

Characterisation of the non-canonical zinc finger protein ZFP263 in mouse

Célia Céline Alice DELAHAYE

**Department of Genetics
University of Cambridge**

Newnham College

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

September 2017

Declaration

The research in this dissertation was carried out in the Department of Genetics under the supervision of Professor Anne Ferguson-Smith. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except when specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the Biology Degree Committee word limit of 60,000 words.

Célia Céline Alice DELAHAYE

September 2017

Characterisation of the non-canonical zinc finger protein ZFP263 in mouse

Célia Céline Alice DELAHAYE

Multicellular organisms are composed of a number of different specialised cells that all carry the same genetic material but are highly divergent in their functions and characteristics. This diversity is only allowed because sets of specific genes are expressed in one type of cells and silent in others. A precise control mechanism is required to fine-tune gene regulation and relies on chromatin structure and regulatory proteins. One of the largest families of DNA-binding factors that influence this in human and mouse is the KRAB zinc finger protein (KZFP) family. KZFPs are thought to have rapidly evolved alongside transposable elements and be mediators of transcriptional repression. The few KZFPs that have been characterised so far have been shown to be involved in a wide range of regulatory and biological processes; hence it is hard to make functional generalisations.

During my PhD, I studied one member of the KZFP family in mouse, ZFP263, with the aim of understanding its mechanism of action in mouse embryonic stem cells (mESCs) and its role in mice. My work has shown that ZFP263 is an ancient protein highly conserved in mammals and under purifying selection. One of its two functional domains however is divergent from the consensus sequence found in most KZFPs and suggests that ZFP263 might have lost the ability to recruit repressive chromatin states. My research identified the targets of ZFP263 binding in mESCs and showed that it does not bind and silence transposable elements. Indeed it targets unique regions of the genome, mostly within transcribed regions of genes. These genes show a wide range of expression levels and are involved in several key biological processes. Surprisingly, binding sites are not associated with the canonical KZFP co-factor but mostly co-localize with active histone marks. My findings lead me to hypothesise that ZFP263 has evolved to target active epigenetic states to unique regions that are positive regulators of transcription, in contrast to the more canonical model of KZFP function. To test this hypothesis, I have generated targeted mutations at *Zfp263* in mice using CRISPR-Cas9 and my preliminary results suggest that *Zfp263* mutants have growth defects indicating a role for this protein in mouse development.

My findings indicate that ZFP263 is a unique KZFP with non-canonical properties and provide novel insights into the evolution and functions of KZFPs in mammals.

Acknowledgments

I am hugely grateful to my supervisor Anne Ferguson-Smith for letting me join the group and work on this project. Thank you for the freedom that you allowed me, for your frankness, for your never-ending ideas and your enthusiasm in the good times, and for your support in the tougher ones. It has been an immense privilege to work with you.

I am deeply grateful to Angela Noon who was my mentor for the first months of my PhD and taught me everything about ChIP-seq, cell culture and in general about being a good scientist. I was also very fortunate to work with Hui Shi on the bioinformatic analyses. Thank you both for your much-valued guidance and patience.

Thank you to all members of the Ferguson-Smith lab, past and present. It has been an enormous privilege to learn from such talented scientists. Thank you for making this lab such a pleasant working environment, for your kindness, for all the cakes (!) and the badminton and pub nights. It has been a great pleasure to work with you all!

Thank you to my advisor Bon-Kyoung Koo who shared his expertise on CRISPR with me and for his kind support. I am very grateful to William Mansfield for his excellent assistance on the zygote injections. Thank you to the PDN animal facility and Wendy Cassidy in particular for her meticulous work.

I am very mindful of the generous support I have received from the Marie Curie EpiHealthNet program. I am grateful for the efforts that have been put into this program for our training and our well-being. I was fortunate to do two secondments: thank you to Li Qibin and his team at BGI for their support; and thank you to Matthew Trotter, Remco Loos and the rest of the team at Celgene to help me learning R and sharing their outstanding science with me.

Finally, thank you to everyone outside this scientific world: to all of my friends for being at my sides no matter what and to the “Herraiz & Co” for supporting our choice to move here.

Papa, Maman, Fabien, thank you for your unconditional love and support, for the regular Skype conversations, for trying to understand my work, for all the good times we shared the 6 of us (and now 7!) in Cambridge, Canéjan or Paris – and for everything else!

Julien, with all my heart, thank you.

Abbreviations

5mC	5-methylcytosine
A	Adenosine
AcCoA	Acetyl Coenzyme A
BC / CB	C57BL6/J x Mus Castaneus / Mus Castaneus x C57BL6/J
BER.....	Base excision repair
bp.....	Base pair
BGI.....	Beijing genomics institute
BTB.....	Broad complex, tramtrack and bric-a-brac
BWA.....	Burrows-wheeler aligner
C	Cytosine
cDNA.....	Complementary DNA
CDS	Coding DNA sequence
CGI.....	CpG island
ChIP-seq	Chromatin immunoprecipitation-sequencing
CpG.....	Cytosine and guanine dinucleotide
CRISPR.....	Clustered regularly interspaced short palindromic repeat
CTD.....	C-terminal domain
Da (kDa).....	Dalton (kiloDalton)
Del.....	Deletion
DMR.....	Differentially methylated region
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
E	Embryonic day
ERV.....	Endogenous retrovirus
G	Guanine
GFP.....	Green fluorescent protein
gRNA	Guide RNA
GO	Gene Ontology
H2A / H2B / H3 / H4	Histone 2A / 2B / 3 / 4
HAT	Histone acetyltransferase
HDAC	Histone deacetylase
HEK293.....	Human embryonic kidney cell
h/mESC.....	Human / mouse embryonic stem cell
HxKy ac / me1/2/3	Histone X lysine Y acetylation / moni- di- trimethylation
HP1	Histone protein 1
Ins	Insertion
Indel	Insertions or deletions
IP	Immunoprecipitation
KD	Knock-down

KAP1	KRAB-associated protein 1
KMT	Lysine methyltransferase
KRAB	Krüppel associated box
KO.....	Knock-out
KZFP.....	KRAB ZFP
LINE	Long interspersed elements
LTR	Long terminal repeat
MACS.....	Model-based analysis for ChIP-seq
MBD	Methyl-CpG binding domain
MEME	Multiple em for motif elicitation
MHR.....	Major homology region
mRNA	Messenger ribonucleic acid
NGS	Next generation sequencing
Panther	Protein analysis through evolutionary relationships
PCR	Polymerase chain reaction
PHD	Plant homeodomain
PLE	Penelope-like element
PRC1/2	Polycomb repressor complex 1/2
PTM	Post-translational modification
qPCR	Quantitative polymerase chain reaction
RBCC.....	RING, B-box, coiled-coil
Rep	Replicate
RNA	Ribonucleic acid
RNA-seq.....	RNA-sequencing
SCAN	SRE_ZBP, CTfin51, AW-1, Number 18 cDNA
SINE.....	Short interspersed elements
siRNA.....	Small interfering RNA
SNP.....	Single nucleotide polymorphism
SOAP	Short oligonucleotide analysis package
SVA.....	SINE/VNTR/Alu
T.....	Thymine
TAD.....	Topologically associated domain
TE	Transposable element
TET	Ten-eleven translocation
TF.....	Transcription factor
TSS.....	Transcription start site
UHRF	Ubiquitin-like containing PHD and ring finger domain
ZF.....	Zinc finger
ZFP	Zinc finger protein

Table of content

CHAPTER 1: INTRODUCTION	10
1.1 REGULATION OF GENE EXPRESSION	10
1.1.1 <i>Chromatin structure</i>	10
1.1.2 <i>Epigenetic regulation of gene expression</i>	13
1.1.3 <i>Transcription factors</i>	23
1.2 ZINC FINGER PROTEINS.....	23
1.2.1 <i>Zinc finger structure and functions</i>	23
1.2.2 <i>KRAB domain structure and functions</i>	26
1.2.3 <i>SCAN domain structure and functions</i>	29
1.2.4 <i>Evolution of KRAB ZFPs</i>	35
1.2.5 <i>KRAB ZFP functions</i>	38
1.3 CHARACTERISATION OF ADDITIONAL MEMBERS OF THE KZFP FAMILY.....	40
1.3.1 <i>ZFP57 is required for maintaining methylation stability at genomic imprints in preimplantation embryos</i>	40
1.3.2 <i>EpiHealth European Project</i>	41
1.3.3 <i>Experimental plan to characterise additional KZFPs</i>	42
1.4 ZFP263 IN MOUSE AND HUMAN.....	43
1.4.1 <i>Structure</i>	43
1.4.2 <i>Function in human</i>	44
1.4.3 <i>Proposed work</i>	47
CHAPTER 2: ZFP263 CONSERVATION AND EXPRESSION.....	49
2.1 INTRODUCTION.....	49
2.1.1 <i>Vertebrate evolution</i>	49
2.1.2 <i>Experimental plan</i>	50
2.2 RESULTS	51
2.2.1 <i>Conservation across species</i>	51
2.2.2 <i>Zfp263 expression in mouse and human</i>	59
2.3 DISCUSSION AND CONCLUSION.....	64
CHAPTER 3: IDENTIFICATION OF ZFP263 TARGETS IN MESCS	66
3.1 INTRODUCTION.....	66
3.1.1 <i>ChIP-seq assay</i>	66
3.1.2 <i>Experimental plan</i>	66
3.2 GENERATION OF FLAG-ZFP263 MOUSE EMBRYONIC STEM CELLS	67
3.3 CHIP-SEQ ANALYSIS PIPELINE.....	69
3.3.1 <i>Beijing Genomics Institute Secondment</i>	69
3.3.2 <i>Pipeline optimisation</i>	69
3.3.3 <i>Mapping Results</i>	73

3.3.4 Peak Calling.....	75
3.4 CHARACTERISATION OF ZFP263 BINDING SITES	77
3.4.1 Identification of ZFP263 binding motif	77
3.4.2 Association with repetitive elements	81
3.4.3 Targeting unique genomic loci.....	84
3.4.4 Epigenetic state of ZFP263 binding loci	88
3.5 EXPERIMENTAL VALIDATION	90
3.6 DISCUSSION AND CONCLUSION	93
3.6.1 Limitations of the study.....	93
3.6.2. ZFP263 is a unique KZFP targeting unique genomic loci	94
CHAPTER 4. ZFP263 FUNCTION IN VIVO	97
4.1 INTRODUCTION.....	97
4.1.1 Objectives	97
4.1.2 Experimental plan.....	97
4.2 GENERATION OF KO MICE	97
4.2.1 Targeting Zfp263 gene	97
4.2.2 CRISPR-Cas9 injection in mouse zygote.....	99
4.3 SCREENING FOR KO MICE	100
4.3.1 Exon 6 mutant mice.....	100
4.3.2 Exon 1 mutant mice.....	105
4.4 PHENOTYPIC CHARACTERISATION OF ZFP263 KO MICE.....	109
4.4.1 Embryo and placenta development.....	109
4.4.2 Post-natal development.....	110
4.5 DISCUSSION AND CONCLUSION	114
4.5.1 Zfp263 is efficiently targeted by the CRISPR/Cas9 methodology.....	114
4.5.2 Phenotypic characterisation of Zfp263 mutants.	115
CHAPTER 5. DISCUSSION.....	118
5.1 ZFP263 IS A HIGHLY CONSERVED MAMMALS-SPECIFIC PROTEIN	118
5.2 ZFP263 IS A UNIQUE NON-CANONICAL KZFP	119
5.2.1 ZFP263 targets unique genomic loci in mESCs.....	119
5.2.2 ZFP263 is associated with active promoter and enhancer characteristics	120
5.2.3 ZFP263 target genes display a large range of expression levels	121
5.2.4 ZFP263 protein interactions	122
5.2.5 Future approaches	122
5.3 ZFP263 REGULATES KEY GENES FOR NORMAL GROWTH IN MICE	124
5.3.1 ZFP263 is involved in growth regulation and placenta development.....	124
5.3.2 Future approaches	125
5.4 CONCLUSION	126
CHAPTER 6. METHODS	128

6.1 DECLARATION OF WORK	128
6.1.1 ChIP-seq assay	128
6.1.2 Animal work	128
6.2 CONSERVATION AND EVOLUTION ANALYSIS	128
6.2.1 Orthologues identification	128
6.2.2 Alignment, Percentage identity and Ka/Ks ratio	129
6.3 CELL CULTURE	129
6.3.1 Mouse embryonic stem cell culture	129
6.3.2 mESCs transduction	129
6.4 CHIP-SEQ ASSAY	130
6.4.1 Chromatin immunoprecipitation	130
6.4.2 Library sample preparation	132
6.4.3 Sequencing platforms	133
6.4.4 ChIP-seq Bioinformatics Analysis	133
6.4.5 ChIP-qPCR	134
6.5 RNA WORK	136
6.5.1 RNA extraction	136
6.5.2 DNaseI treatment	136
6.5.3 cDNA synthesis	136
6.5.4 Quantitative PCR	136
6.6 ANIMAL WORK	137
6.7 CRISPR-CAS9 ZYGOTE INJECTIONS	137
6.7.1 gRNA design	137
6.7.2 Zygote injection	137
CHAPTER 7. REFERENCES	138
8. CHAPTER 8: APPENDIX	151
8.1 APPENDIX CHAPTER 1	151
8.2 APPENDIX CHAPTER 2	157
8.3. APPENDIX CHAPTER 3	162
8.4 APPENDIX CHAPTER 4	180

Chapter 1: Introduction

Multicellular organisms are composed of a number of different cell types that form tissues, assembled into organs that work together in the organism. Vertebrates consist of between 50 and 200 specialised different cell types. In human for example, over 200 cell types have been described, all carrying defined functions. Osteocytes and chondrocytes constitute the skeletal system, red blood cells transport oxygen, neurons transmit signals among the brain, spinal cord and other organs, gametes are reproductive cells and so on. Yet, all of these cells originated from one single cell, the fertilized egg, and they all contain the exact same genetic material, the same set of genes. There are about 20,000 protein-coding genes in human, but only a subset of them is expressed in each cell type. Because of their difference in function, cells only express a fraction of all of their genes, the remaining being silenced.

A precise control mechanism is required to allow the expression of the appropriate subset of genes. All the more so as the set of genes expressed in one cell is not fixed, but can fluctuate according to development stages and changes in the environment. This is why cells continuously receive information on which genes to express and which to turn off. This fine regulation relies on epigenetic mechanisms that coordinate chromatin structure and regulatory proteins – transcription factors (TFs) – to specifically activate or repress expression of genes. Since its first definition by Waddington in the early 1940s, as “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” (Goldberg et al. 2007), epigenetics is now defined as “the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence” (Wu & Morris 2001).

1.1 Regulation of gene expression

1.1.1 Chromatin structure

In eukaryotes, DNA is packaged into a more compact, denser shape around proteins to form the chromatin. This folding of DNA is not only essential to fit the whole genome in the small eukaryotic nucleus but is also a highly regulated process that contributes to genome function. The primary unit of chromatin is the nucleosome which consists of ~ 146 bp DNA wrapped around an octameric configuration of proteins, the histones. Thousands of nucleosomes are linked together by ~ 60 bp stretches of DNA in “beads-on-a-string”-like structure (**Fig. 1.1 A**). Each histone has a protein fold domain which consists of 3 helices linked together by short

loops that allow for heterodimerization: H2A with H2B and H3 with H4. Each dimer can further oligomerize together. The tetrameric H3-H4 occupies the centre of the nucleosome and initiates nucleosome assembly, which is completed by the addition of two H2A-H2B dimers that wrap the remaining DNA (**Fig. 1.1 B**) (Kornberg 1974; Thomas & Kornberg 1975; Kornberg 1977; Hammond et al. 2017; Alberts et al. 2002).

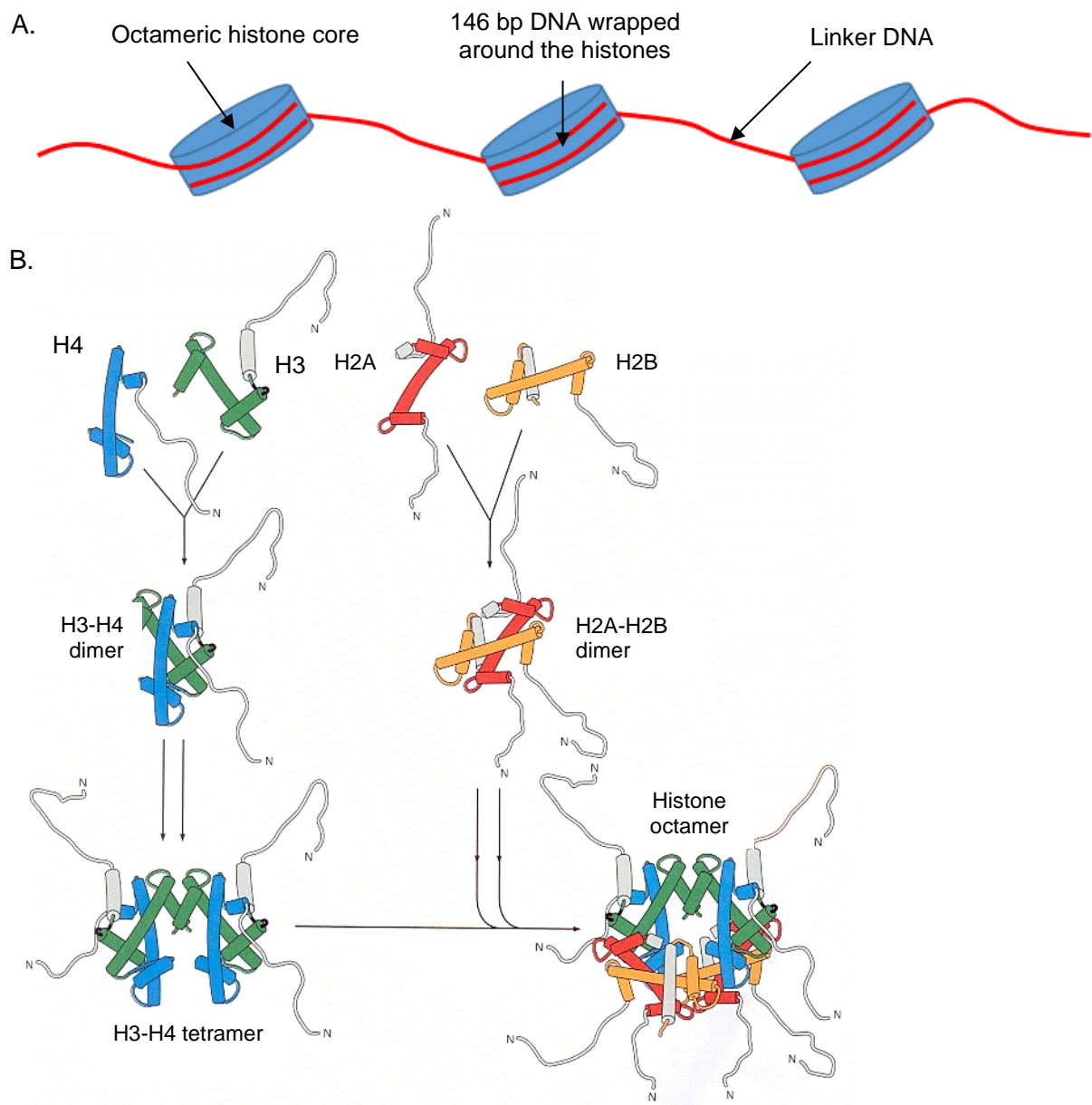


Figure 1.1: Structural organisation of the nucleosome. A. Beads-on-a-string form of chromatin. DNA (red) is wrapped around the histone octamer (blue) to form the nucleosome. B. Assembly of a histone octamer. H3-H4 tetramer forms the scaffold of the octamer onto which two H2A-H2B dimers are added. Blue: H4, Green: H3, Red: H2A, Orange H2B. All N-terminal tails of the histones protrude from the core structure (Alberts 2002).

expressed in a certain cell at a certain time. This regulation is partially provided by the chromatin structure. Indeed, chromatin has classically been considered to be found in two different conformations: the highly condensed heterochromatin and the less condensed euchromatin; however, in reality it is not as simple as this. Genes in canonical heterochromatin are tightly packed and are less accessible to the transcriptional machinery, and thus are silent. In contrast, active genes tend to be contained within euchromatin, where they are accessible and available for transcription. The conformation and function of chromatin is therefore an intimate interaction between epigenetic modifications to DNA and chromatin, the availability of transcription factors and polymerases, and the relationship with other interacting proteins including those that form complexes with the chromatin and its directly interacting partners.

Chromatin is highly dynamic and is subject to structural reorganisation for the regulation of gene expression and other biological processes such as the cell cycle. Recent techniques have unravelled higher order structures of chromatin in the nucleus, such as the topologically-associated domains, or TADs, that are self-interacting globular structures largely conserved across cell-type. Internal TADs structure consists of an array of looping interactions between chromatin contact points (**Fig. 1.2**) (Razin & Ulianov 2017; Dixon et al. 2016). It is suggested that in these conformations, local changes in chromatin conformation can potentially affect distant genes by disrupting loop formation and the accessibility of regulatory elements. Furthermore, organisation of chromatin into distinct domains within the nucleus is important for transcriptional regulation. For example, chromosome territories are partitioned into A and B compartments that are formed by long-range spatial interactions between distant genome loci and that contain active and repressive genome regions, respectively (**Fig. 1.2**) (Gilbert et al. 2004; Razin & Ulianov 2017). The three dimensional arrangement of chromatin and thus gene expression are regulated by epigenetic mechanisms that will be discussed in paragraph 1.1.2.

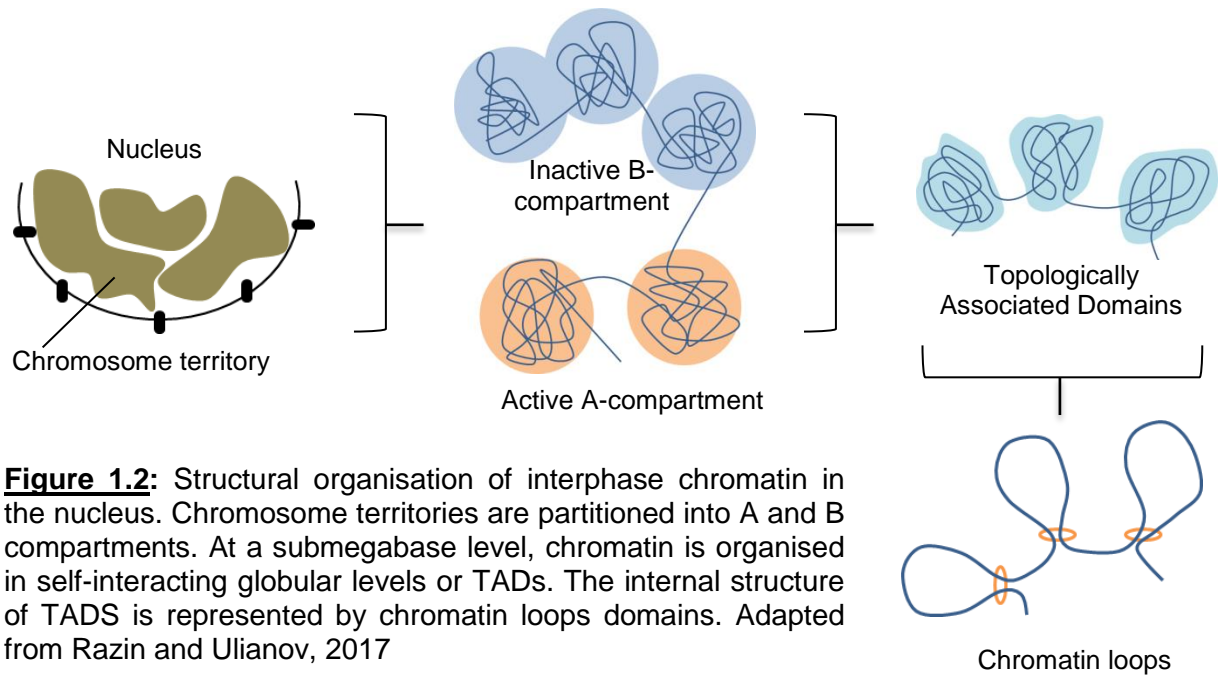


Figure 1.2: Structural organisation of interphase chromatin in the nucleus. Chromosome territories are partitioned into A and B compartments. At a submegabase level, chromatin is organised in self-interacting globular levels or TADs. The internal structure of TADS is represented by chromatin loops domains. Adapted from Razin and Ulianov, 2017

1.1.2 Epigenetic regulation of gene expression

Epigenetic regulation can be mediated by two classes of modifications: covalent modifications to DNA and covalent modifications to histones (Bird 2002). Both are related and function together in multiple layers of control of expression.

1.1.2.1 DNA methylation

DNA methylation is the best characterised modification of DNA, and consists of the addition of a methyl group to the fifth carbon in the cytosine pyrimidine ring (5mC). In vertebrate somatic tissues, this is mostly observed symmetrically on the two strands of CpG dinucleotides. Three classes of enzymes are involved in DNA methylation: the writers catalyse the addition of methyl groups onto cytosine residues, the erasers modify the methyl group and the readers recognize the methylated CpG dinucleotides (Moore et al. 2013).

1.1.2.1.1 Writing DNA methylation

DNA methylation is catalysed by a family of DNA methyltransferases (DNMTs). Establishment of DNA methylation patterns is controlled by two mammalian *de novo* methyltransferases, DNMT3A and DNMT3B, that can methylate naked DNA with no preference for hemi-methylated DNA (**Fig. 1.3**) (Okano et al. 1999). Both proteins are highly similar in structure and function, but they differ in their gene expression pattern in human (Xie

et al. 1999). They are essential for mammalian development, as the knockout of *Dnmt3b* in mice is lethal from embryonic stage E9.5, and homozygous knockout mice for *Dnmt3a* die at about 4 weeks postnatally (Okano et al. 1999). The targeting of *de novo* methylation is mediated by the interaction of DNMT3A and DNMT3B with their homologue DNMT3L that lacks the catalytic domain present in other DNMT enzymes (Aapola et al. 2000; Hata et al. 2002). DNMT3L associates with DNMT3A and DNMT3B and modulates their methyltransferase activity or sequence preference (Suetake et al. 2004; Jia et al. 2007; Strogantsev & Ferguson-Smith 2012). DNMT1, the final member of the DNMT family, preferentially methylates hemi-methylated DNA (Pradhan et al. 1999; Bestor 1992). It is therefore known as the maintenance DNMT because it maintains the pattern of DNA methylation in a replication-dependent manner. Indeed, DNMT1 localises at the replication fork, binds to the newly synthesised DNA and methylates it to mimic the original DNA methylation present before replication (**Fig 1.3**) (Bestor 1992; Hermann et al. 2004; Moore et al. 2013).

1.1.2.1.2 Erasing DNA methylation

The mammalian genome is reprogrammed during early development and two waves of genome-wide DNA demethylation have been described in the germline and in early embryogenesis. This is done by both active and passive DNA demethylation processes. Passive DNA demethylation is replication-dependant and refers to the lack of maintenance of DNA methylation during DNA replication. Absence or inhibition of DNMT1 allows newly incorporated cytosines to remain unmethylated and thus reduces the overall methylation level following cell division (**Fig 1.3**) (Moore et al. 2013). Active demethylation on the other hand refers to an enzymatic process that processes the 5mC in order to revert it back to naked cytosine. During preimplantation development, the maternal genome goes through a replication-dependant methylation dilution process, *i.e.* passive demethylation (Rougier et al. 1998). In contrast, active demethylation is observed on the paternal genome, where the 5mC level dramatically decrease a few hours after fertilisation, when no DNA replication occurred (Mayer et al. 2000; Oswald et al. 2000). Active DNA demethylation arises from multiple pathways and enzymes. Direct removal of the methyl group would involve breaking the strong covalent carbon-to-carbon bond connecting cytosine to its methyl group; therefore demethylation is likely to occur through indirect pathways (Morgan et al. 2004). One active DNA demethylation mechanism has been described to be mediated by the ten-eleven translocation (TET) enzymes (**Fig. 1.3**). Indeed, TET enzymes catalyse 5mC oxidation to 5-hydroxymethylcytosine and thus prohibit maintenance of the existing DNA methylation

pattern as 5hmC is not recognised by DNMT1, eventually leading to passive DNA demethylation during cell division (Tahiliani et al. 2009; Valinluck & Sowers 2007). Subsequent oxidation of 5hmC may generate 5-formylcytosine (5fC) and 5-carboxylcytosine (5-caC) that can be recognised by the base excision repair (BER) pathway to replace the modified base by a naked cytosine (**Fig 1.3**) (Ito et al. 2011; He et al. 2011).

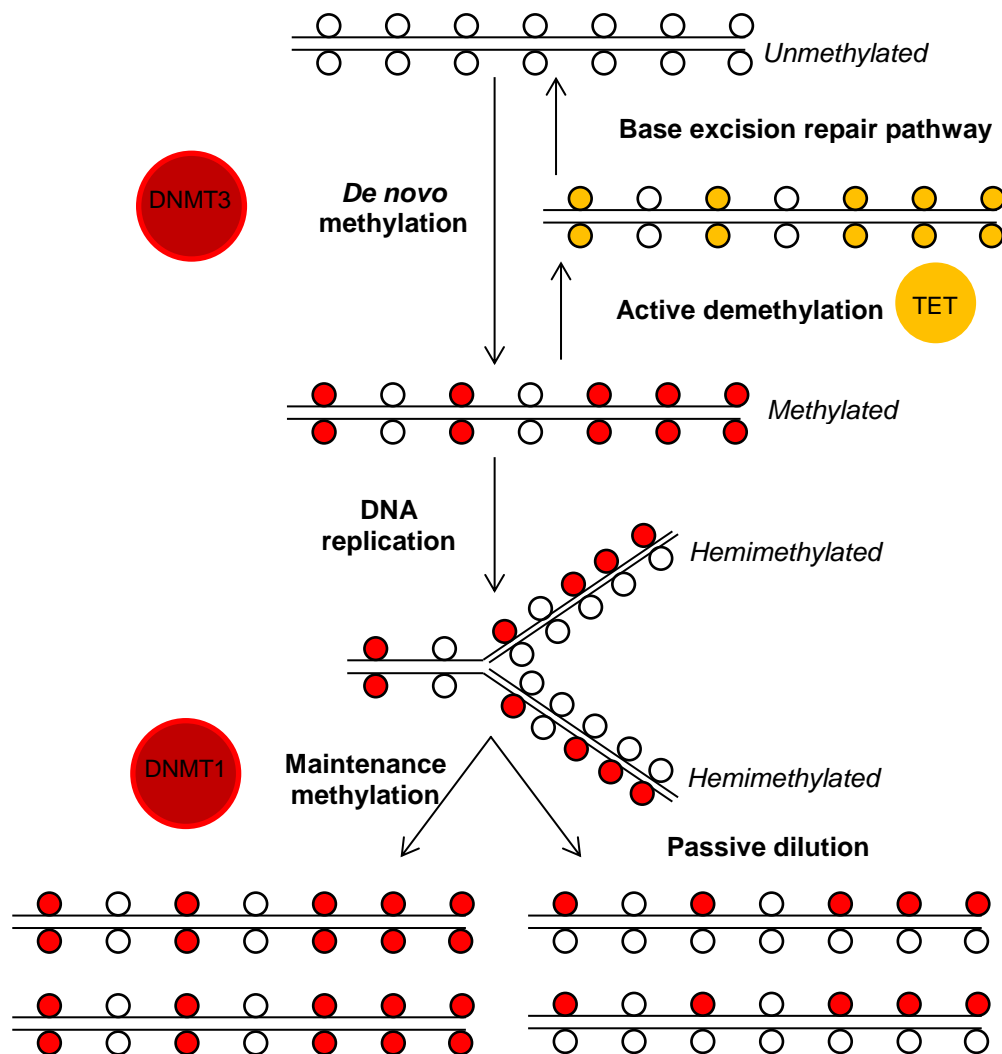


Figure 1.3: *De novo* methylation, maintenance of methylation and active and passive demethylation. A stretch of DNA is represented with CpG dinucleotides as black circles. DNMT3A/B/L symmetrically methylate some CpGs (red). TET enzymes catalyse 5mC into 5hmc, 5fC and 5caC (orange) that are recognised by the base excision repair pathway. DNMT1 completes half-methylated sites and maintains DNA methylation pattern after DNA replication. In the absence of DNMT1, new cytosines remain unmethylated, resulting in passive demethylation. Adapted from (Allis et al. 2015).

1.1.2.1.3 Reading DNA methylation

DNA methylation is recognised by different families of proteins: The methyl-CpG Binding Domain (MBD) proteins, the UHRF proteins (ubiquitin-like, containing PHD and RING finger domain) and the zinc finger proteins. The MBD proteins were the first to be identified and are the best described family. This family includes MeCP2, MBD1, MBD2, MBD3 and MBD4, which all contain a conserved domain that confers a high affinity for fully methylated CpG sites (Nan et al. 1993; Meehan et al. 1989; Lewis et al. 1992; Hendrich & Bird 1998). Each of these also associates with different repressor complexes through their transcriptional repressor domain and can alter chromatin structure (Nan et al. 1998; Ng et al. 1999). MBD proteins are highly expressed in brain and many are important for normal neuronal development and function. For example, mutations in human *MECP2* gene are responsible for the majority of cases of Rett syndrome (Amir et al. 1999). MBD4 is unique as it has enzymatic activity that can selectively remove T from a T-G mismatch in vitro and associates with protein involved in DNA mismatch repair (Hendrich et al. 1999; Millar et al. 2002).

The UHRF proteins play a key role in maintaining epigenetic inheritance patterns through recruitment of DNMT1 to hemimethylated DNA at replication forks. It was shown that UHRF1 proteins first bind to DNMT1 and targets it to hemi-methylated 5mCpG/CpG sites through its SRA domain (Bostick et al. 2007; Sharif et al. 2007; Hashimoto et al. 2008).

Finally, a subset of zinc finger proteins has the ability to specifically recognise 5mC-containing DNA. Kaiso for instance is a BTB/POZ containing-ZFP that preferentially binds to two consecutively methylated CpG sites and recruits chromatin-remodelling machinery to target genes (Daniel et al. 2002). More recently, ZFP57 has been shown to recognise the methylated hexanucleotide motif TGCCGC (Quenneville et al. 2011). The atomic crystal structure of two zinc fingers (ZF2 and ZF3) of mouse ZFP57 in a complex with the methylated TGCCGC site has been solved, defining the mechanism of sequence and methylation-specific recognition of DNA (Y. Liu et al. 2012). As for other KRAB-ZFPs, ZFP57 is known to recruit repressive machinery to its targets via the recruitment of KAP1.

1.1.2.1.4 DNA methylation and transcription

70-90% of mammalian CpGs are thought to be methylated (Ehrlich et al. 1982) with most nonmethylated CpGs found in high-density clusters, the CpG islands (CGI), which are often found at promoters and involved in regulatory functions. Heterochromatic regions and repetitive elements display a high level of DNA methylation that promotes a closed chromatin

structure to prevent expression or recombination (Zamudio et al. 2015). The understanding of DNA methylation function in transcriptional control is more challenging. In general, methylation of promoter regions is often associated with down-regulation of transcription. Nevertheless there is not always a simple correlation between the extent of transcription in the genome and the overall levels of DNA methylation. For instance, studies have shown that in human brain, a much higher percentage of RNA single copy sequence is transcribed than in liver, and yet the human brain DNA is 11% more methylated than human liver DNA (Ehrlich et al. 1982). The CpG density of the promoter partly explains the relationship between promoter methylation and transcription level. In vertebrate genomes, CGI sequences deviate significantly from the average genomic pattern by being CpG-rich and mostly unmethylated (Deaton & Bird 2011). CGIs are associated with the 5' ends of housekeeping genes and many tissue-specific genes, and with the 3' ends of some tissue-specific genes (Gardiner-Garden & Frommer 1987). It is suggested that CGIs at promoters destabilise nucleosomes and attract proteins to create a transcriptionally permissive state (Deaton & Bird 2011). Silencing of CGI promoters is achieved through dense CpG methylation or polycomb recruitment.

1.1.2.2 Histone modifications

The second class of epigenetic modifications to the chromatin is acquired post-translationally and modifies key residues on the histone tails or globular domains, especially those of H3 and H4. Indeed, histones undergo a variety of post-translational modifications (PTM), including acetylation, methylation and ubiquitination of lysine (K) residues, phosphorylation of serine (S) and threonine (T) residues, and methylation of arginine (R) residues. These covalent marks alter the interactions between adjacent nucleosomes or between histones and the DNA, thus potentially changing locally the chromatin structure which can have larger effects on transcription. Such modifications form a “histone code” read by other proteins resulting in the formation of the tightly packed transcriptionally silent heterochromatin, or the euchromatic domains, where the DNA is more accessible to nuclear protein complexes (Strahl & Allis 2000). For simplification, histone modifications are often divided into repressive or active marks according to their correlation with level of transcriptional activity. Acetylation and methylation are the best described histone post-translational modifications.

1.1.2.2.1 Histone acetylation

Histone acetylation marks are written by histone acetyl transferases (HATs), the first enzymes mediating PTM on histones to be identified (**Fig 1.4**) (Brownell et al. 1996; Kleff et al. 1995). Many different HATs have been identified since then and are classified into five subfamilies based on their sequence divergence. They all appear to share a conserved central core region contributing to Acetyl-Coenzyme A (AcCoA) cofactor binding, but diverge in their amino- and carboxy-terminal segments flanking the core. Each

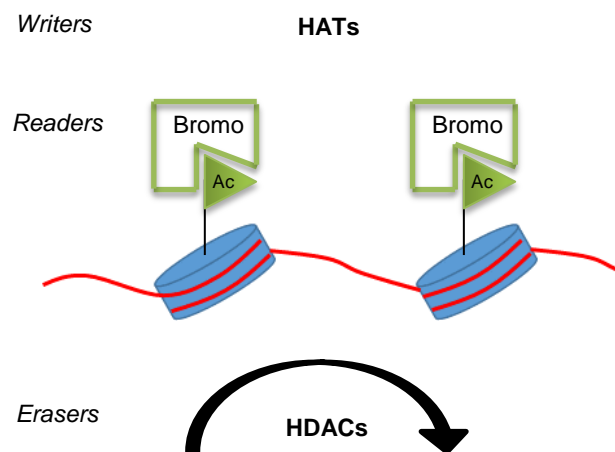


Figure 1.4: Histone acetylation writers, readers and erasers. Adapted from Allis et al. 2015.

HAT subfamily uses a different catalytic strategy to transfer the acetyl group from the AcCoA cofactor to the nitrogen of a lysine side chain within histones. HATs mediate many different biological processes including cell-cycle progression (Weinreich et al. 2004), dosage compensation and repair of DNA damage (Squatrino et al. 2006); and aberrant HAT function is correlated with several human diseases. The acetylation marks on lysine are recognised by small protein domains called bromodomains (**Fig 1.4**). Bromodomains adopt a distinct structural fold with conserved residues within interhelical loops that specifically recognise the acetyllysine. Other residues binding either side of the acetylated lysine contribute to binding specificity. Histone lysine acetylation is highly reversible as acetylation marks can be removed by histone deacetylases (HDACs) (**Fig 1.4 A**). In human, HDACs are traditionally separated into four categories based on their sequence similarities. Structural comparisons amongst HDACs reveal a conserved group of active site residues in Classes I, II and IV, suggesting a common mechanism using zinc to catalyse hydrolysis of the lysine-amino bond (Haberland et al. 2009).

Acetylation is associated with active genes as it neutralizes the positive charge of lysine residue, weakening the electrostatic interaction between negatively charged DNA and histones, thus disrupting the tight packaging of chromatin. Hyper-acetylation mediated by HATs is therefore associated with active transcription and localises at active enhancers and promoters. H3K9ac for example is highly correlated with active promoters and often co-localises with H3K14ac and H3K4me3 that mark transcriptionally active gene promoters (Karmodiya et al. 2012). It has been shown that H3K9ac serves as a substrate for direct

binding and targeting of the super elongation complex to chromatin, thus promoting RNA Pol II pause release. After recruitment of proteins essential for transcription by H3K4me3, H3K9ac may function downstream of transcription initiation and help direct elongation of transcription on chromatin (Gates et al. 2017). Similarly, H3K27 acetylation is enriched at promoter regions of transcriptionally active genes in human T cells (Wang et al. 2008). It is thought to block Polycomb-mediated trimethylation at H3K27, therefore preventing repression (Tie et al. 2009). H3K27ac was also proposed to be an important enhancer mark that could identify active enhancer elements in mouse embryonic stem cells and adult tissues, independently of H3K4me1 enrichment (Creyghton et al. 2010). More recently, acetylation of two lysine residues in the globular domain of H3, H3K64ac and H3K122ac, have been identified to mark active gene promoters and a new subset of active enhancers in mouse embryonic stem cells (Pradeepa et al. 2016). In this study, the authors defined 3 classes of enhancers based on the presence of H3K4me1 overlapping with H3K27ac, H3K64ac and H3K122 acetylation. Group 1 enhancers were positive for H3K4me1 and H3K27ac and marked by high levels of H3K64ac and H3K122 acetylation. Group 3 enhancers were negative for all three acetylation marks. Group 2 enhancers were enriched for H3K4me1, H3K64ac and H3K122ac but lacked H3K27ac. They functionally validated the enhancer activity of some Group 2 enhancers both *in vitro* and *in vivo* and therefore identify a new class of active enhancers lacking H3K27ac (Pradeepa et al. 2016). In contrast, hypo-acetylation, mediated by histone deacetylases, is generally associated with gene silencing, although HDACs have been described to localise at active gene loci as well (Dovey et al. 2010). **Table 1.1** summarises some histone modifications and their effects on transcription.

1.1.2.2.2 Histone methylation

Histone lysine methylation marks are written by lysine methyltransferases (KMTs) (**Fig 1.5**). All known KMTs contain a conserved SET domain of 130 amino acids. This domain was first identified as a shared motif in three proteins in *Drosophila*: suppressor of variegation (Su(var)3-9), enhancer of zeste (E(z)) and homeobox gene regulator trithorax (Trx) (Jenuwein et al. 1998). Mammalian homologs of Su(var)3-9 protein were the first characterised KMTs involved in H3K9 methylation (Peters et al., 2001; Allis et al. 2015). Since then, more than 50 SET domain-containing proteins have been identified in humans with a proven or predicted enzymatic role on methylating lysine on histone tails (Allis et al. 2015). Most SET-containing KMTs are grouped into six categories based on sequence homology within and around the catalytic domain, homology with other additional protein modules, and their structures. KMTs without a SET domain have been described, such as Dot1p, and therefore have an entirely different structural scaffolding and biochemical

properties. Lysine can be demethylated by oxidation by lysine-specific demethylase proteins (Shi et al. 2004) or by hydroxylation by Jumonji-containing demethylases (Shi & Tsukada 2013). Interestingly, there are more distinct domain types recognising lysine methylation than any other modifications. These include the plant homeodomain (PHD) fingers and the so-called “royal family” reader modules, comprising chromodomains, Tudor, PWWP and MBT domains (**Fig 1.5**) (Allis et al. 2015).

The PHD finger is a very common module found amongst chromatin remodelers and the PHD reader often reads a combination of histone post-translational modifications with other reader modules.

Methylation of lysine can be associated with both transcriptional activity and repression depending on the residue and histone modified. It can also occur several times on one residue side chain (mono-, di- or trimethylation), with different biological outcomes for each, adding even more complexity to the histone code (Li et al. 2007). Methylation of H3K9 is largely associated with silencing and repression in many species and is a characteristic hallmark of heterochromatin. H3K9me3 is bound by heterochromatin 1 (HP1) via its chromodomain, mediates chromatin condensation and stabilises heterochromatic subdomains (Lehnertz et al. 2003). Lehnertz et al. also demonstrated a physical and functional link with DNMT3B in mammals, providing a typical example of the interplay between DNA methylation and histone modifications. H3K9me2 is enriched on the inactivated X chromosome, together with H3K27me3 (Rougeulle et al. 2004; Escamilla-Del-Arenal et al. 2013). H3K27me3 is indeed strongly associated with inactive gene promoters and is believed to be critical for the repression of developmental genes. EZH2 is the only H3K27me3 methyltransferase, as part of the Polycomb Repressor Complex 2 (PRC2) (Kuzmichev et al. 2002).

On the other hand, methylation on H3K4 is mostly associated with active or permissive chromatin regions (Fischle et al. 2003). For instance, H3K4me3 is usually found at unmethylated promoters with other characteristics of open chromatin, while H3K4me1 is found at enhancers. H3K4me3 is tightly associated with transcription start sites of active

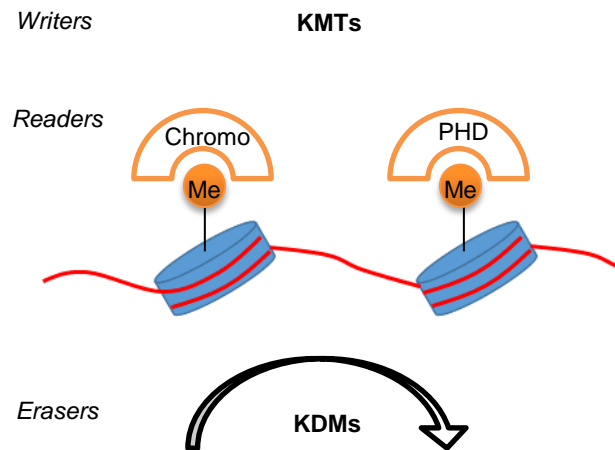


Figure 1.5: Histone Lysine methylation writers, readers and erasers. Adapted from Allis et al. 2015

genes (Barski et al. 2007). It regulates transcription by recruiting positive transcription factors, such as CHD1 and BPTF1 which remodel chromatin in an open state (Flanagan et al. 2005; Li et al. 2006), and prevents the binding of the repressive NuRD and INHAT complexes (Nishioka et al. 2002; Schneider et al. 2004). In mammals, at least 10 known or predicted H3K4 methyltransferases establish methylation through their SET domain (Ruthenburg et al. 2007). Interestingly H3K4me3 has also been found in ESCs in conjunction with H3K27me3. These regions have been described as bivalent promoters that silence developmental genes in ESCs while keeping them poised for activation (Bernstein et al. 2006; Bernstein et al. 2007; Voigt et al. 2013; Azuara et al. 2006). Finally H3K36me3 is enriched on gene bodies where it is thought to prevent acetylation to maintain a hypoacetylated state over the gene, therefore preventing aberrant initiation of transcription (Wagner & Carpenter 2012; Carrozza et al. 2005). It may also be involved in defining exons and influencing alternative splicing by signalling effector proteins to mark exons for inclusion (Luco et al. 2010; de Almeida & Carmo-Fonseca 2012). **Table 1.1** summarises some of the main histone marks described in the last two paragraphs.

1.1.2.2.3 Other post-translational modifications

H2A and H2B histones are more weakly associated with nucleosomal DNA (Wyrick & Parra 2009) and are easily and more frequently displaced from nucleosomes than H3 and H4. Relatively little is known about post-translational modifications of the dimer and its implication in transcription regulation. It is known that H2A and H2B can be monoubiquitinated respectively on lysine 119 and 120 which are found to be accessible to the ubiquitin-conjugating and deubiquitylation machinery. 10% of total higher eukaryotic H2A was reported ubiquitylated but only 1-2% of total H2B (Higashi et al. 2010). Ubiquitylation is deposited by PRC1. H2AK119ub is mostly associated with epigenetic silencing while H2BK120ub correlates with gene expression (Osley 2006).

Many modifications or their associated enzymatic complexes have been found to be involved in both active and repressive transcriptional activity. This suggests that single histone modification may have distinct biological effects depending on the genomic context (Fischle et al. 2003; Berger 2007; Berger 2002; Strahl & Allis 2000). Furthermore, the combinatorial nature of this histone language emphasises the potential for multiple layers of regulation.

Table 1.1: Post-translational modifications on lysine from histone H3, their associated enzymes and role in transcription.

PTM	Location	PTM enzymes (mammals)	Proposed function	References
Acetylation	H3K9	SRC1	Transcriptional activation Located at active promoters with H3K14ac and H3K4me3	(Spencer et al. 1997; Karmodiya et al. 2012; Gates et al. 2017)
	H3K27	P300/CBP	Transcriptional activation Defines active enhancers	(Wang et al. 2008; Tie et al. 2009; Creighton et al. 2010)
	H3K64	?	Transcriptional activation Located at active enhancers with H3K4me1 and with or without H3K27ac	(Pradeepa et al. 2016)
	H3K122			
Methylation	H3K4	SET family, MLL family	Transcriptional activation H3K4me1 enriched at enhancers H3K4me3 enriched at active promoters or at bivalent promoters with H3K27me3	(Ruthenburg et al. 2007; Fischle et al. 2003; Bernstein et al. 2006; Azuara et al. 2006)
	H3K9	SUV39H, G9a, SETDB1	Transcriptional silencing Hallmark of heterochromatin, enriched at inactivated X chromosome, constitutive heterochromatin and repetitive elements.	(Shinkai & Tachibana 2011; Escamilla-Del-Arenal et al. 2013; Rougeulle et al. 2004; Lehnertz et al. 2003)
	H3K27	EZH1/2	Transcriptional silencing Enriched at inactive X chromosome H3K27me3 found at bivalent promoters with H3K4me3	(Kuzmichev et al. 2002; Escamilla-Del-Arenal et al. 2013; Rougeulle et al. 2004; Bernstein et al. 2006; Azuara et al. 2006)
	H3K36	SETD2, NSD family	Transcriptional elongation	(Wagner & Carpenter 2012; Carrozza et al. 2005)
Ubiquitylation	H2AK119	PRC1	Transcriptional silencing	(Higashi et al. 2010; Osley 2006)
	H2BK120	RNF20, RNF40	Transcriptional activation	(Wyrick & Parra 2009; Osley 2006; Minsky et al. 2008)

1.1.3 Transcription factors

In eukaryotes, transcription is regulated by a multitude of proteins called transcription factors. General transcription factors are required at the promoter of a gene to form the initiation complex and recruit RNA Polymerase II, initiating transcription. In addition, diverse gene-specific TFs bind at enhancer or silencer elements to act as activators or repressors, respectively. TFs are DNA-binding proteins that can recognise specific DNA motif. In addition, they often contain an effector domain that interacts with other proteins to inhibit or promote transcription. Transcriptional activators and repressors control gene expression through diverse mechanisms, often in collaboration with co-activators or co-repressors that do not bind DNA directly but are recruited to DNA by the TF. They can interfere or promote the assembly of the transcription machinery, block the binding site for other transcription factors, alter the chromatin context of the genes and regulate transcription from the core promoters of nearby or distant genes through physical contacts that involve looping of the DNA between enhancers and the core promoters (Lee & Young 2000; Roberts 2000; Lee & Young 2013). One of the largest families of DNA-binding factors in vertebrates is the family of the zinc finger proteins, whose characteristics, structure and some of their functions will be introduced in the next section.

1.2 Zinc Finger Proteins

1.2.1 Zinc finger structure and functions

1.2.1.1 Zinc Finger Structure

A zinc finger (ZF) is a small self-contained peptide domain folded into a secondary structure stabilized by a zinc ion (Collins & Sander 2005) (**Fig 1.6 A**). Repetitive zinc-binding domains were first identified in *Xenopus laevis* transcription factor IIIa (TFIIIA) (Miller et al. 1985). TFIIIA was the first eukaryotic TF to be described and was known to allow correct initiation of transcription of the 5S RNA gene, as well as forming a complex with 5S RNA molecules in immature oocytes (Klug 2010). Miller et al. found that TFIIIA contained at least nine zinc ions and that it was formed of compact protein domains of ~3 KDa. They observed a strong and regular pattern of 30-residue repeats in these structures with a unique arrangement of pairs of cysteines and histidines. They proposed that the units were self-sufficient folded domains stabilised by a zinc ion, later called zinc fingers, and that such a structure would explain TFIIIA ability to bind the long internal control region of the 5S RNA gene and transcript (Miller et al. 1985).

The C₂H₂ ZF motif, defined by the consensus sequence CX₂₋₄CX₁₂HX₂₋₆H (with C cysteine, H histidine and X being any amino acids), is one of the most abundant in eukaryotic protein families being found in 2% of all human genes (Edelstein & Collins 2005). An additional characteristic is the presence of hydrophobic residues that are likely to form a structural core of the structure (Miller et al. 1985) (**Fig 1.6 A**). Comparisons of ZF domains with those of other known metalloproteins allowed the development of a detailed three-dimensional model for the ZF, consisting of an antiparallel β -sheet, which contains a loop formed by the two cysteines, followed by an α -helix containing the His-His loop (Berg 1988). Most C₂H₂ zinc fingers are *Krüppel*-type, named after the *Drosophila melanogaster* developmental regulator Krüppel. They are defined by the conserved link (TGEKP(Y/F)X) between the histidine of one finger and the cysteine of the following finger (Collins & Sander 2005; Turner & Crossley 1999). Other zinc-binding domains have been described but are less abundant (Klug & Schwabe 1995; Schwabe & Klug 1994).

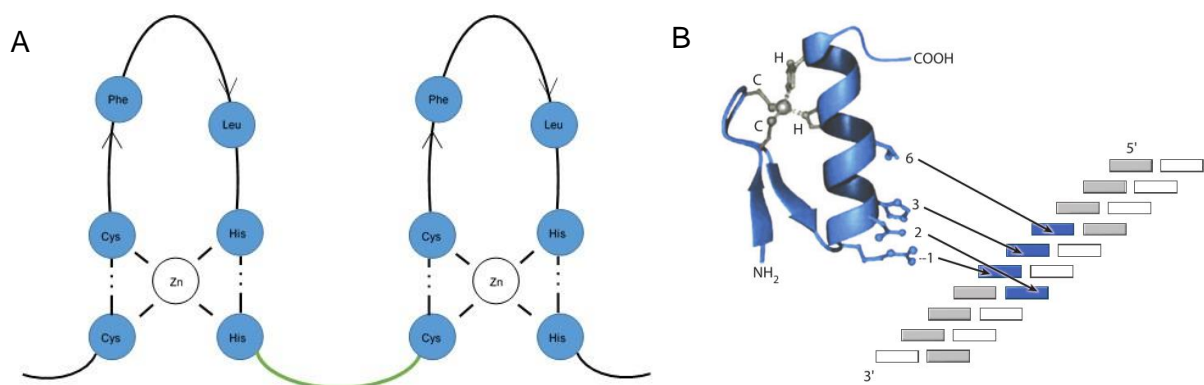


Figure 1.6: Zinc finger structure and interaction with DNA. Adapted from Klug, 2010. A: Folding scheme of two *Krüppel*-type zinc fingers, each centred on tetrahedral arrangement of zinc ligands, Cys2 and His2; and linked by the conserved sequence (green). Two hydrophilic residues are also shown. B: Four amino acids are involved in DNA motif recognition. Amino acids at helical position -1, 3 and 6 contact the coding strand and amino acid at position 2 contacts the other strand



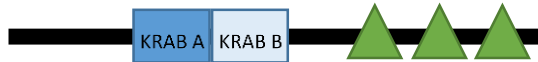

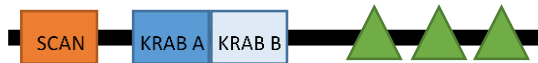
1.2.1.2 Zinc fingers mediate DNA interactions

Zinc fingers mediate specific interactions with DNA or RNA. Crystal structures of classical ZF have been solved in complex with their target DNA binding-sites. An early study on the transcription factor Zif268 (EGR1) found that amino acids at three key positions on the surface of the α -helix play a dominant role in base recognition. Amino acids -1, 3 and 6 relative to the amino-terminus of the helix make a one-to-one contact with 3 consecutive bases on the DNA major groove through specific hydrogen-bond interactions, as shown in

Fig 1.6 B (Klug & Schwabe 1995; Pavletich & Pabo 1991). However a second study showed that all zinc finger proteins did not follow this simplistic rule. They revealed that in a transcriptional regulator in *Drosophila*, amino acid in position 2 is also able to directly interact with the other strand of DNA (Fairall et al. 1993; Rhodes & Klug 1993). It suggests that ZF domains are able to recognise very specifically a precise DNA motif that they bind in a strand-specific manner (**Fig 1.6 B**). However, other elements might affect the DNA-binding specificity. For example, there are a number of phosphate contacts in the structures of zinc finger–DNA complexes that could be important in the energetics of zinc finger–DNA recognition (Wolfe et al. 2000). Similarly, the linker region that connects two zinc fingers is an important structural element and inter-finger organization might be important for DNA-recognition (Wolfe et al. 2000). Furthermore, ZFPs can contain up to 30 ZFs, but in some proteins only a subset of them are required for DNA-recognition. For example, CTCF uses a combination of its 11 ZFs to target a ~50bp motif that displays remarkable sequence variation (Ohlsson et al. 2001). Thus, although zinc fingers are able to strongly and specifically target a DNA sequence, it might be difficult to predict their binding motif, since variations are expected.

Most ZFPs contain other domains at their amino-terminal end. These associated modules include the BTB domain (Broad-Complex, Tramtrack, and Bric-a-brac), the KRAB domain (Krüppel-associated box) and the SCAN domain (SRE_ZBP, CTfin51, AW-1, Number 18 cDNA). **Table 1.2** presents the structure and number of human and mouse ZFPs found with the two latter domains that will be described in more detail in the next section. In human, out of 709 zinc finger protein coding genes, 350 also code for a KRAB domain, and 71 for a SCAN domain. 24 proteins contain both a KRAB and a SCAN domain (Sander et al. 2003). The ratio is similar in mouse although there are only 40 SCAN-containing ZFPs and only 17 SCAN-KRAB-ZFPs (Edelstein & Collins 2005) (**Table 1.2**).

Table 1.2: Primary structure of typical ZFPs and their number in human and mouse genomes. ZFPs possess variable number of zinc fingers (green triangles); some contain additional domains such as the KRAB domain, which consists of the KRAB A box alone or together with the KRAB B box (blue boxes); or the SCAN domain at the N-terminus of the protein (orange box). (Sander et al. 2003; Edelstein & Collins 2005)

	Schematic structure of zinc finger proteins and their associated domains	Protein coding genes in Human	Protein coding genes in Mouse
All C2H2 ZFPs		709	573
KRAB ZFPs	 	350	300
SCAN ZFPs		71	40
SCAN-KRAB ZFPs		24	17

1.2.2 KRAB domain structure and functions

1.2.2.1 KRAB domain structure

The KRAB domain, or Krüppel-Associated Box, is found in almost half of the human ZFPs since the human genome encodes about 350 KRAB ZFP coding genes (**Table 1.1**) (Huntley et al. 2006). The KRAB domain was first described over two decades ago as a highly conserved domain associated with Krüppel-type finger repeats (Bellefroid et al. 1991). It spans approximately 75 amino acids and consists of one or both of the KRAB A and KRAB B boxes (Collins et al. 2001). The two boxes of the KRAB domain are always encoded by separate exons. This structure allows the generation of different products, for example KRAB A only or KRAB A and B proteins, from a single gene, by alternative splicing (Urrutia 2003). The KRAB domain was thought to be restricted to tetrapods and that it evolved to provide vertebrates with a key function that underlies their development (Urrutia 2003). However

recently, KRAB ZFP genes were found in *Latimeria Chalmnae*, the African Coelacanth (Imbeault et al. 2017). Although most of the KRAB ZFPs were primate- or eutheria-restricted, as described before (Liu et al. 2014), the root of the family was reassigned to the Sarcopterygian common ancestor of coelacanth, lungfish and tetrapods (**Fig 2.1**).

1.2.2.2 KRAB and KAP complex

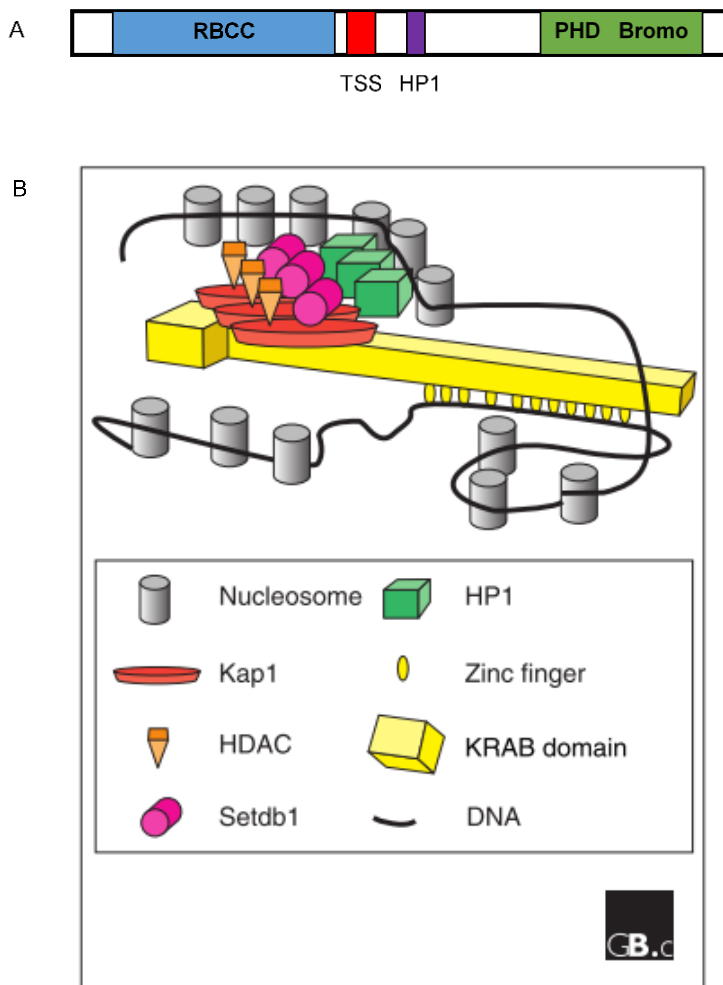


Figure 1.7: KRAB domain structure and recruitment of its co-factor KAP1. A: Structure of KAP1: RBCC domain at the N-terminal end of the protein, TSS domain and HP1 box in the centre, and the PHD finger and bromo-domain at the C-terminus. Adapted from Iyengar and Farnham, 2011. B: From Urrutia, 2003. Current model of the complex formed by KRAB-containing proteins. KAP1 is recruited to DNA via the KZFP and serves as a scaffold for recruitment of HP1, HDACs, and Setdb1, to form heterochromatin.

