Target-specific epigenetic silencing of endogenous retroviruses in mammals



PhD dissertation

by

Gernot Wolf

Department of Molecular Biology and Genetics

Aarhus University

December 2012

**Table of contents**

[Abstract 6](#_Toc343249294)

[Acknowledgements 7](#_Toc343249295)

[Preface 8](#_Toc343249296)

[Abbreviations 9](#_Toc343249297)

[Introduction 10](#_Toc343249298)

[The retroviral life cycle 10](#_Toc343249299)

[Endogenisation of exogenous retroviruses by germ line infection 11](#_Toc343249300)

[The impact of ERVs on their host organisms 12](#_Toc343249301)

[Retroviral restriction factors 12](#_Toc343249302)

[ERV repression by DNA methylation 15](#_Toc343249303)

[ERV silencing by covalent histone modifications 17](#_Toc343249304)

[Target-specific retroviral silencing 19](#_Toc343249305)

[Aims of the study 25](#_Toc343249306)

[Chapter 1: Epigenetic marking and repression of porcine endogenous retroviruses 26](#_Toc343249307)

[Introduction and summary 26](#_Toc343249308)

[Discussion and future perspectives 27](#_Toc343249309)

[Chapter 2: Murine ERV targeting by the TRIM28/ZFP809 repressor complex 29](#_Toc343249310)

[Introduction and summary 29](#_Toc343249311)

[Materials and methods 30](#_Toc343249312)

[Results 33](#_Toc343249313)

[Diversity and repression activity of ERV PBS sequences in the mouse genome. 33](#_Toc343249314)

[Germ line colonisation by two MmERV subgroups using different PBS sequences 35](#_Toc343249315)

[MmERV-Pro and MmERV-Gly elements are differentially expressed 39](#_Toc343249316)

[TRIM28 binding and repressive histone methylations at ERVs in EC cells 41](#_Toc343249317)

[Conservation of the ZFP809 DNA binding domain in mice and non-murine rodents 43](#_Toc343249318)

[Discussion 47](#_Toc343249319)

[Supplementary data for chapter 2 50](#_Toc343249320)

[Chapter 3: Genome-wide analysis of ZFP809 binding sites 55](#_Toc343249321)

[Introduction and summary 55](#_Toc343249322)

[Materials and methods 57](#_Toc343249323)

[Results 59](#_Toc343249324)

[Genome-wide ZFP809 binding site identification by ChIP-seq 59](#_Toc343249325)

[Distribution of ZFP809 binding sites within repetitive elements 61](#_Toc343249326)

[Identification of a ZFP809 target motif in non-repetitive peaks 66](#_Toc343249327)

[ZFP809 target motifs in repetitive elements 68](#_Toc343249328)

[Distribution pattern of non-repetitive ZFP809 binding sites 71](#_Toc343249329)

[Gene ontology of ZFP809 targeted genes 72](#_Toc343249330)

[Confirmation of ZFP809 binding to genomic targets by ChIP-qPCR 73](#_Toc343249331)

[Regulation of gene expression by ZFP809 76](#_Toc343249332)

[Correlation of ZFP809 peaks with available histone modification ChIP-seq data 77](#_Toc343249333)

[Histone methylations and TRIM28 binding at non-repetitive ZFP809 targets 79](#_Toc343249334)

[Repression activity of ZFP809 target motifs incorporated in retroviral vectors 81](#_Toc343249335)

[Discussion 82](#_Toc343249336)

[Supplementary data for chapter 3 89](#_Toc343249337)

[Chapter 4: Generation of a ZFP809 knock-out mouse model 92](#_Toc343249338)

[Introduction and summary 92](#_Toc343249339)

[Materials and methods 93](#_Toc343249340)

[Results 95](#_Toc343249341)

[Generation of homozygous *Zfp890*-/- mice 95](#_Toc343249342)

[Endogenous VL30-Pro elements are derepressed in *Zfp809*-/- MEFs 96](#_Toc343249343)

[ERV and ZFP809 targeted gene expression in *Zfp809*-/- ES cells 98](#_Toc343249344)

[Derepression of ZFP809 targets in hematopoietic cells of *Zfp809*-/- mice 99](#_Toc343249345)

[Characterisation of the *Gm9705* gene and its ZFP809 target site 100](#_Toc343249346)

[Discussion 102](#_Toc343249347)

[Chapter 5: Epigenetic marking and de-repression of human endogenous retroviruses 106](#_Toc343249348)

[Introduction and summary 106](#_Toc343249349)

[Materials and methods 108](#_Toc343249350)

[Results 109](#_Toc343249351)

[Epigenetic marking of HERVs in human EC and ES cells 109](#_Toc343249352)

[Increased H3K9 acetylation at the *CSF1R* THE1B LTR in T-cell lymphomas 113](#_Toc343249353)

[Repressive histone methylations at the *CSF1R* THE1B LTR in lymphomas 115](#_Toc343249354)

[Lack of H3K9 acetylation at the *CSF1R* THE1B LTR in Burkitt’s lymphomas 116](#_Toc343249355)

[Discussion 116](#_Toc343249356)

[Supplementary data for chapter 5 120](#_Toc343249357)

[Concluding remarks and further perspectives 121](#_Toc343249358)

[References 124](#_Toc343249359)

[Appendix: manuscript under submission 138](#_Toc343249360)

# Abstract

Endogenous retroviruses (ERVs) are transcriptionally repressed by epigenetic silencing yet little is known about the factors ensuring target specificity of this mechanism. In murine embryonic stem (ES) cells and embryonic carcinoma (EC) cells, the Krüppel-associated box (KRAB) zinc finger protein 809 (ZFP809) binds to a defined target sequence within the proline primer binding site (Pro-PBS) of *de novo* integrated retroviruses and recruits the KRAB domain binding corepressor TRIM28, which induces transcriptional repression via histone modifying enzymes. Since several murine ERV families that lack the Pro-PBS sequence are derepressed in TRIM28 knock-out ES cells, it has been speculated that various KRAB-ZFPs function as DNA-binding recognition modules that facilitate target-specific TRIM28 recruitment to ERVs. Indeed, KRAB-ZFPs are present at high copy numbers in mammalian genomes and have diversified in different species through gene duplications and positive selection. In this study, we addressed the hypothesis that ERVs are transcriptionally repressed by PBS-targeting KRAB-ZFPs and aimed to investigate whether KRAB-ZFPs have adapted their DNA-binding specificity to repress newly emerging ERV groups in different mammalian species. Furthermore, we performed a genome-wide analysis of ZFP809 binding sites using ChIP-seq technology and generated a ZFP809 knock-out mouse model to study the *in vivo* function of this protein. Although we found that porcine and human ERVs are marked by the same repressive histone methylations as murine ERVs, we did not find evidence for PBS targeting ERV repression mechanisms in porcine and human cell lines. Our evolutionary analysis of ZFP809 revealed that the DNA-binding specificity of this repressor is conserved in mice, hamsters and blind mole rats, which have separated about 25 million years ago whereas ZFP809 targeted ERVs have entered the mouse germ line only relatively recently in history. We further show that ZFP809 is required for efficient transcriptional silencing of a small subgroup of ERVs in murine embryonic fibroblasts and hematopoietic cells, but not in ES or EC cells. Finally, we identified a ZFP809 binding site, introduced by a retroviral insertion, which functions as a regulatory element that represses transcription of a nearby cytochrome P450 gene. In conclusion, our results suggest that that PBS dependent ERV repression is not a common mechanism in mammals and that ZFP809 may have a dual function as retrovirus repressor protein and transcription factor regulating gene expression.

# Acknowledgements

… go to my main supervisor Finn Skou Pedersen and my co-supervisor Anders Lade Nielsen for their great guidance throughout my PhD period. I am grateful that I was given the trust to work independently and do the research I thought was most interesting. I also want to thank Karin Lykke-Hartmann and Jacob Giehm Mikkelsen, who were members of my PhD committee and substantially contributed to my project during extensive and inspiring discussions at our half-annual committee meetings. I further want to thank our cloning experts Annette and Ernst-Martin Füchtbauer, who spent a lot of time and effort to realise a crucial part of my studies. Acknowledgements also go to Dan Hansen Lambertsen and Andreia Machado da Silva, who have been working with me for a while and did an excellent job on their projects. I am also grateful to Hugo Martins for showing me how to run some very user-unfriendly bioinformatics programs and to Steffen Junker for letting me participate in his exciting work.

I would further like to thank our collaborators from outside of Aarhus University, Stoyan Petkov, Stephen Goff, Sharon Schlesinger, Paul Verma and Jun Liu, for sending me urgently needed materials from all over the world. Acknowledgements also go to our technicians, Ane Kjeldsen and Lone Højgaard, and every single member of the FSP lab for their support and occasional cake or breakfast sessions. I apologise to the same people for not bringing cake at any occasion myself. I am deeply ashamed for that and hope I can still make it up before I have to leave Denmark. I sincerely appreciate the numerous social events that have been organised in our lab, especially the infamous annual trips to Finn’s summer house somewhere in the north-west of Jutland, far away from civilisation. I will treasure the long, windy and rainy beach walks and the great after-parties that made it all worth-wile.

Finally, I want to thank my friends in Aarhus for cheering me up during tough times and listening to my whining about work and life in general over a beer or two, and my family and friends in Austria for letting me enjoy wonderful home leaves, far away from the lab. My special thank goes to Ania, for bringing me back to reality day by day after work which kept me going during the last three years.

# Preface

I have started to work as a PhD student in the laboratory of Finn Skou Pedersen in October 2009. Throughout the three years of my studies time I have been working on several projects about epigenetic silencing of endogenous retroviruses in various mammalian species. During my first year, I was working on target-specific silencing mechanism in porcine embryonic germ cells which was the originally proposed project for my entire PhD program. The aim of this project was to evaluate species-specific differences in retrovirus silencing factors in pigs and mice. Although the porcine embryonic germ cells I have received from our collaborator to conduct these studies seemed to be a promising model for an embryonic stem cell-like cell line when this project was initiated, severe concerns were raised about the suitability of these cells soon after I started to work on this project. Although I continued to work on the originally proposed, but modified, project throughout my entire PhD period, the main focus of my studies was shifted to related projects on epigenetic silencing of murine and human endogenous retroviruses.

The results obtained from my work on porcine endogenous retroviruses are presented in form of a manuscript that has been submitted to *Journal of General Virology* in November 2012 (chapter 1). The manuscript was positively assessed by the reviewers with minor revisions suggested and will be submitted for publication in a revised form after I have handed in my PhD thesis.

Chapters 2-4 include results from my work on target-specific silencing of murine endogenous retroviruses. Since I received results from the two main projects (chapter 3 and 4) only within the last six months of my PhD period, I was unable to prepare manuscripts ready for submission before handing in my PhD thesis. The results from chapters 2-4 are planned to be published after further substantiation in form of two manuscripts.

Chapter 5 includes preliminary results on epigenetic marking of endogenous retroviruses in pluripotent human cells and results from a side-project that was conducted in collaboration with Steffen Junker from the Department of Biomedicine at Aarhus University. The results from this side-project will be included in a manuscript that is currently prepared by Steffen Junker. The remaining results in chapter 5 may be used as a starting point for new projects in our lab but are too preliminary to be considered for publication in the near future.

# Abbreviations

5-Aza 5-azacytidine

Aza-dC 5-aza-2´-deoxycytidine

CYP Cytochrome P450

DNMT DNA methyltransferase

EC Embryonic carcinoma (cell)

EBV Epstein-Barr virus

ERV Endogenous retrovirus

ES Embryonic stem (cell)

HDAC Histone deacetylase

HERV Human endogenous retrovirus

HIV-1 Human immunodeficiency virus type 1

HMT Histone methyltransferase

HRS Hodgkin/Reed-Sternberg

IAP Intracisternal A particle

KRAB-ZFP Krüppel-associated box zinc finger protein

LTR Long terminal repeat

LINE Long interspersed nuclear element

MACS Model-based Analysis of ChIP-Seq

MEF Murine embryonic fibroblasts

MBD Methyl-CpG binding domain

MLV Murine leukemia virus

MmERV *Mus musculus* endogenous retrovirus

MS Multiple sclerosis

ORF Open reading frame

PBS Primer binding site

PERV Porcine endogenous retrovirus

piRNA Piwi-interacting RNA

PRC Polycomb-repressive complex

RBS Repressor binding site

SINE Short interspersed nuclear element

TSA Trichostatin A

UTR Untranslated region

VL30 Virus-like 30

TE Transposable elements

TRIM Tripartite motif

TSS Transcription start site

SNP Single nucleotide polymorphism

ZAP Zinc finger antiviral protein

# Introduction

## The retroviral life cycle

All retroviruses share a common replication cycle that includes the transformation of the RNA genome to double-stranded DNA in a process termed reverse transcription. A typical simple retroviral provirus consists of coding regions for Gag, Pol and Env proteins, flanked by two long terminal repeats (LTRs). The two LTRs, which are identical in a newly integrated provirus, are composed of U3 and U5 regions, separated by a repeated repeat (R) region. The U3 region contains binding sites for cellular transcription factors as well as a polyadenylation (poly-A) signal. Whereas the 5’LTR of a provirus mainly serves as a strong promoter element, the 3’LTR, although of identical sequence, provides the poly-A signal. In typical retroviruses, several *cis-*acting elements, e.g. the primer binding site (PBS) and retroviral packaging signals, are located between the 5’LTR and the coding regions. The PBS is complementary to the 3’ end of cellular tRNAs and serves to prime reverse transcription.

In simple retroviruses, such as murine leukemia virus (MLV), full-length viral mRNA is transcribed from the viral promoter in the U3 region of the 5’LTR and polyadenylated at the poly-A signal in the R region of the 3’LTR. The expressed full-length mRNA contains the coding regions for Gag, Pol and Env proteins. Additionally, a spliced mRNA containing only the *env* gene is exported into the cytoplasm. The translated viral proteins form retroviral particles in which two copies of the viral full-length mRNA are packaged. After budding from the plasma membrane, the viral proteins are processed by the viral protease and the virion matures into an infective virus, ready for infection of a new target cell. The released virus can infect new host cells by binding to a proper cellular receptor via the SU domain of its Env proteins. The Env transmembrane domain facilitates fusion of the viral and cellular membranes in order to release the virion core into the target cell. After fusion, the retroviral RNA genome is reverse-transcribed by the viral polymerase within a structure derived from the virion core to form the preintegration complex. The viral double-stranded DNA is then transported to the nucleus, where the viral integrase mediates stable integration of the provirus.

## Endogenisation of exogenous retroviruses by germ line infection

Although retroviruses are believed to primarily infect somatic cells, retroviral infection followed by genomic integration may also occur in cells of the germ line. In case the infected germ cell successfully develops into offspring, the provirus will be present in all cells of the organism and is now referred to as endogenous retrovirus (ERV). ERVs can be passed on from generation to generation as a Mendelian gene and, in case the ERVs are not removed by natural selection or genetic drift, eventually become fixed in a population or an entire species. Consequently, different species share common ERVs if the retroviral germ line colonisation predates speciation.

Since the first discovery of ERVs in chicken and mice more than 40 years ago, a large number of replication competent and degenerated ERVs have been identified in virtually all mammalian genomes that have been investigated [[2](#_ENREF_2)]. The progress in next-generation sequencing technologies within the last few years confirmed the abundance of ERVs in mammalian genomes and led to the identification of a countless number of ERVs that, as known exogenous retroviruses, can be classified as those with generally simple genomes – alpha, beta, gamma and epsilon retroviruses, and those with more complex genomes – lentiviruses and spumaviruses [[2](#_ENREF_2)]. Although several endogenous lenti- and spumaviruses have been identified [[2](#_ENREF_2),[3](#_ENREF_3)], the vast majority of ERVs has simple genomes [[2](#_ENREF_2),[4](#_ENREF_4)]. About 8% of the human genome was estimated to have derived from retroviral germ line insertions [[5](#_ENREF_5)]. However, with the possible exception of HERV-K (HML2) [[6](#_ENREF_6),[7](#_ENREF_7)], all human ERVs are considered to be inactive due to accumulated mutations and deletions in these elements. In contrast, several groups of murine ERVs have been shown to have replication-competent members which are still active in mice [[8](#_ENREF_8)]. Active ERVs have also been identified in pigs [[9](#_ENREF_9),[10](#_ENREF_10)], and in koalas, a currently still on-going germ line colonisation by replication-competent ERVs has been reported [[11](#_ENREF_11)].

After germ line colonisation, ERVs can continue to replicate either by reinfection of germ line cells through forming infectious viral particles or via an intra-cellular retrotransposition process which may include *in trans* complementation of retroviral proteins by exogenous retroviruses [[12](#_ENREF_12)]. ERVs may lose their ability to form infectious particles on their own and switch from replication-competent pathogens to intracellular parasites, as it has been reported for intracisternal A particle (IAP) ERVs in the mouse genome [[13](#_ENREF_13)]. Other ERVs, such as ETn and virus-like 30 (VL30) elements, have lost their entire coding region and replicate by *in trans* complementation or by co-packaging of the viral RNA into particles formed by exogenous retroviruses, respectively [[14](#_ENREF_14),[15](#_ENREF_15)]. Notably, IAP and ETn ERVs are high copy number ERV families that are responsible for most of the insertional germ line mutations that have been described in mice [[16](#_ENREF_16)].

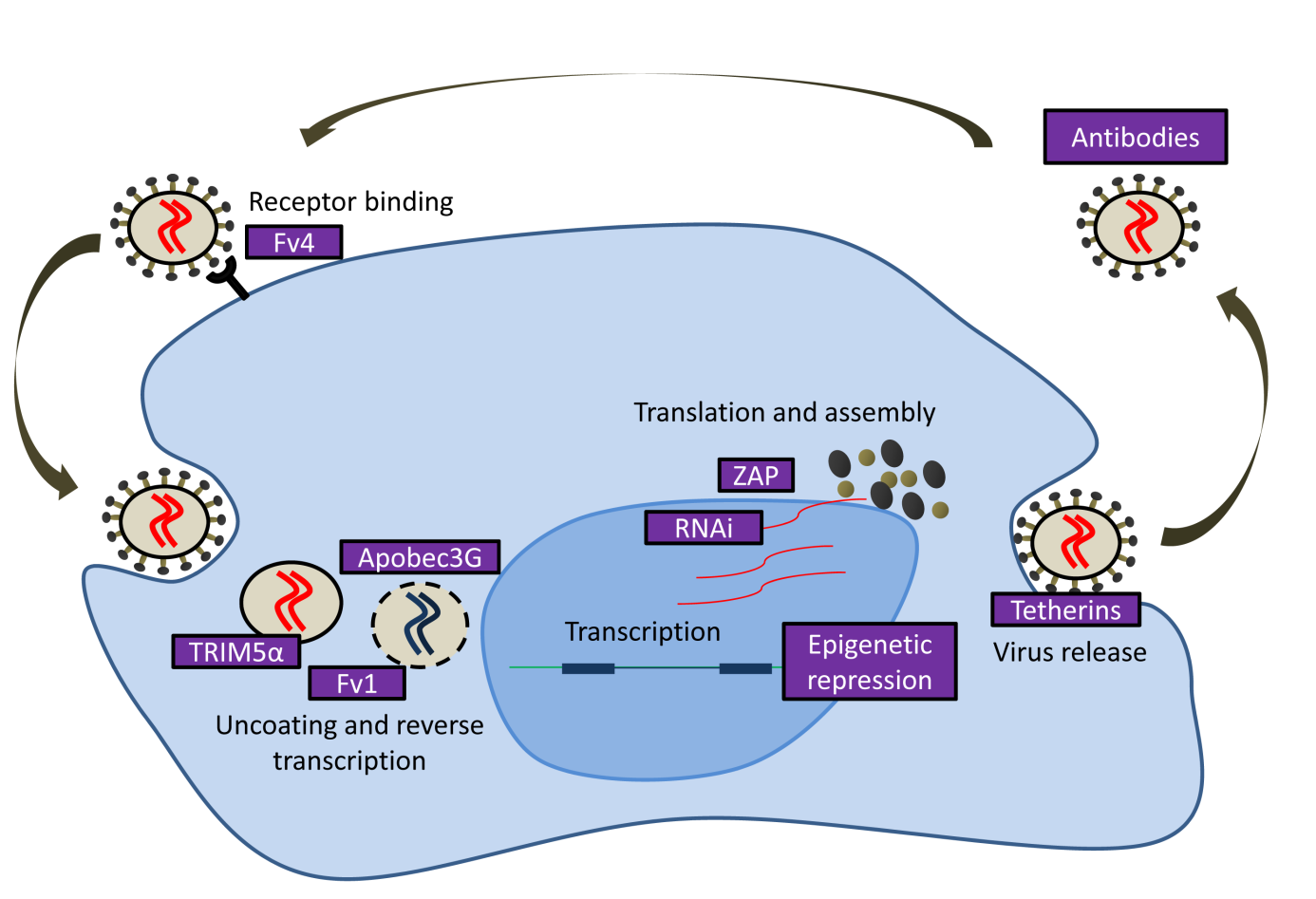
## The impact of ERVs on their host organisms

Originally regarded as purely parasitic DNA, it has become more and more evident that, in numerous cases, ERVs may mediate a positive effect on the host organism. Hosts may benefit from ERVs in several ways: First, ERVs may provide defence mechanisms against infectious exogenous retroviruses (see section below). Second, ERVs can provide *cis-* and *trans*-acting factors that can serve as a dynamic reservoir for new cellular functions. The most prominent examples for the acquisition of a retroviral gene by a host are the *Syncytin* genes in human [[17](#_ENREF_17),[18](#_ENREF_18)] and mice [[19](#_ENREF_19)]. These genes are derived from retroviral *env* genes and are required for placenta formation and immunosuppression during pregnancy [[19](#_ENREF_19)]. Apart from adapted viral genes, regulatory elements on retroviral LTRs can provide new control mechanisms for the expression of cellular genes [[20](#_ENREF_20),[21](#_ENREF_21)]. It has also been suggested that VL30 mRNAs regulate steroidogenesis and possibly other processes by forming a complex with a cellular protein [[22](#_ENREF_22)]. Furthermore, ERVs and other transposable elements (TEs) may play an important role in gene regulation by RNA interference (RNAi) [[23](#_ENREF_23),[24](#_ENREF_24)].

Although more and more cases are known where ERVs fulfil a biological function, there is no doubt that active ERVs represent a threat to the genomic integrity of the infected host. In humans, no active ERVs are known. However, a retroviral LTR was recently identified as an activator of a proto-oncogene in human lymphomas [[25](#_ENREF_25)]. In mice, several active ERVs contribute to an estimated 10% of all *de novo* mutations [[16](#_ENREF_16)]. The mechanisms by which retroviruses can cause defects in cellular functions, are multifaceted and include deletions of host genes, aberrant splicing by providing alternative poly-A signals, overexpression of proto-oncogenes and chromosomal rearrangements by homologous recombination [[16](#_ENREF_16),[20](#_ENREF_20)].

## Retroviral restriction factors

In order to defend their genomes against exogenous and endogenous retroviruses, mammals have developed a wide range of defence mechanisms during evolution. The adaptive immune system, which can react upon infecting parasites such as viruses by producing specific antibodies provides a flexible system that has recently also been associated with ERV restriction [[26](#_ENREF_26)]. Additionally, several antiviral factors that constitute a part of the innate immune system have been described (Fig. 1). As mentioned above, ERVs themselves can provide retroviral resistance factors. Env proteins expressed by ERVs can bind to the receptor that is used by the viruses to enter the host cell. Further infection by exogenous retroviruses that use the same receptor for cell entry is therefore blocked. The best described example for the adaption of a retroviral *env* gene as a retroviral restriction factor is the *Fv-4* gene, which is derived from a truncated endogenous MLV provirus and restricts viral entry of ecotropic MLV [[27](#_ENREF_27)]. A retrovirus-derived gene that provides resistance to infection of a subset of MLV, the *Fv1* gene, was found to encode a Gag-like protein with sequence similarity to a human ERV family (HERV-L) and the MuERV-L family in mice [[28](#_ENREF_28)]. Fv1 targets the viral pre-integration complex and prevents its entry to the nucleus [[29](#_ENREF_29)].



**Figure 1: Steps of the retroviral replication cycle inhibited by retroviral restriction factors**. After entry, the retroviral RNA genome (red) is reverse-transcribed into DNA (blue). After reverse transcription and uncoating the retroviral DNA is integrated into the host cell genome. Viral mRNA is transcribed from the provirus and translated into viral proteins (brown and beige circles). The assembled virus is released from the plasma membrane and can bind to the proper receptor of a new target cell. Several retroviral restriction factors are known to inhibit various steps of the retroviral replication cycle (see text for explanation).

