phosphoenolpyruvate carboxykinase 1 (*Pck1*) gene on mouse chromosome 2, including two clusters of TFs: one immediately proximal to the TSS and another approximately 25 kb upstream of the TSS. Cohesin can be seen colocalising with CTCF as well as with clusters of TFs. These data are quantitatively and qualitatively comparable to other multi-factor experiments in other tissues^[241].

After short read alignment with BWA^[204] and peak calling with SWEml (Wilder et al. in preparation) (see Section 2.5.2), we determined the overlap between sites

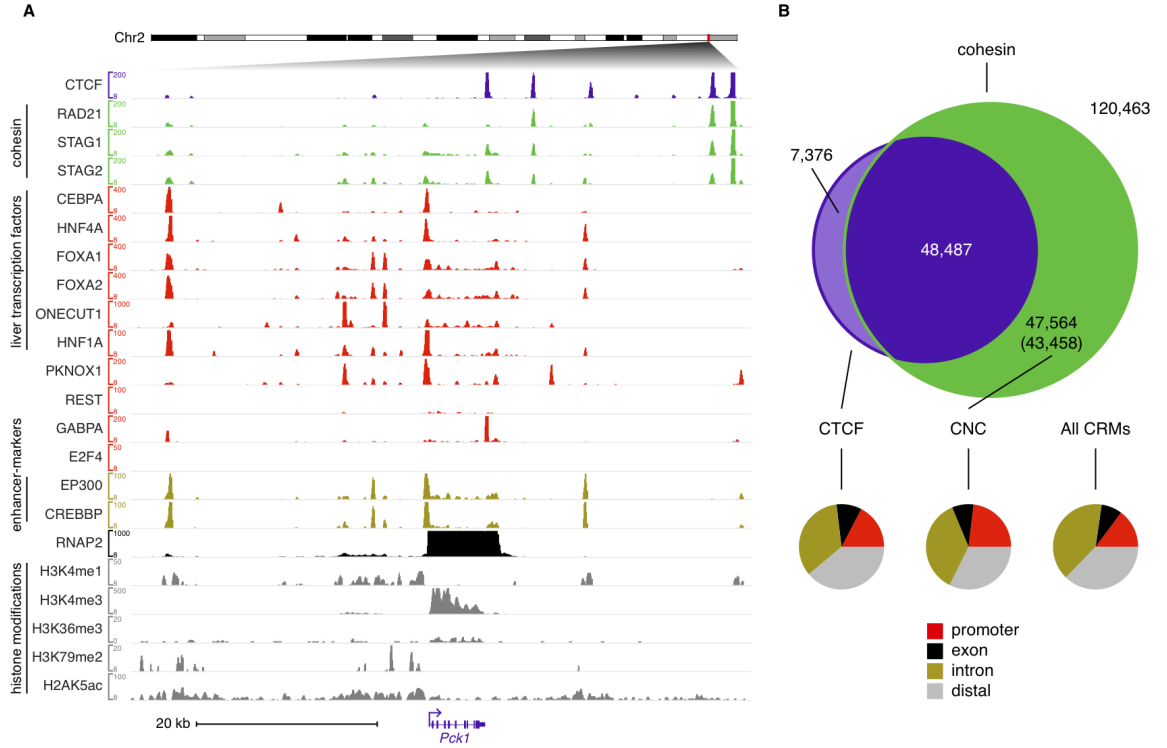


Figure 2.1: Genome-wide distribution of CRMs in primary mouse liver as measured by ChIP-seq. (A) Occupancy of cohesin, CTCF, tissue-specific and ubiquitous TFs near the *Pck1* gene. Cohesin colocalises with CTCF as well as with clusters of transcription factors in the absence of CTCF, one of which can be seen overlapping the TSS of the *Pck1* gene. (B) Venn diagram showing CTCF and cohesin (RAD21, STAG1, STAG2) occurrence within CRMs. The pie charts indicate genomic locations of all CRMs (background), as well as those containing CTCF and CNC. The latter occur within promoter regions at a higher relative frequency compared to the other two classes.

bound by CTCF and the cohesin subunits. As expected, the three assayed cohesin subunits show highly similar patterns of binding with peaks of the RAD21 subunit coinciding with 99% and 94% of STAG1 and STAG2 peaks respectively^[167]. By defining cohesin presence as the occurrence of at least one of its subunits, we find that cohesin colocalises at the majority of CTCF sites (48,487; 87%), but is also present at a similar number of sites independently of CTCF (Figure 2.1B). We define this latter set of 46,471 cohesin binding sites as cohesin-non-CTCF (CNC) sites.

2.3.1 CTCF-independent cohesin binding is associated with master regulators and enhancers

To determine the binding partners of cohesin at both cohesin-CTCF and CNC sites, we defined a set of putative *cis*-regulatory modules (CRMs) by grouping together CTCF and cohesin binding events with overlapping binding events of the ten TFs, and the co-activators EP300 and CREBBP (see Section 2.5.2). The resulting CRMs have a median width of 449 bp, but vary in size depending on the number of factors present ($sd = 346$ bp). The comparatively broad regions of the genome associated with the histone-modifications (H3K4me1, H3K4me3, H3K36me3, H3K79me2, H2AK5ac) and RNAP2 were not used to define the CRMs themselves. Instead they were used to annotate the chromatin state of the CRMs *post hoc* (see Section 2.5.2).

Of the 223,890 CRMs that were identified, 43,458 (19.4%) are identified as CNC sites. These CRMs mostly occur away from annotated TSSs (77%; Figure 2.1B) and tend to coincide with the binding of master regulators, such as HNF4A (69%), and the enhancer-markers EP300 and/or CREBBP (66%). The fraction of promoter-proximal CNCs (23%) is nevertheless higher than that of CTCF (17%), and CNC-containing CRMs are significantly enriched for occurrence near TSSs when compared to all CRMs (Fisher’s Exact Test $P < 10^{-15}$; Figure 2.1B). CNC sites that occur within promoter regions (≤ 2.5 kb from the annotated TSS), are highly enriched for RNAP2 binding compared to cohesin-bound promoters in general (Fisher’s Exact Test $P < 10^{-15}$). These results are similar to those in *Drosophila*, where cohesin lacks a functional interaction with CTCF^[242], but is preferentially detected at promoters of active genes^[243]. Here cohesin selectively binds genes with paused RNAP2 and lacking H3K36me3, a mark associated with transcriptional elongation^[244]. Although we find that cohesin is

associated with increased RNAP2 pausing indices in mouse liver cells (Figure 2.2A), cohesin-bound promoters are also associated with elevated expression levels and an enrichment of H3K36me3 within the gene body (Figure 2.2B, C).

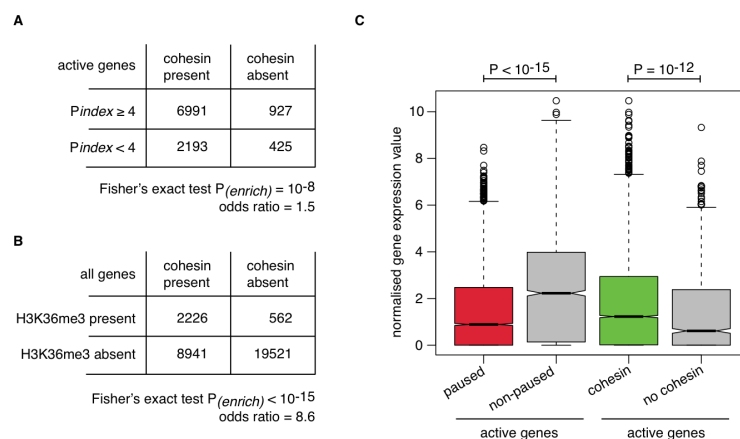


Figure 2.2: Cohesin binding and RNAP2 pausing. (A) Cohesin binding within promoter regions (≤ 2.5 kb from an annotated TSS) is significantly associated with high RNAP2 pausing indices ($Pindex$). Pausing indices were calculated as the length normalised ratio between RNAP2 ChIP fragments mapping within 300 bp of the TSS and those mapping within the entire gene. We considered only active genes, defined as those with RNAP2 peaks overlapping their TSSs, and defined paused promoters as those having $Pindex \geq 4$ ^[245]. (B) Genic H3K36me3 is significantly associated with cohesin-bound promoters genome-wide. (C) Active genes with high RNAP2 pausing indices are associated with lower levels of transcription, whereas cohesin presence at promoters is associated with elevated levels of transcription.

At cohesin sites containing CTCF, we observe a shift in the summit positions of all cohesin subunits with respect to the CTCF summit position when the orientation of the CTCF motif is taken into account (Figure 2.3A). This result is similar to recent reports for RAD21^[246] and supports a direct and directional biochemical interaction between cohesin and CTCF. The observed peak shift may be a result of an interaction between cohesin and a particular CTCF domain that is consistently distal to the DNA-binding region^[162]. The ≈ 20 bp offset of cohesin with respect to CTCF does not preclude close proximity of the proteins given that the DNase 1 footprint of CTCF can be up to 40 bp in length^[247]. The same directional analysis at CNC sites, however, reveals that

the position of cohesin is independent of the peak position and motif orientation of all other sequence-specific factors considered (Figure 2.3B). This demonstrates a specific cohesin-CTCF interaction that is not seen at CNC sites and suggests a different mechanism of cohesin recruitment in the absence of CTCF.

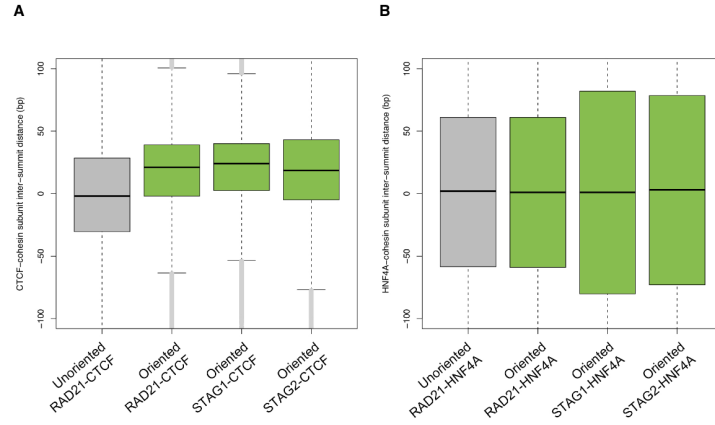


Figure 2.3: Cohesin peak shift with respect to sequence-specific factors. A significant peak shift, suggesting a direct biochemical interaction between cohesin and CTCF, is not present when compared to all other available TFs at CNCs. (A) At sites where CTCF and cohesin colocalise, we observe a shift in the summit position of all cohesin subunits (RAD21, STAG1, STAG2) with respect to the CTCF summit position when the orientation of the CTCF motif is taken into account. The grey boxplot indicates the genomic distance between summit positions of overlapping RAD21 and CTCF ChIP-seq peaks i.e. the sign of the inter-summmit distance is simply a reflection of the chromosomal coordinates of the two summits (unoriented). The green boxplots indicate the genomic distance between summit positions when the orientation (strand) of the best CTCF motif match within its peak was taken into account i.e. the sign of the inter-summmit distance is determined by the strand of the best CTCF motif (oriented). (B) On average, cohesin binding events exhibit no shift in summit position with respect to HNF4A, whether oriented on the strand of the best HNF4A motif match or not. Similar results were obtained using ChIP-seq data for the other nine TFs available (not shown).

To identify the primary interacting partner proteins within the CRMs, we used the within-CRM ChIP fragment count (i.e. the number of mapped ChIP reads, extended to

the estimated fragment length, overlapping the CRM) to measure the binding strength correlation between all ChIP-seq datasets. These correlations highlighted two separate modes of cohesin binding. First, a clear and distinct cluster includes all three cohesin subunits and CTCF (Figure 2.4, purple cluster). Second, cohesin subunits also correlate with tissue-specific factors including FOXA1, HNF4A and HNF1A (green cluster); TFs are not correlated with CTCF. The cohesin/TF cluster is also marked by the active histone modifications H3K4me3, H3K4me1 and H2AK5ac, as well as with the co-activators EP300 and CREBBP, suggesting that cohesin within CNC sites may play a central role in active transcriptional regulation together with a wide range of TFs.

To investigate whether the correlations between CTCF, cohesin and tissue-specific TFs are particular to liver or differentiated tissue, we performed the same analysis for a set of previously published ChIP-seq datasets from mouse ES cells^[169,241,248,249] (see Section 2.5.2). Although the ES cell ChIP-seq dataset contains a different collection of TFs and cohesin subunits, they show patterns highly similar to those observed in primary liver tissue (Figure 2.4). Indeed, the SMC1A and SMC3 cohesin subunits correlate with CTCF (Figure 2.5, purple cluster) while also forming a separate, distinct cluster with key regulators of stem cell identity (POU5F1, SOX2 and NANOG), components of the mediator complex, and RNAP2 (Figure 2.5, green cluster). The cohesin loading factor is absent from the SMC1A/SMC3/CTCF cluster, which is consistent with previous observations of NIPBL’s preferential association with CNC sites and supports the idea of a different mechanism of cohesin recruitment in the absence of CTCF^[169]. Interestingly, CNC sites share a number of characteristics with recently discovered “super-enhancers”^[250,251] or “stretch enhancers”^[252], which are defined as dense clusters of enhancers with particularly high ChIP signals for cell-type-specific master transcription factors and components of the mediator complex. Indeed, the vast majority of super-enhancers in mouse ES cells coincide with a CNC (219, 95%) and most possess at least two (170, 74%). Conversely, less than half of super-enhancers overlap CTCF binding events (109, 47%). These findings are consistent with recent results showing an enrichment of NIPBL, cohesin and condensin at super-enhancers^[253]. Overall, cohesin shows two separate modes of binding that have minimal overlap in two transcriptionally divergent and phenotypically distinct mouse tissues: (i) either with CTCF and showing minimal signs of transcriptional activity, or (ii) with clusters of tissue-specific TFs showing hallmarks of transcriptional activation.

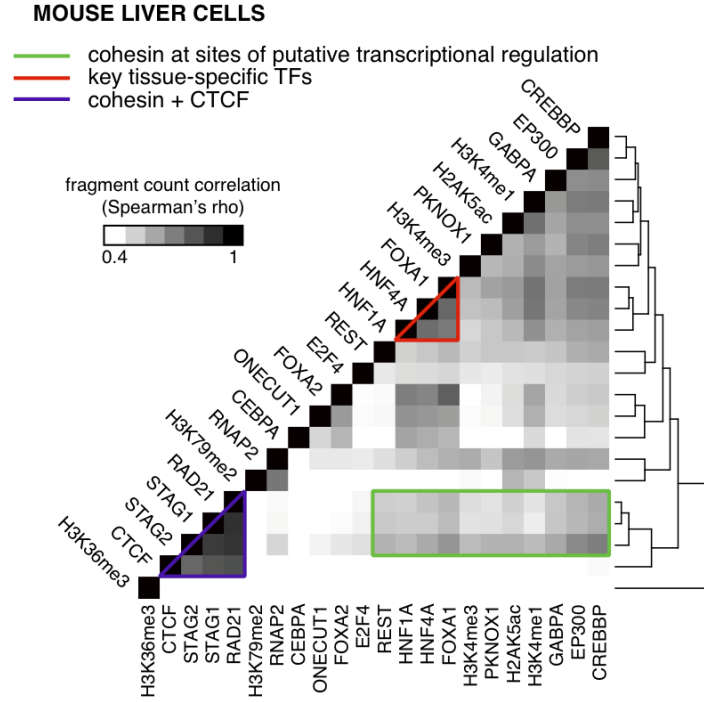


Figure 2.4: Within-CRM binding correlations reveal distinct modes of cohesin binding in mouse liver cells. The number of ChIP fragments (mapped reads extended to the estimated fragment length) overlapping a given CRM was used as a measure of binding strength for each dataset. Factors were clustered along both axes based on the similarity in their colocalisation profiles. Heatmap visualisation of all pair-wise correlations between all ChIP-seq datasets in mouse liver cells illustrates cohesin subunits (RAD21, STAG1, STAG2) clustered with CTCF. Cohesin also correlates with key tissue-specific TFs (FOXA1, HNF4A and HNF1A) independently of CTCF as well as with histone modifications associated with transcriptional activity (H3K4me1, H3K4me3, H2AK5ac) and co-activators (EP300 and CREBBP). Similar results were obtained by performing the correlation analysis separately on CRMs with CNC and CTCF (not shown).

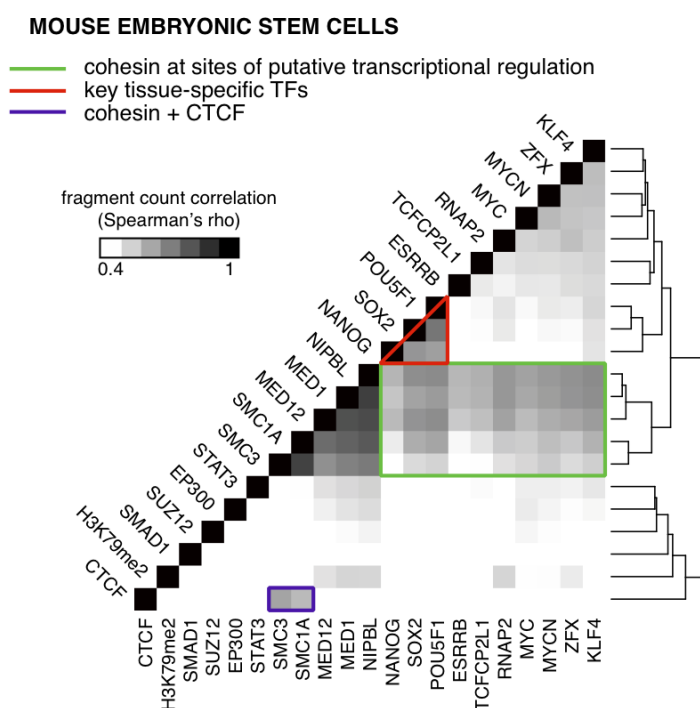


Figure 2.5: Within-CRM binding correlations reveal distinct modes of cohesin binding in mouse ES cells (see Figure 2.4). All pair-wise correlations between previously published ChIP-seq datasets in mouse ES cells. Cohesin binding strength (SMC1A, SMC3) correlates with CTCF while also forming a distinct cluster with key regulators of stem cell identity (POU5F1, SOX2, NANOG, MYC), components of the mediator complex, as well as RNAP2.

2.3.2 CNC sites occur preferentially at multiply-bound *cis*-regulatory modules (CRMs)

To understand the genomic properties of the identified CRMs, we grouped them into similar clusters based either on the normalised ChIP enrichment or binary presence/-absence of the sequence-specific factors using two different clustering methods (K-means and *AutoClass*) (see Section 2.5.2). A primary difference between these two clustering methods is that K-means requires the number of clusters to be defined a priori, whereas *AutoClass* uses a Bayesian probabilistic approach to automatically optimise the properties of each cluster (as well as the number of clusters) to achieve the best separation. Because the overall clustering results were similar between the two methods, we focused our analysis on the results from K-means (with $K = 10$) for ease of interpretation (Figure 2.6). See Section 2.5.2 for a justification of the choice of K and Figure 2.7 for results obtained using *AutoClass*.

The ten clusters totalling 210,067 CRMs, are visualised in Figure 2.6, sorted from left to right by the fraction of CNC-containing CRMs in a given cluster. CRMs with CTCF form a large, distinct cluster at the extreme left (cluster 10; 41,368 CRMs). Most CRMs without CTCF fall into three large groups (clusters 7-9; 102,091 CRMs) with an average of less than two sequence-specific factors (singleton CRMs). The remaining six clusters (66,608 CRMs) have increasing numbers of colocalising TFs, with almost all possessing either HNF4A, FOXA1 or FOXA2 (99%) and nearly half (48%) possessing all three of these factors. Furthermore, these six clusters all show a distinct pattern of chromatin state. For example, compared with singleton CRMs, clusters 1-3 are more strongly enriched for RNAP2, EP300/CREBBP and H3K4me1 ($P < 10^{-15}$), indicating that these clusters are likely to contain active enhancers.

Ranking the CRM clusters by the proportion of CNC sites in each cluster, we observe a strong positive correlation between the average number of distinct TFs present and CNC presence (Spearman's $\rho = 0.95$, $P < 10^{-15}$). In other words CNC sites occur preferentially at multiply-bound CRMs. The most highly bound cluster of CRMs (cluster 1) is also enriched for well conserved TF binding events^[19] compared to the remaining clusters, including CEBPA binding events shared in five species from chicken to human (Fisher's Exact Test $P = 10^{-10}$) and HNF4A binding events shared in human, mouse and dog (Fisher's Exact Test $P < 10^{-15}$). Taken together, these observations

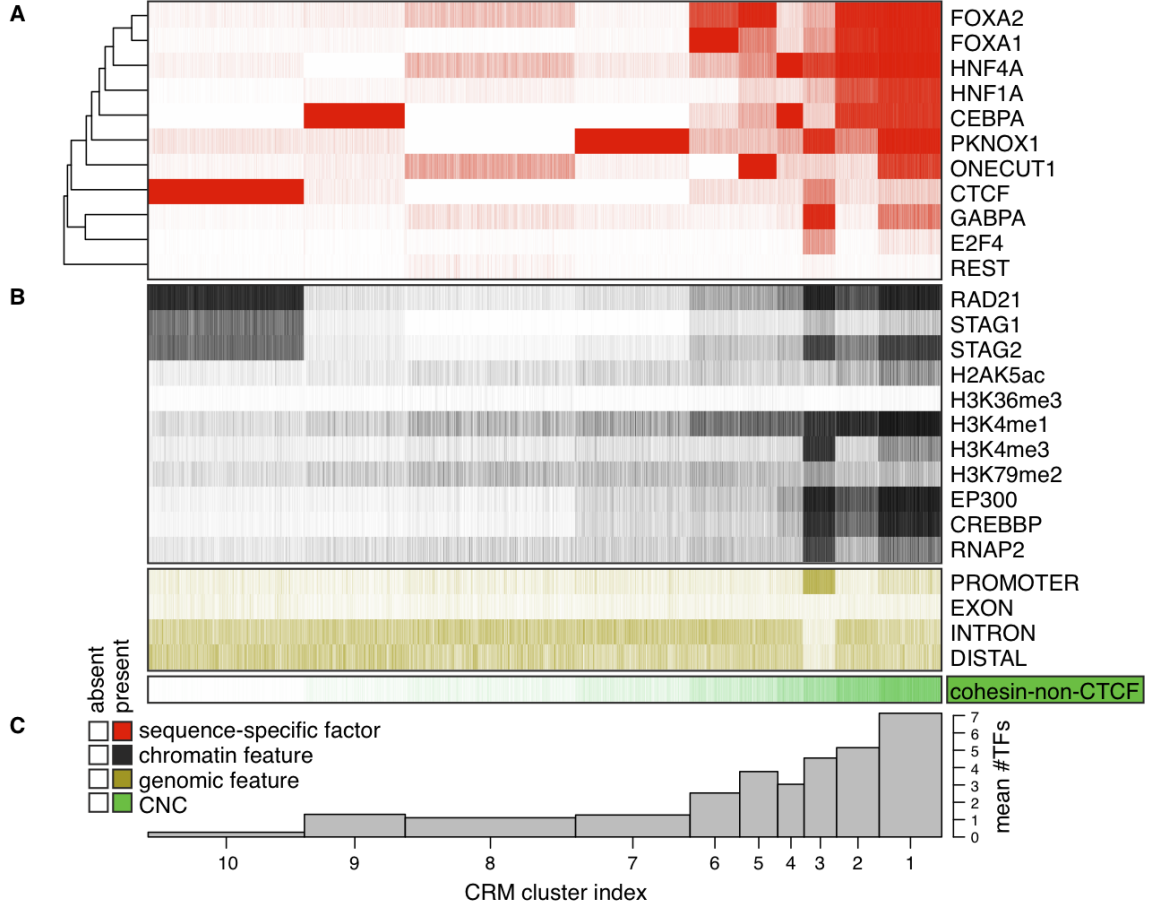


Figure 2.6: Cohesin-non-CTCF (CNC) binding occurs preferentially at multiply-bound CRMs. (A) Results from K-means clustering ($K = 10$) of the binary presence/absence of ChIP-seq peaks corresponding to the eleven sequence-specific factors within 210,067 CRMs containing at least one of these factors. Factors were clustered based on the similarity in their binary occupancy profiles. The clusters were indexed and sorted by the proportion of CRMs with CNC in each cluster (increasing from left to right). (B) The binary presence/absence of ChIP-seq peaks for various chromatin features (non-sequence-specific factors and histone modifications), visualised according to the K-means results in A. Genomic location with respect promoters (≤ 2.5 kb from an annotated TSS), exons, introns, and gene distal regions, is also indicated. The proportion of CRMs with CNC sites in each cluster is indicated at bottom (increasing from left to right). (C) Barplot indicating the mean number of distinct TFs within each CRM cluster. Bar widths correspond to the number of CRMs within each cluster.

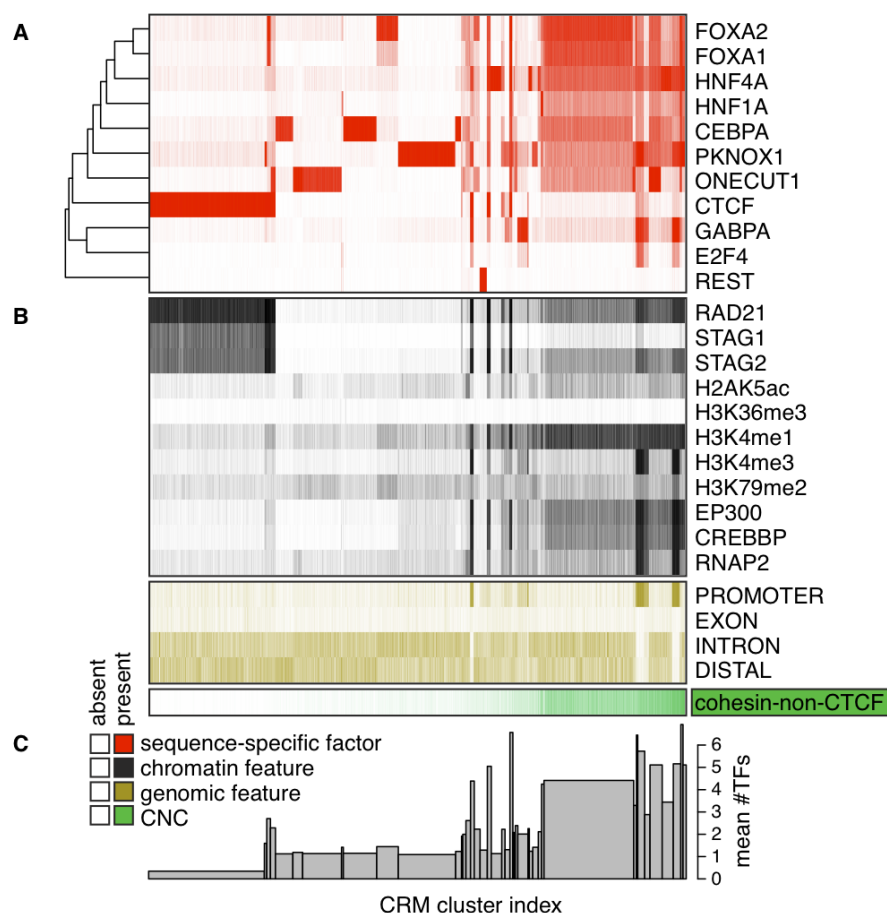


Figure 2.7: *AutoClass* CRM clustering results. (A) Results obtained using the *AutoClass* clustering tool applied to the normalised ChIP enrichment of the eleven sequence-specific factors within 210,067 CRMs, containing ChIP-seq peaks for at least one of these factors. We filtered the clustering results to retain only 142,157 CRMs with class membership posterior probabilities ≥ 0.5 and combined clusters with high correlation ($r > 0.9$) between their median ChIP enrichment profiles. The binary presence/absence of ChIP-seq peaks is displayed instead of the ChIP enrichment score in order to aid visualisation and for consistency with Figure 2.6. The heatmap was clustered along the ordinate based on the similarity of their binary occupancy profiles. The clusters were indexed and sorted along the abscissa by the proportion of CRMs with CNC in each cluster (increasing from left to right). (B) The binary presence/absence of ChIP-seq peaks for various chromatin features (non-sequence-specific factors and histone modifications), visualised according to the *AutoClass* results in A. Genomic location with respect to promoters (≤ 2.5 kb from an annotated TSS), exons (overlap with an exon but not a promoter), introns (overlap with a gene but neither an exon nor a promoter), and gene distal regions (elsewhere), is also indicated. The proportion of CRMs with CNC in each cluster is indicated at bottom (increasing from left to right). (C) Barplot indicating the mean number of TFs within each CRM cluster. Bar widths correspond to the number of CRMs within each cluster.

suggest a role for cohesin in integrating regulatory information from multiple TFs and stabilising the binding of large multi-protein complexes to *cis*-regulatory sequences.

2.3.3 CNC presence is associated with liver-specific gene expression

Results from the unsupervised clustering analysis suggested that there might be a direct correlation between the number of TFs bound within a CRM, CNC presence, and the transcriptional activity of the genomic regions. By explicitly grouping CRMs into classes based purely on the number of distinct TFs present, we see that the proportion of CNC-containing CRMs significantly correlates with the number of bound TFs (Spearman's $\rho = 0.89$, $P = 10^{-3}$), whereas CTCF shows no significant correlation (Spearman's $\rho = 0.49$, $P = 0.13$). Indeed, almost two thirds (62%) of highly occupied CRMs, defined as containing five or more TFs, possess CNC sites. The ratio of CNC- to CTCF-containing CRMs (CNC enrichment) is 0.2 when zero TFs are present, but reaches a maximum of three-fold at seven TFs before returning to equivalence at ten TFs (Figure 2.8A). The proportion of promoter proximal CRMs (≤ 2.5 kb from an annotated TSS) is also correlated with the number of distinct TFs present (Spearman's $\rho = 0.95$, $P < 10^{-15}$), but in contrast to CNC enrichment that peaks at seven TFs, the proportion of both RNAP2 and H3K4me3 increase monotonically from 0-10 TFs (Figure 2.8B). Other signs of transcriptional active chromatin, such as the presence of the coactivators EP300/CREBBP, show a similar consistently increasing trend from 1-10 TFs (not shown).

We next asked how these CRM occupancy patterns may be related to transcriptional output by assigning CRMs to their nearest canonical TSSs and using mouse liver expression data obtained by replicate RNA-seq experiments^[254] (see Section 2.5.2). In addition, we identified 107 genes that are significantly up-regulated in mouse liver^[255]. Median gene expression of CRM-associated genes increases when more than six TFs are present (Figure 2.8B); however only CRMs with between six and nine TFs are significantly enriched for the 107 genes significantly up-regulated in mouse liver cells (Figure 2.8A) (Fisher's Exact Test $P < 0.01$)^[255]. Strikingly, the peak of enrichment for tissue-specific genes at seven TFs coincides precisely with the peak in CNC enrichment at seven TFs. The three-way correspondence between liver-specific gene expression, CRM

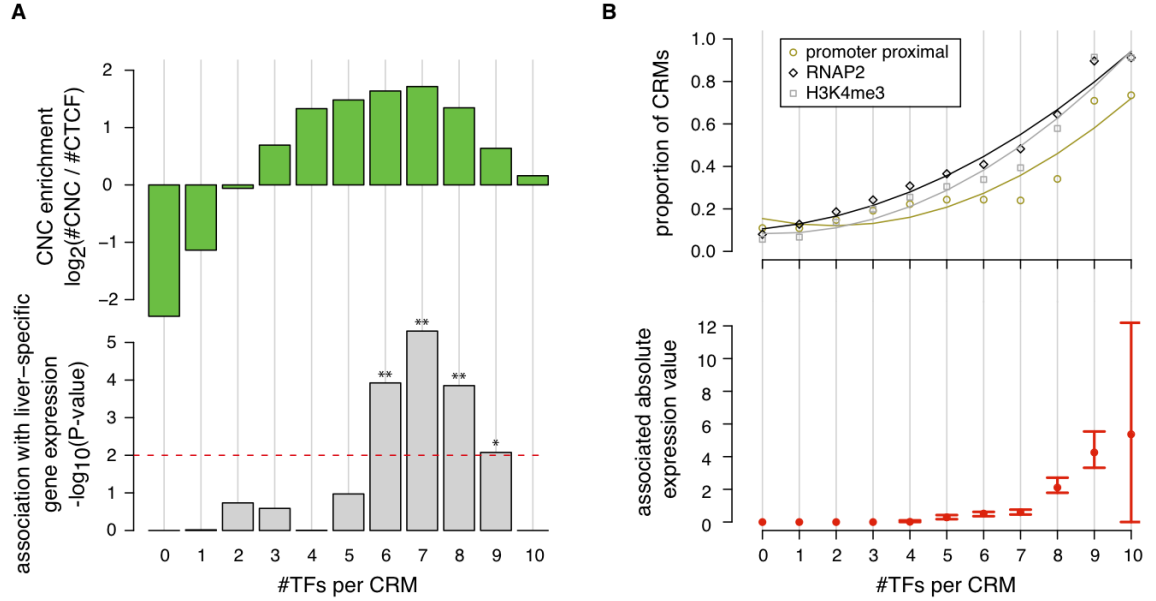


Figure 2.8: CNC sites are associated with liver-specific gene expression. (A) Ratio of CNC-containing CRMs versus those with CTCF (log fold change), for CRM classes with 0-10 TFs. Each class of CRMs was also tested for association with 107 genes significantly up-regulated in mouse liver cells (see Methods). The significance of the association (negative-log-transformed Fisher's exact test P-values) are indicated. $*P < 0.01$; $**P < 0.001$. The enrichment of CNC-containing CRMs reaches three-fold when seven TFs are present and coincides with highly significant enrichment for an association with liver-specific gene expression for the same class. (B) CRMs with high numbers of colocalising TFs are associated with increased promoter proximity (≤ 2.5 kb from an annotated TSS), and characteristics of transcriptional activity (RNAP2 and H3K4me3 ChIP-seq peaks). Likewise, the associated absolute gene expression value increases significantly with the number of bound TFs. Error bars indicate the 95% confidence interval of the median.

occupancy and CNC sites, provides further evidence of cohesin’s CTCF-independent transcriptional regulatory role at regions where multiple TFs assemble to effect tissue-specific expression.

