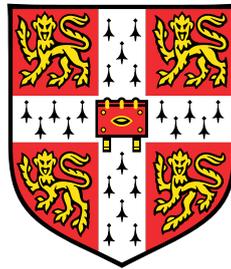


# **Insights into transcriptional regulation from natural and induced variation in closely related species**



**Elissavet Kentepozidou**

EMBL - European Bioinformatics Institute  
University of Cambridge

This dissertation is submitted for the degree of  
*Doctor of Philosophy*



*To my family*



## **Declaration**

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation does not exceed the limit of 60,000 words as specified by the Degree Committee.

Elissavet Kentepozidou

November 2020



## Acknowledgements

First and foremost, I would like to thank my supervisor, Paul Flicek, for having given me the opportunity to carry out my PhD thesis work in his group and develop my research skills engaging in very interesting projects. I am grateful for the guidance and support he has offered throughout this learning process. I am also grateful to Maša Roller, for the support and advice she has provided, her feedback and proofreading throughout this writing process, and for all the positive energy she has been bringing into the lab.

I would also like to acknowledge the members of my thesis advisory committee, Ewan Birney, Judith Jaugg and Elizabeth Murchison for the feedback they provided, as well as for their helpful and friendly attitude. Thanks are also due to all the members of the LCE consortium and the Odom lab for the very interesting discussions, the feedback and support they have provided throughout our collaboration. Special credits should be given to Sarah Aitken for all the enthusiasm, the inspiring talks and the insightful feedback she has been willing to provide any time it was asked for.

In addition, I would like to express my gratitude to Vasavi for being such a nice colleague and office mate, and for all the insightful discussions on scientific ideas or technical details, as well as for her very helpful feedback and support especially throughout my work within the LCE project. Of course, I would like to thank all the members of the Flicek research group—both current and former ones—for the extremely helpful comments and feedback they have provided throughout my work, all the inspiring discussions and the ideas exchange we have had, as well as for every lunch and coffee break and all enjoyable moments in (and out of) the lab. I am definitely very glad to have met and worked with Martina, Maëlle, Rachel, Sarah E., Petra, Osagie, as well as Dhoyazan, Ericca, David T. and David MG.

Additional thanks are due to the EMBL-EBI and EMBL PhD programme for the provided funding as well as the supportive network and vibrant environment, making a PhD experience smoother and more enjoyable. It has also been an honour to be a student at the University of

Cambridge, and especially a member of Darwin College and its dynamic student community.

I would also like to thank all friends from the EBI and the surroundings, with whom I have shared this PhD journey; first and foremost all the predocs, but especially Sushmita, Anna, Hannah, Harald and Claudia for the invaluable and memorable experiences we have shared in the last few years. I am also grateful to all the friends who have been there for me and supported me, each of them in their own way, during these years and especially during my thesis writing time.

Finally, I would like to express my huge gratitude to my family that has supported and encouraged me through every step I have chosen to take and have always been there to share both the challenging times and the most fun and craziest of experiences. For all these I am very grateful to them.

## Abstract

Despite years of research, numerous aspects of gene regulatory mechanisms and their contributions to shaping phenotypic characteristics remain elusive. In this thesis, I gain insights into mechanisms of transcriptional regulation, focusing particularly on the binding activity of transcription factors (TFs). In the comprised projects and analyses, I use models of closely related mouse species, as *in vivo* systems that leverage a pool of molecular variation, in order to study the activity and roles of TF binding within wider functional contexts. The molecular variation is manifested as variation in the sequence context and/or in the binding activity of the TFs, and it may be naturally fixed by evolution among the different species, and/or induced variation among different conditions, for example as a result to carcinogen exposures. The reflection of the underlying molecular variation in the functional roles of TF binding is also tied to some extent to functional outputs and is examined as part of general functional contexts, such as higher-order genome organisation or the neoplastic transformation of a tumour cell in cancer development.

In the first part of the thesis, I explore the evolutionary dynamics of the CTCF (CCCTC-binding factor) binding among mouse species and its interplay with the establishment and maintenance of topologically associating domains (TADs). TADs make up a level of the 3D genome organisation that dictates formation of regulatory landscapes. Although it is known that CTCF plays an important role in TAD formation, evidence of TAD boundaries being robust to CTCF depletion interferences has confounded the understanding of its exact role. My study reveals the occurrence of dynamically evolving clusters of neighbouring CTCF sites at the boundaries of TADs, which contribute to the resilience and flexibility of the TAD structures. My findings also highlight the necessity of examining the binding of CTCF not only as individual binding sites, but also as an ensemble of potentially functional sites that can exert their actions synergistically, additively or interchangeably, and contribute to fostering phenotypic features.

In the second part of the thesis, I present a project I worked on as a member of a collaborative effort of the Liver Cancer Evolution (LCE) consortium. The project makes use of a model of

chemically induced liver carcinogenesis in different mouse species, which offer a controlled biological system with natural genomic and epigenomic variation. Here, I characterise the expression profile of the cell-of-origin, the hepatocyte, of chemically induced liver tumours and the shift of its cell-type specific phenotype along neoplastic transformation, which is underlined by gene dysregulation. In addition, I further explore associations of gene dysregulation in the cell-of-origin with mutagenesis (induced genetic variation as a result of carcinogen exposure) in liver TF binding sites that potentially underlie the onset of hepatocyte dedifferentiation in tumour development. My results disentangle patterns of mutation accumulation among distinct categories of TF binding sites based on their functional characteristics, such as combinatorial binding or association with *cis*-regulatory elements. Also, they reveal sets of hyper-mutated TF binding sites in liver tumours, which are functionally enriched for liver-specific biological processes, cellular stress response and abnormal liver phenotype, and they are associated with dysregulated genes with hepatocyte-specific functions. Finally, these findings also highlight the necessity of studying the activity of TFs bound both in a singleton manner and in concert and within their broader functional context.

# Table of contents

<b>List of figures</b>	<b>xvii</b>
<b>List of tables</b>	<b>xxi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 " <i>DNA makes RNA</i> " - but how is this regulated? . . . . .	1
1.1.1 Central Dogma of Molecular Biology . . . . .	1
1.1.2 Gene expression in brief . . . . .	2
1.1.3 Gene expression regulation . . . . .	2
1.2 Transcriptional regulation . . . . .	4
1.2.1 Transcription initiation and <i>cis</i> -regulatory elements . . . . .	5
Transcription initiation, RNA Pol II and promoters . . . . .	5
Enhancers . . . . .	6
1.2.2 Transcription Factors . . . . .	7
DNA binding specificity as a function of DBDs, TF binding motifs, and combinatorial binding . . . . .	7
Transcriptional regulation through transcription factors (TFs) and TF regulation . . . . .	8
Biological roles of TFs . . . . .	9
TFs in establishment of cell type-specific regulation programmes . . . . .	10
Evolution of TFs, TF binding sites, and regulatory networks . . . . .	11
Implications of TFs in mutagenesis and disease . . . . .	14
1.2.3 Chromatin architecture: small-scale organisation of the genome and epigenome . . . . .	16
DNA methylation . . . . .	17
Nucleosome positioning and histone modifications . . . . .	18
1.2.4 Higher-order chromatin structure . . . . .	19
Chromatin loops . . . . .	20

---

	CTCF and cohesin . . . . .	22
	Topologically associating domains . . . . .	24
	Compartments . . . . .	27
	Revision of the “hierarchical” model of 3D genome organisation . .	28
	Chromosome territories . . . . .	29
1.3	Sequencing technologies and functional genomics assays . . . . .	30
1.3.1	NGS technologies . . . . .	31
	NGS data processing . . . . .	31
1.3.2	Protein-DNA interaction assays . . . . .	32
	ChIP-seq: experimental protocol and data analysis . . . . .	32
	Other assays for protein-DNA interactions . . . . .	34
1.3.3	Chromatin conformation assays . . . . .	35
1.3.4	Gene expression assays . . . . .	36
	RNA sequencing (RNA-seq) . . . . .	36
1.4	Aims of the thesis . . . . .	37
<b>2</b>	<b>Evolutionary dynamics of CTCF binding and higher-order genome structures</b>	<b>39</b>
2.1	Introduction . . . . .	40
2.2	Results . . . . .	41
2.2.1	Conservation of CTCF binding sites and association with TAD borders	41
2.2.2	Evolutionary constraints at TAD-boundary-associated CTCF binding sites . . . . .	45
2.2.3	Representation of LINEs and LINE-derived CTCF sites at TAD boundaries . . . . .	48
2.2.4	Evolutionarily dynamic clusters of CTCF binding sites at TAD borders	51
2.2.5	Localisation of CTCF site clusters with respect to cohesin-occupied sites and genes . . . . .	56
2.2.6	Impact of species-specific CTCF binding event loss on insulating function at TAD boundaries . . . . .	58
2.2.7	Effect of CTCF hemizyosity on CTCF binding and TAD organisation	60
2.3	Discussion . . . . .	61
2.4	Methods . . . . .	64
2.4.1	ChIP-seq experiments and data analysis . . . . .	64
2.4.2	TADs . . . . .	64
2.4.3	Conservation of CTCF binding sites in mice . . . . .	64
2.4.4	Binding affinity and sequence constraint of CTCF motifs . . . . .	65
2.4.5	ChIP enrichment of identified CTCF peaks . . . . .	65

2.4.6	Motif-word usage analysis . . . . .	66
2.4.7	Association of CTCF sites with classes of Transposable Elements . . . . .	66
2.4.8	Representation of TE classes at TAD boundary regions . . . . .	66
2.4.9	Density of CTCF sites at TAD boundaries and clusters of CTCF binding sites . . . . .	67
2.4.10	Clusters in BL6 and cluster conservation analyses . . . . .	67
2.4.11	RNA-seq data . . . . .	68
2.4.12	RNA-seq data processing and analysis . . . . .	68
2.4.13	TAD calling from BL6 MEF Hi-C data . . . . .	69
<b>3</b>	<b>Cell-of-origin and expression profile of chemically induced tumours in liver</b>	<b>71</b>
3.1	Introduction . . . . .	72
3.1.1	Liver Cancer and the LCE project . . . . .	72
3.1.2	Model of chemical carcinogenesis in mouse liver . . . . .	74
3.1.3	Liver cancer cell-of-origin . . . . .	76
3.2	Results . . . . .	77
3.2.1	Bulk RNA-seq libraries from LCE tumour samples . . . . .	77
3.2.2	Single-cell expression data from hepatobiliary cell differentiation in mouse liver . . . . .	78
3.2.3	Cross-mapping expression profiles of bulk LCE tumour samples and single cells from fetal mouse liver . . . . .	80
3.2.4	Dysregulation of cell type marker genes in HCC development . . . . .	85
3.3	Discussion . . . . .	89
3.4	Methods . . . . .	91
3.4.1	Expression quantification in LCE bulk samples and in single cells . . . . .	91
3.4.2	Identification of driver mutations in tumours . . . . .	91
3.4.3	Differential gene expression analyses . . . . .	91
<b>4</b>	<b>Mutational landscapes of distinct transcription factor binding region categories and functional implications in chemically induced liver tumours</b>	<b>93</b>
4.1	Introduction . . . . .	93
4.2	Results . . . . .	96
4.2.1	DEN tumour mutation datasets: abundance of T → N mutations in all three mouse species . . . . .	96
4.2.2	Cistromes of liver TFs and categorisation into sub-cistromes based on combinatorial TF binding and CRE co-occurrence . . . . .	98

---

4.2.3	The majority of TF binding regions contain a combination of TF binding motifs . . . . .	100
4.2.4	Fractions of TF binding regions that are mutated in DEN-induced liver tumours . . . . .	103
4.2.5	Weighted mutation rate in TF cistromes and sub-cistromes . . . . .	106
4.2.6	Pairwise comparisons of mutation spectra per position of TF binding motifs between sub-cistromes . . . . .	112
4.2.7	Pairwise comparisons of sequence context-conditional mutation counts per position of TF binding motifs between sub-cistromes . .	116
4.2.8	Functional analysis of hyper-mutated TFBRs . . . . .	121
4.3	Discussion . . . . .	127
4.3.1	Characterisation of TF sub-cistromes based on functionally relevant features . . . . .	127
4.3.2	Mutation rates in TF (sub-)cistromes, their underlying molecular and evolutionary processes, and hyper-mutated TFBR . . . . .	128
4.3.3	Mutational load of TF binding motifs and their flanking sequences .	131
4.3.4	Importance of studying mutagenesis of TF binding regions and general future perspectives . . . . .	133
4.4	Methods . . . . .	133
4.4.1	Mutation calls in the DEN induced tumour samples . . . . .	133
4.4.2	TF peak calling . . . . .	134
4.4.3	Identification of promoters and enhancers by the LCE consortium .	135
4.4.4	Splitting the TF cistromes in sub-cistromes . . . . .	135
4.4.5	Identification of TF-binding motifs . . . . .	135
4.4.6	Computing and visualizing observed and expected mutations in extended motif regions . . . . .	135
4.4.7	Calculating the expected mutation load of extended motif sequences	136
4.4.8	Functional analyses of hyper-mutated TFBRs . . . . .	137
<b>5</b>	<b>Conclusions</b>	<b>139</b>
5.1	Evolutionary dynamics of CTCF binding at TAD boundaries . . . . .	140
5.2	Mutational landscape of TF binding profiles in hepatocellular carcinoma . .	141
5.3	Final remarks . . . . .	143
	<b>References</b>	<b>145</b>
	<b>Appendix A Supplementary figures and tables for Chapter 2</b>	<b>173</b>

<b>Appendix B</b>	<b>Supplementary figures for Chapter 3</b>	<b>179</b>
<b>Appendix C</b>	<b>Supplementary figures and tables for Chapter 4</b>	<b>183</b>
<b>Appendix D</b>	<b>Publications</b>	<b>185</b>



# List of figures

1.1	Transcription initiation at core promoters . . . . .	6
1.2	Evolutionary mechanisms of TF binding divergence . . . . .	13
1.3	Mutation rate distribution is affected by TF binding and nucleosome positioning that impede access of nucleotide excision repair (NER) machinery at the underlying nucleotide sequence . . . . .	15
1.4	Three dimensional genome organisation presented as a hierarchical model. .	20
1.5	TF-bound enhancer promoting transcription initiation at a core promoter via the formation of a chromatin loop. . . . .	21
1.6	Structure of CTCF domains and CTCF binding motif . . . . .	23
1.7	Organisation of the genome into topologically associating domains (TADs). .	25
1.8	Regulatory landscapes as a function of TAD formation . . . . .	25
1.9	Overview of loop extrusion model . . . . .	26
1.10	Organisation of the genome into compartments . . . . .	28
1.11	ChIP-seq experiment and data analysis . . . . .	33
2.1	<i>Mus</i> -conserved CTCF binding sites commonly occur at TAD borders. . . . .	42
2.2	Fractions of CTCF binding sites of different conservation levels in the studied <i>Mus</i> species. . . . .	43
2.3	Fractions of all <i>Mus</i> CTCF sites of each conservation level that are associated or not with TAD boundaries. . . . .	43
2.4	Fractions of TAD boundaries with CTCF sites of different conservation levels	44
2.5	CTCF binding sites at TAD boundaries are subjected to stronger evolutionary constraints. . . . .	46
2.6	Higher read coverage at TAD boundary-associated CTCF peaks compared to non-TAD-boundary-associated peaks. . . . .	47
2.7	Differences in representation of TE classes and their association with CTCF binding sites between TAD boundaries and other genomic regions. . . . .	50

---

2.8	Clusters of both conserved and divergent CTCF binding sites at TAD boundaries. . . . .	52
2.9	Examples of TAD boundary regions harboring clusters of both conserved and divergent CTCF binding sites. . . . .	54
2.10	Potential occurrence of CTCF site clusters also away from TAD boundaries	55
2.11	Inspection of the CTCF binding profile in BL6 confirms that CTCF sites form clusters in individual species . . . . .	55
2.12	Frequent overlap of clustered CTCF sites with cohesin and localisation close to genes . . . . .	57
2.13	Length distribution of genomic intervals occupied by singleton CTCF sites, “extended” singleton CTCF sites and clusters of CTCF sites . . . . .	57
2.14	Gene expression patterns around TAD boundaries are robust to local species-specific losses of individual CTCF sites . . . . .	59
2.15	CTCF binding at <i>Mus</i> -conserved sites is robust to hemizygous <i>Ctcf</i> deletion	60
3.1	Model of chemical carcinogenesis in mouse strains. . . . .	74
3.2	Outline of hepatobiliary differentiation based on single-cell RNA-seq data from fetal mouse liver . . . . .	79
3.3	Principal component analysis of expression profiles of bulk liver samples (LCE) and single cell data from mouse liver . . . . .	81
3.4	Expression heatmap of the marker gene clusters for BL6 bulk LCE samples and single cells across hepatobiliary cell differentiation . . . . .	82
3.5	Expression heatmap of the marker gene clusters for C3H bulk LCE tumours and single cells across hepatobiliary cell differentiation . . . . .	82
3.6	Expression heatmap of the marker gene clusters for CAST bulk LCE tumours and single cells across hepatobiliary cell differentiation . . . . .	83
3.7	Expression heatmap of the marker gene clusters for CAROLI bulk LCE tumours and single cells across hepatobiliary cell differentiation . . . . .	83
3.8	Distribution of expression levels of genes contained in the marker gene clusters <i>a-d</i> , in "DEN" and "None treatment" cohorts per species . . . . .	84
3.9	Number of tumour samples from each cohort with corresponding driver mutations. . . . .	85
3.10	Dysregulation of cell type marker genes in BL6 DEN-induced tumours . . .	86
3.11	Dysregulation of cell type marker genes in C3H DEN-induced tumours . .	87
3.12	Dysregulation of cell type marker genes in CAST DEN-induced tumours . .	87
3.13	Dysregulation of cell type marker genes in CAROLI DEN-induced tumours	88

---

4.1	Mutational profiles of DEN-induced tumours in each mouse species . . . . .	97
4.2	Identification and characterisation of genome-wide identified binding regions for each of the three liver TFs (CEBPA, FOXA1 and HNF4A) in three mouse species (BL6, CAST and CAROLI) . . . . .	99
4.3	Motif identification in the ChIP-seq TF peaks. . . . .	101
4.4	Content of aggregated TFBRs of each cistrome in canonical CEBPA, FOXA1 and HNF4A binding motifs. . . . .	102
4.5	Relative fractions of mutated and non-mutated TFBRs in each cistrome and sub-cistrome . . . . .	104
4.6	Mutation load per TFBR for each cistrome, in each mouse species . . . . .	105
4.7	Overview of approach developed to calculate weighted mutation rate ( <i>wmr</i> ) for a given cistrome or sub-cistrome, per tumour sample. . . . .	107
4.8	Distribution of weighted mutation rates per sample in each cistrome and sub-cistrome of the studied mouse species. . . . .	109
4.9	Correlation between counts of mutated trinucleotides ( <i>y</i> axis) in each aggregated sub-cistrome of every tumour and total counts of trinucleotides ( <i>x</i> axis) in each aggregated sub-cistrome (in reference genome) . . . . .	110
4.10	Representation of promoter-overlapping, enhancer-overlapping and no-CRE-overlapping TFBRs in the 1TF, 2TF and 3TF sub-cistromes, and vice versa	111
4.11	Mutation spectra per relative position around the CEBPA motif centre for motif occurrences identified in each TF sub-cistrome, in all studied mouse species . . . . .	113
4.12	Mutation spectra per relative position around the FOXA1 motif centre for motif occurrences identified in each TF sub-cistrome, in all studied mouse species . . . . .	114
4.13	Mutation spectra per relative position around the HNF4A motif centre for motif occurrences identified in each TF sub-cistrome, in all studied mouse species . . . . .	115
4.14	Observed versus Expected mutation loads per position of the extended CEBPA motif in the sub-cistromes of each species . . . . .	118
4.15	Observed versus Expected mutation loads per position of the extended FOXA1 motif in the sub-cistromes of each species . . . . .	119
4.16	Observed versus Expected mutation loads per position of the extended HNF4A motif in the sub-cistromes of each species . . . . .	120
4.17	Distribution of hyper-mutated TFBRs across the sub-cistromes . . . . .	121

4.18	Enriched GO terms associated with the hyper-mutated TFBRs in BL6 DEN tumours, as computed by GREAT. . . . .	123
4.19	Enriched GO terms associated with the hyper-mutated TFBRs in CAST DEN tumours, as computed by GREAT. . . . .	124
4.20	Enriched GO terms associated with the hyper-mutated TFBRs in CAROLI DEN tumours, as computed by GREAT. . . . .	125
4.21	Significantly dysregulated genes in DEN-induced tumours that are associated with hyper-mutated TF binding regions . . . . .	126
A.1	CTCF peak reproducibility among replicates of each Mus species. . . . .	175
A.2	Hi-C contact maps from published C57BL/6J liver data. . . . .	176
A.3	There is no evidence of any enrichment of specific motif words at TAD boundary regions among the species. . . . .	177
B.1	Distribution of transformed TPMs ( $\log_2(\text{TPM}+1)$ ) of expressed genes per bulk RNA-seq sample in each species. The tail samples of BL6 are marked red.	179
B.2	Distribution of transformed TPMs ( $\log_2(\text{TPM}+1)$ ) of expressed genes in single cells from fetal mouse liver . . . . .	180
B.3	PCA of expression profiles of bulk liver samples (LCE) and single cell data from mouse liver. LCE samples labeled by Treatment . . . . .	181
B.4	Distribution of expression levels of genes contained in the marker gene clusters <i>a-d</i> in cohort samples per species, grouped by driver mutations . . .	182
C.1	DEN mutational signatures . . . . .	183
C.2	Numbers of identified canonical TF binding motifs in the peaks of each TF cistrome and in CREs . . . . .	184

# List of tables

3.1	Number of available bulk RNA-seq libraries used by LCE . . . . .	78
4.1	Counts of DEN-induced tumour samples, mutations from all tumour samples, and median mutation count per sample in each species . . . . .	96
A.1	Mapping and peak calling statistics for CTCF ChIP-seq data in the five Mus species . . . . .	174
C.1	Genome-wide identified TF peaks in each mouse species . . . . .	183
C.2	Genome-wide identified promoters and enhancers in each mouse species, within the LCE consortium . . . . .	183



# Chapter 1

## Introduction

Contemplating the tremendous range of phenotypes occurring in the living world can be astounding if we consider that a) morphologically dissimilar organisms from different taxa share a substantial part of their gene repertoire, and b) highly variable cells of a single organism have the same genome. Yet similar gene repertoires, or even the same genetic code can give rise to greatly diverse phenotypes among distinct species, but also among functionally specialised tissues of a single organism. It is now known that a major factor underlying diversification of phenotypes are the molecular mechanisms that regulate gene expression. Therefore, unravelling the molecular basis of phenotypic diversity requires a profound understanding of the gene regulatory mechanisms; their components, their functions, and the way they differentiate, among lineages, or among tissues and cells, and even in genetic diseases and cancer.

### 1.1 "DNA makes RNA" - but how is this regulated?

*The genes speak:*

*"Functional units of DNA, we are;*

*Ultimate for all cellular activities;*

*Tailored to express as per tissue demands;*

*Mystery of our molecular actions await unfolding" [1]*

#### 1.1.1 Central Dogma of Molecular Biology

"DNA makes RNA, and RNA makes protein" has probably been the shortest circulated synopsis of the Central Dogma of Molecular Biology. Describing the flow of information

from genes to proteins, the Central Dogma was accurately summarised for the first time by Francis Crick in 1957 and published later in 1970 [2–4]. According to that, the DNA sequences of genes are transcribed into RNA molecules, which in turn provide the information for the synthesis of proteins through the process of translation. The genetic information can be passed on from nucleic acid to nucleic acid<sup>1</sup> and from nucleic acid to protein. As stated by Crick, “The central dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid” [3].

### 1.1.2 Gene expression in brief

The conversion of the genetic information into proteins, or in a broader context the process through which the genotype gives rise to the phenotype, is what we call *gene expression*. It is, in fact, a multi-step process that includes a series of complex, yet well-tuned mechanisms. It all starts with the transcription, the process through which a DNA strand is used as a template by RNA polymerase to generate a messenger RNA (mRNA) strand (described in further detail in 1.2). The generated mRNA molecule can go through further post-transcriptional processing steps, such as mRNA splicing, where introns are spliced out and the remaining exons are linked together into a mature mRNA molecule. The resulting mature mRNA is then transferred from the nucleus of the cell to the cytoplasm and directed to the ribosome to be translated, that is to instruct the polymerisation of amino acids into polypeptide chains with the help of the tRNAs. Synthesized peptides can subsequently fold into secondary structures and then, frequently combined with other peptides, make up proteins, which can often undergo post-translational modifications. The resulting functional proteins are then transported to specific places of the cell or to the extracellular matrix to perform various cellular functions.

### 1.1.3 Gene expression regulation

Eukaryotic cells do not express all of their genes continuously. In fact, different subsets of genes display distinct expression patterns, depending on their functional role in the cell, as well as the intra- and extra- cellular conditions at any given time point. For example, only a subset of genes that are involved in fundamental cellular functions, typically referred to as *housekeeping genes*, are expressed relatively constantly and in all cell types of an organism. On the contrary, genes that code for proteins involved in tissue-specific functions—the so-

---

<sup>1</sup>from DNA to RNA with transcription, from DNA to DNA with DNA replication, or from RNA to DNA with reverse transcription

called *tissue-specific genes*—are expressed only in certain cell types, while *inducible genes* are expressed as a response to environmental stimuli, or at certain phases of the cell cycle. Overall, spatial and temporal differences in gene expression underlie the high level of cell specialisation in multicellular eukaryotic organisms. These complex -yet precise- expression programs, being established during embryonic development, give rise to morphologically and functionally distinct cell types, while they also maintain cell type specificity in adult organs. The establishment and maintenance of these spatial and temporal gene expression patterns is governed by specialised programmes of *gene regulation*. Disruption of these tight regulatory programmes can lead to genetic diseases and cancer. It is therefore clear that unravelling the details of gene expression regulation is of utmost importance in understanding development, regeneration, aging, as well as disease onset.

In addition to phenotypic diversity within a single organism, variation in gene expression patterns as a function of distinct gene regulatory programmes also underlies -to a great extent- phenotypic diversity *between* different individuals, as well as species and lineages. A fair amount of evidence suggests that evolution of organismal complexity has been importantly facilitated by the progressive evolution of highly complex and precise gene regulatory mechanisms [5–7]. The extensive expansion of transcription factor families and the regulation of genes by combinatorial transcription factor binding (reviewed in section 1.2.2), the combinatorial use of *cis*-regulatory elements (reviewed in section 1.2.1) and the deployment of complex elaborate gene regulatory networks are only some of the manifestations of this progressive regulatory complexity [5, 8].

Gene regulation encompasses a range of mechanisms that can act at different levels of gene expression, including the initiation and rate of gene transcription, post-transcriptional processing and stability control of the RNA molecules, translation, post-translational modifications and stability of a synthesized protein [9].

Transcriptional regulation, in particular, is considered to be the most fundamental and frequent level in regulating gene expression [10, 9]. It is orchestrated by chromatin features, transcription factors, and other proteins working in concert to finely tune the amount of RNA being produced through a variety of mechanisms. Besides the various transcriptional regulatory proteins, additional factors, such as the organisation of DNA into chromatin, its modifications and structures, represent further levels of complexity to the transcriptional control [11].

This thesis focuses on the study of mechanisms of transcriptional regulation.

## 1.2 Transcriptional regulation

In eukaryotes, transcription mainly takes place in the nucleus of the cells, where DNA is -at a first level- packaged into chromatin shaping core structural units called nucleosomes, and further organised into higher level chromatin structures. The process of transcription includes three main phases: transcription initiation, elongation of RNA and termination. The newly synthesized RNA molecule can undergo further processing until it is used as a template for protein synthesis or to serve some other cellular function. Different types of RNA can be generated, including messenger RNA (mRNA), which is then used for the assembly of peptide sequences for proteins, transfer RNA (tRNA) and ribosomal RNA (rRNA), which are involved in the translation process, long non-coding RNA (lncRNA) molecules, as well as several classes of small non-coding RNA molecules that play various roles in modulating gene expression.

Transcription is regulated by a variety of mechanisms. These mechanisms involve a number of proteins that interact with specific regulatory DNA sequences and additional co-factor proteins, but they are also a function of the elaborate structure and packaging of DNA in the nucleus. Therefore, regulation of transcription involves information that is stored in the DNA sequence itself, specifically in *cis*-regulatory elements and in genes that encode *trans* factors, but also on top of the DNA sequence as epigenomic information. Epigenomic features are fostered through various mechanisms, such as DNA methylation, post-translational histone modifications, nucleosome positioning, and higher-order chromatin structures, small RNA-mediated regulation, and RNA editing [12].

Below I am presenting basic transcriptional procedures, as well as features of the genome and epigenome that constitute important checkpoints of transcriptional control. I start with transcription initiation, RNA Pol II and regulatory elements (section 1.2.1), transcription factors and their regulatory functions (section 1.2.2), as well as the genome organisation into chromatin (section 1.2.3) and into higher-order structures (1.2.4), which also have a powerful interplay with transcriptional regulation.

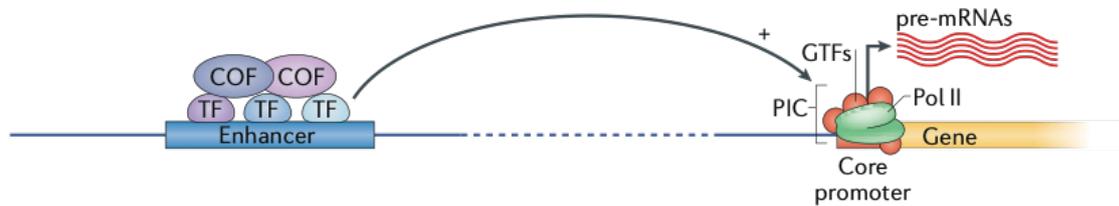
### 1.2.1 Transcription initiation and *cis*-regulatory elements

#### Transcription initiation, RNA Pol II and promoters

Transcription is performed by polymerase, an enzyme that uses as a template the sequence of DNA nucleotides to add one-by-one the complementary RNA nucleotides to a growing RNA strand. In eukaryotes, there are three types of RNA polymerases in the nucleus, each of which specialises in transcription resulting in different types of RNA. RNA polymerase I (RNAPI, or RNA Pol I) catalyses the transcription of ribosomal RNA (rRNA) genes, while RNA polymerase II (RNAPII, or RNA Pol II) transcribes all protein coding genes and some non-coding RNAs (e.g. snRNAs, siRNAs, miRNAs or lnc RNAs), and RNA polymerase III (RNAPIII) is responsible for the transcription of transfer RNAs (tRNAs) genes and small non-coding RNAs, such as 5S rRNA, U6 snRNA, SRP RNA [9]. Each polymerase needs a specified DNA sequence, called *promoter* and a particular set of proteins to initiate gene transcription.

Initiation of transcription is considered a crucial step in transcriptional regulation. It starts with the recruitment of RNA Pol II at the gene promoter region, which together with other proteins, the *general transcription factors* (GTFs; TFIIA, TFIIB, TFIID, TFIIE, TFIIIF and TFIIH), form the *pre-initiation complex* (PIC) [13, 14]. Specifically, first TFIID binds to the promoter and it then mediates recruitment of RNA Pol II, which is followed by PIC assembly. This is facilitated by a carboxyl terminal domain (CTD) in one of the RNA Pol II subunits that is rich in serine residues. When these residues are phosphorylated, the CTD can bind to and bring together different proteins that will assemble the PIC. Subsequent modifications of the CTD peptides throughout the transcription cycle will modulate the activity of RNA Pol II and its binding to other factors, as required, to facilitate elongation and termination of transcription, as well as provide the link to post-transcriptional processing [15]. Transcription ends when the RNA Polymerase encounters a sequence called *terminator*.

The promoter is the region upstream of the gene that harbours regulatory sequences, which can modulate gene transcription. A prerequisite for transcription initiation is for RNA Pol II to specifically bind to the *core promoter* (Fig. 1.1), which includes the transcription start site (TSS) at the 5' end of the gene, and a ~100 base pair window around it. Core promoters can also include short sequence motifs, such as the TATA box or the Initiator (Inr), that are recognised and bound by the GTFs, which subsequently recruit RNA Pol II [14]. The formation of the pre-initiation complex promotes the separation of the two strands in the



**Fig. 1.1 Transcription initiation at core promoters.** Transcription initiates at core promoters upon recruitment of RNA polymerase II (RNA Pol II) and general transcription factors (GTFs) that assemble the pre-initiation complex (PIC). Transcription can be activated by enhancers, typically distal regulatory elements, with the mediation of transcription factors (TF), and cofactors (COF). (Figure adapted from Haberle and Stark [16])

helical DNA, facilitating RNA pol to start the synthesis of a new RNA chain [17].

## Enhancers

The positioning of RNA Pol II at the TSS, although necessary, is usually not sufficient to generate more than just low basal transcriptional activity [14]. Transcription can be enhanced or repressed by distal, *cis*-acting regulatory elements, the enhancers (Fig. 1.1).

Enhancers are usually up to several hundreds of base pairs long and contain short sequence motifs, called transcription factor binding sites, that are recognised and bound by transcription factor (TF) proteins. A combination of TFs can bind at the enhancer regions, and in turn, they can recruit additional cofactors, protein factors that might not be capable of binding to DNA themselves but serve enzymatic functions important in transcriptional regulation, such as post-translational modifications of histones (reviewed in 1.2.3) [18]. Transcriptional cofactors are not required for basal transcription; they can be either activators or repressors, which can respectively increase or decrease the transcription from the core promoters. Co-activators can increase the transcriptional activity by enhancing the interaction of RNA Pol II with the DNA or appropriately regulating the structure of the double-stranded DNA [19, 20], as opposed to repressors, which prevent the binding of RNA Pol II to the promoter. Enhancers can act on their target core promoters, usually irrespectively of their distance and orientation [21, 22]. A number of studies provide evidence that sequence variation in enhancer regions and their TF binding sites is associated with diseases [23].

Tissue-specific expression patterns are largely attributed to the function of enhancers, as most enhancers have been shown to promote tissue-specific gene transcription. An example is

lymphoid cell-specific gene expression, where the synthesis of immunoglobulins is enabled via the association of an enhancer with *Ig* genes between J and C regions [9].

Other long-range regulatory elements are the insulators and silencers. These are also bound by TFs and have repressive effects on gene transcription [24, 25].

### 1.2.2 Transcription Factors

Transcription factors are proteins that recognise and bind specific short DNA sequences, and regulate transcription [26, 27]. TFs contain: a) a DNA-binding domain (DBD) that interacts with specific DNA sequences [28, 29], and b) an activation domain (AD), which bears sites where transcriptional coregulator proteins can bind [30]. It is possible for these two domains to reside on different subunits of the transcription factor complex. Sometimes TFs may also contain a signal-sensing domain (SSD) that can receive external signals, such as ligand binding, and transmit them to the rest of the protein complex generating a cascade for up- or down-regulation of genes. Overall, the ultimate function of TFs is to identify regulatory signals in the DNA sequence and transmit them into the process of gene expression control, as part of cellular functions and response to cellular environments.

#### **DNA binding specificity as a function of DBDs, TF binding motifs, and combinatorial binding**

Interaction of TFs with their target sequences largely depends on the structural features of their DNA-binding domains (DBDs). These features include conformations shaped by alpha helices, beta sheets, or disordered regions [31]. Based on their DBDs, TFs are classified into distinct groups and further into families [32, 27]. The largest fraction of the human and mouse TF repertoires is composed of three TF families: C2H2 zinc-finger TFs, including the CCCTC-binding factor (CTCF), homeodomain TFs and helix-loop-helix TFs [27, 33]. The differences in the structural characteristics of DNA binding domains (DBD) between distinct TF families are reflected to some extent in differentiation of their DNA binding specificities and, sometimes, also in their functional specificities. For example, many TFs that contain homeodomains are associated with developmental processes [32, 27]. On another note, the use of more than one DBD can increase the TF's recognition specificity. This can happen via repetition of tandem DBDs in a single TF or dimerisation of TF proteins.

Through its DNA binding domain (DBD) a TF can interact with specific DNA sequences (sometimes referred to as response elements), which can occur in *cis*-regulatory elements and are associated with specific genes. The interaction of TFs with the nucleotide sequences is the result of electrostatic and Van der Waals forces. Nevertheless, not all nucleotides in the sequences are in direct interaction with the TF. As a result, the binding sites of TFs do not include a specific unique sequence but rather a set of varying, yet closely related sequences, with a varying binding affinity for the TF. These sets of short, specific DNA sequences where TFs preferentially bind, namely TF binding motifs, are usually represented as position weight matrices (PWMs). Generated from the alignment of the multiple motif sequences that bind a TF, a PWM illustrates the probability of each of the four DNA nucleotides being present in each position of the multiple sequence alignment [34].

Since the sequences that are bound by TFs can be variable to some extent, and given their short length, TF binding sites can sometimes occur by chance in the genome. Nevertheless, this does not mean that they will all bind TFs, at least not in a stable or functionally relevant manner. Furthermore, the numerous TF binding sites existing in a genome are not all occupied continuously by TFs. Other components, such as chromatin accessibility or cofactor binding, can affect TF binding at those sites.

Another factor that affects DNA binding specificity is the synergistic activity of distinct TFs through their combinatorial binding, either directly or indirectly, on certain genomic loci [35]. Combinatorial TF binding is also considered to contribute to the specificity and flexibility required for orchestrating transcriptional regulation programmes in different tissues [5]. A possible distinction in a TF's functional role has been suggested based on whether it binds alone or in a combinatorial manner with other TFs. A potential scenario proposes that binding of a single TF might control more fundamental cellular transcription, as opposed to its combinatorial binding with other TFs that might help establish or maintain tissue-specific regulatory programs. Another hypothesis is that the more ubiquitous binding of a single TF can set up a general regulatory pattern for certain genes, while its more specific combinatorial binding can facilitate further fine-tuning of the expression patterns [27]. Several aspects of the combinatorial TF binding phenomenon, though, still require a better understanding [35].

### **Transcriptional regulation through transcription factors (TFs) and TF regulation**

In addition to direct recruitment of RNA Pol II at core promoters, the other main mechanism through which mammalian TFs regulate transcription is the recruitment of cofactors, i.e.

coactivators or corepressors [36]. Cofactors in turn contribute to transcriptional regulation through different mechanisms. A commonly used mechanism for activating transcription of various genes is the recruitment of the Mediator complex, a coactivator that mediates interaction between the TFs and RNA Pol II [37, 38]. Other cofactors include domains that catalyse modifications of RNA Pol II or modification of other TFs. Moreover, some cofactors are involved in remodelling nucleosomes and catalysing histone modifications [39], therefore having an impact on increasing or decreasing chromatin accessibility to the transcriptional machinery (nucleosome structures and histone modifications are reviewed in section 1.2.3). Finally, another way TFs can interfere with gene transcription is simply by occupying genomic sites, thus excluding RNA Pol II or other proteins from the region [40]. Importantly, many TFs bind cooperatively to specific sites in individual enhancers and mediate their physical interaction with distal promoters, which can also accommodate TF binding sites. Such TF-mediated promoter-enhancer interactions are facilitated through the formation of DNA loops and can regulate transcription of proximal or distant genes.

However, how are TF coding genes themselves regulated? One way of regulating TF gene expression is via negative feedback looping, where the TF protein -upon reaching a certain concentration threshold in the cell- can bind to its own gene body and prevent further transcription [41]. Alternatively, the function of TFs synthesized via translation in the cytoplasm can depend on binding a ligand that will help them translocate to the nucleus where they can exert their gene regulatory activity. This ligand binding step, or the lack thereof, can act as a regulatory point of their activity [42]. Another regulation mechanism is TF activation or deactivation through phosphorylation of the TF or activation through ligand binding to the signal-sensing domain of the TF. Furthermore, an important factor that can affect TF activity is also chromatin accessibility, which facilitates binding of TFs to their target sites and is controlled by nucleosome positioning and histone modifications, or methylation of CpG sites (reviewed in section 1.2.3). Finally, the function of a TF often depends on the availability of cofactors or other TFs with which the given TF can co-bind or interact to regulate the transcription of target genes in a synergistic manner.

### **Biological roles of TFs**

Functionally, transcription factors can be divided into those involved in basal transcription and those responsible for differential enhancement or repression of transcription to establish or maintain certain temporal and spatial expression programmes in different cells. The former include the GTFs, while the latter involve numerous TFs that serve a variety of

biological roles. One of these roles is regulation of the cell cycle [43] via control of cell growth and apoptosis by tumour suppressor or proto-oncogene TFs [43, 44], such as the *Myc* oncogene. Other TFs have important roles in cell differentiation and organ morphogenesis during development, through response to cellular stimuli; a well known example is the *Hox* TF family with roles in body pattern formation at embryonic stages [45]. Importantly, TFs also have important roles in cellular responses to endogenous stimuli, for instance hormone release, and environmental stimuli, such as oxygen concentration, as downstream effectors of signalling cascades. A particularly important functional role of TFs in downstream signalling cascades is in immune response [46]. Finally, it has been shown that TFs are also capable of inducing de-differentiation or transdifferentiation, i.e. reprogramming of a differentiated cell into a pluripotent cell or another cell type [47].

### **TFs in establishment of cell type-specific regulation programmes**

The ensemble of TFs that are active in any given cell type regulates transcriptional activation or repression of specific gene sets, defining the cell type-specific expression repertoire. It is generally suggested that the set of TFs involved in setting up and maintaining distinct cellular expression programmes include quite a limited number of TFs. These TFs are often referred to as *master transcriptional regulators* and in some cases, when they are ectopically expressed in other cell types, they are capable of reprogramming cell identity [48–59], (reviewed in Lee and Young [60]). Examples of known master regulators that dictate cell fate determination during development and maintenance of tissue-specific expression programmes in adult vertebrate organs are: the hepatocyte nuclear factor 4a (HNF4A) in liver, the myoblast determination protein 1 (MYOD1) in muscle, as well as the homeobox protein NKX2-5 in heart. A number of studies have suggested that the contribution of master TFs to regulating cell expression programmes is often through transcription elongation of cell type-specific genes (examples are provided by Bai et al., Park et al., Oven et al. [61–63]). An alternative view of expression differences between cells underlines the role of gene regulatory networks (GRNs) in establishing and maintaining cell type-specific expression, considering the concept of “master” regulators as oversimplified – but not mutually exclusive. According to this view, and in contrast to the master regulators scenario, the concept of GRNs takes into consideration the synergy and cooperativity of TFs and context-dependent activity in transcriptional control, pointing out that no single TF can regulate transcription of its target genes alone [46]. As an example in support of this viewpoint, the myogenic transcriptional regulator (MyoD), known to be capable of regulating transdifferentiation of multiple mammalian cell types into muscle cells when expressed ectopically [64], was

shown to in fact be part of a GRN that includes a number of TFs with potential in inducing transdifferentiation [46].

Tissue-specific gene expression patterns are largely conserved among mammalian species [65, 66], which implies that tissue-specific regulatory networks must also be under high evolutionary constraints, probably as a means to safeguard the functional specialisation of organs.

### **Evolution of TFs, TF binding sites, and regulatory networks**

Increase of organismal complexity is, to a large extent, associated with the evolution of highly complex gene regulatory networks, including increased complexity of TF repertoires and their employment in gene regulatory mechanisms.

It has been suggested that evolution of the eukaryotic lineage has been underlined by bursts of acquisition and expansion of TF families [5, 67]. There are a few TF classes, for example those involved in the basal transcription machinery such as the TATA box-binding protein (TBP or TFIIB), that originated in pre-eukaryotic taxa [68, 69]. However, most TFs in eukaryotes have originated in the lineage of the last eukaryotic common ancestor (LECA). This burst likely coincided with structural innovation in the eukaryotic nucleus and organisation of the genome in the form of chromatin [67]. More TF classes emerged after the LECA within different eukaryotic lineages, especially in plant and animal lineages, boosting the potential of cells for developing complex regulatory networks [67]. The major mechanisms mediating expansion of TF classes are gene duplications that resulted in large TF gene families, often with new domains [70, 71], and exaptation of transposable elements (TEs), such as in the case of the MUSTANG and FAR/FHY TF families in plants [72–75]. Finally, emergence of new TF classes via *de novo* gene birth might be another possible, although much less widespread mechanism [67].

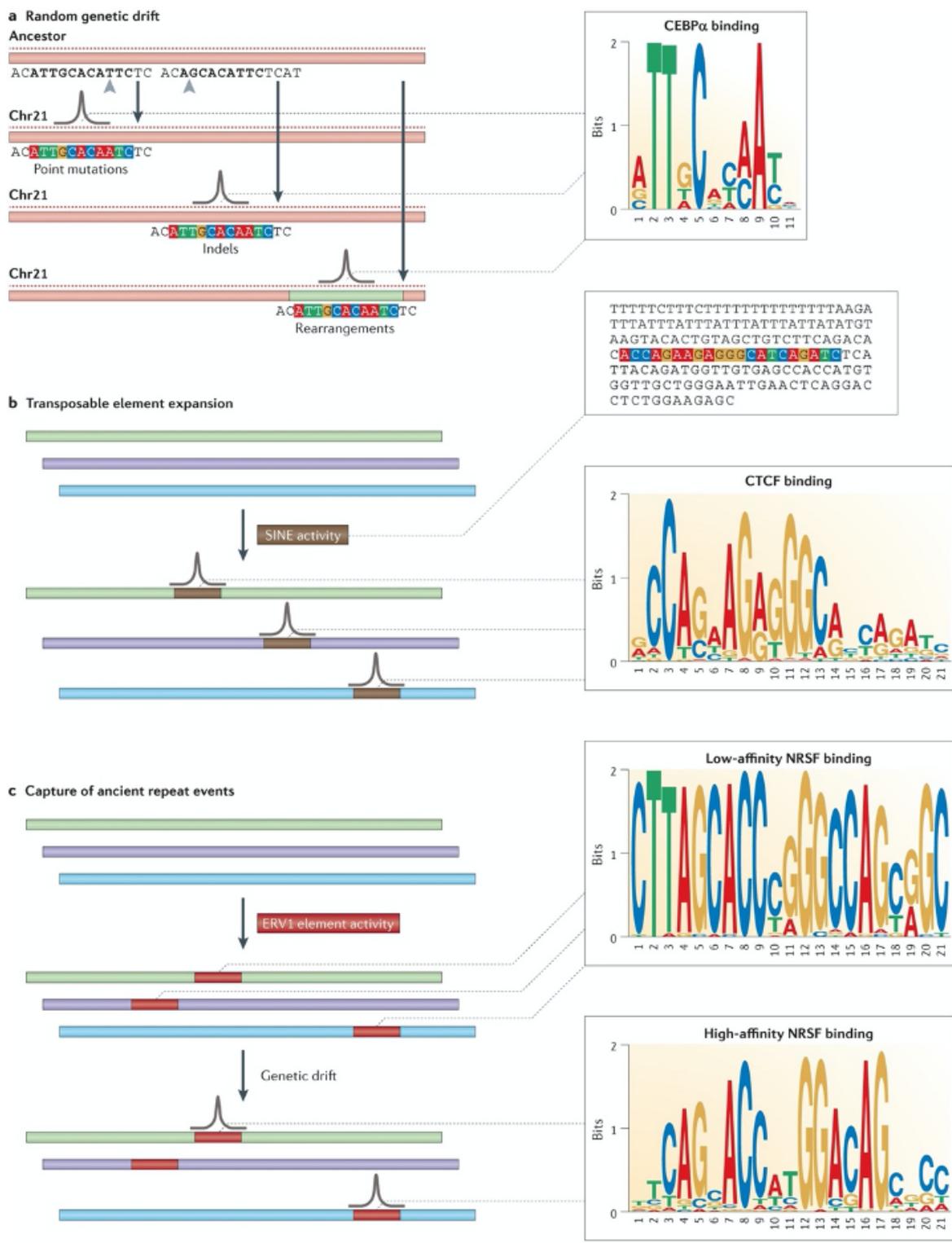
Despite these innovations in TF expansion throughout the evolution of eukaryotes, TFs generally evolve much slower than their regulatory sites [35]. Many TFs are largely conserved from *Drosophila* to human, even displaying highly similar sequence specificities as well as highly conserved physiological roles. One such example is the HOX family of TFs that contribute to body pattern formation during development [45, 35].

However, notwithstanding their functional constraints, the binding sites of the TFs are usually not as conserved, especially within the mammalian lineage. For example, OCT4 and NANOG

TFs display remarkable functional conservation between human and mouse, which, however, is not analogous with the conservation of their binding sites. It has been particularly shown that even though binding of OCT4 and NANOG is enriched around OCT4 target genes (genes downregulated upon OCT4 depletion) in both human and mouse, the exact location of their binding sites is largely nonconserved (reviewed in Villar et al. [76]). Similarly, CEBPA binding has also been shown to be highly variable among vertebrate species [77]. However, in this and other cases, it has been demonstrated that there is an extensive trend of turnover events, i.e. losses of TF binding sites and emergence of compensatory events nearby, especially among different mammalian species. Even over short evolutionary distances, binding of TFs exhibits large extent of variation among mammalian species. For example, a study on the binding of three liver-specific TFs (CEBPA, HNF3A that is encoded by *Foxa1*, and HNF4A) in closely related mouse species has shown substantial diversification of the binding events in orthologous regions among the mouse species [78]. The prevalence of such compensatory events suggests their role in maintaining expression outputs throughout evolution, while their evolution might be in line with the evolution of combinatorial binding of different TFs (reviewed in Villar et al. 2014 [76]). Combinatorial binding of TFs must be evolving in clusters of binding sites in the same vicinity that is co-bound by different TFs. Also, there is a correlation between binding strength and cross-species conservation. It is noted that in other taxa, such as in insects as seen by studies in *Drosophila spp.*, the TF binding profiles are more conserved across species, compared to mammals.

An emerging question is whether variation in TF binding events across species is reflected in variation of the underlying binding sequences. Villar et al. provide a review of relevant studies, and summarise that there is no clear association between TF binding alterations and sequence divergence among species [76]. This means that although a relatively restricted amount of TF binding changes is underlined by differences in the respective bound sequences, most TF binding differences are not associated with any change in the proximal genetic sequences. Nevertheless, they point out that the combination of the whole sum of regulatory sequences that exert their function in concert can be driving TF binding divergence. In some cases, a change of TF binding in absence of nucleotide substitutions in its binding motif sequence can be associated with sequence changes in other proximal TF binding motifs.

In any case, nucleotide substitutions in sequences that bind TFs can lead to degeneration of binding and loss of binding sites, but it is not likely for single substitutions to generate *de novo* binding sites. A widespread mechanism for acquisition of new binding sites, especially in mammalian lineages, is via expansion of different classes of transposable elements



**Fig. 1.2 Evolutionary mechanisms of TF binding divergence.** Birth of a TF binding event driven by: a) changes in the underlying sequence (efficient mostly in cases of birth of short binding motifs), b) expansion of transposable elements that deposit motif sequences at genomic sites, c) employment of ancient repetitive sequences, following sequence changes that increase binding affinity for the TF. Figure adapted from Villar et al. [76]

(TEs). Repeat elements might deposit short sequences with TF binding potential upon their insertion in mammalian genomes (Fig. 1.2). Alternatively, existing degenerate sequences in ancient repeats can at some point be repurposed by the genome and employed as a regulatory sequence in different regulatory networks (Fig. 1.2).

Overall, changes in TF binding, either as a result of sequence alteration or not, can be driven either by purifying and positive selection, but also be the effect of neutral evolution, mediated by genetic drift. Importantly, evolution of TFs is associated with evolution of gene regulatory networks (GRNs), which in turn can be the driving forces of phenotypic diversity.

### **Implications of TFs in mutagenesis and disease**

Disruption of the regulatory function of TFs has been linked to many disease phenotypes, including developmental disorders and cancer. Boyadjiev and Jabs reported in 2000 [79] that one third of developmental disorders in humans are associated with mutations in genes that encode TFs. Moreover, a large subset of tumour suppressors and oncogenes are TFs, whose dysregulated expression can have major implications in various cancer types [44, 80]. These misregulations are attributed to mutations within or near TF genes. However, more recent studies are also revealing a role for mutagenic processes in the binding sites of TFs in inducing development of tumours and other disorders (for examples see Zhou et al., Mazrooei et al. [81, 82]). It is, therefore, not surprising that TFs are of clinical significance, as they make up existing or potential drug targets in cancer therapies.

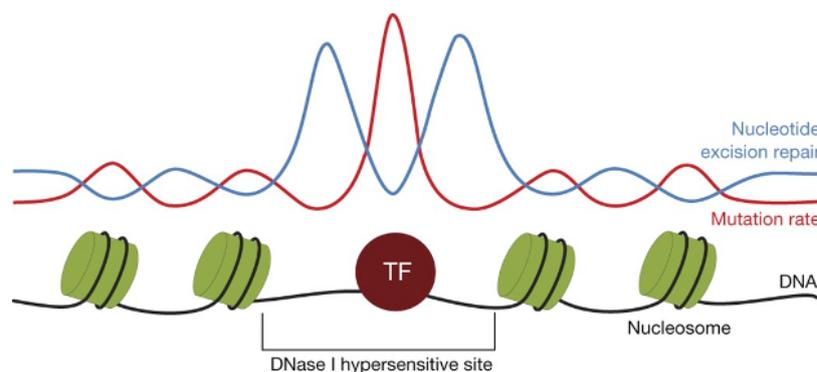
### **TF binding effect on mutational landscape**

Elevated mutation rates in TF binding sites is a common phenomenon in various cancer types [83–88]). This increase of mutation rates is attributed to reduced accessibility of the DNA repair machinery to the underlying TF bound sequences.

How do these mutations occur? It is known that, as all genomes are exposed to environmental or endogenous agents, DNA can get damaged. DNA damage is manifested as lesions, i.e. chemical alterations primarily at the nitrogenous bases, or alternatively at the sugar-phosphodiester backbone of DNA nucleotides [89]. The occurrence of unrepaired lesions can interfere with the process of DNA replication, which can result in the fixation of genomic mutations during synthesis of a new DNA strand. Unrepaired lesions can also be encountered by the transcription machinery and thus hinder the transcription process/transcriptional elon-

gation. Luckily, the cell is equipped with repair mechanisms that can deal with and resolve lesions. Mammals, in particular, recruit the nucleotide excision repair mechanism (NER). NER is a pathway that includes two sub-pathways, the global genome NER (GG-NER, or GGR) that deals with lesions throughout the whole genome and the transcription coupled NER (TC-NER, or TCR), or transcription coupled repair (TCR), that resolves lesions in actively transcribed genomic regions [90].

Sabarinathan et al. [84] have shown that both GGR and TCR enzymes have impaired access and activity at sequences that are occupied by bound TFs. Potential unresolved lesions in these sequences are therefore fixed into permanent mutations. As a consequence of the inefficient local DNA repair, TF bound sites exhibit a peak in their mutation load compared to the surrounding sequences. Similar to the bound TFs, the mutation rate in the flanking sequences also shows some periodicity, which is attributed to the nucleosomal structures hindering the access and action of repair enzymes to some extent (Fig. 1.3) [84].



**Fig. 1.3 Mutation rate distribution is affected by TF binding and nucleosome positioning that impede access of the nucleotide excision repair (NER) machinery at the underlying nucleotide sequence.** Figure adapted from Sabarinathan et al. [84]

### Global genome repair and transcription coupled repair mechanisms

The main difference between the two sub-pathways is the initiating step of lesion recognition and employment of the repair machinery. In eukaryotes, the initiating step of lesion recognition in GG-NER requires the coordinated employment of XPC and UV-DDB proteins [91–93]. In contrast, employment of TC-NER does not involve XPC and UV-DDB but seems to initiate with the stalling of the transcribing RNA Pol II when it encounters a lesion in the template DNA strand. Transcription elongation is hindered as RNA Pol II gets physically blocked from further translocation along the template strand, and it recruits the repair ma-

chinery [94]. However, if the lesion resides on the non-template strand, it will typically be recognised and repaired by the GG-NER mechanism. This is measurable as a difference in the speed of lesion repairs between template and non-template strands, making the repair on the template strand appear faster [95].

After the lesion recognition step, the two different sub-pathways converge to using the same mechanism. In brief, this includes recruitment of TFIIH and the XPA protein that will stabilise TFIIH at the genomic site, and leads to the verification of the damaged strand. TFIIH can then further “open” the damaged DNA strand [96] to facilitate the repair machinery’s access to the region. The endonucleases excision repair cross complementing-1 (XPF-ERCC1) and XPG incise the damaged strand segment at its 5’ and 3’ ends, causing the excision of a ~25-30 nucleotide long segment. The DNA replication machinery is then recruited, and its DNA polymerase components catalyse the synthesis of a short strand that will be incorporated in the DNA molecule by DNA ligases to replace the excised segment [97–100].

### **1.2.3 Chromatin architecture: small-scale organisation of the genome and epigenome**

At its most basic level the first-order structure of DNA, i.e. its nucleotide sequence, encodes information for residue-by residue synthesis of proteins. At a further level though, the local nucleotide composition of DNA sequences is a substrate for certain structural properties of the DNA molecule, which can affect important molecular functions, such as the accessibility of DNA to transcription factors and RNA Polymerase.

For example, poly-A strands of DNA have been shown to be more rigid [101], while AT-rich DNA sequences are known to provide a higher local flexibility to genomic segments. These structural properties can affect nucleosome positioning and the accessibility of chromatin to transcriptional proteins [102], which in turn affect the rate of transcription. In addition, the nucleotide cytosine can be converted to 5-methylcytosine (5mC), or 5-hydroxymethylcytosine (5hmC) via methylation or hydroxymethylation, respectively, a property that makes up an important epigenetic level of information that largely affects transcriptional control.

Beyond the nucleotide composition of the DNA sequence itself, described above, the packaging of the eukaryotic genome in the form of chromatin is itself an important level of

transcriptional control. Chromatin consists of DNA, proteins and RNA. Generally, transcription of genes is facilitated/higher at regions where chromatin is less condensed, thus more accessible to certain protein factors. The level of chromatin compaction is associated to covalent modifications of histone tails and DNA.

### **DNA methylation**

Cytosine methylation is catalysed by the enzyme family of DNA methyltransferases (DNMTs) which covalently add a methyl group to the fifth carbon of cytosine. In mammalian genomes CpGs are relatively sparse and a large fraction of them, ~70-80%, are methylated [103]. Cytosine is typically methylated in the context of CG dinucleotides (CpG), which occur with a high frequency in short DNA sequences, namely CpG islands (CGI) [104]. However, despite its pervasiveness, mammalian CpG methylation is considered to have different functional impacts on different genomic regions.

Typically, open chromatin regions are accessible to transcription factor binding, are more transcriptionally active and display low levels of DNA methylation [105]. In contrast, high levels of CpG methylation -which also have a reported role in heterochromatin formation [106]- are associated with decrease or loss of transcriptional activity, due to the local occupancy of methylated DNA by methylcytosine binding proteins. As a consequence transcription factors, such as the CCCTC-binding factor (CTCF), and other proteins do not have access in the methylated regions and are prevented from interacting with their specialised binding sites [107, 108]. So, there have been reports of decreased binding of particular TFs to their corresponding binding sites on the genome when they are highly methylated [109]. Impediment of TF binding and subsequent transcription activation by DNA methylation is a regulatory mechanism in the promoters of genes, which often occur in CGI regions. Nevertheless, CGI promoter methylation is not common, and promoter inactivation is usually mediated through a histone modification, H3K27me3 (explained in the following section), which is a more easily reversible form of transcription repression than CpG methylation [110].

It is important to note that, despite the generalised statement that TFs are excluded from CpG methylated sites, emerging evidence suggests that there are cases of TFs, especially in development, that do bind to methylated sequences [109, 111].

Seemingly contrary to the silencing role of DNA methylation in promoters, gene bodies are often enriched in DNA methylation but positively correlated with transcriptional activity,

especially nucleosome associated genic segments that are exon-rich, [112–114]. Proposed explanations of this phenomenon are that DNA methylation within the gene bodies might play a role in transcription elongation and mRNA splicing, or that it is involved in repression of alternative cryptic promoters that may lie within the gene [115].

DNA methylation has been shown to play essential roles in transcriptional repression of transposable elements, the monoallelic expression patterns of specific-parent-origin imprinted genes, and the inactivation of the X chromosome [116].

### **Nucleosome positioning and histone modifications**

Besides DNA methylation, chromatin accessibility is also modulated by nucleosome positioning and post translational modifications of the nucleosome histones. These are both key components of chromatin architecture, which not only pack the ~2m long DNA in the cell nucleus, is a primary component of its functional organisation with implications in regulating selective access to its regulatory regions by regulatory factors of transcription [117, 118].

Chromatin organisation changes dynamically during the cell cycle phases, super-coiling the DNA into coerced chromosomes during mitosis to facilitate the division of the chromosome copies into the daughter cells, while getting a loose but complex distribution during the interphase to allow for the regulated transcription of genes, as well as DNA recognition and repair.

The nucleosome is the structural unit of chromatin organisation. The double-helical DNA wraps around octamers of histone proteins to form nucleosomes. A nucleosome typically includes ~147bp of DNA sequence, with up to ~90bp long linker DNA segments bridging the nucleosomes, and the histone octamers include a dimer of each of the histone proteins H2A, H2B, H3, and H4 [119, 120]. Each histone has an amino-terminal end, overhanging from the nucleosome, that can undergo various post-transcriptional modifications (PTMs), the most important and well-studied of which are methylation, acetylation, phosphorylation and ubiquitylation [121]. The different types of histone modifications affect the way histones interact with the nucleosome DNA and can be either positively or negatively correlated with transcriptional activation, although transcriptional regulation is usually not exclusively dependent upon a single histone PTM type. Acetylation on the lysine residues (K) of the amino-terminal histone tails reduces the positive charges of the tails and decreases their binding affinity to DNA, which is negatively charged, resulting in looseness of the nucleosome

structure and local increase of chromatin accessibility [9]. Methylation of lysines in histone tails can either increase or decrease the affinity between histones and DNA depending on the lysine residue that is subject to the modification and the number of methyl groups added to it. For example, core promoters of transcriptionally active genes are typically enriched for H3K4me3 and H3K27ac, while active enhancers are associated with H3K4me1 and H3K27ac. Within gene bodies, active transcription is associated with H3K36me3, while transcriptional repression is paired with H3K9me3 or H3K27me3 [122]. H3K9me3 is also involved in repair of double-strand breaks of DNA during homologous recombination, by increasing local accessibility to repair enzymes [123].