Furthermore, there are several examples of mammalian retroviral restriction genes that are not of retroviral origin and often conserved in different species. TRIM5α, a member of the tripartite motif (TRIM) family of proteins, targets incoming capsids of the human immunodeficiency virus type 1 (HIV-1) in infected cells of Old World monkeys and blocks reverse transcription [[30](#_ENREF_30),[31](#_ENREF_31)]. Interestingly, human TRIM5α fails to restrict HIV-1, but was shown to successfully inhibit a resurrected 4-million year old ERV from the chimpanzee genome [[32](#_ENREF_32)]. This may reflect selective changes of the TRIM5α protein in different primate lineages resulting in slight modifications that determine antiviral specificity. Members of the APOBEC family of cytidine deaminases form a restriction system that is conserved in many mammals including mice, pigs and primates. This enzymatic defence system targets incoming retroviruses by cytidine (C) deamination, leading to a C to uracil (U) mutation in the minus DNA strand that is formed during reverse transcription of the retroviral RNA genome. In turn, a guanine (G) to adenine (A) mutation is introduced in the synthesised sense strand [[33](#_ENREF_33),[34](#_ENREF_34),[35](#_ENREF_35)]. Similar to TRIM5α in primates, adaptive evolution has shaped APOBEC3 in the *Mus* genus, resulting in modified antiviral specificities of this restriction factor in different *Mus* species [[36](#_ENREF_36)].

Retroviral restriction factors can also target late replication steps of infecting retroviruses. The zinc finger antiviral protein (ZAP) was identified in a screen for genes that confer resistance to infection by MLV in mice [[37](#_ENREF_37)]. ZAP is a CCCH-type zinc finger protein that targets expressed viral mRNA and promotes its degradation by the RNA processing exosome [[38](#_ENREF_38)]. Tetherin was found to restrict HIV-1 particle release in the absence of the antagonising HIV-1 encoded protein U (Vpu) [[39](#_ENREF_39)]. This interferon-inducible trans-membrane protein causes an accumulation of viral particles in endosomal vesicles and the retention of viral particles at the cell surface [[39](#_ENREF_39)]. Sequence analysis of Tetherins in various mammals revealed positive selection on the region interacting with viral protein, indicating that also this restriction factor has adapted to new targets in different species [[40](#_ENREF_40)].

The restriction factors mentioned above have been shown to restrict the activity of exogenous retroviruses but may also control certain ERVs that replicate as their exogenous counterparts. However, ERVs can adapt to an intracellular life cycle that removes some of the steps that can be targeted by restriction factors (e.g. virus entry and viral particle release). Furthermore, the adaptive immune system may fail to restrict the activity of ERVs since their expressed proteins might not be recognised as foreign antigens. Nevertheless, ERV amplification depends on the transcription of their genome, and this step seems to be the target of cellular defence mechanisms that have evolved to repress ERVs and other TEs, as discussed below.

## ERV repression by DNA methylation

DNA methylation in mammalian genomes is catalysed by DNA methyltransferases (DNMTs) and occurs almost exclusively at cytosine residues at CpG dinucleotides. CpG methylation can repress transcription either by directly interfering with the binding of transcription factors [[41](#_ENREF_41)] or by recruiting methyl-CpG binding domain (MBD) proteins which can induce transcriptional silencing [[42](#_ENREF_42)].

In mammals, four functional DNMTs have been identified. DNMT1 is the major maintenance DNMT that is responsible for methylation of hemi-methylated DNA after replication [[43](#_ENREF_43)]. Through this mechanism, methylation is inheritable and, in concert with other factors, provides a faithful system of retaining transcriptional silencing from the germ cell throughout the entire life cycle of an organism [[44](#_ENREF_44)]. DNMT3a and DNMT3b are *de novo* DNMTs that can methylate unmethylated, as well as hemi-methylated DNA. DNMT3l has no enzymatic activity itself, but is required for the full activity of DNMT3a and DNMT3b [[45](#_ENREF_45)].

The observation that the majority of methylated CpG dinucleotides lies within TEs led to the speculation that DNA methylation evolved as a defence mechanism against mobile DNA elements [[46](#_ENREF_46)]. Indeed, treatment of dividing cells with the DNA methylation inhibitors 5-azacytidine (5-Aza) or 5-aza-2´-deoxycytidine (Aza-dC) leads to an upregulation of several retrotransposons including ERVs [[47](#_ENREF_47),[48](#_ENREF_48)]. Knock-out of the maintaining DNMT1 was accompanied by a loss of the methylation patterns on endogenous IAP genomes and transcriptional ERV activation in mouse embryos [[49](#_ENREF_49)]. Furthermore, in germ cells of DNMT3l deficient animals, several retrotransposons have been shown to be demethylated and upregulated [[50](#_ENREF_50)]. Expression levels of DNMT3a and DNMT3b are particularly high in ES cells but decrease dramatically upon differentiation. In accordance, proviruses become rapidly methylated and silenced after integration during the pre-implantation stage of early mouse embryos [[51](#_ENREF_51)], whereas retroviral genomes integrating in differentiated cells, such as fibroblasts, remain unmethylated [[52](#_ENREF_52)]. Apart from interfering with transcription factor binding and recruiting MBD proteins, methylated cytosines are also subjected to a higher mutation rate than unmethylated cytosines, leading to an accelerated degeneration of TEs [[53](#_ENREF_53)].

The mechanisms by which *de novo* DNMTs in mammals are targeted to ERVs are poorly understood. However, some indications came from the study of transposon silencing in invertebrate animals. In *Drosophila*, a class of small RNAs, termed Piwi-interacting RNAs (piRNAs), which join Piwi proteins and guide them to specific targets, repress transcription of mobile elements [[54](#_ENREF_54)]. Transposon-derived mRNA that is targeted by the piRNA-Piwi complex is cleaved and also forms a secondary piRNA, which is derived from the mRNA. The secondary piRNA can again target transposon mRNA, transcribed from endogenous loci and produce original anti-sense piRNA that targets other transposon RNAs. This amplification cycle, which was termed “ping-pong cycle”, allows efficient targeting of transcribed transposons [[54](#_ENREF_54)]. Importantly, since the targeting piRNAs are derived from TE themselves, this adaptive system also allows repression of newly emerging TEs.

Several components of the piRNA system are also conserved in vertebrates. In mice, dispersed copies of retrotransposons serve as templates for the production of piRNAs [[55](#_ENREF_55)]. Knock-down of the mouse Piwi protein MILI resulted in derepression and hypomethylation of L1 and IAP retrotransposons in foetal testes [[55](#_ENREF_55),[56](#_ENREF_56)]. Interestingly, derepression of IAP and LINE retrotransposons in DNMT3l depleted mice seems to be linked to an increase of piRNAs, derived from these elements [[57](#_ENREF_57)]. However, the mechanism by which piRNAs repress TEs in mammals is little understood and it remains unknown whether piRNAs are targeting ERVs other than IAP.

Although DNA methylation represents a faithful ERV repression mechanism in most tissues and cell types, it may fail during certain developmental stages in mammals. After fertilisation, the paternal and maternal pronuclei undergo rapid active and slow passive demethylation, respectively, followed by passive demethylation of both genomes. At the morula stage, the global DNA methylation reaches its lowest level, followed by de novo DNA methylation during development to the blastocyst stage. Global DNA demethylation also takes place in female and male primordial germ cells after they start migrating to the genital ridge around E7.5 [[58](#_ENREF_58)]. DNA demethylation is a little understood process. No enzyme is known that catalyses the demethylation of cytosines in a single step reaction. However, TET proteins catalyse oxidation of methylated cytosine (5mC) to 5-hydroxylmethylcytosine (5hmC) and further oxidation products such as 5-carboxylcytosine (5caC). Since thymine-DNA glycosylase (TDG) can excise 5caC and, thus, allowing replacement of the modified base with unmethylated cytosine through base excision repair, 5hmC was proposed to represent an intermediate step in the DNA demethylation process that occurs in the preimplantation embryo [[59](#_ENREF_59)]. These DNA demethylation cycles render the host vulnerable to ERV invasion and amplification since ERV amplification in ES cells and germ cells may also lead to germ line amplification of these elements. Furthermore, Since low levels of 5mC have been detected at IAP elements in murine ES cells [[60](#_ENREF_60)], it has been speculated that some ERVs may be capable of escaping DNA methylation and repression via recruitment of TET proteins [[59](#_ENREF_59)].

Mammals seem to have developed additional repression mechanisms that protect the organism during developmental stages in which DNA methylation may fail to repress ERVs. Indeed, *de novo* integrated proviruses are transcriptionally silenced in ES cells several days before DNA methylation can be detected [[58](#_ENREF_58)] and it has been recently shown that ERV repression in ES cells does not depend on DNA methylation [[61](#_ENREF_61),[62](#_ENREF_62)]. In the following section, I will discuss DNA methylation-independent ERV repression by histone modifications.

## ERV silencing by covalent histone modifications

In eukaryotic cells, genomic DNA is packaged with histones and other proteins to form chromatin. The basic unit of chromatin in the nucleosome consists of an octamer of the four core histones (H3, H4, H2A, H2B) which are wrapped around by 146 base pairs of DNA [[63](#_ENREF_63)]. So far more than 60 different residues on histones have been described that can be enzymatically modified by various covalent modifications [[64](#_ENREF_64)]. Besides function in replication, DNA repair and general chromatin orchestration, histone modifications are known to constitute a transcriptional control mechanism that allows eukaryotic cells to regulate gene expression. Histone modifications can affect transcription either by altering the compactness of the chromatin structure or by recruiting transcription factors and proteins that specifically bind to modified histones. Histone acetylation at lysine residues of histones introduces a negative charge to the nucleosome, thus neutralizing the basic charge of the lysine and decreasing the affinity of the nucleosome to the negatively charged DNA. This leads to an open chromatin stage that is accessible for transcription factors and polymerases [[65](#_ENREF_65)]. Reversal of lysine acetylation by histone deacetylases (HDACs) is therefore correlated with transcriptional repression. Histone deacetylation was also associated with ERV silencing since murine embryonic fibroblasts (MEFs) treated with an inhibitor of histone deacetylation, showed a strong upregulation of VL30 ERVs [[48](#_ENREF_48)].

Histone methylation can either activate or repress transcription, depending on the methylated residue. H3K4 methylation is associated with active gene promoters whereas H3K9 methylation is generally found in tightly packed heterochromatin that excludes transcription factor binding and is associated with transcriptional silencing [[64](#_ENREF_64)]. As the majority of histone modifications, lysine methylation is reversible and can be removed by histone demethylases such as LSD1. Concordantly, LSD1 can either repress transcription by demethylating H3K4 [[66](#_ENREF_66)] or activate gene expression by demethylation of H3K9 [[67](#_ENREF_67)]. In LSD1 knock-out ES cells, a strong upregulation of endogenous MERVL elements was observed [[68](#_ENREF_68)], indicating that H3K4 demethylation may play an important role in ERV silencing.

H3K9 methylation is a hallmark of transcriptionally silent heterochromatin which is known to contribute to several biological processes such as X chromosome inactivation, gene imprinting, nuclear organisation and transcriptional regulation of gene expression [[69](#_ENREF_69)]. However, heterochromatin silencing has been suggested to have originally evolved as a control mechanism to prevent recombination between repetitive DNA elements [[69](#_ENREF_69)]. Earlier studies that have described repressive histone methylations at LTR elements in murine ES cells [[70](#_ENREF_70)] have been later confirmed by chromatin immunoprecipitation (ChIP) in combination with high-throughput sequencing (ChIP-seq). ChIP-seq analysis revealed that ERVs are generally marked by both the repressive histone modifications trimethylation of lysine-9 on histone 3 (H3K9me3) and trimethylation of lysine-20 on histone 4 (H4K20me3) in murine ES cells but not in differentiated MEFs and neural progenitor cells [[71](#_ENREF_71)]. A recent study re-evaluated the raw sequencing data obtained by Mikkelsen *et al.* [[71](#_ENREF_71)] using improved computational methods for the enrichment of repetitive elements in ChIP-seq data and revealed that ERV-K and ERV1 families are preferably marked by H3K9me3 and H4K20me3 in ES cells whereas the ERV-L and MaLR families are rather marked by histone H3 lysine-27 trimethylation (H3K27me3) [[72](#_ENREF_72)]. The authors also found that several ERV types in MEFs but not neural progenitor cells remain significantly enriched in the H3K9me3 mark, indicating that ERV silencing by histone methylation is not strictly restricted to undifferentiated ES cells [[72](#_ENREF_72)].

H3K27 trimethylation is catalysed by Polycomb complexes and has been associated with transcriptional repression of developmental regulators in murine and human ES cells [[73](#_ENREF_73),[74](#_ENREF_74)]. Recently, it has been shown that depletion of both of the Polycomb-repressive complexes PRC1 and PRC2 leads to a strong upregulation of endogenous MLV elements (which belong to the ERV1 subfamily) [[75](#_ENREF_75)] yet upregulation of other ERVs has not been investigated in these cells so far.

Although knock-out of the histone methyltransferases (HMTases) Suv39h1 and Suv39h2 resulted in a loss of the H3K9me3 mark at major satellite repeats, no H3K9me3 reduction was observed at ERVs, indicating that different HMTs may be required for various heterochromatin targets [[76](#_ENREF_76)]. Recently, knock-out of the HMTase SETDB1 (also known as ESET or KMT1E) was shown to result in a general loss of the H3K9me3 mark at ERV genomes in murine ES cells, accompanied by transcriptional upregulation of various ERVs such as IAP and MLV [[77](#_ENREF_77)]. SETDB1 knock-out also resulted in a reduction of the H4K20me3 mark, which was shown to exhibit a similar distribution in murine ES cells as H3K9me3 [[71](#_ENREF_71)]. However, although knock-out of the H4K20 HMTases Suv420h1 and Suv420h2 in murine ES cells led to a loss of the H4K20me3 mark at ERVs, no H3K9me3 reduction or ERV upregulation was observed in these cells [[77](#_ENREF_77)], indicating that trimethylation of H4K20 is rather a secondary effect downstream of H3K9 trimethylation and not essential for ERV silencing. Importantly, the H3K9me3 mark at ERVs remained unaffected in ES cells depleted of all three catalytically active DNMTs [[77](#_ENREF_77)], indicating that H3K9 trimethylation is not merely a consequence of DNA methylation. On the other hand, DNA methylation of some ERVs was moderately reduced in SETDB1 knock-out ES cells [[77](#_ENREF_77)], suggesting that H3K9 trimethylation by SETDB1 might also affect DNA methylation of ERVs.

The mechanism, by which H3K9 trimethylation represses ERV transcription, is not fully understood yet. A large number of proteins (e.g. HP1) that are known to bind to methylated histones were found to be dispensable for ERV silencing in ES cells [[78](#_ENREF_78)]. It was therefore speculated that H3K9 trimethylation simply prevents acetylation at this residue and, thus, transcriptional activation by chromatin opening [[78](#_ENREF_78)].

## Target-specific retroviral silencing

Most TEs, such as long and short nuclear interspersed elements (LINEs and SINEs), are engaged in a long-term evolutionary relationship with their hosts, allowing them to evolve defence mechanism that restrict TE activity. Unlike LINEs and SINEs, many ERVs express Env proteins, which enable these elements to transfect new host cells, individuals and even new species via horizontal transmission. Due to this ability, ERVs can be introduced to species which have not been confronted with these elements before and may therefore be unable to recognise and silence these ERVs. Furthermore, ERVs may return to replicate as exogenous retroviruses that, after replication cycles in which they mutate at a much higher frequency than their hosts, can re-infect the germ line. ERVs therefore need to be considered as rather unstable elements that may not be sufficiently recognised and repressed by the mechanisms that have evolved to restrict other TEs such as SINEs and LINEs. ERVs may have forced mammals and other vertebrates to develop flexible defence mechanisms that can rapidly adapt to newly emerging ERV groups in the genome.

In the late 1970s, it was observed that murine EC cells, a widely used cell model for ES cells, are non-permissive for expression of *de novo* integrated MLV [[79](#_ENREF_79)]. Although infection of EC cells led to efficient integration of the provirus, expression from the viral promoter was heavily impaired. This is partially due to the lack of the proper LTR enhancer binding transcription factors in these cells [[80](#_ENREF_80)]. More importantly, a 17 bp long sequence, named the repressor binding site (RBS), which is contained within the 18 bp of the MLV proline PBS (Pro-PBS), was found to be the major target of MLV repression in EC cells [[81](#_ENREF_81),[82](#_ENREF_82)]. A virus with a single G to A mutation in the RBS/PBS, known as the B2 mutation, was isolated and found to be relieved from repression [[81](#_ENREF_81)].

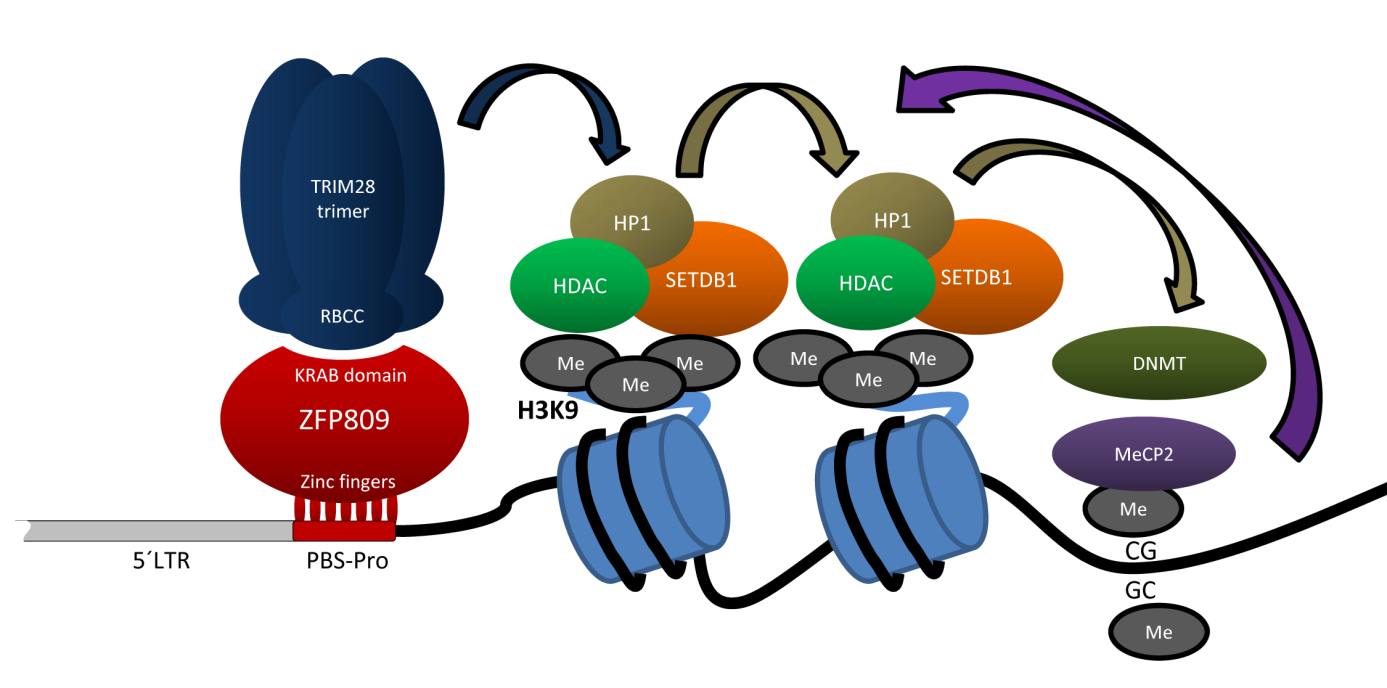
Several years later, a repressor complex was found to bind to the Pro-PBS *in vitro* [[83](#_ENREF_83),[84](#_ENREF_84)]. A screen of various PBS sequences for their ability to repress transcription in EC cells revealed that also the lysine1,2 PBS (Lys1,2-PBS), used by Mason-Pfizer monkey virus, visna virus and spuma virus, contains a RBS that is targeted by a repressor complex *in vivo* and *in vitro* [[84](#_ENREF_84)]. Importantly, the two repressor complexes targeting the Pro-PBS and Lys1,2-PBS did not cross-compete in *in vitro* binding studies, suggesting that they were not identical [[84](#_ENREF_84)]. In later studies, Pro-PBS targeted repression was also observed in primary murine hematopoietic stem cells and embryonic fibroblasts [[85](#_ENREF_85)]. Importantly, the Pro-PBS is able to induce transcriptional repression of various viral and cellular promoters, independent of its orientation or the vector backbone [[84](#_ENREF_84),[85](#_ENREF_85)]. Although it has been claimed that the Pro-PBS is targeted in human hematopoietic cells from the umbilical cord [[85](#_ENREF_85)], the repression activity in these cells was rather weak and this hypothesis has not been supported by further evidence so far.

In 2007, the tripartite motif-containing protein 28 (TRIM28), also known as KAP1 or TIF1-β, was identified as an integral component of the repressor complex targeting the Pro-PBS in mice [[86](#_ENREF_86)]. TRIM28 was also identified as a component of the repressor complex that targets the Lys1,2-PBS [[87](#_ENREF_87)]. TRIM28 is known to represses transcription via recruitment of histone modifying factors such as the NuRD HDAC complex and SETDB1 via its bromodomain and COOH-terminal plant homeodomain (PHD), respectively [[88](#_ENREF_88),[89](#_ENREF_89)]. TRIM28 also recruits the heterochromatin protein 1 (HP1) through a PxVxL motif in the HP1 binding domain [[90](#_ENREF_90)], which was shown to be essential for Pro-PBS targeted silencing [[91](#_ENREF_91)].

TRIM28 lacks DNA binding domains but can be recruited to genomic targets via interaction of its RING-B box-coiled coil (RBCC) domain with the Krüppel-associated box (KRAB) domains of KRAB-zinc finger proteins (KRAB-ZFPs) [[92](#_ENREF_92)]. The KRAB-ZFP family constitutes the largest group of transcriptional repressors in higher organisms and has rapidly expanded and diverged through multiple rounds of gene duplications after first appearing in early tetrapods [[92](#_ENREF_92),[93](#_ENREF_93)]. KRAB-ZFPs contain up to more than 30 C2H2 zinc finger domains, each of them contacting three to four nucleotides, allowing these transcription factors to bind to genomic regions in a highly specific manner [[92](#_ENREF_92)]. Indeed, a KRAB-ZFP, ZFP809, was identified as the recognition module that targets the Pro-PBS and recruits TRIM28 [[94](#_ENREF_94)]. Knock-down of ZFP809 in murine F9 EC cells led to a loss of Pro-PBS repression activity and ZFP809 overexpression in human 293A cells was sufficient to induce Pro-PBS targeted silencing in these cells [[94](#_ENREF_94)]. Soon after TRIM28 and ZFP809 had been identified as components of the Pro-PBS targeting repressor complex, TRIM28 knock-out has been shown to be accompanied by a strong upregulation of various ERV groups in murine ES cells [[95](#_ENREF_95)]. Whereas TRIM28 enrichment at endogenous retroviruses was not impaired in SETDB1 knock-out ES cells, TRIM28 depletion led to a loss of SETDB1 at ERVs, indicating that SETDB1 acts downstream of TRIM28 [[77](#_ENREF_77)]. TRIM28 binding to ERVs was unaffected in DNMT depleted ES cells [[77](#_ENREF_77)], indicating that TRIM28 mediated silencing in ES cells is independent of DNA methylation.

In agreement with the model of TRIM28 mediated silencing by histone modifications, endogenous IAP elements in TRIM28 knock-out ES cells show reduced levels of H3K9me3 [[95](#_ENREF_95)]. IAP elements use various PBS sequences that are complementary to phenylalanine tRNAs, but these sequences were not sufficient to induce silencing of a reporter gene [[95](#_ENREF_95),[96](#_ENREF_96)]. However, a 500 bp long fragment within the 5’UTR of IAP elements induced TRIM28 dependent silencing in EC cells at comparable levels as the Pro-PBS [[95](#_ENREF_95)], indicating that retroviral repression can be targeted by TRIM28 containing repressor complexes to retroviral sequences other than the PBS.

The TRIM28/KRAB-ZFP silencing complex can act over long distances [[97](#_ENREF_97)] since heterochromatin marks and heterochromatin associated proteins are able to spread out via a self-propagating loop (Fig. 2). PBS targeted ERV silencing can therefore lead to heterochromatin formation throughout the entire retroviral promoter, as well as viral coding regions and flanking genomic regions. HP1 was found to be a crucial factor for heterochromatin spreading. It is targeted to H3K9me3 marks by the histone K9 methyltransferase Suv39h1 and, in turn, recruits more Suv39h1 [[98](#_ENREF_98),[99](#_ENREF_99)]. Surprisingly, HP1 was found to be dispensable for H3K9 trimethylation of ERVs in ES cells [[78](#_ENREF_78)], indicating that the mechanisms by which TRIM28 represses newly integrated retroviruses and ERVs might be not identical. TRIM28/KRAB-ZFP mediated silencing may also lead to DNA methylation via HP1 and Suv39h1, which can both recruit DNMTs [[100](#_ENREF_100),[101](#_ENREF_101)], and thus mediate inheritable transcriptional repression (Fig. 2). Importantly, the MDB protein MeCP2 was shown to associate with methylated MLV proviruses and recruit HDACs [[102](#_ENREF_102)], during heterochromatin formation at these elements.