2.3.4 Maximally occupied CRMs show similar properties to HOT regions

A total of 34 CRMs contain all ten assayed TFs. These regions have similar characteristics to recently identified high occupancy target (HOT) regions^[256,257,258]. These CRMs have high ChIP signal for all of the ten TFs, are highly enriched in promoter-proximal regions (Fisher’s Exact Test $P = 10^{-13}$) and are associated with genes having high absolute expression value – yet none of these are liver-specific genes. However, due to the low number of CRMs with all ten factors, the confidence intervals for the expression value are large.

The group of genes associated with these HOT regions includes *Polr2a*, which encodes the largest subunit of the RNA polymerase II complex, and *Ccnl1* a gene whose product (cyclin L1) participates in the regulation of the pre-mRNA splicing process^[259]. Another gene with a nearby HOT region, *Grif1*, encodes a transcription factor that binds to the promoter region of the glucocorticoid receptor (*Nr3c1*), a gene that is expressed in almost all cell types^[260].

These observations support the idea that HOT regions tend to occur near genes with house-keeping functions and consist of constitutively open chromatin^[257]. Although the number of TFs in this study is limited, and many of those that were included have tissue-specific functions, these are the first HOT regions to be identified in vertebrates with similar properties to those described in the model organisms *D. melanogaster* and *C. elegans*.

2.3.5 Cohesin intensity explains disparities between motif score and ChIP signal

The resolution of ChIP-seq data lends itself to the problem of finding TF binding site motifs as the actual binding site is typically within ≈ 50 bp of the peak summit. Nonetheless, the presence of the canonical motif usually explains only a fraction of the original ChIP-seq peaks^[261]. Although the proportion of peaks with a motif match is dependent on the chosen score threshold, some ChIP-positive sequence regions have no recognisable similarity to the canonical motif^[262,263]. Furthermore, quantitative TF binding, as measured by either ChIP-chip or ChIP-seq enrichment, is only weakly correlated with motif strength, as measured by the PWM log-odds score^[19,264].

In order to investigate these phenomena, we asked whether there was an unexpected correlation between a given factor's motif score and another factor's ChIP signal, within our identified CRMs. Briefly, for each sequence-specific factor, we first determined the PWM score of the best motif match within each corresponding peak. We then compared this motif score to the ChIP signal of all other datasets within CRMs containing that peak. Similarly, motif score correlations were calculated for the occupancy count (i.e. the number of distinct TFs present) and the distance to the nearest canonical TSS (Figure 2.9A).

As expected due to their roles at the core promoter, high motif scores for both GABPA and E2F4 are most associated with H3K4me3 ChIP signal and tend to occur near to annotated TSSs^[265]. In addition, CRM occupancy count is anti-correlated with motif scores of all factors except E2F4, indicating that when TFs occur in the absence of other potential binding partners, their binding is more likely to coincide with a high-scoring motif match. However, for only four out of the eleven sequence-specific factors we tested, the factor's motif score is most strongly correlated with its own ChIP signal. In fact, the strength of motif score for four different factors (ONECUT1, FOXA1, FOXA2 and HNF4A) is most strongly associated with HNF4A ChIP signal.

Interestingly, cohesin ChIP signal is also anti-correlated with motif scores of all assayed factors except CTCF (Spearman's $\rho = 0.11$) and E2F4 (Spearman's $\rho = 0.13$); in other words, stronger cohesin binding is associated with lower-quality motif matches for co-bound TFs. The two exceptions to this rule, i.e. positive correlations with E2F4

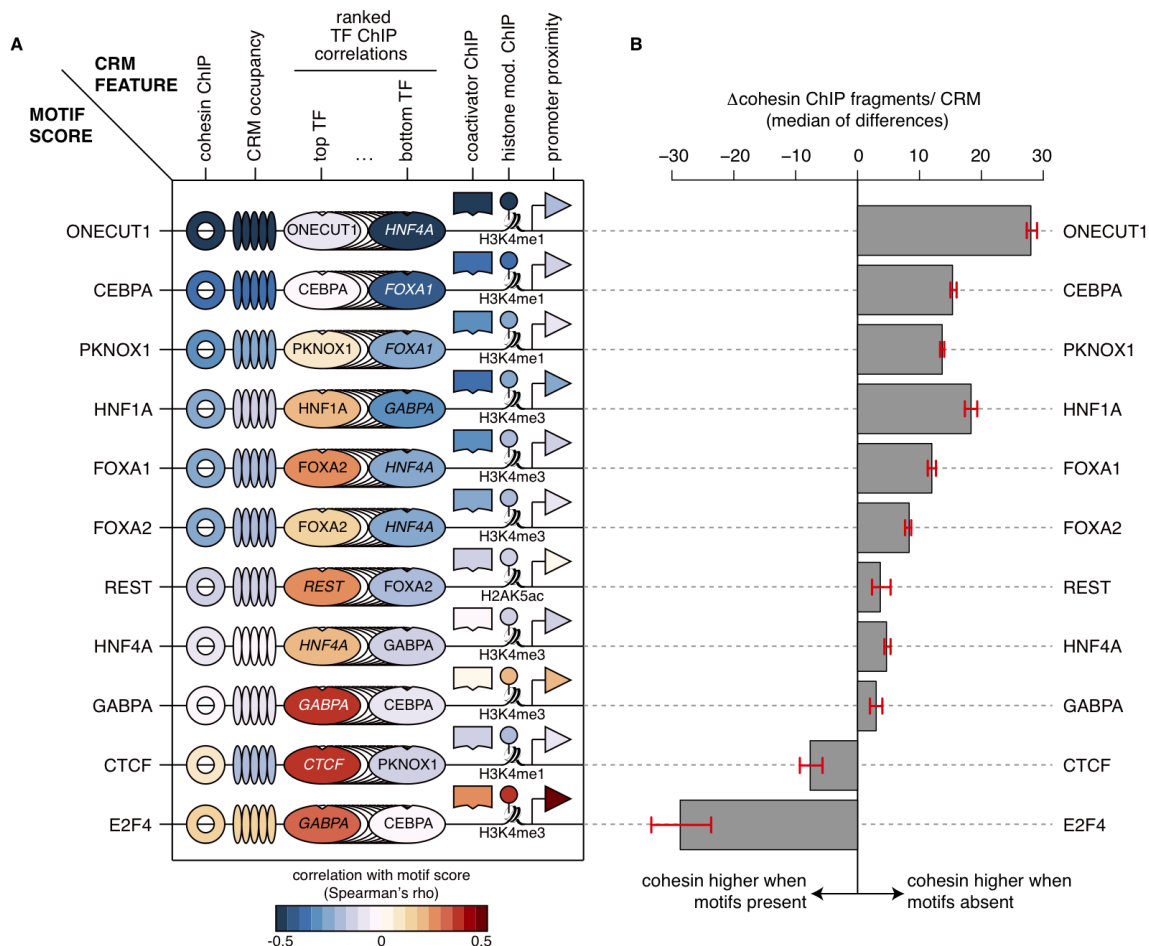


Figure 2.9: Cohesin ChIP signal is significantly associated with TF motif score. (A) Cartoon heatmap representation of correlations between each sequence-specific factor's motif score and the ChIP signal of all available ChIP-seq datasets. Correlations with CRM occupancy (number of distinct TFs present) and promoter proximity (distance to the nearest canonical TSS) are also shown. For each factor, the motif score correlation was calculated on the set of CRMs that contained a ChIP-seq peak for the same factor. Correlations with cohesin and coactivator ChIP signal were averaged over subunits (RAD21, STAG1, STAG2) and family members (EP300, CREBBP) respectively. Heatmap rows were ordered by increasing correlation with cohesin ChIP signal (from top to bottom). As a visual summary, only the highest- and lowest-ranking correlations involving TFs are shown. (B) Increased cohesin ChIP signal at TF binding events without motifs. For each sequence-specific factor, the number of cohesin ChIP fragments within CRMs without high-scoring motifs was compared to that of CRMs with motifs. 95% confidence intervals shown are based on a normal approximation of the HodgesLehmann estimate (median of all possible differences).

and CTCF, are unsurprising since strong E2F4 motifs and binding are found in highly occupied CRMs and CTCF binding has previously been shown to correlate well with motif quality and there is evidence that CTCF recruits cohesin to sites where they co-occur^[160]. For all other factors, stronger cohesin ChIP signals are associated with lower motif scores, particularly for the ONECUT1 motif (Spearman's $\rho = -0.5$) and CEBPA motif (Spearman's $\rho = -0.38$).

We also compared levels of cohesin ChIP signal between explicit groups of CRMs: those with, and those without high-scoring motifs according to a minimum PWM score threshold. For all sequence-specific factors except CTCF and E2F4, we observe higher levels of cohesin in the absence of high-scoring motifs (Figure 2.9B).

To determine whether cohesin presence could help to explain the discrepancy between TF ChIP signal and motif score, we trained logistic regression classifiers to predict the presence of high-scoring motifs for each sequence-specific factor, with and without cohesin ChIP signal information. For ONECUT1, CEBPA, HNF1A, PKNOX1, FOXA1, FOXA2, REST and E2F4, cohesin ChIP information markedly improved the performance of the classifier. For GABPA, HNF4A and CTCF there is minimal improvement in performance with the inclusion of cohesin in the model (Figure 2.10). These results suggest that cohesin presence is able to partially decouple ChIP signal from motif score for a significant number of TFs including those that are often found at enhancer elements.

2.3.6 ONECUT1 ChIP signal is reduced at weak motifs in heterozygous *Rad21*^{+/-} mouse liver cells

In order to determine whether cohesin plays an active role in the binding of TFs to their target sequences, particularly in the absence of high-scoring motifs, we used liver tissue from mice with only one functional allele of the *Rad21* gene. Homozygous knockouts of *Rad21* are lethal early in embryogenesis suggesting that at least one wild type *Rad21* allele is essential for normal development in mammals. Although heterozygous *Rad21*^{+/-} mice are viable, they possess a number of defects including hypersensitivity to ionising radiation and impaired DNA repair capacity^[210]. To confirm that the level of cohesin binding is reduced and to determine whether TF binding is consequently affected, we

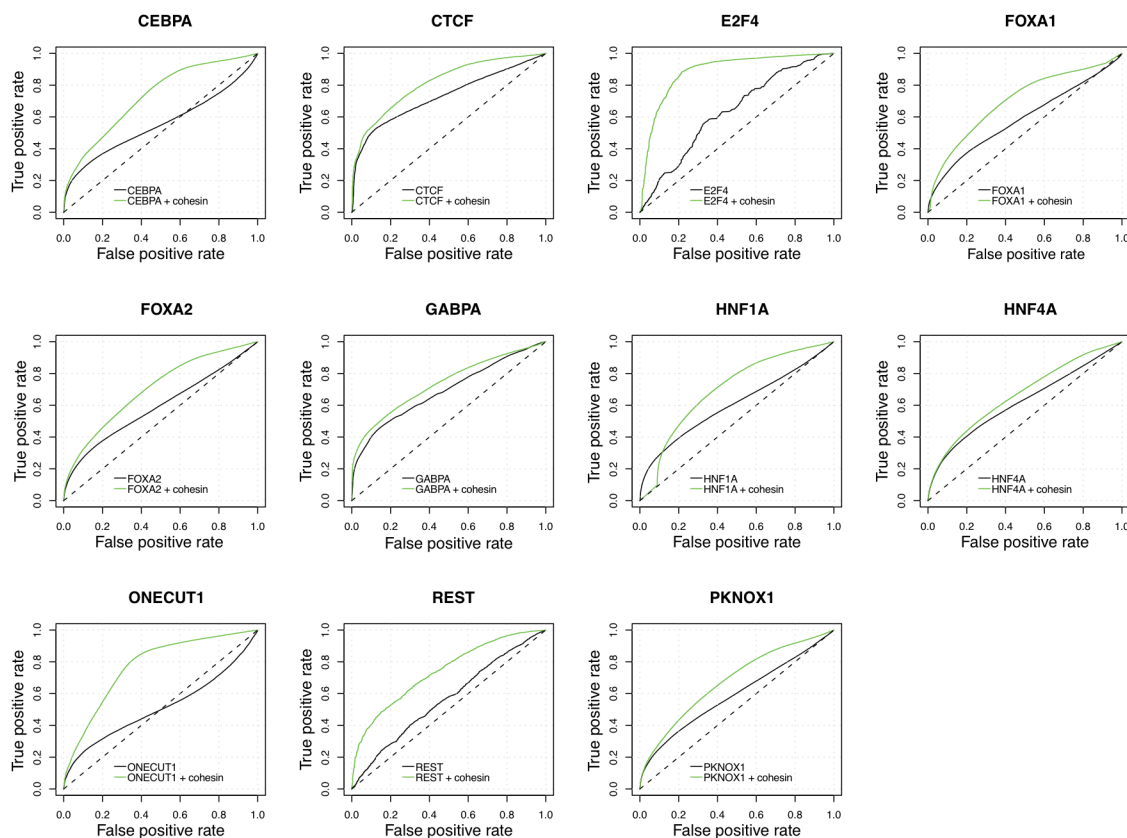


Figure 2.10: Performance results of all motif classifiers. Ten-fold cross-validation performance results of classifiers trained to predict the presence of high-scoring motif matches using the same factor's ChIP signal with (solid green curve) and without (solid black curve) cohesin ChIP signal information (see inset Legend). Positive deviation from the dashed black line indicates performance exceeding that expected by a random classifier (i.e. area under the curve, $AUC = 0.5$).

mapped RAD21, ONECUT1, CEBPA and HNF4A in heterozygous *Rad21*^{+/-} mouse liver cells using ChIP-seq.

The total number of binding events for all TFs is reduced in heterozygous *Rad21*^{+/-} cells (ONECUT1 45%, CEBPA 63%, HNF4A 18%) and, as expected, the reduction is most severe for RAD21 (14%). 78,625 CRMs lose RAD21 binding according to the absence of an overlapping peak in heterozygous *Rad21*^{+/-} cells. We focus the remainder of our analysis on these sites. In terms of peak loss, CRMs without high-scoring motifs are enriched for binding events lost in heterozygous *Rad21*^{+/-} cells (responsive binding events) for all three assayed TFs (CEBPA odds ratio: 2.6, HNF4A odds ratio: 5.3, ONECUT1 odds ratio: 5.6; Fisher's Exact Test $P < 10^{-15}$). We also performed statistically robust differential binding analysis on replicate ONECUT1 and CEBPA ChIP-seq data in order to determine ChIP signal differences between wild type and heterozygous *Rad21*^{+/-} cells (see Section 2.5.2). Similar to the results from the peak-level analysis, CRMs exhibiting significantly reduced ONECUT1 ChIP signal in mutant cells are enriched for ONECUT1 peaks without high-scoring motifs (Fisher's Exact Test $P = 10^{-4}$; Figure 2.11B). However differential binding analysis for CEBPA revealed no significant differences in ChIP signal.

Interestingly, promoter-proximal CRMs tend to be associated with both reduced ONECUT1 motif scores (Figure 2.9A) and *Rad21*^{+/-} responsive ONECUT1 binding events (Fisher's Exact Test $P < 10^{-15}$). This suggests that cohesin may help to stabilise the binding of ONECUT1 near promoters in particular. One such region is shown in Figure 2.11A overlapping the *BC031353* promoter, where all but one of the remaining ONECUT1-containing CRMs displayed retain ONECUT1 binding in heterozygous *Rad21*^{+/-} cells (resistant binding events). Note that a high-scoring motif is absent from the *Rad21*^{+/-} responsive ONECUT1 binding event overlapping the TSS, although the effect on *BC031353* expression was not assessed.

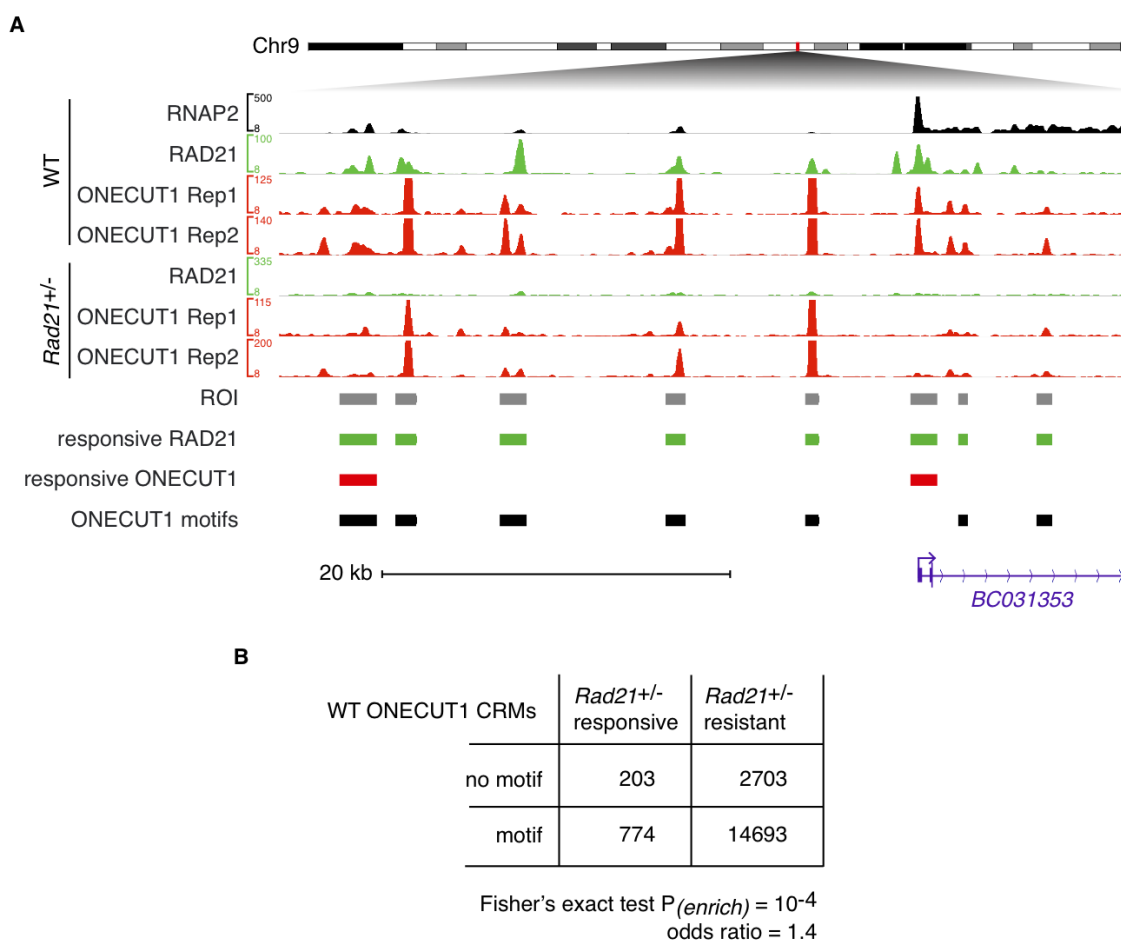


Figure 2.11: ONECUT1 ChIP-seq in heterozygous *Rad21*^{+/-} mouse liver cells reveals preferential loss of TF binding events at sites without motifs. (A) Sample region near the *BC031353* gene showing overall reduction in RAD21 ChIP signal in heterozygous *Rad21*^{+/-} cells (responsive RAD21) and associated significant reduction in ONECUT1 ChIP signal within two CRMs (responsive ONECUT1). The ONECUT1 binding event overlapping the TSS contains no ONECUT1 motif. (B) WT ONECUT1 CRMs without motifs show a preferential decrease in ChIP signal ($FDR < 0.1$) in heterozygous *Rad21*^{+/-} mouse liver cells (Fisher's Exact Test $P = 10^{-4}$). Regions of interest (ROI) are those CRMs where RAD21 binding was ablated in heterozygous *Rad21*^{+/-} mouse liver cells (responsive RAD21).

2.3.7 Mirrored binding of CTCF near transcription start sites and cohesin-bound enhancers is associated with elevated expression levels

Cohesin has been shown to be crucial for two distinct types of chromatin interactions: (i) looping between individual CTCF binding events^[163,164,165,166] and, (ii) interactions between promoters and CNC-containing enhancers^[167,169,170]. Reports of long-range chromatin looping mediated by CTCF have suggested that CTCF may influence transcription by facilitating enhancer-promoter interactions^[266]. In this model, interactions between promoter-proximal and distal CTCF binding events connect enhancers to their target genes by looping out the intervening DNA, thereby reducing the effective distance and increasing the probability of interactions between linearly distant genomic regulatory regions.

We therefore searched for genes where this configuration has the potential to occur i.e. genes with CTCF/cohesin binding events both nearby the TSS and proximal to their associated enhancers (Figure 2.12B). If these consistent binding patterns have biological relevance, we expect their presence to be associated with increased expression levels of the corresponding genes. To test this, we first compiled a list of putative liver-specific enhancers, defined as CRMs more than 5 kb from their nearest canonical TSS that possess (i) a CNC site, (ii) the liver master regulator HNF4A, (iii) the EP300 enhancer marker, and (iv) the histone signature H3K4me1, but (v) not H3K4me3. In the absence of high resolution chromatin conformation data that could be used to infer enhancer-promoter pairs, we then assigned each identified enhancer to the nearest gene based on distance to the TSS, such that each enhancer is assigned to only one gene.

Of the 5,364 genes with nearby enhancers as defined above, 532 genes have CTCF binding events both 2.5 kb from the TSS and nearby their enhancers (≤ 2.5 kb). Indeed, these genes have significantly greater expression values than genes with CTCF binding near the TSS but not nearby their enhancers, or vice-versa (Figure 2.12A, Mann-Whitney U Test $P < 10^{-3}$).

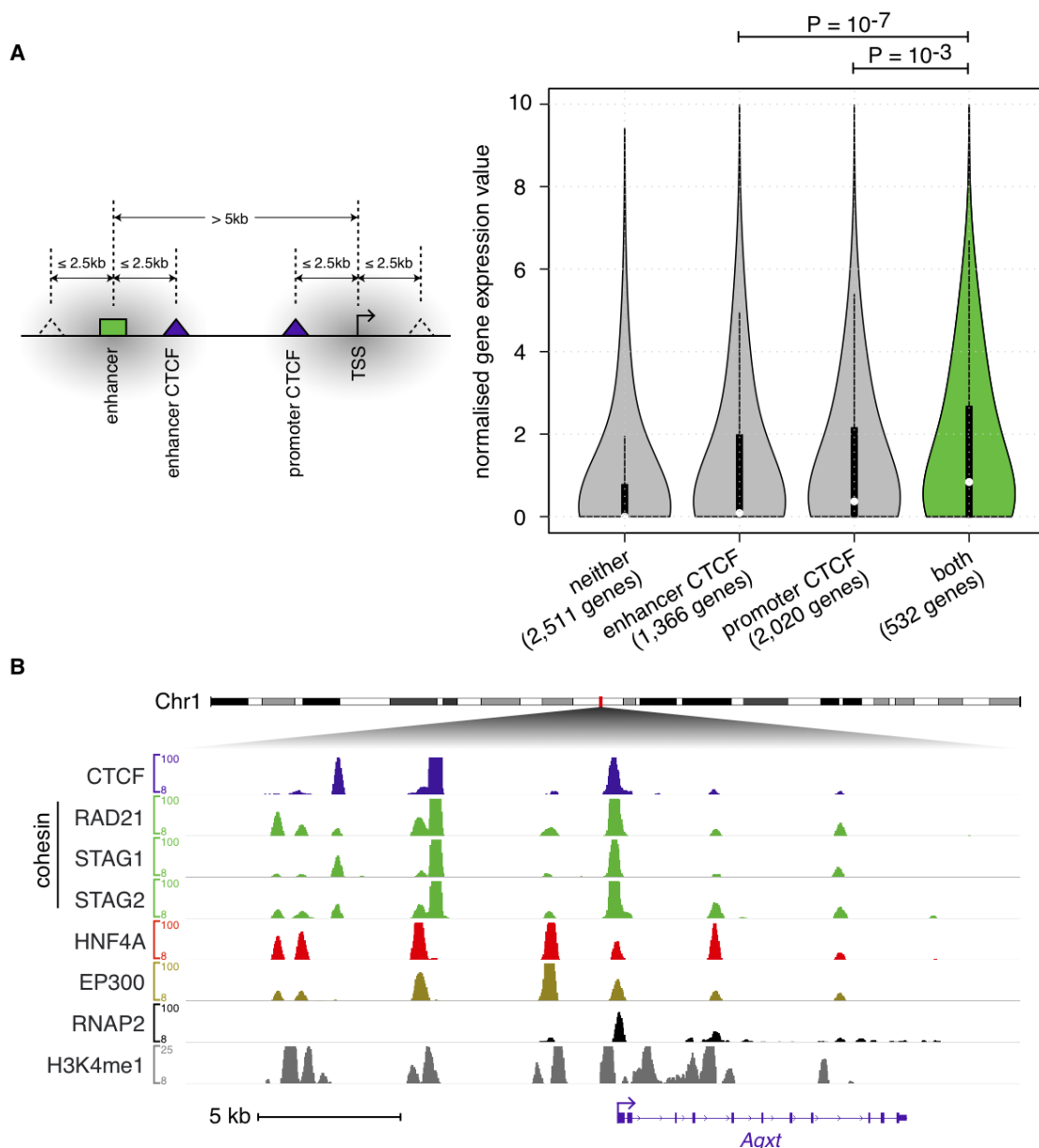


Figure 2.12: Simultaneous CTCF binding within promoters and nearby enhancers is associated with elevated expression levels. (A) Violin plots showing gene expression distributions. Genes with CTCF binding events both within their promoters and nearby their associated enhancers, show significantly elevated expression levels over those of the other three indicated classes (Mann-Whitney U test $P < 10^{-3}$). (B) Sample region near the liver-expressed *Agxt* gene, where CTCF binds within the core promoter, as well as near putative upstream cohesin-bound enhancers. Note that while CTCF is absent from the enhancers (CNC), it co-binds with HNF4A and EP300 within the *Agxt* promoter.

2.4 Discussion

Cohesin has multiple, vital functions in mammalian cells including well-established roles in sister chromatid segregation in mitosis and meiosis. Recent results have implicated cohesin in the regulation of gene expression. Because cohesin has no known DNA binding domain, the mechanism of this transcriptional regulation is assumed to arise from cohesin's ability to stabilise higher-order chromatin structure through interactions with chromatin organisation proteins such as CTCF^[163,165]. We and others have shown that cohesin plays a role in tissue-specific transcriptional regulation and that this role is at least partially characterised by CTCF-independent cohesin localisation with master regulators in several tissues. To better understand cohesin's contribution to gene regulation, we collected genome-wide localisation data of 10 TFs, several histone modifications and other functional DNA-protein interactions in primary mouse liver. These data provide a comprehensive map of cohesin's two known roles: one associated with CTCF and another CTCF-independent role in tissue-specific transcriptional regulation. We show that these roles are functionally similar across multiple tissues, by demonstrating that cohesin's presence at binding events of liver-specific TFs mirrors its localisation with ES cell-specific factors.

To further characterise cohesin's tissue-specific regulatory role, we focussed on the properties of cohesin-non-CTCF (CNC) sites. By clustering the binding patterns of sequence-specific factors within CRMs and ranking these clusters by the fraction that overlap CNC sites, we demonstrate that CNC sites occur preferentially at CRMs containing multiple TFs and are less likely to be found at CRMs with singleton binding events that represent the majority of regions bound by any given factor. The class of CRMs with the most TFs is also highly enriched for binding events that are persistent across hundreds of millions of years of evolution^[19], suggesting that the conservation of these events – and possibly those of other tissue-specific TFs – is attributable to their highly bound state and putative functional context. However, only 5% of the CRMs in the maximally occupied cluster have a deeply shared binding event (i.e. five-species CEBPA or three-species HNF4A). Thus, the colocalisation of a large number of TFs does not mean, *a priori*, that a binding event will be invariant over evolutionary time.

We observe a striking relationship between CNC enrichment and liver-specific gene expression for CRMs with submaximal numbers of distinct TFs bound. In particular,

CRMs with between six and eight of our assayed TFs are more than twice as likely to possess CNC sites than CTCF, and are also the most highly enriched for liver-specific gene expression. Although both CNC enrichment and association with liver-specific expression peaks when CRMs have seven TFs, we find no evidence in the mouse ES cell data that this is a (just right) “goldilocks” number of TFs. However our results in mouse liver are consistent with previous research in other species showing that regions with low-to-moderate numbers of transcription factors are most significantly enriched for annotated enhancers and signs of active transcriptional regulation^[258]. Therefore, although the distribution of tissue-specific and ubiquitous factors is different in the ES cell experiments, this does not rule out the attractive hypothesis that a specific and relatively small number of TFs binding together and stabilised by cohesin is a fundamental characteristic of mammalian tissue-specific gene regulation.

Intriguingly, the most highly occupied CRMs containing all ten of our assayed TFs are neither associated with liver-specific genes nor CNC enrichment. Instead these regions seem to be nearby constitutively active genes and have characteristics that are similar to recently described HOT regions^[256,257,258]. To our knowledge these are the first HOT regions to be described in vertebrates.

The DNA sequence preferences of TFs are typically described using position weight matrices (PWMs), and referred to as binding site motifs. These motifs remain a challenge to discover computationally despite the large number of de novo motif discovery algorithms that have been developed to infer these sequence preferences^[267]. Previous results have demonstrated that while ChIP-seq data is useful for identifying the specific regions of the genome bound by a given TF, there remains a subset of binding events with either weak or non-existent motif matches. This lack of a clear relationship between DNA sequence content and TF recruitment has been described as a result of indirect or cooperative binding and recent approaches tailored specifically to ChIP-seq data have subsequently focussed on finding these candidate co-factors^[268].

Using both computational and experimental methods, we show that the presence of cohesin likely explains the inverse relationship between ChIP signal and motif score observed for a number of our assayed factors. These TFs bind to stronger motifs in the absence of cohesin. Stated alternatively, we observe higher levels of cohesin in the absence of high-scoring motifs. These results suggest that cohesin enables TFs to bind

to sub-optimal motif sequences either by stabilising large protein-DNA complexes at highly occupied CRMs or by inducing binding through specific chromatin contortions. Importantly, we show that computational classifiers trained to predict high-scoring motif occurrence exhibit markedly improved performance when cohesin is incorporated into the model. Furthermore, using ChIP-seq in the livers of a *Rad21*-cohesin haploinsufficient mouse model, we show that heterozygous loss of *Rad21* results in the loss of 86% of RAD21 binding events found in the wild type. This is accompanied by a reduction in ChIP-seq peak numbers for ONECUT1, CEBPA and HNF4A that disproportionately affects binding events without high-scoring motifs for these TFs. Similarly, we find that sites both without RAD21 peaks and showing a significant loss of ONECUT1 ChIP signal in heterozygous *Rad21*^{+/-} cells, are also significantly depleted for high-scoring ONECUT1 motifs. Taken together with our observations in wild type cells that cohesin is more abundant at highly occupied CRMs and at those without high-scoring motifs, these results point towards a role for cohesin in stabilising the binding of TFs to *cis*-regulatory sequences particularly near promoters. Alternatively expression level differences of the TFs themselves caused by the loss of cohesin may contribute to the overall reduction in binding events observed in heterozygous *Rad21*^{+/-} cells.

Promoter regions are important sites of TF binding where multiple regulatory signals are integrated to coordinate cell-type-specific expression programs. Both CTCF and cohesin have been shown to modulate chromatin structure in order to enable promoter-proximal factors to respond to signals from distant *cis*-regulatory elements, such as enhancers. However our results indicate that the majority of highly occupied CRMs, which show typical characteristics of enhancers, possess cohesin in the absence of CTCF (CNC sites). An attractive hypothesis is that CTCF may set up indirect chromatin interactions as the primary step towards enabling enhancer-promoter communications^[266]. We tested whether the dual presence of CTCF binding events both nearby TSSs and their corresponding enhancers is associated with increased expression levels. Using this simple approach, we observe genome-wide patterns that support the model that concerted CTCF binding to linearly distant regulatory regions is associated with significantly elevated expression levels. Further investigations using 3C-based chromatin conformation assays would be needed to determine whether these patterns are indeed associated with functional chromatin looping interactions between enhancers and promoters.

2.5 Methods

2.5.1 Experimental methods

2.5.1.1 ChIP sequencing

The experiments described in this paragraph were performed by Dominic Schmidt and Michael D. Wilson. ChIP experiments were performed with wild type primary mouse (C57BL/6 and/or C57BL/6xA/J) liver tissue and antibodies against CTCF (2 replicates, 2 individuals; antibody: upstate, 07729), STAG1 (3 replicates, 2 individuals; antibody: abcam, ab4457), STAG2 (singlicate; antibody: abcam, 4464), RAD21 (singlicate; antibody: abcam, ab992), CEBPA (6 replicates, 2 individuals; antibody: santa cruz, sc9314), HNF4A (2 replicates, 1 individual; antibody: aviva systems biology, ARP31946), FOXA1 (2 replicates, 2 individuals; antibody: abcam, ab5089), FOXA2 (4 replicates, 2 individuals; antibody: santa cruz, sc6554), ONECUT1 (6 replicates, 2 individuals; antibody: santa cruz, sc13050), HNF1A (3 replicates, 1 individual; antibody: santa cruz, sc6547), PKNOX1 (singlicate; antibody: santa cruz, sc6245), REST (singlicate; antibody: santa cruz, sc25398), GABPA (2 replicates, 1 individual; antibody: santa cruz, sc22810), E2F4 (singlicate; antibody: santa cruz, sc1082), EP300 (2 replicates, 2 individuals; antibody: santa cruz, sc585), CREBBP (singlicate; antibody: santa cruz, sc369), RNAP2 (2 replicates, 2 individuals; antibody: abcam, ab5408), H3K4me1 (singlicate; antibody: abcam, ab8895), H3K4me3 (singlicate; antibody: abcam, ab8580), H3K36me3 (singlicate; antibody: abcam, ab9050), H3K79me2 (singlicate; antibody: abcam, ab3594) and H2AK5ac (singlicate; antibody: abcam, 1764) as recently described (Schmidt et al. 2009). Briefly, the immunoprecipitated DNA was end-repaired, A-tailed, ligated to the sequencing adapters, amplified by 18 cycles of PCR, and size selected (200-300 bp) followed by single-end sequencing on an Illumina Genome Analyzer according to the manufacturer's recommendations.