The KRAB domain has been described as a transcriptional repression module (Margolin et al. 1994; Witzgall et al. 1994). The KRAB A box has been shown to play a key role in transcriptional repression by binding the universal co-repressor KRAB-associated protein 1 (KAP1), also known as TRIM28 or TIF1 β (Friedman et al. 1996), while the B box is thought to enhance repression mediated by the A box although the mechanism is unknown (Urrutia 2003). KAP1 is the founding member of a small family of cofactors designated as the transcriptional intermediary factor 1 (TIF) in human and mouse (Venturini et al. 1999). KAP1 is a 97-kDa nuclear protein and contains a RBCC domain (a ring finger / two B boxes / a coiled-coil domain) at its N-terminus. The central region includes a TSS domain and HP1 box, and at the C-terminus, a PHD finger and a bromo-domain (**Fig 1.7 A**). The RBCC domain is essential for binding to the KRAB domain and participates in multimerization of KAP1 (Lechner et al. 2000).

KAP1 has been shown to initiate heterochromatin formation by recruiting heterochromatin protein 1 (HP1), H3K9 methyltransferase SETDB1, and the nucleosome remodelling and deacetylation (NuRD) histone deacetylase complex (Lechner et al. 2000; Schultz et al. 2000; Schultz et al. 2002). Lechner et al. showed that the chromoshadow domain in HP1 is required to bind KAP1, and a 15-residue sequence in the middle of KAP1 is necessary and sufficient for association with HP1. Two later studies by Schultz et al. provided evidence that KAP1 mediates interaction between the KRAB box and the NuRD complex via its PHD and bromo-domain (Schultz et al. 2000). The tandem PHD finger and bromo-domain of KAP1 forms a cooperative unit that is required for transcriptional repression, as they interact with the Mi-2 α subunit of the histone deacetylase complex NuRD. Furthermore, they showed that KAP1 binds to SETDB1, a H3K9 methyltransferase, and thus shed light on the mechanisms that both target and coordinate H3K9 methylation and HP1-mediated heterochromatin (**Fig 1.7 B**) (Schultz et al. 2002). Finally KAP1 has been shown to associate with DNA methyltransferases and to target them to imprinting control regions via ZFP57 (Quenneville et al. 2011). KAP1 appeared as a molecular scaffold targeted by KRAB ZFPs to specific loci, coordinating histone methylation, histone deacetylation and HP1 deposition to repress gene expression. Therefore, the large number of KRAB ZFP present in the mammalian genome gives rise to a complex, locus-specific regulatory network that targets genes for silencing (Takahashi et al. 2015).

KAP1 is a critical protein in the regulation of normal mouse development and differentiation. Mice lacking KAP1 fail to gastrulate and die at embryonic day E5.5 (Cammass et al. 2000) and the protein is required to control convergent extension and morphogenesis of extra-embryonic tissues (Shibata et al. 2011). KAP1 is also involved in a broad range of biological processes such as maintenance of pluripotency (Hu et al. 2009), mESC differentiation (Cammass et al. 2004; Cammas et al. 2002), epigenetic stability during mouse oocyte to embryo transition (Messerschmidt et al. 2012), anxiety disorders and tumour development (Iyengar & Farnham 2011). KAP1 is also associated with the repression of endogenous retroviruses and is required more broadly to safeguard the transcriptional dynamics of early embryos (Rowe et al. 2010; Rowe et al. 2013). Furthermore KAP1 has been suggested to regulate apoptosis, to suppress recombination and to be implicated in DNA repair when phosphorylated (Goodarzi et al. 2008; Iyengar & Farnham 2011).

KAP1 genomic binding sites have been identified using ChIP-seq experiment. Several studies showed that KAP1 binding sites are enriched in 3' coding exons of zinc finger genes and the promoter region of zinc finger genes or other genes. Since KAP1 is recruited to the

DNA via interaction with KRAB ZFPs, it has been suggested that expression of KRAB ZFP might be controlled by an auto-regulatory mechanism involving KAP1 (O'Geen et al. 2007). It has also been shown that the genes most responsive to *KAP1* knock down in cells were indirect targets of KAP1, suggesting a role for KAP1 that extended beyond direct transcriptional regulation at the majority of its strongest binding sites (Iyengar et al. 2011).

Most of the evidence supports a role for KRAB/KAP1 as a transcriptional repressor. Site-directed mutagenesis within the KRAB domain was performed by Margolin et al. in 1994. They targeted the most highly conserved amino acids between several KZFPs and identified highly conserved residues that are critical for KAP1 recruitment and repressive function. Substitution at the DV, EEW and MLE sequences significantly inhibited the ability to repress transcription in their *in vitro* system (**Fig 1.8**) (Margolin et al. 1994; Friedman et al. 1996). It suggests that some KRAB ZFP might not be able to recruit KAP1 to act as repressors if they contain one of these mutations.

LVTFK **DV** FVDFTR **EEW** KLLDTAQQIVYRNV **MLE** NYKNLVSLGYQLT **KP** **DVILR** **LE** KGEEPW

Figure 1.8: KOX1 KRAB domain amino acids sequence. Margolin et al. performed mutagenesis at highly conserved residues (squared). Mutation in KP and LE (black and white squares) had little effect on repression, while substitutions in DV, EEW and MLE decreased the ability to repress transcription (red squares) Adapted from (Margolin et al. 1994)

1.2.3 SCAN domain structure and functions

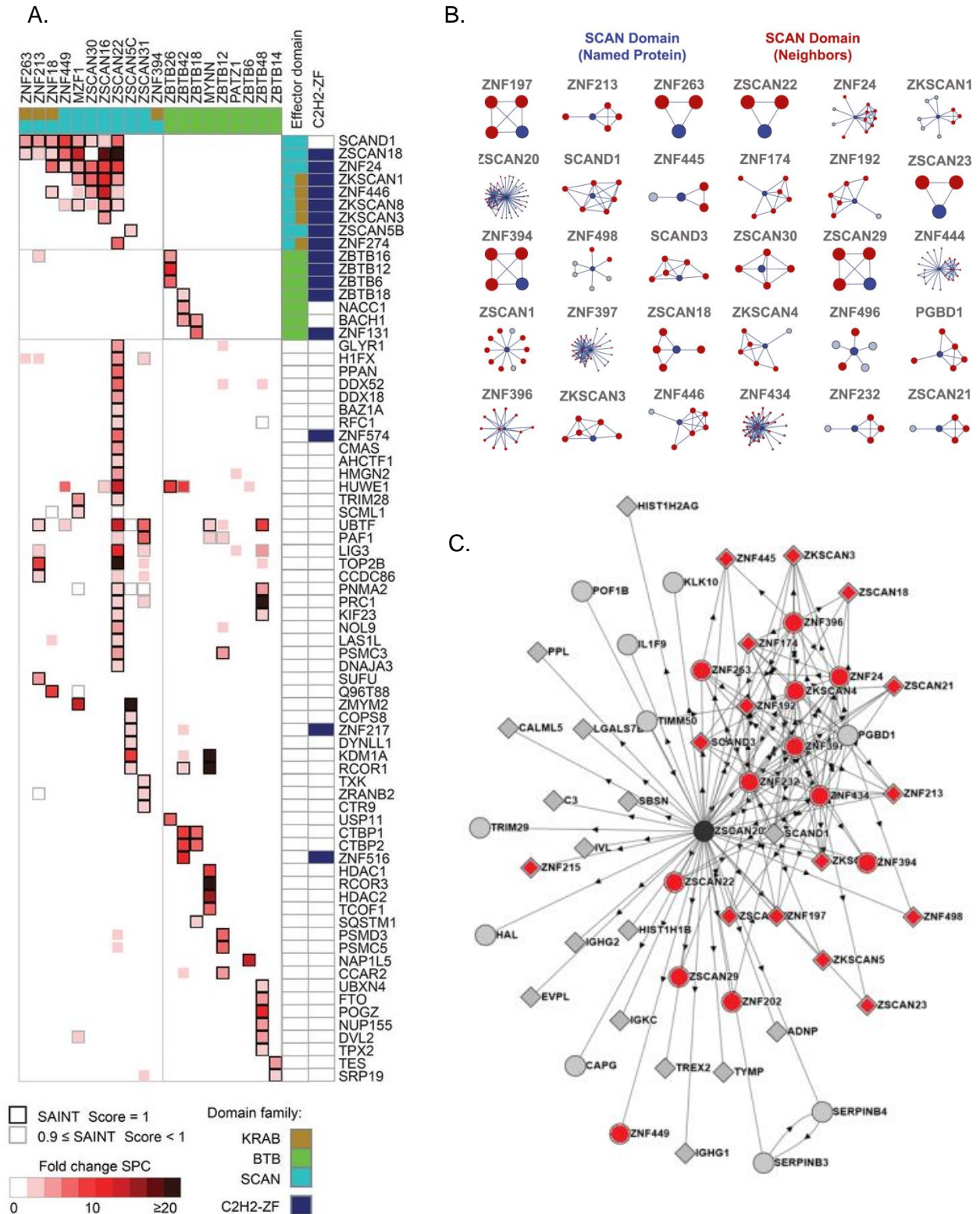
1.2.3.1 Domain identification

The SCAN domain was first identified in the zinc finger transcription factor ZNF174 (Williams et al. 1995). This new domain was identified by its homology with similar elements in other ZFPs, and named from the first letters of the proteins where it was originally found: SRE-ZBP, CTfin-51, AW-1 and Number 18 cDNA. Based on the sequence of its hydrophobic and negatively charged residues, an early study suggested that the SCAN domain may contain amphipathic α -helices, which could permit multimerization with itself or other proteins containing similar domains (Williams et al. 1995). Further analyses refined the definition of the SCAN domain as an 84 residue motif mostly found at the N-terminus of C2H2 ZFPs (Edelstein & Collins 2005).

1.2.3.2 The SCAN domain is a protein interaction motif

The SCAN domain itself does not have transcriptional activation or repression abilities, but acts as an interaction motif (Collins et al. 2001). A mammalian two-hybrid assay demonstrated that the SCAN domain functioned as an oligomerization domain mediating self-association or association with other proteins bearing a SCAN domain (Williams et al. 1999). They showed that not all SCAN domains are able to self-associate, that ZNF174 SCAN domain self-association required an entire intact SCAN domain, and that ZNF174 can selectively bind other SCAN members. Another study using a yeast two-hybrid system identified two SCAN proteins that interact with the SCAN-containing protein ZNF202 and confirmed that SCAN motifs have the ability to associate selectively with each other (Schumacher et al. 2000).

Large-scale protein-protein interactions studies helped identify more interactors for specific SCAN-containing ZFPs. Schmitges et al for example examined protein-protein interactions by affinity purification followed by mass-spectrometry for 118 ZFPs in HEK293 cells. They showed that many ZFPs display a unique interaction profile and identified novel highly diverse interaction partners (Schmitges et al. 2016). As expected, 9 of the 11 SCAN-containing ZFPs they used as baits interacted with other SCAN ZFPs (**Fig 1.9 A**). Interestingly, some interacting proteins were common to multiple SCAN ZFPs, such as SCAND1, ZSCAN18 and ZNF24 that interact specifically with 8, 7 and 6 of the bait ZFPs respectively (**Fig 1.9 A**). In contrast, other interactors were highly specific to only one ZFP in their dataset. Additional specific interactions were also found with non-SCAN proteins (**Fig 1.9 A**) (Schmitges et al. 2016). Another study used high-throughput affinity-purification mass spectrometry and identified interacting partners for 2,594 human proteins in HEK293T cells, resulting in the BIOPLEX network with 23,744 interactions amongst 7,668 proteins (Huttlin et al. 2015). They showed that SCAN-containing proteins self-associate, as expected, and that some SCAN-ZFPs interact with several others (**Fig 1.9 B**). This is the case for ZSCAN20 that interacts with multiple SCAN ZFPs (**Fig 1.9C**).



1.2.3.3 Domain structure

In 2002, Nam et al defined a smaller minimal functional unit of the SCAN domain. Based on multiple sequence alignment of SCAN-containing proteins, they identified a conserved region of 58 amino acids on the N-terminus site of the SCAN domains. They predicted the SCAN functional unit as a bundle of three α -helices folded to a core structure and divided by conserved proline residues (Nam et al. 2004) (**Fig 1.10 A**). The amino terminal helix revealed the highest diversity measure among the three helices offering critical surface-exposed binding residues. Thus this helix is likely to contain key components that determine selective dimerization patterns (Nam et al. 2004). Similarly, *in vitro* binding studies indicated that the SCAN domain in ZFP206 could selectively associate with other members of the SCAN TF family (**Fig 1.10 B**) (Liang et al. 2012). They solved the SCAN domain crystal structure and showed that the N-terminal helix 1 was critical for selective heterodimerization. These findings confirm that SCAN domains are modules that enable dimerization in a highly-selective manner.

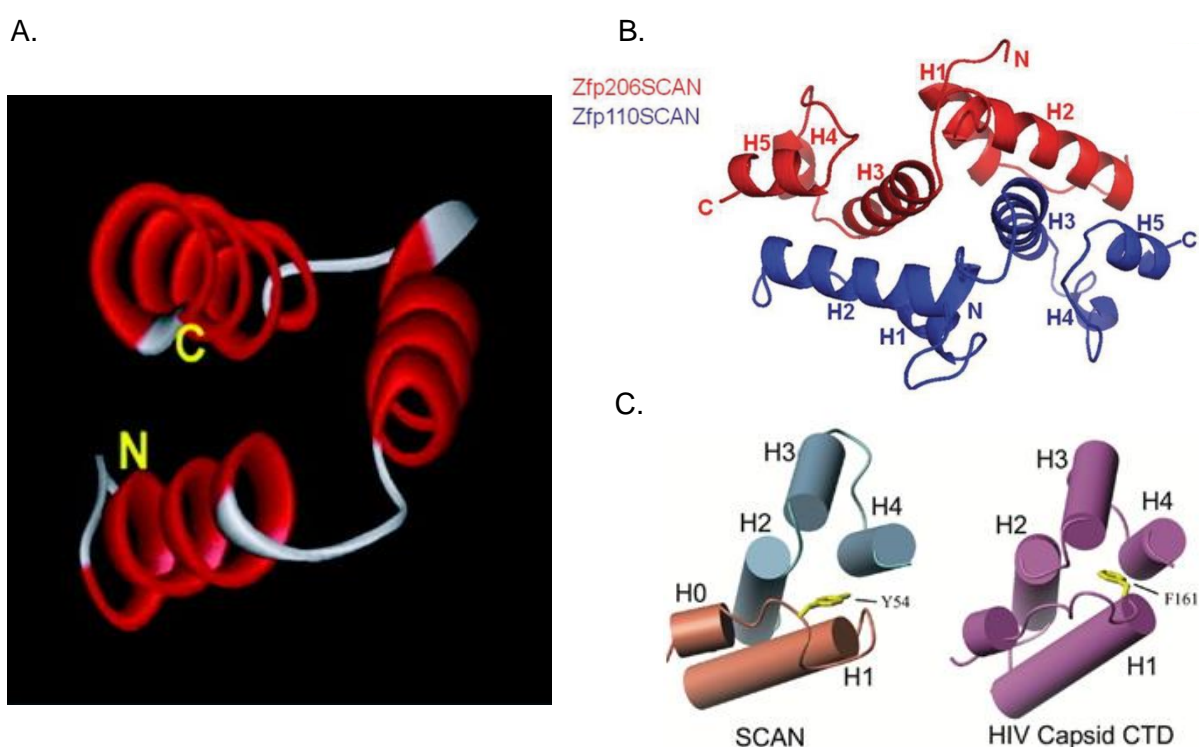


Figure 1.10: SCAN domain structure predictions. A: The fold prediction of the minimal functional unit for SCAN-domains is represented as a bundle of three alpha helices (Nam et al. 2004). B: Model for ZFP206-SCAN/ZFP110-SCAN heterodimer (red and blue respectively) (Liang et al. 2012). C: SCAN dimer (right – orange and grey) compared with the structure of the C-terminal domain of HIV-1 capsid protein (left – pink) showing similar structure (Ivanov et al. 2005)

1.2.3.4 Emergence and evolution

A structural homology search revealed similarity of the SCAN domain to the C-terminal domains (CTD) of retroviral capsid proteins (Ivanov et al. 2005). The CTD of capsid proteins contain critical determinants of Gag oligomerization as well as a conserved sequence motif, the major homology region (MHR). These authors found that the C-terminal domain of the HIV-1 capsid protein exhibits essentially the same structure as the SCAN domain of ZNF174, despite the absence of any detectable amino acid sequence homology (**Fig 1.9 C**). The capsid SCAN domain however uses a different dimerization surface caused by swapping the MHR-like element between the monomers (Ivanov et al. 2005). They suggest that the modern SCAN domain could be a descendant of the CTD-like domain of a retrovirus.

A later study reported that the SCAN domain was derived from the C-terminal portion of the gag capsid protein from the Gmr1-like family of Gypsy/Ty3-like retrotransposons (Emerson & Thomas 2011). Gmr1-like elements are a class of Ty3/Gypsy long terminal repeat (LTR) retrotransposons similar to most other Ty3/Gypsy elements but with a different protein domain order within the *pol* gene. They propose that the SCAN domain was adaptively co-opted for a novel function by C2H2 zinc finger proteins. Imbeault et al. described recently that SCAN containing ZFPs were older than KRAB ZFPs and shared deeper roots with marsupials and sauropsids.

1.2.3.5 SCAN-containing proteins in human and mouse

To determine the total number of genes predicted to encode SCAN domains in the human genome, genome databases were screened with a representative human SCAN domain. SCAN domains were identified and adjacent DNA sequences were annotated to predict the cDNA structures. This screen revealed the presence of 71 SCAN-containing genes in the human genome, 64 in complete open reading frames (**Table 1.2**) (Sander et al. 2003). Thus the SCAN domain proteins constitute approximately 10% of all human ZFPs. Out of the 71 SCAN-containing human genes, 14 are isolated single genes, but the majority (80%) is found in clusters on human chromosomes. Some of these genes are arrayed in tandem and present a target for mispairing and unequal crossover, which could result in duplication and divergence of the genes. These local duplications may account for the high degree of sequence similarity shared by neighbouring genes, as observed in the clustered group on human chromosome 19q13.4, where four neighbouring genes share a highly conserved

SCAN domain but more variable zinc fingers (Sander et al. 2003). This suggests that the genes have acquired mutations after a duplication event facilitating functional divergence.

In mice, the same approach has been taken and identified 40 mouse SCAN domain proteins (**Table 1.2**) (Edelstein & Collins 2005). Most of the mouse SCAN proteins have a putative orthologue on a conserved fragment of the syntenic human chromosomes. Sometimes an orthologue assignment was not possible because homologous SCAN family members within the human clusters were indistinguishable from each other when compared to the mouse (Edelstein & Collins 2005). Interestingly, it was found that human SCAN clusters are represented by a smaller number of SCAN genes in the conserved syntenic regions of the mouse. These findings provide evidence for human-specific cluster expansions of SCAN family members and argues that some genes within the SCAN family are lineage-specific and may have been selected independently since the divergence of primate and rodent lineages (Edelstein & Collins 2005). Together with the Imbeault results showing that SCAN-containing ZFPs are old and conserved proteins, this suggests that old SCAN proteins are highly conserved across evolution, and that duplication events have allowed fast evolution of emergent lineage-specific SCAN proteins.

Thirty-four members of the human SCAN family contain a novel motif at the very N-terminus of the predicted protein. This region is 13 residues in length but does not correspond to any established protein modules. A search of the human genome with the consensus sequence indicated that this region may be restricted to this subset of SCAN proteins (Sander et al. 2003). The domain contains a sequence for conjugation with the small-ubiquitin-related modifier, SUMO. Post-translational modification of the SCAN domain may have a substantial effect on the function of some of these family members (Collins & Sander 2005). Similarly, 25 out of the 40 mouse SCAN proteins contain this new conserved motif (Edelstein & Collins 2005).

Twenty-four members of the human SCAN proteins also contain a KRAB domain, half of them contain both KRAB A and B boxes while the other half only have the A box (**Table 1.2**) (Sander et al. 2003). In mice, 17 out of 40 contain a KRAB domain (**Table 1.2**) (Edelstein & Collins 2005). It is unclear whether one domain influences the function of the other. Interestingly, it was shown that six human SCAN-KRAB-ZFPs were KAP1-independent transcriptional repressors (Itokawa et al. 2009). The KRAB domains of these ZFP were not able to associate with KAP1 despite retaining transcriptional repression activity. Imbeault *et*

al. also mentioned that SCAN-containing ZFP were less prone to recruit KAP1, although this data is not shown in their publication (Imbeault et al. 2017). This is suggesting that the KRAB domain might be impaired in its ability to bind KAP1, and potentially its repressor activity, in the presence of a SCAN domain.

Although the majority of SCAN domains associate with C2H2 zinc fingers, six human SCAN proteins were identified without any zinc fingers. These SCAN domain-only sequences might represent novel genes without zinc fingers or splice forms of larger transcripts (Collins & Sander 2005). A mouse SCAN-only protein has also been described (Edelstein & Collins 2005).

1.2.4 Evolution of KRAB ZFPs

1.2.4.1 Transposable Elements

More than 40% of the mammalian genome is composed of repetitive sequences, of which about a quarter is derived from endogenous retroviruses (ERV). Transposable elements (TE) have been found in all eukaryotic species and display extreme diversity. TEs have been classified in two distinct groups. Class I contain the retrotransposons and class II the DNA transposons. Class II DNA transposons have the ability to leave their DNA locus and reintegrate themselves somewhere else. This is a “cut and paste” strategy (Wicker et al. 2007). The class I retrotransposons are able to transpose via an RNA intermediate and insert themselves randomly in the genome. These retrotransposons can be divided into five classes: LTR retrotransposons, DIRS-like elements, Penelope-like elements (PLEs), LINEs and SINES. The LTR retroelements are very abundant in plants but less in animals. In human, most of them are now inactive. Retroviruses and LTR retrotransposons are evolutionary closely related. Retroviruses might have evolved from LTR elements by acquiring a set of additional proteins, such as an envelope protein, and adopted a viral status. On the contrary, retroviruses that mutate and lose their infectious properties can become LTR retrotransposons, and are classified as Endogenous Retro Viruses (Wicker et al. 2007). Human ERVs encompass 1% of the human genome and have retained the ability to retrotranspose (Löwer et al. 1996). LINEs are long retroelements that lack LTRs but contain all domains that are necessary for transposition, and so are called “autonomous” retroelements. SINES, in contrast, are short and non-autonomous elements. They possess an internal Pol III promoter that allow them to be expressed, but they rely on LINEs for trans-acting transposition (Wicker et al. 2007).

1.2.4.2 The arms race model

TEs are often seen as a threat to the host organism. Indeed they can retrotranspose and integrate numerous copies randomly into the genome, thus impacting genomic structure and function (Löwer et al. 1996). They can for example disrupt genes via alternative splicing, truncation or insertion of new exon, or modify their expression by interfering with promoters, enhancers or repressors. They also often underlie recombination events due to their repetitive nature, can alter short- and long-range chromatin interactions, and provide novel open reading frames (Friedli & Trono 2015; Ecco et al. 2017). TE insertions or deregulations have been shown to be responsible for disease, including cancers, haemophilia, and other congenital or acquired human diseases. One cause of human breast cancer is for example the insertion of a primate-specific Alu SINE into the BRCA1 and BRCA2 genes (Ecco et al. 2017; Anwar et al. 2017)

KRAB ZFPs were thought to have emerged and evolved in parallel with ERVs as a way of suppressing their expression and retrotransposition and protect the host transcriptional dynamics (Emerson & Thomas 2009; Thomas & Schneider 2011; Hurst & Magiorkinis 2017). Indeed, the modular structure of ZFPs makes them well suited for rapid adaptive evolution. Several studies support this hypothesis. For example, ERVs are targeted and silenced during early embryogenesis by the KRAB-KAP1 silencing machinery to prevent retrotransposition (Rowe et al. 2010; Rowe et al. 2013; Rowe & Trono 2011). Similarly, in KAP1-depleted mESCs, repressive chromatin marks at ERVs are replaced by histone modifications normally associated with active enhancers, affecting nearby genes expression (Rowe et al. 2013). In human ES cells, a range of human-specific endogenous retroelements are controlled by KAP1 that acts as a scaffold for DNA-methylation-inducing factors (Turelli et al. 2014). Two KZFPs have been shown to target specific types of retroelements. ZFP809 has been shown to repress murine leukemia viruses in mESCs and to be involved in silencing of ERV transposition during embryonic development (Wolf et al. 2015; Wolf & Goff 2007). Gm6871 has been characterised as binding L1 elements in mESCs (Castro-Diaz et al. 2014). A study by Jacobs et al. brought further evidence that the KZFP family expanded to repress newly emerged retrotransposons (Jacobs et al. 2014). They showed that two primate-specific KZFPs evolved to repress two types of TEs shortly after they began to spread in the ancestral genome. ZNF91 acquired a number of structural changes that enabled it to repress SVA elements. ZNF93 evolved to repress the primate L1 lineage, until one subfamily escaped the restriction through the loss of its ZNF93 binding sites. They therefore suggested that repression of newly emerged retrotransposons by KZFPs, followed by mutations in the TE to evade repression, could explain the rapid expansion of lineage-specific KZFPs. It is

also suggested that newly emerged TEs are first repressed by DNA methylation-inducing small RNA-based mechanisms before a targeting KZFP evolved to repress them more stably by recruiting KAP1 (Rowe et al. 2013; Imbeault & Trono 2014).

1.2.4.3 Domestication model

Although this arms race model fits for specific examples KZFPs, it cannot entirely explain the evolution of this large family. First of all, several studies have shown that ancient retroelements have integrated themselves into the core cell circuitry and have acquired functions as alternative gene promoters and enhancers (Feschotte 2008; Ecco et al. 2017; Gifford et al. 2013; Elbarbary et al. 2016). Indeed, they can provide their hosts with new binding sites for transcription factors and provide a beneficial source of genetic variations. There is increasing evidence that endogenous retroelements have contributed to a variety of host biological functions, in particular in immunity and antiviral defence (Garcia-Perez et al. 2016). Noncoding sequences derived from ERVs may also act as enhancers of antiviral or pro-inflammatory genes (Frank & Feschotte 2017). Furthermore, chromatin marks deposited by the KRAB/KAP1 complex can spread out and affect regulation of neighbouring genes. Therefore KZFPs could be involved in the regulation of more specific cellular processes via the recognition of endogenous retroviruses that provide a platform for epigenetic regulation of other physiological processes (Ecco et al. 2016). Thus, retrotransposition events could be drivers of evolution by creating genetic diversity and new regulatory platforms, suggesting a gradual domestication of endogenous retroviruses by their hosts for adaptive purposes (Schlesinger & Goff 2014; Ecco et al. 2017; Imbeault & Trono 2014).

Moreover, new evidence confirmed that KZFPs exploit evolutionary conserved fragments of TEs as regulatory platforms long after the arms race against these genetic invaders has ended (Imbeault et al. 2017). Indeed, several studies found that there was a selective pressure to maintain KZFP binding on old and inactive TEs (Castro-Diaz et al. 2014; Imbeault et al. 2017). Cases were also observed where KZFPs appeared to have preceded the emergence of their targets. For instance, ZNF649 binds the L1PA promoter and dates back to the time of mammalian radiation, 60 million years before any of its targets had appeared (Ecco et al. 2017; Najafabadi et al. 2015). Finally, Imbeault et al. reported that if the majority of human KZFPs binds to retroelements, a third of them is recruited to other types of targets, such as promoters or simple repeats. Therefore they concluded that KZFPs partner with transposable elements to build a species-restricted layer of epigenetic regulation.

1.2.5 KRAB ZFP functions

There has been an increasing interest in zinc finger proteins over the past decade, leading to numerous studies that have started to shed light on the roles of these proteins. However, because of their large number in human and mouse, KZFPs remain largely uncharacterised and it is difficult to make functional generalizations since their mechanisms of actions may act in very different ways. The few KZFPs that have been described so far are involved in a large range of biological processes.

1.2.5.1 KRAB ZFPs in development and TE regulation

As discussed above, KZFPs are known to mediate heterochromatin formation via the recruitment of KAP1 and other co-factors. During early embryogenesis in mouse and human ES cells, the KRAB and KAP machinery silences TEs to prevent retrotransposition events (Rowe et al. 2010; Rowe & Trono 2011; Rowe et al. 2013; Turelli et al. 2014). KZFPs display very specific patterns of expression during embryogenesis, reflecting their importance in tightly regulating TE-derived loci during this period (Corsinotti et al. 2013; Fort et al. 2014; Gifford et al. 2013; Ecco et al. 2017). ZFP809 in particular has been shown to repress murine leukemia viruses in mESCs and to be involved in ERV silencing early in development, while ZNF91 and ZNF93 were shown to repress SVAs and LINE-1 respectively (Wolf & Goff 2007; Wolf et al. 2015; Jacobs et al. 2014). ZFP57 is a well-known ZFP that recognises a methylated hexanucleotide and interacts with KAP1 at imprinted control regions to maintain maternal and paternal methylation imprints after fertilization (X. Li et al. 2008; Quenneville et al. 2011; Strogantsev et al. 2015).

Other examples of KZFPs acting in ESCs and development include ZFP322a that regulates mESC pluripotency and enhances reprogramming efficiency when used in combination with the Yamanaka factors (Ma et al. 2014); ZFP568 regulates convergent extension in the mouse embryo and is required in embryonic-derived tissue for yolk sac and placenta morphogenesis (García-García et al. 2008; Shibata & García-García 2011; Shibata et al. 2011); and others play a role in erythropoiesis, osteogenesis and mammary gland development (Lupo et al. 2013; Ecco et al. 2017) via interactions with unique regions of the genome emphasising the functions of these proteins outside a role in the regulation of transposable elements. The evolutionary relationships between the unique and repeat functions of KZFPs are not understood.

1.2.5.2 KRAB ZFPs in metabolism

A number of KZFPs have been described in metabolic pathways, such as ZFP69 that was reported as mediating liver fat accumulation and mild insulin resistance in transgenic mice overexpressing *Zfp69*, or ZNF224 associated with glycolysis and oxidative metabolism (Lupo et al. 2013; Ecco et al. 2017). Recently a study on ZFP423 provided another example of a KZFP involved in metabolism. ZFP423 is critical for the maintenance of white adipocytes in adult mice, and is essential for the terminal differentiation of subcutaneous white adipocytes during foetal adipose tissue development (Shao et al. 2017)

1.2.5.3 KRAB ZFPs in apoptosis and cancer

Several KZFPs have been described in cancer and apoptosis. ZNF224 for instance has been shown to play a pro-apoptotic role through the interaction with different molecular partners (Lupo et al. 2013). Overexpression of this gene in breast cancer cells is also observed and it was recently associated with apoptosis resistance in chronic lymphocyte leukemia (Busiello et al. 2016). Other functional studies showed that ZNF545 is involved in the suppression of cancer cell growth and inducing apoptosis. A recent work on ovarian cancer tissues showed a significantly lower expression of *ZFP403* compared with normal ovarian tissues and cells. Its overexpression in ovarian cancer cells suppressed cell proliferation, suggesting that the protein may serve as a tumour suppressor in ovarian cancer (Zhu et al. 2017).

1.2.5.4 Genome-wide analysis of KZFPs genomic targets

Advances in NGS technologies have enabled genome-wide studies to identify the genomic targets of hundreds of KZFPs repressors (Najafabadi et al. 2015; Schmitges et al. 2016; Imbeault et al. 2017). These works are immensely useful to understand the evolution of the family and confirm that even though KZFPs belong to the same family and share structural domains, they exhibit a vast array of functions and mechanisms. More importantly, these works provide evidence that KZFPs do not solely repress transposition events but have evolved more diverse functions. Some KZFPs target unique genomic loci and interact with a unique spectrum of co-activators and co-repressors (Najafabadi et al. 2015; Schmitges et al. 2016; Imbeault et al. 2017). However, many questions remain to be addressed to better understand the role and mechanism of individual KZFP. More targeted studies *in vitro* and using animal models are therefore essential.

1.3 Characterisation of additional members of the KZFP family

1.3.1 ZFP57 is required for maintaining methylation stability at genomic imprints in preimplantation embryos

Genomic imprinting is an epigenetically regulated process that leads to monoallelic expression of genes according to their parental origin, and has been described in fungi, plants and animals (Ferguson-Smith 2011; Surani 1998; Martienssen 1998). Unlike most genes that are biallelically expressed from both maternal and paternal copies, imprinted genes are marked by differential methylation of CpG-rich domains, resulting in the silencing of one copy in a parent-of-origin dependant manner. As a consequence, the imprinted genes show either maternal or paternal expression (Ferguson-Smith 2011). Many known imprinted genes localise in clusters regulated by a cis-acting imprinted control region (ICR), which acquires differential methylation between the two parental chromosomes in the germline (Edwards & Ferguson-Smith 2007). During embryonic development and gonadal sex determination, primordial germ cells undergo genome-wide demethylation to erase previous parental-specific methylation marks that regulate imprinted gene expression (Strogantsev & Ferguson-Smith 2012; Edwards & Ferguson-Smith 2007). Paternal methylation is established during spermatogenesis whereas maternal imprints are established at a later stage, after birth, in growing oocytes. After fertilization, the paternal genome is actively demethylated, while the maternal genome undergoes passive demethylation. Genome-wide remethylation occurs on both parental genomes around implantation. However, imprinting is maintained throughout this post-fertilisation reprogramming, allowing for inheritance of parental-specific monoallelic expression in somatic tissues throughout adulthood (Morgan et al. 2005).

In 2008, in collaboration with the Leder lab (Harvard), the Ferguson-Smith lab discovered that a member of the KRAB ZFPs family, ZFP57, was required for maintaining methylation stability at genomic imprints in preimplantation embryos (X. Li et al. 2008). This work identified ZFP57 as a KRAB-containing protein interacting with KAP1. Generation of *Zfp57* mutant mice indicated that loss of zygotic function of *Zfp57* resulted in partial lethality, while loss of both maternal and zygotic functions resulted in a highly penetrant embryonic lethality. Finally, they showed that differential DNA methylation was lost at several imprinted region in homozygous maternal-zygotic mutants embryos, and concluded that *Zfp57* is an essential maternal-zygotic effect gene maintaining both maternal and paternal methylation imprints after fertilization at multiple imprinted regions (X. Li et al. 2008). That same year, recessive mutations in *ZFP57* were reported in individuals with a pattern of DNA hypomethylation at

imprinted loci throughout the genome and presenting a conserved range of clinical features associated with loss of imprinting, notably transient neonatal diabetes (Mackay et al. 2008). This provided evidence for a mechanistic link between factors mediating epigenetic stability at their target sequences and human health. The Ferguson-Smith lab hypothesised that other KZFPs may have the potential to mediate interactions between sequence-specific genomic loci and epigenetic modifications machinery thus influencing transcription. Therefore, it was proposed to identify and characterise additional members of this family that could modulate epigenetic stability.

1.3.2 EpiHealth European Project

In 2011, the Ferguson-Smith's group joined the collaborative project EpiHealth funded by the European Union framework project FP7. The main goal of the project was to improve human health by understanding the mechanisms and pathways in early development that have a long term effect on the health of individuals across their lifespan. Genetic and epigenetic mechanisms early in life create biological variation and can affect and programme ageing and adult life. The two working hypotheses of the proposal were that (1) the critical window for this programming is during peri-conception oocyte and embryo development and that (2) molecular pathways involved in embryo metabolic and stress adaptation restrict health and longevity in adult life. EpiHealth focused on these early events in several models to decipher some of the most important pathways and potentially offer options for early intervention to avoid adverse health effects. Specific goals included (1) identifying the main genetic pathways affecting the health of developing embryos in a diabetic or obese maternal environment, (2) identifying the main genetic and metabolic pathways affected, and epigenetic imprinting perturbations arising in human and model pre-implantation embryos and assisted reproductive technologies models that may compromise health of the progeny, (3) identifying the key genes and pathways affecting epigenetic and imprinting sensitivity in early stages of development, in order to create intervention tools against epigenetic misprogramming, (4) using bioinformatics tools to link health related pathways with early epigenetic perturbations in order to explain how early events influence the health and lifespan of individuals, and (5) studying the possibility of early intervention to ameliorate the maternal environment. The Ferguson-Smith's lab was involved in the third sub-program objectives through the characterisation of proteins involved in epigenetic control and imprinting sensitivity in mice and mESC, in particular the KZFPs. The hypothesis that the environmental influence on the developmental programme can be mediated epigenetically

was strengthened by the identification of factors contributing to the maintenance of epigenetic stability during development.

1.3.3 Experimental plan to characterise additional KZFPs

Based on ZFP57 results, the Ferguson-Smith lab hypothesised that other KZFPs might influence transcription by targeting epigenetic modifications machinery to their genomic loci, and therefore proposed to develop a project to identify new KZFPs mediating these interactions. This 4-year project fit into the third sub-program objectives of EpiHealth through the characterisation of proteins involved in epigenetic control and imprinting sensitivity during the periconceptual period in mice and in mESC, in particular the KRAB ZFPs. Two research associates – Dr Noon and Dr Shi – initiated the project in 2012 and developed a standard ES cell protocol and *in vivo* analysis to assay genome-wide epigenetic states and identify additional KZFPs that act in the early embryo to modulate epigenetic stability. Dr Noon identified 44 KRAB-ZFPs specifically expressed in mouse ES cells that become down regulated upon differentiation (**Fig 1.11**) (Cloonan et al., 2008, Guttman et al., 2010). The project aimed to (i) identify novel KRAB ZFPs that influence epigenetic states, (ii) identify their genomic targets and understand the relationship between DNA sequence, maintenance of DNA methylation and the recruitment of repressive chromatin complexes, (iii) understand how epigenetic states are stably established at specific regions, and (iv) develop strategies for potential therapeutic targeted modulation of epigenomes. Dr Noon selected 9 candidates that were highly conserved across species (**Fig 1.11**) and proposed to identify their targets in mESCs using chromatin immunoprecipitation followed by sequencing (ChIP-seq) and investigate their roles in the targeted regulation of epigenetic states using knockout mESCs and ChIP-seq for histone modifications in compromised cell lines. Dr Noon created a pipeline for ChIP-seq using tagged KZFPs in mESCs. The details of the experimental plan are presented in Chapter 3.

I joined this project in October 2013 and worked alongside Dr Noon and Dr Shi. The clones overexpressing FLAG-tagged proteins had been generated and 7 samples had been sent to sequencing. The ChIP-seq results became available in December 2013. Only 1 replicate had been sent to sequencing, therefore I generated more FLAG-tagged KZFPs overexpressing clones and prepared the second replicate for 8 samples. I analysed both replicates following Dr Shi guidance and starting the characterisation of the binding sites in mESCs for 8 KZFPs. I decided to focus on one protein to study in more depth and characterise *in vitro* and *in vivo* using a mouse mutant model. Based on the ChIP-seq results that are presented in Chapter

3, I decided to focus on ZFP263 which seemed to be an atypical KZFP. In this dissertation I will present my findings on *Zfp263* only. A brief description of the other initial candidate KZFPs is presented in **Appendix 8.1 – Table 8.1.1**. Further preliminary data on other KZFPs are presented in Chapter 3.

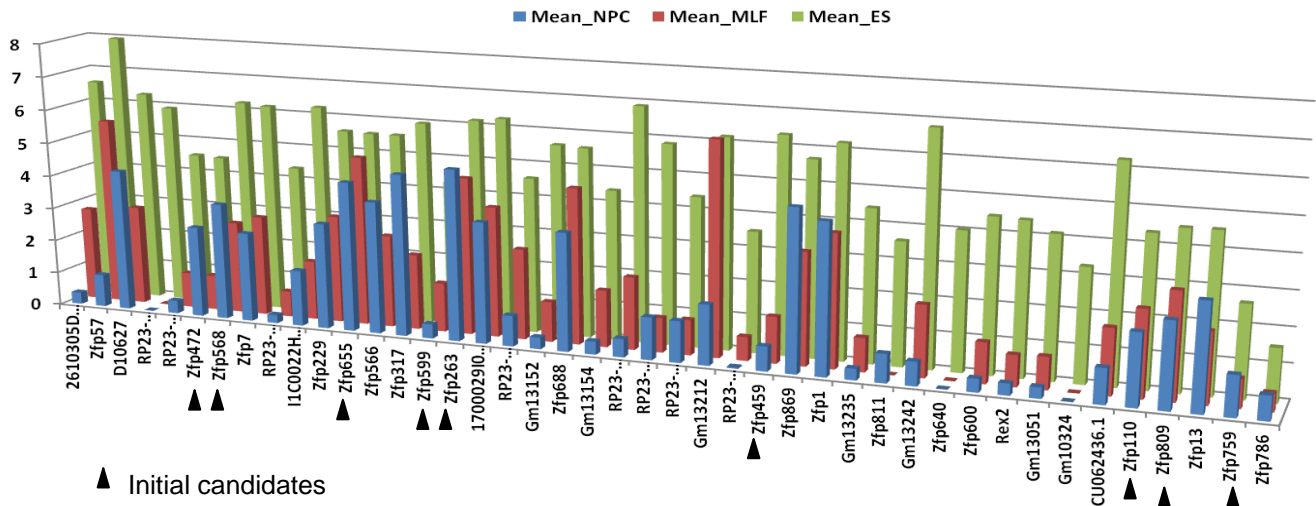


Figure 1.11: Mean KZFPs expression in mouse embryonic stem cells, mouse neural progenitor cells (NPC) and mouse lung fibroblasts (MLF) based on RNA-seq samples from Guttman et al 2010 (Noon A., unpublished).

1.4 ZFP263 in mouse and human

1.4.1 Structure

Human ZNF263 and mouse ZFP263 share 82% identity at the amino acid sequence level. The protein contains a SCAN domain, a KRAB domain and nine zinc fingers (**Fig 1.12**). The SCAN domains are highly conserved between the human and mouse protein (93% identity). They are also very similar to the consensus sequence from Pfam (~85% similarity) (**See Chapter 2**). The KRAB domain is composed of the KRAB A box only and is less conserved between human and mouse (74%). It is also very divergent from the Pfam consensus sequence (34% identity) (**See Chapter 2**). The zinc fingers are C2H2 zinc fingers. ZF1 is isolated from the rest of the ZFs. ZF2 to 4 are Krüppel-type ZFs and are linked together by the consensus sequence TGEK/RPY. ZF4 and ZF5 are linked by a longer sequence, and ZF5 to 9 are again Krüppel-type linked by TGEK/RPY motif.

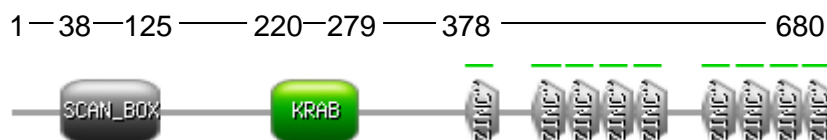


Figure 1.12: Schematic structure of human and mouse ZFP263 from PROSITE (Sigrist et al. 2012) with mouse amino acids corresponding to different domains. The protein contains a SCAN box at its N-terminus, the KRAB A box and 9 zinc fingers at its C-terminus.

1.4.2 Function in human

1.4.2.1 Genomic targets of human ZNF263 in K562 cells

Human ZNF263 has been studied in human chronic myelogenous leukemia cells (K562) (Fietze et al. 2010). They have identified target sites of ZNF263 in K562 by ChIP-seq using a commercial antibody against the endogenous human protein. They found that the binding sites were mostly between -2kb and +2kb of the transcription start site (TSS) or at intragenic regions. A large percentage of the intragenic sites were found within 10kb of the TSS, and 76% of the intragenic sites were found in introns. They identified a 24-nucleotide binding motif that was present in 86% of the TSS binding sites and in 73% of the intragenic category binding sites. Therefore, it seems that ZNF263 is recruited to the intragenic sites using the same motif as used in the core promoter region recruitment. To assess whether ZNF263 acted as a repressor, like most of the KRAB-ZFPs, they assessed target gene expression levels using Affymetric expression microarrays and found that most genes were moderately expressed in wild-type K562 cells, with the set of targets having a similar overall expression as a randomly selected set of genes - even the genes bound by ZNF263 in the promoter region. Fietze *et al* then performed a ZNF263 knock-down (KD) experiment using siRNA treatment in HeLa cells. Upon reduction of ZNF263 level, 195 genes were up-regulated, 61 of which had been identified as ZNF263 targets in K562 cells; and 118 genes were down-regulated, 37 of which were identified as ZNF263 targets in K562. Those findings suggest that binding of ZNF263 to a regulatory region of a target gene can either positively or negatively affect transcription. Using a Gene Ontology analysis with DAVID, they showed that one of the largest categories of genes whose expression decreased upon ZNF263 KD was “Cellular component organization and biogenesis”. The largest categories of genes whose expression increased upon loss of ZNF263 were “negative regulation of biological processes” and “negative regulation of cellular processes”. The authors conclude that ZNF263 may play a critical role in maintaining cell structure and proliferation.

This study did not clarify a transcriptional function for ZNF263 and did not provide evidence for ZNF263 being a repressor as most KZFPs. The genes targeted by ZNF263 in K562 cells display a wide range of expression in wild-type cells, and upon reduction of the protein level some of these targets are up-regulated while some are down-regulated. As the authors suggest, this finding may suggest that ZNF263 is able to act both as a repressor or as an activator, depending on the context. ZNF263 might recruit different co-factors to mediate transcriptional regulation at different targets and may depend on chromatin conformation context or epigenetics state at the targets. Another possibility is that ZNF263 is not directly involved in transcriptional regulation. It may control other regulation pathway by interacting with molecular partners that will in turn impact transcription. That could explain the different outcomes in transcription de-regulation of ZNF263 targets.

1.4.2.2 A possible role for ZNF263 in stress-related diseases

ZNF263 has more recently been mentioned in a stress-related study in children (Nätt et al. 2015). Natt et al. worked on hair samples from a cohort of 48 5-year-old healthy children. They combined hair cortisol measurements, a well-documented biomarker for chronic stress, with whole-genome DNA-methylation sequencing. They found that high levels of cortisol, associated with high level of stress, was generally associated with hypomethylation of differentially methylated regions (DMRs) in SINEs and genes important for calcium transport, phenomena commonly affected in stress-related disease and ageing. They found that 39% of the identified DMRs shared a consensus DNA sequence. The authors compared this sequence with predicted TF binding sites and identified three ZFPs with significantly similar binding motif: ZNF263, EGR1 and SP1. By comparing their DMRs locations with the TF binding sites, they found that ZNF263 was associated with the hypomethylated DMRs in a proximal region, while EGR1 bound more distally. From previous ChIP-seq data they also found that SINEs were the most abundant repeats in ZNF263 binding, as in their DMRs, and so concluded that high cortisol is associated with a decrease in DNA methylation at ZNF263 binding sites and targets SINEs across the genome. Because KRAB ZFPs are mostly known for their interaction with KAP1 and their repressive effects, they hypothesised that the loss of methylation at ZNF263 binding sites in stressed children could be directly mediated by the loss of ZNF263 itself.

This work clearly lacks functional validation and their conclusion is very much based on hypothetical function of ZNF263. The authors claim that “high cortisol is associated with a decrease in DNA methylation at ZNF263 binding sites”, whereas in reality they only found

that some regions bound by ZFP263 were close to the identified DMRs. They did not actually verify the DNA methylation level at all known ZNF263 binding sites. They do not show a direct overlap between ZNF263 binding sites and DMRs either. The DMR motif also differs from ZNF263 binding motif identified by Fietze et al. Furthermore, available ChIP-seq data from Fietze et al. did not show any significant enrichment of ZNF263 with SINE elements. Finally, Natt et al. hypothesise that ZNF263 is a repressor of transcription like most KZFPs, whereas results from Fietze et al. did not provide evidence for a clear role in transcriptional repression. For these reasons, their hypothesis that hypomethylation in stressed children would be mediated by loss of ZNF263 itself seems unlikely.

1.4.2.3 Large-scale studies

In the past two years, two groups have studied KZFPs on a larger scale and provided information on human ZNF263. In HEK293 cells, Schmitges et al. performed ChIP-seq on 131 tagged ZFPs, while Imbeault et al. performed ChIP-exo on 257 tagged-KZFPs. The ZNF263 recognition motif was similar to the one found in K562 cells by Fietze et al. Both papers discuss that SCAN-containing ZFPs do not target endogenous retroelements. Imbeault et al. also suggest that SCAN ZFPs have lost the ability to recruit KAP1, although the data is not shown. However, they suggest that ZNF263 co-localizes with another SCAN protein, ZKSCAN2, although the peak for this latter protein is much broader and seems less convincing. A protein-protein interaction analysis shows that ZNF263 interacts with SCAND1 and ZKSCAN18 in HEK293 cells, but not with KAP1 (**Fig 1.9 A**) (Schmitges et al. 2016). The BIOPLEX network identifies 57 partners for ZNF263 (**Fig 1.13**), including SCAND1 and ZSCAN20 that interacts with most SCAN-containing proteins. Interestingly, it seems that ZNF263 interacts with a lot of non ZFPs, but not with KAP1.

Although these studies provide useful information on ZNF263, more work remains to be done to fully characterise this ZFP and elucidate its function. The work from Fietze *et al.* has been performed in established abnormal tissue culture cells only, did not provide a clear indication of function and no mechanistic study was carried out. The second work only provides a hypothesis regarding ZNF263 involvement in stress-related disease, but does not provide any evidence for a real implication in this process. The last two studies provide to the community NGS data that can be used and analysed in more details, but do not bring more information on the role and mechanism of individual ZFPs.

training, obtain new transferable skills, and be exposed to different working environments. I went to the Beijing Genomic Institute in Shenzhen, China, for six weeks as a member of Professor Li Qibin's team. For my second secondment, I worked at Celgene in Seville, Spain, for three weeks as a member of the team of Dr. Matthew Trotter. A brief report was written for each project and submitted to the European Project Officer – both reports are presented in **Appendix 8.1.2 and 8.1.3.**

As briefly explained in paragraph 1.3.3, I decided to focus on one member of the KZFP family in mouse, ZFP263. This decision was based on preliminary results from ChIP-seq data that Dr Shi and Dr Noon initiated. Some of these results are presented in Chapter 3. My aim was to characterise ZFP263 by understanding its mechanism of actions in mESCs and its role in mice. I first proposed to investigate *Zfp263* evolution in different species and across tissues in mouse and human. I assessed *Zfp263* expression in different tissues and studied its conservation across different species to gain insight into its function. I aimed at identifying ZFP263 genomic targets and the key genes and pathways that it regulated. To answer this question, my objective was to perform a ChIP-seq experiment in mESC and characterise its binding sites. I developed a pipeline to analyse the epigenetic states around the target sites and developed hypotheses about ZNF263 function *in vitro*. Finally, I aimed to understand ZFP263 function *in vivo*, during embryogenesis and adult development. To do this, I generated *Zfp263* KO mice using the CRISPR-Cas9 technology with the goal of investigating their development and phenotype to shed light on ZFP263 role *in vivo*.

Chapter 2: *Zfp263* conservation and expression

2.1 Introduction

2.1.1 Vertebrate evolution

The first vertebrate appeared more than 500 million years ago. Vertebrates are chordates that share several common anatomical features such as a vertebral column, a head with the brain and a concentration of other sense organs and a skeleton made of bone and cartilage, amongst other characteristics. Emergence of the vertebrate lineage was also accompanied by acquisition of the neural crest, whose cells contribute to the development of diverse structures by undergoing an epithelial-to-mesenchymal transition and migrating to multiple sites (Green et al. 2015). Vertebrates include fish, amphibians, birds, reptiles and mammals. The first vertebrates to appear were the jawless fishes, almost completely extinct today except for the hagfishes and the lampreys. The vertebrates also include the cartilaginous and bony fishes, the amphibians who first colonised land, the reptiles and birds who lay hard-shelled eggs, and the mammals (**Fig 2.1**). Mammals have fur, mammary glands, and split into three different groups: the monotremes with only three living species and who are the only mammals to lay eggs; the marsupials whose young develop first in the mother's uterus and then postnatally in the pouch; and the eutherian, or "true placental mammals".

Until recently, the KRAB domain was thought to be restricted to tetrapods, but Imbeault *et al.* found KZFP genes in *Latimeria chalumnae*, the African coelacanth, and thus reassigned the root of the family to the Sarcopterygian common ancestor of coelacanth, lungfish and tetrapods (**Fig 2.1**) (Imbeault et al. 2017). They found that most KRAB ZFPs were restricted to primates or eutherian, but those with a SCAN domain were older and often shared with marsupials or sauropsids.

It is suggested that environmental stimuli can promote epigenetic changes that might result in phenotypic changes, thus shaping epigenomes and impacting genome function over evolutionary time (Varriale & Annalisa 2014). Similarly, repetitive DNA and retrotransposons are now considered to be potential drivers of evolution, for example by changing chromatin structure or rewiring transcriptional networks (Zuckerandl & Cavalli 2007; Imbeault et al. 2017). KZFPs are DNA-binding proteins targeting epigenetic states to their genomic locations, rapidly evolving, and targeting mainly transposable elements, which suggests they could be key players of vertebrate evolution. Therefore it is of value to study *Zfp263* evolution and conservation across species, to learn more about its history and potential role.

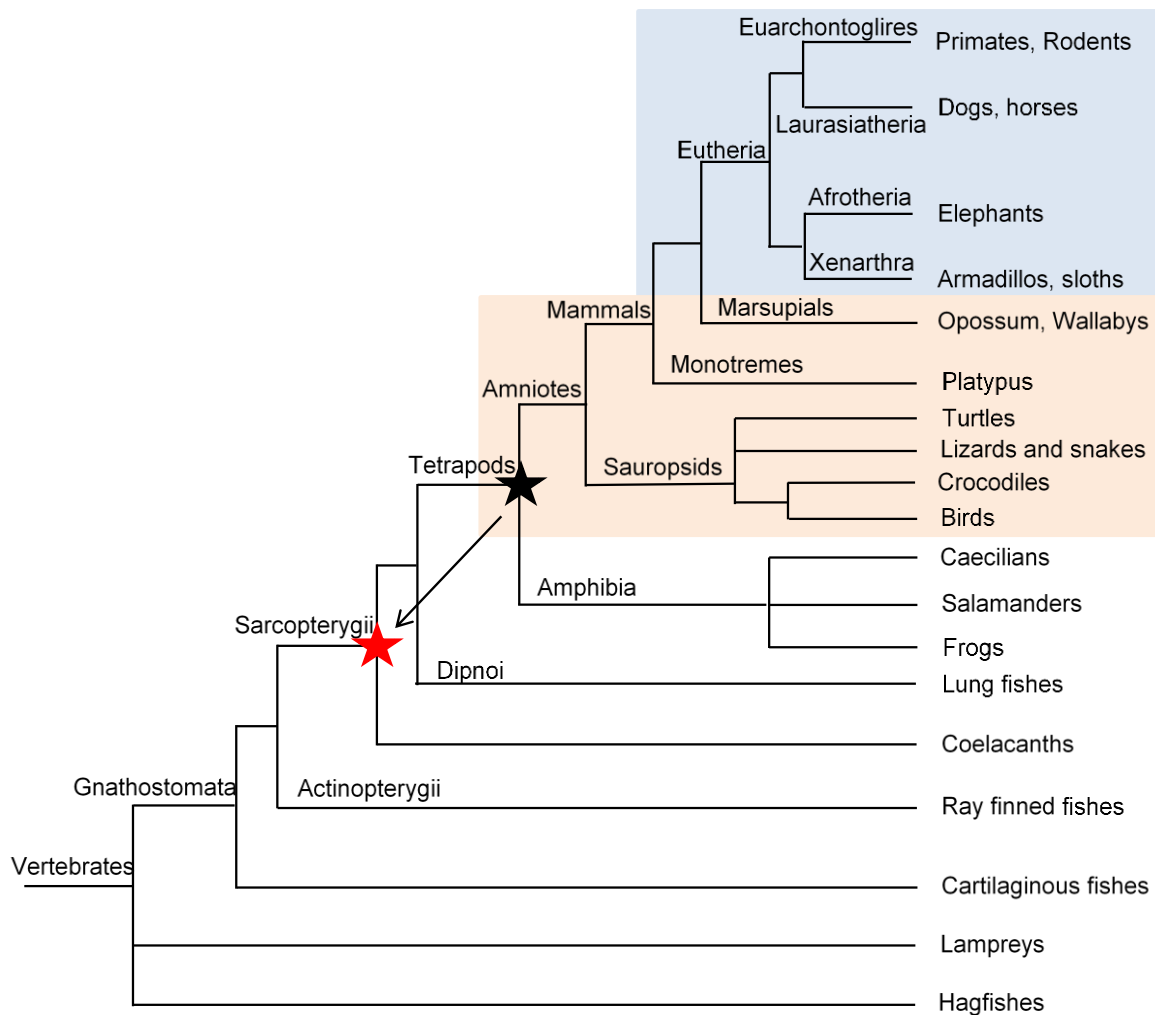


Figure 2.1: Phylogenetic tree of vertebrates, adapted from (Irisarri et al. 2017) The root of the ZFP family has been reassigned to Sarcopterygii instead of Tetrapods (stars and arrows). Human ZFPs with a KRAB domain were found to be primate- or eutheria-restricted (light purple) but those with a SCAN domain were shared with marsupials and Sauropsids (light orange) (Imbeault et al. 2017).

2.1.2 Experimental plan

First, *ZNF263* orthologues were identified and comparatively analysed to trace back its origin and first time of emergence. Their CDS and amino acids sequences were aligned and used to build a maximum likelihood phylogenetic tree that illustrates the relative amino acids difference between species. To investigate the evolutionary relationship between them, percentage identity and similarity between coding sequences and amino acid sequences were calculated. Two sequences with a high percentage of identity or similarity are thought to share a common ancestor or function. The amino acids sequences were also used to determine a hypothesis around whether the protein has similar roles in all species or whether it evolved to perform different functions. Second, the Ka/Ks ratios were calculated indicating

the balance between neutral mutations, purifying selection and beneficial mutations between homologous genes. This gives a good estimation of how conserved the protein has been during evolution (Li et al. 1985; Hurst 2002).

Finally *Zfp263* expression was assessed in different adult and embryonic mouse and human tissues by quantitative PCR and using publically available datasets. Its expression pattern has the potential to provide insight into its function and provides a framework for future experiments.

2.2 Results

2.2.1 Conservation across species

This analysis has been done under the guidance of Dr Carol Edwards from the Ferguson-Smith's lab.

Human *ZNF263* orthologues were identified using NCBI Gene and the Ensembl Genome Browser (Yates et al. 2016) (**See Methods**). The number of exons, the size of the protein and the surrounding genes were checked to verify the accuracy of the orthologues. *ZNF263* orthologues were found in 101 species on NCBI and 42 species in Ensembl. No fish, birds or reptiles contained *ZNF263* gene. From Ensembl, *ZNF263* was found in one species of Monotreme, 3 species of Marsupial and in all groups of Eutherian mammals: *ZNF263* was identified in 2 species of Xenarthra, 2 species of Afrotheria, 14 species of Laurasiatheria, 12 primates species and 8 rodents (**Table 2.1**), showing that the gene is well represented in all orders and clades of mammals. Over 80% of the homologous sequences from each species match the human sequence (Target %id) except for the few species from Xenarthra, Marsupials and Monotremes. However these latter sequences were most likely not annotated properly, as they were found on NCBI with better percentage identity.

Table 2.1 (next page): Human *ZNF263* orthologues from Ensembl. For each species, the type and name of the orthologue is given, with the percentage of the homologous sequence matching the human sequence (Target %id), the percentage of the human sequence matching the sequence of the orthologue (Query %id) and the clade or order of each species. Species underlined were used to represent each clade or order in further analyses.