**Figure2: Model of Pro-PBS targeted retroviral silencing and heterochromatin spreading**. ZFP809 binds specifically to the RBS within the retroviral Pro-PBS via its seven zinc fingers and recruits TRIM28 through the KRAB domain. TRIM28 recruits HDACs and SETDB1, which catalyse H3K9 deacetylation and trimethylation, respectively. Trim28 also interacts with HP1, which recruits more SETDB1, HDACs and HP1, leading to heterochromatin spreading. HP1 and SETDB1 are able to recruit DNMTs, which methylate cytosines at CpG sites. Methylated CpG dinucleotides are bound by CpG binding proteins (e.g. MeCP2), which in turn, can recruit more histone modifying enzymes. ZFP809/TRIM28 initiated heterochromatin can spread in both directions and may lead to heterochromatin formation over the entire provirus.

Although retroviral vectors containing a Pro-PBS target site for ZFP809 binding are silenced faster and more efficiently than vectors lacking this sequence, it has been shown that also proviral DNA without the Pro-PBS is transcriptionally repressed by histone modifications and DNA methylation in EC and ES cell over time [[77](#_ENREF_77),[103](#_ENREF_103)], indicating that ZFP809 independent repression mechanisms are active in these cells.

The finding that ZFP809 targets a retroviral sequence [[94](#_ENREF_94)] and the observed upregulation of ERVs in murine TRIM28 knock-out ES cells [[95](#_ENREF_95)] led to the speculation that ERVs might be generally repressed in a target-specific manner by KRAB-ZFPs in mammalian germ cells [[95](#_ENREF_95),[104](#_ENREF_104)]. Indeed, KRAB-ZFPs are the most diverse transcription factors in mammals and only few of these proteins are well characterised [[92](#_ENREF_92)]. Furthermore, signs of positive selection at the binding specificity conferring residues of primate KRAB-ZFPs have been found [[95](#_ENREF_95)] and a correlation between the abundance of ERVs and the number of KRAB-ZFPs in mammalian genomes has been recently reported [[104](#_ENREF_104)]. Although KRAB-ZFPs are beyond dispute not solely restricted to retroviral repression, it is tempting to speculate that the rapid evolution of at least some KRAB-ZFPs may be the result of an evolutionary arms race between retroviruses and KRAB-ZFPs.

However, replicating exogenous retroviruses evolve rapidly and may easily escape KRAB-ZFP mediated repression. Even a functional KRAB-ZFP targeted retroviral element, such as the PBS, which is relatively conserved within an ERV group, may be easily replaced by alternative sequences to escape repression [[96](#_ENREF_96),[105](#_ENREF_105),[106](#_ENREF_106)]. Furthermore, it has been shown that replication competent MLV using PBS that are not repressed by ZFP809 replicate at similar rates to the Pro-PBS containing wild-type in infected mice [[105](#_ENREF_105)], suggesting that ZFP809 does not restrict retroviruses during the course of a regular infection. Unlike exogenous retroviruses, ERVs mutate at the same rate as the host genome and may therefore be less prone to escape target-specific repression. One might further speculate that repression of infecting retroviruses does not favour retroviral escape mutants if the repression mechanism is absent in the primary target cells in which the retrovirus replicates during the course of an infection. It was therefore assumed that ERVs that amplify within the host germ line after initial colonisation and reach a critical number of members will exert a selective pressure on the host organism that drives the evolution of a target-specific KRAB-ZFP restricting the ERV group [[104](#_ENREF_104)]. Furthermore, since new ERV groups are repeatedly introduced to a host genome over time, new KRAB-ZFPs need to evolve to repress these elements. Alternatively, ERV targeting KRAB-ZFPs may adapt to those new targets if their originally targeted ERV group has lost its potential to harm the host.

# Aims of the study

Different mammalian species have been colonised by different groups of ERVs and are therefore expected to have evolved their own set of KRAB-ZFPs that specifically target these ERVs for repression. The main goal of this study was to understand the evolutionary and functional aspects of adaptive, target-specific ERV repression by KRAB-ZFPs.

We first aim to identify PBS sequences frequently used by porcine and human ERVs and test whether these sequences represent targets for KRAB-ZFP/TRIM28 mediated repression in murine, porcine and human cells. Furthermore we want to identify and analyse the retroviral PBS sequences that are most abundant in the mouse genome and may therefore be good candidates for previously unknown targets for murine repression factors. To investigate the role of ZFP809 in ERV silencing we will identify murine ERV groups which use a Pro-PBS and analyse TRIM28 binding and repressive histone modifications at these elements. We also aim to investigate whether ZFP809 has undergone positive selection to adapt its binding-specificity to ERV groups that have colonised the mouse genome. Since it is possible that ZFP809 does not exclusively targets retroviral Pro-PBS sequences, we will perform a genome-wide binding analysis of the genomic ZFP809 binding sites using ChIP-seq technology to identify potential alternative ZFP809 target sites. Furthermore, expression data available on the BioGPS database (http://biogps.org) indicate that ZFP809 is expressed in a large number of tissues at comparable levels as in ES cells, indicating that germ line restriction of ERVs might be only one out of several functions of ZFP809. We therefore aim to generate a ZFP809 knock-out mouse that should allow us to study the *in vivo* functions of ZFP809. This model will also be used to investigate the functional consequences of ZFP809 binding to genomic targets identified by ChIP-seq analysis.

# Chapter 1: Epigenetic marking and repression of porcine endogenous retroviruses

## Introduction and summary

Porcine ERVs (PERVs) are of special interest since viral particles derived from PERV elements are able to infect and replicate in human cells [[10](#_ENREF_10),[107](#_ENREF_107)]. Furthermore, it has been shown that PERVs can infect immunodeficient mice after transplantation of foetal pig pancreatic cells [[108](#_ENREF_108)]. The reported trans-species infection of PERVs is a major concern since xenotransplantations from pigs to humans are widely investigated to alleviate the shortage of human donor organs [[109](#_ENREF_109)]. Although trans-species infections of PERVs to humans have not been reported in porcine cell or organ transplantation trials [[110](#_ENREF_110),[111](#_ENREF_111)], concerns remain about the safety of xenotransplantation using porcine organs [[109](#_ENREF_109)]. Surprisingly, very little is known about the putative epigenetic repression mechanisms controlling PERV expression.

Initially we attempted to identify PERV associated PBS sequences that are targeted by transcriptional silencing in porcine embryonic germ cells (pEGCs) and investigate species-specific differences in the restriction factors that bind these sequences in pigs and mice. We assumed that pigs have developed a different set of KRAB-ZFPs from mice since these two species have been colonised by different ERV families. However, none of the tested PERV PBS sequences induced transcriptional silencing in the investigated porcine cells. Although the pEGCs that have been chosen for this study seemed to be a promising model for an ES-like porcine cell line when the study was initiated, major concerns about these cells were raised when it was later shown that they do not express many of the core pluripotency markers [[112](#_ENREF_112)].

We therefore aimed to investigate more general transcriptional repression mechanisms that have been studied for murine and human, but not porcine ERVs. In the attached manuscript, we conducted a genome-wide *in silico* screen of the pig genome to identify retroviral elements and characterise several PERV families of which some have only been poorly described before. Using the sequence information obtained by our *in silico* screen, we designed primers specifically amplifying several PERV families. These primers were used to analyse PERV expression levels in porcine cells treated with inhibitors of DNA methylation and histone deacetylation. Furthermore we analysed chromatin modifications at these elements via ChIP.

In the attached manuscript (see Appendix), we show that PERV-B, but not PERV-A, PERV-C or two unrelated PERV families are upregulated in primary porcine fibroblast upon inhibition of DNA methylation. We further show that PERVs are marked by repressive histone modifications that have previously been identified at murine ERVs. Interestingly, we found that the only PERV family that is known to be still active in pigs was marked by higher levels of repressive histone methylations than inactive PERV families which are expressed at several magnitudes’ lower levels. Although not conclusive, our results indicating a lack of PBS dependent repression in porcine cells were also included in the manuscript.

## Discussion and future perspectives

Although we have shown that, as for murine ERVs, some PERVs are upregulated upon DNA methylation and marked by repressive histone methylations, we were unable to identify a retroviral target sequence that may initiate transcriptional repression of integrated exogenous retroviruses. Due to the limitations of the used cell lines, the observed lack of PBS-dependent silencing mechanisms is inconclusive since putative repressor proteins may simply not be expressed in these cells. Nevertheless, the absence of a ZFP809-like repressor that targets the Pro-PBS in pigs would not be unexpected since we found that only one subgroup of a low copy PERV-γ1 family (PERV-C) contains intact Pro-PBS sequences. PERV-C elements have entered the pig germ line only recently as indicated by the observation that PERVs have not been fixed in pigs [[113](#_ENREF_113),[114](#_ENREF_114)]. Therefore, the presence of a Pro-PBS binding repressor in pigs seems unlikely since the evolution of a target-specific retroviral restriction factor would require a selective pressure that constantly affects the fitness of a species for many generations. However, the oldest PERV-γ1 and PERV-γ2 elements are believed to have colonised the pig genome several million years ago [[115](#_ENREF_115),[116](#_ENREF_116)], but their commonly used glycine-PBS sequences are not targeted by repression in the analysed cells either.

When this study was initiated by the end of 2009, the recent identification of a PBS targeting KRAB-ZFP suggested that the PBS may be the preferred target sequence for retroviral repressor proteins. However, in 2010 it was shown that IAPS are silenced by TRIM28-dependent repression, targeted to a putative target sequence in the 5´UTR in these elements [[95](#_ENREF_95)]. If KRAB-ZFP/TRIM28 repressor complexes also target retroviral sequences that are not overlapping with the retroviral PBS, the number of possible PERV target sequences becomes simply too large to test all these sequences via transduction assays, as performed in this study.

We have also attempted to detect repressor complex binding in porcine cells using electrophoretic mobility shift assays (EMSA). Although we could readily detect the previously described Pro-PBS binding complex in murine F9 cells, no such shift was observed in pEGCs (these results were excluded from the submitted manuscript). Although we detected some binding activity in pEGC nuclear extracts to other PBS sequences, these complexes were of a significantly higher mobility than the originally described repressor complex and the same shift were observed in murine F9 cells. The identification of proteins and complexes binding to oligo probes in EMSA is technically challenging and, in our opinion, not promising since we did not observe repression activity of these sequences in transduction assays. Furthermore, as for the retroviral transduction assays, the results obtained by EMSA are inconclusive since we did not have a reliable and stable ES-like porcine cell line and the large number of potential target sequences remains a major technical obstacle.

Altogether, the approach to identify target- and possibly species-specific transcriptional repressor proteins in pigs by transduction assays and EMSA was not only hampered by the lack of a suitable cell line but is also unlikely to succeed due to the enormous number of possible target sequences that may be targeted by repression. A perhaps more promising approach could be the generation of TRIM28 knock-out porcine ES cells, as it has been described for murine ES cells [[95](#_ENREF_95)], and detect upregulated PERV sequences by large scale sequencing. However, this approach is biased since only PERVs, which retained intact promoter and enhancer elements necessary for efficient transcription would be identified as targets of TRIM28-dependent silencing. Moreover, a suitable stable porcine ES cell line is not available so far and the generation of TRIM28 deficient knock-out ES cells is technically challenging since these cells lose the ability of self-renewal and, thus, their pluripotent characteristics [[95](#_ENREF_95)]. Further research on epigenetic repression of PERVs should therefore rather focus on medically relevant aspects of PERV silencing in porcine cells and organs that could potentially be used for transplantation into humans. Our study may inspire such studies and ultimately contribute to a better understanding of epigenetic PERV silencing and the potential risk of PERV derepression in xenotransplantation.

# Chapter 2: Murine ERV targeting by the TRIM28/ZFP809 repressor complex

## Introduction and summary

The identification of a KRAB-ZFP as a retroviral restriction factor together with the observed redundancy and diversity of KRAB-ZFPs in mammals led to the speculation that at least some KRAB-ZFPs have evolved to repress potentially hazardous TEs such as ERVs [[94](#_ENREF_94),[95](#_ENREF_95)]. This hypothesis was supported by the observation that TRIM28 knock-out ES cells show elevated expression of several ERVs such IAP and MusD [[95](#_ENREF_95)]. Furthermore, a correlation between the abundance of ERVs and the number of KRAB-ZFPs in mammalian genomes has been reported recently [[104](#_ENREF_104)].

In this chapter, we address the hypothesis that murine ERVs are repressed by PBS-targeting repression mechanisms. We first systematically analysed the PBS usage of murine ERVs by computational data mining and tested frequently used PBS sequences for repression activity. Although all tested PBS sequences efficiently primed reverse transcription of retroviral vectors, none of them, except for the Pro-PBS control, induced repression in F9 EC cells. To assess the impact of ZFP809-dependent silencing on ERVs, we analysed expression levels, epigenetic modifications and germ line amplification rates of proviral elements of the *Mus musculus* ERV (MmERV) group using a combination of computational and experimental analysis methods. Phylogenetic reconstruction revealed that MmERV elements form two closely related subgroups using two different PBS types, of which only the Pro-PBS induced transcriptional repression. Nevertheless, both subgroups have colonized the mouse genome at comparable rates and were both marked by repressive histone modifications in ES and EC cells. Finally, we identified conserved ZFP809 orthologues in wild *Mus* species and two non-murine rodents and showed that the target specificity of ZFP809 has evolved prior to the MmERV germ line colonization. In conclusion, our results indicate that: (i) ZFP809 is involved in, but not crucial for epigenetic marking of some ERVs, (ii) germ line colonization by the MmERV family was not prevented by PBS-dependent repression and (iii) ZFP809 did not adapt to new ERV targets by positive selection.

## Materials and methods

#### Cell culture

F9 EC cells were grown on gelatine-coated culture plates in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% FBS and antibiotics. NIH/3T3 fibroblasts were cultivated in DMEM supplemented with 10% new born calf serum and antibiotics. PlatE packaging cells [[117](#_ENREF_117)] were grown in same medium as F9 cells but on uncoated culture plates. Murine CJ7 ES cells were grown on inactivated feeder cells as described previously [[118](#_ENREF_118)]. ES cells were transferred to gelatine-coated feeder cell-free plates two days before analysis.

#### Plasmids and vectors

The Akv-MLV derived retroviral vectors pPBS-Pro and pPBS-B2 are described elsewhere [[96](#_ENREF_96)]. The vectors using alternative PBS sequences were constructed by an overlapping PCR approach as described in chapter 1 (see attached manuscript). PBS sequences (mLeu1, mLeu1.2, mLys3 and mThr3) are given in Table 2.1. The additional tested PBS sequences were as follows: mHis9 (TGGTGCCGTGACTCGGAT), mGlu13 (TGGTTCCCTGACCGGGAA), mLeu7 (TGGTGTCAGAAGTGGGAT), hGly2/hArg2 (TGGTGCATTGGCCAGGAA) and hSeC1 (TGGTGACCACAAAGGGAC). The vector expressing N-terminal FLAG-tagged ZFP809(1-353) on a pCDNA3.1 background was a kind gift of Stephen Goff [[94](#_ENREF_94)]. Since transgene expression levels from this vector were found to be unsatisfying in F9 cells, the *ZFP809*(1-353) ORF was cut out with *NheI* and *NotI* and cloned into the pIRESpuro3 expression vector (Clontech). To improve integration efficiency, a fragment including the full ZFP809(1-353) ORF and a part of the IRES was then amplified from the pIRESpuro vector and inserted between the *NcoI* and *ApaI* restriction sites of the Sleeping Beauty (SB) transposon-based vector pSB/PGIP which was previously generated by replacing the RSV promoter of pSB/RGIP [[119](#_ENREF_119)] with a PGK promoter. By inserting the ZFP809 ORF, the GFP cassette of pSB/PGIP was removed. From the resulting vector, pSB/FLAG-MmZFP809, the FLAG-tagged *Zfp809*(1-353) and puromycin genes were expressed as a bicistronic mRNA from the PGK promoter. The promoter and the expression cassette were flanked by SB inverted repeats allowing transposition of the expression unit by the hyperactive SB100X transposase, expressed from a pCMV(CAT)T7 vector backbone (pCMV-SB100X) [[120](#_ENREF_120)], a kind gift of Zoltán Ivics. The vector expressing a recombinant ZFP809 protein harbouring the mutations in the DNA-binding domain of the ZFP809 homologue in *M. pahari*, pSB/FLAG-MpZFP809, was generated by replacing the DNA-binding domain (consisting of seven C2H2-type zinc fingers) of the Zfp809 ORF on the pSB/FLAG-MmZFP809 vector with the corresponding fragment amplified with high fidelity *Pfu* DNA Polymerase (Fermentas) from *M. pahari* genomic DNA using an overlapping PCR approach. All vectors and plasmids were sequenced to confirm successful cloning.

#### Viral titer assays

For viral transductions, 2.5 × 106 PlatE packaging cells were transfected with 10 µg of retroviral vector DNA in 57 cm2 culture dishes using the calcium phosphate transfection method as described elsewhere [[96](#_ENREF_96)]. Growth medium of PlatE cells was renewed one day after transfection, and cells were incubated for one more day before harvesting viral particles. Virus supernatants were harvested 48 hours post-transfection, filtered through a sterile 0.45 µm filter, and transferred serially diluted to target cells (seeded at a density of 1 × 104 cells/cm2 in 6-well dishes one day prior to transduction) in the presence of 6 µg/ml Polybrene. G418-containing selection medium was added 24 hours post transduction at 0.4 mg/ml (F9) or 0.6 mg/ml (NIH/3T3). Resistant colonies were counted after 10 to 14 days of selection to determine viral titers (CFU/ml).

#### Genomic PCR screening of wild Mus species

Genomic *Mus musculus* and *Mus dunni* DNA was extracted from NIH/3T3 fibroblasts or immortalised *Mus dunni* tail fibroblasts, respectively, using the TRIzol reagent (Invitrogen). The following DNA samples from wild *Mus* species were purchased from The Jackson Laboratory: *Mus caroli*/EiJ, *Mus pahari*/Ei, PANCEVO/Ei (*Mus spicilegus*) and SPRET/Ei (*Mus spretus*). PCR reactions were performed using 50 to 100 ng of genomic DNA as template and *Taq*-Polymerase (Invitrogen). Amplified partial MmERV *pol* sequences were sub-cloned using the TOPO® TA Cloning® Kit (Invitrogen) and individual clones were sequenced using the BigDye Terminator v3.1 Cycle sequencing kit (Qiagen). Quantitative PCR (qPCR) on genomic DNA was conducted using Platinum SYBR Green qPCR SuperMIX UDG (Invitrogen). The DNA-binding region on exon 5 of Zfp809 was amplified using primers annealing up- and down-stream of the zinc finger domain with *Taq*-Polymerase (Invitrogen) and sequenced without further cloning. Primer sequences are shown in Supplementary Table S2.1.

#### Inhibitor treatment

To optimise inhibitor treatment, Aza-dC or Trichostatin A (TSA) were applied to logarithmically growing NIH/3T3 cells at concentrations ranging from 0.3 µM to 10 µM or 0.1 µM to 0.8 µM, respectively. Aza-dC treatment was for 24 hours followed by one day of recovery in normal medium. TSA was applied for 24 hours without recovery. The optimal concentrations of Aza-dC and TSA for ERV derepression were determined by reverse-transcription quantitative PCR (RT-qPCR) with primers specific for VL30 elements that have been shown to be upregulated in MEFs [[48](#_ENREF_48)]. A new batch of NIH/3T3 cells was then grown and treated with the determined optimal inhibitor concentrations (10 µM Aza-dC and 0.4 µM TSA) or a combination of both drugs at these concentrations. Double treatment with TSA and Aza-dC was performed by growing cells in medium with 10 μM of Aza-dC for 24 hours, followed by treatment with 0.4 µM TSA for additional 24 hours.

#### Reverse transcription quantitative PCR (RT-qPCR)

Total RNA was extracted from logarithmically growing cells using the TRIzol reagent (Invitrogen) according to the manual. Isolated RNA was purified from possible DNA contaminants using the DNA-free kit (Ambion). Reverse transcription was carried out with the RevertAid H Minus First strand cDNA synthesis kit (Fermentas) using 1 to 2 µg of total RNA and random hexamer primer or a MmERV *gag* specific antisense primer. For quantification of viral mRNAs, qPCR was performed on a Mx3000P thermocycler (Stratagene) using Platinum SYBR Green qPCR SuperMIX UDG (Invitrogen). The amplification efficiency of each primer pair was determined using dilution series of a cDNA template or RT-PCR products. Primer sequences are shown in Supplementary Table S2.1.

#### ChIP-qPCR

Preparation of soluble chromatin was carried out using the MAGnify Chromatin Immunoprecipitation kit (Invitrogen) according to the user manual. Briefly, cells were trypsinised at 80-95% confluence, washed and resuspended in phosphate buffered saline buffer, and cross-linked for 10 minutes with formaldehyde at a final concentration of 1%. After cross-linking, cells were lysed and chromatin was sheared in a Covaris S2 sonicator using the following settings: 30 cycles of 30 seconds sonication with 200 cycles/burst, 20% duty cycle and intensity set to 8. Fragment length of reverse-cross-linked sheared chromatin was analysed by electrophoresis and found to be in the range of 300 to 1000 bp. Sheared chromatin of 1 × 105 cells was used per immunoprecipitation. Enrichment of specific ERV sequences was determined by qPCR using an input sample as control. Primer efficiency was determined by dilution series of the input control. Antibodies for immunoprecipitation of modified histones were purchased from Abcam (cH3 (ab1791), H3K9me3 (ab8898), H4K20me3 (ab9053) and H3K27me3 (ab6002)). The monoclonal TRIM28 antibody used in this study has been described previously [[121](#_ENREF_121)]. Monoclonal ANTI-FLAG® M2 antibody was purchased from Sigma-Aldrich. An unspecific rabbit IgG antibody was used as a control to determine background signals caused by non-specific chromatin binding. Per ChIP, 1 µg of antibody or IgG was applied.

#### Generation of F9 EC cells stably expressing FLAG-tagged ZFP809

F9 EC cells were seeded at a density of 1 × 104 cells/cm2 in a 6-well dish one day prior to transfection. Cells were then cotransfected with 2 µg of pSB/FLAG-MmZFP809 or pSB/FLAG-MpZFP809 and 2 µg of pCMV-SB100X using the Turbofect reagent (Fermentas). One day post transfection, F9 cells were trypsinised and seeded into 9 cm culture plates. After one additional day, medium containing 1 µg/ml puromycin was applied to start selection of stably transfected cells. After 9 days of selection, puromycin-resistant colonies (>100) were pooled and seeded into new culture plates for further selection. Pooled cells were split every 2 to 3 days and kept under selection for one more week before analysis.

## Results

### Diversity and repression activity of ERV PBS sequences in the mouse genome.

The majority of retroviruses in mammals utilise PBS sequences that are complementary to the 18 most 3’ nucleotides of mature cellular tRNAs. To assess the diversity of PBS sequences in murine ERVs, we generated a library of potential PBS sequences derived from the 3´ end of all known murine tRNA sequences available at the genomic tRNA database (http://gtrnadb.ucsc.edu). Since a CCA trinucleotide is attached to the 3’ end of transcribed tRNAs during processing in eukaryotes, this sequence was also added to the tRNAs in our library. The resulting 147 non-redundant potential PBS sequences were then screened for genomic copies by BLAST search against the mouse genome under settings allowing only perfect matches. Out of these putative PBS sequences, 54 were found in the mouse genome with copy numbers ranging from one to more than 1000 (Table 2.1). To confirm the retroviral origin of these sequences in the murine genome, 500 bp of the downstream flanking regions of all PBS sequences that were found in more than ten genomic copies were screened for repetitive elements with the *Repeatmasker* program (www.repeatmasker.org). *Repeatmasker* screens sequences for repetitive elements collected in the *Repbase* database where each ERV family is deposited as two or more separate entries, one or more long terminal repeats (LTRs) and the internal regions between the LTRs.