The experiments described in this paragraph were performed by Stephen Watt and Huiling Xu. ChIP experiments were performed with heterozygous *Rad21*^{+/-} primary mouse liver tissue and antibodies against RAD21 (2 replicates, 2 individuals; antibody: abcam, ab992), CEBPA (2 replicates, 2 individuals; antibody: santa cruz, sc9314), HNF4A (2 replicates, 2 individuals; antibody: aviva systems biology, ARP31946), ONECUT1 (2 replicates, 2 individuals; antibody: santa cruz, sc13050) as above.

2.5.2 Computational methods

2.5.2.1 Read mapping and peak calling

All ChIP sequencing reads from each replicate were aligned to the mouse reference genome assembly (NCBI37/mm9) using BWA^[204] with default parameters. After pooling replicate data for each factor/histone-modification, the reads were then filtered to remove low quality mappings (Phred-scaled mapping quality < 10), multiple reads mapping to the same genomic location and strand, as well as those mapping to the mitochondrial genome. Peaks were then called on all datasets using matched input data and a dynamic programming algorithm (SWEml) with $-R\ 0.005$ as recently described^[167].

2.5.2.2 Cohesin-non-CTCF site definition and peak clustering

Firstly, overlapping ChIP-seq peaks for CTCF and the cohesin subunits (STAG1, STAG2, RAD21) were merged to form a set of disjoint genomic regions. Our definition of cohesin-non-CTCF (CNC) sites required the presence of at least one cohesin subunit peak and the absence of CTCF. In order to obtain a high-confidence set of CNC sites, in the absence of significant CTCF ChIP enrichment that may have escaped peak-detection, we required that these sites also satisfied the following criterion: $\log((\text{norm_CTCF_ChIP})/(\text{norm_Input})) < 0.68$. This cut-off corresponds to the fifth percentile of ChIP enrichment scores within CTCF peaks.

Overlapping peak regions of the sequence-specific factors (CTCF, CEBPA, HNF4A, FOXA1, FOXA2, ONECUT1, HNF1A, PKNOX1, REST, GABPA, E2F4), as well as cohesin (STAG1, STAG2, RAD21), CNC sites and the coactivators EP300/CREBBP, were merged to define putative *cis*-regulatory modules (CRMs)^[13]. A single-linkage clustering approach was used, where a peak overlap of ≥ 1 bp with at least one other peak within a CRM is sufficient for membership within the CRM. The presence or absence of a particular histone-modification (H3K4me1, H3K4me3, H3K36me3, H3K79me2, H2AK5ac) or RNAP2 binding within a CRM was then determined *pos hoc* by satisfaction of either of the following criteria (i) the presence of an overlapping peak, or (ii) ChIP enrichment within the entire CRM region of at least three-fold, where the number of ChIP reads overlapping the CRM ≥ 8 .

2.5.2.3 Motif analysis and selection

We used MEME^[211] and NestedMica^[212] to perform *de novo* motif discovery for each sequence-specific factor using peak regions with the top one thousand scores. In each case, 50 bp of DNA sequence surrounding the SWEmbl summit was used to find five frequently occurring sequence motifs up to 25 bp in length (MEME parameters: `-nmotifs 5 -maxw 25 -revcomp`; NestedMica parameters: `-numMotifs 5 -minLength 5 -maxLength 25 -revComp -backgroundOrder 1 -backgroundClasses 4`). We scanned all bound (positive) regions for each factor with PWMs for all five NestedMica motifs as well as the top-scoring MEME motif, to determine the score of the best motif match in each case. We repeated this using equally sized unbound (negative) regions, which were randomly sampled from the repeat- and exon-masked genome. The optimal motif for each factor, which was retained for further analysis, was defined as that best able to discriminate between positive and negative regions according to the AUC (area under ROC curve) performance measure (Figure 2.13). All *de novo* motifs closely match corresponding previously published PWMs in TRANSFAC^[213] where available.

2.5.2.4 Mouse ES cell data analysis

Publicly-available ChIP-seq datasets from mouse ES cells were downloaded, re-processed and analysed using a similar procedure to that described above: CTCF, MYC, ESRRB, KLF4, MYCN, SMAD1, STAT3, TCF7L1, ZFX, EP300, SUZ12^[241], NANOG, POU5F1, SOX2, H3K79me2^[248], RNAP2^[249], NIPBL, SMC1A, SMC3, MED1, MED12^[169].

2.5.2.5 CRM clustering and analysis

To restrict our analysis to sites with possible patterns of combinatorial TF binding, we filtered our data to retain only CRMs containing a binding event of at least one sequence-specific factor. We used two independent methods (K-means and AutoClass) to group CRMs into similar clusters.

- (I) We used K-means to group CRMs into K similar clusters based on the binary presence/absence of the eleven sequence-specific factors within each CRM. In












	Motif Logo	Motif Score Cut-off				Motif Logo	Motif Score Cut-off		
		PWM Score	FDR	Peaks with Motif			PWM Score	FDR	Peaks with Motif
CEBPA		1.76	0.40	92%	HNF1A		2.54	0.40	39%
		2.12	0.35	86%			2.80	0.35	34%
		2.33	0.30	82%			3.05	0.30	30%
		2.60	0.25	74%			3.26	0.25	26%
		2.88	0.20	65%			3.48	0.20	23%
		3.18	0.15	48%			3.78	0.15	19%
		3.57	0.10	33%			4.11	0.10	14%
CTCF		4.23	0.05	9%	HNF4A		4.62	0.05	8%
		PWM Score	FDR	Peaks with Motif			PWM Score	FDR	Peaks with Motif
		0.85	0.40	99%			2.56	0.40	69%
		1.19	0.35	98%			2.87	0.35	55%
		1.49	0.30	97%			3.14	0.30	41%
		1.77	0.25	95%			3.45	0.25	28%
		2.04	0.20	93%			3.83	0.20	14%
E2F4		2.33	0.15	90%	HNF6		4.23	0.15	5%
		2.69	0.10	86%			4.62	0.10	1%
		3.22	0.05	78%			4.67	0.05	0%
		PWM Score	FDR	Peaks with Motif			PWM Score	FDR	Peaks with Motif
		1.26	0.40	98%			1.75	0.40	91%
		1.27	0.35	96%			1.97	0.35	86%
		1.43	0.30	94%			2.14	0.30	82%
FOXA1		1.74	0.25	92%	NRSF/REST		2.30	0.25	79%
		1.74	0.20	92%			2.50	0.20	74%
		1.75	0.15	89%			2.81	0.15	68%
		2.26	0.10	87%			3.58	0.10	49%
		3.12	0.05	75%			4.18	0.05	18%
		PWM Score	FDR	Peaks with Motif			PWM Score	FDR	Peaks with Motif
		2.55	0.40	71%			1.03	0.40	94%
FOXA2		2.87	0.35	58%	PREP1		1.37	0.35	90%
		3.20	0.30	43%			1.69	0.30	86%
		3.52	0.25	28%			1.97	0.25	82%
		3.88	0.20	15%			2.31	0.20	75%
		4.26	0.15	6%			2.64	0.15	68%
		5.11	0.10	0%			3.07	0.10	59%
		5.11	0.05	0%			3.59	0.05	49%
GABPA		PWM Score	FDR	Peaks with Motif			PWM Score	FDR	Peaks with Motif
		2.44	0.40	80%			2.09	0.40	77%
		2.77	0.35	68%			2.41	0.35	65%
		3.11	0.30	53%			2.74	0.30	52%
		3.45	0.25	36%			3.04	0.25	41%
		3.79	0.20	21%			3.36	0.20	31%
		4.14	0.15	10%			3.72	0.15	21%
		4.50	0.10	3%			4.11	0.10	13%
		5.18	0.05	0%			4.69	0.05	6%
		PWM Score	FDR	Peaks with Motif					
		1.80	0.40	89%					
		2.06	0.35	82%					
		2.33	0.30	73%					
		2.58	0.25	64%					
		2.76	0.20	55%					
		3.05	0.15	42%					
		3.41	0.10	28%					
		3.89	0.05	12%					

Figure 2.13: Motifs and motif statistics. The optimal motifs for each sequence-specific factor obtained using ChIP-seq peak sequences as input to MEME/NestedMica. We scanned all bound (positive) regions for each factor with the corresponding PWMs to determine the score of the best motif match in each case. We repeated this using equally sized unbound (negative) regions, which were randomly sampled from the repeat- and exon-masked genome. The PWM score cut-off and percentage of (positive) peaks with the motif, are reported for false discovery rates (FDR) of 0.4, 0.35, 0.3, 0.25, 0.20, 0.15, 0.1, 0.05, where $FDR = FP/(FP + TP)$. Motif score cut-offs corresponding to $FDR = 0.4$ were chosen to determine motif presence/absence.

order to choose an appropriate value for K , we ran the clustering algorithm on a random subset of 20,000 CRMs and determined the median within-cluster sum of squares (WCSS) over 10 replicates of each value of K in the range [2-50]. The WCSS tends to decrease as the number of clusters K increases, but the decrease flattens slightly for values of K near 10 (Figure 2.14). We used this “elbow” method to choose a value of $K = 10$ when running the algorithm on the entire dataset.

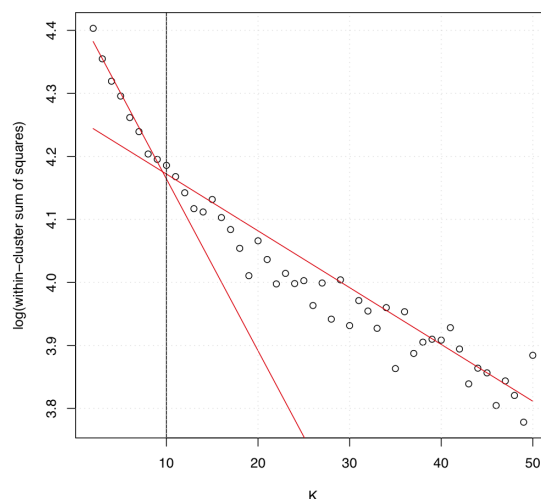


Figure 2.14: Choice of K for K-means CRM clustering analysis. We ran the K-means clustering algorithm on a random subset of 20,000 CRMs and determined the median within-cluster sum of squares (WCSS) over 10 replicates of each value of K in the range [2-50]. The red lines indicate linear regression results using WCSS values in the range [2-10] and [40-50]. The decrease in WCSS is less pronounced for values of $K > 10$. We therefore used $K = 10$ when clustering the entire dataset.

- (II) We used *AutoClass*^[269] to group CRMs into similar classes based on the normalised ChIP enrichment of the eleven sequence-specific factors within each CRM. *AutoClass* uses a Bayesian probabilistic approach to automatically optimise the properties of each class (as well as the number of classes) to achieve the best separation. An advantage of this “fuzzy” clustering approach, not provided by other traditional clustering methods such as K-means, is the availability of a

measure (posterior probability) to assess the confidence that each CRM belongs to its assigned class. The *AutoClass* C command-line program was used with the following primary settings: (i) data model: `single_normal_cn` (factor ChIP enrichments follow conditionally independent normal variables); (ii) convergence criterion: `converge_3` (most stringent); (iii) initial values for the number of class: 2, 3, 5, 7, 10, 15, 25, 35, 45, 55, 65, 75, 85, 95, 105; (iv) absolute error of input data: 10% for all factors. We filtered the CRMs in the resulting classification to retain only those with posterior probabilities ≥ 0.5 and combined classes with high correlation (Spearman's $\rho > 0.9$) between their median ChIP enrichment profiles.

Other CRM attributes such as the ChIP peak presence/absence of other factors/-histone-modifications not used in the original clustering were added to aid visualisation of the clustering results. Gene annotation information from Ensembl version 60^[270] was used to add genomic localisation information for each CRM, where “Promoter” was defined as occurring ≤ 2.5 kb from an annotated TSS, “Exon” corresponds to overlap with an exon but not a promoter, “Intron” corresponds to overlap with a gene but neither an exon nor a promoter, and “Distal” for localisation elsewhere. Gene annotation information for pseudogenes was ignored throughout the analysis.

2.5.2.6 Expression analysis

The analysis described in this paragraph was performed by Ângela Gonçalves. We used previously published RNA-seq data from mouse liver to obtain absolute expression estimates for all genes^[254]. Briefly, the raw reads were truncated to 35-mers and aligned to mouse transcript sequences (cDNA sequences from Ensembl version 60, NCBI37/mm9) using Bowtie version 0.12.7^[205] with default parameters. Normalised gene expression estimates were obtained using MMSEQ^[225] and summarised by taking the replicate mean.

We used a previously published dataset consisting of expression measurements from 40 diverse mouse tissues to determine sets of genes with liver-specific patterns of expression^[255]. The processed data (ArrayExpress accession: E-MTAB-25) was obtained from the Gene Expression Atlas^[271] where up-regulation in a particular tissue with respect to the remainder was assessed using a t-test and $P < 0.05$.

2.5.2.7 Motif presence prediction

For each sequence-specific factor, we trained logistic regression classifiers to predict the presence of high-scoring motif matches using the ChIP signals (estimated number of ChIP fragments overlapping a given CRM) of various factors. Models were trained using: (a) ChIP signal of the corresponding factor, and (b) both ChIP signal of the corresponding factor and ChIP signals of the cohesin subunits (RAD21, STAG1, STAG2). Motif score cut-offs corresponding to $FDR = 0.4$ were chosen to determine high-scoring motif match presence/absence. Ten-fold cross-validation was performed using CRMs containing a peak for the factor of interest, where 50% of these CRMs were randomly selected for the training set and the remaining 50% formed the test set.

2.5.2.8 Wild type versus heterozygous *Rad21*^{+/-} differential binding analysis

Read mapping and filtering for CEBPA and ONECUT1 was carried out as described above for both wild type and heterozygous *Rad21*^{+/-} ChIP-seq datasets, except reads for biological replicates were handled separately (technical replicates were pooled). The DiffBind package^[218] was used with default parameters to determine CRMs with significantly lower ChIP signal in heterozygous *Rad21*^{+/-} mouse liver cells versus wild type liver cells ($FDR_{threshold} = 0.1$).

Chapter 3

Cohesin-based chromatin interactions enable regulated gene expression within pre-existing architectural compartments

3.1 Summary

Chromosome conformation capture approaches have shown that interphase chromatin is partitioned into spatially segregated Mb-sized compartments and sub-Mb-sized topological domains. This compartmentalisation is thought to facilitate the matching of genes and regulatory elements, but its precise function and mechanistic basis remain unknown. Cohesin controls chromosome topology to facilitate DNA repair and chromosome segregation in cycling cells, and also associates with active enhancers, promoters and with CTCF to form long-range interactions important for gene regulation. Although these findings suggest an important role for cohesin in genome organisation, this has not been assessed on a global scale. Unexpectedly, we find that architectural compartments – a major feature of interphase chromatin organisation – are maintained in non-cycling mouse thymocytes after genetic depletion of cohesin *in vivo*. Cohesin is however required for specific long-range interactions within compartments where

cohesin-regulated genes reside. Cohesin depletion diminishes interactions between cohesin binding events, while alternative interactions between chromatin features associated with transcriptional activation and repression become more prominent, with corresponding changes in gene expression. Our findings indicate that cohesin-mediated long-range interactions facilitate discrete gene expression states within pre-existing chromosomal compartments.

This study is the result of a collaboration between Dr. Matthias Merckenschlager’s laboratory at the Medical Research Council Clinical Sciences Centre and Dr. Paul Flicek’s research group at the EMBL European Bioinformatics Institute. Dr. Vlad Seitan performed most of the experiments for this project and I carried out the computational analysis, except where otherwise specified. This chapter is based on a manuscript that has recently been accepted for publication in *Genome Research* and is adapted here with the publisher’s permission.

3.2 Introduction

The regulated transcription of mammalian genomes packaged into nuclei six orders of magnitude smaller than the length of chromosomal DNA requires a complex organisation. The underlying mechanisms are beginning to be explored in terms of the biophysical properties of the DNA polymer and associated chromatin^[99], as well as nuclear landmarks such as the nuclear lamina that provide genomic scaffolds^[110,272]. Of particular interest are functional interactions that operate within these constraints to facilitate long-range interactions between gene regulatory elements^[52].

Genome-scale chromosome conformation capture has shown that interphase chromatin is organised into Mb-scale compartments that tend to be “open”, gene-rich, highly transcribed and interactive (A) or “closed” (B), gene-poor, and less transcriptionally active^[22]. This architecture is thought to facilitate the regulation of gene expression by constraining the number of regulatory elements a given gene is likely to encounter to those that are co-located within the same compartment or domain^[85]. However it is unknown how compartments and domains are built, and how they contribute to the precise regulation of gene expression.

The cohesin protein complex is essential for genome integrity in cycling cells where it facilitates post-replicative DNA repair and sister chromatid cohesion by controlling chromosome topology^[148,273]. In addition, cohesin contributes to the regulation of gene expression by mechanisms thought to involve long-range interactions between its binding sites at regulatory elements associated with CTCF^[142,159,160,161] or with active promoters and enhancers^[167,169,238,243,274]. Taken together these properties suggest a role for cohesin in genome organisation.

Elucidating the global contribution of cohesin to the organisation of the genome remains a challenge, not least because the depletion of cohesin from proliferating cells interferes with DNA replication, DNA repair and chromosome segregation^[148]. Here we use a genetic approach^[170] to define the contribution of cohesin to the organisation of the genome and the regulation of gene expression in non-cycling cells in vivo. Previous work has shown that cohesin-deficient thymocytes differentiate with reduced efficiency as a result of defective chromatin architecture, transcription and impaired rearrangement at the Tcr α locus^[170]. However, the genome-wide impact of cohesin depletion has yet to be determined.

Contrary to expectation we find no major role for cohesin in the maintenance of architectural features of genome organisation. However, cohesin depletion reduces long-range interactions between cohesin-bound sites. The resulting chromosomal interaction landscape is characterised by alternative interactions between chromatin features associated with transcriptional activation and repression. Interestingly, this re-organisation of long-range interactions is accompanied by changes in gene expression. Similarly to previous observations in *Drosophila*^[275], genes at the lower end of the expression spectrum are preferentially up-regulated, while genes at the higher end of the expression spectrum are preferentially down-regulated. Our data indicate that the organisation of the genome into architectural compartments and the random assortment of genes and regulatory elements within them are insufficient for the precise regulation of gene expression. Rather, cohesin enables discrete gene expression states by promoting cohesin-based interactions within a pre-existing architectural framework.

3.3 Results

In our experimental system the locus encoding the cohesin subunit RAD21 is deleted by the activity of a CD4Cre transgene when developing thymocytes exit the cell cycle as part of their developmental program in vivo. Developing thymocytes arrest at the G1 phase of the cell cycle when DNA is unreplicated and chromosomes are present as single copies, not as sister chromatids. In contrast to some model organisms, homologous chromosomes are not paired in mammalian interphase, and cohesin is not tasked with holding sister chromatids or homologs together. This resulted in *Rad21^{lox/lox}* CD4Cre *CD4⁺ CD8⁺* double positive thymocytes (referred to as “cohesin-deficient thymocytes” below) with a 75-80% reduction in RAD21 protein expression^[170]. CTCF binding was not significantly different between control and cohesin-deficient cells. This approach allowed us to investigate the role of cohesin in gene expression regulation and structural organisation of the genome in non-cycling cells without resorting to cell lines or in vitro culture systems.

We prepared Hi-C libraries from two biological replicates of control and cohesin-deficient thymocytes, obtaining a total of $\approx 392M$ unique valid pairs: $\approx 203M$ for control and $\approx 188M$ for cohesin-deficient thymocytes. To define chromosomal compartments we applied eigenvector analysis of chromosomal organisation to iteratively corrected Hi-C maps^[236] at 140 kb resolution. Strikingly, the assignment of chromosomal compartments is highly correlated for control and cohesin-deficient thymocytes (Figure 3.1A, B). We conclude that, unexpectedly, Mb-scale architectural compartments assigned by eigenvector analysis of chromosomal organisation are resilient to a 75-80% reduction of RAD21 protein expression in interphase.

3.3.1 Cohesin binding predicts perturbed long-range interactions in cohesin-deficient thymocytes

We applied the HOMER pipeline^[235] to our Hi-C data (see Section 3.5.2.3). We first identified 100 kb genomic regions that interacted significantly with each other in either control or cohesin-deficient cells ($FDR = 0.1$), using an approach that takes both linear genomic distance and locus-specific sequencing depth differences into account. Interactions between these 100 kb regions were then compared between control and

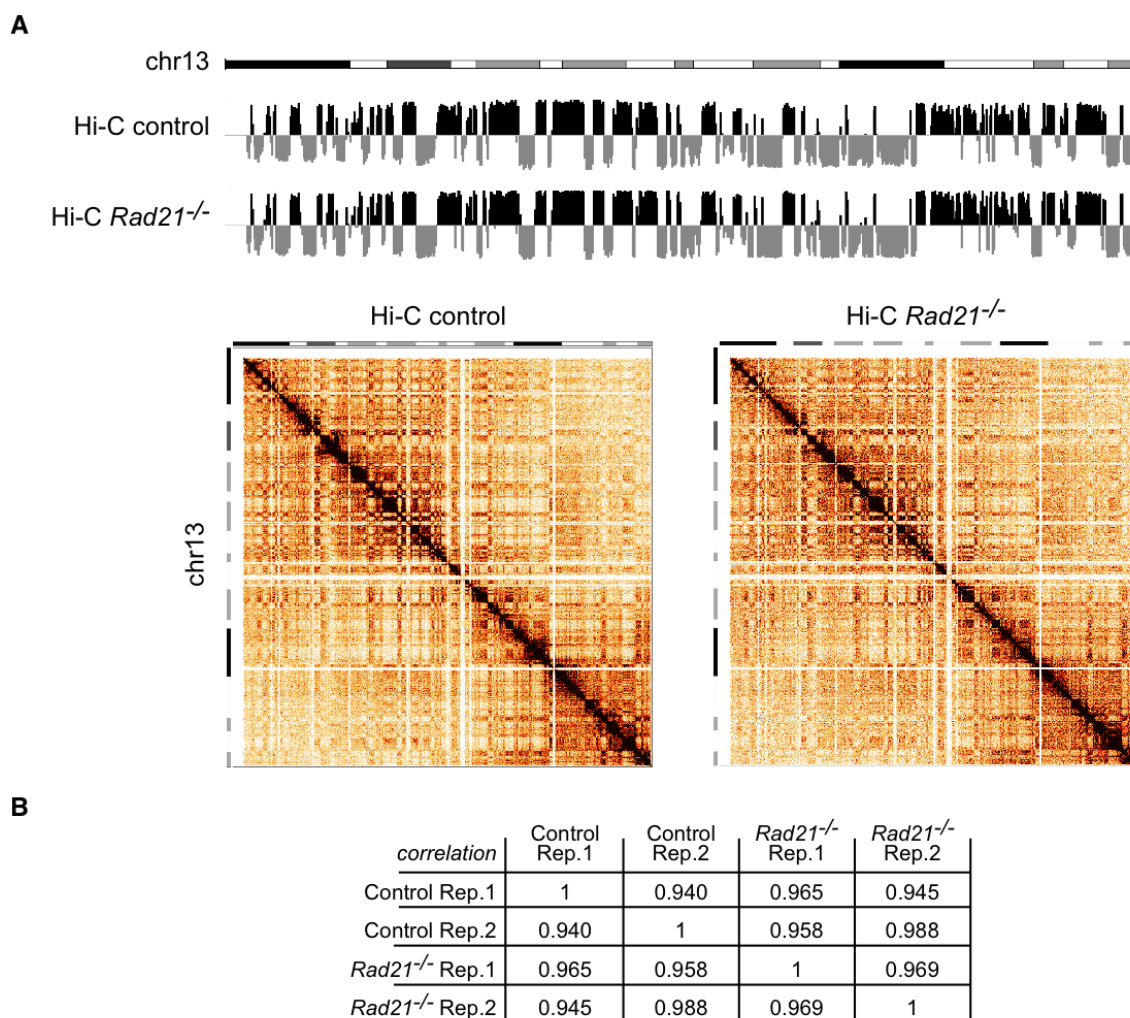


Figure 3.1: Chromosomal compartments are resilient to the depletion of the cohesin subunit RAD21 from non-cycling thymocytes in vivo. (A) Compartment tracks (top) and interaction matrices (bottom) for chr13 at 140 kb resolution in control and cohesin-deficient thymocytes. Of the 17,548 regions evaluated, only two show a consistent compartment change from A to B and two from B to A (change in eigenvector value > 0.03). (B) Table showing that Hi-C compartment data are highly correlated for control and cohesin-deficient thymocytes.

cohesin-deficient cells. Although the majority of significant interactions are unchanged in cohesin-deficient cells, 10,917 interactions are significantly altered in the pooled data, of which 1,476 are found in replicate 1 and 5,004 in replicate 2 ($P < 0.05$). There is a highly significant overlap of 502 differential interactions between replicates ($P < 10^{-15}$, odds ratio of 5.25 for increased interactions and 9.96 for decreased interactions). Compared to interactions that are unaffected by the depletion of cohesin, 278 of the 502 differential interactions are decreased and 224 are increased, where all differential interactions are intrachromosomal (Figure 3.2A). The great majority of reduced interactions are contained within individual A compartments and a fraction of gained interactions occur within individual B compartments (Figure 3.2B).

To explore the relationship between cohesin binding and cohesin-dependent differential interactions we focused on the 946 distinct 100 kb genomic regions that participated in the 502 differential interactions. These differentially interacting regions are significantly enriched for the binding of the cohesin subunit RAD21, CTCF, the cohesin loading factor NIPBL and the mediator subunit MED1 (Figure 3.3A). At the 100 kb level, decreased interactions in particular are enriched for cohesin, both with and without CTCF (cohesin-non-CTCF or CNC in Figure 3.3A), CTCF and NIPBL (Figure 3.3A). Increased interactions are enriched for marks of transcriptional activity, including MED1, H3K4me3 and RNA polymerase II (RNAP2) (Figure 3.3A). The observed decrease in interactions between regions rich in cohesin and CTCF binding is intuitive based on reduced interactions between cohesin and CTCF binding sites in cohesin-depleted cells^[163,169,170]. Increased interactions between regions rich in features associated with active genes (MED1, H3K4me3 and RNAP2) could suggest a role for transcriptional activity in such interactions^[276], although the 100 kb resolution achieved in the analysis of such regions is insufficient to pinpoint the role of individual features in driving differential interactions (see below).

Interactions that decrease upon cohesin depletion are strong in control cells, while interactions that increase following cohesin depletion are significantly weaker than average in control cells where cohesin is present (Figure 3.3B).

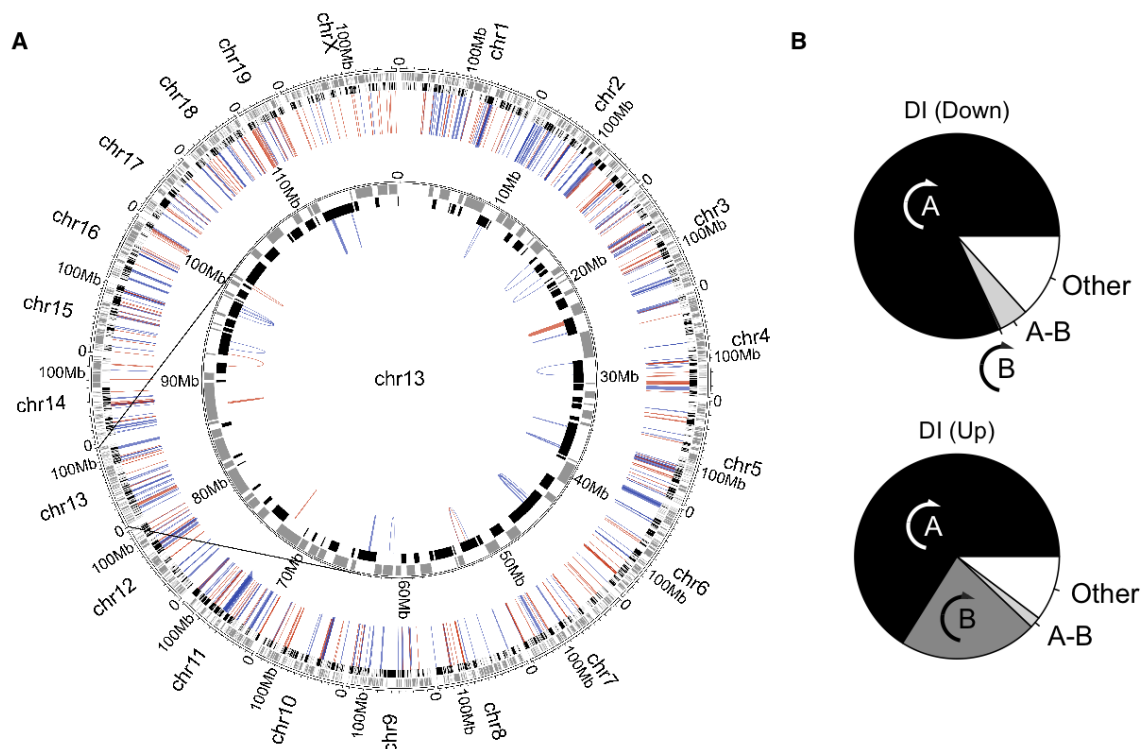


Figure 3.2: Cohesin depletion perturbs long-range interactions in cohesin-deficient thymocytes. (A) Circos plot illustrating the chromosomal position of differential interactions in the context of chromosomal compartments. The HOMER software suite was used to determine significant interactions between 100 kb genomic regions in either control or cohesin-deficient thymocytes ($FDR = 0.1$; replicates pooled). Of 10,917 interactions that are significantly altered in the pooled samples, 1,476 interactions change in replicate 1 and 5,004 in replicate 2 ($P < 0.05$). Of 502 differential interactions that are shared between replicates, 278 are decreased (blue) and 224 are increased (red). All differential interactions are intra-chromosomal. Compartment A and B assignment is indicated in the Circos plots by black and grey rectangles respectively. (B) Differential interactions in cohesin-deficient cells are largely contained within pre-existing chromosomal compartments. A: differential interactions entirely contained within the same A compartment. B: differential interactions entirely contained within the same B compartment. A-B: differential interactions bridging A and B compartments. Other: interacting regions are either unassigned or bridge two distinct A or B compartments. Down-regulated interactions (top) and up-regulated interactions (bottom) are shown separately.

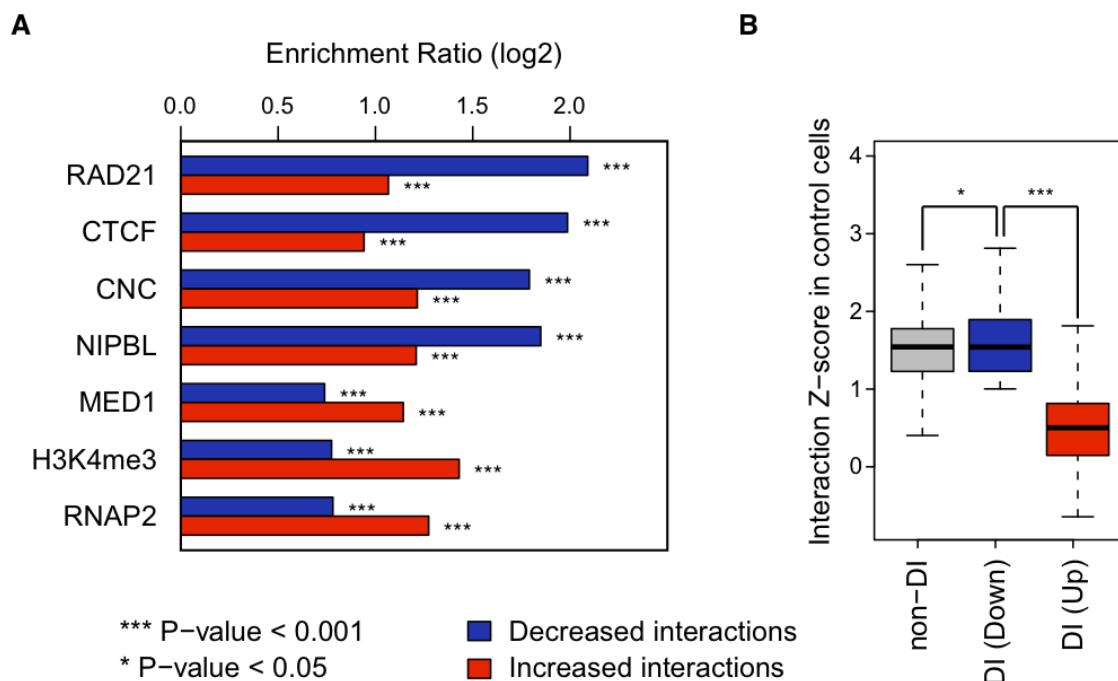


Figure 3.3: Cohesin binding predicts perturbed long-range interactions in cohesin-deficient thymocytes. (A) Features enriched in 100 kb regions that show differential interactions in cohesin-deficient thymocytes. Differential interactions between control and cohesin-deficient samples involve 946 unique 100 kb regions (510 involved in decreased interactions, 427 in increased interactions and 9 involved in both) that participate in 502 unique differential interactions (278 that are decreased and 224 that are increased) and that are shared between replicates ($P < 0.05$). We tested whether differentially interacting 100 kb regions are enriched for the presence of RAD21, CTCF, NIPBL, MED1, H3K4me3 and RNAP2 binding events. Differentially interacting regions are significantly enriched for the binding of the cohesin subunit RAD21, both with and without CTCF (cohesin-non-CTCF; CNC) and of the cohesin-associated factors CTCF and NIPBL, and features of transcriptional activity including MED1, H3K4me3 and RNAP2. Differential interactions are further classified into decreased and increased interactions. (B) Strength distribution of cohesin-dependent interactions. Using the number of Hi-C reads as an indicator of the strength of interactions, differential interactions that are decreased in cohesin-depleted thymocytes are similar in strength to unchanged interactions in control cells, whereas increased interactions tend to be weak before cohesin depletion (Mann-Whitney U test $P < 10^{-15}$). Outliers are not depicted.