Species	Type	Orthologue	Target %id	Query %id	Class / order
Bushbaby (<i>Otolemur garnettii</i>)	1-to-1	ZNF263	89.62%	89.75%	Primates
<u>Chimpanzee</u> (<i>Pan troglodytes</i>)	1-to-1	ZNF263 (ENSPTRG00000007692)	99.27%	99.27%	
Gibbon (<i>Nomascus leucogenys</i>)	1-to-1	ZNF263 (ENSNLEG00000009603)	97.95%	97.95%	
<u>Gorilla</u> (<i>Gorilla gorilla gorilla</i>)	1-to-1	ZNF263 (ENSGGOG00000010468)	99.12%	99.12%	
Macaque (<i>Macaca mulatta</i>)	1-to-1	TIGD7 (ENSMMUG00000017973)	95.89%	95.61%	
<u>Marmoset</u> (<i>Callithrix jacchus</i>)	1-to-1	ZNF263 (ENSCJAG00000019612)	96.05%	96.19%	
Mouse Lemur (<i>Microcebus murinus</i>)	1-to-1	ZNF263 (ENSMICG00000014148)	92.40%	92.53%	
Olive baboon (<i>Papio anubis</i>)	1-to-1	ZNF263 (ENSPANG00000017591)	97.09%	97.66%	
<u>Orangutan</u> (<i>Pongo abelii</i>)	1-to-1	ZNF263 (ENSPPYG00000007038)	98.98%	98.98%	
Tarsier (<i>Tarsius syrichta</i>)	1-to-1	ZNF263 (ENSTSYG00000012027)	74.18%	63.10%	
Vervet-AGM (<i>Chlorocebus sabaeus</i>)	1-to-1	ZNF263 (ENSCSAG00000010072)	98.10%	98.24%	
Guinea Pig (<i>Cavia porcellus</i>)	1-to-many	ENSCPOG00000009776	93.41%	35.29%	Rodents
Guinea Pig (<i>Cavia porcellus</i>)	1-to-many	ENSCPOG00000012343	87.95%	50.22%	
Kangaroo rat (<i>Dipodomys ordii</i>)	1-to-1	Zfp263 (ENSODRG00000001453)	74.93%	74.38%	
<u>Mouse</u> (<i>Mus musculus</i>)	1-to-1	Zfp263 (ENSMUSG00000022529)	84.41%	84.04%	
Pika (<i>Ochotona princeps</i>)	1-to-1	ZNF263 (ENSOPRG00000012974)	86.03%	85.65%	
Rabbit (<i>Oryctolagus cuniculus</i>)	1-to-1	ZNF263 (ENSOCUG00000012210)	87.12%	87.12%	
<u>Rat</u> (<i>Rattus norvegicus</i>)	1-to-1	Zfp263 (ENSRNOG00000007678)	84.12%	83.75%	
Squirrel (<i>Ictidomys tridecemlineatus</i>)	1-to-1	ZNF263 (ENSSTOG00000013118)	91.82%	92.09%	
Tree Shrew (<i>Tupaia belangeri</i>)	1-to-1	ZNF263 (ENSTBEG00000012439)	81.80%	79.65%	
Alpaca (<i>Vicugna pacos</i>)	1-to-1	ZNF263 (ENSVPAG00000011425)	83.06%	59.59%	Laurasiatheria
Cat (<i>Felis catus</i>)	1-to-1	ENSFCAG00000028681	95.00%	16.69%	
<u>Cow</u> (<i>Bos taurus</i>)	1-to-1	ZNF263 (ENSBTAG00000018625)	88.94%	89.46%	
<u>Dog</u> (<i>Canis lupus familiaris</i>)	1-to-1	ZNF263 (ENSACFG00000019292)	92.11%	92.24%	
Dolphin (<i>Tursiops truncatus</i>)	1-to-1	ZNF263 (ENSTTRG00000011649)	90.09%	74.52%	
Ferret (<i>Mustela putorius furo</i>)	1-to-1	ZNF263 (ENSMPUG00000015125)	73.63%	67.06%	
Hedgehog (<i>Erinaceus europaeus</i>)	1-to-1	ZNF263 (ENSEEUG00000009569)	86.01%	66.62%	
<u>Horse</u> (<i>Equus caballus</i>)	1-to-1	ZNF263 (ENSECAG00000024249)	91.53%	91.80%	
Megabat (<i>Pteropus vampyrus</i>)	1-to-1	ZNF263 (ENSPVAG00000001368)	78.77%	78.77%	
Microbat (<i>Myotis lucifugus</i>)	1-to-1	ENSMLUG00000024399	86.47%	50.51%	
Panda (<i>Ailuropoda melanoleuca</i>)	1-to-1	ZNF263 (ENSAMEG00000010624)	91.81%	91.95%	
Pig (<i>Sus scrofa</i>)	1-to-1	ZNF263 (ENSSSCG00000007963)	88.89%	89.02%	
Sheep (<i>Ovis aries</i>)	1-to-1	ENSOARG00000001622	84.94%	47.88%	
Shrew (<i>Sorex araneus</i>)	1-to-1	ZNF263 (ENSSARG00000000228)	79.09%	79.21%	
<u>Elephant</u> (<i>Loxodonta africana</i>)	1-to-1	ZNF263 (ENSLAFG00000000525)	88.18%	88.43%	Afrotheria
Hyrax (<i>Procavia capensis</i>)	1-to-1	ZNF263 (ENSPCAG00000003912)	58.42%	58.42%	
Lesser hedgehog tenrec (<i>Echinops telfairi</i>)	1-to-1	ZNF263 (ENSETEG00000020049)	85.80%	85.80%	Xenarthra
Sloth (<i>Choloepus hoffmanni</i>)	1-to-1	ZNF263 (ENSCHOG00000000602)	63.78%	63.69%	
<u>Armadillo</u> (<i>Dasypus novemcinctus</i>)	1-to-1	ENSDNOG00000038909	34.73%	18.16%	Marsupials
<u>Opossum</u> (<i>Monodelphis domestica</i>)	1-to-1	ZNF263 (ENSMODG00000016443)	70.32%	70.42%	
Tasmanian devil (<i>Sarcophilus harrisii</i>)	1-to-1	ENSSHAG00000002634	58.87%	22.84%	
Wallaby (<i>Macropus eugenii</i>)	1-to-1	ZNF263 (ENSMEUG00000006986)	68.97%	68.67%	Monotreme
<u>Platypus</u> (<i>Ornithorhynchus anatinus</i>)	1-to-1	ENSOANG00000030916	45.42%	16.69%	

ZNF263 orthologues from 14 organisms were chosen to represent each group (**Table 2.1**) for further analyses and their CDS and amino acids sequences were aligned (**Appendix 8.2.1 and Fig 2.2**). It is quite striking that *ZFP263* is very highly conserved, and found to contain nine C2H2 zinc fingers in all organisms. As presented in the Chapter 1, it has been suggested that 4 amino acids from a zinc finger are particularly involved in the binding of DNA as interactions with DNA are made through specific hydrogen-bond interactions from amino acids at helical position -1, 2, 3 and 6 to four consecutive bases on both strands of the DNA. These amino acids have been proposed to define a zinc finger “fingerprint” and were used to identify new ZFPs orthologues (Liu et al. 2014). Interestingly, for *ZFP263*, these amino acids are identical in all species (**Fig 2.2 – green residues**). The SCAN and KRAB domains are shown on **Fig 2.2** with a blue and orange arrow respectively, and the alignment is zoomed in **Fig 2.3**. The consensus sequences were also downloaded from the Pfam database and were aligned to the SCAN and KRAB domains of the 14 orthologues. The SCAN domain is highly conserved from one species to another and highly similar to the consensus sequence (**Fig 2.3 A**). According to the structural study of different SCAN domain by Nam et al, 2010, *ZFP263* SCAN domain could form 3 α -helices on the N-terminus of the domain (**Fig 2.3A Blue**). Interestingly, these 3 sequences are identical in almost all species. The KRAB domain however seems less conserved between species and more divergent from the consensus sequence (**Fig 2.3 B**). Three regions within the KRAB domain have been shown to be essential for KAP1 recruitment (Margolin et al. 1994). Site-directed substitution at these sites impaired the KRAB domain in its ability to repress transcription (**Fig 1.8**). These three regions are highlighted in red in **Fig 2.3 B**. Interestingly, the first domain is conserved in Platypus, Opossum, Armadillo and Marmoset, but in all other species the second Valine is replaced with a Methionine. This substitution may have occurred during evolution in Afrotheria. The second domain is conserved only in primates but not in other species. The last domain is not conserved at all in any of the species analysed here, with either all three amino acids replaced or two out of three.

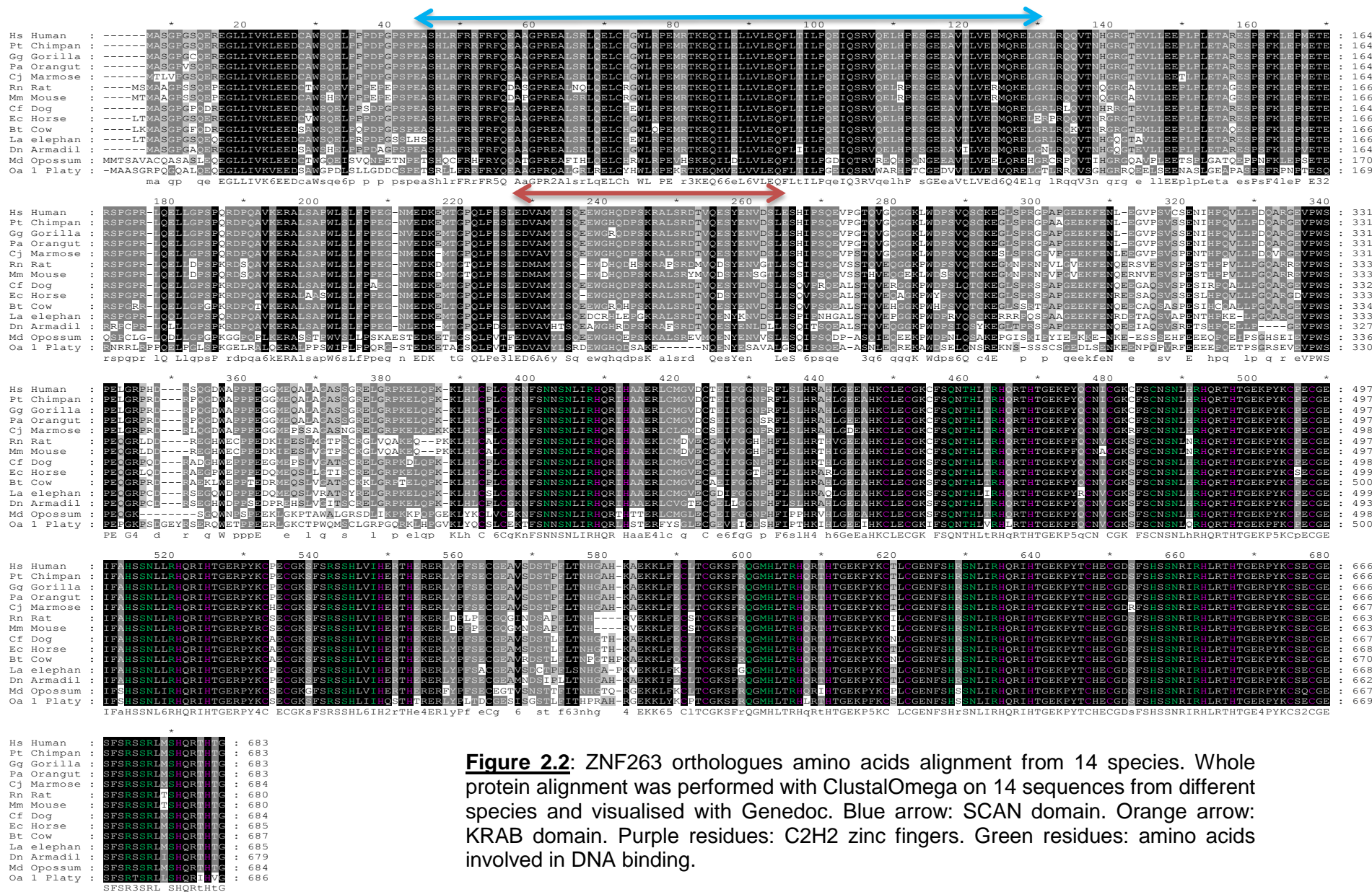


Figure 2.2: ZNF263 orthologues amino acids alignment from 14 species. Whole protein alignment was performed with ClustalOmega on 14 sequences from different species and visualised with Genedoc. Blue arrow: SCAN domain. Orange arrow: KRAB domain. Purple residues: C2H2 zinc fingers. Green residues: amino acids involved in DNA binding.

Pfam SCAN : PEASRQFRFRQFRYQEAEGPREALSRLRELRCQWLRPEVHTKEQILELLLVLEQFLTILPPELQAWVREKKPESGEEAVALAEDLEREL : 87
 Hs Human S : PEASHLRFRRFRFRQEAAGPREALSRLQELCHGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQREL : 87
 Pt Chimpan : PEASHLRFRRFRFRQEAAGPREALSRLQELCHGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQREL : 87
 Gg Gorilla : PEASHLRFRRFRFRQEAAGPREALSRLQELCHGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQREL : 87
 Pa Orangut : PEASHLRFRRFRFRQEAAGPREALSRLQELCHGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQREL : 87
 Cj Marmose : PEASHLRFRRFRFRQEAAGPREALSRLQELCHGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQREL : 87
 Rn Rat SCA : PEASHLRFRRFRFRQDASGPREALNQLOELCRGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELRPESGEEAVTLVERMQKEL : 87
 Mm Mouse S : PEASHLRFRRFRFRQDAPGPREALSRLQELCRGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELRPESGEEAVTLVERMQKEL : 87
 Cf Dog SCA : PEASHLRFRRFRFRQEAAGPREALSRLQELCHEWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQREL : 87
 Ec Horse_S : PEASHLRFRRFRFRQEAAGPREALSRLQELCHGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQREL : 87
 Bt_Cow_SCA : PEASHLRFRRFRFRQEAAGPREALSRLQELCHGWLOPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQKEL : 87
 La elephant : LHSHLRFRRFRFRQEAAGPREALSRLQELCHEWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVTLVEDMQREL : 87
 Dn Armadil : PEASHLRFRRFRFRQEAAGPREALSRLQELCHGWLRPEMRTKEQILELLLVLEQFLTILPQEIQSRVQELHPESGEEAVILVEDMQREL : 87
 Md Opossum : PETSHQCFRHFRYQQAATGPREAFIHLQELCHRWLRPEVHSKEQILDLLVLEQFLTILPGDIQTRVREQHPQNGEEAVTLVEELQREH : 87
 Oa 1 Platy : PETSRLLFRFRFRYQEAAGPRQALGRLRELCHYHWLKPEKRTKEQMVELVLEQFLTILPGEIQSRVWARHPTCGEDVVTLEDVQREL : 87
 peaShlrFRrFR5Q A GPR2AlsrLqELC WL PE r3KEQ66eL6VLEQFLtILP e6Q rVqelhP sGEeaVtLvEd624E1

Pfam : FEDVAVDFTQ^{*}EWEELLDPAQEFLYRDVMLENYRN---- : 34
Hs Human K : LEDVAMYISOEWGHHQDPSKRALSRTTVQESYENVDSL : 38
Pt_Chimpan : LEDVAMYISOEWGHHQDPSKRALSRTTVQESYENVDSL : 38
Gg_Gorilla : LEDVAMYISOEWGRQDPSKRALSRTTVQESYENVDSL : 38
Pa_Orangut : LEDVAMYISOEWGHHQDPSKRALSRTTVQESYENVDSL : 38
Cj_Marmose : LEDVAVYISOEWGHHQDPSKRALSRTTVQESYENVDSL : 38
Rn Rat KRA : LEDMAMYISQ-EWDHQDH^{*}SKRAPSRDMVQDSYENVGTL : 37
Mm Mouse K : LEDMAMYISQ-EWDHQDPSKRALSRMYMVQDSYENSGLT : 37
Cf_Dog_KRA : LEDVAMYISOEWGHHQDPSKRALSRTTVQESYENVDSL : 38
Ec_Horse_K : LEDVAMYISQ-EWGHQDPSKRALSRTTVQDSYENVDSL : 37
Bt_Cow KRA : LEDVAMYISOEWGGROHPSKRALSRTTVQESYENVDSL : 38
La_elephan : LEDVAMXLSEEDCRHLLEPGKRALSRTTVQENYKNVDSL : 38
Dn_Armadil : LEDVAVHTSQEAWGHRDPSKRAF^{*}SRTTVQESYENLDLL : 38
Md_Opossum : FEDVAVXLSQEAWGHEPSSKALSRREVMOENYENVVSLL : 38
Oa_Platypu : FEDVAVYLSDREWGHQDSAKE-----NOENYESAVAL : 32

1ED6A6y 3q ew hqdpsk alsrd ge Yen 1

55

The maximum-likelihood phylogenetic tree was built from the alignment with 500 bootstraps (**Fig 2.4**). It recapitulates the current mammalian tree model and the branch lengths indicate the changes of amino acids. Most amino acids changes happened before the Afrotheria clade. In Eutheria, the two rodent species show the longest branch, indicating the greater number of residues changes.



Figure 2.4: Phylogenetic maximum likelihood tree with 500 bootstraps based on ZFP263 amino acids sequences from 14 species.

To precisely evaluate the level of similarity between sequences, the percentage of identical and similar amino acids between representatives was then calculated. For the full length protein, the percentage of identity and similarity is very high between orthologues (**Table 2.2 A**). They are all more than 70% similar and even above 90% from the Xenarthra super-order. The SCAN domain is also very well conserved across species. The SCAN domains from platypus and armadillo share 70 to 80% similarities with the other species. From the afrotheria clade, SCAN domains are all 90% identical between each other. They are also all very similar to the consensus sequence from Pfam, sharing more than 80% similar amino acids (**Table 2.2 C**). On the other hand, the KRAB domain is less conserved and much more divergent from the consensus sequence, since they all share only 30 to 50% similar amino acids with the consensus sequence. The platypus KRAB domain only shares 34% identity with the consensus KRAB sequence, suggesting that the protein first emerged with an atypical KRAB domain. The opossum KRAB domain mutated and increased its similarity to the consensus sequence, but the %identity and similarity decreased again after the opossum, suggesting that this domain has not been highly conserved and was subjected to mutations. The KRAB domains are also more divergent between the different species and must have emerged for the first time before the platypus in a non-canonical version (**Table 2.2 D**).

Table 2.2: Percentage of identical and similar amino acids between human ZNF263 and its orthologues calculated by Ident and Sim and their Ka/Ks ratio calculated in R. A. Percentage of identical and similar amino acids in the full length protein between orthologues. B. Ka/Ks ratio between orthologues. C. Percentage of identical and similar amino acids in the SCAN domain between orthologues and the Pfam consensus sequence. D. Percentage of identical and similar amino acids in the KRAB domain between orthologues and the Pfam consensus sequence.

A	Whole Protein	chimpanzee	gorilla	orangutan	marmoset	rat	mouse	dog	horse	cow	elephant	armadillo	opossum	platypus
	human	99.2/99.4	99.1/99.4	98.9/99.4	95.9/96.7	82.5/86.7	82.8/87.1	92.1/94.3	91.4/93.7	88.9/91.9	88/90.5	88.3/90.6	68.9/76.5	61.8/71.3
	chimpanzee		99.5/99.7	99.4/99.7	95.9/96.6	82.8/87.1	82.9/87.3	91.6/94.1	91.2/93.5	88.9/91.8	87.9/90.3	88.1/90.6	68.9/76.7	61.6/71.1
	gorilla			99.2/99.7	95.7/96.6	82.5/86.9	82.8/87.3	91.9/94.3	91.1/93.5	89.2/91.9	87.4/90.1	88.1/90.7	68.6/76.4	61.5/71.1
	orangutan				95.9/96.6	82.6/86.9	82.9/87.3	91.8/94.1	91.1/93.5	88.7/91.8	87.4/90.1	88/90.6	68.7/76.4	61.8/71.3
	marmoset					81.9/86.3	82.2/86.7	90.5/93.4	89.3/92.2	87.2/90.8	86/89.2	87.2/90.3	68.7/76.4	61.2/70.8
	rat						96.4/97.0	83.1/88.2	84.1/88.6	81.2/86.3	79.0/84.7	80.3/85	65.2/73.9	59.6/69.1
	mouse							82.9/88	83.8/88.6	81.2/86.1	79.0/84.4	80.4/85.3	64.6/73.3	59.8/69.3
	dog								93.4/95.6	92.4/94.7	87.9/90.8	89/91.8	69.8/77.1	63.1/71.6
	horse									91.2/93.8	87.6/90.6	88.1/90.9	68.3/77.0	62.2/71.4
	cow										86/89	86.6/89.9	67.9/76.3	61.5/71
	elephant											85.7/87.6	68.2/75.9	60.8/70.7
	armadillo												66.9/74.2	61.9/70.8
	opossum													64.3/74.8
	platypus													

B	Ka / Ks	chimpanzee	gorilla	orangutan	marmoset	rat	mouse	dog	horse	cow	elephant	armadillo	opossum	platypus
	human	0.48	0.21	0.12	0.17	0.20	0.19	0.11	0.15	0.16	0.19	0.19	0.24	0.26
	chimpanzee		0.22	0.10	0.19	0.20	0.19	0.12	0.15	0.16	0.19	0.20	0.24	0.26
	gorilla			0.11	0.18	0.20	0.18	0.12	0.15	0.16	0.19	0.19	0.24	0.26
	orangutan				0.20	0.21	0.19	0.12	0.16	0.17	0.20	0.20	0.23	0.26
	marmoset					0.20	0.19	0.13	0.17	0.18	0.19	0.19	0.22	0.25
	rat						0.11	0.16	0.17	0.19	0.23	0.20	0.22	0.26
	mouse							0.16	0.17	0.18	0.21	0.19	0.21	0.23
	dog								0.13	0.12	0.18	0.16	0.21	0.25
	horse									0.18	0.22	0.20	0.24	0.26
	cow										0.22	0.21	0.24	0.27
	elephant											0.27	0.23	0.26
	armadillo												0.24	0.26
	opossum													0.23
	platypus													

C

SCAN Domain	human	chimpanzee	gorilla	orangutan	marmoset	rat	mouse	dog	horse	cow	elephant	armadillo	opossum	platypus
Pfam	75.8/86.2	75.8/86.2	75.8/86.2	75.8/86.2	75.8/86.2	71.2/82.7	73.5/85.0	75.8/87.3	75.8/86.2	73.5/85.0	72.4/83.9	74.7/86.2	67.8/82.7	64.3/79.3
human		100/100	100/100	100/100	100/100	90.8/95.4	93.1/97.7	98.8/98.8	100/100	97.7/98.8	95.4/95.4	97.7/97.7	71.2/83.9	73.5/82.7
chimpanzee			100/100	100/100	100/100	90.8/95.4	93.1/97.7	98.8/98.8	100/100	97.7/98.8	95.4/95.4	97.7/97.7	71.2/83.9	73.5/82.7
gorilla				100/100	100/100	90.8/95.4	93.1/97.7	98.8/98.8	100/100	97.7/98.8	95.4/95.4	97.7/97.7	71.2/83.9	73.5/82.7
orangutan					100/100	90.8/95.4	93.1/97.7	98.8/98.8	100/100	97.7/98.8	95.4/95.4	97.7/97.7	71.2/83.9	73.5/82.7
marmoset						90.8/95.4	93.1/97.7	98.8/98.8	100/100	97.7/98.8	95.4/95.4	97.7/97.7	71.2/83.9	73.5/82.7
rat							96.5/96.5	89.6/94.2	90.8/95.4	90.8/94.2	86.2/90.8	88.5/93.1	67.8/82.7	66.6/79.3
mouse								91.9/96.5	93.1/97.7	93.1/96.5	88.5/93.1	90.8/95.4	67.8/82.7	67.8/80.4
dog									98.8/98.8	96.5/97.7	96.5/96.5	96.5/96.5	71.2/83.9	73.5/82.7
horse										97.7/98.8	95.4/95.4	97.7/97.7	71.2/83.9	73.5/82.7
cow											93.1/94.2	95.4/96.5	68.9/82.7	72.4/81.6
elephant												93.1/93.1	68.9/82.7	71.2/81.6
armadillo													68.9/81.6	71.2/81.4
opossum														62/83.9
platypus														

D

KRAB Domain	human	chimpanzee	gorilla	orangutan	marmoset	rat	mouse	dog	horse	cow	elephant	armadillo	opossum	platypus
Pfam	44.7/47.3	44.7/47.3	44.7/47.3	44.7/47.3	47.3/50	31.5/39.4	34.2/42.1	44.7/47.3	39.4/44.7	42.1/44.7	42.1/55.2	42.1/44.7	50/60.5	34.2/39.4
human		100/100	97.3/100	100/100	97.3/97.3	76.3/81.5	76.3/81.5	100/100	94.7/97.3	94.7/97.3	76.3/84.2	78.9/81.5	73.6/84.2	47.3/55.2
chimpanzee			100/100	100/100	97.3/97.3	76.3/81.5	76.3/81.5	100/100	94.7/97.3	94.7/97.3	76.3/84.2	78.9/81.5	73.6/84.2	47.3/55.2
gorilla				97.3/100	94.7/97.3	76.3/81.5	76.3/81.5	97.3/100	92.1/97.3	97.3/97.3	76.3/84.2	76.3/81.5	71.0/84.2	44.7/55.2
orangutan					97.3/97.3	76.3/81.5	76.3/81.5	100/100	94.7/97.3	94.7/97.3	76.3/84.2	78.9/81.5	73.6/84.2	47.3/55.2
marmoset						73.6/78.9	73.6/78.9	97.3/97.3	92.1/94.7	92.1/94.7	73.6/81.5	81.5/84.2	76.3/86.8	50/57.8
rat							89.1/89.1	76.3/81.5	81.0/83.7	71.0/78.9	55.2/68.4	60.5/65.7	55.2/73.6	39.4/50
mouse								76.3/81.5	81.0/83.7	71.0/78.9	55.2/68.4	60.5/63.1	57.8/73.6	39.4/47.3
dog									94.7/97.3	94.7/97.3	76.3/84.2	78.9/81.5	73.6/84.2	47.3/55.2
horse										89.4/94.7	71.0/81.5	73.6/78.9	68.4/81.5	44.7/52.6
cow											73.6/81.5	73.6/78.9	71.0/81.5	42.1/52.6
elephant												63.1/68.4	65.7/73.6	36.8/50
armadillo													60.5/71	42.1/50
opossum														57.8/65.7
platypus														

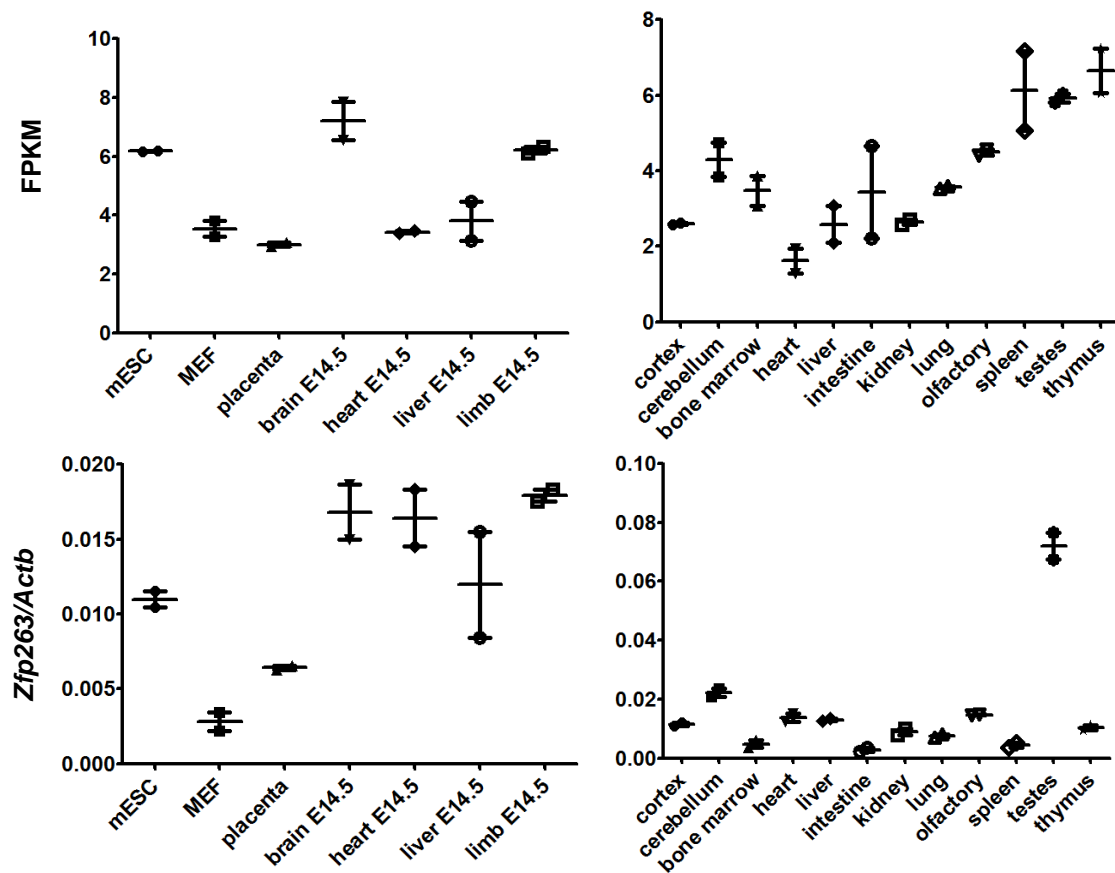
Finally the Ka/Ks ratio was calculated. Ka is the number of nonsynonymous mutation per nonsynonymous sites in a given period of time, while Ks is the number of synonymous mutations per synonymous sites. Synonymous mutations are neutral for the protein function, so a ratio equals to 1 indicates a neutral selection. A ratio greater than 1 shows that there are more nonsynonymous mutations than synonymous mutations, which indicates a positive selection in favour of mutations. For ZFP263, the ratio is very close to 0, indicating a purifying selection (**Table 2.2 B**). This means that natural selection is acting against mutation, maintaining conservation and stabilizing the protein sequence and its function.

2.2.2 *Zfp263* expression in mouse and human

Next, *Zfp263* expression was assessed to provide insight into its function in vivo. Publically available RNA-seq dataset from the mouse ENCODE project was analysed to assess *Zfp263* expression in mouse tissues. RNA-seq was performed on tissues from E14.5 embryos and 8-week old male C57Bl/6 mice, as well as on mESC line Bruce 4 and mouse embryonic fibroblasts (MEF) derived from E13.5 C57Bl/6 embryos (Shen et al. 2012). In this dataset, *Zfp263* is expressed in all embryonic and adult tissues tested at a low/medium level, from 2 to 8 FPKM depending on the tissues (**Fig 2.5 TOP**). No tissues were found with no *Zfp263* expression at all. The bottom panel of **Figure 2.5** shows the relative expression of *Zfp263* compared to β -actin. To confirm these results, mRNA was extracted from E16.5 embryonic and adult mouse tissues to assess *Zfp263* expression level relative to β -actin. *Zfp263* is expressed in all embryonic tissues and placenta but at a much lower level than found using the RNA-seq data (**Fig 2.5B**). Similarly, in adult tissues, *Zfp263* is expressed at a very low level compared to β -actin.

Based on the RNA-seq analysis (Shen et al. 2012) and the qPCR experiment, it seems that *Zfp263* is expressed in a large range of mouse tissues at a low or medium level. In embryos, it is expressed the most in brain, heart and liver. In adult, it is expressed the most in kidney and likely in testes, although this has not been validated by qPCR. An absolute quantification of the transcript could be performed instead of the relative expression level compared to β -actin. This would give a better confirmation of the range of *Zfp263* expression. Quantitative analysis of the protein would provide more accurate insights into the extent to which the protein is translated in different tissues. However, the numerous anti-ZFP263 antibodies tested and optimised for immunoblot appeared highly unspecific to ZFP263 and multiple non-specific bands were observed in whole protein extract from tissues. Part of the on-going optimisation work is presented in **Appendix 8.2.2 – 8.2.4**.

A *Zfp263* absolute and relative expression (RNA-seq) in E14.5 embryos and adult mouse



B *Zfp263* relative expression in E16.5 embryos *Zfp263* relative expression in adult

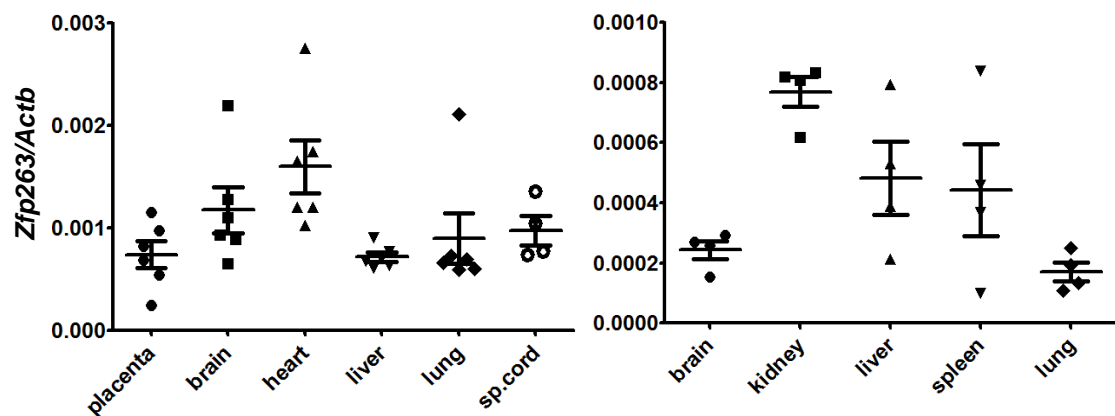
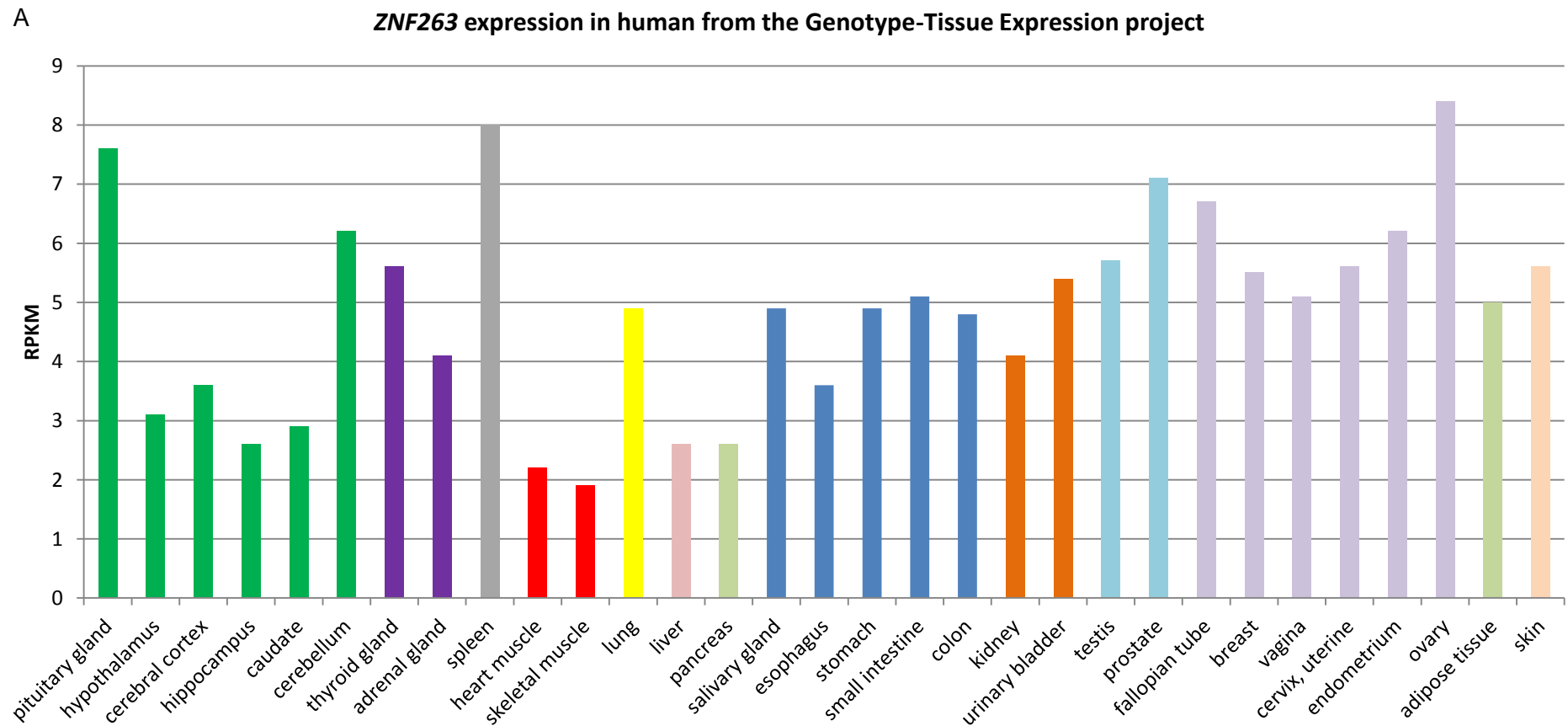


Figure 2.5: *Zfp263* expression in mouse embryonic and adult tissues. A. Average absolute expression (TOP panel) and relative to β -actin (BOTTOM panel) in E14.5 embryos (left panel) and adult tissues (right panel) from 2 replicates RNA-seq (Shen et al. 2012). B. *Zfp263* expression in different mouse tissues at E16.5 (left) and 3 months old (right) from qPCR data. RNA was extracted from tissues from 4 to 5 different individuals using TriReagent and synthesised into cDNA using oligo(dT). *Zfp263* expression was normalised to β -actin expression.

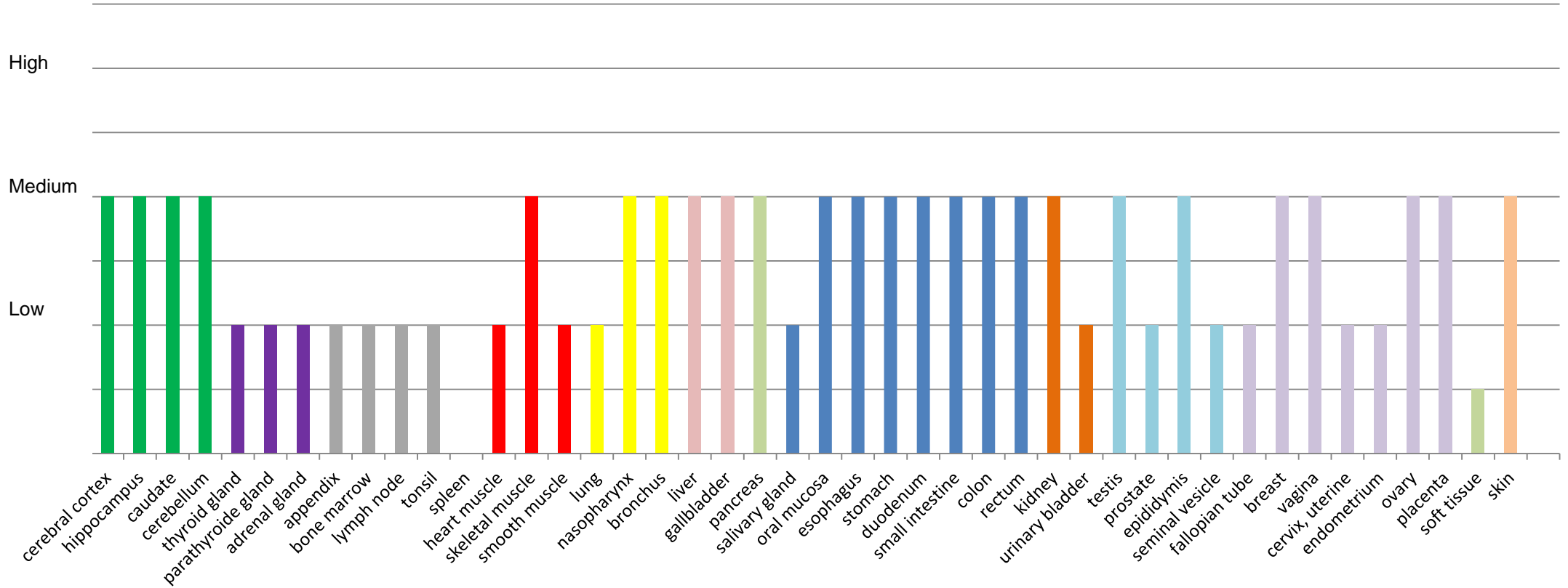
Human ZNF263 data are available on the Human Protein Atlas (www.proteinatlas.org). This program aims to map all the human protein in cells, tissues and organs by compiling various omics technology (Uhlen et al. 2010). The tissues Atlas shows the distribution of the proteins across major human tissues and organs based on RNA-seq and immunochemistry data. Histology-based profiles on multiple human cell types describe the spatial distribution, cell type specificity and relative abundance of proteins in these tissues. The immunohistochemical staining profile is matched for each gene and each tissue with mRNA data to yield the “protein expression” profile. In **Figure 2.6 A** is shown the expression data from one RNA-seq experiment in several human tissues. *ZNF263* is expressed in all tissues tested at a similar level than *Zfp263*, at a low to medium level between 2 to 8 RPKM depending on tissues. Interestingly, protein quantification data is available for human tissues and shows that the protein is being translated at a low to medium level in all tissues (**Fig 2.6 B**). Spleen is the only tissue where the protein is not detected, whereas the mRNA is transcribed.

Figure 2.6: *ZNF263* gene expression level (A) and protein level (B – next page) in human tissues from the Human Atlas project (Uhlen et al. 2010). The gene expression level is quantified in RPKM from RNA-seq data from the Genotype-Tissue Expression Project. The protein expression level is quantified based on immunochemistry data and mRNA level for each tissue.



B

ZNF263 protein expression from the Human Protein Atlas



2.3 Discussion and Conclusion

We have shown that *Zfp263* is a gene that first appeared in the platypus, suggesting that the gene first emerged 180 million years ago. This result is consistent with Imbeault *et al.* where they found that SCAN-containing KZFPs had deeper roots than the other KZFPs (Imbeault *et al.* 2017). Interestingly, they also found that these old and conserved SCAN-KZFPs were more prone to bind to promoters and unique genomic elements rather than the more canonical binding to retrotransposons.

The zinc fingers have been extremely well conserved during evolution, and the 4 amino acids predicted to be involved in DNA binding are found to be identical between species. It suggests that the protein can potentially recognize the same DNA sequence in all species, and so may undertake the same role. However other surrounding amino acids may also play a role in the binding, and all ZFs might not be involved at the same time, but rather function in different combinations of ZFs. The ZFs involved in binding might also change during development, between tissues and of course between species, so it cannot be confirmed from this analysis that ZFP263 carries the same function in all species but rather that it can theoretically bind to the same genomic locations. The Ka/Ks ratio shows that the protein is under purifying selection, protecting the coding sequence and acting against mutations. This means the protein was selected to be conserved across evolution to preserve its structure and function.

The SCAN domain is also conserved and very similar to the consensus sequence, suggesting that it was protected against mutations and that its function could be conserved across evolution. The three sequences that could potentially form 3 helices on the N-terminus of the domain are almost identical in all orthologues. These helices are likely to be key to determine the highly selective dimerization patterns of the protein. Therefore ZFP263 is likely to interact with similar molecular partners in all species.

The KRAB domain however is divergent from the consensus sequence with only about 40% similarity and is also less conserved between species. Witzgall *et al.* showed that mutations at three highly conserved regions of the KRAB domain were critical for recruiting KAP1. This first domain is only conserved in old species and diverges from the Afrotheria clade. The second domain is only conserved in primates. The third domain is not conserved in any of the species. This suggests that the protein might be severely compromised in its ability to recruit KAP1, and is therefore less likely to interact as a KAP1-dependant transcriptional

repressor. Interestingly, it seems that most of the SCAN-containing KZFPs do not have the ability to bind KAP1 (Imbeault et al. 2017).

Finally *Zfp263* transcripts were found in all tested tissues in mouse and human as well as in embryonic stem cells although the gene is not very highly expressed in any tissues. The human protein has been quantified in several tissues and detected at a medium level in most tissues. Quantifying tissue-specific protein levels in mouse tissues would provide further insights into whether these transcripts are being translated. These expression results and the human protein data suggest that ZFP263 does not act in a tissue-specific way but rather that it is present at a low and constant level at different developmental stages and throughout adult life.

To learn more about the role of ZFP263, it was decided to study the gene further in the mouse since this is a relatively tractable animal model. Mouse embryonic stem cells are a good *in vitro* model where *Zfp263* is highly expressed and because it is an easy to manipulate and well-described system. Based on this first set of results, we hypothesised that mouse ZFP263 binding sites and motifs should be similar to the ones identified in human, because of their identical zinc fingers. We also hypothesised that ZFP263 KRAB domain is unable to recruit KAP1, and that ZFP263 will not therefore be specifically and/or solely targeting retrotransposons. To test these two hypotheses, we proposed to identify ZFP263 targets in mESCs with a ChIP-seq experiment, of which the results are discussed in Chapter 3.

Chapter 3: Identification of ZFP263 targets in mESCs

3.1 Introduction

3.1.1 ChIP-seq assay

As discussed in Chapter 2, the next objective was to identify ZFP263 binding sites in order to better understand ZFP263 function. Several bioinformatics programs have been implemented to predict ZFP binding motifs *in silico* and thus potentially describe their genomic targets. However, because the mechanism of DNA recognition by zinc fingers is not entirely understood, it is difficult to make accurate predictions. As discussed in Chapter 1, it is expected that 3 or 4 amino acids are particularly involved in DNA binding, but the neighbouring amino acids or domains of the protein are also likely to affect DNA binding specificity although their level of influence remain unknown. Likewise, combinations of ZFs are expected to be involved in DNA recognition, rather than all ZFs at once, complicating again the use of bioinformatic tools to predict the motif bound by such proteins. Finally, even though a strong tool might identify an accurate binding sequence, it is unlikely that all the sequences in a genome matching the predicted motif are in fact occupied *in vivo* by the protein (Garton et al. 2015) or indeed, occupancy does not always predict function. Functional validation would be required and thus a more direct method is preferred.

A better method to identify transcription factor binding sites *in vivo* or *in vitro* on a genome-wide scale is the chromatin immunoprecipitation assay. The development of high-throughput sequencing technologies and the optimisation of NGS data analysis tools have made the ChIP-seq a powerful method to identify ZFP targets. The different steps of a ChIP-seq assay have been extensively documented and successfully used for histone modifications and transcription factors (Geen et al. 2010; Pepke et al. 2009; Landt et al. 2012; Furey 2012).

3.1.2 Experimental plan

A ChIP-seq experiment was performed in mESCs, where the gene is highly expressed, to test the hypotheses given in Chapter 2. In this Chapter, first the design of the experiment is presented, as well as the bioinformatics analysis pipeline optimisation that was completed in collaboration with Professor Li Qibin from the Beijing Genomics Institute in China, where I spent 6 weeks on secondment. Finally, the results of the ChIP-seq and the characterisation of the binding sites are presented and discussed.

3.2 Generation of FLAG-ZFP263 mouse embryonic stem cells

This work was initiated by a post-doc in the lab, Dr Noon, who designed the experimental plan, generated the first clones and the first ChIP-seq library as part of a higher-throughput screen. Dr Strogantsev designed the GFP-T2A-FLAG plasmid. Mouse reciprocal hybrid ESC lines had been generated in collaboration with Bowen Sun and Prof Roger Pedersen (Sun et al. 2012) and were maintained in the Ferguson-Smith lab. As presented in Chapter 1.3, I initially worked on 6 different KZFPs and I generated the second clones and the second replicate ChIP-seq library myself and analysed all the results. Only the results about ZFP263 are presented in this dissertation. Expression patterns of the other KZFP candidate genes in the selected clones used for the ChIP-seq experiment are shown in **Appendix 8.3.1, 8.3.2 and 8.3.3.**

Since very few antibodies exist for KRAB-ZFPs it was decided to work with FLAG tagged-proteins and express them in mESCs. The FLAG tag is an eight amino acid peptide (AspTyrLysAspAspAspLys) that was developed as a marker sequence for purification of recombinant proteins (Hopp et al. 1988). It is a highly hydrophilic sequence that expresses strong antigenicity and should adopt a highly exposed three-dimensional conformation on the surface of the protein, so it can readily interact with its ligand and is less likely to interfere with the protein structure and function (Einhauer & Jungbauer 2001). Although the FLAG tag is typically positioned at the 5' of the gene of interest to ensure good translational initiation, it was decided in this case to fuse a 3xFLAG at the C-terminus of the protein. A cleavable GFP reporter was added at the 5' end of the gene to allow the selection of infected clones. A T2A peptide was inserted to release the GFP and prevent it from interfering with the immunoprecipitation protocol. A lentivirus system was used to permanently integrate into the genome the gene of interest or a control sequence coding for the GFP reporter and the 3xFLAG tag without any other coding sequences (**Fig 3.1 A**). Reciprocal hybrid mESCs from a C57BL6/J x Mus Castaneus (BC/CB) cross were used to identify strain-specific binding motif and help us explore the properties of the binding site (**Fig 3.1 B**). Vectors coding for ZFP263 and control vectors were transfected to PLAT-E cells and reciprocal hybrid mESCs were infected by lentiviruses. GFP-positive single clones were picked and expanded. Two BC control lines (BC_ContA and BC_ContC) and one CB control line (CB_ContA) were selected based on their GFP expression. **Fig 3.1 C** shows BC_ContA and CB_ContA clones that are highly fluorescent. BC_263C and CB_263A however are much less bright. This could suggest that the control construct is more easily integrated leading to higher fluorescence intensity in the control lines.

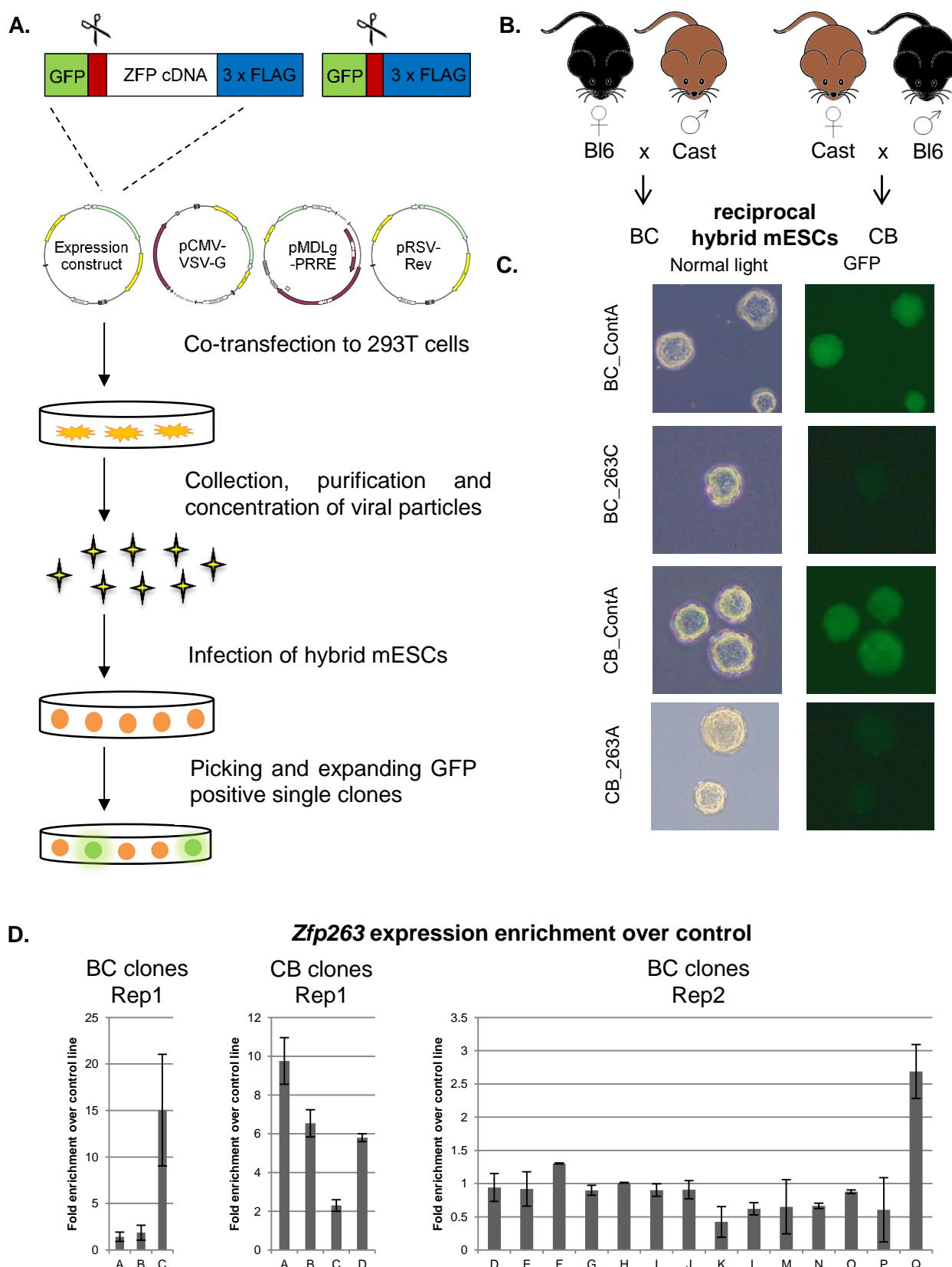


Figure 3.1: Generation of FLAG-ZFP263 mESCs. A. Diagram of the experimental protocol to infect mESCs with lentiviral particles. B. Diagram of the experimental design to generate hybrid mESCs. C. Pictures of 4 infected hybrid clones under normal light and GFP fluorescence. D. Fold enrichment of *Zfp263* in infected clones compared to control lines by qRT-PCR. Values were normalised with β -actin. $n = 3$, error bar = standard deviation.

Expression of *Zfp263* was assessed in clones infected with *Zfp263* cDNA and control cDNA, using primers targeting the 6th exon of the mRNA. In total, three BC clones and four CB clone were screened for the first replicate by Dr. Noon. Fourteen BC clones were screened for the second replicate (**Fig 3.1 D**). Overexpression of *Zfp263* compared to control lines meant that the exogenous gene was being stably expressed. BC_263C and BC_263Q expressed *Zfp263* about 15 and 2.6 times more than in the control line respectively and therefore were selected for the ChIP-seq experiment. CB_263A expressed the gene about 16 times more than in the control line and was used for the experimental target validation.

ChIP-seq experiments were performed on the four BC_263 and BC_Cont selected clones. Crosslinking steps, sonication time and immunoprecipitation protocol were optimised by Dr Noon (**See Methods**). The monoclonal antibody anti-FLAG M2 (Sigma-Aldrich F3165) was used for immunoprecipitation. Anti-FLAG M2 antibody had been raised for use in affinity purification of FLAG fusion proteins and is efficient for both N-terminal and C-terminal FLAG fusion proteins. Elution was performed by competition with synthetic peptides. ChIP-seq libraries were prepared following the Illumina protocol. The two replicates were sequenced at different times on a HiSeq2500 platform for the first one and a HiSeq4000 for the second set. The sequencing was 100bp paired-end.

3.3 ChIP-seq analysis pipeline

3.3.1 Beijing Genomics Institute Secondment

As part of the Marie Curie Training Network EpiHealthNet, I did a 6-week secondment in BGI in Shenzhen, China. The aim of the secondment was to analyse the two ChIP-seq replicates. I was supervised by Pr. Li Qibin and his team, in the BGI-Tech organisation. I followed an online training course to use Linux, and learnt the different steps to analyse NGS data. In Cambridge, the bioinformatics work was performed under the guidance of Dr Hui Shi, from the Ferguson-Smith's lab. My work in BGI provided an opportunity to carry out a comparison between different bioinformatic pipeline methodologies, (outlined in paragraph 3.2.2) and choose the optimal tools to perform the most robust analysis.

3.3.2 Pipeline optimisation

Raw reads generated from ChIP-sequencing have to undergo several processes before being analysed (**Fig 3.2**). The first step is to assess their quality directly after sequencing and

decide whether they are suitable for further analysis. The next step is to filter out low quality reads and then to align them to a reference genome. Finally reads are used for peak calling and further annotation and characterisation.

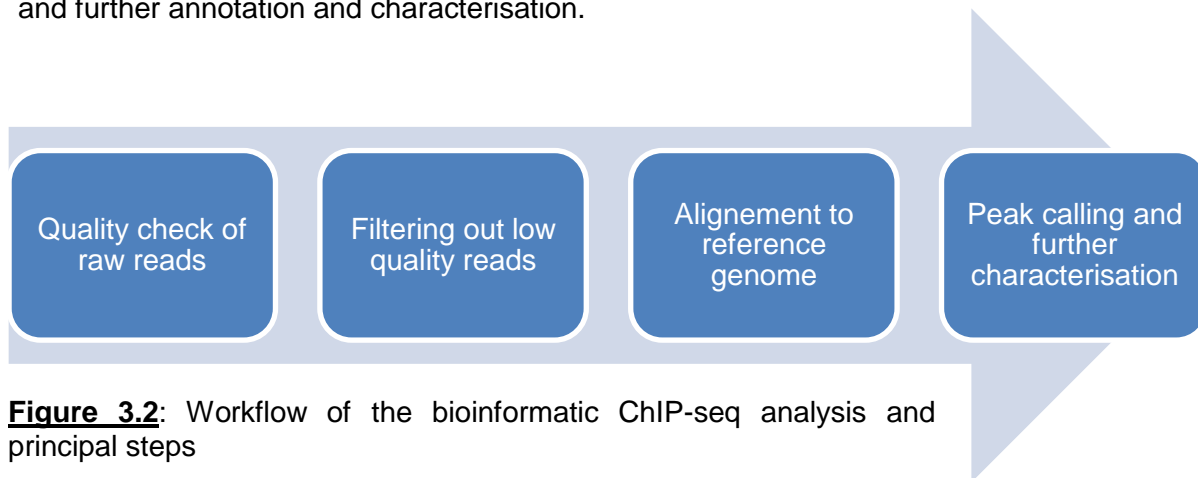


Figure 3.2: Workflow of the bioinformatic ChIP-seq analysis and principal steps

For quality check, the FastQC tool developed by Babraham Informatics was used (**Fig 3.3**). Because the samples were multiplexed in one lane of sequencing, they were ligated with adaptors to allow their tracking. Therefore the second step is to trim adaptors from the reads. At this stage low quality reads and low quality bases with a phred quality score lower than Q20, meaning that the base call accuracy was less than 99%, can also be discarded. In BGI, the approach was to discard the whole read if more than half of the adaptor was found in the read, if more than 50% bases had quality lower than Q20 or if more than 10% of bases in the read were undefined. This method led to removal of a lot of reads hence a more subtle method was adopted. Instead TrimGalore!, also developed by Babraham Informatics, was used. With this tool, adaptor sequences and bases with a quality score lower than Q20 were trimmed off, but the reads were retained. Only reads shorter than 20bp were discarded. Read quality was assessed again after trimming. All samples displayed good quality reads and could be used for further analyses.

In BGI, the alignment was done with SOAP2 (Short Oligonucleotide alignment Program). SOAP has been designed by BGI for the alignment of short oligonucleotides (R. Li et al. 2008). SOAP allows a certain number of mismatches, 2 in this case, or continuous gap for aligning a read onto the reference sequence. The best hit of each read with minimal number of mismatches or smaller gap will be reported (R. Li et al. 2008). This tool was initially designed for short reads, typically below 50bp. However several improved versions had been released to work on longer reads (Li et al. 2009; C.-M. Liu et al. 2012). The major issue with the BGI pipeline was that it has been designed to function on a reference genome only. However, since the ChIP-seq was performed in reciprocal hybrid mESCs, it was suboptimal

for our data. Therefore Dr Shi in the Ferguson-Smith lab generated an indexed hybrid genome for C57BL6/J-Mus Castaneus. The use of the hybrid genome is essential in this case because it allows strain specific alignment. Without the hybrid genome, because mismatches are allowed during the alignment step, a SNP between the two strains would not be taken into account, and strain-specific reads would be able to align to the other genome. Thus information about strain-specificity would be lost and indeed the accuracy of the alignment compromised. The alignment was eventually performed with the aligner BWA (Li & Durbin 2009). Two algorithms were tested – BWA-backtrack and BWA-MEM – and mapping efficiency was compared. BWA-MEM is the latest version and is recommended for high-quality queries, is faster and more accurate. The algorithm works by seeding alignments with maximal exact matches and then extends seeds. It performs local alignment and may produce multiple primary alignments for different part of the query sequence. BWA-backtrack works with the first bases on the 5' end. If a good match is found, it will try to align the rest of the read. It can be an issue if the first bases are wrong or less accurate, thus causing mistakes in the rest of the alignment. BWA-MEM should be better since it can work on any part of the read. However, BWA-MEM allows trimming of bases if the alignment is not perfect, and can trim a lot, resulting in shorter reads and wrong alignment. Here, the mapping rate was lower using BWA-MEM so BWA-backtrack was used with 1 mismatch allowed and using the first 20bp from the 5' end of the read.

After the alignment to the indexed hybrid genome, three files were generated (**Fig 3.3**). Unmapped reads, with a mapping quality below MAPQ20, were excluded from further analysis. C57BL6/J- and Cast-specific reads were split and sorted according to their coordinates. Reads that aligned to both genomes or to multiple locations were assigned a mapping quality of 0. These reads were extracted and mapped again to individual C57BL6/J and Mus Castaneus genomes separately. Similarly, they were sorted out according to their coordinates. Reads that aligned to multiple locations on a single genome were assigned a mapping quality of 0 and were separated from the rest of the analysis. These multi-mapped reads are likely to bind repetitive elements because the same sequence is found at multiple locations in the genome and thus cannot be assigned to one location only in the genome. Duplicates reads were removed from uniquely-aligned using Picard. Uniquely aligned reads to the same genome were finally merged together. This analysis generated five different types of reads – reads that align uniquely to C57BL6/J genome only, reads that align uniquely to Castaneus genome only, reads that align uniquely to both genomes, reads that align to multiple places in one genome, and reads that cannot be mapped – and two output files – C57BL6/J alignment and Castaneus alignment – for each sample of each replicate.