**Table 2.1: Diversity and copy numbers of PBS sequences frequently found in murine ERV**

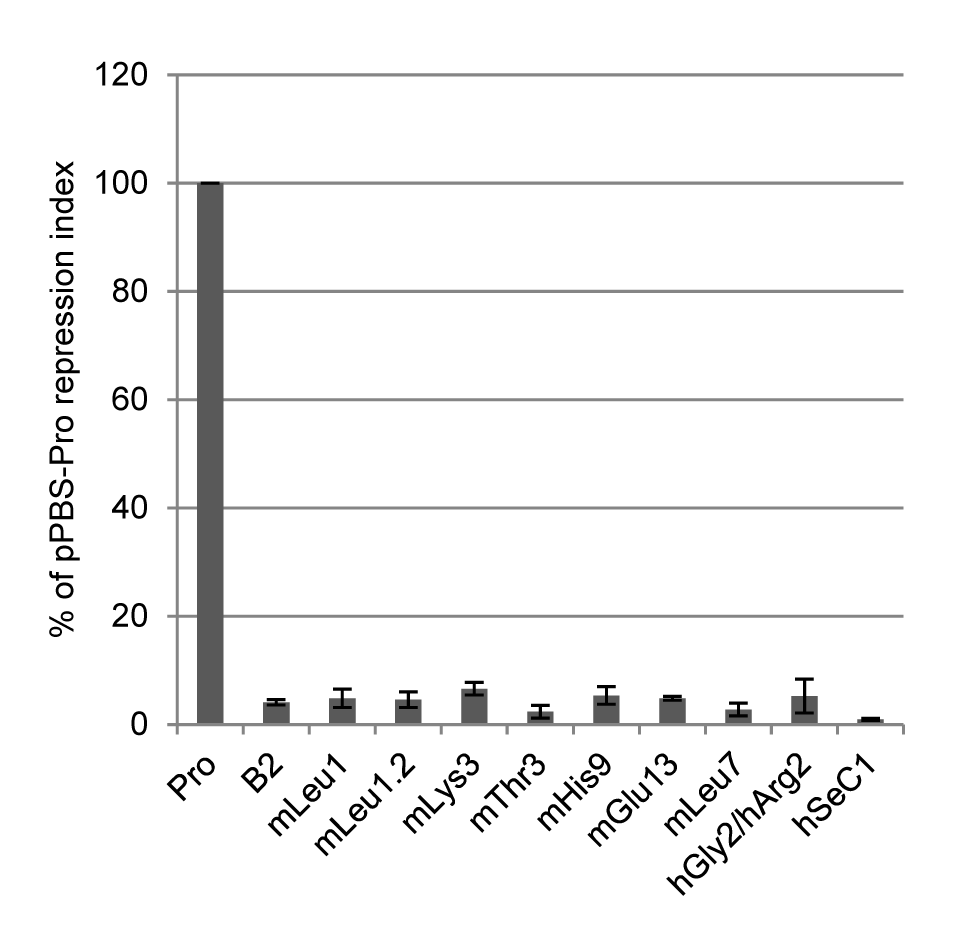
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PBS | Sequence | Associated ERV elements identified by Repeatmasker# | copy number | repression |
| mPhe6/mSer1 (Phe) juhp(P((Phe) | TGGTGCCGAAACCCGGGA | IAPEz(462), RLTR10(435), MMERVK10C(55), other(80) | 1065 | - [[84](#_ENREF_84)], [[95](#_ENREF_95)] |
| mIle9 (I) | TGGTGGCCCGTACGGGGA | RLTR10(618), IAP-d(9), RLTR44(7), other(3) | 643 | - [[84](#_ENREF_84)] |
| mLeu1 | TGGTACCAGGAGTGGGGT | MT(292), ORR1A1(108), ORR1A3(65), other (105) | 592 | - Fig. 2.1 |
| mPhe1 (Phe1) | TGGTGCTGAAACCCGGGA | IAPEz(169), RLTR10(184), MMERVK10C(22), other(33) | 420 | - [[95](#_ENREF_95)] |
| mLeu1.2 | TGGTACCAGAAGTGGGGT | MT(113), ORR1A1(41), MERVL(20), other ERV(39) | 220 | - Fig. 2.1 |
| mGly14 (Gly1) | TGGTGCATTGGCCGGGAA | RLTR6(69), MMVL30(76), RLTR1B(11), other(6) | 193 | - Chapter 1 |
| mLeu1/mSup1 | TGGTACTGGGAGTGGGGT | ORR1A1(126), MT(29), ORR1A3(7), other(7) | 172 | nd\* |
| mPro14 (Pro) | TGGGGGCTCGTCCGGGAT | RLTR6(91), MuRRS(18), MMVL30(15), other(18) | 149 | + [[81](#_ENREF_81)], [[84](#_ENREF_84)], [[94](#_ENREF_94)] |
| mGln10 | TGGAGGTCCCACCGAGAT | MMERGLN(37), RLTR4\_MM(29), RLTR1B(21), MuLV(2) | 91 | nd |
| mLys10 (Lys1,2) | TGGCGCCCAACGTGGGGC | MMETn(35), RLTR45(15), MMERVK9C\_I(4), other(9) | 64 | + [[84](#_ENREF_84)], [[96](#_ENREF_96)], [[87](#_ENREF_87)] |
| mLys6 (Lys3) | TGGCGCCCGAACAGGGAC | ETnERV(32), ERVB2\_1A-I\_MM(14), ETnERV3(3), other(6) | 60 | - [[84](#_ENREF_84)], [[96](#_ENREF_96)] |
| mGln4 (Gln) | TGGAGGTTCCACCGAGAT | RLTR4\_MM(32), MMERGLN(20), RLTR1B(5), MuLV(2) | 59 | - [[96](#_ENREF_96)], [[85](#_ENREF_85)] |
| mGly7 (Gly2) | TGGTGCGTTGGCCGGGAA | RLTR1B(25), MMVL30(16), RLTR6(8) | 55 | - Chapter 1 |
| mIle1 | TGGTGGCCCATACGGGGA | RLTR10(32), RLTR44(10), IAP-d(5), other(3) | 52 | nd |
| mLys3 | TGGCGCCTGAACAGGGAC | ETnERV(23), ERVB2\_1A-I\_MM(18), MMERVK9E\_I(2) | 47 | - Fig. 2.1 |
| mThr3 | TGGAGGCCCCAGCGAGAT | RLTR1B(16), MMERGLN(10), RLTR4\_MM(3), other(3) | 44 | - Fig. 2.1 |
| mGly1 | TGGTGCATTGGCTGGGAA | MERV1\_I(6), MMVL30(4), RLTR6(3), other(6) | 24 | nd |
| mPro1 | TGGGGGCTCGTCTGGGAT | MuRRS(14), RLTR6(4), other(4) | 24 | nd |
| mSeC1 | TGGCGCCCAATGTGGGGC | MMETn(7), RLTR45(5), other(6) | 19 | nd |
| mMet9 | TGGTAGCAGAGGATGGTT | RMER17A(8), MurERV4\_19(1) | 16 | nd |
| mPro1.2 | TGGGGGCTCATCCGGGAT | MuRRS(4), MuRRS4(3), MMERGLN(2), other(2) | 13 | nd |
| mLys1 | TGGCACCCAACGTGGGGC | MMETn(3), RLTR45(2), other(4) | 11 | nd |

# Elements are internal ERV regions, indicated by the suffix “-int” in Repeatmasker (not shown in the table)

\* nd: not determined

Putative PBS sequences derived from the 3´ end of murine tRNAs were screened for genomic copy numbers by BLAST search of the mouse genome (MGSCv37 assembly). The name of the PBS indicates the type and number of matching murine tRNAs. The names of PBS sequences that have been analysed in previous studies are shown in parentheses. The most frequent associated ERVs (internal regions of LTR-class repetitive elements) identified by *Repeatmasker* are shown for each PBS. The number of genomic PBS sequences associated with each ERV family is shown in parentheses. Only PBS sequences that are present in more than ten genomic copies are shown in the table.

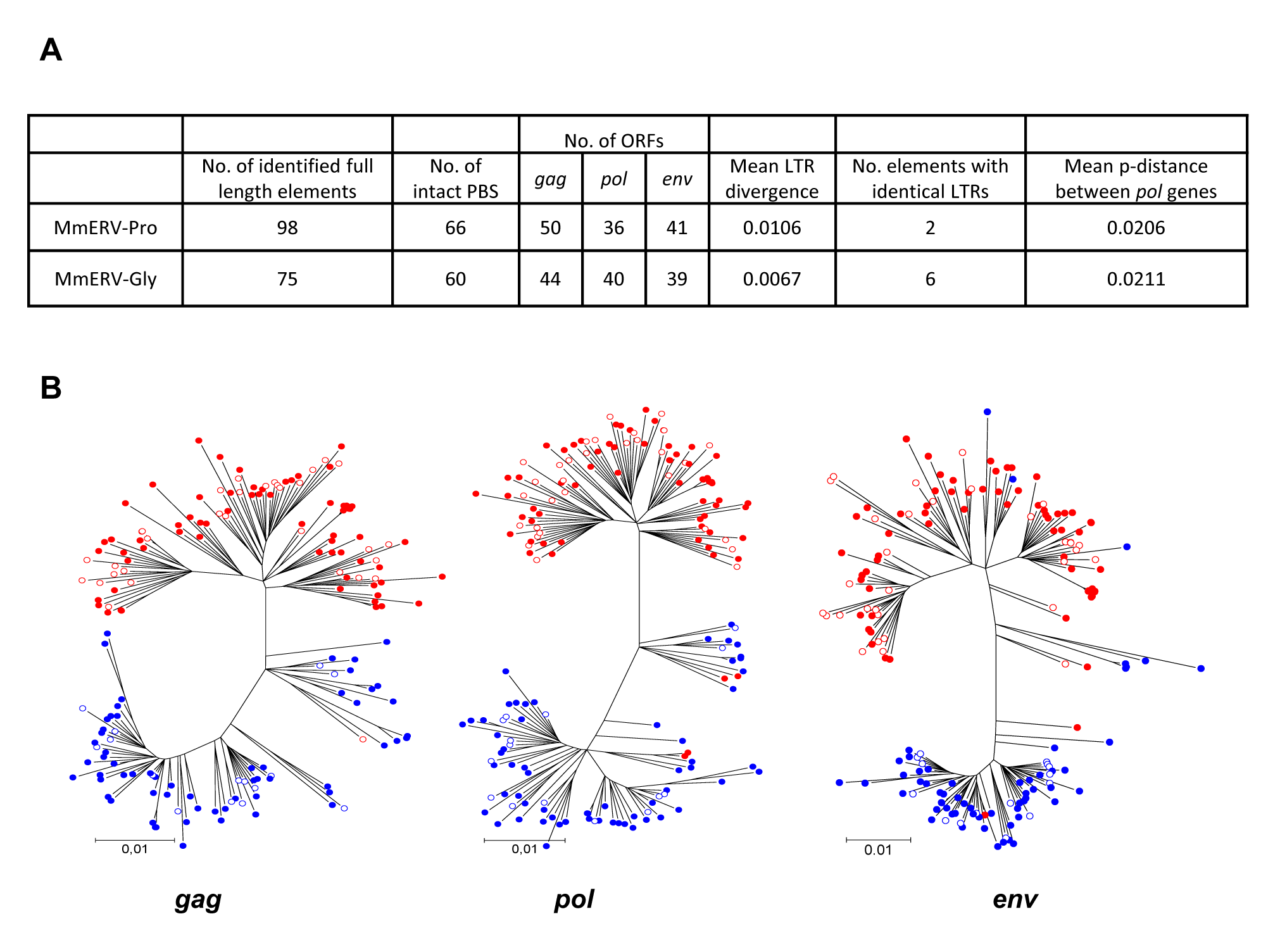
The fetched 500 bp regions downstream of the PBS are expected to contain internal ERV regions, thus, only matching internal elements of the LTR class were considered in our analysis. We found that the vast majority of these PBS sequences were indeed located within annotated ERV elements, indicating that these sequences have been used by various retroviruses to prime reverse transcription (Table 2.1). We speculated that the PBS sequences that are abundant in the mouse genome would represent promising targets for retroviral repression and created retroviral vectors with various, previously untested PBS sequences. Additionally, several potential PBS sequences (mHis9, mGlu13 and mLeu7) which were not found in the mouse genome, but at high copy numbers in the human genome were tested in the assay. We also included two vectors with PBS sequences (hGly2/hArg2 and hSeC1) that matched to human but not murine tRNAs. Since a packaging cell line of human origin was used, we confirmed that all PBS sequences contained in the vectors match to human tRNAs and therefore can prime reverse transcription in the assembled virus particles. The vectors were tested for repression activity in F9 EC cells as described previously [[96](#_ENREF_96)]. As shown in Fig. 2.1, none of the tested vectors were repressed at levels that were comparable to that of the Pro-PBS wild-type vector. By isolating and sequencing genomic DNA from pooled resistant F9 and NIH/3T3 colonies, we confirmed that the original PBS sequence was retained after reverse transcription and integration.



**Figure 2.1: Repression activity of PBS sequences in F9 EC cells**. NIH/3T3 fibroblasts and F9 EC cells were transduced in parallel with serial dilutions of retroviral particles that utilise different PBS sequences for reverse transcription. Viral titers (CFU/ml) for each vector were determined in both cell lines by counting resistant colonies after 10 to 14 days of selection with G418. The repression index for each vector was defined as the ratio of the NIH/3T3 titer to the F9 titer. Repression indices were normalised to the pPBS-Pro vector (100%). Transductions were performed in triplicates; standard deviations are shown as error bars.

### Germ line colonisation by two MmERV subgroups using different PBS sequences

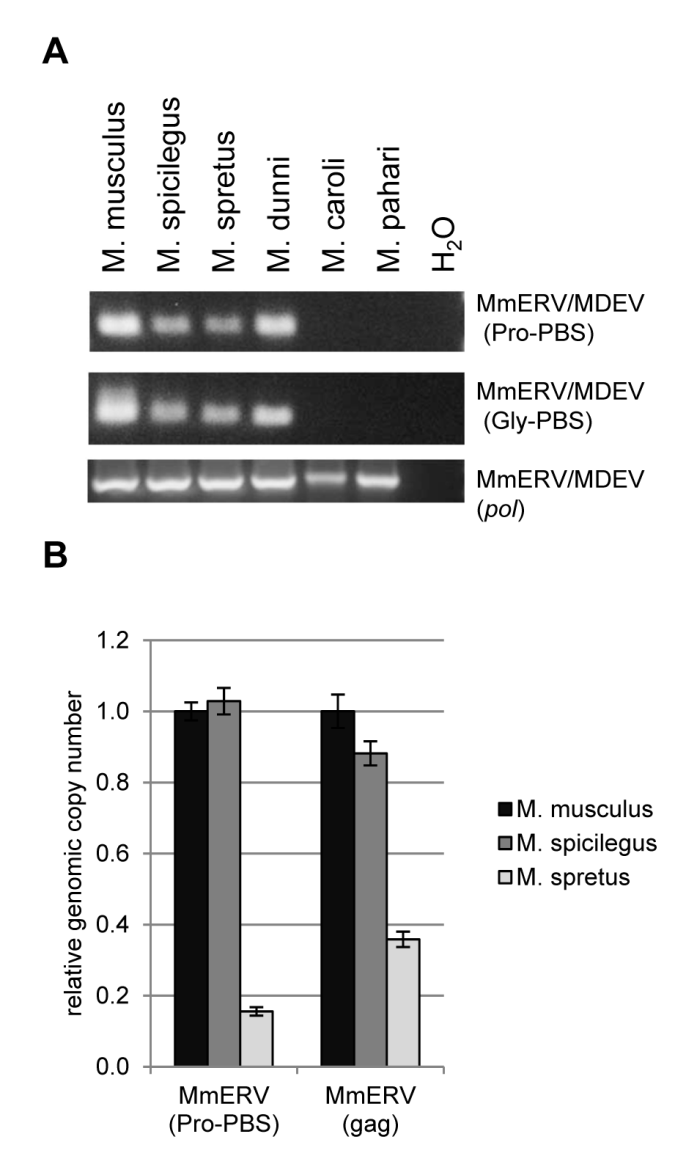
As shown in Table 2.1, about two thirds of all mPro14-PBS sequences, hereafter referred to as Pro-PBS, in the mouse genome are located within *Mus musculus* ERV (MmERV) elements (*Repbase* name: RLTR6). This ERV group is closely related to *Mus dunni* ERV (MDEV) and was first discovered by data mining of the partially sequenced mouse genome [[122](#_ENREF_122)]. As MDEV, MmERV elements show homologies in their LTR region to endogenous virus-like 30 (VL30) elements, a family of non-autonomous LTR elements that lack coding regions for retroviral proteins. The originally described MmERV element contains a Pro-PBS [[122](#_ENREF_122)]. However, we found that also a large number of glycine-PBS (mGly14) sequences, hereafter referred to as Gly-PBS, were located within MmERV elements (Table 2.1). The MmERV Gly-PBS is identical to the PBS that is used by many PERVS and we have shown that this PBS is not transcriptionally repressed in F9 EC cells (chapter 1, see manuscript). The MmERV group therefore represents an interesting opportunity to assess the impact of the presence or absence of a strong repressor target sequence within an ERV. To evaluate the phylogenetic relationships of these MmERV elements, we performed a genome-wide BLAST search with the *gag* gene from a published MmERV sequence [[122](#_ENREF_122)]. Matching sequences including their genomic 5’ and 3’ flanking regions were extracted from the genome and aligned with the MmERV reference genome to retrieve full-length viral genomes. Only elements that shared homologies with MmERV over the entire length were kept for further analysis. The PBS of each MmERV element was annotated manually based on the alignment information. The majority of the elements contained either a Pro- or Gly-PBS, and these elements are hereafter referred to as MmERV-Pro and MmERV-Gly, respectively (Fig. 2.2A). The remaining MmERV elements contained PBS sequences differing in only one or a few nucleotides from either the Pro-PBS or the Gly-PBS. The majority of the mutations in these PBS sequences were distributed randomly and most of the mutated PBS sequences did not perfectly match to any murine tRNA indicating that most of these mutations occurred after retroviral integration. The MmERV coding regions for the retroviral Gag, Pol and Env proteins were determined based on annotations on the MmERV reference genome [[122](#_ENREF_122)] and aligned independently. Phylogenetic trees reconstructed from the MmERV coding region alignments indicated that the MmERV group consists of two closely related subgroups using either a Pro-PBS or Gly-PBS (Fig. 2.2B). For some elements, the PBS type did not correlate with the sequence similarity of the coding regions. Closer inspection revealed that these elements could be assigned to one of the two subtypes depending on which regions were used for phylogeny reconstruction and therefore most likely represent recombinants between members of the two MmERV subgroups (data not shown).



**Figure 2.2: Comparative analysis of two MmERV subgroups**.(A) Characterisation of identified MmERV-Pro and MmERV-Gly subgroups using different PBS sequences. Intact PBS refers to the PBS sequences, predominantly found in the two subgroups. Only ORFs with a length of at least 60% of the length of the corresponding ORF in the MmERV reference genome were counted. The mean LTR divergence is the mean of the p-distances between the 5´ and the 3´ LTR of all elements. (B) Phylogenetic analysis of MmERV elements in the mouse genome. Coding regions for Gag, Pol and Env proteins were extracted from identified MmERV elements and aligned using the MUSCLE alignment tool in MEGA5. The phylogenetic relationship was inferred using the Neighbor-Joining method. The branch lengths indicate the p-distances between sequences (number of amino acid differences per site). Shown are the bootstrap consensus trees for each coding region. Sequences belonging to elements containing a Pro- or Gly-PBS are indicated by red and blue markers, respectively. Filled circles represent intact PBS sequences; empty circles indicate imperfect Pro- and Gly-PBS sequences.

To analyse the relationship of the putative founder viruses that gave rise to the two MmERV subgroups, two consensus sequences were derived from aligned elements of both MmERV subgroups. By comparing the MmERV-Pro and MmERV-Gly consensus sequences, we determined a similarity of 99.5% in the *pol* gene and synonymous to non-synonymous mutation rates (dS/dN) of 13.8, 9.3 and 3.9 in the *gag*, *pol* and *env* gene, respectively. This indicated that purifying selection, possibly during cycles of active replication as exogenous retroviruses, shaped the genomes of the potential exogenous founder viruses of both MmERV subgroups after their split from each other.

Since the 5’ and 3’ LTR of a retrovirus are identical at the time of integration, analysis of the LTR substitution rate allows an approximate estimation of the integration time. We found that elements of both subgroups had highly similar and in some case even identical LTRs (Fig. 2.2A), indicating that both MmERV subgroups have integrated into the mouse genome relatively recently in history. To further determine the approximate time of MmERV germ line integration, we performed PCR on genomic DNA of wild *Mus* species with primer pairs amplifying a short fragment in the 5’ untranslated region (UTR) of MmERV. To allow specific detection of the highly similar MmERV subgroups, two forward primers annealing to the Pro-PBS or Gly-PBS, which differ in 6 nucleotides, and a common reverse primer annealing to a conserved region in the 5´UTR of both MmERV subgroups were designed. Both primer pairs amplified PCR products in *M. musculus, M. spicilegus, M. spretus and M. dunni* (Fig. 2.3A, upper panels). However, since MDEV is closely related to MmERV, we could not exclude that the used primer pairs also amplified MDEV elements. To determine whether these elements in wild *Mus* species are more closely related to MmERV or MDEV, another primer pair, amplifying a 630 bp region in the *pol* gene of MmERV and MDEV was designed and used for PCR on genomic DNA. PCR products of the expected size could be amplified from genomic DNA of all tested *Mus* species (Fig. 2.3A, lower panel). The *pol* amplicons were then sub-cloned into vectors and six clones per species were sequenced. Sequencing revealed that amplicons from *M. spicilegus* and *M. spretus* were highly similar (>95% identity) to the published MmERV *pol* gene [[122](#_ENREF_122)]. All sequences amplified from *M. dunni* DNA were highly similar to MDEV (>94% identity), whereas sequences from *M. caroli* and *M. pahari* were about 75 % homologous to both MmERV and MDEV, and therefore represented more distantly related ERVs. Supplementary Figure S2.1 shows a phylogenetic tree of the amplified partial *pol* sequences derived from different *Mus* species and the corresponding sequences of the published MmERV and MDEV genomes.



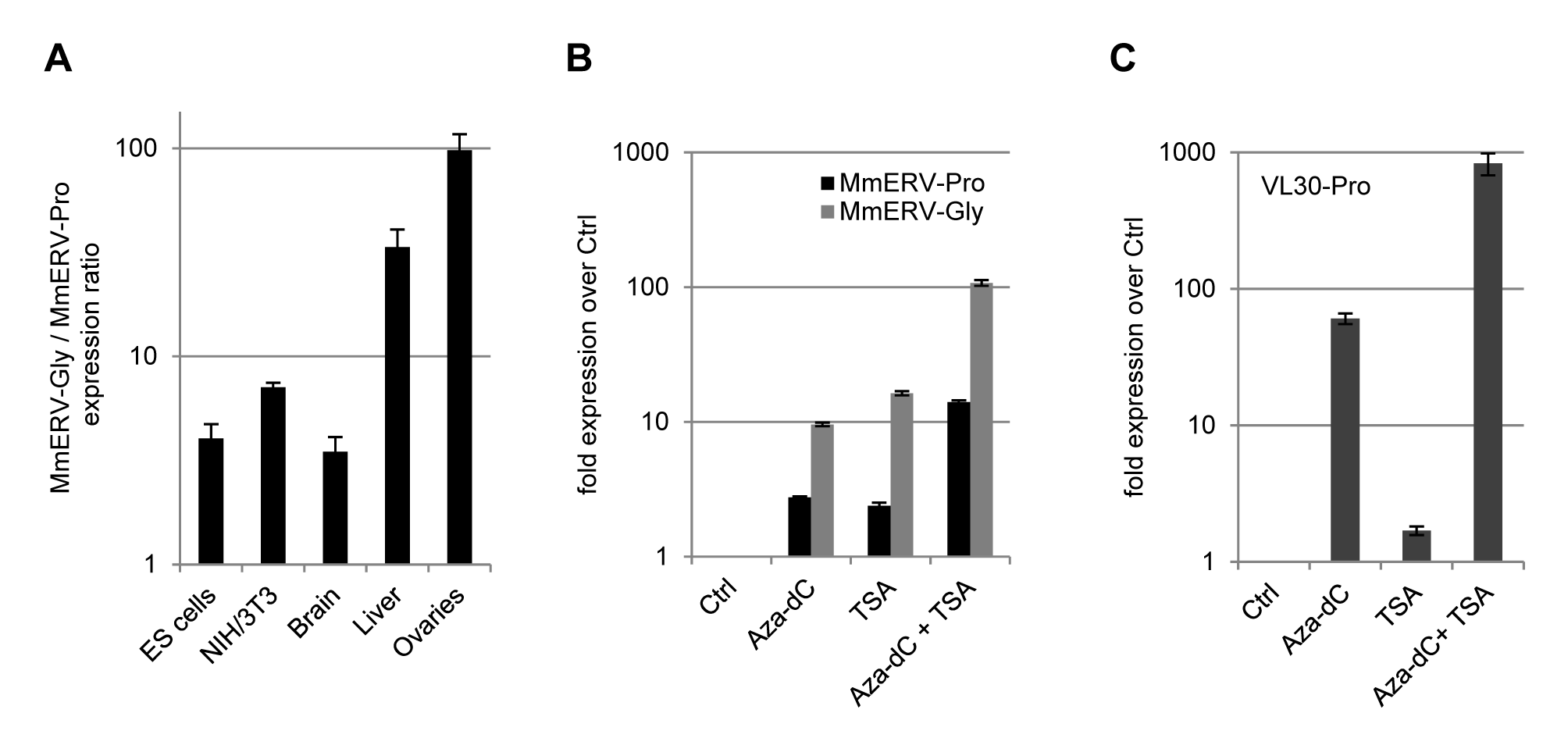
**Figure 2.3: Prevalence of MmERV elements in wild *Mus* species**. (A) PCR with primer pairs specific for MmERV and possibly MDEV with Pro-PBS or Gly-PBS sequences or primers annealing to conserved regions in the *pol* gene of MmERV and MDEV was performed on genomic DNA from various *Mus* species. (B) Genomic *Mus* DNA was screened for MmERV insertions by qPCR using primers amplifying either only the MmERV-Pro subgroup or both MmERV types (*gag*). Results were normalised using primers amplifying a short region within the *Zfp809* DNA binding region that was found to be conserved in these *Mus* species. The histogram shows the mean relative amount of MmERV targets in wild *Mus* species relative to *M. musculus* with standard deviations (n = 3, technical replicates).