3.3.2 Cohesin depletion perturbs gene expression in open compartments

RNA-seq analysis of two independent biological replicates at a depth of 144M total reads identified 1,153 genes that are differentially expressed between control and cohesin-deficient thymocytes ($FDR = 0.05$; 703 up-regulated, 450 down-regulated; Figure 3.4A). We validated 15 of these by quantitative reverse transcriptase polymerase chain reaction (qRT-PCR) and find that both methods are in close agreement (Figure 3.4B). Comparison of this RNA-seq data with our Hi-C compartment results show that 98% of deregulated genes reside in “open” (A-type) compartments (Figure 3.5A; $P < 10^{-15}$, odds ratio=5.49).

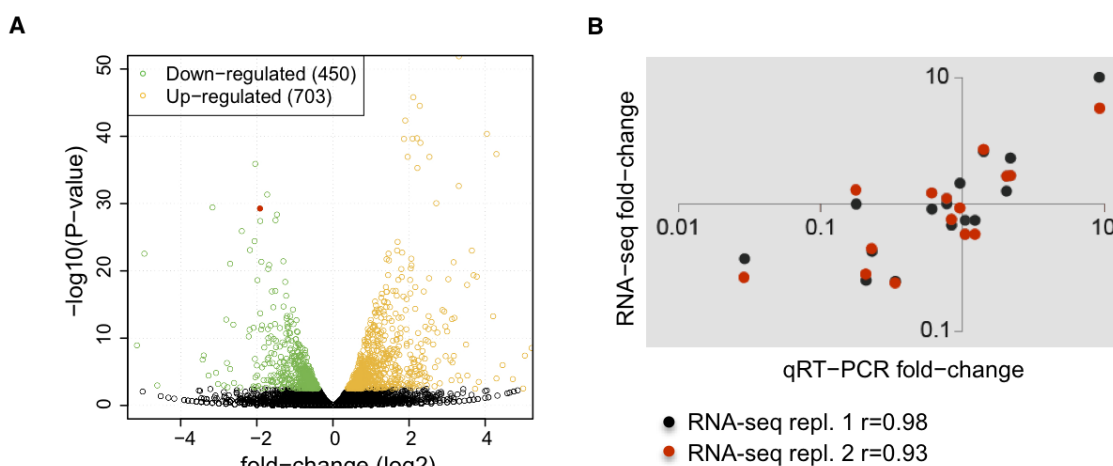


Figure 3.4: Differentially expressed genes in cohesin-deficient thymocytes as assayed by RNA-seq. (A) Volcano plot of RNA-seq data. Of the 17,849 genes assayed, 1,153 are significantly differentially expressed in cohesin-deficient thymocytes ($FDR = 0.05$), 450 of which are up-regulated (orange) and 703 down-regulated (green). The red dot represents *Rad21*. Eight genes above the $-\log_{10}(P\text{-value}) = 50$ threshold (all up-regulated) were omitted from the plot. (B) Validation of RNA-seq data for 15 transcripts over a wide range of expression levels. Shown is fold-change measured by qRT-PCR (X-axis, two independent biological replicates) against the fold-change measured by RNA-seq (Y-axis) and the Pearson correlation coefficient (r) for two independent RNA-seq experiments.

3.3.3 Genes that are sensitive to cohesin dosage are bound by cohesin, CTCF and NIPBL

The great majority of genes that are deregulated in cohesin-deficient thymocytes are bound by NIPBL (96%, $P < 10^{-15}$, odds ratio=5.56), RAD21 (82%, $P < 10^{-15}$, odds ratio=2.45), and CTCF (92%, $P < 10^{-13}$, odds ratio=2.16; Figure 3.5B), indicating that most of these genes are direct targets of cohesin-mediated regulation. In agreement with this, the most highly enriched gene ontology (GO) terms (adjusted $P < 10^{-8}$) include functions related to development, transcription, signal transduction, lymphocyte activation and differentiation as well as haematopoiesis and the immune system (Figure 3.5C). There is no enrichment for cell cycle related terms including cell cycle, DNA replication, chromosome segregation and checkpoint activation. Terms relating to DNA damage, DNA repair, and cell division that are highly enriched in results from previous studies using RNAi-mediated depletion of cohesin from dividing ES cells^[169] are not enriched in cohesin-deficient thymocytes, validating our rationale for an experimental system based on non-dividing cells. Taken together with the observation that architectural compartments remain largely intact in cohesin-deficient thymocytes, these results suggest that deregulated gene expression occurs at the level of individual loci, and is not secondary to a global collapse in genome organisation.

3.3.4 Predictive features of genes that show cohesin-dependent expression

To delineate factors associated with gene expression changes in cohesin-deficient thymocytes, we used a regression model that integrated gene expression, Hi-C and ChIP-seq data. We assigned genes to one of three classes (up-regulated, down-regulated or unchanged) and tested for the presence or absence of ChIP-seq peaks near each gene promoter ($\text{TSS} \pm 2.5$ kb) as well as gene location within 100 kb regions that interact differentially in control and cohesin-deficient thymocytes. Differential interactions were further divided into interactions that were stronger (DI region, Up) or weaker (DI region, Down) in cohesin-deficient thymocytes. We also considered the presence of the H3K4me3 histone modification, the binding of RAD21, CTCF, NIPBL and Mediator, RNAP2, paused RNAP2 at the promoter^[245], the presence of promoter CpG islands (CGI) and gene length. We determined the relative importance of each variable in

A

Compartment	#DE genes	odds ratio	significance
A	1,121 (98%)	5.49	$P<10^{-15}$
B	10 (1%)	0.12	$P<10^{-15}$
Unassigned	11 (1%)	0.40	$P<10^{-3}$

B

Factor	Location	#DE genes	odds ratio	sig.
RAD21	TSS±2.5 kb	411 (36%)	2.27	$P<10^{-15}$
	gene±10 kb	942 (82%)	2.45	$P<10^{-15}$
NIPBL	TSS±2.5 kb	1,022 (89%)	4.04	$P<10^{-15}$
	gene±10 kb	1,105 (96%)	5.56	$P<10^{-15}$
CTCF	TSS±2.5 kb	715 (62%)	1.77	$P<10^{-15}$
	gene±10 kb	1,065 (92%)	2.16	$P<10^{-13}$

C

Gene Ontology (GO) term	ID	sig.
Reg. of trans. from Pol II prom.	GO:0006357	$P<10^{-11}$
Transcription from Pol II prom.	GO:0006366	$P<10^{-10}$
Leukocyte differentiation	GO:0002521	$P<10^{-10}$
Lymphoid organ development	GO:0048534	$P<10^{-9}$
Immune system development	GO:0002520	$P<10^{-9}$
System development	GO:0048731	$P<10^{-9}$
Reg. of metabolic process	GO:0019222	$P<10^{-9}$
T cell activation	GO:0042110	$P<10^{-9}$
Haemopoiesis	GO:0030097	$P<10^{-9}$
Developmental process	GO:0032502	$P<10^{-9}$
Lymphocyte activation	GO:0046649	$P<10^{-9}$
Cell activation	GO:0001775	$P<10^{-9}$
Reg. of signal transduction	GO:0009966	$P<10^{-8}$

Figure 3.5: Characteristics of cohesin-sensitive genes. (A) Genes affected by cohesin depletion are associated with open compartments (compartment A). Approximately 98% of differentially expressed genes that could be assigned to compartments (1,121 of 1,142 autosomal and X-linked genes) reside in open compartments, compared with 91.2% (16,255 of 17,819) of genes included in our analysis. Genes spanning more than one compartment were assigned to A when they overlapped at least partially with open compartments; genes overlapping compartments that could not be clearly defined as A or B (unassigned) are denoted as “unassigned”. Only 21 deregulated genes are found outside open compartments. Of these, 11 are located in unassigned compartments and just 10 of 1,142 deregulated genes are located in B-type compartments. This corresponds to a highly significant depletion of deregulated genes in B-type compartments ($P < 10^{-15}$, odds ratio=0.12). (B) Cohesin-regulated genes are bound by cohesin and associated factors. Associations are shown for ChIP-seq peaks for RAD21, NIPBL and CTCF^[277] within 2.5 kb of transcription start sites (TSS) and within 10 kb of canonical gene bodies. Note that differentially expressed genes are highly enriched for nearby cohesin binding events, with NIPBL showing stronger association than CTCF. (C) Gene Ontology (GO) analysis of genes that are differentially expressed in cohesin-deficient thymocytes and bound by cohesin (as detected by RAD21 ChIP-seq). Representative GO Biological Process terms with adjusted $P < 10^{-8}$ are shown

the multivariate model and ranked the corresponding coefficients by their statistical significance. The presence of a CGI at the promoter, gene length, the presence near promoters of RAD21, particularly without CTCF (CNC), and promoter-associated RNAP2 emerge as important variables (Figure 3.7; see Figure 3.6 for the results of a univariate analysis considering each variable separately). Interestingly, location within regions that show differential interactions in cohesin-deficient thymocytes – but not a control set of interacting regions that show no differences between control and cohesin-deficient thymocytes (“Random DI region” in Figure 3.7) – is predictive of gene expression changes. Decreased gene expression in particular is associated with differential interactions ($P < 10^{-11}$, odds ratio=2.63).

Investigating variables highlighted by the regression model we first probed the relationship between deregulated gene expression and differential long-range interactions. Results from this analysis show a strong association between decreased interactions and down-regulated gene expression (Figure 3.8A). Further analysis confirmed that genes with increased promoter CpG density are more likely to be up- than down-regulated in cohesin-deficient thymocytes (Figure 3.8B, left), and also more likely to coincide with differentially interacting regions in cohesin-deficient cells (Figure 3.8B, right).

Moreover, longer genes are more likely to be down-regulated in cohesin-deficient cells (Figure 3.8C, left), and are found preferentially in regions of reduced interactions (Figure 3.8C, right). The transcriptional control of long genes may be particularly complex, requiring cis-regulatory elements to coordinate and exert their effects over large genomic distances. For example, the formation of local regulatory neighbourhoods or looping structures^[80] could pose a particular challenge in the case of these genes. A set of complex loci of significantly greater length than average (Mann-Whitney U test $P < 10^{-15}$) overlap ultra-conserved non-coding elements, regulatory blocks or archipelagos^[73,278,279]. These loci are preferentially down-regulated in cohesin-deficient thymocytes (Figure 3.8D), which is consistent with a role for cohesin in long-range regulation of genes with complex regulatory inputs.

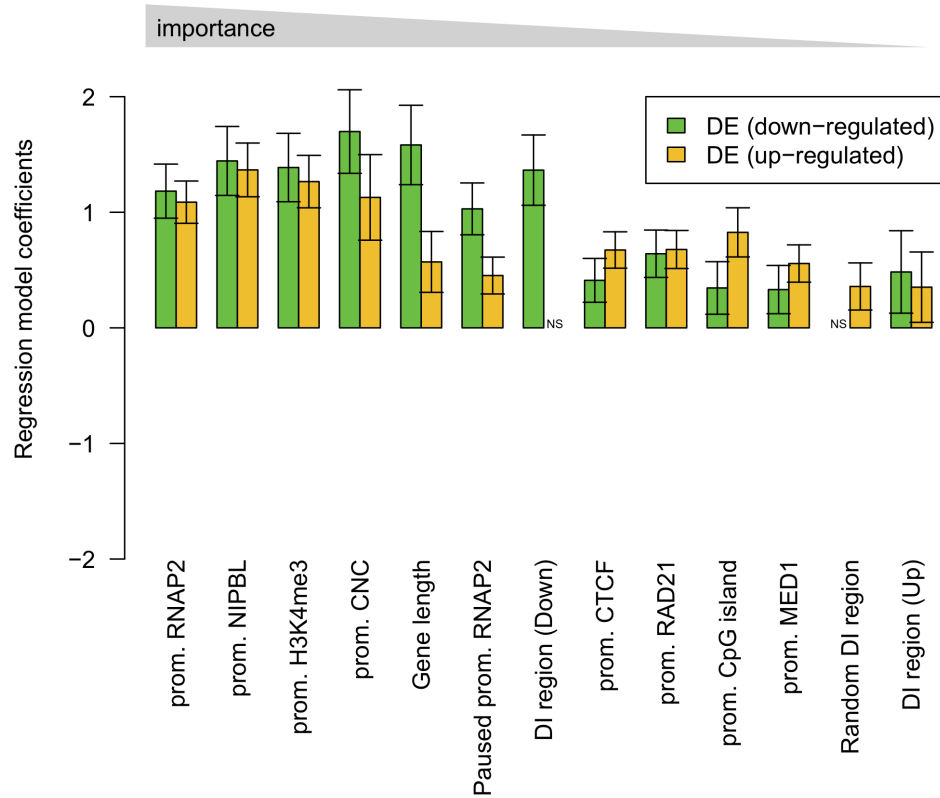


Figure 3.6: Univariate analysis of predictors of differential gene expression in cohesin-deficient thymocytes. Multinomial logistic regression model integrating gene expression, Hi-C and ChIP-seq data to predict up-regulated, down-regulated and unchanged genes. Only one variable was included in the model at a time (univariate analysis). Error bars represent 95% confidence intervals. Variables are ranked by coefficient significance from left to right.

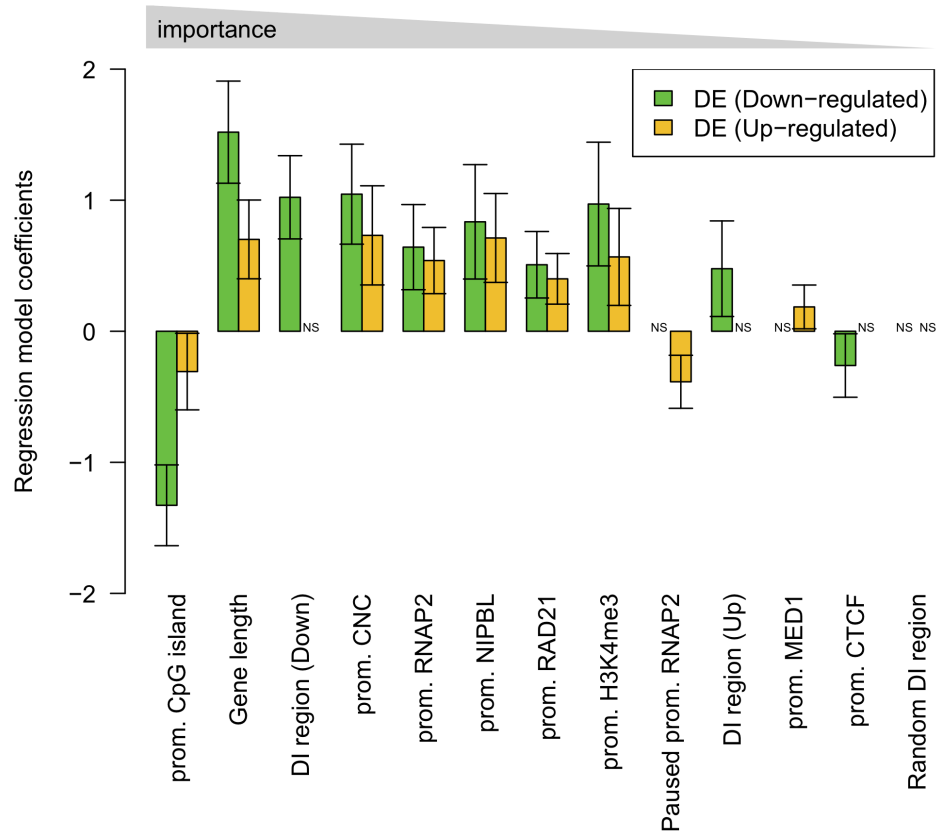


Figure 3.7: Multivariate analysis of predictors of differential gene expression in cohesin-deficient thymocytes. Multinomial logistic regression model integrating gene expression, Hi-C and ChIP-seq data to predict up-regulated, down-regulated and unchanged genes. Error bars represent 95% confidence intervals. Variables are ranked by coefficient significance from left to right.

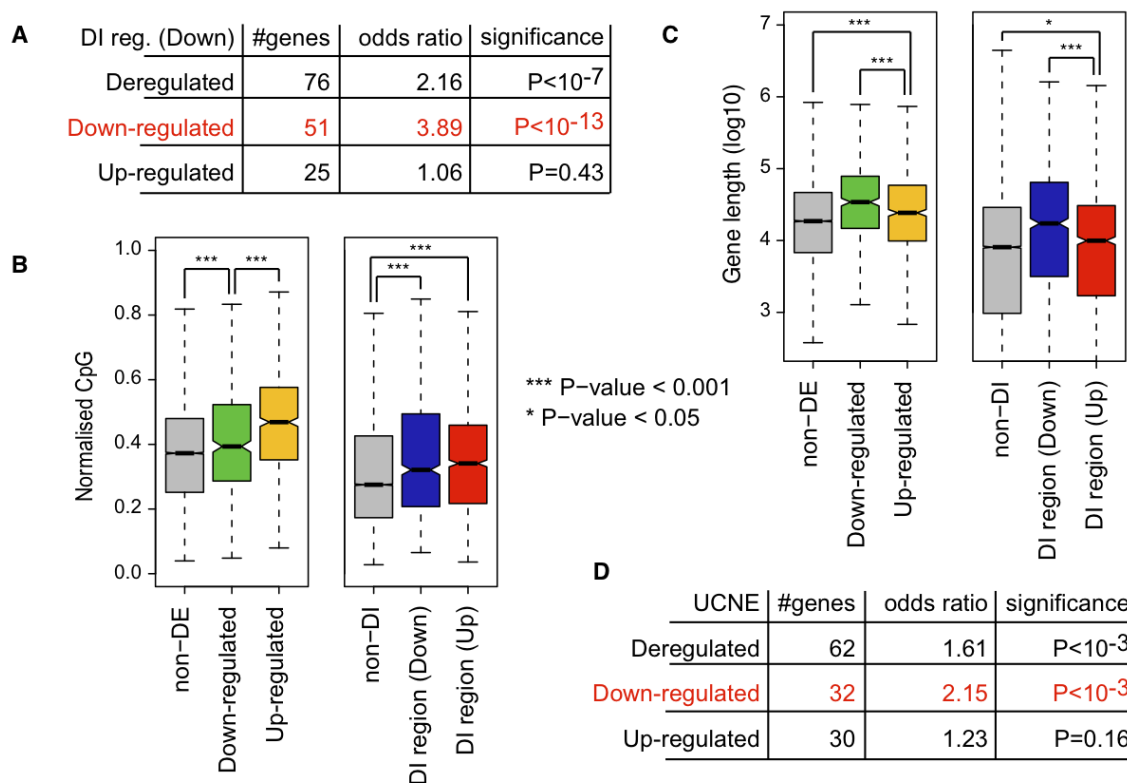


Figure 3.8: Investigating top predictors of differential gene expression in cohesin-deficient thymocytes. (A) Down-regulated genes are enriched for weakened chromatin interactions in cohesin-deficient cells. (B) Promoter CpG content of deregulated genes (DE) and genes participating in differential interactions (DI) in cohesin-deficient thymocytes. Up-regulated genes are associated with higher levels of promoter CpG dinucleotides (TSS \pm 2.5 kb) than both down-regulated and non-differentially expressed genes. Likewise, genes participating in increased interactions in cohesin-deficient thymocytes tend to have higher promoter CpG content. Outliers are not depicted. (C) Gene length of deregulated genes (DE) and genes participating in differential interactions (DI) in cohesin-deficient thymocytes. Down-regulated genes are significantly longer than both up-regulated and non-differentially expressed genes. Likewise, genes involved in decreased interactions in cohesin-deficient thymocytes tend to be longer. Outliers are not depicted. (D) Genes with complex regulatory inputs require cohesin for full expression. UCNE, ultra-conserved non-coding elements^[279].

3.3.5 Cohesin depletion perturbs long-range interactions within architectural compartments and compresses the dynamic range of gene expression

We used Structured Interaction Matrix Analysis (SIMA)^[89] to obtain a high-resolution view of interactions between specific chromatin features within the chromosomal compartments assigned by eigenvector analysis of chromosomal organisation. Although the number of Hi-C reads that map to any one specific feature – for example a cohesin-bound site – is too low to assign interactions with confidence, this approach is designed to combine information associated with multiple feature occurrences^[89]. We focused on open (A-type) compartments where the great majority of cohesin binding sites and cohesin-regulated genes reside (Figure 3.5A) and selected a range of features for analysis. These include ChIP-seq peaks for cohesin and associated factors, histone modifications indicative of active (H3K4me3) and repressed (H3K27me3) chromatin states, enhancers, transcription start sites of active and silent genes, as well as control sites that do not overlap with these features. For each compartment we counted Hi-C reads connecting features of the same type (homotypic interactions) and Hi-C reads connecting different features (heterotypic interactions), associating Hi-C reads that mapped within 10 kb of each feature^[280]. To determine the impact of cohesin on these interactions we compared interactions in control and cohesin-deficient cells for each feature (and pair of features) within between corresponding open compartments (see Section 3.5.2).

This approach reveals that interactions between RAD21 and CTCF sites are reduced in cohesin-deficient thymocytes (Figure 3.9), and the analysis of interactions between all pairs of features reveals that RAD21-RAD21, RAD21-enhancer and CTCF-RAD21 interactions are most strongly decreased (Figure 3.10A). This result is consistent with previous 3C experiments, which indicate reduced interactions between such sites in cohesin-depleted cells^[163,169,170].

Interactions that remain – or even increase – in cohesin-deficient cells include features of active transcription such as NIPBL, H3K4me3 and RNAP2 and extend to interactions between activation-associated features and repressive H3K27me3 marks (Figure 3.10A, B). Upon cohesin depletion, promoters of genes that are silent in control cells also show detectable interactions with a range of features linked to transcriptional

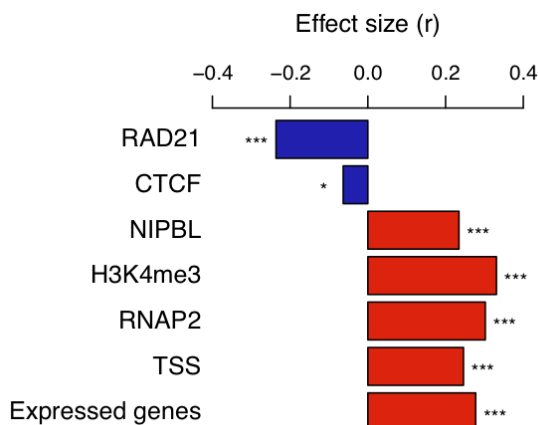


Figure 3.9: Impact of cohesin-deficiency on homotypic cohesin-based and alternative interactions. SIMA was used to determine the enrichment of Hi-C reads in interactions connecting “like” features in control and cohesin-deficient thymocytes. The difference between normalised enrichment ratios in each condition was assessed by the Wilcoxon signed-rank test (see Section 3.5.2).

activation (Figure 3.10A).

To ask whether these interactions in cohesin-deficient cells are selective, we identified control sites that are at least 10 kb removed from the other features considered. Interactions involving these “random” sites showed little cohesin dependence, suggesting that increased interactions preferentially involve sites that are characterised by the presence of the features analysed, thereby providing a measure of selectivity (Figure 3.11A).

Since SIMA effectively aggregates Hi-C data over all occurrences of a specific feature, we do not know whether the increased representation of alternative interactions in cohesin-deficient cells is absolute or relative to the loss of cohesin-based interactions. Either way, the data indicate a shift in the chromatin landscape in cohesin-deficient cells from cohesin-based to alternative interactions (Figure 3.10B).

As a proxy for the scale of differential interactions with decreased and increased representation in our Hi-C data, we stratified SIMA results by compartment size. Interestingly, the reduction of cohesin-based interactions is most pronounced in compart-

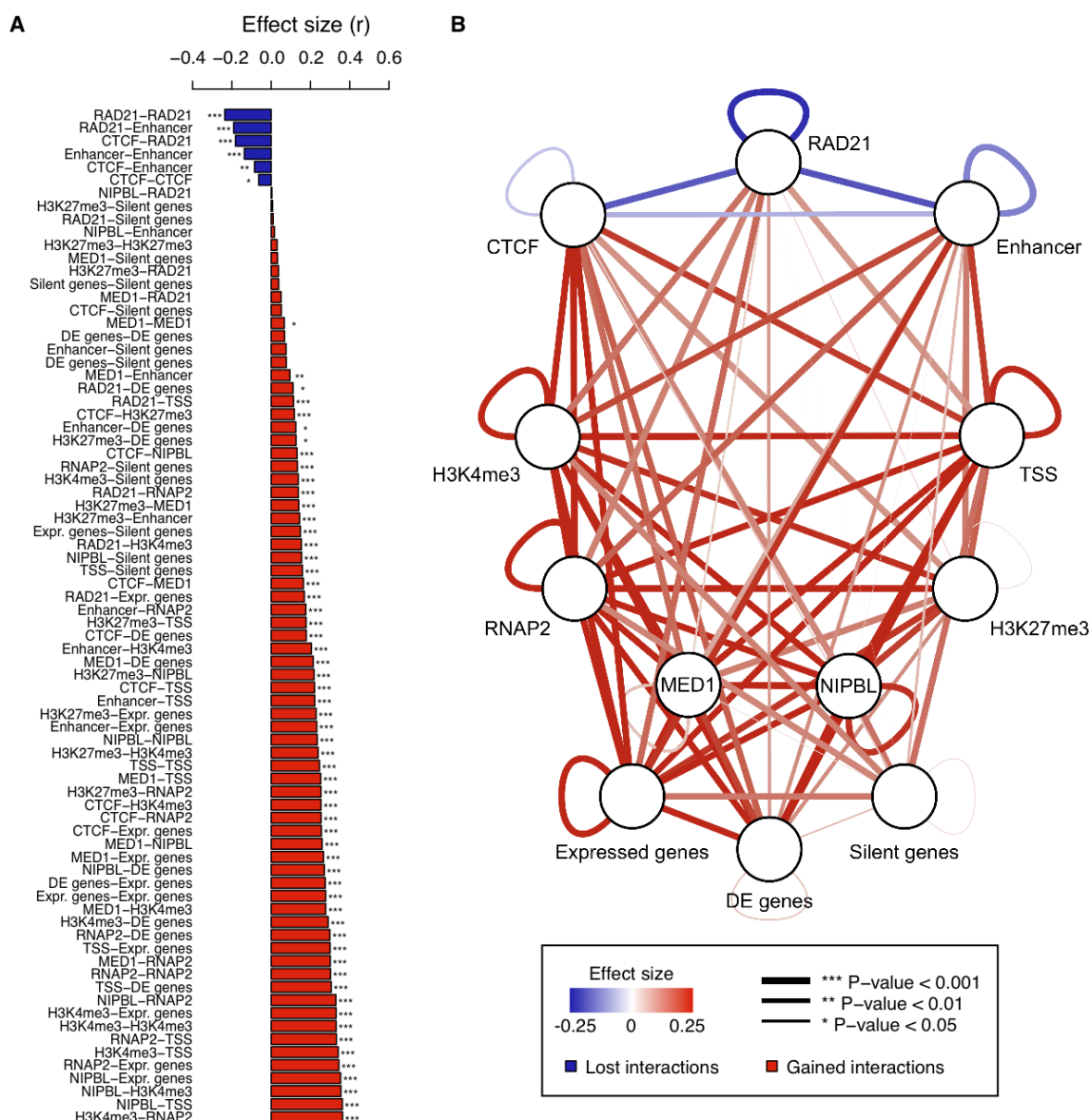


Figure 3.10: Impact of cohesin-deficiency on all pair-wise feature-based interactions. (A) SIMA results for homotypic interactions are shown together with those associated with interactions connecting different features (heterotypic interactions). Refer to Figure 3.9 for details. Homotypic RAD21-RAD21 interactions are the most decreased, followed by RAD21-enhancer and CTCF-RAD21 interactions. Interactions between features associated with active transcription are strongly increased in cohesin-deficient thymocytes, as are interactions involving the repressive histone modification H3K27me3^[281] with marks of active transcription. (B) Cytoscape representation of SIMA results in (A). Edge colour and width correspond to the Wilcoxon signed-rank test effect size and significance, respectively.

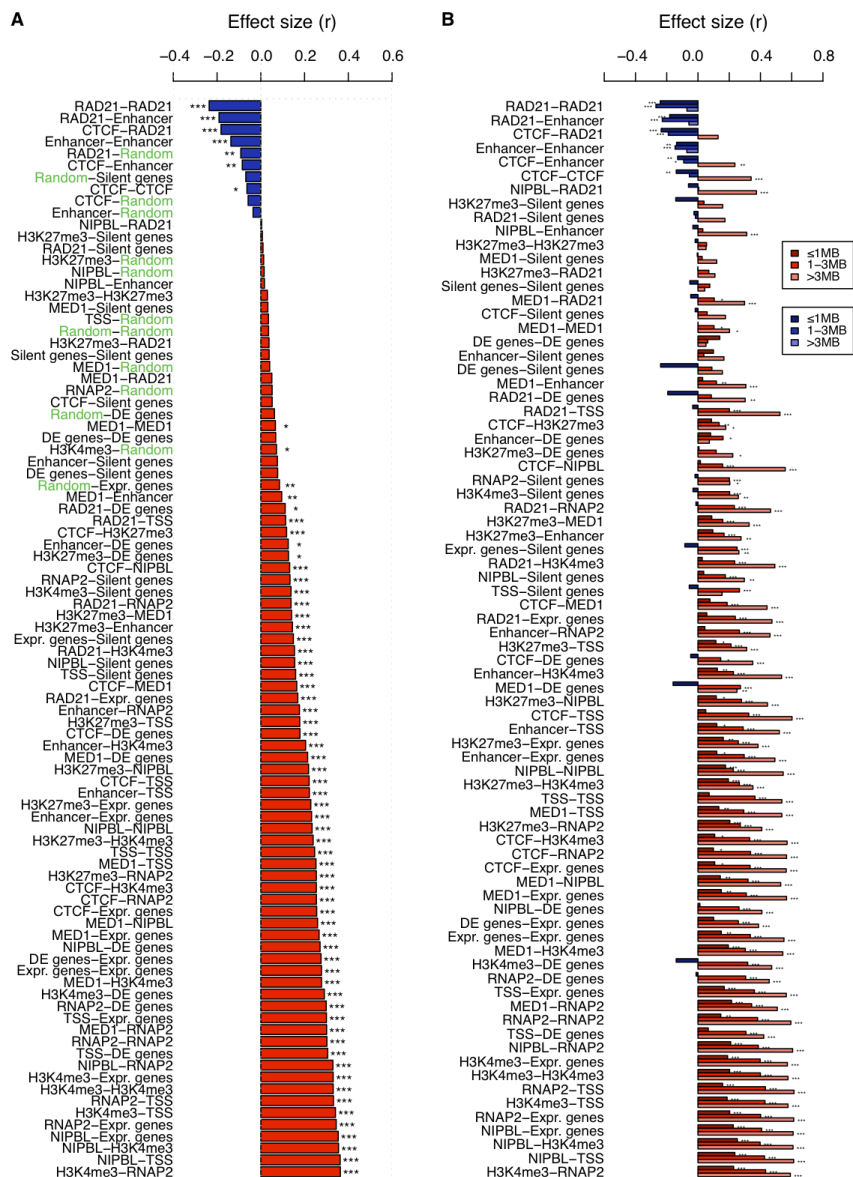


Figure 3.11: Selectivity of increased interactions and the effect of compartment size on all pair-wise feature-based interactions. (A) Impact of cohesin-deficiency on all pair-wise feature-based interactions as assayed by SIMA (see Section 3.5.2 and Figure 3.9). “Random” sites (green text) were defined as at least 10 kb removed from other features. Note that interactions involving these random sites show little cohesin dependence. (B) Compartment size and the impact of cohesin-deficiency on pair-wise feature-based interactions. Interactions were assayed by SIMA and stratified by compartment size (see Section 3.5.2 and Figure 3.9). The effect of reduced cohesin-based interactions is most pronounced within the more numerous smaller compartments (300 kb-3 Mb), whereas increased alternative interactions dominate within larger compartments (3-5 Mb).

ments < 1 Mb in size, suggesting an upper limit for cohesin-based interactions within compartments (Figure 3.11B and Figure 3.12). On the other hand, the alternative interactions detected in cohesin-deficient cells increase with compartment sizes > 1 Mb, suggesting that they preferentially occur over larger distances or in larger compartments, which may exhibit a greater degree of complexity (Figure 3.11B and Figure 3.12).