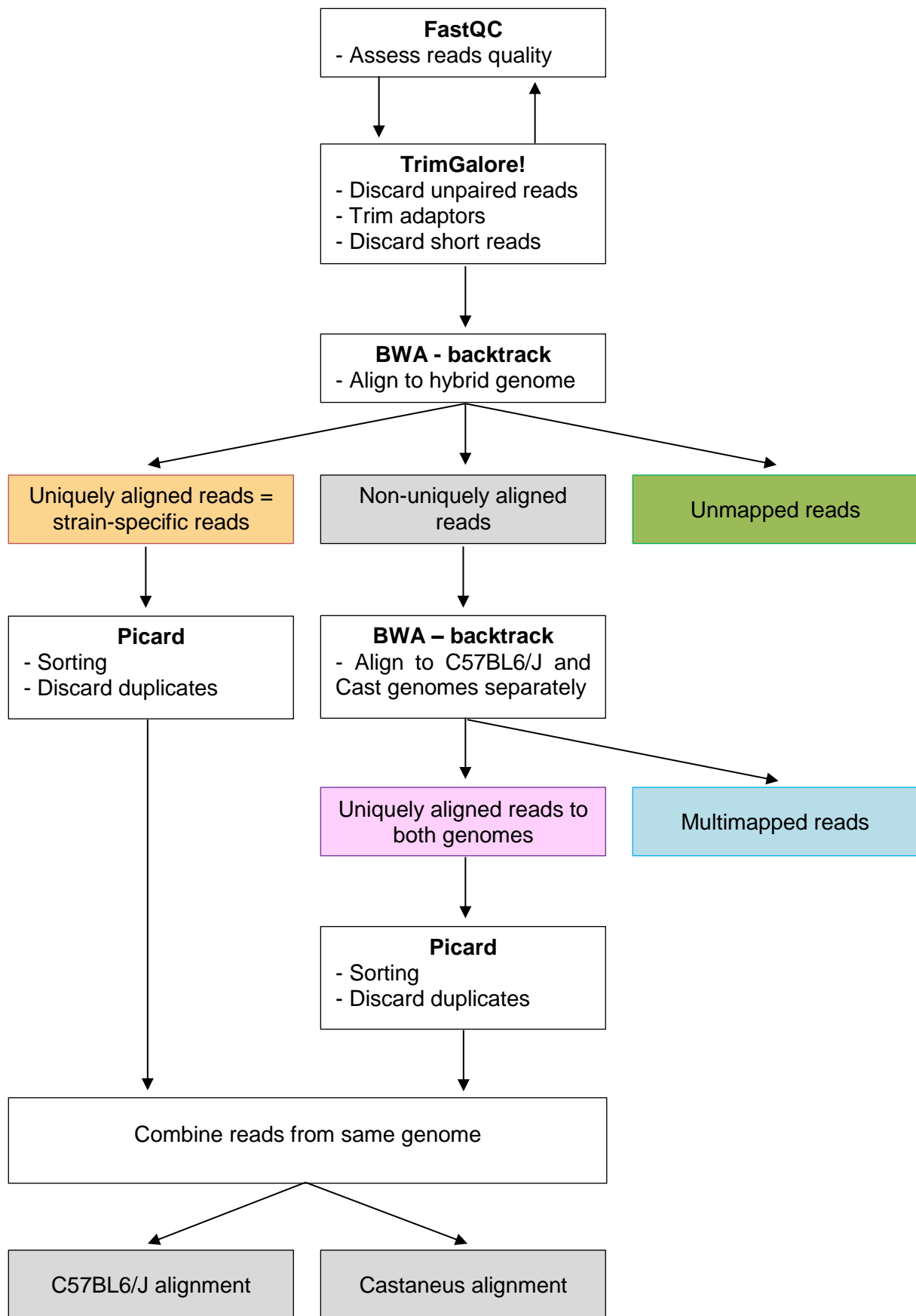


Figure 3.3: ChIP-seq analysis pipeline with the tools used and the different alignment files generated. The colours relate to **Table 3.1**.

3.3.3 Mapping Results

Table 3.1 presents the number of reads for each category and each sample. For mammalian transcription factors, 20 million reads per sample is adequate for accurate analysis (Bailey et al. 2013; Pepke et al. 2009; Landt et al. 2012). The first replicate gave slightly less than 20 million reads (**Table 3.1 col. A**). The second replicate has much more reads because of the sequencing platform improvement. In total, less than 0.4% and 0.04% of total reads were discarded for the first and second replicates respectively (**col. C**), which is a proof of good quality reads. The total of uniquely mapped reads (**col. G**) is the total of C57BL6/J-specific reads (**col. D**) plus Cast-specific reads (**col. E**) plus the number of uniquely-aligned reads to both genomes (**col F**).The total number of mapped reads (**col K**) includes the total of uniquely mapped reads (**col G**) and the multi-mapped reads (**col J**). In human and mouse, it is normal to have above 70% of uniquely mapped reads, whereas less than 50% may be a problem (Bailey et al. 2013). Other reviews suggest that at least 80% of reads should be mapped to distinct genomic locations (Furey 2012). Here, 3 samples show a mapping percentage above 80% (**col. L**), with maximum 7.4% of multi-mapped (**col. J**). The sample 263_Rep1 has a lower mapping percentage, of 62.6% and 67% of uniquely mapped and all mapped respectively (**col. H and L**). One should bear in mind however, that KRAB-ZFPs are known to target repetitive elements and therefore we retained an interest in the multi-mapped reads and, at this point, were interested in the finding that these were not generally highly represented.

Duplication is expected in ChIP-seq data and can arise during the library preparation process. If the amount of immunoprecipitated chromatin is low, it will result in a large amount of duplication during the PCR amplification step. A low amount of chromatin can be due to poorly efficient antibody or if the protein binds only to a few places in the genome. Our data show a relatively high duplication, in particular 263_Rep1 with 70% of duplication (**col. M**). This was a worrying result emphasising the need for experimental target validation of results.

Overall, the data generated are of good quality, except for 263_Rep1 that shows a higher number of unmapped reads and of duplication levels. Aligned reads were used to call peaks and identify ZFP263 targets.

Table 3.1: Number of sequencing reads for each category for the 4 samples. The colours relate to **Fig 3.3**. Orange: number of C57BL6/J- and Cast-specific reads. Purple: number of reads that are aligned at a unique location in both C57BL6/J and Cast genomes. The total of uniquely mapped reads is the total of BL6 specific plus Cast-specific plus uniquely mapped to both genomes. Blue: number of reads that align to multiple locations in one genome. The total number of mapped reads is the total of uniquely mapped plus multi-mapped. Green: number of reads that are mapped with a score lower than Q20 and so are excluded from the analysis.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	Raw reads (PE)	Reads post-TrimGalore (PE)	% discarded	BL6 specific	Cast specific	Uniquely mapped - both genome	Total uniquely mapped	% uniquely mapped	Multi mapped	% multi mapped	Total mapped reads	% Mapping rate	% Duplication rate	Unmapped reads	% unmapped reads
CONT Rep1	17490596	17424673	0.38	3927679	3622776	5492099	13042554	74.9	1186839	6.8	14229393	81.7	35.1%	3195280	18.3
263 Rep1	18403122	18345482	0.31	3194626	2917143	5370665	11482434	62.6	816261	4.4	12298695	67.0	71.9%	6046787	33.0
CONT Rep2	54035202	54012256	0.04	11922389	10987249	18019539	40929177	75.8	3978239	7.4	44907416	83.1	42.8%	9104840	16.9
263 Rep2	59099562	59071041	0.05	14080043	12154315	19616640	45850998	77.6	3987005	6.7	49838003	84.4	30.0%	9233038	15.6

3.3.4 Peak Calling

Seqmonk was used to visualize ChIP-seq data. Seqmonk is a tool developed by the Babraham Institute's informatics team to visualise and analyse high throughput mapped sequence data. Peaks were called using MACS (Model-based Analysis for ChIP-Seq) in Seqmonk. The control reads came from the control clones that were transfected with a GFP-3xFLAG vector and were subjected to the same experimental plan as the ZFP263 samples. Thus control reads represent background noise from the ChIP-seq experiment. Peaks were called for each genomic alignment of each sample and normalised using the control reads. They were quantified using the read count quantitation correcting per million reads and log transformed ($\log(\text{rpm})$). 1300 peaks were called in Rep1 in C57BL6/J and 1290 in Rep1 in Cast. 4000 and 4669 peaks were found in Rep2 in C57BL6/J and Cast respectively (**Fig 3.4 A**). 675 peaks were called in both genetic backgrounds in Rep1, and 2150 in Rep2. In C57BL6/J, 183 peaks overlapped between Rep1 and Rep3. In Castaneus, 195 peaks overlapped between Rep1 and Rep3 (**Fig 3.4 B**). In total, 103 peaks were called in all replicates and genetic backgrounds. 120 peaks were called in three samples: 60 were called in C57BL6/J in both replicates and in one or the other replicate in Castaneus, and 60 were called in both replicates in Castaneus but only in one or the other replicate in C57BL6/J. These peaks were included in the common genetic background peaks. Finally, 20 peaks were called in C57BL6/J in both replicates but not in Castaneus, and 32 in Castaneus in both replicates but not in C57BL6/J (**Fig 3.4 C**). These constituted genetic-background specific targets.

Examples of peaks are presented as screenshots from the UCSC Genome Browser in **Figure 3.4 D** and **Appendix 8.3.4**. The ChIP-seq signals are shown for both replicates and both genetic backgrounds for the control and ZFP263, as well as the UCSC genes, repetitive elements from RepeatMasker and locations of histone modifications. **Appendix 8.3.4** presents screenshots for 7 different peaks called in both genetic backgrounds in different genomic contexts. **Appendix 8.3.5** presents screenshots of C57BL/6-specific peaks, and **Appendix 8.3.6** of Castaneus-specific peaks. Peaks from the second replicate are consistently lower than from the first replicate. There remains some ambiguity in the allele-specific binding sites. For example, in **appendix 8.3.4 C**, there seems to be a signal in Rep1 BL6 although it is not called as a peak. Similarly for Castaneus-specific sites, **appendix 8.3.5 D** shows a signal in Rep1 BL6. These will require further experimental validation.

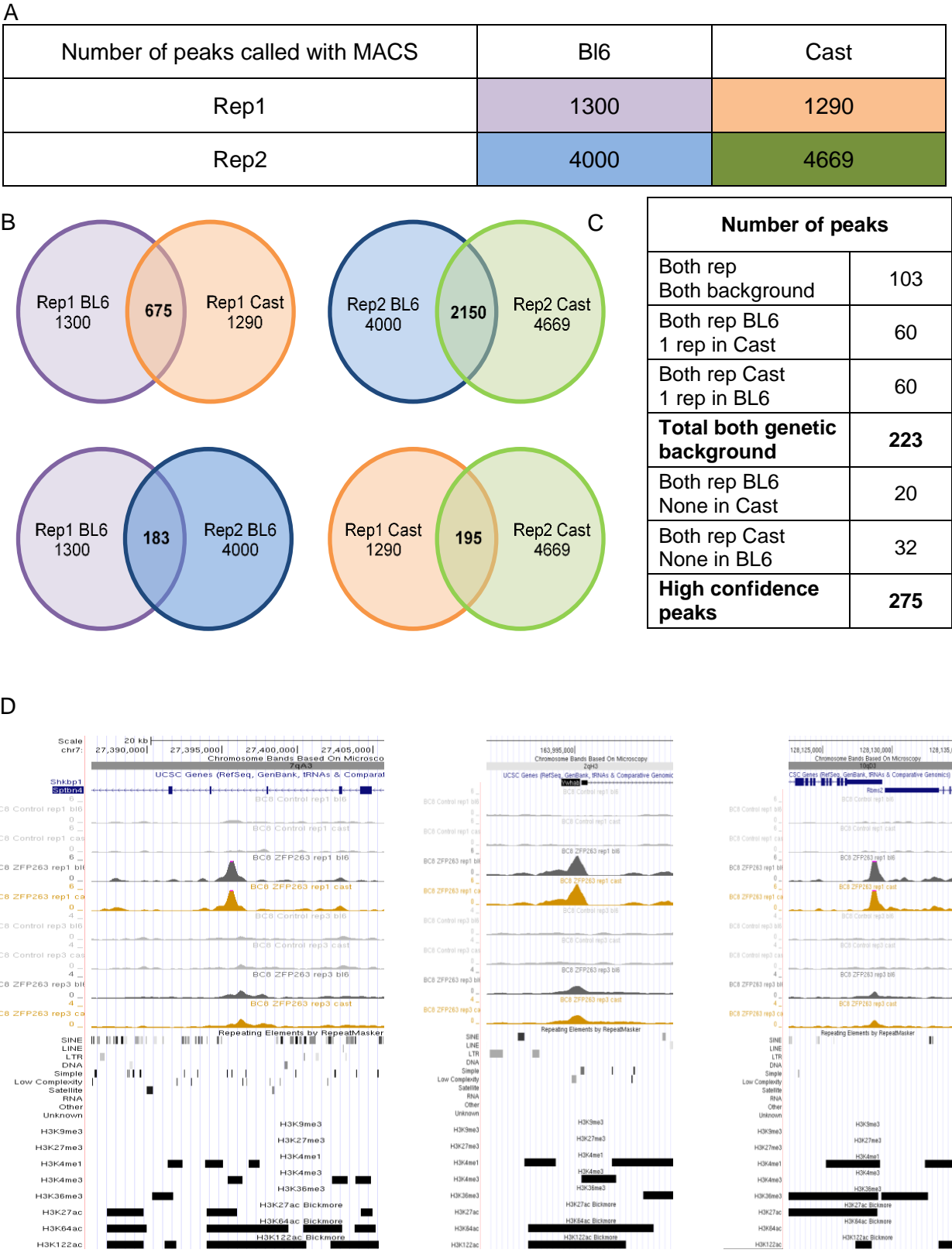


Figure 3.4: legend next page

Figure 3.4 (previous page): Number of peaks called using MACS peak caller in C57BL6/J x Mus Castaneus hybrid ES cells. A: Table with the number of peaks identified in each replicate for each genetic background. B: Venn diagrams showing the overlap of peaks between genetic backgrounds in the same replicate (top panel) or between two replicates in the same genetic background (bottom panel). C: Table presenting the overlap of peaks between replicates and genetics background with the total of highly confidence peaks. D: Screenshots of ChIP-seq signal from the UCSC Genome browser. The ChIP-seq signal is shown for both replicates (“rep1” and “rep3”), for the control and ZFP263 for each genetics background. The different tracks show the UCSC genes, the repetitive elements by RepeatMasker, 5 histone modification domains from mENCODE project, and 3 histone modifications domains from Pradeepa et al. 2015

The difference in number of peaks between the two replicates is quite striking. Rep2 had many more reads than Rep1, because of the sequencing depth. Rep1 also had more duplication events, suggesting that the amount of chromatin for library preparation was quite low. This might explain why we cannot call as many peaks as in Rep2. In total, 275 highly confidence peaks were called, *i.e.* peaks that are found in both replicates. 223 peaks are independent of genetic background, called in both replicates in one genetic background and in one or both replicates in the other background, while 20 and 32 are C57BL6/J- and Cast-specific respectively. Overall, there is a relatively low overlap between the 2 replicates suggesting that the parameters are too stringent and that, while keeping false positive low, relevant peaks may have been discarded. On the other hand, the high confidence peaks are less likely to be false-positives and should be biologically relevant. This was confirmed using experimental validation (see section 3.5).

3.4 Characterisation of ZFP263 binding sites

3.4.1 Identification of ZFP263 binding motif

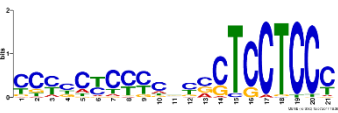
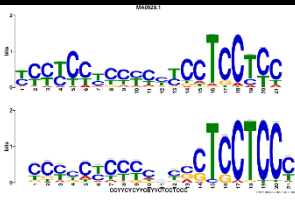
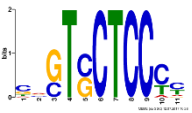
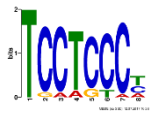
3.4.1.1 ZFP263 binding sites are enriched for the human ZNF263 DNA binding motif

To describe ZFP263 binding sites and gain more insight into their characteristics and properties, a consensus DNA sequence was first investigated using the MEME-ChIP portal. In Chapter 2, I showed that the human ZNF263 and the mouse ZFP263 were highly conserved, in particular the zinc fingers, of which the amino acids involved in DNA recognition are identical between the two orthologues. The binding motif depends on the combination of zinc fingers involved, post-translational modifications of the protein and other genomic contexts. Nevertheless, our hypothesis is that ZFP263 binding motif in mESCs

could be very similar to the consensus DNA sequence identified for the human ZNF263 in Fietze et al.

Coordinates of the high confidence peaks were uploaded to MEME-CHIP, and three motifs were identified as significantly enriched in the dataset, as presented in **Table 3.2**. The first one is a 21-nucleotide sequence highly similar to ZNF263 binding site identified by Fietze et al. The other two are shorter motifs, 11 bp and 8 bp, identified in 89 and 83 sites respectively. They are not being recognised as ZNF263 motif by TOMTOM, but it is interesting to note that they look like a truncated version of the longer motif. In total, 106 peaks had one of the three versions of the motif.

Table 3.2: DNA motifs found by MEME significantly enriched in the set of 275 high-confidence peaks, of which 102 sites had one or the other motif. TOMTOM identifies the first motif as similar to the human ZNF263 binding motif identified by Fietze et al. The two other motifs look like a truncated motif.

		Motif MEME	sites	e-value MEME	Alignment TOMTOM ZNF263
275 peaks	MEME1		96	$3.0e^{-139}$	
	MEME2		89	$1.8e^{-049}$	
	MEME3		83	$1.1e^{-009}$	

MEME-CHIP was run a second time using separately the coordinates from the 183 peaks common to Rep1 and Rep2 in a BL6 background, or the coordinates from the 195 peaks common to Rep1 and Rep2 in a Castaneus background. Similar results were found: the full length ZNF263 binding motif was significantly enriched in both datasets, as well as a shorter version of the sequence. No other DNA sequence was significantly enriched. By combining these results, 253 sites were identified to have a significant enrichment in the full length or the truncated version of the human ZNF263 binding motif, *i.e.* 92% of the high-confidence

peaks contained ZNF263 motif. Only 22 binding sites did not contain any motif at all. Finally, the entire set of binding sites in different backgrounds and different replicates was also submitted to MEME-CHIP separately: 1300 loci from Rep1 BL6, 4000 from Rep2 BL6, 1290 from Rep1 Cast and 4669 from Rep2 Cast. As expected, the full length and truncated version of ZNF263 binding motif were significantly enriched in each dataset.

Two other interesting results came out of the analysis. First of all, one other motif was significantly enriched in all dataset and was found in between 40 and 80 sites according to the dataset. This motif resembles a simple repeat element. This could suggest that ZFP263 does in fact target repeat elements, which would be consistent with the evolutionary understanding of ZFPs functions. However, this result is not recapitulated in any of the common peaks datasets, therefore it is more likely to be a non-specific motif. Likewise, another motif was significantly enriched on these datasets, although only at a very few sites (in 30 sites maximum) and was not identified in the other analysis. Finally, this analysis enabled me to identify a few loci that contained a ZNF263 binding motif but that were not part of the 275 high-confidence peaks. This means that these peaks had been called in only one replicate and were therefore excluded from the functional characterisation. However, this could also suggest that the peak calling parameters were too stringent and that biologically relevant peaks were excluded. It is therefore important to bear in mind that some true binding sites may have been excluded, although only a few peaks are affected in this context and thus this is unlikely to alter the global results and general interpretations.

3.4.1.2 Allele-specific binding sites

The use of hybrid cell lines enabled the analysis of genetic background-specific binding. As presented in 3.2.4, allele-specific ZFP263 binding revealed a subset of monoallelic peaks that were associated with one or the other allele: 20 binding sites are C57BL/6J specific and 32 are Mus Castaneus-specific. One possible explanation for this specificity is strain-specific genetic variation in the binding motif, such as indels or SNPs that could alter DNA recognition or create new binding sites for ZFP263 or other transcription factors. The binding motif was carefully analysed at these loci.

First of all, two BL6-specific and 7 Cast-specific sites did not contain the ZNF263 binding motif or any other consensus DNA sequence. This questioned the biological relevance of these particular peaks. SNPs within those that contained the binding motif were identified in 7 BL6-specific peaks and in 5 Cast-specific peaks. **Fig 3.5 A** shows the consensus binding motif with the locations of SNPs resulting in allele-specific binding. These disrupted sequences might directly alter the DNA recognition and specificity by ZFP263 on one of the alleles, or impair a protein-protein interaction between ZFP263 and one of its interactors resulting in the absence of binding of ZFP263 on one of the two chromosome homologues. Interestingly, 11 C57BL6/J-specific and 20 Castaneus-specific peaks did not present genetic variation between the two strains within the motif or in its close vicinity. Therefore, in theory, ZFP263 should be able to recognise these motifs independently of the genetic background. The allele-specific binding of these sites despite an intact DNA binding motif might represent a differential functional interaction with other proteins perhaps with a neighbouring polymorphic binding site. Indeed, SNPs further away from the motif could disrupt binding sites of the molecular partners of ZFP263 that might be necessary for the recruitment of ZFP263 to its targets. Likewise, genetic variations further from the motif could also create new binding motif for proteins that would prevent ZFP263 either by direct functional interactions or simply by reducing ZFP263 accessibility to the DNA. Understanding more about these types of potential interactions might provide insights into the biochemical properties and regulation by ZFP263. Interestingly, 6 binding sites common to the C57BL6/J and Cast background also presented a SNP (**Fig 3.5 B**) within the binding motif without impairing ZFP263 ability to bind to both alleles. In the binding motif, three locations (14, 18 and 21, orange stars) are targeted by substitutions that can either have no impact on the DNA binding (**Fig 3.5 B**), or that can result in allele-specific binding (**Fig 3.5 A**) providing insights into the relative importance of particular sites in the motif. I observed that the substitutions were not identical for the SNPs at position 14 and 21, suggesting a very precise mechanism of binding regulation. At location 18, the same nucleotide substitution results in two different outcomes. This supports the hypothesis of another mechanism that influence ZFP263 binding such as an interaction with a molecular partner.

In summary, 92% of the high-confidence peaks were enriched with the human ZNF263 DNA recognition motif. This confirms that the two orthologues are very similar and might regulate the same targets. It is important to remember that if the majority of the sites contain the full length motif, about 50 of them only display a truncated DNA motif and thus might impact ZFP263 specificity and its regulation. It is noticeable that the motif does not contain any CpG

and thus the protein is unlikely to be methylation-sensitive. 22 sites from the high-confidence peaks did not contain the consensus ZFP263 binding motif nor any other motif, suggesting that these peaks might be false-positives or that other factors might mediate their binding, such as interactions with other DNA-targeting proteins. The use of hybrid cells did not fully allow us to decipher the binding properties of the protein but strongly suggests an implication for other proteins influencing the ZFP263 binding to its targets. Further analysis and extensive validation is required to do justice to this powerful approach. Indeed, SNPs within or near the binding motif might explain the genetic-background specificity for some BL6- and Cast- specific peaks, but not for all of them. Transcriptome analysis in comparison with non-hybrid cells would provide a more confident interpretation of strain specific interactions.

Figure 3.5: ZFP263 consensus DNA binding motif. The stars represent the locations where SNPs were observed within the motif. A: The SNPs were present in genetic background-specific binding sites. B: The SNPs were present in binding sites common to both genetic backgrounds. The orange stars highlight three locations where SNPs at the same location within the motif did not result in the same outcome.

3.4.2 Association with repetitive elements

As presented in Chapter 1.2.4, KZFPs are understood to have evolved in parallel with retroelements to suppress their activity. In the dataset, only 4 to 7% of all reads were multi-mapped reads, meaning that they map to multiple locations in the genome, and thus are likely to represent repetitive elements. These reads could be mapped and randomly allocated to one genomic locus to be analysed together with the uniquely mapped reads. However, this method could potentially modify the results of the analysis and alter their true biological significance. This low percentage suggests that repeat elements are not highly represented in this dataset and that analysis of these reads alone would be problematic for normalisation

and overall would not provide much information on the dataset. Finally, because the sequencing was 100bp paired-end, it was possible to “rescue” most repetitive elements that would be targeted by the protein. If part of the read was multi-mapped but the other part uniquely mapped, it was then possible to map this read at a unique location. Therefore, it was decided to exclude the multi-mapped reads from the analysis, without impairing the possibility to analyse the binding of repeat elements.

The overlap between repetitive elements and the binding sites were identified within using RepeatMasker. Out of the 275 high-confidence binding sites, 201 sites, (73%) contained one or several repetitive elements, whereas 74 did not contain any (**Fig 3.6 A**). 275 random genomic loci of the same average length as ZFP263 binding sites were subjected to the same analysis. 77% also overlapped with one or more repeats (**Fig 3.6 A**). The large size of binding sites due to the nature of the ChIP-seq assay (average size is ~500bp) and the highly repetitive nature of the mouse genome easily explains this result. A more precise analysis of individual binding sites refined the result, and showed that the binding motif was rarely located within the repetitive element. The different screenshots in **appendix 8.3.4, 8.3.5 and 8.3.6** show that the centres of the peaks are depleted of transposable elements. Only 49 (17%) binding sites represented a direct overlap between the binding motif and a repetitive element, and all other repetitive elements were on the edge of the peaks. Repetitive elements at ZFP263 binding sites are SINEs, LTRs, simple repeats or regions of low DNA complexity (**Fig 3.6 B**). Randomly selected control genomic loci also show an enrichment in the same repetitive elements, with the exception of the “low DNA complexity” category that appear significantly enriched in ZFP263 binding sites compared to random genomic loci (**Fig 3.6 B**).

The low complexity DNA sites are primarily poly-purine/poly-pyrimidine stretches or regions of high AT or GC content, as described in RepeatMasker. Therefore they include promoter regions and CpG islands. This result suggests that ZFP263 does not target TEs. This is consistent with new recent studies that shed light on alternative roles for KZFPs. First of all, because TEs have acquired new functions within their host genomes, KZFPs might have also evolved to carry more diverse biological functions, alongside their TE-derived target loci. Second, the “arms race” model cannot explain the evolution of this protein family on its own, supporting the hypothesis that KZFPs are involved in a variety of other biological processes. Finally, until recently, very few KZFPs had been described to target unique genomic loci. ZFP57 was the first to be characterised as targeting imprinted control region in ESCs. However, very recent larger-scale studies suggest that about a third of the human KZFPs are

not targeting TEs but rather simple repeats or other unique genomic features; especially the older KZFPs that contain SCAN domains (Imbeault et al. 2017). ZFP263 is therefore likely to target unique genomic loci, which will be further characterised below.

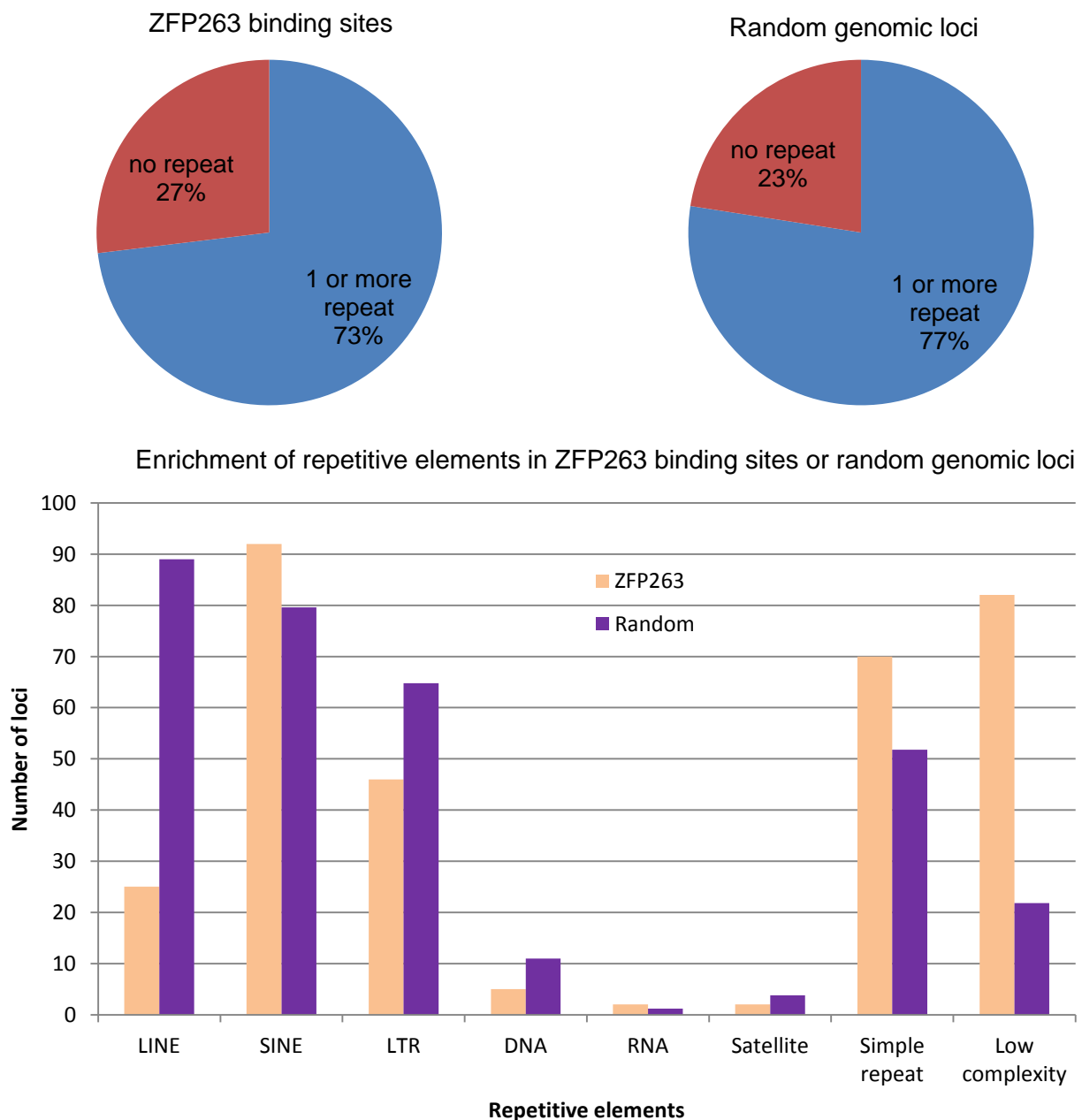


Figure 3.6: Enrichment of repetitive elements in ZFP263 binding sites compared to random genomic loci. A. Pie charts of the overlap between ZFP263 binding sites (left) or random genomic locations (right), of which the size averages 500bp, with repetitive elements from RepeatMasker. B. Detailed analysis of number of sites enriched in each repetitive element.

3.4.3 Targeting unique genomic loci

3.4.3.1 Genomic location relative to genes

To assess whether ZFP263 indeed binds to unique genomic locations, its binding loci were mapped relative to genes. 200 high confidence peaks overlap with a gene, while 75 are intergenic. Within the intergenic peaks, 29 and 25 are found 20kb downstream and upstream of a gene respectively, and 21 are not found within 20kb of any gene (**Fig 3.7**). Almost half of the intragenic peaks are overlapping the gene transcription start site, while the others are mostly within introns or overlapping exon-intron junctions (**Fig 3.7**). Not all peaks are consistent between replicates and hence require experimental validation (See section 3.5). Screenshots in **Figure 3.8** show 2 examples of peaks where the association with gene promoters is less clear. The first screenshot shows the signal quite clearly in the Tmem240 5'UTR, but screenshot B could potentially target the promoter of two different genes in different orientations. The binding site in screenshot 3 could be associated with either Naa35 promoter or Recql5 intron. Interestingly, around a quarter of the genes targeted by ZFP263 in mESCs were also identified as ZNF263 targets in human tissue culture cells (Fietze et al 2010).

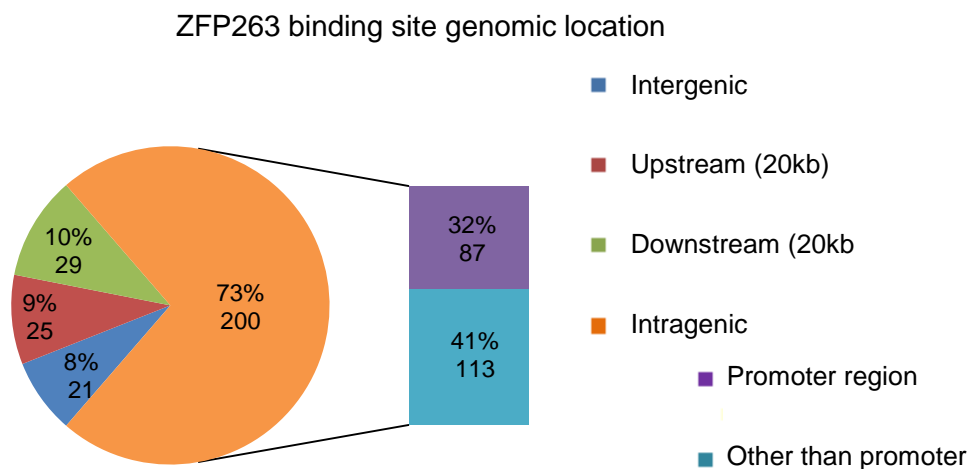
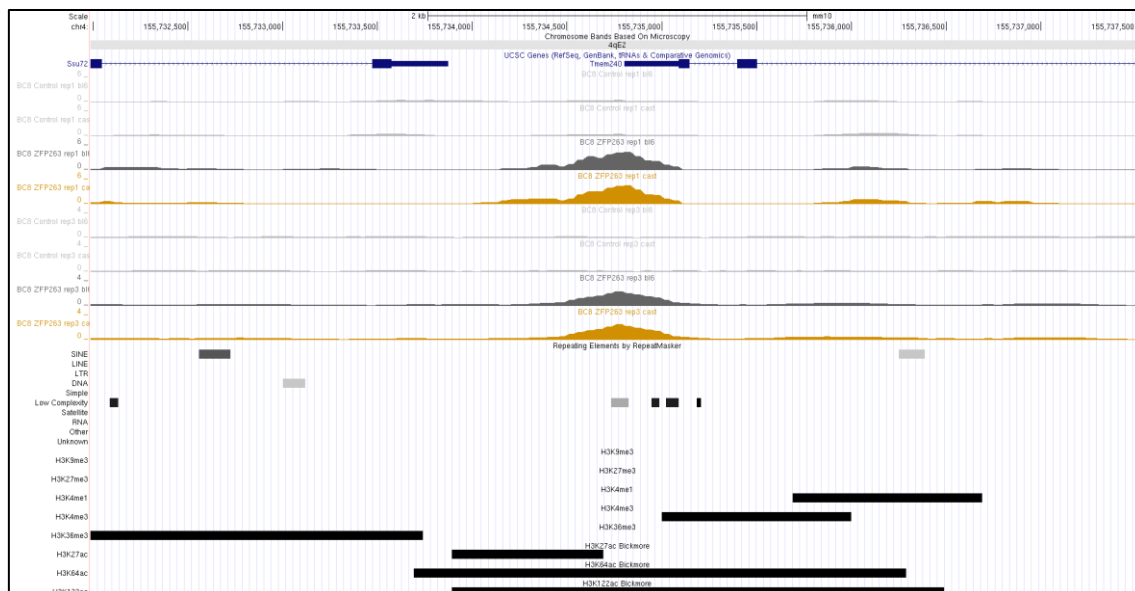


Figure 3.7: Pie chart of ZFP263 binding site location relative to a gene. 70% of the binding sites are intergenic (orange), with almost half of them overlapping with a promoter region.

These observations confirm that ZFP263 is involved in targeting unique regions of the genome. The association with promoter regions is likely to explain the additional enrichment at some low complexity repetitive DNA sequences. The overlap of mouse and human genes targeted by ZFP263/ZNF263 indicates that the two proteins might have conserved functions in the two species, as already suggested by the binding motif analysis, although they are also likely to have some species-specific functions. Finally, these results suggest that the protein could be involved in transcription regulation of its targets.

A



B



C

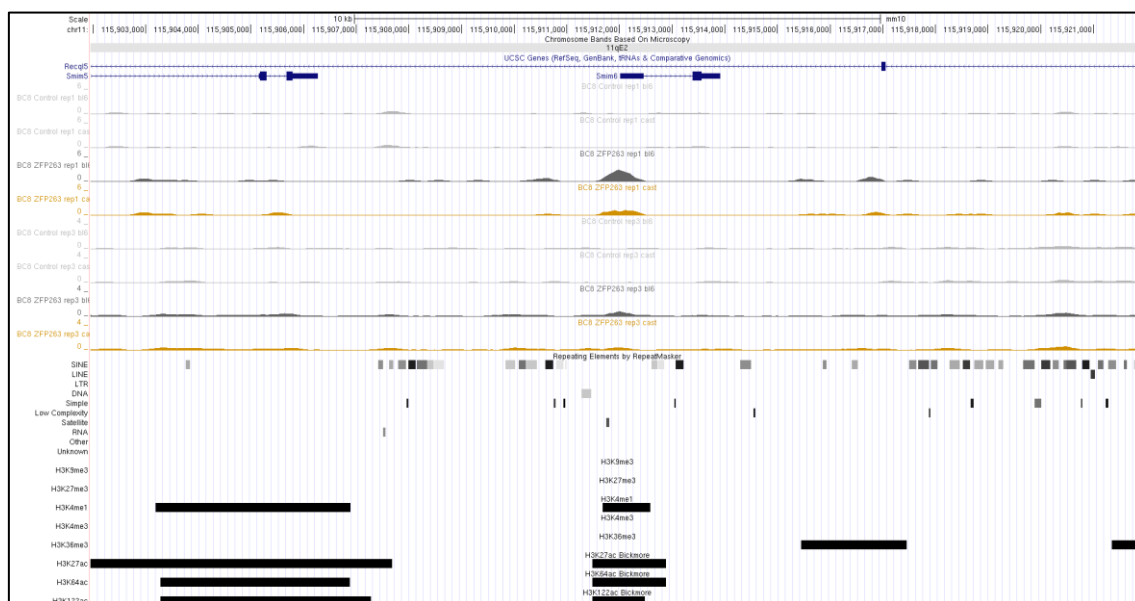


Figure 3.8: Screenshots of peaks associated with promoter regions.

3.4.3.2 Target Gene Ontology analysis

The target genes were subjected to a Gene Ontology analysis using Panther and DAVID, to assess whether ZFP263 was involved in the regulation of one or more particular biological pathways. Interestingly, the targets were only significantly enriched in one Biological Processes “negative regulation of transcription from RNA polymerase II promoter” (**Fig 3.9 red**). This is perhaps surprising given the association with activating epigenetic modifications (See section 3.4.4.2). They were not significantly enriched in any other Gene Ontology term. The targets are instead involved in several key biological processes and molecular functions, such as regulation of transcription, transport, regulation of cell cycle or development (**Fig 3.9**). This result emphasises that the protein may not have evolved to regulate one specific pathway but rather several key genes involved in different processes.

3.4.3.3 Target expression

Expression of the genes targeted by ZFP263 was analysed from publically available RNA-seq data in mESCs (Abad et al. 2013). Interestingly, the targeted genes display very different levels of expression, some being highly expressed and others completely repressed (**Fig 3.10**). This is unexpected considering the nature of KRAB ZFPs that appear to have evolved as repressors. However, ZFP263 seems to be a non-canonical KZFP since it does not preferentially bind repetitive elements but rather has unique targets. Furthermore, Fietze et al. suggested that human ZNF263 could act both as an activator and a repressor. Our finding is therefore consistent with previous studies, and reveals the unique character of this protein. It also suggests that other interacting proteins are likely to contribute to the function of ZFP263 and that understanding complexes including ZFP263 is likely to be a useful approach to understand function.

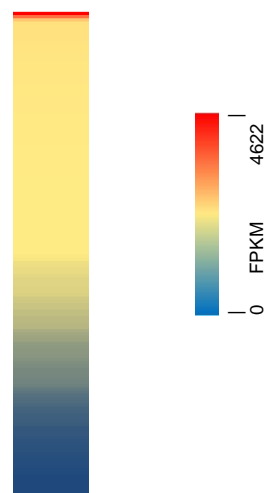
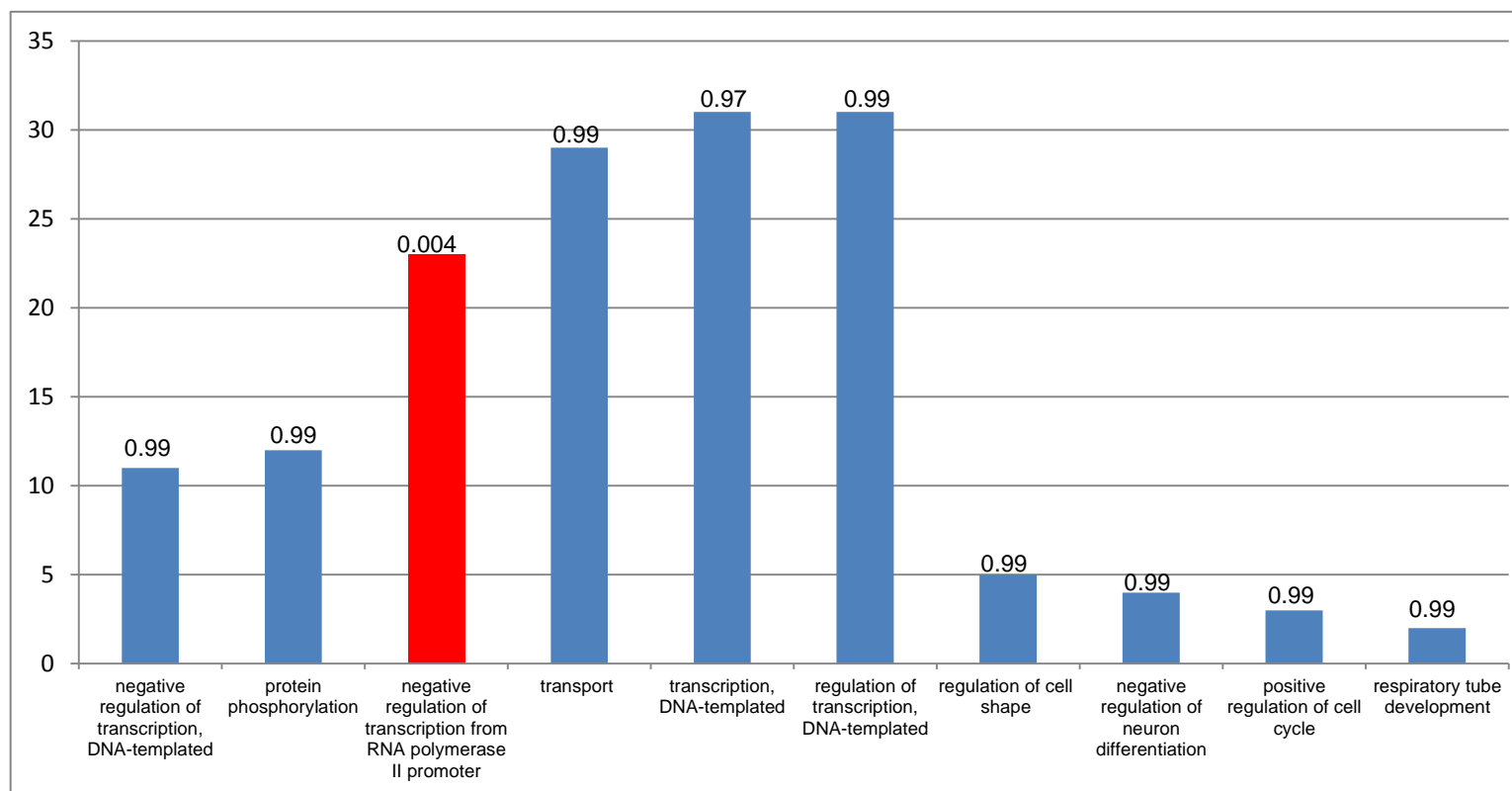


Figure 3.10: Heatmap of the target genes expression level in mESCs. They display a large range of expression level.

Figure 3.9: Number of target genes involved in Gene Ontology Terms for biological processes with corrected p-values from DAVID. There is no significant enrichment in any of the categories (p-values = 0.99) except for negative regulation of transcription (red, p-value<0.05)



3.4.4 Epigenetic state of ZFP263 binding loci

KZFPs are known to target epigenetic modifications at their genomic loci via the interaction with other co-factors. Most of them are described as transcriptional repressors via the recruitment of their canonical co-factor KAP1 and the assembling of heterochromatin-initiating complex.

3.4.4.1 Association with KAP1

An analysis to identify overlap between ZFP263 binding sites and previously published KAP1 binding sites in mESCs was performed (Rowe et al. 2010). Only four ZFP263 binding sites were associated with KAP1 binding. This result is not surprising considering the nature of its KRAB domain that diverges from the consensus sequence. Indeed, the amino acid sequence analysis of ZFP263 performed in Chapter 2 revealed that the KRAB domain of the ZFP263 had extensively diverged from the consensus KRAB sequence, and so probably before its first appearance in platypus. In particular, the mouse KRAB domain diverges from the consensus sequence in places that have been shown critical for the recruitment of KAP1. This is a surprising finding considering the current model of understanding of KZFPs actions, and it may highlight a new mode of action for part of the family. The most recent work from Imbeault et al. suggested that some of the older KZFPs, such as ZFP263, are not able to recruit KAP1. Although the authors do not show any data to support this hypothesis, their suggestion is consistent with my result.

3.4.4.2 Association with histone modifications

Histone marks in mESCs associated with ZFP263 binding sites were identified using dataset from the ENCODE database and publically available ChIP-seq datasets (Pradeepa et al. 2016). Strikingly, 240 sites were associated with one or more types of histone modifications and only 35 binding sites out of 275 were not associated with any in the ENCODE database. For comparison, on a set of random genomic loci, only 39 sites are associated with histone marks. This result shows that ZFP263 binding sites are significantly enriched in histone post-translational modifications, which is not surprising considering the general understanding of KZFPs as transcription factors. What is more surprising is the composition of the modifications. Indeed, none of the binding sites overlapped with H3K9me marks and only 8 were associated with H3K27me3. Rather, the target sites for ZFP263 were enriched for one or more active histone marks, such as H3K4me1/3, H3K36me3, H3K27ac, H3K64ac or H3K122ac (**Fig 3.11**).

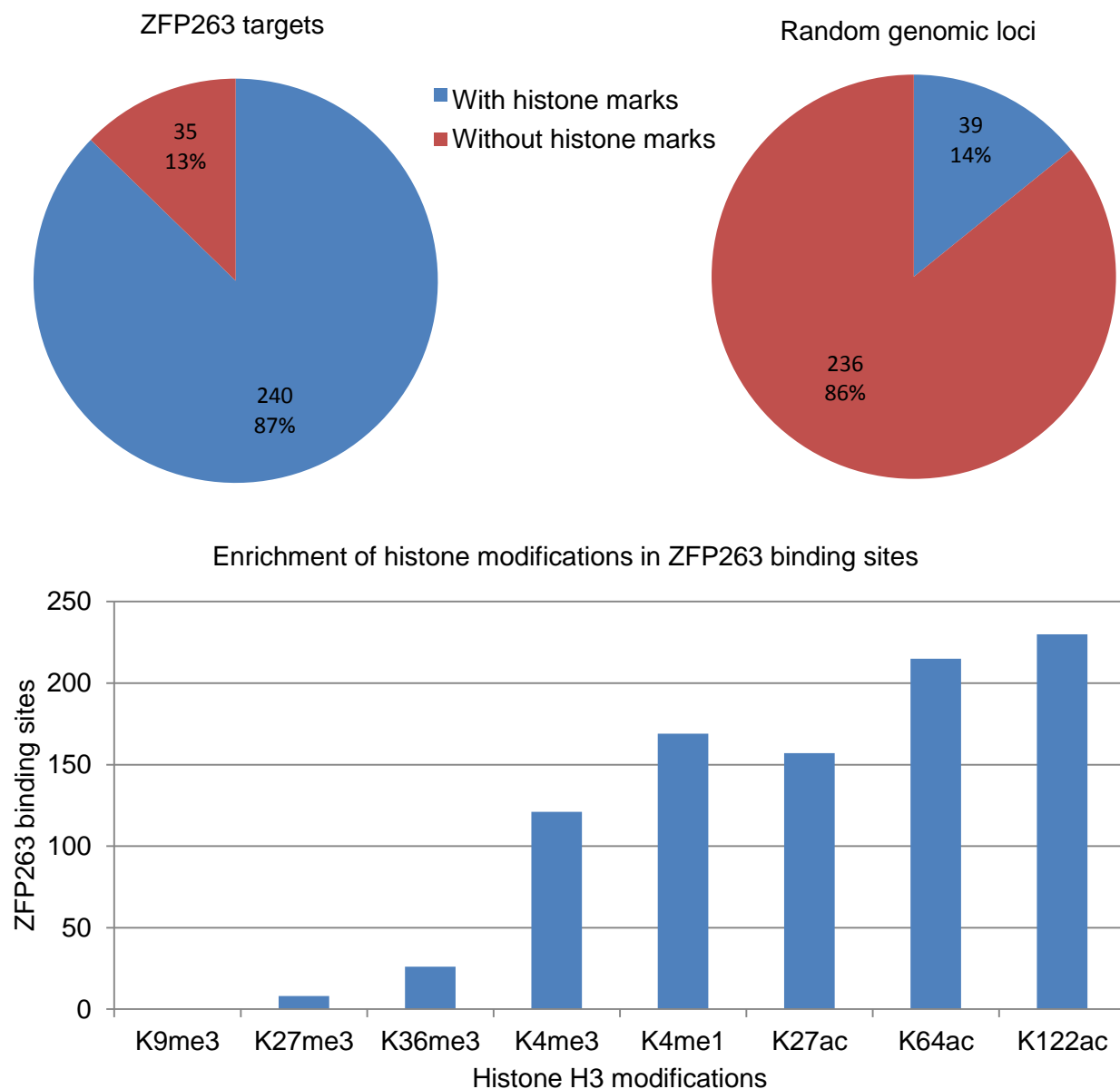


Figure 3.11: Pie charts of the histone contents in ZFP263 binding sites (left) or in random genomic locations of the same average size (right). ZFP263 target sites are significantly enriched in histone modifications, especially in marks associated with transcriptionally active chromatin

H3K4me3 is often found at promoters and can be observed at bivalent promoters in ECS in association with H3K27me3. Here, the absence of H3K27me3 suggests that bivalent promoters are not targeted by ZFP263. Instead, the binding sites are enriched for modifications often observed at enhancers, such as H3K4me1. Moreover, acetylation of the globular domain of histone H3 (H3K64ac and H3K122ac) was recently described as a new mark for active promoters and enhancers (Pradeepa et al. 2016). Interestingly the author of

this study identified a new subset of active enhancers marked by H3K122ac but lacking H3K27ac. This suggests that ZFP263 might target active enhancers.

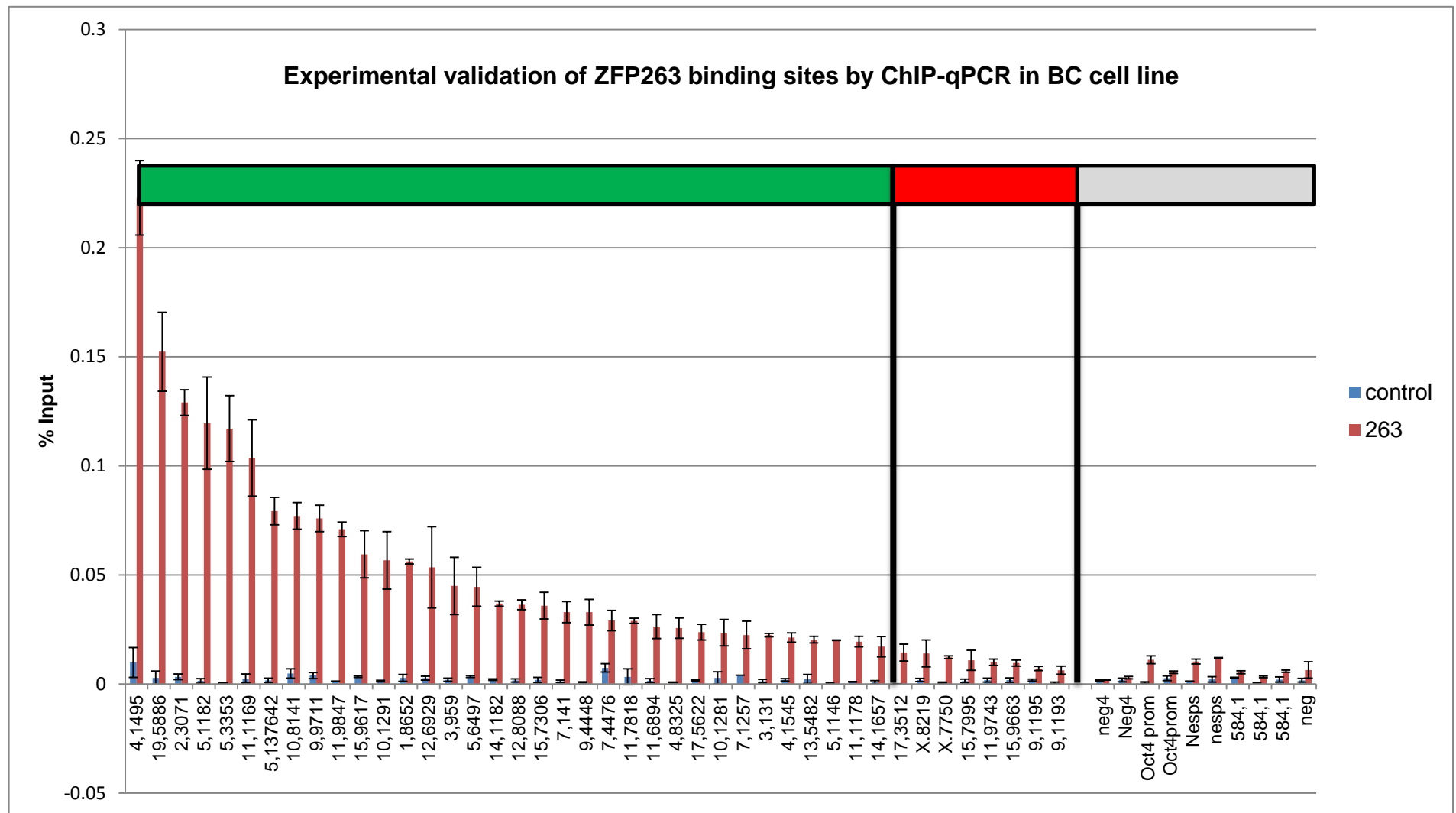
3.5 Experimental validation

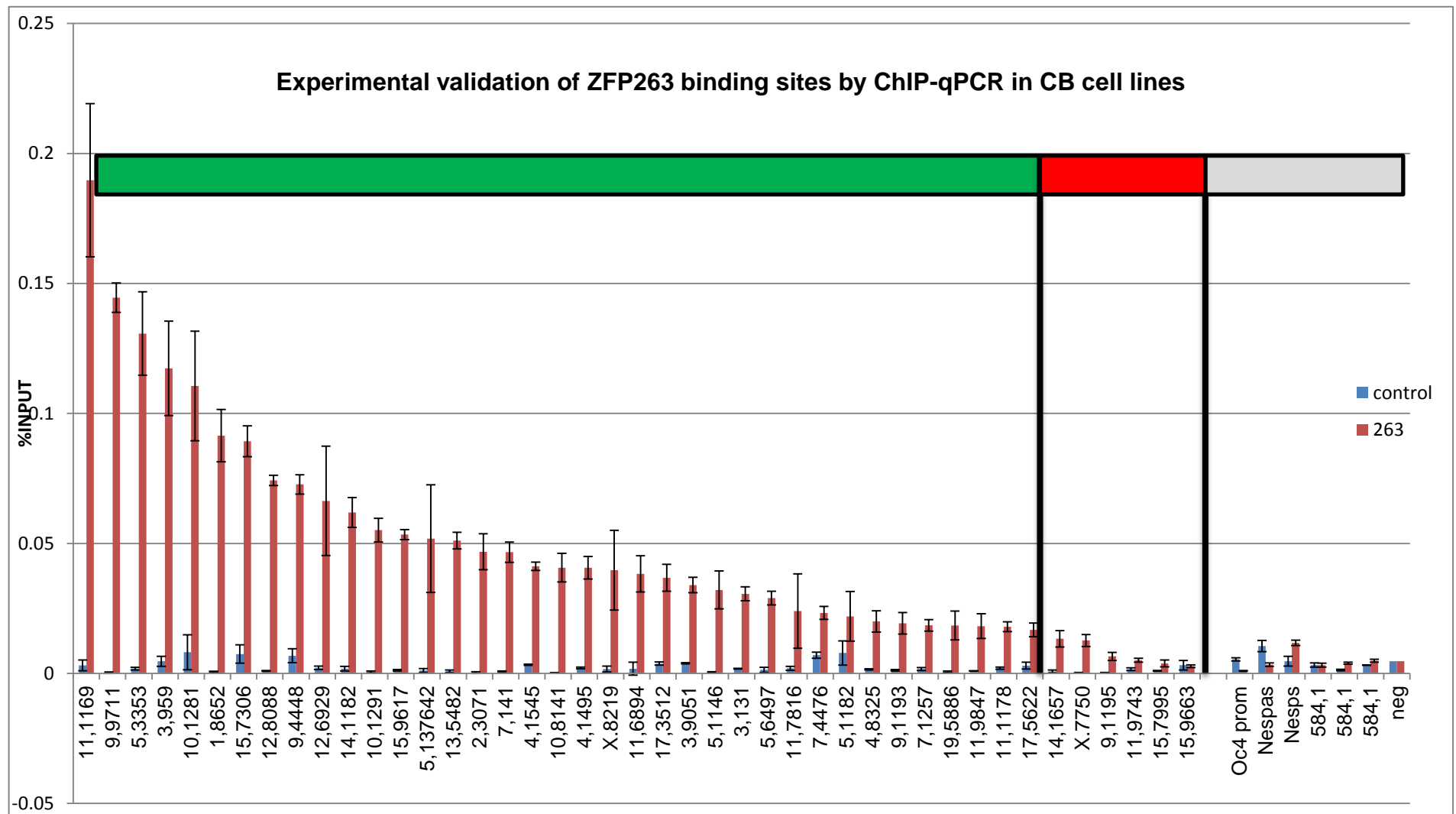
The binding sites were experimentally validated by ChIP-qPCR and using the %INPUT calculation (see Methods). 15% of the 275 high confidence peaks were selected and were tested in both hybrid cell lines BC and CB. Four different sites that were not targeted by ZFP263 in our dataset were used as negative controls. Regions that were at least 3 times above the mean of the negative controls were considered as validated.

In total, out of 42 tested binding sites, only 5 did not validate in both hybrid cell lines. These 5 sites were common to both genetic background and were called in both replicates. They all have a full length motif with no SNPs between the two mice strains. They overlap with genes and are associated with active histone modifications. 4 other sites validated in only one cell line but not in the other. Similarly, these binding sites are not specific for one genetic background or the other; they contain the full length motif without SNPs between the strains.

Importantly, 80% of the tested binding sites were validated as true binding sites in both reciprocal hybrid cell lines. This confirms that the analysis pipeline was adequate to call peaks and it strengthens the results observed.

Figure 3.12 (next 2 pages): Validation of 15% of the high confidence ZFP263 binding sites by ChIP-qPCR in BC and CB cell lines. Four regions not targeted by ZFP263 were tested multiple times as negative control (grey bar). Regions that present a %INPUT less than 3 times the mean of the negative control are not validated (red bar). 38 sites are validated (green bar) as true binding sites. Error bars: standard deviation.





3.6 Discussion and Conclusion

3.6.1 Limitations of the study

The second objective of this project was to identify ZFP263 binding sites in mESCs in order to better understand the evolution and function of the protein. A cell culture system rather than *ex vivo* tissues was used as well as a tagged exogenous protein. This represents the limitations of the study and should be borne in mind when interpreting the results.

First of all, cell culture is an artificial system that does not necessarily recapitulate *in vivo* functions and mechanisms. Furthermore, single clones were screened and selected, but it is well known that cells are heterogeneous within a population, and therefore a single clone might not truly recapitulate the results for an entire cell population. The overexpression of an exogenous protein might also affect cellular processes and again alter the significance of the results. Overexpression of the gene might result in non-specific binding to genomic loci and generate false positive results. The selected clones were overexpressing the gene 15 and 2.5 times more than the endogenous gene for the first and second replicate respectively. This difference in overexpression could also affect the results. As ZFP263 is a SCAN-containing protein, it can potentially homodimerize, and the increase of protein amount in the cells could impact on its ability to bind DNA or to recruit its other co-factors.

Second, the two replicates were performed at different time on two different sequencing platforms. The second sequencing replicate generated more reads than the first one, due to the use of a more recent and more powerful sequencing platform. Therefore, the increase number of reads for the second replicate is not biologically relevant but reflects a technical element. However, this discrepancy between replicates might affect the results. Indeed, many more peaks were called using MACS in the second replicate, but it is noticeable that the peaks look less convincing than the peaks called in the first replicate, possibly due to the lower expression of the tagged construct compared to the first replicate. The increase number of reads might dilute the true positive reads and therefore flatten the peaks. It would be interesting to analyse the two replicates with the same number of reads by taking randomly selected reads from the second replicate, and run the analysis again to compare both sets of results.