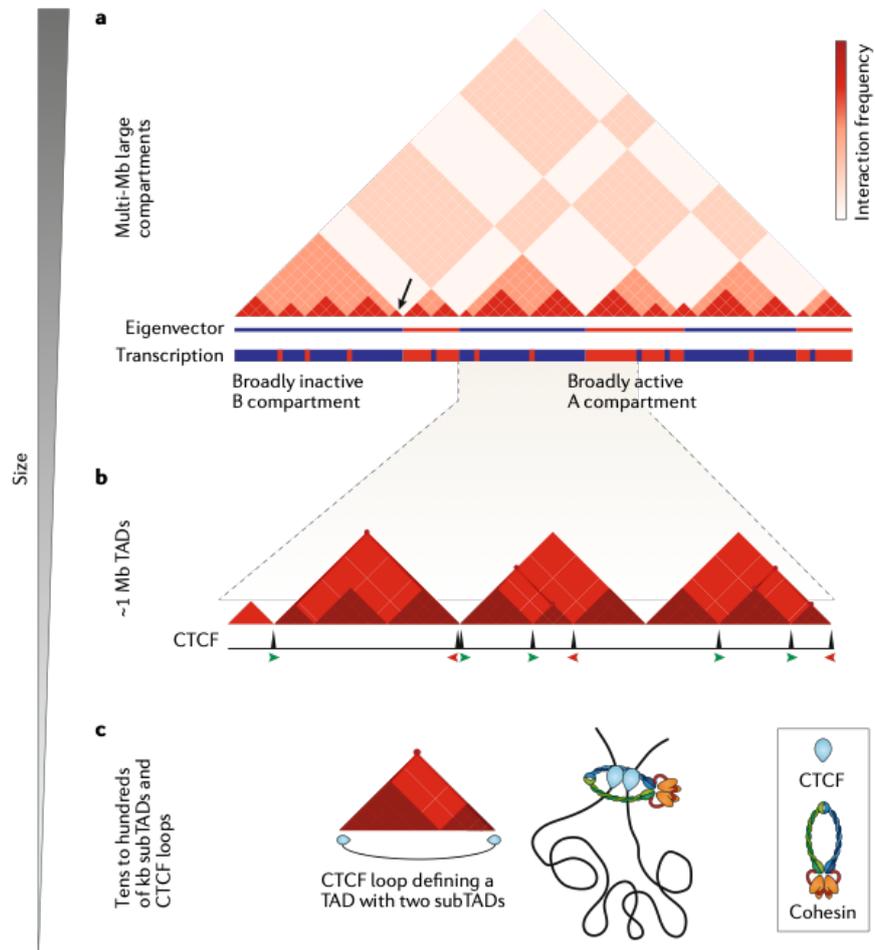
The positioning of nucleosomes also has implications in transcription regulation. Specifically, accessibility of genomic regions to specific proteins can be modulated by nucleosome remodelling or repositioning. This is performed by chromatin remodelling complexes that can rotate the double helix of DNA around the nucleosomes to unwind TF binding sites, as well as displace or slide nucleosomes along the chromatin structure to allow the interaction of TFs with regulatory sequences around gene TSSs [124–126].

### 1.2.4 Higher-order chromatin structure

Eukaryotic genomes are organised in a refined three-dimensional conformation comprising distinct structural layers, which associate with important regulatory functions, and range across various scales [127–130]). These include chromatin loops, topologically associating domains (TADs), compartmental domains and chromosome territories [131–135]. These structural formations organise the regulatory information in distinct regulatory landscapes, which promote specific interactions of regulatory elements and disable others [136–138]. Querying the characteristics of this structure has become easier and more accurate with the advent of 3C-based (Chromosome-Conformation-Capture) techniques, especially with Hi-C [131, 139]. Although the 3D genome structure has been thought of as a determinant of genomic interactions, and by extension of transcriptional regulation, newer research shows it can also be the result of functional components of chromatin and transcription [140].

Below, I will describe the traditional hierarchical layers of 3D genome structure (Fig. 1.4) from lower to higher scales. In addition, I will provide further details for some of their important architectural proteins components, such as CTCF and cohesin. Finally, I will refer to open questions about the mechanisms of 3D genome structure formation, as well as to a recent proposal for a revision of the hierarchical model towards the adoption of a

non-hierarchical model.

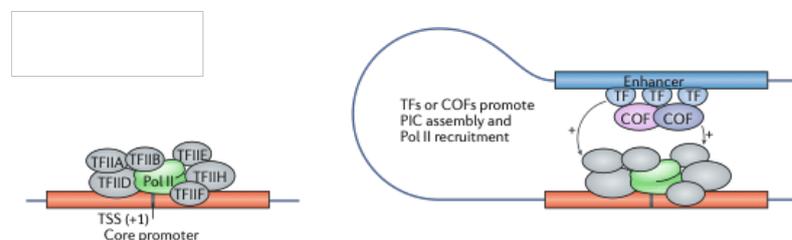


**Fig. 1.4 Three dimensional genome organisation presented as a hierarchical model.** The panel shows a cartoon version of a Hi-C contact map of relatively low resolution. a) At the multi-megabase scale the genome is organised into A and B compartments (respectively red and blue). b) At the sub-megabase scale the genome is organised in topologically associating domains (TADs), which further contain smaller sub-TADs and frequently coincide with loops between interacting CTCF sites with convergent orientation; CTCF loops are identified as strong punctate signals. c) A TAD that coincides with a CTCF loop and contains two sub-TADs (Figure adapted from Rowley and Corces [140])

### Chromatin loops

As explained in the previous sections, enhancers are essential players in transcriptional control, as they bind TFs and interact with promoters to regulate the expression of specific

genes. In a similar way, insulators can also act on gene promoters to block an enhancer's activity [21, 141, 142]. However, it is known that interacting *cis*-regulatory elements can be located thousands of bases apart on the linear genome. Their interaction, which has an effect on the transcriptional output, is enabled through the formation of chromatin loops that bring these distal regulatory elements in close proximity in the 3D space of the nucleus [143–145, 134], and the mediation of the transcription factors that are bound to them (Fig. 1.5). As an extension to an individual loop, more complex chromatin loop structures can be formed when multiple genomic contacts occur simultaneously. Such structures/conformations have been referred to as active chromatin hubs (ACH) (reviewed by Pombo and Dillon [146]).



**Fig. 1.5 TF-bound enhancer promoting transcription initiation at a core promoter via the formation of a chromatin loop.** Left: Assembly of pre-initiation complex (PIC) and recruitment of RNA Pol II at a core promoter. Right: Enhancer promoting PIC assembly and RNA Pol II placement at a core promoter via recruiting transcription factors (TFs) and cofactors (COFs) and upon chromatin looping. Figure adapted from Haberle and Stark [16]

In the human genome almost 10,000 loops have been identified, with many of them sharing the same boundaries, or so-called loop anchors. The majority of loop anchor loci, i.e. the distant genomic loci that interact, are less than 2Mb apart [134]. Moreover, a number of loops are also shared between cell types, while others are cell type-specific. Genes whose promoters are anchored by a loop display generally higher expression levels than genes not associated with loops. By extension, cell type-specific loops were suggested to be associated with expression differences between cell types [134]. Generally, chromatin loops are dynamic and can differ among the cells of a population.

Rao et al. [134] also reported that 86% loops in the human genome are anchored by the CCCTC-binding factor (CTCF) and/or cohesin. Indeed, CTCF together with cohesin have been identified as essential factors in promoting interactions between distal genomic loci [147–149]. The majority of CTCF-anchored loops -92% based on a Hi-C experiment and 65% based on ChIA-PET data- have been reportedly shaped by pairs of CTCF binding motifs with convergent orientations, i.e. pointing to each other [134, 150] (see more details on

CTCF motifs in the following section). Indeed, CRISPR -mediated inversion of a CTCF binding motif was shown to cause ectopic loop formation and introduction of new interactions between regulatory elements [151].

## **CTCF and cohesin**

### **Cohesin**

Cohesin is a protein complex with a ring-like structure and is made up of the subunits, SMC1, SMC3, SCC3 and RAD21/SCC1 [152–154]. It is known to be involved in sister chromatid cohesion for chromosome segregation, homologous recombination and DNA damage repair action. As a function of its shape, cohesin can entrap two chromatin segments in *cis* resulting in the formation of a chromatin loop. It has been shown to establish or maintain chromatin loops between *cis*-regulatory elements and facilitate the binding of TFs at enhancer regions [37, 155]. Cohesin-occupied sites on the genome largely colocalise with CTCF binding sites [156], reflecting also their joined contribution to establishing three-dimensional genome structures.

### **CTCF**

CTCF is an 11 zinc-finger transcription factor that contains a deeply conserved DNA-binding domain (DBD), from *Drosophila* to human [157–159] (Fig. 1.6a). In mammals, this DBD is flanked by two loosely structured tails: an amino- and a carboxyl- terminal end [160, 161]. CTCF is ubiquitously expressed in all cell types. Initially, it was identified as a repressor of the *C-MYC* oncogene [162, 159]. It was long considered just as an insulator protein, binding to insulators between pairs of enhancers and genes and thus blocking gene expression [163]. However, numerous studies over the last decades have shown that it is an unusual transcription factor with multiple biological roles, including promoting interactions between regulatory elements, activating or suppressing genes, importantly contributing to gene imprinting during development and X chromosome inactivation. One of its most important properties is facilitating chromatin loop formation together with cohesin, explaining its common characterisation as an architectural protein. As described above this property of CTCF is essential in establishing and maintaining higher order genome structures, such as TADs and compartments.

CTCF preferentially binds to a 19 bp long primary motif (Fig. 1.6b), or sometimes to an extended 40bp-long sequence, including its primary motif and shorter secondary ones [164]. Upon binding, it can bend the double helix of DNA in different ways [165]. It has been demonstrated that only four out of the eleven Zn-fingers of its DNA binding domain are essential for CTCF binding to the core 12 nucleotides of its primary motif [158]. Its ability to bind chromatin is greatly affected by local CpG methylation, as CTCF preferentially binds to hypomethylated regions [166].

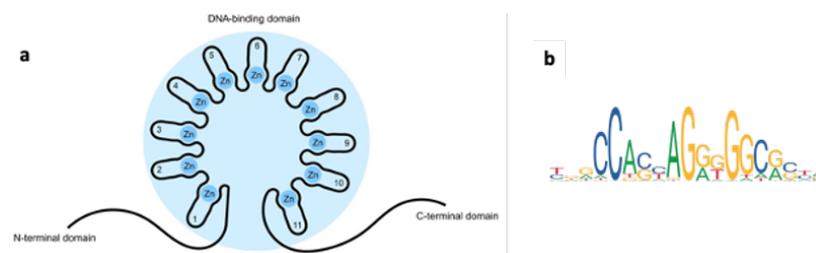


Fig. 1.6 **Structure of CTCF domains and CTCF binding motif.** a) N-terminus domain, DNA binding domain with eleven zinc fingers, and C-terminal domain. Figure adapted from Arzate-Mejia et al. [167]. b) Position weight matrix (PWM) of CTCF binding motif (MA0139.1), derived from JASPAR database [168]

High throughput interrogation of mammalian genomes via chromatin immunoprecipitation has revealed the approximately 40,000–90,000 sites that bind CTCF, with the exact number depending on antibody specificity and the computational parameters used in binding site identification. About 30-60% of these sites are cell-type specific with around half of them being in intergenic regions and half in promoter elements [169, 156, 170–173]. It has been reported that during transcription, RNA Pol II can dislocate CTCF molecules from their binding sites [174].

Cross-species comparisons of vertebrate CTCF binding profiles have shown that CTCF binding sites are conserved to a greater extent among mammalian species, compared to other, tissue-specific TFs [77, 175, 164, 76]. Different lines of evidence, especially but not limited to studies in rodent species, support that a major evolutionary mechanism of CTCF binding site expansion has been through insertion waves of transposable element families that carried and deposited potential CTCF-binding sequences, which were then exapted by the mammalian genomes [176, 175, 164, 177]. The relatively high species-conservation level of CTCF binding profiles likely reflects the multitude of CTCF biological functions, which might also be conserved to some extent, as well as its importance in organizing regulatory

landscapes on chromatin.

CTCF can form homodimers or heterodimers, by physically interacting with other CTCF molecules [148, 147, 178], cohesin or other proteins, respectively. A recent study has shown that CTCF also contains an internal RNA binding region (“RBRi”), which can often be utilised for RNA-mediated interaction between separate CTCF molecules that can come together and form small CTCF hubs, i.e. CTCF protein clusters [179]. Among CTCF anchored loops, some are RBRi-dependent, i.e. disappear upon RBRi-deletion from the CTCF molecules [179], and some are RBRi-independent. Dysregulation of CTCF or its binding is also known to be implicated in the emergence of diseases, such as developmental and neurological disorders, immune disorders, and cancer. Interestingly, CTCF, as well as cohesin subunits genes display some of the highest mutation rates in various cancer types [180, 181].

### **Topologically associating domains**

A number of studies have reported partitioning of the eukaryotic genome at a sub-megabase scale into topologically associating domains (TADs) [133, 135, 182], referred to also as topological domains [133], physical domains [182] or contact domains [134]. These regions, ranging from tens to hundreds of kilobases in length, are characterised by high frequency of self-interactions, while little or no interaction is observed between neighbouring regions. This means that a genomic locus within a TAD is more likely to interact with another genomic locus of the same TAD, rather than loci outside of this domain. The fact that these contact frequency patterns are not continuous, but rather appear as distinct genomic blocks, indicates that TADs have certain boundaries and are insulated from neighbouring TADs [183] (Fig. 1.7). The fact that TADs have been identified in taxa as diverse as humans, mice and flies, demonstrates that chromatin organisation in such structures is an evolutionarily conserved genomic feature [183]. It has also been suggested that TADs are highly conserved across species and cell-types [133, 184].

As a consequence of their structure, TADs modulate interactions between regulatory elements and, thus, play a prominent role in transcriptional control (Fig. 1.8) [135, 186, 187, 146]. Studies have revealed cases where disruption of the boundaries of TADs can lead to ectopic *cis*-regulatory contacts that cause diseases [188–190]. Altogether, these findings underline the importance of establishment and maintenance of these chromatin structures.

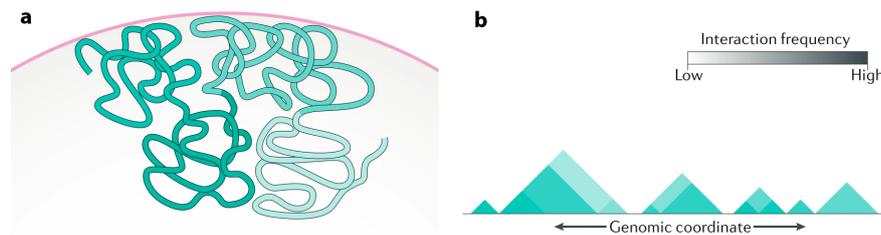


Fig. 1.7 **Organisation of the genome into topologically associating domains (TADs).** Cartoon illustration of a) chromatin conformation in TAD organisation, in the cell nucleus, and b) Hi-C heatmap of TADs that contain intra-TAD genomic contacts. Figure adapted from van Steensel and Furlong [185].

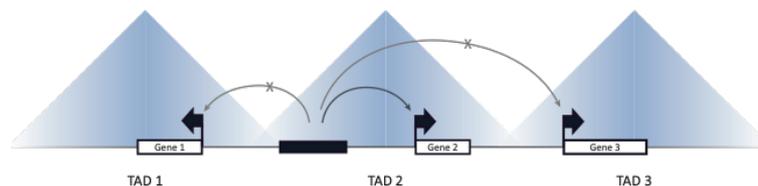
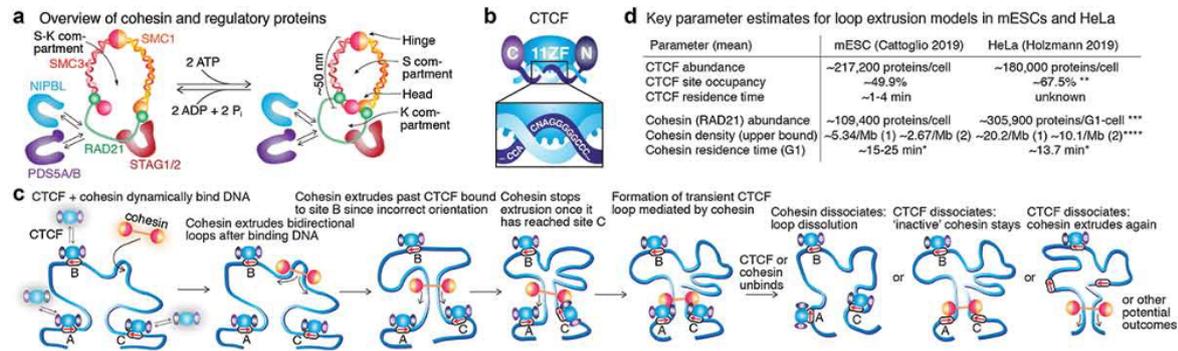


Fig. 1.8 **Regulatory landscapes as a function of TAD formation.** Enhancers can, typically, interact with gene promoters lying within the same TAD, and not in different TADs. Figure adapted from Zabidi and Stark [191]

A question that has emerged with the first reports of genome organisation into TADs was how these domain structures are formed, or more specifically, how their boundaries are defined. Evidence from these first studies on TADs had already shown that their boundaries are enriched for CTCF sites [133, 135], indicating an important role of this architectural protein in TAD formation. The proposed mechanism of TAD formation is via loop extrusion: the cohesin protein complex acts in *cis* on chromatin and slides on it forming a growing loop, until it meets two bound CTCF motifs that have convergent orientations and prevent cohesin from further sliding (Fig. 1.9) [192–195]. The apparent requirement of CTCF sites to be convergently oriented is often referred to as the convergent rule. The combined action of CTCF and cohesin maintain the TAD as a chromatin loop structure until these two architectural proteins dissociate from DNA. This indicates the dynamic nature of both the loop establishment via the extrusion mechanism and loop maintenance [193, 173, 128].

### Role of cohesin and CTCF in TAD formation

The importance of CTCF and cohesin in forming and maintaining TADs is supported by reports of major disruption of TAD organisation of the genome upon acute depletion of these



**Fig. 1.9 Overview of loop extrusion model.** a-b) Structure of mammalian cohesin (a) and CTCF (b). c) Illustration of loop extrusion mediated by cohesin and convergently oriented CTCF sites. (d) Parameters that can constrain loop extrusion models in mouse embryonic stem cells and human HeLa cells. Figure adapted from Hansen [179].

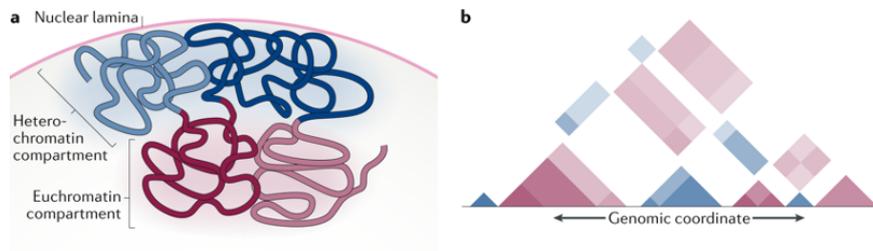
two proteins [196–200]. Specifically, Schwarzer and colleagues (2017) deleted the *Nipbl* cohesin loading factor to achieve four-fold to six-fold decrease of the total amount of cellular cohesin that is loaded on chromatin, and they reported clear disappearance of TAD structures [198] even though the CTCF site occupancy was not affected. In consistency with these findings, Rao and colleagues (2017) also observed disappearance of TAD loop structures upon cohesin loss and re-formation of TADs after cohesin recovery [197]. Furthermore, Gassler and colleagues (2017) reported substantial or entire loss of TADs and loops upon genetic knockout of the *SCC3* protein (which forms cohesin's ring structure that extrudes chromatin) in zygotes, while the cohesin unloading factor *WAPL* affected the length of the loops [200]. In another study, Wutz and colleagues (2017) once more underlined the necessity of cohesin for TAD and loop formation, based on effects observed upon auxin-induced proteolysis of the cohesin subunit *SCC1* [199]. In addition, they also degraded CTCF using auxin and observed some effect on the TAD and loop structures, although clearly weaker compared to the effect of cohesin depletion. In a separate study, Nora and colleagues (2017) reported a rather strong effect of acute auxin-driven CTCF depletion on TAD structures; they observed loss of contact insulation between TADs, although some TAD boundaries remained visible on the contact maps even after the depletion of CTCF [196].

The vast majority (i.e. more than 90%) of identified cohesin-occupied genomic sites in mammalian genomes have been found to co-localise with CTCF binding sites [156, 173, 201–203] (reviewed by Hansen [171]). Insight from other studies shows that CTCF is essential for localising and stabilising cohesin at specific genomic sites for loop formation, but not for loading cohesin onto chromatin [196, 201, 201, 204] (reviewed by Hansen [171]).

Despite the firm indications of the important role of CTCF binding in TAD formation, several details about its functional implication remain unresolved and obscure its exact role. As described above, the acute depletion of CTCF does affect TAD boundary determination but to variable degrees. It is also known that CTCF binding is so abundant in the genome that only a small fraction (15%) of CTCF binding sites coincide with TAD boundaries, meaning that CTCF binding alone is not enough to demarcate the boundaries of these self-associating regions [133]. With respect to the so-called convergent rule, not all CTCF sites oriented in a convergent manner demarcate TADs, and, vice versa, not all TAD boundaries are marked by convergent CTCF sites [134, 205, 151, 150]. In fact, although most TADs apparently are defined by CTCF loops, as shown by a number of studies that measured the degree of interaction bias upstream or downstream of a genomic region [133] to computationally predict TADs [206, 133, 135, 182, 207], a small number of them are not flanked by CTCF binding sites [140]. Such non CTCF-dependent TADs have also been called ordinary domains, suggested to be “passively” formed as a consequence of the formation of adjacent CTCF-flanked TADs or transitions between active and inactive chromatin states [134]. Nevertheless, the appearance of ordinary domains might be a result of the dynamic nature of TADs in a population of relatively heterogeneous cells, rather than being a simple passively appearing phenomenon [128].

### **Compartments**

Chromatin is also organised into two compartments, A and B, defined as sets of chromosomal regions with distinct biochemical features, which display long-range interactions with regions from the same set more frequently than expected by the “random polymer conformation of a chromosome” [131, 132, 183]. Compartments were defined for the first time as multi-megabase-scale structures, based on Hi-C contact maps that revealed a plaid pattern of long-range chromosomal interactions (Fig. 1.10) [131]. Importantly, compartments A and B have been found to associate with distinct biochemical features of chromatin. Compartment A regions are enriched in active histone marks, such as H3K36me3, are more gene rich and associated with increased transcriptional activity, while they also contain some genes that are silenced, and are more accessible to DNase I. They generally overlap the so-called open chromatin regions, as opposed to regions in compartment B, which are less rich in genes, associate with inactive chromatin marks and low transcription rates [131]. The genome organisation into A and B compartments also reflects the segregation of chromosomes respectively into euchromatin and heterochromatin [183].



**Fig. 1.10 Organisation of the genome into compartments.** Cartoon illustration of a) compartment organisation in the cell nucleus as a result of aggregation of regions with similar biochemical and functional features. The main compartments are heterochromatin and euchromatin, and b) Hi-C heatmap that shows self-association of euchromatin and heterochromatin domains over long genomic distances. Figure adapted from van Steensel and Furlong [185].

### Revision of the “hierarchical” model of 3D genome organisation

As outlined above, the different structural layers of spatial chromatin organisation -mainly TADs and compartments- have been identified as structures at different size scales. Due to this difference in relative sizes, TADs and compartments have been “traditionally” viewed as components of a nested hierarchical model of genome structure [208, 209]. This view was based on the first reports identifying and characterising TADs [133] and compartments [131] based on Hi-C data that generated contact maps of megabase resolution. Later studies that improved the sample sizes, sequencing depth and algorithms for computational identification of these features, reported Hi-C-based contact maps of higher resolutions, for example binning the human genome up to 1kb [134]. That resulted in the identification of self-associating contact domains -i.e. TADs of smaller size than previously thought, just a few kilobases long on average. Some studies also reported sub-divisions of TADs into smaller sub-TADs [210, 134]. Similarly, combined PCA and Hi-C data-based contact maps of resolution as high as 10kb have identified compartments of smaller size than initially reported [211, 212]. Specifically, in *Drosophila melanogaster* A compartmental domains (i.e. A compartment regions of small size, as a result of high resolution) were identified with a median size of 15kb, which also corresponded to transcriptionally active chromatin domains [211]. Furthermore, high-resolution Hi-C contact maps allowed the identification of more sets of chromosomal regions based on their epigenome marks and their interaction patterns with each other, leading to the characterisation of six sub-compartments, rather than just two major compartments: subcompartments A1 and A2 with mostly active histone marks, and subcompartments B1, B2, B3 and B4 with features matching those of heterochromatin [134]. These findings imply an apparent increase of the number and decrease of the size of identified TADs, as well as of compartment types and regions, as the resolution of the

contact maps increases [134, 213, 214, 183]. This shows that the characterisation of distinct structural components based on their appearance on contact maps can be fluid and dependent on the experimental and technical parameters, such as the sample size, the sequencing depth, the experimental methods and algorithms used to identify and report them.

Another important aspect to account for is that the state of CTCF binding, cohesin stabilisation on chromatin and the three-dimensional genomic structures are not static, but rather dynamic features of the genome. They all re-model during the cell cycle, interplaying with other dynamic molecular processes, such as chromatin accessibility and transcriptional regulation [215, 216, 128, 217]. In addition, both the ChIP-seq profiles and the Hi-C maps we obtain using the respective experimental methods represent only snapshots of the genome state in the nuclei of populations of variable cells at a given time point.

Reviewing these aspects of chromatin organisation, Rowley and Corces propose the consideration of the 3D genome more as a dynamic balance between compartmental domains and CTCF loops and their functions, rather than a hierarchical structure model [140]. In another review with respect specifically to TADs, Elzo de Wit suggests that “a more constructive path will be to explain TADs in the light of the mechanisms that form them, rather than describing TADs as we see them” [218], which can also more broadly be applied to compartments and their interplay with TADs.

### **Chromosome territories**

At the largest scale, interphase chromosomes occupy distinct spatial territories in the nucleus in a preferential manner [219]. This preferential chromosome localisation can vary among cell types [220], but has been shown to be conserved among higher primate species [221]. In general, chromosomes are spatially organised in the nucleus such that gene-rich chromosomal regions tend to be located in the inner part of the nucleus, while regions that are relatively poorer in genes tend to locate in the nuclear periphery. For example, the lamina associated domains (LADs) of chromatin are characterised by a relatively low gene density as well as decrease transcriptional activity (reviewed in [146]). However, it has been demonstrated that there is some extent of intermingling between chromosome territories, which is associated with repositioning of genes outside of their territories and transcriptional activation [222].

As in the previous layers of three-dimensional chromatin organisation, chromosome territories also indicate that the 3D structure of the genome is closely associated with its function

in gene expression regulation.

### **1.3 Sequencing technologies and functional genomics assays**

Over the last few decades, technological advancements in molecular biology research have generated great potential in performing large scale studies on the genome and its functions. Specifically, the advent of numerous high throughput technologies has improved the accuracy and enabled the automation of experimental procedures, allowing an unimaginable extent of experiment reproducibility, generation of biological data and high powered analyses.

An outstanding milestone in the establishment of advanced molecular technologies has been the development of the Sanger sequencing method in the late 70's, by Frederic Sanger and his colleagues [223, 224]. This method was the prevalent sequencing technology for more than two decades. Implemented by the so-called “first generation” DNA sequencing machines, it enabled the sequencing of the first whole genomes, including the human genome within the massive collaborative effort of the Human Genome Project. This motivated the development of more advanced, high-throughput Next Generation Sequencing (NGS) platforms that are highly scalable and operate with higher accuracy and at a lower cost; these include Roche 454, SOLiD, Illumina (former Solexa), Ion Torrent. In recent years, a new generation of high-throughput technologies that perform long-read sequencing has been developed and is being used in “third generation” sequencing platforms. These include the SMRT and Oxford Nanopore sequencing technologies.

HTS technologies have facilitated not just sequencing of whole genomes, but also their functional characterisation, as they have provided the basis for the development of further technologies and applications, including whole exome sequencing, transcriptome sequencing, genome and genome-protein interactions identification (ChIP-seq, ChIA-PET, Hi-C), supporting also epigenome profiling (ChIP-seq, ATAC-seq, bisulfite sequencing). These technological advancements have also spurred the development of new computational tools and approaches for the analyses of the emerging large-scale data.

### 1.3.1 NGS technologies

In the last decade NGS sequencing methods have introduced massive parallel sequencing of vast amounts of short reads (i.e. sequence fragments) revolutionising genome analysis. The underlying principle is a series of repeated cycles of DNA molecule synthesis, where fluorescently labeled nucleotides (dNTPs) get incorporated into a DNA template strand. As the dNTPs are being incorporated, they are identified by fluorophore excitation.

The vast majority of available sequencing data worldwide have been generated using the Illumina sequencing platforms.

#### **Illumina Sequencing**

Illumina sequencing is performed in three main steps. These are: library preparation, cluster generation and sequencing by synthesis. During the library preparation, the DNA is purified and randomly sheared into short fragments. The resulting DNA fragments are ligated with adapters at their 5' and 3' ends. The adapter ligated sequence fragments are then amplified with polymerase chain reaction (PCR). Next, the prepared library is passed on the cluster amplification step by getting loaded onto a chip with thousands of unlabeled short synthetic DNA oligonucleotides. The DNA fragments hybridise with the oligos that are complementary to their ligated adapters. Subsequently, each cluster is amplified into distinct clonal clusters, meaning that a high number of copies of each read is generated. The reads are then separated into single strands and are ready to be sequenced. Sequencing by synthesis is performed by adding fluorescently labeled reversible terminators, primers and DNA polymerase and then detecting the emitted fluorescence after laser excitation.

The generated Illumina sequencing data can then be analyzed and used in multiple approaches. The first step of the analysis is usually the alignment onto the reference genome of the organism they originate from. Depending on the type of the study, they can be used for various purposes.

#### **NGS data processing**

##### **Quality control and filtering**

NGS technologies generate high throughput sequencing reads with a high accuracy. However, sequence artefacts such as base calling errors, mainly at the 3' ends of reads and at a

rate 0.5-2%, as well as adapter contamination can be common. Evaluation of the extent of such errors, and thus the quality of reads can be done via performing quality controls of the read dataset. Poor quality per base, low sequence content diversity, high duplication levels, and specific sequence overrepresentation are usually associated with high base calling error rates, PCR over-amplification, or adapter contamination. A very commonly used software package for quality control of high-throughput sequencing data is FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). FastQC calculates various quality metrics for the sequenced reads, including per-base quality scores, sequence content, levels of read duplication which can be a result of PCR over-amplification, or sequence over-representation which can be a result of adapter contamination. Based on the evaluation of these metrics, read quality can be improved by using specialised software packages to eliminate sequences with poorer quality. A versatile software toolkit is the fastx-toolkit, which, among other, includes fastx-trimmer that can be used for trimming low quality called bases at the 3' end of the reads, and fastx-clipper that can be used to trim off adapter sequences from the reads. PCR and optical read duplicates can be identified or removed using the MarkDuplicates tool from the picard tool suite [225].

### **Read mapping**

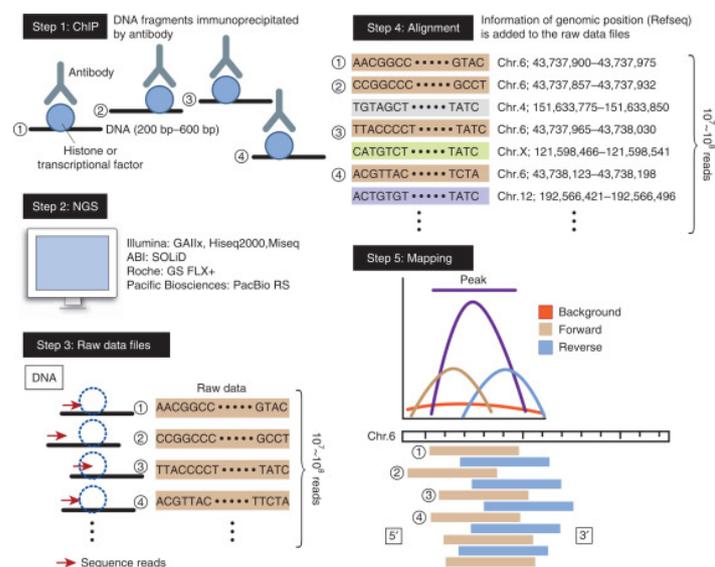
Once it is ensured that the sequenced reads are of good quality, they can be mapped against the reference genome of the corresponding organism. There are a number of available algorithms for read alignment on genome assemblies, which can differ in their indexing strategies, memory efficiency, mapping sensitivity and speed. Due to their computational efficiency, BWA [226] and Bowtie2 [227], which are based on the Burrows-Wheeler transform (BWT [228]) and the Ferragina Manzini index (FM index [229]), are some of the most commonly used read mapping algorithms.

## **1.3.2 Protein-DNA interaction assays**

### **ChIP-seq: experimental protocol and data analysis**

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a useful method for profiling epigenomic features across a genome, as it enables identification and analysis of protein-DNA interactions. It can be used for genome wide mapping of the binding sites of a transcription factor, or more generally the genomic loci occupied by proteins, such as polymerases or histones with specific modifications.

The first step of the experimental method includes the use of formaldehyde for tissue fixation and cross-linking of the protein-chromatin interactions, followed by homogenisation of the fixed tissue and cell lysis. In this way, the chromatin is obtained and then fragmented into small segments, usually 300-500bp long. The next step is chromatin immunoprecipitation (ChIP) (Fig. 1.11). This includes the use of an antibody during incubation and centrifugation. It is important that the antibody is specific for the protein of interest, so that the DNA fragments that are bound by the target protein can be isolated. These immunoprecipitated DNA fragments are then purified by reversing the cross-link between DNA and proteins using heating. They are also size-selected and then used to prepare the sequencing library. Specifically, oligonucleotide adaptors are added to the ends of the ChIP-ed DNA fragments, which allows their massive parallel sequencing, commonly by Illumina Sequencing (Fig. 1.11).



**Fig. 1.11 ChIP-seq experiment and data analysis.** Step1: Chromatin is immunoprecipitated using an antibody specific for the TF or the modified histone, followed by isolation of DNA fragments that were bound by the protein. Step 2: High throughput sequencing of the DNA fragments. Step 3: Raw sequencing data is obtained from the sequencer. Step 4: Reads are aligned against the reference genome assembly. Step 5: Peaks of enriched read mapping are identified with using a peak calling algorithm. Figure adapted from Mimura et al. [230]

The sequenced DNA fragments correspond to all the genomic locations the protein of interest was bound to. However, ChIP-seq output is affected by sequence content, chromatin state and DNA sonication bias, as GC-rich regions and open chromatin regions typically yield higher read coverage in sequencing, while heterochromatic regions, which are more resistant to shearing, can be under-represented in the sequencing output. To account for the effect

of these factors in a ChIP-seq experiment, it is important to use a control, also known as the input, sample. The same experimental protocol used for the ChIP-ed samples is applied to the input, except for the step of immunoprecipitation. The resulting dataset is used as a reference against which the ChIP-enriched regions are compared in the downstream analysis to control for the relevant biases. Additional factors that might affect the efficiency of the experiment are: the antibody specificity and the sequencing depth. For identifying genomic interactions with proteins that occupy more or longer, on average, genomic intervals, a higher sequencing depth is required.

The sequenced fragments from the ChIP-seq experiment correspond to genomic loci that interact with the protein of interest. The first step in the processing of these sequenced reads is to perform read quality controls and filtering, as described in section 1.3.1, and read mapping using one of the available alignment algorithms, as described in section 1.3.1. Then, the read alignment output can be passed on to a peak calling algorithm, so as to identify peaks of ChIP-seq read enrichment, which represent genomic sites where the protein of interest localises. A very popular peak caller is MACS2, which uses a window-based approach to determine significantly enriched regions and randomly shuffles ChIP-ed and Input reads to calculate an empirical false discovery rate (FDR). The identified peak regions with read enrichment can be visualised on genome browsers, such as Ensembl and IGV. Moreover, following the identification of ChIP-seq peaks, the underlying genomic intervals can then be scanned to identify short sequence motifs that bind TFs. MEME suite is a commonly used software that includes a range of tools for motif identification, for instance MEME for *de novo* motif discovery, or FIMO for identification of specified motif matches in the scanned sequences. A number of databases with identified sequence motifs can also be used to complement motif search and analysis; one of them is JASPAR [168]. Finally, further functional analyses of the identified ChIP-seq peaks can be performed to identify associations of the ChIP-marked (regulatory) regions with genes and gene ontology terms.

### **Other assays for protein-DNA interactions**

An alternative method, widely used before the establishment of the ChIP-seq protocol, is the ChIP-on-ChIP. That includes the ChIP wet lab protocol just as in ChIP-seq, but is followed by hybridisation of the purified DNA fragments with probe oligos on a DNA microarray, instead of high throughput sequencing. Two more recently developed methods are the “cleavage under targets and release using nuclease” followed by sequencing (CUT&RUN-sequencing [231]) and “cleavage under target and tagmentation” followed by sequencing (CUT&Tag-

sequencing [232]). These are based on cleavage of the antibody-targeted genomic sites by using micrococcal nucleases or a nuclease-transposase fusion, respectively, and they both require less material and offer a higher signal-to-noise ratio with shallower sequencing.

### 1.3.3 Chromatin conformation assays

Querying of the three dimensional genome structure has been enabled through the development of technologies based on Chromosome-Conformation-Capture (3C) [233, 234]. The main underlying principle of these techniques is the use of proximity ligation to identify physical interactions between genomic regions. Some of the 3C-based protocols, specifically chromatin interaction analysis with paired-end tag sequencing (ChIA-PET) and ChIP-loop (also referred to as 6C) include an immunoprecipitation step to detect genomic interactions that are mediated by specific proteins (reviewed in [139]). In contrast, the 3C, 4C, 5C and Hi-C protocols determine chromatin interactions in a protein independent manner (reviewed in [127]). Following the 3C-based technologies, genome architecture mapping (GAM) has also been developed as a novel method to identify genome wide contacts based on cryosectioning and DNA sequencing from thin nuclear sections, rather than using ligation [235].

The Hi-C protocol is very commonly used and allows for identification of chromatin interactions across the whole genome, thus supporting the construction of genome-wide contact frequency maps. In brief, Hi-C includes crosslinking of cells to form covalent links between interacting genome fragments, DNA digestion by a restriction enzyme, ligation of the interacting genomic fragments with biotin, DNA shearing and quantification of the ligated products via high-throughput sequencing and subsequent data analyses. These analyses estimate the frequency of genomic contacts [131]. Hi-C maps have revealed the organisation of the genome into compartments A and B [131, 132], as well as in TADs [133, 135, 182]. Besides characterisation of global chromatin conformation features, Hi-C can be used for identifying specific looping interactions that may also take place between genomic sequences and that may have regulatory interactions. Popular software packages are HiCUP [236] for pre-processing and HOMER [237] for quality control and analysis of Hi-C data. Hi-C data analysis aims to quantify contact frequency maps between interacting genomic regions. These can be visualised as two dimensional matrices where the row and column indices represent genomic bins and the values in the matrix cells indicate the frequency of contacts between the corresponding genomic bins. Therefore, the size of the genomic bins and the sequencing depth are factors that affect the resolution of Hi-C contact maps.

### 1.3.4 Gene expression assays

#### RNA sequencing (RNA-seq)

RNA sequencing has been established as a *de facto* approach for studying whole transcriptomes and quantifying gene expression. The majority of RNA-seq data over the last decade were generated with Illumina short-read sequencing platforms. The typical RNA-seq protocol includes extraction of RNA, RNA fragmentation, reverse transcription to cDNA, adaptor ligation, PCR amplification and high throughput sequencing, which results in cDNA reads usually shorter than 200bp (reviewed by Stark [238]). The standard protocol often includes a selection step for transcripts that contain poly-A tails, thus representing mostly protein-coding genes and being largely depleted of small non-coding RNAs or enhancer RNAs. If the poly-A selection step is removed, the library is usually also depleted of rRNA [239]. A number of other modifications can also be applied to the standard protocol, depending on the aspect of the transcriptome each study focuses on. For example cap analysis of gene expression (CAGE) can be used to enrich for 5' cDNA fragments and support identification of promoters, TSSs and transcribed enhancers [240, 241]. Another innovation is the addition of unique molecular identifiers (UMIs) [242, 243] to the cDNA fragments before amplification in order to, subsequently, identify and remove PCR duplicates accurately.

The generated sequencing data can be passed on to downstream analysis to be assigned to gene transcripts, quantify expression and/or compare between sample cohorts that correspond to different biological conditions, i.e. differential gene expression analysis. The analysis starts with quality control and filtering in the same way as for DNA sequencing data. Then reads are mapped on the reference genome using read alignment algorithms such as TopHat [244] or STAR [245] that can produce spliced alignments, and subsequently the mapped reads are assigned to genomic features and quantified with tools such as featureCounts [246] or Htseq [247]. Alternatively, the read mapping and quantification steps can be performed at once by more recently developed methods, such as kallisto [248] and Salmon [249], which do not use a reference genome, but rather a transcriptome assembly. After read counts are computed, differential expression analysis is performed, including normalisation across samples and statistical analyses of significant differences between samples. Popular software packages for that are DESeq2 [250] and edgeR [251].

## 1.4 Aims of the thesis

Despite years of research, numerous aspects of gene regulatory mechanisms and their contributions to shaping phenotypic characteristics are not well understood. This thesis aims to gain insights into mechanisms of transcriptional regulation, focusing particularly on the binding activity of transcription factors (TFs). The individual projects and analyses make use of *in vivo* systems that leverage a pool of regulatory variation to study the activity and roles of TF binding within a wider functional context. The *in vivo* systems used include closely related mouse species where molecular variation is manifested as changes in the sequence context and/or in the binding activity of the TFs. Also, these changes may be naturally fixed by evolution in the different species (natural variation), or induced by various factors, for example mutagenesis as a result of carcinogen exposures. The functional implications of the underlying molecular variation in TF binding is also evaluated as part of broader functional contexts, such as higher-order genome organisation, or the neoplastic transformation of a tumour cell in cancer development.

In the following chapter, I present my study on the functional contribution of CTCF binding to establishing higher-order genome structures. There, I leverage natural genetic variation among five mouse species to study the interplay between the CTCF binding evolutionary dynamics and the demarcation of boundaries of topologically associating domains (TADs). My main findings reveal the association of TAD boundaries with dynamically evolving clusters of neighbouring CTCF sites, which contribute to the resilience of TAD structures and the shaping of regulatory landscapes.

In the third and fourth chapters, I present work carried out as part of a bigger collaborative project of the Liver Cancer Evolution (LCE) consortium. The project makes use of a model of chemically induced liver carcinogenesis in different mouse species, to study mutagenesis mechanisms and the relative contributions of the genome and epigenome to liver cancer development.

In the third chapter, I present my work on inspecting bulk expression data of monoclonal liver tumours. There, I characterise the expression profile of the liver tumour cell-of-origin, the hepatocyte, by leveraging developmental single-cell expression datasets from mouse liver. I also identify gene dysregulation in the clonal tumour cells that underlies the shift from their hepatocyte specific to a more dedifferentiated expression phenotype along neoplastic transformation.

In the fourth chapter, I further explore associations of gene dysregulation and the hepatocyte phenotypic shift along liver cancer development with mutagenesis at liver TF binding sites in the tumour cell-of-origin. My results disentangle different patterns of mutation accumulation among the distinct categories of TF binding sites based on their functional characteristics, such as combinatorial binding or association with *cis*-regulatory elements, and show that TF binding sites with substantial regulatory roles are generally less mutated. However, I also reveal subsets of hyper-mutated TF binding sites in liver tumours, which are enriched for liver-specific biological processes, cellular stress response and abnormal liver phenotypic features, and they are associated with dysregulated genes that have hepatocyte-specific functions.

## Chapter 2

# Evolutionary dynamics of CTCF binding and higher-order genome structures

This chapter describes my study on evolutionary dynamics of CTCF binding among mouse species and its role in establishment and maintenance of topologically associating domains -a layer of the higher order chromatin organisation. The data used in this study have been mostly generated in the Odom lab, by Sarah Aitken, Christine Feig and Klara Stefflova, while some of the datasets were retrieved from previously published studies. Except as noted, all analyses described here were carried out by myself. The manuscript was also written by myself, edited collectively and agreed upon by all co-authors. The results of this study have been published in the following paper:

**Kentepozidou, E.**, Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M., Flicek, P. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol* **21**, 5 (2020). [252]

Also, part of results from this study, with respect to the effect of *Ctcf* hemizyosity on CTCF binding and TAD structures, have complemented analyses in a project led by Sarah J. Aitken and Ximena Ibarra-Soria. These were included in the following paper:

Aitken, S.J., Ibarra-Soria, X., **Kentepozidou, E.**, Flicek, P., Feig, C., Marioni, J.C., Odom, D.T. CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biol* **19**, 106 (2018). [253]

## 2.1 Introduction

At a sub-megabase scale the genome is organised into topologically associating domains (TADs), which define regulatory landscapes via their self-interaction patterns [254, 135, 146, 187]. Despite their conservation and presumed functional importance, the mechanisms underlying their stability and evolution remain unclear. Accumulating evidence supports a model where the CCCTC-binding factor (CTCF), colocalised with the cohesin protein complex, plays a causal role in the formation and maintenance of TADs [210, 255, 256]. CTCF is a ubiquitously expressed protein with a deeply conserved DNA-binding domain [157, 158]. It is responsible for diverse regulatory functions including transcriptional activation and repression as well as promoter and enhancer insulation. Its diverse functions are based on its role in promoting interactions between distant genomic elements by mediating chromatin loop formation [162, 159, 170]. A loop extrusion mechanism of TAD formation has been proposed wherein the cohesin protein complex slides along chromatin forming a growing loop until it meets two CTCF molecules bound with convergent orientation. This architecture then prevents cohesin from sliding further, demarcating the TAD boundaries [193, 192]. This model explains why these boundaries usually harbor CTCF binding sites. Nevertheless, there are ubiquitous CTCF-bound regions with diverse functions throughout the genome, while only a small fraction of them occur at TAD boundaries [133]. In addition, a small number of TADs are not flanked by CTCF binding sites [135, 207, 182, 206]. This has made it challenging to delineate the precise role of CTCF binding in establishing and stabilizing TAD structures.

Several recent perturbational studies experimentally provide some insights into the role of CTCF in determining local and genome-wide three-dimensional chromatin organisation. Local disruption of CTCF binding can lead to abrogation of TAD insulation and formation of ectopic *cis*-regulatory interactions between neighbouring TADs [257, 151, 135, 170, 146], although TAD structures have been reported to remain intact [258, 135, 192]. Local TAD disruptions may also lead to disease [259, 188–190]. Upon acute, transient genome-wide depletion of CTCF there is marked disruption to chromatin loop and TAD structures [196, 199, 260], but the degree of TAD destabilisation remains controversial. The impact of this CTCF-mediated insulation on gene expression remains poorly understood. Indeed, experimental approaches that disrupt CTCF binding remain limited by the fundamental roles of CTCF in development and cell viability.

The binding profiles of CTCF in present-day mammalian genomes are shaped by repeated waves of transposable element insertions carrying CTCF binding sequences across mammalian genomes [176, 164, 261, 262]. Mammalian-conserved sites resulted from ancestral expansions, while recent expansions have established lineage-specific binding patterns. For example, the B2 family of short interspersed nuclear elements (SINEs) active in the mouse-rat ancestor shaped the CTCF binding profile of all Muridae species and specific members of the B2 family remain active in a lineage-specific manner [176, 164, 261]. The human and macaque genomes also share a large fraction of

CTCF-associated transposable elements despite the absence of recent large-scale insertional activity [262]. Moreover, mammals representing distinct vertebrate orders share conserved CTCF binding sites at their TAD borders [133, 134, 184].

The evolutionary history of CTCF binding facilitates a complementary approach to understanding the role of CTCF in TAD stability. Specifically, we can leverage the natural genetic variation between species as opposed to experimental approaches using targeted or systemic CTCF binding disruption. We can thus investigate the consequences of CTCF binding changes stably fixed by evolution as a version of an *in vivo* mutagenesis screen [263]. A unique and important advantage of this approach is that the physiological cellular system can be assumed to be in stable and homeostatic equilibrium [264]. CTCF is ideally suited to such an evolutionary approach because in each species the CTCF binding profile is composed of substantial numbers of both deeply conserved and evolutionarily recent sites [164, 261].

In this study, I used CTCF ChIP-seq data from five mouse strains and species, which have similar genomes and transcriptional profiles, to give insight into the establishment and stability of TADs. My analysis of the genome-wide CTCF binding leverages natural genetic variation between species to assess the evolutionary dynamics of TAD boundary demarcation. I also investigated the impact of local losses of CTCF binding on gene expression in adjacent TADs, as well as the effect of CTCF hemizygoty on TAD structures. The results revealed that TAD borders are characterised by clusters of both evolutionarily old and young CTCF binding sites. In addition, CTCF bound regions at TAD borders, regardless of age, exhibit increased levels of sequence constraint compared with CTCF binding sites not associated with TAD boundaries. Such clusters are consistent with a model of TAD boundaries in a dynamic equilibrium between selective constraints and active evolutionary processes. As a result, they apparently retain a redundancy of CTCF binding sites that give resilience to the three-dimensional genome structure.

## 2.2 Results

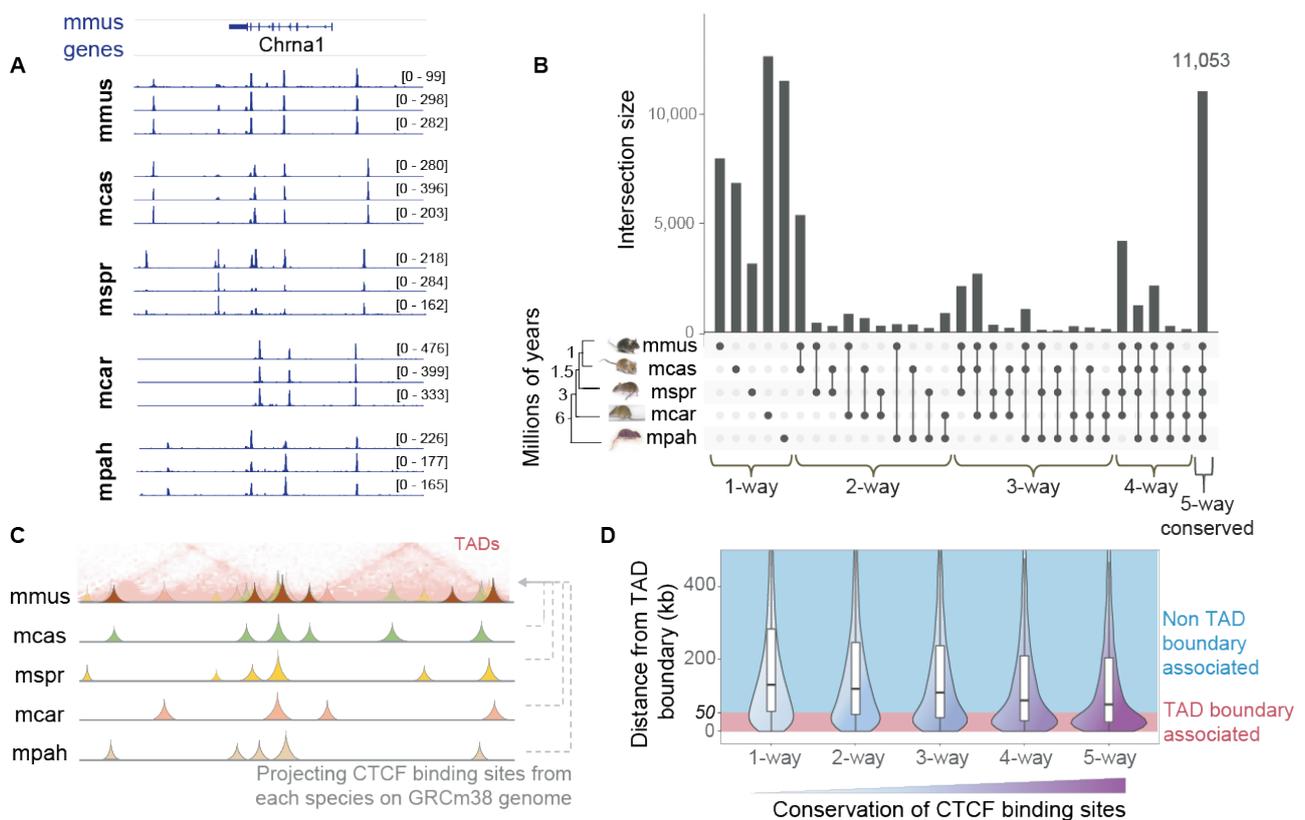
### 2.2.1 Conservation of CTCF binding sites and association with TAD borders

To investigate the evolutionary dynamics of CTCF binding, I first determined CTCF binding profiles in closely related species. In particular, I used ChIP-seq data to identify CTCF enriched regions in the livers of five mouse species<sup>1</sup>: *Mus musculus domesticus* (C57BL/6J, hereafter referred to as BL6), *M.*

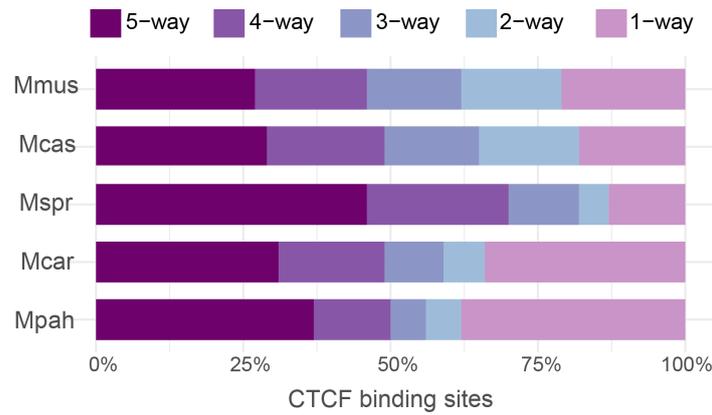
---

<sup>1</sup>Although BL6 (*Mus musculus domesticus*) and CAST (*Mus musculus castaneus*) are subspecies of the same species, *Mus musculus*, hereafter they will be conventionally referred to as *species*.

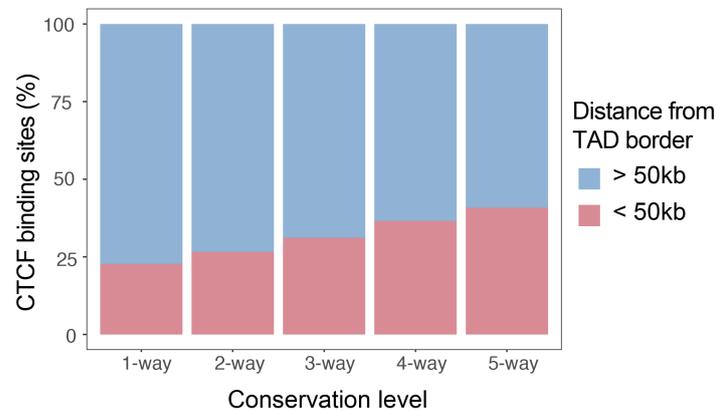
*musculus castaneus* (CAST), *M. spretus* (SPRET), *M. caroli* (CAROLI), and *M. pahari* (PAHARI) (Fig. 2.1A, Fig. A.1 - Appendix A). Specifically, for CAST and SPRET I used ChIP-seq data generated in the Odom lab, while for CAROLI and PAHARI I retrieved previously published ChIP-seq data [261]. I processed the raw data and determined CTCF peaks in three biological replicates of each species (see Methods). Peaks present in at least two of the replicates were defined as reproducible and were used for downstream analyses (Table A.1 - Appendix A, Fig. A.1 - Appendix A).



**Fig. 2.1 *Mus*-conserved CTCF binding sites commonly occur at TAD borders.** (A) CTCF ChIP-seq enrichment tracks around the *Chrna1* locus in BL6 and in orthologous regions of the other *Mus* species. Raw data from three biological replicates are shown for each species. (B) Conservation of CTCF binding sites across the five studied *Mus* species. Conservation levels at the bottom of the panel indicate the number of species CTCF sites are shared in (phylogenetic distances are from Thybert et al. [261]). (C) Graphical representation of using orthologous alignments of the CTCF sites identified in each *Mus* species to project them on the genome of BL6 (GRCm38) where TADs are available. (D) Distances of CTCF sites with different conservation levels to their closest TAD boundary. CTCF sites with a distance  $\leq 50\text{kb}$  are considered TAD-boundary associated, while sites with a distance  $> 50\text{kb}$  are referred to as non-TAD-boundary associated.



**Fig. 2.2 Fractions of CTCF binding sites of different conservation levels in each of the studied *Mus* species.**

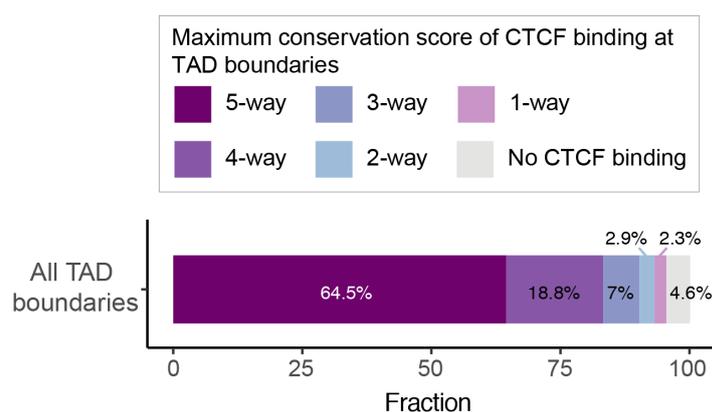


**Fig. 2.3 Fractions of all *Mus* CTCF sites of each conservation level that are associated ( $d \leq 50\text{kb}$ ) or not associated ( $d > 50\text{kb}$ ) with TAD boundaries.**

To characterise the conservation of the identified CTCF binding sites among the mouse species, I lifted over the binding site regions from one species to another (see Methods). I identified conserved sites between species as orthologous regions that were ChIP-enriched in each of the compared species. That way, I determined the conservation level of each CTCF binding site based on whether it is shared by all species (*Mus*-conserved or 5-way), fewer than five species (4-way, 3-way, 2-way) or are species-specific (1-way) (Fig. 2.1B). *Mus*-conserved and species-specific CTCF binding sites were the most numerous among the different conservation level groups (Fig. 2.1B, 2.2). A total of ~11,000 CTCF binding sites were found to be *Mus*-conserved. To put this number in context, it corresponded to more than a quarter (~27%) of the total number of CTCF sites identified in BL6 (Fig. 2.2). This is consistent with previous studies that reported high CTCF binding conservation across eutherian mammals, especially compared with other transcription factors such as HNF4A and CEBPA [175, 77, 164].

To evaluate the CTCF binding evolutionary dynamics with respect to TAD borders, I firstly intersected the CTCF binding profiles with TAD borders identified from published Hi-C in BL6 liver (Fig. A.2 - Appendix A) [184]. Although I used Hi-C data for only one of the five species, it has been shown that TADs are largely conserved across species and cell types [133, 184, 209]. For these closely related mouse species with very similar genomes, transcriptomes and CTCF binding patterns, I expected that this assumption is valid to a great extent. Therefore, I projected the CTCF sites identified in each of the five *Mus* species onto the BL6 genome (GRCm38/mm10) (Fig. 2.1C). Having grouped all the CTCF sites by conservation level, I measured the distance of each site in these groups to its closest TAD boundary. Based on this distance and the resolution of the TAD map used, I further sub-categorised CTCF sites as TAD-boundary-associated ( $d \leq 50\text{kb}$ ) and non-TAD-boundary-associated ( $d > 50\text{kb}$ ). It is noted that, in this way,  $\sim 13\%$  of the genome was classified as TAD-boundary-associated. Although CTCF sites of all conservation levels were found to occur at the boundaries of TADs, more highly conserved CTCF sites were, on average, located closer to TAD boundaries (Fig. 2.1D). Previous studies have highlighted the overlap of TAD boundaries with sites conserved in more distantly related species, such as human and mouse [134] or mouse and dog [184]. Yet, an association between CTCF sites of progressively increasing conservation level and proximity to TAD boundaries is evident even when zooming in in evolutionary time, examining closely related species. This is exemplified if I consider that 41% of the *Mus*-conserved CTCF sites, as compared to 23% of species-specific sites, were found to lie within 50kb of TAD boundaries (Fig. 2.3).

Shifting the perspective from CTCF bound regions to TAD boundaries, the majority of TAD borders were found to overlap with highly-conserved CTCF binding sites. Nevertheless, a small fraction of the boundaries did not harbour any *Mus*-conserved CTCF binding events. In particular, twelve percent of them contained CTCF sites that were conserved only in one, two or three out of the five studied



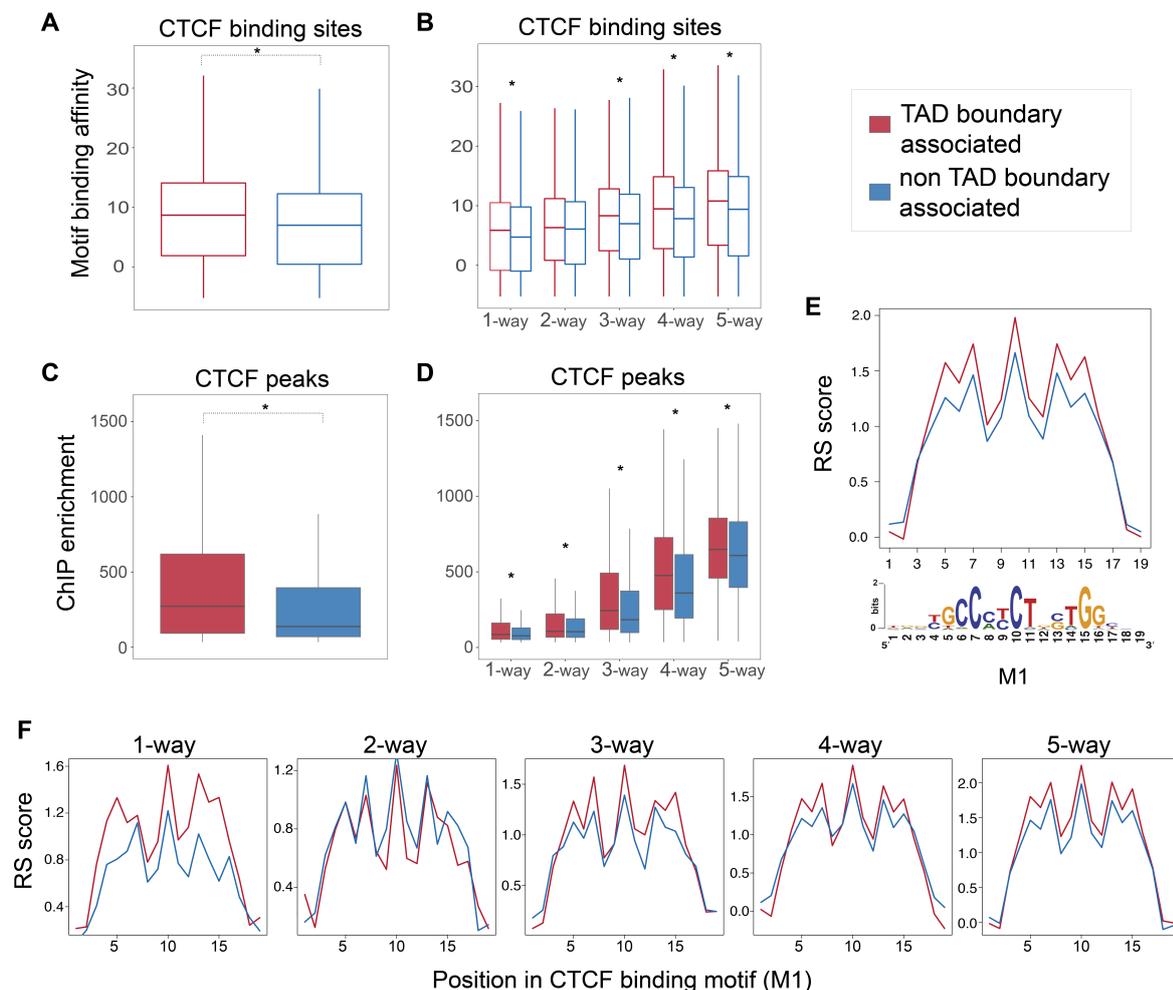
**Fig. 2.4 Fractions of TAD boundaries with CTCF sites of different conservation levels.** Most TAD boundaries (64%) harbour at least one *Mus*-conserved (5-way) CTCF site. Lower percentages of TAD borders do not contain any *Mus*-conserved CTCF site but overlap with less conserved sites or do not bind CTCF at all.

mouse species (Fig. 2.4). Furthermore, nearly 5% of TAD boundaries apparently do not overlap with any CTCF bound site at all (Fig. 2.4). Other studies have also identified TAD boundaries deprived of CTCF bound sites and have proposed that even though the connection between CTCF binding and TAD boundaries was consistently observed, it may not be a strictly necessary feature for demarcation of TAD boundaries. It has also been suggested that a few TADs, whose borders are not demarcated by CTCF sites, may be passively shaped between other CTCF-anchored TADs [128, 134].