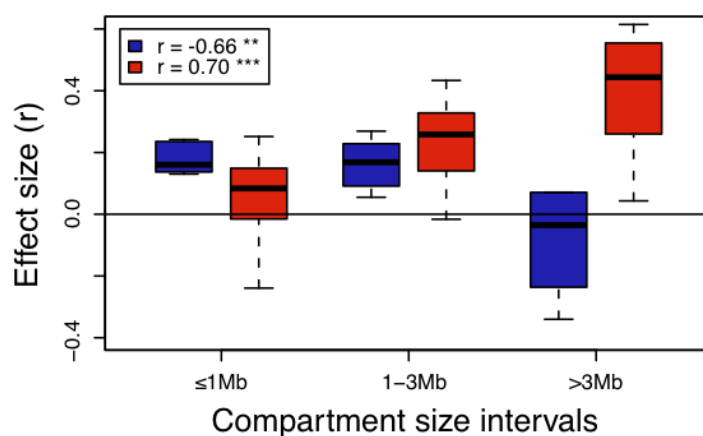


Figure 3.12: Length scale of lost and gained cohesin-dependent feature-based interactions. Boxplots are based on SIMA comparisons stratified into three classes according to compartment size (≤ 1 Mb, 1-3 Mb, > 3 Mb). Effect sizes for decreased interactions (blue: RAD21-RAD21, CTCF-RAD21, CTCF-CTCF) and increased interactions (red: the remainder) were grouped and are indicated separately (see legend). The effect of reduced cohesin-based interactions is most pronounced within smaller compartments and decreases when larger compartments are considered (Pearson correlation coefficient $r = -0.66$, $P < 0.01$), whereas alternative interactions increase with compartment size ($r = 0.7$, $P < 0.001$). Outliers are not depicted.

The prominence of interactions between features associated with both active gene expression and silencing in cohesin-deficient cells (Figure 3.10A) resonates with models where cohesin not only facilitates specific interactions, but also provides separation between genes and regulatory elements^[78,91,110,282]. To determine the effect of the homogenisation of long-range interactions on gene expression, we stratified genes ac-

cording to their level of expression. Consistent with the altered chromatin interaction landscape described above, gene expression is perturbed across the entire range of the expression spectrum (Figure 3.13A). In addition, genes with low expression are more often up-regulated while genes with high expression are more often down-regulated (Figure 3.13A). This results in a systematic skewing of gene expression away from the extremes of the dynamic range and towards average values (Pearson’s correlation coefficient $r = -0.16$, $P < 10^{-7}$; Figure 3.13A). Accordingly, the proportion of up-regulated genes is higher at the lower end of the expression spectrum, while the proportion of down-regulated genes is lower at the higher end of the expression spectrum (Pearson’s correlation coefficient $r = -0.97$, $P < 10^{-4}$; Figure 3.13B). Hence, genes showed a more uniform pattern of expression in cohesin-deficient cells.

3.4 Discussion

The cohesin protein complex provides physical linkage between sister chromatids from the time of chromosome duplication in S-phase until chromosomes segregate in cell division^[148] and can form long-range interactions that link gene regulatory elements with their targets in interphase^[163,169,170]. Unexpectedly, our Hi-C analysis shows that cohesin is not required for the maintenance of compartments in non-dividing mammalian cells, at least not at a level of 75-80% cohesin depletion in our experimental system^[170]. This reduction is far greater than the 20-30% reduction known to cause severe developmental abnormalities in model organisms and in human Cornelia de Lange Syndrome (CdLS), suggesting that a breakdown of genome organisation at the level of compartments may not be the cause for deregulated gene expression in CdLS^[283,284]. The maintenance of architectural compartments may be dependent on other factors with chromatin organising potential in thymocytes, such as SATB1^[285], although their ubiquitous presence in normal cells suggests that they are instead a reflection of nuclear lamina interactions^[110,272] or simply genome-wide differences in transcriptional activity. In contrast to the preservation of architectural genome organisation, we find that cohesin depletion alters long-range chromosomal interactions within compartments and results in a more uniform expression of genes affected by cohesin depletion.

We find that the presence of cohesin binding sites coincides with differential interactions of 100 kb regions that show both increased and decreased interactions in

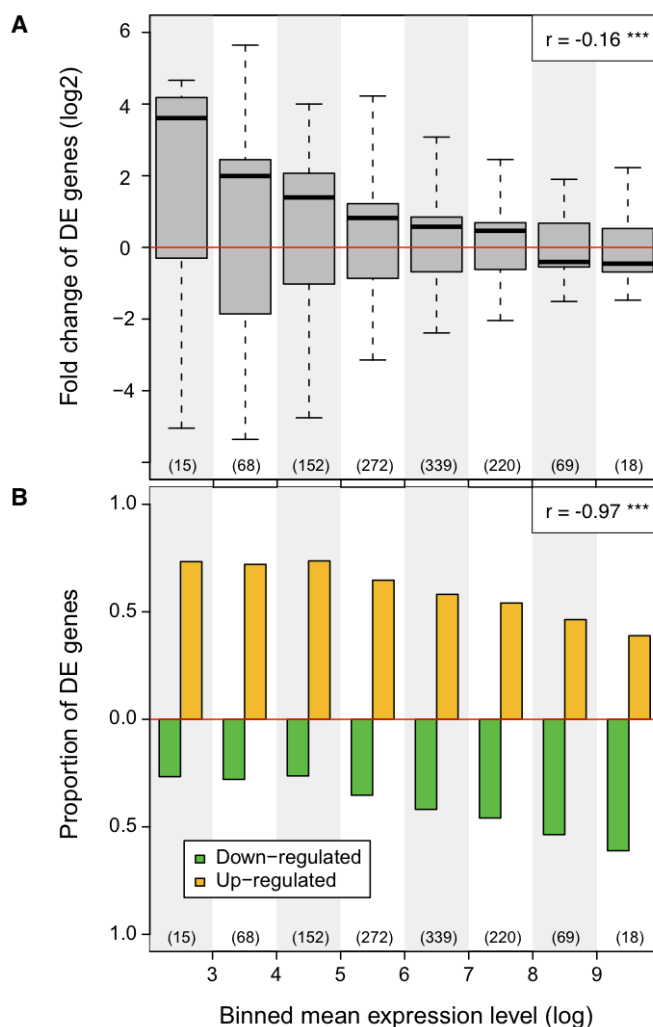


Figure 3.13: Cohesin depletion compresses the dynamic range of gene expression. (A) Genes were stratified into ten equally sized log intervals from low (0 – 1) to high (> 9) based on the average gene expression of control and cohesin-deficient thymocytes. Boxplots indicate the distribution of gene expression fold changes in cohesin-deficient thymocytes. The number of genes in each bin is indicated (bins 1 and 2 are empty). Note that lowly expressed genes are frequently up-regulated, whereas highly transcribed genes tend to be down-regulated. P-values are based on one-sample Wilcoxon signed-rank tests and indicate significant difference from 0 (no change). Genes with zero mean expression in both cohesin-deficient and control thymocytes are excluded and outliers are not depicted. (B) Barplot indicating the proportion of up- and down-regulated genes in each gene expression interval (see A). The proportion of up-regulated genes is anti-correlated with interval rank (Pearson’s correlation coefficient $r = -0.97$, $P < 10^{-4}$).

cohesin-deficient thymocytes. Our analysis at the local level shows that cohesin depletion reduces long-range interactions locally within 10 kb of RAD21 and CTCF binding sites, as well as near enhancers. Other interactions remain, or even increase, in cohesin-deficient cells. These alternative interactions include features of transcriptionally active and repressed chromatin. Using compartment size as a proxy for the length scale of interactions, cohesin-based interactions and the alternative interactions detected in cohesin-deficient cells appear to differ in scale. Consistent with correlative data^[286,287], cohesin-dependent interactions within compartments are mostly confined to distances < 1 Mb, i.e. the scale of topologically associated domains (TADs). In contrast, the alternative interactions detected in cohesin-deficient cells increase with compartment sizes above 1 Mb. This result suggests that TAD-scale cohesin-based interactions prevent the detection of longer-range alternative interactions. Hence, disrupting cohesin-based interactions may lead to a degree of mixing between chromosomal domains beyond the TAD scale (> 1 Mb).

Since SIMA effectively aggregates Hi-C data over all occurrences of a specific feature we do not know whether the observed increases are absolute, or relative to the loss of cohesin-based interactions. Either way, the data indicate a shift in the chromatin landscape in cohesin-deficient cells from cohesin-based to alternative interactions. Importantly, the homogenisation of interactions that results from this shift is consistent with the pattern of deregulated gene expression in cohesin-deficient thymocytes where genes at the lower end of the expression spectrum are preferentially up-regulated, while genes at the higher end of the expression spectrum are preferentially down-regulated. Perhaps this is the genome-scale equivalent of the ability of CTCF and CTCF-associated cohesin to mediate insulator and boundary functions. Although such functions have been predicted based on correlative analysis of histone modifications^[134], lamin-associated domains^[110] and long-range interactions maps^[91], their impact on gene expression was previously documented only in reporter assays^[78,282] and at individual loci, based on the manipulation of individual CTCF sites^[288] or the deletion of CTCF^[289]. We speculate that cohesin-based interactions limit the extent of alternative interactions to enable discrete gene expression states.

Our conclusion that cohesin contributes to functional interactions within pre-existing chromosomal compartments contrasts with cohesin's structural role in providing stable cohesion between sister chromatids^[148]. On closer inspection, however, this con-

clusion is consistent with a substantial body of existing knowledge on how cohesin interactions with chromatin are regulated during the cell cycle. In the G1 phase of the cell cycle, chromatin-associated cohesin shows relatively rapid turnover and has a half-life of minutes^[290,291,292,293]. It is only during S-phase that a subset of cohesin complexes is stabilised by the acetylation of specific lysine residues in the SMC3 subunit, which displace the cohesin unloading factor WAPL and allow association of the cohesin-stabilising factor sororin to extend the half-life of a subset of cohesin complexes^[290,291,294]. This subset of long-lived cohesin complexes is thought to have a structural role by mediating cohesion between sister chromatids from S-phase until the transition between metaphase and anaphase, which can be hours or even decades in the case of human oocytes^[295]. However, a population of cohesin complexes with high turnover remains even after DNA replication, and perhaps it is those complexes that continue to contribute to the regulation of gene expression during the S- and G2-phases of the cell cycle. The depletion of the cohesin removal factor WAPL demonstrates the consequences of rendering all cohesin complexes stable^[294]. Interestingly, loss of WAPL causes dramatic changes in the structure of interphase chromatin and perturbs the regulation of gene expression^[294]. It therefore appears that dynamic cohesin turnover is essential for regulated gene expression, and we speculate that this behaviour is reflected in the contribution of cohesin to genome organisation in interphase that is described by our data.

3.5 Methods

3.5.1 Experimental methods

The experiments described in this paragraph were performed by Vlad Seitan, Ye Zhan and Rachel Patton McCord. The conditional *Rad21* allele crossed to CD4Cre and methods for RT- and genomic PCR, chromosome conformation capture and ChIP-seq have been described^[170]. ChIP was performed using Abcam ab992 rabbit polyclonal antibody to RAD21^[170], Bethyl Laboratories A300-793A rabbit polyclonal antibody to MED1^[169] and Bethyl Laboratories A301-779A rabbit polyclonal antibody to NIPBL^[169]. RNA-seq was carried out as previously described^[170] except that we used Truseq kits according to the manufacturers' instructions (Illumina). Hi-C libraries were prepared as previously described^[22,296].

3.5.2 Computational methods

3.5.2.1 ChIP-seq read mapping and peak calling

Raw read alignment, filtering and peak calling for RAD21, MED1, NIPBL, CTCF^[277] and definition of CNCs was carried out as previously described^[238]. ChIP-seq data for H3K27me3^[281] was similarly processed, except CCAT version 3.0^[210] was used to identify the relatively broad regions occupied by this mark (pre-compiled histone modification configuration). Single base-pair summit positions for H3K4me3, RNAP2 and enhancers were obtained from the Mouse ENCODE Project^[297] were extended to a width of 200 bp to define peak regions.

3.5.2.2 RNA-seq data analysis

Raw reads for each condition and replicate were independently aligned to mouse transcript sequences (cDNA sequences from Ensembl version 66^[206], NCBI37/mm9) using Bowtie version 0.12.8^[205] with default parameters. Gene expression estimates and normalised count equivalents were obtained using MMSEQ version 0.11.2^[225]. We used the Bioconductor R package DESeq version 1.6.1^[217] to determine significantly differentially expressed genes in cohesin-deficient thymocytes versus control cells ($FDR = 0.05$). Empirical gene expression dispersion values were estimated in a condition-specific fashion (method="per-condition") and used to fit a dispersion-mean relationship, where only the fitted values were used (sharingMode="fit-only"). T cell receptor gene segments, pseudogenes, ribosomal genes and genes with an aggregate exon mappability score in the lowest ten percentile (UCSC genome browser track "wgEncodeCrgMapabilityAlign50mer") were excluded from the analysis. Expressed genes were defined as those having $\log(expression_level + 1) \geq 1$ in control cells; otherwise genes were considered silent.

3.5.2.3 Hi-C data analysis

Iterative error correction of Hi-C data was performed as described^[236]. The HOMER Hi-C software analysis pipeline^[235] was used to determine significant interactions, differ-

ential interactions and to perform Structured Interaction Matrix Analysis (SIMA)^[89]. Briefly, paired-end reads were trimmed to remove sequence following the canonical HindIII ligation junction sequence (1 bp mismatch allowed to account for potential star activity). Trimmed reads were aligned independently to the mouse reference genome assembly (NCBI37/mm9) using BWA^[204]. Paired-end reads were merged and filtered to remove duplicate read pairs (`-tbp 1`), paired-end reads likely representing continuous genomic fragments or re-ligation events (`-removePEbg`), self-ligations (`-removeSelfLigation`) and reads originating from regions with unusually high tag density (`-removeSpikes 10000 5`). Additionally, only read-pairs where both ends mapped near restriction sites were retained (`-both`).

To identify differential interactions, we first determined a “universe” of interactions on which to focus the analysis, defined as significant interactions between 100 kb genomic regions in either control or cohesin-deficient thymocytes (replicates pooled; $FDR = 0.1$). HOMER uses the binomial distribution to determine significant deviations above the expected number of Hi-C reads occurring between two loci, where the background model takes into account linear genomic distance and locus-specific sequencing depth differences. For each replicate, high scoring differential interactions ($P < 0.05$) were then determined by comparing Hi-C read levels between control and cohesin-deficient thymocytes within this subset of all possible interactions. Only differential interactions consistently identified in both Hi-C replicates were retained for downstream analysis.

To determine genomic features associated with chromatin interactions, we used a method that pools Hi-C information associated with a given set of genomic regions within a specified set of domains (SIMA)^[89]. We used default resolution (`-res 2500`) and optimal Hi-C interaction search space parameters (`-superRes 10000`). Domains of interest were defined as merged adjacent 140 kb regions within open compartments at least 500 kb (`-minDsize 500000`) and not more than 5 Mb in length. Within-domain associations were assessed independently in control and cohesin-deficient thymocytes for all peak sets (RAD21, MED1, NIPBL, CTCF, H3K4me3, H3K27me3, RNAP2, enhancers) as well as all canonical TSSs (excluding pseudogenes; Ensembl version 66), promoters of silent genes, expressed genes and significantly differentially expressed genes. Normalising by the number of expected Hi-C reads under the background model and comparing to the randomised average (after shuffling feature positions 10,000 times)

we obtained an enrichment ratio for each genomic region indicating the association of each feature (or feature pair) with interactions contained within that region. To determine the impact of cohesin on these interaction associations we compared observed/randomised enrichment ratios in control and cohesin-deficient cells for each feature (or feature pair) within each compartment. The size and direction of change in these ratios in cohesin-depleted thymocytes was compared using a paired statistical test (Wilcoxon signed-rank test) to provide a measure for the cohesin dependence of long-range interactions between specific features.

3.5.2.4 Multinomial logistic regression model

We used gene features derived from ChIP-seq, RNA-seq, Hi-C, genomic sequence and annotation to predict gene expression changes in cohesin-deficient thymocytes using a multinomial logistic regression model (nnet R package version 7.3-1^[298]). Model coefficients were estimated using the quasi-Newton BFGS optimisation algorithm. Binary variables used included RAD21 (overlapping CTCF), cohesin-non-CTCF (CNC), CTCF, NIPBL, MED1, H3K4me3 and RNAP2 peak and CpG island (CGI)^[299] overlap within the gene promoter ($TSS \pm 2.5$ kb), paused promoter RNAP2 ($Pindex \geq 4$ ^[245]), gene body overlap with 100 kb differentially interacting (DI) regions (Up, increased; Down, decreased) as well as a randomly selected “control” set of interacting regions. We also included total gene length as a continuous variable. The three-class categorical response variable was encoded as follows: “0” non-differentially expressed, “-1” significantly down-regulated, “1” significantly up-regulated.

Chapter 4

The CTCF paralog CTCFL/BORIS regulates gene expression and displaces cohesin from promoter-proximal sites when expressed in somatic cells

4.1 Summary

CTCFL (also known as BORIS) is the paralog of CTCF, a highly conserved and ubiquitously expressed DNA-binding protein with multiple roles in gene regulation and genome organisation in conjunction with cohesin. Although CTCF and CTCFL possess highly similar 11 zinc finger DNA-binding domains, their C- and N-terminal domains show no significant similarity (see Figure 4.1) and there is growing evidence that they fulfil equivalently divergent – possibly even antagonistic – cellular functions. To investigate the impact of CTCFL on global gene expression and its uncharacterised relationship with cohesin, we expressed a FLAG-tagged version of CTCFL in mouse ES cells and generated genome-wide binding maps of CTCF, CTCFL and cohesin (RAD21) by ChIP-seq. While CTCF occurs primarily at promoter-distal regulatory elements and recruits cohesin to these sites, we find that CTCFL binding occurs within a strikingly different DNA sequence and chromatin context, preferentially interacting

with GC-rich active promoters. The functional significance of CTCFL binding to active promoters is supported by CTCFL-associated changes in the expression of many of its target genes. Mechanistically, we find evidence that CTCFL modulates gene expression at least in part by reducing the association of cohesin with promoter-proximal regulatory elements, thereby potentially perturbing cohesin-mediated regulatory interactions. Hence our results show that the paralogs CTCF and CTCFL have evolved divergent strategies for cohesin-based gene expression regulation.

This study is the result of a collaboration between Dr. Matthias Merkenschlager's laboratory at the Medical Research Council Clinical Sciences Centre and Dr. Paul Flicek's research group at the EMBL European Bioinformatics Institute. Dr. Hegias Mira-Bontenbal performed most of the experiments for this project and I carried out the computational analysis, except where otherwise specified.

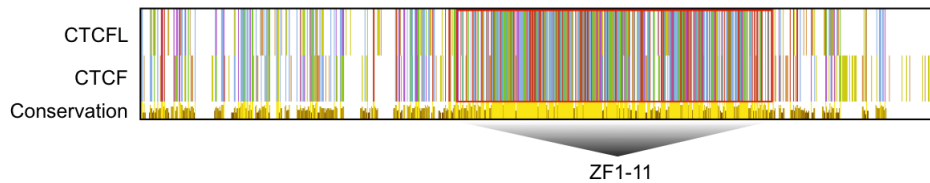


Figure 4.1: Schematic alignment of mouse CTCFL and CTCF protein sequences. Amino acid sequences of both proteins are shown (Clustal X colour scheme) together with the conservation score (bottom), where bar height and colour indicate the level of conservation. The highly conserved 11 zinc finger DNA-binding domains are indicated.

4.2 Introduction

Compared to CTCF, there has been less attention focussed on its sole paralog, CTCFL (also known as BORIS), and consequently its cellular function remains poorly understood. Open questions regarding the function of CTCFL binding include whether it has a role in regulating imprinted genes, whether it competes with CTCF for binding sites and the extent of its effect on global gene expression and DNA methylation, particularly in the context of oncogenesis. The expression of CTCFL in cancer cells and

lately its detection at lower levels in normal cells^[176] suggest a more widespread functional role outside of germ cell development. A recent genome-wide study showed that CTCFL associates with promoter-proximal nucleosome free regions, but it remains unclear whether this binding is opportunistic or functionally relevant^[175]. Furthermore, considering that CTCF and cohesin are co-dependent, it is of particular interest to determine whether CTCFL interferes with this relationship and/or recruits cohesin to its own binding sites.

4.3 Results

4.3.1 CTCFL associates with the majority of active promoters in human K562 leukaemia cells and mouse ES cells

To determine the binding preferences and possible functions of CTCFL expression in somatic cells, we analysed previously published ChIP-seq datasets in human K562 leukaemia cells from the ENCODE project^[15]. As expected, we found that CTCFL and CTCF co-bind many target regions (16,622), but a large fraction of binding events for each factor were occupied in isolation (Figure 4.2). The vast majority of CTCF binding events occur distal to promoters ($\text{TSS} \pm 2.5 \text{ kb}$; 80.9%), whereas CTCFL occurs more frequently within promoter regions (39.9%), particularly in the case of CTCFL-only binding events (55.9%; Figure 4.3). This is consistent with a previous report of promoter-proximal binding of CTCFL over-expressed in mouse ES cells^[175].

Visualising the status of chromatin within regions corresponding to the three distinct sets of CTCFL/CTCF-bound loci shown in Figure 4.2, we observe a number of striking differences (Figure 4.3). CTCFL binding is associated with increased ChIP signal for H3K4me2/3, H3K79me2 and RNA polymerase II (RNAP2), all of which are characteristics of active chromatin. Di- and trimethylation of H3K4 and RNAP2 are associated with active promoters, and H3K79 dimethylation has been implicated in transcriptional elongation^[300], suggesting that CTCFL may preferentially associate with TSSs of active genes.

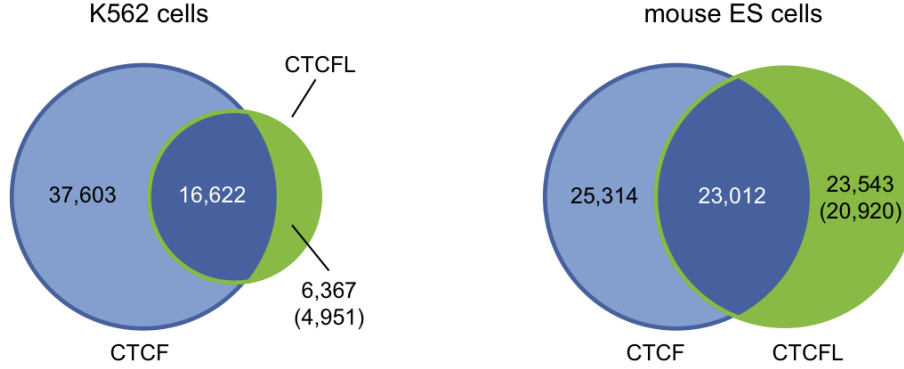


Figure 4.2: Venn diagrams showing the numbers of CTCF-only, CTCFL-only and CTCFL+CTCF (1bp peak overlap) binding events in K562 and mouse ES cells. The values in parentheses indicate the number of high-confidence CTCFL-only sites with no significant CTCF ChIP enrichment ($P < 0.05$).

The localisation of FLAG-CTCFL expressed in mouse ES cells (Figure 4.4B) mirrored the distribution of endogenous CTCFL in K562 cells in terms of partial overlap with CTCF (Figure 4.2) and promoter-proximal binding (36.6% and 53.3% for total CTCFL and CTCFL-only respectively; Figure 4.3). Combining our data with a re-processed collection of previously published ChIP-seq datasets in mouse ES cells^[54,56,241,248,249], we observe that the preference of CTCFL for marks of active transcription and the presence of RNAP2 is recapitulated in this system (Figure 4.3). Importantly, these maps of chromatin state were compiled using mouse ES cells prior to FLAG-CTCFL transfection. We can therefore infer that these marks are not simply a consequence of CTCFL binding, but more likely represent conditions favourable for its recruitment to chromatin.

Restricting our analysis to promoters only ($TSS \pm 2.5$ kb), we see that CTCFL targets a subset of these regions with characteristics of active transcription (Figure 4.5A). Defining active promoters as those possessing both H3K4me3 and RNAP2 ChIP-seq peaks ($TSS \pm 2.5$ kb), in K562 cells we see that endogenous CTCFL associates with the majority (67%) of active but with only a minority (13.6%) of inactive promoters (Fisher's Exact Test $P < 10^{-15}$, odds ratio=12.9). Similarly, 73.9% of active but only 11.6% of inactive promoters bound FLAG-CTCFL in mouse ES cells (Fisher's Exact Test $P < 10^{-15}$, odds ratio=21.6; Figure 4.5B). Hence, CTCFL binds most active pro-

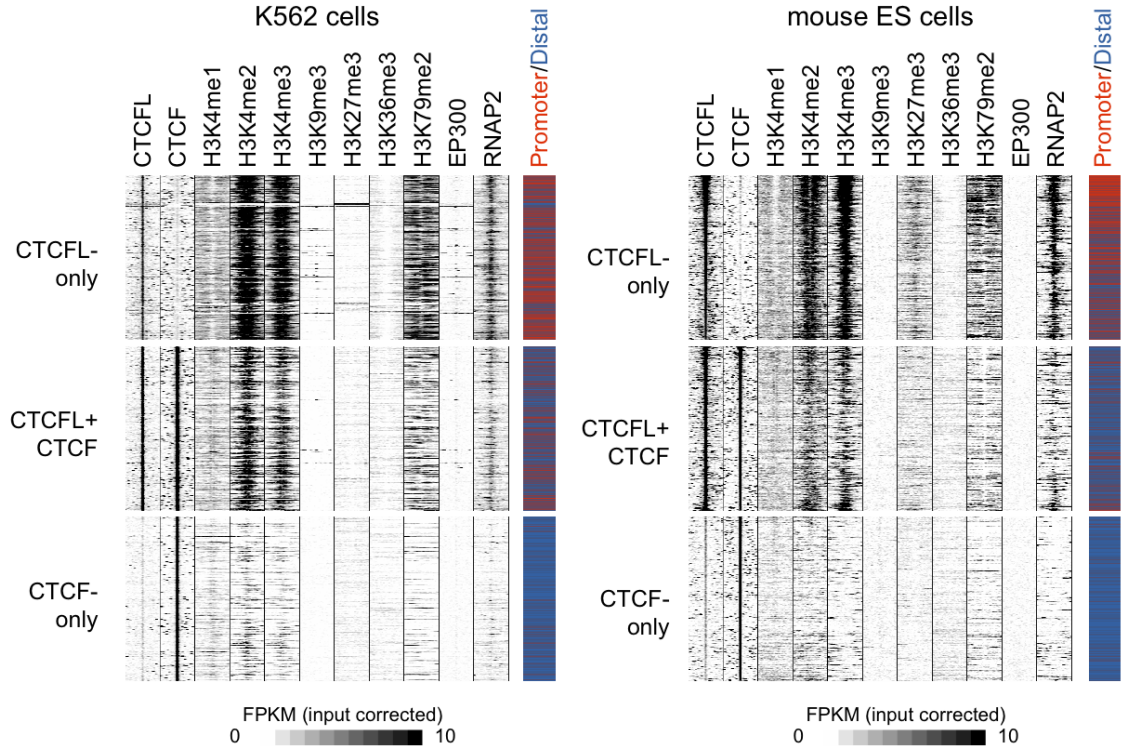


Figure 4.3: Heatmap representation of ChIP-seq fragment profiles indicates CTCFL binding events are characterised by an active chromatin status and promoter proximity. Profiles consist of input corrected FPKM for the indicated ChIP-seq datasets calculated at 100 bp resolution and shown within a 5 kb window centred on either the CTCFL or CTCF peak summit position. “Promoter” associated binding events (red) were defined as occurring within 2.5 kb of an annotated TSS; occurrence elsewhere classified as “Distal” (blue).

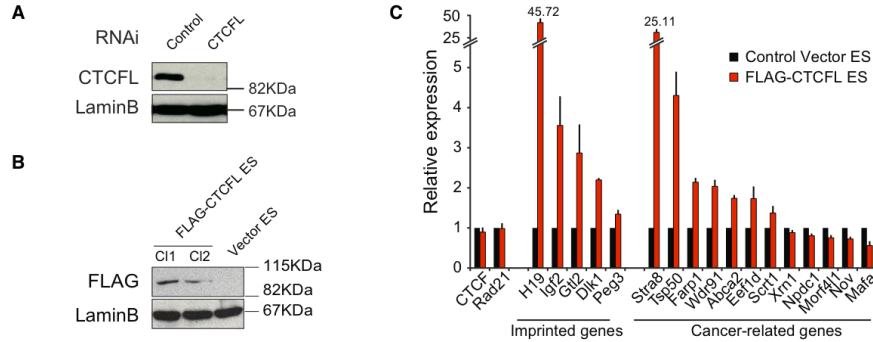


Figure 4.4: Validation of RNAi-mediated knockdown of CTCFL in K562 cells and characterisation of FLAG-CTCFL ES cells. (A) CTCFL immunoblot analysis of K562 cells transfected for 48h with control or CTCFL siRNA. Lamin B is the loading control. (B) Anti-FLAG immunoblot analysis of ES cells stably transfected with FLAG-CTCFL (clone 1, 2) or control vector. Lamin B is the loading control. (C) PCR validation of CTCFL-regulated genes in mouse ES cells. qRT-PCR analysis of a selection of imprinted and cancer-related genes that change expression in FLAG-CTCFL ES cells. *Ctcf* and *Rad21* expression remained unchanged and are shown as control.

motors when present in somatic cells and the observed differences in chromatin state at sites of CTCFL binding (Figure 4.3) cannot be entirely explained by its preference for promoter regions in general. The extent to which CTCFL associates with active promoters was not apparent from a previous study^[175], not least because the number of CTCFL binding sites identified was too small to match the number of active promoters.

4.3.2 In contrast to CTCF, CTCFL binds clusters of GC-rich motifs and low complexity repeats

De novo motif-finding (see Section 4.5.2) recovers a CTCFL motif which is highly similar to the canonical CTCF motif in both K562 cells and mouse ES cells (Figure 4.6). The only difference in the 14-mer CTCF consensus (GCGCCCCCTGGTGG) is the replacement of the right-most thymine with cytosine in that of CTCFL (GCGCCCC-CTGGCGG).

Despite similarities in their genome-wide motifs, CTCFL binds genomic loci with

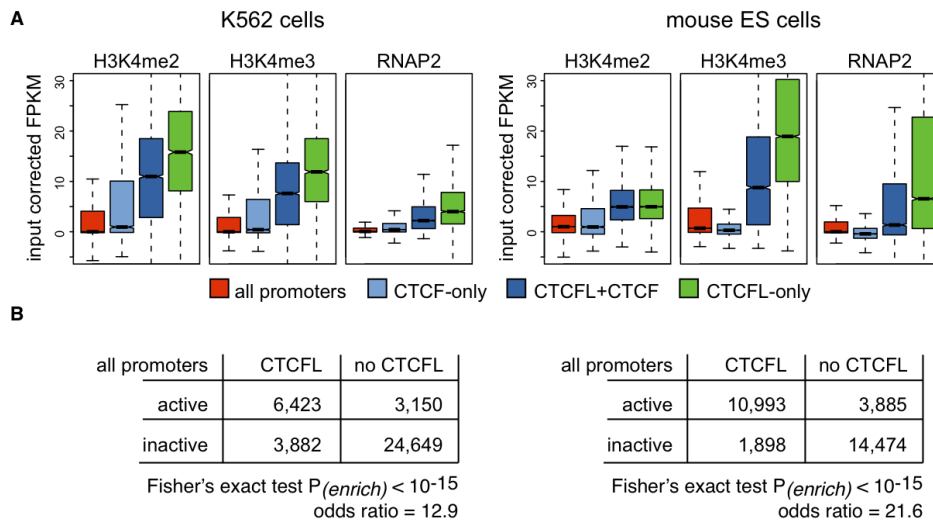


Figure 4.5: CTCFL associates with the majority of active promoters in human K562 leukaemia cells and mouse ES cells. (A) Promoter-proximal CTCFL binding is associated with elevated H3K4me2/3 and RNAP2 ChIP signal, particularly in the absence of CTCF. Outliers are not depicted. (B) Numbers of active and inactive promoters bound by CTCFL (TSS \pm 2.5 kb). Active promoters are defined as those with both H3K4me3 and RNAP2 ChIP-seq peaks within 2.5 kb of the TSS. Fisher's Exact Test P-values indicate the statistical significance of the positive association between active promoters and CTCFL occupancy.

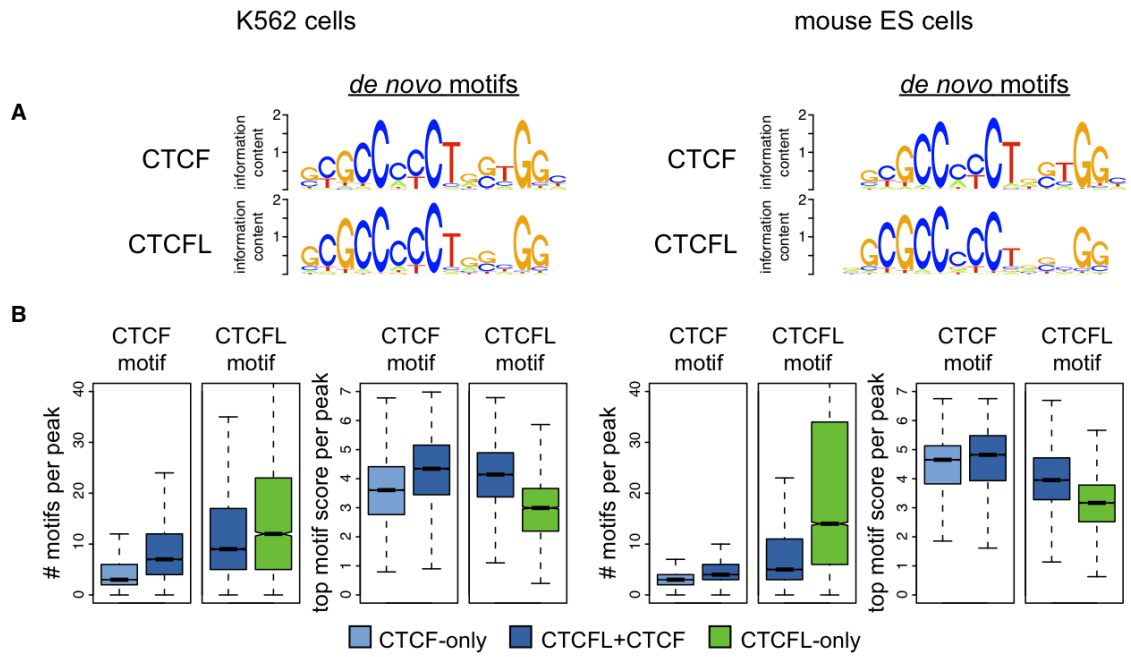


Figure 4.6: *De novo* motif characterisation and distribution within CTCF and CTCFL-bound regions. (A) Optimal motifs from *de novo* motif discovery using CTCF and CTCFL ChIP-seq peak sequences as input to MEME/NestedMica (see Methods). (B) In the absence of CTCF, CTCFL binds to regions of the genome with higher numbers of lower-scoring motifs.

higher numbers of motif instances (or words) than CTCF. Whereas most CTCF-only binding events contain 1-3 motifs in both K562 cells and mouse ES cells (*median* = 3), the median number of motifs within CTCFL-only peaks is 12 in K562 cells and 14 in mouse ES cells (Figure 4.6B). Co-bound CTCFL+CTCF regions show intermediate numbers of motif instances. Additionally, considering only the single highest-scoring motif per peak, CTCFL-only regions show the weakest match to the corresponding *de novo* motif (Figure 4.6B). These results show a preference of CTCFL for regions with multiple lower-scoring motif instances when bound independently of CTCF.