Finally, we performed quantitative PCR (qPCR) on genomic *Mus* DNA using a MmERV-Pro specific primer pair as described above and primers annealing to conserved regions in the *gag* gene of both MmERV types to determine relative MmERV copy numbers in different *Mus* species. As shown in Fig. 2.3B, we detected about five-fold less MmERV-Pro copies in *M. spretus* compared to *M. musculus* and *M. spicilegus*, indicating that MmERV was still active after the split of these species which was estimated to have occurred about 1.5 million years ago [[123](#_ENREF_123)].

Although, we cannot exclude that MmERV is also present in more distant *Mus* species, altogether the data presented above suggest that MmERV integrated in the germ line of the common ancestor of *M. musculus, M. spicilegus* and *M. spretus* after the split from *M. dunni*, which has been colonised by MDEV. Interestingly, both MmERV subgroups expanded to similar copy numbers in *M. musculus* after germ line colonisation despite using PBS sequences with different repression activities.

### MmERV-Pro and MmERV-Gly elements are differentially expressed

To analyse subgroup-specific MmERV expression, primers annealing to the Pro-PBS or Gly-PBS, as described above, were used for RT-qPCR expression analysis. Since the primer pair designed for detecting the MmERV-Gly subgroup was also predicted by *in silico* PCR analysis to amplify a few VL30 elements which share homologies with MmERV in the 5’UTR and contain the same Gly-PBS, RNA isolated from cell lines and tissues was reverse-transcribed using a primer annealing to a region of the MmERV *gag* gene which is absent in VL30 elements. As shown in Fig. 2.4A, MmERV-Gly mRNAs were detected at higher levels than MmERV-Pro in all tested cell lines and tissues although the MmERV-Pro specific primer pair was predicted to amplify twice as many elements as the MmERV-Gly primers by in *silico* PCR (Supplementary Table S1). Interestingly, the difference in expression was only moderate in ES cells in which strong PBS-Pro-mediated transcriptional repression was reported previously [[85](#_ENREF_85)]. In contrast, MmERV-Gly elements were expressed at about 30 and 100 fold higher levels than MmERV-Pro in liver and ovaries, respectively. As indicated by the cycle numbers necessary to detect the PCR products, this difference seemed to be due to a stronger expression of MmERV-Gly rather than a decrease of MmERV-Pro expression in these tissues (data not shown).



**Figure 2.4: Expression analysis of MmERV elements using different PBS sequences.** (A) RNA from murine cells and tissues was reverse-transcribed with a specific primer binding at the MmERV *gag* region to exclude amplification of unspecific targets. The amount of MmERV-Pro and MmERV-Gly mRNAs was determined by RT-qPCR using subgroup-specific primers annealing to the PBS sequences. Histogram shows relative MmERV-Gly expression levels normalised to MmERV-Pro expression with standard deviations (n = 3, technical replicates). (B) Logarithmically growing NIH/3T3 cells were treated with 3 µM Aza-dC, 1.5 µM TSA or both inhibitors combined. Extracted RNA was reverse-transcribed with a MmERV *gag* specific primer as described above. Expression levels of MmERV subgroups were determined by qRT-PCR and normalised to total input RNA used for reverse transcription. Histogram shows the mean fold increase in expression over untreated NIH/3T3 cells. (C) RNA from NIH/3T3 cells treated with Aza-dC and/or TSA as described above was reverse transcribed using random hexamer primer. Fold increase in VL30-Pro expression levels was determined by RT-qPCR and normalised to *ActinB* expression.

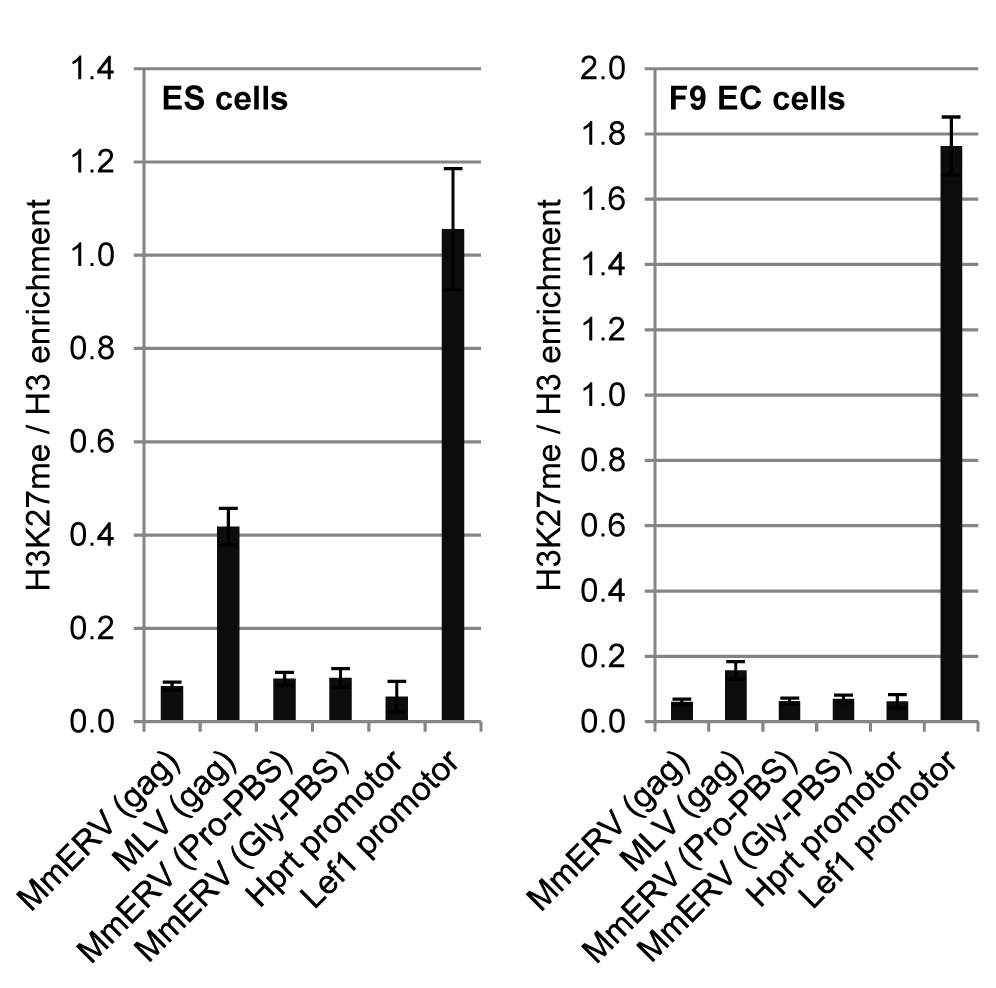
To assess whether the lower expression rate of MmERV-Pro elements in NIH/3T3 fibroblasts was due to epigenetic repression such as histone modifications and DNA methylation, NIH/3T3 cells were treated with the histone deacetylase inhibitor TSA, the DNA methylase inhibitor Aza-dC, or both compounds combined. Inhibitor treatment resulted in an overall upregulation of MmERV that was synergetic when a combination of both inhibitors was applied (Fig. 2.4B). Interestingly, the synergistic upregulation of MmERV-Gly elements was almost 10 times higher than the one observed for MmERV-Pro. This led us speculate that the Pro-PBS may be responsible for a more stable transcriptional repression that cannot be counteracted by TSA and Aza-dC. However, expression of VL30 elements containing a Pro-PBS was almost three magnitudes higher in NIH/3T3 cells treated with TSA and Aza-dC compared to untreated cells (Fig. 2.4C), indicating that epigenetic silencing of ERVs containing a Pro-PBS was reversible by treatment with these inhibitors. To investigate whether MmERV upregulation was limited to one or a few elements or affected a larger number of ERV copies, we amplified a 630 bp amplicon within the MmERV *pol* gene from cDNA of untreated NIH/3T3 cells and NIH/3T3 cells treated with TSA and Aza-dC. PCR products were then sub-cloned into vectors, and eight elements of each cDNA sample were sequenced. Sequence analysis revealed that the MmERV *pol* amplicons from treated cells were highly polymorphic. We counted between eight and 27 nucleotide differences between these sequences, indicating that all PCR-amplicons were derived from RNA expressed from different genomic loci. A phylogentic tree of the amplified *pol* sequences is shown in Supplementary Figure S2.2.

### TRIM28 binding and repressive histone methylations at ERVs in EC cells

*De novo* integrated retroviruses and vectors containing a Pro-PBS are transcriptionally repressed by a complex consisting of TRIM28, HP1 and ZFP809 in F9 EC cells [[86](#_ENREF_86),[91](#_ENREF_91),[94](#_ENREF_94)]. Furthermore, TRIM28 enrichment was observed at several classes of ERVs and other TEs in EC cells [[77](#_ENREF_77),[95](#_ENREF_95)]. We therefore wanted to evaluate the epigenetic differences between the two identified MmERV subgroups in order to determine the effect of the presence or absence of a strong transcriptional repression target in otherwise highly similar ERVs. As assessed by ChIP, MmERV and VL30 elements were clearly bound by TRIM28 at their Pro-PBS (Fig. 2.5). However, also ERVs without a known repressor target site, such as MmERV-Gly or endogenous MLV with a glutamine-PBS (Gln-PBS), were bound by TRIM28 at the PBS to some extent, as compared to the *Hprt* promoter. In accordance with these findings, modest enrichment of TRIM28 at various ERVs without a Pro-PBS and non-LTR retrotransposons has been reported previously [[77](#_ENREF_77)]. MmERV-Pro elements were marked by higher levels of H3K9me3 and H4K20me3 than MmERV with a Gly-PBS (Fig. 3.5). Both of these histone modifications have been recently associated with ERV silencing in ES cells [[71](#_ENREF_71),[77](#_ENREF_77)]. No difference in the nucleosome density, as assessed by chromatin immunoprecipitation with H3 antibody, was observed between the two MmERV subtypes showing that the higher enrichment of methylated histones is not due to a general higher histone density at these regions (Fig. 2.5). Despite different degrees of histone methylations at various ERVs, all ERV regions analysed showed marks of transcriptionally inactive, tightly packed heterochromatin. To confirm our findings, ChIP was performed with chromatin from murine ES cells. Also in these cells, MmERV-Pro elements were bound by TRIM28 and marked by higher degrees of H3K9me3 and H4K20me3 than MmERV-Gly elements (Supplementary Figure S2.3). Recently, also Polycomb complexes and H3K27 trimethylation have been associated with retroviral silencing in ES cells [[75](#_ENREF_75)]. To investigate whether MmERV elements are repressed by these histone modifications, ChIP assays using an antibody specific for H3K27me3 were performed with F9 EC and ES cells. As shown in Fig. 2.6, we observed the previously reported enrichment of H3K27me3 at the *gag* gene of endogenous MLV elements in ES cells [[75](#_ENREF_75)] and in F9 EC cells, although to a lesser extent, but we did not detect any enrichment of this histone modification at the *gag* gene or PBS regions of MmERV.



**Figure 2.5: TRIM28 binding and histone methylations at ERV loci in F9 EC cells.** Cross-linked chromatin isolated from F9 EC cells was immunoprecipitated with antibodies against histone 3 (H3), H3K9me3, H4K20me3, TRIM28 or an unspecific IgG control antibody. Enrichment of target sequences was determined by qPCR using a chromatin input control for normalisation. The promoter of the constitutively expressed *Hprt* gene was included as a control region for open chromatin. The *Mest* promoter is a known target of TRIM28 in F9 EC cells and the *Lef1* promoter is known to be repressed by trimethylation of lysine 27 on histone 3 (H3K27me3). A genomic region within the DNA binding domain of *Zfp809* was included as control for a region without promoter activity. Graphs show mean ChIP enrichment values as percentage of the input control with standard deviations (n = 3, technical replicates). Enrichment of target sequences in chromatin samples immunoprecipitated with IgG was less than 0.01% of input and is not shown.



**Figure 2.6: Trimethylation of H3K27 at ERV loci in F9 and ES cells.** Cross-linked chromatin was isolated from F9 and ES cells and immunoprecipitated with antibodies against H3 or H3K27me3. Relative enrichment of target sequences in immunoprecipitated DNA was determined by qPCR using an input sample as normalisation control. H3K27me3 enrichment was calculated as ratio over H3 enrichment to normalise for nucleosome density at the analysed regions. Histogram shows mean H3K27m3/H3 ratios (n = 3, technical replicates) with standard deviation (error bars)

### Conservation of the ZFP809 DNA binding domain in mice and non-murine rodents

Our findings that two subgroups of MmERV using PBS sequences with different repression activities amplified to similar copy numbers in the mouse genome led us to speculate that ZFP809 might have recently, possibly after the MmERV germ line colonisation, evolved to repress ERVs with a Pro-PBS. Indeed, initially we could not find any potential ZFP809 homologues in other species than mouse by BLAST searches of all available NCBI nucleotide and protein databases. To investigate the recent evolution of ZFP809, primers annealing up- and downstream of the *Zfp809* DNA-binding domain consisting of seven C2H2-type zinc fingers were applied to amplify this region from genomic DNA of wild *Mus* species. DNA from rat and a deer mouse (*Peromyscus*) was included in the analysis but despite trying various primer sets we were unable to amplify a product from these DNA templates. PCR products amplified from *Mus* DNA samples were sequenced and analysed using bioinformatics tools. Only after the sequencing analysis was conducted, we eventually identified two potential ZFP809 orthologues by BLAST search. These sequences have been published recently and therefore remained undetected in our initial screens. The first matching sequence (NW\_003624827) was an unplaced genomic scaffold derived from the Chinese hamster (*Cricetulus griseus*) whole genome shotgun sequencing project and the corresponding predicted partial protein (XP\_003516040). The genomic sequence is truncated at the 5´ end but could be aligned to the *Zfp809* locus in the mouse genome over a stretch of about 2.8 kb including the entire DNA-binding domain on exon 5 as well as exon 4 and two introns. Although the predicted partial protein was annotated as “zinc finger protein 84-like” it showed a much higher similarity to ZFP809 in the DNA binding region (more than 80%) than to any other protein in the NCBI database as confirmed by BLAST search. The observed homology in exons and introns strongly suggests that this sequence represents a ZFP809 orthologue in *C. griseus*.

The second matching nucleotide sequence (JO002639) was deposited as an mRNA sequence that has been identified by next generation sequencing of the transcriptome of the blind mole rat (*Nannospalax galili*). This sequence contains a potential DNA-binding region with seven C2H2 zinc fingers that is about 80% identical on nucleotide level to the corresponding ZFP809 DNA binding region but it is missing the 5’ end of the mRNA. By alignment of the sequence to the *M. musculus* *Zfp809* mRNA (NM\_172763), we found that a frame shift mutation, caused by a single nucleotide deletion in the fourth zinc finger domain, introduced premature stop codons in the *N. galili* sequence. However, the deletion was located after 6 consecutive adenine bases and might therefore be merely the result of a sequencing error. Manual correction of the deletion restored an ORF ended by a stop codon after the seventh zinc finger domain, as in the *Zfp809* mRNA. The DNA binding domain of the translated protein was 76% identical to the corresponding ZFP809 domain but less than 60% identical to any other protein sequence in the NCBI database. The DNA binding regions of all ZFP809 sequences were then aligned and the ratio of synonymous to non-synonymous substitution rates (dS/dN) relative to *M. musculus* ZFP809 determined.

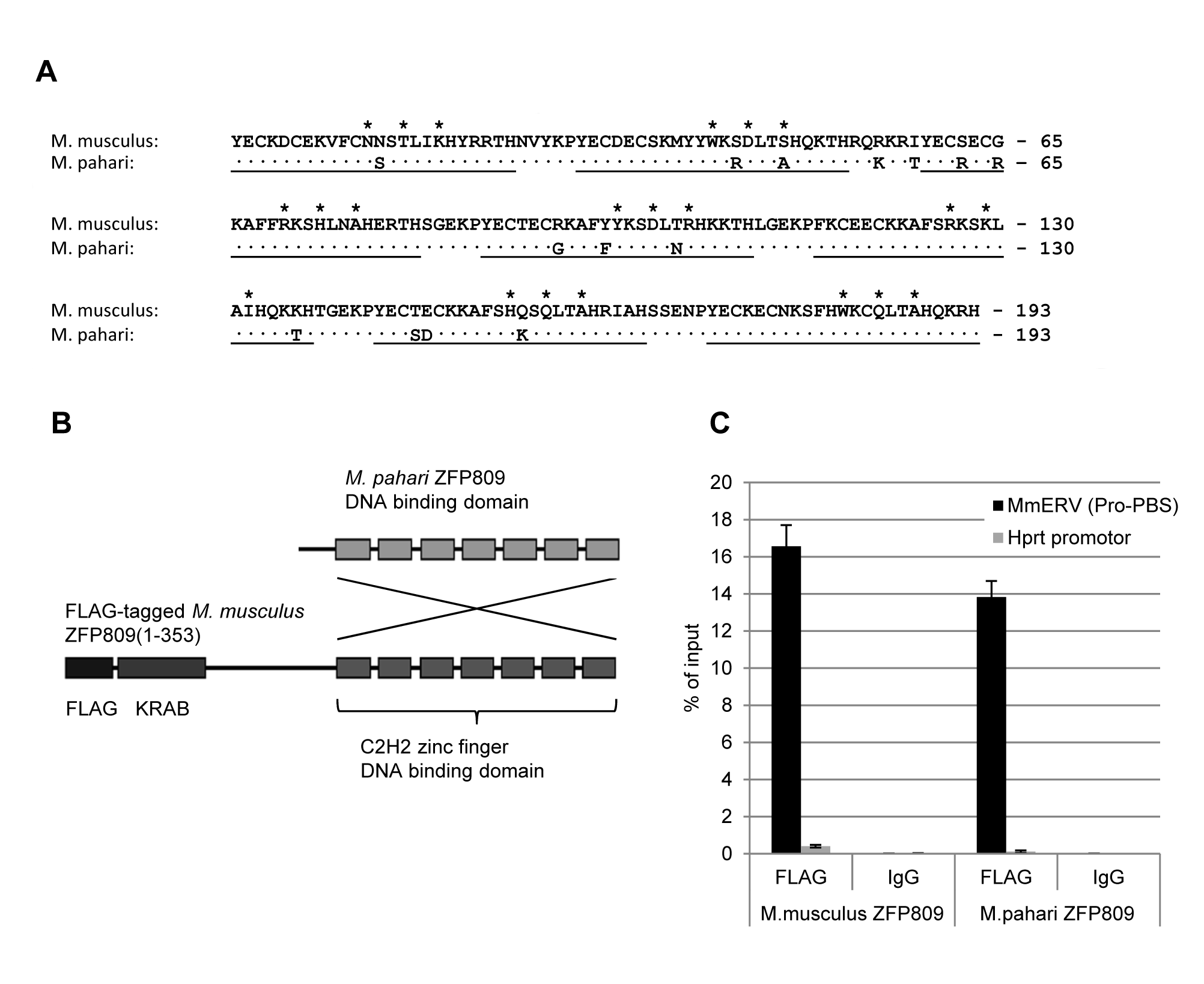
**Table 2.2: Conservation and substitution rates of putative ZFP809 orthologues**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | % homology (DNA) | % homology (protein) | dS | dN | dS/dN | No. of mutated specificity aa | dS/dN (specificity aa codons) |
| *M. spicilegus* | 98.3 | 99.5 | 0.076 | 0.002 | 34.9 | 0 | - |
| *M. spretus* | 98.8 | 99.5 | 0.050 | 0.002 | 23.2 | 0 | - |
| *M. caroli* | 97.9 | 99.0 | 0.084 | 0.004 | 19.4 | 0 | - |
| *M. dunni* | 97.2 | 97.9 | 0.101 | 0.009 | 11.6 | 0 | - |
| *M. pahari* | 94.6 | 92.7 | 0.139 | 0.032 | 4.4 | 1 | 3.6 |
| *C. griseus* | 81.9 | 80.2 | 0.463 | 0.108 | 4.3 | 1 | 21.5 |
| *N. galili* | 80.7 | 76.6 | 0.479 | 0.119 | 4.0 | 3 | 8.7 |

Sequence homology and synonymous to non-synonymous (dS/dN) substitution rates of identified ZFP809 DNA binding domains, spanning seven C2H2 zinc finger, relative to *M. musculus* ZFP809 (NM\_172763) were determined. The three specificity amino acids (aa) in each zinc finger that are mainly responsible for sequence specific DNA binding were identified analysed for mutations and substitution rates.

As shown in Table 2.2, all sequences mutated faster at synonymous sites, indicating that purifying selection removed mutations that would change the amino acid sequence of the translated protein. Interestingly, the dS/dN ratio was especially high in closely related *Mus* species. For example, the ZFP809 orthologues of *M. musculus* and *M. spicilegus* differed in ten positions in the DNA binding region but only one of these mutations was non-synonymous. Since each C2H2-type zinc finger contains three amino acids at defined positions that directly interact with the DNA and therefore mainly determine binding specificity [[93](#_ENREF_93)], the conservation of these amino acids in the seven ZFP809 zinc fingers was analysed in detail. Apart from one mutation in the second zinc finger of *M. pahari* ZFP809 (Fig. 2.7A), all orthologues were conserved at these positions (Supplementary Figure S2.4). Also the putative ZFP809 orthologues in hamster and blind mole rat were mainly conserved at these positions with only one or three mutations, respectively, compared to *M. musculus* ZFP809.

To test whether the 14 non-synonymous mutations (including one mutation of a specificity conferring amino acid that was also found in *N. galili*) of the *M. pahari* Zfp809 orthologue would affect the binding specificity of this protein, we replaced the DNA binding domain of a FLAG-tagged ZFP809 protein with the DNA binding domain of *M. pahari* (Fig. 2.7B). Transposon-based vectors expressing FLAG-tagged *M. musculus* ZFP809 or the modified ZFP809 harbouring the 14 amino acid substitutions of *M. pahari* were used to stably transfect F9 cells, and ChIP with an anti-FLAG antibody was performed. As shown in Fig. 2.7C, both ZFP809 variants were able to bind the Pro-PBS of MmERV in a highly specific manner and with comparable efficiencies indicating that the ZFP809 binding specificity did not significantly change since the split of *M. musculus* and *M. pahari* which occurred about 5 million years ago [[123](#_ENREF_123)].



**Figure 2.7: Conservation and target specificity of ZFP809 in *M. musculus* and *M. pahari*.** (A) Alignment of the *M. musculus* and *M. pahari* ZFP809 DNA binding domains. Conserved amino acids in *M. pahari* are shown as dots. The seven C2H2 zinc finger domains are underlined and the main specificity conferring amino acids are marked by asterisks on top of the alignment. (B) The DNA binding domain of a FLAG-tagged ZFP809 protein was replaced by the corresponding zinc finger domains of *M. pahari* to generate a recombinant ZFP809 protein that harbours the 14 amino acid substitutions in the DNA binding domain. (C) F9 EC cells were stably transfected with transposon based vectors expressing FLAG-tagged ZFP809 or the recombinant ZFP809 protein in which the DNA binding domain was replaced. Cross-linked chromatin was immunoprecipitated with an anti-FLAG antibody or an unspecific IgG control antibody. Enrichment of a putative ZFP809 target (Pro-PBS of MmERV) or a non-target control region (*Hprt* promoter) in the precipitated samples was determined by qPCR using an input control for data normalisation. Graphs show mean enrichment values over input control (n=3, technical replicates) with standard deviation (error bars).

Having shown that the ZFP809 DNA binding region is conserved in non-murine rodents, we wanted to investigate whether those species share common Pro-PBS containing ERV families with mice. Since the genomes of *C. griseus* and *N. galiIi* have not been fully sequenced yet, we screened the available genome of the more closely related rat (*Rattus norvegicus*) for Pro-PBS sequences and flanking retroviral sequences as described above. We identified a total of 528 intact Pro-PBS sequences in the rat genome (Baylor 3.4/rn4 assembly) and the majority of these sequences (86%) were located within or directly upstream of internal ERV regions. More than 99% of these ERV elements belonged to 13 different class1 ERV (ERV1) families. However, as reported by the *Repeatmasker* program, none of these families was also clearly identified in the mouse genome indicating that ZFP809 targets different ERV families in different species.

## Discussion

Transcriptional repression of ERVs is believed to constitute a part of the innate immunity that evolved to control the activity of both exogenous as well as endogenous retroviruses. Although a large number of cellular factors and epigenetic modifications involved in ERV silencing are identified, the targeting mechanisms that ensure heterochromatin formation and transcriptional silencing at specific loci are only vaguely understood. The identification of the DNA-binding protein ZFP809 that targets the host silencing machinery to an ERV-specific sequence did not only reveal the mechanisms behind transcriptional silencing of MLV in EC and ES cells but also led to the hypothesis that ERVs might be repressed in a similar manner. However, only a small subset of murine ERVs use a Pro-PBS and is therefore potentially targeted by ZFP809, yet the majority of murine ERVs is marked by repressive histone modifications in ES cells [[71](#_ENREF_71)].