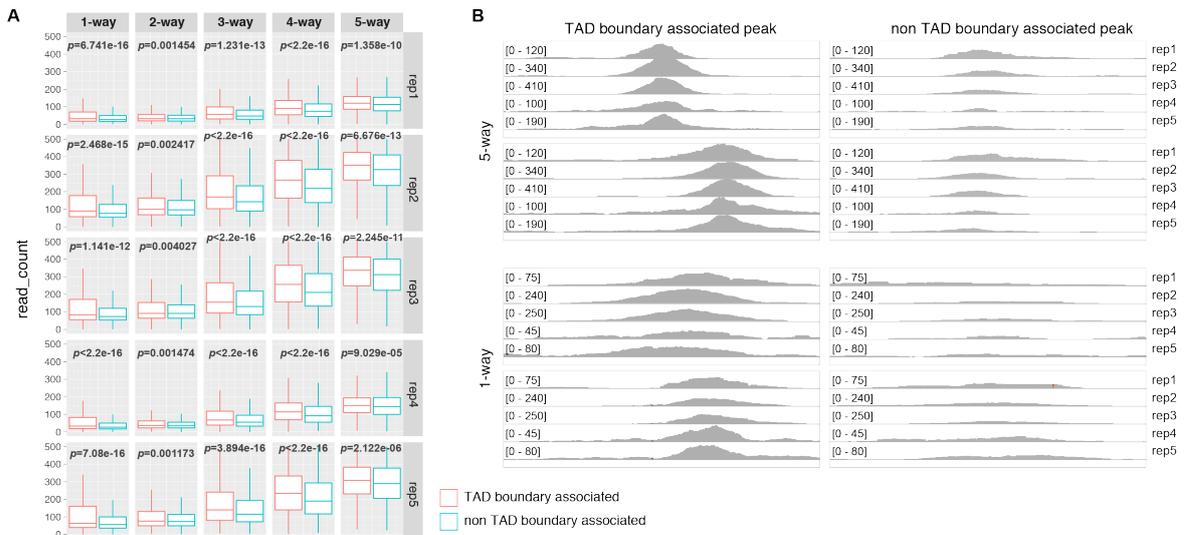
To summarise, CTCF binding sites were found to be conserved to a large extent across the five studied *Mus* species. In addition, 41% of the *Mus*-conserved CTCF binding sites were associated with a TAD boundary, while the vast majority (>95%) of all TAD boundaries were shown to harbour at least one CTCF binding site.

## 2.2.2 Evolutionary constraints at TAD-boundary-associated CTCF binding sites

CTCF is known to bind to a 33/34-bp region of the genome consisting of a primary sequence motif (M1) and a shorter secondary motif (M2). I found that overall binding affinity, as computationally predicted from the motif sequence, was significantly greater for boundary-associated CTCF sites compared to non-boundary-associated sites (Mann-Whitney U test,  $p < 2.2e16$ ) (Fig. 2.5a). I went on to investigate whether other characteristics of CTCF binding sites -beyond the cross-species conservation of binding activity- are coupled with TAD boundary association. To this aim, I first assessed the relationship among CTCF conservation level, TAD boundary association, and CTCF motif strength. It is noted that CTCF preferentially binds to a 19 base pair-long motif sequence (primary CTCF binding motif, M1), or to an extended version of it that also includes a secondary, shorter motif (M2) and spans 33/34 base pairs in total [164]. I scanned our ChIP-seq peak regions, identified occurrences of the CTCF binding motif, and calculated the binding affinity of each of these motif sequences (see Methods). Binding affinity was, overall, significantly higher for boundary-associated CTCF sites compared to non-boundary-associated sites (Mann-Whitney U test,  $p < 2.2e-16$ ) (Fig. 2.5A). However, according to the previous observation, TAD boundaries are frequently associated with *Mus*-conserved sites, while it is also known that species-conserved sites tend to have motifs of higher sequence constraints and thus higher binding affinity for CTCF than less conserved sites [78]. Based on these, could the observed increase in CTCF binding affinity at TAD-boundary-associated sites simply reflect a bias of many *Mus*-conserved CTCF sites occurring at TAD boundaries? To address this, I grouped the sites based on their conservation level, and repeated the comparison of TAD-boundary-associated with non-TAD-boundary-associated sites within each conservation level group. As expected, motif binding affinity increased with the CTCF binding site conservation level. Yet, TAD-boundary-associated CTCF binding sites of any given conservation level consistently had greater binding affinity than non-boundary-associated sites (Mann-Whitney U tests between TAD-



**Fig. 2.5 CTCF binding sites at TAD boundaries are subjected to stronger evolutionary constraints.** (A) CTCF-bound sites at TAD boundaries contain motifs with higher binding affinity for CTCF than non-TAD-boundary-associated sites (Mann-Whitney U test:  $p$ -value  $< 2.2e-10$ ). (B) Although the binding affinity of CTCF sites is generally proportional to the conservation level of the site (how many species it is shared by), CTCF sites at TAD boundaries have stronger binding affinity than non-TAD-boundary-associated sites, independent of their conservation level (Mann-Whitney U tests between TAD-boundary-associated and non-TAD-boundary-associated sites:  $p_{1\text{-way}} = 0.001$ ,  $p_{2\text{-way}} = 0.06$ ,  $p_{3\text{-way}} = 6.1e-07$ ,  $p_{4\text{-way}} = 5.2e-13$ ,  $p_{5\text{-way}} = 3.9e-11$ ). (C) TAD-boundary-associated CTCF peaks display higher ChIP enrichment scores, as calculated by MACS, than non-TAD-boundary-associated peaks (Mann-Whitney U test:  $p$ -value  $< 2.2e-10$ ). (D) TAD-boundary-associated CTCF peaks, at every conservation level, display stronger ChIP enrichment than non-TAD-boundary-associated peaks (Mann-Whitney U tests:  $p_{1\text{-way}} < 2.2e-16$ ,  $p_{2\text{-way}} = 0.002316$ ,  $p_{3\text{-way}} < 2.2e-16$ ,  $p_{4\text{-way}} < 2.2e-16$ ,  $p_{5\text{-way}} = 2.047e-12$ ). (E) The most information-rich bases of the primary CTCF M1 motif at TAD boundaries display higher rejected substitution (RS) scores compared to non-TAD-boundary-associated motifs. The bottom panel shows the position weight matrix of the CTCF M1 motif from Schmidt et al. [164]. (F) The observation in (E) is independent of the conservation level of the CTCF sites, as shown for subsets of sites at each conservation level.



**Fig. 2.6 Higher read coverage at TAD boundary-associated CTCF peaks compared to non-TAD-boundary-associated peaks.** (A) Mapped read counts for TAD boundary-associated versus non-TAD-boundary-associated peaks in five biological replicates of BL6 liver. The two groups were compared using Mann Whitney U tests. (B) ChIP-seq read coverage for example loci of 5-way conserved and species-specific peaks at TAD boundaries (left) compared, respectively, to 5-way and species-specific (1-way) peaks at non boundary regions (right). The read coverage of TAD boundary-associated peaks is higher than at the non-TAD-boundary-associated peaks, independently of whether the peaks are conserved or species-specific. The observations are consistent among the replicates.

boundary-associated and non-TAD-boundary-associated sites:  $p_{5\text{-way}} = 3.9e-11$ ,  $p_{4\text{-way}} = 5.2e-13$ ,  $p_{3\text{-way}} = 6.1e-07$ ,  $p_{2\text{-way}} = 0.06$ ,  $p_{1\text{-way}} = 0.001$ ) (Fig. 2.5B). In addition to that, CTCF binding sites at TAD borders were shown to have higher ChIP enrichment than non-TAD-boundary-associated CTCF sites, independently of conservation level (Fig. 2.5C, 2.5D)<sup>2</sup>. I further validated this, by using additional ChIP-seq libraries from BL6 (see Methods) to increase the power. Specifically, I mapped and quantified ChIP-seq reads at TAD-boundary-associated and at non-TAD-boundary-associated sites of each conservation level group. Sites at TAD boundary regions had higher ChIP-seq read coverage (Fig. 2.6). These observations were consistent with the stronger predicted affinity of the corresponding CTCF motif sequences. Overall, these results give new insight into the observation that

<sup>2</sup>The lower ChIP enrichment of 1-way and 2-way conserved sites compared to more deeply conserved ones, as shown in Fig. 2.5D, might raise a question about how likely it is that the 1-way and 2-way sites are functional. It is possible that a (small) fraction of them may represent occasionally and transiently bound sites, or even correspond to false positive binding sites resulting from ChIP-seq artefacts. However, it is likely that most of them do have some functionality, as: a) they have been identified against input libraries, and independently in at least two biological replicates, which is in line with the ENCODE Consortium standards for ChIP-seq based binding site discovery [265]. Secondly, it has been shown that evolutionarily young CTCF binding sites—even subspecies-specific ones—generally show functional signatures similar to more deeply conserved binding sites [266].

mammalian-conserved CTCF sites have higher motif affinity than species-specific sites [164, 184]. Importantly, for all CTCF binding sites, including species-specific ones, proximity to a TAD boundary was associated with an increase in binding affinity (Fig. 2.5B, 2.5D). This implies that CTCF binding motifs at TAD boundaries may be under stronger selective constraint than the motif sequences of non-TAD-boundary-associated CTCF peaks.

To further evaluate this hypothesis, I explored evolutionary sequence constraint of the CTCF binding motif itself. For this purpose, I measured rejected substitution (RS) scores at each position of every 19 base-long primary CTCF binding motif (M1) and compared the score between (a) TAD-boundary-associated and (b) non-TAD-boundary-associated regions (Fig. 2.5E, 2.5F). RS score is a measure of sequence constraint. It reflects the number of base substitutions that were rejected at a specific genomic position as a result of purifying selection, compared to the number of substitutions that would have occurred if the sequence was evolving under neutral selection [267]. The M1 motif in TAD-boundary-associated sites displayed higher RS scores compared to the motifs of non-TAD-boundary-associated sites (Fig. 2.5E). I further compared the mean RS score per base between the two categories for CTCF sites at every conservation level and confirmed the generality of this observation (Fig. 2.5F). In addition, I established that this observation was not caused by an enrichment of specific motif instances at TAD boundaries (Fig. A.3 - Appendix A).

Taken together, these results showed that CTCF binding sites at TAD boundaries are subject to stronger evolutionary constraints than the CTCF binding sites located further away, and that this relationship is independent of evolutionary origin of the sites.

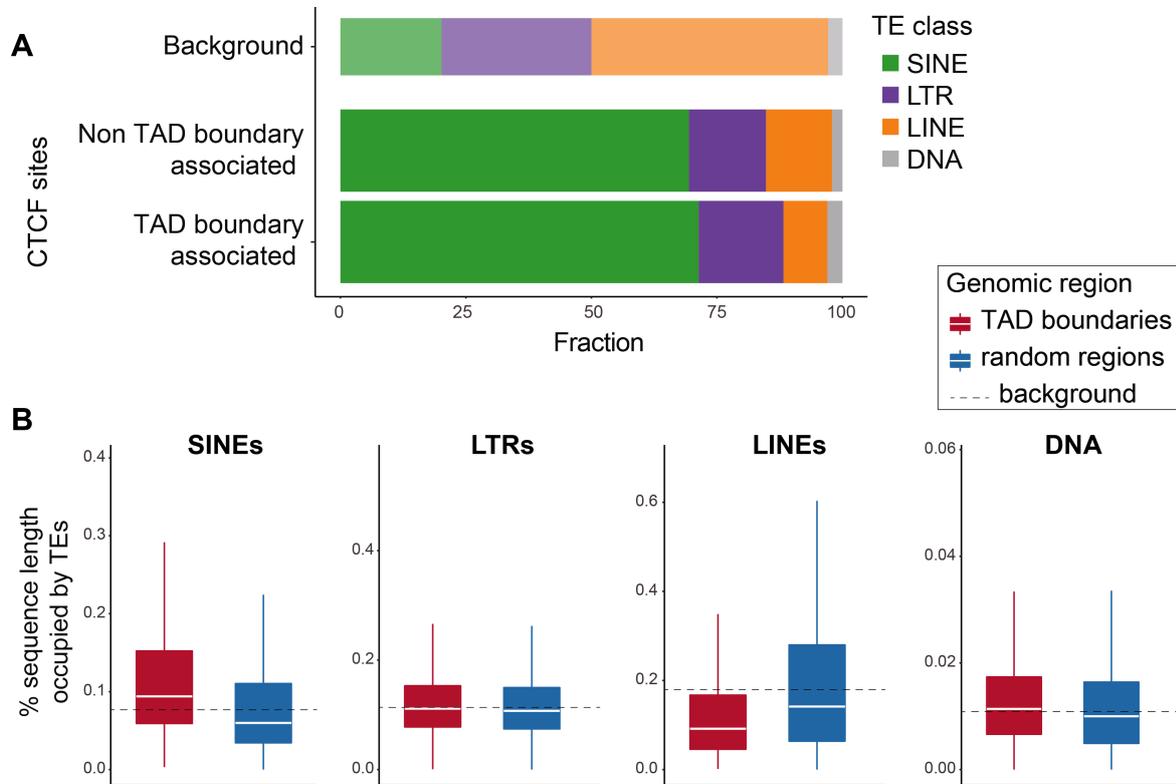
### **2.2.3 Representation of LINEs and LINE-derived CTCF sites at TAD boundaries**

The above observations reveal that occurrence at TAD boundaries is associated with sequence and functional conservation of CTCF binding sites. I questioned whether distinct evolutionary mechanisms underlie CTCF binding expansion near TAD boundaries as compared to the background genome. According to previous studies, the genome-wide binding profile of CTCF in mammals is, to a large extent, driven by expansion of repeat elements [176, 164, 177, 261]. I therefore sought to characterise the enrichment of transposon classes that drive CTCF binding expansion at TAD boundaries compared to the whole genome. After distinguishing CTCF sites based on whether they locate at TAD boundary regions<sup>3</sup> or not, and for each group I calculated the number of CTCF peak centers that were embedded in SINEs, long terminal repeats (LTRs), long interspersed nuclear elements (LINEs), and DNA transposons. As expected, the largest fraction of CTCF sites in both

<sup>3</sup>A TAD boundary region, here, is defined between two adjacent TADs as the first nucleotide of the downstream TAD +/- a 50kb window.

categories were found to be SINE-derived (Fig. 2.7A) [176]. The fraction of SINE-derived CTCF sites at TAD borders was slightly, but not significantly, larger than in the rest of the genome ( $\chi^2$  test without Yates correction:  $p = 0.01$ ), implying that SINEs may have uniform potential to establish a CTCF site at both TAD boundaries and other genomic regions. Similarly, CTCF sites of LTR origin did not show significant differences between the two categories ( $\chi^2$ :  $p = 0.015$ ). In contrast, the relative proportion of DNA transposon-derived CTCF sites was increased at TAD boundaries ( $\chi^2$ :  $p = 0.0003$ ) but accounted for less than 3% of the TEs that contribute to CTCF binding (Fig. 2.7A). The depletion of LINE-derived CTCF binding sites at TAD boundaries compared to the background genome was the most striking difference ( $\chi^2$ :  $p = 3.147e-15$ ; Fig. 2.7A) suggesting that CTCF binding site formation via LINE expansion is significantly less common at TAD borders than genome-wide.

In addition to inspecting CTCF sites derived from transposable elements of different classes, I evaluated the representation of different transposon classes themselves around TAD boundaries, irrespectively of whether they contained CTCF binding sites. In particular, I determined the corresponding fraction of the 100kb TAD border regions occupied by SINE, LTR, LINE, and DNA transposon sequences and compared these with random genomic regions of similar size and distribution. I found that SINE sequences were significantly enriched at TAD boundaries (Mann-Whitney U test:  $p < 2.2e-16$ ; Fig. 2.7B) [133], which is in line with the observed enrichment of CTCF sites at TAD borders and the known role of SINEs in CTCF binding expansion. The fraction of LTR-derived sequences at TAD boundaries was only marginally higher than random genomic regions ( $p=0.005$ ), and the fraction of DNA transposon sequences was also slightly higher at TAD borders ( $p = 9.72e-14$ ; Fig. 2.7B). In contrast, LINE sequences were significantly under-represented at TAD boundaries, compared to random genomic regions (Mann-Whitney U test:  $p < 2.2e-16$ ; Fig. 2.7B), suggesting that TAD boundaries are depleted of LINES. That may also explain why LINE-derived CTCF sites appear under-represented at TAD boundaries (Fig. 2.7A), as shown above. Considering that LINE elements typically represent longer sequences compared to the other transposons, this observation potentially indicates that LINE insertion is negatively selected at TAD borders as a consequence of their characteristic long insertion fragments that might be disruptive for functionally important regions harboured at TAD boundaries. This result is complementary to recent reports of selection against long sequence deletions at the functional regions of TAD boundaries, within a wider context of TAD boundary disruptions being under purifying selection [268]. Moreover, it extends our previous observations and reinforces the hypothesis that in addition to TAD-boundary-associated CTCF sites being subjected to stronger sequence and functional constraints, TAD boundary regions as a whole are under stronger evolutionary pressure [268].



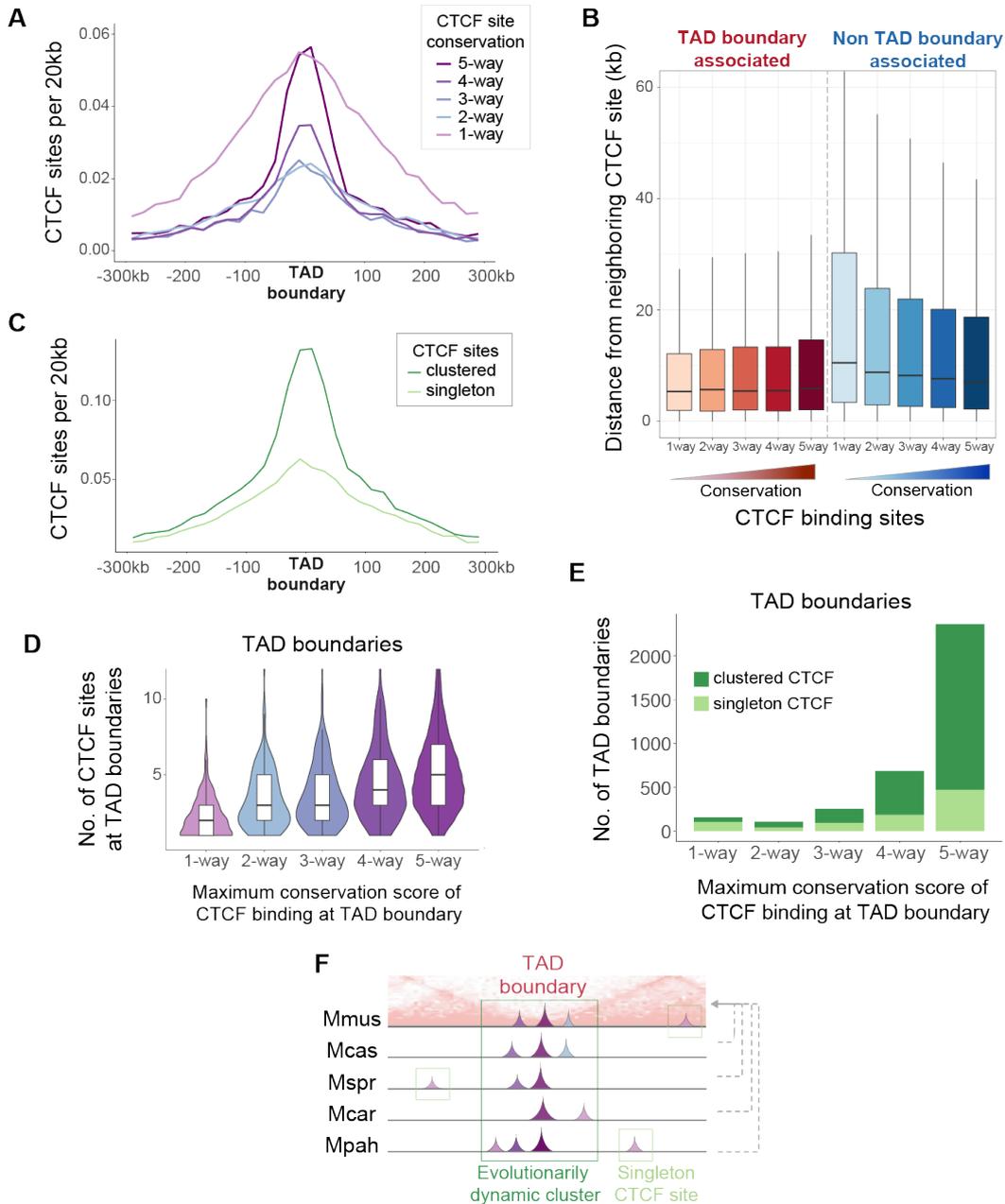
**Fig. 2.7 Differences in representation of TE classes and their association with CTCF binding sites between TAD boundaries and other genomic regions.** (A) Fractions of TAD-boundary-associated versus non-TAD-boundary-associated CTCF binding sites that are embedded in different TE classes. LINE-embedded CTCF-sites are under-represented at TAD boundaries ( $\chi^2$  test without Yates correction:  $p = 3.12e-15$ ), while DNA-transposon-embedded CTCF sites are over-represented ( $\chi^2$  test:  $p = 0.0003$ ), although accounting for just 3% of the TAD-boundary-associated sites. SINE-derived CTCF sites ( $\chi^2$  test:  $p = 0.01$ ) and LTR-associated CTCF sites ( $\chi^2$  test:  $p = 0.015$ ) show no significant differences between the two categories. The top bar (Background) shows the percentage of the BL6 genome sequence that corresponds to each TE class, for reference. (B) Fraction of sequence length of TAD boundary regions (TAD boundary  $\pm$  50kb) occupied by each TE class, compared to random genomic regions of equal length. SINE sequences are significantly over-represented (Mann-Whitney U test:  $p < 2.2e-16$ ), while LINEs are significantly depleted at TAD boundaries ( $p < 2.2e-16$ ). DNA transposons are slightly, but significantly, enriched at TAD borders ( $p = 9.72e-14$ ), although they account for only 1% of the sequences of the studied regions on average. Representation of LTR sequences shows no significant difference between TAD boundaries and random genomic regions ( $p = 0.005$ ; significance threshold: 0.001).

### 2.2.4 Evolutionarily dynamic clusters of CTCF binding sites at TAD borders

Aiming to gain further insight into the architecture of TAD boundaries, I sought to characterise the organisation of their contained CTCF binding sites, which had different conservation levels. In particular, after grouping all CTCF binding sites based on conservation level, I examined their density, i.e. their proximity to each other, considering also their distance from the TAD boundary. As expected, TAD borders were shown to be highly enriched for *Mus*-conserved CTCF binding events (Fig. 2.8A). However, species-specific CTCF sites were, surprisingly, also enriched at TAD boundaries (Fig. 2.8A). CTCF sites shared by 2-4 species also seemed to be moderately enriched around TAD boundaries. Thus, TAD boundaries harbour numerous conserved CTCF binding sites, but also a high concentration of species-specific CTCF sites in the examined, closely related, rodent species. On a further note, regardless their conservation level, the average distance of any TAD-boundary-associated site to its closest neighbouring site was consistently shorter than the corresponding average of non-TAD-boundary-associated sites, with a median distance equal to ~5.3kb-5.9kb (Fig. 2.8B). In contrast, CTCF binding sites not associated with a TAD boundary, besides being further apart from each other (Mann-Whitney U test:  $p < 2.2e-16$ ), their proximity to their closest neighbouring site was dependent on conservation level; the less conserved a site was, the further it was located from other sites. The median distance from the closest neighbouring site was 7kb for 5-way conserved CTCF sites, as opposed to a median of 10.5kb for species-specific ones (Fig. 2.8B).

Further querying whether CTCF sites organise in specific conformations at TAD borders, I also investigated potential ancestral clusters from the whole set of *Mus* CTCF binding sites projected to the BL6 genome ( $n = 56,625$ ; Fig. 2.1C). I searched for potential clusters of neighbouring CTCF binding sites, defining a cluster as a group of at least two CTCF sites being located less than 10kb apart from each other on the genome. A bit more than half of all CTCF sites (57%) were found to be part of a cluster; these were 32,393 sites in total ("clustered" sites), making up 11,507 clusters. The remaining 43% of the sites were "singletons", i.e. not belonging to any cluster. Interestingly, clustered CTCF sites appeared to be significantly enriched at TAD borders compared to singleton CTCF sites (Fig. 2.8C). This finding strongly implies that clusters of proximal CTCF binding sites are a fundamental architectural feature of TAD boundaries.

Then, I further inspected the observed CTCF binding clusters at TAD borders, specifically looking into potential associations between CTCF binding site clustering, redundancy and presence of conserved and nonconserved binding events at TAD boundaries. I observed that TAD borders that harboured at least one 5-way conserved CTCF site also contained a higher number of CTCF sites overall (Fig. 2.8D) and these sites were mainly clustering with each other (Fig. 2.8E). Generally, a progressive association between presence of individual conserved CTCF sites with a higher abundance of CTCF sites at a TAD boundary was outlined. In summary, TAD boundaries seemed to usually contain at least one



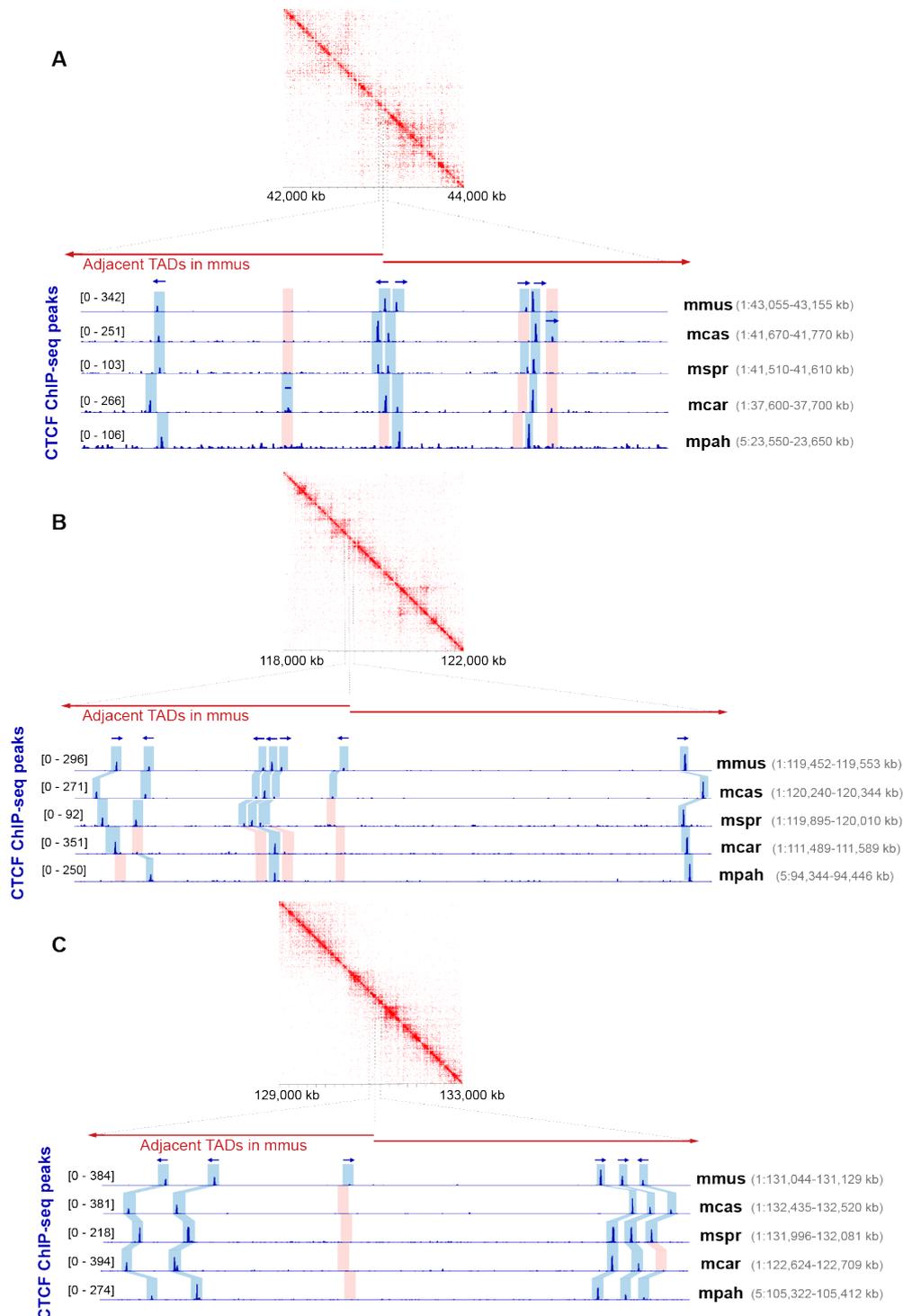
**Fig. 2.8 Clusters of both conserved and divergent CTCF binding sites at TAD boundaries.** (A) Both *Mus*-conserved and species-specific CTCF binding sites are highly enriched around TAD boundaries. (B) TAD-boundary-associated sites lie significantly closer to each other compared to non-TAD-boundary-associated CTCF sites (Mann-Whitney U test:  $p < 2.2e-16$ ). (C) CTCF binding sites that belong to a cluster (“clustered”) are more enriched at TAD boundaries than singleton CTCF sites. (D) The violin plots correspond to TAD boundaries categorised according to the maximum conservation level of CTCF binding they contain. Each violin plot shows the distribution of the total number of CTCF sites that occur at the TAD boundaries in the category. TAD boundaries with at least one *Mus*-conserved site (right-most violin plot) also have a higher number of CTCF sites overall. In contrast, TAD boundaries that do not contain any species-conserved CTCF sites (left-most violin plot) have low numbers of CTCF binding sites.

(E) The bars correspond to TAD boundaries categorised according to the maximum conservation level of CTCF binding they contain. Dark green demarcates TAD boundaries with clustered CTCF sites; light green shows TAD boundaries with only singleton sites. TAD boundaries that harbor species-conserved CTCF sites also contain CTCF site clusters. (F) Schematic representation of evolutionarily dynamic clusters of CTCF sites that commonly occur at TAD boundaries. TAD borders usually have at least one 5-way conserved CTCF site that is clustered with other sites of lower conservation, including species-specific ones. These CTCF clusters preserve CTCF binding potential at TAD boundaries.

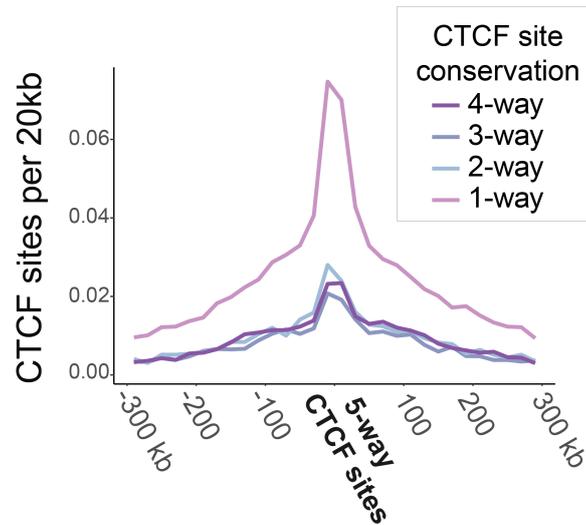
conserved CTCF site surrounded by other sites, frequently species-specific ones, and all the sites were forming clusters together. Taken together, these findings show that TAD boundaries harbour clusters of both *Mus*-conserved CTCF sites and more recently evolved CTCF binding sites (Fig. 2.8F, Fig. 2.9).

I questioned whether this phenomenon is solely a characteristic of TAD boundaries or is it also found in other parts of the genome. To address that, I firstly identified 5-way conserved CTCF sites that were not associated with TAD boundaries and inspected the CTCF binding profile around them. These were defined as sites with a distance  $d > 80\text{kb}$  from the closest TAD border, so as to ensure that, in case of a cluster forming around the given site, the entire cluster would be further than 50kb from the closest TAD border. Additional CTCF sites of various conservation levels, including several species-specific CTCF sites, were generally found to accumulate around these defined, anchor *Mus*-conserved sites (Fig. 2.10). This shows that *Mus*-conserved CTCF binding events are usually part of CTCF binding clusters, rather than appearing as singleton sites. Moreover, although the clusters are apparently stably anchored at 5-way CTCF sites, the cluster as a whole seems to be evolving dynamically, allowing for integration of many evolutionarily younger lineage-specific binding sites or loss of binding events in the one or the other mouse species.

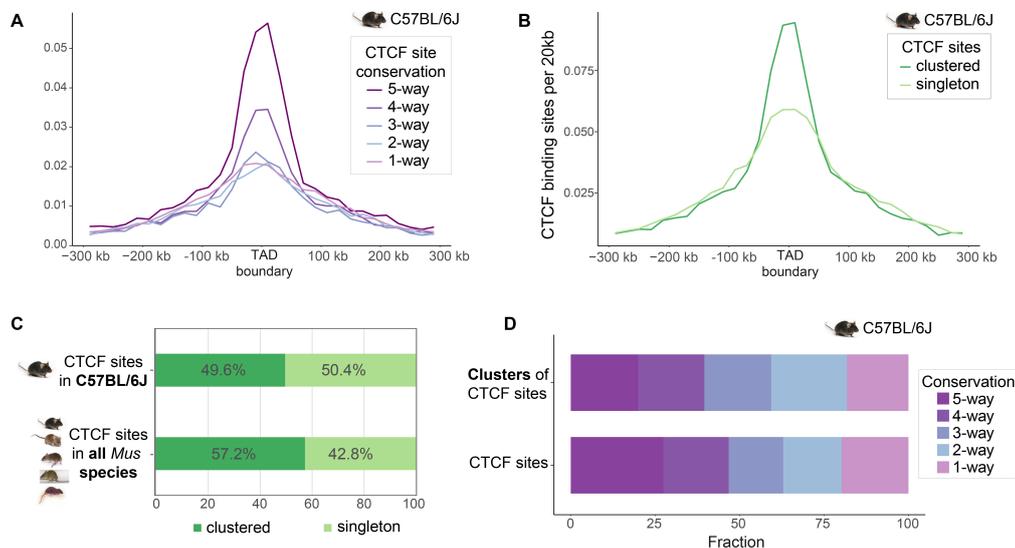
The above described CTCF binding features at TAD boundaries were outlined by inspecting the whole set of *Mus* CTCF sites projected on the BL6 genome. This allowed for an integrated overview of CTCF binding in mouse species, focusing on TAD boundaries, and evaluation of its evolutionary dynamics. I queried whether these evolutionary characteristics of clustered CTCF binding, which were concluded by assessing the aggregation of CTCF sites in all five species, were also recapitulated in a single mouse species, specifically in BL6. In that case, I evaluated the aforementioned CTCF binding features by inspecting only the identified ChIP-seq enriched sites in BL6. In that case, orthologous aligned regions of putative CTCF binding sites from other species were not considered. This inspection confirmed that BL6 CTCF sites of any conservation level were enriched at TAD boundaries (Fig. 2.11A) and that clustered CTCF sites in BL6 were also more frequent at TAD boundaries than singleton CTCF sites (Fig. 2.11B), as observed in all *Mus* species (Fig. 2.11A, 2.11C). Moreover, I found that half of BL6 CTCF binding sites were clustered, similar to the full set of *Mus* CTCF binding regions (Fig. 2.11C). It was also shown that the conservation of whole clusters of CTCF sites in BL6



**Fig. 2.9 Examples of TAD boundary regions harboring clusters of both conserved and divergent CTCF binding sites.** (A–C) CTCF ChIP-seq tracks illustrating three examples of TAD boundary regions harboring clusters of closely located CTCF binding sites. Although some of the sites are conserved across species, there are also often lineage-specific gains or losses in the vicinity. Blue shadow boxes highlight the statistically significant peaks identified by MACS, while pink shadow boxes mark CTCF binding losses (orthologous regions with no significant peaks). Arrows indicate the orientations of the CTCF binding motif identified within each peak. In case of more than one motif identified in a peak, the orientation shown corresponds to the motif with the lowest  $p$  value. The contact maps were visualised using Juicebox [269]



**Fig. 2.10 Potential occurrence of CTCF site clusters also away from TAD boundaries.** Enrichment of CTCF sites of different conservation levels around *Mus*-conserved CTCF sites that are not associated with TAD boundaries (distance from closest TAD border:  $d > 80$ kb). A high number of species-specific (1-way) CTCF sites are present around these “anchor” 5-way conserved sites, showing that sites of mixed conservation levels can be clustered together.



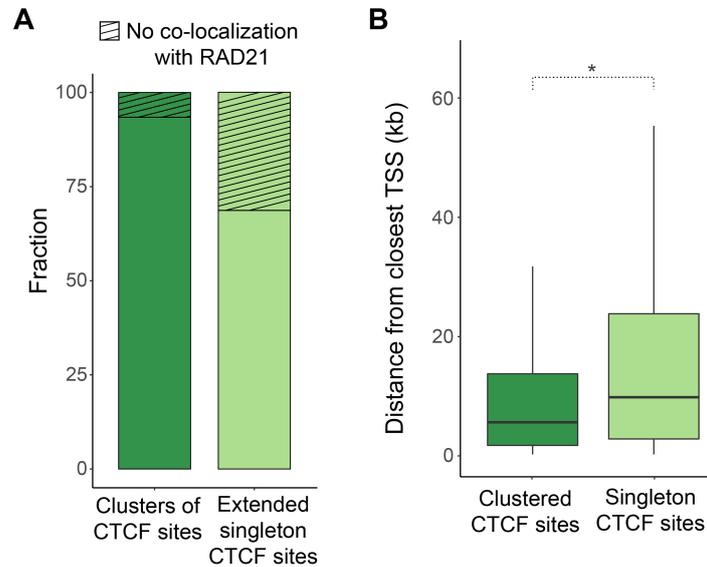
**Fig. 2.11 Inspection of the CTCF binding profile in BL6 confirms that CTCF sites form clusters in individual species.** (A) Enrichment of BL6 CTCF sites of different conservation levels at TAD boundaries. (B) Clustered BL6 CTCF sites are more highly enriched than singleton sites at TAD borders. (C) The fraction of clustered CTCF sites in BL6 is similar to that of CTCF sites belonging to ancestral *Mus* clusters. (D) The conservation pattern of CTCF site clusters, as distinct functional entities, resembles that of individual CTCF binding sites.

was similar to that of individual CTCF binding sites (Fig. 2.11D). This implies that clusters of CTCF sites are evolving under selective pressure similar to that underlying the conservation of individual CTCF binding sites.

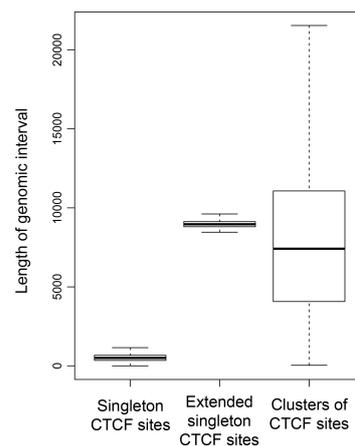
In summary, CTCF binding clusters comprising both species-conserved and species-specific sites are a common characteristic of TAD boundaries. These clusters are maintained by dynamic evolutionary processes underlined by deployment of evolutionarily young CTCF sites around older, conserved ones. In addition, CTCF clusters with similar characteristics can also be found at long distances from currently identified TAD borders, which suggests a broader role in genome function. That could be to generally provide a CTCF binding buffer for ensuring CTCF binding at regions where it is of important functionality.

### **2.2.5 Localisation of CTCF site clusters with respect to cohesin-occupied sites and genes**

CTCF is known to interact with cohesin to form chromatin loops [170, 202, 203, 270, 201, 271] and to help establish TAD boundaries [193, 192]. In addition, CTCF is known to frequently bind near gene promoters [272]. Aiming to gain further insight into possible additional functional roles of CTCF binding site clusters, I determined the frequency of co-localisation of clustered and singleton CTCF sites with the RAD21 cohesin subunit, as well as their distance from gene transcription start sites (TSS) in BL6. To address the first question, I identified cohesin enriched regions on the BL6 genome using ChIP-seq data for the cohesin subunit RAD21 from BL6, which were provided by the Odom lab. Then I intersected them with the CTCF clusters and singleton sites. CTCF site clusters were significantly more likely to overlap with RAD21-enriched regions compared with singleton CTCF sites; the respective fractions of clusters and of singleton sites co-localizing with cohesin were 93% and 69% ( $\chi^2$  test,  $p < 2.2e-16$ ) (Fig. 2.12A). It is noted that since clusters represent longer genomic regions than singleton CTCF sites, to control for the length difference of the compared region sets, I extended the genomic intervals around singleton CTCF sites so that the mean of their length distribution was equal to that of the CTCF site clusters (Fig. 2.13). This ensured that the higher overlap frequency of CTCF clusters compared to singletons was not an artefact of their larger length. Generally, this observation suggests that clusters of closely located CTCF binding sites help stabilise cohesin at anchors of chromatin loops and/or TAD boundaries. Concerning the characterisation of CTCF sites with respect to their proximity to genes, I measured the distance of each CTCF site belonging to a cluster to the nearest gene TSS and compared this distribution to the corresponding distances for singleton CTCF sites. It was demonstrated that CTCF sites belonging to a cluster are generally located significantly closer to TSSs (Median distance = 5.3kb) than singleton CTCF sites (Median distance = 10.9kb) (Mann-Whitney U test,  $p < 2.2e-16$ ; Fig. 2.12B). This suggests that



**Fig. 2.12 Clustered CTCF sites overlap more frequently with cohesin and locate closer to genes, compared to singleton CTCF binding sites.** (A) 93.7% of the clusters of CTCF binding sites colocalise with the cohesin subunit RAD21, while the respective fraction of extended singleton CTCF sites is 69% ( $\chi^2$  test:  $p < 2.2e-16$ ). The singleton CTCF binding regions were extended by a few kilobases prior to intersection with RAD21-enriched regions to ensure the mean of their length distribution is equal to the mean length distribution of clusters of CTCF sites. (B) CTCF sites that belong to clusters (clustered) are located closer to gene TSSs (Median distance = 5.3kb) than singleton CTCF sites (Median distance = 10.9kb) (Mann-Whitney U test:  $p < 2.2e-16$ ).



**Fig. 2.13 Length distribution of genomic intervals occupied by singleton CTCF sites, “extended” singleton CTCF sites and clusters of CTCF sites.** The extended singleton CTCF sites represent genomic windows of singleton CTCF sites that were extended so that the mean of their length distribution becomes equal to that of the length distribution for the CTCF clusters.

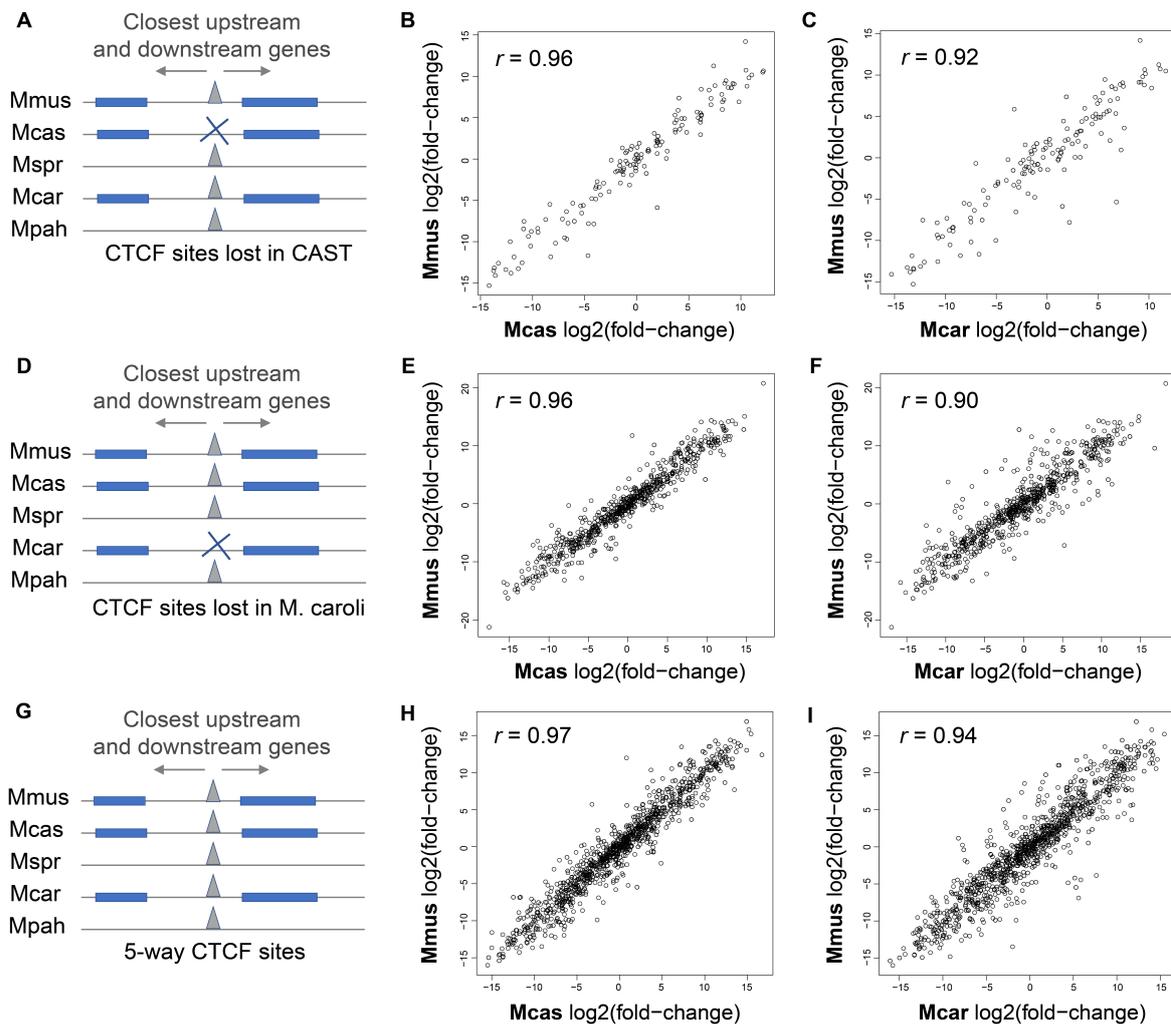
clusters of CTCF sites may also play an integral role in regulating gene expression.

### 2.2.6 Impact of species-specific CTCF binding event loss on insulating function at TAD boundaries

CTCF binding sites lying at TAD boundaries are proposed to buffer contact insulation between *cis*-regulatory elements of adjacent TADs [187]. Independent studies have shown that local disruption of CTCF binding at TAD boundaries can lead to ectopic interactions between promoters and enhancers [259, 151, 135]. Nevertheless, the impact of such disruptions on local gene expression has not been systematically investigated. In this study, I leveraged the natural genetic variation in the examined, closely related mouse species and our own CTCF binding data to study the effect of CTCF binding site loss in a model fixed by evolution. This approach offers significant advantages over many other experimental approaches, such as disruption of specific CTCF sites [258, 151, 189, 135], hemizygous *Ctcf* deletion [273], or transient acute depletion systems [260, 196, 199] in which there are global disruptions of cellular equilibrium.

Particularly, I investigated the instances at TAD boundaries where a CTCF binding event was conserved in all but one of the five study species, i.e. there was a species-specific loss of a TAD-boundary-associated CTCF binding event. I estimated the impact of these changes on the expression of proximal genes using RNA sequencing (RNA-seq) in BL6, CAST and CAROLI. In the first place, I identified either CAST-specific (Fig. 2.14A) or CAROLI-specific losses of individual CTCF binding events at TAD boundaries (Fig. 2.14D). For each of these lost CTCF sites, I determined the closest upstream and the closest downstream one-to-one orthologous gene in all three species (Fig. 2.14A, 2.14D) and calculated the relative gene expression of this gene pair (expressed as  $\log_2(\text{fold-change})$ ) in each of the species (see Methods). I then compared these relative expression patterns among the three species.

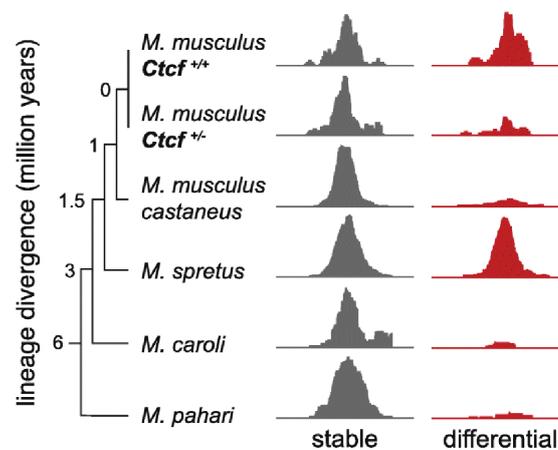
There was no detected impact on insulating function due to species-specific losses of individual CTCF binding events at TAD borders (Fig. 2.14). This suggests that expression patterns of genes at the borders of TADs are robust to the losses of individual CTCF binding even in cases where the binding event is preserved in multiple other closely related species. Interpreting that in view of our observed CTCF binding site clusters, this may underline an interchangeable or additive function of binding sites contained in a CTCF cluster, which contributes to the maintenance of this functional resilience.



**Fig. 2.14 Gene expression patterns around TAD boundaries are robust to local species-specific losses of individual CTCF sites.** (A) Identified CAST-specific CTCF sites losses at TAD boundaries and estimated gene expression patterns around them, expressed as  $\log_2(\text{fold change})$  between the closest downstream to the closest upstream gene. (B, C) Comparisons of  $\log_2(\text{fold change})$  values of gene pairs flanking the CAST-specific losses of CTCF sites between BL6 and CAST, with inconsistent CTCF binding, as well as between BL6 and CAROLI, with consistent CTCF binding. Only genes that have a one-to-one orthologous relationship and similar gene lengths among BL6, CAST and CAROLI were used. (D) CAROLI-specific CTCF sites losses at TAD boundaries and estimated gene expression patterns around them, expressed as  $\log_2(\text{fold change})$  between the closest downstream to the closest upstream gene. (E, F) Comparisons of  $\log_2(\text{fold change})$  values of gene pairs flanking the CAROLI-specific losses of CTCF sites between BL6 and CAST, with consistent CTCF binding, as well as between BL6 and CAROLI, with inconsistent CTCF binding. (G) For reference, *Mus*-conserved CTCF sites and gene expression patterns around them (computed  $\log_2(\text{fold change})$  of the closest downstream to the closest upstream gene) in each of the species are shown. (H, I) Comparisons of  $\log_2(\text{fold-change})$  values of gene pairs flanking the examined *Mus*-conserved CTCF sites between BL6 and CAST, as well as between BL6 and CAROLI.

### 2.2.7 Effect of CTCF hemizyosity on CTCF binding and TAD organisation

I also leveraged the gained insights into *Mus* CTCF binding evolution to help further inform the cellular effects of CTCF haploinsufficiency in BL6. It is known that although homozygous deletion of the *Ctcf* gene in mouse embryos or in specific cell types causes, respectively, lethality [274] or dramatic organ dysfunction [275–278], mice with heterozygous *Ctcf* deletion can develop normally. However, they are—reportedly—prone to developing cancer [273], while *Ctcf* has also been identified as a haploinsufficient tumour suppressor gene in human cancer types [273, 158, 157].



**Fig. 2.15 CTCF binding at *Mus*-conserved sites is robust to hemizygous *Ctcf* deletion.** We found that the identified differentially bound sites in *Ctcf* hemizygous cells are depleted of *Mus*-conserved CTCF sites. Example genome tracks showing CTCF ChIP-seq enrichment at a stable CTCF binding site (grey; chr6:120,736,800) and a differentially bound CTCF site (red; chr2:31,887,060) in BL6 *Ctcf*<sup>+/-</sup>. CTCF ChIP-seq enrichment at the corresponding orthologous regions in the other mouse species is also shown.

Our collaborators, Sarah J. Aitken and Ximena Ibarra-Soria compared genomic features between cohorts of wild type (WT) and *Ctcf* hemizygous (*Ctcf*<sup>+/-</sup>) mouse embryonic fibroblasts (MEF) from BL6 mice. Their focus was on investigating differential CTCF binding and its functional implications in *Ctcf* hemizygous cells compared to WT. Sarah J Aitken performed experiments and Ximena Ibarra-Soria processed the generated data. They reported that, despite deletion of one of the *Ctcf* gene copies, there is some compensation of the mRNA (63%) and CTCF protein levels (73%) in the *Ctcf*<sup>+/-</sup> MEF hemizygous cells. With respect to the hemizyosity effects on the binding activity of CTCF, they identified 787 differentially bound CTCF sites motifs of lower binding affinity in MEF hemizygous cells, as compared to WT. I further found that the *Mus*-conserved sites, which I identified and reported above, were depleted among the set of differentially bound sites in *Ctcf*<sup>+/-</sup> cells (Fig. 2.15). This shows that CTCF binding at conserved binding sites is robust to low CTCF protein levels

resulting from hemizyosity.

Motivated by the previously reported evidence that *Ctcf*<sup>+/-</sup> mice are prone to cancer development, I also aimed to address whether the identified differentially bound sites upon *Ctcf* hemizyosity are recapitulated by CTCF site mutagenisation in human cancer. Thus, I compared the differentially bound loci in BL6 *Ctcf*<sup>+/-</sup> MEF cells against CTCF binding sites containing mutation hotspots in human gastrointestinal tumors [279]. Specifically, I projected the identified recurrently mutated CTCF sites ( $n=17$ ) from the human genome to mouse (BL6) to find their orthologous regions and examined whether they coincided with differentially bound sites in *Ctcf* hemizyosity. However, I did not find any overlap between the two orthologous sets of sites. Thus, these particular findings of differentially bound *Ctcf*<sup>+/-</sup> sites were not proven to recapitulate the suggested recurrent, driver CTCF mutations in human gastric cancer.

To investigate the impact of *Ctcf* hemizyosity on large-scale genome organisation, I called TADs using normalised Hi-C data that were generated and pre-processed, respectively, by Sarah J. Aitken and Ximena Ibarra Soria. After calling the TADs, we compared TAD intervals between the two genotypes and found that the vast majority of self-interacting domains (95%) remained stable in the *Ctcf* hemizygous MEFs compared to WT. Ximena Ibarra-Soria further estimated the correlation coefficient between the interaction profiles of the two genotypes, which was ~0.9.

These results further support the robustness of TAD structures to CTCF binding perturbations, such as the differential binding caused by *Ctcf* hemizyosity. Given that TAD boundaries frequently harbour conserved CTCF sites, their resilience seems to be attributed to the robustness of CTCF binding at conserved sites, while CTCF site clustering may also enhance this effect.

## 2.3 Discussion

In this study, I leveraged natural genetic variation in the CTCF binding events of five closely related species to investigate and characterise features of CTCF binding at TAD boundaries. The findings highlight that CTCF binding sites at the boundaries of TADs are generally subject to stronger sequence constraints compared to CTCF sites in the background genome. Nevertheless, the CTCF binding profile at TAD borders seems to also be evolving under the effect of dynamic evolutionary processes. This is indicated by numerous gains of new species-specific CTCF binding sites close to species-conserved ones, giving rise to mixed clusters with both evolutionarily old and young CTCF binding sites.

The described analyses demonstrate that CTCF binding is conserved to a large extent across *Mus* species, which is consistent with prior studies showing a high level of CTCF binding conservation across more distantly related mammalian species [175, 77, 164]. Also in agreement with previous reports of co-occurring TAD boundaries and CTCF sites that are conserved in mammalian lineages [134, 184], our data reveal a high occurrence frequency of *Mus*-conserved CTCF sites at the boundaries of TADs. In addition to that, we show that a significant fraction of species-specific CTCF sites also localise in the vicinity of TAD borders. Moreover, we demonstrate that CTCF binding sites at TAD boundaries have both stronger sequence constraints and stronger binding affinity, independently of their conservation across species. Our data also reveal discrepancies in the expansion of TE classes at TAD boundary regions compared to the background genome. Specifically, TAD boundaries are relatively depleted of both LINE elements and LINE-derived CTCF binding sites, which may be evidence of negative selection against insertions of long - potentially disrupting - sequences at TAD boundaries. This is complementary to observed structural variant depletion at TAD boundaries as an effect of purifying selection [268]. Overall, these observations suggest that the functional role of CTCF binding at TAD boundary regions is maintained by multiple evolutionary mechanisms including local sequence constraint, new site acquisition, and rejection of insertions and deletions.

Our results show that dynamically conserved regions that contain clusters of CTCF sites are another common characteristic of TAD boundaries. These clusters comprise both conserved CTCF binding events, which were apparently fixed at TAD boundary regions in the common ancestor, and divergent sites, which are the result of more recent gains or losses within the distinct mouse lineages. These clusters create the potential that local turnover events will largely preserve TAD structure and function. Indeed, a recent study has demonstrated CTCF binding site turnover at loop anchors mediated by TEs and suggested that this is a common mechanism of contributing to conserved genome folding events between human and mouse [280]. Based on these observations, we conclude that the formation of CTCF binding site clusters serves as an additional evolutionary buffering mechanism to preserve the CTCF binding potential of TAD boundaries and ensure resilience of higher order chromatin structure by maintaining a dynamic redundancy of CTCF binding sites.

Evolutionarily conserved clusters of CTCF binding site may help explain previous observations of TAD structures remaining intact upon experimental disruption of individual or multiple CTCF sites, assuming that such clustered CTCF binding sites can be used interchangeably to provide higher order resilience against local disruptions. For example, Nora et al. show that the deletion of a TAD boundary is followed by ectopic *cis*-interactions locally but adjacent TADs do not merge; they hypothesize that there must be additional elements within TADs that “act as relays when a main boundary is removed” [135]. Furthermore, Barutcu et al. demonstrated that TAD structures are preserved upon deletion of the CTCF-rich *Firre* locus from a TAD boundary [258]. They hypothesize that additional CTCF binding sites outside the *Firre* locus may serve to recruit CTCF and thus help maintain the TAD boundary. Thus, neighbouring CTCF sites may support CTCF binding activity in a collective

way. I also found that gene expression around TAD boundaries in cases of species-specific losses of individual CTCF sites is highly robust. As a whole, these results strongly suggest that the dynamic conservation of genomic regions harboring clusters of CTCF sites is an important feature of CTCF binding evolution, which is critical to the functional stability of higher order chromatin structure. Interestingly, such clusters are also found in other genomic regions apart from TAD borders. It is possible that these regions are related to the establishment of higher order chromatin structure, potentially representing unidentified TAD boundaries or loop anchors, or other functional and regulatory roles of CTCF.

Further insight into the functional implications of CTCF site clusters come from our results that CTCF clusters co-localise with the cohesin subunit RAD21 to a greater frequency than singleton CTCF sites. Moreover, we demonstrate that clustered CTCF sites are located significantly closer to TSSs than singleton sites. Together, these suggest that clusters play an important role in stabilizing cohesin at specific genomic regions, as well as in transcriptional regulation. These observations may provide new mechanistic insight to the previously proposed dynamic loop maintenance complex (LMC) model, in which cohesin associates with a genomic region for a significantly longer time than CTCF molecules [173]. Specifically, our observations of clustered CTCF binding sites support the proposed rapid unloading and rebinding of CTCF molecules in close genomic proximity, which facilitates rapid cohesin translocation on DNA between CTCF binding sites that act as occasionally permeable boundary elements [281, 173]. This process apparently facilitates gene transcription by allowing RNA polymerase II to push cohesin along gene bodies [282, 281, 283].

Finally, it is tempting to speculate a connection between our identified clusters of closely located CTCF binding sites on the genome and the reportedly observed “clusters” -or “hubs”- of CTCF protein molecules in 3D [179, 171]. In particular, Hansen et al. have proposed a guided mechanism where an RNA strand can bind to and gather together multiple CTCF protein molecules near cognate binding sites. These CTCF molecule “hubs” apparently enhance the search for target binding sites, increase the binding rate of CTCF to its cognate sites (also as part of the LMC model) and are often implicated in chromatin loop formation [179, 171]. It is possible that our identified CTCF site clusters facilitate this mechanism by providing the cognate sites for the concentrated CTCF molecules to bind.

In conclusion, we identified dynamic evolutionary clusters of CTCF binding sites as a feature of TAD boundary architecture and we propose that these likely contribute to the remarkable resilience of TAD structures and gene expression to losses and gains of individual CTCF binding sites. Thus, further studies of seeking a definitive understanding of the functional roles of CTCF might require consideration of extended regions that harbor clusters of multiple CTCF sites.

This study exemplifies how natural molecular variation in closely related species can be used to gain mechanistic and functional insights into mechanisms of transcriptional regulation.

## 2.4 Methods

### 2.4.1 ChIP-seq experiments and data analysis

C. Feig performed chromatin immunoprecipitation experiments for CTCF using adult liver tissue from CAST and SPRETUS. ChIP-seq libraries from three biological replicates of each species were prepared as described in Schmidt et al. [164]. Subsequently, libraries were sequenced on a HiSeq2000 to produce 100bp paired-end sequence fragments.

In addition, I obtained published CTCF ChIP-seq data from the livers of BL6, CAROLI and PAHARI [261]. Three biological replicates from each species were used.

I aligned sequenced reads from CAST and SPRET to the reference genome assemblies CAST\_EiJ\_v1 and SPRET\_EiJ\_v1, respectively, using BWA 0.7.12 [284]. I also mapped the retrieved raw ChIP-seq reads from BL6, CAROLI and PAHARI to the genomes GRCm38 (mm10), CAROLI\_EIJ\_v1.1 and PAHARI\_EIJ\_v1.1, using the same method for the sake of performing matched analyses in all species. CTCF enrichment peaks were called with MACS 1.4.2 [285]. For downstream analyses, I used peaks identified in at least two replicates of each species.

S. Aitken also performed ChIP-seq in BL6 liver to enable identification genomic regions enriched for the cohesin subunit RAD21. Sample preparation and chromatin immunoprecipitation was performed as described in Schmidt et al. [164] using 10  $\mu$ g RAD21 antibody (Abcam, ab992, lot GR12688-8). Immunoprecipitated DNA and 50 ng of input DNA was used for library preparation using the ThruPLEX DNA-Seq library preparation protocol (Rubicon Genomics, UK). Library fragment size was determined using a 2100 Bioanalyzer (Agilent). Libraries were quantified by qPCR (Kapa Biosystems). Pooled libraries were deeply sequenced on a HiSeq2500 (Illumina) according to manufacturer's instructions to produce single-end 50 bp reads. I obtained 41,260,604 sequenced reads and mapped them on the mouse genome assembly GRCm38 using BWA 0.6.1 [284]. I then called RAD21 peaks using MACS2 2.1.2.1 [285].

### 2.4.2 TADs

I used the boundaries of mouse liver TADs published by Vietri Rudan et al. [184].

### 2.4.3 Conservation of CTCF binding sites in mice

To investigate the conservation of CTCF binding across the studied *Mus* species, I first found the orthologous alignments of the CTCF sites in the genomes of the other species. These orthologous

CTCF regions across mice were obtained using an extended version of the eutherian mammal Endo-Pecan-Ortheus (EPO) multiple genome alignment that also included the genomes of CAST, SPRET, CAROLI and PAHARI [261]. Once the orthologous regions of CTCF sites were identified in all *Mus* species, I cross-validated the binding of CTCF in each species using the corresponding ChIP-seq data. Specifically, I considered that a CTCF site was conserved if a) it had orthologous alignments across species and b) the orthologous alignments also contained a CTCF peak.

#### 2.4.4 Binding affinity and sequence constraint of CTCF motifs

To identify CTCF binding motifs, I retrieved the FASTA sequences of all CTCF peaks in BL6, using bedtools getfasta [286] and scanned these sequences for the primary CTCF binding motif (M1) using FIMO (Find Individual Motif Occurrences) from the MEME suite [287, 288] with default parameters. I extended the identified 19 base-long M1 motifs to include 20 bases upstream and 20 bases downstream in order to ensure I captured the extended version of the motifs (M1 and M2) if present. Finally, I calculated the binding affinity of these sequences for CTCF using DeepBind [289], and compared the distributions of the affinity values between motifs found in TAD boundary-associated and in non-TAD boundary-associated CTCF peaks of each conservation level performing Mann Whitney U tests.