In order to further investigate the common and specific binding preferences of CTCFL and CTCF, we compiled a catalog of distinct 14-mer motif words bound at least five times by either factor. By counting the number of times a specific word is bound, and normalising by the total number of occurrences in the entire genome, we obtain a score indicating the genome-wide preference of a given factor for a given motif word. Figure 4.7 shows these normalised word frequencies for CTCFL and CTCF grouped into three classes based on the relative scores for each factor. In total we found 1,560 highly bound words (normalised frequency ≥ 0.5) in K562 cells and 15,704 such words in mouse ES cells. Motif representations of the collection of words in each class are also shown (Figure 4.7). As expected, the words for which CTCF shows a stronger preference (1,024 and 575 in K562 cells and mouse ES cells respectively) closely match the *de novo* motif and together recapitulate the canonical consensus. Words bound with equal preference between CTCFL and CTCF (83 and 126) show similar results. On the other hand, motif words that are more highly bound by CTCFL (456 and 15,006) comprise a heterogenous mix of sequences, which, in ensemble, result in a relatively weak GC-rich summary motif with reduced information content.

Consistent with the above results, we find that CTCFL is enriched within low sequence complexity regions as defined by RepeatMasker^[301], specifically GC-rich repeats (Figure 4.8). Interestingly, CTCFL is particularly depleted at B2 repeats, which are retroelements responsible for rodent-specific expansions of CTCF binding^[117].

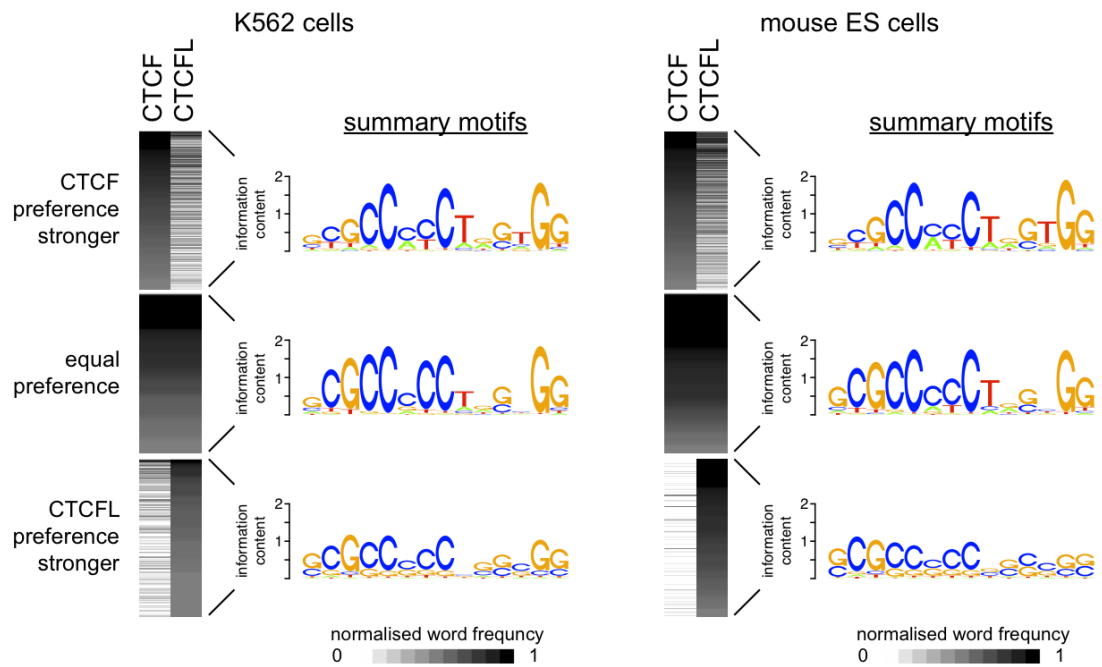


Figure 4.7: CTCFL binding preferences are stronger than CTCF for a diverse set of GC-rich motif words. Heatmaps show the normalised frequency of highly bound motif words (CTCF or CTCFL normalised frequency ≥ 0.5) occurring ≥ 5 times within CTCF or CTCFL ChIP-seq peaks. Summary motifs derived from the corresponding motif word subsets are shown.

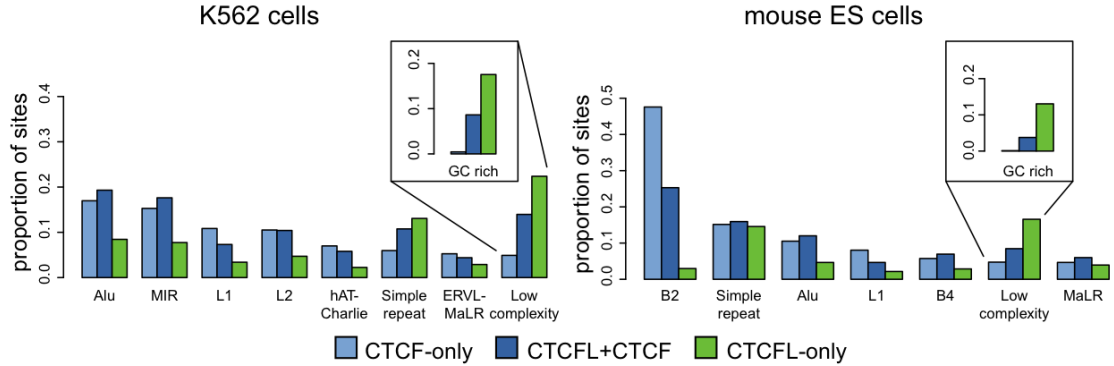


Figure 4.8: CTCFL binding is enriched in regions with high GC content and low DNA sequence complexity. Association of CTCF-only, CTCFL+CTCF and CTCFL-only binding events with repeat families are shown. Only repeat families with at least 5% overlap with at least one binding event class are shown.

4.3.3 CTCFL-regulated genes are enriched for active, CGI promoters, imprinting and relevance to cancer

As suggested by Sleutels *et al.*^[175], the localisation of CTCFL to active promoters could be a result of “opportunistic” binding i.e. largely due to the generally permissive, nucleosome-depleted chromatin status of these regions. CTCFL binding in this scenario is expected to be of little or no consequence for gene expression. Alternatively, such binding could reflect functional CTCFL-promoter interactions that affect the expression of the corresponding target genes. To distinguish between these two possibilities, we depleted endogenous CTCFL from K562 cells and enforced CTCFL expression in mouse ES cells, and quantified the associated changes in global gene expression.

Acute depletion of endogenous CTCFL in K562 cells resulted in the deregulation of 1,037 transcripts ($FDR = 0.05$; 636 down, 401 up; Figure 4.9; Figure 4.4A). Genes that bind CTCF, RAD21 and CTCFL are preferentially deregulated, as are genes with active, CpG island-containing (CGI) promoters^[299] (Figure 4.10). Perturbed genes in response to CTCFL knockdown are also enriched for putative house-keeping functions^[302], which is unsurprising considering that the vast majority of these genes have CGI promoters (92.8%). On the other hand, only four testis-specific genes and 74 genes

with tissue-specific patterns of expression were deregulated, neither of which correspond to a significant enrichment.

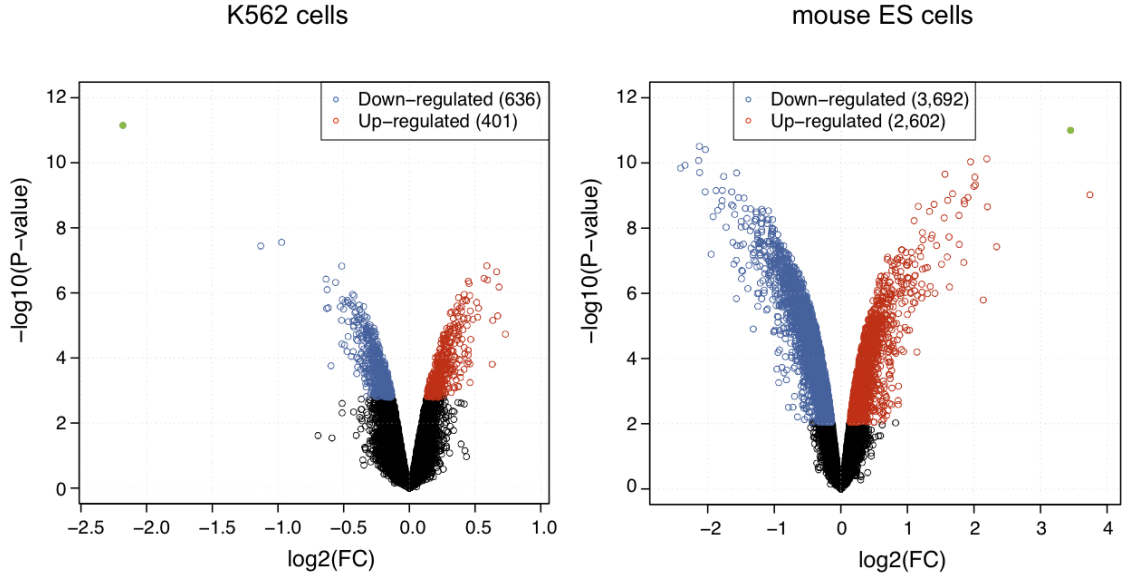


Figure 4.9: Impact of CTCFL on gene expression in K562 cells and ES cells. Volcano plots of microarray expression results. *Left:* Of the 32,321 transcripts assayed in siCTCFL K562 cells, 1,037 are significantly differentially expressed ($FDR = 0.05$), 401 of which are up-regulated (red) and 636 down-regulated (blue). *Right:* Of the 34,760 transcripts assayed, 6,294 are significantly differentially expressed in FLAG-CTCFL mouse ES cells ($FDR = 0.05$), 2,602 of which are up-regulated (red) and 3,692 down-regulated (blue). The green dots represent *Ctcfl* in each cell type respectively.

Stable expression of FLAG-CTCFL in mouse ES cells resulted in CTCFL binding at 12,891 active promoters (Figure 4.5) and CTCFL-bound genes are highly enriched among the 4,064 differentially expressed genes (Figure 4.10). Similarly to K562 cells, deregulated genes in mouse ES cells are enriched for CGI promoters, CTCF and RAD21 promoter-proximal binding, as well as house-keeping functions. We also tested a set of recently identified imprinted mouse genes^[303] and a set of putative human oncogenes^[304] for preferential deregulation by CTCFL. Using orthologs (obtained from Ensembl version 66^[206]) in human and mouse respectively, we find that these

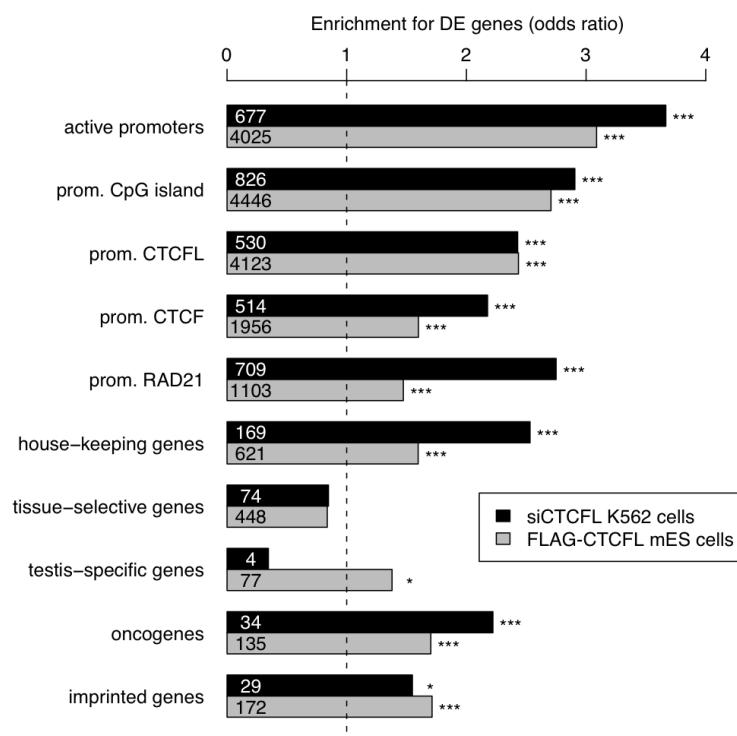


Figure 4.10: Impact of CTCFL on gene class expression in K562 cells and ES cells. Gene sets deregulated in siCTCFL K562 cells and FLAG-CTCFL mouse ES cells. Association of the indicated gene sets with membership in the list of genes significantly differentially expressed in each cell type was assessed using Fishers Exact Test (bars represent odds ratio; $*P < 0.05$, $***P < 0.001$). Genes with active promoters were defined as those with both H3K4me3 and RNAP2 ChIP-seq peaks within 2.5 kb of the TSS. Gene sets were also defined based on promoter overlap ($TSS \pm 2.5$ kb) with a CpG island^[299], as well as with either CTCFL, CTCF and RAD21 ChIP-seq peaks. The number of deregulated genes in each class is indicated at the base of the bars.

classes of genes are sensitive to CTCFL concentration in both of these systems (Figure 4.10). Validation by qRT-PCR confirmed differential expression of selected imprinted and cancer-related genes (Figure 4.4C). We conclude from these findings in K562 and mouse ES cells that CTCFL regulates many of the genes that it binds to, and that CTCFL binding to promoter-proximal sites is functional.

4.3.4 The relationship between CTCFL and cohesin binding

Following from the observation that cohesin (RAD21) is enriched at the promoters of genes deregulated by CTCFL, particularly in K562 cells (Figure 4.10), we decided to test the association of these genes with previously defined candidate genes for cohesin-mediated regulation in mouse ES cells^[169]. Kagey *et al.*^[169] used RNAi to decrease the expression of the cohesin subunit SMC1A in mouse ES cells, however five days after knockdown the expression of most ($> 10,000$) tested genes was affected. In an attempt to enrich for likely targets, the authors restricted this set of genes to those additionally affected by RNAi-mediated knockdown of the cohesin loading factor NIPBL and the Mediator subunit MED12, which has been shown to interact with cohesin^[169]. Similarly to the authors, we further defined direct targets as gene promoters also bound by cohesin (RAD21), NIPBL and Mediator (MED12 or MED1) as assayed by ChIP-seq. These criteria define a set of 224 potential cohesin-regulated genes, which overlap significantly with genes regulated by CTCFL in mouse ES cells (Fisher's Exact Test $P < 10^{-15}$, odds ratio=3.26). Interestingly, we find an even stronger association with a more comprehensive set of 964 genes significantly affected by acute genetic depletion of the cohesin subunit RAD21 in mouse ES cells (Fisher's Exact Test $P < 10^{-15}$, odds ratio=4.22; unpublished data).

To directly investigate the impact of CTCFL on cohesin binding we next carried out ChIP-seq experiments to map the cohesin subunit RAD21 in FLAG-CTCFL mouse ES cells. Comparison of cohesin maps showed a moderate reduction in the number of RAD21 peaks from 28,444 RAD21 peaks in WT ES cells to 22,711 RAD21 peaks in FLAG-CTCFL ES cells. The complete loss of RAD21 peaks from gene promoter regions is significantly associated with differential expression of the corresponding genes (Fisher's Exact Test $P < 10^{-15}$, odds ratio=1.76). Focussing on promoter-proximal

binding events, we find that although cohesin recruitment is only marginally affected at CTCF-only sites, co-bound CTCFL+CTCF, and CTCFL-only sites in particular, show markedly reduced cohesin binding in FLAG-CTCFL mouse ES cells (Figure 4.11). Distal RAD21 binding events are affected to a lesser degree (Figure 4.12). Reduced cohesin binding was confirmed by ChIP-qPCR at a selection of CTCFL target regions (Figure 4.13).

4.4 Discussion

By comparing genome-wide ChIP-seq profiles of CTCFL and CTCF with maps of chromatin state in K562 and mouse ES cells, we show that the paralogs have overlapping yet far from identical binding preferences, despite their highly conserved DNA binding domains.

Our results indicate that CTCFL binds to the majority of active promoters and, in contrast to CTCF, shows greater diversity in its preferred DNA sequence context. While CTCFL's binding site motif is almost indistinguishable from the canonical CTCF motif and the ChIP signal of both paralogs is stronger at regions where they co-occur (Figure 4.3 and 4.11), independent CTCFL binding seems to be governed by a different set of "rules". Apart from CTCF, CTCFL localises within regions containing clusters of relatively low-scoring motifs, which points towards mechanisms such as cooperativity (with itself or with other cofactors) to promote binding to these otherwise sub-optimal and potentially lower-affinity sequences. Evidence that this binding is indeed functional – as opposed to opportunistic – is provided by the fact that CTCFL is highly enriched within the promoters of genes that it regulates (Figure 4.10). The depletion of CTCFL-only sites within B2 and Alu repeats, which are both lineage-specific transposable elements (short interspersed elements, SINEs) responsible for reorganising regulatory landscapes in mouse and human respectively, may reflect distinct mechanisms of binding site evolution for CTCF and CTCFL. The binding of CTCF to B2 retroelements may protect them from the silencing effects of DNA methylation and thereby constitute an escape strategy^[52]. However, the depletion of CTCFL at these sites, together with its enrichment within promoter CGIs, supports the idea that CTCF and CTCFL differ in terms of their effect on methylated DNA and perhaps also their respective contribution to transposable element retention.

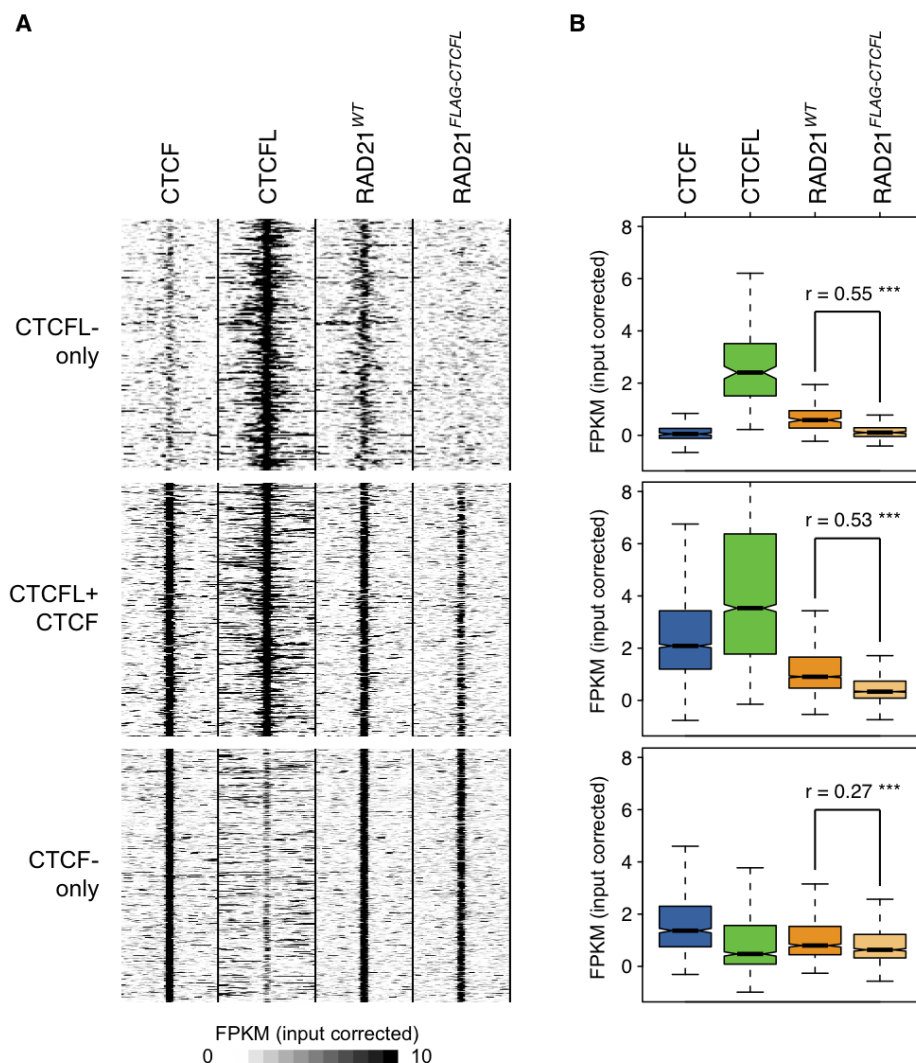


Figure 4.11: (A) Heatmap representation of ChIP-seq fragment profiles indicates preferential RAD21 depletion at promoter-proximal CTCFL binding events in FLAG-CTCFL mouse ES cells. Profiles consist of input corrected FPKM for the indicated ChIP-seq datasets calculated at 100 bp resolution and shown within a 5 kb window centred on either the CTCFL or CTCF peak summit position. Only profiles associated with promoter-proximal binding events are shown ($TSS \pm 2.5$ kb). (B) Boxplots showing ChIP enrichment for the indicated ChIP-seq datasets corresponding to the windows described in A. Wilcoxon signed-rank test effect size (r) and significance are indicated for the difference between RAD21 in WT and FLAG-CTCFL cells ($***P < 0.001$).

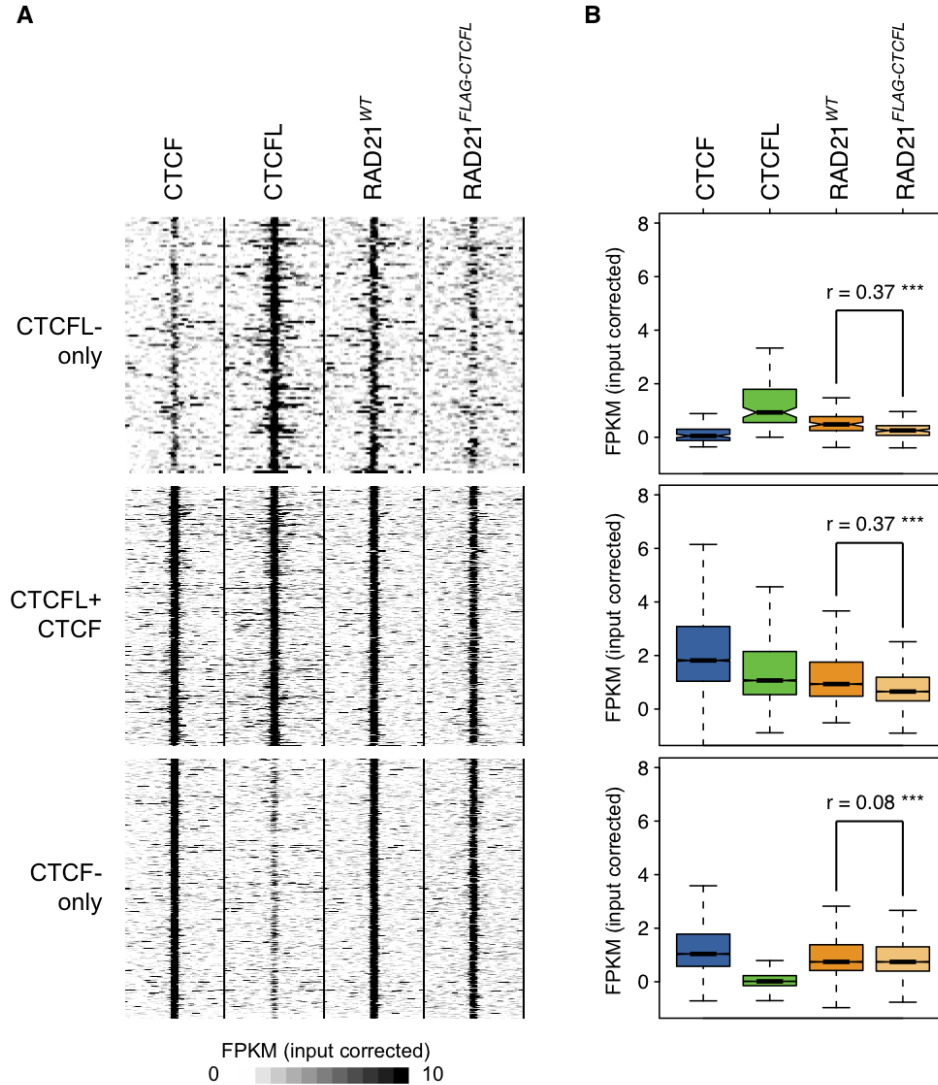


Figure 4.12: (A) Heatmap representation of ChIP-seq fragment profiles indicates RAD21 depletion at promoter-distal CTCFL binding events in FLAG-CTCFL mouse ES cells. Profiles consist of input corrected FPKM for the indicated ChIP-seq datasets calculated at 100 bp resolution and shown within a 5 kb window centred on either the CTCFL or CTCF peak summit position. Only profiles associated with promoter-distal binding events are shown (> 2.5 kb). (B) Boxplots showing ChIP enrichment for the indicated ChIP-seq datasets corresponding to the windows described in A. Wilcoxon signed-rank test effect size (r) and significance are indicated for the difference between RAD21 in WT and FLAG-CTCFL cells (*** $P < 0.001$).

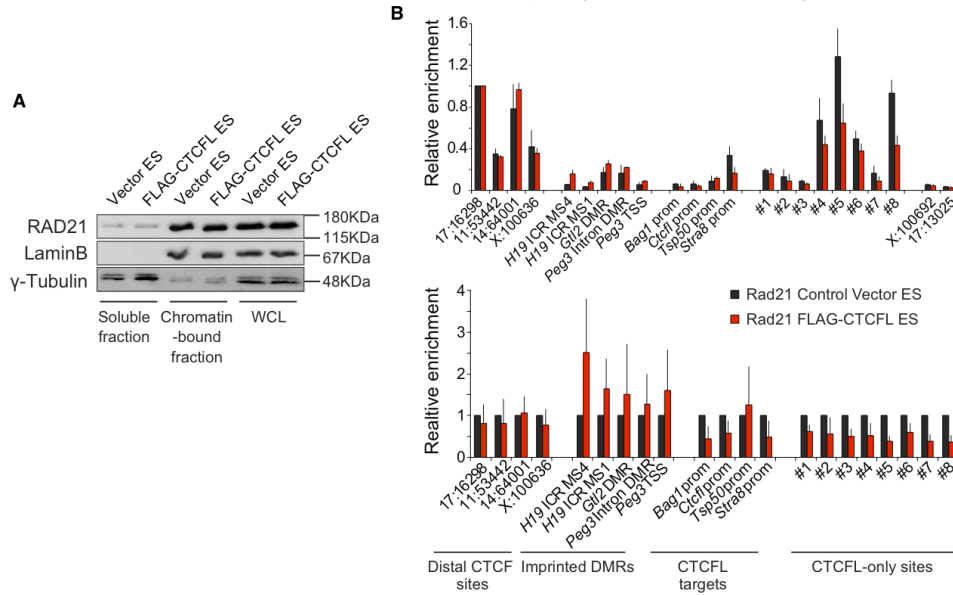


Figure 4.13: Validation of reduced RAD21 binding at CTCFL target sites. (A) CTCFL does not alter the expression of RAD21 protein or its association with chromatin in FLAG-CTCFL ES cells. Whole cell lysates of control vector and FLAG-CTCFL ES cells were fractionated into soluble and chromatin-bound and analysed by immunoblotting for RAD21, γ -tubulin (soluble) and Lamin B (nuclear insoluble). (B) The enrichment of RAD21 at specific loci was assayed by ChIP-qPCR in ES cells stably transfected with control vector (black) and FLAG-CTCFL (red). Sites include promoter-distal CTCF-RAD21 sites, differentially methylated regions (DMRs) of imprinted loci, known CTCFL target loci, promoter-proximal CTCFL-only sites identified in this study as well as the negative control sites chrX:100692 kb and chr17:13025 kb. Above: data normalised to the chr17:16298 kb CTCF-only binding site (mean \pm SD, n=3). Below: same data normalised to control vector ES cells to show the enrichment or depletion of RAD21 at each locus in FLAG-CTCFL ES cells.

Surprisingly, the number of highly bound CTCFL/CTCF motif words is more than an order of magnitude greater in mouse ES cells than in K562 cells. This disparity is largely attributable to additional CTCFL-only bound words in mouse ES cells and may be related to the many more CTCFL-only peaks in this cell type. We speculate that CTCFL binding in these regions, and in turn the extent of bound motif word diversity, may be concentration-dependent. Dose-dependent effects of CTCFL on gene expression, chromatin organisation and cancer cell viability have been demonstrated previously^[305,306].

The binding of CTCFL to GC-rich repeats and its preferential perturbation of genes with promoter CpG islands suggests that CTCFL may play a role in determining or maintaining the epigenetic status of these regions, which are known sites of dynamic transcriptional regulation by DNA methylation in the germline^[307]. The striking association of CTCFL with H3K4me2 in K562 cells (Figure 4.5) is much reduced in mouse ES cells where this modification was assayed before FLAG-CTCFL transfection, indicating that CTCFL may be involved in the deposition of this mark. Interestingly, SET1A, a H3K4 methyltransferase, has been shown to interact with CTCFL and co-bind targets upstream of *Myc* and *BRCA1*, thereby increasing local H3K4me2 and expression of these oncogenes^[178]. Additionally, there is evidence for a relationship between H3K4me2 and the protection of inactive unmethylated CpG islands from DNA methylation^[308]. CTCFL expression in cancer cells as well as in developing germ cells is correlated with global genome-wide changes in DNA methylation – physiological and aberrant respectively. Therefore, our findings that CTCFL tends to bind and regulate many essential house-keeping and imprinted genes as well as those with oncogenic potential is consistent with a role for CTCFL in the epigenetic control of transcription. Importantly, these results are consistent between the two somatic cell types in our study, supporting their generality.

In addition to CTCF sites, cohesin is also found at active enhancers and promoters, where its binding is associated with tissue-specific transcription factors and Mediator components^[167,169,238]. Together, these two classes of cohesin binding events are thought to form a network of long-range interactions that reflects and promotes cell-type-specific gene expression programmes^[52]. Cohesin binding to active genes has been linked to transcriptional regulation^[169,170] and an unexpected finding of our study is that CTCFL binding reduced the recruitment of cohesin to promoter-proximal regions.

Reduced cohesin recruitment is not explained by altered cohesin expression at RNA or protein level, or a change in the fraction of chromatin-associated cohesin. Although CTCFL expression does not (or only marginally) affect cohesin recruitment to CTCF-only sites, both co-bound CTCF+CTCFL and CTCFL-only sites show significantly reduced cohesin binding in FLAG-CTCFL ES cells.

The eviction of cohesin from promoter-proximal sites represents an interesting mechanism for CTCFL-mediated gene expression that is expected to significantly impact the biology of cells expressing the protein. This novel function adds to the catalogue of potential antagonistic properties of CTCF and CTCFL. It remains to be explored whether other particular peculiarities of CTCFL expression, such as its dynamic cellular localisation^[176,182] and multiple splice isoforms^[309], influence its ability to impact CTCF binding and cohesin recruitment.

4.5 Methods

4.5.1 Experimental methods

The experiments described in this section were performed by Hegias Mira-Bontenbal.

4.5.1.1 ChIP sequencing

Methods for RT- and genomic PCR, and ChIP-seq for RAD21, CTCF and FLAG-CTCFL were carried out as previously described^[170], except elution of the FLAG ChIP was done using a 3xFLAG elution peptide (0.2 μ g/ μ l final concentration), 30min 4°C, twice, then proceeding to normal reverse cross-linking as with the other ChIP and Input samples.

4.5.1.2 Microarray experiments

For K562 experiments, 10⁶ cells were transfected with CTCFL siRNAs (four individual siGENOME, Dharmacon) using DharmFECT reagent 4 following the manufacturers instructions (Dharmacon). SiGENOME Non-Targeting siRNA #2 was used as Control. Total RNA and protein samples were generated from three independent experiments. Knockdown was confirmed by immunoblot 48h after siRNA transfection. Microarray samples were prepared following Affymetrix instructions and a GeneChip Human Gene

1.0 ST Array was used. For mouse ES cell experiments, microarray samples of three biological replicates of total RNA of Vector Only mouse ES and FLAG-CTCFL ES cells were prepared following Affymetrix instructions and a GeneChip Mouse Gene 1.0 ST Array was used.