It has been previously shown that many of the PBS sequences used by ERVs are not repressed in murine F9 EC cells [[84](#_ENREF_84),[96](#_ENREF_96)]. This observation is further substantiated in the current study. We also found that the MmERV group exists as two closely related subgroups in the genome, only one of them containing the strong PBS-Pro repressor target. Nevertheless, both groups amplified at similar rates despite evolution of the ZFP809 repressor prior to the germ line integration of these ERVs. On the other hand, we show that a FLAG-tagged ZFP809 can indeed bind to the Pro-PBS of MmERV and these loci are bound by TRIM28 and marked by high levels of repressive histone methylations in EC and EC cells. It is therefore feasible that targeted repression of ERVs cannot efficiently prevent ERV amplification in the germ line. A few ERV copies that escape repression due to genomic integration sites in open chromatin [[81](#_ENREF_81)] might provide enough viral mRNA and proteins to allow continuous amplification of an ERV family. Alternatively, ERVs that are transcriptionally repressed in ES cells could be expressed in other cell types or tissues of the developed organism and amplify via re-infection of germ cells, a mechanism that has been proposed for the amplification of some human ERVs [[124](#_ENREF_124)].

ZFP809-dependent repression most likely represents just one out of several targeting mechanisms involved in ERV silencing. This hypothesis is supported by the observed histone methylations at ERVs without Pro-PBS. The putative Pro-PBS-independent repression mechanism seems to depend on H3K9me3 and H4K20me3 modifications and possibly TRIM28 recruitment as indicated by the modest enrichment of this corepressor at MmERV-Gly and other ERVs without Pro-PBS. However, we and others [[77](#_ENREF_77),[95](#_ENREF_95)] observed low levels of TRIM28 binding to all of the heterochromatin regions that have been analysed, suggesting that low levels of TRIM28 enrichment is rather a general feature of heterochromatin. TRIM28 binding to these regions might be independent of sequence-specific DNA binding factors and instead be based on TRIM28 interactions with heterochromatin-associated proteins such as HP1 [[125](#_ENREF_125)].

Furthermore, other retroviral sequences than the PBS might be targets for transcriptional repression by unknown proteins. It was shown that the 5’UTR of certain IAP elements is able to induce efficient TRIM28-dependent silencing when introduced into a vector [[95](#_ENREF_95)]. Yet unknown KRAB-ZFPs may bind to various ERV regions and recruit TRIM28. The initiated heterochromatin formation may then spread over large regions once initiated at the target sequence [[126](#_ENREF_126)], explaining the observed repressive histone methylations at all tested ERV regions. Nevertheless, the presence of a strong repressor target, such as the Pro-PBS, seems to enhance ERV repression by histone methylations.

The *Muroidea* superfamily of rodents consists of more than 1,500 species and is the most diverse and geographically widespread sub-ordinal clade of mammals [[127](#_ENREF_127)]. The *Spalacidae* family, including *Nannospalax galili*, separated the from *Eumuroida* clade, which includes *Muridae* (mice and rat) and *Cricetinae* (hamsters), early in the *Muroidea* radiation around 25 million years ago [[128](#_ENREF_128),[129](#_ENREF_129)]. We identified sequences, most likely representing ZFP809 orthologues, in species belonging to *Muridae*, *Cricetinae*, and *Spalacidae*, indicating that this protein has evolved before the radiation of *Muroidea*. Interestingly, ZFP809 seems to have retained target specificity over time in different *Muroidea* species that have been colonised by different ERV families without undergoing positive selection to adapt towards new targets. This suggests that either a constant selective pressure exerted by ubiquitous exogenous and/or endogenous retroviruses with Pro-PBS sequences preserved the ZFP809 binding specificity, or that the protein is involved in cellular processes that are conserved in *Muroidea*. In the following two chapters, we will aim to investigate the *in vivo* functions of ZFP809 as ERV repressor and potential regulator of cellular gene expression by genome-wide binding analysis of ZFP809 by ChIP-seq and the generation of a ZFP809 knock-out mouse model.

## Supplementary data for chapter 2



**Supplementary Figure S2.1: Neighbor-Joining tree of MmERV *pol* sequences in wild *Mus* species**. PCR on genomic DNA from wild *Mus* species was performed using primers annealing to conserved regions in the *pol* gene of MmERV and MDEV. PCR products of 630 bp length were sub-cloned and sequenced for phylogenetic analysis. Sequences were aligned with MUSCLE and a Neighbor-Joining tree based on the p-distance between the nucleotide sequences was created in MEGA5. The corresponding *pol* regions of MmERV and MDEV (AF053745) were included in the analysis. The MmERV *pol* reference sequence was fetched from the originally described MmERV provirus which is located between nucleotides 112341 and 121005 of a *M. musculus* PAC clone (AC005743). Shown is the bootstrap consensus tree inferred from 100 replicates. Bootstrap values higher than 80 are indicated at the branch points of the tree.



**Supplementary Figure S2.2: Neighbor-Joining tree of expressed MmERV pol sequences.** Partial MmERV *pol* sequences were amplified from cDNA prepared from untreated NIH/3T3 cells (black circles) or NIH/3T3 cells after treatment with Aza-dC and TSA (grey triangles). Elements were aligned with MUSCLE and a Neighbor-Joining tree was created in MEGA5. Bootstrap values (n=100) are indicated at the branch points of the tree.



**Figure S2.3: TRIM28 binding and histone methylations at MmERV elements in ES cells.** Cross-linked chromatin isolated from ES cells was immunoprecipitated with antibodies against histone 3 (H3), histone 3 trimethylated at lysine 9 (H3K9me3), histone 4 trimethylated at lysine 20 (H4K20me3), TRIM28 or an unspecific IgG control antibody. Enrichment of target sequences was determined by qPCR using a chromatin input control for normalisation. The promoter of the constitutively expressed *Hprt* gene was included as control region for open chromatin. The *Mest* promoter is a known target of TRIM28 in F9 EC cells. Histogram shows the mean ChIP enrichment values as percentage of the input control. Enrichment of all target sequences in chromatin samples immunoprecipitated with IgG was less than 0.01% of input and is not shown.



**Figure S2.4:** **Conservation of the DNA binding domain of putative ZFP809 orthologues**. Zfp809 DNA binding domains were translated into proteins and aligned with MUSCLE. The positions of the seven C2H2 zinc finger domains are highlighted by black bars on top of the alignment. The amino acids mainly responsible for binding specificity are marked by asterisks on top of the alignment.

Supplementary Table S2.1: Primer sequences

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Target | FW/RV | Sequence (5´ to 3´) | Application | Target copy number# | Amplicon length | Reference |
| MmERV (Pro-PBS) | FW | TGGGGGCTCGTCCGGGAT | ChIP/qPCR/RT-qPCR | 93 | 122 |  |
| RV | ACTGCGATCGCACCAAACTTAG |
| MmERV (Gly-PBS) | FW | TGGTGCATTGGCCGGGAATCA | ChIP/RT-qPCR | 46 | 123 |  |
| RV | ACTGCGATCGCACCAAACTTAG |
| MmERV pol | FW | GGGGACTGCTGGCTTTTGTAGA | PCR (sequencing) | 176 | 630 |  |
| RV | CAGGGTTGGTCAGTCAGGTC |
| MmERV gag | FW | TACATTGTGACCTGGCAGAATC | ChIP/qPCR | 190 | 165 |  |
| RV | GAGGAGGTCGTCAATCTCTGG |
| VL30 (Pro-PBS) | FW | TGGGGGCTCGTCCGGGAT | ChIP/RT-qPCR | 15 | 78 |  |
| RV | ATTACCAAGCGACAGAACTTACC |
| MusD (5' UTR) | FW | GTGCTAACCCAACGCTGGTTC | ChIP | 114 | 175 | [[130](#_ENREF_130)] |
| RV | CTCTGGCCTGAAACAACTCCTG |
| ETnII (5' UTR) | FW | GTGCTAACCCAACGCTGGTTC | ChIP | 25 | 218 | [[130](#_ENREF_130)] |
| RV | ACTGGGGCAATCCGCCTATTC |
| IAP (PBS) | FW | CGTGAGAACGCGTCGAATAA | ChIP | 5 | 95 - 110 | [[77](#_ENREF_77)] |
| RV | TTCTGGTTCTGGAATGAGGG |
| MLV (Gln-PBS) | FW | GAGGTCCCACCGAGATTTGGA | ChIP | 32 | 153 |  |
| RV | AGCTAACTAGTACAGACGCAGG |
| MLV (gag) | FW | AGTCTAACCTTGCAGCACTGG | ChIP | 59 | 137 | [[75](#_ENREF_75)] |
| RV | GTACCATCCTGAGGCCATCC |
| *Lef1* promoter | FW | ATCAGTCATCCCGAAGAGGA | ChIP | 1 | 107 | [[75](#_ENREF_75)] |
| RV | AGCTGCCCACTCACCTCAT |
| *Hprt* promoter | FW | CACCGCCCGGCTGGATCT | ChIP | 1 | 151 | [[131](#_ENREF_131)] |
| RV | CCCAGAGGATTCCCAGATAATCACT |
| *Mest* promoter | FW | CAGCAGCTTCTGGCATGTGG | ChIP | 1 | 454 | [[95](#_ENREF_95)] |
| RV | AACCCCAGATTCTAGTGAAG |
| *Zfp809* (exon 5) | FW | TCGAAAGTCCAAACTCGCTATAC | ChIP/qPCR | 1 | 132 |  |
| RV | TTCTCAGATGAATGAGCAATCCG |
| *Zfp809* (DNA binding domain) | FW | GGCTCGACAGGAAGTCTCCAA | PCR (sequencing) | 1 | 824 |  |
| RV | GGGATTGTGTGGTATACCGGTTAC |
| *ActB* mRNA | FW | CTTCCTCCCTGGAGAAGAGC | RT-qPCR | nd | 124 | [[132](#_ENREF_132)] |
| RV | ATGCCACAGGATTCCATACC |

# Target copy number in the mouse genome (mm9 assembly) was determined by the *in silico* PCR tool of the UCSC Genome Browser.

# Chapter 3: Genome-wide analysis of ZFP809 binding sites

## Introduction and summary

As described in chapter 2, we have shown that the retroviral repressor protein ZFP809 is mainly conserved in mice, hamsters and blind mole rats. Considering the diversity of PBS sequences in exogenous and endogenous retroviruses, it seems surprising that ZFP809 did not adapt to new targets in any of these species. We therefore speculated that ZFP809 might have functions apart from retroviral silencing such as gene regulation, which could possibly explain the observed lack of positive selection in the ZFP809 DNA binding domain.

Although the mechanism of gene silencing by KRAB-ZFP/TRIM28 complexes has been intensively investigated previously [[89](#_ENREF_89),[97](#_ENREF_97),[125](#_ENREF_125),[133](#_ENREF_133)], these studies were focused on functional aspects of heterochromatin formation using experimental setups in which TRIM28 is artificially tethered to the genome (e.g. fusion of a KRAB domain to the DNA binding domain of the Tet repressor). However, a recent genome-wide TRIM28 binding analysis in human EC cells revealed thousands of TRIM28 binding sites in the genome [[134](#_ENREF_134)]. TRIM28 knock-down in human EC cells led to a downregulation of approximately 140 genes but also upregulation of more than 360 genes. A large proportion of these genes were found to be bound by TRIM28, indicating that the change in expression was mediated by TRIM28 binding and not by secondary effects. However, the majority of dysregulated genes were not targeted by TRIM28 at the promoter region and it remains unknown how TRIM28 regulates expression of these genes [[134](#_ENREF_134)]. It has also been shown that TRIM28 binding to gene promoters in HEK293 cells was independent of the RBCC domain that is known to interact with the KRAB domain of KRAB-ZFPs [[135](#_ENREF_135)]. In contrast, TRIM28 binding sites at the 3’ coding exons of zinc finger genes was dependent on the RBCC domain but expression of these genes was not controlled by TRIM28 [[135](#_ENREF_135)].

So far, few studies have been performed to investigate genome-wide binding sites of KRAB-ZFPs. The genomic binding sites of the human KRAB-ZFP ZNF274 were found to co-localise with TRIM28, SETDB1 and the H3K9me3 mark in human EC and 293 cells [[136](#_ENREF_136)]. However, these binding sites were almost exclusively located at the 3’ coding exons of KRAB-ZFP genes on which the zinc finger domains are located. It was therefore speculated that ZNF274 does not primarily function as a transcriptional repressor but rather maintains heterochromatin structures at homologous KRAB-ZFP DNA binding domains to prevent chromosomal recombination [[136](#_ENREF_136)]. In contrast, the binding sites of the human ZNF263 were found to be preferably located near transcriptional start sites (TSS) and core promoter regions of genes [[137](#_ENREF_137)] in human EC and 293 cells. However, no overlap between ZNF263 binding sites and H3K9me3 marks has been reported [[136](#_ENREF_136),[137](#_ENREF_137)] and ZNF263 targeted genes were either up- or downregulated upon ZNF263 knock-down [[137](#_ENREF_137)], indicating that KRAB-ZFPs and TRIM28 may not exclusively function as repressor proteins that induce H3K9 trimethylation. Indeed, TRIM28 has also been associated with transcriptional activation yet the mechanism by which this activation is mediated remains largely unknown [[138](#_ENREF_138)].

ChIP in combination with next-generation sequencing techniques (ChIP-seq) is a powerful tool to detect genome-wide binding sites of transcription factors and has rapidly replaced hybridisation-based assays during the last few years. We have generated F9 cells overexpressing FLAG-tagged ZFP809 and have shown that this protein binds efficiently and highly specifically to genomic target sequences containing the Pro-PBS repressor binding sequence (chapter 2). In this chapter, we identified genomic ZFP809 targets in repetitive elements, such as ERVs, and in the non-repetitive fraction of the mouse genome using ChIP-seq technology.

We found that ZFP809 primarily targets Pro-PBS containing ERVS but also shows some affinity to several ERV target sequences that are similar but clearly distinct from the Pro-PBS. Furthermore, we identified a large number of ZFP809 binding sites in the non-repetitive fraction of the mouse genome and found that these binding sites are preferably located near TSS of genes. ZFP809 knock-down in F9 cells did not lead to an upregulation of Pro-PBS containing ERVs. Also the majority of genes that were targeted by ZFP809 in their promoter region were not upregulated upon ZFP809 knock-down. Furthermore, these genes were not marked by heterochromatin markers at the ZFP809 binding sites. However, we detected TRIM28 binding at some of these regions and overexpression of ZFP809 led to transcriptional silencing by heterochromatin formation of some of the ZFP809 targeted genes. Finally, we identified a gene with a strong ZFP809 binding site in the promoter region which is marked by repressive histone methylations in ES cells. ZFP809 knock-down led to transcriptional upregulation of this gene, indicating that ZFP809 may function as a transcription factor regulating gene expression.

## Materials and methods

#### Library construction for ChIP-seq

The generation of F9 EC cells stably expressing *M. musculus* FLAG-tagged ZFP809 (F9 [pSB/KRAB-ZFP809]) is described in chapter 2. ChIP was performed according to a protocol that that has been optimised for ChIP-seq with KRAB-ZFPs [[139](#_ENREF_139)]. Briefly, 1 × 108 F9 [pSB/KRAB-ZFP809] cells were harvested by trypsinisation, washed and cross-linked with formaldehyde. Chromatin was sonicated for 20 minutes on a Covaris S2 sonicator using the following settings: 200 cycles/burst, 20% duty cycle and intensity set to 8. After this treatment, the size of the sonication fragments was between 100 and 300 bp, as confirmed by gel electrophoresis. Chromatin was then immunoprecipitated with 10 µg of monoclonal ANTI-FLAG® M2 antibody (Sigma-Aldrich). After reverse-crosslinking, DNA was purified and eluted in 12 µl of elution buffer. Immunoprecipitated DNA (ca. 20 ng), together with an input control (DNA purified from sonicated and reverse cross-linked chromatin without immunoprecipitation), was submitted for library construction and Illumina high-throughput sequencing at the BGI sequencing facility (http://www.genomics.cn). High quality reads (adapter sequences and reads with quality values <=20 in more than 50% of the sequenced nucleotides and reads with more than 10% “N” content were removed) of 49 nucleotides’ length were generated. Sequencing quality was confirmed by the *Fastqc* tool on the Galaxy platform (https://main.g2.bx.psu.edu/).

#### Read mapping and peak calling

Reads from both samples were mapped to the mouse genome (mm9) using the Bowtie alignment tool available at the Galaxy platform under the pre-set commonly used settings. Briefly, the seed length for each read alignment was set to 28 bp allowing two mismatches. Since the reads have already been filtered and trimmed and we confirmed the high quality of the sequenced reads, all positions of the reads were used for global alignment. Peak calling was performed using the Model-based Analysis of ChIP-Seq (MACS) (version 1.0.1) [[140](#_ENREF_140)] which is available on the Galaxy platform. The program was run at default settings except the Mfold parameter, setting the minimum enrichment of ChIP sample reads over the control reads to find peaks for building the shifting model, which was set to 10. The following pre-set parameters have been used: Band width: 300 bp; p-value cut-off for peak detection: 1e-05; three levels of regions around the peak region to calculate the maximum lambda as local lambda: 1000, 5000 and 10000. To prevent false peak calling caused by multiple tags at the exact same position and strand orientation, often a result of the PCR amplification step during library construction, MACS calculates the maximum tags at the exact same location based on a binominal distribution using 1e-5 as p-value cut-off. For each enriched peak region, MACS calculates the ChIP-seq read enrichment over the input reads as well as the peak summit that represents the position of the peak in which the read enrichment was at its maximum.

#### Target motif identification

The 100 top-scored (as calculated by MACS) non-repetitive (<20% repetitive sequence content) ChIP-seq peaks were trimmed to 100 bp up- and downstream of the peak summit and used for motif identification by the *CisFinder* program (http://lgsun.grc.nia.nih.gov/CisFinder/) using the following settings: False discovery rate (FDR): 1.E-04; Minimum enrichment ratio: 1.5; Match threshold for clustering: 0.55. The identified motifs were clustered by similarity and linkage.

#### Genomic correlation studies

All correlation studies were performed using tools and semi-automated workflows on the Galaxy platform. Core promoter regions of Refseq genes were defined as 4 kb regions around the TSS of Refseq genes that are annotated in the UCSC Genome Browser. The genomic coordinates of previously identified ChIP-seq peaks were downloaded as BED files from the UCSC Table Browser and used for correlation studies performed on the Galaxy platform.

#### ZFP809 knock-down by short hairpin (sh) RNAs

F9 cells stably transfected with a pSuper-retro vector expressing a shRNA directed against the *Zfp809* mRNA or a non-specific scramble shRNA were kindly provided by Sharon Schlesinger and have been described previously [[94](#_ENREF_94),[103](#_ENREF_103)]. Cells were received as viable freeze stocks, seeded on gelatine coated plates and grown for two passages. Three days before RNA isolation, culture medium was changed to puromycin-containing selection medium to ensure stable shRNA expression of transduced cells.

#### RT-qPCR and ChIP-qPCR

ERV and gene expression analysis by RT-qPCR in F9 cells was performed as described in chapter 2. ChIP-qPCR in F9 cells overexpressing FLAG-ZFP809 was performed as in chapter 2.

#### Viral titer assays

Putative ZFP809 target sequences were inserted into the previously described LJ-QAdMLPEnh- vector [[1](#_ENREF_1)] (see Fig. 3.15A). For this, double-stranded (ds) 29 bp long oligos were created by annealing single-stranded (ss) sense and antisense oligos. The sense and antisense ss-oligos contained flanking nucleotides 5’ and 3’ of the 29 bp target sequences that generated a short *BamHI* or *NotI* compatible ss-overhang in the 5’ and 3’ end, respectively, of the annealed ds-oligo. The pLJ-QAdMLPEnh- plasmid vector was digested with *BamH1* and *Not* and annealed ds-oligos were inserted by ligation. Transduction of F9 and NIH/3T3 cells was performed as described in chapter 2. Viral titers (CFU/ml) were determined by counting resistant colonies after 10 to 12 days of selection.

## Results

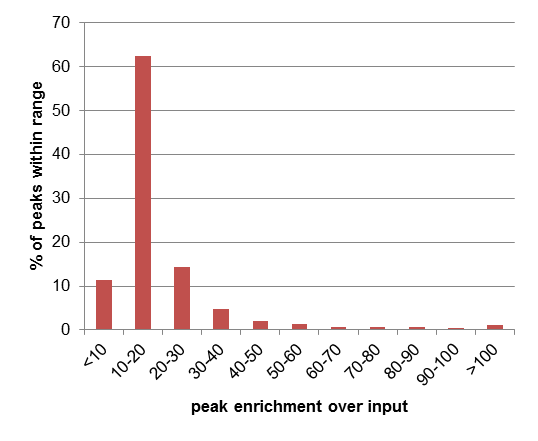
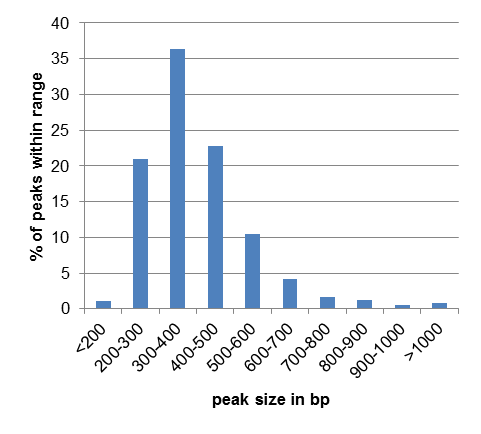
### Genome-wide ZFP809 binding site identification by ChIP-seq

To detect genomic binding sites of ZFP809, we performed a genome-wide binding analysis by ChIP-seq. For this, cross-linked chromatin of F9 EC cells stably expressing FLAG-tagged ZFP809 (as described in chapter 2) was immunoprecipitated with an anti-FLAG antibody and enriched DNA samples were submitted for library construction and Illumina/Solexa sequencing. Approximately 1.3 × 107 high quality reads were generated from both the immunoprecipitated sample as well as a control sample which contained DNA that had been purified from sonicated and reverse cross-linked chromatin without immunoprecipitation. The sequenced reads generated from both samples were mapped to the mouse genome (mm9) using the Bowtie mapping tool. For reads that could be aligned to multiple genomic loci, only one valid alignment, which was selected among all possible alignments in an unbiased manner using a pseudo-random number generator by Bowtie, was allowed.

**Table 3.1: Mapping efficiency of sequenced ChIP-seq reads to the mouse genome**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Read length | No. of produced high quality reads | No. reads mapped to the mouse genome | Percentage of reads that could be mapped |
| FLAG-ZFP809 | 49 bp | 12,962,032 | 12,459,245 | 96.12% |
| Input control | 49 bp | 12,844,971 | 12,115,145 | 94.32% |

As shown in Table 3.1, the majority of the ChIP-seq reads could be successfully mapped to the mouse genome. Next, we determined genomic regions in which reads sequenced from the FLAG-ZFP809 sample were significantly overrepresented as compared to reads from the input control using the Model-based Analysis of ChIP-Seq (MACS) algorithm. Using the settings described in the methods section, a total of 9011 ChIP-seq peaks were detected by MACS and analysed for peak length distribution and enrichment over input control. As shown in Fig. 3.1A, the majority of the peaks were between 300 and 600 bp long, as expected for a sequence-specific DNA binding protein. The read enrichment over the input control was less than 30-fold for about 90% of the peaks, yet several peaks were enriched more than a 100-fold over input (Fig. 3.1B). Since the analysis was not performed in duplicates which would allow identifying *bona fide* ZFP809 binding sites with a higher fidelity, we attempted to remove low quality peaks that may represent false hits. For this, we filtered out all peaks that did not match the following criteria: At least ten FLAG-ZFP809 ChIP-seq reads mapped within the peak region, at least ten-fold read enrichment over the input control and peak size less than 2,500 bp. After applying these filtering criteria, 6499 peaks remained for further analysis. Since we observed that the peak regions varied greatly in their enrichment values (Fig. 3.1B), which may be relevant for their biological significance, all filtered 6499 peaks were categorised into four groups: Very strong peaks (n=105, enrichment: >100); strong peaks (n=333, enrichment: 50-100); moderate peaks (n=1918, enrichment: 20-50) and weak peaks (n=4143, enrichment: 10-20).