Finally, the bioinformatic analysis is a potential source of bias, as in any other analysis. However, the careful optimisation of the parameters and the extensive experimental

validation of the data strengthen the results of the experiment. Several key findings should be highlighted.

3.6.2. ZFP263 is a unique KZFP targeting unique genomic loci

First, ZFP263 recognises a DNA sequence very similar to the human ZNF263 binding motif (Fietze et al, 2010), suggesting that both orthologues might have very similar functions in their respective organisms. Therefore, the mouse is a good model organism to decipher ZFP263 functions. Interestingly, three types of motif were identified: the full-length 21-bp motif, or 2 shorter motifs that could be a truncated version of the longer motif. This might suggest that ZFP263 is able to target 2 types of DNA sequences, a long and a short version. The other possibility is that the longer motif actually represents two short motifs side by side potentially targeted by two proteins. The assessment of transcription of genes targeted with either the short or long version of the motif would provide interesting insight into the significance of the binding in both cases. Interestingly, no other motifs were significantly enriched in the set of 275 high-confidence peaks. This suggests that there are not any other common DNA-binding proteins targeting the same loci.

The use of hybrid cell lines enabled the identification of genetic background-specific binding. However, nine allele-specific sites did not contain the common binding motif, and therefore might be false positive peaks. The careful observation of the ChIP-seq signal in the UCSC Genome Browser also questioned the relevance of these binding sites, as some of them appear to also have a signal on the other allele but that was not called as a peak by MACS peak caller. SNPs within the motif could explain the strain-specificity in some cases by altering the DNA recognition of other proteins necessary for the binding of ZFP263 to its targets. Interestingly, the same SNP within the motif had two different outcomes: one site was strain-specific whereas the other one was common to both genetic backgrounds. This could suggest an issue during the peak calling step, or might be explained by the presence of another ZFP263 binding motif in close vicinity, that would enable the binding on both alleles. However in some instances, no variations were observed within the motif, suggesting that the specificity of these binding sites might be due to interaction with other proteins with a neighbouring polymorphic site. The analysis of strain-specific binding sites suggests that other proteins influence the specificity of ZFP263 binding sites, or that multiple ZFP263 could target the same loci, supporting the hypothesis that the long 21bp motif would be 2 smaller motifs side by side.

Strikingly, ZFP263 is not associated with transposable elements in mESCs, but targets unique regions of the genome, and in particular intragenic regions. ZFP263 binding sites are significantly enriched in low-complexity DNA, which is the result of ZFP263 binding preferentially at gene promoters. ZFP263 binding sites were also enriched within genes, within introns and at exon-intron junction. Very few KZFPs have been studied and even fewer has been shown to target unique genomic loci, although a recent study suggests that around one third of human KZFPs does not target transposable elements. Therefore, this is a new and exciting concept that is not well described or understood and merits further investigation if we are to truly understand the evolution and function of this class of proteins. Interestingly, about a quarter of the genes targeted by ZFP263 in mESCs were also identified as ZNF263 targets in human, which supports the hypothesis that the two orthologues exert similar function in both species as well as species-specific functions.

Another fascinating finding is that ZFP263 binding sites are not associated with KAP1. This is a very unusual result considering our understanding of KZFPs structure and mechanism of actions. However, it is consistent with the analysis performed in Chapter 2, whereby the KRAB domain of ZFP263 was found to be very divergent from the consensus KRAB sequence and lacking key residues for the recruitment of KAP1. The ChIP-seq results strengthen the hypothesis that ZFP263 is compromised in its ability to recruit KAP1. It is therefore less surprising that ZFP263 is not targeting any transposable elements, and that the binding sites are enriched in active histone marks at genes displaying different levels of expression. The nature of the histone modifications together with the location of the binding sites at promoters or within introns suggests that ZFP263 might regulate gene transcription by targeting their promoters or by recruiting histone modifications. Its binding sites may also be enhancers or other positive regulators of transcription. It is known that introns, where the protein preferentially binds, can act as enhancers. Thus ZFP263 might be a co-activator of its direct targets or of other genes since enhancers can be located up to 1 Mb away from the gene they regulate (Benabdallah et al. 2016). The binding location of ZFP263 at exon/intron junction also suggests that the protein could be involved in splicing, which would be consistent with the presence of H3K36me3 that might be involved in alternative splicing.

The gene ontology analysis showed that ZFP263 has evolved to regulate different biological processes rather than one particular pathway. However, as just mentioned, if ZFP263 acts as an enhancer, it could regulate genes other than its direct targets. If the latter is true it is not surprising that the GO analysis does not give any significant enrichment in biological pathways, since the genes regulated by ZFP263 would not necessarily be its sole direct

targets. Similarly, if ZFP263 regulates other genes than its direct targets, it could also explain the large range of expression of ZFP263 targets. The difference in target genes expression could also mean that ZFP263 is not directly involved in transcription regulation. ZFP263 could be indirectly involved in gene regulation by recruiting different co-factors depending on the chromatin or cellular context, or it could be involved in a different type of regulation that does not influence directly transcription.

Overall, we propose that ZFP263 is a unique KZFP, one of the first to be described as targeting unique genomic loci and that does not associate with the canonical co-factor KAP1. We hypothesise that ZFP263 acts as a co-activator for regulation of its targets and associated genes.

Chapter 4. ZFP263 function *in vivo*

4.1 Introduction

4.1.1 Objectives

I showed in Chapter 3 that ZFP263 is a unique KZFP that appears to positively regulate unique genomic loci through the association with active histone marks. The GO analysis did not show any significant enrichment in one particular biological pathway regulated by ZFP263. Therefore, to gain insight into the role of this atypical protein, I decided to extend my studies to an *in vivo* analysis. As shown in Chapter 2, the human and mouse ZFP263 are highly similar, and the result of the ChIP-seq experiment confirmed that the murine protein had similar binding target sequences as in the human. Hence, the mouse is a model of choice to study this protein function further. The objective was to generate mutant mice depleted of the functional protein and follow their development and phenotypic characterisation.

4.1.2 Experimental plan

Briefly, the generation of mutant mice was performed using the CRISPR-Cas9 technology in the zygote; the process is detailed in chapter 4.2. The mice were screened for mutations within the *Zfp263* gene and crossed to generate homozygous mutants, as detailed in chapter 4.3. The embryonic development of such crosses was followed, as well as their adult development; the results are presented in chapter 4.3.

4.2 Generation of KO mice

4.2.1 Targeting *Zfp263* gene

Two protein coding isoforms of the mouse *Zfp263* gene are described (**Fig 4.1 A**). One codes for the full length protein with the SCAN and KRAB domains and the nine zinc fingers while the other encodes the nine zinc fingers only, without any of the additional putative functional domains. The SCAN domain is encoded in the first exon (**Fig 1.4 B pink**), the KRAB domain in exon 4 (**Fig 1.4 B red**) and the nine zinc fingers in exon 6 (**Fig 1.4 B blue**). Two strategies were adopted to mutate the *Zfp263* gene. The different guide RNAs (gRNAs) were designed using the CRISPR design tool by the Feng Zhang laboratory (<http://crispr.mit.edu/>). First, to fully impair the binding capacity of the protein, it was decided

to target both isoforms. The gRNA was therefore designed to target exon 6 upstream of the zinc finger coding regions (**Fig 4.1 B orange arrows**). The objective was to induce a frameshift mutation to create a STOP codon upstream of the zinc fingers to generate a truncated protein unable to exert a function through the binding of its target. Two sets of gRNAs were designed to optimise the chances of obtaining frameshift mutations.

The second strategy targeted *Zfp263* in the first exon within the SCAN domain coding region (**Fig 4.1 B green arrows**). Similarly to the first strategy, the objective was to create a frameshift mutation to create a STOP codon very early on the transcript to generate a small peptide without any of the functional domain. A STOP codon in the first exon could also lead to the degradation of the transcript through the nonsense mediated decay. The gRNAs were chosen based on their scores and to limit off-target mutations.

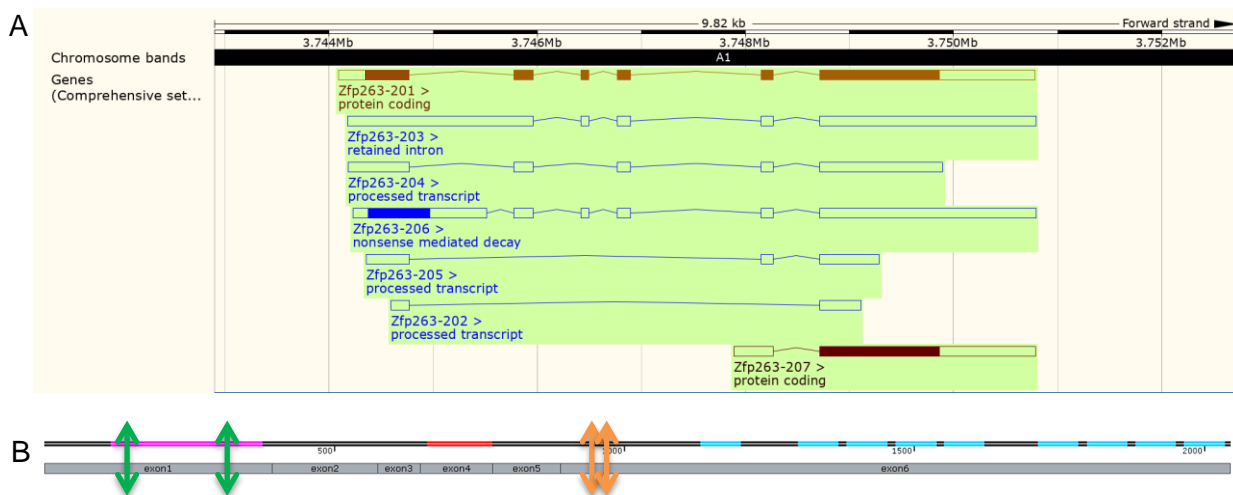


Figure 4.1: Design of gRNAs to target *Zfp263* gene. A: Screenshot from Ensembl displaying the mouse *Zfp263* isoforms. Isoforms 201 and 207 are protein coding. B: *Zfp263* gene encoding for the SCAN domain in exon 1 (pink), the KRAB domain in exon 4 (Red) and nine zinc fingers in exon 6 (blue). Two gRNAs were designed targeting exon 1 (green arrows) and exon 6 (orange arrows).

All gRNAs were chosen to limit possible off-target effects. The CRISPR design tool gives a list of potential genome-wide off-target sites with a score of likelihood targeting, the number of mismatches and their location within the off-target sequence. A mismatch close to the PAM sequence is likely to impair the Cas9 protein binding, and the more mismatches in the off-target sequence the less likely the targeting. The chosen gRNAs have a limited number of potential off-target sites with a very low score (<1.5), at least 3 mismatches within the off-target sequence and none were on chromosome 16, the same chromosome as *Zfp263*. Very few were located within exons, as shown in **Appendix 8.4.1**.

4.2.2 CRISPR-Cas9 injection in mouse zygote

The CRISPR-Cas9 technology originates from bacterial adaptive immunity against viruses and plasmids. The Cas9 protein is an endonuclease using a guide sequence within an RNA duplex to target DNA sequences and induce double-strand break in the DNA. The technology has been engineered so that a single guide RNA binds to Cas9 protein and leads it to the DNA sequence of interest (Doudna & Charpentier 2014). This technology has been extensively used in a wide array of cells and organisms, including in mouse zygotes to generate mutant mice (Yang et al. 2014). For this project the microinjections were performed in the Cambridge Stem Cell Facility in collaboration with William Mansfield. Hybrid mice (C57Bl/6 x CBA) were used to facilitate female superovulation and increase the number of potential mutants. After mating, hundreds of zygotes were collected and the gRNA together with the Cas9 mRNA or Cas9 protein were injected into one pronucleus. The embryos were then transferred to a surrogate mother and the pups genotyped after birth (**Figure 4.2**).

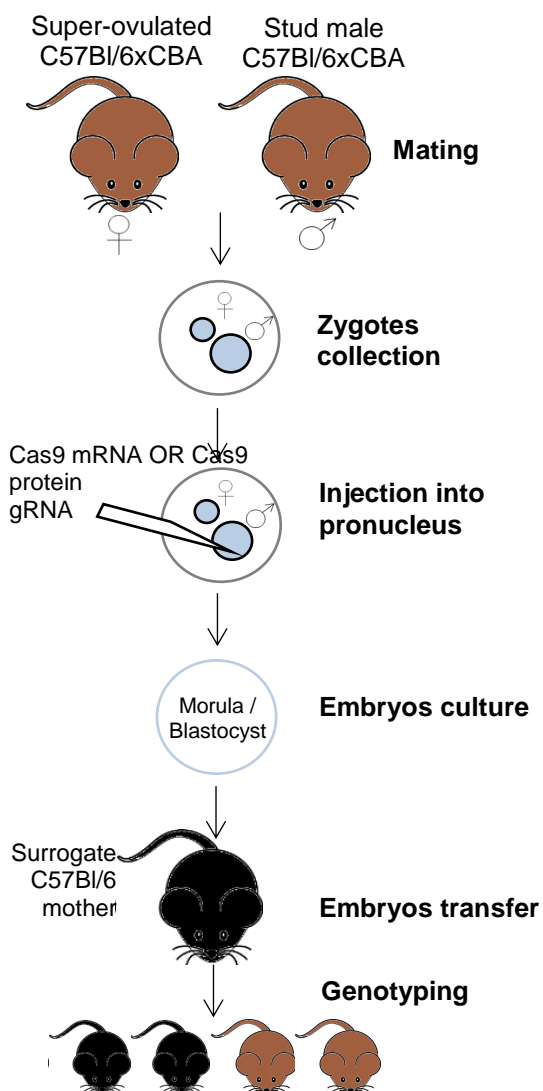


Figure 4.2: Schematic pipeline for the microinjections of gRNA/Cas9 in mouse zygotes. Zygotes were collected after mating and injected with the gRNA and the Cas9mRNA or Cas9 protein. Embryos were cultured and transferred to a surrogate mother. The generated mice were finally genotyped and further characterised.

4.3 Screening for KO mice

All mice resulting from the zygote injections and successful embryo transfer were genotyped after birth. Briefly, genomic DNA was extracted from ear notches and a ~500bp fragment was amplified around the site targeted by the gRNA. The amplified fragment DNA was run on an agarose gel, purified and either sequenced or cloned before sequencing. The different mutations for each experiment are summarised in the next paragraphs.

4.3.1 Exon 6 mutant mice

4.3.1.1 Genotyping

The first experiment targeted exon 6 of *Zfp263* gene. The first gRNA was mixed with Cas9 mRNA whereas the second gRNA was mixed with Cas9 protein. For the first set of injection with gRNA1, 80 fertilised eggs were collected from 6 super-ovulated females, 60 were injected with gRNA1/Cas9 mRNA, and 28 were transferred into two surrogate mothers who gave birth to nine pups in total. For the second set of injection, 78 fertilised eggs were injected with gRNA2/Cas9 protein and 50 were transferred to three surrogate mothers who gave birth to ten pups in total. **Table 4.1** presents the genotype status of the 19 founder mice targeted in exon 6 and the initial chromatograms for each founder are presented in **Appendix 8.4.2**.

From the first injection with gRNA 1 and Cas9 mRNA, only three mice were mutants while six were WT. From the second experiment with the second gRNA and the Cas9 protein, 10 out of 10 mice carried a mutation. This suggests that microinjection with the Cas9 protein increases the efficiency of mutagenesis. Most of the mice were heterozygous with one mutation, or compound heterozygous with two different mutations, one on each of the two chromosome homologues. Some mutations were found in more than one mouse. In total, this mutagenesis experiment generated a series of 16 different mouse mutant lines. **Table 4.2** presents these different mutations.

Before further analysis, it was verified whether the mutations could be transmitted to the next generation by crossing the founders with pure C57BL/6J mice. All mutations except two were successfully passed to the next generation (**Table 4.2 grey**). Furthermore, because Cas9 injection was performed in a mixed background optimised for embryo manipulation, the

mutants were being back-crossed with pure C57BL/6J anyway in order to generate KO mutants on a pure genetic background.

Table 4.1: Individual mice born from zygote injection targeting *Zfp263* exon 6

experiment	individual	sex	genotype	mutations
Exon 6 gRNA1 Cas9 mRNA 3 recipients 2 litters	1.1A	M	WT	
	1.1B	F	WT	
	1.2A	M	WT	
	1.2B	M	WT	
	1.2C	M	WT	
	1.2D	M	WT	
	1.2E	F	Heterozygous	5bp del
	1.2F	F	Compound heterozygous	5bp del / 1bp ins
	1.2G	F	Heterozygous	5bp del
Exon 6 gRNA2 Cas9 protein 3 recipients 3 litters	2.1A	M	Compound heterozygous	17bp del / 113bp del + 62bp ins
	2.1B	M	Compound heterozygous	1bp ins / 3bp del
	2.1C	F	Homozygous	3bp del
	2.1D	F	Compound heterozygous	1bp ins / 2bp del
	2.1E	F	Compound heterozygous	17bp del / 5bp del
	2.2A	M	Compound heterozygous	16bp del / 43bp del
	2.2B	M	Compound heterozygous	17bp del / 15bp del + 3bp ins
	2.2C	M	Compound heterozygous	7bp del / 1bp inversion
	2.3A	F	Compound heterozygous	12bp del / 5bp del
	2.3B	F	Compound heterozygous	16bp del / 1bp ins

Table 4.2: 16 types of mutations in *Zfp263* gene generated by microinjection of gRNA and Cas9 mRNA or protein into mouse zygote. Deletions in grey were not passed on the next generation. In red and highlighted in blue are the mutations used for embryos and placentas weight and pups weight respectively.

Frameshift mutations: STOP codon located upstream of zinc fingers	Deletions	2bp deletion
		5bp deletion - A
		5bp deletion - B
		5bp deletion - C
		5bp deletion - D
		7bp deletion
		16bp deletion
		17bp deletion
		43bp deletion
	Insertions	1 insertion - A
		1 insertion - B
Insertions and deletions	113bp deletion + 62bp insertion	
In frame mutations	Insertions and deletions	15bp deletion + 3bp insertion
	Deletions	3bp deletion
		12 bp deletion
Synonymous mutation	Substitution	Synonymous substitution

Most mutations are small indels, the two largest mutations being a 43bp deletion and a 113bp deletion with 62bp insertion. One mutation is a single nucleotide switch resulting in the same amino acid and this is therefore a silent mutation. Three mutations do not induce a frameshift; two of them do not induce a premature STOP codon (3bp and 12bp deletion). The third in-frame mutation (15bp deletion and 3bp insertion) however results in a STOP codon in exon 6. Twelve other mutations show insertions, deletion or both and result in a frameshift mutation upstream of the zinc finger coding region. The DNA sequence alignment is shown

in **Appendix 8.4.3** with the premature STOP codon. One mutation with 113bp deletion and 62bp insertion is likely to cause the retention of the last intron, as it removes the splicing acceptor site (**Fig. 4.3 green**). The retention of the last intron would give premature STOP codon (**Fig 4.3 red**).

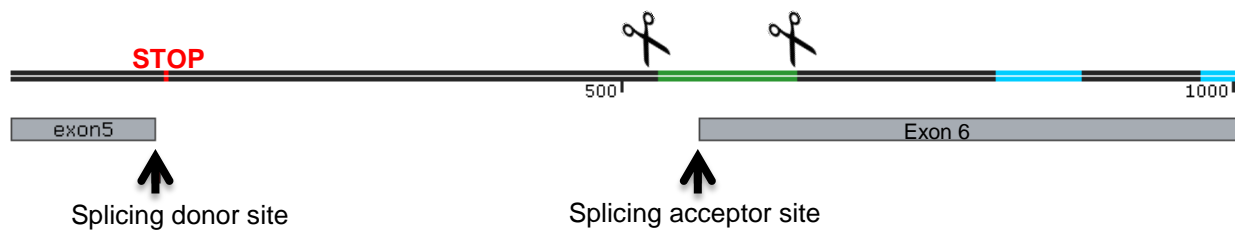


Figure 4.3: Genotyping of the 113bp deletion and 62bp insertion. The deletion (green and scissors) removes the splicing acceptor site and is likely to induce intron retention. The intron retention would lead to a premature STOP codon (red).

4.3.1.2 *Zfp263* transcription

The transcription level of *Zfp263* was also assessed in the founders. Indeed, three forms of co-translational mRNA surveillance mechanisms have been reported in order to clear mutant transcripts: the nonsense-mediated decay (NMD), the no-go decay (NGD) and the nonstop decay (NSD) processes. NMD specifically targets mRNAs containing a premature termination codon, NSD targets mRNAs lacking a termination codon and NGD targets mRNAs containing a range of potential stall-inducing sequences (Shoemaker & Green 2012). Premature termination codons are generally recognised by their proximity to exon-junction complexes deposited near exon junctions during pre-mRNA splicing, or by the lack of proximity with the poly(A) tail. Here, the STOP codon is located in the last exon, upstream of the zinc fingers. Therefore it is unlikely that the nonsense mediated decay is activated in this case, and the transcripts should not be degraded. Indeed, I observed that there is no difference between the WT and the mutants in the level of *Zfp263* transcription in tail samples, suggesting that the transcripts are being protected from degradation despite a premature termination codon (**Fig 4.4**). It is interesting to note that in KO4, with 16bp and 43bp deletions, the transcript is almost three times more abundant than in WT, suggesting that *Zfp263* was more expressed in this mutant.

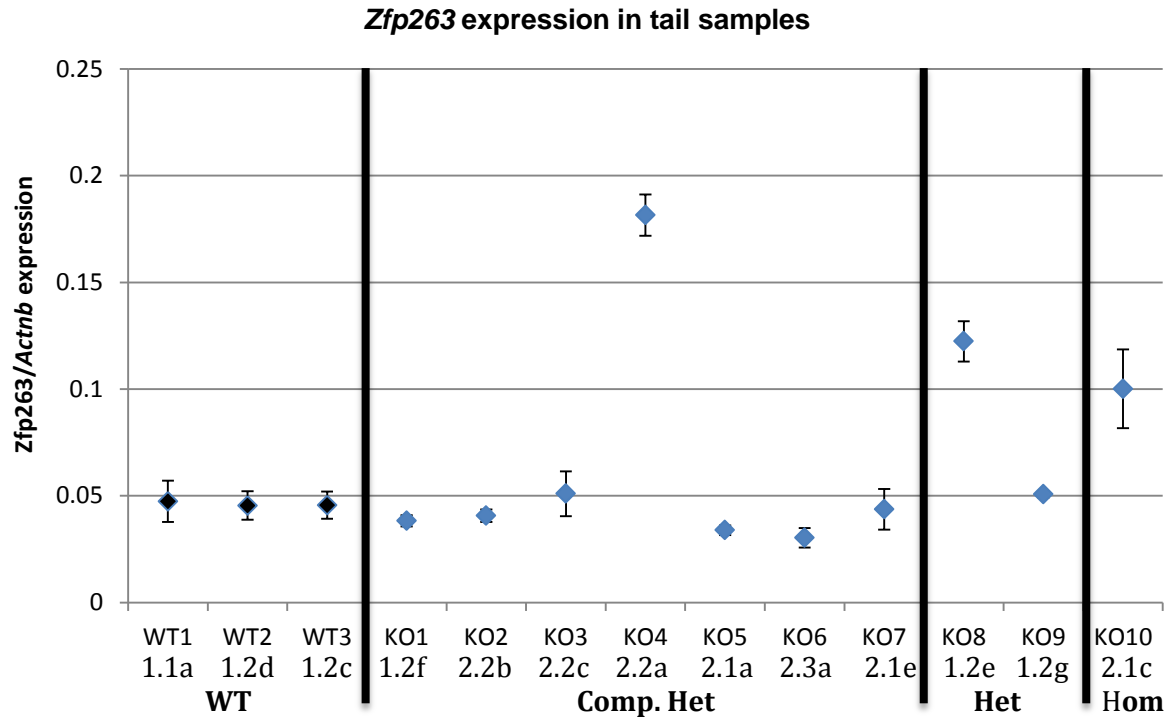


Figure 4.4: *Zfp263* expression level in tail samples normalised with β -actin in three wild-type animals and 10 compound heterozygous. Error bars: standard deviation from 3 technical replicates.

4.3.1.3 Prediction of protein translation

It is possible to predict the protein sequence based on the genotype. The single nucleotide switch mutation is a silent mutation and will translate the wild-type protein. The 3bp and 12bp deletion will translate the protein lacking one and four amino acids respectively, upstream of the zinc fingers coding region. As these amino acids are neither within a ZF nor within the KRAB or SCAN domain, these mutations are unlikely to affect the protein function. All of the other mutations induce a STOP codon upstream of the zinc fingers and could therefore translate a ~33kDa protein with a full SCAN and KRAB domains but without zinc fingers (**Fig 4.5**). Thus ZFP263 might lose its binding ability and these mice would lack a targetable protein. However, the 33kDa product could potentially have a function, perhaps with a dominant negative mode.



Figure 4.5: Predicted structure of the truncated protein in *Zfp263* exon6 mutant mice. The protein contains a full SCAN and KRAB domains but lacks its zinc fingers.

Quantitative analysis of the translated protein would provide information on whether the protein is being truncated or degraded in different tissues. I have optimised immunoblot protocol for several anti-ZFP263 antibodies on protein nuclear extracts (**Appendix 8.2.4**) but some optimisation work remains to be conducted to assess the level of protein translated in the mutant tissues.

4.3.2 Exon 1 mutant mice

4.3.2.1 Genotyping

The second KO strategy targeted exon 1 of *Zfp263* gene. Two different gRNAs were designed in exon 1 and both were injected into fertilised eggs with Cas9 protein. None of the eggs transferred with the first gRNA survived. This was most likely due to a technical problem during the micro-injection where they used a larger needle which may have damaged the eggs more than the previous injections. The second gRNA was injected into 100 zygotes, of which 55 were transferred into three recipients. 19 pups were born from three litters and genotyped.

Only two mice were wild-type, one was homozygous, one heterozygous and 15 were compound heterozygotes. In total, this experiment generated a series of 18 different mouse mutant lines as shown in **Table 4.3**. All mutations are located within the SCAN domain coding region. One in-frame mutation deletes 123bp in the second half of the SCAN domain. All of the other mutations result in a frameshift in the SCAN domain and lead to a premature STOP codon further on in the first exon. DNA alignments of all mutations compared to WT are shown in **Appendix 8.4.4**.

It was decided to keep only four lines (Table 4.3 orange). These mutations were all successfully transmitted to the next generation by crossing the founders with pure C57BL/6J mice. Similar to the first experiment targeting exon 6, the Cas9 injection was performed in a mixed background optimised for embryo manipulation. Therefore these four mutant lines are being back-crossed with pure C57BL/6J to generate KO mutants with a pure genetic background.

Table 4.3: 18 types of mutations in exon 1 of *Zfp263* gene generated by microinjection of gRNA and Cas9 protein into mouse zygote. Highlighted in orange are the mutations back-crossed to pure C57Bl/6 mice and that will be further analysed.

Frameshift mutations: STOP codon in the SCAN domain coding region	Deletions	20bp
		20bp
		19bp
		14bp
		14bp
		14bp
		14bp
		8bp
		11bp
		55bp
	Insertions	1bp
		1bp
		1bp
		113bp
	Deletions and insertions	24 del 4 ins
		22 del 8 ins
		29 del 4 ins
In Frame mutation	Deletion	123bp

4.3.2.2 Zfp263 transcription and protein translation

In this experiment, the premature STOP codon caused by frameshift mutations occurs in the first exon of *Zfp263* gene. Therefore, the nonsense-mediated decay, which is a mechanism that degrades aberrant transcript as described in 4.3.1.2, might be activated to clear the short transcript. Tissues have been collected to assess the presence of *Zfp263* transcript in different tissues.

If the transcript is not being degraded, a shorter protein will be translated with a partial SCAN domain but without the KRAB domain and without any of the zinc fingers. The amino acid alignment is shown in **Figure 4.6** with the SCAN domain in pink and the premature STOP codon in red. As explained in Chapter 1, the SCAN domain is predicted to fold into three helices that could confer the selective dimerization pattern (**Fig 4.6 BLUE**). Interestingly, the protein translated from the three frame-shift mutations (20bp and 55bp deletion, 113bp insertion) would retain the first half of the SCAN domain and thus the residues involved in the formation of helices.

The last mutation is an in-frame mutation with a 123bp deletion and does not result in a premature STOP codon (**Fig. 4.6**). Therefore the transcript should not be recognised as an aberrant transcript by the nonsense-mediated decay mechanism and should not be degraded. The translated protein from this transcript would lack 41 amino acids within the SCAN domain and the whole protein should be 4.5kDa smaller than the WT protein. It is interesting to note that the predicted protein would retain the residues folding into the first two helices of the core SCAN domain structure, but would lack the residues forming the third helix. The central helix is very well conserved between different SCAN-containing proteins, whereas the amino terminal helix reveals the highest diversity. Nam et al. 2004 suggested that the first helix might therefore contain key elements to determine the dimerization pattern (see Chapter 1.2.3.3).

20 bp deletion

WT	MTMAAGPSSQEPEGLLIVKLEEDCAWSHEVPPPEPEPSPEASHLRFRFRFQDAPGPREA
20del	MTMAAGPSSQEPEGLLIVKLEEDCAWSHEVPPPEPEPSPEASHLRFRFRFQDAPGPREA
WT	LSRLQELCRGWLRLPEMRTKEQILELLVLEQFLTILPQEIQSRVQELRPESGEEAVTLVER
20del	LSRLQELCRGWLRLPEMRTKEQILELLVLEQFLTILPQEIQSRVRRRSSHSCGAYAERTWE
WT	MQKELGKLRQQVTNQGRGAEVLLPEPLPLETAGESPSFKLEPMETERSPGPRLQELLDPS
20del	TEATGHKPRAGSRSAFGGAFATGNSRRVTELQAGANGDSTOP

55 bp deletion

WT	MTMAAGPSSQEPEGLLIVKLEEDCAWSHEVPPPEPEPSPEASHLRFRFRFQDAPGPREA
55del	MTMAAGPSSQEPEGLLIVKLEEDCAWSHEVPPPEPEPSPEASHLRFRFRFQDAPGPREA
WT	LSRLQELCRGWLRLPEMRTKEQILELLVLEQFLTILPQEIQSRVQELRPESGEEAVTLVER
55del	LSRLQELCRGWLRLPEMRTKEQILELLVLEQFLTKKQSLWSVCRKNLGNSTOP

113 bp insertion

WT	MTMAAGPSSQEPEGLLIVKLEEDCAWSHEVPPPEPEPSPEASHLRFRFRFQDAPGPREA
113ins	MTMAAGPSSQEPEGLLIVKLEEDCAWSHEVPPPEPEPSPEASHLRFRFRFQDAPGPREA
WT	LSRLQELCRGWLRLPEMRTKEQILELLVLEQFLTILPQEIQSRVQELRPESGEEAVTLVER
113ins	LSRLQELCRGWLRLPEMRTKEQILELLVLEQFLTILPQEIQSRVQELRPEKLIRLTIDCLF
WT	MQKELGKLRQQVTNQGRGAEVLLPEPLPLETAGESPSFKLEPMETERSPGPRLQELLDPS
113ins	SALVLELEIVSSTOP

123 bp deletion

WT	MTMAAGPSSQEPEGLLIVKLEEDCAWSHEVPPPEPEPSPEASHLRFRFRFQDAPGPREA
123del	MTMAAGPSSQEPEGLLIVKLEEDCAWSHEVPPPEPEPSPEASHLRFRFRFQDAPGPREA
WT	LSRLQELCRGWLRLPEMRTKEQILELLVLEQFLTILPQEIQSRVQELRPESGEEAVTLVER
123del	LSRLQELCR-----GEEAVTLVER
WT	MQKELGKLRQQVTNQGRGAEVLLPEPLPLETAGESPSFKLEPMETERSPGPRLQELLDPS
123del	MQKELGKLRQQVTNQGRGAEVLLPEPLPLETAGESPSFKLEPMETERSPGPRLQELLDPS

Figure 4.6: Protein alignment of four different mutant lines. All mutations occur within the SCAN domain (pink). The first three mutations result in a premature STOP codon (red) whereas the last one is an in-frame mutation resulting in the deletion of 41 amino acids.

4.4 Phenotypic characterisation of *Zfp263* KO mice

The phenotypic analysis was carried on mutant lines targeted in exon 6 only, as the exon 1 mutagenesis experiment was performed only a few months before the writing of this dissertation

4.4.1 Embryo and placenta development

After a first backcross with pure C57BL6/J mice, het x het crosses were set up as preliminary experiments to assess whether null mutants were viable at the embryonic stage. Crosses for 3 different frameshift mutations were set up: deletions of 17 bp, 5 bp and a mutation of 15bp deletion and 3bp insertion. These mutations all result in a premature termination codon upstream of the zinc fingers coding region, and potentially producing a shorter protein lacking its binding ability. Embryos were dissected at E16.5 and genotyped. All the embryos were viable and similar in their developmental stage. After genotyping, it appears that WT, heterozygous and homozygous embryos were viable at E16.5, and the mutants were not developmentally retarded. Table 4.4 summarises the genotyping.

Table 4.4: Genotyping of E16.5 embryos from het x het crosses for 3 different mutant lines

	WT	Het	HOM	
5bp del	1	5	2	+1 resorbed placenta
17bp del	2	4	2	
15bp del + 3bp ins	1	4	1	

The embryos and placentas were weighted individually. The wet weights for each mutant line are presented in Appendix 8.4.4. The pooled weights are presented in **Figure 4.7**, showing that at E16.5, there is no difference in weight between WT and heterozygous or homozygous embryos carrying one of the three mutations in exon 6 of *Zfp263*. Similarly there is no weight difference in placenta between WT and heterozygous, but there is a significant decrease in placenta weight between WT and homozygous. The numbers are very low therefore they should be considered preliminary, but they suggest that *Zfp263* might be involved in placenta development, and that the lack of the functional DNA-binding protein could affect placentogenesis.

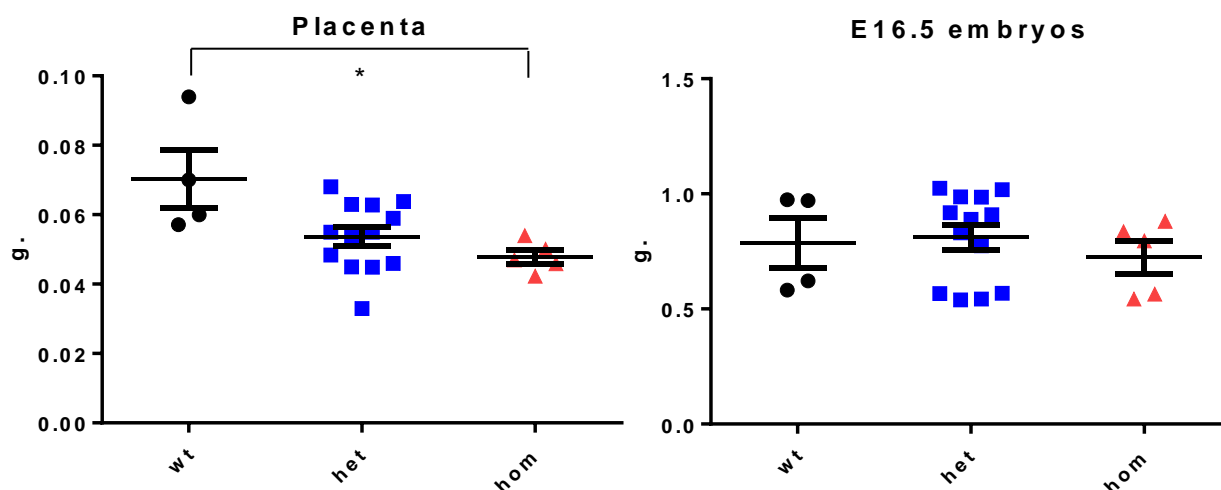


Figure 4.7: Weight of E16.5 embryos and placentae. Wild type, heterozygous and homozygous embryos were dissected from het x het crosses for 3 different frameshift mutations triggering a premature STOP codon: 2 types of deletions (17bp and 5 bp) and one deletion and insertion mutation (15bp del + 3bp insertion). There is a significance decrease between WT and Hom placentas (1 way ANOVA, p-value < 0.05)

4.4.2 Post-natal development

Intercrosses between heterozygous were set up for other mutations to assess whether null mutants were viable after birth. **Table 4.5** presents the genotype of individuals from het x het crosses around day 10 after birth. These individuals were generated from several litters and from different parents for each mutant line. It is striking to observe that the het x het crosses did not generate as many homozygous as expected according to the Mendelian ratios in particular for 113+62 and 15+3 mutation. All mutations should result in a STOP codon upstream of the zinc fingers, and overall for these 5 different mutant lines, 40 WT were born, 85 heterozygous, and only 30 homozygous. This suggests that although the homozygous are viable, they are subjected to a severe phenotype.

The development and weight of mice generated from different crosses were also monitored (**Fig. 4.8**). Weights from two het x het crosses with 1bp insertion mutation show that the heterozygous and homozygous pups tend to be smaller than their WT littermates. The homozygous show a greater variability in weight at day 5 compared to the WT. Similarly, homozygous with a 43bp deletion mutation are smaller than their heterozygous littermates, although they do not look developmentally retarded, suggesting a growth defect. In one litter with 16bp deletion mutation, the heterozygote was much smaller than the WT, did not catch up after weaning and died at day 31. The numbers are very low and these results are only preliminary, but there seems to be a trend for homozygous to be smaller than WT and heterozygous.

Table 4.5: Number of WT, heterozygous and homozygous from 5 intercrosses 10 days after birth. The number of litters is indicated for each mutation

Type of mutation	WT	heterozygous	Homozygous
1 bp insertion Day 10 – 5 litters	10	20	8
113 bp del + 62 bp ins Day 10 – 5 litters	16	23	7
15 bp del + 3 bp ins Day10 – 3 litters	5	18	6
43 bp deletion Day 10 – 5 litters	6	20	9
16bp deletion Day 10 – 1 litter	3	4	0

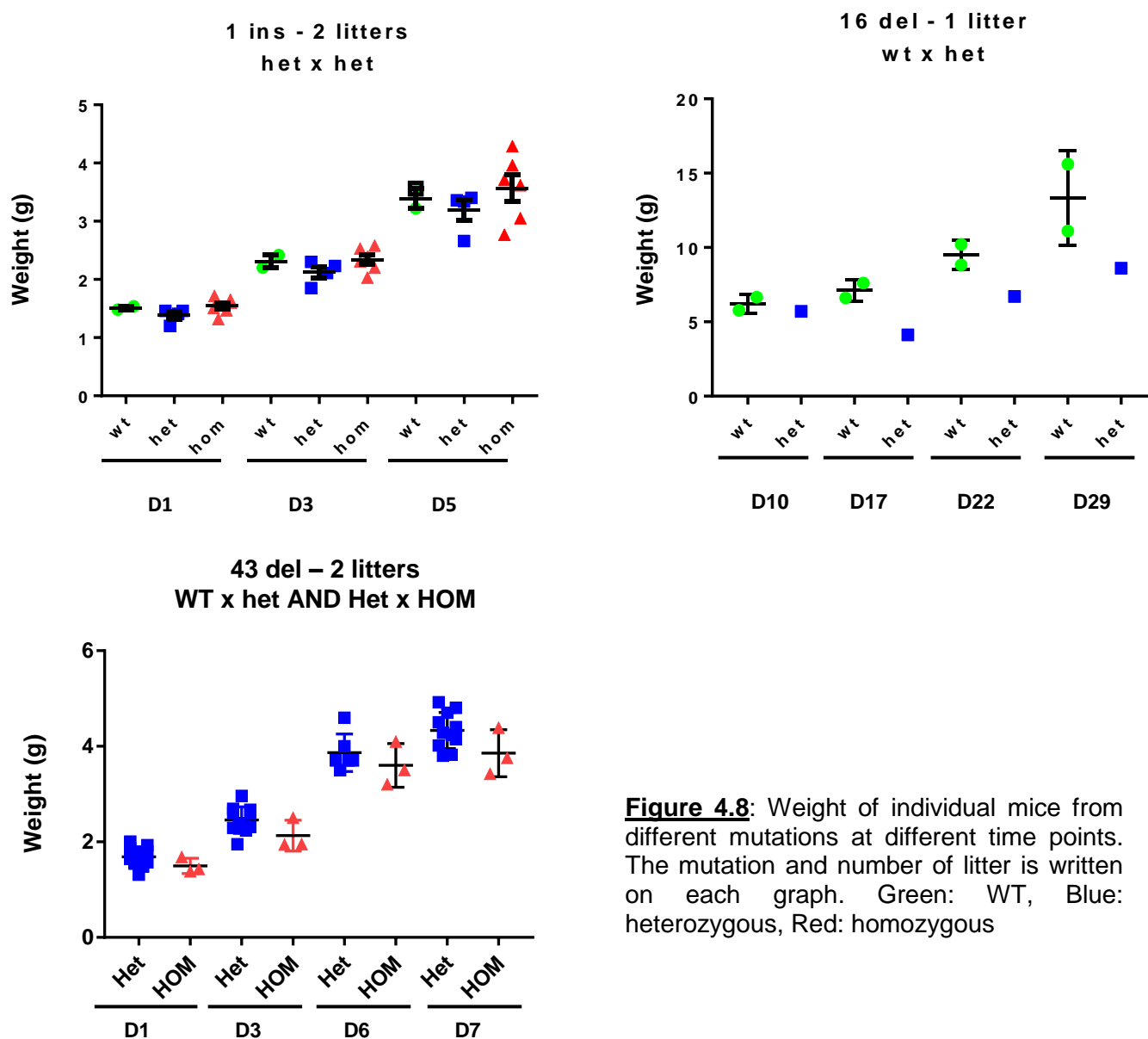


Figure 4.8: Weight of individual mice from different mutations at different time points. The mutation and number of litter is written on each graph. Green: WT, Blue: heterozygous, Red: homozygous

Very interestingly, It can also be observed that heterozygous individuals carrying the same mutation but born from a different mother do not show the same phenotype. The mutation 113 del + 62 insertion shows that, at day 9, heterozygous pups from a Het mother weighed above 6g (**Fig 4.9 Top Left**) whereas heterozygous from a HOM mother weighed around 4g (**Fig 4.9 Top Right**). The same is true between the HOM pups; the one from a HOM mother is much smaller than the one from a Het mother at day 9 and day 15. A similar pattern is observed for the 15 deletion + 3 insertion mutation, where homozygous from a homozygous mother (MZhomo) are smaller than the homozygous with a heterozygous mother at the same time points (**Fig. 4.9 Bottom**). It suggests that there may be a more severe phenotype for the offspring when the mother is itself depleted of the protein perhaps indicating a maternal effect. The numbers are very low and no statistical test can be performed, therefore these data are only preliminary, but they strongly support a role for ZFP263 in growth and development.

Together, these highly preliminary results show that *Zfp263* KO mice are viable but might exhibit a phenotype conferred by the mutation. The placentas are significantly smaller in homozygous embryos and homozygous pups tend to be born smaller. They do not present obvious signs of developmental retardation but the very low number of homozygous animals born compared to the heterozygous and WT offspring from heterozygous intercrosses is a sign that they are strongly affected by the loss of ZFP263 function. Furthermore, there might be a more severe phenotype in the maternal-zygotic homozygous offspring and this has not been assessed.

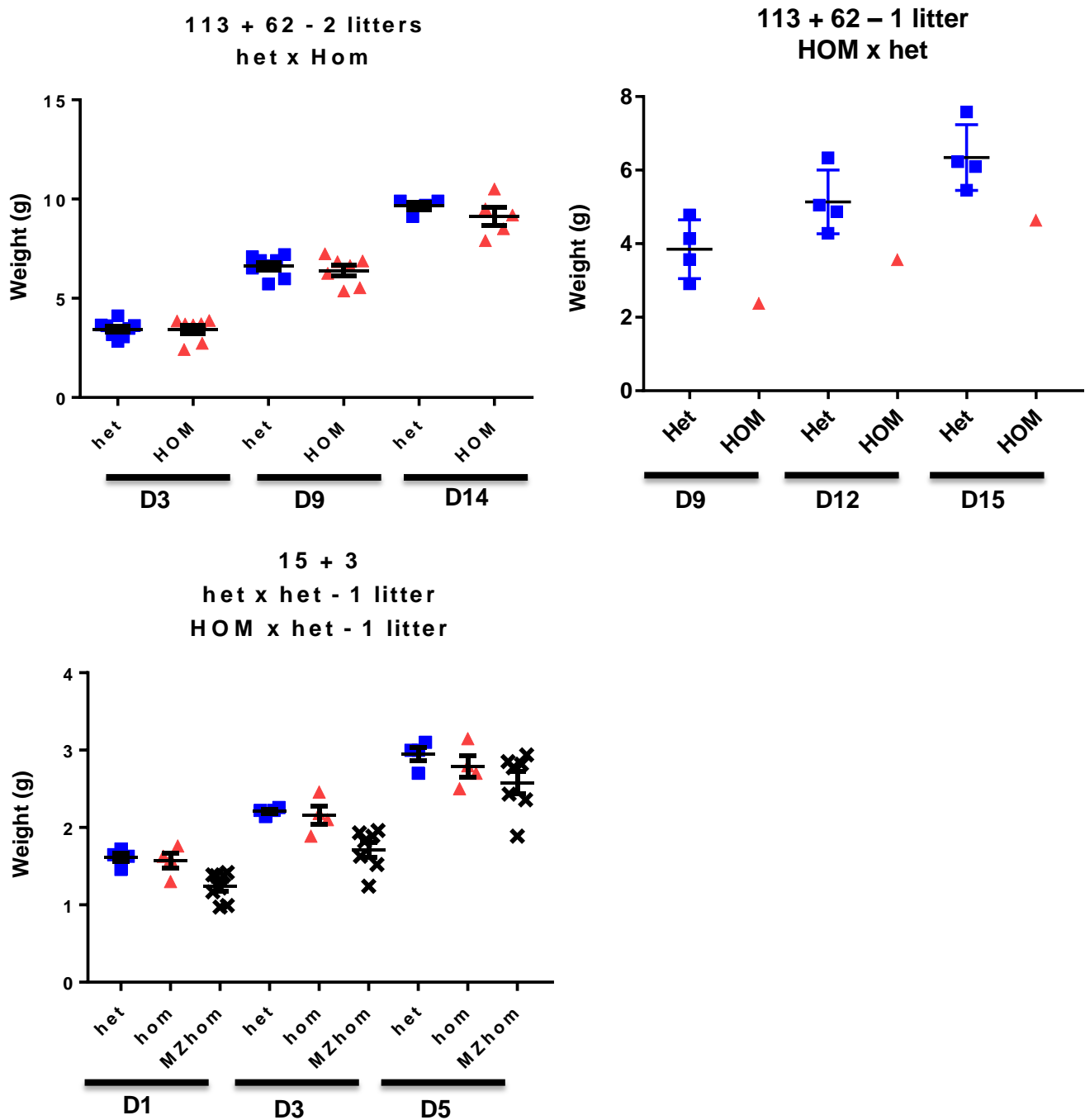


Figure 4.9: Weight of individual mice from different mutations at different time points. The mutation and number of litter is written on each graph. The top panel shows weight for the 113+62 mutation, with a het mother on the left graph and a HOM mother on the right graph. The bottom graph shows weight for the 15+3 mutation. Blue: heterozygous, Red: homozygous, X: maternal-zygotic HOM from a homozygous mother.

4.5 Discussion and Conclusion

4.5.1 *Zfp263* is efficiently targeted by the CRISPR/Cas9 methodology

A mutagenesis experiment using CRISPR/Cas9 in mouse zygotes successfully generated 16 viable *Zfp263* mutants in exon 6, and 18 mutant lines in exon 1. The injection of the Cas9 protein appeared much more efficient than with Cas9 mRNA. The mRNA might be indeed more fragile and sensitive to degradation during storage and at the time of injection, as well as *in vivo*. The protein is probably more stable and should be active straight after the injection, whereas the mRNA has to be processed and translated before the technology can be functional potentially leading to greater mosaicism and reduced germline transmission. Although the gRNA and Cas9 protein were injected into one pronucleus, most founder mice were compound heterozygotes. This suggests that the protein and the gRNA complex are stable for long enough to target both maternal and paternal genome. Only a few founders were heterozygous, where the Cas9 was only able to target one allele. All founders bred with WT passed on their mutations to the next generation except for 2 mutations. This means that these individuals were mosaic and their germ cells were not carrying the mutations. In this case, the DNA mutation probably occurred at or after the 2-cell stage zygote. However, no mice were found to be mosaic with more than 2 mutations, suggesting that the allele cannot be targeted again after a first round of mutation and that in most cases the gRNA/Cas9 complex is being diluted or degraded before the first cell division.

The mutations in exon 6 occur upstream of the portion coding for the zinc fingers and most generate premature termination codons. Despite the premature STOP codon, *Zfp263* transcript in the initial mutant mice was present at a similar level than in WT mice. This means that the transcript is not being recognised as an aberrant transcript by any mechanism, and therefore the protein is likely to be translated. One founder mouse (KO4) carrying two different deletions (16bp and 43 bp) showed a higher *Zfp263* expression than the WT. This could suggest a negative self-regulation by the protein itself on *Zfp263* expression, although none of the other mice showed a similar effect despite similar mutations. The two mutations (16bp and 43bp deletions) and their effect on transcription and phenotype are being studied in more details. The mutations in exon 1 could lead to the degradation of the transcript by the nonsense-mediate decay, as the premature termination codon occurs in the first exon, upstream of the exon-exon junction. One mutation however does not change the open reading frame and therefore the transcript is very unlikely to be degraded, unless the large deletion (123bp) creates aberrant secondary structures or recruits

another degradation mechanism. This will be assessed by measuring the level of *Zfp263* transcript in different tissues.

It is possible to predict the mutated protein structure based on the DNA sequence. Mutations in exon 6 could produce a truncated protein without its zinc fingers, therefore potentially unable to achieve its function through its binding capacity. However the SCAN and KRAB domains should be intact, therefore the truncated protein could still interact with its molecular partners, but would not be able to target them towards specific DNA sequences. Mutations in exon 1, if the transcript is not being degraded, would produce a very short peptide with a truncated SCAN domain. The N-terminal part of the domain should be intact with the residues folding into helices and involved in the dimerization process. This minimal core structure should be sufficient to confer the dimerization property of the SCAN domain, although this is not certain. Therefore, the peptide might be able to interact with its partners, but would lack the KRAB domain and the zinc fingers. The last mutation in exon 1 deletes 41 amino acids and disrupts the residues involved in the third helix formation. It is unknown whether the first two helices will be able to form in the absence of the third one. Furthermore, although the specificity of the dimerization pattern is exerted by the first helix, the core structure itself will be disrupted and therefore might affect the protein dimerization properties. Thus, this truncated protein would have functional zinc fingers and a WT KRAB domain, but might be impaired in its ability to recruit its partners through its SCAN domain, which might affect its function. However, the lack of quantitative protein analysis in mutant compared to WT tissues makes any of these suggestions and the following results very hypothetical. Optimising the protocol is now a priority.

4.5.2 Phenotypic characterisation of *Zfp263* mutants.

The results of this chapter were generated very recently, using only mutants in exon 6 of *Zfp263* shortly before the writing of this dissertation and the numbers are very low so the findings can only be considered preliminary. However, some phenotypic trends are observed. First of all, the weight of HOM placentas are significantly decreased compared the WT placentas, suggesting a role for ZFP263 in placenta growth. It would be interesting to perform morphometrics on these placentas to try and decipher which cell types of the placenta are affected, which might provide insight into any functional effects. The embryos that were collected at E16.5 did not show any growth retardation or defect and the homozygotes were not significantly smaller than their WT littermates at this stage. However

only 3 mutations were tested with only one litter each, so the numbers are very low but would suggest that the lack of DNA-binding ZFP263 does not affect embryo development.

However, the reduced homozygosity compared to the heterozygous is recapitulated in several litters and from different types of frameshift mutations. Though numbers are low, this suggests partial embryonic lethality, although that was not recapitulated in the embryo collection experiment. The numbers were very low for the embryos collection whereas the heterozygous intercrosses were performed for 3 to 5 litters for 5 different mutations. Some homozygous are viable and healthy, therefore the lack of the WT ZFP263 protein is not lethal, but the low ratio of homozygous to WT clearly indicates reduced viability of the animals before birth and suggests that there might be variation in the expressivity of the phenotype between individuals. It would be interesting to perform more embryonic experiments to understand at which stage of development the embryos are affected. Another possibility explaining this low ratio would be that the homozygous might be dying very early after birth, before genotyping. However the litters are checked very regularly, the number of pups followed from day 0 and a dead body would be removed and genotyped. Unless the homozygous die straight after birth and are completely eaten by the parents, this hypothesis is less likely to explain the results.

Finally, there is trend of growth retardation in homozygous and some heterozygous compared to WT. This suggests that the lack of functional ZFP263 triggers a growth phenotype. Some pups are already smaller at day 0, suggesting an embryonic phenotype. This was not recapitulated with the E16.5 embryo collection experiment, but again this experiment was performed with only 3 mutations and the numbers were very low, therefore the results cannot be very conclusive. Some homozygous are born only slightly smaller than their littermates but do not put on weight as much as the WT and do not catch up before or after weaning. Some heterozygous also display the same phenotype and some are severely affected. This suggests that there might be an accumulation of small effects due to the truncated protein that do not necessarily induce lethality but do affect growth and development. Very interestingly, it was observed for two different mutations that the pups born from a homozygous mother (maternal-zygotic mutants) have a more severe phenotype than the ones born from a heterozygous mother (zygotic mutants). This result could indicate a maternal effect - that the maternal ZFP263 is essential in the fertilised oocyte to regulate gene expression or epigenetic status very early on after fertilisation, and that the absence of the functional maternal protein has an influence on growth and development later on. The maternal zygotic heterozygous and homozygous pups do not catch up in weight after birth as

much as the zygotic heterozygous and homozygous. The lack of WT ZFP263 in the mother might also affect the physiology of the mother and the care she provides to her litter, whether it is a change in behaviour or a lactation issue for example.

More work remains to be done to decipher ZFP263 function in vivo. More crosses will need to be monitored to increase the numbers and allow statistical analyses and more detailed analyses on the mutant animals are in progress.

Chapter 5. Discussion

5.1 ZFP263 is a highly conserved mammals-specific protein

ZFP263 is part of the huge family of zinc finger proteins. In addition to nine zinc fingers, the protein contains a KRAB domain and a SCAN domain which makes it one of the 17 SCAN- and KRAB-containing ZFPs in mouse. Most KZFPs are restricted to tetrapods, but those containing a SCAN domain are older and shared with marsupials or sauropsids. *Zfp263* has indeed orthologues in 101 species of mammals but none in fish, birds or reptiles, and therefore appeared relatively recently on the evolutionary scale, some 180 million years ago in the platypus. The gene may have replaced an ancient gene that would then have been lost, or evolved a mammal-specific function and could have been involved in specific mammalian phenotype, metabolism or development. One could also suggest that the gene evolved alongside mammalian characteristics, without being directly involved in their regulation. Until recently, the KZFPs were indeed thought to have evolved alongside retrotransposon elements to repress their activity and protect the host organism. *Zfp263* might have evolved in platypus to target a new retrovirus and suppress its activity, and might have been kept since then. However this arm race hypothesis cannot entirely explain the expansion of this large family of DNA-binding proteins, and recent studies showed that older SCAN-KZFPs are more prone to bind to unique regions and did not target retrotransposons. This suggests that ZFP263 could have a unique function not related to retrotransposon regulation.

In the intervening time over evolution, *Zfp263* experienced very few changes. Its zinc fingers and in particular the three amino acids involved in DNA binding have been extremely well conserved. Although other factors can affect the DNA interactions, this suggests that the orthologue proteins could potentially target the same regions in different species. Similarly the SCAN domain, and in particular the residues forming the three structural loops involved in the interactions with other molecular partners, is very similar in all orthologues, suggesting again that the proteins from different species could function within the same network of proteins. Protein-protein interactions studies have identified some 57 partners to ZFP263 in cells, and it would be interesting to assess whether these proteins are also conserved in the same species as ZFP263. The protein has been under purifying selection across evolution and thus has been strongly protected against mutations to preserve its structure and most likely its function. This means that the protein is likely to have retained its initial function when it first emerged in platypus. Interestingly, the old retrotransposons have today lost their retrotransposition activity as they have mutated across evolution, and therefore are not a

threat to the host organism anymore. Therefore it would be surprising that ZFP263 was conserved over hundreds of millions of years to target an inactive retrovirus. This strongly suggests that ZFP263 initially evolved its function at unique genomic loci.

Finally, ZFP263 KRAB domain is more divergent from the consensus sequence and residues critical for KAP1 recruitment are absent in all orthologues. This supports the hypothesis that ZFP263 should not be considered as a canonical KZFP acting as a repressor targeting retroelements to repress them through the recruitment of KAP1. It suggests instead that ZFP263 has probably evolved a unique mammal-specific function in platypus and has been conserved since then to exert the same function today.

Expression of *Zfp263* across tissues in mouse and human shows that the gene is widely expressed in several human and mice tissues. The protein quantification in human tissues also shows a low to medium level in most tissues. It suggests that the protein does not act in a tissue-specific way but that it is present at different developmental stages and throughout adult life. The protein might target the same loci in different tissues, but this is very uncertain as it depends on the cellular and chromatin context in one tissue or on the combination of zinc fingers involved in the DNA contacts. It would be interesting to assess whether the identified interacting factors for ZFP263 are also expressed in different tissues and whether ZFP263 could act within the same network of proteins in different tissues.

5.2 ZFP263 is a unique non-canonical KZFP

5.2.1 ZFP263 targets unique genomic loci in mESCs

The identification of ZFP263 binding sites in mESCs revealed several interesting results. First of all, 275 high-confidence binding sites were identified; 92% of them contained a shared consensus DNA sequence, highly similar to the human ZNF263 binding motif in K562 cells (Fietze et al. 2010). This confirms that the human and mouse orthologues target a similar DNA motif and the same combination of zinc fingers are likely to be used in DNA recognition in the two species. I showed in Chapter 2 that the amino acids involved in DNA recognition are identical in ZFP263 orthologues from 14 species, even in its most ancestral form in platypus. Therefore it is possible that the protein targets the same DNA sequence in all of the species genomes. Furthermore, the protein has been under purifying selection

across evolution to protect it from mutations, hence it may have retained its initial function when it first emerged before platypus.

The second striking result is that ZFP263 targets are not significantly enriched in transposable elements. The protein preferentially targets intragenic unique loci and do not overlap with KAP1 binding sites in mESCs. These results are not consistent with the “arms race” model that suggests that KZFPs evolved rapidly alongside retrotransposons to silence their activity via KAP1 recruitment. However recent genome-wide studies on human KZFPs showed that SCAN-containing KZFPs are old proteins shared with marsupials or sauropsids (Imbeault et al. 2017). Plus, the authors observed that SCAN-KZFPs were more prone to bind unique genomic loci such as promoters, and were impaired in their ability to recruit KAP1. We hypothesise that a subset of very old and highly conserved SCAN-KZFPs are distinct from the other “canonical” KZFPs, and that ZFP263 is a representative of this subfamily. These proteins must have emerged in sauropsids or monotremes and have been conserved across evolution to protect their initial functions. They are unlikely to be involved in transposable elements regulation due to their inability to bind KAP1.

5.2.2 ZFP263 is associated with active promoter and enhancer characteristics

The meticulous characterisation of ZFP263 binding sites confirmed the highly unusual nature of the protein. More than 70% of the binding sites are located within genes, either at promoters, within introns or at intron/exon junction. A large subset of binding sites is enriched in H3K4me3 and H3K27ac, which are associated with active chromatin region. H3K4me3 is found at promoters of active genes and regulates transcription by recruiting positive transcription factors. It can also be found at bivalent promoters in association with H3K27me3, but this is not the case in this dataset. Similarly, H3K27ac is enriched at promoter regions of transcriptionally active genes in human T cells and is thought to prevent repression by blocking trimethylation at H3K27. A subset of genes is also enriched in H3K36me3, which is found on actively transcribed gene bodies. ZFP263 binding sites are therefore preferentially located at promoters, and enriched in histone marks associated with actively transcribed gene promoters, suggesting that ZFP263 could positively regulate target gene transcription.

A large subset of binding sites also co-localises with H3K4me1, H3K27ac, H3K64ac and H3K122ac, which are all hallmarks for active enhancers. Some sites co-localise with all four marks while others only display H3K64ac and H3K122ac. Recently, acetylation on K64 and

K122 of histone 3 was identified as a signature to mark a subset of new enhancers lacking H3K27ac. This shows that ZFP263 is associated with active enhancers and might be involved in transcriptional activity through interactions with this particular class of enhancers providing an interesting aspect of specificity to its function.

Finally, a small subset of sites is located at intron-exon junctions and enriched in H3K36me3. H3K36me3 is known to be a histone modification enriched on gene bodies and particularly on exons. It was suggested that H3K36me3 could be involved in alternative splicing regulation by signalling exons to the splicing machinery. This suggests that ZFP263 might be involved in splicing regulation of its targets. Nevertheless we cannot explain whether ZFP263 binds to H3K36me3 to recruit splicing effectors, or whether ZFP263 mediates the establishment of the histone marks at its targets.

These results are very exciting as they question our current understanding of KZFP functions, and support the hypothesis that ZFP263 is a non-canonical KZFP. As the other human SCAN-KZFPs have been shown to prefer binding promoters and be less associated with KAP1 than the more conventional KZFPs, it supports the theory that the SCAN-KZFPs are distinct from the more conventional KZFPs in their functions, and I hypothesise that other SCAN-KZFPs in mammals have active roles in transcriptional regulation.