To retrieve Rejected Substitution (RS) scores for each position of every identified 19 base-long M1 motif in BL6, I fetched pre-calculated conservation scores for each nucleotide of these mouse M1 sequences from Ensembl [290]. These scores have been computed by retrieving genomic alignments blocks from a multiple genome alignment of eutherian mammals –including mouse- and running the GERP algorithm [267] on them. The “rejected substitution” (RS) score of a genomic position was calculated as the difference of observed to expected substitutions. I then averaged the RS score per position among all motifs and compared these averaged RS scores of TAD boundary-associated M1 motifs with non-TAD boundary-associated motifs (Fig. 2.5E, 2.5F).

#### 2.4.5 ChIP enrichment of identified CTCF peaks

The CTCF sites that I identified in each species were the intersection of the CTCF peaks called in at least 2 biological replicates. I calculated the ChIP enrichment of each CTCF site by averaging the ChIP enrichment scores, reported by MACS, over the replicates. The ChIP enrichment scores represent the pile up of ChIP-seq fragments mapping to each peak region. I then compared the distributions of average ChIP enrichment between TAD boundary-associated and non-TAD boundary-associated CTCF sites of each conservation level (Fig. 2.5C, 2.5D).

### 2.4.6 Motif-word usage analysis

I scanned all CTCF peaks from each of the five species for the primary CTCF binding motif (M1) using FIMO from the MEME suite [287, 288]. From the M1 motif instances identified in each species I retrieved the central, most informative 14mer and estimated its frequency of occurrence as: the number of occurrences of the 14mer word in CTCF binding regions divided by the number of occurrences of the word in the whole genome of the species using the procedure of Schmidt et al., [164]. I filtered out any motif word that occurred fewer than five times in the whole genome. I illustrated the occurrence frequency of the motif words in each species on a heatmap which is sorted by distance to the closest TAD border (Fig. A.3).

### 2.4.7 Association of CTCF sites with classes of Transposable Elements

I used the full set of CTCF sites identified in all species and projected them on to the BL6 genome (GRCm38), as well as published transposable elements in BL6 [261] (<ftp://ftp.ebi.ac.uk/pub/databases/vertebrategenomics/FOG21/transposableElements/>). I intersected the center of each CTCF binding sites with the transposable elements and reported the number of CTCF site centers that overlapped with each TE class. The overall representation of each TE class in the whole genome that is shown as a reference (marked as “background” in Fig. 2.7A) was calculated as: the total length of all TE class (e.g. SINE, etc) sequences divided by the total genome length.

### 2.4.8 Representation of TE classes at TAD boundary regions

I defined TAD boundary regions as genomic windows of 50kb upstream and 50kb downstream of the boundaries of TADs (Fig. 2.7B). To evaluate the representation of each TE class in each of these regions, I summed the length of sequences that occurred within each TAD boundary region and corresponded to each TE class, and divided that by the total length of the TAD boundary region, i.e. 100kb. To retrieve random genomic regions of similar length and distribution, I shuffled the TAD boundary regions using bedtools shuffle, having first excluded chromosome Y, genome scaffolds, and chromosome ends, where TADs are not called. I repeated the same calculation for these shuffled TAD boundaries, which correspond to random genomic regions. I then plotted the distribution of these values for TAD boundary regions and random genomic regions. To determine the representation of each TE class in the background genome (dotted line in Fig. 2.7B), I divided again -as in Fig. 2.7A- the total length of all sequences that correspond to each TE class by the total BL6 genome (GRCm38) length.

### 2.4.9 Density of CTCF sites at TAD boundaries and clusters of CTCF binding sites

To determine the enrichment of CTCF binding sites in TAD boundary regions (compared to the surrounding genome) I measured the distance of each CTCF binding site to its closest TAD boundary using bedtools closest. I then categorised the CTCF sites based on their conservation level. For each CTCF site category, I grouped all distance values up to +/- 300kb in bins of 20kb and plotted the number of CTCF sites in each bin divided by the length of the bin, i.e. 20kb (Fig. 2.8A). To further characterise the density of CTCF sites at TAD boundaries, I also grouped CTCF sites according to their conservation level and association with a TAD boundary (vs non association with any TAD boundary), and for each of these categories I found the distance of each CTCF site from its closest CTCF site using bedtools closest (Fig. 2.8B).

To identify clusters of CTCF binding sites, I used the full set of CTCF binding sites of all five *Mus* species projected onto the BL6 genome (GRCm38), as shown in Fig. 2.1C. I identified instances of consecutive CTCF sites that were up to 10kb apart from each other, using bedtools cluster. I then determined and compared the enrichment of clustered and singleton CTCF sites at TAD boundaries using the same approach as in Fig. 2.8A but having categorised the CTCF sites based on whether they belong to a cluster ("clustered") or not ("singletons") (Fig. 2.8C).

For Figures 2.8D, 2.8E, I again defined TAD boundary regions as TAD boundary +/- 50kb. For each region I determined the highest conservation level of CTCF sites it contains and based on this I categorised the TAD boundary region in one of five categories. Subsequently, for each category of TAD boundary regions I found the total number of CTCF sites it harbored (Fig. 2.8D), as well as the number of these TAD boundary regions with clustered CTCF sites, as opposed to the ones without CTCF clusters but only singleton sites (Fig. 2.8E).

For Fig. 2.10, I defined *Mus*-conserved (5-way) CTCF sites with a distance to the closest TAD border >80kb as non-TAD boundary-associated. I calculated the enrichment of 1-way (species-specific), 2-way, 3-way and 4-way conserved CTCF sites in their vicinity in the same way as in Fig. 2.8A, but using as anchor the non-TAD boundary-associated 5-way CTCF sites themselves, instead of the TAD boundaries.

### 2.4.10 Clusters in BL6 and cluster conservation analyses

I identified clusters of CTCF binding sites in BL6 (Fig. 2.11) in the same way as for Fig. 2.8C but using only CTCF peaks called in BL6. I used the same methods as for Fig. 2.8A and 2.8C to determine the enrichment of CTCF sites of different conservation levels at TAD borders (Fig. 2.11A),

as well as the enrichment of clustered versus singleton CTCF sites (Fig. 2.11B).

To estimate the conservation of CTCF sites clusters (Fig. 2.11D), I identified all the genomic regions that correspond to clusters of CTCF sites in each of the five species separately. I then projected (i.e. found orthologous alignments of) the cluster regions of each species onto the BL6 genome and determined whether they overlap with the orthologous cluster regions of the other species.

### 2.4.11 RNA-seq data

I retrieved published liver-derived RNA-seq data from six biological replicates for each of the species BL6 and CAST [291], as well as from four biological replicates of CAROLI [292]. In addition, K. Stefflova generated and sequenced two additional RNA-seq libraries for CAROLI in order to have the same number of replicates in each species, using the methods described in Goncalves et al. [291] and Wong et al. [292]. Briefly, total RNA was extracted from two independent liver samples using Qiazol (Qiagen) and DNase treated with DNA-free (Ambion). Polyadenylated mRNA was enriched, directional double-stranded cDNA was generated and then fragmented by sonication and prepared for sequencing. Each of the two libraries were sequenced on an Illumina GAIIX to generate 75bp-long paired-end reads.

### 2.4.12 RNA-seq data processing and analysis

Adapter sequences were trimmed off with reaper from the Kraken tool suite [293]. The paired-end RNA-seq reads from each replicate of BL6, CAST and CAROLI were mapped to the corresponding species genomes (see “ChIP-seq experiments and data analysis” section of Methods) using STAR 1.5.2 [245] with default settings. Raw reads mapping to annotated genes were counted using htseq-count [247]. I then used the raw read counts to perform differential expression analyses with DESeq2 1.20.0 [250] and default settings.

To determine gene expression patterns around instances of 5-way conserved CTCF sites or species-specific CTCF site losses at TAD boundaries, I first identified their closest upstream and downstream genes in each species, as shown in Fig. 2.14A,D,G, using the genesets from Ensembl version 95 [294] and then calculated the relative gene expression of downstream to upstream gene in each species. I were not interested in the relative expression of the gene pair flanking a CTCF site per se, but in whether this ratio for each CTCF site is consistent between species especially when the in-between CTCF binding changes. For this reason, I only used CTCF sites that were flanked by 1:1 orthologous genes between the three species. I went on to use DESeq2 [250] in order to compute the log<sub>2</sub>(fold change) between the downstream and upstream gene -as a measure of the relative expression of

genes flanking each CTCF site- in each species, and to subsequently compare this  $\log_2(\text{fold change})$  between species. Since DESeq2 is not originally designed to normalise for gene lengths, and our aim was just to get as comparable expression pattern estimations as possible between the species, I also required all the orthologous genes that I used to have a similar length among the three species ( $0.7 < \text{len\_ratio} < 1.3$ , where  $\text{len\_ratio}$  is: length of gene in species A) length of its orthologous gene in species B). Finally, I compared the calculated  $\log_2(\text{fold-change})$  values for each gene pair in BL6 with the corresponding value of its orthologous gene pair in CAST (Fig. 2.14B,E,H) and in CAROLI (Fig. 2.14C, 2.14F, 2.14I).

### 2.4.13 TAD calling from BL6 MEF Hi-C data

Hi-C experiments were performed by Sarah J. Aitken in three biological replicates from each MEF cohort, i.e. WT and *Ctcf*<sup>+/-</sup>. The corresponding Hi-C libraries were sequenced on a HiSeq4000. The raw Hi-C data of all three biological replicates from each cohort were merged and Ximena Ibarra-Soria processed the raw data of each replicate pool, using HiCUP [236] and HOMER [237], to generate normalised contact matrices between genomic loci at 20kb genomic bin resolution. Subsequently, I called TADs using the normalised contact matrices and the tadTool software [295] tuning the parameters as `window_size=300kb` and `cut_off=11.5`.



## Chapter 3

# Cell-of-origin and expression profile of chemically induced tumours in liver

This chapter outlines my work within a broader collaborative project on liver cancer evolution (LCE). The LCE project is led by the LCE consortium, which comprises multiple research groups: Flicek group (EMBL-EBI, Cambridge), Odom group (CRUK, Cambridge and DKFZ, Heidelberg), Taylor and Semple groups (MRC Human Genetics Unit, Edinburgh) and the Lopez-Bigas group (IRB, Barcelona). The project makes use of a chemical carcinogenesis model in mouse species to study mutational processes, as well as the relative contributions of genomic and epigenomic variation in liver tumorigenesis. To put my own work within the broader context of the project, I firstly provide an overview of the motivation and the work carried out in the LCE study, as well as the available datasets. The LCE data presented here have been generated by the LCE consortium at Cancer Research UK (CRUK) and have also mostly been pre-processed by the LCE consortium. Then I describe my work on initial exploratory analyses of LCE bulk expression datasets and then combined use of these sets with external single cell expression data to profile the expression of the cell-of-origin in liver tumours. The analyses were led by myself, often making use of the LCE consortium pre-processed data. The distinction between my contributions and the contribution of others is noted in each section. Based on the analyses presented in this chapter, I drew further research directions to characterise mutagenesis at transcription factor binding sites in liver tumourigenesis, which will be described in chapter 4.

My work described below was part of the supporting analyses to a particular study of the LCE consortium that investigated mechanisms of mutagenesis in liver cancer development and tumour heterogeneity. This first LCE study was published in the following paper: Aitken, S.J., Anderson, C.J., Connor, F., Pich, O., Sundaram, V., Feig, C., Rayner, T.F., Lukk, M., Aitken, S., Luft, J., **Kentepozidou, E.**, Arnedo-Pac, C., Beentjes, S.V., Davies, S.E., Drews, R.M., Ewing, A. Kaiser, V.B., Khamseh, A. López-Arribillaga, E., Redmond, A.M., Santoyo-Lopez, J., Sentís, I., Talmane, L., Yates, A.D., Liver Cancer Evolution Consortium, Semple, C.A., López-Bigas, N., Flicek, P., Odom,

D.T., Taylor M.S., (2020). Pervasive lesion segregation shapes cancer genome evolution. *Nature* **583**, 265–270. [296]

## 3.1 Introduction

### 3.1.1 Liver Cancer and the LCE project

Primary liver cancer is the sixth most common among human malignancies and, by mortality rate, it ranks fourth. Recent estimates report more than 840,000 cases and 780,000 deaths per year, with higher incidence rates among men than among women [297, 298]. Hepatocellular carcinoma (HCC) is the most common histological type of primary liver cancer, accounting for more than 75% of all liver cancer cases [299], while the second most common type -intrahepatic cholangiocarcinoma-accounts for 12-15% of the cases [299]. HCC usually occurs in the context of liver diseases, such as non-alcoholic fatty liver disease (NAFLD), which is frequently associated with diabetes and obesity, chronic viral hepatitis B or C (HBV or HCV), or life style factors, such as increased alcohol consumption, smoking and dietary aflatoxin exposure [300, 298]. These conditions can lead to chronic liver disease or cirrhosis, which involve accumulated liver injury, and can eventually cause HCC development [301]. More than half of the HCC cases are diagnosed at a relatively late stage, which renders liver tumours resistant to currently available therapies [302]. As in other cancer types, the potential of cancer cells to evade pharmacological effects is, to a large extent, attributed to the intrinsic tumour heterogeneity. This heterogeneity is manifested by distinct genetic and phenotypic characteristics within the same tumour, or between tumours of a single individual or of different patients [303, 304].

Generally, tumour development in liver, as in other organs, can be viewed as an evolutionary process that encompasses mutagenesis and selection. This process draws from healthy liver tissue through the establishment of a clonal cell aggregation -the tumour- that has acquired a selective advantage. Thanks to this advantage, the clonal tumour can escape the tight regulation programmes that dictate normal cell behaviour, as well as exogenous constraints on cell proliferation posed by the cellular environment. It, thus, can display aberrant cell proliferation and outgrow adjacent cell populations. Further subclones can then be established through subsequent processes of mutagenesis and selection (reviewed in [305]).

It becomes obvious why, at the molecular level, development of cancer -including HCC- is associated with the occurrence of somatic mutations. Somatic mutagenesis is caused by exposure to a variety of endogenous and exogenous DNA damaging factors and can involve faulty DNA replication, enzymatic modification of DNA, or inefficiency of DNA repair mechanisms [306, 307]. A handful of these somatic mutations, referred to as *driver* mutations, have a causal role in driving oncogenesis, as

they provide the cell with a selective advantage and underline neoplastic transformation [306, 307]. Nevertheless, a cancer genome bears much more than just a small number of driver mutations. The large amounts of the rest, so-called *passenger* mutations, although not directly implicated in driving oncogenesis, reflect the series of mutational processes, which occur throughout tumourigenesis and tumour progression, and can cause biological perturbations [306, 308, 309]. Hence, they are important for understanding the historical molecular context of tumour development. In addition, although driver mutations confer a selective advantage to the tumour cell-of-origin, they do not necessarily dictate malignant transformation of the cancer cells; in fact this can happen, but it is not always the case [305]. Therefore, to understand malignancy development, it is useful to study the ensembles of mutational patterns in the genomes of the cancer cells. Overall, cancer genomes contain several distinct mutational patterns, or *mutational signatures*, which are generated as a result of different mutational processes and are associated with different mutagenic factors.

In HCC genomes, a variety of distinct driver mutations and mutational signatures have been identified. Most known driver mutations have been identified in coding genes, as it is more straightforward to infer the functional impact of changes in coding sequences, rather than in non-coding ones. Some of the most frequent driver mutations are those in: the *Tert* gene promoter, the tumour suppressor gene *Tp53*, the transcriptional regulator *Cttnb1*, genes *Arid1a* and *Arid2*, which are implicated in chromatin remodeling, etc [310]. With respect to the different mutational signatures identified in human HCC, they are associated with various patient factors, such as aging, alcohol abuse, smoking or aflatoxin exposure [311–313]. These mutational signatures arise as tumour mutations accumulate over a patient's entire lifetime and are a consequence of inter-individual differences in genome sequence, environmental exposures, DNA damage, and repair efficiency [313, 314, 312].

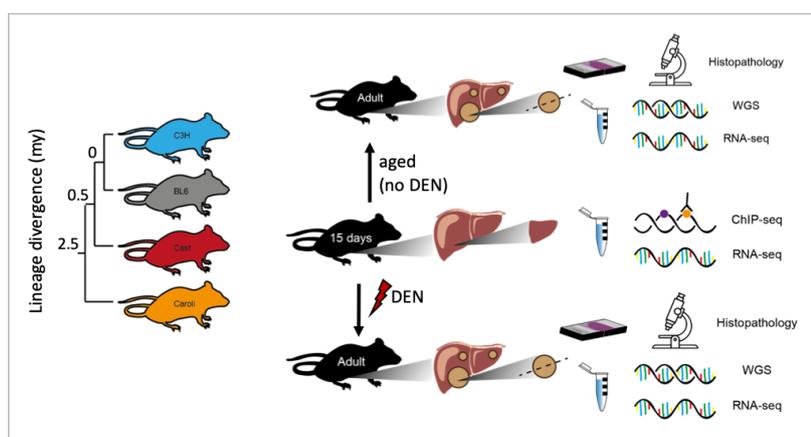
Therefore, the multiplicity of driver genes and mutational patterns identified in HCC genomes demonstrate the diversity of aetiologies and implicated molecular processes. Moreover, it manifests the molecular (spatial and temporal) heterogeneity of liver tumours as a response to the various exogenous and endogenous exposures, germline genetic variation, epigenomic variation and somatic mutational processes. All this intrinsic tumour heterogeneity and the correlative nature of the mutational signatures poses major challenges in deconvoluting overlapping mutational patterns in cancer, and in delineating their underlying DNA damage and repair biases, as well as their molecular functional impact. The challenge further extends to obstructing the interpretation of tumour clonal evolution dynamics and phenotypic alteration along tumour progression.

This motivated the use of a controlled, physiological cancer model system, by the LCE consortium, to study liver tumorigenesis while overcoming limitations posed by molecular heterogeneity and diverse aetiologies in liver cancer development. Within the LCE project, liver cancer evolution was re-run multiple times via a single chemical carcinogen insult in inbred mice of four different strains and/or species. These included two strains of the species *Mus musculus domesticus*, namely C57BL/6J

(BL6) and C3H, a subspecies of the same species, namely *M. m. castaneus* (CAST), and another related mouse species, *Mus caroli* (CAROLI). Although these include species, sub-species and strains, hereafter I will conventionally refer to them as *species*, yet bearing in mind that they are very closely related. The LCE project generalises an experimental system used for a study of the mutational patterns in the exomes of C3H mice [315]. In the LCE project, the use of inbred mice from each species provides a pool of genetically identical individuals, where tumorigenesis was replicated many times providing high resolution of mutational patterns and increased power for statistical analyses of the effect of the epigenetic differences between species. The different, yet closely related, genomes and epigenomes of the mouse species provide a pool of natural genetic and epigenetic variation that could potentially trigger different responses to carcinogenic factors.

The main aims of the LCE project focus on studying mutagenesis mechanisms in liver cancer initiation and evolution, as well as on disentangling and evaluating the relative contributions of the genome and the epigenome to liver cancer development.

### 3.1.2 Model of chemical carcinogenesis in mouse liver



**Fig. 3.1 Model of chemical carcinogenesis in mouse strains that was used in the LCE project.** A single dose of diethylnitrosamine (DEN) was administered to 15 day-old mice from strains/species: *Mus musculus domesticus* C57BL/6J (BL6) and C3H/HeOuJ (C3H), *M. m. castaneus* CAST/EiJ (CAST), and *M. caroli* (CAROLI). DEN-treated mice were aged until they develop liver tumours. A small number of individuals were not treated with DEN and were aged to develop spontaneous tumours. Samples were collected from DEN-induced tumours, 15 day-old healthy liver tissue and from a small number of spontaneously induced tumours, to generate molecular data. Healthy liver tissue from adult mice had also been collected to be used as a reference.

Liver tumours were chemically induced using a single dose of diethylnitrosamine (DEN) in 15-day old, male individuals of four mouse strains and species: *Mus musculus domesticus* strains C57BL/6J (BL6) and C3H/HeOuj (C3H), sub-species *M. m. castaneus* CAST/EiJ (CAST), and species *M. caroli* (CAROLI). DEN is a known carcinogenic compound that gets metabolically activated in liver by cytochrome P450 enzymes. Upon its bioactivation, it produces metabolites with a prominent genotoxic effect [316]. The DEN treated mice were aged to develop liver tumours. The time course from DEN administration to tumour development and collection varied among the different mouse species; for BL6, C3H, CAST and CAROLI, it corresponded to 35-36 weeks, 23-25 weeks, 39 weeks and 58-68 weeks, respectively. DEN-induced tumours that were resected from the livers of mice from each species included mostly dysplastic nodules and a small number of hepatocellular carcinomas (HCCs). It is noted that dysplastic nodules are lower grade tumours compared to the more developed HCCs. Moreover, tumours were selected by temporal and morphological criteria (early collection after onset of tumour development and homogeneity) via histopathological examination, so that they are monoclonal. Separate untreated mouse cohorts from each species were aged to develop spontaneous liver tumours. A small number of spontaneously developed tumours was also collected. Based on subsequent analyses the collected tumours of each species were found to be driven by different mutations in genes of the EGFR-RAF-RAS signaling pathway. In addition, samples of healthy liver tissue were collected from adult mice. In detail, samples were collected and data were generated (Fig. 3.1) as following:

- To determine the genomic, epigenomic and expression profile of the liver tissue at the time point of the carcinogen insult, liver samples were collected from 15 day-old mice of the four genotypes, prior to DEN treatment. Whole genome sequences and expression data (bulk RNA-seq) were generated from these samples. Epigenomic marks (CTCF, H3K4me3 and H3K27ac ChIP-seq) were also mapped using pools of ten liver samples so as to have enough material.
- A number of samples that had undergone DEN treatments but showed healthy liver phenotype were collected only from adult normal BL6 mice. Whole genome sequences and RNA-seq data were derived from these samples.
- DEN-induced tumours, including dysplastic nodules and HCCs, were collected from adult mice of the four genotypes. Whole genome sequences and RNA-seq data were derived from these samples.
- Matched normal liver samples were also collected from non-DEN-treated adult mice of the four strains. Whole genome sequences and RNA-seq data were derived from these samples.
- A small number of spontaneously developed dysplastic nodules and HCCs (no DEN treatment applied) were collected from aged BL6, C3H and CAST mouse cohorts. Whole genome sequences and RNA-seq data were derived from these samples.

Overall, the chemically-induced liver carcinogenesis model used by the LCE project encompasses the repetition of tumour initiation in individuals of four mouse strains and species upon a single DEN insult in 15 day-old mice. Observed differences in the latency periods from the carcinogen insult until tumour development among the strains/species reflect their differential susceptibility to hepatocarcinogenesis, as has been previously reported [317]. Among the four strains, C3H displays the highest susceptibility [318] and CAROLI the lowest. In addition, although driver mutations in the tumours of each strain were identified in the EGFR-RAS-RAF signaling pathway [318, 319, 315, 296], they were found to occur in different genes of the pathway. The initial phases of the project are focused on establishing molecular characterisations in each species and examining the reproducibility of the observed patterns and mechanisms of mutagenesis in the other species, while explicit cross-species comparisons are to be performed at further stages of the study.

Given this background and data sets and with the aim to understand the molecular characteristics of the liver tumours, I set out to investigate the expression profile of the cell-of-origin of the collected tumours using the bulk RNA-seq produced by the LCE consortium and externally produced developmental single cell RNA-seq data from mouse liver (described below). Following that, I also assessed mutation rates in transcription factor binding sites and their associations with the liver tumour expression output (Chapter 4).

### 3.1.3 Liver cancer cell-of-origin

As mentioned above, a main focus of the LCE project is to delineate the roles of genomic and epigenomic landscape in the cell's response to the carcinogen compound. Thus, a fundamental starting point of the study is the molecular characterisation of the cell that -through mutagenesis and selection- acquires the growth advantage, and leads to tumour development through neoplastic transformation and clonal expansion; that is the tumour cell-of-origin. Characterisation of the cell-of-origin would include determination of its cell type and comprehensive profiling of its expression, genomic, and epigenomic features.

With regards to the cell-of-origin of HCC, it is typically considered to be the hepatocyte. Yet, this has generally been a debated subject. A main reason is that the risk of an organ to develop tumours has been associated with cells that have high generative capacity, such as damage-induced stem cells in the adult liver [320]. Considering that hepatocytes, the most abundant liver cells, and cholangiocytes, the second most abundant, are already fully differentiated cells, HCC development has been proposed to involve activation of hepatic progenitor cells (HPCs, also known as hepatic stem cells) [321]. Generally, HPCs are considered capable of initiating primary liver tumours because they are typically bipotential cells (hepatoblasts) that, during liver development, differentiate into either hepatocytes or cholangiocytes. Such progenitor cells, with stemness features, are also present in the adult liver to

ensure tissue repair upon tissue injury by various damaging factors [322]. Therefore, progenitor cells can also serve as the cell-of-origin of HCCs that display progenitor-like features. However, recent studies have provided evidence that the adult hepatocyte is the cell-of-origin of HCC. This can happen in cases of acute genomic damage that result in direct degeneration of adult hepatocytes into HCC. Alternatively, adult hepatocytes can dedifferentiate into progenitor-like cells, which will eventually transform into HCC expressing genes that typically mark progenitor cells. Finally, adult hepatocytes can also transdifferentiate into biliary-like cells that can give rise to iCCA [323–326]. On another note, cholangiocytes, which display more limited plasticity, can give rise only to iCCA [327].

Overall, characterising the cell type of the cell-of-origin in our collected liver tumours, as well as its genomic, epigenomic and transcriptional state at the time point of the carcinogen insult, is important in order to understand the cell's response to the mutagenic effect of DEN that leads to neoplastic transformation and HCC progression. In addition, considering the fact that even among the different species, but also among inbred mice of the same mouse strain, we had observed some extent of variation in the selected driver gene, the contribution of the two DEN mutational signatures, as well as the latency period until the tumours develop, delineating the molecular features of the cell-of-origin and how they vary among independent tumours, would help understand the factors that contribute to perturbation of cell homeostasis, selection of driver gene and lead to tumorigenesis.

Therefore, I sought to characterise the cell type and gene expression profile of the cell-of-origin of the liver tumours by leveraging their derived bulk RNA-seq libraries. A challenge in that case was that, as the available bulk expression data represent the average expression signal from all cells of a population, they do not provide high resolution with respect to the profile of the specific cell type. Nevertheless, given that our collected liver tumours were monoclonal, they were expected to be highly homogeneous. I reasoned that we can leverage this characteristic of the bulk RNA-seq libraries and cross-map the expression data of the tumours with expression profiles of single cells from liver, so as to gain more detailed insights into the transcriptional profile of their cell-of-origin.

## 3.2 Results

### 3.2.1 Bulk RNA-seq libraries from LCE tumour samples

To gain insight into the cell-of-origin of our collected liver tumours, I used corresponding bulk RNA-seq data that were generated by Sarah J. Aitken and pre-processed by the LCE consortium (see Methods). This expression dataset included a different number of bulk RNA-seq libraries derived from independent tumours of different individuals from each mouse species (Table 3.1). The available tumour RNA-seq samples included mostly DEN-induced tumours, as well as a small number of spontaneously developed tumours (labeled as "None" for Treatment in Table 3.1). In addition, I

SPECIES	BL6	BL6	C3H	C3H	CAST	CAST	CAROLI	CAROLI
TREATMENT	DEN	None	DEN	None	DEN	None	DEN	None
liver dysplastic nodule	55	1	175	22	84	6	73	0
HCC	11	2	1	3	0	2	4	0
normal liver	12	0	0	0	0	0	0	0
tail tissue (control)	2	0	0	0	0	0	0	0

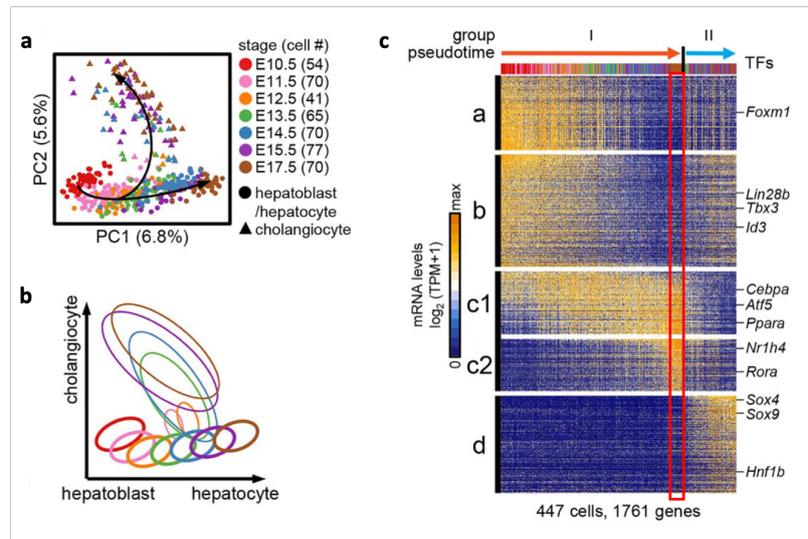
Table 3.1 **Number of available bulk RNA-seq libraries used by LCE.** These include mostly DEN-induced and to a small extent spontaneously developed (labeled as "None" for Treatment) tumours, aka dysplastic nodules and HCCs. For BL6, non-tumour liver samples treated with DEN and two tail samples to be used as controls are also available.

included in the analysis samples of normal liver tissue that was treated with DEN but was not part of tumours, which were available only for BL6. Finally, I also included two available tail samples from BL6 to be used as controls (Table 3.1). The expression levels of all expressed genes (sum(TPM) across all samples > 1) are shown in Fig. B.1, where the tail samples appear as outliers.

### 3.2.2 Single-cell expression data from hepatobiliary cell differentiation in mouse liver

To profile the expression of the bulk tumours as close as possible to cell type resolution, I mapped them onto expression profiles of single cells derived from liver. For this purpose, I used an appropriately informative published single-cell RNA-seq data from fetal mouse liver [328]. The single cells in the set had been derived from fetal livers of F1 progenies of BL6 and C3H mice, at different developmental stages (embryonic days: E10.5, E11.5, E12.5, E13.5, E14.5, E15.5, and E17.5). After fluorescent activated cell sorting (FACS), Yang and colleagues [328] had performed single-cell RNA-seq using the Smart-Seq2 platform. A total of 447 cells passed quality controls. These included undifferentiated, proliferating cells at the first embryonic stages of the sampling time course, followed by cells at intermediate embryonic stages that were either along a hepatoblast-to-hepatocyte differentiation pathway, or a hepatoblast-to-cholangiocyte differentiation pathway. Finally, cells at the latest sampling time points -the last one being E17.5- corresponded to either fully differentiated hepatocytes or differentiated cholangiocytes. The hepatoblast-to-hepatocyte transition had been temporally identified between E13.5 and E15.5, while the hepatoblast-to-cholangiocyte transition had been found to take place between E11.5 and E14.5. A principal component analysis (PCA) of all single cells by Yang et al. [328] outlined how cells from the different embryonic stages were distributed across the main hepatoblast-to-hepatocyte differentiation pathway, and the branching hepatoblast-to-cholangiocyte differentiation pathway (Fig. 3.2).

I retrieved the RNA-seq data for these 447 single cells from Gene Expression Omnibus (GEO series: GSE90047), and I quantified their gene expression following the same methods that had been used



**Fig. 3.2 Outline of hepatobiliary cell differentiation based on single-cell RNA-seq data from fetal mouse liver.** Figure adapted from [328]. a,b) PCA of single cells showing the two branching differentiation pathways from hepatoblasts at earlier embryonic stages to differentiated cholangiocytes or hepatocytes in the latest embryonic stages. c) Distinct expression profiles of marker gene clusters, clusters *a-d*, along the differentiating cells. Differentiated hepatocytes that are characterised by high expression of cluster *c* genes are marked in red frame. cluster *a*: genes related to cell proliferation and expressed in early embryonic stages, cluster *b*: genes expressed in hepatoblasts, cluster *c*: genes expressed in hepatocytes, cluster *d*: genes expressed in cholangiocytes

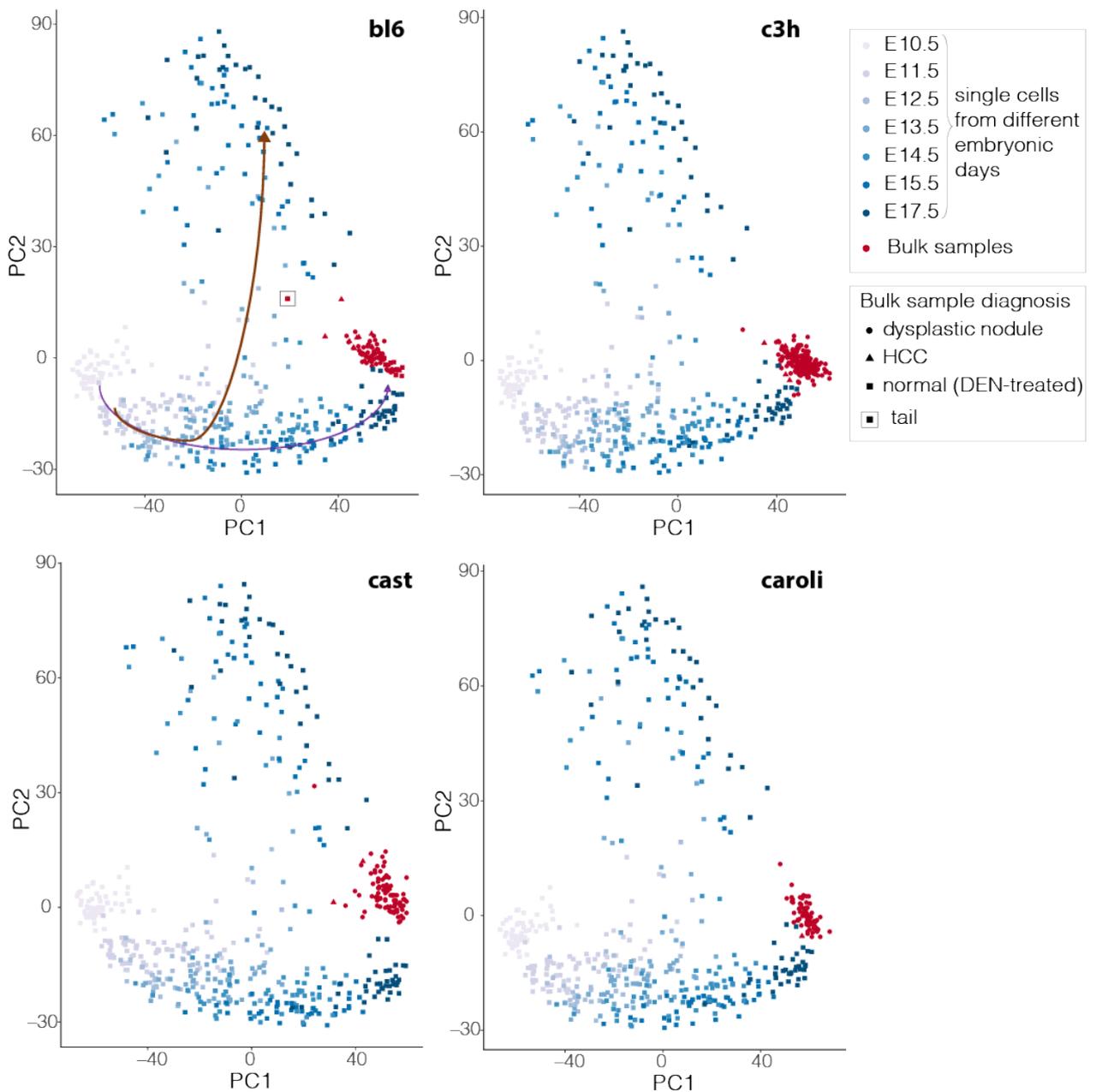
for the corresponding analysis of bulk RNA-seq samples. Specifically, I used kallisto to calculate TPM values. Visualisation of the average expression of all genes per single cell was variable (Fig. B.2a) and lower than that in bulk RNA-seq samples (Fig. B.2a, Fig. B.1). This was expected, as single-cell RNA-seq measures the distribution of expression levels of each gene across a number of single cells, while bulk RNA-seq provides an estimate of the average expression level of each gene across a population of cells.

Yang and colleagues had already identified 1,761 genes that were heterogeneously expressed among putative cell types along liver development. These genes were grouped in four clusters, with each one of them marking a distinct cell type. Gene cluster *a* marked proliferating cells in early embryonic stages, cluster *b* included genes that are typically expressed in hepatoblasts, gene cluster *c* marked differentiated hepatocytes and genes from cluster *d* marked differentiated cholangiocytes (Fig. 3.2c [328]). In further analyses, to deal with the convoluted expression signal of many thousands of genes, I used only this informative subset of genes that were heterogeneously expressed among cell types (Fig. B.2b).

### 3.2.3 Cross-mapping expression profiles of bulk LCE tumour samples and single cells from fetal mouse liver

To cross-map the bulk with the single-cell expression profiles, based on the expression of the marker gene clusters *a-d*, I performed a principal component analysis (PCA) of both the bulk LCE samples and the fetal liver single-cells. Overall, the bulk tumours, including both dysplastic nodules and HCCs, clustered at the end of the hepatoblast-to-hepatocyte differentiation pathway, closest to the differentiated hepatocytes of the latest sampling embryonic day (E17.5). This indicates that their expression profile based on cell-type marker genes (gene clusters *a-d*) resembled that of differentiated hepatocytes, rather than that of cholangiocytes, hepatoblasts, or undifferentiated proliferating cells at early developmental stages. As mentioned above, in the BL6 cohort, samples of normal liver tissue treated with DEN were also included, as well as two tail samples. The normal liver samples clustered together with the monoclonal liver tumours. Given that 70%-80% of the cells in liver are hepatocytes [329], this is not surprising, as the expression signal from the bulk normal liver samples is mostly drawn from the most abundant cell type. The tail samples, which had been kept as controls, were further from the main cluster of bulk samples (Fig. 3.3). In BL6, two more samples were slightly off of the main cluster of bulk expression profiles and closer to differentiating hepatoblasts from earlier embryonic days. These two samples corresponded to HCCs. This might reflect the onset of a more evident divergence of the cell phenotype from the hepatocyte phenotype as the tumour is progressing. In fact it seems to reflect dedifferentiation of the hepatocytes as they go through neoplastic transformation. In CAST, a few tumour samples that were slightly far off of the clustered bulk tumours corresponded to spontaneously developed tumours ("None" Treatment, Fig. B.3). This observation seems to reflect differences in the development of tumours under the effect of a chemical carcinogen insult and the spontaneously developed ones. In the former case, there is a single carcinogen hit at 15 day-old mice which causes a main mutation burst. In the latter case, the mutations accumulated in liver probably result from chronic liver damage that accumulates along the aging of mice. Different aetiologies of liver tumour development are associated with different molecular processes and are reflected in differences among mutational patterns and expression perturbations.

To better characterise the expression profiles of tumours, I inspected the expression levels of genes included in the marker gene clusters, **a-d**, in the different liver-derived bulk sample cohorts. On average, I observed clearly increased expression levels of cluster *c* genes, genes with hepatocyte related functions, compared to the genes of clusters *a*, *b* and *d* (Fig. 3.4, 3.5, 3.6, 3.7, Fig. 3.8, Fig. B.4). I also performed hierarchical clustering of the ensemble of both bulk samples and single cells using unweighted pair group method with arithmetic mean (UPGMA). All bulk liver-derived samples, including DEN-treated normal liver, in BL6, and tumours (dysplastic nodules and HCCs) clustered together and their closest single cells in the hierarchical clustering were the differentiated hepatocytes of the latest sampling embryonic day (E17.5) (Fig. 3.4, 3.5, 3.6, 3.7). The two tail samples from BL6 had a distinctively different expression profile and appeared as outgroups in the hierarchical clustering



**Fig. 3.3 Principal component analysis of expression profiles of both bulk liver samples (LCE) and single cell data from mouse liver.** Expression profiles are based on expression of marker gene clusters *a-d* [328]. The BL6 cohort includes DEN-treated normal liver samples, liver tumours (dysplastic nodules and HCCs), as well as two samples of tail tissue. The cohorts of the other species include only liver tumours. The bulk sample expression profiles mostly map on the latest stage of differentiated hepatocytes. The arrows in the first panel show the main hepatoblast-to-hepatocyte differentiation pathway (purple) and the hepatoblast-to-cholangiocyte differentiation pathway (brown). For interpretation also see Fig. 3.2a, 3.2b.

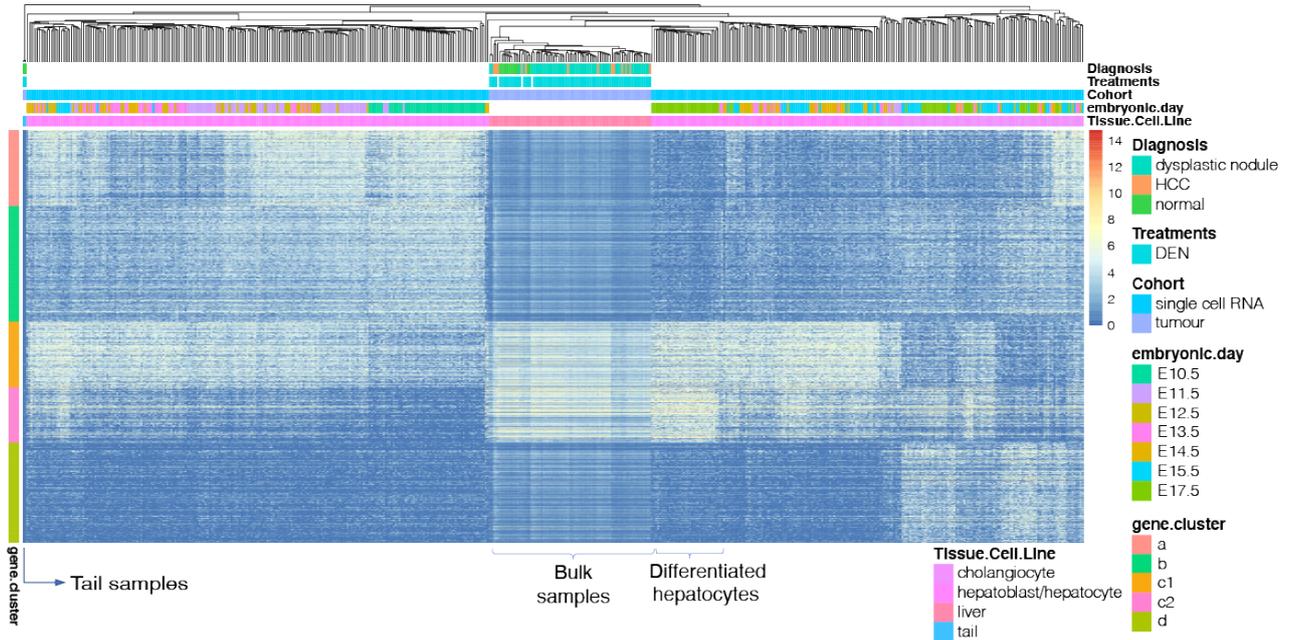


Fig. 3.4 Expression heatmap of the four marker gene clusters, *a-d*, for BL6 bulk LCE samples and single cells from fetal mouse liver. Bulk samples (including mostly monoclonal tumours) cluster with differentiated hepatocytes from the latest embryonic stage of the single cell sampling course.

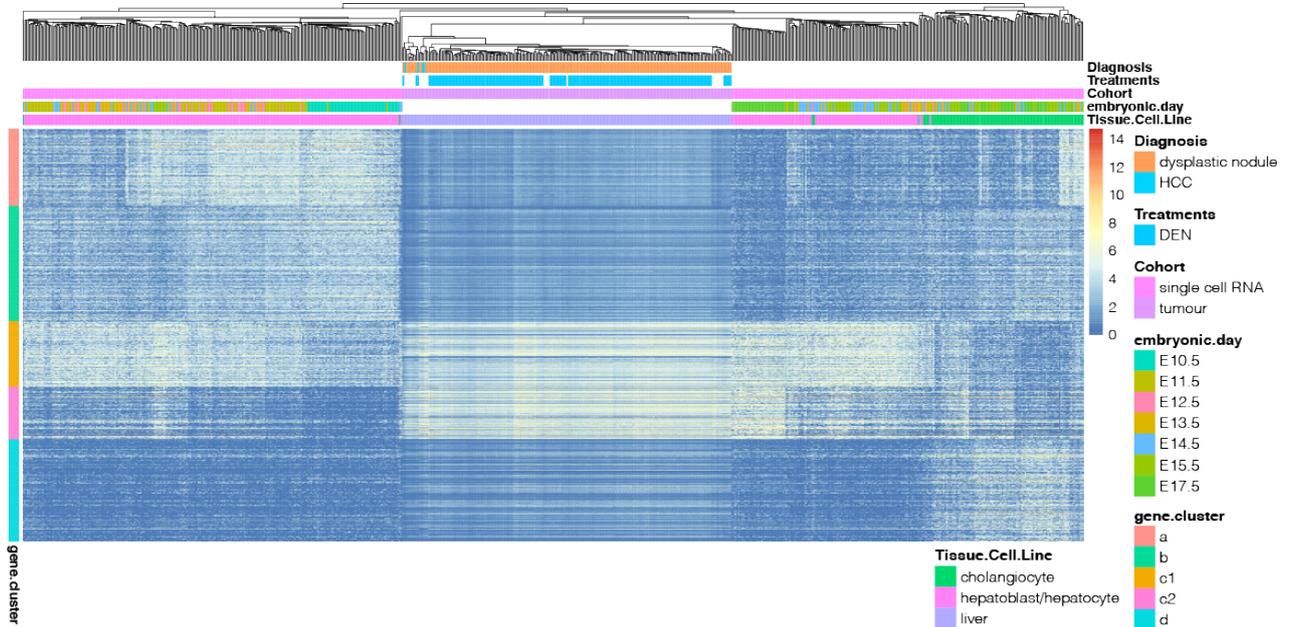


Fig. 3.5 Expression heatmap of the four marker gene clusters, *a-d*, for C3H bulk monoclonal tumours (LCE) and single cells from fetal mouse liver. Monoclonal bulk tumours cluster with differentiated hepatocytes.

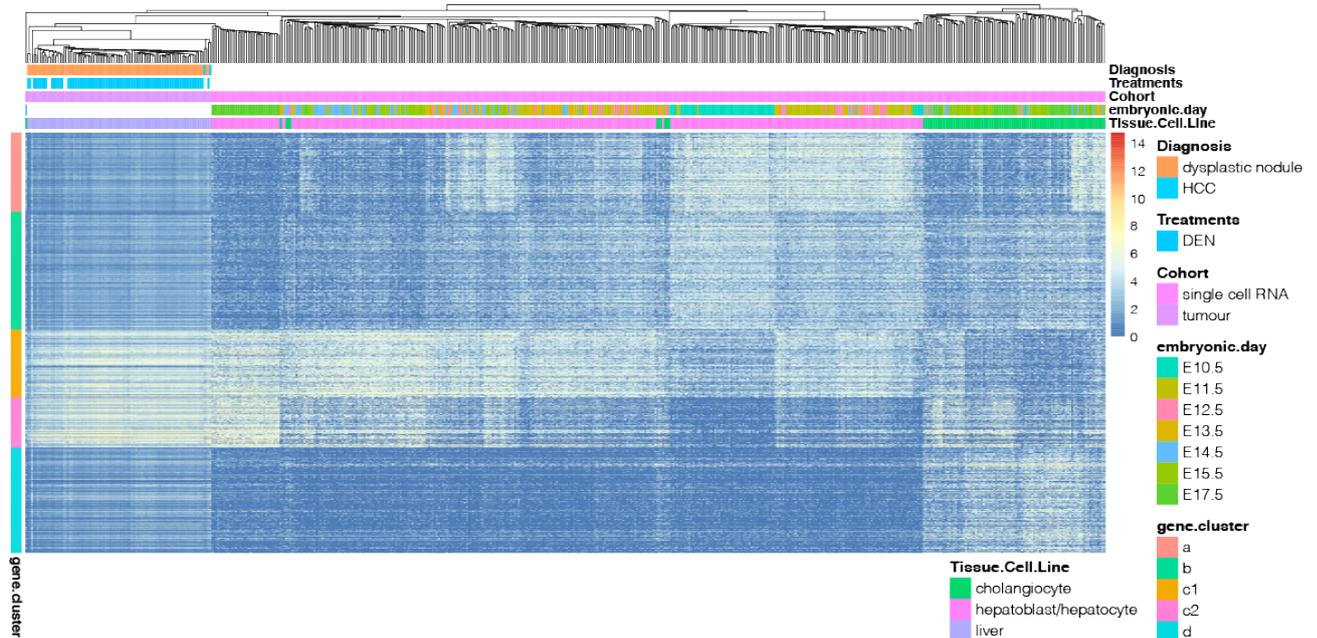


Fig. 3.6 Expression heatmap of the four marker gene clusters, *a-d*, for CAST bulk monoclonal tumours (LCE) and single cells from fetal mouse liver. Monoclonal bulk tumours cluster with differentiated hepatocytes.

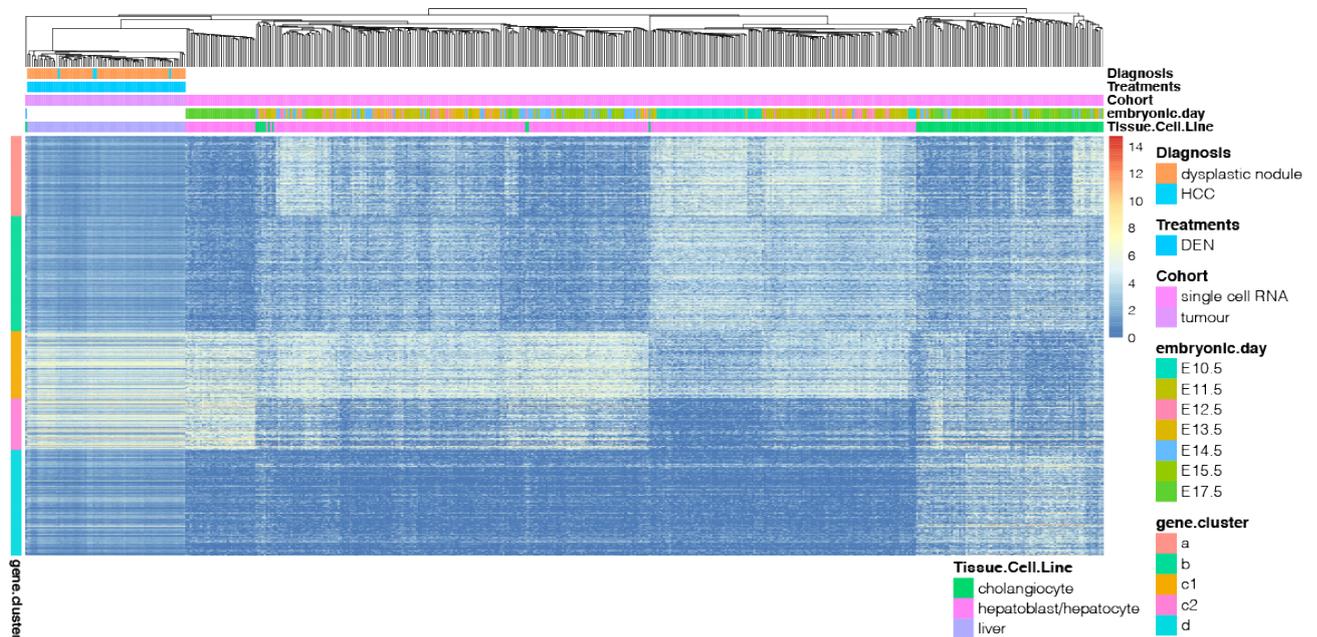
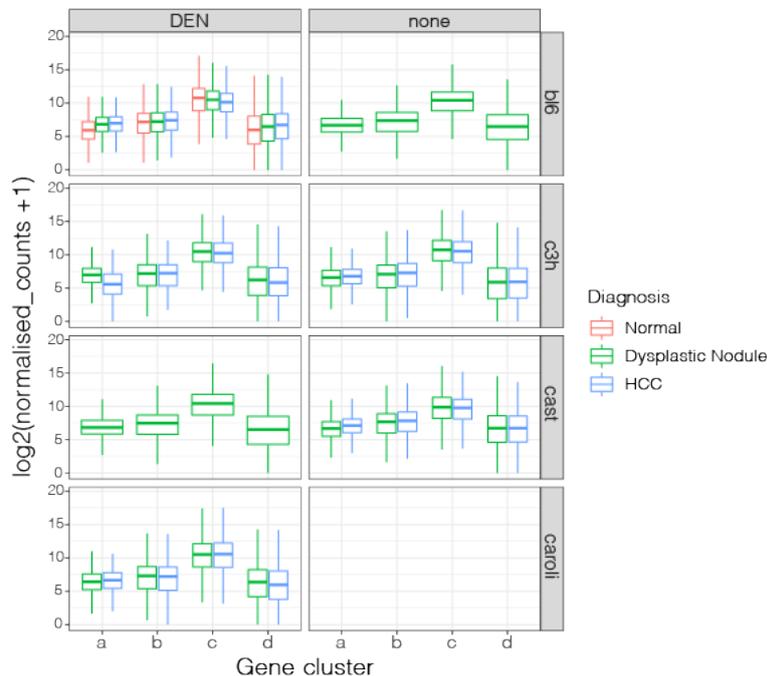


Fig. 3.7 Expression heatmap of the four marker gene clusters, *a-d*, for CAROLI bulk monoclonal tumours (LCE) and single cells from fetal mouse liver. Monoclonal bulk tumours cluster with differentiated hepatocytes.

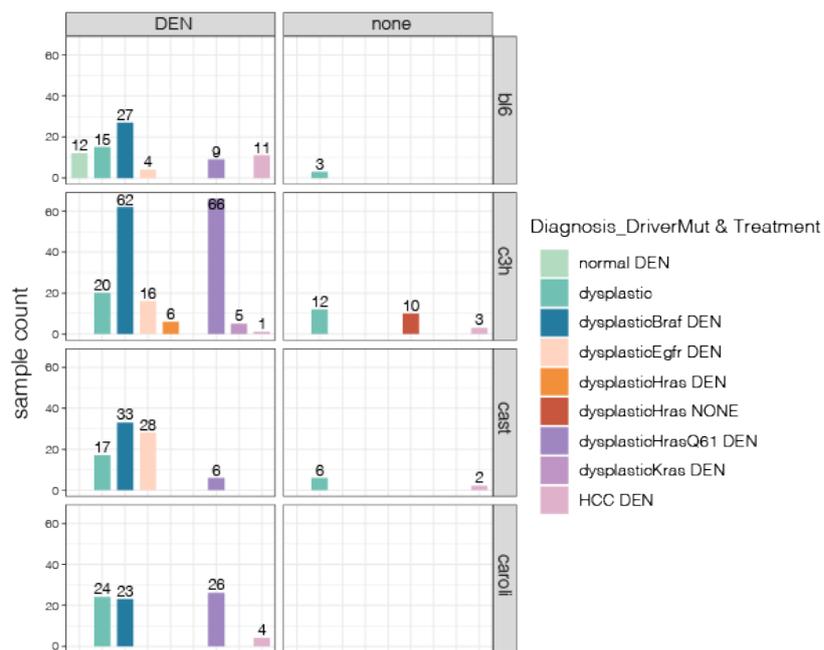
(Fig. 3.4). Another general observation within these initial exploratory analyses was that the average expression levels of cluster *c* genes (hepatocyte markers), albeit higher than the genes of clusters *a*, *b* and *d*, showed a decreasing trend along tumour progression, i.e. from dysplastic nodules to HCC in BL6 and C3H, as well as from DEN-treated normal liver to dysplastic nodules in BL6 (Fig. 3.8). It is noted that this was a general observation within initial, exploratory analyses, rather than a statistically tested and substantially confirmed trend. Yet, it triggered questions about expression perturbations of genes that typically establish and maintain the normal cell phenotype; in this case, specifically, the hepatocyte.



**Fig. 3.8 Distribution of expression levels (measured as log-transformed normalised counts) of genes contained in the marker gene clusters *a-d*, in "DEN" and "None treatment" cohort samples per species.** Raw read counts were retrieved from the kallisto output and normalised across samples and species by the LCE consortium. Samples from all cohorts display distinctly higher expression of cluster *c* genes, which are hepatocyte markers. In BL6, a drop of normalised count distribution of cluster *c* genes is outlined along tumour progression (from normal liver to dysplastic nodules and subsequently to HCCs). Distribution shifts are noticed also in the other gene clusters.

### 3.2.4 Dysregulation of cell type marker genes in HCC development

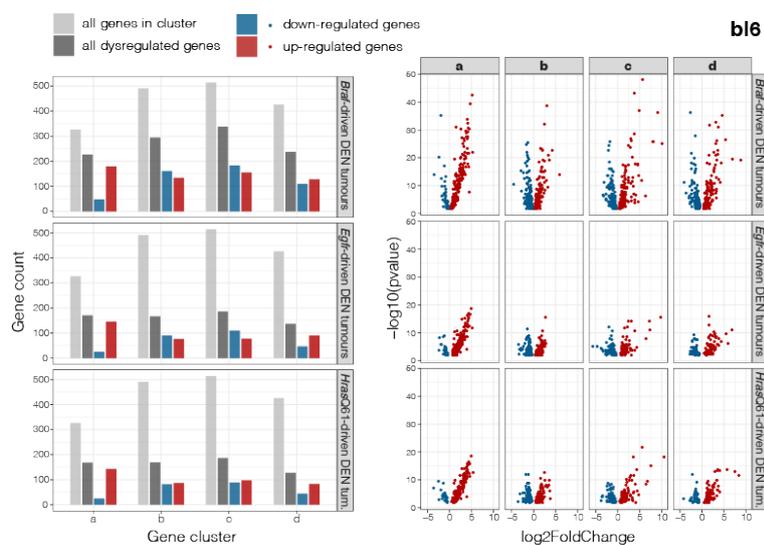
Having noticed presumed shifts in the distributions of the expression levels for different marker gene clusters, I queried whether the genes included in these clusters are significantly dysregulated in the liver-induced tumours as compared to healthy liver tissue. To address this, I utilised results of differential expression analyses that had been performed by the LCE consortium for each of the mouse species. Differential gene expression had been estimated between healthy liver tissue from adult individuals and DEN-induced tumours that had been grouped according to their identified driver mutations (Fig. 3.9). Specifically, tumour driver mutations had been found in known driver genes of the EGFR-RAF-RAS signaling pathway. Yet, a number of different known driver genes were identified among the tumours of each species, and they were usually mutually exclusive [315, 296]. The few spontaneously developed tumours (labeled as "none" for treatment in Fig. 3.9) had been excluded from the differential expression estimation, so as to keep the tumour dataset homogeneous and avoid confounding differential expression results due to potential differences between spontaneous and chemically-induced tumourigenesis processes.



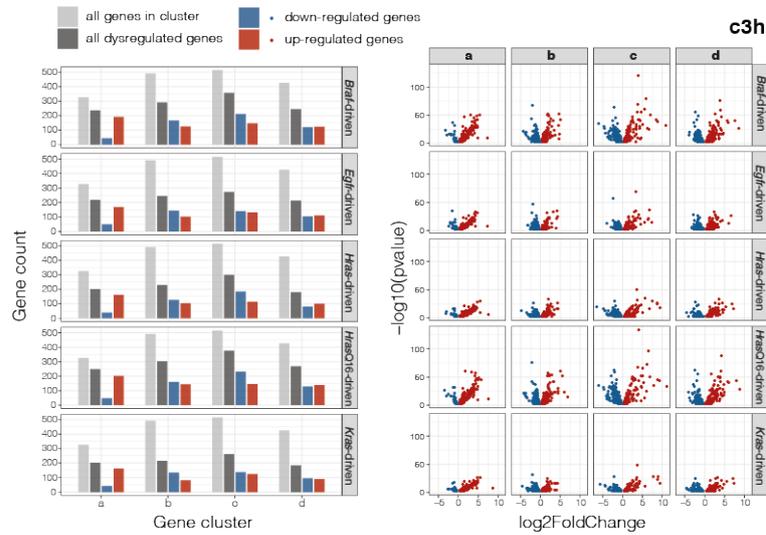
**Fig. 3.9 Number of tumour samples from each cohort with corresponding driver mutations.** The identified known driver mutations in the samples were usually mutually exclusive. "Dysplastic", without indicated driver mutation, labels samples for which no known driver mutation was identified or had a mix of potential driver mutations.

For each species, I retrieved and visualised the differential expression results for the cell type marker genes (i.e. genes from clusters *a-d*) in the distinct tumour cohorts that were driven by different mutations. Indeed in most cohorts over half of the genes from each cluster were significantly dysregulated

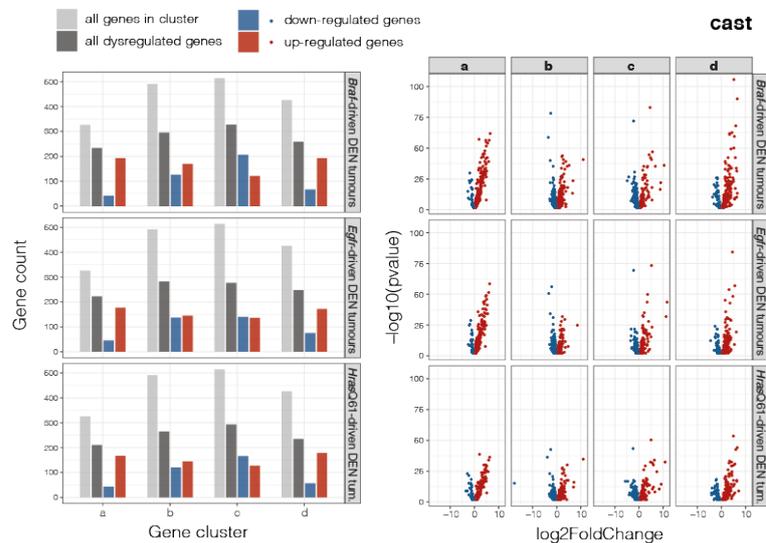
(see barplots in Fig. 3.10, 3.11, 3.12, 3.13). The fractions of dysregulated genes from the marker gene clusters *a*, *b*, *c* and *d*, corresponded, respectively, to 51-67%, 33-63%, 36-73% and 29-64% for the different tumour cohorts (Fig. 3.10, 3.11, 3.12, 3.13) as compared to normal liver. Thus, overall, there seemed to be consistent dysregulation of genes that typically establish and maintain the phenotype of the specific cell type, in this case the hepatocyte phenotype, along tumour development. This type of dysregulation seems to underlie the onset of the cell phenotype shifting from the differentiated hepatocyte phenotype towards a more dedifferentiated state, a phenomenon that is generally observed in tumour progression.



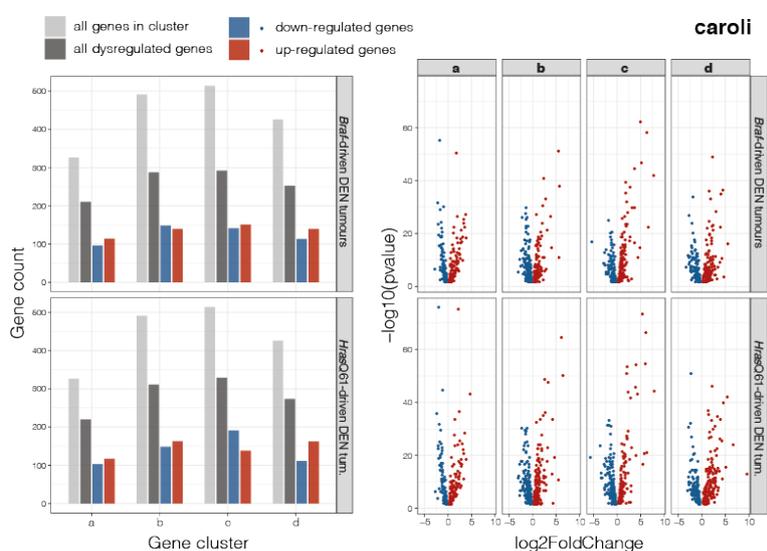
**Fig. 3.10 Dysregulation of cell type marker genes in BL6 DEN-induced tumours with different identified driver mutations.** Left: Number of significantly dysregulated genes (up- and down-regulated) from each cell-type marker gene cluster (clusters *a-d*) for BL6. Right: Volcano plots of the corresponding, significantly up- and down-regulated genes.