**A**

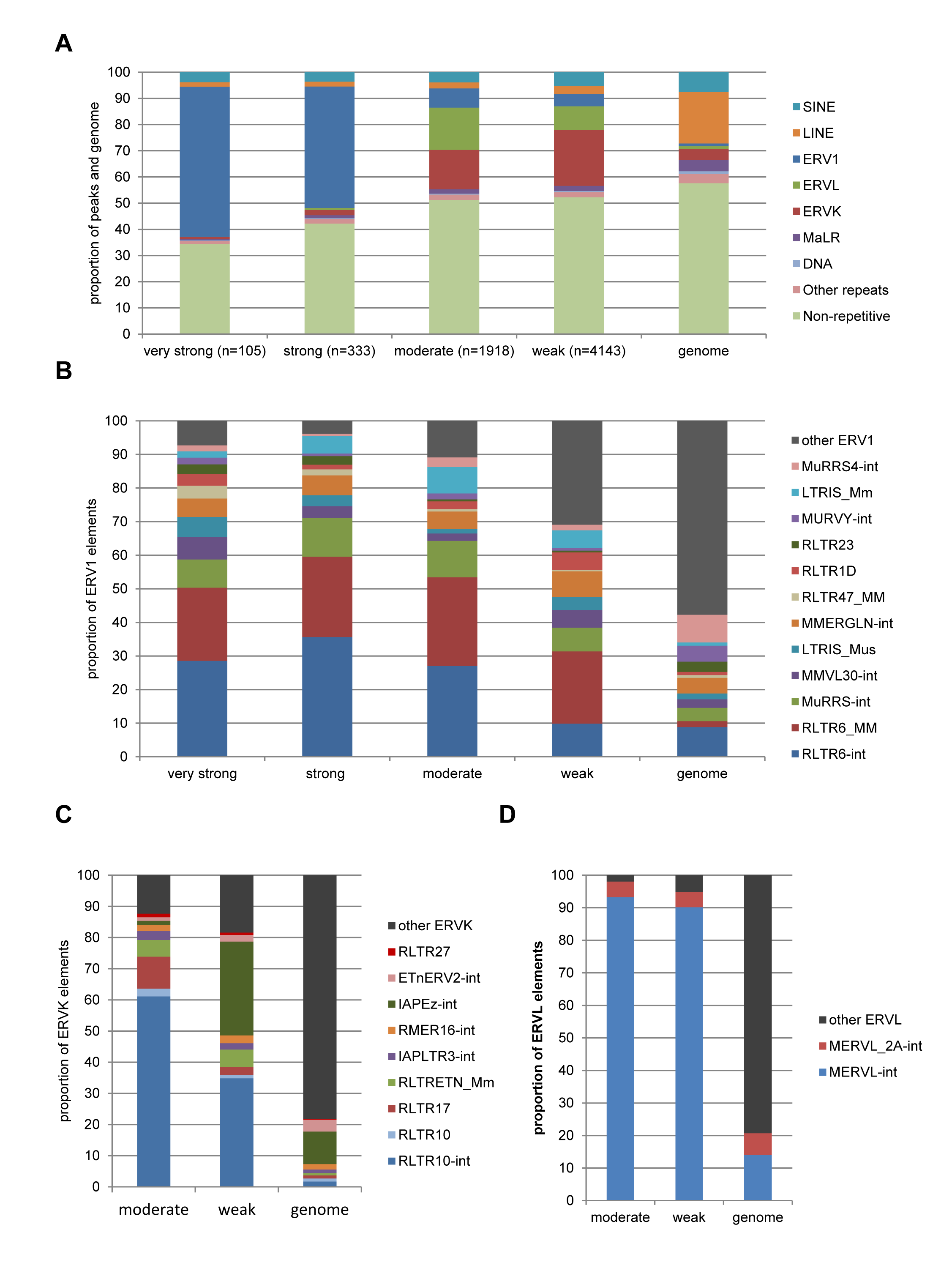
**B**

**Figure 3.1: Peak size and enrichment distribution.** Histograms show the proportion of FLAG-ZFP809 ChIP-seq peaks that fall within the specified range of (A) peaks size and (B) relative enrichment over input read density.

### Distribution of ZFP809 binding sites within repetitive elements

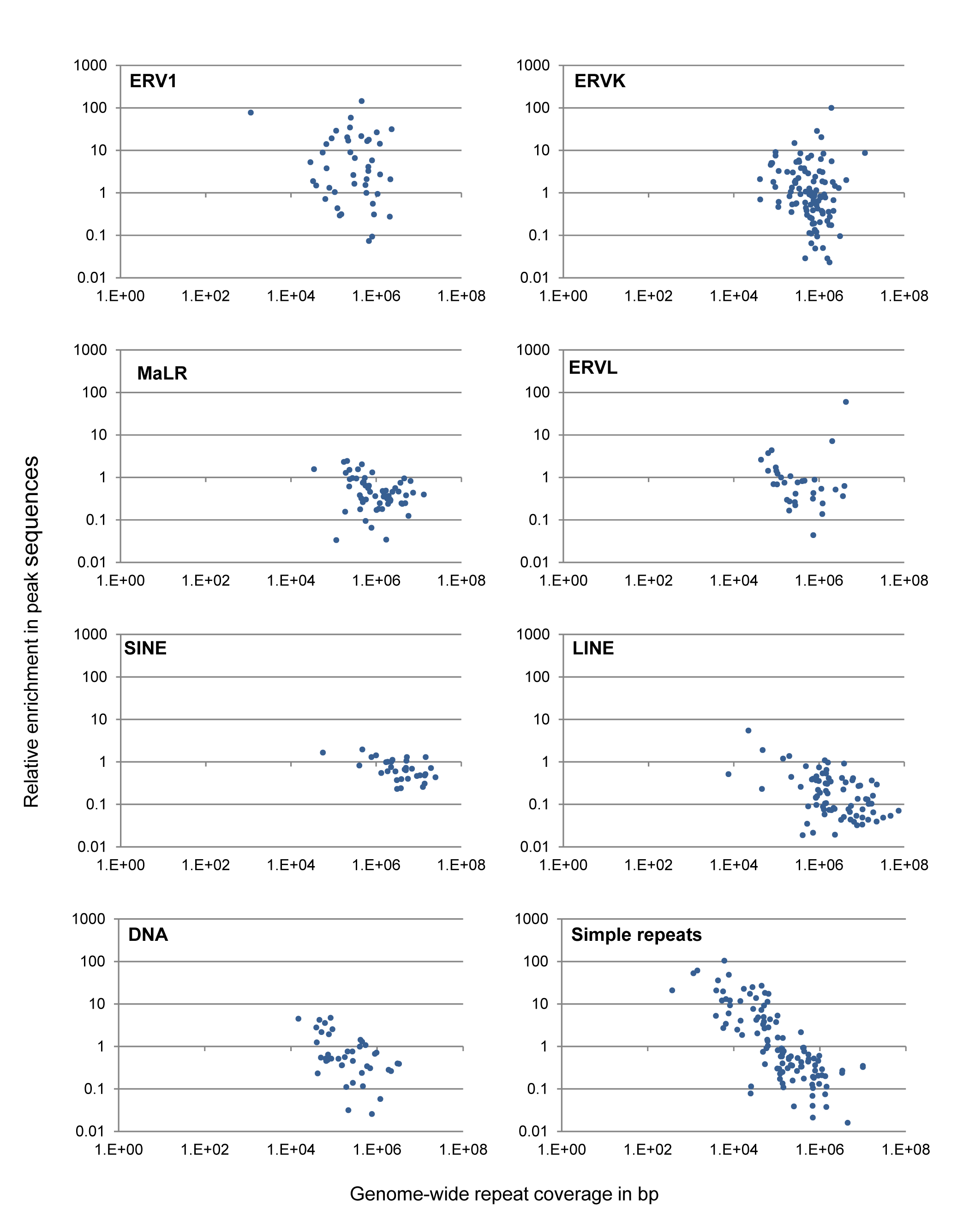
Since ZFP809 is known to bind to a retroviral PBS sequence, we first determined the content of repetitive elements in the categorised peak regions by correlating the genomic peak coordinates with the coordinates of all murine repetitive elements annotated in the in the Repeatmasker track of the UCSC Table Browser. As shown in Fig. 3.2A, peaks with a highly enriched FLAG-ZFP809 read density were almost exclusively located within annotated ERVs of the ERV1 family or in the non-repetitive fraction of the genome. Since ERV1 elements only constitute less than 1% of the mouse genome, this indicates that these elements are targeted in a highly specific manner. ERV1 elements were also found to be enriched in moderate and weak peaks (Fig. 3.2A). Additionally, ERVK and ERVL elements were overrepresented in moderate and weak peaks. In contrast, SINE and LINE retrotransposons and MaLR ERVs were generally underrepresented in all peak sequences.

Next, we determined which ERV groups are most abundant found in the peak fraction that is located within ERV1, ERVK and ERVL elements. As shown in Fig. 3.2B, several groups of ERV1 elements could be found within the peak regions. Internal RLTR6 (RLTR6-int) regions together with their associated LTR elements (RLTR6\_Mm) constituted between 30 and 60% of the peak regions generally located within ERV1 elements in all peak categories. As we have shown in chapter 2, this ERV, which is also referred to as MmERV, contains the majority of the genomic Pro-PBS sequences in the mouse genome and was therefore predicted to be one of the main targets of ZFP809. Since the PBS is located at the 5´end of the internal RLTR6 region, a peak covering this sequence is expected to overlap with the corresponding 5´LTR sequence and the enrichment of RLTR6 LTR regions in the peak regions therefore does not necessarily indicate that these LTRs are directly bound by ZFP809. In chapter 2, we have shown that Pro-PBS sequences can also be found in VL30, MuRRS, MMERGLN, MuRRS4 and MURVY elements. Indeed, these elements and their LTRs constituted a major fraction of the highly enriched peak regions (Fig. 3.2B). Peaks overlapping with RLTR47\_MM, RLTR1D and RLTR23 elements were re-examined by determining which other repetitive elements are co-localised with these LTRs. We found that virtually all of the peaks containing these repeats also overlapped with internal regions of ERV groups which contain Pro-PBS sequences, indicating that these elements may be located close Pro-PBS sequences and not be targeted by ZFP809 themselves. Since ERVK and ERVL elements were overrepresented in moderate and weak peaks, we next analysed whether these elements are generally or rather selectively enriched. A shown in Fig. 3.2C, more than 60% of the ERVK elements found in moderate peaks belonged to internal RLTR10 regions, which only constitute less than 2% of all ERVK elements in the mouse genome. Also the corresponding LTR elements (RLTR10) were frequently found in these peaks indicating that the ZFP809 binding site might be located close the 5’ or 3’ end of the internal region. The second and third largest fractions of the moderate peaks could be assigned to RLTR27 and RLTRETN\_MM elements, respectively. The majority of the peaks overlapping with these elements did not match to any other repeats indicating that these LTRs might be ZFP809 targets. RLTR10 elements also constituted the largest ERVK fraction in weak peaks; however, an approximately equal proportion of these peaks were located within internal regions of IAPEz elements. About 95% of the ERVL elements in the moderate and weak peak regions could be assigned to internal regions of two ERVK elements, MERVL and MERVL\_2A (Fig. 3.2D).



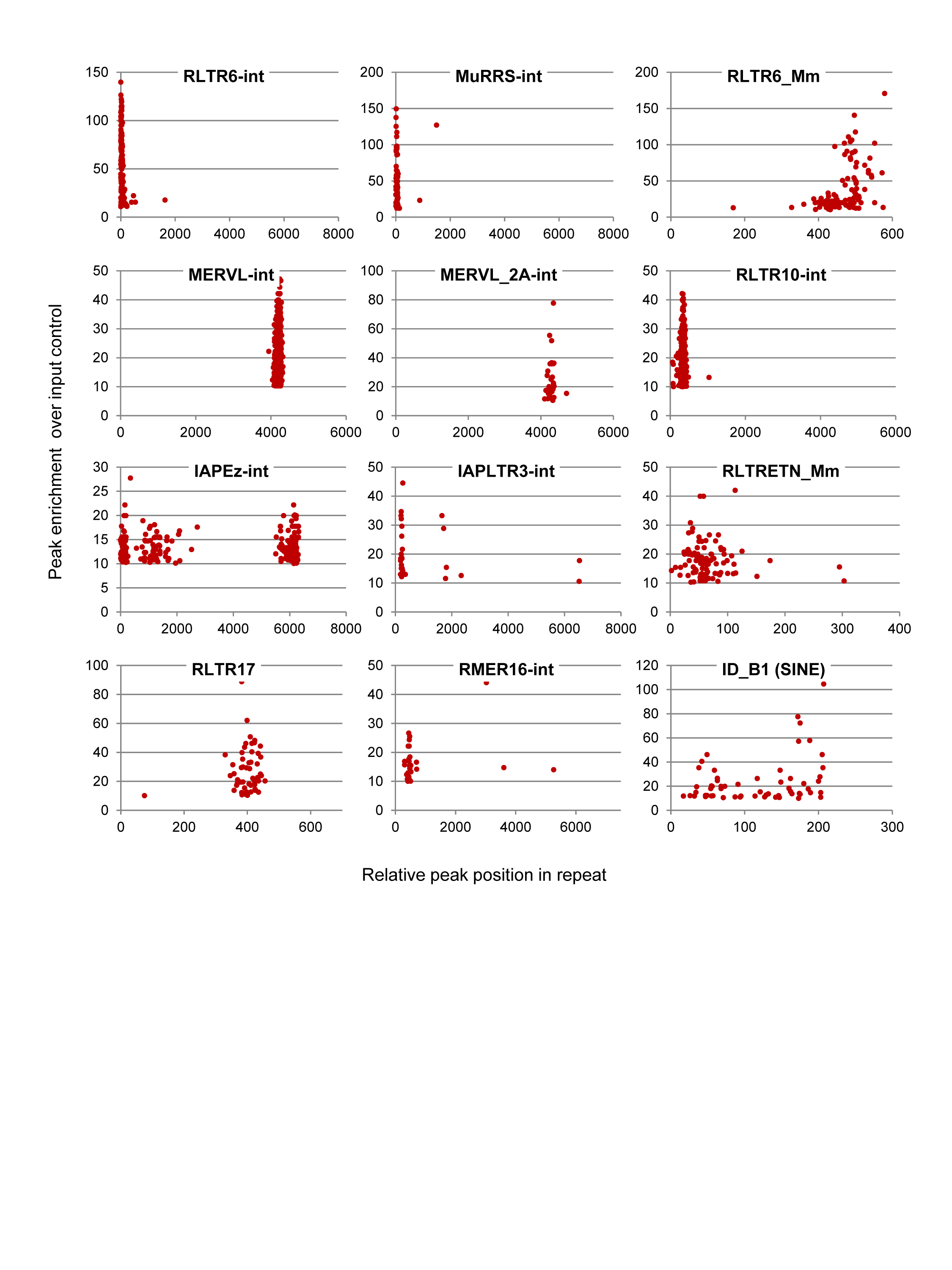
**Figure 3.2: Repeat content in ZFP809 ChIP-seq peaks**. All peaks that passed the filtering criteria were categorised into four groups according to their enrichment values and analysed for overlapping Repbase elements annotated in the UCSC Genome Browser. (A) Repeat content in ChIP-seq peaks and mouse genome by repeat class and ERV families. (B) Proportion of ERV1 elements in peaks and mouse genome. (C) Proportion of ERV-K elements in moderate and weak peaks. (D) Proportion of ERVL elements in moderate and weak peaks.

So far, we focused on identifying abundant repetitive elements in peak regions that may represent major ZFP809 targets. However, also low copy elements might be targeted in a sequence-specific manner yet those elements would only constitute a small fraction of the peaks and, thus, remain undetected. We therefore determined the relative enrichment for each repeat type that overlapped with at least one of the 6499 peak regions. The enrichment was calculated as the ratio of the proportion of peaks covered by the repeat type to the proportion of the mouse genome covered by the repeat. The relative enrichment of each repetitive element was plotted against the genomic coverage of the same element. As shown in Fig. 3.3, the majority of ERV1 elements overlapping with peaks were proportionally overrepresented in these regions. Also a large number of ERVK elements were overrepresented in the peaks, whereas the majority of MaLR and ERVL elements were found at lower frequency in the peaks than expected by chance. SINE and LINE retrotransposons as well as DNA transposons were generally underrepresented in peaks, indicating that those elements were not targeted by ZFP809. Interestingly, we also identified several low copy simple repeats that were highly overrepresented in peaks regions.



**Figure 3.3: Relative enrichment of repetitive elements overlapping with ZFP809 ChIP-seq peaks.** Each data point represents a repeat type that was plotted according to the total genomic coverage and the fold enrichment of the repeat in FLAG-ZFP809 ChIP-seq peaks. The enrichment values of repeats that were at least five fold enriched in peaks are shown in Supplementary Tables S3.1 and S3.2.

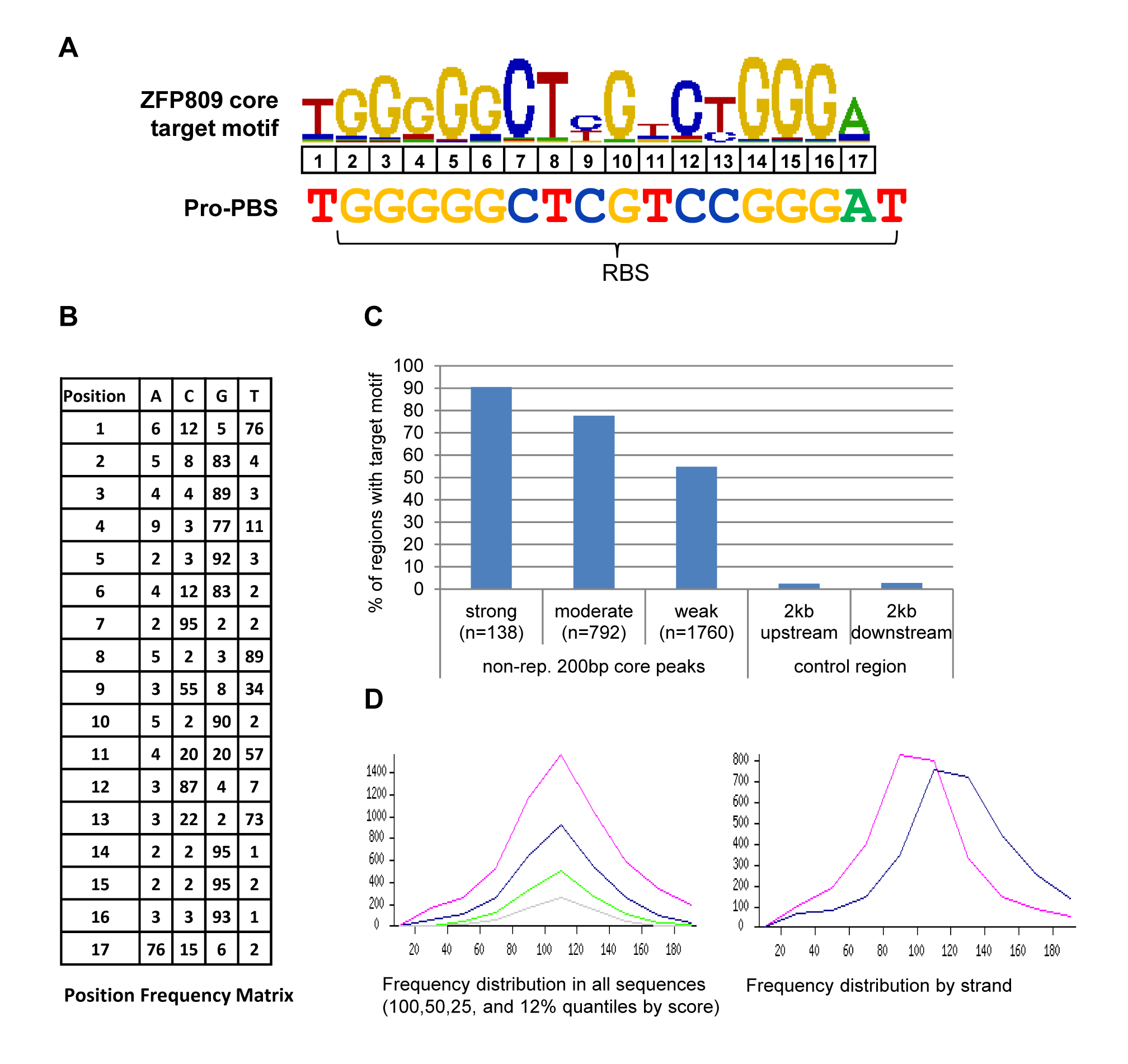
For all repetitive elements annotated in the UCSC Genome Browser, information about their relative position to the determined repeat consensus sequence is available in the UCSC Repeatmasker track. It is therefore possible to determine whether regions located within repetitive elements of the same type are found at the same relative position as the repeat consensus sequence. To determine whether the identified targeted repetitive elements are bound by ZFP809 in a sequence-specific manner, the genomic coordinates of the exact peak summits were correlated with their relative position within the annotated repetitive elements. The majority of the peak summits within internal regions of RLTR6 and MuRRS elements were located at the 5’ end of the internal region (Fig. 3.4). This was expected since also the Pro-PBS of these elements is located at this position. Also peak summits within internal regions of other ERV1 elements containing Pro-PBS sequences, such as MuRRS, VL30 and MMEREGLN, were almost exclusively located at the PBS (data not shown). To confirm that Pro-PBS containing ERVs are generally targeted by ZFP809, we screened the peak regions for Pro-PBS sequences and we found that virtually all of the Pro-PBS sequences in the mouse genome are covered by highly enriched peaks (data not shown). However, also several other ERVs in which we did not identify Pro-PBS sequences seemed to be targeted by ZFP809 in a sequence-specific manner (Fig. 3.4). In RLTR10-int elements the peak summits were almost exclusively located at a position about 300 bp downstream of the PBS and in MERV-L-int elements, the peaks were concentrated around a region about 4 kb downstream of the PBS. However, the peaks in these ERVs were significantly less enriched over input indicating that these binding sites have a lower affinity to the ZFP809 DNA binding domain.



**Figure 3.4: Relative position of ZFP809 ChIP-seq peak summits in a selection of repeat types.** Data points represent peak summit positions of peaks that overlapped with the indicated repeat type and were plotted according to their relative position in the repeat as annotated in the UCSC *Genome Browser* and the enrichment of the peak in which the summit was located. The range of the x-axis represents the approximate length of the repeat consensus sequences.

### Identification of a ZFP809 target motif in non-repetitive peaks

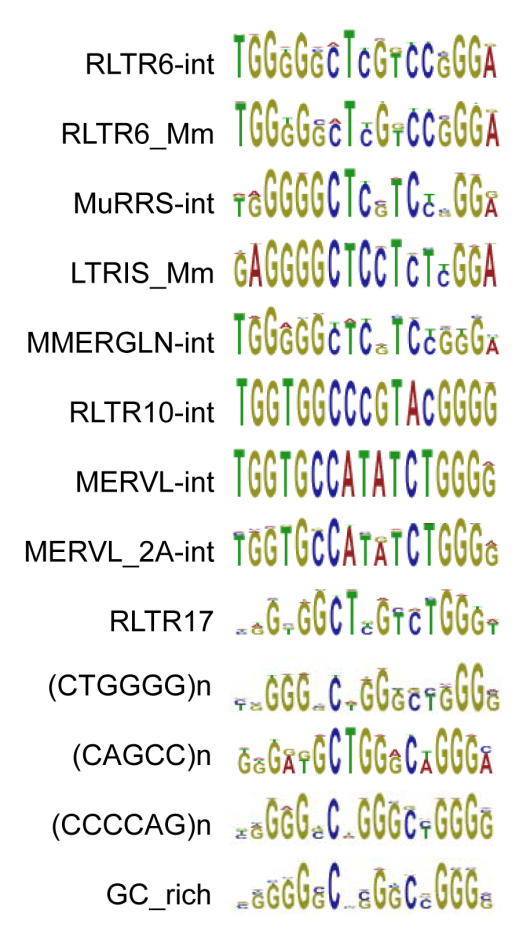
The identification of commonly occurring target motifs based on peaks in repetitive sequences is nearly impossible since such peaks regions share sequence homologies over the entire peak region that would be mistakenly identified as target motifs by motif prediction tools. We therefore employed peak regions in the non-repetitive fraction of the genome to identify potential ZFP809 target motifs. For this, we filtered out all peaks that contained more than 20% repetitive elements and of the remaining 2690 non-repetitive peak regions, the 100 top-scored peaks were used for motif identification. Since target motifs are expected to be located close to the peak summit, these peaks were trimmed to a 200 bp region around the peak summit prior to motif identification using the *CisFinder* tool. The program was run at various settings, however, a virtually identical 17 nucleotide target motif was found in more than 94% of the non-repetitive peaks that were used for motif identification. As shown in Fig. 3.5A, the identified motif consensus is highly similar to the Pro-PBS sequence, with only one nucleotide (T instead of C at position 13) difference and two positions (9 and 11), in which the nucleotides were less conserved in the individual motifs. This indicates that ZFP809 is recruited to non-repetitive targets using the same motif as used for targeting the retroviral Pro-PBS sequence. To determine motif occurrence in the remaining peak regions, we trimmed all peak regions with less than 20% repeat content to 200 bp core peak sequences and divided them into three categories based on the enrichment values as described above. Since less than 50 of the non-repetitive peaks were enriched more than 100 fold over input, these peaks were combined with the peaks with enrichment values between 50 and 100 and categorised together as “strong peaks”. The trimmed peaks were screened for the identified ZFP809 target motif using the *CisFinder* toolunder settings allowing only one false positive match per 10 kb. In approximately 90% of the 138 strong non-repetitive peaks, the inferred target motif was identified. The percentage of peaks that contain a target motif was lower in moderate and weak peaks but still clearly higher than in the control regions up- and downstream of the peaks, indicating that ZFP809 recruitment to these regions depends on a target motif.