4.5.2 Computational methods

4.5.2.1 ChIP-seq read mapping and peak calling

Raw read alignment, filtering and peak calling for CTCFL, CTCF, H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K79me2, EP300, RNAP2 and RAD21^[15] in K562 cells (GRCh37/hg19); H3K4me1, H3K4me2^[56], H3K4me3, H3K9me3, H3K27me3, H3K36me3^[54], H3K79me2^[248], EP300^[241] and RNAP2^[249] in mouse ES cells (NCBI37/mm9) and the definition of CTCFL-non-CTCF (CTCFL-only) sites was done as previously described^[238] (see Section 2.5.2).

The computational analysis described in this paragraph was performed by Adam Giess. For CTCF, FLAG-CTCFL, RAD21 in WT ES cells and RAD21 in FLAG-CTCFL ES cells, raw reads were aligned with Eland (Illumina GA pipeline version 1.4.1; NCBI37/mm9). Peaks were called using SWEml with `-R 0.01`, only uniquely mapping reads and fragment length estimates (`-f`) of 110 (FLAG-CTCFL), 121 (CTCF), 108 (RAD21^{WT}) and 130 (RAD21^{FLAG-CTCFL}).

4.5.2.2 Motif analysis

De novo motif finding and selection for CTCF and CTCFL/FLAG-CTCFL were carried out as previously described^[238] (see Section 2.5.2). Similarly, motif score cut-offs corresponding to $FDR = 0.4$ were chosen to determine motif presence/absence. For motif word analysis, the *de novo* motifs for CTCF and CTCFL were used to scan their respective peak regions to identify all 14-mer motif occurrences bound at least five times by either CTCF or CTCFL. We then estimated the genomic frequency of all words using Nucleotide-Nucleotide BLAST version 2.2.26+ (`blastn`) and only retaining perfect full-length matches: `-value 20 -word.size 7 -reward 1 -penalty -3 -gapopen 5 -gapextend 2 -dust no`. The normalised word frequency was calculated for each word by dividing the number of bound occurrences by the total number of occurrences in the genome. Limiting the analysis to only highly bound words

(normalised word frequency ≥ 0.5), words were grouped into three classes based on their relative CTCF/CTCFL frequencies i.e. $CTCF = CTCFL$ (equal preference), $CTCF > CTCFL$ (CTCF preference stronger), $CTCFL > CTCF$ (CTCFL preference stronger). Summary PWMs and associated motifs were constructed using the Bioconductor R package seqLogo version 1.14.0.

4.5.2.3 Microarray analysis

Microarray datasets for both siCTCFL K562 cells and FLAG-CTCFL mouse ES cells were processed using the Bioconductor R package vsn version 3.16.0^[220]. Differential expression was determined at the transcript level using the Bioconductor R package limma version 3.4.4^[221,222], which employs empirical Bayes methods to borrow information between genes. Significant differential expression of at least one transcript was defined as sufficient for differential expression of the associated gene. For this reason genes can be simultaneously classified as significantly up- and down-regulated (one gene in K562 cells, 19 genes in mouse ES cells).

Chapter 5

Conclusions and future work

In this thesis I have investigated key DNA-associated proteins with closely related roles in the control of mammalian chromatin architecture and gene expression. The three independent studies presented here rely on different experimental and computational approaches, but share a number of similarities. These include *in vivo* maps of cohesin and CTCF binding in different mouse cell types (liver, thymus, ES) and results with implications for genome-wide chromatin organisation. Common computational themes include the use of machine learning techniques as well as established and novel visualisation methods to navigate these high-dimensional datasets, thereby helping to generate hypotheses for follow-up experiments.

Methods in computational biology can be broadly classified into two categories: hypothesis-driven and so-called “hypothesis-free” (exploratory) approaches. The former – embodied by the statistical hypothesis test – represent the classical approach to scientific research, where data is collected or analysed for the express purpose of proving or disproving a predefined theory. On the other hand, the aim of the latter is to discover previously unknown properties of biological systems using machine learning techniques such as unsupervised clustering, *de novo* sequence motif discovery and PCA. A more appropriate term to describe these explorative approaches is “hypothesis-relaxed” as they tend to rely on general, implicit hypotheses about the presence of structure in the data. Both approaches were used in analyses presented in this thesis, but there are clear benefits of exploratory methods when confronted with high-dimensional datasets. Clustering can aid visualisation by summarising genome-wide data in an easily-interpretable manner. This in turn allows for (human) domain expert-based pattern discovery and the generation of novel hypotheses, which can then

be evaluated formally using statistical hypothesis tests. In this way, these two approaches can be combined into a sequential knowledge discovery “pipeline” consisting of hypothesis-generation (exploratory) and hypothesis-testing phases.

In Chapter 2 I describe results obtained using such a procedure. Insights from the integration of a large catalogue of genome-wide data comprising TF binding, chromatin state and gene expression, further define the role of cohesin at loci apart from CTCF. Here cohesin preferentially colocalises with clusters of TFs that possess signatures of active transcriptional regulation. Evidence that cohesin stabilises the binding of these TFs to tissue-specific CRMs is supported by comparisons to data obtained from mouse liver cells with only one functional copy of a gene encoding a crucial cohesin subunit (*Rad21*). In Chapter 3 I present results from the analysis of experiments conducted in a developing mouse thymocyte system free from cell division-related biases, where the same subunit was subject to conditional homozygous genetic deletion. This enabled the genome-wide study of cohesin’s proven ability to facilitate chromatin contacts between CTCF binding events – and others – once again in the context of cell-type-specific transcriptional regulation. Here, results from the computational analysis of maps of global chromatin conformation in both normal and cohesin-deficient cells, support a model where cohesin collaborates with CTCF to constrain functional interactions within active compartments.

In Chapter 4 I describe results from the study of enforced CTCFL expression in mouse ES cells, which suggest that the protein can perturb the relationship between CTCF, cohesin and chromatin. CTCFL binding correlates with widespread gene expression changes and comparisons to results from knockdown experiments in a human cancer cell line (K562) indicate common properties of affected targets. CTCFL misexpression results in preferential deregulation of active, house-keeping genes with promoter-associated cohesin in both cell types, but particularly so in cancer cells. Coupled with preliminary results showing cohesin depletion at CTCFL-bound sites in ES cells, this points toward a possible mechanism of action i.e. transcriptional regulation by disrupting cohesin recruitment. Cohesin deregulation is associated with pleiotropic effects due to its diverse functions, and therefore it is revealing that enforced CTCFL expression, as well as silencing, results in large numbers of both up- and down-regulated genes in different cell types.

Results from the study of ectopic CTCFL in somatic cells may help to understand the role of transient CTCFL expression in germ cells. Developing sperm cells undergo global chromatin reorganisation involving the near complete replacement of histones with protamines, which dramatically condense DNA into toroidal structures and are essential for normal sperm functioning^[310]. However, a subset of promoters retain their modified nucleosomes in sperm cells and the corresponding genes are enriched in specialised functions related to embryonic development^[311]. Apart from CTCFL’s hypothesised role in the regulation of global DNA methylation, further investigation into this alternative epigenetic mechanism seems promising. Also, in view of cohesin’s lately proposed role in the inheritance of accessibility^[312], it would be interesting to explore whether CTCFL is capable of interfering in this process, thereby helping to reset patterns of open chromatin in the germline.

Chapters 2 and 3 relate to seemingly distinct putative functions for cohesin – stabilisation of TF binding and stabilisation of chromatin interactions respectively – the reconciliation of which is an important consideration. Firstly, the former may be directly related to chromatin topology, where looping interactions between distinct regulatory elements (in *cis* or *trans*) may facilitate indirect binding of TFs to chromatin at mutually contacting sites. In addition to reduced contacts between CTCF binding events, cohesin depletion in non-dividing thymocytes also decreases interactions between tissue-specific enhancers, as well as heterotypic enhancer-cohesin and enhancer-CTCF interactions. Indeed, the link between chromatin looping events, accessibility and “phantom” signals in ChIP-seq (indicative of indirect binding events) has recently been explored in terms of cotranscriptional splicing^[313].

Secondly, these potentially unrelated roles may be performed during distinct stages of the cell cycle. Cohesin itself is subject to multiple levels of cell cycle-dependent regulation, involving differential loading (NIPBL), unloading (WAPL), stabilisation (sororin), post-translational modification (SMC3 acetylation by ECO1) and cleavage (separase). Although the experiments in thymocytes used cells arrested in interphase and, likewise, the vast majority of adult hepatocytes are non-proliferating^[314], the latter are likely to have suffered from cell cycle-related effects during prior development in the case of *Rad21* haploinsufficiency. Similar to results presented in this work, the authors of a recently published study involving the most comprehensive ChIP-seq profiling of TFs in a single cell type to date^[312] (human colorectal cancer LoVo cells), ob-

serve CTCF-independent cohesin binding together with almost all TF clusters (CRMs). Consistent with a role in the stabilisation of TF binding, the authors also provide evidence that cohesin promotes chromatin accessibility and TF recruitment, suggesting that cohesin persistence at these sites during mitosis (when most TFs dissociate from chromatin) serves as an epigenetic “bookmark”. In this model, cohesin ensures stable TF binding across cell division events – possibly involving contacts in *trans* between homologous CRMs – but this does not exclude an interphase role in the stabilisation of transcriptional regulatory contacts in *cis*.

There are a number of promising avenues for further purely computational investigations of themes established in this thesis. A publicly available high resolution Hi-C dataset from human fibroblasts^[315] could be used to independently test hypotheses from Chapter 2. For example, the presence of mirrored CTCF binding events near transcription start sites and cohesin-bound enhancers is predicted to be associated with “bridging” interactions leading to elevated expression levels of the corresponding target genes. The enhancer-promoter interactions identified in this Hi-C dataset could also be used to determine whether putative indirect TF binding – inferred from the absence of motifs – is indeed associated with chromatin looping events. Furthermore, incorporating data describing dynamic chromatin state, temporal gene expression changes during development^[281] and TADs identified using the existing thymocyte Hi-C data is likely to improve the regression model described in Chapter 3.

During interphase, CTCF’s role is tightly linked to that of cohesin and it would be valuable to determine what proportion of co-bound sites facilitate chromatin looping, enhancer and/or insulator functions. Studying the effects of artificially introduced CTCF motifs on chromatin architecture at different genomic loci could have implications for our understanding of the evolution of its binding sites. Although CTCF-independent cohesin binding is investigated in this thesis, the functions of the low number of cohesin-independent CTCF sites is still unknown. Similarly, other combinations involving cohesin-associated proteins may perform specific, as yet undefined, functions. A recent study in this direction attempted to disentangle subclasses of CTCF, cohesin and mediator binding events by correlating these results with chromatin loops identified by 5C in ES and neural progenitor cells^[286]. The high resolution 5C assay, as well as the reduced representation ChIA-PET method, may also help to answer questions about the connectivity of particular regulatory elements in the genome, i.e. how fre-

quently enhancers, promoters and insulators connect with each other and themselves to coordinate their effects on target genes.

Although unbiased (genome-wide) high resolution chromatin conformation assays are currently impractical, it is now possible to detect protein-DNA interactions at high resolution. ChIP-exo enables TF binding events to be determined at close to single base pair resolution and is expected to trivialise tasks like motif discovery, indirect binding event identification and integration with nucleotide-level DNA methylation datasets^[316]. Such assays could be used to address open questions regarding CTCFL binding and motif distribution, as well as the possibly indirect nature of TF binding within cohesin-occupied CRMs. Other techniques that will benefit research of high-order chromatin structure are assays to determine RNA-chromatin interactions, as demonstrated by RNA antisense purification (RAP) recently used to map Xist lncRNA binding locations in the context of X-chromosome inactivation^[317]. The list of ncRNAs involved in controlling chromatin architecture is growing and a possible relationship between enhancer-bound cohesin and the generation of enhancer RNAs^[318] (eRNAs) with topological functions is an intriguing one.

More broadly, with the plummeting cost of DNA sequencing and emergence of microfluidics technologies comes the promise of potentially measuring the state of entire cell populations or tissues at single-cell resolution^[319]. Apart from bringing nearer into view the seemingly impossible task of constructing realistic bottom-up computational models of whole organisms, ChIP-seq and 3C-based assays of individual cells would enable researchers to better understand interaction signals. Current approaches relying on population averages cannot distinguish whether signal differences are attributable to alterations in interaction frequency (across all cells) or to cell-to-cell heterogeneity. Progress in this direction has been made by a recent study that mapped the conformation of X chromosomes in single cells by Hi-C, confirming the existence of megabase-sized domains despite high levels of cell-to-cell variability at larger scales^[320]. Combining these types of experiments with single-cell gene expression measurements will be a promising approach to study the contribution of cell-specific chromatin interactions to inter-cell gene expression variability, which may play a role in autonomous cell-fate decisions^[321].

Lastly, as high-throughput technologies relentlessly improve in terms of speed and

cost, perceptions of wastefulness towards the collection of enormous quantities of biological data will likely continue to be challenged by results from increasingly sophisticated hypothesis-generation methods, which naturally complement hypothesis-driven (or hypothesis-limited^[322]) approaches. These methods will be crucial to interpret future maps of genome state at higher resolutions in space and time, thus enabling more complete functional characterisations of key factors involved in the organisation of chromatin structure.

Publications

The following are publications arising from this thesis:

- Faure, A. J., Schmidt, D., Watt, S., Schwalie, P. C., Wilson, M. D., Xu, H., Ramsay, R. G., Odom, D. T., and Flicek, P. (2012). **Cohesin regulates tissue-specific expression by stabilising highly occupied *cis*-regulatory modules.** *Genome Research*.
- Seitan, V., Faure, A. J., Zhan, Y., McCord, R. P., Lajoie, B. R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A., Flicek, P., Dekker, J., Merckenschlager, M. (2013). **Cohesin-based chromatin interactions enable regulated gene expression within pre-existing architectural compartments.** *Genome Research*.

References

- [1] Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921. 1
- [2] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., et al. (1996). Life with 6000 Genes. *Science (New York, NY)*, 274(5287):546–567. 1
- [3] The Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408(6814):796–815. 1
- [4] C. elegans Sequencing Consortium (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science (New York, NY)*, 282(5396):2012–2018. 1
- [5] Adams, M. D. (2000). The Genome Sequence of *Drosophila melanogaster*. *Science (New York, NY)*, 287(5461):2185–2195. 1
- [6] Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–197. 2
- [7] Ideker, T., Galitski, T., and Hood, L. (2001). A NEW APPROACH TO DECODING LIFE: Systems Biology. *Annual review of genomics and human genetics*, 2(1):343–372. 2
- [8] Petrov, D. A. (2001). Evolution of genome size: new approaches to an old problem. *Trends in Genetics*, 17(1):23–28. 2
- [9] Sears, K. E., Behringer, R. R., Rasweiler, J. J., and Niswander, L. A. (2006). Development of bat flight: morphologic and molecular evolution of bat wing digits. *Proceedings of the National Academy of Sciences*, 103(17):6581–6586. 2
- [10] Roeder, R. G. and Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, 224(5216):234–237. 3
- [11] Roeder, R. G. (2005). Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS letters*, 579(4):909–915. 3, 4, 5
- [12] Fuda, N. J., Ardehali, M. B., and Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–192. 3, 4

-
- [13] Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70. 4, 70
- [14] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263. 4
- [15] The ENCODE Project Consortium (2011). A User’s Guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biology*, 9(4). 4, 105, 123
- [16] Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., et al. (2009). Unlocking the secrets of the genome. *Nature*, 459(7249):927–930. 4
- [17] Saunders, A., Core, L. J., and Lis, J. T. (2006). Breaking barriers to transcription elongation. *Nature reviews Molecular cell biology*, 7(8):557–567. 5
- [18] Kim, Y. J., Björklund, S., Li, Y., Sayre, M. H., and Kornberg, R. D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell*, 77(4):599–608. 5
- [19] Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., et al. (2010). Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science (New York, NY)*, 328(5981):1036–1040. 5, 44, 52, 58, 66
- [20] Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics*, 42(7):631. 5, 22
- [21] Schones, D. E. and Zhao, K. (2008). Genome-wide approaches to studying chromatin modifications. *Nature Reviews Genetics*, 9(3):179–191. 7
- [22] 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 (New York, NY)*, 326(5950):289–293. 7, 17, 19, 22, 37, 39, 40, 77, 99
- [23] Nelson, H. C., Finch, J. T., Luisi, B. F., and Klug, A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature*, 330(6145):221–226. 6
- [24] Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778. 6
- [25] Choi, J. K. and Kim, Y.-J. (2009). Intrinsic variability of gene expression encoded in nucleosome positioning sequences. *Nature Genetics*, 41(4):498–503. 6
- [26] Branco, M. R., Ficz, G., and Reik, W. (2012). Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews Genetics*, 13(1):7–13. 6

-
- [27] Bird, A. (2011). The dinucleotide CG as a genomic signalling module. *Journal of molecular biology*, 409(1):47–53. 6
- [28] Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R. W., Deaton, A., Andrews, R., James, K. D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, 464(7291):1082–1086. 8
- [29] Tate, P. H. and Bird, A. P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. *Current opinion in genetics & development*, 3(2):226–231. 8
- [30] Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature Genetics*, 33 Suppl:245–254. 8
- [31] Riggs, A. D. and Pfeifer, G. P. (1992). X-chromosome inactivation and cell memory. *Trends in Genetics*, 8(5):169–174. 8
- [32] Smallwood, S. A. and Kelsey, G. (2012). De novo DNA methylation: a germ cell perspective. *Trends in Genetics*, 28(1):33–42. 8
- [33] Jones, P. A. and Baylin, S. B. (2007). The Epigenomics of Cancer. *Cell*, 128(4):683–692. 9
- [34] Gaffney, D. J., McVicker, G., Pai, A. A., Fondufe-Mittendorf, Y. N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J. K. (2012). Controls of nucleosome positioning in the human genome. *PLoS Genetics*, 8(11):e1003036. 9
- [35] Liu, X., Lee, C.-K., Granek, J. A., Clarke, N. D., and Lieb, J. D. (2006). Whole-genome comparison of Leu3 binding in vitro and in vivo reveals the importance of nucleosome occupancy in target site selection. *Genome Research*, 16(12):1517–1528. 10
- [36] Mohrmann, L. and Verrijzer, C. P. (2005). Composition and functional specificity of SWI2/SNF2 class chromatin remodeling complexes. *Biochimica et biophysica acta*, 1681(2-3):59–73. 10
- [37] Fu, Y., Sinha, M., Peterson, C. L., and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genetics*, 4(7):e1000138. 10
- [38] Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature*, 474(7352):516–520. 10
- [39] Yuan, G.-C., Liu, Y.-J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J., and Rando, O. J. (2005). Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science (New York, NY)*, 309(5734):626–630. 10
- [40] Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*, 132(2):311–322. 10

-
- [41] Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R., and Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Research*, 17(6):877–885. 10, 31
- [42] Auerbach, R. K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrançois, P., Struhl, K., Gerstein, M., and Snyder, M. (2009). Mapping accessible chromatin regions using Sono-Seq. *Proceedings of the National Academy of Sciences*, 106(35):14926–14931. 10, 31
- [43] Chen, X., Sabo, P. J., Sandstrom, R., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S., and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, 6(4):283–289. 10
- [44] Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21(3):447–455. 10
- [45] Tan, M., Luo, H., Lee, S., Jin, F., Yang, J. S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., et al. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, 146(6):1016–1028. 10
- [46] Hon, G., Wang, W., and Ren, B. (2009). Discovery and Annotation of Functional Chromatin Signatures in the Human Genome. *PLoS Computational Biology*, 5(11):e1000566. 10, 12, 13
- [47] Ernst, J. and Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8):817–825. 10
- [48] Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., and Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1):77–88. 10
- [49] Li, B., Carey, M., and Workman, J. L. (2007). The Role of Chromatin during Transcription. *Cell*, 128(4):707–719. 11, 12
- [50] Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21(3):381–395. 11
- [51] Sneppen, K., Micheelsen, M. A., and Dodd, I. B. (2008). Ultrasensitive gene regulation by positive feedback loops in nucleosome modification. *Molecular Systems Biology*, 4:182. 11
- [52] Merkenschlager, M. and Odom, D. T. (2013). CTCF and Cohesin: Linking Gene Regulatory Elements with Their Targets. *Cell*, 152(6):1285–1297. 11, 77, 117, 121
- [53] Zhou, V. W., Goren, A., and Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics*, 12(1):7–18. 11, 12, 13
- [54] Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560. 12, 106, 123

-
- [55] Fischle, W., Wang, Y., Jacobs, S. A., Kim, Y., Allis, C. D., and Khorasanizadeh, S. (2003). Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. *Genes & development*, 17(15):1870–1881. 12
- [56] Meissner, A., Mikkelsen, T., Gu, H., and Wernig, M. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*. 12, 106, 123
- [57] Zilberman, D., Coleman-Derr, D., Ballinger, T., and Henikoff, S. (2008). Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, 456(7218):125–129. 12
- [58] Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics*, 39(3):311–318. 12
- [59] Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936. 12
- [60] Giles, K. E., Gowher, H., Ghirlando, R., Jin, C., and Felsenfeld, G. (2010). Chromatin boundaries, insulators, and long-range interactions in the nucleus. *Cold Spring Harbor symposia on quantitative biology*, 75:79–85. 13
- [61] Goren, A. and Cedar, H. (2003). Replicating by the Clock. *Nature reviews Molecular cell biology*, 4(1):25–32. 13
- [62] Saurin, A. J., Shiels, C., Williamson, J., Satijn, D. P., Otte, A. P., Sheer, D., and Freemont, P. S. (1998). The human polycomb group complex associates with pericentromeric heterochromatin to form a novel nuclear domain. *The Journal of Cell Biology*, 142(4):887–898. 13
- [63] Sutherland, H. and Bickmore, W. A. (2009). Transcription factories: gene expression in unions? *Nature Reviews Genetics*, 10(7):457. 13, 21
- [64] Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403(6765):41–45. 13
- [65] Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, 41(2):827–841. 13
- [66] Hübner, M. R., Eckersley-Maslin, M. A., and Spector, D. L. (2013). Chromatin organization and transcriptional regulation. *Current opinion in genetics & development*, 23(2):89–95. 13
- [67] Stefflova, K., Thybert, D., Wilson, M. D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D. J., Talianidis, I., Marioni, J. C., et al. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–540. 14

-
- [68] Noonan, J. P. and McCallion, A. S. (2010). Genomics of Long-Range Regulatory Elements. *Annual review of genomics and human genetics*, 11(1):1–23. 14
- [69] Bulger, M. and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–339. 14
- [70] Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell*, 10(6):1453–1465. 14
- [71] Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics*, 12(14):1725–1735. 15
- [72] Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. *Cell*, 147(5):1132–1145. 15, 16
- [73] Montavon, T. and Duboule, D. (2012). Landscapes and archipelagos: spatial organization of gene regulation in vertebrates. *Trends in cell biology*, 22(7):347–354. 15, 87
- [74] Hakim, O., Sung, M.-H., and Hager, G. L. (2010). 3D shortcuts to gene regulation. *Current opinion in cell biology*, 22(3):305–313. 15
- [75] Sexton, T., Bantignies, F., and Cavalli, G. (2009). Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation. *Seminars in cell & developmental biology*, 20(7):849–855. 16
- [76] Palstra, R.-J., Simonis, M., Klous, P., Brasset, E., Eijkelkamp, B., and de Laat, W. (2008). Maintenance of long-range DNA interactions after inhibition of ongoing RNA polymerase II transcription. *PLoS ONE*, 3(2):e1661. 16
- [77] Amano, T., Sagai, T., Tanabe, H., Mizushima, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Developmental Cell*, 16(1):47–57. 16
- [78] Phillips, J. E. and Corces, V. G. (2009). CTCF: master weaver of the genome. *Cell*, 137(7):1194–1211. 16, 22, 43, 95, 98
- [79] Sofueva, S. (2012). Cohesin-mediated chromatin interactions—into the third dimension of gene regulation. *Briefings in Functional Genomics*. 16
- [80] Tan-Wong, S. M., Zaugg, J. B., Camblong, J., Xu, Z., Zhang, D. W., Mischo, H. E., Ansari, A. Z., Luscombe, N. M., Steinmetz, L. M., and Proudfoot, N. J. (2012). Gene Loops Enhance Transcriptional Directionality. *Science (New York, NY)*. 16, 87
- [81] Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N. F., Horton, A., Andrews, S., Kukuruti, S., Mitchell, J. A., Umlauf, D., Dimitrova, D. S., et al. (2010). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nature Genetics*, 42(1):53–61. 16

-
- [82] Hu, Q., Kwon, Y.-S., Nunez, E., Cardamone, M. D., Hutt, K. R., Ohgi, K. A., Garcia-Bassets, I., Rose, D. W., Glass, C. K., Rosenfeld, M. G., et al. (2008). Enhancing nuclear receptor-induced transcription requires nuclear motor and LSD1-dependent gene networking in interchromatin granules. *Proceedings of the National Academy of Sciences*, 105(49):19199–19204. 16
- [83] Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., Poh, H. M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98. 16
- [84] Gheldof, N., Smith, E. M., Tabuchi, T. M., Koch, C. M., Dunham, I., Stamatoyannopoulos, J. A., and Dekker, J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. *Nucleic Acids Research*, 38(13):4325–4336. 17
- [85] Gibcus, J. H. and Dekker, J. (2013). The Hierarchy of the 3D Genome. *Molecular Cell*, 49(5):773–782. 17, 77
- [86] Sanyal, A., Lajoie, B. R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113. 17
- [87] Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*, 14(6):390–403. 18, 37, 38, 39
- [88] Zhang, Y., McCord, R. P., Ho, Y.-J., Lajoie, B. R., Hildebrand, D. G., Simon, A. C., Becker, M. S., Alt, F. W., and Dekker, J. (2012). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, 148(5):908–921. 17
- [89] Lin, Y. C., Benner, C., Mansson, R., Heinz, S., Miyazaki, K., Miyazaki, M., Chandra, V., Bossen, C., Glass, C. K., and Murre, C. (2012). Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate. *Nature immunology*, 13(12):1196–1204. 17, 39, 40, 41, 91, 101
- [90] Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell*, 148(3):458–472. 19
- [91] 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(7398):376–380. 19, 22, 95, 98
- [92] 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(7398):381–385. 19
- [93] Marshall, W. F., Straight, A., Marko, J. F., Swedlow, J., Dernburg, A., Belmont, A., Murray, A. W., Agard, D. A., and Sedat, J. W. (1997). Interphase chromosomes undergo constrained diffusional motion in living cells. *Current Biology*, 7(12):930–939. 19

-
- [94] Chubb, J. R., Boyle, S., Perry, P., and Bickmore, W. A. (2002). Chromatin motion is constrained by association with nuclear compartments in human cells. *Current Biology*, 12(6):439–445. 19
- [95] Baù, D., Sanyal, A., Lajoie, B. R., Capriotti, E., Byron, M., Lawrence, J. B., Dekker, J., and Marti-Renom, M. A. (2011). The three-dimensional folding of the α -globin gene domain reveals formation of chromatin globules. *Nature structural & molecular biology*, 18(1):107–114. 19
- [96] Umbarger, M. A., Toro, E., Wright, M. A., Porreca, G. J., Baù, D., Hong, S.-H., Fero, M. J., Zhu, L. J., Marti-Renom, M. A., McAdams, H. H., et al. (2011). The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Molecular Cell*, 44(2):252–264. 19
- [97] Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A., and Noble, W. S. (2010). A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367. 19
- [98] Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature Biotechnology*, 30(1):90–98. 19
- [99] Fudenberg, G. and Mirny, L. A. (2012). Higher-order chromatin structure: bridging physics and biology. *Current opinion in genetics & development*, 22(2):115–124. 19, 77
- [100] Cremer, T. and Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, 2(4):292–301. 20
- [101] Boyle, S., Gilchrist, S., Bridger, J. M., Mahy, N. L., Ellis, J. A., and Bickmore, W. A. (2001). The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human molecular genetics*. 20
- [102] Chaumeil, J., Le Baccon, P., Wutz, A., and Heard, E. (2006). A novel role for Xist RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes & development*, 20(16):2223–2237. 20
- [103] Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951. 20, 22
- [104] Reddy, K. L., Zullo, J. M., Bertolino, E., and Singh, H. (2008). Transcriptional repression mediated by repositioning of genes to the nuclear lamina. *Nature*, 452(7184):243–247. 20
- [105] 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. *Molecular Cell*, 38(4):603–613. 20

-
- [106] Zullo, J. M., Demarco, I. A., Pique-Regi, R., Gaffney, D. J., Epstein, C. B., Spooner, C. J., Luperchio, T. R., Bernstein, B. E., Pritchard, J. K., Reddy, K. L., et al. (2012). DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. *Cell*, 149(7):1474–1487. 20
- [107] Brown, C. R., Kennedy, C. J., Delmar, V. A., Forbes, D. J., and Silver, P. A. (2008). Global histone acetylation induces functional genomic reorganization at mammalian nuclear pore complexes. *Genes & development*, 22(5):627–639. 20
- [108] Capelson, M., Liang, Y., Schulte, R., Mair, W., Wagner, U., and Hetzer, M. W. (2010). Chromatin-bound nuclear pore components regulate gene expression in higher eukaryotes. *Cell*, 140(3):372–383. 20
- [109] Kalverda, B., Pickersgill, H., Shloma, V. V., and Fornerod, M. (2010). Nucleoporins directly stimulate expression of developmental and cell-cycle genes inside the nucleoplasm. *Cell*, 140(3):360–371. 20
- [110] Van Steensel, B. (2011). Chromatin: constructing the big picture. *The EMBO journal*, 30(10):1885–1895. 20, 77, 95, 96, 98
- [111] 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. *Nature Genetics*, 36(10):1065–1071. 21
- [112] Németh, A., Conesa, A., Santoyo-Lopez, J., Medina, I., Montaner, D., Péterfia, B., Solovei, I., Cremer, T., Dopazo, J., and Längst, G. (2010). Initial genomics of the human nucleolus. *PLoS Genetics*, 6(3):e1000889. 21
- [113] Fedoriw, A. M., Stein, P., Svoboda, P., Schultz, R. M., and Bartolomei, M. S. (2004). Transgenic RNAi reveals essential function for CTCF in H19 gene imprinting. *Science (New York, NY)*, 303(5655):238–240. 21
- [114] Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., and de Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development*, 20(17):2349–2354. 21
- [115] Heath, H., Ribeiro de Almeida, C., Sleutels, F., Dingjan, G., van de Nobelen, S., Jonkers, I., Ling, K.-W., Gribnau, J., Renkawitz, R., Grosveld, F., et al. (2008). CTCF regulates cell cycle progression of alphabeta T cells in the thymus. *The EMBO journal*, 27(21):2839–2850. 21
- [116] Moon, H., Filippova, G., Loukinov, D., Pugacheva, E., Chen, Q., Smith, S. T., Munhall, A., Grewe, B., Bartkuhn, M., Arnold, R., et al. (2005). CTCF is conserved from *Drosophila* to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO reports*, 6(2):165–170. 21
- [117] Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. *Cell*. 21, 22, 111

-
- [118] Ohlsson, R., Renkawitz, R., and Lobanenko, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genetics*, 17(9):520–527. 21
- [119] Baniahmad, A., Steiner, C., Köhne, A. C., and Renkawitz, R. (1990). Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. *Cell*, 61(3):505–514. 21
- [120] Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., and Lobanenko, V. V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and cellular biology*, 16(6):2802–2813. 21
- [121] Lobanenko, V. V., Nicolas, R. H., Adler, V. V., Paterson, H., Klenova, E. M., Polotskaja, A. V., and Goodwin, G. H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, 5(12):1743–1753. 21
- [122] Klenova, E. M., Nicolas, R. H., Paterson, H. F., Carne, A. F., Heath, C. M., Goodwin, G. H., Neiman, P. E., and Lobanenko, V. V. (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. *Molecular and cellular biology*. 21
- [123] Awad, T. A., Bigler, J., Ulmer, J. E., Hu, Y. J., Moore, J. M., Lutz, M., Neiman, P. E., Collins, S. J., Renkawitz, R., Lobanenko, V. V., et al. (1999). Negative Transcriptional Regulation Mediated by Thyroid Hormone Response Element 144 Requires Binding of the Multivalent Factor CTCF to a Novel Target DNA Sequence. *The Journal of biological chemistry*. 21
- [124] Burcin, M., Arnold, R., Lutz, M., Kaiser, B., Runge, D., Lottspeich, F., Filippova, G. N., Lobanenko, V. V., and Renkawitz, R. (1997). Negative protein 1, which is required for function of the chicken lysozyme gene silencer in conjunction with hormone receptors, is identical to the multivalent zinc finger repressor CTCF. *Molecular and cellular biology*. 21
- [125] Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *Cell*, 98(3):387–396. 21
- [126] Felsenfeld, G. and Bell, A. C. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, 405(6785):482–485. 21
- [127] Hark, A. T., Schoenherr, C. J., Katz, D. J., Ingram, R. S., Levorse, J. M., and Tilghman, S. M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature*, 405(6785):486–489. 21
- [128] Kanduri, C., Pant, V., Loukinov, D., Pugacheva, E., Qi, C. F., Wolffe, A., Ohlsson, R., and Lobanenko, V. V. (2000). Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Current Biology*, 10(14):853–856. 21