**Fig. 3.11 Dysregulation of cell type marker genes in C3H DEN-induced tumours with different identified driver mutations.** Left: Number of significantly dysregulated genes (up- and down-regulated) from each cell-type marker gene cluster (clusters *a-d*) for C3H. Right: Volcano plots of the corresponding, significantly up- and down-regulated genes.



**Fig. 3.12 Dysregulation of cell type marker genes in CAST DEN-induced tumours with different identified driver mutations.** Left: Number of significantly dysregulated genes (up- and down-regulated) from each cell-type marker gene cluster (clusters *a-d*) for CAST. Right: Volcano plots of the corresponding, significantly up- and down-regulated genes.



**Fig. 3.13 Dysregulation of cell type marker genes in CAROLI DEN-induced tumours with different identified driver mutations.** Left: Number of significantly dysregulated genes (up- and down-regulated) from each cell-type marker gene cluster (clusters *a-d*) for CAROLI. Right: Volcano plots of the corresponding, significantly up- and down-regulated genes.

### 3.3 Discussion

As part of first exploratory analyses within the LCE project, I have characterised the expression profile of cell-type marker genes in the collected monoclonal liver tumours of the different species, by cross-mapping these profiles with analogous ones of single cells from different hepatobiliary cell differentiation stages. The expression profiles of the monoclonal tumours of all species match with those of differentiated hepatocytes, providing further evidence that the HCC cell-of-origin in our chemical carcinogenesis model is the hepatocyte. Considering that DEN is bioactivated by cytochrome P450 enzymes (CYP2E1) in centrilobular hepatocytes [330], it is expected that hepatocytes will be particularly burdened with DEN-caused lesions, many of which will be fixed as somatic mutations. It is therefore reasonable to establish that, upon a single large mutagenic burst due to DEN exposure, it is a hepatocyte that acquires a selective advantage, undergoes neoplastic transformation and gives rise to a DEN-induced clonally expanding tumour in the studied mouse species. A number of studies have considered the impaired replicative capacity of hepatocytes in chronic liver diseases or massive liver injury, in human as well as in mouse liver injury models, and have supported a main contribution of bipotential progenitor cells in initiating liver tumorigenesis [331–336]. Nevertheless, these claims have been controversial [337, 338]. On the contrary, recent studies indicate the hepatocyte as the HCC cell-of-origin [339–342, 326]. A number of them have used fate-tracing experiments to show that HCC originates from hepatocytes in chemically induced carcinogenesis in mice [340, 341], as well as in cases of models that do not implicate carcinogen treatments [341, 339]. In addition, the potential of hepatocytes to give rise to tumour development is supported by their remarkable plasticity. It is known that mature hepatocytes have the capacity to dedifferentiate into a cell state with progenitor-like features and replenish hepatocytes upon liver injury [325]. It has been shown that their dedifferentiation can take place upon loss-of-function mutations in the WNT or NOTCH pathways that might silence the tumour suppressor gene *Tp53* [343]. On another note, mature hepatocytes can also transdifferentiate into biliary-like cells and subsequently give rise to iCCA [324]. Mu et al. provide evidence that HCC originates from hepatocytes and any observed progenitor-like molecular features in HCC are attributed to acquired features of transformed hepatocytes, rather than to initiation of HCC from hepatic progenitor cells [339]. Our collected evidence based on expression profiles of tumours supports that the DEN-induced liver tumours, including dysplastic nodules and HCCs, in the mouse strains originate from hepatocytes.

However, despite the clearly prominent expression signal from genes related to hepatocyte functions, as opposed to the relatively lower expression levels of genes related to cell proliferation in early embryonic stages, hepatoblast or cholangiocyte functions, a substantial fraction of genes from all cell-type marker gene clusters show consistent significant dysregulation in tumours with different driver mutations and in all four species in the LCE study. This may reflect the onset of hepatocyte transformation, which drives the phenotype of the cell to diverge from the normal hepatocyte phenotype to a dysplastic phenotype. It is generally important to understand what aberrations of molecular

processes can take place throughout the phenotypic shift, destabilise homeostasis and increase the risk of developing liver cancer and/or the potential of malignant transformation. Such aberrations may be downstream of the oncogenic driver mutation hit, yet have the capacity to further perturb the cellular environment and enhance the growth advantage of the cell, eventually making it more resistant to cancer therapies. Shift of a mature cell phenotype to a less differentiated (or dedifferentiated) cell type is generally associated with tumour progression, aggressiveness and resistance to medical treatments [344–346]. In addition, emerging evidence shows that loss of cell differentiation along tumourigenesis is associated with the capacity of the transforming tumour cells to evade recognition and destruction by the immune system (reviewed in [347]).

Taken together, this evidence highlights that understanding and dealing with mechanisms of liver carcinogenesis requires a profound molecular characterisation of the phenotypic shift from a differentiated cell type towards a dedifferentiated state. This phenotypic shift is associated with gene expression alterations in the transforming hepatocyte, which in turn must be underlined by corresponding changes in the regulatory programme of the cell. Therefore, to molecularly characterise the hepatocyte neoplastic transformation we would need to determine the relevant changes in gene regulatory mechanisms.

Understanding the molecular changes that take place within the context of a specific cell-type environment governed by a specific regulatory programme, such as that of the hepatocyte in HCC development, is also relevant with respect to elucidating the conditions and factors that determine the choice of the cancer driver gene. It is known that mutations in driver genes exhibit a spectrum of remarkable specificity in distinct cancer types. Only few driver mutations, such as in *TP53* or *TERT*, show a broad distribution in a range of cancer types. In contrast, most driver mutations promote cancer development in specific tissues. For instance, chronic myeloid leukemia is strikingly associated with the *BCR-ABL* gene fusion mutation, while retinoblastoma is usually driven by loss-of-function mutations of particular tumour suppressor genes [348]. Other cancer types appear to be frequently linked with recurrent mutations in a set of genes involved in particular signalling pathways, such as the EGFR-RAS-RAF pathway that is also perturbed in cases of HCC or lung adenocarcinoma [349] (reviewed in [350]). Another example is the case of mutations in *BRCA1* and *BRCA2* that predispose specifically for breast and ovarian cancer, even though these genes are ubiquitously expressed in a range of tissues [351]. In most cases, the underlying molecular mechanisms of tissue-specificity for distinct sets of driver genes remain elusive. A prevailing hypothesis is that tissue-specificity of driver mutations is associated with particular aspects of the distinct regulatory programme that is employed by each tissue in order to establish and maintain the differentiated cell phenotype and suppress the pluripotent, stem cell gene expression repertoire that can be implicated in tumourigenesis (reviewed in [352]). This is another aspect that highlights the importance of characterising regulatory state in the tumour cell-of-origin, monitoring how it might change throughout tumourigenesis and outline the susceptibilities to cancer initiation this might confer to the cell. However, hereafter I will focus

particularly on aspects of regulatory changes through the angle of identifying potential implications in the shift of the cell phenotype along HCC development.

## 3.4 Methods

### 3.4.1 Expression quantification in LCE bulk samples and in single cells

Raw RNA-seq data from bulk samples were pre-processed by the LCE consortium. Specifically, transcriptome indices were generated from cDNA sequences in each species (Ensembl 91) and transcripts per million (TPM values) were calculated for each transcript isoform using kallisto (v0.43.1) [248] with the options *plaintext*, *bias*, and *pseudobam*.

To quantify gene expression in the single cell RNA-seq libraries (GEO series: GSE90047) [328], I followed the same approach used for expression quantification in the bulk RNA-seq libraries by the LCE. used kallisto v0.43.1 [248]. First, I used kallisto index to generate a transcriptome index of cDNA sequences in each species (Ensembl 91) also including 92 ERCC spike-in sequences that were used in the RNA sequencing of the single cells. To calculate TPM values for each transcript isoform, I used kallisto quant with the options *plaintext*, *bias*, *single*, *pseudobam*, *fragment-length = 350* and *sd = 50*. Finally, to calculate TPM values at gene level, I used tximport (v1.10.1) [353]. To enable comparisons between the single-cells and bulk tumours, I also calculated TPMs at gene levels for the bulk samples, using tximport (v1.10.1).

### 3.4.2 Identification of driver mutations in tumours

Driver mutations were identified in each tumour by the LCE consortium using OncodriveFML (v.2.2.0) [354] and OncodriveCLUSTL (v.1.1.1) [355].

### 3.4.3 Differential gene expression analyses

Differential gene expression analyses were performed by the LCE consortium between each cohort of DEN-induced tumours with a specified driver mutation and normal liver samples. Differential expression for coding genes was then called using the Wald test from DESeq2 [250] and raw count values that had been calculated by kallisto and aggregated to the gene level by tximport.



# Chapter 4

## **Mutational landscapes of distinct transcription factor binding region categories and functional implications in chemically induced liver tumours**

This chapter describes my work on characterising DEN-caused mutational patterns in distinct groups of liver TF binding regions, and assessing their functional implications in liver cancer development. It follows up from the expression profiling of the tumour cell-of-origin discussed in the previous chapter. To perform the analyses described below, I have used datasets from two different sources: a) mutations in DEN-induced tumours from three mouse species, which were called by the LCE consortium using WGS data from the tumours, and b) raw, published ChIP-seq data for three liver TFs in the same mouse species, which I analysed. Except as noted, all of the analyses described below were carried out by me.

### **4.1 Introduction**

In the previous chapter, I profiled the expression of cell-type marker genes in liver tumours with different driver mutations, in all four mouse species used in the LCE project. Despite the fact that monoclonal tumours showed an expression profile that matched that of a hepatocyte, there was observed consistent dysregulation of hundreds of genes that typically associate with establishment and maintenance of the hepatocyte-specific phenotype. These differential expression patterns manifest a dedifferentiation process the hepatocyte undergoes as it transforms to HCC. Understanding and dealing with liver carcinogenesis requires a profound characterisation of the molecular changes that are implicated in disruption of the cellular equilibrium and loss of the cell phenotype specificity. Given that establishment and maintenance of the cell phenotype depends to a great extent on cell-type

specific regulatory networks, it is possible that gene expression shifts associated with regulatory changes may contribute to the observed shift of the cellular phenotype. Addressing the cause of gene dysregulation in HCC development requires to determine how the gene regulatory programme changes in the transforming hepatocyte.

It is known that important players in establishing gene regulatory networks in distinct cell types are the transcription factors (TFs) expressed and the genomic sites where they bind and exert their regulatory function. Therefore, among other factors, it may be useful to identify changes in TF activity with respect to hepatocyte gene dysregulation along neoplastic transformation. Modifications of TF activity can encompass changes in at least one of two factors: the expression levels of TF genes themselves, and the binding of TFs to their corresponding binding sites that, in turn, affects regulation of the TF target genes. Given that, in general, tumourigenesis is to a great extent a consequence of somatic mutations, tumour-associated changes in regulatory mechanisms, including TF binding, must be largely a result of somatic mutations. Mutations that affect TF activity may be in TF-coding genes or in genes whose products interact with TFs. Alternatively, they can be in binding sites of TFs, having an effect on the binding affinity.

Although mutational processes in protein-coding genes and their oncogenic effect have been extensively studied, little is known about the implication of non-coding mutations—especially within TF binding regions—in cancer development. The best known examples are mutations in the promoter of the *TERT* oncogene that create novel binding sites for ETS TFs leading to upregulation of TERT [356, 357]. Similarly, it has been shown that creation of a new TF binding site driven by non-coding mutations leads to overexpression of the *TALI* oncogene reportedly via regulation by a super-enhancer [358]. Despite the elucidation of this handful of non-coding driver examples, non-coding mutations remain largely uncharacterised. Yet, the majority of mutations in cancer occur in non-coding regions [359]. Most of them are considered passenger mutations, i.e. not directly driving oncogenesis. However, there is growing evidence that putative passenger mutations have an aggregated contribution to the whole mutational landscape in a cancer genome, and they can have a collective impact on molecular functions [360, 361]. It is also suggested that some putative passengers may have a weak effect on the fitness of cancerous cells, thus contributing to or impeding tumour growth [361, 362]. Recent studies further suggest that putative passengers, even if seemingly not implicated in tumour initiation, can confer fitness advantages to the tumour cells at subsequent stages of tumour progression [363, 360]. In addition, it was shown that the collective impact of putative passenger mutations can be used as a predictor to distinguish between healthy and cancer phenotypes, especially in tumours with no known driver mutations [360]. Yet, it is acknowledged that small sets of supposedly passenger mutations may, in fact, represent changes with driver functions [360].

With respect to mutations in TF binding sites, in a recent study, partitioning the genome into cistromes<sup>1</sup>, showed convergence of known risk single nucleotide variants (SNVs) and somatic SNVs in the cistromes of the TFs AR, FOXA1 and HOXB13 that act as master regulators in prostate tumour development. In addition, some of these mutations were shown to change the expression output from the associated regulatory elements [82]. Moreover, in a complementary study by the same group, inspection of the regulatory plexus of *FOXA1*<sup>2</sup> resulted in identification of a set of mutations in six CREs that regulate *FOXA1* expression. These mutations were shown to have an impact on the transactivation potential of the respective CREs [81].

Motivated by this background and considering that DEN-induced liver tumour development is very much driven by acute DEN mutagenic effect, we reasoned that HCC-associated expression changes may be attributed to somatic regulatory mutations, including mutations in TF binding. We therefore sought to investigate potential associations between TF binding activity alterations and genomic mutations in the DEN-induced liver tumours. As mutations in genic sequences and detailed gene expression analyses—including TF-coding genes—are being investigated extensively by other consortium members, I focused on investigating mutagenesis in TF binding regions in the DEN-induced tumours. To this end, I utilised the mutations identified in WGS data from the collected LCE tumours, as well as previously published CHIP-seq data for three TFs expressed in liver, namely CCAAT/enhancer-binding protein alpha (CEBPA), forkhead box A1 (FOXA1), and hepatocyte nuclear factor 4 alpha (HNF4A) from three of the mouse species studied in the LCE project, i.e. BL6, CAST and CAROLI [78]. By combined analyses of the CHIP-seq and the tumour mutation datasets in the three mouse species, I characterised the mutational profiles of the binding region ensembles, or cistromes, of the corresponding liver TFs. I further evaluated the relative mutation rates among the three TF cistromes in each species, as well as among functional sub-categories of cistromes after splitting them based on a) their potential of combinatorial TF binding and b) their association with *cis*-regulatory elements. The results revealed different mutation rates among distinct sets of TF binding regions, potentially reflecting differential effects of purifying selection according to the functional importance of the corresponding regions. In addition, the analyses revealed a set of hyper mutated TF binding regions, largely in *cis*-regulatory elements, that are associated with dysregulation of genes driving the cell phenotype. The mutational and gene dysregulation patterns were largely reproducible among the three mouse species. These results exemplify exploration of non-coding mutation sets with functional implications in liver tumour development.

---

<sup>1</sup>A cistrome is defined as "the set of *cis*-acting targets of a *trans*-acting factor on a genome-wide scale, also known as the *in vivo* genome-wide location of transcription factor binding sites, or histone modifications" [364]

<sup>2</sup>The plexus of a gene is defined as its "cell-type-specific three-dimensional gene-regulatory neighborhood, inferred using Hi-C chromosomal interactions and chromatin state annotations" [365]

## 4.2 Results

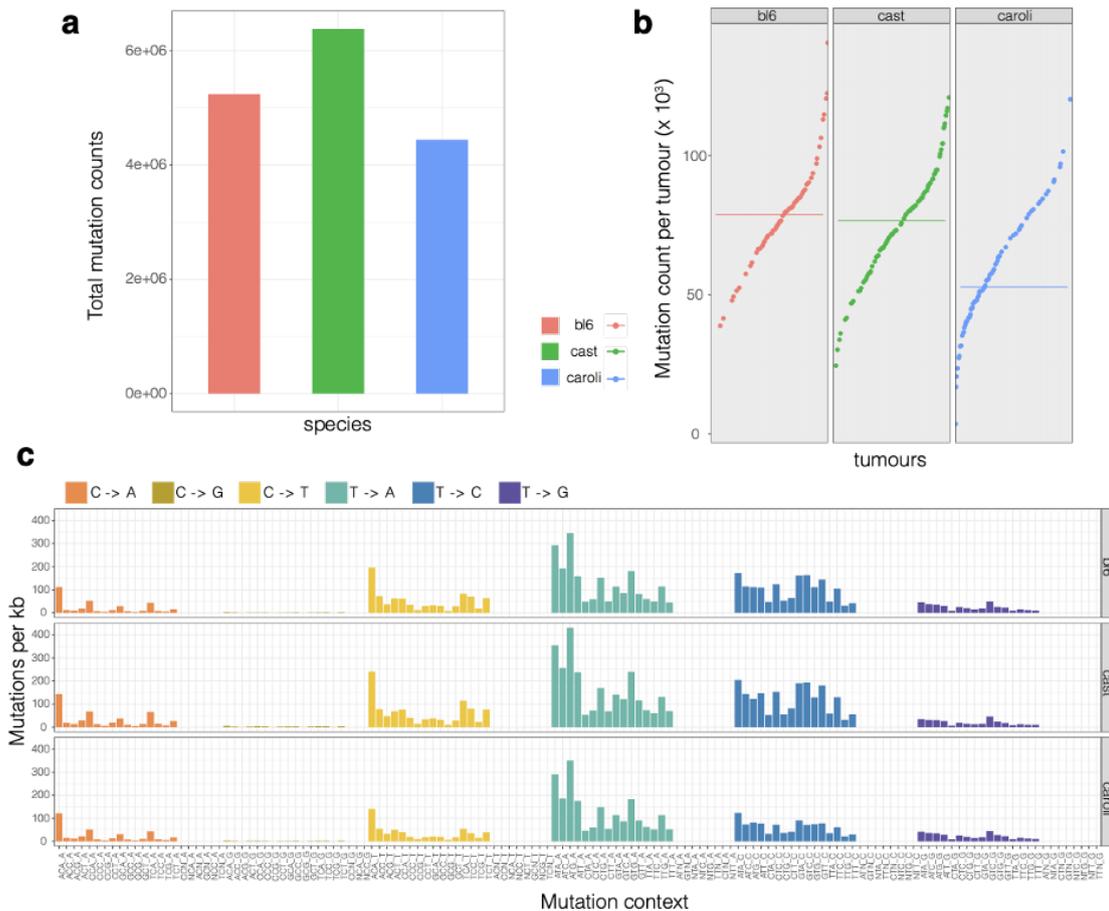
### 4.2.1 DEN tumour mutation datasets: abundance of T → N mutations in all three mouse species

The mutation data I used in this study had been generated, processed and quality filtered by the LCE consortium. I retrieved the mutation datasets for the BL6, CAST and CAROLI liver tumour cohorts, and selected only the Single Nucleotide Variants (SNVs)—excluding other types of mutations—that were identified in DEN-induced tumours. Therefore, hereafter, the use of the term “mutations” refers to SNVs. Mutation data from C3H were not considered, because there were no available TF ChIP-seq for this strain. My selected mutation dataset included in total 66, 84 and 79 DEN-induced liver tumours from BL6, CAST and CAROLI, respectively (Table 4.1 - Appendix C). The total number of mutations from all the tumour samples of each species were  $\sim 5.2 \times 10^6$ ,  $\sim 6.4 \times 10^6$ ,  $\sim 4.4 \times 10^6$ , with a median number of mutations per tumour being approximately  $78.8 \times 10^3$ ,  $76.6 \times 10^3$ ,  $52.8 \times 10^3$  for BL6, CAST and CAROLI respectively (Fig. 4.1a, Fig. 4.1b, Table 4.1). The higher total number of mutations in CAST than in BL6 (Fig. 4.1a) seemed to reflect the higher number of available CAST tumour samples. However, the median mutation count per tumour (Fig. 4.1b) was very similar between BL6 and CAST. In contrast, the median number of mutations per tumour was slightly lower in CAROLI, even when accounting also for genome sizes. Whether this is associated with the longer latency period between the DEN insult and liver tumour development in CAROLI (as mentioned in chapter 3) is not clear based on the current analyses. Yet it provides a future research direction for the study.

	<b>BL6</b>	<b>CAST</b>	<b>CAROLI</b>
Number of tumour samples	66	84	79
Total mutation counts from all tumours	5,240,706	6,381,768	4,441,825
Median mutation count per tumour	78,843	76,596	52,819
Median mut. count per tumour divided by genome size	2.99e-05	2.90e-05	2.11e-05

Table 4.1 Counts of DEN-induced tumour samples, aggregated mutations from all tumour samples, and median mutation count per sample in each species.

To get an overview of the most frequent mutations in the DEN tumours of each species and how they compare to each other, I calculated and visualised the corresponding mutation spectra. These showed a similar mutation distribution pattern among the three species, with thymine-to-adenine (T → A) mutations being the most frequent, followed by thymine-to-cytosine (T → C) mutations, and subsequently by cytosine-to-thymine (C → T) and cytosine-to-adenine (C → A) mutations (Fig. 4.1c). A high frequency of mutated T’s was generally expected, as it is known that DEN mostly



**Fig. 4.1 Mutational profiles of DEN-induced tumours in each mouse species.** a) Total number of mutations identified in all DEN tumour samples. b) Mutation counts (per kb) per DEN tumour sample, ranked over  $x$  axis. Lines mark median values. Number of tumour samples (data points) per species shown in Table 4.1. c) Mutational spectra from the aggregated mutations of all DEN tumour samples in each species.

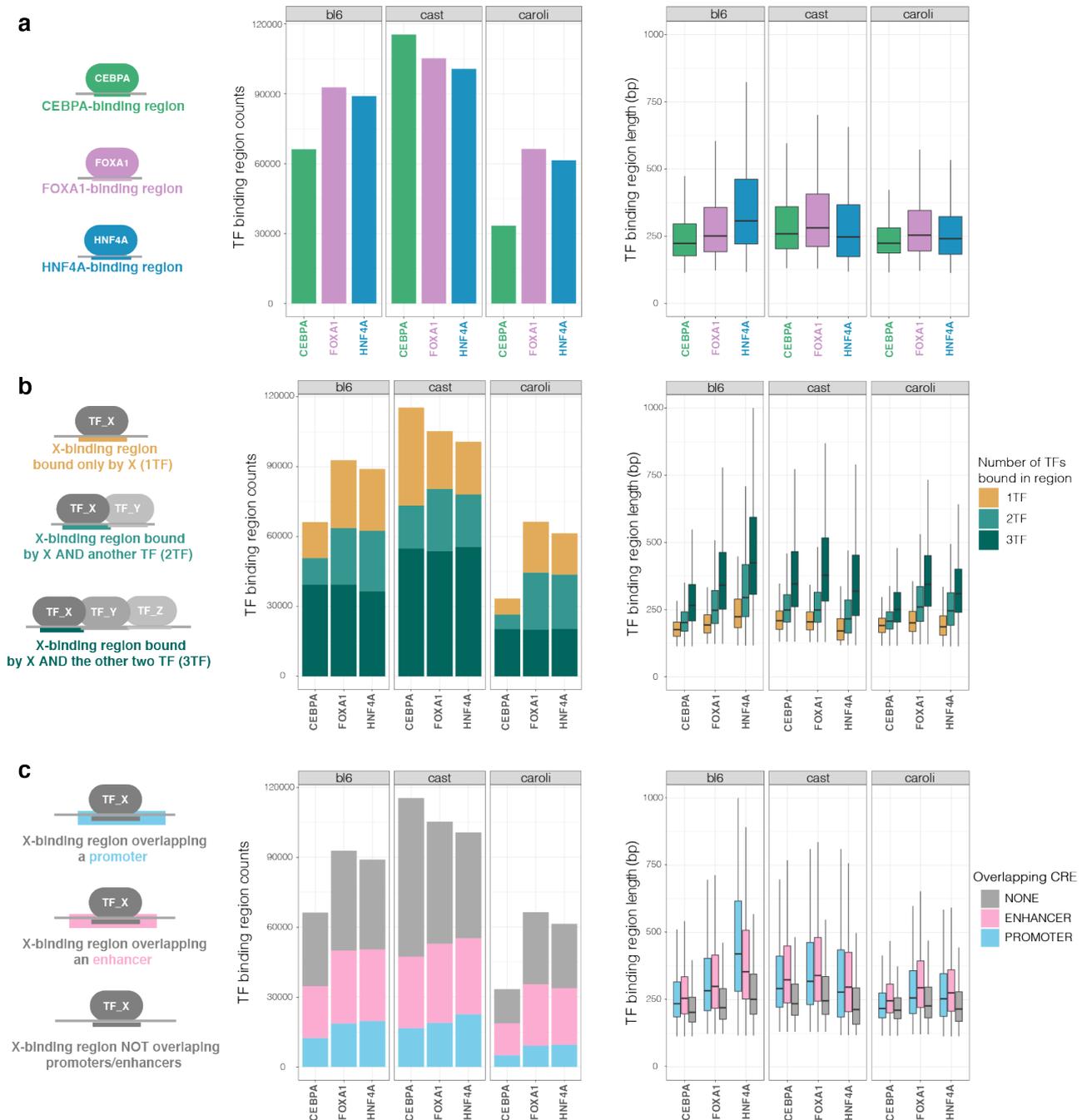
causes lesions on thymines. Specifically, formation of a  $O_4$ -ethyl-deoxythymidine adduct on the DNA thymines is the most common effect of DEN bioactivation [316]. That often leads to fixed thymine mutations ( $T \rightarrow N$ ), or the complement  $A \rightarrow N$  mutations, upon translesion replication or translesion transcription of the damaged DNA strand segment. This is a result of impaired recognition of the template nucleotide by the DNA replication machinery when encountering a nucleotide with a lesion, which can lead to a nucleotide mismatch in the newly synthesized DNA strand. Similarly, the second most frequent mutations, i.e.  $C \rightarrow N$ , or their complementary guanine mutations ( $G \rightarrow N$ ), can also be explained as a result of the formation of  $O_6$ -ethyl-2-deoxyguanosine lesions on guanines upon DEN bioactivation [316]. The lack of striking differences in the mutation spectra among the mouse strains is not surprising, as the spectra of nucleotide substitutions basically reveal the DEN mutational signature. In our previous study of DEN-induced tumours in mouse strains, a detailed inspection of

mutational signatures was carried out by the LCE consortium, in C3H and in CAST derived liver tumours. This revealed the presence of a primary DEN mutational signature (DEN1), predominated by  $T \rightarrow N$  mutations (or the complementary  $A \rightarrow N$ ), mostly  $T \rightarrow C$  and  $T \rightarrow A$ , as well as a secondary mutational signature (DEN2), predominated by  $C \rightarrow T$  mutations (or the complementary  $G \rightarrow A$ ). DEN1 was clearly more prominent in the DEN tumours, while DEN2 had a minor contribution to the mutational landscape of most tumours (Fig. C.1 - Appendix C) [296]. The overview of mutation spectra presented here seems to reflect the predominant contribution of the DEN1 signature - as inferred by the abundance of  $T \rightarrow N$  mutations- and a minor contribution of the DEN2 signature in all three species (Fig. 4.1c).

## 4.2.2 Cistromes of liver TFs and categorisation into sub-cistromes based on combinatorial TF binding and CRE co-occurrence

As mentioned in the introduction, in this study I also used published ChIP-seq datasets for three liver TFs, namely CEBPA, FOXA1 and HNF4A, in BL6, CAST and CAROLI [78]. I retrieved and reanalysed the raw ChIP-seq data to determine the genome wide binding profiles of the three TFs of interest, CEBPA, FOXA1 and HNF4A. Specifically, I mapped the raw ChIP-seq reads on the updated genome assemblies used by the LCE consortium and identified TF binding regions (TFBRs), hereby also referred to as TF peaks meaning the peaks of piled mapped reads at the bound sites showing up in the computational analysis (see details in Methods). Hereafter, the terms TF peaks and TF binding regions are used interchangeably. In all species, tens of thousands of binding regions were identified for each TF, exceeding a hundred thousand regions in CAST (Fig. 4.2a, Table C.1 - Appendix C). Small differences in the number of identified peaks among the species are possibly attributed to experimental or technical differences such as antibody specificity or genome assembly quality. Overall, there were fewer peaks identified for all TFs in CAROLI, especially for CEBPA, compared to the other species. The low peak number specifically for CEBPA seems to be partly due to very poor read quality in one of the CEBPA ChIP-seq libraries that were used. Specifically, the read mapping rate for this low quality library was  $< 3\%$ , strikingly lower than the average mapping rate of  $\sim 80\%$  that was observed for all the other ChIP-seq libraries. At the same time, this poor quality CEBPA library from CAROLI had a much higher sequencing depth ( $\sim 71$  million reads), compared to the other libraries ( $\sim 27$  million reads on average).

A major focus of my work has been on evaluating the effect of combinatorial binding of distinct TFs on the mutational patterns of TFBRs. Thus, I also characterised the TFBRs contained in each TF cistrome based on whether they present binding by only one of the three studied TFs (i.e. CEBPA, FOXA1 or HNF4A binding alone), a combination of two of the TFs, or a combination of all three TFs. The corresponding groups were respectively labelled as “1TF”, “2TF”, or “3TF” groups (Fig. 4.2b). The criterion used to assign a binding region of a given TF (as identified by the ChIP-seq enrichment signal) to one of the three groups was whether this region overlaps with any binding region of the



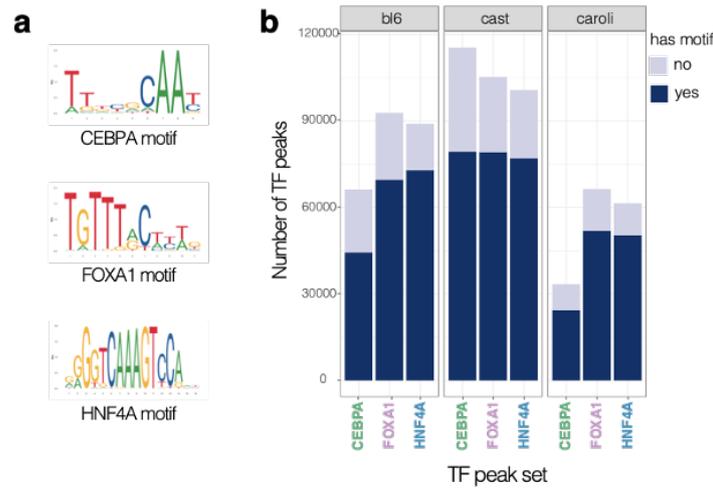
**Fig. 4.2 Identification and characterisation of genome-wide identified binding regions for each of the three liver TFs (CEBPA, FOXA1 and HNF4A) in three mouse species (BL6, CAST and CAROLI).** The first column schematically shows the grouping of TFBRs, the second the number of peaks in each category, and the third the length distribution of the corresponding TFBRs in each (sub-)cistrome, at different levels of classification: a) TFBRs per cistrome, b) TFBRs of each cistrome grouped in categories based on the total number of liver TFs bound to the underlying genomic region. TF\_X, TF\_Y and TF\_Z represent one of CEBPA, FOXA1 and HNF4A. c) Grouping of the TFBRs of each cistrome based on their overlap with *cis*-regulatory elements.

other TFs (Fig. 4.2b). This type of grouping revealed that the majority of the identified TFBRs have combinatorial binding of either two or three of the TFs, while the regions that are bound by one TF alone were relatively fewer, in line with previous studies [366].

TFs are known to often bind within *cis*-regulatory elements (CREs), such as promoters and enhancers. To investigate this functional feature of TFs, I also categorised the binding regions of each of the three TFs based on whether they overlap with any promoter, enhancer, or none of these CREs (Fig. 4.2c). To assign the TFBRs in one of the three groups, I intersected each one of them with promoters and enhancers identified in livers of healthy BL6, CAST and CAROLI mice by the LCE consortium. Promoters and enhancers had been determined as genomic regions marked by promoter-associated or enhancer-associated histone modifications, via analysis of corresponding ChIP-seq data (see Methods). Promoter-overlapping TFBRs represented the smallest of the three categories, corresponding to fractions ~14-22% of each cistrome, showing small variety among the cistromes of the different TFs and species. The following smaller category was the enhancer-overlapping TFBRs, representing 26-40% of the total TF binding regions per cistrome, while the rest of the TF binding regions were not found to overlap with any of these CREs (Fig. 4.2c). The increased fraction of TFBRs overlapping with enhancers, as compared to overlap with promoters, seems to reflect the higher total number of identified enhancers in the mouse genomes, compared to the total count of promoters (Table C.2 - Appendix C). Other CRE characterisations have similarly reported ratios of 2.5 enhancers per promoter in mammalian species [367]. Overall, around half of the binding regions in each set were found to overlap with regulatory elements, either promoters or enhancers.

### 4.2.3 The majority of TF binding regions contain a combination of TF binding motifs

To complement the characterisation of the liver TF cistromes, I also identified TF binding motifs within their contained peaks. Given the extensive combinatorial binding that had been observed, a particular aspect to be addressed was whether this extent of TF co-binding was also recapitulated in the motif content of the TF peaks. In the first place, I estimated the relative amount of identified ChIP-seq TF peaks that contain the binding motif of the corresponding TF. Specifically, I retrieved the underlying genomic sequences of the CEBPA, FOXA1 and HNF4A peak sets, as identified by ChIP-seq, and scanned them searching for occurrences of the canonical CEBPA-binding, FOXA1-binding and HNF4A-binding motif respectively [168] (see Methods) (Fig. 4.3a, b). Most peaks of each set (67-77%) were indeed found to contain at least one occurrence of the binding motif of the corresponding TF, while a minority of them did not have any motif (Fig. 4.3b). Overall, some tens to hundreds of thousands of occurrences of each canonical TF binding motif were identified in the aggregation of peaks of the corresponding TF cistrome (Fig. C.2a - Appendix C). The total number of identified motifs in each cistrome was generally proportional to the number of TF peaks (Fig. 4.3b), although there was not a one-to-one relationship, as some peaks might contain more than one motif,

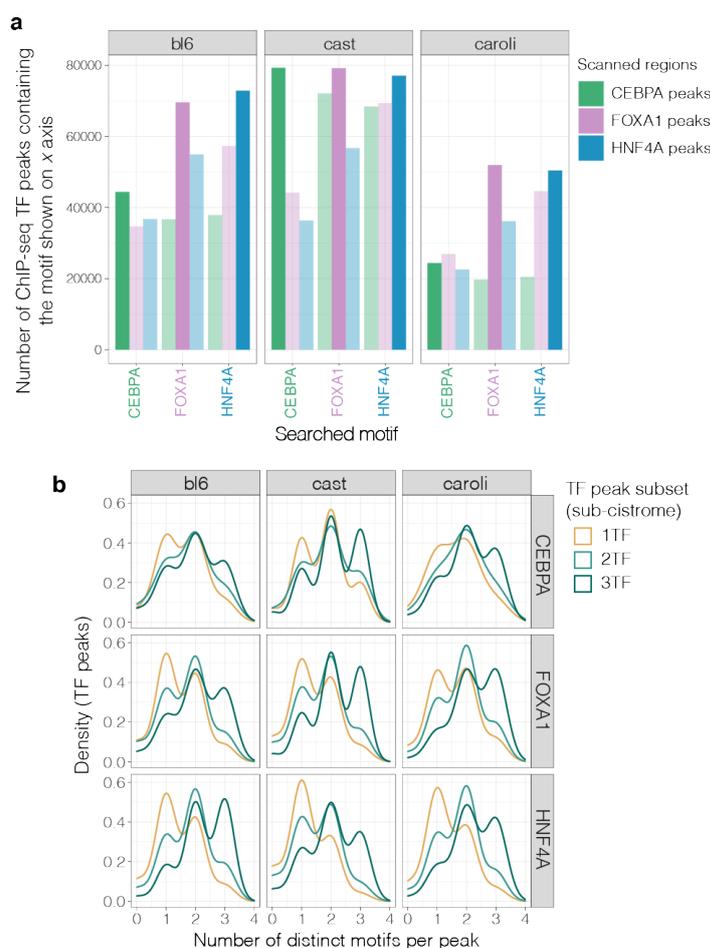


**Fig. 4.3 Motif identification in the ChIP-seq TF peaks.** a) Position weight matrices of the CEBPA, FOXA1 and HNF4A motifs that were searched within the ChIP-enriched regions for the corresponding TF [168]. b) Number of TF peaks that contain, or not, occurrences of the searched motif.

while also some others might not contain any. Cross-species differences in the amount of motifs identified in each TF peak set may be an effect of the different number of peaks included in each of these sets. For instance, the lower number of CEBPA motifs that were identified in the CAROLI CEBPA peaks than in the CAST CEBPA peaks is probably attributed to the low number of CEBPA peaks identified in CAROLI.

Next, I also searched for occurrences of each of the three TF-binding motifs in the cistromes of the other TFs, for example I searched for CEBPA-binding motif occurrences in the FOXA1 peaks. Interestingly, numerous occurrences of each motif were also found in the ChIP-seq peaks of the other TFs, for example thousands of FOXA1-binding motifs were identified not only in FOXA1 peaks but also in CEBPA peaks (Fig. C.2a - Appendix C). This is indicative of the potential of each TFBR to bind a combination of TFs, either simultaneously or not. On another note, as a subset of the identified TFBRs occur within CREs (Fig. 4.2c), also a subset of the identified motifs in the TFBRs occur in CRE peaks (Fig. C.2b - Appendix C).

To gain further insight into the potential of combinatorial binding at the TFBRs of each cistrome, I sought to report also the number of peaks from each TF cistrome that contain the binding motif of other TFs. I observed that, besides the canonical motif of their corresponding TF, the ChIP-seq peaks of each TF cistrome also contained, to a large extent, the canonical motifs of the other liver TFs (Fig. 4.4a). This further supports that most of the TF peaks have the potential to bind different TFs, very likely in combinations.



**Fig. 4.4 Content of aggregated TFBRs of each cistrome in canonical CEBPA, FOXA1 and HNF4A binding motifs.** CEBPA, FOXA1 and HNF4A motifs were searched in the underlying sequences of all peaks from each set. a) Total number of TF peaks (y axis) containing at least one occurrence of each canonical motif indicated on x axis, b) Density plots showing the number of distinct TF binding motifs (CEBPA, FOXA1 and HNF4A) contained per TF peak, with the peaks of each cistrome grouped as 1TF-, 2TF- or 3TF-bound based on ChIP-seq enrichment for the liver TFs.

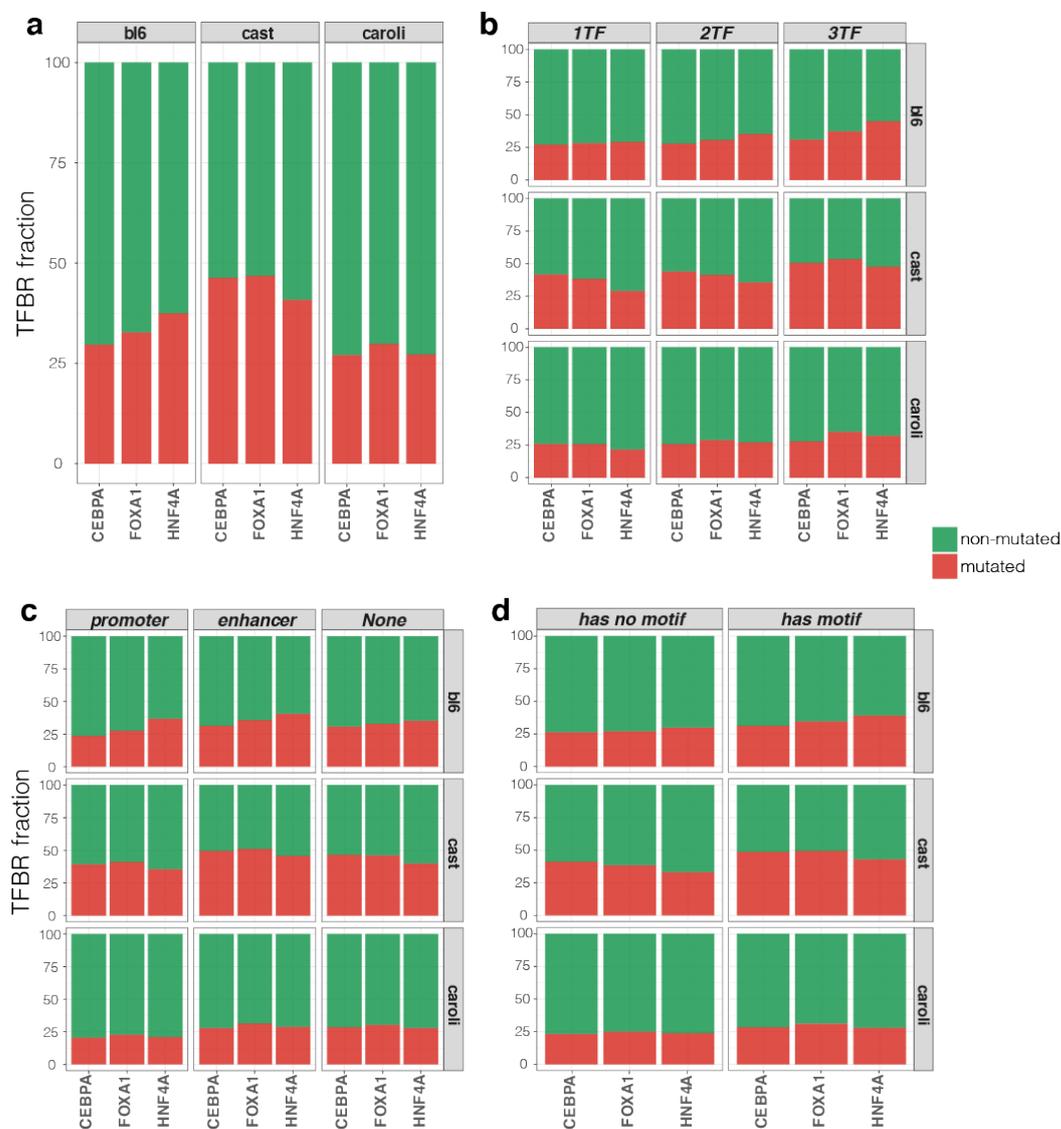
A particular observation in this analysis was that the amount of CEBPA peaks containing CEBPA-binding motifs was very similar to the reported number of CEBPA peaks containing FOXA1 and HNF4A motifs (CEBPA peaks are shown as green bars in Fig. 4.4a). Similarly, the total amount of occurrences of the canonical FOXA1 and HNF4A motifs in CEBPA peaks was very close to the total amount of CEBPA-binding motifs in the CEBPA peaks (CEBPA peaks containing occurrences of different motifs are also shown as green bars in Fig. C.2a - Appendix C). Considering also the relative fraction of combinatorially bound CEBPA TFBRs (2TF and 3TF groups in Fig. 4.2b), as compared to the fewer singly bound CEBPA TFBRs (1TF group in Fig. 4.2b) for BL6 and CAROLI, these obser-

vations may indicate that CEBPA co-binds with the other two TFs at a particularly increased frequency.

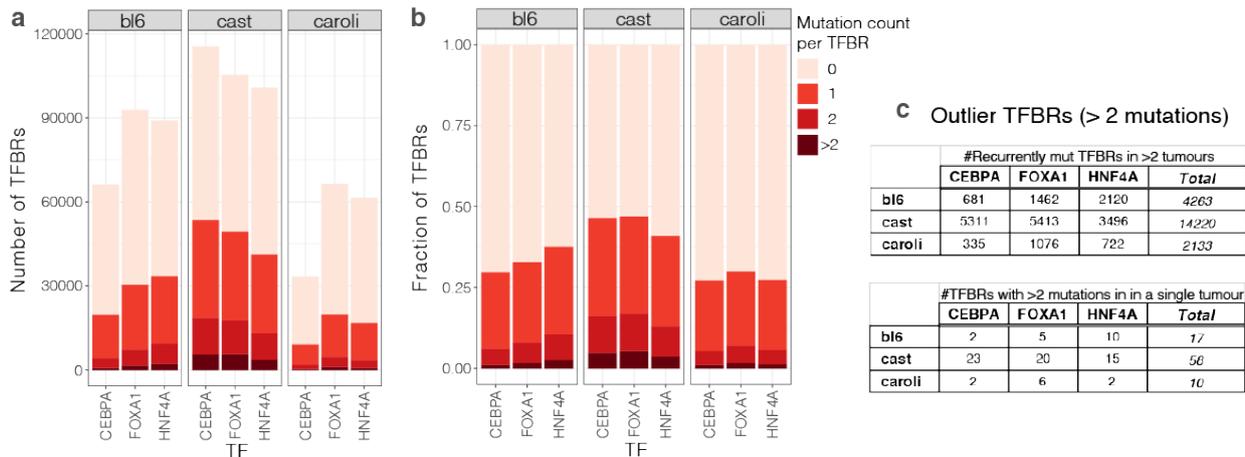
Up to this point, I reported total numbers of motif occurrences in the aggregated peaks of each TF cistrome. In addition to that, I summarised whether the sub-grouping of the peaks of each TF cistrome into 1TF-, 2TF- and 3TF-bound sub-cistromes (based on ChIP-seq enrichment for the corresponding numbers of TFs) was recapitulated by their content of canonical motifs of one TF, two TFs or three liver TFs in total. I found that 1TF-bound peaks of the FOXA1 and HNF4A cistromes indeed contained the canonical motif of only one liver TF (median) (Fig. 4.4b). In contrast, CEBPA peaks that had been grouped as 1TF contained motifs of a median of 2 TFs, perhaps underlying again a higher potential of CEBPA to co-bind with the other liver TFs (Fig. 4.4b). Another exception was also FOXA1 1TF peaks in CAROLI, which also contained motifs of 2 TFs (median). Peaks that had been grouped as 2TF-bound were found to contain canonical motifs of a median of 2 TFs, consistently across all TF cistromes and all species. Finally, the 3TF peaks in all cases contained motifs of a median of 2 TFs, although the corresponding distributions were skewed towards 3 (Fig. 4.4b). This implies that a TFBR can possibly get bound by three different liver TFs, even if it does not explicitly contain the binding motifs of all the three corresponding TFs. Overall, the extent of combinatorial TF binding identified by ChIP-seq enrichment was largely recapitulated by the presence of canonical motifs for distinct TFs within the combinatorially bound ChIP-seq peak regions.

#### **4.2.4 Fractions of TF binding regions that are mutated in DEN-induced liver tumours**

A major aim of this project was to evaluate the mutation loads of the liver TF binding profiles in DEN-induced liver tumours. To address this, for each individual species, I performed combined analyses of both the liver TF cistrome datasets and the corresponding mutation datasets. In the first place, I intersected each TFBR set with the aggregation of mutations from all DEN tumours in each species. Overall, less than half of the regions of each set were found to be mutated (Fig. 4.5a, Fig. 4.6). The percentages of mutated TFBRs showed little variation among the peak sets of the different TFs, corresponding to 27-38%, 41-46% and 27-30% for BL6, CAST and CAROLI respectively (Fig. 4.5a). Differences in the mutated TFBR fractions between corresponding cistromes of distinct species may reflect differences in the total number of TFBRs identified in the species, the total number of identified mutations, the available tumour samples, the genome assemblies, or the biology of the species. However, it is noted that the current phase of the study focuses on characterising the TFBR mutational patterns in each species separately and check their reproducibility among species, rather than performing direct detailed cross-species comparisons of measured variables. For the reference, CAST, which had a higher fraction of mutated TFBRs than BL6 and CAROLI, was also the species that had the highest total number of identified mutations, the highest number of available tumour samples from which the mutations were called ( $N=84$ , Table 3.1), as well as the highest number of



**Fig. 4.5 Relative fractions of mutated and non-mutated TFBRs in each cistrome and sub-cistrome.** a) all TFBRs per cistrome, b-d) TFBRs grouped in sub-cistromes based on: the number of (co-)bound TFs in region (b), their overlap with CREs (c), and whether they contain the corresponding TF motif (d).



**Fig. 4.6 Mutation load per TFBR for each cistrome, in each mouse species.** a) Numbers, and b) Fractions of individual TFBRs in each cistrome that harbour 0, 1, 2, or >2 mutations (mutation counts per TFBR are based on the aggregated mutation set from all DEN tumours in each species). The majority of TFBRs do not contain any mutation, while most of the TFBRs that are mutated harbour only one mutation. TFBRs with >2 mutations are referred to as *outliers*. c) Most of the *outliers* are TFBRs recurrently mutated in >2 independent tumour samples, while very few of them have >2 mutations in a single tumour sample.

identified TFBRs. In general, the larger size of the CAST data sets were speculated to provide higher statistical power in relevant analyses.

With respect to the number of mutations occurring in each TFBR, as described above, most TFBRs did not contain any mutation. Among the mutated TFBRs, the vast majority contained only one mutation, while there were a number (a few thousands) of outlier TFBRs with higher numbers of mutations (Fig. 4.6). These corresponded mostly to TFBRs that were recurrently mutated in independent tumours of each species, while only very few of them harboured a relatively high number of mutations in a single tumour.

As part of exploratory analyses, I furthered investigated whether the relative proportions of mutated TFBRs differed strikingly between TFBRs sets with distinct functional features. Examining the mutation loads in TFBRs after splitting each cistrome into sub-cistromes based on the number of (co-)bound TFs (Fig. 4.5b), their overlap with regulatory elements (Fig. 4.5c), as well as whether they contain the binding motif of the respective TF or not (Fig. 4.5d), did not show any strikingly dissimilar distributions of mutated versus non mutated TFBRs among the different subsets. Little differences in the fractions of mutated TFBRs among sub-cistromes were difficult to evaluate at this stage, as the different TFBR sets were not corrected for potential biases in their total number of mutations (given that they were called based on different numbers of tumour samples), the total number of their contained TFBRs, the length distribution of the TFBRs, as well as their sequence

context. Therefore, this overview of mutated TFBR fractions is rather part of the comprehensive description of the dataset, while comparison of mutation loads between (sub-)cistromes would require normalisation for potential inherent differences in their included regions.

#### 4.2.5 Weighted mutation rate in TF cistromes and sub-cistromes

As explained in the previous section, to evaluate the relative mutation rates of the liver TF cistromes and sub-cistromes, it was necessary to first correct for potential inherent biases of these different TFBR sets. Thus, to enable comparison of mutation loads between distinct (sub-)cistromes, I developed an approach to calculate mutation rates that included correction for potential batch effects.

The first issue I accounted for was that the vast majority of TFBRs contain no or only one mutation. A simplistic approach to estimate and compare mutation rates between cistromes would be to calculate the mutation density of each TFBR in a cistrome, i.e. divide the number of mutations in a TFBR by the length of the TFBR, and average across all TFBRs of the cistrome. However, the fact that—as shown above—the vast majority of TFBRs contains only one mutation would be problematic in the calculation of mutation density per TFBRs. The reason would be that since most TFBR mutation density fractions would have a numerator (number of mutations) equal to one, the denominator (TFBR length) would be the defining factor of the mutation density value, thus making TFBRs with larger length (aka large denominator) appearing as more lowly mutated. This means that different mutation rates between distinct TFBR sets could reflect differences in their contained TFBR length distributions rather than differences in their mutation burdens. Therefore, to deal with this issue, rather than calculating mutation rate per individual TFBR in a (sub-)cistrome, I calculated the mutation rate in the aggregation of all TFBRs in a (sub-)cistrome (Fig. 4.7).

An additional parameter I took into account to calculate mutation rates across different sets of TFBRs was the sequence context. The sequence context can affect the observed mutation rate, because not all nucleotides have the same probability of being mutated. For example, as explained above, the mutagenic effect of DEN is more prominent on thymines. This means that mutation rates between distinct (sub-)cistromes might appear different as a result of their different nucleotide composition. In that case, it would be difficult to distinguish when a mutation rate difference between two TF peak sets represents a nucleotide composition bias, or it is attributed to differences in biological processes, such as repair efficiency in the underlying regions or negative selection effects. Therefore, in my calculation of mutation rates, I also corrected for the sequence context biases of the regions included in the aggregated (sub-)cistromes.

Taken together, I calculated the mutation rate in each aggregated (sub-)cistrome, weighting it by trinucleotide context frequency (*weighted mutation rate, wmr*). Specifically, I first retrieved from the

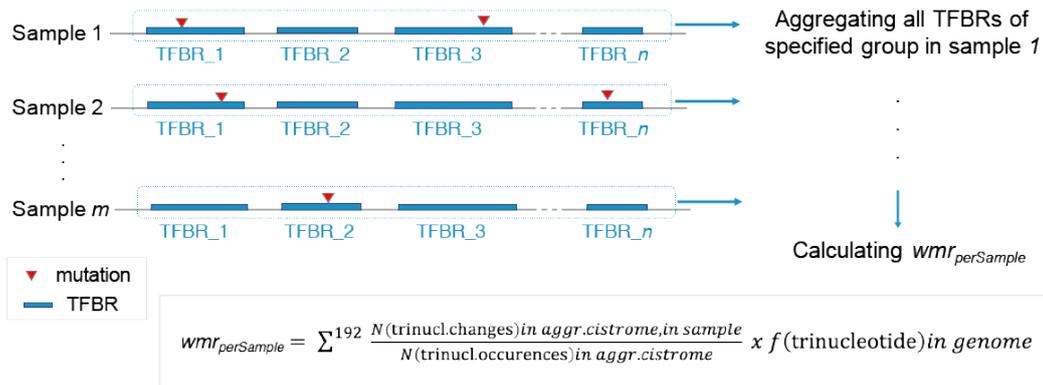
reference genome all TFBR sequences of a specified (sub-)cistrome and aggregated them. Within the sequence aggregation, I counted the occurrences of each possible trinucleotide, creating a vector of 64 trinucleotide counts for the aggregated (sub-)cistrome. Then, I determined the number of mutations per trinucleotide that occurred within the specified (sub-)cistrome, in each individual tumour sample. This means that, in each sample, I counted the occurrences of each mutated nucleotide, conditioned on its 5' and 3' nucleotides, i.e. accounting for its trinucleotide context. That resulted in a vector of size 192, containing the occurrence count of each of the 192 possible trinucleotide changes, in the aggregated (sub-)cistrome in each individual sample (Fig. 4.7). I then calculated the mutation rate for each of the 192 trinucleotide changes as the number of occurrences of the specific trinucleotide mutation in the aggregated cistrome in a given sample ( $N(\text{trinucleotide.changes})_{\text{in\_aggr.cistrome.in\_sample}}$ ), divided by the total count of the corresponding trinucleotide in the aggregated cistrome ( $N(\text{trinucleotide.occurrences})_{\text{in\_aggr.cistrome}}$ ), and then weighted that fraction by multiplying by the frequency of occurrence of this trinucleotide in the genome of the corresponding species ( $f(\text{trinucleotide})_{\text{in\_genome}}$ ):

$$wmr_{\text{trinucleotide}} = \frac{N(\text{trinucleotide.changes})_{\text{in\_aggr.cistrome.in\_sample}}}{N(\text{trinucleotide.occurrences})_{\text{in\_aggr.cistrome}}} \times f(\text{trinucleotide})_{\text{in\_genome}} .$$

The sum of the mutation rates for all possible trinucleotide changes in a given tumour sample was the weighted mutation rate of the aggregated cistrome in the sample:

$$wmr_{\text{perSample}} = \sum^{192} \frac{N(\text{trinucleotide.changes})_{\text{in\_aggr.cistrome.in\_sample}}}{N(\text{trinucleotide.occurrences})_{\text{in\_aggr.cistrome}}} \times f(\text{trinucleotide})_{\text{in\_genome}} \text{ (Fig. 4.7).}$$

Using the above described method, I calculated the weighted mutation rate for each defined TF cistrome (CEBPA, FOXA1 and HNF4A) and sub-cistrome (1TF, 2TF, 3TF and promoter-overlapping, enhancer-overlapping, no-CRE-overlapping) in every tumour sample of each species (Fig. 4.8a).



**Fig. 4.7 Overview of approach developed to calculate weighted mutation rate ( $wmr$ ) for a given cistrome or sub-cistrome, per tumour sample.** The reference sequences of all TF binding regions contained in the (sub-)cistrome were aggregated. First the  $wmr$  of each of the 192 possible trinucleotide changes was calculated in the specified cistrome, in a given sample. The final  $wmr$  for the aggregated cistrome in the given sample was total sum of all 192 trinucleotide context  $wmrs$ .

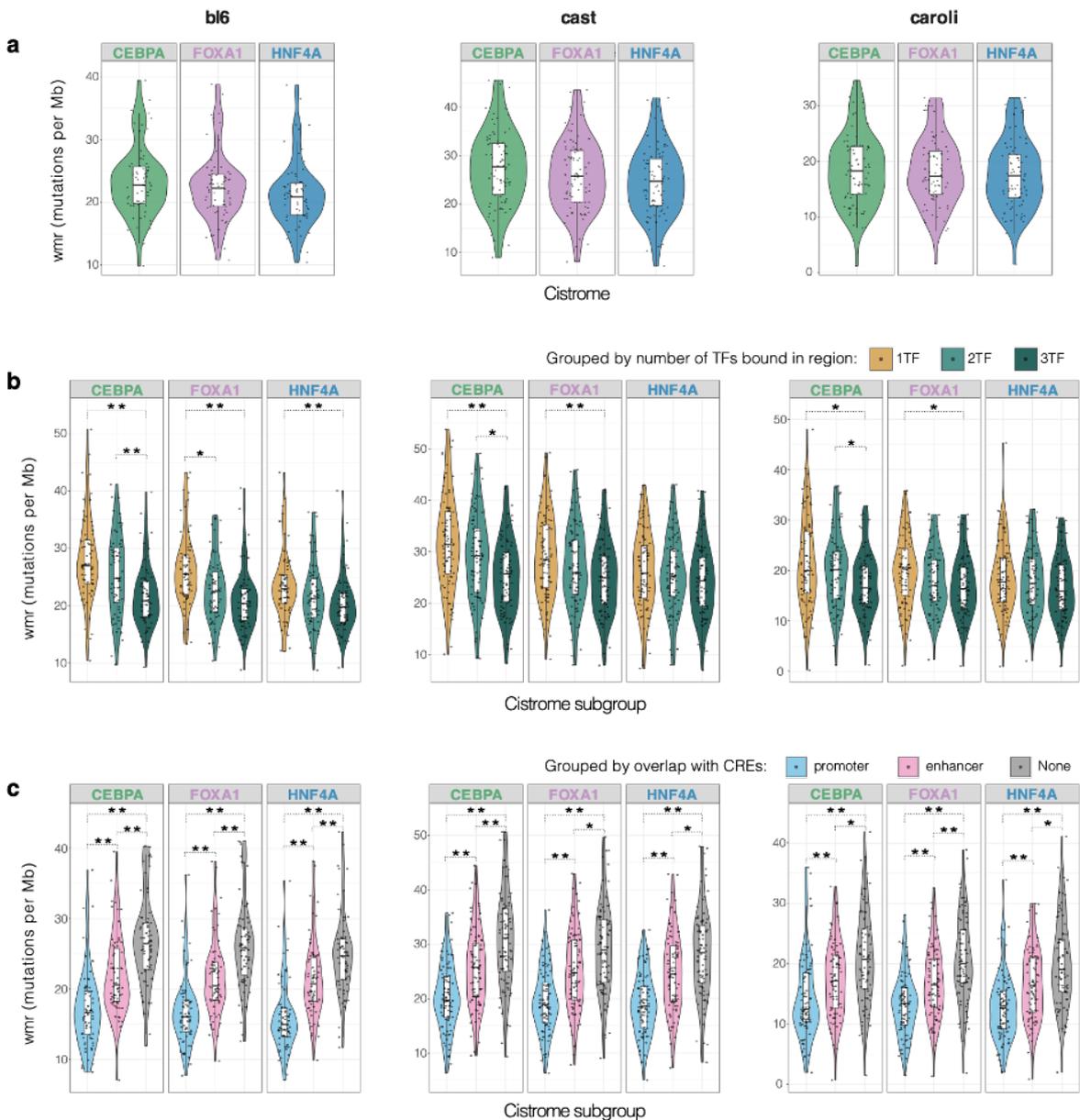
Within each species, pairwise comparisons were performed between each pair of cistromes and pairs of sub-cistromes, using Mann-Whitney U tests.

Consistently in all the three species, no significant difference was observed between the cistromes of the liver TFs (Fig. 4.8a). This suggests that none of the three TF cistromes shows particularly higher or lower tolerance to accumulating mutations in the DEN-induced liver tumours.

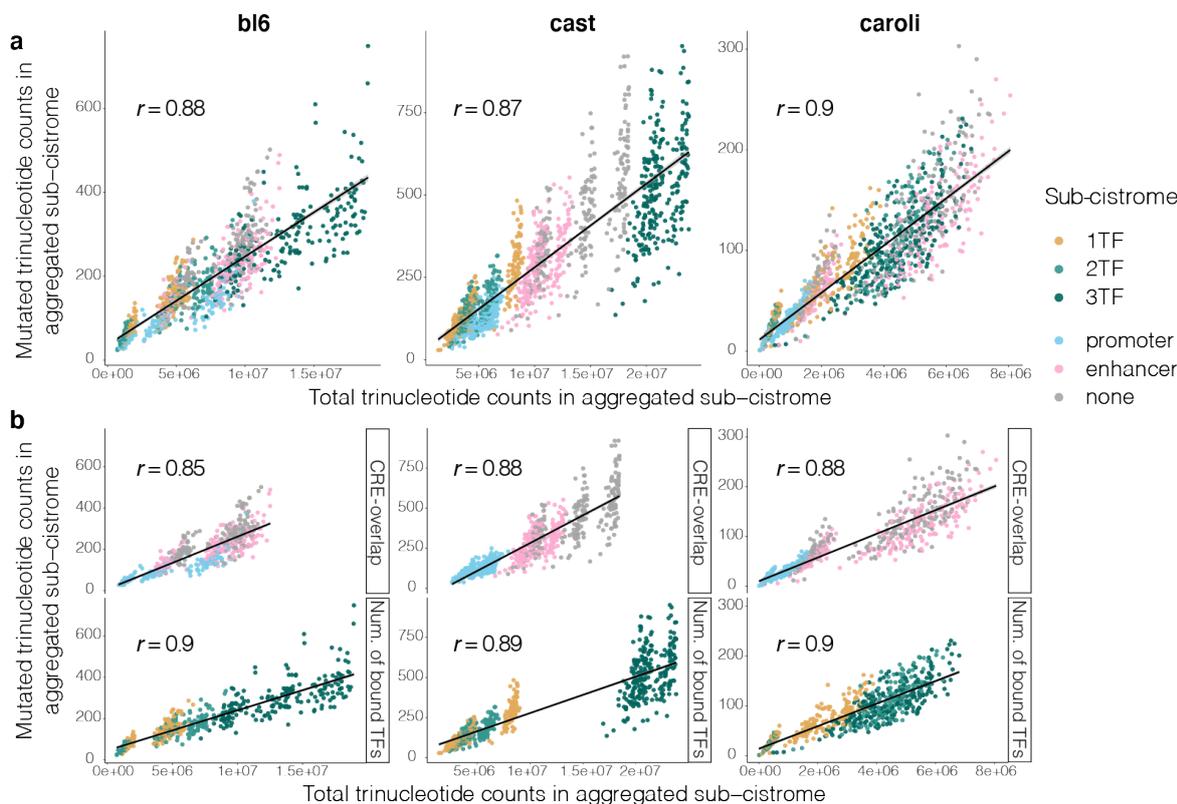
Then, I performed pairwise comparisons among the *wmrs* of the 1TF, 2TF and 3TF sub-cistromes of each cistrome. The 1TF-bound TFBR groups of the CEBPA and FOXA1 cistromes, showed significantly higher mutation rates than the corresponding 3TF-bound groups. This pattern was reproducible in all three species, although in CAROLI the *p*-values were slightly higher (*p*-values < 0.01, as opposed to the other significant *p*-values < 0.001) (Fig. 4.8b). Moreover, only in BL6, 1TF-bound regions of the HNF4A cistrome also showed significantly increased *wmr* compared to the 3TF-bound HNF4A regions (Fig. 4.8b), while the corresponding comparison in the other species did not show significant differences. Interestingly, in all species, the CEBPA binding regions that were co-bound by a total of 2TFs (as inferred by the ChIP enrichment analysis) also showed significantly higher mutation rates than the 3TF-co-bound CEBPA regions (Fig. 4.8b). Finally, pairwise comparisons between 1TF- and 2TF-bound regions of the different cistromes did not show significant differences. The only exception was the 1TF-bound FOXA1 regions specifically in BL6 that exhibited higher mutation rates than the corresponding 2TF-bound regions of the FOXA1 cistrome (Fig. 4.8b). Overall, mutation rate profiling revealed a progressive reduction of mutation loads from regions bound by one TF alone to regions that are co-bound by different TFs.