5.2.3 ZFP263 target genes display a large range of expression levels

ZFP263 target genes display various levels of expression in mESCs. First, this could suggest, as found before in the human study, that ZFP263 might act both as an activator and a KAP1-independent repressor depending on the genomic context. ZFP263 might be able to recruit different co-factors depending on the chromatin or cellular context and thus have a different effect on transcription. Nevertheless, because ZFP263 binding sites are associated with active enhancer histone marks, it may act as a repressor by modulating the accessibility of an active enhancer. ZFP263 might influence transcription of other genes through its enhancer binding without targeting them directly. Finally, one could also argue that ZFP263 might not be involved in transcription regulation at all considering the inconsistency between the expressions of its targets. It might be involved in different processes not directly related to transcription regulation. As the SCAN-containing proteins are known to dimerise with other SCAN domains, one might hypothesise that ZFP263 could function to shape chromatin structure, by bringing into proximity DNA sequences targeted by two different SCAN-ZFPs.

ZFP263 could be involved in structural organisation of the chromatin, for example at a higher level in chromatin compartments formation, or in the generation of structural loops.

5.2.4 ZFP263 protein interactions

Existing proteomics dataset identified ZFP263 as part of an interacting network with 57 other proteins, some of them being other SCAN-containing proteins, though not the majority. The careful analysis of allele-specific binding sites suggests that genomic-variation between the two mouse strains might directly alter the ZFP263 DNA recognition and specificity or impair a protein-protein interaction between ZFP263 and one of its interactors. However some allele-specific binding did not contain a disrupted motif and might represent a differential functional interaction with other proteins. Moreover the duality of ZFP263 function as activators and repressors might be influenced by interactions with other proteins/transcription factors, although I showed in Chapter 3 that there was no other significantly enriched motif in ZFP263 binding motif. If ZFP263 interacts with other DNA-binding proteins, the DNA sequences targeted by these potential co-factors were not sequenced with ZFP263 binding sites. This supports the hypothesis of trans-acting co-factors that could potentially be involved in chromatin structural conformation. Understanding more about these types of potential interactions might provide insights into the biochemical properties and regulation by ZFP263.

5.2.5 Future approaches

I propose to use ZFP263 as a model to shed light on the unusual SCAN-containing KZFPs functions and mechanisms of action. First, identification of ZFP263 targets *in vivo* would add important insight into ZFP263 function to determine whether the binding repertoire is dynamic across development. ChIP-seq could be performed in liver, a tissue that provides less heterogeneity than other tissues. Second, analysis of transcriptomes from mutant tissues and cell culture and comparative analysis of perturbed transcripts with and without binding sites would provide insight into which transcripts are directly and indirectly regulated by ZFP263. I propose to assess genome-wide transcription in WT and KO mESCs and liver by performing RNA-seq transcriptome analysis using qPCR for validation. Gene expression changes would be quantified as well as changes in splicing isoforms. These transcriptional effects would be correlated with alterations in epigenetic state in regulatory regions and at ZFP263 target sites in particular H3K4 methylation and H3K64/122 acetylation, using ChIP-grade antibodies.

Finally I have hypothesised that other interacting proteins are likely to contribute to the function of ZFP263. Therefore understanding complexes including ZFP263 is likely to be a useful approach to understand function and biochemical analysis could be performed to identify potential interacting partners. To identify the interacting partners of ZFP263 in an unbiased fashion in mESCs, we could combine proximity-dependent biotin identification (BioID) with mass-spec (Roux et al. 2012). The BioID protocol applies a biotin ligase (BirA) fused to the ZFP263 protein which acts as a bait. The construct which includes a nuclear localisation signal, will be stably expressed in mESCs and when stimulated with biotin the BirA biotinylates proteins in its proximity – in this case those interacting with ZFP263. Biotinylated proteins are then isolated by streptavidin affinity capture and identified by mass spectrometry. Control constructs lacking BirA will also be run in parallel, and the experiment also conducted in cells with the endogenous ZFP263 KO. As interaction partners and neighbours are marked by stable covalent modifications, it is unnecessary to maintain protein complexes throughout the purification step. Harsh lysing condition can be employed to effectively solubilize most cellular proteins (Lambert et al. 2015). Recent studies show that, for a given protein, BioID recapitulates the detection of previously known interaction partners but also enables the identification of novel protein-protein interactions. We hypothesise that ZFP263 interacting partners will likely to be other SCAN-containing proteins, because the SCAN domain is understood to function as a protein interaction domain.

The SCAN domain is of particular interest because it is poorly described and its implications in the protein functions are not well understood. To further elucidate its role, a mutagenesis and complementation approach in the ZFP263 KO mESC cell culture system could be taken. ZFP263 function would be rescued in null mESC by transfecting wild type *Zfp263*, or various *Zfp263* truncations/mutations, *i.e.* without the KRAB domain, or without the SCAN domain or without either of the domains, or if appropriate, with other more subtle mutations in these or other parts of the protein. In these contexts, genome-wide transcription by RNA-seq will be assessed as well as the associated epigenetic environment surrounding the targets by histone ChIP-qPCR in rescued and potentially partially rescued mESCs. I hypothesise that the SCAN domain will be a key player in the ZFP263 interaction network, thus differences in transcription and chromatin states in the protein lacking all or part of the SCAN domain are likely to be observed. Network analysis might also provide previously undetermined functional pathways that this DNA binding protein might regulate.

5.3 ZFP263 regulates key genes for normal growth in mice

5.3.1 ZFP263 is involved in growth regulation and placenta development

Zfp263 mutant mice were generated using the CRISPR-Cas9 technology. The frameshift mutations induce a premature termination codon before the zinc finger coding regions, and the mutated transcripts should produce a truncated protein without its zinc fingers, although this has not been checked yet. The protein would thus lack its binding ability. The ChIP-seq experiment showed that ZFP263 is likely to regulate key processes throughout development and adulthood in several tissues, rather than being very specific to one pathway or one tissue, and might also regulate other genes indirectly. This makes phenotypic analysis somewhat challenging, although some phenotypic trends are observed.

Het x het intercrosses were generated for different frameshift mutations and revealed that there is a significant reduction of E16.5 homozygous placenta weight compared to the WT. Morphometrics analyses on these smaller placentas would be useful to assess whether their functions could be affected and whether it could impact embryos development, but it does suggest a role for ZFP263 in placenta development; yet the genes identified as target genes in the ChIP-seq experiment were not enriched in a Gene Ontology term related to placenta development. However, as discussed in 5.2.3, ZFP263 could regulate additional genes without directly targeting them, either through its binding to enhancers, or by arranging chromatin conformation in partnership with other DNA-binding proteins. Furthermore, it is interesting to note that *Zfp263* is present in species of monotremes and marsupials, where there is no placentation. I showed that the gene has been strongly conserved and protected against mutation across evolution, and I hypothesised that *Zfp263* must have conserved its function in different species. It is thus bewildering to identify a potential role in placenta development for ZFP263. ZFP263 might have evolved a different role from marsupials to eutherians, without changing dramatically its sequence and structure, or the protein may regulate another process and indirectly impact placenta development.

Het x het Intercrosses also revealed reduced homozygosity compared to the heterozygotes after birth, which was recapitulated in several types of *Zfp263* mutations and for several litters. Nevertheless, some homozygous were viable and healthy, meaning that the lethality phenotype is variably penetrant. These results suggest that the presence of a truncated ZFP263 induces partial embryonic lethality or reduced viability just after birth. ZFP263 may therefore be involved in transcriptional regulation of key genes for normal development.

Finally, although the numbers are very low and the results very preliminary a phenotypic trend was observed. The homozygous pups tend to be smaller than their heterozygous littermates. Some of them were born smaller and others were not putting on weight as much as their littermates after birth and after weaning. This suggests that the lack of functional ZFP263 triggers an embryonic phenotype and a growth phenotype after birth. It would be interesting to know whether the reduced weight of the homozygous pups at birth could be due to a smaller and potentially impaired placenta. The partial lethality and the phenotype observed in some heterozygous as well suggest that there is probably an accumulation of effects and a variation in the expressivity of the phenotype between individuals.

Very interestingly, the more severe phenotype in maternal-zygotic mutants compared to zygotic mutants might indicate a role for the protein in the fertilised oocyte at the very early stages after fertilisation. The lack of the full-length protein could impact gene expression, epigenetic status or chromatin conformation and affect growth and development later on. Similarly, because the maternal zygotic heterozygous are also more affected than the zygotic heterozygous after birth, it could indicate a role for ZFP263 in the physiology of the mother. It is of interest to note that *Zfp263* evolved in mammals but not in birds, fish or reptiles, and therefore I hypothesised that ZFP263 could be involved in a mammal-specific function. Lactation for instance could be affected by the lack of ZFP263, as lactation is a characteristic of mammals, including monotremes and marsupials. A lactation issue might affect development of pups after birth and could explain the observed phenotype, but more analyses remain to be done to fully understand ZFP263 role *in vivo*.

5.3.2 Future approaches

First of all, more crosses are being monitored to increase the numbers and allow statistical analyses and more detailed analyses of the mutant animals are in progress. More heterozygous intercrosses are being generated using individuals following several rounds of back-crossing with WT C57Bl/6 mice. Indeed, the initial intercrosses were set up using F1 individuals with a partial hybrid genome, with 75% C57Bl/6 and 25% CBA genetic background. Phenotypes can sometimes differ according to the genetic background and therefore all the experiments need to be repeated using an individual from later generations with a pure C57Bl/6 background. Nine backcrosses are necessary to generate a 99.9% C57Bl/6 background.

We are following development and phenotype over the life-course of the mutant mice and include wild type, heterozygous, and homozygous stock as well as maternal-zygotic mutants lacking this KZFP in the oocyte with potential impact for the establishment and maintenance of the earliest epigenetic programme. Prenatal growth, placentation and comparative developmental analysis will be conducted on conceptuses. Mice will be assessed postnatally for growth, viability, behaviour, and metabolic state through a gross phenotypic pipeline. Further more detailed analyses will be applied to characterise effects identified in the initial phenotypic characterisation. These same experiments are also being conducted using the second mutant lines generated targeting exon 1 of *Zfp263* gene and the phenotypes will be carefully compared.

5.4 Conclusion

The aim of this piece of research was to characterise ZFP263 in mouse, a member of the huge KZFP family. This project was part of the seventh framework programme of the European Union EpiHealthNet, which aimed at improving human health by understanding the mechanisms and pathways in early development that have a long term effect on the health of individuals across their lifespan, and in particular by identifying the key genes and pathways affecting epigenetic and imprinting sensitivity in early stages of development, in order to create intervention tools against epigenetic misprogramming.

ZFP263 is a KRAB- and SCAN-containing protein with nine zinc fingers that was studied using an *in vitro* system in human cells (Fietze et al, 2010). The authors of that study conducted a ChIP-seq experiment and a knock-down experiment and concluded that ZNF263 could act both as an activator and a repressor, and might play a critical role in maintaining cell structure and proliferation. A second study mentioned ZNF263 where they found that stressed children lost methylation at ZNF263 sites. They suggested that loss of ZNF263 itself in these children could mediate the loss of methylation.

Here, I identified ZFP263 as a highly conserved protein specific to mammals, expressed throughout development and adult life in mouse and human. I showed that ZFP263 is an unusual KZFP, as it does not bind transposable elements nor is associated with KAP1 binding sites. It targets unique genomic loci and is associated with active histone marks characteristics of promoter and enhancer activity. Its target genes display a large range of expression and are involved in several key biological processes including “negative

regulation of transcription from RNA polymerase II promoter". Based on the mutagenesis analysis in mice, I showed that ZFP263 is likely to be involved in growth regulation and placenta development, although these are preliminary data.

In summary, I hypothesise that *Zfp263* emerged in platypus to exert a mammal-specific function not related to the regulation of transposable elements through the recruitment of KAP1. Instead, the work presented in this dissertation allows me to draw two hypotheses about the mode of action ZFP263. First, ZFP263 may actively regulate gene transcription of its direct targets through binding to their promoters, or of indirect targets through binding to active enhancers. Second, ZFP263 might act as a structural protein to arrange chromatin conformation through interactions with other SCAN-ZFPs, and therefore might not be directly involved in transcriptional regulation.

I hypothesise that ZFP263 defines a new class of KZFPs. I propose that a subset of old and conserved SCAN-containing KZFPs does not follow the typical model of KRAB ZFPs recruiting KAP1 and mediating repressive chromatin states. These proteins surely merit further investigation to better understand their functions and shed light on the very poorly understood SCAN domain role.

Chapter 6. Methods

6.1 Declaration of work

The experiments listed below were done by collaborators.

6.1.1 ChIP-seq assay

Ruslan Strogantsev designed the GFP-T2A-ZFP-FLAG plasmids. Angela Noon generated the first FLAG tagged ZFP mESC clones and the first replicate of the ChIP-seq library. She optimised the mESCs transduction protocol and the ChIP protocol. Mouse reciprocal hybrid ESC lines had been generated in collaboration with Bowen Sun and Prof Roger Pedersen (Sun et al. 2012) and were maintained in the Ferguson-Smith lab.

Hui Shi performed the reads trimming, generated the hybrid genome C57BL6/J-Mus Castaneus, performed the alignment and generated the files with the deduplicated and aligned reads for each genomic background and each replicate.

6.1.2 Animal work

The CRISPR-Cas9 microinjections in mouse zygote were performed by William Mansfield from the Stem Cell Facility.

6.2 Conservation and evolution analysis

6.2.1 Orthologues identification

Orthologues to the human ZNF263 were identified with Ensembl Genome Browser (Yates et al. 2016) – Ensembl release 89 – and the NCBI Gene resource (NCBI Resource Coordinators 2017; Brown et al. 2015) . The location of the gene, the number of exons and the surrounding genes were verified for each orthologue. CDS and amino acids sequences were retrieved from Ensembl or NCBI. 14 orthologues from 14 different species were analysed in more details.

6.2.2 Alignment, Percentage identity and Ka/Ks ratio

The CDS and amino acid sequences from 14 orthologues were aligned using the Multiple Sequence Alignment tool ClustalOmega with default parameters (Sievers et al. 2014). Alignments were visualised in Genedoc. Percentage of identical and similar amino acids between orthologues were calculated using Ident and Sim from the Sequence Manipulation Suite (Stothard 2000). The maximum-likelihood phylogenetic tree was built with PhyML with 500 bootstraps (Guindon et al. 2010). Synonymous versus nonsynonymous mutations ratio (Ka/Ks) was calculated using the package seqinr in R (<https://CRAN.R-project.org/package=seqinr>).

6.3 Cell culture

6.3.1 Mouse embryonic stem cell culture

Flasks or dishes were coated for one hour with 0.2% gelatine and dried under the hood or coated with poly-L-ornithine for twenty 20 minutes, rinsed twice with water and coated for 2 hours with 1X laminin. Flasks and dishes were rinsed with 1X PBS and used immediately or stored at 4°C. Cell vials were thawed at 37°C and grown in 2i-LIF medium (NDiff N2B27, 50µg/mL Gentamycin, 1µM PD0325901, 3µM CHIR 99021 and 500U Mouse LIF Medium). Medium was changed every day or every other day. When ~70% confluent, cells were split. Cells were rinsed first with warm 1X PBS and coated with accutase for 5 minutes at 37°C. Cell clones were thoroughly dissociated, pelleted and resuspended in fresh media. For freezing, the same process was followed and wells but cells were resuspended in fresh medium with 10% DMSO and slowly frozen at -80°C before being stored in liquid nitrogen.

6.3.2 mESCs transduction

6.3.2.1 HEK293 cells transfection

On day 1 HEK293 cells were plated in 12 well plates. On day 2, cells were transfected with 0.5 µg of expression construct, 0.375 µg of the gag/pol elements plasmid (pMDLg/pRRE) and the packaging plasmid (pRSV-Rev), 0.25 µg of the envelope protein plasmid for producing lentiviral particles (pCMV-VSV-G). For 12-well plate wells, the plasmids were mixed with 50 µL of OptiMEM medium without serum. 3 µL of Metafectene Pro (Biontex) were diluted with 50 µL OptiMEM without serum. DNA solution was added to Metafectene Pro solution and was incubated for 20 min at RT. 100 µL of the complexes were added to

HEK293 cells. After ~6 hours medium was changed with fresh medium and cells were left for 2 days at 37°C. All materials in contact with the virus particles or plasmids/metafectene/optimum were treated with 3% virkon before being disposed of.

6.3.2.2 Lentivirus purification and concentration

On day 4, supernatant from HEK293T was collected. Cells in suspension in the supernatant were pelleted by centrifugation, and the supernatant was filtered through a Sartorius Minisart Syringe filter. One volume of Lenti X concentrator (Clontech) was added to 3 volumes of supernatant and the mixture was incubated for at least 30 minutes at 4°C. Viral particles were concentrated by centrifugation (1500 x g for 45 minutes at 4°C) and resuspended in 100 µL of fresh medium. Viruses were used immediately or stored at 4°C for 2 days maximum.

6.3.2.3 Transduction and single clone selection

mESCs were grown separately during the virus production period. On the day of transduction, 1×10^5 ESC for 1 well of a 12-well plate were mixed with 5 µg/mL final concentration of polybrene. Cells were infected with 50 µL of viral suspension and incubated for 2 days. Cells were finally transferred to 10-cm dishes when confluent and checked for GFP expression. Single GFP positive clones were manually picked, dissociated in accutase and transferred to a flat bottom 96-well plate. When nearly confluent, cells were transferred to 12- or 6-well plates and T25 flasks for further expansion.

6.4 ChIP-seq assay

6.4.1 Chromatin immunoprecipitation

6.4.1.1 Crosslinking

Cells were grown in 10 10-cm dishes and cross linked when 70% confluent. 9 dishes were used for crosslinking and 1 dish for cells counting. 1% final concentration formaldehyde was added directly to cell media and dishes were shaken for 10 minutes at RT. The reaction was stopped by adding 0.125M glycine final concentration. Media was removed and cells rinsed with cold 1X PBS. 5 mL of cold lysis buffer (0.25% Triton X-100, 10 mM EDTA, 0.5 mM EGTA, 10 mM Tris pH8, filtered and stored at 4°C. 1 protease inhibitors tablet was added in

50 mL before use) was added to the dish. Plates were kept on a cold tray and cells were scraped into lysis buffer. Cells were pelleted at 2000 rpm for 5 min at 4°C. Supernatant was removed and the pellet was transferred in residual buffer into a clean tube. Pellet was frozen in liquid nitrogen or used for sonication.

6.4.1.2 Sonication

Crosslinked chromatin was thawed on ice and spun at 2000 rpm for 5 min at 4°C. Residual lysis buffer was removed and the pellet was resuspended in ChIP RIPA buffer (1X PBS, 1% NP-40, 0.5% Sodium Deoxycholate, 0.1% SDS, filtered and stored at 4°C. Complete Protease inhibitors were added before use). Chromatin was transferred in siliconized tubes and sonicated in a Bioruptor PLUS for 15 minutes in cycles of 30 seconds ON, 30 seconds OFF, twice. Lysates were spun at 14000 rpm for 15 min at 4°C. Supernatant was pooled in fresh tubes and frozen in liquid nitrogen or used for immunoprecipitation.

6.4.1.3 Immunoprecipitation

Sonicated chromatin was thawed on ice. Lysate from 1×10^7 cells was transferred into a fresh tube for the IP step and about 50 µL for the input. TBS and protease inhibitor was added up to 1 mL. 50µL (PGV) protein A beads per IP were washed with 1XPBS three times. Chromatin was added with 10 µg non-immune rabbit IgG. Chromatin was pre-cleared for 1h on rotator at 4°C. FLAG beads were thawed and 100 µL (PGV) beads per IP were washed three times in 1X TBS. Pre-cleared samples were added to FLAG beads and incubated overnight on a rotator at 4°C. The next day, supernatant was discarded and beads were washed twice in buffer 1 (1% Triton X-100, 0.1% SDS, 2 mM EDTA, 150 mM NaCl, 20 mM Tris pH8, filtered and stored at 4°C), twice in buffer 2 (1% Triton X-100, 0.1% SDS, 2 mM EDTA, 500 mM NaCl, 20 mM Tris pH8, filtered and stored at 4°C), twice in Lithium Chloride buffer (100 mM Tris pH7.5, 500 mM LiCl, 1% NP-40, 1% sodium Deoxycholate, filtered and stored at 4°C) and once in 1X TBS buffer (50 mM Tris HCl, 150mM NaCl, pH7.4, filtered). Chromatin was incubated with elution buffer (0.5M Tris HCl pH7.5, 1 M NaCl, filtered. 3XFLAG peptides were diluted to 150µg/mL final concentration) for 30 minutes at 4°C twice. Samples were frozen to -20°C or reverse crosslinked.

6.4.1.4 Reverse crosslinking

Immunoprecipitated and input samples were thawed on ice and incubated at 65°C with 0.1 mg/mL proteinase K overnight. The next day samples were frozen at -20°C or purified.

6.4.1.5 DNA purification

Samples were thawed on ice. One volume of 25:24:1 phenol/chloroform/isoamyl alcohol was added and incubated at room temperature for 10 min. Samples were spun at 12000g for 8 minutes. One volume of chloroform/isoamyl alcohol was added to the aqueous supernatant. After incubation and spinning the second aqueous supernatant was collected and mixed with 0.1 volume of 3 M sodium acetate pH 5.2, 2.5 volumes of 100% ethanol and 1 µL of 20 mg/mL glycogen per 700 µL. The samples were incubated for 1h at -80°C or overnight at -20°C and spun at 12000g for 30 min at 4°C. DNA pellet was washed twice with 70% ethanol, dried, dissolved in double distilled water and kept at -20°C.

6.4.2 Library sample preparation

The library samples were prepared following the TrueSeq ChIP sample preparation guide from Illumina. Briefly, because the protocol requires between 5 and 10 ng of ChIP DNA as starting material, DNA from three independent ChIP were pooled together. The first step converted the overhangs of ChIPed DNA into blunt ends using an End Repair Mix. The 3' to 5' exonuclease activity of this mix removes the 3' overhangs and the polymerase activity fills in the 5' overhangs. DNA was purified with AMPure XP Beads and the 3' ends were next adenylated to prevent them from ligating to one another. Straight after the 3' adenylation step the adapters were ligated to the ends of the DNA fragments, preparing them for hybridization onto a flow cell for sequencing. At the end of this process DNA was purified twice with AMPure XP Beads. The next step purifies the product of the ligation reaction on a gel and removes unligated adapters, as well as any adapters that might have ligated to one another, and selects a narrow 250–300 bp size-range of DNA fragments for ChIP library construction appropriate for cluster generation. Samples were loaded on a 2% agarose with SyBr Gold gel using 1X TAE buffer. The gel was run at 120 V for 10 minutes, then 60 V for 180 minutes. A gel slice of the sample lane was excised at exactly 250–300 bp using the markers as a guide. DNA was purified following the instructions in the MinElute Gel Extraction Kit except that the gel slices in the QG solution were incubated at RT and not at 50°C. DNA was purified using the MinElute column and eluted in 25 µL of QIAGEN EB solution. Finally PCR

was performed to enrich DNA fragments that have adapter molecules on both end and to amplify to amount of DNA in the library. The PCR program was as follow: 98°C for 30 seconds; 18 cycles of: — 98°C for 10 seconds — 60°C for 30 seconds — 72°C for 30 seconds; 72°C for 5 minutes. Enriched DNA was purified twice with the AMPure XP Beads.

Library size was verified on a BioAnalyser and quantified with the KAPA library quantification kit. Samples quantity was normalised and samples were pooled for sequencing.

6.4.3 Sequencing platforms

The samples were sequenced on a HiSeq2500 platform in The Babraham Institute for the first replicate and on a HiSeq4000 for the second replicate in BGI. The sequencing was 100bp paired-end.

6.4.4 ChIP-seq Bioinformatics Analysis

6.4.4.1 Reads mapping and peak calling

Quality of the ChIP-seq dataset was assessed by FastQC. Adaptors were trimmed using TrimGalore!. The reads with a phred quality score lower than Q20 were discarded as well as reads shorter than 20 bp. The reads were aligned to the hybrid genome C57BL6/J-Mus Castaneus with BWA-backtrack allowing for one mismatch and using the first 20 bp from 5' end of the read. Duplicate reads were discarded using Picard. Peaks were called in the samples using MACS in Seqmonk normalising with the control reads and using default parameters.

6.4.4.2 Binding sites characterisation

DNA motifs were identified with the MEME-ChIP portal. The parameters in MEME were set to look for zero or one occurrence of the motif in each sequence and to look for 5 motifs in total. Motifs are found on both strand of the sequence and are then automatically submitted to TOMTOM that looks for a similar known motif. Repetitive elements associated with the binding sites were identified using the RepeatMasker portal. The Gene Ontology analysis was performed using Panther (Mi et al. 2017). The exon and intron coordinates were downloaded from the UCSC Table Browser – mm10 mouse Genome (Karolchik et al. 2004).

MESCs ChIP-seq datasets for histone modifications were downloaded from ENCODE/LICR – mm9 mouse genome – and transformed to mm10 coordinates. KAP1 ChIP-seq dataset was downloaded from Rowe et al. 2010. The overlap between ZFP263 binding sites and other datasets was performed using Galaxy (Afgan et al. 2016)

6.4.5 ChIP-qPCR

For validation of the sequencing data, a ChIP-qPCR was performed and the %INPUT was calculated. The qPCR was performed on Input sample, which is a crosslinked and sonicated sample but that was not immunoprecipitated, and on the ChIPed samples: the actual sample and the control sample. Four regions not bound by ZFP263 were used as negative controls. Primers below were used

Name	Forward	Reverse
11.1169	CAGTGCCAGCATTGTAGCC	ACCAAAGAGACCCTTAACCAAGA
9.9711	CAGTGGTAGCCTTGGAAGCA	CCATGGGGAGAGGGAGAGAA
5.3353	ACTTCAAGGCCACCTATTCCA	CTGGGATGAGGGAGATCTGAG
3.959	CACTCACGGCTGCGTACTAA	GGTCACCACATCCCTTTGTGA
10.1281	TCATTGCTACTGGAGGTGC	CTTGAGGGGCAGTGAGGATG
1.8652	CTTACAGCAGACGCTGACCA	TGACTGTGACTCGCACCTTC
15.7306	ACATGGGTTGACAAGGGACC	ATCTTAGGCCGAAGCACGAC
12.8088	TCCTTCCACTAGGCCTGTCATA	TGATAAACGGCAGTGGTGTC
9.4448	CCTACCACCTTAGCGCAGAG	GGCAGATGACCCAACCTCTC
12.6929	TGGGCGAGGAATTGGGTTTA	CTGAGTCCTCCCAACATCCG
14.1182	CGCCACCTGCGAGTGTT	ACTCCACTCGCCTAACCTG
10.1291	CTTTCCACACAACCACAGCC	AGAGGTGATGAGTCACTTGAGAG
15.9617	GACAGGAGGAGAGGAGGAGG	TCTCTCTCCAACCCGGAAGT
5.137642	GGGTGGGGAGATCCATTGAG	CAGAGCGGAATTCCTAAGGCT
13.5482	TAGCTTAGGGTCCCAGTCCAA	TGCCCTTGTTGAGACCACAC
2.3071	CCTCTGCCTAACATCCTTCCC	AGTGAAAACCACACCCGCA
7.141	CTAGCTGTTCCGGAGCCAAA	GAGTGTTGTGGGGATGAGGG

4.1545	CACCCTTATTCTCACC GCC	CTTCAAGGCCTCAGGGCAAT
10.8141	CAGAGGTCAGGGGTTCATTCA	GAGCAGAGTGGACTGGGAAA
4.1495	CCTCAGTTCTCAGCCTGACC	GTGCCGCCCTGAAGTCTATG
X.8219	GGCTACTGTGGTGCCCATTA	ACTTGTGGGCAAAGGAGATCA
11.6894	ACTCCTCGAGAAAGGGGCT	TTAAGTGGGTGTTGCAGGGC
17.3512	AGGCCCTAGAGGAAGCCG	CCTGGCGAGAGAACCAGGAT
3.9051	TTCAGTGGAGCCCTTGTGAC	TGAGAATGGTAACCGAGCCG
5.1146	GCAGGGGGATCACAAGTTCA	ATGTGCAGTACCCATGGTGG
3.131	ATACTCCTCCCCACACGCAT	AACTACTCCGTCGTCTGGGA
5.6497	GGGCCATACCATCAGTGTCC	AGAGGCTTCTCGGAATGACC
11.7816	GTGGGAGAGGAAGAGGGGAT	TTTCTCCCTCCTCCCTCTGG
7.4476	TAGTCTGAGGGCCTACACCT	TGCTGAATGGCAGTTAGGAGG
5.1182	CGCCACCTGCGAGTGTT	ACTCCACTCGCCTAAACCTG
4.8325	CACTACAGGAGCCGGGATTC	CCCTCATGTCTGCTTGTGGT
9.1193	CCAACAAACAGTGGTCTGCT	AGTTGGTCGACAGCAAGCAT
7.1257	CACCTCGTCCAACAACGACT	CGGACCAATGGAGGAGGAAC
19.5886	GGAGCCTCCGGTGCTAAATC	TCCTGGTGCGCTAATGACAG
11.9847	TTGACCTTTGCCTTCACCGT	CCCTGTGAGATGAGGGATGC
11.1178	GGGCCTGTCCAGTCTTCTTC	AATGCCTTGAGCCCTGAGAG
17.5622	CCTGTCACCGCATCACTCAA	TAGTGGGAGTGGTCTGCACT
14.1657	CTCAGCGTGCAAACAATGCC	CTTGCGGAGTGCTTTTACGG
X.7750	AGGCCGAAAGGGATGTAGTG	CTGAAGGCGAACGCTTGTG
9.1195	ACCACTCAAGAGTCCTCCCTT	GGTGTGTGAACTCAGGCAGAT
11.9743	CCTGGGTGTTTACTTTGTGGC	CCCTCTGGGAATGGACAACA
15.7995	GTGACAGTCTGTGGCTTCCA	ACCCAAATGATCCAGGCAGG
15.9663	TCACCAGTCTGCACTCACAA	GAAAAGGGACCGGATAGGCT

6.5 RNA work

6.5.1 RNA extraction

Cells were dissociated in accutase, rinsed in PBS and resuspended in 0.5 – 1 mL per flask or 10-cm dish in TriReagent. Tissues were homogenised in 0.5 - 1mL TriReagent a MagNA Lyser instrument. 0.1mL of 1-bromo-3-chlopropane was added per mL of TriReagent used, mixed vigorously for 15 sec, incubated for 5 min at room temperature and centrifuged at 12000g for 15 min at 4°C. Aqueous supernatant was mixed with 0.2 mL of 2-propanol, mixed and incubated for 5 min at room temperature. RNA was pelleted by centrifugation at 12000g for 10 min at 4°C, washed with 75% ethanol twice, dried and dissolved in double distilled water. Concentration was estimated with a Biodrop and samples were used for DNaseI treatment or stored at -80°C.

6.5.2 DNaseI treatment

All RNA samples were treated with DNaseI to remove contaminating DNA. 2 µg were treated with DNaseI for 30 min at 37°C. Reaction was stopped with 50 mM EDTA at 65°C for 10 min.

6.5.3 cDNA synthesis

1 µg of DNaseI treated RNA was mixed with 1 µL of oligo d(T) and incubated at 65°C for 5 min. Samples were mixed with reaction buffer, RNase inhibitor, 10 mM dNTP and reverse transcriptase or water for the negative controls, incubated at 42°C for 1 h and 70°C for 5 min. RNaseH was added and incubated for 30 min at 37°C. cDNA were diluted 1/20 in water and stored at -20°C.

6.5.4 Quantitative PCR

qPCR was performed using SYBR green and 1uM primers final concentrations in a LightCycler 480. Technical replicates were performed for each sample. Genomic DNA or cDNA was diluted five times 1:5 for the standard curve. Analysis was performed with the Light Cycler program to determine the Ct values. Relative quantification using the standard curve was performed, and all values were corrected for the each primer amplification efficiency. Only primers with a standard efficiency of 2 plus or minus 20% were used. Melting curves were checked for a single sharp peak and for the presence of non-specific

amplification, contamination or primer dimers. Relative quantity values were calculated for both genes of interest and the housekeeping control, *β-actin*.

6.6 Animal work

All animal work, dissection and weighing is licenced by the UK Government Home Office (Number: PC9886123: expiry date February 2022). Mice are housed in an approved facility with controlled temperature, humidity, and light-dark cycle (07:00-19:00).

For RNA and protein extraction, tissues were snap frozen in liquid nitrogen following dissection and never allowed to thaw before extraction.

6.7 CRISPR-Cas9 zygote injections

6.7.1 gRNA design

The gRNAs were designed using the CRISPR MIT tool from the Zhang lab (crispr.mit.edu). They were synthesised following the GeneArt Precision gRNA kit protocol. 50ng/μL of gRNA were mixed with 42ng/μL Cas9 protein (Gene Art Platinum Cas9 nuclease)

6.7.2 Zygote injection

The zygote injection was performed at the Cambridge Stem Cell Facility. Briefly, superovulated female B6D2F1 mice were mated to B6D2F1 stud males. The gRNA and Cas9 mRNA/protein mix was injected into fertilized eggs, and 15–25 blastocysts were transferred into uterus of pseudopregnant females.

Chapter 7. References

- Aapola, U. et al., 2000. Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. *Genomics*, 65(3), pp.293–8.
- Abad, M. et al., 2013. Reprogramming in vivo produces teratomas and iPS cells with totipotency features. *Nature*, 502(7471), pp.340–5.
- Afgan, E. et al., 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1), pp.W3–W10.
- Alberts, B. et al., 2002. Chromosomal DNA and Its Packaging in the Chromatin Fiber.
- Allis, C.D. et al., 2015. *Epigenetics* Cold Sprin.,
- de Almeida, S.F. & Carmo-Fonseca, M., 2012. Design principles of interconnections between chromatin and pre-mRNA splicing. *Trends in Biochemical Sciences*, 37(6), pp.248–253.
- Amir, R.E. et al., 1999. Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, 23(2), pp.185–188.
- Anwar, S.L., Wulaningsih, W. & Lehmann, U., 2017. Transposable Elements in Human Cancer: Causes and Consequences of Dereglulation. *International Journal of Molecular Sciences*, 18(5).
- Azuara, V. et al., 2006. Chromatin signatures of pluripotent cell lines. *Nature Cell Biology*, 8(5), pp.532–538.
- Bailey, T. et al., 2013. Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLoS computational biology*, 9(11), p.e1003326.
- Barski, A. et al., 2007. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), pp.823–37.
- Bellefroid, E.J. et al., 1991. The evolutionarily conserved Krüppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 88(9), pp.3608–12.
- Benabdallah, N.S. et al., 2016. SBE6: a novel long-range enhancer involved in driving sonic hedgehog expression in neural progenitor cells. *Open Biology*, 6(11), p.160197.
- Berg, J.M., 1988. Proposed structure for the zinc-binding domains from transcription factor IIIA and related proteins. *Biophysics*, 85, pp.99–102.
- Berger, S.L., 2002. Histone modifications in transcriptional regulation. *Current Opinion in Genetics & Development*, 12(2), pp.142–148.
- Berger, S.L., 2007. The complex language of chromatin regulation during transcription. *Nature*, 447(7143), pp.407–412.
- Bernstein, B.E. et al., 2006. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. *Cell*, 125(2), pp.315–326.
- Bernstein, B.E., Meissner, A. & Lander, E.S., 2007. The Mammalian Epigenome. *Cell*, 128(4), pp.669–681.
- Bestor, T.H., 1992. Activation of mammalian DNA methyltransferase by cleavage of a Zn binding regulatory domain. *The EMBO journal*, 11(7), pp.2611–7.

- Bird, A., 2002. DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1), pp.6–21.
- Bostick, M. et al., 2007. UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells. *Science*, 317(5845), pp.1760–1764.
- Brown, G.R. et al., 2015. Gene: a gene-centered information resource at NCBI. *Nucleic Acids Research*, 43(D1), pp.D36–D42.
- Brownell, J.E. et al., 1996. Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, 84(6), pp.843–51.
- Busiello, T. et al., 2016. Role of ZNF224 in cell growth and chemoresistance of chronic lymphocytic leukemia. *Human Molecular Genetics*, 17(2), p.ddw427.
- Cammas, F. et al., 2004. Association of the transcriptional corepressor TIF1beta with heterochromatin protein 1 (HP1): an essential role for progression through differentiation. *Genes & development*, 18(17), pp.2147–60.
- Cammas, F. et al., 2002. Cell differentiation induces TIF1 β association with centromeric heterochromatin via an HP1 interaction. *Journal of Cell Science*, 115(17).
- Cammas, F. et al., 2000. Mice lacking the transcriptional corepressor TIF1beta are defective in early postimplantation development. *Development (Cambridge, England)*, 127(13), pp.2955–63.
- Carrozza, M.J. et al., 2005. Histone H3 Methylation by Set2 Directs Deacetylation of Coding Regions by Rpd3S to Suppress Spurious Intragenic Transcription. *Cell*, 123(4), pp.581–592.
- Castro-Diaz, N. et al., 2014. Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes & development*, 28(13), pp.1397–409.
- Collins, T. & Sander, T.L., 2005. The Superfamily of SCAN Domain Containing Zinc Finger Transcription Factors. In *Zinc Finger Proteins*. Boston, MA: Springer US, pp. 156–167.
- Collins, T., Stone, J.R. & Williams, A.M.Y.J., 2001. MINIREVIEW All in the Family : the BTB / POZ , KRAB , and SCAN Domains. , 21(11), pp.3609–3615.
- Corsinotti, A. et al., 2013. Global and stage specific patterns of Krüppel-associated-box zinc finger protein gene expression in murine early embryonic cells. C. J. Wilusz, ed. *PloS one*, 8(2), p.e56721.
- Creyghton, M.P. et al., 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), pp.21931–6.
- Daniel, J.M. et al., 2002. The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. *Nucleic acids research*, 30(13), pp.2911–9.
- Deaton, A.M. & Bird, A., 2011. CpG islands and the regulation of transcription. *Genes & Development*, 25(10), p.1010.
- Dixon, J.R., Gorkin, D.U. & Ren, B., 2016. Chromatin Domains: The Unit of Chromosome Organization. *Molecular cell*, 62(5), pp.668–80.
- Doudna, J.A. & Charpentier, E., 2014. The new frontier of genome engineering with

- CRISPR-Cas9. *Science*, 346(6213), pp.1258096–1258096.
- Dovey, O.M., Foster, C.T. & Cowley, S.M., 2010. Emphasizing the positive: A role for histone deacetylases in transcriptional activation. *Cell Cycle*, 9(14), pp.2700–2701.
- Ecco, G. et al., 2016. Transposable Elements and Their KRAB-ZFP Controllers
- Ecco, G., Imbeault, M. & Trono, D., 2017. KRAB zinc finger proteins. *Development*, 144(15), pp.2719–2729.
- Edelstein, L.C. & Collins, T., 2005. The SCAN domain family of zinc finger transcription factors. *Gene*, 359, pp.1–17.
- Edwards, C. a & Ferguson-Smith, A.C., 2007. Mechanisms regulating imprinted genes in clusters. *Current opinion in cell biology*, 19(3), pp.281–9.
- Ehrlich, M. et al., 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic acids research*, 10(8), pp.2709–21.
- Einhauer, A. & Jungbauer, A., 2001. The FLAGTM peptide, a versatile fusion tag for the purification of recombinant proteins. *Journal of Biochemical and Biophysical Methods*, 49(1–3), pp.455–465.
- Elbarbary, R.A., Lucas, B.A. & Maquat, L.E., 2016. Retrotransposons as regulators of gene expression. *Science*, 351(6274).
- Emerson, R.O. & Thomas, J.H., 2009. Adaptive evolution in zinc finger transcription factors. S. Myers, ed. *PLoS genetics*, 5(1), p.e1000325.
- Emerson, R.O. & Thomas, J.H., 2011. Gypsy and the Birth of the SCAN Domain. *Journal of Virology*, 85(22), pp.12043–12052.
- Escamilla-Del-Arenal, M. et al., 2013. Cdyl, a New Partner of the Inactive X Chromosome and Potential Reader of H3K27me3 and H3K9me2. *Molecular and Cellular Biology*, 33(24), pp.5005–5020.
- Fairall, L. et al., 1993. The Crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition. *Nature*, 336.
- Ferguson-Smith, A.C., 2011. Genomic imprinting: the emergence of an epigenetic paradigm. *Nature reviews. Genetics*, 12(8), pp.565–75.
- Feschotte, C., 2008. Transposable elements and the evolution of regulatory networks. *Nature reviews. Genetics*, 9(5), pp.397–405.
- Fischle, W., Wang, Y. & Allis, C.D., 2003. Histone and chromatin cross-talk. *Current Opinion in Cell Biology*, 15(2), pp.172–183.
- Flanagan, J.F. et al., 2005. Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature*, 438(7071), pp.1181–1185.
- Fort, A. et al., 2014. Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nature Genetics*, 46(6), pp.558–566.
- Frank, J.A. & Feschotte, C., 2017. Co-option of endogenous viral sequences for host cell function. *Current Opinion in Virology*, 25, pp.81–89. Friedli, M. & Trono, D., 2015. The Developmental Control of Transposable Elements and the Evolution of Higher Species. *Annual review of cell and developmental biology*.

- Friedman, J.R. et al., 1996. KAP-1, a novel corepressor for the highly conserved KRAB repression domain. *Genes & development*, 10(16), pp.2067–78.
- Frietze, S. et al., 2010. Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. *The Journal of biological chemistry*, 285(2), pp.1393–403.
- Furey, T.S., 2012. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nature reviews. Genetics*, 13(12), pp.840–52.
- García-García, M.J., Shibata, M. & Anderson, K. V, 2008. Chato, a KRAB zinc-finger protein, regulates convergent extension in the mouse embryo. *Development (Cambridge, England)*, 135(18), pp.3053–62.
- Garcia-Perez, J.L., Widmann, T.J. & Adams, I.R., 2016. The impact of transposable elements on mammalian development. *Development*, 143(22).
- Gardiner-Garden, M. & Frommer, M., 1987. CpG Islands in vertebrate genomes. *Journal of Molecular Biology*, 196(2), pp.261–282.
- Garton, M. et al., 2015. A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity. *Nucleic Acids Research*, 43(19), pp.9147–9157.
- Gates, L.A. et al., 2017. Acetylation on histone H3 lysine 9 mediates a switch from transcription initiation to elongation. *The Journal of biological chemistry*, 292(35), pp.14456–14472.
- Geen, H.O., Frietze, S. & Farnham, P.J., 2010. Using ChIP-seq Technology to Identify Targets of Zinc Finger Transcription Factors J. P. Mackay & D. J. Segal, eds. , 649, pp.437–455.
- Gifford, W.D., Pfaff, S.L. & Macfarlan, T.S., 2013. Transposable elements as genetic regulatory substrates in early development. *Trends in Cell Biology*, 23(5), pp.218–226.
- Gilbert, N., Gilchrist, S. & Bickmore, W.A., 2004. Chromatin Organization in the Mammalian Nucleus. , 242, pp.283–336.
- Goldberg, A.D., Allis, C.D. & Bernstein, B.E., 2007. Epigenetics: A Landscape Takes Shape. *Cell*, 128(4), pp.635–638.
- Goodarzi, A.A. et al., 2008. ATM Signaling Facilitates Repair of DNA Double-Strand Breaks Associated with Heterochromatin. *Molecular Cell*, 31(2), pp.167–177.
- Green, S.A., Simoes-Costa, M. & Bronner, M.E., 2015. Evolution of vertebrates as viewed from the crest. *Nature*, 520(7548), pp.474–482.
- Guindon, S. et al., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3), pp.307–321.
- Haberland, M., Montgomery, R.L. & Olson, E.N., 2009. The many roles of histone deacetylases in development and physiology: implications for disease and therapy. *Nature Reviews Genetics*, 10(1), pp.32–42.
- Hammond, C.M. et al., 2017. Histone chaperone networks shaping chromatin function. *Nature reviews. Molecular cell biology*, 18(3), pp.141–158. Hashimoto, H. et al., 2008. The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature*, 455(7214), pp.826–829.

- Hata, K. et al., 2002. Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. *Development (Cambridge, England)*, 129(8), pp.1983–93.
- He, Y.-F. et al., 2011. Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. *Science*, 333(6047), pp.1303–1307.
- Hendrich, B. et al., 1999. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*, 401(6750), pp.301–304.
- Hendrich, B. & Bird, A., 1998. Identification and characterization of a family of mammalian methyl-CpG binding proteins. *Molecular and cellular biology*, 18(11), pp.6538–47.
- Hermann, A., Goyal, R. & Jeltsch, A., 2004. The Dnmt1 DNA-(cytosine-C5)-methyltransferase Methylates DNA Processively with High Preference for Hemimethylated Target Sites. *Journal of Biological Chemistry*, 279(46), pp.48350–48359.
- Higashi, M., Inoue, S. & Ito, T., 2010. Core histone H2A ubiquitylation and transcriptional regulation. *Experimental Cell Research*, 316(17), pp.2707–2712.
- Hopp, T.P. et al., 1988. A Short Polypeptide Marker Sequence Useful for Recombinant Protein Identification and Purification. *Bio/Technology*, 6, pp.1204–1210.
- Hu, G. et al., 2009. A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. *Genes & development*, 23(7), pp.837–48.
- Huntley, S. et al., 2006. A comprehensive catalog of human KRAB-associated zinc finger genes: Insights into the evolutionary history of a large family of transcriptional repressors. *Genome Research*, 16(5), pp.669–677.
- Hurst, L.D., 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in Genetics*, 18(9), pp.486–487.
- Hurst, T.P. & Magiorkinis, G., 2017. Epigenetic Control of Human Endogenous Retrovirus Expression: Focus on Regulation of Long-Terminal Repeats (LTRs). *Viruses*, 9(6).
- Huttlin, E.L. et al., 2015. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, 162(2), pp.425–440.
- Imbeault, M., Helleboid, P.-Y. & Trono, D., 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, 543(7646), pp.550–554.
- Imbeault, M. & Trono, D., 2014. As Time Goes by: KRABs Evolve to KAP Endogenous Retroelements. *Developmental Cell*, 31(3), pp.257–258.
- Irisarri, I. et al., 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nature Ecology & Evolution*.
- Ito, S. et al., 2011. Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science*, 333(6047), pp.1300–1303.
- Itokawa, Y. et al., 2009. KAP1-independent transcriptional repression of SCAN-KRAB-containing zinc finger proteins. *Biochemical and biophysical research communications*, 388(4), pp.689–94.
- Ivanov, D. et al., 2005. Mammalian SCAN Domain Dimer Is a Domain-Swapped Homolog of the HIV Capsid C-Terminal Domain. *Molecular Cell*, 17(1), pp.137–143.

- Iyengar, S. et al., 2011. Functional Analysis of KAP1 Genomic Recruitment. *Molecular and Cellular Biology*, 31(9), pp.1833–1847.
- Iyengar, S. & Farnham, P.J., 2011. KAP1 Protein: An Enigmatic Master Regulator of the Genome. *Journal of Biological Chemistry*, 286(30), pp.26267–26276.
- Jacobs, F.M.J. et al., 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature*.
- Jenuwein, T. et al., 1998. SET domain proteins modulate chromatin domains in eu- and heterochromatin. *Cellular and molecular life sciences : CMLS*, 54(1), pp.80–93.
- Jia, D. et al., 2007. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature*, 449(7159), pp.248–51.
- Karmodiya, K. et al., 2012. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics*, 13(1), p.424.
- Karolchik, D. et al., 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(90001), p.493D–496.
- Kleff, S. et al., 1995. Identification of a gene encoding a yeast histone H4 acetyltransferase. *The Journal of biological chemistry*, 270(42), pp.24674–7.
- Klug, A., 2010. The Discovery of Zinc Fingers and Their Applications in Gene Regulation and Genome Manipulation.
- Klug, A. & Schwabe, J.W., 1995. Protein motifs 5. Zinc fingers. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 9(8), pp.597–604.
- Kornberg, R.D., 1974. Chromatin structure: a repeating unit of histones and DNA. *Science (New York, N.Y.)*, 184(4139), pp.868–71.
- Kornberg, R.D., 1977. Structure of Chromatin. *Annual Review of Biochemistry*, 46(1), pp.931–954.
- Kuzmichev, A. et al., 2002. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes & Development*, 16(22), pp.2893–2905.
- Landt, S.G. et al., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9), pp.1813–31.
- Lechner, M.S. et al., 2000. Molecular determinants for targeting heterochromatin protein 1-mediated gene silencing: direct chromoshadow domain-KAP-1 corepressor interaction is essential. *Molecular and cellular biology*, 20(17), pp.6449–65.
- Lee, T.I. & Young, R.A., 2000. Transcription of Eukaryotic Protein-Coding Genes. *Annual Review of Genetics*, 34(1), pp.77–137.
- Lee, T.I. & Young, R.A., 2013. Transcriptional regulation and its misregulation in disease. *Cell*, 152(6), pp.1237–51.
- Lehnertz, B. et al., 2003. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. *Current biology : CB*, 13(14), pp.1192–200.

- Lewis, J.D. et al., 1992. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*, 69(6), pp.905–14.
- Li, B., Carey, M. & Workman, J.L., 2007. The Role of Chromatin during Transcription. *Cell*, 128(4), pp.707–719.
- Li, H. et al., 2006. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature*, 442(7098), pp.91–5.
- Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), pp.1754–60.
- Li, R. et al., 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5), pp.713–714.
- Li, R. et al., 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15), pp.1966–1967.
- Li, W.H., Wu, C.I. & Luo, C.C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution*, 2(2), pp.150–174.
- Li, X. et al., 2008. A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. *Developmental cell*, 15(4), pp.547–57.
- Liang, Y. et al., 2012. Structural analysis and dimerization profile of the SCAN domain of the pluripotency factor Zfp206. *Nucleic Acids Research*, 40(17), pp.8721–8732.
- Liu, C.-M. et al., 2012. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics*, 28(6), pp.878–879.
- Liu, H. et al., 2014. Deep Vertebrate Roots for Mammalian Zinc Finger Transcription Factor Subfamilies. *Genome Biology and Evolution*, 6(3), pp.510–525.
- Liu, Y. et al., 2012. An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence. *Genes & development*, 26(21), pp.2374–9.
- Löwer, R., Löwer, J. & Kurth, R., 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 93(11), pp.5177–84.
- Luco, R.F. et al., 2010. Regulation of alternative splicing by histone modifications. *Science (New York, N.Y.)*, 327(5968), pp.996–1000.
- Lupo, A. et al., 2013. KRAB-Zinc Finger Proteins: A Repressor Family Displaying Multiple Biological Functions. *Current genomics*, 14(4), pp.268–78.
- Ma, H. et al., 2014. Zfp322a Regulates mouse ES cell pluripotency and enhances reprogramming efficiency. *PLoS genetics*, 10(2), p.e1004038.
- Mackay, D.J.G. et al., 2008. Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57. *Nature genetics*, 40(8), pp.949–51.
- Margolin, J.F. et al., 1994. Kruppel-associated boxes are potent transcriptional repression domains. *Biochemistry*, 91, pp.4509–4513.
- Martienssen, R., 1998. Chromosomal imprinting in plants. *Current opinion in genetics & development*, 8(2), pp.240–4.

- Mayer, W. et al., 2000. Demethylation of the zygotic paternal genome. *Nature*, 403(6769), pp.501–2.
- Meehan, R.R. et al., 1989. Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs. *Cell*, 58(3), pp.499–507.
- Messerschmidt, D.M. et al., 2012. Trim28 is required for epigenetic stability during mouse oocyte to embryo transition. *Science (New York, N.Y.)*, 335(6075), pp.1499–502.
- Mi, H. et al., 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1), pp.D183–D189.
- Millar, C.B. et al., 2002. Enhanced CpG Mutability and Tumorigenesis in MBD4-Deficient Mice. *Science*, 297(5580), pp.403–405.
- Miller, J., McLachlan, A.D. & Klug, A., 1985. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *The EMBO journal*, 4(6), pp.1609–14.
- Minsky, N. et al., 2008. Monoubiquitinated H2B is associated with the transcribed region of highly expressed genes in human cells. *Nature Cell Biology*, 10(4), pp.483–488.
- Moore, L.D., Le, T. & Fan, G., 2013. DNA methylation and its basic function. *Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology*, 38(1), pp.23–38.
- Morgan, H.D. et al., 2004. Activation-induced Cytidine Deaminase Deaminates 5-Methylcytosine in DNA and Is Expressed in Pluripotent Tissues. *Journal of Biological Chemistry*, 279(50), pp.52353–52360.
- Morgan, H.D. et al., 2005. Epigenetic reprogramming in mammals. *Human molecular genetics*, 14 Spec No(suppl_1), pp.R47–58.
- Najafabadi, H.S. et al., 2015. C2H2 zinc finger proteins greatly expand the human regulatory lexicon. *Nature biotechnology*, 33(5), pp.555–562.
- Nam, K., Honer, C. & Schumacher, C., 2004. Structural components of SCAN-domain dimerizations. *Proteins: Structure, Function, and Bioinformatics*, 56(4), pp.685–692.
- Nan, X. et al., 1998. Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature*, 393(6683), pp.386–389.
- Nan, X., Meehan, R.R. & Bird, A., 1993. Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2. *Nucleic acids research*, 21(21), pp.4886–92.
- Nätt, D. et al., 2015. High cortisol in 5-year-old children causes loss of DNA methylation in SINE retrotransposons: a possible role for ZNF263 in stress-related diseases. *Clinical epigenetics*, 7(1), p.91.
- NCBI Resource Coordinators, 2017. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 45(D1), pp.D12–D17.
- Ng, H.-H. et al., 1999. MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nature Genetics*, 23(1), pp.58–61.
- Nishioka, K. et al., 2002. Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation. *Genes & Development*, 16(4), pp.479–489.

- O'Geen, H. et al., 2007. Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS genetics*, 3(6), p.e89.
- Ohlsson, R., Renkawitz, R. & Lobanenkov, V., 2001. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends in Genetics*, 17(9), pp.520–527.
- Okano, M. et al., 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3), pp.247–57
- Osley, M.A., 2006. Regulation of histone H2A and H2B ubiquitylation. *Briefings in Functional Genomics and Proteomics*, 5(3), pp.179–189.
- Oswald, J. et al., 2000. Active demethylation of the paternal genome in the mouse zygote. *Current biology: CB*, 10(8), pp.475–8.
- Pavletich, N. & Pabo, C., 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science*, 252(5007).
- Pepke, S., Wold, B. & Mortazavi, A., 2009. Computation for ChIP-seq and RNA-seq studies. *Nature Methods Supplement*, 6(11).
- Pradeepa, M.M. et al., 2016. Histone H3 globular domain acetylation identifies a new class of enhancers. *Nature Genetics*, 48(6), pp.681–686.
- Pradhan, S. et al., 1999. Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *The Journal of biological chemistry*, 274(46), pp.33002–10.
- Quenneville, S. et al., 2011. In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions. *Molecular cell*, 44(3), pp.361–72.
- Razin, S. V & Ulianov, S. V, 2017. Gene functioning and storage within a folded genome. *Cellular & molecular biology letters*, 22, p.18.
- Rhodes, D. & Klug, A., 1993. Zinc fingers. *Scientific American*, 268(2), pp.56–9, 62–5. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/8430296> [Accessed April 4, 2018].
- Roberts, S.G.E., 2000. Mechanisms of action of transcription activation and repression domains. *Cellular and Molecular Life Sciences*, 57(8), pp.1149–1160.
- Rougeulle, C. et al., 2004. Differential Histone H3 Lys-9 and Lys-27 Methylation Profiles on the X Chromosome. *Molecular and Cellular Biology*, 24(12), pp.5475–5484.
- Rougier, N. et al., 1998. Chromosome methylation patterns during mammalian preimplantation development. *Genes & development*, 12(14), pp.2108–13.
- Rowe, H.M. et al., 2010. KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*, 463(7278), pp.237–40.
- Rowe, H.M. et al., 2013. TRIM28 repression of retrotransposon-based enhancers is necessary to preserve transcriptional dynamics in embryonic stem cells. *Genome research*, 23(3), pp.452–61.
- Rowe, H.M. & Trono, D., 2011. Dynamic control of endogenous retroviruses during development. *Virology*, 411(2), pp.273–87.
- Ruthenburg, A.J., Allis, C.D. & Wysocka, J., 2007. Methylation of lysine 4 on histone H3:

- intricacy of writing and reading a single epigenetic mark. *Molecular cell*, 25(1), pp.15–30.
- Sander, T.L. et al., 2003. The SCAN domain defines a large family of zinc finger transcription factors. *Gene*, 310, pp.29–38.
- Schlesinger, S. & Goff, S.P., 2014. Retroviral Transcriptional Regulation and Embryonic Stem Cells: War and Peace. *Molecular and cellular biology*.
- Schmitges, F.W. et al., 2016. Multiparameter functional diversity of human C2H2 zinc finger proteins. *Genome research*, 26(12), pp.1742–1752.
- Schneider, R. et al., 2004. Direct Binding of INHAT to H3 Tails Disrupted by Modifications. *Journal of Biological Chemistry*, 279(23), pp.23859–23862.
- Schultz, D.C. et al., 2002. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes & development*, 16(8), pp.919–32.
- Schultz, D.C., Friedman, J.R. & Rauscher Iii, F.J., 2000. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2_α subunit of NuRD. *Genes and Development*, 15:428-443.
- Schumacher, C. et al., 2000. The SCAN Domain Mediates Selective Oligomerization. *Journal of Biological Chemistry*, 275(22), pp.17173–17179.
- Schwabe, J.W.R. & Klug, A., 1994. Zinc mining for protein domains. *Nature Structural & Molecular Biology*, 1(6), pp.345–349.
- Shao, M. et al., 2017. Fetal development of subcutaneous white adipose tissue is dependent on Zfp423. *Molecular metabolism*, 6(1), pp.111–124.
- Sharif, J. et al., 2007. The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, 450(7171), pp.908–912.
- Shen, Y. et al., 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409), pp.116–120.
- Shi, Y. et al., 2004. Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell*, 119(7), pp.941–953.
- Shi, Y.G. & Tsukada, Y., 2013. *The discovery of histone demethylases* Cold Spring.,
- Shibata, M. et al., 2011. TRIM28 is required by the mouse KRAB domain protein ZFP568 to control convergent extension and morphogenesis of extra-embryonic tissues. *Development (Cambridge, England)*, 138(24), pp.5333–43.
- Shibata, M. & García-García, M.J., 2011. The mouse KRAB zinc-finger protein CHATO is required in embryonic-derived tissues to control yolk sac and placenta morphogenesis. *Developmental biology*, 349(2), pp.331–41.
- Shinkai, Y. & Tachibana, M., 2011. H3K9 methyltransferase G9a and the related molecule GLP. *Genes & Development*, 25(8), pp.781–788.
- Shoemaker, C.J. & Green, R., 2012. Translation drives mRNA quality control. *Nature Structural & Molecular Biology*, 19(6), pp.594–601.

- Sievers, F. et al., 2014. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1), pp.539–539.
- Spencer, T.E. et al., 1997. Steroid receptor coactivator-1 is a histone acetyltransferase. *Nature*, 389(6647), pp.194–8.
- Squatrito, M., Gorrini, C. & Amati, B., 2006. Tip60 in DNA damage response and growth control: many tricks in one HAT. *Trends in Cell Biology*, 16(9), pp.433–442.
- Stothard, P., 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques*, 28(6), pp.1102, 1104.
- Strahl, B.D. & Allis, C.D., 2000. The language of covalent histone modifications. *Nature*, 403(6765), pp.41–45.
- Strogantsev, R. et al., 2015. Allele-specific binding of ZFP57 in the epigenetic regulation of imprinted and non-imprinted monoallelic expression. *Genome biology*, 16(1), p.112.
- Strogantsev, R. & Ferguson-Smith, A.C., 2012. Proteins involved in establishment and maintenance of imprinted methylation marks. *Briefings in functional genomics*, 11(3), pp.227–39.
- Suetake, I. et al., 2004. DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. *The Journal of biological chemistry*, 279(26), pp.27816–23.
- Sun, B. et al., 2012. Status of Genomic Imprinting in Epigenetically Distinct Pluripotent Stem Cells. *STEM CELLS*, 30(2), pp.161–168.
- Surani, M. a, 1998. Imprinting and the initiation of gene silencing in the germ line. *Cell*, 93(3), pp.309–12.
- Tahiliani, M. et al., 2009. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science (New York, N.Y.)*, 324(5929), pp.930–5.
- Takahashi, N. et al., 2015. ZFP57 and the Targeted Maintenance of Postfertilization Genomic Imprints. *Cold Spring Harbor symposia on quantitative biology*, 80, pp.177–87.
- Thomas, J.H. & Schneider, S., 2011. Coevolution of retroelements and tandem zinc finger genes. *Genome research*, 21(11), pp.1800–12.
- Thomas, J.O. & Kornberg, R.D., 1975. An octamer of histones in chromatin and free in solution. *Proceedings of the National Academy of Sciences of the United States of America*, 72(7), pp.2626–30.
- Tie, F. et al., 2009. CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. *Development*, 136(18), pp.3131–3141.
- Turelli, P. et al., 2014. Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. *Genome research*, 24(8), pp.1260–70.
- Turner, J. & Crossley, M., 1999. Mammalian Krüppel-like transcription factors: more than just a pretty finger. *Trends in Biochemical Sciences*, 24(6), pp.236–240.
- Uhlen, M. et al., 2010. Towards a knowledge-based Human Protein Atlas. *Nature biotechnology*, 28(12), pp.1248–50.
- Urrutia, R., 2003. Protein family review KRAB-containing zinc-finger repressor proteins.