-
- [129] Murrell, A., Heeson, S., and Reik, W. (2004). Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nature Genetics*, 36(8):889–893. 22
 - [130] Kurukuti, S., Tiwari, V. K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenko, V., Reik, W., and Ohlsson, R. (2006). CTCF binding at the *H19* imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to *Igf2*. *Proceedings of the National Academy of Sciences*, 103(28):10684–10689. 22
 - [131] Yoon, Y. S., Jeong, S., Rong, Q., Park, K.-Y., Chung, J. H., and Pfeifer, K. (2007). Analysis of the *H19*ICR insulator. *Molecular and cellular biology*, 27(9):3499–3510. 22
 - [132] Engel, N., Raval, A. K., Thorvaldsen, J. L., and Bartolomei, S. M. (2008). Three-dimensional conformation at the *H19*/*Igf2* locus supports a model of enhancer tracking. *Human molecular genetics*, 17(19):3021–3029. 22
 - [133] Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenko, V. V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, 128(6):1231–1245. 22
 - [134] Cuddapah, S., Jothi, R., Schones, D. E., Roh, T.-Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Research*, 19(1):24–32. 22, 43, 98
 - [135] 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 Research*, 22(9):1680–1688. 22
 - [136] Botta, M., Haider, S., Leung, I. X. Y., Lio, P., and Mozziconacci, J. (2010). Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. *Molecular Systems Biology*, 6(1). 22
 - [137] Cuylen, S. and Haering, C. H. (2010). A new cohesive team to mediate DNA looping. *Cell stem cell*, 7(4):424–426. 23
 - [138] Choufani, S., Shuman, C., and Weksberg, R. (2010). Beckwith-Wiedemann syndrome. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 154C(3):343–354. 22
 - [139] Klenova, E., Scott, A. C., Roberts, J., Shamsuddin, S., Lovejoy, E. A., Bergmann, S., Bubb, V. J., Royer, H.-D., and Quinn, J. P. (2004). YB-1 and CTCF differentially regulate the 5-HTT polymorphic intron 2 enhancer which predisposes to a variety of neurological disorders. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 24(26):5966–5973. 22
 - [140] Fiorentino, F. P. and Giordano, A. (2011). The tumor suppressor role of CTCF. *Journal of Cellular Physiology*, 227(2):479–492. 22, 26
 - [141] Holdorf, M. M., Cooper, S. B., Yamamoto, K. R., and Miranda, J. J. L. (2011). Occupancy of chromatin organizers in the Epstein-Barr virus genome. *Virology*, 415(1):1–5. 22

-
- [142] Stedman, W., Kang, H., Lin, S., Kissil, J. L., Bartolomei, M. S., and Lieberman, P. M. (2008). Cohesins localize with CTCF at the KSHV latency control region and at cellular c-myc and H19/Igf2 insulators. *The EMBO journal*, 27(4):654–666. 22, 24, 78
- [143] Walsh, C. P. and Bestor, T. H. (1999). Cytosine methylation and mammalian development. *Genes & development*. 22
- [144] Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H.-H., et al. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research*, 18(11):1752–1762. 22
- [145] Peters, J.-M., Tedeschi, A., and Schmitz, J. (2008). The cohesin complex and its roles in chromosome biology. *Genes & development*, 22(22):3089–3114. 24
- [146] Anderson, D. E., Losada, A., Erickson, H. P., and Hirano, T. (2002). Condensin and cohesin display different arm conformations with characteristic hinge angles. *The Journal of Cell Biology*, 156(3):419–424. 24
- [147] Haering, C. H., Löwe, J., Hochwagen, A., and Nasmyth, K. (2002). Molecular architecture of SMC proteins and the yeast cohesin complex. *Molecular Cell*, 9(4):773–788. 24
- [148] Nasmyth, K. and Haering, C. H. (2009). Cohesin: its roles and mechanisms. *Annual Review of Genetics*. 24, 78, 96, 98
- [149] Rollins, R. A., Korom, M., Aulner, N., Martens, A., and Dorsett, D. (2004). Drosophila nipped-B protein supports sister chromatid cohesion and opposes the stromalin/Scc3 cohesion factor to facilitate long-range activation of the cut gene. *Molecular and cellular biology*, 24(8):3100–3111. 24
- [150] Rollins, R. A., Morcillo, P., and Dorsett, D. (1999). Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes. *Genetics*, 152(2):577–593. 24
- [151] Donze, D., Adams, C. R., Rine, J., and Kamakaka, R. T. (1999). The boundaries of the silenced HMR domain in *Saccharomyces cerevisiae*. *Genes & development*, 13(6):698–708. 24
- [152] Bénard, C. Y., Kébir, H., Takagi, S., and Hekimi, S. (2004). mau-2 acts cell-autonomously to guide axonal migrations in *Caenorhabditis elegans*. *Development*, 131(23):5947–5958. 24
- [153] Pauli, A., Althoff, F., Oliveira, R. A., Heidmann, S., Schuldiner, O., Lehner, C. F., Dickson, B. J., and Nasmyth, K. (2008). Cell-type-specific TEV protease cleavage reveals cohesin functions in *Drosophila* neurons. *Developmental Cell*, 14(2):239–251. 24
- [154] Horsfield, J. A., Anagnostou, S. H., Hu, J. K.-H., Cho, K. H. Y., Geisler, R., Lieschke, G., Crosier, K. E., and Crosier, P. S. (2007). Cohesin-dependent regulation of Runx genes. *Development*, 134(14):2639–2649. 24

-
- [155] Zhang, B., Jain, S., Song, H., Fu, M., Heuckeroth, R. O., Erlich, J. M., Jay, P. Y., and Milbrandt, J. (2007). Mice lacking sister chromatid cohesion protein PDS5B exhibit developmental abnormalities reminiscent of Cornelia de Lange syndrome. *Development*, 134(17):3191–3201. 24
- [156] Krantz, I. D., McCallum, J., DeScipio, C., Kaur, M., Gillis, L. A., Yaeger, D., Jukofsky, L., Wasserman, N., Bottani, A., Morris, C. A., et al. (2004). Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of *Drosophila melanogaster* Nipped-B. *Nature Genetics*, 36(6):631–635. 24
- [157] Vega, H., Waisfisz, Q., Gordillo, M., Sakai, N., Yanagihara, I., Yamada, M., van Gosliga, D., Kayserili, H., Xu, C., Ozono, K., et al. (2005). Roberts syndrome is caused by mutations in ESCO2, a human homolog of yeast ECO1 that is essential for the establishment of sister chromatid cohesion. *Nature Genetics*, 37(5):468–470. 24
- [158] Sumara, I., Vorlaufer, E., Gieffers, C., Peters, B. H., and Peters, J. M. (2000). Characterization of vertebrate cohesin complexes and their regulation in prophase. *The Journal of Cell Biology*, 151(4):749–762. 24
- [159] Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., et al. (2008). Cohesin mediates transcriptional insulation by CCTC-binding factor. *Nature*, 451(7180):796–801. 24, 78
- [160] Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H. C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell*, 132(3):422–433. 24, 60, 78
- [161] Rubio, E. D., Reiss, D. J., Welsh, P. L., Distech, C. M., Filippova, G. N., Baliga, N. S., Aebersold, R., Ranish, J. A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences*, 105(24):8309–8314. 24, 78
- [162] 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. *Molecular and cellular biology*, 31(11):2174–2183. 24, 47
- [163] Hadjur, S., Williams, L. M., Ryan, N. K., Cobb, B. S., Sexton, T., Fraser, P., Fisher, A. G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature*, 460(7253):410–413. 24, 43, 64, 66, 81, 91, 96
- [164] Mishiro, T., Ishihara, K., Hino, S., Tsutsumi, S., Aburatani, H., Shirahige, K., Kinoshita, Y., and Nakao, M. (2009). Architectural roles of multiple chromatin insulators at the human apolipoprotein gene cluster. *The EMBO journal*, 28(9):1234–1245. 24, 64
- [165] Nativio, R., Wendt, K. S., Ito, Y., Huddleston, J. E., Uribe-Lewis, S., Woodfine, K., Krueger, C., Reik, W., Peters, J.-M., and Murrell, A. (2009). Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus. *PLoS Genetics*, 5(11):e1000739. 24, 64, 66
- [166] Hou, C., Dale, R., and Dean, A. (2010). Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proceedings of the National Academy of Sciences*, 107(8):3651–3656. 24, 64

-
- [167] Schmidt, D., Schwalie, P. C., Ross-Innes, C. S., Hurtado, A., Brown, G. D., Carroll, J. S., Flicek, P., and Odom, D. T. (2010). A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Research*, 20(5):578–588. 25, 46, 64, 70, 78, 121
- [168] Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. B., Orlov, Y. L., Velkov, S., Ho, A., Mei, P. H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64. 25, 37
- [169] Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, 467(7314):430–435. 25, 49, 64, 71, 78, 81, 85, 91, 96, 99, 116, 121
- [170] Seitan, V. C., Hao, B., Tachibana-Konwalski, K., Lavagnolli, T., Mira-Bontenbal, H., Brown, K. E., Teng, G., Carroll, T., Terry, A., Horan, K., et al. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature*, 476(7361):467–471. 25, 43, 64, 78, 79, 81, 91, 96, 99, 121, 122
- [171] Dorsett, D. and Ström, L. (2012). The ancient and evolving roles of cohesin in gene expression and DNA repair. *Current Biology*, 22(7):R240–50. 25
- [172] Koonin, E. V. (2005). ORTHOLOGS, PARALOGS, AND EVOLUTIONARY GENOMICS 1. *Annual Review of Genetics*, 39(1):309–338. 25
- [173] Loukinov, D. I., Pugacheva, E., Vatolin, S., Pack, S. D., Moon, H., Chernukhin, I., Manan, P., Larsson, E., Kanduri, C., Vostrov, A. A., et al. (2002). BORIS, a novel male germline-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma. *Proceedings of the National Academy of Sciences*, 99(10):6806–6811. 25, 26
- [174] Hore, T. A., Deakin, J. E., and Graves, J. A. M. (2008). The Evolution of Epigenetic Regulators CTCF and BORIS/CTCF in Amniotes. *PLoS Genetics*, 4(8):e1000169. 25, 26
- [175] Sleutels, F., Soochit, W., Bartkuhn, M., Heath, H., Dienstbach, S., Bergmaier, P., Franke, V., Rosa-Garrido, M., van de Nobelen, S., Caesar, L., et al. (2012). The male germ cell gene regulator CTCFL is functionally different from CTCF and binds CTCF-like consensus sites in a nucleosome composition-dependent manner. *Epigenetics & chromatin*, 5(1):8. 26, 105, 108, 113
- [176] Jones, T. A., Ogunkolade, B. W., Szary, J., Aarum, J., Mumin, M. A., Patel, S., Pieri, C. A., and Sheer, D. (2011). Widespread expression of BORIS/CTCF in normal and cancer cells. *PLoS ONE*, 6(7):e22399. 26, 105, 122
- [177] de Necochea, R., Ghochikyan, A., Josephs, S. F., Zacharias, S., Woods, E., Karimi-Busheri, F., Alexandrescu, D. T., Chen, C.-S., Agadjanyan, M. G., and Carrier, E. (2011). Expression of the Epigenetic factor BORIS (CTCF) in the Human Genome. *Journal of translational medicine*, 9(1):213. 26

-
- [178] Nguyen, P., Bar-Sela, G., Sun, L., Bisht, K. S., Cui, H., Kohn, E., Feinberg, A. P., and Gius, D. (2008). BAT3 and SET1A Form a Complex with CTCFL/BORIS To Modulate H3K4 Histone Dimethylation and Gene Expression. *Molecular and cellular biology*, 26, 121
- [179] Sun, L., Huang, L., Nguyen, P., Bisht, K. S., Bar-Sela, G., Ho, A. S., Bradbury, C. M., Yu, W., Cui, H., Lee, S., et al. (2008). DNA Methyltransferase 1 and 3B Activate BAG-1 Expression via Recruitment of CTCFL/BORIS and Modulation of Promoter Histone Methylation. *Cancer Research*, 68(8):2726–2735. 26
- [180] Renaud, S., Loukinov, D., Alberti, L., Vostrov, A., Kwon, Y.-W., Bosman, F. T., Lobanenko, V., and Benhattar, J. (2011). BORIS/CTCF-mediated transcriptional regulation of the hTERT telomerase gene in testicular and ovarian tumor cells. *Nucleic Acids Research*. 26
- [181] Renaud, S., Pugacheva, E. M., Delgado, M. D., Braunschweig, R., Abdullaev, Z., Loukinov, D., Benhattar, J., and Lobanenko, V. (2007). Expression of the CTCF-paralogous cancer-testis gene, brother of the regulator of imprinted sites (BORIS), is regulated by three alternative promoters modulated by CpG methylation and by CTCF and p53 transcription factors. *Nucleic Acids Research*, 35(21):7372–7388. 26
- [182] Rosa-Garrido, M., Ceballos, L., Alonso-Lecue, P., Abaira, C., Delgado, M. D., and Gandarillas, A. (2012). A cell cycle role for the epigenetic factor CTCF-L/BORIS. *PLoS ONE*, 7(6):e39371. 26, 122
- [183] Hutchison, C. A. (2007). DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*. 27
- [184] Heller, M. J. (2002). DNA microarray technology: devices, systems, and applications. *Annual review of biomedical engineering*, 4:129–153. 27
- [185] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63. 27, 35, 36
- [186] Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G., et al. (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Research*, 14(3):331–342. 27
- [187] International HapMap Consortium, Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., et al. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861. 28
- [188] Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680. 28
- [189] Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., Bignell, H. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59. 28, 33

- [190] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80. 28
- [191] Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46. 29
- [192] Strausberg, R. L., Levy, S., and Rogers, Y. H. (2008). Emerging DNA sequencing technologies for human genomic medicine. *Drug discovery today*. 30
- [193] Krueger, F., Kreck, B., Franke, A., and Andrews, S. R. (2012). DNA methylome analysis using short bisulfite sequencing data. *Nature Methods*, 9(2):145–151. 29
- [194] Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., and Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, 33(18):5868–5877. 31
- [195] Ficiz, G., Branco, M. R., Seisenberger, S., Santos, F., Krueger, F., Hore, T. A., Marques, C. J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, 473(7347):398–402. 31
- [196] Boyle, A., Davis, S., Shulha, H., and Meltzer, P. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell*. 31
- [197] Schones, D. E., Cui, K., Cuddapah, S., Roh, T.-Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–898. 31
- [198] Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., and Odom, D. T. (2009). ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods (San Diego, Calif)*, 48(3):240–248. 31
- [199] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831. 31, 34
- [200] Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nature Methods*, 4(8):613–614. 32
- [201] Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. 33
- [202] Morgan, M., Lawrence, M., and Anders, S. *ShortRead: Base classes and methods for high-throughput short-read sequencing data*. 33
- [203] Li, H., Ruan, J., and Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*, 18(11):1851–1858. 33

-
- [204] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–1760. 33, 44, 70, 101
- [205] Langmead, B., Trapnell, C., and Pop, M. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. 33, 74, 100
- [206] Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., et al. (2012). Ensembl 2013. *Nucleic Acids Research*, 41(D1):D48–D55. 33, 100, 114
- [207] Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, 29(1):24–26. 33
- [208] Wilbanks, E. G. and Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS ONE*, 5(7):e11471. 33
- [209] Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9):R137. 34
- [210] Xu, H., Handoko, L., Wei, X., Ye, C., Sheng, J., Wei, C.-L., Lin, F., and Sung, W.-K. (2010). A signal-noise model for significance analysis of ChIP-seq with negative control. *Bioinformatics (Oxford, England)*, 26(9):1199–1204. 34, 60, 100
- [211] Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, 2:28–36. 34, 71
- [212] Down, T. A. and Hubbard, T. J. P. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Research*, 33(5):1445–1453. 34, 71
- [213] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34(Database issue):D108–10. 34, 71
- [214] Harris, M., Clark, J., Ireland, A., and Lomax, J. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research*. 34
- [215] McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501. 34
- [216] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1):139–140. 35, 36
- [217] Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106. 35, 36, 100

-
- [218] 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*. 35, 75
- [219] Garber, M., Grabherr, M. G., Guttman, M., and Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6):469–477. 35, 36
- [220] Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S96–104. 35, 124
- [221] Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical applications in genetics and molecular biology*, 3(1). 35, 124
- [222] Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors, *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, New York. 35, 124
- [223] Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829. 35
- [224] Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., Mungall, K., Lee, S., Okada, H. M., Qian, J. Q., et al. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, 7(11):909–912. 35
- [225] Turro, E., Su, S.-Y., Gonçalves, Â., Coin, L. J., Richardson, S., and Lewin, A. (2011). Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biology*, 12(2):R13. 36, 74, 100
- [226] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics (Oxford, England)*, 26(4):493–500. 36
- [227] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515. 36
- [228] Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., Cremer, C., Fakan, S., and Cremer, T. (2002). Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence in situ hybridization (3D-FISH). *Experimental cell research*, 276(1):10–23. 37
- [229] Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. *Science (New York, NY)*. 37

-
- [230] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., Van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, 38(11):1348–1354. 37
- [231] Würtele, H. and Chartrand, P. (2006). Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. *Chromosome Research*, 14(5):477–495. 37
- [232] Zhao, Z., Tavoosidana, G., Sjölander, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, 38(11):1341–1347. 37
- [233] 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. *Nature Methods*, 9(10):969–972. 37
- [234] Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Research*, 16(10):1299–1309. 37
- [235] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4):576–589. 39, 40, 79, 100
- [236] Imakaev, M., Fudenberg, G., McCord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., Dekker, J., and Mirny, L. A. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*. 39, 40, 79, 100
- [237] Yaffe, E. and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 43(11):1059–1065. 40
- [238] Faure, A. J., Schmidt, D., Watt, S., Schwalie, P. C., Wilson, M. D., Xu, H., Ramsay, R. G., Odom, D. T., and Flicek, P. (2012). Cohesin regulates tissue-specific expression by stabilising highly occupied cis-regulatory modules. *Genome Research*. 43, 78, 100, 121, 123
- [239] Misteli, T. (2007). Beyond the sequence: cellular organization of genome function. *Cell*, 128(4):787–800. 43
- [240] Gard, S., Light, W., Xiong, B., Bose, T., McNairn, A. J., Harris, B., Fleharty, B., Seidel, C., Brickner, J. H., and Gerton, J. L. (2009). Cohesinopathy mutations disrupt the subnuclear organization of chromatin. *The Journal of Cell Biology*, 187(4):455–462. 43

-
- [241] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., et al. (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell*, 133(6):1106–1117. 44, 49, 71, 106, 123
- [242] Bartkuhn, M., Straub, T., Herold, M., Herrmann, M., Rathke, C., Saumweber, H., Gillfillan, G. D., Becker, P. B., and Renkawitz, R. (2009). Active promoters and insulators are marked by the centrosomal protein 190. *The EMBO journal*, 28(7):877–888. 46
- [243] Misulovin, Z., Schwartz, Y. B., Li, X.-Y., Kahn, T. G., Gause, M., MacArthur, S., Fay, J. C., Eisen, M. B., Pirrotta, V., Biggin, M. D., et al. (2007). Association of cohesin and Nipped-B with transcriptionally active regions of the *Drosophila melanogaster* genome. *Chromosoma*, 117(1):89–102. 46, 78
- [244] Fay, A., Misulovin, Z., Li, J., Schaaf, C. A., Gause, M., Gilmour, D. S., and Dorsett, D. (2011). Cohesin selectively binds and regulates genes with paused RNA polymerase. *Current Biology*, 21(19):1624–1634. 46
- [245] Hendrix, D. A., Hong, J.-W., Zeitlinger, J., Rokhsar, D. S., and Levine, M. S. (2008). Promoter elements associated with RNA Pol II stalling in the *Drosophila* embryo. *Proceedings of the National Academy of Sciences*, 105(22):7762–7767. 47, 85, 102
- [246] Nitzsche, A., Paszkowski-Rogacz, M., Matarese, F., Janssen-Megens, E. M., Hubner, N. C., Schulz, H., de Vries, I., Ding, L., Huebner, N., Mann, M., et al. (2011). RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS ONE*, 6(5):e19470. 47
- [247] Quitschke, W. W., Taheny, M. J., Fochtmann, L. J., and Vostrov, A. A. (2000). Differential effect of zinc finger deletions on the binding of CTCF to the promoter of the amyloid precursor protein gene. *Nucleic Acids Research*, 28(17):3370–3378. 47
- [248] Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J., et al. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134(3):521–533. 49, 71, 106, 123
- [249] Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A., and Sharp, P. A. (2008). Divergent transcription from active promoters. *Science (New York, NY)*, 322(5909):1849–1851. 49, 71, 106, 123
- [250] Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319. 49
- [251] Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I., and Young, R. A. (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334. 49
- [252] Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences*. 49

-
- [253] Downen, J. M., Bilodeau, S., Orlando, D. A., Hübner, M. R., Abraham, B. J., Spector, D. L., and Young, R. A. (2013). Multiple Structural Maintenance of Chromosome Complexes at Transcriptional Regulatory Elements. *Stem Cell Reports*, 1(5):371–378. 49
- [254] Kutter, C., Brown, G. D., Gonçalves, Â., Wilson, M. D., Watt, S., Brazma, A., White, R. J., and Odom, D. T. (2011). Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. *Nature Genetics*, 43(10):948–955. 55, 74
- [255] Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences*, 101(16):6062–6067. 55, 74
- [256] Moorman, C., Sun, L. V., Wang, J., de Wit, E., Talhout, W., Ward, L. D., Greil, F., Lu, X. J., White, K. P., Bussemaker, H. J., et al. (2006). Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, 103(32):12027–12032. 57, 67
- [257] Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al. (2010). Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science (New York, NY)*, 330(6012):1775–1787. 57, 67
- [258] Nègre, N., Brown, C. D., Ma, L., Bristow, C. A., Miller, S. W., Wagner, U., Kheradpour, P., Eaton, M. L., Loriaux, P., Sealfon, R., et al. (2011). A cis-regulatory map of the *Drosophila* genome. *Nature*, 471(7339):527–531. 57, 67
- [259] Dickinson, L. A., Edgar, A. J., Ehley, J., and Gottesfeld, J. M. (2002). Cyclin L is an RS domain protein involved in pre-mRNA splicing. *The Journal of biological chemistry*, 277(28):25465–25473. 57
- [260] Adcock, I. M. and Caramori, G. (2001). Cross-talk between pro-inflammatory transcription factors and glucocorticoids. *Immunology and Cell Biology*, 79(4):376–384. 57
- [261] Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglu, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature Methods*, 5(9):829–834. 58
- [262] Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, NY)*, 316(5830):1497–1502. 58
- [263] Boyle, A. P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., Iyer, V. R., Crawford, G. E., and Furey, T. S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, 21(3):456–464. 58
- [264] Wilczyński, B. and Furlong, E. E. M. (2010). Dynamic CRM occupancy reflects a temporal map of developmental progression. *Molecular Systems Biology*, 6:383. 58

-
- [265] Conboy, C. M., Spyrou, C., Thorne, N. P., Wade, E. J., Barbosa-Morais, N. L., Wilson, M. D., Bhattacharjee, A., Young, R. A., Tavaré, S., Lees, J. A., et al. (2007). Cell cycle genes are the evolutionarily conserved targets of the E2F4 transcription factor. *PLoS ONE*, 2(10):e1061. 58
 - [266] Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W. H., Ye, C., Ping, J. L. H., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genetics*, 43(7):630–638. 64, 68
 - [267] Nguyen, T. T. and Androulakis, I. P. (2009). Recent Advances in the Computational Discovery of Transcription Factor Binding Sites. *Algorithms*, 2(1):582–605. 67
 - [268] Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics (Oxford, England)*, 27(12):1653–1659. 67
 - [269] Cheeseman, P. and Stutz, J. (1996). Bayesian Classification (AutoClass): Theory and Results. *Advances in knowledge discovery and data mining American Association for Artificial Intelligence, Menlo Park, CA*. 73
 - [270] Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2010). Ensembl 2011. *Nucleic Acids Research*, 39(Database):D800–D806. 74
 - [271] Kapushesky, M., Emam, I., Holloway, E., Kurnosov, P., Zorin, A., Malone, J., Rustici, G., Williams, E., Parkinson, H., and Brazma, A. (2010). Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Research*, 38(Database issue):D690–8. 74
 - [272] McCord, R. P., Nazario-Toole, A., Zhang, H., Chines, P. S., Zhan, Y., Erdos, M. R., Collins, F. S., Dekker, J., and Cao, K. (2013). Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome. *Genome Research*, 23(2):260–269. 77, 96
 - [273] Aragon, L., Martinez-Perez, E., and Merckenschlager, M. (2013). Condensin, cohesin and the control of chromatin states. *Current opinion in genetics & development*. 78
 - [274] Dorsett, D. and Merckenschlager, M. (2013). Cohesin at active genes: a unifying theme for cohesin and gene expression from model organisms to humans. *Current opinion in cell biology*. 78
 - [275] Schaaf, C. A., Misulovin, Z., Sahota, G., Siddiqui, A. M., Schwartz, Y. B., Kahn, T. G., Pirrotta, V., Gause, M., and Dorsett, D. (2009). Regulation of the Drosophila Enhancer of split and invected-engrailed gene complexes by sister chromatid cohesion proteins. *PLoS ONE*, 4(7):e6202. 78
 - [276] Fraser, P. and Bickmore, W. (2007). Nuclear organization of the genome and the potential for gene regulation. *Nature*, 447(7143):413–417. 81
 - [277] Shih, H.-Y., Verma-Gaur, J., Torkamani, A., Feeney, A. J., Galjart, N., and Krangel, M. S. (2012). Tcra gene recombination is supported by a Tcra enhancer- and CTCF-dependent chromatin hub. *Proceedings of the National Academy of Sciences*. 86, 100

-
- [278] Akalin, A., Fredman, D., Arner, E., Dong, X., Bryne, J. C., Suzuki, H., Daub, C. O., Hayashizaki, Y., and Lenhard, B. (2009). Transcriptional features of genomic regulatory blocks. *Genome Biology*, 10(4):R38. 87
- [279] Dimitrieva, S. and Bucher, P. (2012). UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic Acids Research*, 41(D1):D101–D109. 87, 90
- [280] Lin, H., Gupta, V., Vermilyea, M. D., Falciani, F., Lee, J. T., O’Neill, L. P., and Turner, B. M. (2007). Dosage compensation in the mouse balances up-regulation and silencing of X-linked genes. *PLoS Biology*, 5(12):e326. 91
- [281] Zhang, J. A., Mortazavi, A., Williams, B. A., Wold, B. J., and Rothenberg, E. V. (2012). Dynamic Transformations of Genome-wide Epigenetic Marking and Transcriptional Control Establish T Cell Identity. *Cell*, 149(2):467–482. 93, 100, 128
- [282] Wallace, J. A. and Felsenfeld, G. (2007). We gather together: insulators and genome organization. *Current opinion in genetics & development*, 17(5):400–407. 95, 98
- [283] Strachan, T. (2005). Cornelia de Lange Syndrome and the link between chromosomal function, DNA repair and developmental gene regulation. *Current opinion in genetics & development*, 15(3):258–264. 96
- [284] Kawauchi, S., Calof, A. L., Santos, R., Lopez-Burks, M. E., Young, C. M., Hoang, M. P., Chua, A., Lao, T., Lechner, M. S., Daniel, J. A., et al. (2009). Multiple organ system defects and transcriptional dysregulation in the Nipbl(+/-) mouse, a model of Cornelia de Lange Syndrome. *PLoS Genetics*, 5(9):e1000650. 96
- [285] Lee, B.-K. and Iyer, V. R. (2012). Genome-wide Studies of CCCTC-binding Factor (CTCF) and Cohesin Provide Insight into Chromatin Structure and Regulation. *The Journal of biological chemistry*, 287(37):30906–30913. 96
- [286] Phillips-Cremins, J. E., Sauria, M. E. G., Sanyal, A., Gerasimova, T. I., Lajoie, B. R., Bell, J. S. K., Ong, C.-T., Hookway, T. A., Guo, C., Sun, Y., et al. (2013). Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*. 98, 128
- [287] de Wit, E., Bouwman, B. A. M., Zhu, Y., Klous, P., Splinter, E., Verstegen, M. J. A. M., Krijger, P. H. L., Festuccia, N., Nora, E. P., Welling, M., et al. (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*. 98
- [288] Guo, C., Yoon, H. S., Franklin, A., Jain, S., Ebert, A., Cheng, H.-L., Hansen, E., Despo, O., Bossen, C., Vettermann, C., et al. (2011). CTCF-binding elements mediate control of V(D)J recombination. *Nature*, 477(7365):424–430. 98
- [289] Ribeiro de Almeida, C., Stadhouders, R., de Bruijn, M. J. W., Bergen, I. M., Thongjuea, S., Lenhard, B., van Ijcken, W., Grosveld, F., Galjart, N., Soler, E., et al. (2011). The DNA-binding protein CTCF limits proximal V κ recombination and restricts κ enhancer interactions to the immunoglobulin κ light chain locus. *Immunity*, 35(4):501–513. 98

-
- [290] Gause, M., Misulovin, Z., Bilyeu, A., and Dorsett, D. (2010). Dosage-sensitive regulation of cohesin chromosome binding and dynamics by Nipped-B, Pds5, and Wapl. *Molecular and cellular biology*, 30(20):4940–4951. 99
- [291] Gerlich, D., Koch, B., Dupeux, F., Peters, J.-M., and Ellenberg, J. (2006). Live-Cell Imaging Reveals a Stable Cohesin-Chromatin Interaction after but Not before DNA Replication. *Current Biology*, 16(15):1571–1578. 99
- [292] McNairn, A. J. and Gerton, J. L. (2009). Intersection of ChIP and FLIP, genomic methods to study the dynamics of the cohesin proteins. *Chromosome Research*, 17(2):155–163. 99
- [293] Onn, I. and Koshland, D. (2011). In vitro assembly of physiological cohesin/DNA complexes. *Proceedings of the National Academy of Sciences*, 108(30):12198–12205. 99
- [294] Kueng, S., Hegemann, B., Peters, B. H., Lipp, J. J., Schleiffer, A., Mechtler, K., and Peters, J.-M. (2006). Wapl controls the dynamic association of cohesin with chromatin. *Cell*, 127(5):955–967. 99
- [295] Jessberger, R. (2012). Age-related aneuploidy through cohesion exhaustion. *EMBO reports*, 13(6):539–546. 99
- [296] Belton, J.-M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: A comprehensive technique to capture the conformation of genomes. *Methods (San Diego, Calif)*, 58(3):268–276. 99
- [297] Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenko, V. V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature*. 100
- [298] Venables, W. N. and Ripley, B. D. (2002). *Modern applied statistics with S*. Springer Verlag. 102
- [299] Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R. W., James, K. D., Turner, D. J., Smith, C., Harrison, D. J., Andrews, R., and Bird, A. P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genetics*, 6(9). 102, 113, 115
- [300] Nguyen, A. T. and Zhang, Y. (2011). The diverse functions of Dot1 and H3K79 methylation. *Genes & development*, 25(13):1345–1358. 105
- [301] Smit, A. F. A., Hubley, R., and Green, P. (1996). RepeatMasker Open-3.0 <http://www.repeatmasker.org>. 111
- [302] Chang, C. W., Cheng, W.-C., Chen, C.-R., Shu, W.-Y., Tsai, M.-L., Huang, C.-L., and Hsu, I. C. (2011). Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS ONE*, 6(7):e22859. 113
- [303] Gonçalves, Â., Leigh-Brown, S., Thybert, D., Stefflova, K., Turro, E., Flicek, P., Brazma, A., Odom, D. T., and Marioni, J. C. (2012). Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research*. 114

-
- [304] Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M. R. (2004). A census of human cancer genes. *Nature reviews Cancer*, 4(3):177–183. 114
- [305] Dougherty, C. J., Ichim, T. E., Liu, L., Reznik, G., Min, W.-P., Ghochikyan, A., Agadjanyan, M. G., and Reznik, B. N. (2008). Selective apoptosis of breast cancer cells by siRNA targeting of BORIS. *Biochemical and biophysical research communications*, 370(1):109–112. 121
- [306] Gaykalova, D., Vatapalli, R., Glazer, C. A., Bhan, S., Shao, C., Sidransky, D., Ha, P. K., and Califano, J. A. (2012). Dose-dependent activation of putative oncogene SBSN by BORIS. *PLoS ONE*, 7(7):e40389. 121
- [307] Smallwood, S. A., Tomizawa, S.-I., Krueger, F., Ruf, N., Carli, N., Segonds-Pichon, A., Sato, S., Hata, K., Andrews, S. R., and Kelsey, G. (2011). Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature Genetics*, 43(8):811–814. 121
- [308] Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, 39(4):457–466. 121
- [309] Pugacheva, E. M., Suzuki, T., Pack, S. D., Kosaka-Suzuki, N., Yoon, J., Vostrov, A. A., Barsov, E., Strunnikov, A. V., Morse, H. C., Loukinov, D., et al. (2010). The structural complexity of the human BORIS gene in gametogenesis and cancer. *PLoS ONE*, 5(11):e13872. 122
- [310] Balhorn, R. (2007). The protamine family of sperm nuclear proteins. *Genome Biology*, 8(9):227. 127
- [311] Hammoud, S. S., Nix, D. A., Zhang, H., Purwar, J., Carrell, D. T., and Cairns, B. R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. *Nature*. 127
- [312] Yan, J., Enge, M., Whittington, T., Dave, K., Liu, J., Sur, I., Schmierer, B., Jolma, A., Kivioja, T., Taipale, M., et al. (2013). Transcription Factor Binding in Human Cells Occurs in Dense Clusters Formed around Cohesin Anchor Sites. *Cell*, 154(4):801–813. 127
- [313] 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. *Nature Genetics*, 45(8):852–859. 127
- [314] Séverine Celton-Morizur, G. M. D. C. and Desdouets, C. (2010). Polyploidy and liver proliferation: Central role of insulin signaling. *Cell Cycle*, 9(3):460–466. 127
- [315] Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., Yen, C.-A., Schmitt, A. D., Espinoza, C. A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*. 128

- [316] Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419. 129
- [317] Engreitz, J. M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E. S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science (New York, NY)*, 341(6147):1237973. 129
- [318] Natoli, G. and Andrau, J.-C. (2012). Noncoding transcription at enhancers: general principles and functional models. *Annual Review of Genetics*, 46:1–19. 129
- [319] Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630. 129
- [320] Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. 129
- [321] Noordermeer, D., de Wit, E., Klous, P., van de Werken, H., Simonis, M., Lopez-Jones, M., Eussen, B., de Klein, A., Singer, R. H., and de Laat, W. (2011). Variegated gene expression caused by cell-specific long-range DNA interactions. *Nature Cell Biology*, 13(8):944–951. 129
- [322] Goodman, L. (1999). Hypothesis-Limited Research. *Genome Research*. 130