In addition, for each species, I compared the weighted mutation rates of the promoter-overlapping, enhancer-overlapping and no-CRE-overlapping sub-cistromes of each TF cistrome (Fig. 4.8c). All possible pairwise comparisons between promoter-overlapping, enhancer-overlapping and no-CRE-overlapping sub-cistromes revealed statistically significant differences. Specifically, in all species and for all three liver TFs, binding regions in promoters showed the lowest mutation rates, followed by enhancer-overlapping binding regions, and lastly by TFBRs that do not overlap with CREs, which had the highest mutation rates (Fig. 4.8c). This suggests a strong effect of overlap with promoters or enhancers on the mutation burden of TF binding regions.

Finally, I revisited the issue of low and relatively invariable mutation counts per TFBR that I described at the beginning of this section (i.e. the fact that the vast majority of TFBRs contained no or only one mutation). I, specifically, sought to determine whether this issue had been effectively corrected for by aggregating the TFBRs of each cistrome. The motivation for this arose from the fact that the 3TF sub-cistromes, which were shown to have the lowest mutation rates compared to the 1TF and 2TF groups, were also observed to generally contain longer TFBRs than the 2TF and 1TF-bound TF peak regions (Fig. 4.2b). In addition to that, the 3TF groups in most cases were shown to

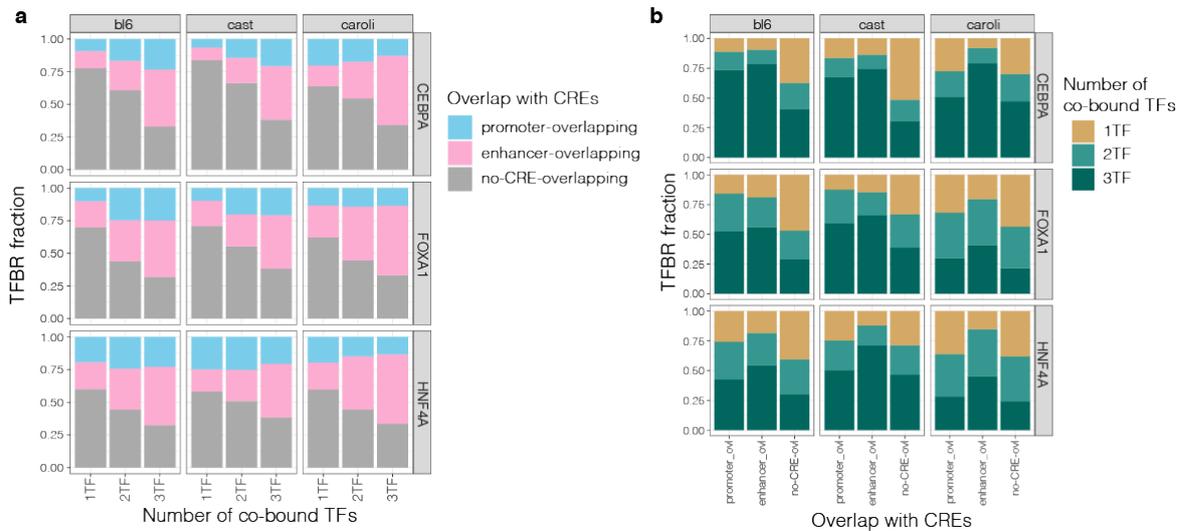


**Fig. 4.8 Distribution of weighted mutation rates per sample in each cistrome and sub-cistrome of the studied mouse species.** Distribution of *wmr* per sample for a) the CEBPA, FOXA1 and HNF4A cistrome in each species, b) the 1TF, 2TF and 3TF sub-cistrome of each cistrome shown in (a), and c) the promoter-overlapping, enhancer-overlapping and no-CRE-overlapping sub-cistromes of each cistrome shown in (a). Each data point represents the *wmr* in a single DEN tumour sample. Pairwise comparisons between cistromes or sub-cistromes were performed using Mann Whitney U tests. \*\* indicates  $p$ -values  $< 0.001$ , \* indicates  $p$ -values  $< 0.01$ . Non-significant comparison results are not labelled. Number of tumour samples (data points) for each species are shown in Table 3.1



**Fig. 4.9 Correlation between counts of mutated trinucleotides (y axis) in each aggregated sub-cistrome of every tumour and total counts of trinucleotides (x axis) in each aggregated sub-cistrome (in reference genome).** a) Each possible sub-cistrome in a sample is shown as a data point, b) same data points as in (a) but split in sub-cistromes by overlap with regulatory elements (upper panel) and sub-cistromes by number of (co-)bound TFs (lower panel). A linear model is fitted in each case and Pearson's correlation of mutated trinucleotide counts with total trinucleotide counts is shown.

include a higher number of TFBRs than the 1TF or 2TF groups (Fig. 4.2a), which would result in the corresponding sequence aggregation of sub-cistromes of the 3TF groups having a larger total length than the aggregated 2TF and 1TF aggregations. If the “invariable mutation counts” problem was persistent even after the aggregation of TFBRs within (sub-)cistromes, i.e. the total mutation count was consistently low among all aggregated cistromes and the total length was presenting large differences, that would be reflected in a relatively invariable numerator (mutated trinucleotides) across the different cistromes, as opposed to a highly variable denominator (total trinucleotide counts in the aggregated cistrome) in the *wmr* fractions (equation in Fig. 4.7). That would cause inflation of mutation rates in (sub-)cistromes with smaller total length of the aggregated sequences, thus smaller total trinucleotide counts. To gain further evidence that the *wmr* differences between aggregated sub-cistromes are not determined only by the length differences of their contained (aggregated) regions, I went on to examine the relationship of mutated trinucleotides with the total number of trinucleotides



**Fig. 4.10 Representation of promoter-overlapping, enhancer-overlapping and no-CRE-overlapping TFBRs in the 1TF, 2TF and 3TF sub-cistromes (a), and vice versa (b).**

(which is proportional to aggregated cistrome length) in each aggregated sub-cistrome per sample. In case of persisting "invariable mutation counts" problems, i.e. relatively stable mutation counts across the (sub-)cistromes, yet different lengths of the aggregated sub-cistromes, the two variables would not be changing linearly and plotting them would reveal a plateau trend. I confirmed that this was not the case. Instead, mutated trinucleotide counts were shown to increase linearly with total trinucleotide counts in aggregated sub-cistromes, having also a high correlation (Pearson's  $r = 0.85 - 0.9$ ) (Fig. 4.9). This increased the confidence that wmr differences do not reflect region length biases.

Overall, among the TF peak groups with different number of (co-)bound TFs, the 3TF-bound regions had the lowest mutation rates (Fig. 4.8b), while among the TF peak groups with different CRE associations, the promoter-overlapping regions also exhibited the lowest mutation rates (Fig. 4.8c). This observation raised the following question: could the 3TF binding region groups coincide with the promoter-overlapping region groups of each cistrome? In other words, could the reduced mutation rate of regions that display combinatorial binding of liver TFs be explained by the fact that these regions coincide with promoter regions and/or vice versa? To address this question, I further investigated the association of the 1TF, 2TF and 3TF groups with the promoter-overlapping, enhancer-overlapping and none-overlapping groups. Specifically, I computed and reported the representation of each of the latter groups in each of the former ones (Fig. 4.10). I observed that indeed promoter-overlapping binding regions were very often more highly represented in 3TF groups, and this over-representation was often in cases where the 3TF group had been shown to have significantly lower mutation rate than the 1TF group. However, that was not always the case; for example, in CAROLI CEBPA and FOXA1 cistromes, the 3TF groups were not more enriched for promoter-overlapping regions, but they were significantly less mutated than the 1TF groups. Nevertheless, the 3TF groups were more

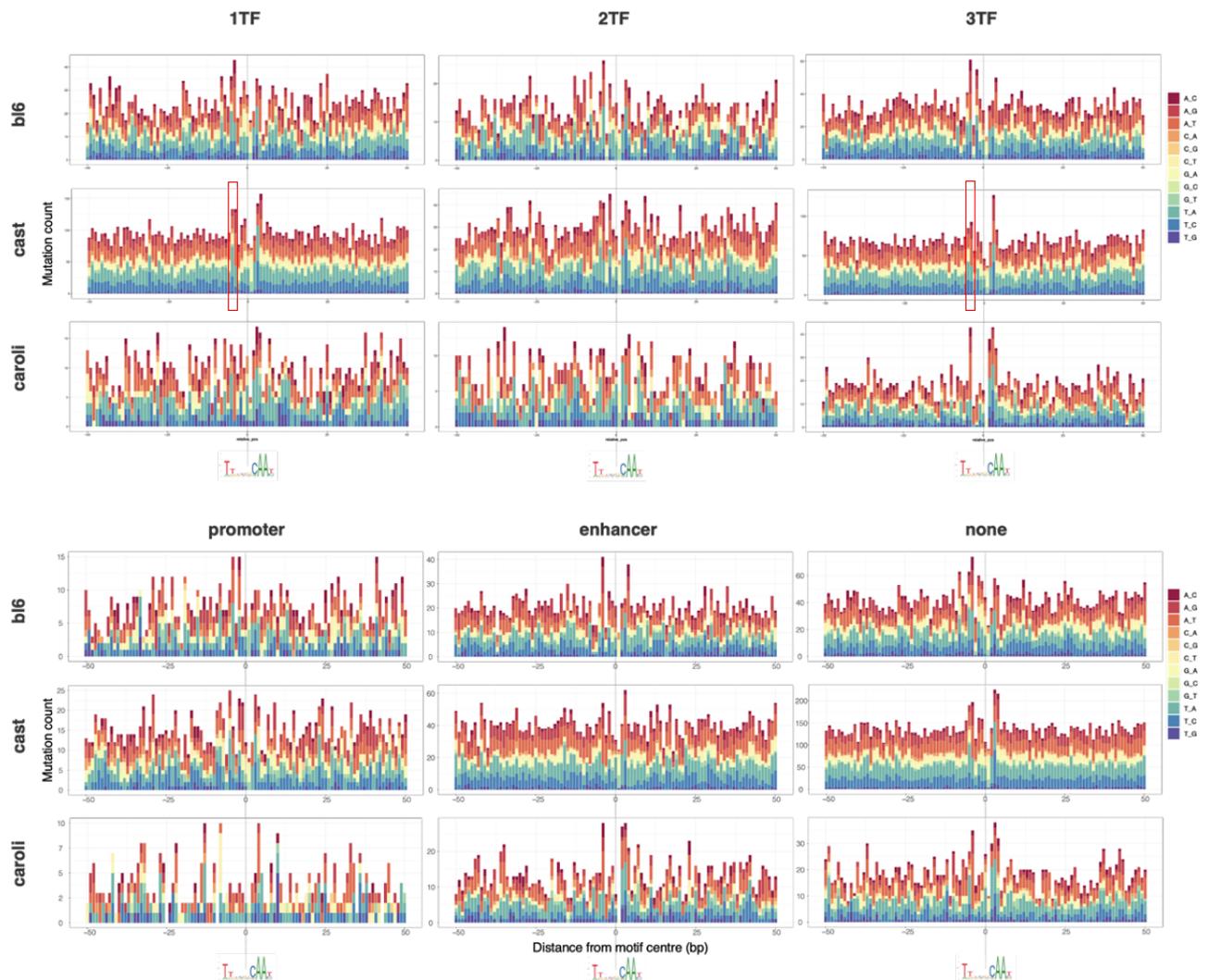
enriched for enhancer-overlapping regions. Overall, there seemed to be a progressively increasing representation of binding regions associated to enhancer and to promoters with increasing number of (co-)bound TFs (Fig. 4.10).

#### 4.2.6 Pairwise comparisons of mutation spectra per position of TF binding motifs between sub-cistromes

As already mentioned, an important feature of TFs is their ability to recognise and bind specific motifs. To gain further insight in the mutational patterns of the TF binding regions in different groups, I characterised the mutational profile of their binding motifs. In case the binding motif in the more highly mutated TF peak sets was accumulating mutations at a specific position, this could have an impact on the TF binding and, therefore, potential functional implications.

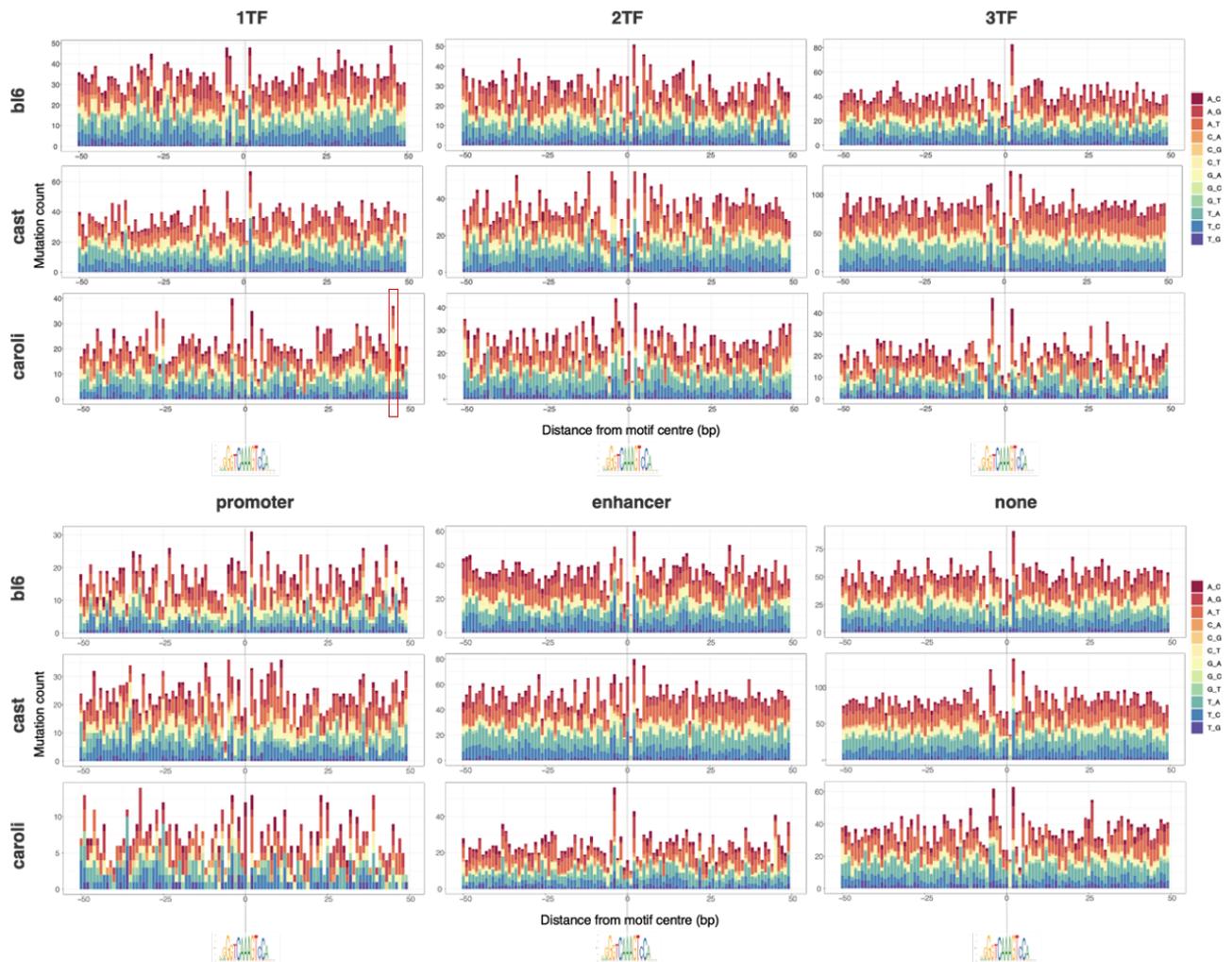
Therefore, I profiled the mutational spectra of the motif sequences, which were identified in the ChIP-seq peaks of each group, and their direct flanking sequences. For each ChIP-seq peak set, I collected the identified motif occurrences for the corresponding TF, e.g. for the identified CEBPA ChIP-seq peaks, I considered their contained CEBPA motif occurrences (motifs identified as explained in 4.2.3, Fig. 4.3, also see Methods for motif identification details), independently of whether the underlying region was also bound by other TFs. Moreover, for peaks containing more than one motif occurrence, I considered only the one with the lowest  $p$ -value in the FIMO output (see Methods for motif identification). That was done to avoid including overlapping motif sequences, which would result in counting their contained mutations multiple times. A motif occurrence was assigned to one of the 1TF, 2TF or 3TF groups, based on the corresponding grouping of the TF peak it was identified in. Similarly, a motif was assigned to one of the promoter-overlapping, enhancer-overlapping, or no-CRE-overlapping groups based on the corresponding categorisation of the TF peak it was identified in.

I retrieved the FASTA sequences of the identified motifs, centred them at their middle nucleotide and extended by  $N=50$ bp both upstream and downstream. Having all the extended motif sequences of equal length, I stacked them together accounting for their orientation, so that their corresponding positions are aligned. For each relative position around the motif centre I calculated the raw mutation counts for each possible nucleotide change and, putting all calculations together, I visualised the mutation spectra per position (Fig. 4.11, 4.12, 4.13). Then I performed pairwise comparisons of the mutation spectra (i.e. the different nucleotide change contexts) of corresponding motif positions between the different groups of TF binding motifs, by applying Fisher's exact tests. Specifically, for every pairwise comparison between two motif groups (e.g. 1TF vs 3TF group), I ran a Fisher's exact test for each nucleotide change context at every relative motif position between the two groups that were compared. The contingency tables I used included the nucleotide change counts at a given



**Fig. 4.11 Mutation spectra per relative position around the CEBPA motif centre for motif occurrences identified in each TF sub-cistrome, in all studied mouse species.** CEBPA motif occurrences were identified in each cistrome, extended +/- 50bp and stacked together, centred at the motif centre and in the same orientation. Pairwise comparisons were performed for mutation spectra at each corresponding position between groups, by running Fisher's exact tests and Bonferroni multiple testing corrections. In red are marked positions with significantly different enrichment of mutation spectra between two groups.





**Fig. 4.13 Mutation spectra per relative position around the HNF4A motif centre for motif occurrences identified in each TF sub-cistrome, in all studied mouse species.** HNF4A motif occurrences were identified in each cistrome, extended +/- 50bp and stacked together, centred at the motif centre and in the same orientation. Pairwise comparisons were performed for mutation spectra at each corresponding position between groups, by running Fisher's exact tests and Bonferroni multiple testing corrections. In red are marked positions with significantly different enrichment of mutation spectra between two groups.

position and the reference nucleotide counts at the same position in the two compared groups. For example, the contingency table for the comparison of T → A mutations at position 2 of a given extended motif between 1TF with the 3TF groups would be:

	1TF	3TF
Pos 2: T > A (nucleotide change context)	Raw counts	Raw counts
Pos 2: T (reference nucleotide)	Raw counts	Raw counts

Given that there are twelve possible nucleotide change contexts (A→C, A→G, A→T, C→A, C→G, C→T, G→A, G→C, G→T, T→A, T→C, and T→G), for every pairwise comparison between motif groups, twelve independent Fisher's exact tests were performed for each position of the stacked extended motif sequences. This way I performed all possible pairwise comparisons between the motif groups 1TF, 2TF and 3TF, as well as the promoter-overlapping, enhancer overlapping and no-CRE-overlapping groups. To correct the *p*-values of the multiple performed tests, I applied a Bonferroni multiple testing correction including all pairwise Fisher's exact tests. None of the pairwise comparisons of nucleotide change context enrichment per position passed the Bonferroni multiple testing correction.

However, this method of comparing mutation spectra per position of the extended motif had an important caveat: it did not take into account the probability of each individual position to be mutated based on its sequence context. Even if there were identified certain motif positions with significantly different mutation loads between two TFBR groups, it would be hard to conclude whether this could reflect potential sequence context biases between the groups, especially considering that DEN causes numerous lesions mostly on T's. Therefore, I sought to apply another method that would take into account the sequence context.

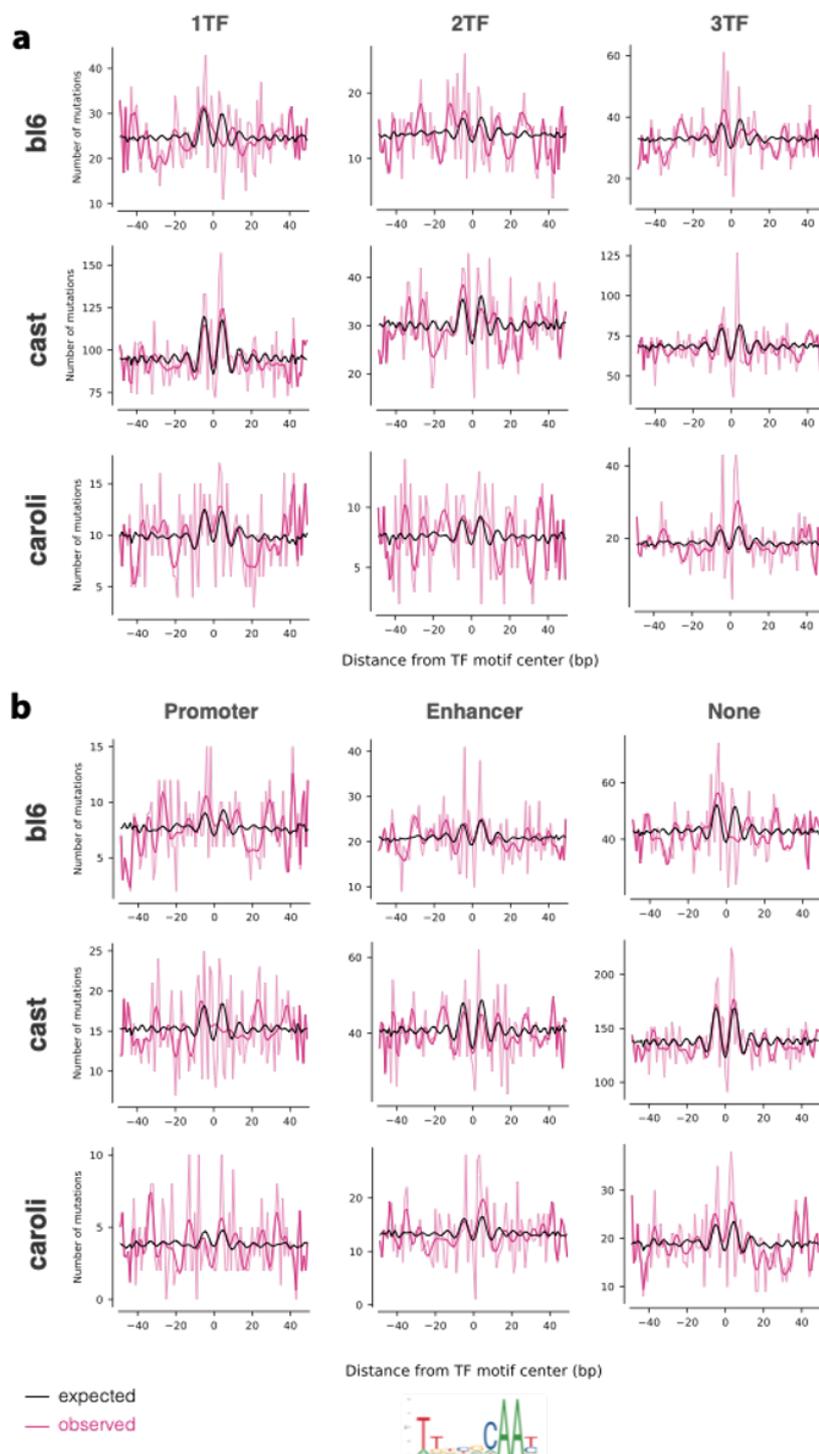
#### **4.2.7 Pairwise comparisons of sequence context-conditional mutation counts per position of TF binding motifs between sub-cistromes**

Following from the approach described above, I compared sub-cistrome groups with respect to their mutation load per motif position, normalised by sequence context. I stacked, as described above, the extended motif sequences, but besides the observed mutation counts per position, I also calculated the expected mutation load (i.e. mutation counts) of each position conditioned on its probability to be mutated (see Methods). The expected mutation load could then be used as a reference to evaluate the observed mutation load. To compute the observed mutation load per position, I counted all mutations occurring at a position, without splitting the mutation counts by nucleotide change context as was done in the mutation spectra analysis above. The total mutation count at a given position of the stacked motifs would give an estimation of this motif position being susceptible to mutation accumulation in general, independently of the nucleotide change context of the occurring mutations

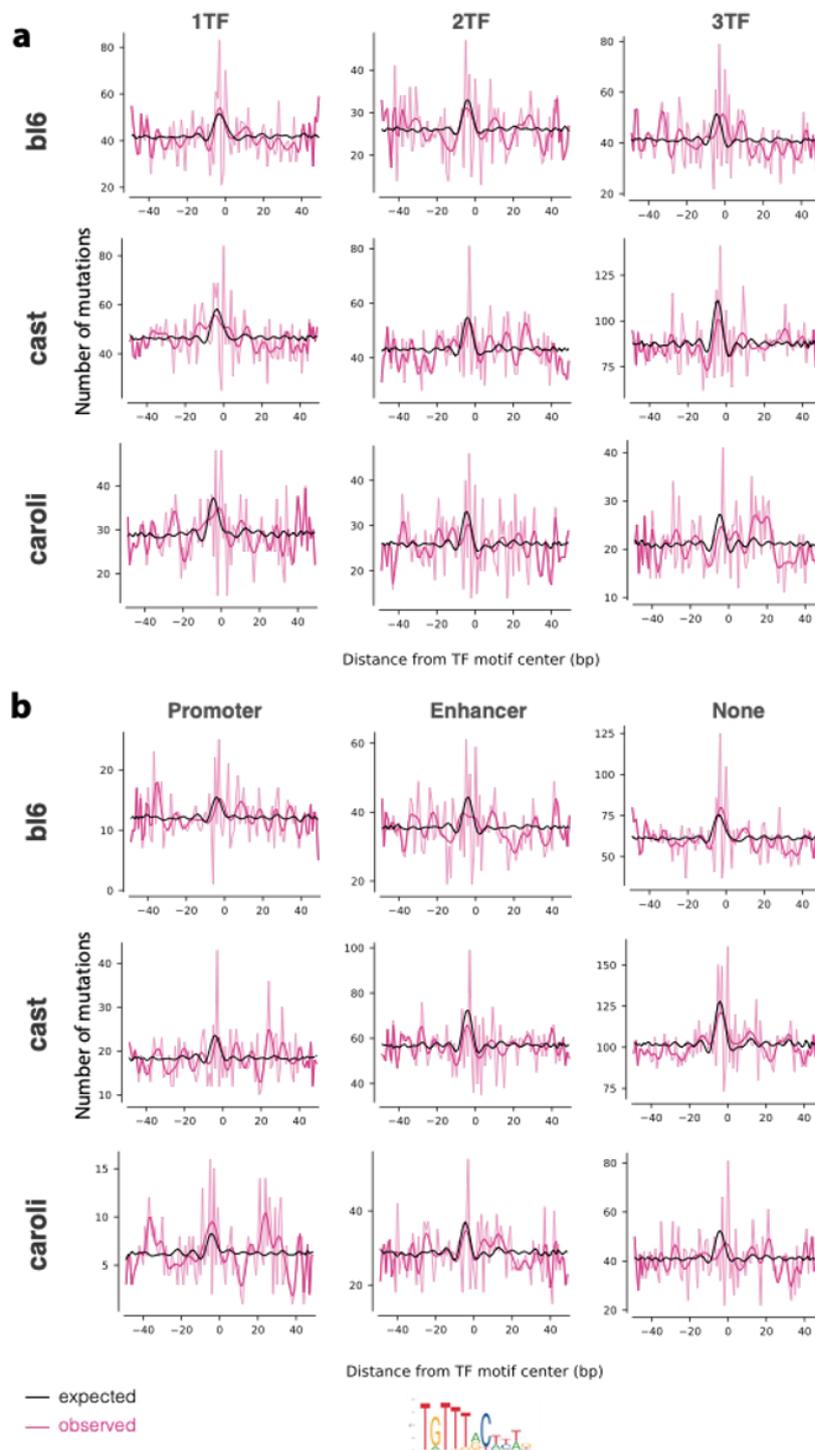
(i.e. independently of whether the occurring mutations would, for example, be mostly  $A \rightarrow T$ , or  $G \rightarrow A$ , etc). In addition, it would increase the statistical power, as the number of mutations would not have to be split in smaller sets by nucleotide change context. To compute the expected mutation counts of a position, I did take into consideration the probability of every position to be mutated based on its trinucleotide composition. For this, I precalculated mutation probabilities of each trinucleotide context (see Methods). This approach was an adaptation of a computational method from Frigola et al. (2017) [368] that calculates observed and expected mutation loads in aggregations of stacked genomic sequences. By adapting and implementing this method, I generated and overlaid the observed and expected mutation counts per position in the stacked extended motif sequences (Fig. 4.14, 4.15, 4.16). Then, I performed pairwise comparisons of mutation loads between corresponding positions of motif sets from the different sub-cistromes. For the pairwise comparisons, I applied Fisher's exact tests, but this time in the contingency tables I included the observed and expected mutation counts at a given position. For example, to compare the mutation burden at position 2 of stacked extended motif sequences between the 1TF and 3TF groups, I used the following contingency table in the corresponding Fisher's exact test:

	1TF	3TF
Observed mutation counts at pos. 2	Raw counts	Raw counts
Expected mutation counts at pos. 2	Raw counts	Raw counts

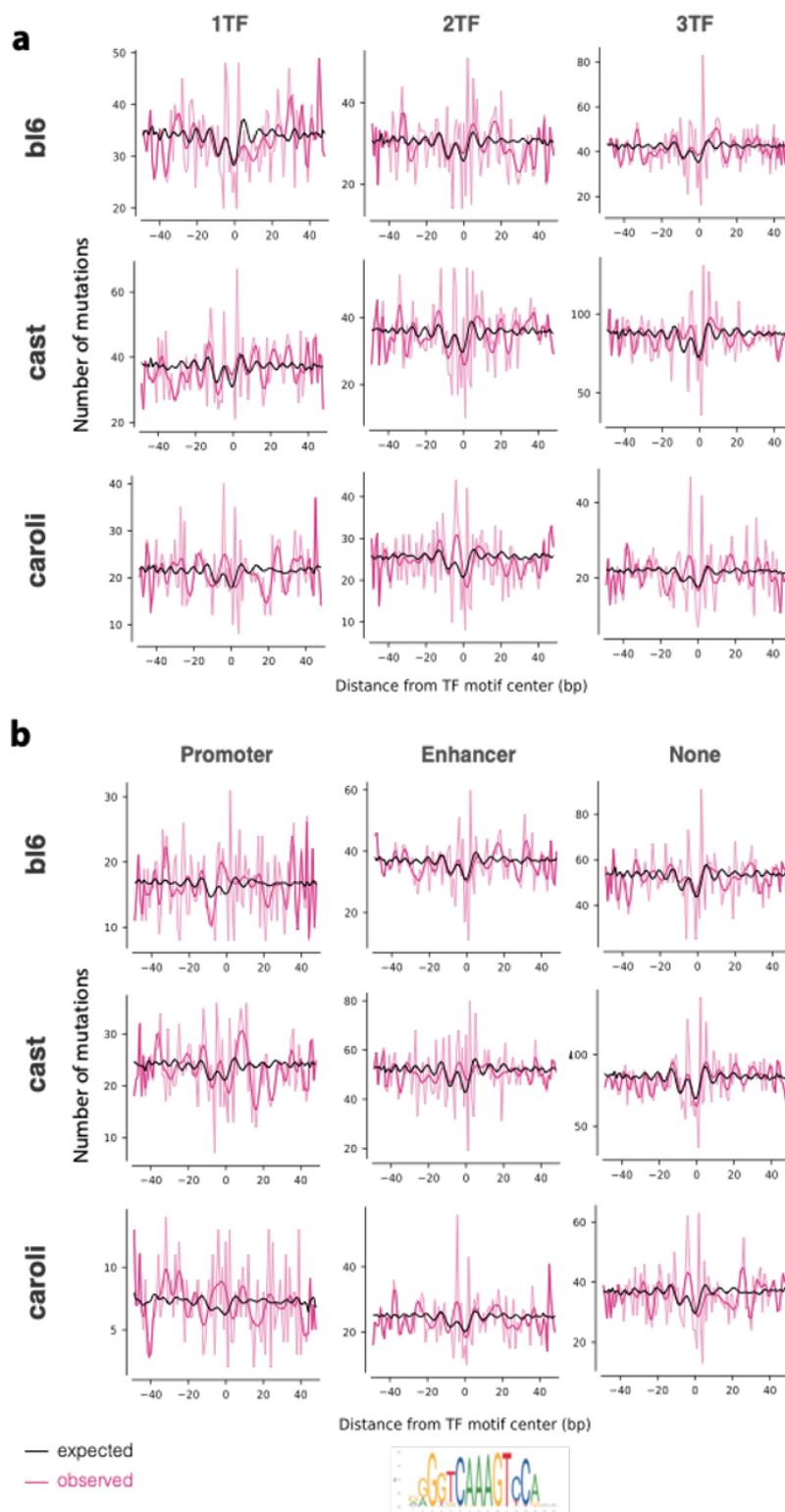
Although visualisation of observed and expected mutations showed some positions where the deviation of the observed mutation from the expected mutation load seemed to differ between different sub-cistromes, the Fisher's exact tests did not show any statistically significant results. This means that, based on this analysis, there was not identified any specific position of the extended TF binding motifs (CEBPA, FOXA1 and HNF4A motifs) that is more susceptible to accumulating mutations.



**Fig. 4.14 Observed versus Expected mutation loads per position of the extended CEBPA motif in the sub-cistromes of each species.** The sequences of identified CEBPA motif occurrences within each sub-cistrome were stacked, centred on the motif middle nucleotide. Observed mutation loads represent raw mutation counts at each relative position around the motif centre. Calculation of expected mutation loads accounts for the mutational probability of each position, based on its trinucleotide context.



**Fig. 4.15 Observed versus Expected mutation loads per position of the extended FOXA1 motif in the sub-cistromes of each species.** The sequences of identified FOXA1 motif occurrences within each sub-cistrome were stacked, centred on the motif middle nucleotide. Observed mutation loads represent raw mutation counts at each relative position around the motif centre. Calculation of expected mutation loads accounts for the mutational probability of each position, based on its trinucleotide context.



**Fig. 4.16 Observed versus Expected mutation loads per position of the extended HNF4A motif in the sub-cistromes of each species.** The sequences of identified HNF4A motif occurrences within each sub-cistrome were stacked, centred on the motif middle nucleotide. Observed mutation loads represent raw mutation counts at each relative position around the motif centre. Calculation of expected mutation loads accounts for the mutational probability of each position, based on its trinucleotide context.

### 4.2.8 Functional analysis of hyper-mutated TFBRs

As mentioned in section 4.2.4, although the vast majority of mutated TFBRs were hit by only one mutation from the whole mutation aggregation of all tumour samples in a species, there were some thousands of TFBRs with more than two mutations (appearing as outliers in Fig. 4.6a). Most of these TFBRs were recurrently mutated in different tumour samples, while very few of them had more than two mutations in a single sample (Fig. 4.6c). I questioned whether these *hyper-mutated* TFBRs can be associated with expression changes in HCC or with any dysregulated pathway. I first sought to profile how these TFBRs distribute across the defined sub-cistromes of each TF. Interestingly, the majority of the hyper-mutated TFBRs were found to fall in the category of 3TF bound regions, and more than half of them also occurred in CREs, primarily enhancers (Fig. 4.17). This was not expected as TFBRs in the 3TF category had lower mutation rates than the 1TF and 2TF groups of TFBRs (Fig. 4.8b), while CRE overlapping TFBRs also had relatively low mutation rates (Fig. 4.8c) overall. However, it should be noted that, although the *wmr* calculation accounts for the total number of sliding trinucleotides in the aggregated cistromes (denominator in formula in Fig. 4.7), which in fact represents the length of the corresponding TFBRs, it is possible that large length differences among TFBRs may still be a confounding factor. For example, 3TF-bound TFBRs with exceedingly high length may be more likely to accumulate a high number of mutations, thus showing up as hyper-mutated TFBRs.

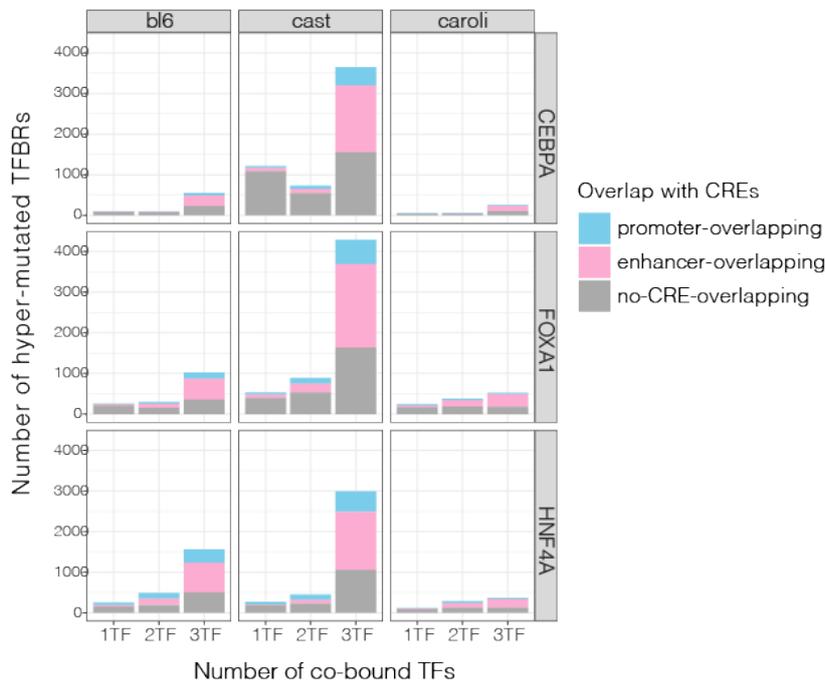


Fig. 4.17 Distribution of hyper-mutated TFBRs (TFBRs with more than 2 mutations) across the sub-cistromes.

I further sought to gain insights into the functional roles of these hyper-mutated TFBRs. For this purpose, I used GREAT (Genomic Regions Enrichment of Annotations Tool) [369] to perform Gene Ontology (GO) enrichment analyses and support biological interpretation of the observations. In brief, GREAT is suitable for associating non-coding regions that have potential *cis*-regulatory functions—such as ChIP enriched regions—with genes. Specifically, for each gene in the whole genome annotation, GREAT determines by default a "gene regulatory domain" as a genomic window of 5kb upstream and 1kb downstream of the gene TSS. It then extends this basal regulatory domain to both directions until it meets the nearest gene, or until it reaches a maximum extension of length 1,000kb. It also performs further refinement by using global control regions with experimentally validated regulatory domains. Finally, it performs GO enrichment analyses using binomial tests [369], accounting for the variability of regulatory domain sizes and considering the genome fraction that is associated with a given ontology term, as well as the number of input regions that are associated with it. Statistical enrichments calculated by GREAT are therefore relative to a background set, which can be either the whole genome or a control set of certain regions. I performed a GREAT analysis for the hyper-mutated TFBRs in each species, using as a background set the corresponding non-mutated TFBRs. Finally, GREAT supports analyses on the BL6 genome assembly (mm10), but not on the CAST and CAROLI genomes. Thus, to enable analyses also in these two mouse species, I lifted over their hyper-mutated TFBRs onto the BL6 genome assembly (see Methods).

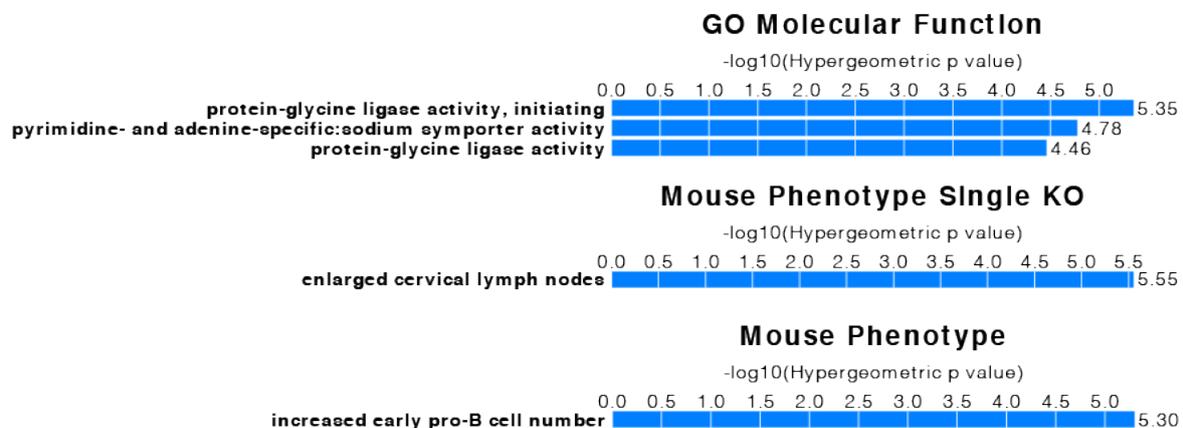
GREAT analysis of the hyper-mutated TFBRs in BL6 showed enrichment of terms related to catalytic activity (*protein-glycine ligase activity*, *protein-glycine ligase activity*, *initiating*) and transmembrane transport (*pyrimidine- and adenine-specific:sodium symporter activity*). In addition, the results included phenotype terms related to abnormal B-cell differentiation and abnormal morphology of peripheral lymph nodes (*increased early pro-B cell number*, *enlarged cervical lymph nodes*), which generally indicate abnormal morphology of the immune system (Fig. 4.18).

The hyper-mutated TFBRs in CAST showed association with catalytic activity (*chitinase activity*), transmembrane transport and signalling (*negative regulation of iron ion transmembrane transport*, *glucagon-like peptide 1 receptor activity*, *trace-amine receptor activity*, *lysophospholipid transporter activity*, *JUN kinase kinase kinase activity*, *transmembrane transporter activity*), metabolic processes (*peptide cross-linking*), response to stress (*negative regulation of transcription from RNAP II promoter in response to stress*), as well as abnormal organ development and morphology, including abnormal cardiovascular development (*absent cardiac neural crest cells*, *absent coronary sinus*, *abnormal Meissner's corpuscle morphology*), abnormal kidney morphology (*tubular nephritis*) nervous system morphology (*abnormal pacinian corpuscle morphology*, *abnormal Meissner's corpuscle morphology*), as well as abnormal immune system morphology (*lymphatic vessel hyperplasia*) (Fig. 4.19). Finally, the hyper-mutated TFBRs also showed association with specific genes involved in metabolic processes (*Plcx3*) and signalling receptor activity and response to stimulus (e.g. *Ly6c2*, *Ly6c1*, *Ly6g*, *Vmn2r20*, *Glp1r*) (Fig. 4.19). Interesting is also the gene *Cited2* (Cbp/p300-interacting transactivator 2), which

encodes a transcriptional co-activator involved in cellular processes that are known to be perturbed in carcinogenesis, such as transforming growth factor beta (TGFB) signalling and stimulation of the tumour suppressor WT1-mediated transcription activation. A more detailed investigation of these results would be of great interest.

Finally, CAROLI hyper-mutated TFBRs were associated with terms related to catalytic activity (*cholesterol 24-hydroxylase activity*, genes *Gal3st2b*, *Cyp46a1*), regulation of transcription (*chromatin silencing at rDNA*), nucleosome assembly (*DNA replication-dependent nucleosome assembly*), and response to external stimuli and stress (*defense response to bacterium*, *defense response to protozoan*, gene *Gm4951*) (Fig. 4.20). The Cholesterol 24-hydroxylase gene (*Cyp46a1*) that encodes a member of the Cytochrome P450 (CYP450) enzyme superfamily would also be of particular interest, as CYP450 is known to metabolise DEN in the hepatocytes.

Overall, the hyper-mutated TFBRs in all three species showed association with transmembrane transport and signalling, as well as catalytic activity. The results in BL6 and CAST also included immune system aberrations. Additional terms that were enriched in one or two of the species were related to metabolic processes, abnormal organ morphology, nucleosome assembly and transcriptional regulation.



**Fig. 4.18 Enriched GO terms associated with the hyper-mutated TFBRs in BL6 DEN tumours, as computed by GREAT.**

To further investigate the functional impact of the mutation load in the hyper-mutated TFBRs, I aimed to explore associations of these TFBRs with dysregulated genes in liver cancer. To address this, I utilised the differential gene expression (DGE) analyses results between tumours with different drivers and normal adult liver in each mouse species. The DGE analyses had been previously performed by the LCE consortium using DESeq2 [250]. To infer whether the hyper-mutated TFBRs are associated with dysregulated genes, I first identified the closest gene of each of the hyper-mutated TFBRs. Then,

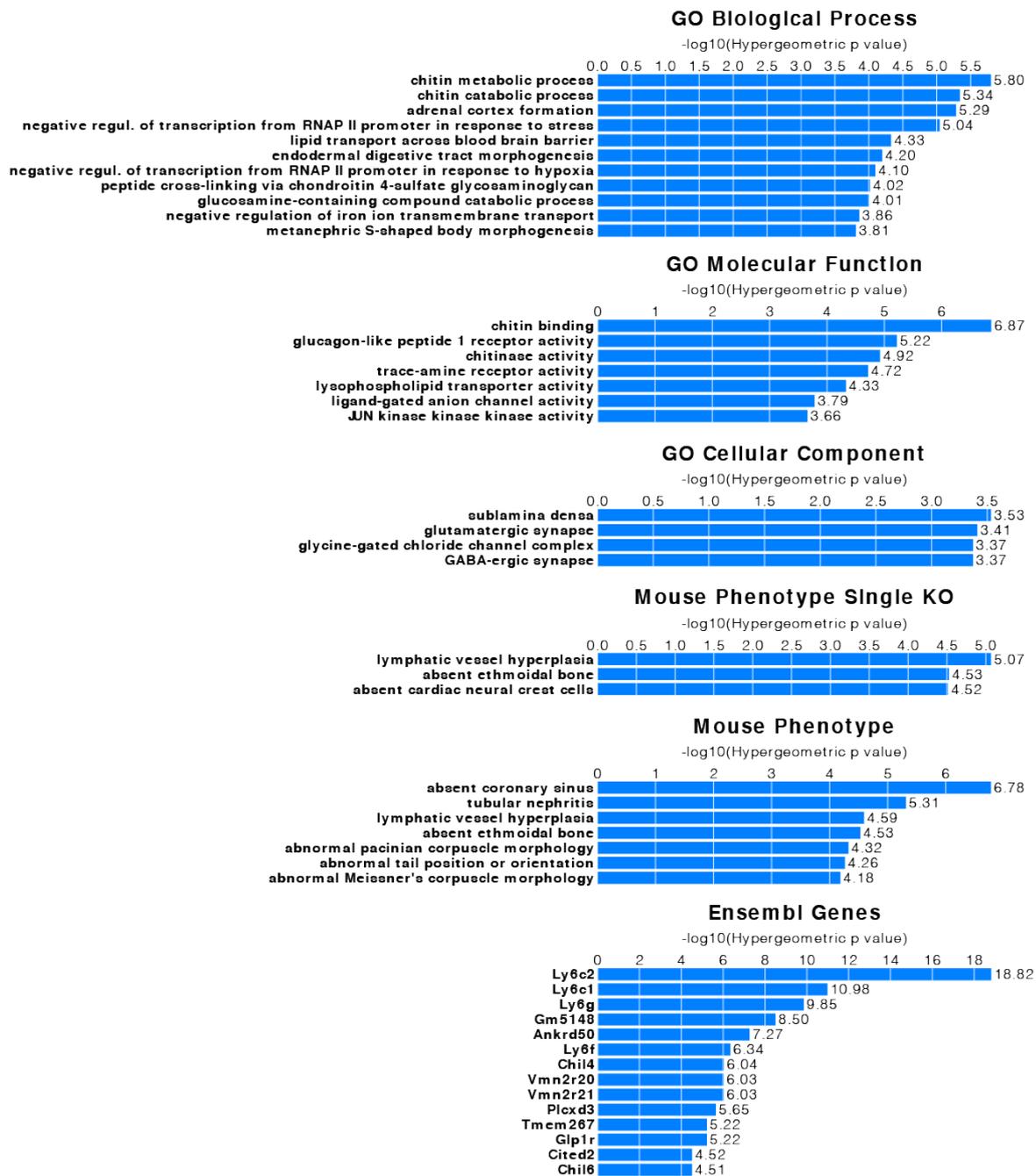


Fig. 4.19 Enriched GO terms associated with the hyper-mutated TFBRs in CAST DEN tumours, as computed by GREAT.

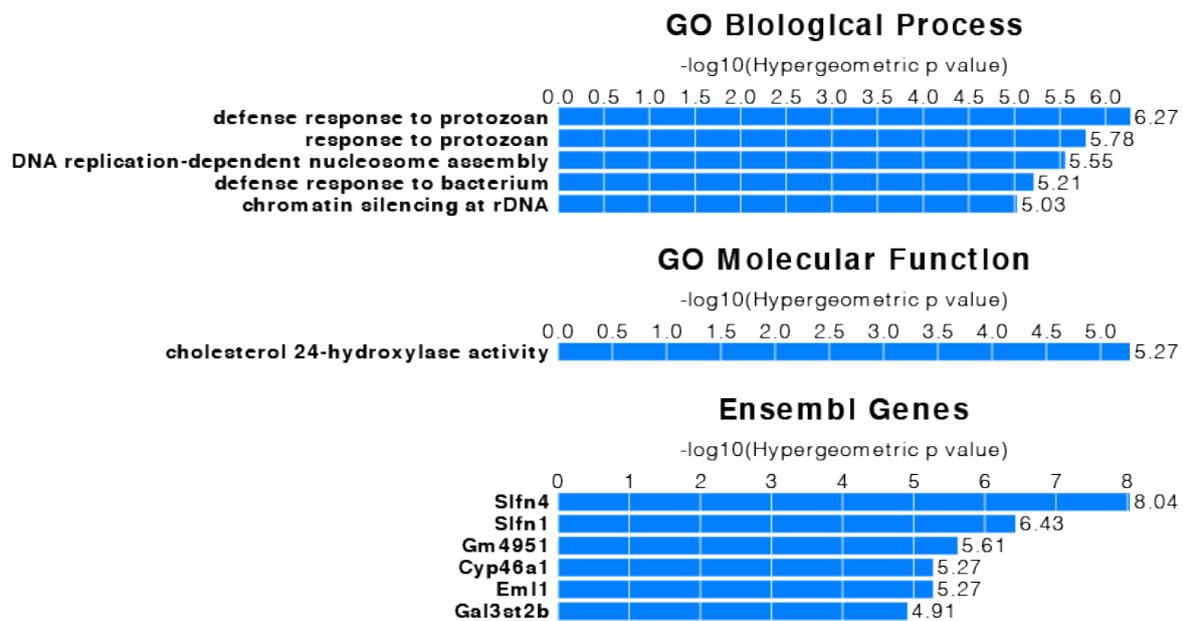
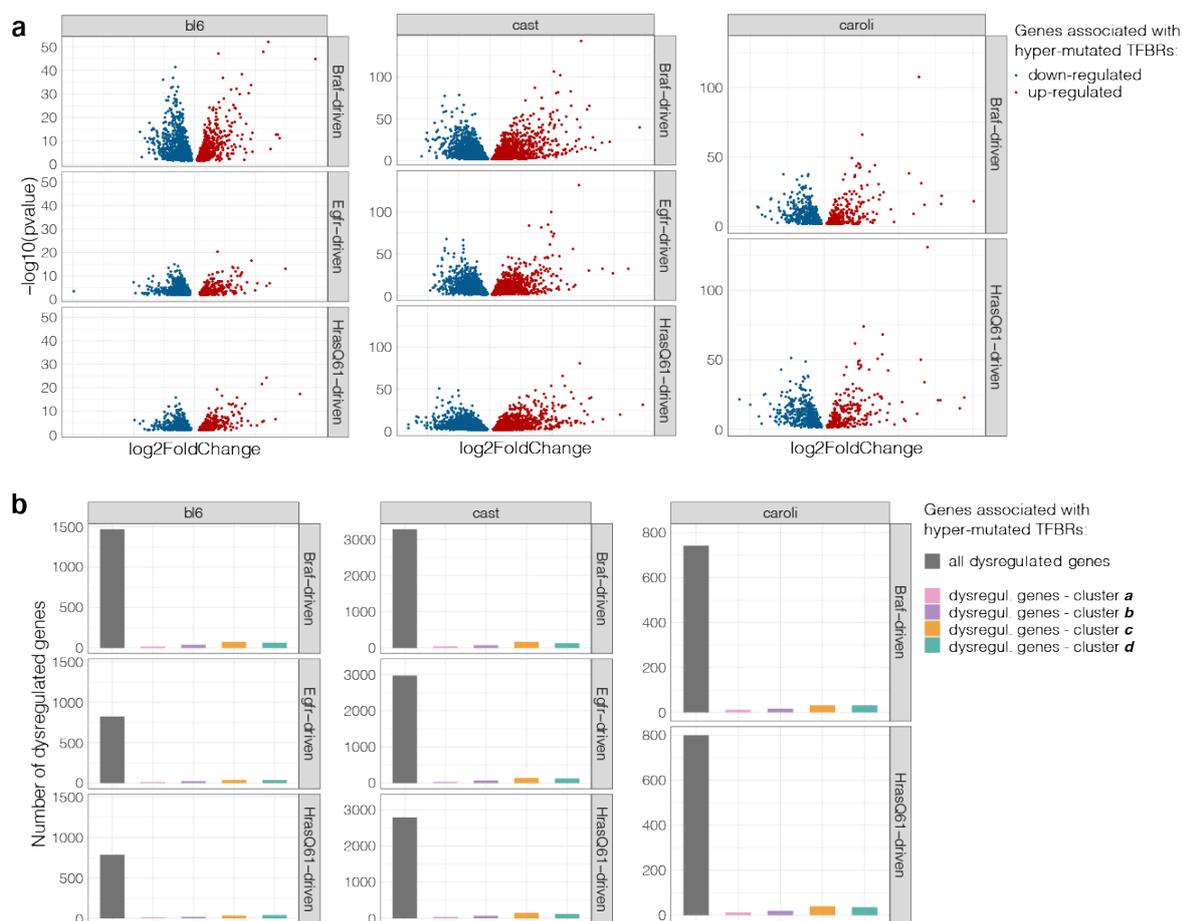


Fig. 4.20 Enriched GO terms associated with the hyper-mutated TFBRs in CAROLI DEN tumours, as computed by GREAT.

I retrieved the differential expression analysis results for the identified closest genes and visualised them (Fig. 4.21a).

Indeed, a few hundreds to thousands of genes that were associated with hyper-mutated TFBRs were significantly dysregulated in the liver tumours of each species (Fig. 4.21a). This is indicative of the functional impact of high mutation loads in this specific set of TFBRs. In addition, small subsets of these dysregulated genes were included in the cell-type marker gene clusters, clusters *a-d*, that had been described in Chapter 3 (Fig. 4.21b). This observation suggests that the mutational patterns of the specific TFBR set may be associated with gene dysregulation that is relevant to the shift of the cell phenotype.



**Fig. 4.21 Significantly dysregulated genes in DEN-induced tumours that are associated with hyper-mutated TF binding regions.** a) For each species, volcano plots of genes that are associated with hyper-mutated TFBRs and are also significantly upregulated (red) or downregulated (blue) in tumours with different driver mutations. b) Number of corresponding dysregulated genes. A small subset of all dysregulated genes that associated with hyper-mutated TFBRs are included in the cell-type marker gene clusters that were described in Chapter 3: clusters *a*, *b*, *c* and *d* correspond respectively to genes involved in cell proliferation at early embryonic stages, hepatoblast-specific functions, hepatocyte-specific functions and cholangiocyte-specific functions.

## 4.3 Discussion

In this chapter, following up the identified dysregulation of cell-type marker genes in the transforming/dedifferentiating hepatocyte throughout HCC development in three mouse species (BL6, CAST, CAROLI), I set out to explore the mutational landscapes of transactivated regulatory regions and their potential association with expression perturbations and the cell phenotype shift along neoplastic transformation. To this end, I determined the cistrome of each of three liver TFs, CEBPA, FOXA1 and HNF4A, and characterised their respective mutation rates in the DEN-induced liver tumours of each mouse species. A focus has been on distinguishing sets of TF binding regions (TFBRs) within each cistrome based on their functionally relevant features that may be associated with shaping the corresponding mutational landscapes; these were their combinatorial TF binding potential and their co-occurrence with CREs (promoters and enhancers). Evaluation of the relative mutation rates of sub-cistromes showed that combinatorially bound, as well as CRE-overlapping TFBRs display significantly lower mutation rates than other TFBR sets. Nevertheless, a set of hyper-mutated TFBRs was also identified, which was unexpectedly enriched for combinatorially bound and CRE-overlapping regions. Interestingly, functional analyses of these hyper-mutated regions demonstrated their association with transmembrane transport and signalling, response to stress, catalytic activity, as well as immune system aberrations. In addition, many of their associated genes—as inferred by linear genomic distance—were significantly dysregulated in the liver tumours. Running the analyses independently on the three different mouse species confirmed the reproducibility of the observed mutational patterns in the TFBR ensembles and revealed commonalities among the functional analyses. A potential bias source that should be taken into account before drawing final conclusions are the length differences among TFBRs and TFBR sets. Even though the TFBR lengths are accounted for (as total trinucleotide counts) in mutation rates calculation, particularly large length differences could possibly still act as a source of bias, especially when identifying TFBRs with particularly higher mutation load than other (hyper-mutated TFBRs). A better establishment of the importance of length differences with respect to the observed results would further improve the present study.

### 4.3.1 Characterisation of TF sub-cistromes based on functionally relevant features

To separate out distinct mutational patterns within identified TF cistromes, I sub-categorised each cistrome according to functionally relevant features of its contained TFBRs. In particular, based on the number of different TFs they (co-)bind, the binding regions of each TF were distinguished into 1TF-, 2TF-, or 3TF- bound. In a parallel categorisation based on whether they co-occur with *cis*-regulatory elements, they were also distinguished into promoter-overlapping, enhancer-overlapping or non-CRE-overlapping. The majority of the regions were found to present combinatorial binding of either three (to a larger extent) or two of the studied liver TFs: CEBPA, FOXA1 and HNF4A. That was

confirmed both by CHIP-seq enrichment signal for the different TFs and by the content of the TFBRs in binding motifs for each TF. In addition, about half of the binding regions of each cistrome were found to bind to CREs. Overall, liver TFs were shown to largely bind in concert to genomic regions. Evidence from previous studies using knock out experiments of one of the TFs CEBPA, FOXA1 and HNF4A in the liver of mice has shown that stabilisation of one factor at typically combinatorially bound regions depends on the binding of the other TFs [78]. In addition, compared to singly TF bound regions, combinatorially bound ones tend to be more evolutionarily conserved, as shown by the cross-species conservation of their binding activity [78, 370], as well as by lower rates of nucleotide substitutions in their motifs compared to other regions [78]. On top of that, it has been shown that species-conserved TFBRs, including many combinatorially bound TFBRs, tend to display stronger TF binding [78]. Taken together, these findings suggest that combinatorial TF binding is associated with more stable binding of the individual TFs. Also, the underlying combinatorially bound sequences are less tolerant to nucleotide substitutions, which probably reflects their functional importance. Another indication of their functional importance is the large extent of overlap ( 75%) between the set of 3TF co-bound regions with the TFBRs coinciding with promoters (Fig. 4.17). Therefore, liver TFs appear to mainly bind and act at regulatory elements in concert. The importance of combinatorial binding of TFs in governing spatial and temporal gene expression patterns has been highlighted in a number of previous studies [371, 372, 370, 373, 374, 78, 375, 376]. Especially, tissue-specific gene regulation is, to a great extent, a function of tissue-specific combinations of distinct TFs that act in concert via binding to specific TFBR sets, and thus forming tissue-specific regulatory networks [372, 376].

### **4.3.2 Mutation rates in TF (sub-)cistromes, their underlying molecular and evolutionary processes, and hyper-mutated TFBR**

Profiling the mutation rates of the whole binding region sets of the three TFs in all mouse species showed that none of the three cistromes seems to accumulate mutations at a particularly higher or lower rate in HCC development, compared to the other TFs. Although each of the three TF cistromes includes a number of singly bound TFBRs (classified in the 1TF group), the three liver TFs were shown to largely co-bind, which indicates the extent of their synergistic activities and involvement in similar regulatory functions. Therefore, the fact that they are affected in similar ways by DEN-caused mutagenesis is not surprising. Interestingly, a recent study on the binding profiles of 13 distinct TFs showed convergence of somatic mutations with germline variants at cistromes of a particular subset of these TFs that act as master transcriptional regulations in prostate cancer. This TF set also included FOXA1 [82]. It would be interesting to implement a similar approach of combining somatic with germline SNVs for the three studied liver TFs, as well as to examine and compare a larger number of TF cistromes with respect to their mutational patterns in HCC development, in order to see if any more specific patterns would be observed as in the case of prostate cancer.

With respect to comparisons among the sub-cistromes of each TF, the 3TF combinatorially bound regions displayed significantly lower mutation rates compared to regions bound by one TF alone (1TF category), in all cases except for HNF4A in CAST and CAROLI (Fig 4.2b). Similarly to the lowly mutated 3TF bound regions, significantly lower mutation rates were observed in TFBRs occurring in promoters compared to TFBRs within enhancers, which in turn were significantly less mutated than no-CRE-overlapping TFBRs. It is generally expected that the mutational landscapes of TFBRs in HCC development are forged by a number of distinct molecular processes, which may have opposing effects. These include DNA damage upon bioactivation of DEN, inefficiency of DNA repair mechanisms, as well as purifying selection/evolutionary constraints.

With respect to the DEN-caused damage, it is not considered to differentially affect distinct sets of genomic regions based on their characteristics. The load of DEN-caused lesions must be relatively homogeneous across the genome, with the only established bias being in the sequence context (e.g. more lesions on T's). It is noted, though, that mutation rates calculated in this study were normalised for sequence context biases.

Regarding the potential reduced repair activity at TFBRs, it would be expected to inflate local mutation rates. A number of studies have shown that TF binding regions exhibit high mutation rates in various cancer types [88, 84, 83]. These high TFBR mutation rates are attributed to impaired access of DNA repair enzymes to the TF-occupied genomic regions and, thus, incapacity to repair DNA lesions caused by mutagenic factors, such as DEN [88, 84, 83]. Unrepaired lesions can then be fixed to somatic mutations via faulty DNA replication during cell division.