- Valinluck, V. & Sowers, L.C., 2007. Endogenous Cytosine Damage Products Alter the Site Selectivity of Human DNA Maintenance Methyltransferase DNMT1. *Cancer Research*, 67(3), pp.946–950.
- Varriale, A. & Annalisa, 2014. DNA methylation, epigenetics, and evolution in vertebrates: facts and challenges. *International journal of evolutionary biology*, 2014, p.475981.
- Venturini, L. et al., 1999. TIF1gamma, a novel member of the transcriptional intermediary factor 1 family. *Oncogene*, 18(5), pp.1209–17.
- Voigt, P., Tee, W.-W. & Reinberg, D., 2013. A double take on bivalent promoters. *Genes & Development*, 27(12), pp.1318–1338.
- Wagner, E.J. & Carpenter, P.B., 2012. Understanding the language of Lys36 methylation at histone H3. *Nature Reviews Molecular Cell Biology*, 13(2), pp.115–126.
- Wang, Z. et al., 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7), pp.897–903.
- Weinreich, M., Palacios DeBeer, M.A. & Fox, C.A., 2004. The activities of eukaryotic replication origins in chromatin. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, 1677(1–3), pp.142–157.
- Wicker, T. et al., 2007. A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics*, 8(12), pp.973–82.
- Williams, A.J. et al., 1995. Isolation and characterization of a novel zinc-finger protein with transcription repressor activity. *The Journal of biological chemistry*, 270(38), pp.22143–52.
- Williams, A.J., Blacklow, S.C. & Collins, T., 1999. The zinc finger-associated SCAN box is a conserved oligomerization domain. *Molecular and cellular biology*, 19(12), pp.8526–35.
- Witzgall, R. et al., 1994. The Krüppel-associated box-A (KRAB-A) domain of zinc finger proteins mediates transcriptional repression. *Proceedings of the National Academy of Sciences of the United States of America*, 91(10), pp.4514–8.
- Wolf, D. & Goff, S.P., 2007. TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. *Cell*, 131(1), pp.46–57.
- Wolf, G. et al., 2015. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes & development*, 29(5), pp.538–54.
- Wolfe, S.A., Nekludova, L. & Pabo, C.O., 2000. DNA Recognition by Cys 2 His 2 Zinc Finger Proteins. *Annual Review of Biophysics and Biomolecular Structure*, 29(1), pp.183–212.
- Wu, C. -t. & Morris, J.R., 2001. Genes, Genetics, and Epigenetics: A Correspondence. *Science*, 293(5532), pp.1103–1105.
- Wyrick, J.J. & Parra, M.A., 2009. The role of histone H2A and H2B post-translational modifications in transcription: A genomic perspective. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(1), pp.37–44.
- Xie, S. et al., 1999. Cloning, expression and chromosome locations of the human DNMT3 gene family. *Gene*, 236(1), pp.87–95.
- Yang, H., Wang, H. & Jaenisch, R., 2014. Generating genetically modified mice using CRISPR/Cas-mediated genome engineering. *Nature protocols*, 9(8), pp.1956–68.

Yates, A. et al., 2016. Ensembl 2016. *Nucleic Acids Research*, 44(D1), pp.D710–D716.

Zamudio, N. et al., 2015. DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. *Genes & development*, 29(12), pp.1256–70.









Zhu, Z. et al., 2017. ZFP403, a novel tumor suppressor, inhibits the proliferation and metastasis in ovarian cancer. *Gynecologic Oncology*.

Zuckerandl, E. & Cavalli, G., 2007. Combinatorial epigenetics, “junk DNA”, and the evolution of complex organisms. *Gene*, 390(1–2), pp.232–242.

8. Chapter 8: Appendix

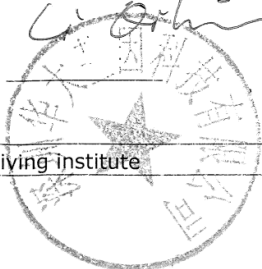
8.1 Appendix Chapter 1

Appendix 8.1.1: Summary of the 8 initial KZFP candidates in addition to ZFP263. Dr Noon selected these candidates based on their high expression in mESCs and lower expression in differentiates cells. The table summarises the structure of the protein and presence or absence of additional domains (KRAB and SCAN domains). The conservation across species is given as well as a very brief description of their functions if known.

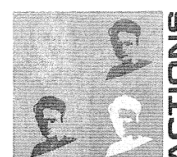
ZFP	structure	conservation	Literature
ZFP472	59 kDa KRAB box 12 ZFs 	Conserved in rodents only	- ZFP472 orthologue in rat, KRIM1, can form a complex with c-Myc and negatively regulates Myc functions (Hennemann et al. 2003).
ZFP568	74 kDa KRAB box 11 ZFs 	Conserved in Eutheria Orthologues in 2 species of fish, the anole lizard and the Chinese softshell turtle.	- Regulates convergent extension in the mouse embryo - Required in embryonic-derived tissue for yolk sac and placenta morphogenesis (García-García et al. 2008; Shibata & García-García 2011; Shibata et al. 2011). - Represses placental-specific Igf2 transcript (Yang et al. 2017).
ZFP655	60/20 kDa 7 ZFs KRAB box 	Conserved in Eutheria Absent in birds, reptiles and fish	- ZFP655 human orthologue is proposed to be involved in cell-cycle regulation (Houlard et al. 2005).
ZFP599	60 kDa KRAB box 11 ZFs 	Conserved in some primates and rodents	/
ZFP459	29 kDa KRAB box 4 ZFs 	Conserved in some primates and rodents	- Described as a potential marker for immunosuppression as its expression increased in mice treated with immunosuppressive chemicals (Oshida et al. 2011)
ZFP110	91 kDa KRAB box SCAN box 5 ZFs 	Conserved in some primates, rodents and <u>Laurasatheria</u>	- Transcription regulator required for p75-mediated apoptosis of sympathetic neurons and a potential regulator of neuronal cholesterol biosynthesis genes (Korade et al. 2009; Linggi et al. 2005; Kendall et al. 2003; Geetha et al. 2005).
ZFP809	55 kDa KRAB box 7 ZFs 	Conserved in some primates and rodents	- Repress murine leukemia viruses in mESCs - Involved in ERV silencing early in development, (Wolf & Goff 2007; Wolf et al. 2015).
ZFP759	79 kDa KRAB box 22 ZFs 	Conserved in some primates, rodents and <u>Laurasatheria</u>	/

Appendix 8.1.2: Official certificate of a 6-week secondment in the Beijing Genomic Institute, Shenzhen, China

Category 3, Research and Training, Marie Curie ITN; Official certificate

Name Fellow	Celia Delahaye
ESR/ ER	ESR5
Host institute	University of Cambridge, Department of Genetics, Anne Ferguson-Smith's lab
Receiving institute	BGI tech-Shenzhen
Supervisor	Pr Li Qibin
Period	Sept-Oct 2014
Duration	6 weeks
Sort of training / Research	Bioinformatics training
Main Subject	Analysis of ChIP-seq data (Chromatin immunoprecipitation followed by sequencing)
Summary of activities	<p>The main goal was to analyse data from 2 ChIP-seq experiments. The first replicate had already been analysed in Cambridge and was analysed as a control to learn the different steps of the pipeline. The data from the second replicate was generated during my secondment, analysed in BGI, and compared to the first set of results.</p> <p>I also had the opportunity to attend a training on writing a scientific manuscript for non-English speakers, organised by BGI-Shenzhen.</p>
What did you learn	<p>I first learnt the basic commands on Linux.</p> <p>I also learnt the different steps to analyse ChIP-seq data and how to set the parameters.</p> <p>Finally I learnt how to use the analysis pipeline developed by BGI.</p>
Achievements	<p>During my secondment, I managed to generate the standard analysis for the 2 ChIP-seq experiments.</p> <p>I started the comparison between the 2 sets of results, but further analysis need to be done in Cambridge to complete this work</p>
Date <u>03/11/14</u> Name <u>Anne Ferguson-Smith</u> Fellow Signature <u>[Signature]</u>	Date <u>29/10/14</u> Name <u>Li Qibin</u> Supervisor Signature <u>[Signature]</u> 
Logo host institute	Logo receiving institute

Department of Physiology
University of Cambridge
Downing Street
Cambridge CB2 3EG



Appendix 8.1.3: Official certificate of a 3-week secondment in the Celgene Institute for Translational Research Europe, Seville, Spain

Category 3, Research and Training, Marie Curie ITN; Official certificate

Name Fellow	Celia Delahaye
ESR/ ER	ESR5
Host institute	University of Cambridge, Department of Genetics, Anne Ferguson-Smith's lab
Receiving institute	Celgene Institute for Translational Research Europe, Seville, Spain
Supervisor	Dr Matthew Trotter, Dr Remco Loos
Period	Sept 2016
Duration	3 weeks
Sort of training / Research	Bioinformatics training
Main Subject	Analysis of transcriptome data with R



<p>Summary of activities</p>	<p>The main goal of the secondment was to learn how to analyse data with R, which is a free software environment for statistical computing and graphics. I focused the first days on learning the basic commands with an on-line tutorial (https://www.datacamp.com/home). After an introduction to the R language, I learnt how to import and structure data, and how to use packages and functions to interpret the results.</p> <p>I focused in particular on using the package gplots to draw heatmaps and extracts clusters, on correlation analysis, and on gene enrichment methods, such as Enrichr, GSEA and clusterProfiler.</p> <p>My main project, apart from learning computational techniques, focused on the expression of human KRAB Zinc Finger Proteins (ZFPs) in the context of cancer, in particular in acute myeloid leukemia (AML). Transcriptome data from a cohort of patients with AML is publically available and has been use in Celgene (Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia, The Cancer Genome Atlas Research Network, 2013). This study identified 7 clusters of patients with distinct gene expression profiles. They found association between these clusters and overall survival, association with subtypes currently in use for diagnosis and treatment (FAB subtypes), and correlation between some expression signatures and the stage of myeloid differentiation</p> <p>I first looked at ZNF263 expression in these patients. ZNF263 is the human ortholog of Zfp263, the gene I am focused on for my PhD project. I have noticed that its expression decreased in 2 clusters, while it was stable in the others. I wondered whether other ZFPs behaved the same way, but I could not find any strong correlations. Therefore I looked at other genes and found 656 genes correlating with ZNF263. The genes that are positively-correlated are mostly ZFPs, involved in regulation of transcription. The gene anti-correlated with ZNF263 are significantly enriched in pathways involved in innate and adaptive immune system, hemostasis and response to infection. Further analysis would be needed to interpret these results in more details and understand their biological meaning.</p> <p>Finally, I had a chance to have an overview of the main differences between working in industry and academia. I attended CITRE group meetings and discussed with Matthew Trotter and the rest of the team to understand the characteristics and challenges of research in industry.</p>
------------------------------	--



What did you learn	<p>I learnt how to use R to analyse large dataset: use of factors, lists, dataframes, loops, functions, importing dataset, data visualization with gplots, gene set enrichment with clusterProfiler.</p> <p>With the AML dataset, I have learnt how to structure data and extract information to answer biological questions. I gained insight into the behaviour of KRAB ZFPs in acute myeloid dataset and in the genes correlating with human ZNF263. I will now be able to analyse other sort of data, and will be more comfortable to use other packages and functions in R.</p> <p>Finally, being in a biotech company for 3 weeks allowed me to appreciate the distinctions between industry and academia, the different demands and motivations.</p>
Achievements	<p>In three weeks, I have acquired the basic knowledge about R and analysed the expression of KRAB ZFP genes in transcriptome data. Studying human ZNF263 expression in leukemia patients was the opportunity for me to consider the relevance of my project in a disease context, and gave me the main keys and tools to analyse data with R. I should be able to put into practice this learning to other publically available data.</p>
Date <u>27/10/2016</u> Name <u>C. Delahaye</u> Fellow Signature <u>[Signature]</u>	Date <u>30-9-2016</u> Name <u>Renico Loos</u> Supervisor Signature <u>[Signature]</u>
Logo host institute	Logo receiving institute




8.2 Appendix Chapter 2

Appendix 8.2.1: Coding DNA sequence alignment of orthologues to ZNF263 from 14 species.

Ha_Human	:	-----	20	40	60	80	100	120	140	160	:	151
Pt_Chimpan	:	-----	20	40	60	80	100	120	140	160	:	151
Gg_Gorilla	:	-----	20	40	60	80	100	120	140	160	:	151
Fa_Orangut	:	-----	20	40	60	80	100	120	140	160	:	151
Cj_Marmose	:	-----	20	40	60	80	100	120	140	160	:	151
Rn_Rat	:	-----	20	40	60	80	100	120	140	160	:	157
Mm_Mouse	:	-----	20	40	60	80	100	120	140	160	:	157
Cf_Dog	:	-----	20	40	60	80	100	120	140	160	:	157
Ec_Horse	:	-----	20	40	60	80	100	120	140	160	:	157
Bt_Cow	:	-----	20	40	60	80	100	120	140	160	:	157
La_elephan	:	-----	20	40	60	80	100	120	140	160	:	157
Ch_Sloth	:	-----	20	40	60	80	100	120	140	160	:	151
Md_Opossum	:	-----	20	40	60	80	100	120	140	160	:	169
Me_Wallaby	:	-----	20	40	60	80	100	120	140	160	:	130
Oa_Platytyp	:	-----	20	40	60	80	100	120	140	160	:	145
atggc cggcgc ggtccctc gaga c gaggctctcggc tct aa tggcagga gactg c tggag caggag t cc c cc ga cc c cc agccc gaa cctccc tggccttc gagggtccctcga g												
Ha_Human	:	-----	180	200	220	240	260	280	300	320	:	320
Pt_Chimpan	:	-----	180	200	220	240	260	280	300	320	:	320
Gg_Gorilla	:	-----	180	200	220	240	260	280	300	320	:	320
Fa_Orangut	:	-----	180	200	220	240	260	280	300	320	:	320
Cj_Marmose	:	-----	180	200	220	240	260	280	300	320	:	320
Rn_Rat	:	-----	180	200	220	240	260	280	300	320	:	326
Mm_Mouse	:	-----	180	200	220	240	260	280	300	320	:	326
Cf_Dog	:	-----	180	200	220	240	260	280	300	320	:	326
Ec_Horse	:	-----	180	200	220	240	260	280	300	320	:	326
Bt_Cow	:	-----	180	200	220	240	260	280	300	320	:	326
La_elephan	:	-----	180	200	220	240	260	280	300	320	:	326
Ch_Sloth	:	-----	180	200	220	240	260	280	300	320	:	320
Md_Opossum	:	-----	180	200	220	240	260	280	300	320	:	328
Me_Wallaby	:	-----	180	200	220	240	260	280	300	320	:	314
Oa_Platytyp	:	-----	180	200	220	240	260	280	300	320	:	299
agac ctcc cccc ga cctccagctggg ca gacitctcga ggggttc cgcct cagatggcga aagcagcagac tggagctgctgtgtgc agt cctccc cccagagatcagcagcagctgacga g												
Ha_Human	:	-----	360	380	400	420	440	460	480	500	:	489
Pt_Chimpan	:	-----	360	380	400	420	440	460	480	500	:	489
Gg_Gorilla	:	-----	360	380	400	420	440	460	480	500	:	489
Fa_Orangut	:	-----	360	380	400	420	440	460	480	500	:	489
Cj_Marmose	:	-----	360	380	400	420	440	460	480	500	:	489
Rn_Rat	:	-----	360	380	400	420	440	460	480	500	:	495
Mm_Mouse	:	-----	360	380	400	420	440	460	480	500	:	495
Cf_Dog	:	-----	360	380	400	420	440	460	480	500	:	495
Ec_Horse	:	-----	360	380	400	420	440	460	480	500	:	495
Bt_Cow	:	-----	360	380	400	420	440	460	480	500	:	495
La_elephan	:	-----	360	380	400	420	440	460	480	500	:	495
Ch_Sloth	:	-----	360	380	400	420	440	460	480	500	:	489
Md_Opossum	:	-----	360	380	400	420	440	460	480	500	:	507
Me_Wallaby	:	-----	360	380	400	420	440	460	480	500	:	483
Oa_Platytyp	:	-----	360	380	400	420	440	460	480	500	:	468
gAgcgcCA GAAG Gt AC CTtGcGAGgataTcAGagaga CttggAGAc gaa CaaCAGGTcCaAaaCAtggcGgGca c GaagTcTtTtgAGAGcCttTgCC CTGaaAaCgAc agATcCcaagCTtCAAGcGAGcTgagact												
Ha_Human	:	-----	520	540	560	580	600	620	640	660	:	645
Pt_Chimpan	:	-----	520	540	560	580	600	620	640	660	:	645
Gg_Gorilla	:	-----	520	540	560	580	600	620	640	660	:	645
Fa_Orangut	:	-----	520	540	560	580	600	620	640	660	:	645
Cj_Marmose	:	-----	520	540	560	580	600	620	640	660	:	642
Rn_Rat	:	-----	520	540	560	580	600	620	640	660	:	651
Mm_Mouse	:	-----	520	540	560	580	600	620	640	660	:	651
Cf_Dog	:	-----	520	540	560	580	600	620	640	660	:	651
Ec_Horse	:	-----	520	540	560	580	600	620	640	660	:	651
Bt_Cow	:	-----	520	540	560	580	600	620	640	660	:	651
La_elephan	:	-----	520	540	560	580	600	620	640	660	:	651
Ch_Sloth	:	-----	520	540	560	580	600	620	640	660	:	666
Md_Opossum	:	-----	520	540	560	580	600	620	640	660	:	666
Me_Wallaby	:	-----	520	540	560	580	600	620	640	660	:	642
Oa_Platytyp	:	-----	520	540	560	580	600	620	640	660	:	611
gAGcgaagccctggccccaagctg caggagctcctgagcctcc GcCCc AaagGgacCCcCa cTgtcTAaAGgAGAgcact tct cctccc tgg ttttctt tttct c a g a c gaaga aa ga a gag ggc cccagttg												
Ha_Human	:	-----	680	700	720	740	760	780	800	820	:	814
Pt_Chimpan	:	-----	680	700	720	740	760	780	800	820	:	814
Gg_Gorilla	:	-----	680	700	720	740	760	780	800	820	:	814
Fa_Orangut	:	-----	680	700	720	740	760	780	800	820	:	814
Cj_Marmose	:	-----	680	700	720	740	760	780	800	820	:	811
Rn_Rat	:	-----	680	700	720	740	760	780	800	820	:	817
Mm_Mouse	:	-----	680	700	720	740	760	780	800	820	:	817
Cf_Dog	:	-----	680	700	720	740	760	780	800	820	:	814
Ec_Horse	:	-----	680	700	720	740	760	780	800	820	:	817
Bt_Cow	:	-----	680	700	720	740	760	780	800	820	:	817
La_elephan	:	-----	680	700	720	740	760	780	800	820	:	820
Ch_Sloth	:	-----	680	700	720	740	760	780	800	820	:	814
Md_Opossum	:	-----	680	700	720	740	760	780	800	820	:	832
Me_Wallaby	:	-----	680	700	720	740	760	780	800	820	:	808
Oa_Platytyp	:	-----	680	700	720	740	760	780	800	820	:	756
cctg gaggctt GAgagctGgc tctac tcttccAGaggg tgggg CatCagga cctagtTAAGAgggcCct TccagGga a gTgCAGAGa tTatgagAAtgtg c cact GagtG tTccagctCa ga g c gaacca gtgg ac												
Ha_Human	:	-----	860	880	900	920	940	960	980	1000	:	974
Pt_Chimpan	:	-----	860	880	900	920	940	960	980	1000	:	974
Gg_Gorilla	:	-----	860	880	900	920	940	960	980	1000	:	974
Fa_Orangut	:	-----	860	880	900	920	940	960	980	1000	:	974
Cj_Marmose	:	-----	860	880	900	920	940	960	980	1000	:	974
Rn_Rat	:	-----	860	880	900	920	940	960	980	1000	:	980
Mm_Mouse	:	-----	860	880	900	920	940	960	980	1000	:	980
Cf_Dog	:	-----	860	880	900	920	940	960	980	1000	:	977
Ec_Horse	:	-----	860	880	900	920	940	960	980	1000	:	980
Bt_Cow	:	-----	860	880	900	920	940	960	980	1000	:	980
La_elephan	:	-----	860	880	900	920	940	960	980	1000	:	983
Ch_Sloth	:	-----	860	880	900	920	940	960	980	1000	:	977
Md_Opossum	:	-----	860	880	900	920	940	960	980	1000	:	984
Me_Wallaby	:	-----	860	880	900	920	940	960	980	1000	:	979
Oa_Platytyp	:	-----	860	880	900	920	940	960	980	1000	:	984
ag ag aagc tgg a cc agt t caga t caagga g ga cca a cccag t cag g a aa ga ga t tctg tga ca cctg g t c c g ca c												
Ha_Human	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1140
Pt_Chimpan	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1140
Gg_Gorilla	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1140
Fa_Orangut	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1140
Cj_Marmose	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1140
Rn_Rat	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1140
Mm_Mouse	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1140
Cf_Dog	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1140
Ec_Horse	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1143
Bt_Cow	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1146
La_elephan	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1149
Ch_Sloth	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1146
Md_Opossum	:	-----	1020	1040	1060	1080	1100	1120	1140	1160	:	1143
Me_Wallaby	:	-----	1020	1040	1060	1080	1100	1120	1140			

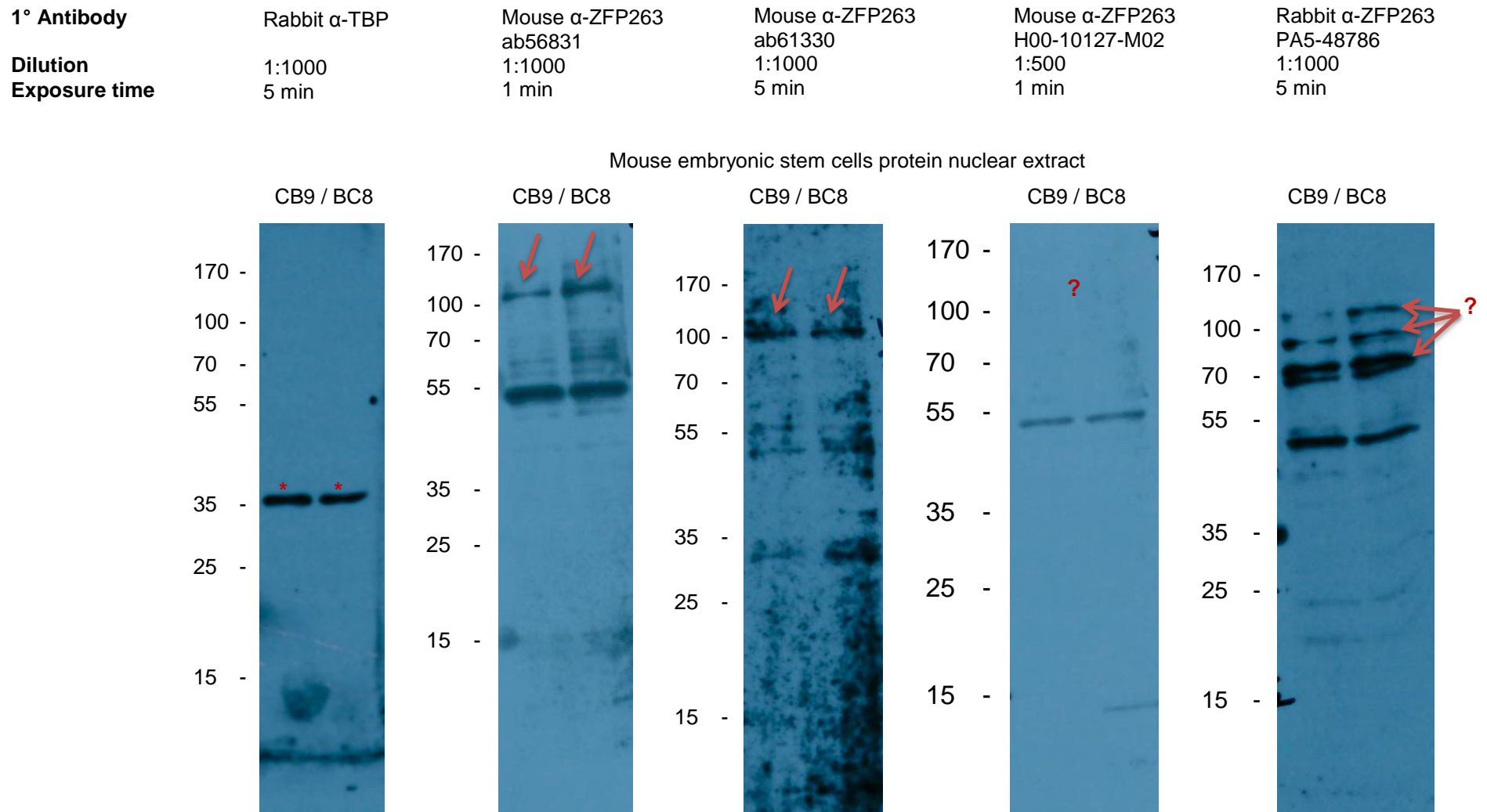
158

Appendix 8.2.2: List of anti-ZFP263 antibodies tested with their immunogen peptide, the % identity with the mouse protein and the location of the targeted amino acids sequence on the protein (red box)

Antibody	Host	Immunogen	Identity with mouse protein	Notes
Abcam ab56831	Mouse monoclonal	Human 201-299 amino acids (KRAB domain)	74% identity with mouse protein sequence Human PEGNMEDEKMTGQQLPESLEDVAMYISQEEWGHQDPSKRALSRDTVQESYENVDSLESHI Mouse PEGN+EDK+MTG QLPESELD+AMYISQE W HQDPSKRALSR VQ+SYEN +LES I PEGNVEDKDMTGTQLPESELDMAMYISQE-WDHQDPSKRALSRVMVQDSYENSGTLESSI	No longer available
Abcam ab61330	Mouse monoclonal	PEGNMEDEKEM TGPQLPESLE DVAMYISQEE WGHQDPSKRA LSRDTVQESY ENVDSLESHI PSQEVPGTQV GQGGKLWDPS VQSCKEGLSP RGPAPGEE	Human PSQEVPGTQVGGGKLWDPSVQSCKEGLSPRGPAPGEE Mouse PSQEV T V QG KLWD SVQ+CKEG++PR P PG E PSQEVSTHVEQGEKLWDSSVQTCKEGMNPRNPVPGVE	No longer available
Novus Biologicals H00010127-M02	Mouse monoclonal		1—38—125 — 220—279 — 378 — 680 	
ThermoFisher PA5-48786	Rabbit polyclonal	Human 318-346 amino acids VPSVCSENIH PQVLLPDQAR GEVPWSPELG RPHDRSQGDW APPPEG	64% identity with mouse protein sequence Human VPSVCSENIHPQVLLPDQARGEVPWSPELGRPHDRSQGDWAPPPE Mouse V SV E+ HP VLLP QAR EVPWSPE GR DR +G W PPE VESVSPSTHPPVLLPGQARREVWSPSEQRLDDR-EGHWECPPE	
ThermoFisher PA5-57150	Rabbit polyclonal	Human 295-382 amino acids PGEKGFENLE GVPSVCSENI HPQVLLPDQA RGEVPWSPEL GRPHDRSQGD WAPPPEGGME QALAGASSGR ELGRPKEQLP KKLHLCPL	60 % identity with mouse protein sequence Human PGEKGFENLE-GVPSVCSENIHPQVLLPDQARGEVPWSPELGRPHDRSQGDWAPPPEGGM Mouse PG EKFN E V SV E+ HP VLLP QAR EVPWSPE GR DR +G W PPE + PGVEKFNQERNVESVSPSTHPPVLLPGQARREVWSPSEQRLDDR-EGHWECPEDKI Human EQALAGASSGRELGRPKEQLPKKLHLCPL Mouse E++L G S + L + KE QPKKLHLC L EESLVGTPSCKGLVQAKE-QPKKLHLCAL	Not tested yet

Appendix 8.2.3: Additional Material and Methods for Western Blot

Cells from a T25 flask or a 10-cm dish were rinsed in PBS, coated with accutase for 5 minutes at 37°C, re-suspended in PBS and pelleted for 4 minutes at 1000rpm. The pellet was re-suspended in 200µL cold lysis buffer 1 (0.25% Triton X-100, 10mM EDTA, 0.5mM EGTA, 10mM Tris pH8, added with complete EDTA-free protease inhibitors before use) and incubated on ice for 20 minutes. Cells were pelleted at 14000rpm for 5 minutes at 4°C. The supernatant was collected as the membrane and cytoplasmic fraction. The pellet was rinsed once with 50µL of cold lysis buffer 1 and spun down. The supernatant was added to the cytoplasmic fraction. The pellet was re-suspended in 100µL cold lysis buffer 2 (1X PBS, 1% NP-40, 0.5% Sodium deoxycholate, 0.1% SDS, added with complete EDTA-free protease inhibitors before use) and incubated on ice for 20 minutes. Cells were pelleted at 14000rpm for 15 minutes at 4°C and the supernatant was collected as the nuclear fraction. One volume of 2X Laemmli buffer (4% SDS, 20% glycerol, 0.004% bromophenol blue, 0.125M Tris-HCl pH6.8, 10% 2-mercaptoethanol) was added to one volume of protein extract. Samples were boiled for 5 minutes, aliquoted and frozen down. Protein samples SDS-Polyacrylamide Gel Electrophoresis was performed and the transfer on PVDF membrane was performed using semi-dry western blot at 200mA for 1 hour. Membranes were activated 1 min in methanol, washed 5 min in water and 10 min in transfer buffer (25mM Tris-base, 14.4g/L glycine, 10% methanol). After the transfer membranes were incubated for 30 min in blocking buffer (8g/L NaCl, 0.2 g/L KCl, 3 g/L Tris), 0.05% Tween, 3% BSA). Primary antibodies are diluted in 5 ml of blocking buffer according to the supplier instruction for each antibody. Membranes are incubated on a shaker in antibody solution over-night at 4°C. Membranes are incubated 3 x 10 min in washing buffer (1X TBS, 0.05% Tween). Horseradish Peroxidase-conjugated antibodies were diluted according to the supplier instruction in blocking buffer. Membranes were incubated with the secondary antibody solution for 1h at room temperature and signal was detected using chemiluminescence.

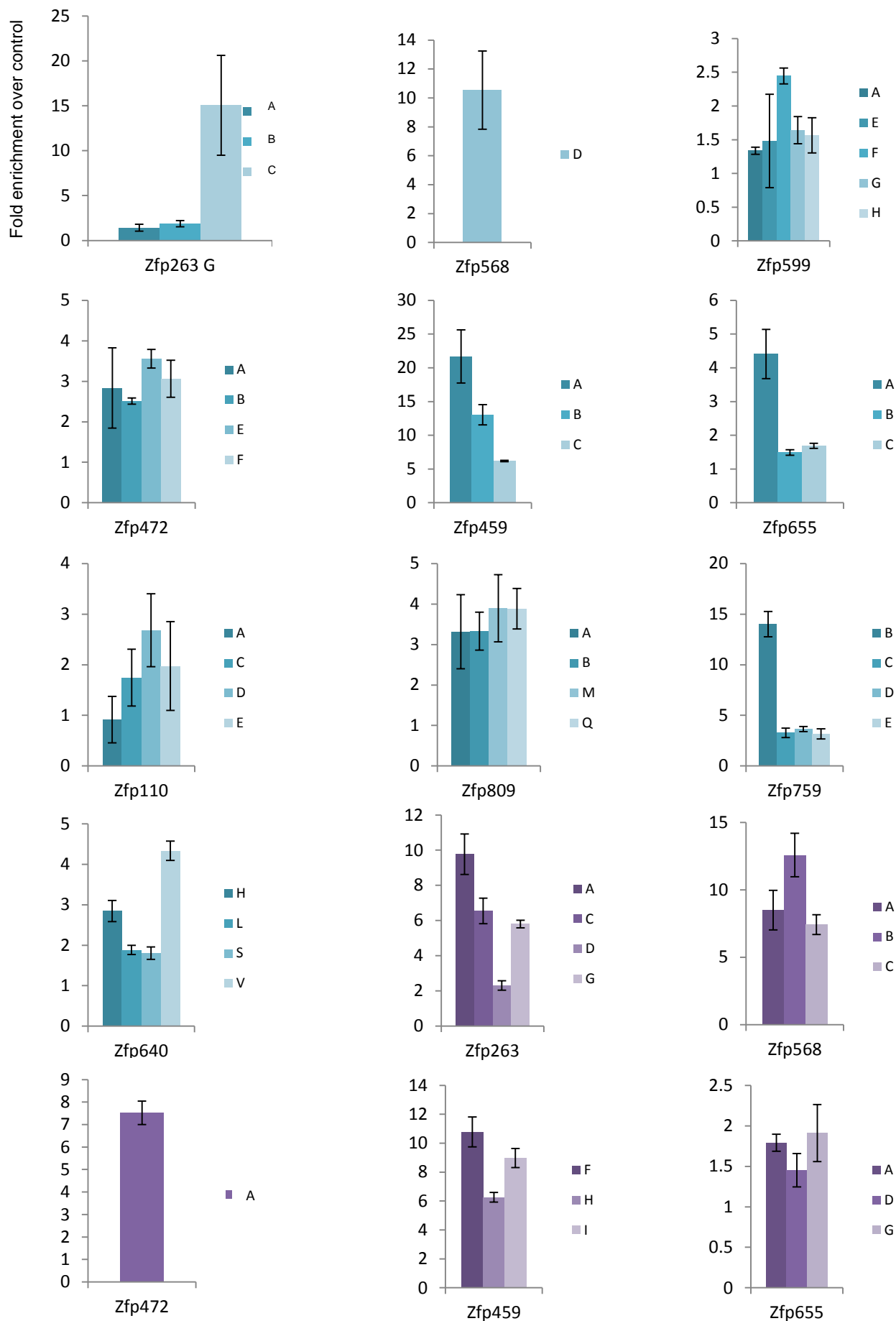


Appendix 8.2.4: Images of immunoblots. Proteins from nuclear extracts from two lines of mouse embryonic stem cells were run on a polyacrylamide gel and transferred onto a membrane. The same amount of protein was loaded for each lane. The membranes were incubated with antibodies targeting the TATA-binding protein (TBP) as a control (star) or with different antibodies targeting ZFP263 (arrow).

8.3. Appendix Chapter 3

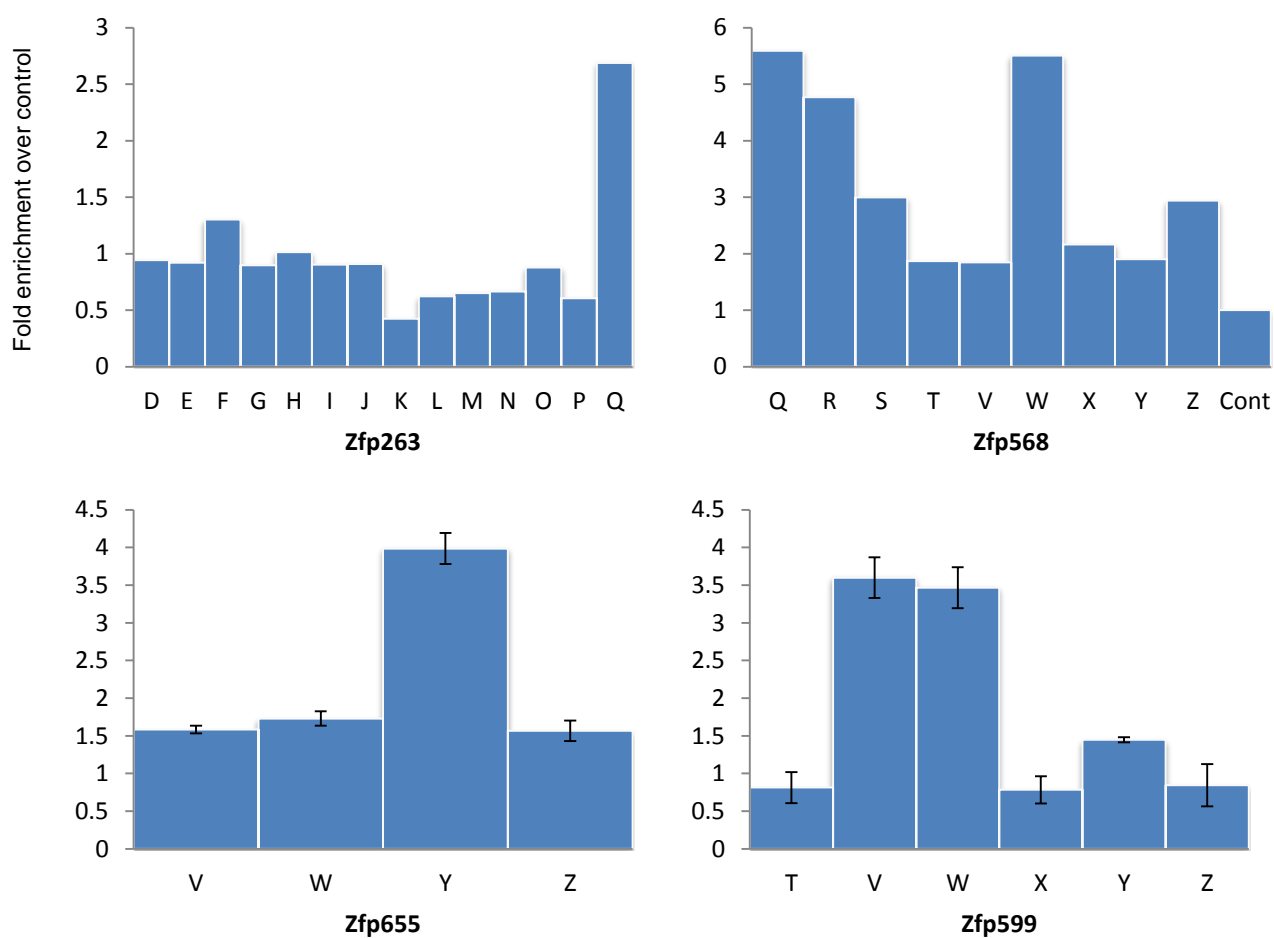
Appendix 8.3.1: Table summarising selection for clones overexpressing FLAG-tagged - candidate KZFPs. Dr Noon initiated the project, generated and screened the first clones that were used for the first ChIP-seq replicate. I performed the second screen for four KZFPs.

	<i>Zfp</i>	Clones screened (Dr Noon)	2 nd screen	overexpressing	Rep1	Rep2	validation
BC8	263	A, B, C	D - Q (18)	C	G	Q	G and Q
	568	D	Q - Z (9)	D	D	W	
	599	A, E, F, G, H	T - Z (6)	F	F	V	
	472	A, B, E, F		B, E, F	E	B	
	459	A, B, C		A, B, C	A	B	
	655	A, B, C	V, W, Y, Z	A	A	Y	
	110	A, C, D, E		D, E			
	809	A, B, M, Q		A, B, M, Q	C		
	759	B, C, D, E		B, C, D, E	B		
	640	H, L, S, V		H, V			
CB9	263	A, B, C, D		A, B, D			A
	568	A, B, C		A, B, C	B		
	472	A		A			
	459	F, H, I		F, H, I	F		
	655	A, D, G		A, G			



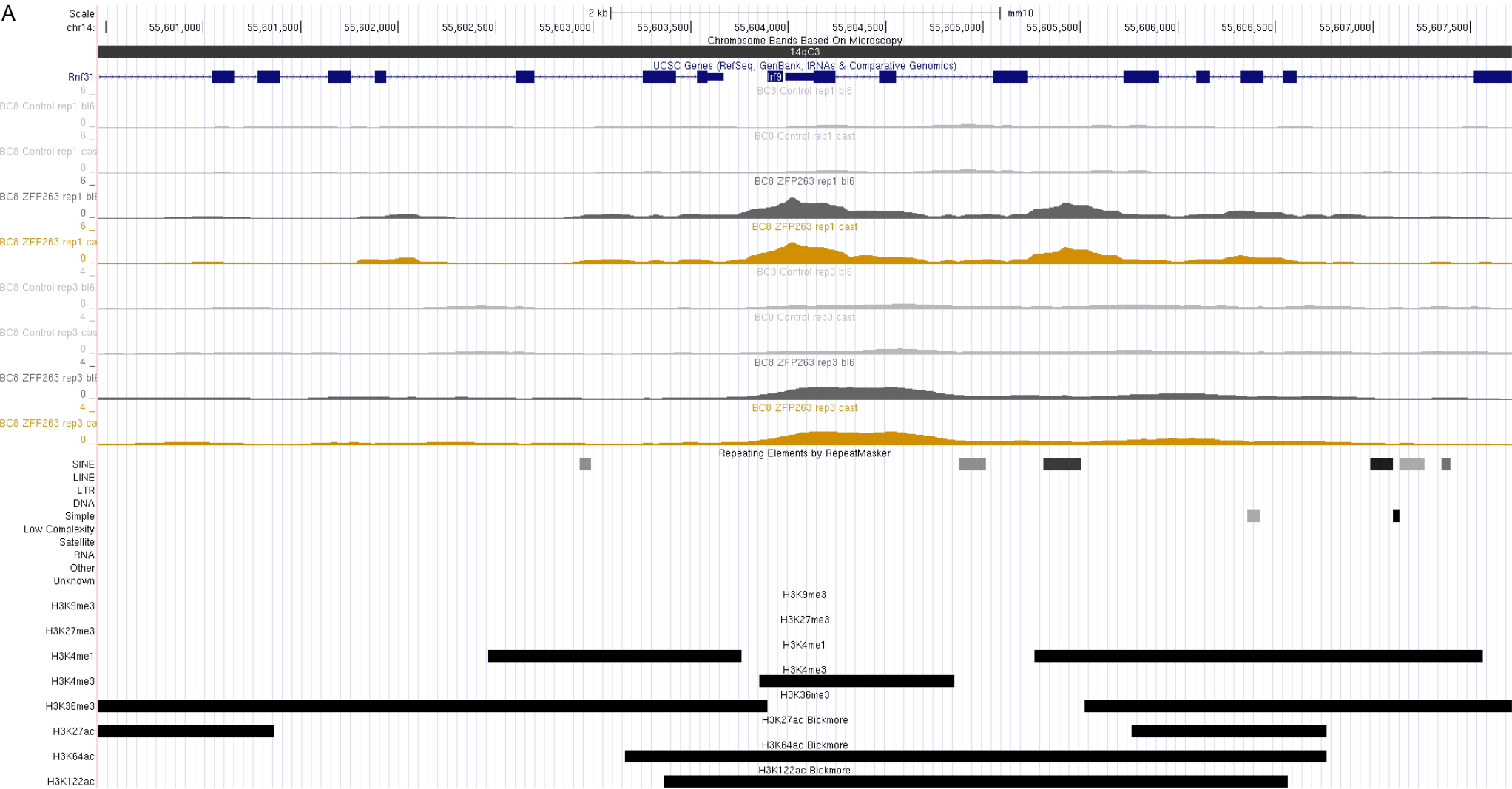
Appendix 8.3.2: Legend next page

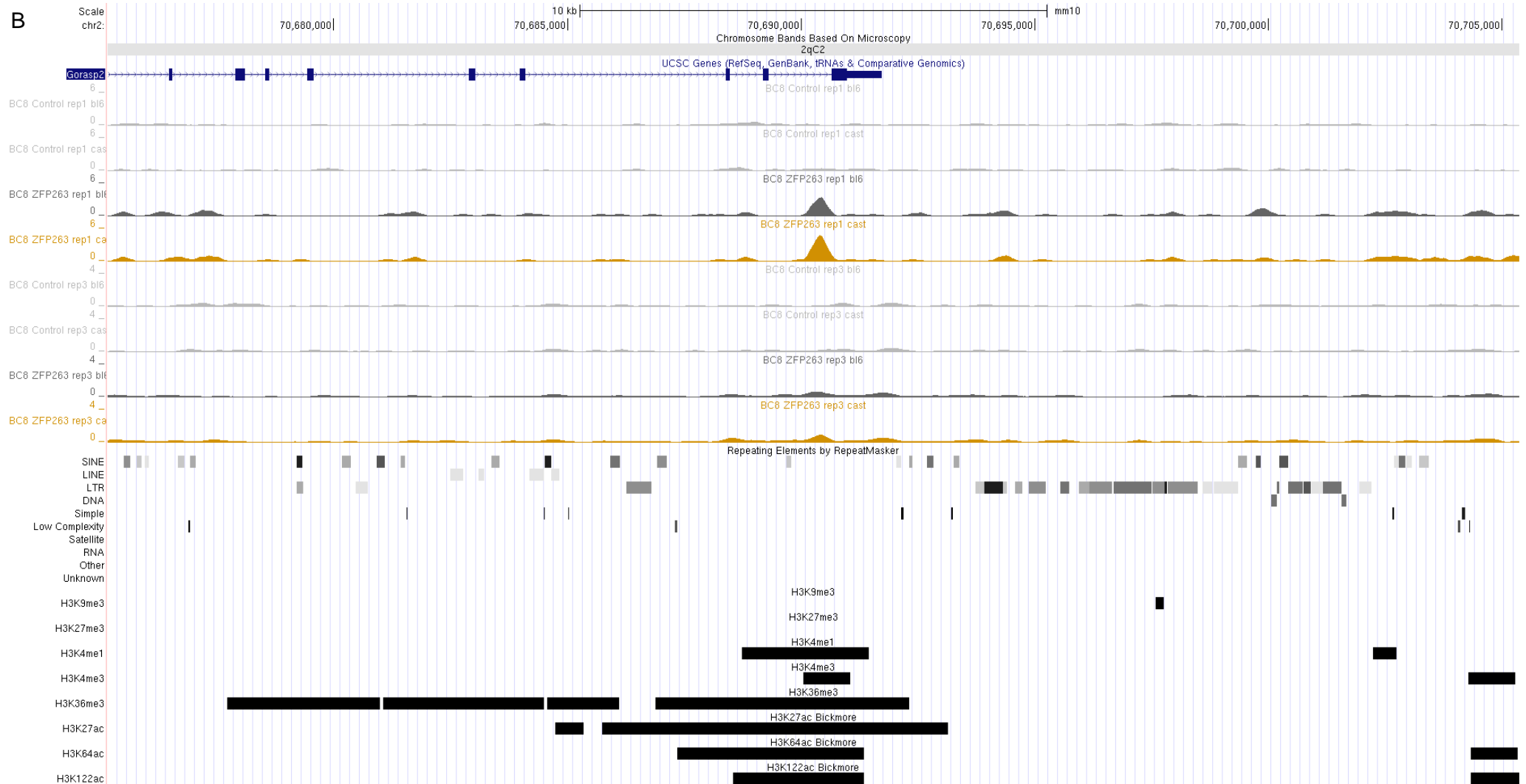
Appendix 8.3.2 (previous page): Fold enrichment of different ZFP genes in infected clones compared to control lines, screened by Dr Noon. Values were normalised with β -actin. n = 3, error bar = standard deviation. In light blue are the BC clones, in purple the CB clones.



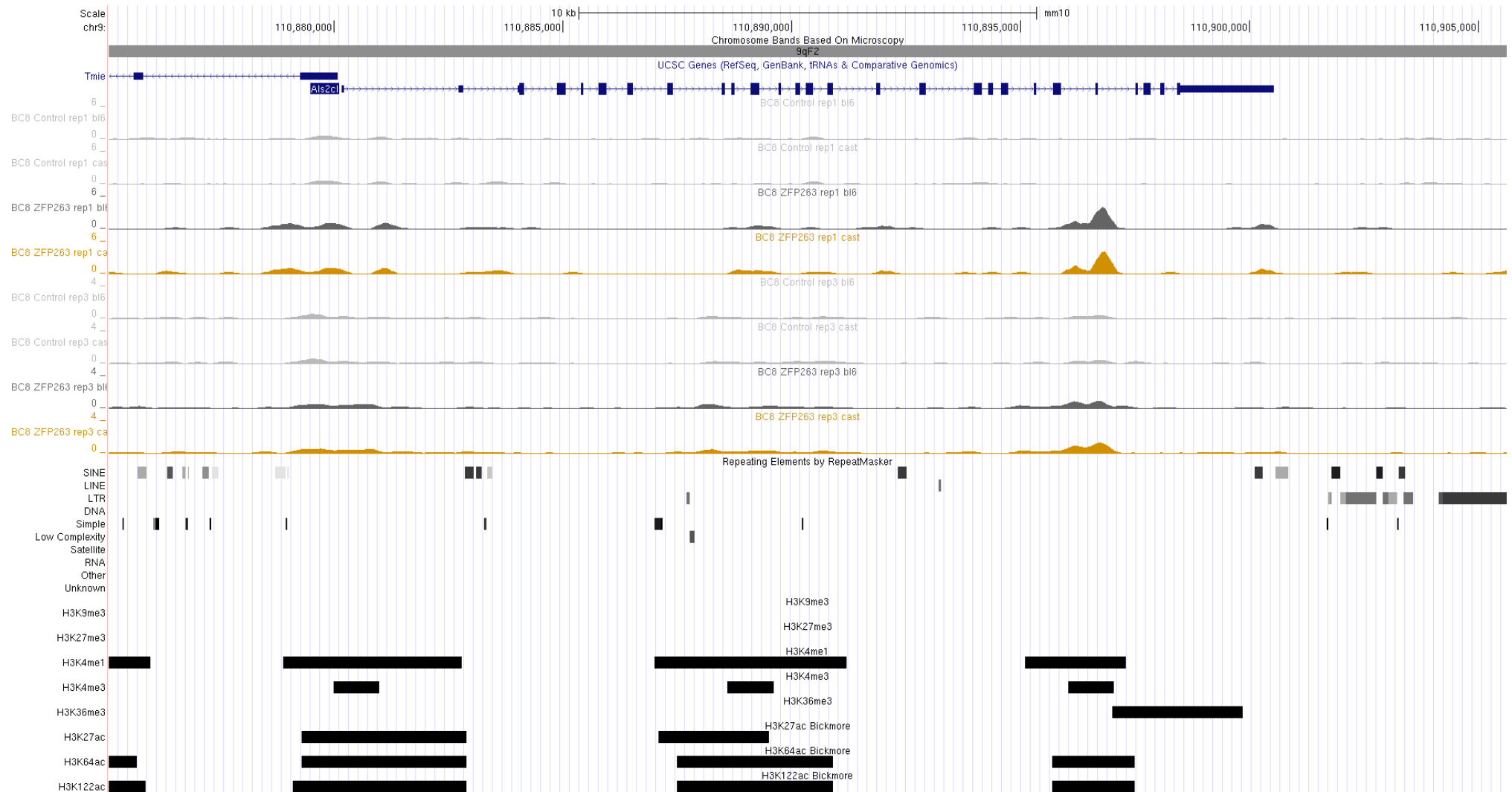
Appendix 8.3.3: Fold enrichment of different ZFP genes in infected clones compared to control lines. Values were normalised with β -actin. n = 3, error bar = standard deviation.

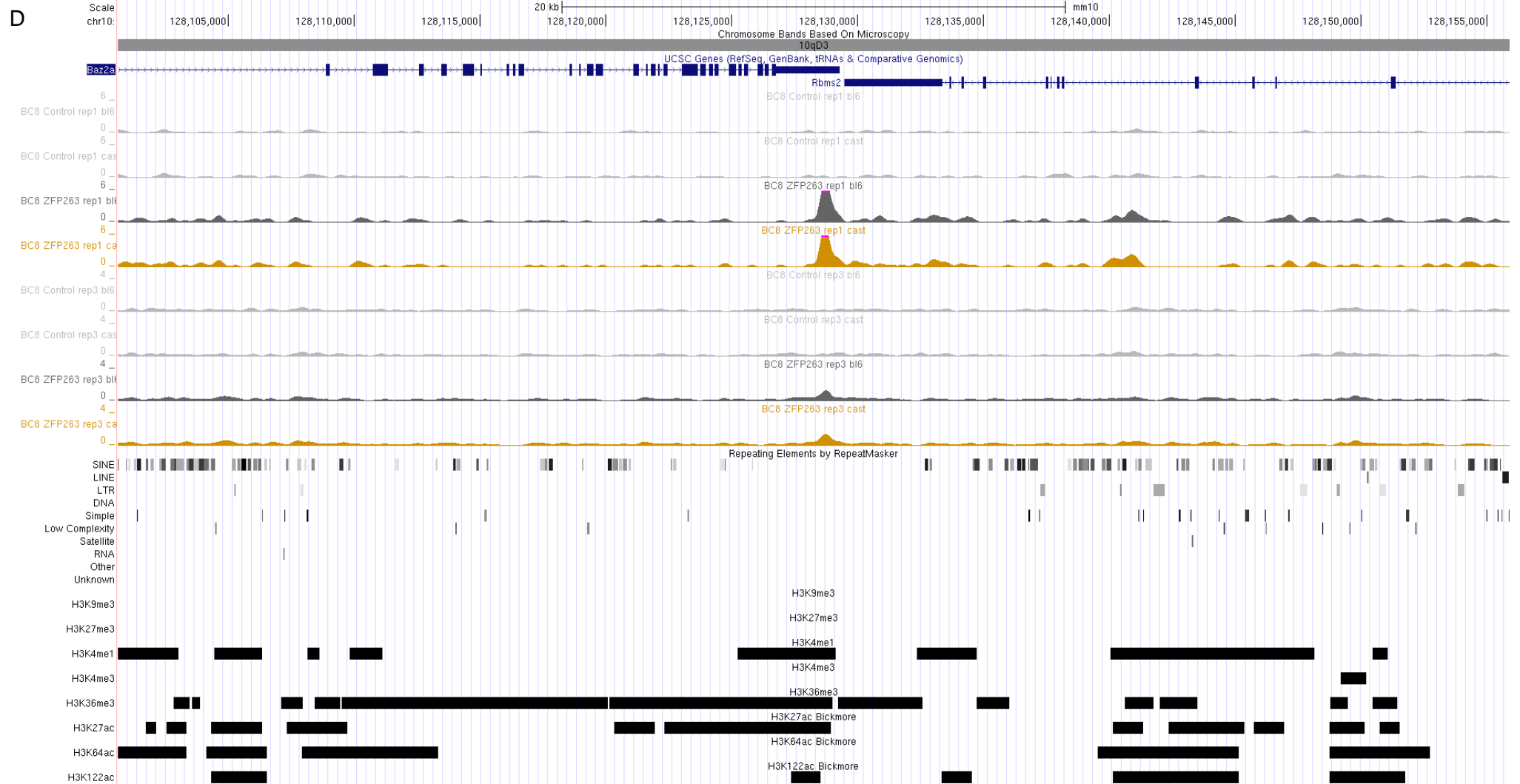
Appendix 8.3.4 (next 8): Screenshots of ChIP-seq signal from the UCSC Genome browser in different genome context. The ChIP-seq signal is shown for both replicates (“rep1” and “rep3”), for the control and ZFP263 for each genetics background (bl6 and cast). The different tracks show the UCSC genes, the repetitive elements by RepeatMasker, 5 histone modification domains from mENCODE project, and 3 histone modifications domains from Pradeepa et al. 2015.