On the other hand, assuming that many TFBRs—especially the combinatorially bound and CRE-overlapping ones— have essential regulatory roles, their observed low mutation rates may reflect the effect of purifying selection against potentially deleterious mutations in these functionally important regions. Nevertheless, this assumption would require further investigation before conclusions are drawn.

The differences in mutation rates of distinct TFBRs may reflect differences in the relative contributions of the molecular and evolutionary processes, such as lesion repair efficiency and selective pressures, that are in effect in each TFBR set with distinct functional characteristics. For example, the negative selection against mutations may be weaker in TFBRs with lower implication in gene regulatory networks, such as the 1TF-bound regions, or the ones that do not overlap with CREs. Many of these regions may even represent occasionally and transiently bound regions, without fundamental regulatory roles. A number of them could also correspond to false positive TFBRs resulting from ChIP-seq artefacts. On the contrary, functionally important TFBRs with essential regulatory roles may be under stronger evolutionary constraints, thus less tolerant to accumulating mutations. Deleterious mutations in these regions may be lethal and, thus, do not make it through tumour evolution. That is also a

reason why it would be difficult to measure the effect of purifying selection. In the case of coding regions, that could be done by comparing the rates of non-synonymous with synonymous mutations [368]. However, this approach is not as easily applicable in the case of non-coding regions that are not translated. Overall, determining the extent to which the mutation landscapes in these regions are attributed to reduced repair efficiency or to neutral selection would require further investigation.

Considering the observed low mutation rates in presumably functionally important regions, such as 3TF bound regions and CREs, the enrichment of the hyper-mutated TFBR set for such regions appears as a paradox. There are thousands of regions recurrently mutated across independent tumours in each species, and though very few are mutated in a single tumour, they harbour a remarkably high load of mutations given their short length. These hyper-mutated regions are associated with functions related to transmembrane transport and signalling, response to stress, catalytic activity, abnormalities of the immune system and metabolic processes. Perturbation of such cellular processes, e.g. immune response, metabolic processes, signalling pathways and response to stress, are generally known to characterise the development of various cancer types. In addition, the observed association with catalytic activities could be an interesting point to further examine with respect to potential enzyme activity perturbations in chemically induced HCC development. Especially the implication of the *Cyp46a1* gene, in CAROLI, that encodes for a member of the DEN-metabolising CYP450 would be of particular interest. Furthermore, most genes associated with these hyper-mutated TFBRs are found to be dysregulated in DEN tumours. Given these insights into potential functional associations of the hyper-mutated TFBRs, it is possible that these regions are associated with some functional impact along the neoplastic cell transformation. As long as mutations in these TFBRs are not lethal for the cell, relatively higher mutation loads could be tolerated. An emerging question is whether mutations in these regions could even confer fitness to the tumour cell. This could potentially be addressed via investigating whether there is a signal of positive selection in the mutations of the hyper-mutated TFBRs. Although it may be challenging to study the effects of selective pressure in non-coding regions, it is possible that machine learning could be used in order to distinguish sequence patterns that are associated to function. Such an approach has been recently implemented to detect positive selection signals in enhancers and in TF binding regions [377] [378]. This could also be adapted to investigate potential positive selection effects in the hyper-mutated TFBRs that are associated with dysregulated genes in HCC development. However, it should be highlighted again that more detailed analyses to rule out the effect of potential length and sequence biases would be encouraged in order to provide more rigorous identification and functional characterisation of hyper-mutated TFBRs and infer robust conclusions.

Other future directions of the study could focus on a more comprehensive characterisation of the set of hyper-mutated TFBRs. Within this frame, a first point to be further addressed would be to establish in more detail the association of each hyper-mutated TFBR with regulatory elements. In the case of promoter-overlapping hyper-mutated TFBRs, it would be interesting to examine whether the gene

associated to the corresponding promoter is highly or lowly expressed at the time of carcinogenic insult. This direction is motivated by the fact that the local effects of transcription coupled repair (TCR) on lesions are related to the expression levels in the region; TCR activity is known to be high at actively transcribed genomic regions. Therefore, regions with genes that are actively transcribed at the time of the carcinogen insult in the tumour cell-of-origin will exhibit a lower mutation rate in their vicinity, because lesions will be more efficiently repaired by TCR, compared to other regions with silenced genes where TCR will not be recruited. This could provide a link between the expression state of the cell-of-origin at the time of the carcinogen insult with the finally observed mutational landscape in the tumours.

Another possible direction to be pursued is a better establishment of associations between the individual hyper-mutated TFBRs and their target genes, on which they exert some regulatory effect. The approach that has been followed here was based on the linear genomic distance of each TFBR from its closest gene. However, this could be further improved by setting biologically relevant threshold on the allowed distances, or by identifying plexuses of the TFBRs of interest with their target genes. Such plexuses can be inferred from Hi-C data analysis and identification of looping interactions, particularly relevant for TFBRs within putative enhancers. In addition, comprehensive pathway enrichment analyses of the identified target genes should be performed.

One limitation of the study is the lack of TF binding profiles in tumours. Although we were able to identify genome-wide profiles of liver TF binding in normal liver and can investigate mutagenisation of these regions, we are not able to directly conclude whether there is differential TF binding at the mutated regions. In addition, we may miss newly created TF binding events at other genomic regions or losses of TF binding at others as a result of mutations, which may be implicated in the neoplastic transformation of the hepatocytes. Still, although not directly, conclusions can be drawn via inferring associations between TFBRs and dysregulated genes, as I have described in this chapter.

### **4.3.3 Mutational load of TF binding motifs and their flanking sequences**

Analyses of the TF binding motifs identified in distinct sub-cistromes with different functional characteristics (1TF, 2TF, 3TF, and promoter-overlapping, enhancer-overlapping, no-CRE-overlapping) have not shown enrichment of recurrent occurrence of mutations at particular motif positions. The approach of comparing mutation spectra in corresponding positions of motifs assigned to different functional TFBR categories was limited by the fact that it did not account for the sequence context. A possible correction would be to also calculate the expected mutation count for each nucleotide change context (based on its mutation probability) and use it as a reference to evaluate the corresponding observed mutation counts per nucleotide change context per position. However, such analysis would have reduced power, as splitting the motifs into functional categories according to the TFBRs they are

found in, combined with splitting the mutations at each motif position by nucleotide change context, results in many subsets with few data points. Therefore, an approach for calculating observed versus expected mutation counts per motif position was implemented without splitting the mutations by their nucleotide change context, yet accounting for the conditional probability of each position to be mutated based on its trinucleotide context. However, this analysis did not show any significant differences between functional categories of TFBRs with respect to mutation enrichment at any of their individual extended motif positions.

Nevertheless, even if no particular motif position of the studied TFs is susceptible to accumulating mutations in one functional TFBR category compared to another, it is possible that the whole motif sequence of a particular TFBR category is more tolerant to mutations than the motifs of another TFBR category. An approach to test this hypothesis would be to calculate weighted mutation rates (*wmr*) only for motif sequences identified in the different sub-cistromes, besides calculating the *wmr*s over aggregation of the whole TF bound regions (defined by ChIP-seq peaks) as was described in this chapter. An alternative method to evaluate mutation load in motifs and compare among distinct TFBR categories, would be to look at the relative comparison of the observed versus expected overall mutation load of the whole motif sequence, rather than looking at each individual position of a motif. In any case, it is important to conclude whether the short motif sequences that bind TFs are disrupted by mutagenesis in cancer or, in our case, in chemically induced liver cancer. Given that the TF binding motifs are the actual sequences that are explicitly recognised and bound by the TFs, dysregulation of gene expression as a result of mutagenesis in TFBRs should, in principle, be mediated by differential binding affinity of the TFs to their target motifs because of changes in their nucleotide sequence. However, another aspect to consider is that the observed phenotypic changes (cell phenotype shift in neoplastic transformation) mediated by gene dysregulation do not necessarily have to narrow down to mutation enrichment in a very particular motif sequences set of a particular functional category of TFBRs. As shown here, the identified hyper-mutated TFBR set is enriched for functional categories of TFBRs that overall have low mutation rates (3TF, CRE-overlapping). This exemplifies how individual regions with different functional characteristics can be mutated and may have a collective impact on the expression output, and in turn on the phenotype. If these regions are studied only within their predefined functional categories (i.e. within the whole set of 3TF or promoter-overlapping TFBRs), the signal of their particular mutation patterns—deviating from the "rule"—is not possible to be detected.

Overall, there might be mutations in TF binding motif sequences that mediate dysregulation of genes via differential TF binding although identifying them is challenging. Approaches such as the use of machine learning to detect sequence patterns associated with function, and thus potential signals of positive selection (as described above), may be relevant here.

### 4.3.4 Importance of studying mutagenesis of TF binding regions and general future perspectives

Overall, the results of this study disentangle different patterns of mutation accumulation among diverse categories of TF binding regions. In addition, they provide insights into how gene dysregulation via mutagenesis of TF binding regions can lead to distortion of the hepatocyte phenotype in HCC development. Importantly, they also highlight the necessity of studying the activity of TFs bound both in a singleton manner and in concert, and within their broader functional context. Partitioning the binding regions of TFs into functionally relevant categories is also very useful for identifying inherent variation in the binding activity and the underlying sequence contexts of each TF. Yet examination and comparison of TF binding regions in ensembles (such as functionally relevant categories of TFBRs) can be complemented by investigation of individual regions or region sets that seemingly deviate from the typical molecular patterns observed in the distinct TFBR categories. Finally, combined studies of somatic mutagenesis in the cisomes of higher numbers of distinct TFs, as well as in combination with germline variants are expected to provide more comprehensive insights into the role of non-coding, regulatory mutations, such as in TFBRs during cancer development.

Besides elucidating aspects of cell-type specific gene dysregulation and distortion of the cell phenotype (cell dedifferentiation) along the tumour progression, this system also provides an appropriate frame for better understanding the effect of cisome SNVs on gene expression regulation and for gaining insights into complex tissue-specific gene regulatory networks. Though not a specific cell type such as those of a cell line, the liver is quite a homogeneous tissue (75% composition of hepatocytes) which makes it a suitable system for studying cell type specific regulatory networks *in vivo* and, in a wider context, regulatory mechanisms in vertebrates.

## 4.4 Methods

### 4.4.1 Mutation calls in the DEN induced tumour samples

The SNVs had been previously identified in the mouse tumours, within the frame of the LCE consortium work, as described in Aitken et al. 2020 [296]. In brief, whole genome sequencing had been performed for the different tumour samples derived from BL6, CAST and CAROLI livers, and sequencing reads had been mapped to the reference genome assemblies GRCm38, C3H\_HeJ\_v1, CAST\_EiJ\_v1, CAROLI\_EiJ\_v1.1 (Ensembl release version 90) [379, 261] with bwa-meme (v0.7.12) [226]. Identification of single nucleotide variants (SNVs) was performed using Strelka (v2.8.4) [380] with default settings. SNVs were subjected to multiple filtering steps. Firstly, low-confidence SNV calls were identified and removed by using the gatk-tools package (version 0.2.2, <https://github.com/crukciobioinformatics/gatk-tools>). Filtering was based on low mapping and

base quality scores, proximity to alignment ends, and low absolute read counts. Germline variants (defined as those shared exclusively by tumours originating from a single mouse, or within a single litter of mice) were also removed. Also, SNVs with an allele frequency smaller than 2.5% were removed.

In addition, a number of genomic regions had been found to display low read coverage during the process of SNV calling, resulting on disability to call mutations in these regions. I considered the so-identified low coverage regions and also excluded them from all my downstream analyses, where I identified TF peaks etc.

#### 4.4.2 TF peak calling

Transcription factor binding regions were identified in the mouse genomes by analysing raw ChIP-seq data from Stefflova et al. 2013 [78]. These included two biological replicates from each mouse species. The ChIP-seq data were downloaded from ArrayExpress (E-MTAB-1414). I called TFBRs on the updated reference genome assemblies (Ensembl 90) using the raw reads from each of the two ChIP-seq libraries. Following the Stefflova et al. approach for TF peak calling, I also generated a third, technical, replicate by pooling together the reads from the two biological replicates. I performed read quality controls using FastQC (v0.11.5) (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>), which overall showed a high read quality in all libraries. The only exception was one of the two CEBPA ChIP-seq libraries from the biological replicates in CAROLI. This was probably reflected in the lower number of CEBPA peaks identified in CAROLI, compared to the other species (Fig. 4.2), in the downstream analysis. I then mapped the reads of the two biological and the one technical replicate on the corresponding genome of each species (Ensembl 90) using bwa (v0.7.17) [226]. I also discarded multiply mapped reads setting the MAPQ threshold in samtools (v1.9) equal to 1. I then called ChIP-seq peaks for each replicate using MACS2 (v2.1.2) [285]. To distinguish the real ChIP-seq signal of TF binding from the artefact signal due to genome repeats and open chromatin I used, as controls, input ChIP-seq libraries from each species, which were also retrieved from Stefflova et al. (2013) [78], and processed the same way as the ChIP-seq libraries for the different TFs. To further eliminate the ChIP-seq signal noise, Vasavi Sundaram performed an additional filtering on the identified peaks using ChIP-seq greylists that she had generated while identifying epigenomic/regulatory regions from ChIP-seq data within the LCE project. The grey lists had been generated with GreyListChIP package from R Bioconductor [381]. Finally, I identified reproducible peaks among the three replicates. In particular, I determined as reproducible the peaks that were shared by at least two of the three replicates.

### 4.4.3 Identification of promoters and enhancers by the LCE consortium

Promoters and enhancers had been determined, within the LCE consortium, by performing ChIP-seq experiments for histone modifications in pooled livers of healthy, 15day-old mice from each of the studied mouse species. The generated ChIP-seq data were analysed and genomic regions marked by H3K4me3 and H3K27ac, as well as regions with only H3K4me3 were identified as promoters. Regions with only H3K27ac were identified as enhancers.

### 4.4.4 Splitting the TF cistromes in sub-cistromes

I assigned every TFBRs in one of the 1TF, 2TF or 3TF groups as following: I used bedtools intersect (v2.29.2) [286] to identify if the called ChIP-seq peak of a given TF overlaps with at least one ChIP-seq peak of another TF (2TF category) or at least one ChIP-seq peak from each of the other two TFs (3TF category). It is noted that each TFBR is assigned uniquely to one of the groups. Similarly, to conclude whether each TF binding region coincides with promoters or enhancers, I checked whether there is any overlap of each ChIP-seq TF peak with any promoter or any enhancer using bedtools intersect (v2.29.2) [286].

### 4.4.5 Identification of TF-binding motifs

TF-bound motifs were identified within the ChIP-seq peak regions of each TF, using FIMO from the MEME suite (v 5.0.5) [287], with  $p$ -value threshold = 0.001, providing the genome of the corresponding species in markov background model format (option -bgfile), and setting the parameter -max\_stored\_matches = 1,000,000. In particular, the ChIP-seq peaks of CEBPA, FOXA1 and HNF4A were scanned for the position weight matrices MA0102.2 (CEBPA-binding motif), MA0148.1 (FOXA1-binding motif), MA0114.3 (HNF4A-binding motif), which I retrieved from the JASPAR database [168].

### 4.4.6 Computing and visualizing observed and expected mutations in extended motif regions

To evaluate the mutational load in the motif sequences and their flanking regions, I worked as in the mutation spectra analysis. I collected the FIMO identified motifs from the ChIP-seq peak regions, determined their middle nucleotide and extended them by  $N = 50$ bp on each side, upstream and downstream. I then intersected these extended motif regions of equal length with the mutations from

DEN-induced tumours. I retrieved the extended motif sequences with their contained mutations, and stacked them together also taking into account their orientation. That means that motifs with a reverse orientation were flipped to be added onto the stack of motif sequences with forward orientation, so that they all have the same directionality.

Having the motif sequences stacked so that the individual motif positions are aligned, I calculated the observed mutation load by summing up the number of mutations occurring at each position of the aligned aggregated sequences. As in the mutation spectra analysis, potentially overlapping motifs—especially after their up- and down-stream extension—could share a number of mutations and thus cause inflation of the mutation count estimates if their contained mutations were counted multiple times. Therefore, I aimed to make sure that each mutation was accounted for only once. To achieve this, I only considered one motif per ChIP-seq peak, that was the motif with the lowest  $p$ -value in the FIMO output.

#### 4.4.7 Calculating the expected mutation load of extended motif sequences

To calculate the expected mutation load of each position, considering also the sequence context, I first calculated the occurrence probability of each trinucleotide change in the genome. By trinucleotide change, here, I refer to the mutation -i.e. single nucleotide substitution- but considering it as a triplet with its 5' and 3' nucleotide,  $A_iX_jC_k \rightarrow A_iX_lC_k$ .

The calculation of the mutation probability per trinucleotide was performed as following:

- First, I calculated the trinucleotide counts in the genome, that means the number of occurrences of each of the 64 possible trinucleotides in the whole genome sequence,  $N(A_iX_jC_k)$ .
- Second, I computed the counts of each trinucleotide alteration in the whole genome, that is the count of a trinucleotide  $A_iX_jC_k$  (reference trinucleotide) changing into  $A_iX_lC_k$  (altered trinucleotide),  $N(A_iX_jC_k \rightarrow A_iX_lC_k)$  That resulted in a vector of 192 count values, one for every possible trinucleotide alteration.
- Then I calculated the relative frequency of each trinucleotide alteration,  $f(A_iX_jC_k)$ , by dividing the trinucleotide alteration count by the count of the reference trinucleotide in the genome  $N(A_iX_jC_k)$ :  $f(A_iX_jC_k \rightarrow A_iX_lC_k) = \frac{N(A_iX_jC_k \rightarrow A_iX_lC_k)}{N(A_iX_jC_k)}$
- Finally, I normalised the frequency of each trinucleotide alteration by dividing by the sum of the relative frequencies of all 192 trinucleotide alterations:  $\overline{f(A_iX_jC_k \rightarrow A_iX_lC_k)} = \frac{f(A_iX_jC_k \rightarrow A_iX_lC_k)}{\sum^{192} f(A_iX_jC_k \rightarrow A_iX_lC_k)}$
- Given that a trinucleotide can have three possible alterations (with its middle nucleotide changing), I determined the mutation probability of each trinucleotide as the sum of its three

normalised trinucleotide alteration frequencies. These are the mutation probabilities of the 64 trinucleotides that I used in the downstream calculations to estimate the expected mutation load

$$\text{of the extended motif regions: } P_{mut}(A_iX_jC_k) = \sum_1^{n=3} f(A_iX_jC_k \rightarrow A_iX_lC_k)$$

To calculate the expected mutation load at each position of a stacked set of extended motif sequences, first I counted the total number of observed mutations in the stacked motif region. Then I retrieved the reference fasta sequence of the region and identified all its sliding trinucleotides. To each of the determined trinucleotides, I assigned its corresponding mutation probability from the calculations described above  $P_{mut}(A_iX_jC_k)$ . Then I summed up the assigned mutation probabilities of all trinucleotides included in the fasta sequence

$$\sum_1^{seq-length} P_{mut}(A_iX_jC_k)$$

Then I divided the mutation probability

of each individual trinucleotide of the sequence by this sum of all probabilities  $\left(\frac{P_{mut}(A_iX_jC_k)}{\sum_1^{seq-length} P_{mut}(A_iX_jC_k)}\right)$ .

In that way, I converted the mutation probabilities into a system where they add up to 1. Finally, I multiplied the mutation probability of each position in the motif sequence by the total number of mutations occurring in the sequence:  $\frac{P_{mut}(A_iX_jC_k)}{\sum_1^{seq-length} P_{mut}(A_iX_jC_k)} * N(\text{mutations in sequence})$ . That is the expected mutation load of the given position in the extended motif region.

To visualise the observed together with the expected mutation counts across the extended motif regions, I made use of a python script available by Frigola et al. (2017) [368] that, besides the raw mutation counts, also plots smoothed observed and expected mutation counts using a polynomial fit.

#### 4.4.8 Functional analyses of hyper-mutated TFBRs

As GREAT (Genomic Regions Enrichment of Annotations Tool) does not support analyses on CAST and CAROLI, to enable GREAT analyses of the hyper-mutated TFBRs in these species, I lifted over the corresponding TFBR intervals from each species onto the BL6 genome assembly (mm10). For the liftovers I used the `halLiftOver` tool [382] and a multiple whole genome alignment (HAL alignment) for the mouse species, which is available within the LCE consortium. The orthologous alignments of a number of the genomic intervals from each species onto the BL6 genome contained short gaps. To avoid accounting the projection of any given TFBR interval as multiple intervals, I merged orthologously aligned TFBR intervals on BL6 genome that were lying at a distance  $d \leq 100bp$  from each other.



# Chapter 5

## Conclusions

Unraveling the big picture of phenotypic diversity among lineages, individuals, as well as within organisms requires a profound understanding of the underlying gene regulatory mechanisms and their variation. Transcriptional regulation is the most frequently employed level of gene regulation, which further encompasses a variety of mechanisms and components. Important players among them are the transcription factor (TF) proteins that bind to regulatory sequences of the genome and contribute to gene expression control.

In my thesis, I gained insights into the binding activities of particular TFs, their evolutionary dynamics and their contribution to broader functional contexts and phenotypic outputs, by investigating natural and induced variation in their binding profiles. Although I studied TF binding in different contexts, the approaches I used were based on a number of similar principles. The first was the use of *in vivo* systems, specifically mouse models, that allow studying molecular components within their complex cellular environments, and evaluation of their functional roles within a system that includes synergies and interactions with other components. Besides that, the vast majority of the data analysed here were derived from mouse liver, which makes a good model organ for studies of regulatory networks, as it is highly homogenous consisting from hepatocytes to its largest volume [375, 329]. The second important underlying principle of the approaches used was the leverage of molecular variation, manifested as variation in the sequence context and/or in the binding activity of the TFs, either among closely related species of mice (natural variation) or among different conditions under the effect of exogenous factors, such as tumour development upon exposure to a carcinogen with mutagenic effects (induced variation). The last core principle has been that TF binding was studied with respect to its role within a broader functional context or phenotypic output.

Following these principles, in the first part of my thesis, I leveraged natural variation among five mouse species to examine the evolutionary dynamics of CTCF binding, within the functional context of establishing and maintaining higher order chromatin structures and regulatory landscapes. In

the second part of my thesis, I investigated a specific phenotypic output, the hepatocyte neoplastic transformation along chemically induced HCC development, in four mouse species. I characterised the expression profile of the hepatocyte, the tumour cell-of-origin, and identified dysregulated genes along neoplastic transformation. Following this characterisation, I investigated mutational patterns in the sequence context of liver TF binding profiles, upon a carcinogen insult, and their functional implications. Below, for each of these projects, I summarise the main findings, the limitations, the future perspectives and general conclusions.

## **5.1 Evolutionary dynamics of CTCF binding at TAD boundaries**

The first study in this thesis, on evolutionary dynamics of CTCF binding within the context of three dimensional genome organisation, was motivated by limitations in the current understanding of CTCF's role in demarcating TAD boundaries. Although it is known that CTCF, together with cohesin, plays an important role in TAD establishment, evidence of TAD boundaries being robust to CTCF depletion has confounded the understanding of its exact role. To gain insights into the role of CTCF binding in TAD establishment, I assessed how CTCF binding patterns stably fixed by evolution in five mouse species contribute to the establishment and evolutionary dynamics of TAD boundaries. This study revealed that CTCF binding is maintained at TAD boundaries under a balance of selective constraints and dynamic evolutionary processes. Particularly, regardless of how conserved they are across species, CTCF binding sites at TAD boundaries were found to be subject to stronger sequence and functional constraints compared to other CTCF binding sites in the genome. The study also provided a characterisation of transposable elements enrichment at TAD boundaries compared to the background genome. Indicated depletion of LINEs and LINE-derived CTCF sites at TAD boundaries likely also reflected purifying selection against insertions of long sequences at TAD boundaries, further supporting the evolutionary constraints at these regions. Nevertheless, TAD boundaries were also shown to frequently harbour dynamic clusters of both evolutionarily old and young CTCF sites, which contribute to the resilience and flexibility of the TAD structures. Specifically, clusters of CTCF sites are often found at orthologous regions of the studied mouse species, in the vicinity of TAD boundaries. These clusters can contain both a number of species-conserved CTCF sites and divergent CTCF sites that result from repeated acquisition of new species-specific sites close to conserved ones, or species-specific losses of individual binding events. Finally, identification of pronounced co-localisation of CTCF site clusters with cohesin and their frequent proximity to gene TSSs, compared to non-clustered CTCF sites, suggests that CTCF clusters particularly contribute to cohesin stabilisation and transcriptional regulation. Overall, the study showed that dynamic conservation of CTCF site clusters is an important feature of CTCF binding evolution and apparently critical to the functional stability of higher order chromatin structure. It also highlights the necessity of examining the binding of CTCF not only as

individual binding sites, but also as an ensemble of potentially functional sites that can exert their actions synergistically, additively or interchangeably, and contribute to fostering regulatory landscapes and in turn phenotypic features.

Although the study provides novel insights into how the architecture and evolutionary dynamics of CTCF binding sites at TAD boundaries contributes to the establishment and resilience of the three-dimensional genome organisation, the extent and the depth of the study has been restricted by the lack of Hi-C in all the mouse species. In addition, the available Hi-C data in one of the species (BL6) is of relatively limited resolution. With the improvement and upscaling of the underlying technologies and experimental methods, a larger number of high resolution datasets are becoming available. More extensive comparisons of both CTCF and cohesin binding profiles combined with the use of distinct, high resolution Hi-C maps for different species could provide an ever clearer outline of the interplay of these architectural proteins with all levels of three dimensional genome structures and their functional features. This work also motivates a more detailed characterisation of the composition of distinct families and sub-families of transposable elements at TAD boundaries. This profiling would complement the understanding of evolutionary forces that drive CTCF binding and cohesin in these regions and help clarify their importance. Finally, besides the role of CTCF and cohesin in defining TAD boundaries, a general topic that requires further clarification is the extent of conservation of TADs across more evolutionarily distant lineages [383] and across tissues. Further cross-species and cross-tissue comparisons of genome-wide contact maps would further elucidate the evolutionary dynamics and functional importance of TADs.

## **5.2 Mutational landscape of TF binding profiles in hepatocellular carcinoma**

The second study included in this thesis is part of a collaborative project within the Liver Cancer Evolution (LCE) consortium. The LCE project makes use of a model of chemically induced liver carcinogenesis in different mouse species, which offers a controlled biological system with natural genomic and epigenomic variation. One of the project's main goals is to evaluate the relative contributions of the genome and the epigenome of the tumour cell-of-origin to determining the mutational landscape induced upon carcinogen exposure and how this landscape gives rise to the liver cancer phenotype. Here, I first presented my work on characterising the expression profile of the hepatocyte, as the cell-of-origin of chemically-induced liver tumours in mice. There I outlined dysregulation of genes in HCC that is associated with changes in the cell phenotype along neoplastic transformation. Following that, I explored the mutational landscapes of liver TF binding regions that underlie dysregulation of genes with hepatocyte related functions, reflected in shift of the hepatocyte phenotype along HCC development. Specifically, I characterised the mutation rates in binding region

ensembles of three liver TFs (CEBPA, FOXA1 and HNF4A) in chemically induced mouse liver tumours, considering their potential of combinatorial TF binding and their overlap with *cis*-regulatory elements (CREs; specifically promoters and enhancers). I reported that binding regions that present combinatorial TF binding and/or occur in CREs are significantly less mutated than others, possibly as a result of negative selection against mutations in these functionally important regions. However, I identified a few thousand hyper-mutated binding regions that are enriched in the combinatorially bound and CRE-overlapping binding region sets, and are associated with gene expression perturbations. These findings disentangle different mutation rates among distinct functional categories of TF binding regions in liver cancer development, and provide insights into potential associations of liver TF binding region mutagenesis with neoplastic transformation of hepatocytes. Finally, they highlight once more the necessity of studying the activity of TFs bound both in a singleton manner and in concert and within their broader functional context.

More accurate evaluation of the contribution of hyper-mutated TFBS to tumour evolution requires further evidence on their functional implications and their potential to enhance the fitness of tumour cells throughout the tumour evolution. Further functional TFBS characterisations can be performed via more precisely establishing associations of these TFBS with their target genes and with *cis*-regulatory elements, for example through additional assays of epigenomic state and/or chromatin conformation. Further evidence on whether these hyper-mutated TFBS enhance tumour fitness can be obtained by searching for signals of positive selection in the underlying mutated sequences of these TFBS. Finally, combined analyses of somatic with germline variants would be useful in addressing the interplay of somatic mutagenisation due to environmental exposures with genetic predisposition for specific cancer types.

A limitation of the study has been the lack of TF binding profiles in the liver tumours. Performing differential TF binding analyses between healthy and tumour liver would help establish the direct impact of DEN caused mutations (or other somatic mutations in tumours) on TF binding and gene expression disruption. However, TF binding profiles corresponding to healthy tissues at the carcinogenic insult can be used to infer functional implications of mutations, as shown in this study. Yet, addressing how the TF binding at mutagenised genomic regions changes in tumour would provide further proof of their functional impact.

Finally, the results presented in this study could be further expanded by including additional binding profiles of more TFs. Here, I have defined combinatorial binding based on the ChIP enrichment of three specific TFs, which does not exclude co-binding of one of these TFs with another TF not included in the study. It would also be particularly interesting to characterise binding regions of liver TFs according to co-binding with architectural TFs, such as CTCF, and in combination with chromatin looping data. That would help gain further insights in the regulatory networks in healthy tissues and

their potential disruption and implications in cancer development.

### **5.3 Final remarks**

The findings reported in this thesis provide insights into a) the role of CTCF binding and its evolutionary dynamics in establishing and maintaining topologically associating domains, and b) the mutational landscape of the binding profiles of liver TFs and their implications in neoplastic transformation and tumour evolution. Taken together, these insights highlight how the regulatory functions of a given TF binding site are exerted via a great extent of synergistic activity with other sites, co-binding of others TFs, chromatin modifications indicative of promoters and enhancers, and three dimensional chromatin conformation. Furthermore, the binding sites of TFs are evolving units that—to a large extent—contribute to the gene regulatory programme of a cell collectively. Therefore, the presented studies emphasise the necessity of examining TFs and their binding sites, both as singletons and in concert.

These general concepts can also be applied to the wider context of studying gene regulation. Overall, elucidation of regulatory mechanisms requires, on the one hand, disentanglement of their individual components and interrogation of their characteristics and functions. On the other hand, it is equally important to examine these components and their activities as an ensemble. These approaches combined can account for the context of each component's activity and for potential synergistic forces that underline relevant molecular functions that, in turn, affect the phenotype. Implementation of such approaches can enhance our understanding of the mechanisms and functions of gene regulation at a fundamental level, their evolution, as well as their roles in genetic diseases and cancer, and open new perspectives for medical applications.



# References

- [1] Satyanarayana, U. & Chakrapani, U. *Biochemistry, 5th Edition (Updated and Revised Edition)-E-Book* (Elsevier Health Sciences, 2020). URL <https://books.google.co.uk/books?id=PljwDwAAQBAJ>.
- [2] CRICK, F. H. On protein synthesis. *Symposia of the Society for Experimental Biology* **12**, 138–63 (1958). URL <http://www.ncbi.nlm.nih.gov/pubmed/13580867>.
- [3] Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–3 (1970). URL <http://www.nature.com/articles/227561a0http://www.ncbi.nlm.nih.gov/pubmed/4913914>.
- [4] Cobb, M. 60 years ago, Francis Crick changed the logic of biology. *PLOS Biology* **15**, e2003243 (2017). URL <https://dx.plos.org/10.1371/journal.pbio.2003243>.
- [5] Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003). URL <http://www.nature.com/articles/nature01763>.
- [6] Davidson, E. H. Gene Regulatory Networks and the Evolution of Animal Body Plans. *Science* **311**, 796–800 (2006). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1113832>.
- [7] Chen, K. & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs (2007).
- [8] Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics* **13**, 613–626 (2012). URL <http://www.nature.com/articles/nrg3207>.
- [9] Miglani, G. S. *Eukaryotic Gene Regulation* (Alpha Science International Limited, 2013).
- [10] Darnell, J. E. Variety in the level of gene control in eukaryotic cells. *Nature* **297**, 365–371 (1982). URL <http://www.nature.com/articles/297365a0>.
- [11] Cooper, G. M. The Cell , 2nd edition. In *The Cell: A Molecular Approach* (2000).
- [12] Margueron, R. & Reinberg, D. Chromatin structure and the inheritance of epigenetic information. *Nature Reviews Genetics* **11**, 285–296 (2010). URL <http://www.nature.com/articles/nrg2752>.
- [13] Hampsey, M. Molecular Genetics of the RNA Polymerase II General Transcriptional Machinery. *Microbiology and Molecular Biology Reviews* **62**, 465–503 (1998). URL <https://mmbr.asm.org/content/62/2/465>.

- [14] Kadonaga, J. T. Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology* **1**, 40–51 (2012). URL <http://doi.wiley.com/10.1002/wdev.21>.
- [15] Corden, J. L. RNA Polymerase II C-Terminal Domain: Tethering Transcription to Transcript and Template. *Chemical Reviews* **113**, 8423–8455 (2013). URL <https://pubs.acs.org/doi/10.1021/cr400158h>.
- [16] Haberle, V. & Stark, A. Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology* **19**, 621–637 (2018). URL <http://www.nature.com/articles/s41580-018-0028-8>.
- [17] Watson, J. D. *et al.* *Molecular Biology of the Gene (7th edition)* (2015).
- [18] Brivanlou, A. H. Signal Transduction and the Control of Gene Expression. *Science* **295**, 813–818 (2002). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1066355>.
- [19] Roeder, R. G. Transcriptional regulation and the role of diverse coactivators in animal cells. *FEBS letters* **579**, 909–15 (2005). URL <http://doi.wiley.com/10.1016/j.febslet.2004.12.007><http://www.ncbi.nlm.nih.gov/pubmed/15680973>.
- [20] Green, M. R. Eukaryotic transcription activation: right on target. *Molecular cell* **18**, 399–402 (2005). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276505012803><http://www.ncbi.nlm.nih.gov/pubmed/15893723>.
- [21] Banerji, J., Rusconi, S. & Schaffner, W. Expression of a  $\beta$ -globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981). URL <https://linkinghub.elsevier.com/retrieve/pii/009286748190413X>.
- [22] Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics* **15**, 272–286 (2014). URL <http://www.nature.com/articles/nrg3682>.
- [23] Miguel-Escalada, I., Pasquali, L. & Ferrer, J. Transcriptional enhancers: functional insights and role in human disease. *Current Opinion in Genetics & Development* **33**, 71–76 (2015). URL <https://linkinghub.elsevier.com/retrieve/pii/S0959437X15000921>.
- [24] Maeda, R. K. & Karch, F. Gene expression in time and space: additive vs hierarchical organization of cis-regulatory regions. *Current Opinion in Genetics & Development* **21**, 187–193 (2011). URL <https://linkinghub.elsevier.com/retrieve/pii/S0959437X11000335>.
- [25] Yang, J. & Corces, V. G. Chromatin Insulators: A Role in Nuclear Organization and Gene Expression. In *Advances in Cancer Research*, 43–76 (2011). URL <https://linkinghub.elsevier.com/retrieve/pii/B9780123864697000037>.
- [26] Fulton, D. L. *et al.* TFCat: the curated catalog of mouse and human transcription factors. *Genome Biology* **10**, R29 (2009). URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r29>.
- [27] Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics* **10**, 252–263 (2009). URL <http://www.nature.com/articles/nrg2538>.

- [28] Mitchell, P. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science* **245**, 371–378 (1989). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.2667136>.
- [29] Ptashne, M. & Gann, A. Transcriptional activation by recruitment. *Nature* **386**, 569–577 (1997). URL <http://www.nature.com/articles/386569a0>.
- [30] Warnmark, A., Treuter, E., Wright, A. P. H. & Gustafsson, J.-A. Activation Functions 1 and 2 of Nuclear Receptors: Molecular Strategies for Transcriptional Activation. *Molecular Endocrinology* **17**, 1901–1909 (2003). URL <https://academic.oup.com/mend/article/17/10/1901/2747428>.
- [31] Pabo, C. O. & Sauer, R. T. Transcription Factors: Structural Families and Principles of DNA Recognition. *Annual Review of Biochemistry* **61**, 1053–1095 (1992). URL <http://www.annualreviews.org/doi/10.1146/annurev.bi.61.070192.005201>.
- [32] Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. An overview of the structures of protein-DNA complexes. *Genome biology* **1**, REVIEWS001 (2000). URL <http://www.ncbi.nlm.nih.gov/pubmed/11104519><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC138832>.
- [33] Gray, P. A. Mouse Brain Organization Revealed Through Direct Genome-Scale TF Expression Analysis. *Science* **306**, 2255–2257 (2004). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1104935>.
- [34] Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. Use of the ‘Perceptron’ algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research* **10**, 2997–3011 (1982). URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/10.9.2997>.
- [35] Lambert, S. A. *et al.* The Human Transcription Factors (2018).
- [36] Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics & Development* **43**, 73–81 (2017). URL <https://linkinghub.elsevier.com/retrieve/pii/S0959437X17300059>.
- [37] Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010). URL <http://www.nature.com/articles/nature09380>.
- [38] Malik, S. & Roeder, R. G. The metazoan Mediator co-activator complex as an integrative hub for transcriptional regulation. *Nature Reviews Genetics* **11**, 761–772 (2010). URL <http://www.nature.com/articles/nrg2901>.
- [39] Fietze, S. & Farnham, P. J. Transcription Factor Effector Domains. In *Sub-Cellular Biochemistry*, 261–277 (2011). URL <http://link.springer.com/10.1007/978-90-481-9069-0{ }12>.
- [40] Akerblom, I., Slater, E., Beato, M., Baxter, J. & Mellon, P. Negative regulation by glucocorticoids through interference with a cAMP responsive enhancer. *Science* **241**, 350–353 (1988). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.2838908>.
- [41] Pan, G. *et al.* A negative feedback loop of transcription factors that controls stem cell pluripotency and self-renewal. *The FASEB Journal* **20**, 1730–1732 (2006). URL <https://onlinelibrary.wiley.com/doi/abs/10.1096/fj.05-5543fje>.

- [42] Whiteside, S. T. & Goodbourn, S. Signal transduction and nuclear targeting: regulation of transcription factor activity by subcellular localisation. *Journal of cell science* **104** ( Pt 4, 949–55 (1993). URL <http://www.ncbi.nlm.nih.gov/pubmed/8314906>.
- [43] Simon, I. *et al.* Serial Regulation of Transcriptional Regulators in the Yeast Cell Cycle. *Cell* **106**, 697–708 (2001). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867401004949>.
- [44] Furney, S. J., Higgins, D. G., Ouzounis, C. A. & López-Bigas, N. Structural and functional properties of genes involved in human cancer. *BMC Genomics* **7**, 3 (2006). URL <https://bmcbgenomics.biomedcentral.com/articles/10.1186/1471-2164-7-3>.
- [45] Bürglin, T. R. Homeodomain Subtypes and Functional Diversity. In *Sub-Cellular Biochemistry*, 95–122 (2011). URL [http://link.springer.com/10.1007/978-90-481-9069-0\\_{\\_}5](http://link.springer.com/10.1007/978-90-481-9069-0_{_}5).
- [46] Singh, H., Khan, A. A. & Dinner, A. R. Gene regulatory networks in the immune system. *Trends in Immunology* **35**, 211–218 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S147149061400057X>.
- [47] Takahashi, K. & Yamanaka, S. A decade of transcription factor-mediated reprogramming to pluripotency. *Nature Reviews Molecular Cell Biology* **17**, 183–193 (2016). URL <http://www.nature.com/articles/nrm.2016.8>.
- [48] Buganim, Y. *et al.* Direct Reprogramming of Fibroblasts into Embryonic Sertoli-like Cells by Defined Factors. *Cell Stem Cell* **11**, 373–386 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S1934590912004778>.
- [49] Huang, P. *et al.* Induction of functional hepatocyte-like cells from mouse fibroblasts by defined factors. *Nature* **475**, 386–389 (2011). URL <http://www.nature.com/articles/nature10116>.
- [50] Ieda, M. *et al.* Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors. *Cell* **142**, 375–386 (2010). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867410007713>.
- [51] Kajimura, S. *et al.* Initiation of myoblast to brown fat switch by a PRDM16–C/EBP- $\beta$  transcriptional complex. *Nature* **460**, 1154–1158 (2009). URL <http://www.nature.com/articles/nature08262>.
- [52] Marro, S. *et al.* Direct Lineage Conversion of Terminally Differentiated Hepatocytes to Functional Neurons. *Cell Stem Cell* **9**, 374–382 (2011). URL <https://linkinghub.elsevier.com/retrieve/pii/S1934590911004334>.
- [53] Pang, Z. P. *et al.* Induction of human neuronal cells by defined transcription factors. *Nature* **476**, 220–223 (2011). URL <http://www.nature.com/articles/nature10202>.
- [54] Sekiya, S. & Suzuki, A. Direct conversion of mouse fibroblasts to hepatocyte-like cells by defined factors. *Nature* **475**, 390–393 (2011). URL <http://www.nature.com/articles/nature10263>.
- [55] Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126**, 663–676 (2006). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867406009767>.
- [56] Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010). URL <http://www.nature.com/articles/nature08797>.

- [57] Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise Reprogramming of B Cells into Macrophages. *Cell* **117**, 663–676 (2004). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867404004192>.
- [58] Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J. & Melton, D. A. In vivo reprogramming of adult pancreatic exocrine cells to  $\beta$ -cells. *Nature* **455**, 627–632 (2008). URL <http://www.nature.com/articles/nature07314>.
- [59] Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000 (1987). URL <https://linkinghub.elsevier.com/retrieve/pii/009286748790585X>.
- [60] Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867413002031>.
- [61] Bai, X. *et al.* TIF1 $\gamma$  Controls Erythroid Cell Fate by Regulating Transcription Elongation. *Cell* **142**, 133–143 (2010). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867410005647>.
- [62] Park, K.-S. *et al.* Transcription Elongation Factor Tcea3 Regulates the Pluripotent Differentiation Potential of Mouse Embryonic Stem Cells Via the Lefty1 -Nodal-Smad2 Pathway. *STEM CELLS* **31**, 282–292 (2013). URL <http://doi.wiley.com/10.1002/stem.1284>.
- [63] Oven, I. *et al.* AIRE Recruits P-TEFb for Transcriptional Elongation of Target Genes in Medullary Thymic Epithelial Cells. *Molecular and Cellular Biology* **27**, 8815–8823 (2007). URL <https://mcb.asm.org/content/27/24/8815>.
- [64] Weintraub, H. The MyoD family and myogenesis: Redundancy, networks, and thresholds. *Cell* **75**, 1241–1244 (1993). URL <https://linkinghub.elsevier.com/retrieve/pii/0092867493906103>.
- [65] Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343–348 (2011). URL <http://www.nature.com/articles/nature10532>.
- [66] Cardoso-Moreira, M. *et al.* Gene expression across mammalian organ development. *Nature* **571**, 505–509 (2019). URL <http://www.nature.com/articles/s41586-019-1338-5>.
- [67] de Mendoza, A. & Seb e-Pedr os, A. Origin and evolution of eukaryotic transcription factors. *Current Opinion in Genetics & Development* **58-59**, 25–32 (2019). URL <https://linkinghub.elsevier.com/retrieve/pii/S0959437X1830128X>.
- [68] Iyer, L. M. & Aravind, L. Insights from the architecture of the bacterial transcription apparatus. *Journal of Structural Biology* **179**, 299–319 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S1047847711003613>.
- [69] Talbert, P. B., Meers, M. P. & Henikoff, S. Old cogs, new tricks: the evolution of gene expression in a chromatin context (2019).
- [70] Grau-Bov e, X. *et al.* Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife* (2017).
- [71] B urglin, T. R. & Affolter, M. Homeodomain proteins: an update. *Chromosoma* **125**, 497–521 (2016). URL <http://link.springer.com/10.1007/s00412-015-0543-8>.

- [72] Joly-Lopez, Z., Hoen, D. R., Blanchette, M. & Bureau, T. E. Phylogenetic and Genomic Analyses Resolve the Origin of Important Plant Genes Derived from Transposable Elements. *Molecular Biology and Evolution* **33**, 1937–1956 (2016). URL <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw067>.
- [73] Lin, R. *et al.* Transposase-Derived Transcription Factors Regulate Light Signaling in Arabidopsis. *Science* **318**, 1302–1305 (2007). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1146281>.
- [74] Babu, M. M., Iyer, L. M., Balaji, S. & Aravind, L. The natural history of the WRKY–GCM1 zinc fingers and the relationship between transcription factors and transposons. *Nucleic Acids Research* **34**, 6505–6520 (2006). URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl888>.
- [75] Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* **9**, 397–405 (2008). URL <http://www.nature.com/articles/nrg2337>.
- [76] Villar, D., Flicek, P. & Odom, D. T. Evolution of transcription factor binding in metazoans — mechanisms and functional implications. *Nature Reviews Genetics* **15**, 221–233 (2014). URL <http://www.nature.com/articles/nrg3481>.
- [77] Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science (New York, N.Y.)* **328**, 1036–40 (2010). URL <http://science.sciencemag.org/content/328/5981/1036.abstract>.
- [78] Stefflova, K. *et al.* Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. *Cell* **154**, 530–540 (2013). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867413008416>.
- [79] Boyadjiev, S. & Jabs, E. Online Mendelian Inheritance in Man (OMIM) as a knowledgebase for human developmental disorders. *Clinical Genetics* **57**, 253–266 (2001). URL <http://doi.wiley.com/10.1034/j.1399-0004.2000.570403.x>.
- [80] Bhagwat, A. S. & Vakoc, C. R. Targeting Transcription Factors in Cancer. *Trends in Cancer* **1**, 53–65 (2015). URL <https://linkinghub.elsevier.com/retrieve/pii/S2405803315000023>.
- [81] Zhou, S. *et al.* Noncoding mutations target cis-regulatory elements of the FOXA1 plexus in prostate cancer. *Nature Communications* **11**, 441 (2020). URL <http://www.nature.com/articles/s41467-020-14318-9>.
- [82] Mazrooei, P. *et al.* Cistrome Partitioning Reveals Convergence of Somatic Mutations and Risk Variants on Master Transcription Regulators in Primary Prostate Tumors. *Cancer Cell* (2019).
- [83] Kaiser, V. B., Taylor, M. S. & Semple, C. A. Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS Genetics* (2016).
- [84] Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* (2016).
- [85] Poulos, R. C. *et al.* Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. *Cell Reports* **17**, 2865–2872 (2016). URL <https://linkinghub.elsevier.com/retrieve/pii/S221112471631628X>.

- [86] Perera, D. *et al.* Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. *Nature* **532**, 259–263 (2016). URL <http://www.nature.com/articles/nature17437>.
- [87] Umer, H. M. *et al.* A Significant Regulatory Mutation Burden at a High-Affinity Position of the CTCF Motif in Gastrointestinal Cancers. *Human Mutation* **37**, 904–913 (2016). URL <http://doi.wiley.com/10.1002/humu.23014>.
- [88] Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nature Genetics* **advance on**, 818–21 (2015). URL <http://dx.doi.org/10.1038/ng.3335> { % } 5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26053496.
- [89] Pagès, V. & Fuchs, R. P. How DNA lesions are turned into mutations within cells? *Oncogene* **21**, 8957–8966 (2002). URL <http://www.nature.com/articles/1206006>.
- [90] Marteijn, J. A., Lans, H., Vermeulen, W. & Hoeijmakers, J. H. J. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature Reviews Molecular Cell Biology* **15**, 465–481 (2014). URL <http://www.nature.com/articles/nrm3822>.
- [91] Gillet, L. C. J. & Schärer, O. D. Molecular Mechanisms of Mammalian Global Genome Nucleotide Excision Repair. *Chemical Reviews* **106**, 253–276 (2006). URL <https://pubs.acs.org/doi/10.1021/cr040483f>.
- [92] Scrima, A. *et al.* Structural Basis of UV DNA-Damage Recognition by the DDB1–DDB2 Complex. *Cell* **135**, 1213–1223 (2008). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867408013834>.
- [93] Yang, W. Structure and mechanism for DNA lesion recognition. *Cell Research* **18**, 184–197 (2008). URL <http://www.nature.com/articles/cr2007116>.
- [94] Svejstrup, J. Q. Mechanisms of transcription-coupled DNA repair. *Nature Reviews Molecular Cell Biology* **3**, 21–29 (2002). URL <http://www.nature.com/articles/nrm703>.
- [95] Mellon, I. & Hanawalt, P. C. Induction of the Escherichia coli lactose operon selectively increases repair of its transcribed DNA strand. *Nature* (1989).
- [96] Evans, E. Mechanism of open complex and dual incision formation by human nucleotide excision repair factors. *The EMBO Journal* **16**, 6559–6573 (1997). URL <http://emboj.embopress.org/cgi/doi/10.1093/emboj/16.21.6559>.
- [97] Moser, J. *et al.* Sealing of Chromosomal DNA Nicks during Nucleotide Excision Repair Requires XRCC1 and DNA Ligase III $\alpha$  in a Cell-Cycle-Specific Manner. *Molecular Cell* **27**, 311–323 (2007). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276507004042>.
- [98] Ogi, T. *et al.* Three DNA Polymerases, Recruited by Different Mechanisms, Carry Out NER Repair Synthesis in Human Cells. *Molecular Cell* (2010).
- [99] Scharer, O. D. Nucleotide Excision Repair in Eukaryotes. *Cold Spring Harbor Perspectives in Biology* **5**, a012609–a012609 (2013). URL <http://cshperspectives.cshlp.org/lookup/doi/10.1101/cshperspect.a012609>.
- [100] Vermeulen, W. & Fousteri, M. Mammalian transcription-coupled excision repair. *Cold Spring Harbor Perspectives in Biology* (2013).

- [101] Nelson, H. C. M., Finch, J. T., Luisi, B. F. & Klug, A. The structure of an oligo(dA)-oligo(dT) tract and its biological implications. *Nature* **330**, 221–226 (1987). URL <http://www.nature.com/articles/330221a0>.
- [102] Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006). URL <http://www.nature.com/articles/nature04979>.
- [103] Jabbari, K. & Bernardi, G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* **333**, 143–149 (2004). URL <https://linkinghub.elsevier.com/retrieve/pii/S0378111904000836>.
- [104] Gardiner-Garden, M. & Frommer, M. CpG Islands in vertebrate genomes. *Journal of Molecular Biology* **196**, 261–282 (1987). URL <https://linkinghub.elsevier.com/retrieve/pii/0022283687906899>.
- [105] Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490–495 (2011). URL <http://www.nature.com/articles/nature10716>.
- [106] Ng, H.-H. *et al.* MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nature Genetics* **23**, 58–61 (1999). URL <http://www.nature.com/articles/ng0999{ }58>.
- [107] Tate, P. H. & Bird, A. P. Effects of DNA methylation on DNA-binding proteins and gene expression. *Current Opinion in Genetics & Development* **3**, 226–231 (1993). URL <https://linkinghub.elsevier.com/retrieve/pii/0959437X9390027M>.
- [108] Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011). URL <http://www.nature.com/articles/nature10442>.
- [109] Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.aaj2239>.
- [110] Marasca, F., Bodega, B. & Orlando, V. How Polycomb-Mediated Cell Memory Deals With a Changing Environment. *BioEssays* **40**, 1700137 (2018). URL <http://doi.wiley.com/10.1002/bies.201700137>.
- [111] Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nature Reviews Genetics* **17**, 551–565 (2016). URL <http://www.nature.com/articles/nrg.2016.83>.
- [112] Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009). URL <http://www.nature.com/articles/nature08514>.
- [113] Bender, C. M. *et al.* Roles of Cell Division and Gene Transcription in the Methylation of CpG Islands. *Molecular and Cellular Biology* **19**, 6690–6698 (1999). URL <https://mcb.asm.org/content/19/10/6690>.
- [114] Varley, K. E. *et al.* Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Research* **23**, 555–567 (2013). URL <http://genome.cshlp.org/cgi/doi/10.1101/gr.147942.112>.
- [115] Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology* **20**, 590–607 (2019). URL <http://www.nature.com/articles/s41580-019-0159-6>.

- [116] Edwards, J. R., Yarychkivska, O., Boulard, M. & Bestor, T. H. DNA methylation and DNA methyltransferases. *Epigenetics & Chromatin* **10**, 23 (2017). URL <http://epigeneticsandchromatin.biomedcentral.com/articles/10.1186/s13072-017-0130-8>.
- [117] Johnson, D. G. & Dent, S. Y. Chromatin: Receiver and Quarterback for Cellular Signals. *Cell* **152**, 685–689 (2013). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867413000688>.
- [118] Stavreva, D. A. & Hager, G. L. Chromatin structure and gene regulation: a dynamic view of enhancer function. *Nucleus* **6**, 442–448 (2015). URL <https://www.tandfonline.com/doi/full/10.1080/19491034.2015.1107689>.
- [119] Richmond, T. J. & Davey, C. A. The structure of DNA in the nucleosome core. *Nature* **423**, 145–150 (2003). URL <http://www.nature.com/articles/nature01595>.
- [120] Kornberg, R. D. & Lorch, Y. Twenty-Five Years of the Nucleosome, Fundamental Particle of the Eukaryote Chromosome. *Cell* **98**, 285–294 (1999). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867400819583>.
- [121] Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867407001845>.
- [122] Kimura, H. Histone modifications for human epigenome analysis. *Journal of Human Genetics* **58**, 439–445 (2013). URL <http://www.nature.com/articles/jhg201366>.
- [123] Wei, S. *et al.* Histone methylation in DNA repair and clinical practice: new findings during the past 5-years. *Journal of Cancer* **9**, 2072–2081 (2018). URL <http://www.jcancer.org/v09p2072.htm>.
- [124] Clapier, C. R. & Cairns, B. R. The Biology of Chromatin Remodeling Complexes. *Annual Review of Biochemistry* **78**, 273–304 (2009). URL <http://www.annualreviews.org/doi/10.1146/annurev.biochem.77.062706.153223>.
- [125] Lee, C.-K., Shibata, Y., Rao, B., Strahl, B. D. & Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics* **36**, 900–905 (2004). URL <http://www.nature.com/articles/ng1400>.
- [126] Liu, G., Liu, G.-J., Tan, J.-X. & Lin, H. DNA physical properties outperform sequence compositional information in classifying nucleosome-enriched and -depleted regions. *Genomics* **111**, 1167–1175 (2019). URL <https://linkinghub.elsevier.com/retrieve/pii/S088875431830140X>.
- [127] Phillips-Cremins, J. E. Unraveling architecture of the pluripotent genome. *Current Opinion in Cell Biology* **28**, 96–104 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S0955067414000489>.
- [128] Hansen, A. S., Cattoglio, C., Darzacq, X. & Tjian, R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9**, 20–32 (2018). URL <https://doi.org/10.1080/19491034.2017.1389365>.
- [129] Merkenschlager, M. & Nora, E. P. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annual Review of Genomics and Human Genetics* **17**, 17–43 (2016). URL <http://www.annualreviews.org/doi/10.1146/annurev-genom-083115-022339>.

- [130] Ruiz-Velasco, M. & Zaugg, J. B. Structure meets function: How chromatin organisation conveys functionality. *Current Opinion in Systems Biology* **1**, 129–136 (2017). URL <http://dx.doi.org/10.1016/j.coisb.2017.01.003>.
- [131] Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009). URL <http://www.sciencemag.org/cgi/doi/10.1126/science.1181369>. arXiv:1011.1669v3.
- [132] Zhang, Y. *et al.* Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908–921 (2012).
- [133] Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012). URL <http://dx.doi.org/10.1038/nature11082>. 1206.5533.
- [134] Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014). 1206.5533.
- [135] Nora, E. P. *et al.* Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature* **485**, 381–385 (2012). URL <http://www.nature.com/articles/nature11049>.
- [136] Cremer, T. & Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics* **2**, 292–301 (2001). URL <http://www.nature.com/articles/35066075>.
- [137] Sexton, T., Schober, H., Fraser, P. & Gasser, S. M. Gene regulation through nuclear organization. *Nature Structural & Molecular Biology* **14**, 1049–1055 (2007). URL <http://www.nature.com/articles/nsmb1324>.
- [138] Bickmore, W. A. The Spatial Organization of the Human Genome. *Annual Review of Genomics and Human Genetics* **14**, 67–84 (2013). URL <http://www.annualreviews.org/doi/10.1146/annurev-genom-091212-153515>.
- [139] van Steensel, B. & Dekker, J. Genomics tools for unraveling chromosome architecture. *Nature Biotechnology* **28**, 1089–1095 (2010). URL <http://www.nature.com/doi/10.1038/nbt.1680>. NIHMS150003.
- [140] Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nature reviews. Genetics* **19**, 789–800 (2018). URL <http://www.nature.com/articles/s41576-018-0060-8><http://www.ncbi.nlm.nih.gov/pubmed/30367165><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6312108>.
- [141] Blackwood, E. M. Going the Distance: A Current View of Enhancer Action. *Science* **281**, 60–63 (1998). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.281.5373.60>.
- [142] Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews Genetics* **7**, 703–713 (2006). URL <http://www.nature.com/articles/nrg1925>.
- [143] Schleif, R. DNA Looping. *Annual Review of Biochemistry* **61**, 199–223 (1992). URL <http://biochem.annualreviews.org/cgi/doi/10.1146/annurev.biochem.61.1.199>.
- [144] Wijgerde, M., Grosveld, F. & Fraser, P. Transcription complex stability and chromatin dynamics in vivo. *Nature* **377**, 209–213 (1995). URL <http://www.nature.com/articles/377209a0>.

- [145] Dillon, N., Trimborn, T., Strouboulis, J., Fraser, P. & Grosveld, F. The Effect of Distance on Long-Range Chromatin Interactions. *Molecular Cell* **1**, 131–139 (1997). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276500800143>.
- [146] Pombo, A. & Dillon, N. Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology* **16**, 245–257 (2015). URL <http://dx.doi.org/10.1038/nrm3965><http://www.nature.com/articles/nrm3965>.
- [147] Splinter, E. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & Development* **20**, 2349–2354 (2006). URL <http://www.genesdev.org/cgi/doi/10.1101/gad.399506>.
- [148] Hou, C., Zhao, H., Tanimoto, K. & Dean, A. CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proceedings of the National Academy of Sciences* **105**, 20398–20403 (2008). URL <http://www.pnas.org/cgi/doi/10.1073/pnas.0808506106>.
- [149] Phillips, J. E. & Corces, V. G. CTCF: Master Weaver of the Genome. *Cell* **137**, 1194–1211 (2009). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867409006990>.
- [150] Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* (2015).
- [151] Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900–910 (2015). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867415009150>.
- [152] Nasmyth, K. & Haering, C. H. Cohesin: Its Roles and Mechanisms. *Annual Review of Genetics* **43**, 525–558 (2009). URL <http://www.annualreviews.org/doi/10.1146/annurev-genet-102108-134233>.
- [153] Gligoris, T. G. *et al.* Closing the cohesin ring: Structure and function of its Smc3-kleisin interface. *Science* **346**, 963–967 (2014). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1256917>.
- [154] Huis in 't Veld, P. J. *et al.* Characterization of a DNA exit gate in the human cohesin ring. *Science* **346**, 968–972 (2014). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1256904>.
- [155] Faure, A. J. *et al.* Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Research* **22**, 2163–2175 (2012). URL <http://genome.cshlp.org/cgi/doi/10.1101/gr.136507.111>.
- [156] Pugacheva, E. M. *et al.* CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proceedings of the National Academy of Sciences of the United States of America* (2020).
- [157] Filippova, G. N. *et al.* A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers. *Genes, Chromosomes and Cancer* **22**, 26–36 (1998). URL [https://onlinelibrary.wiley.com/doi/10.1002/\(SICI\)1098-2264\(199805\)22:1{ }3C26::AID-GCC4{ }3E3.0.CO;2-9](https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1098-2264(199805)22:1{ }3C26::AID-GCC4{ }3E3.0.CO;2-9).
- [158] Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease (2001).

- [159] Lobanenkov, V. V. *et al.* 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**, 1743–53 (1990). URL <http://www.ncbi.nlm.nih.gov/pubmed/2284094>.
- [160] Martinez, S. R. & Miranda, J. L. CTCF terminal segments are unstructured. *Protein Science* **19**, 1110–1116 (2010). URL <http://doi.wiley.com/10.1002/pro.367>.
- [161] Bonchuk, A. *et al.* N-terminal domain of the architectural protein CTCF has similar structural organization and ability to self-association in bilaterian organisms. *Scientific Reports* **10**, 2677 (2020). URL <http://www.nature.com/articles/s41598-020-59459-5>.
- [162] Baniahmad, A., Steiner, C., Köhne, A. C. & Renkawitz, R. Modular structure of a chicken lysozyme silencer: Involvement of an unusual thyroid hormone receptor binding site. *Cell* **61**, 505–514 (1990). URL <https://linkinghub.elsevier.com/retrieve/pii/009286749090532J>.
- [163] Bell, A. C., West, A. G. & Felsenfeld, G. The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *Cell* **98**, 387–396 (1999). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867400819674>.
- [164] Schmidt, D. *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* **148**, 335–48 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867411015078><http://www.ncbi.nlm.nih.gov/pubmed/22244452><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3368268>.
- [165] Arnold, R., Burcin, M., Kaiser, B., Muller, M. & Renkawitz, R. DNA Bending by the Silencer Protein NeP1 Is Modulated by TR and RXR. *Nucleic Acids Research* **24**, 2640–2647 (1996). URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/24.14.2640>.
- [166] Mukhopadhyay, R. The Binding Sites for the Chromatin Insulator Protein CTCF Map to DNA Methylation-Free Domains Genome-Wide. *Genome Research* **14**, 1594–1602 (2004). URL <http://www.genome.org/cgi/doi/10.1101/gr.2408304>.
- [167] Arzate-Mejía, R. G., Recillas-Targa, F. & Corces, V. G. Developing in 3D: the role of CTCF in cell differentiation. *Development* **145**, dev137729 (2018). URL <http://dev.biologists.org/lookup/doi/10.1242/dev.137729>.
- [168] Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research* **46**, D260–D266 (2018). URL <http://academic.oup.com/nar/article/46/D1/D260/4621338>.
- [169] Kim, T. H. *et al.* Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell* **128**, 1231–1245 (2007). URL <https://linkinghub.elsevier.com/retrieve/pii/S009286740700205X>.
- [170] Ong, C.-T. & Corces, V. G. CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics* **15**, 234–246 (2014). URL <http://www.nature.com/doifinder/10.1038/nrg3663>. 15334406.
- [171] Hansen, A. S. CTCF as a boundary factor for cohesin-mediated loop extrusion: evidence for a multi-step mechanism. *Nucleus* **11**, 132–148 (2020). URL <https://www.tandfonline.com/doi/full/10.1080/19491034.2020.1782024>.

- [172] Nakahashi, H. *et al.* A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Reports* **3**, 1678–1689 (2013). URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124713002064>.
- [173] Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife* **6** (2017). URL <https://elifesciences.org/articles/25776>.
- [174] Lefevre, P., Witham, J., Lacroix, C. E., Cockerill, P. N. & Bonifer, C. The LPS-Induced Transcriptional Upregulation of the Chicken Lysozyme Locus Involves CTCF Eviction and Noncoding RNA Transcription. *Molecular Cell* **32**, 129–139 (2008). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276508005467>.
- [175] Kunarso, G. *et al.* Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nature Genetics* (2010).
- [176] Bourque, G. *et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Research* **18**, 1752–1762 (2008). URL <http://genome.cshlp.org/cgi/doi/10.1101/gr.080663.108>.
- [177] Sundaram, V. *et al.* Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome research* **24**, 1963–76 (2014). URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.168872.113><http://www.ncbi.nlm.nih.gov/pubmed/25319995><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4248313>.
- [178] Yusufzai, T. M., Tagami, H., Nakatani, Y. & Felsenfeld, G. CTCF Tethers an Insulator to Subnuclear Sites, Suggesting Shared Insulator Mechanisms across Species. *Molecular Cell* **13**, 291–298 (2004). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276504000292>.
- [179] Hansen, A. S. *et al.* Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF. *Molecular Cell* **76**, 395–411.e13 (2019). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276519305945>.
- [180] Hnisz, D., Schuijers, J., Li, C. H. & Young, R. A. Regulation and Dysregulation of Chromosome Structure in Cancer. *Annual Review of Cancer Biology* **2**, 21–40 (2018). URL <http://www.annualreviews.org/doi/10.1146/annurev-cancerbio-030617-050134>.
- [181] Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014). URL <http://www.nature.com/articles/nature12912>.
- [182] Sexton, T. *et al.* Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell* **148**, 458–472 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867412000165>.
- [183] Eagen, K. P. Principles of Chromosome Architecture Revealed by Hi-C. *Trends in Biochemical Sciences* **43**, 469–478 (2018). URL <https://linkinghub.elsevier.com/retrieve/pii/S0968000418300604>.
- [184] Vietri Rudan, M. *et al.* Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Reports* **10**, 1297–1309 (2015).
- [185] van Steensel, B. & Furlong, E. E. M. The role of transcription in shaping the spatial organization of the genome. *Nature Reviews Molecular Cell Biology* (2019). URL <http://www.nature.com/articles/s41580-019-0114-6>.