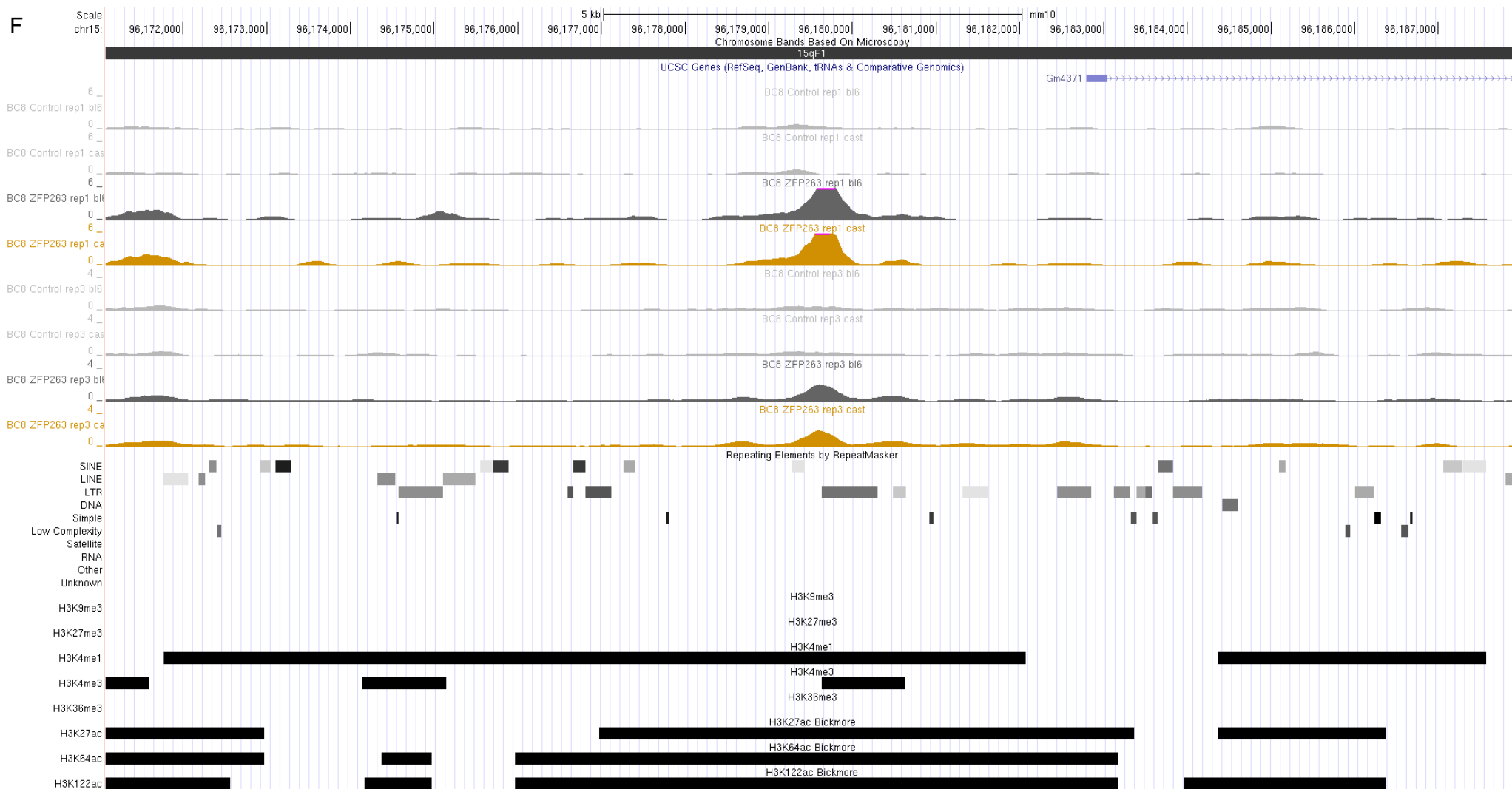


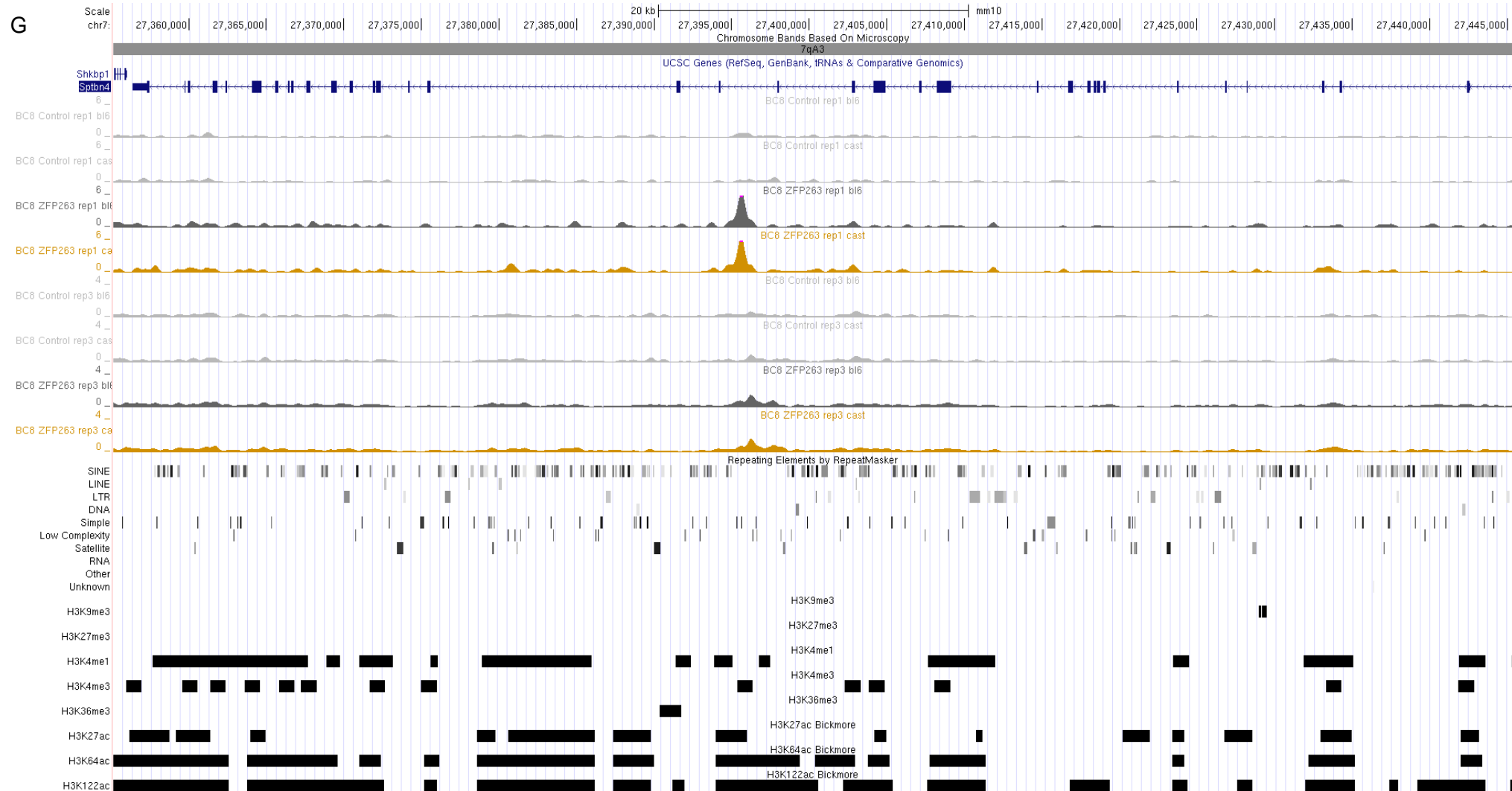
C

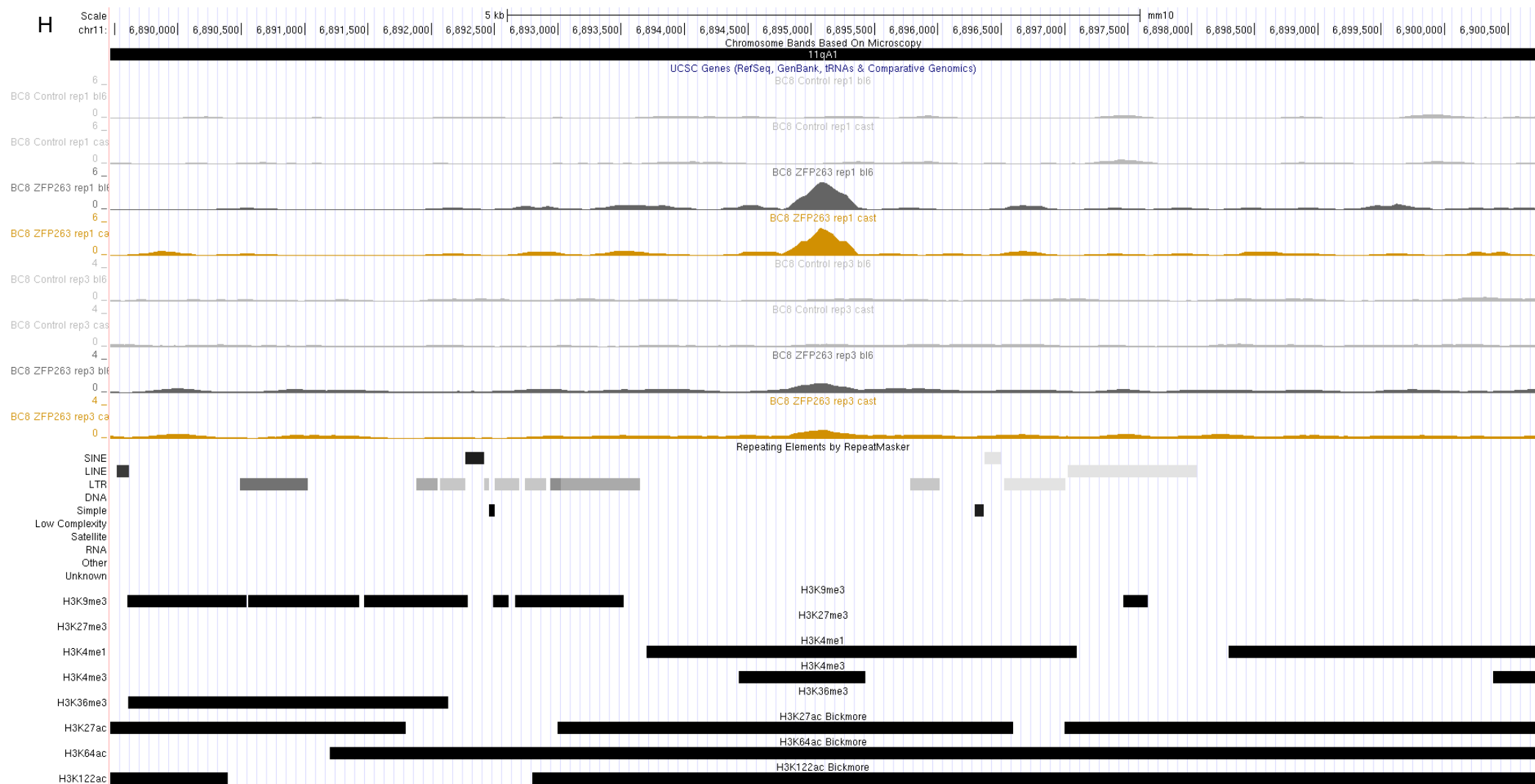




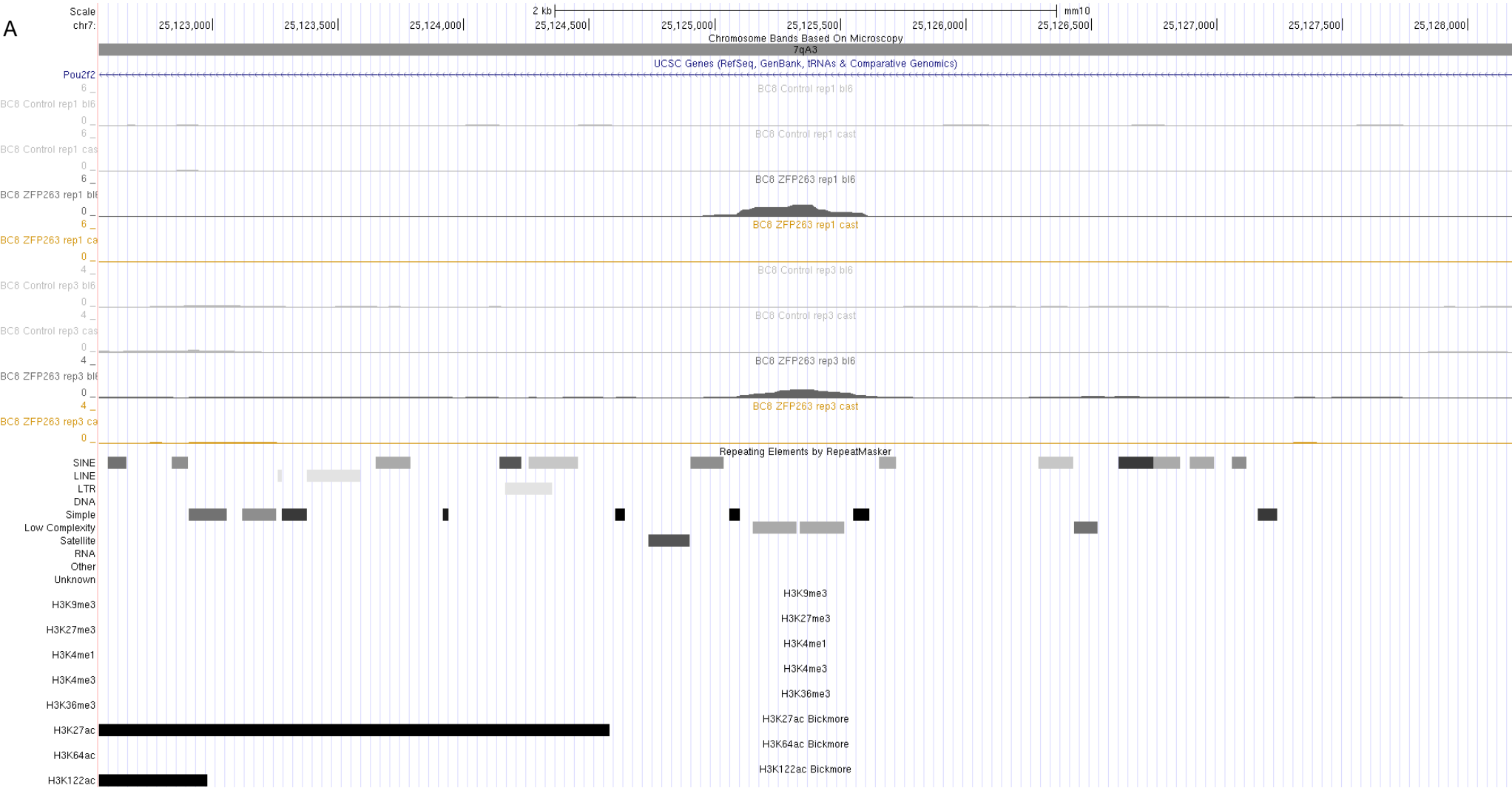




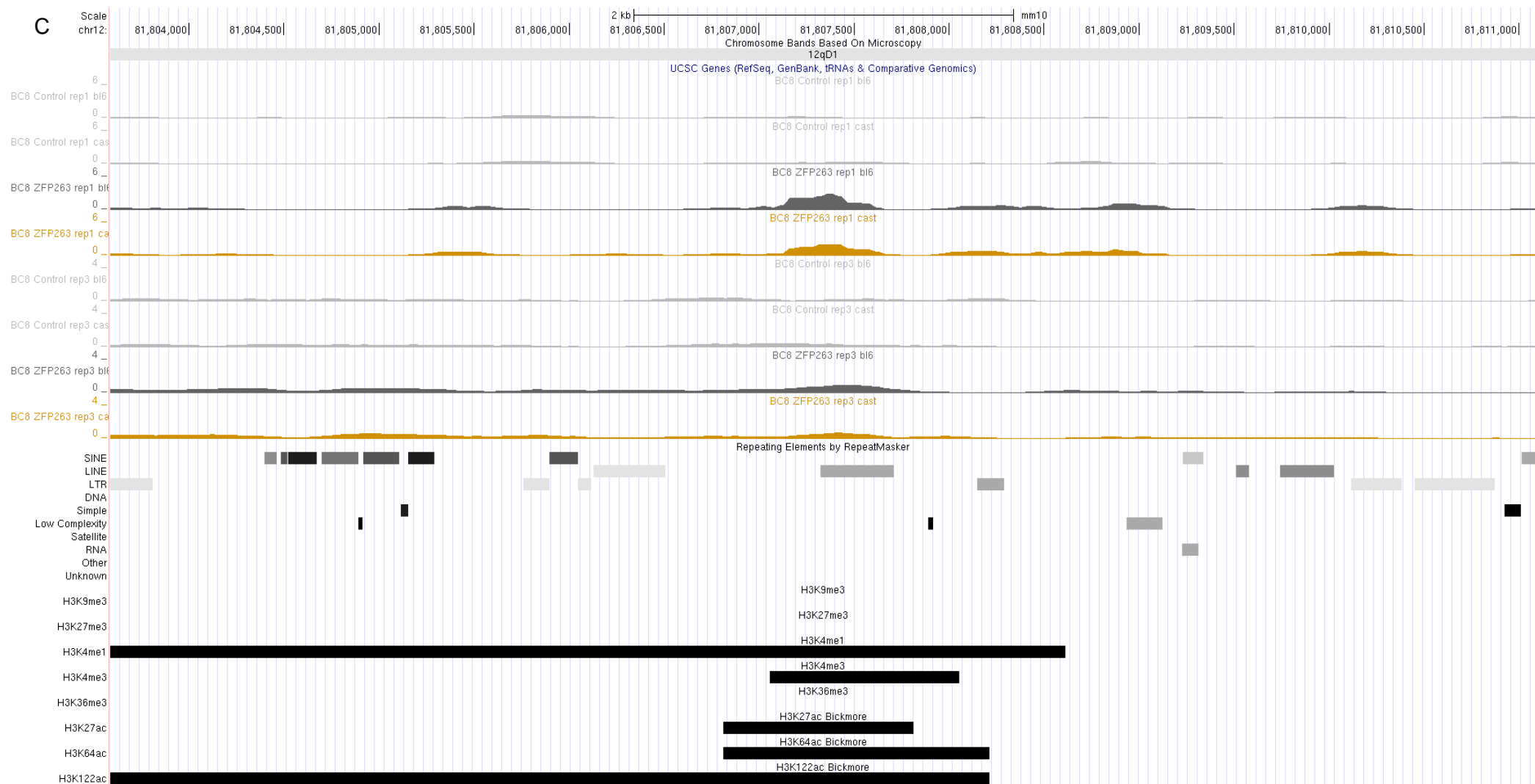




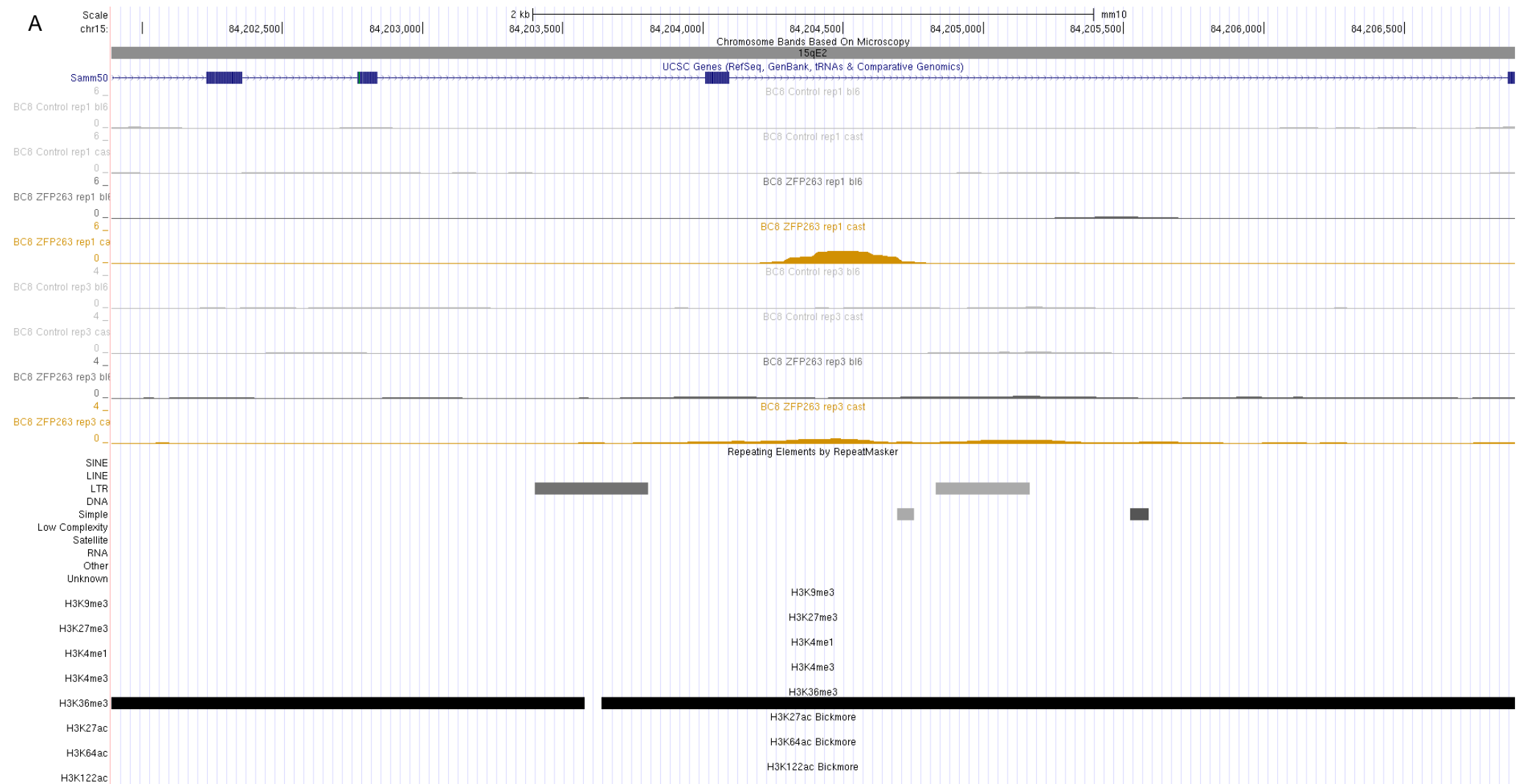
Appendix 8.3.5 (next 3): Screenshots of ChIP-seq signal from the UCSC Genome browser in different genome context. The tracks are identical to the Appendix 8.3.4. The peaks shown here are called only in C57BL/6 and are therefore allele-specific.

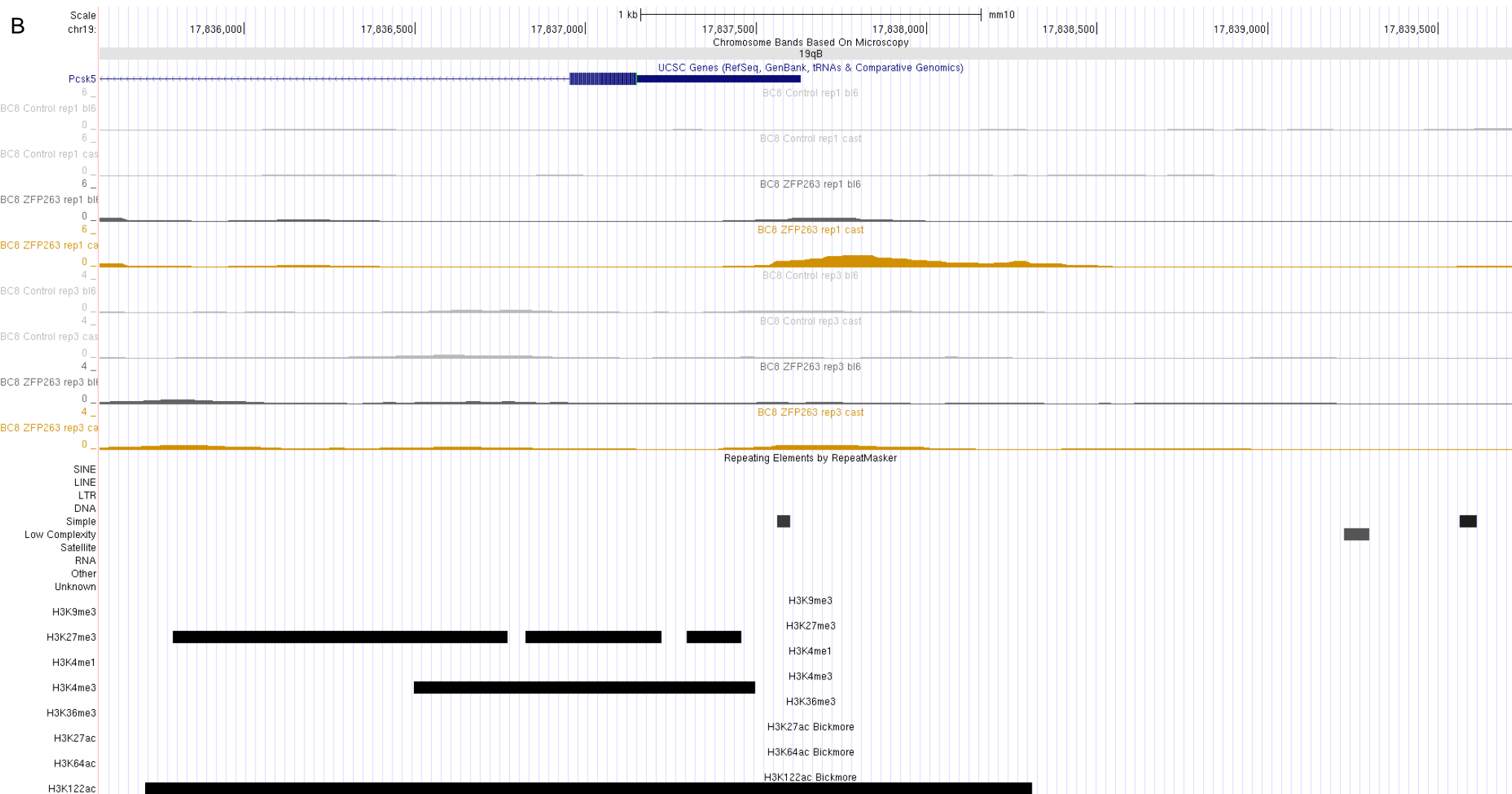


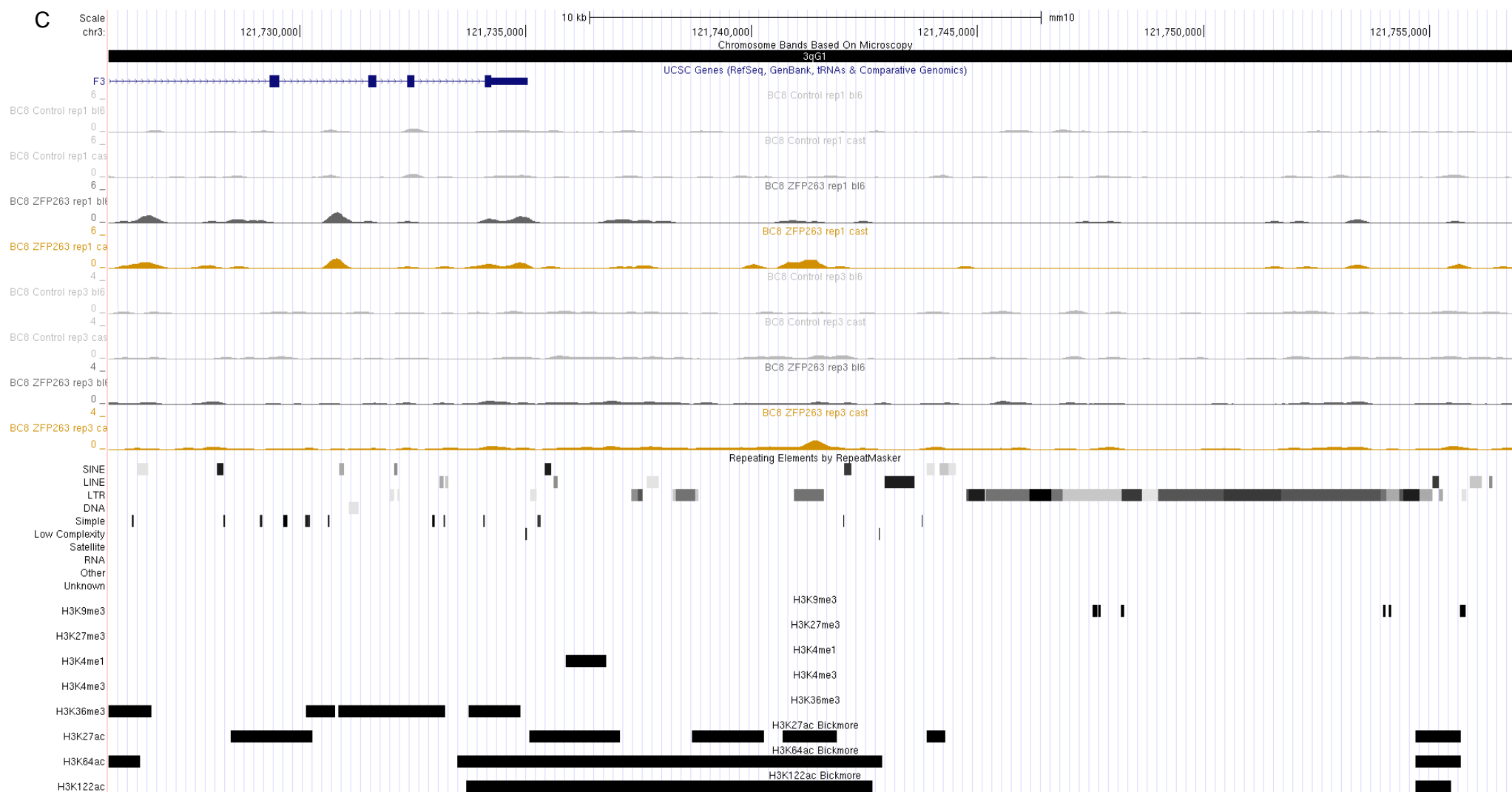


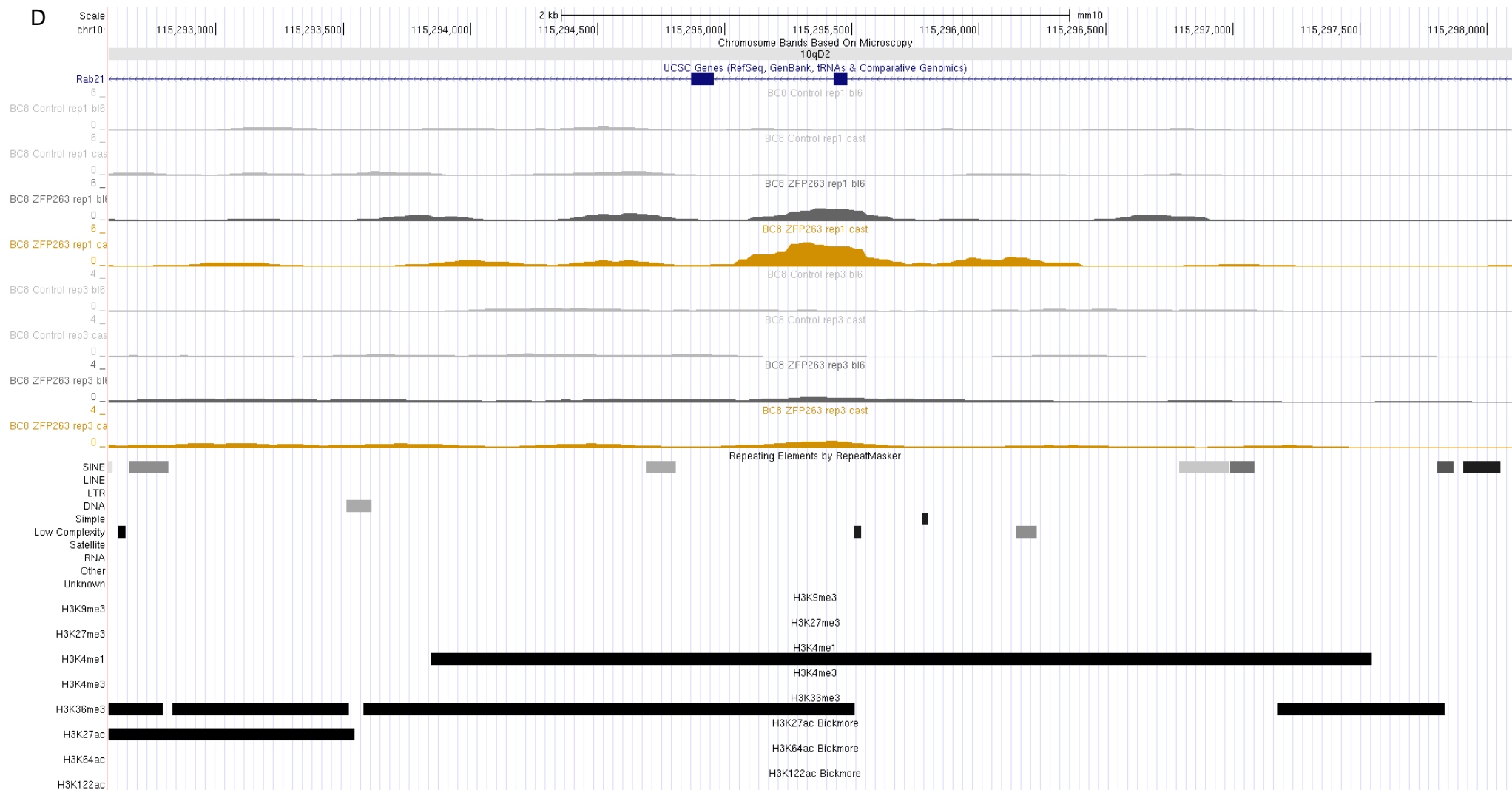


Appendix 8.3.6 (next 3): Screenshots of ChIP-seq signal from the UCSC Genome browser in different genome context. The tracks are identical to the Appendix 8.3.4. The peaks shown here are called only in Castaneus and are therefore allele-specific.









8.4 Appendix Chapter 4

Appendix 8.4.1 (next page): Designed gRNAs targeting exon 6 (A) and exon 1 (B). Highlighted in green are the two gRNAs chosen for each experiment. All of the exonic potential off-target sites are shown for each highlighted gRNA. The sequence of the potential off-target is given, as well as the score of the off-target likelihood, the number of mismatches and their locations within the off-target sequence, the corresponding gene and chromosome locus.

A

Top 10 gRNAs for *Zfp263* exon 6

	score	sequence	
Guide #1	86	GCCTGGCCAGGCAATAGTAC	TGG
Guide #2	79	CATCTCCAGTACTATTGCC	TGG
Guide #3	79	TGGCCAGGCAATAGTACTGG	AGG
Guide #4	79	TCCAGTACTATTGCTGGCC	AGG
Guide #5	78	GTGGGAACCCCAAGTTGTAA	AGG
Guide #6	78	CACCACTCCTTTACAACTTG	GGG
Guide #7	76	TATTGCTGGCCAGGCTAGA	AGG
Guide #8	75	GCACCACTCCTTTACAACTT	GGG
Guide #9	73	TGCACCACTCCTTTACAACT	TGG
Guide #10	73	ACTGGAGATGGGTACTTTC	AGG

sequence

ACCTGTCAAGGCAATAGTAGGAG
GGCCGCCAGGCAAGTAGTACCAG
GCTTGGCCAGGCAATGGTAGGG
GACTGGACAGGCAATAGTACCAG
GGCTGGCCAGGCTAGTAGTATAG
GCCTGGCCCGGCGATGGAACCAAG
GCCTGGCCAGGCAATGGTAGGAG

sequence

TCCGCTACCACTTGGCTGGCCAG
TCCATTACTCTGCTAGCCCGG
TCCACTGCCATTGCTTGGCCAGG
TTCAGCAATTTCTGCTGGCCAG
GCCAGTACTTTAGCTTGGCCCTAG
TACAGTACTCTTGGCTTGGCCAGG
GCCTGGCACTATTGACTGGCCCTAG
TCCAGTCTAGTCTCTGGCCCGAG
TCCAGTACTATTATCTGCCCCAG
TCCAGTACCACTTGGCTGCTAGG

All exonic potential off-target sites for exon 6

score	mismatches	UCSC gene	locus
0.4	4MMs [1:6:8:20]	NM_176980	chr11:+72697502
0.2	4MMs [2:4:6:15]	NM_177759	chr5:+116622394
0.2	3MMs [3:16:20]	NM_029116	chr8:-15028996
0.1	4MMs [2:7:14:17]	NM_175391	chr15:+77355512
0.0	4MMs [2:13:14:20]	NM_009539	chr1:-36839349
0.0	4MMs [9:13:16:18]	NM_198599	chr16:-20241328
0.0	4MMs [14:16:18:19]	NM_025725	chr5:+36827348

score	mismatches	UCSC gene	locus
1.5	3MMs [4:5:9]	NM_029116	chr8:+15028998
0.3	4MMs [5:10:11:17]	NM_011700	chr9:+118966881
0.2	4MMs [5:7:9:17]	NM_021313	chr1:-74642275
0.2	4MMs [2:6:9:13]	NM_001081165	chr10:+33759803
0.2	4MMs [1:10:12:15]	NM_001160145	chr1:+137931205
0.2	4MMs [2:10:17:19]	NM_001114174	chr19:-24052684
0.1	4MMs [1:4:6:14]	NM_001081020	chr13:+105234169
0.0	4MMs [7:11:13:14]	NM_175015	chr2:+73747922
0.0	3MMs [13:14:18]	NM_001044308	chr15:-80210767
0.0	4MMs [9:14:18:20]	NM_007831	chr18:+71606038

B

Top 10 gRNAs for *Zfp263* exon 1

	score	sequence	
Guide #1	97	ATTTCGACGGTTCCGCTTTC	AGG
Guide #2	96	GGAACCGTCGAAATCGCAGA	TGG
Guide #3	95	GAACCGTCGAAATCGCAGAT	GGG
Guide #4	92	TTCAGGAGCTTTGCCGCGGG	TGG
Guide #5	92	CCTCCCATCTGCGATTTCGA	CGG
Guide #6	92	GCTCAGCGGGGGTACTCTCG	TGG
Guide #7	90	TTCGCTTTTCAGGACGCCGCC	GGG
Guide #8	90	CGCTCGAAATCGCAGATGGG	AGG
Guide #9	90	AAGCTCCTGAAGCGGCTAA	GGG
Guide #10	90	GGAGGCCCTTAGCCGGCTTC	AGG

sequence

GTTTAGACGGTTTCTCTTCCAG
AATTCCGACAGTTCTGATTTCCAGG
CTTTCCGACGGTACAGCTCTCCAG

sequence

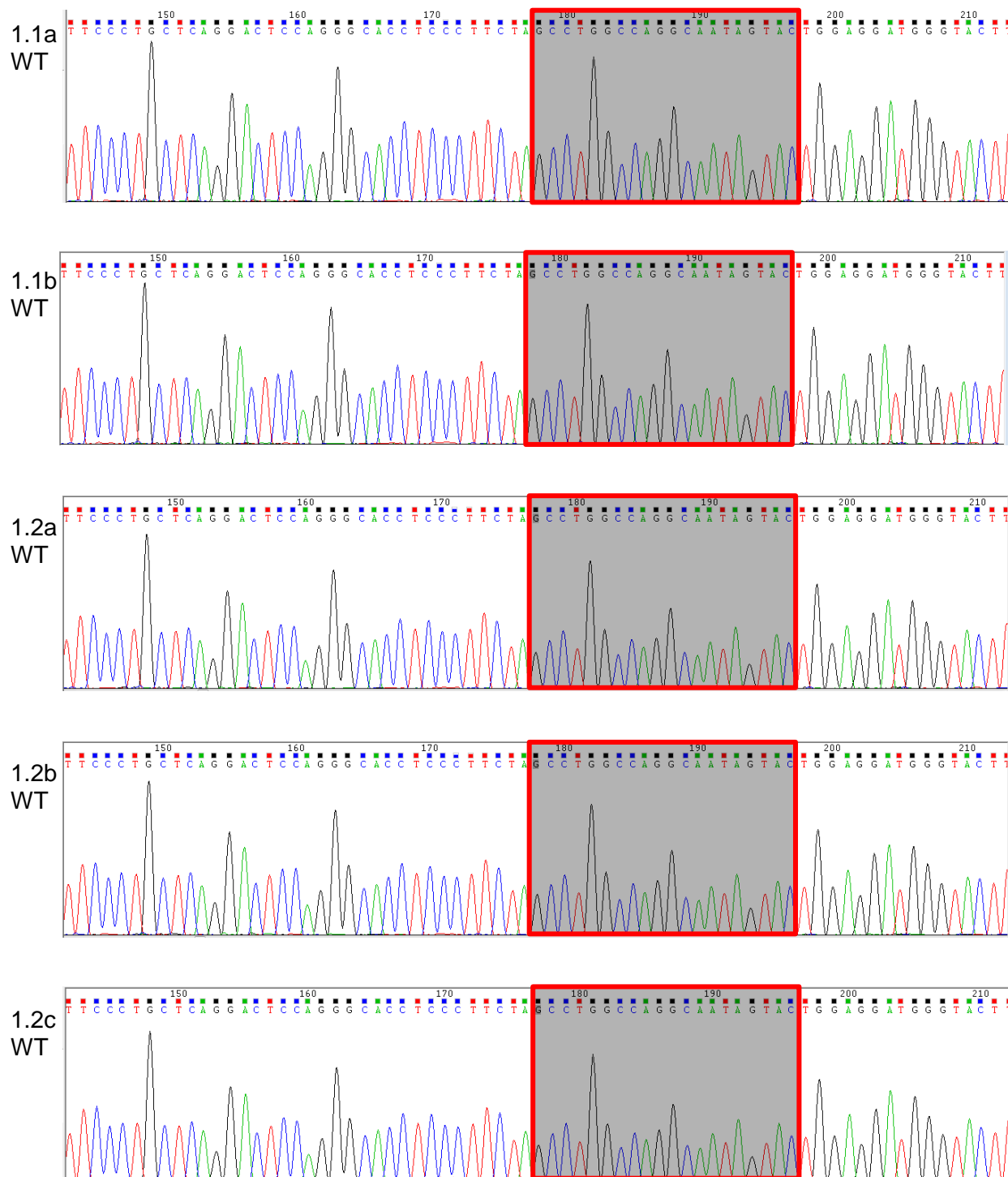
ACCGCTTCTTCACCGCGCTCAAG
ACCACTTCTTCGCTGCTCTCTGG
ACGCTTCTTCCCACTCTCGGG
ACTGCCTCCTCGCCACTCTCGGG
ACCGCTTCTTCCGCTCTCGGG
ACAGCTTCTTCACCGCTCCCAAG
CCTGCTTCTTCCGCGCTGGGG
AGTCTTCTTCGCGCCCTCCCTGG
TCTGCCTCCTCGCTGCTCTCGGG
GCTGCTTCTTAGGCTCTCCAGG
ACTGCTTATTTCGCGCTGTCAAG
ACTGCTTCTTCGCGCCCGCAAG
ACTCCTTCTTCCGCTTTCAAG
TCTGCTTCTTAGAGGCTCTCAAG
ACAGCTTCTTCGCAAGTCTCAGG
ACTGCTTCTTCAAGGCTCTCAGG
ACTGCTTCTTCCCACTGTCTAG

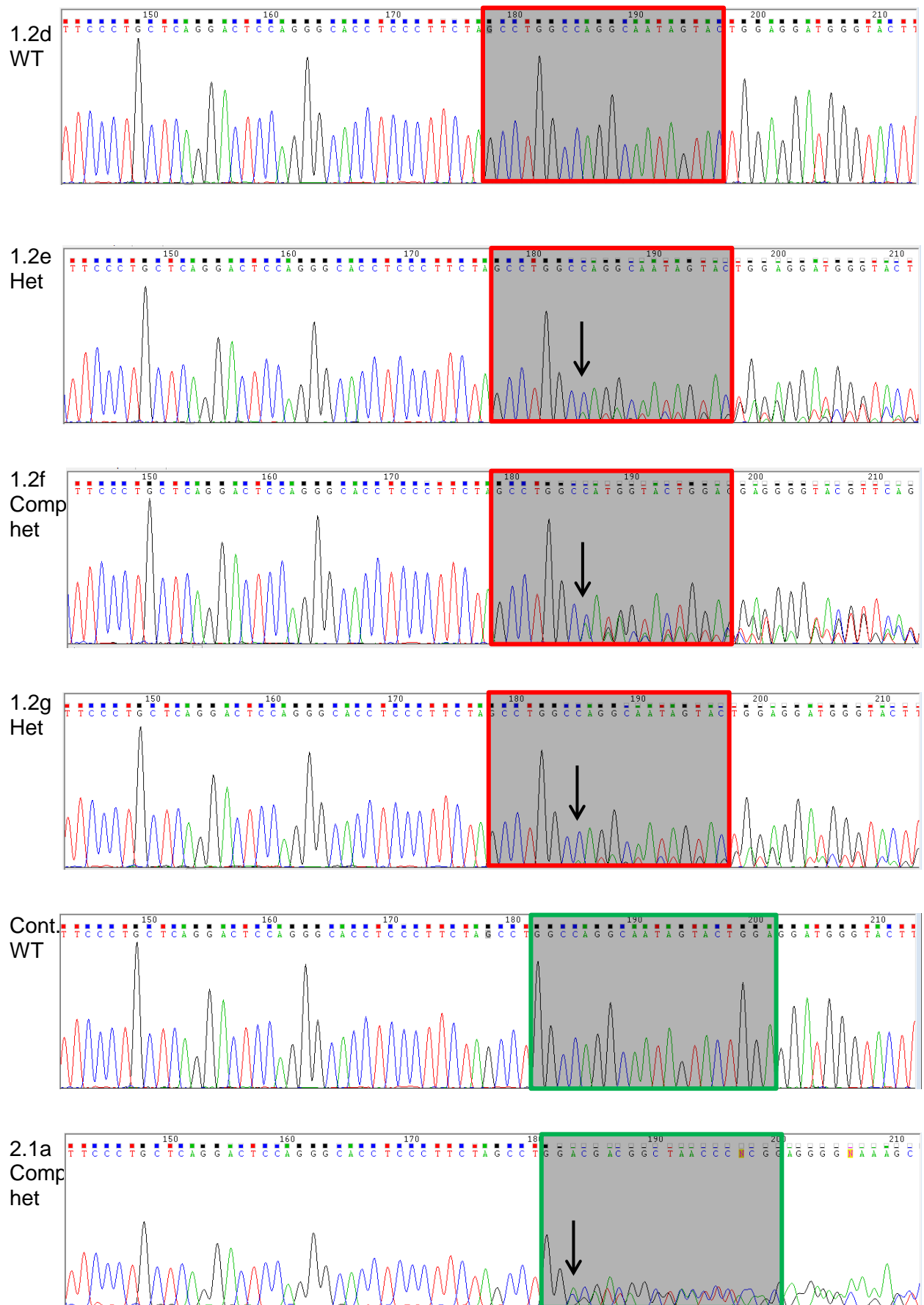
All exonic potential off-target sites for exon 1

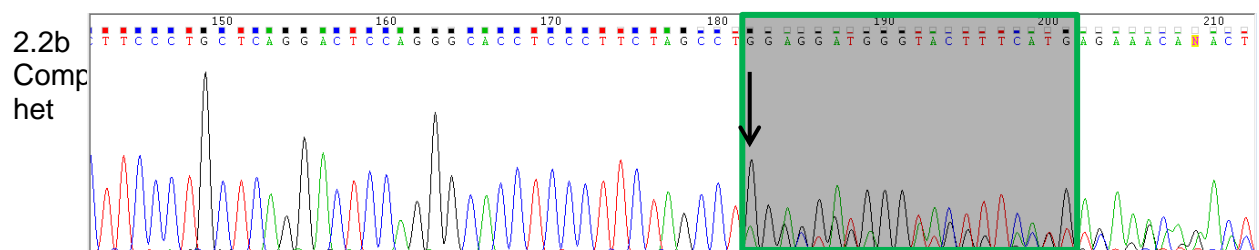
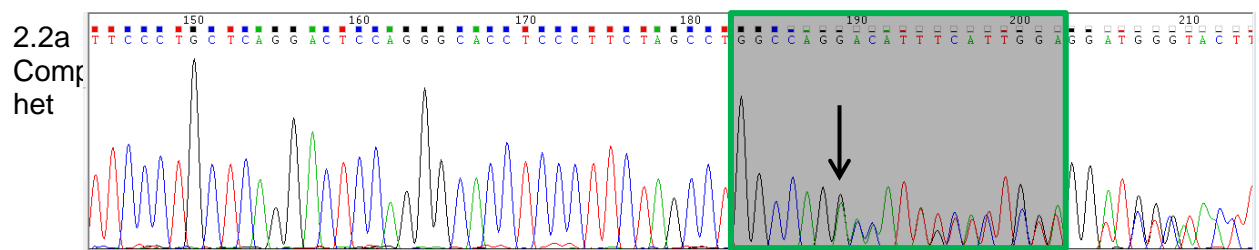
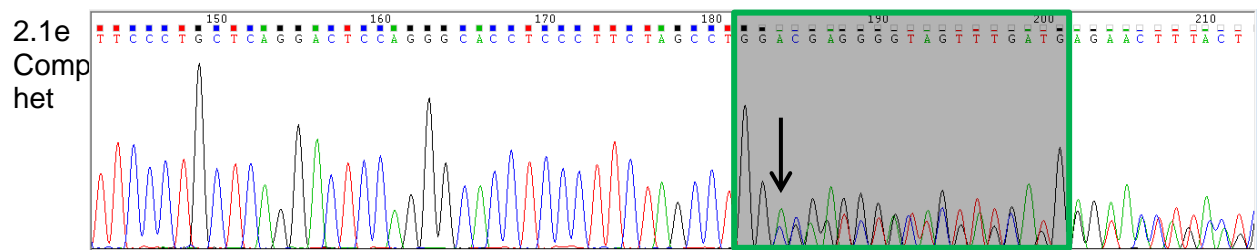
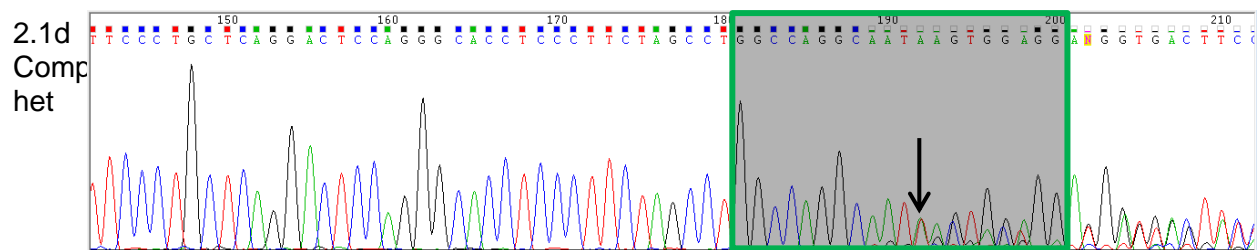
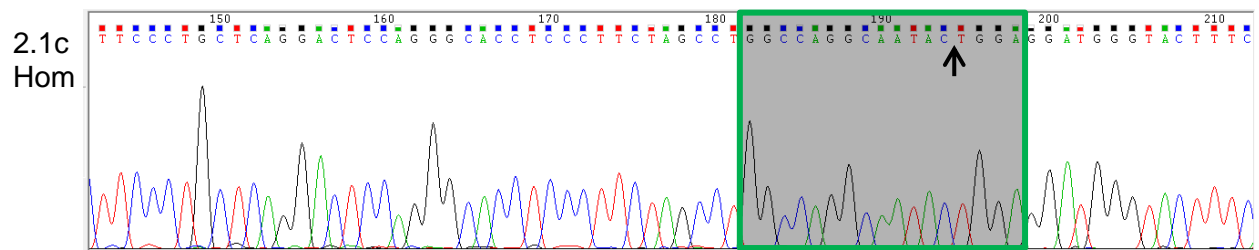
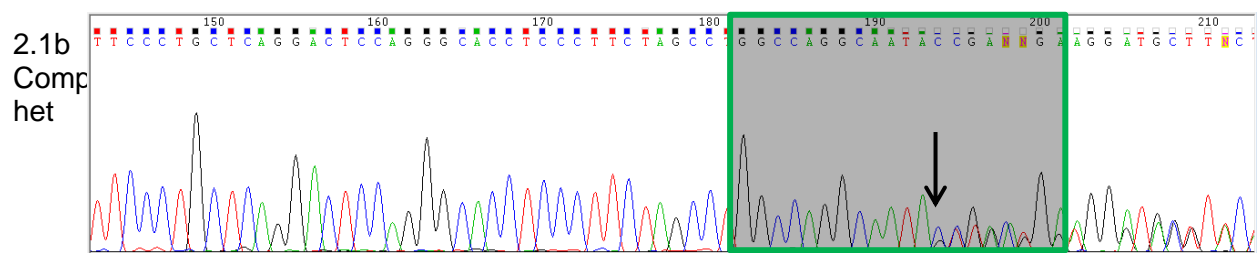
score	mismatches	UCSC gene	locus
0.2	4MMs [1:5:13:15]	NM_026816	chr14:+76296674
0.0	4MMs [2:9:14:16]	NM_173865	chr1:-133744899
0.0	4MMs [1:12:14:18]	NM_029649	chr11:+102015932

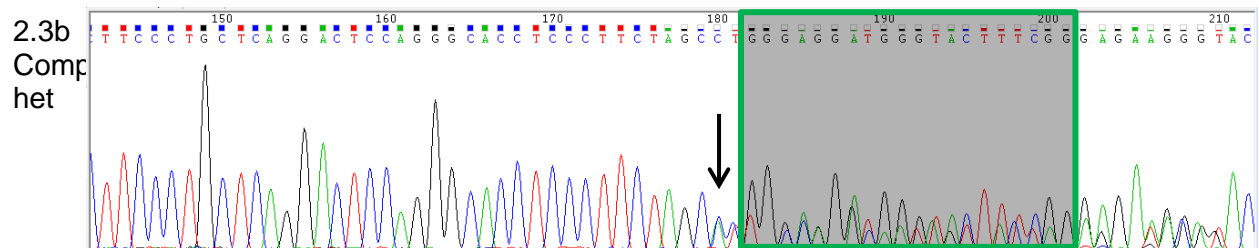
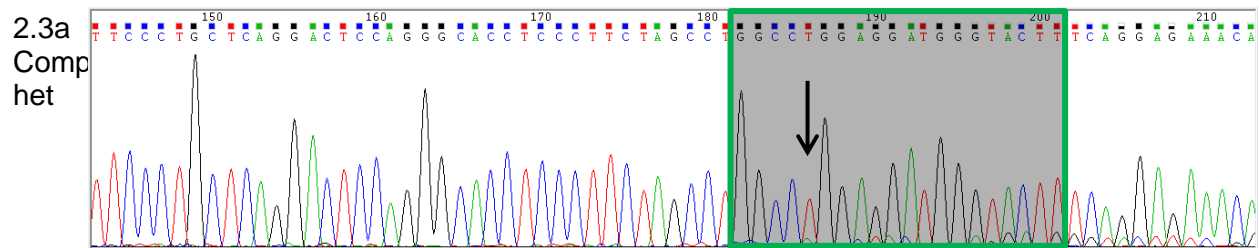
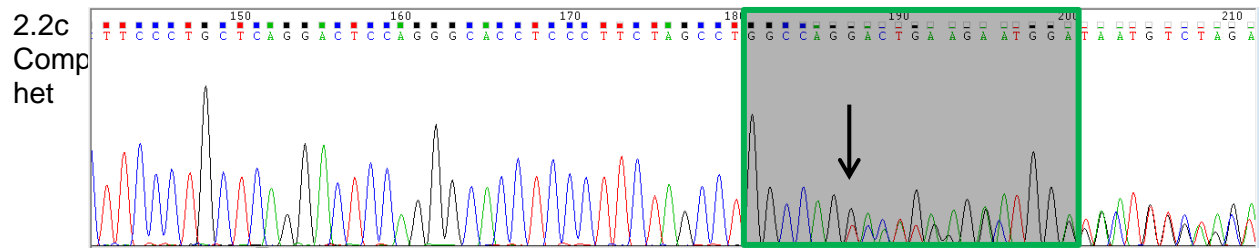
score	mismatches	UCSC gene	locus
0.6	3MMs [3:12:17]	NM_027090	chr9:-63246691
0.4	3MMs [3:4:14]	NM_133208	chr11:+62541826
0.4	3MMs [3:12:15]	NM_016683	chr5:-145966501
0.3	3MMs [6:9:15]	NM_030713	chr9:-40015203
0.3	4MMs [3:6:9:12]	NM_177758	chr4:+128281496
0.2	4MMs [3:9:12:19]	NM_001033496	chr17:+23698160
0.1	4MMs [1:12:17:20]	NM_146218	chr8:-113823995
0.1	4MMs [2:10:15:19]	NM_016896	chr11:+103085491
0.1	4MMs [1:6:9:14]	NM_001034862	chr8:+14033483
0.1	4MMs [1:11:15:19]	NM_001111102	chr3:+88290820
0.1	4MMs [8:11:12:18]	NM_019400	chr11:+70658256
0.1	4MMs [7:9:17:19]	NR_001592	chr7:+149762768
0.1	4MMs [4:12:13:18]	NM_007659	chr10:+68808884
0.0	4MMs [1:11:13:14]	NM_009945	chr9:-79606203
0.0	4MMs [3:9:15:16]	NM_001177505	chr9:-122797994
0.0	4MMs [6:9:12:14]	NM_009553	chr7:-88008489
0.0	4MMs [10:12:14:18]	NR_040416	chr5:+120135854

Appendix 8.4.2: Chromatograms of the genomic DNA sequencing for each mouse generated from the first experiment targeting exon 6 of *Zfp263*. Highlighted in red and green are the sequences targeted by gRNA1 and gRNA2 respectively within exon 6. A black arrow indicates a mixed signal in the sequencing, meaning a mutation occurred in one or both alleles.









Appendix 8.4.3: DNA alignment of all mutations generated in exon 6 of *Zfp263*. The premature STOP codon caused by the mutation is highlighted in red.

IN FRAME mutations

Inversion

WT	GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATTGCCTGGCCAGGCTAGAAGGGAGGTG
Sing. nuc	GTTTCTCCTGAAAGTACCCATCCTCCAGTCTATTGCCTGGCCAGGCTAGAAGGGAGGTG

Deletions

WT	GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATTGCCTGGCCAGGCTAGAAGGGAGGTG
12 del	GTTTCTCCTGAAAGTACCCATCCTCCAG-----GCCAGGCTAGAAGGGAGGTG
3 del	GTTTCTCCTGAAAGTACCCATCCTCCAG---TATTGCCTGGCCAGGCTAGAAGGGAGGTG

Insertions and Deletions

WT	GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATTGCCTGGCCAGGCTAGAAGGGAGGTG
15del+3in	GTTTCTCCTGAAAGTACCCATCCTCCAG-----CCTAGGCTAGAAGGGAG

FRAMESHIFT mutations

Deletions

WT	GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATTGCCTGGCCAGGCTAGAAGGGAGGTG
2 del	GTTTCTCCTGAAAGTACCCATCCTCCAGTA--ATTGCCTGGCCAGGCTAGAAAGGGAGGTG
5 del A	GTTTCTCCTGAAAGTACCCATCCTCCAGTA-----GCCTGGCCAGGCTAGAAAGGGAGGTG
5 del B	GTTTCTCCTGAAAGTACCCATCCTCCAGTACTA-----TGGCCAGGCTAGAAAGGGAGGTG
5 del C	GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATT-----GCCAGGCTAGAAAGGGAGGTG
5 del D	GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATTGCCT-----GGCTAGAAAGGGAGGTG
7 del	GTTTCTCCTGAAAGTACCCATCCTCCAGTACT-----GGCCAGGCTAGAAGGGAGGTG CCCTGGAGTCCTGAGCAGGGAAGACTTGATGACAGAGAAGGACATTGGGAATGTCCCCAGAGGACAAAAATAGAG
16 del	GTTTCTCCTGAAAGTACCCATCCTCCAGT-----GCTAGAAGGGAGGTG CCCTGGAGTCCTGAGCAGGGAAGACTTGATGACAGAGAAGGACATTGGGAATGTCCCCAGAGGACAAAAATAGAG
17 del	GTTTCTCCTGAAAGTACCCATCCT-----CCAGGCTAGAAAGGGAGGTG
WT	CCCAGAAACCCAGTTCCAGGAGTGGAAGTTTGAGAACCAAGAAAGAAATGTTGAGTCT
43 del	CCCAGAAACCCAGTTCCAGGAGTGGAAGTTTGAGAACCAAGAAAGAAATGTT-----
WT	GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATTGCCTGGCCAGGCTAGAAGGGAGGTG
43 del	-----CTGGCCAGGCTAGAAGGGAGGTG CCCTGGAGTCCTGAGCAGGGAAGACTTGATGACAGAGAAGGACATTGGGAATGTCCCCAGAGGACAAAAATAGAG

Insertions

WT

GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATTGCC-TGGCCAGGC TAGAAGGGAGGT
|||||
1 ins A GTTTCTCCTGAAAGTACCCATCCTCCAGTACTATTGCCTTGGCCAGGC TAGAAGGGAGGT
|||||

WT

GTTTCTCCTGAAAGTACCCATCCTCCAGTA-CTATTGCCTGGCCAGGC TAGAAGGGAGGT
|||||
1 ins B GTTTCTCCTGAAAGTACCCATCCTCCAGTAGCTATTGCCTGGCCAGGC TAGAAGGGAGGT
|||||

1 ins B'

GTTTCTCCTGAAAGTACCCATCCTCCAGTACCTATTGCCTGGCCAGGC TAGAAGGGAGGT
|||||

1 ins B''

GTTTCTCCTGAAAGTACCCATCCTCCAGTAACTATTGCCTGGCCAGGC TAGAAGGGAGGT
|||||

Insertions and Deletions

WT	TTTTTTTTTCTCTGTAAAATAATCTAGGGCTTTATTTTTTCTCTCTTACTAGGAGTGGAA
113del+62	TTTTTTTTTCTCTGTAAA-----
WT	AAAGTTTGAGAACCAAGAAAGAAATGTTGAGTCTGTTTCTCCTGAAAGTACCCATCCTCC
113del+62	-----
WT	AGTACTATTGCC-----
113del+62	-----AAGAGAAGCTTTTCCACAGGCATTGCACTGGAAGGGCTTTTCCCCTG
WT	-----TGGCCAGGCTAGAAGGGAGGTGCCCTGGAGTCCTGAGCAGGGAAGA
113del+62	TGTGGGTGCGCTGATGGCCAGGCTAGAAGGGAGGTGCCCTGGAGTCCTGAGCAGGGAAGA

Appendix 8.4.4: DNA alignment of all mutations generated in exon 1 of *Zfp263*. The premature STOP codon caused by the mutation is not highlighted here because it is occurring further on in the sequence except for the 113 insertion mutation

IN FRAME mutations

Deletion

WT	CTTAGCCGGCTTCAGGAGCTTTGCCGCGGGTGGCTGCGGCCGGAGATGCGCACCAAGGAA
123del	CTTAGCCGGCTTCAGGAGCTTTGCCGCGG-----
WT	CAGATCCTGGAGCTGTTGGTGCTGGAGCAGTTCTTGACTATCCTTCCCCAGGAGATTCAG
123del	-----
WT	AGCAGGGTGCAGGAGCTGCGCCCAGAGAGCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
123del	-----CGAAGAAGCAGTCACTCTTGTGGAGCGT

FRAMESHIFT mutations

Deletions

WT	AGCAGGGTGCAGGAGCTGCGCCCAGAGAGCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
20delA	AGCAGGGT-----GCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
20delB	AGCAGGGTGC-----GCGGAATAAGAAGTCACTCTTGTGGAGCGT
19del	AGCAGGGTGCA-----GGCGAAGAAGCAGTCACTCTTGTGGAGCGT
14delA	AGCAGGGTGCA-----AGAGCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
14delB	AGCAGGGTGCA-----GAGCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
14delC	AGCAGGGTGCAGGAGCT-----GCGAAGAAGCAGTCACTCTTGTGGAGCGT
14delD	AGCAGGGTGCAGGAGCTGCGC-----AGAAGCAGTCACTCTTGTGGAGCGT
8del	AGCAGGGTGCAGGAGCT-----AGAGCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
11del	AGCAGGGTGCAGGAGCT-----GCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
WT	CAGATCCTGGAGCTGTTGGTGCTGGAGCAGTTCTTGACTATCCTTCCCCAGGAGATTCAG
55del	CAGATCCTGGAGCTGTTGGTGCTGGAGCAGTTCTTGAC-----
WT	AGCAGGGTGCAGGAGCTGCGCCCAGAGAGCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
55del	-----GAAGAAGCAGTCACTCTTGTGGAGCGT

Insertions

WT	AGCAGGGTGCAGGAGCTGCGCCCAGAG-AGCGGCGAAGAAGCAGTCACTCTTGTGGAGCG
1insA	AGCAGGGTGCAGGAGCTGCGCCCAGAGAAGCGGCGAAGAAGCAGTCACTCTTGTGGAGCG
WT	AGCAGGGTGCAGGAGCTGCGCCCAGAGA-GCGGCGAAGAAGCAGTCACTCTTGTGGAGCG
1insB	AGCAGGGTGCAGGAGCTGCGCCCAGAGAGGCGGCGAAGAAGCAGTCACTCTTGTGGAGCG
WT	AGCAGGGTGCAGGAGCTGCGCCCAGAGAG-CGGCGAAGAAGCAGTCACTCTTGTGGAGCG
1insC	AGCAGGGTGCAGGAGCTGCGCCCAGAGAGACGGCGAAGAAGCAGTCACTCTTGTGGAGCG
WT	AGCAGGGTGCAGGAGCTGCGCCCAGAGA-----
113ins	AGCAGGGTGCAGGAGCTGCGCCCAGAGAAATTAATACGACTCACTATAGACTGCTTATTC
WT	-----
113ins	TTTATTCTCCGCTCTCGTTTTAGAGCTAGAAATAGTAAGT TAA AATAAGGGTAGTCCGTT
WT	-----GCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT
113ins	ATCAACTTGAAAAAGTGGCACCGAGTCGGCGGCGAAGAAGCAGTCACTCTTGTGGAGCGT

Insertions and Deletions

WT	AGCAGGGTGC----AGGAGCTGCGCCCAGAGAGCGGCGAAGAAGCAGTCACTCTTGTGGA
24del+4	AGCAGGGTGCAGGCA-----AAGAAGCAGTCACTCTTGTGGA
WT	AGCAGGGTGCAGGAGCTGCG-----CCCAGAGAGCGGCGAAGAAGCAGTCACTCTTG
22del+4	AGCAGGGTGCAGGAGCTGCGAACAAACC-----GTCACTCTTG
WT	AGCAGGGTGCAGGA----GCTGCGCCCAGAGAGCGGCGAAGAAGCAGTCACTCTTGTGGA
29del+4	AGCAGGGTGCAGGACCAC-----TCACTCTTGTGGA

Appendix 8.4.5: Weight of placenta (TOP) and E16.5 embryos (BOTTOM) for three different mutant lines from het x het crosses.