- [186] Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome Research* **24**, 390–400 (2014).
- [187] Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research* **25**, 582–597 (2015). URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.185272.114>.
- [188] Ibn-Salem, J. *et al.* Deletions of chromosomal regulatory boundaries are associated with congenital disease. *Genome biology* **15**, 423 (2014).
- [189] Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).
- [190] Lupiáñez, D. G., Spielmann, M. & Mundlos, S. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics* **32**, 225–237 (2016). URL <http://dx.doi.org/10.1016/j.tig.2016.01.003>.
- [191] Zabidi, M. A. & Stark, A. Regulatory Enhancer–Core-Promoter Communication via Transcription Factors and Cofactors. *Trends in Genetics* **32**, 801–814 (2016). URL <https://linkinghub.elsevier.com/retrieve/pii/S0168952516301214>.
- [192] Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences* **112**, E6456–E6465 (2015). URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1518552112>.
- [193] Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports* **15**, 2038–2049 (2016). URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124716305307>. 15334406.
- [194] Alipour, E. & Marko, J. F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Research* (2012).
- [195] Ganji, M. *et al.* Real-time imaging of DNA loop extrusion by condensin. *Science* **360**, 102–105 (2018). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.aar7831>.
- [196] Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930–944.e22 (2017). URL <http://www.ncbi.nlm.nih.gov/pubmed/28525758><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5538188>. 095802.
- [197] Rao, S. S. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* (2017).
- [198] Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017). URL <http://www.nature.com/articles/nature24281>.
- [199] Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal* **36**, 3573–3599 (2017). URL <https://onlinelibrary.wiley.com/doi/abs/10.15252/embj.201798004>.
- [200] Gassler, J. *et al.* A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *The EMBO Journal* **36**, 3600–3618 (2017). URL <https://onlinelibrary.wiley.com/doi/abs/10.15252/embj.201798083>.
- [201] Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* (2008).

- [202] Parelho, V. *et al.* Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell* (2008).
- [203] Rubio, E. D. *et al.* CTCF physically links cohesin to chromatin. *Proceedings of the National Academy of Sciences* **105**, 8309–8314 (2008). URL <http://www.pnas.org/cgi/doi/10.1073/pnas.0801273105>.
- [204] Busslinger, G. A. *et al.* Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* **544**, 503–507 (2017). URL <http://www.nature.com/articles/nature22063>.
- [205] de Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell* **60**, 676–684 (2015). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276515007625>.
- [206] Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker’s guide to Hi-C analysis: Practical guidelines. *Methods* **72**, 65–75 (2015). URL <https://linkinghub.elsevier.com/retrieve/pii/S1046202314003582>.
- [207] Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Molecular cell* **48**, 471–84 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276512007757><http://www.ncbi.nlm.nih.gov/pubmed/23041285><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3496039>.
- [208] Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nature Reviews Genetics* **17**, 661–678 (2016). URL <http://www.nature.com/articles/nrg.2016.112>.
- [209] Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology* **17**, 743–755 (2016). URL <http://www.nature.com/articles/nrm.2016.104>.
- [210] Phillips-Cremins, J. E. *et al.* Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell* **153**, 1281–1295 (2013).
- [211] Rowley, M. J. *et al.* Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Molecular Cell* **67**, 837–852.e7 (2017). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276517305464>.
- [212] Dong, Q. *et al.* Genome-wide Hi-C analysis reveals extensive hierarchical chromatin interactions in rice. *The Plant Journal* **94**, 1141–1156 (2018). URL <http://doi.wiley.com/10.1111/tpj.13925>.
- [213] Eagen, K. P., Aiden, E. L. & Kornberg, R. D. Polycomb-mediated chromatin loops revealed by a subkilobase-resolution chromatin interaction map. *Proceedings of the National Academy of Sciences* **114**, 8764–8769 (2017). URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1701291114>.
- [214] Cubeñas-Potts, C. *et al.* Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Research* **45**, 1714–1730 (2017). URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1114>.
- [215] Naumova, N. *et al.* Organization of the Mitotic Chromosome. *Science* **342**, 948–953 (2013). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1236083>.

- [216] Nagano, T. *et al.* Cell-cycle dynamics of chromosomal organization at single-cell resolution. *Nature* **547**, 61–67 (2017). URL <http://www.nature.com/articles/nature23001>.
- [217] Oomen, M. E., Hansen, A. S., Liu, Y., Darzacq, X. & Dekker, J. CTCF sites display cell cycle-dependent dynamics in factor binding and nucleosome positioning. *Genome Research* **29**, 236–249 (2019). URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.241547.118>.
- [218] de Wit, E. TADs as the Caller Calls Them. *Journal of Molecular Biology* **432**, 638–642 (2020). URL <https://linkinghub.elsevier.com/retrieve/pii/S0022283619305923>.
- [219] Schardin, M., Cremer, T., Hager, H. D. & Lang, M. Specific staining of human chromosomes in Chinese hamster x man hybrid cell lines demonstrates interphase chromosome territories. *Human Genetics* **71**, 281–287 (1985). URL <http://link.springer.com/10.1007/BF00388452>.
- [220] Parada, L. A., McQueen, P. G. & Misteli, T. Tissue-specific spatial organization of genomes. *Genome biology* (2004).
- [221] Tanabe, H. *et al.* Evolutionary conservation of chromosome territory arrangements in cell nuclei from higher primates. *Proceedings of the National Academy of Sciences* **99**, 4424–4429 (2002). URL <http://www.pnas.org/cgi/doi/10.1073/pnas.072618599>.
- [222] Branco, M. R. & Pombo, A. Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations. *PLoS Biology* **4**, e138 (2006). URL <https://dx.plos.org/10.1371/journal.pbio.0040138>.
- [223] Sanger, F. & Coulson, A. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* **94**, 441–448 (1975). URL <https://linkinghub.elsevier.com/retrieve/pii/0022283675902132>.
- [224] Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463–5467 (1977). URL <http://www.pnas.org/cgi/doi/10.1073/pnas.74.12.5463>.
- [225] Broad Institute. Picard toolkit (2019).
- [226] Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp324>.
- [227] Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**, 357–9 (2012). URL <http://www.ncbi.nlm.nih.gov/pubmed/22388286><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3322381>.  $\{\backslash\}\{\#\}\{\backslash\}\}14603$ .
- [228] Burrows, M. & Wheeler, D. A block-sorting lossless data compression algorithm. *Algorithm, Data Compression* (1994). 0908.0239.
- [229] Ferragina, P. & Manzini, G. Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, 390–398 (IEEE Comput. Soc, 2000). URL <http://ieeexplore.ieee.org/document/892127/>.
- [230] Mimura, I., Kanki, Y., Kodama, T. & Nangaku, M. Revolution of nephrology research by deep sequencing: ChIP-seq and RNA-seq. *Kidney International* **85**, 31–38 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S008525381556153X>.

- [231] Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6** (2017). URL <https://elifesciences.org/articles/21856>.
- [232] Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications* **10**, 1930 (2019). URL <http://www.nature.com/articles/s41467-019-09982-5>.
- [233] de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes & Development* **26**, 11–24 (2012). URL <http://genesdev.cshlp.org/cgi/doi/10.1101/gad.179804.111>.
- [234] Dekker, J., Marti-Renom, M. A. & Mirny, L. A. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics* **14**, 390–403 (2013). URL <http://www.nature.com/articles/nrg3454>.
- [235] Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017). URL <http://www.nature.com/articles/nature21411>.
- [236] Wingett, S. W. *et al.* HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015). URL <https://f1000research.com/articles/4-1310/v1>.
- [237] Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **38**, 576–589 (2010). URL <https://linkinghub.elsevier.com/retrieve/pii/S1097276510003667>.
- [238] Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nature Reviews Genetics* **20**, 631–656 (2019). URL <http://www.nature.com/articles/s41576-019-0150-2>.
- [239] Morlan, J. D., Qu, K. & Sinicropi, D. V. Selective Depletion of rRNA Enables Whole Transcriptome Profiling of Archival Fixed Tissue. *PLoS ONE* **7**, e42882 (2012). URL <https://dx.plos.org/10.1371/journal.pone.0042882>.
- [240] Murata, M. *et al.* Detecting Expressed Genes Using CAGE. In *Methods in Molecular Biology*, 67–85 (2014). URL [http://link.springer.com/10.1007/978-1-4939-0805-9\\_{\\_}7](http://link.springer.com/10.1007/978-1-4939-0805-9_{_}7).
- [241] Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014). URL <http://www.nature.com/articles/nature12787>.
- [242] Fu, G. K., Hu, J., Wang, P.-H. & Fodor, S. P. A. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences* **108**, 9026–9031 (2011). URL <http://www.pnas.org/cgi/doi/10.1073/pnas.1017621108>.
- [243] Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods* **9**, 72–74 (2012). URL <http://www.nature.com/articles/nmeth.1778>.
- [244] Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36 (2013). URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r36>.
- [245] Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013).
- [246] Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014). URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt656>. 1305.3347.

- [247] Anders, S., Pyl, P. T. & Huber, W. HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics* (2015).
- [248] Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* **34**, 525–527 (2016). URL <http://www.nature.com/articles/nbt.3519>.
- [249] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* **14**, 417–419 (2017). URL <http://www.nature.com/articles/nmeth.4197>.
- [250] Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014). URL <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>.
- [251] Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010). URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp616>.
- [252] Kentepozidou, E. *et al.* Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biology* **21**, 5 (2020). URL <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1894-x>.
- [253] Aitken, S. J. *et al.* CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biology* **19**, 106 (2018). URL <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1484-3>.
- [254] Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature genetics* **47**, 598–606 (2015). URL <http://www.nature.com/articles/ng.3286><http://www.ncbi.nlm.nih.gov/pubmed/25938943>.
- [255] Sofueva, S. *et al.* Cohesin-mediated interactions organize chromosomal domain architecture. *The EMBO Journal* **32**, 3119–3129 (2013). URL <http://emboj.embopress.org/cgi/doi/10.1038/emboj.2013.237>.
- [256] Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. USA* **111**, 996–1001 (2014).
- [257] Gómez-Marín, C. *et al.* Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proceedings of the National Academy of Sciences* **112**, 7542–7547 (2015). URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1505463112>.
- [258] Barutcu, A. R., Maass, P. G., Lewandowski, J. P., Weiner, C. L. & Rinn, J. L. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nature Communications* **9** (2018). URL <http://dx.doi.org/10.1038/s41467-018-03614-0>.
- [259] Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016). URL <http://www.nature.com/articles/nature16490>.
- [260] Kubo, N. *et al.* Preservation of Chromatin Organization after Acute Loss of CTCF in Mouse Embryonic Stem Cells. *bioRxiv* **118737** (2017).

- [261] Thybert, D. *et al.* Repeat associated mechanisms of genome evolution and function revealed by the *Mus caroli* and *Mus pahari* genomes. *Genome Research* **28**, 448–459 (2018).
- [262] Schwalie, P. C. *et al.* Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biology* **14**, R148 (2013). URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-12-r148>.
- [263] Heinz, S. *et al.* Effect of natural genetic variation on enhancer selection and function. *Nature* **503**, 487–492 (2013). URL <http://www.nature.com/articles/nature12615>.
- [264] Gasch, A. P., Payseur, B. A. & Pool, J. E. The Power of Natural Variation for Model Organism Biology. *Trends in Genetics* **32**, 147–154 (2016). URL <https://linkinghub.elsevier.com/retrieve/pii/S0168952515002218>.
- [265] Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**, 1813–31 (2012). URL <http://genome.cshlp.org/cgi/doi/10.1101/gr.136184.111> <http://www.ncbi.nlm.nih.gov/pubmed/22955991> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3431496>.
- [266] Azazi, D., Mudge, J. M., Odom, D. T. & Flicek, P. Functional signatures of evolutionarily young CTCF binding sites. *BMC biology* **18**, 132 (2020). URL <https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-020-00863-8> <http://www.ncbi.nlm.nih.gov/pubmed/32988407> <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7520972>.
- [267] Cooper, G. M. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Research* **15**, 901–913 (2005). URL <http://www.genome.org/cgi/doi/10.1101/gr.3577405>.
- [268] Fudenberg, G. & Pollard, K. S. Chromatin features constrain structural variation across evolutionary timescales. *Proceedings of the National Academy of Sciences* **116**, 2175–2180 (2019). URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1808631116>.
- [269] Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems* (2016).
- [270] Stedman, W. *et al.* Cohesins localize with CTCF at the KSHV latency control region and at cellular *c-myc* and H19/Igf2 insulators. *The EMBO Journal* **27**, 654–666 (2008). URL <http://emboj.embopress.org/cgi/doi/10.1038/emboj.2008.1>.
- [271] Xiao, T., Wallace, J. & Felsenfeld, G. 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**, 2174–2183 (2011). URL <http://mcb.asm.org/cgi/doi/10.1128/MCB.05093-11>.
- [272] Chen, H., Tian, Y., Shu, W., Bo, X. & Wang, S. Comprehensive Identification and Annotation of Cell Type-Specific and Ubiquitous CTCF-Binding Sites in the Human Genome. *PLoS ONE* **7**, e41374 (2012). URL <http://dx.plos.org/10.1371/journal.pone.0041374>.
- [273] Kemp, C. J. *et al.* CTCF Haploinsufficiency Destabilizes DNA Methylation and Predisposes to Cancer. *Cell Reports* **7**, 1020–1029 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124714002915>.

- [274] Fedoriw, A. M. Transgenic RNAi Reveals Essential Function for CTCF in H19 Gene Imprinting. *Science* **303**, 238–240 (2004). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1090934>.
- [275] Wan, L. B. *et al.* Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. *Development* (2008).
- [276] Ribeiro de Almeida, C. *et al.* The DNA-Binding Protein CTCF Limits Proximal V $\kappa$  Recombination and Restricts  $\kappa$  Enhancer Interactions to the Immunoglobulin  $\kappa$  Light Chain Locus. *Immunity* **35**, 501–513 (2011). URL <https://linkinghub.elsevier.com/retrieve/pii/S1074761311003992>.
- [277] Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N. & Yagi, T. CTCF Is Required for Neural Development and Stochastic Expression of Clustered Pcdh Genes in Neurons. *Cell Reports* **2**, 345–357 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S221112471200174X>.
- [278] Lee, D. P. *et al.* Robust CTCF-Based Chromatin Architecture Underpins Epigenetic Changes in the Heart Failure Stress–Gene Response. *Circulation* **139**, 1937–1956 (2019). URL <https://www.ahajournals.org/doi/10.1161/CIRCULATIONAHA.118.036726>.
- [279] Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nature Communications* **9**, 1520 (2018). URL <http://www.nature.com/articles/s41467-018-03828-2>.
- [280] Choudhary, M. N. *et al.* Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biology* **21**, 16 (2020). URL <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1916-8>.
- [281] Davidson, I. F. *et al.* Rapid movement and transcriptional re-localization of human cohesin on DNA. *The EMBO Journal* **35**, 2671–2685 (2016). URL <https://onlinelibrary.wiley.com/doi/abs/10.15252/embj.201695402>.
- [282] Borrie, M. S., Campor, J. S., Joshi, H. & Gartenberg, M. R. Binding, sliding, and function of cohesin during transcriptional activation. *Proceedings of the National Academy of Sciences* **114**, E1062–E1071 (2017). URL <http://www.pnas.org/lookup/doi/10.1073/pnas.1617309114>.
- [283] Heinz, S. *et al.* Transcription Elongation Can Affect Genome 3D Structure. *Cell* **174**, 1522–1536.e22 (2018). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867418309759>.
- [284] Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biology* **11**, R22 (2010). URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-2-r22>.
- [285] Zhang, Y. *et al.* Model-based analysis of CHIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
- [286] Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
- [287] Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202–W208 (2009). URL <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkp335>.
- [288] Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011). URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr064>.

- [289] Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* **33**, 831–838 (2015). URL <http://dx.doi.org/10.1038/nbt.3300>. 9605103.
- [290] Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016**, bav096 (2016). URL <https://academic.oup.com/database/article-lookup/doi/10.1093/database/bav096>.
- [291] Goncalves, A. *et al.* Extensive compensatory cis-trans regulation in the evolution of mouse gene expression. *Genome Research* **22**, 2376–2384 (2012). URL <http://genome.cshlp.org/cgi/doi/10.1101/gr.142281.112>.
- [292] Wong, E. S. *et al.* Decoupling of evolutionary changes in transcription factor binding and gene expression in mammals. *Genome Research* **25**, 167–178 (2015). URL <http://genome.cshlp.org/lookup/doi/10.1101/gr.177840.114>.
- [293] Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N. & Enright, A. J. Kraken: A set of tools for quality control and analysis of high-throughput sequence data. *Methods* **63**, 41–49 (2013). URL <https://linkinghub.elsevier.com/retrieve/pii/S1046202313002399>.
- [294] Cunningham, F. *et al.* Ensembl 2019. *Nucleic Acids Research* **47**, D745–D751 (2019). URL <https://academic.oup.com/nar/article/47/D1/D745/5165265>.
- [295] Kruse, K., Hug, C. B., Hernández-Rodríguez, B. & Vaquerizas, J. M. TADtool: visual parameter identification for TAD-calling algorithms. *Bioinformatics* **32**, 3190–3192 (2016). URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw368>.
- [296] Aitken, S. J. *et al.* Pervasive lesion segregation shapes cancer genome evolution. *Nature* **583**, 265–270 (2020). URL <http://www.nature.com/articles/s41586-020-2435-1>.
- [297] Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* **68**, 394–424 (2018). URL <http://doi.wiley.com/10.3322/caac.21492>.
- [298] Dasgupta, P. *et al.* Global Trends in Incidence Rates of Primary Adult Liver Cancers: A Systematic Review and Meta-Analysis. *Frontiers in Oncology* **10** (2020). URL <https://www.frontiersin.org/article/10.3389/fonc.2020.00171/full>.
- [299] Petrick, J. L. & McGlynn, K. A. The Changing Epidemiology of Primary Liver Cancer. *Current Epidemiology Reports* **6**, 104–111 (2019). URL <http://link.springer.com/10.1007/s40471-019-00188-3>.
- [300] Yang, J. D. *et al.* A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nature Reviews Gastroenterology & Hepatology* **16**, 589–604 (2019). URL <http://www.nature.com/articles/s41575-019-0186-y>.
- [301] Llovet, J. M., Burroughs, A. & Bruix, J. Hepatocellular carcinoma. In *Lancet* (2003).
- [302] Rao, C. V., Asch, A. S. & Yamada, H. Y. Frequently mutated genes/pathways and genomic instability as prevention targets in liver cancer. *Carcinogenesis* **38**, 2–11 (2017). URL <https://academic.oup.com/carcin/article-lookup/doi/10.1093/carcin/bgw118>.
- [303] Amirouchene-Angelozzi, N., Swanton, C. & Bardelli, A. Tumor Evolution as a Therapeutic Target. *Cancer Discovery* **7**, 805–817 (2017). URL <http://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-17-0343>.

- [304] Losic, B. *et al.* Intratumoral heterogeneity and clonal evolution in liver cancer. *Nature Communications* **11**, 291 (2020). URL <http://www.nature.com/articles/s41467-019-14050-z>.
- [305] Müller, M., Bird, T. G. & Nault, J.-C. The landscape of gene mutations in cirrhosis and hepatocellular carcinoma. *Journal of Hepatology* **72**, 990–1002 (2020). URL <https://linkinghub.elsevier.com/retrieve/pii/S016882782030060X>.
- [306] Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome (2009).
- [307] Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nature Reviews Genetics* **15**, 585–598 (2014). URL <http://www.nature.com/articles/nrg3729>.
- [308] Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979–993 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867412005284>.
- [309] Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149**, 994–1007 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867412005272>.
- [310] Lee, J.-S. The mutational landscape of hepatocellular carcinoma. *Clinical and Molecular Hepatology* **21**, 220 (2015). URL <http://e-cmh.org/journal/view.php?doi=10.3350/cmh.2015.21.3.220>.
- [311] Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013). URL <http://www.nature.com/articles/nature12477>.
- [312] Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020). URL <http://www.nature.com/articles/s41586-020-1943-3>.
- [313] Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature Communications* **8**, 1315 (2017). URL <http://www.nature.com/articles/s41467-017-01358-x>.
- [314] Supek, F. & Lehner, B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* **521**, 81–84 (2015). URL <http://www.nature.com/articles/nature14173>.
- [315] Connor, F. *et al.* Mutational landscape of a chemically-induced mouse model of liver cancer. *Journal of Hepatology* (2018).
- [316] Verna, L. N-Nitrosodiethylamine mechanistic data and risk assessment: Bioactivation, DNA-adduct formation, mutagenicity, and tumor initiation. *Pharmacology & Therapeutics* **71**, 57–81 (1996). URL <https://linkinghub.elsevier.com/retrieve/pii/0163725896000629>.
- [317] Maronpot, R. R. Biological Basis of Differential Susceptibility to Hepatocarcinogenesis among Mouse Strains. *Journal of Toxicologic Pathology* **22**, 11–33 (2009). URL [https://www.jstage.jst.go.jp/article/tox/22/1/22\\_{\\_}1\\_{\\_}11/{\\_}article](https://www.jstage.jst.go.jp/article/tox/22/1/22_{_}1_{_}11/{_}article).
- [318] Hanigan, M. H., Kemp, C. J., Ginsler, J. & Drinkwater, N. R. Rapid growth of preneoplastic lesions in hepatocarcinogen-Sensitive C3H/HeJ male mice relative to C57BL/6J male mice. *Carcinogenesis* **9**, 885–890 (1988). URL <https://academic.oup.com/carcin/article-lookup/doi/10.1093/carcin/9.6.885>.

- [319] Buchmann, A., Karcier, Z., Schmid, B., Strathmann, J. & Schwarz, M. Differential selection for B-raf and Ha-ras mutated liver tumors in mice with high and low susceptibility to hepatocarcinogenesis. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **638**, 66–74 (2008). URL <https://linkinghub.elsevier.com/retrieve/pii/S0027510707003417>.
- [320] Zhu, L. *et al.* Multi-organ Mapping of Cancer Risk. *Cell* (2016).
- [321] Tummala, K. S. *et al.* Hepatocellular Carcinomas Originate Predominantly from Hepatocytes and Benign Lesions from Hepatic Progenitor Cells. *Cell Reports* (2017).
- [322] Tang, D. G. Understanding cancer stem cell heterogeneity and plasticity. *Cell Research* **22**, 457–472 (2012). URL <http://www.nature.com/articles/cr201213>.
- [323] Tanimizu, N. *et al.* Hepatic biliary epithelial cells acquire epithelial integrity but lose plasticity to differentiate into hepatocytes in vitro during development. *Journal of Cell Science* (2013).
- [324] Chen, Y., Wong, P. P., Sjeklocha, L., Steer, C. J. & Sahin, M. B. Mature hepatocytes exhibit unexpected plasticity by direct dedifferentiation into liver progenitor cells in culture. *Hepatology* **55**, 563–574 (2012). URL <http://doi.wiley.com/10.1002/hep.24712>.
- [325] Tarlow, B. D. *et al.* Bipotential Adult Liver Progenitors Are Derived from Chronically Injured Mature Hepatocytes. *Cell Stem Cell* **15**, 605–618 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S1934590914003993>.
- [326] Sia, D., Villanueva, A., Friedman, S. L. & Llovet, J. M. Liver Cancer Cell of Origin, Molecular Class, and Effects on Patient Prognosis. *Gastroenterology* **152**, 745–761 (2017). URL <https://linkinghub.elsevier.com/retrieve/pii/S0016508516355299>.
- [327] Guest, R. V. *et al.* Cell lineage tracing reveals a biliary origin of intrahepatic cholangiocarcinoma. *Cancer Research* (2014).
- [328] Yang, L. *et al.* A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. *Hepatology* **66**, 1387–1401 (2017). URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/hep.29353>.
- [329] Marcos, R., Monteiro, R. A. F. & Rocha, E. Design-based stereological estimation of hepatocyte number, by combining the smooth optical fractionator and immunocytochemistry with anti-carcinoembryonic antigen polyclonal antibodies. *Liver International* **26**, 116–124 (2006). URL <http://doi.wiley.com/10.1111/j.1478-3231.2005.01201.x>.
- [330] OINONEN, T. & LINDROS, O. K. Zonation of hepatic cytochrome P-450 expression and regulation. *Biochemical Journal* **329**, 17–35 (1998). URL <https://portlandpress.com/biochemj/article/329/1/17/33476/Zonation-of-hepatic-cytochrome-P450-expression-and>.
- [331] Fausto, N. Liver regeneration and repair: Hepatocytes, progenitor cells, and stem cells. *Hepatology* **39**, 1477–1487 (2004). URL <http://doi.wiley.com/10.1002/hep.20214>.
- [332] Miyajima, A., Tanaka, M. & Itoh, T. Stem/Progenitor Cells in Liver Development, Homeostasis, Regeneration, and Reprogramming. *Cell Stem Cell* **14**, 561–574 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S1934590914001477>.
- [333] Lu, W.-Y. *et al.* Hepatic progenitor cells of biliary origin with liver repopulation capacity. *Nature Cell Biology* **17**, 971–983 (2015). URL <http://www.nature.com/articles/ncb3203>.

- [334] Español-Suñer, R. *et al.* Liver Progenitor Cells Yield Functional Hepatocytes in Response to Chronic Liver Injury in Mice. *Gastroenterology* **143**, 1564–1575.e7 (2012). URL <https://linkinghub.elsevier.com/retrieve/pii/S0016508512012449>.
- [335] Furuyama, K. *et al.* Continuous cell supply from a Sox9-expressing progenitor zone in adult liver, exocrine pancreas and intestine. *Nature Genetics* **43**, 34–41 (2011). URL <http://www.nature.com/articles/ng.722>.
- [336] Petersen, B. Mouse A6-positive hepatic oval cells also express several hematopoietic stem cell markers. *Hepatology* **37**, 632–640 (2003). URL <http://doi.wiley.com/10.1053/jhep.2003.50104>.
- [337] Schaub, J. R., Malato, Y., Gormond, C. & Willenbring, H. Evidence against a Stem Cell Origin of New Hepatocytes in a Common Mouse Model of Chronic Liver Injury. *Cell Reports* **8**, 933–939 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S2211124714005701>.
- [338] Yanger, K. *et al.* Adult Hepatocytes Are Generated by Self-Duplication Rather than Stem Cell Differentiation. *Cell Stem Cell* **15**, 340–349 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S1934590914002513>.
- [339] Mu, X. *et al.* Hepatocellular carcinoma originates from hepatocytes and not from the progenitor/biliary compartment. *Journal of Clinical Investigation* **125**, 3891–3903 (2015). URL <https://www.jci.org/articles/view/77995>.
- [340] Shin, S. *et al.* Genetic lineage tracing analysis of the cell of origin of hepatotoxin-induced liver tumors in mice. *Hepatology* **64**, 1163–1177 (2016). URL <http://doi.wiley.com/10.1002/hep.28602>.
- [341] Jörs, S. *et al.* Lineage fate of ductular reactions in liver injury and carcinogenesis. *Journal of Clinical Investigation* **125**, 2445–2457 (2015). URL <http://www.jci.org/articles/view/78585>.
- [342] Marquardt, J. U. Deconvolution of the cellular origin in hepatocellular carcinoma: Hepatocytes take the center stage. *Hepatology* **64**, 1020–1023 (2016). URL <http://doi.wiley.com/10.1002/hep.28671>.
- [343] Tschaharganeh, D. F. *et al.* p53-Dependent Nestin Regulation Links Tumor Suppression to Cellular Plasticity in Liver Cancer. *Cell* **158**, 579–592 (2014). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867414008186>.
- [344] Yuan, S., Norgard, R. J. & Stanger, B. Z. Cellular Plasticity in Cancer. *Cancer Discovery* **9**, 837–851 (2019). URL <http://cancerdiscovery.aacrjournals.org/lookup/doi/10.1158/2159-8290.CD-19-0015>.
- [345] Hölzel, M., Bovier, A. & Tüting, T. Plasticity of tumour and immune cells: a source of heterogeneity and a cause for therapy resistance? *Nature Reviews Cancer* **13**, 365–376 (2013). URL <http://www.nature.com/articles/nrc3498>.
- [346] Boumahdi, S. & de Sauvage, F. J. The great escape: tumour cell plasticity in resistance to targeted therapy. *Nature Reviews Drug Discovery* **19**, 39–56 (2020). URL <http://www.nature.com/articles/s41573-019-0044-1>.
- [347] Li, J. & Stanger, B. Z. How Tumor Cell Dedifferentiation Drives Immune Evasion and Resistance to Immunotherapy. *Cancer Research* **80**, 4037–4041 (2020). URL <http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-20-1420>.

- [348] Murphree, A. & Benedict, W. Retinoblastoma: clues to human oncogenesis. *Science* **223**, 1028–1033 (1984). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.6320372>.
- [349] Network, C. G. A. R. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014). URL <http://www.nature.com/articles/nature13385>.
- [350] Yiu Chan, C. W., Gu, Z., Bieg, M., Eils, R. & Herrmann, C. Impact of cancer mutational signatures on transcription factor motifs in the human genome. *BMC Medical Genomics* **12**, 64 (2019). URL <https://bmcmmedgenomics.biomedcentral.com/articles/10.1186/s12920-019-0525-4>.
- [351] Elledge, S. J. & Amon, A. The BRCA1 suppressor hypothesis: An explanation for the tissue-specific tumor development in BRCA1 patients. *Cancer Cell* **1**, 129–132 (2002). URL <https://linkinghub.elsevier.com/retrieve/pii/S1535610802000417>.
- [352] Haigis, K. M., Cichowski, K. & Elledge, S. J. Tissue-specificity in cancer: The rule, not the exception. *Science* **363**, 1150–1151 (2019). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.aaw3472>.
- [353] Sonesson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521 (2016). URL <https://f1000research.com/articles/4-1521/v2>.
- [354] Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: A general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biology* (2016).
- [355] Arnedo-Pac, C., Mularoni, L., Muiños, F., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUSTL: a sequence-based clustering method to identify cancer drivers. *Bioinformatics* **35**, 4788–4790 (2019). URL <https://academic.oup.com/bioinformatics/article/35/22/4788/5522012>.
- [356] Killela, P. J. *et al.* TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proceedings of the National Academy of Sciences* **110**, 6021–6026 (2013). URL <http://www.pnas.org/cgi/doi/10.1073/pnas.1303607110>.
- [357] Huang, F. W. *et al.* Highly Recurrent TERT Promoter Mutations in Human Melanoma. *Science* **339**, 957–959 (2013). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1229259>.
- [358] Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a non-coding intergenic element. *Science* **346**, 1373–1377 (2014). URL <https://www.sciencemag.org/lookup/doi/10.1126/science.1259037>.
- [359] Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nature Reviews Genetics* **17**, 93–108 (2016). URL <http://www.nature.com/articles/nrg.2015.17>.
- [360] Kumar, S. *et al.* Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell* (2020).
- [361] McFarland, C. D., Korolev, K. S., Kryukov, G. V., Sunyaev, S. R. & Mirny, L. A. Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences* **110**, 2910–2915 (2013). URL <http://www.pnas.org/cgi/doi/10.1073/pnas.1213968110>.

- [362] Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nature Reviews Cancer* **15**, 680–685 (2015). URL <http://www.nature.com/articles/nrc3999>.
- [363] Warrell, J. & Gerstein, M. Cyclic and multilevel causation in evolutionary processes. *Biology & Philosophy* **35**, 50 (2020). URL <http://link.springer.com/10.1007/s10539-020-09753-3>.
- [364] Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biology* **12**, R83 (2011). URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-8-r83>.
- [365] Sallari, R. *et al.* Convergence of dispersed regulatory mutations predicts driver genes in prostate cancer. *bioRxiv* (2016).
- [366] Nie, Y., Shu, C. & Sun, X. Cooperative binding of transcription factors in the human genome. *Genomics* **112**, 3427–3434 (2020). URL <https://linkinghub.elsevier.com/retrieve/pii/S0888754319302587>.
- [367] Roller, M. *et al.* LINE elements are a reservoir of regulatory potential in mammalian genomes. *bioRxiv* (2020). URL <https://www.biorxiv.org/content/early/2020/05/31/2020.05.31.126169>.
- [368] Frigola, J. *et al.* Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics* **49**, 1684–1692 (2017). URL <http://www.nature.com/articles/ng.3991>.
- [369] McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology* **28**, 495–501 (2010). URL <http://www.nature.com/articles/nbt.1630>.
- [370] He, Q. *et al.* High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nature Genetics* **43**, 414–420 (2011). URL <http://www.nature.com/articles/ng.808>.
- [371] Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M. & Furlong, E. E. M. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **462**, 65–70 (2009). URL <http://www.nature.com/articles/nature08531>.
- [372] Ravasi, T. *et al.* An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell* **140**, 744–752 (2010). URL <https://linkinghub.elsevier.com/retrieve/pii/S0092867410000796>.
- [373] Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012). URL <http://www.nature.com/articles/nature11245>.
- [374] Hardison, R. C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nature Reviews Genetics* **13**, 469–483 (2012). URL <http://www.nature.com/articles/nrg3242>.
- [375] Ballester, B. *et al.* Multi-species, multi-transcription factor binding highlights conserved control of tissue-specific biological pathways. *eLife* **3** (2014). URL <https://elifesciences.org/articles/02626>.
- [376] Zhou, Q. *et al.* A mouse tissue transcription factor atlas. *Nature Communications* **8**, 15089 (2017). URL <http://www.nature.com/articles/ncomms15089>.

- [377] Liu, J. & Robinson-Rechavi, M. Robust inference of positive selection on regulatory sequences in human brain. *bioRxiv* (2020). URL <https://www.biorxiv.org/content/early/2020/09/26/2020.03.09.984047>.
- [378] Liu, J. *et al.* The hourglass model of evolutionary conservation during embryogenesis extends to developmental enhancers with signatures of positive selection. *bioRxiv* (2020). URL <https://www.biorxiv.org/content/early/2020/11/02/2020.11.02.364505>.
- [379] Lilue, J. *et al.* Sixteen diverse laboratory mouse reference genomes define strain-specific haplotypes and novel functional loci. *Nature Genetics* **50**, 1574–1583 (2018). URL <http://www.nature.com/articles/s41588-018-0223-8>.
- [380] Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods* **15**, 591–594 (2018). URL <http://www.nature.com/articles/s41592-018-0051-x>.
- [381] GreyListChIP: Grey Lists – Mask Artefact Regions Based on ChIP Inputs. R package version 1.12.0 (2018).
- [382] Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013). URL <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt128>.
- [383] Eres, I. E. & Gilad, Y. Opinion A TAD Skeptic : Is 3D Genome Topology Conserved ? *Trends in Genetics* (2020).



# **Appendix A**

## **Supplementary figures and tables for Chapter 2**

Table A.1 Mapping and peak calling statistics for CTCF ChIP-seq data in the five *Mus* species

Species	Factor	Individual	Unique ID	#Input Reads	# Mapped Reads	%Mapped Reads	#MACS Peaks (per replicate)	#Reproducible peaks ( $\leq 2$ replicates)
<i>Mus musculus domesticus</i>	CTCF	685304.0	do3065	37,067,110	32,106,748	87	27,927	40,289
<i>Mus musculus domesticus</i>	CTCF	685305.0	do3066	63,240,749	53,082,610	84	45,718	
<i>Mus musculus domesticus</i>	CTCF	85922.0	do3073	49,861,065	45,543,182	91	44,202	
<i>Mus musculus domesticus</i>	input	85304.0	do3071	46,823,684	44,348,418	95	NA	NA
<i>Mus musculus castaneus</i>	CTCF	74636.0	do3075	68,333,659	65,956,183	97	38,833	38,373
<i>Mus musculus castaneus</i>	CTCF	74874.0	do3076	37,740,031	36,120,900	96	39,542	
<i>Mus musculus castaneus</i>	CTCF	76993.0	do3069	69,732,659	60,847,415	87	40,009	
<i>Mus musculus castaneus</i>	input	74636.0	do3079	66,991,879	54,675,669	82	NA	NA
<i>Mus spretus</i>	CTCF	1a	do3180	80,500,930	76,032,454	94	37,042	24,183
<i>Mus spretus</i>	CTCF	2c	do3181	72,366,179	67,433,478	93	17,974	
<i>Mus spretus</i>	CTCF	3e	do3182	48,076,327	44,763,098	93	24,185	
<i>Mus spretus</i>	input	3e	do3185	30,940,333	29,778,137	96	NA	NA
<i>Mus caroli</i>	CTCF	76791.0	do3177	84,228,625	78,277,357	93	41,519	37,027
<i>Mus caroli</i>	CTCF	76792.0	do3178	75,520,746	71,218,023	94	38,604	
<i>Mus caroli</i>	CTCF	78713.0	do3179	118,560,403	113,114,517	95	38,467	
<i>Mus caroli</i>	input	76791.0	do3184	43,083,604	41,392,442	96	NA	NA
<i>Mus pahari</i>	CTCF	66009.0	do3174	75,214,636	66,852,515	89	33,462	29,924
<i>Mus pahari</i>	CTCF	82898.0	do3176	70,433,073	64,809,159	92	28,353	
<i>Mus pahari</i>	CTCF	82903.0	do3175	70,754,788	64,160,019	91	32,802	
<i>Mus pahari</i>	input	66009.0	do3183	50,780,061	47,658,093	94	NA	NA

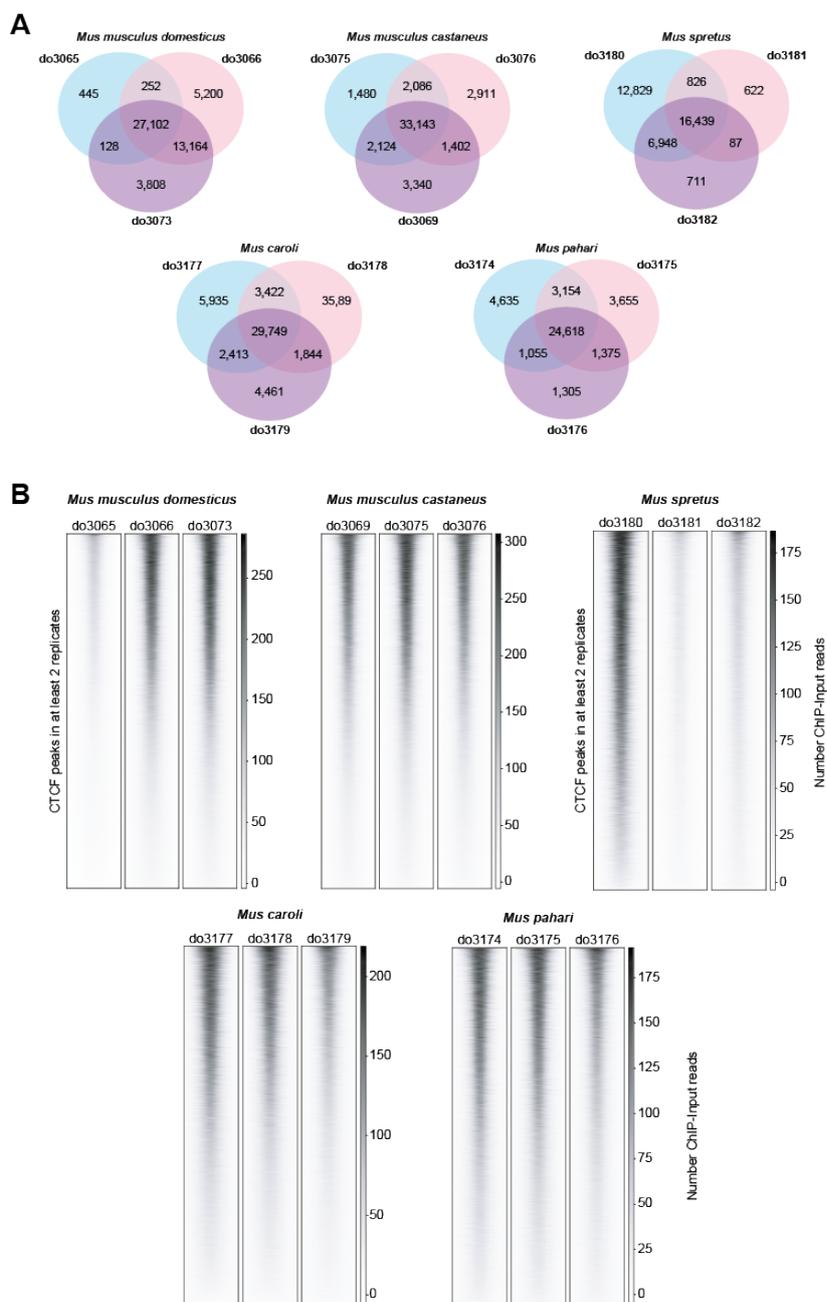


Fig. A.1 CTCF peak reproducibility among replicates of each *Mus* species. Venn diagrams (A) and binding heatmaps (B) showing reproducibility of peak calling among the three biological replicates from each species. We used peaks that were identified in at least two of the three replicates.

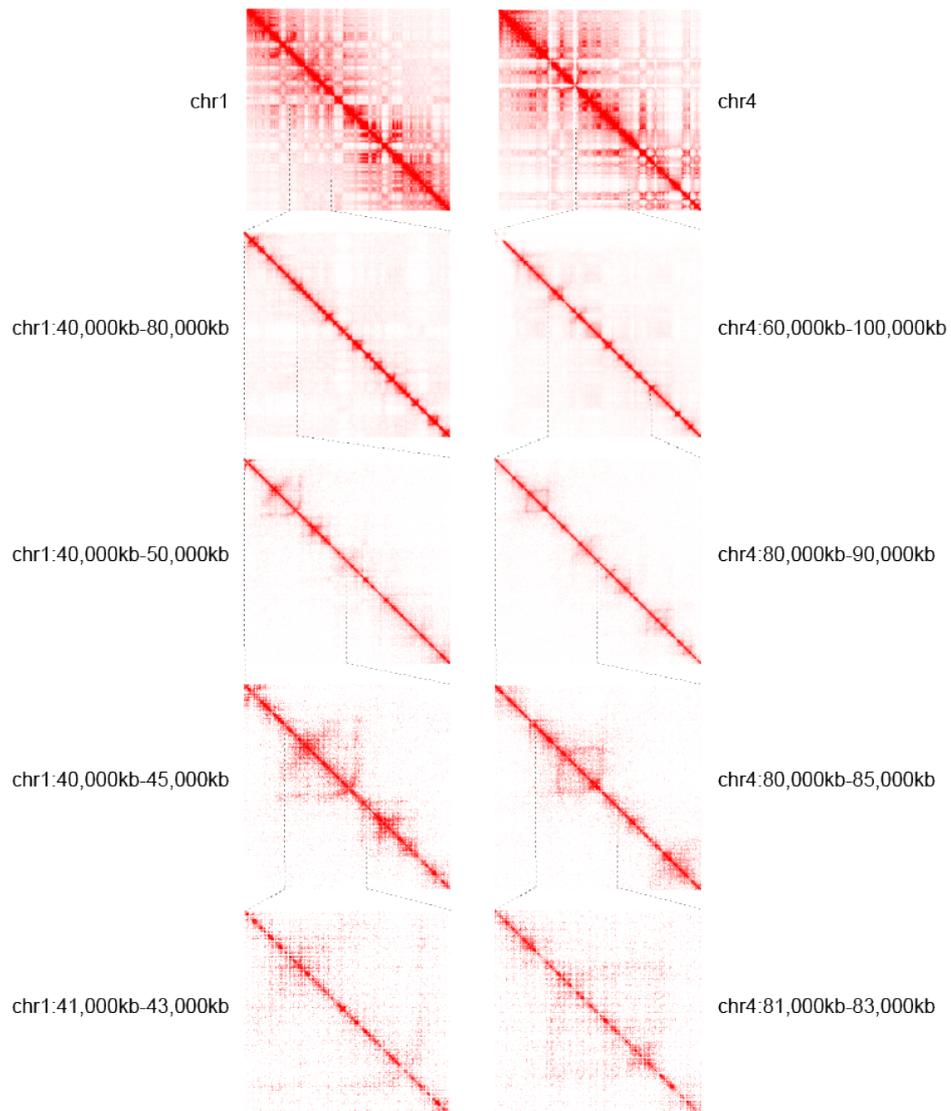


Fig. A.2 Hi-C contact maps from published C57BL/6J liver data. The contact maps generated from the published C57BL/6J Hi-C data [184] were visualized using Juicebox [269]. They show example regions from chromosomes 1 and 4, zoomed at different scales.

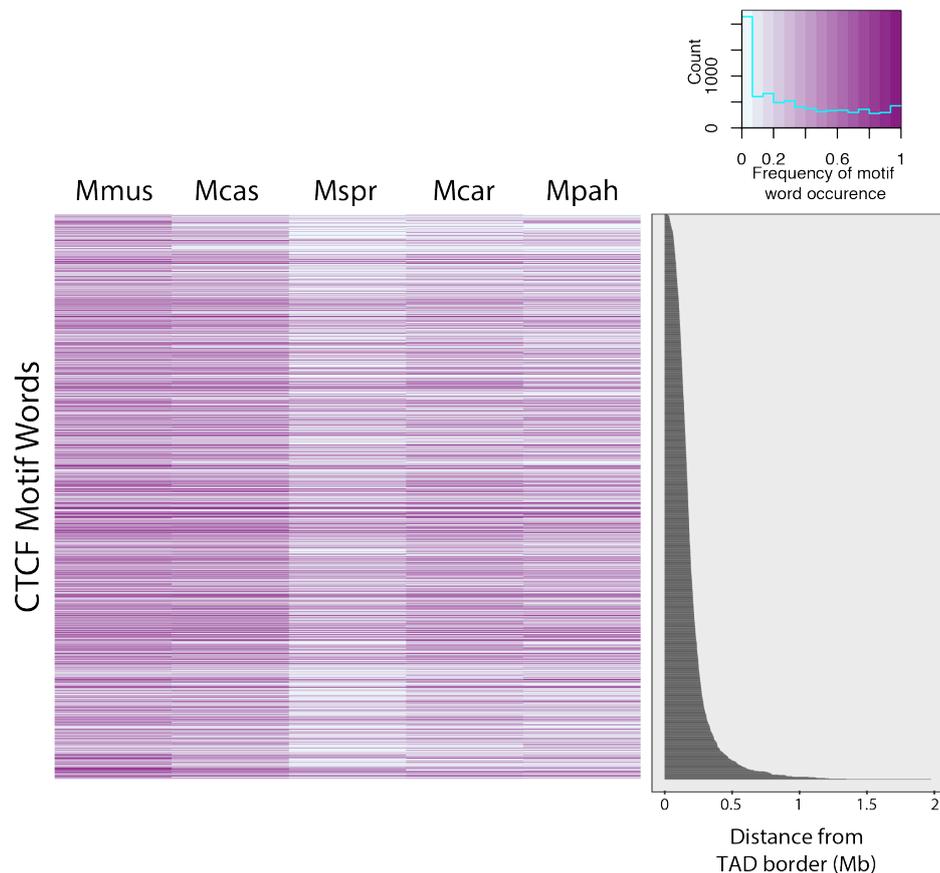


Fig. A.3 There is no evidence of any enrichment of specific motif words at TAD boundary regions among the species. Heatmap of the ~1,500 motif words found in CTCF peaks in the five *Mus* species. Each row corresponds to a motif word, while the color density represents its frequency of occurrence. The occurrence frequency of each motif word in the CTCF peaks is normalized by the number of its occurrences in the whole genome for the respective species. Motif words in the heatmap are sorted based on their distance to the closest TAD boundary. There is no evidence of any selected set of motif words being used with significant frequency at TAD boundaries among the species. The lower density of motif words in *M. spretus* reflects the smaller number of CTCF binding sites identified in that species.



# Appendix B

## Supplementary figures for Chapter 3

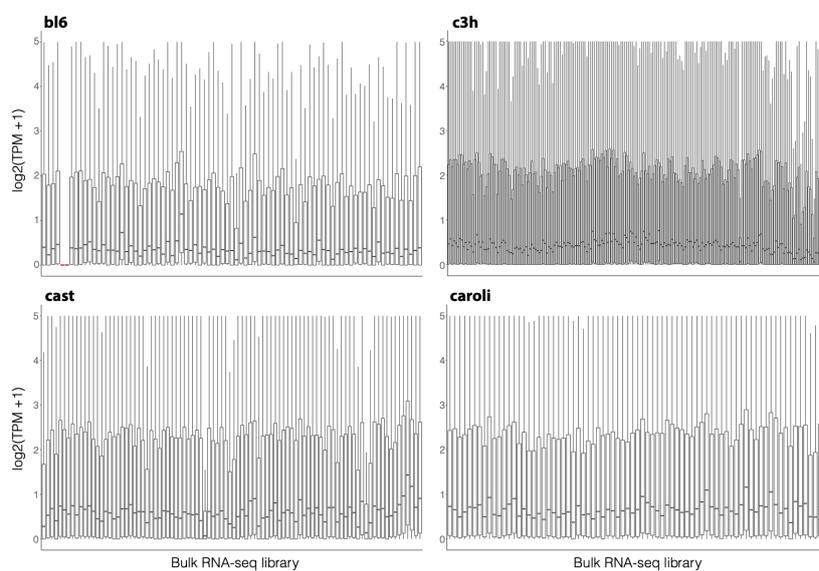


Fig. B.1 Distribution of transformed TPMs ( $\log_2(\text{TPM}+1)$ ) of expressed genes per bulk RNA-seq sample in each species. The tail samples of BL6 are marked red.

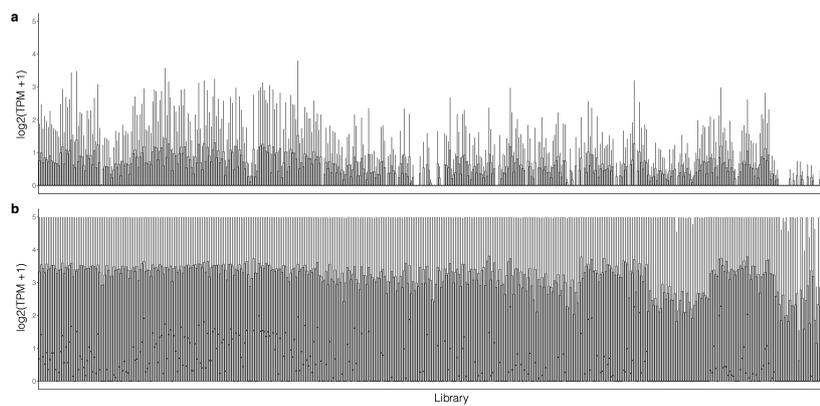


Fig. B.2 Distribution of transformed TPMs ( $\log_2(\text{TPM}+1)$ ) of expressed genes in single cells from fetal mouse liver (Yang et al. 2017). a) All expressed genes per single cell, b) Only genes from gene clusters *a-d*, which mark distinct putative cell types in liver.

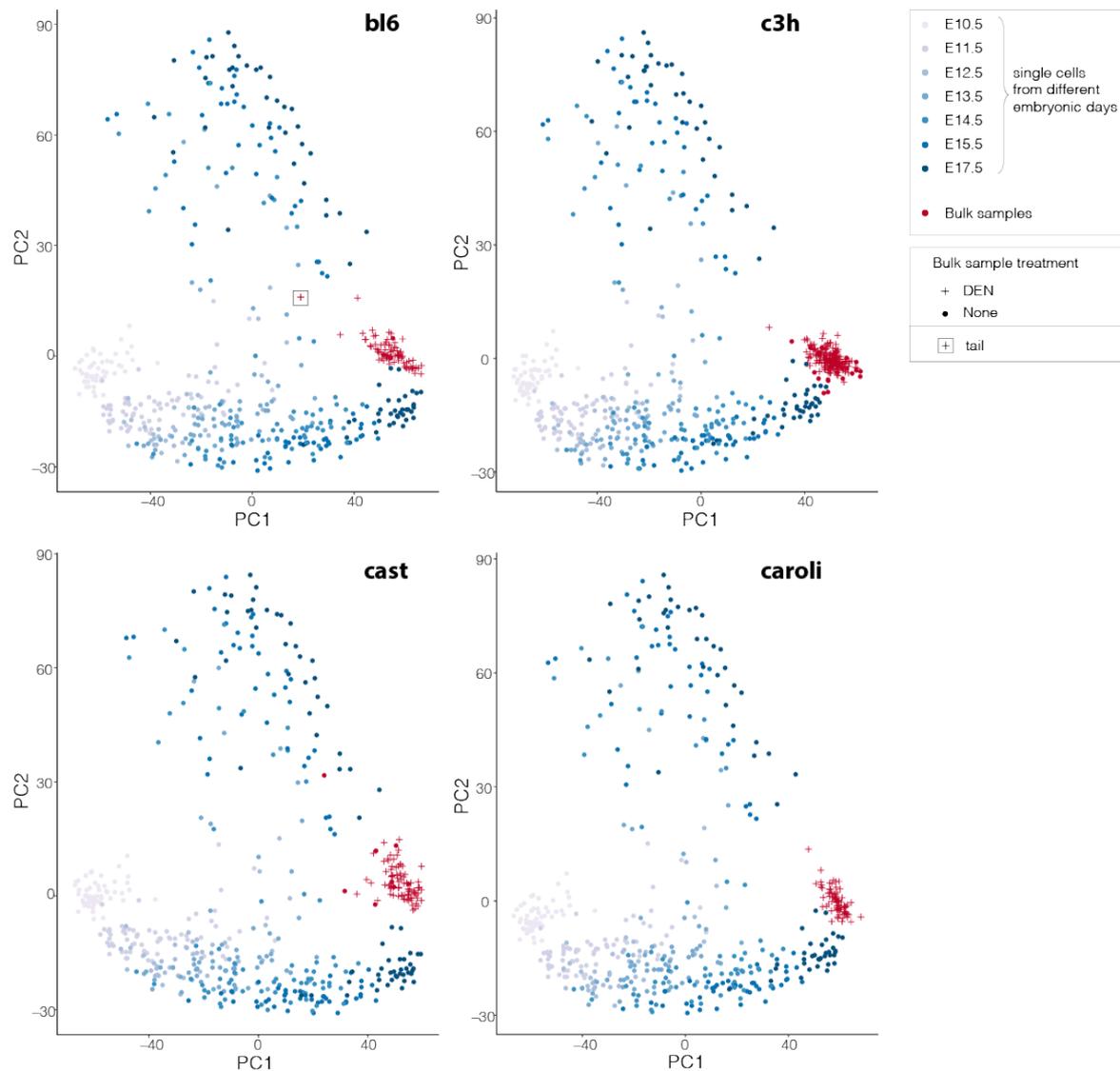


Fig. B.3 Principal component analysis of expression profiles of both bulk liver samples (LCE) and single cell data from mouse liver [328]. Expression profiles are based on expression of marker gene clusters *a-d*. The bl6 cohort includes DEN-treated normal liver samples, liver tumours (dysplastic nodules and HCCs), as well as two samples of tail tissue. The cohorts of the other species include only liver tumours. The bulk sample expression profiles mostly map on the latest stage of differentiated hepatocytes. Same data as in Fig. 3.3, albeit labeled by Treatment.

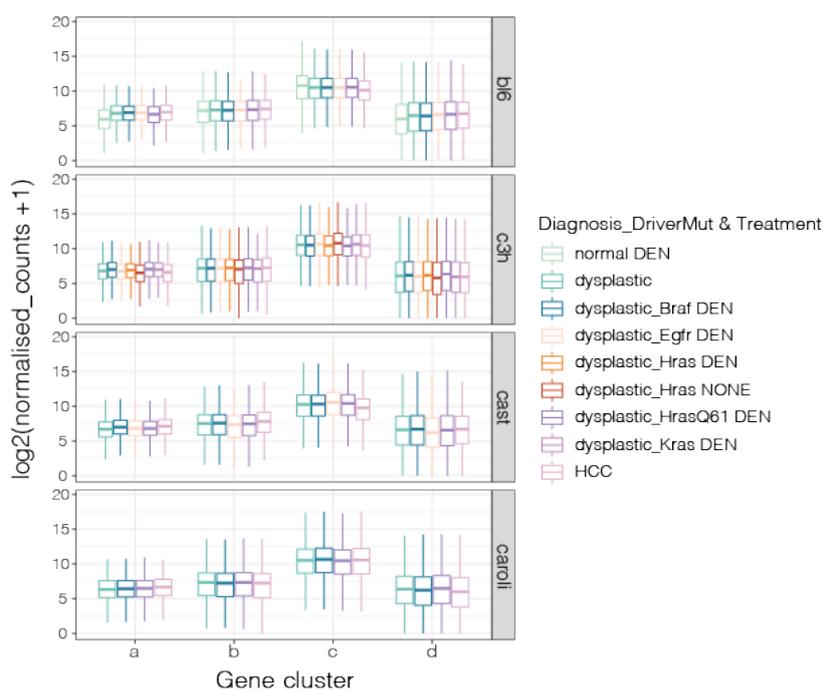


Fig. B.4 Distribution of expression levels (measured as transformed normalised counts across samples and species) of genes contained in the marker gene clusters *a-d*, in "DEN" and "None treatment" cohort samples per species. Samples from all cohorts display distinctly higher expression of cluster *c* genes, which are hepatocyte markers. In bl6, a drop of normalized count distribution of cluster *c* genes is outlined along tumour progression (from normal liver to dysplastic nodules and subsequently to HCCs). Distribution shifts are noticed also in the other gene clusters.

# Appendix C

## Supplementary figures and tables for Chapter 4

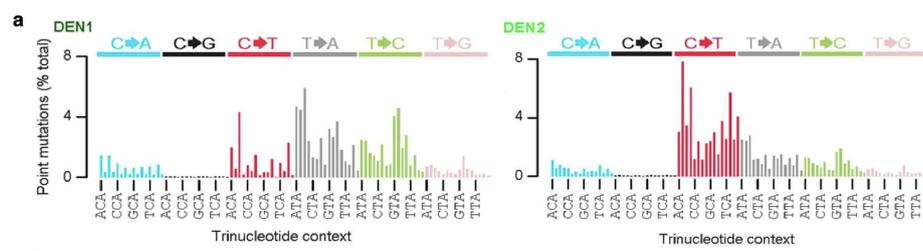


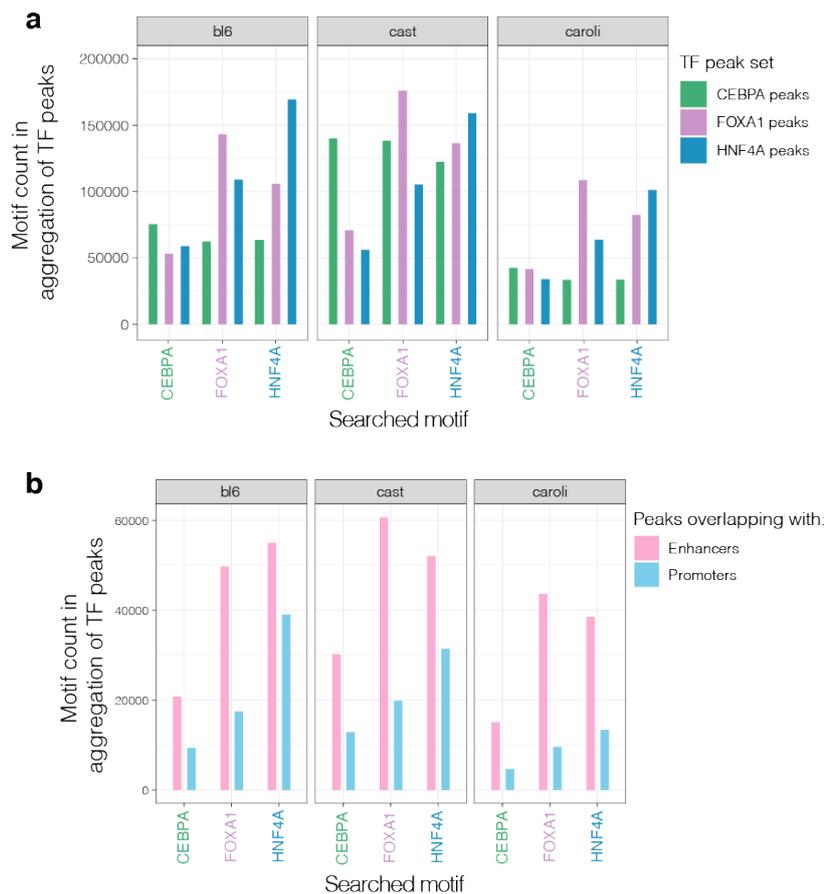
Fig. C.1 DEN1 and DEN2 mutational signatures, as identified from DEN-induced tumours of C3H mice and reproduced in CAST [296]. Underlying analyses of this figure were run in the LCE consortium in parallel with mine and figure is adapted from Aitken et al. (2020) [296]

	<b>bl6</b>	<b>cast</b>	<b>caroli</b>
CEBPA	67,237	117,069	33,865
FOXA1	94,408	106,850	67,455
HNF4A	90,940	102,133	62,432

Table C.1 Number of genome-wide identified TF peaks in each mouse species.

	<b>bl6</b>	<b>cast</b>	<b>caroli</b>
promoters	16,005	17,032	14,434
enhancers	45,165	41,660	34,808

Table C.2 Number of genome-wide identified promoters and enhancers in each mouse species, within the LCE consortium.



**Fig. C.2 Numbers of identified canonical TF binding motifs in the peaks of each TF cistrome and in CREs.** Canonical CEBPA, FOXA1 and HNF4A motifs were searched in the underlying sequences of all peaks from each set. a) Number of identified motifs in the ChIP-seq peaks of each cistrome (CEBPA, FOXA1, HNF4A peaks), b) Number of identified motifs in the intersection of the TF peak sets with *cis*-regulatory elements (CREs)

# Appendix D

## Publications

### Publications during this thesis

**Kentepozidou, E.**, Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M., Flicek, P. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol* **21**, 5 (2020). [252]

Aitken, S.J., Ibarra-Soria, X., **Kentepozidou, E.**, Flicek, P., Feig, C., Marioni, J.C., Odom, D.T. CTCF maintains regulatory homeostasis of cancer pathways. *Genome Biol* **19**, 106 (2018). [253]

Aitken, S.J., Anderson, C.J., Connor, F., Pich, O., Sundaram, V., Feig, C., Rayner, T.F., Lukk, M., Aitken, S., Luft, J., **Kentepozidou, E.**, Arnedo-Pac, C., Beentjes, S.V., Davies, S.E., Drews, R.M., Ewing, A. Kaiser, V.B., Khamseh, A. López-Arribillaga, E., Redmond, A.M., Santoyo-Lopez, J., Sentís, I., Talmane, L., Yates, A.D., Liver Cancer Evolution Consortium, Semple, C.A., López-Bigas, N., Flicek, P., Odom, D.T., Taylor, M.S. Pervasive lesion segregation shapes cancer genome evolution. *Nature* **583**, 265–270 (2020). [296]

### Manuscripts in preparation

**Kentepozidou, E.**, Aitken, S.J., Sundaram, V., LCE consortium, Flicek, P. Mutational landscape of transcription factor binding sites in chemically-induced liver tumours.