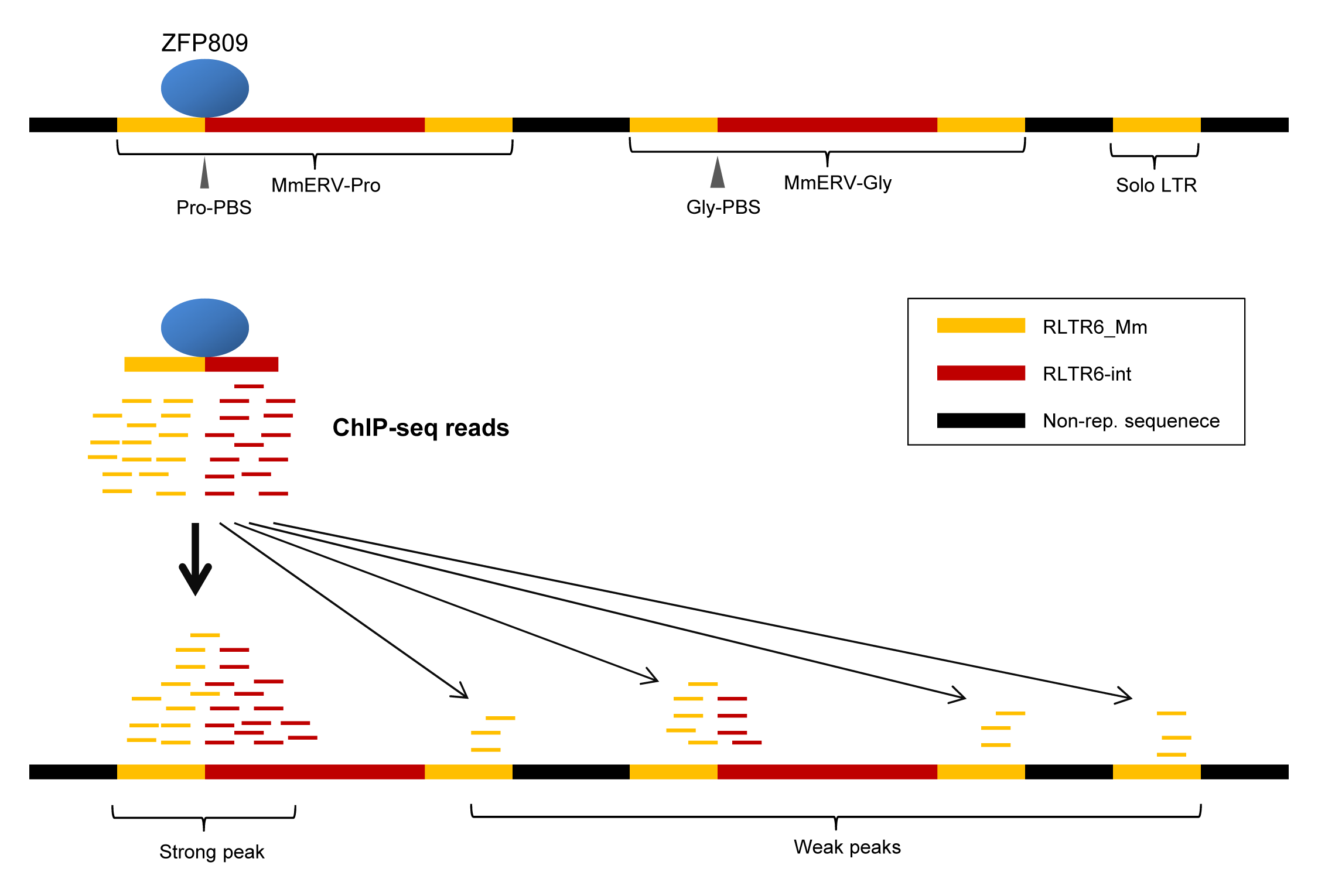
**Figure 3.5: Target motif identification in non-repetitive ZFP809 ChIP-seq peaks.** (A) ZFP809 consensus core target motif derived from the 100 top-scored non-repetitive 200 bp ChIP-seq core peaks by the *CisFinder* program in comparison to the Pro-PBS sequence. The RBS, originally determined as target for ZFP809/TRIM28 complex binding, is indicated with a brace (B) Position frequency matrix showing the frequency of nucleotide occurrence at each position of the determined core target motif. (C) Target motif occurrence in non-repetitive peaks categorised by their enrichment values and in 200 bp control regions located 2 kb up- and downstream of all non-repetitive peaks. (D) Frequency distribution of the target motif position in all non-repetitive 200 bp core peaks by score and by strand.

### ZFP809 target motifs in repetitive elements

Next, we wanted to identify the inferred ZFP809 target motif in repetitive elements. For this all 6499 peak regions were trimmed to the 200 bp core sequence around the peak summit and screened for the ZFP809 target motif under less stringent settings that allowed 5 false hits per 10 kb. A total of 5904 motif sequences were detected in 4960 of all 6499 peaks. The genomic coordinates of these motifs were then associated with all annotated repetitive elements in the UCSC Table Browser to identify target motifs in peaks that overlap with repeats. Supplementary Tables S3.2 and S3.3 show a summary of this analysis. To distinguish repeats that are directly bound by ZFP809 from repetitive elements that are merely located close to other elements containing a target motif and are therefore enriched in peak regions, we determined the number of peaks that are overlapping with a certain repeat type as well as the number of peak summits that are located within the repeat. A major discrepancy between these numbers indicates that either the target motif is located close to the end of a repeat or within a nearby located element. This is the case for Pro-PBS containing elements such as RLTR6, VL30 or LTRIS, in which the Pro-PBS target is located at the 5’ end of the internal region, directly adjacent to the 5’LTR (Supplementary Table S3.1). In contrast, the majority of the peaks within RLTR10-int and MERVL-int elements also had their summits located in these elements. As shown in Supplementary Table S3.1, 153 peaks summits were located within internal RLTR6 regions and 125 motifs were identified within these peaks. Since we counted all motifs that overlap with elements of a given repeat type by at least one bp, motifs that are located at the boundary of two repeats could be counted twice due to incorrect repeat annotation on the UCSC Genome Browser. Indeed the determined consensus motifs identified in RLTR6-int elements and RLTR6\_Mm LTRs were nearly identical (Fig. 3.6) and we could not detect a ZFP809 target motif in the RLTR6\_Mm consensus sequence, indicating that some Pro-PBS sequences are at least partially wrongly annotated as RLTR6\_Mm elements. Of the 125 ZFP809 target motifs in RLTR6-int elements, 74 were identical with the Pro-PBS sequence. We also found several mutated Pro-PBS and intact Gly-PBS sequences the ChIP-seq peaks within RLTR6-int elements. Since these sequences are similar to the Pro-PBS, they were identified as motifs by the *CisFinder* program. However, this does not necessarily imply that all of these sequences are targeted by ZFP809 and might be merely due to the sequence similarity of MmERV-Pro and MmERV-Gly elements in the LTR and 5’UTR regions. The Bowtie mapping tool distributes reads that can be aligned to multiple genomic loci randomly. Therefore, many reads that were actually derived from MmERV-Pro elements are likely to be aligned to other MmERV elements that lack the ZFP809 binding site (Fig. 3.7). To overcome this problem only reads that can be aligned to a unique genomic locus may be considered for peak calling. However, this would introduce a strong bias since peaks in well conserved elements would not be identified at all. Nevertheless, we observed that peaks on intact RLTR6-int Pro-PBS sequences are clearly more strongly enriched over input than peaks on RLTR6-int Gly-PBS sequences (data not shown). We observed the same effect for other ERVs groups of which only certain members contain a Pro-PBS. To test whether Gly-PBS sequences are targeted by ZFP809 or only appear within ZFP809 ChIP-seq peaks due to the homology of their flanking sequences to RLTR6-int elements with Pro-PBS sequences, we analysed Gly-PBS sequences in other ERV groups and found that none of them was overlapping with a ChIP-seq peak (data not shown). This indicates that only MmERV-Pro but not MmERV-Gly elements are targeted by ZFP809 and highlights the importance of a careful interpretation of ChIP-seq peaks in repetitive elements. The motif logos derived from aligned target motifs in RLTR6-int and other repeat types (Fig. 3.6) therefore do not necessarily indicate that a large number of different target motifs sequences is bound by ZFP809 in these repeats but may merely reflect the sequence diversity at the position at which ZFP809 binds in at least some elements. Notably, the target motifs in RLTR10 and MERV-L elements were well conserved and we did not find Pro-PBS sequences in any of these elements, indicating that these target motifs may indeed be bound by ZFP809. However, we cannot exclude that only a small subset of these ERVs are actually bound by ZFP809 and randomly mapped reads are responsible for the large number of peaks at these elements (see Fig. 3.7).



**Figure 3.6: ZFP809 target motifs in repetitive elements**. Identified target motifs that were found in peaks overlapping with repetitive elements in more than 20 copies were aligned according to their matching repeat type. Shown are the consensus sequence logos of the aligned motifs.



**Figure 3.7: False peak calling by randomly mapped repetitive ChIP-Seq reads**. ZFP809 binds to the Pro-PBS in the internal region of MmERV-Pro proviruses. After ChIP, reads are sequenced and mapped to the genome. Reads that cannot be uniquely aligned to a genomic region are randomly distributed among all possible mapping sites, leading to enriched read density at regions, such as MmERV-Gly proviruses or solo LTRs, and potentially peak calling at regions that are not targeted by ZFP809.

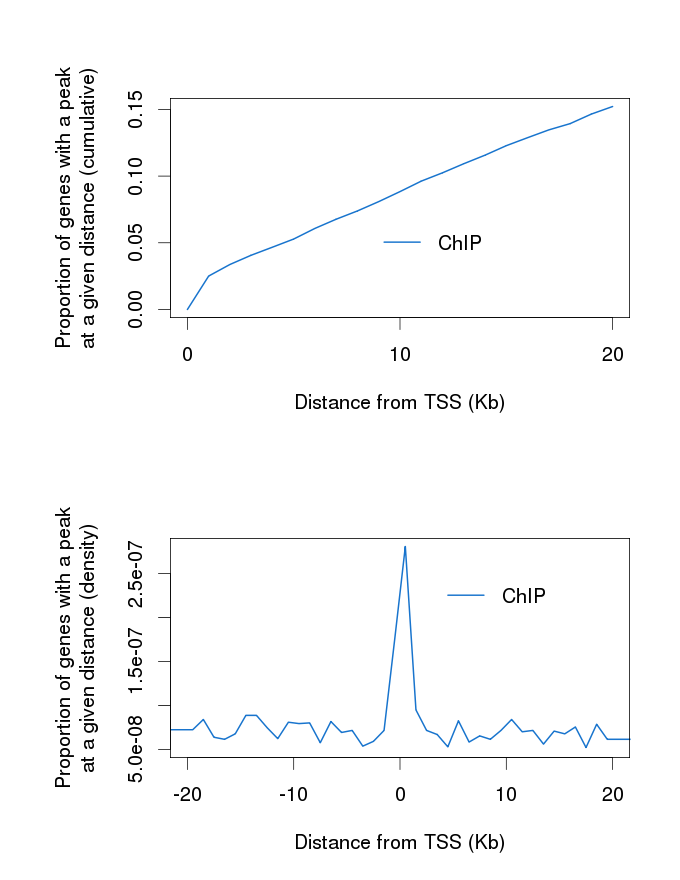
Interestingly, the identified motif in RLTR10-int elements, which differs from the Pro-PBS in four nucleotides, is complementary to the 3’ end of nine murine isoleucine-tRNAs (referred to as mIle9 in chapter 2). This putative PBS was found at high copy numbers in the mouse genome (chapter 2, Table 2.1). However, the sequence is located about 300 bp downstream of the conserved RLTR10 PBS, which is complementary to a phenylalanine tRNA, and it remains unknown whether this sequence serves as an alternative PBS for reverse transcription. The motif in MERV-L elements was only about 50% homologous to the Pro-PBS. As mentioned above, the peak regions found in RLTR10 and MERV-L elements were clearly less enriched over input than peaks in Pro-PBS containing ERVs (Fig. 3.4), indicating that the target sites in these elements do not efficiently recruit ZFP809.

As shown in Supplementary Table S3.1, no target motifs were found in a large number of peaks overlapping with other ERV elements such as RLTRETN\_MM and IAPEz-int, indicating that the target binding sites responsible for ZFP809 recruitment to these loci are either too degenerated to be recognised, or sequence independent mechanisms contribute to ZFP809 recruitment to these loci. Indeed, in IAPEz-int elements the relative peak summit positions were scattered over several loci and belonged to peaks of low enrichment values (Fig. 3.4). The motifs identified in simple repeats were only remotely similar to the Pro-PBS and the weakly enriched peaks overlapping with these elements indicate that these elements are not efficiently bound by ZFP809 (Fig. 3.6 and Supplementary Table S3.2).

### Distribution pattern of non-repetitive ZFP809 binding sites

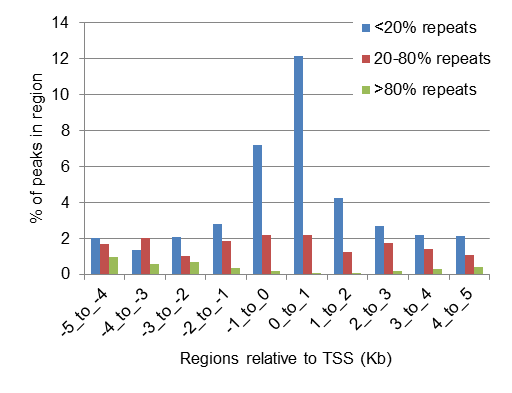
DNA-binding transcription factors are often located close to the transcription start sites (TSS) of genes where they can regulate gene expression by recruiting transcriptional cofactors or directly modifying the chromatin structure. To investigate whether ZFP809 binding sites are associated with TSS of annotated genes, all 6499 peak regions were analysed using a peak annotation tool on the Nebula platform (http://nebula.curie.fr/root). Surprisingly, the FLAG-ZFP809 peaks were indeed preferably located around TSS of Refseq genes (Fig. 3.8A). Since about half of the peak regions were located within the non-repetitive fraction of the genome, whereas the other half was found in ERV elements (Fig. 3.2A), we wanted to evaluate whether peaks associated with repetitive elements or non-repetitive genomic regions were responsible for the observed peak distribution. For this, we divided the peaks into three groups: Non-repetitive peaks (n=2690) with less than 20% repeat content, repetitive peaks (n=2602) with more than 80% repeat content and intermediate peaks (n=1205) with a repeat content between 20 and 80%. Peaks of each category were associated with 1kb long regions up- and downstream of all Refseq gene TSS to determine overlapping regions. The peaks in the non-repetitive fraction of the genome were preferably located around the TSS of genes, whereas repetitive peaks were generally excluded in regions close to TSS (Fig. 3.8B). Intermediate peaks were distributed rather randomly around TSS. Notably, we observed a highly similar distribution pattern for weak, moderate and strong non-repetitive peaks (data not shown).

### Gene ontology of ZFP809 targeted genes



**A**

**B**



**Figure 3.8: ZFP809 ChIP-seq peak distribution around TSS of Refseq genes.** (A) All filtered 6499 FLAG-ZFP809 ChIP-seq peaks were analysed using the peak annotation tools on the Nebula platform. (B) Peaks were categorised according to their repeat content and correlated with 1 kb long regions around Refseq TSS. The histogram shows the proportion of repetitive, intermediate and non-repetitive peaks overlapping with the indicated genomic regions around TSS of Refseq genes.

We next wanted to investigate whether genes targeted by ZFP809 share common molecular and biological functions using the *Genomic Regions Enrichment of Annotations Tool* (*GREAT*) [[141](#_ENREF_141)] which employs gene annotations from several ontologies to associate genomic regions and genes with annotations. *GREAT* therefore facilitates the association of a set of input regions, such as ChIP-seq peaks, with gene annotation terms and, thus, the identification of terms that are preferably associated with these input regions. The pre-set gene association rule in *GREAT* associates each input region to the closest annotated gene whose TSS is located less than 1,000 kb apart from the peak. Under this setting we did not observe any significantly enriched gene annotation terms. Since KRAB-ZFPs are believed to act as transcription factors by binding to a specific target sequence, and we observed that ZFP809 binding sites are preferably located close to the TSS of genes, we set the maximum distance of the peak regions to gene TSS to 5 kb. Under these settings, the 6,499 ZFP809 peak regions hit approximately 5% of all genes annotated in *GREAT*. As shown in Table 3.2, the set of genes whose 10kb regions around the TSS contain ZFP809 peaks was enriched in SMAD binding and sequence specific DNA binding proteins.

**Table 3.2: Gene ontology annotation of Genes with ZFP809 binding sites close to the TSS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| GO Molecular function term | Hyper FDR Q-Val | Hyper Observed Gene Hits | Hyper Total Genes | Hyper Fold Enrichment |
| SMAD binding | 9.92E-03 | 12 | 54 | 4.23 |
| Transcription regulatory region sequence-specific DNA binding | 4.78E-02 | 14 | 86 | 3.10 |
| Transcription regulatory region DNA binding | 3.29E-02 | 30 | 270 | 2.12 |
| Regulatory region DNA binding | 2.85E-02 | 31 | 279 | 2.12 |
| Sequence-specific DNA binding transcription factor activity | 3.59E-05 | 82 | 828 | 1.89 |
| Nucleic acid binding transcription factor activity | 2.67E-05 | 82 | 830 | 1.88 |
| Sequence-specific DNA binding | 1.93E-03 | 63 | 653 | 1.84 |
| DNA binding | 6.66E-03 | 140 | 1873 | 1.42 |
| Protein binding | 3.96E-05 | 367 | 5448 | 1.28 |
| Binding | 2.25E-02 | 647 | 11144 | 1.11 |

The set of 6,499 ChIP-seq peak regions picked 1,062 (5%) of all 20,221 genes. Genes were associated with peak regions located not more than 5 kb up or downstream of the TSS. Genes hit by several peaks were only counted once. Shown are enriched terms in the category “GO Molecular function term”, as calculated by *GREAT*.

SMAD proteins are transcription factors that transduce extracellular signals and regulate gene expression. Apart from molecular function terms, also a large number of other terms were significantly enriched in this set of 1,065 genes. For instance, seven of the 20 genes that are annotated with the term “embryonic digestive tract morphogenesis” were hit by a ZFP809 ChIP-seq peak within 5 kb to the TSS. However, when we filtered out peaks that were less than 50 times enriched over input no terms were significantly enriched in the 60 genes that were hit by these peaks near the TSS, indicating that genes with strong ZFP809 binding sites do not share common characteristics.

### Confirmation of ZFP809 binding to genomic targets by ChIP-qPCR

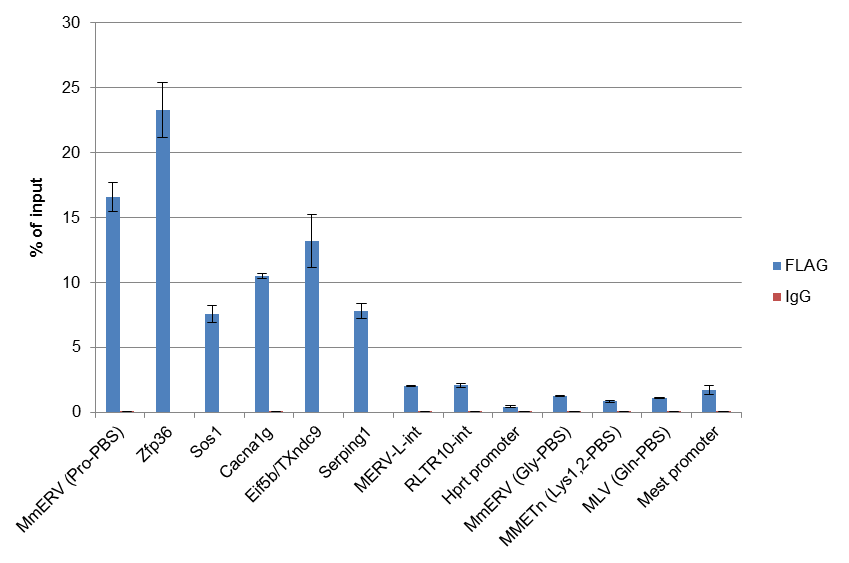
We next wanted to investigate some of the strong ZFP809 binding sites close to TSS in more detail and created a list with all genes that contain a ZFP809 peak which was more than 50 times enriched over input and located not more than 3 kb up- or downstream of the TSS of Refseq genes annotated in *GREAT* (Table 3.3). To confirm ZFP809 binding activity at some of the identified genomic promoter regions, we re-examined anti-FLAG immunoprecipitated DNA samples derived from the ChIP experiment with FLAG-ZFP809 expressing F9 cells described in the previous chapter (Fig. 2.7C).

**Table 3.3: Genes with ZFP809 ChIP-seq peaks in proximity to their TSS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Chr. | Peak start | Peak end | Peak enrichment | Target gene(s) | ZFP809 target motif |
| chr17 | 32756794 | 32756994 | 175 | Gm9705 (-1370) | **TGGGGGCTTGTCTGGGA** |
| chr17 | 80882328 | 80882528 | 138 | Sos1 (-2635) | **CGGGGGCGCGTCTGTAC** |
| chr2 | 74511243 | 74511443 | 121 | Hoxd12 (-1744) | **GGGAGGCTTGTCTGGGA (+1)** |
| chr16 | 32280301 | 32280501 | 113 | Rnf168 (+2854) | **CGGGGGCTCGTCTGGGA** |
| chr2 | 84614954 | 84615154 | 102 | Serping1 (+532) | **TGGGGGCGGGCCCGGGA** |
| chr11 | 94334414 | 94334614 | 102 | Cacna1g (+998) | **GTGGGGCTCGTCTGGGG** |
| chr1 | 38053799 | 38053999 | 97 | Eif5b (-956); Txndc9 (+154) | **GAGGGCCTCGTCTGGGC (+1)** |
| chr1 | 58564262 | 58564462 | 91 | Orc2 (-2599) | **TGGGGGCTTGTCCGGGA** |
| chr7 | 29164267 | 29164467 | 88 | Zfp36 (-120) | **CGGGGGCGCGTCCCGGA** |
| chr8 | 107155338 | 107155538 | 86 | Rrad (-217) | **AGGAGGCTCGCGCGGGA** |
| chr11 | 69579816 | 69580016 | 84 | Zbtb4 (+502) | **TGGGGGTTGGGTTGGGG (+1)** |
| chr19 | 4989088 | 4989288 | 84 | Npas4 (+783) | **-** |
| chr2 | 77118227 | 77118427 | 82 | Sestd1 (+355) | **CGGGGGCTTGCCTGGGA** |
| chr11 | 88936280 | 88936480 | 78 | Gm525 (+1225) | **CATGGGCTCGTCCGGGA** |
| chr6 | 29479180 | 29479380 | 75 | Irf5 (+2547) | **TGGGGGCAGGGCTGGGA** |
| chr4 | 147242301 | 147242501 | 74 | Fv1 (-687); Miip (+427) | **GGGTGGCAGGGCTGGGA (+1)** |
| chr2 | 92241024 | 92241224 | 72 | Mapk8ip1 (+296) | **TGGGGGCTCGGCCGGGA** |
| chr12 | 80010105 | 80010305 | 69 | Plek2 (-2280) | **TGGGGGCTTGTCTGGGA** |
| chr1 | 174136931 | 174137131 | 67 | Pea15a (-119) | **TGGGGGCAGGACTGGGA** |
| chr7 | 136400267 | 136400467 | 65 | Ppapdc1a (-241) | **CAGGGGCGCGTCAGGGA** |
| chr3 | 8923783 | 8923983 | 63 | Mrps28 (-26) | **AGTAGGCTCGCCTGGGA (+2)** |
| chr8 | 91566771 | 91566971 | 62 | Sall1 (+1190) | **AGCATGCTCGCCCGGGA (+1)** |
| chr7 | 148647006 | 148647206 | 62 | Efcab4a (+113) | **TGGGGGCTTGGTTGGGG (+1)** |
| chr18 | 4635009 | 4635209 | 62 | 9430020K01Rik (+182) | **TGGGGGCGCGCGCGGGA (+1)** |
| chr14 | 63456668 | 63456868 | 62 | Wdfy2 (+241) | **AGGGGGCGCGCCCGGGG** |
| chr18 | 21811017 | 21811217 | 61 | Klhl14 (-248) | **AGGGTCCTCGTCTGGGG** |
| chr3 | 67386828 | 67387028 | 60 | Mfsd1 (+238) | **GCGGGGCTCGGCTGGGG** |
| chr5 | 30950698 | 30950898 | 55 | 4930471M23Rik (+486) | **TGGGGGCCTGCCTGGGA** |
| chr5 | 37442140 | 37442340 | 55 | Jakmip1 (+144) | **TGAGTCCTCGGCTGGGA** |
| chr1 | 191552569 | 191552769 | 55 | Ptpn14 (+522) | **TGCTCGCTCGTCCGGGA** |
| chr17 | 28217188 | 28217388 | 55 | Tcp11 (+241) | **TGGGGGCTCGTCAGGGC** |
| chr13 | 109001780 | 109001980 | 55 | Elovl7 (-2718) | **GAGGGGCTTGCCTGGGA** |
| chr19 | 44144313 | 44144513 | 55 | Erlin1 (-237) | **TTGGTCCTCGTCTGGGC** |
| chr10 | 62801973 | 62802173 | 54 | Sirt1 (-290) | **CGCTTGCTCGTCCGGGA** |
| chr19 | 45309622 | 45309822 | 53 | Lbx1 (+4) | **TGGGGGCCCGGCCGGGG (+1)** |
| chr2 | 119613211 | 119613411 | 52 | Rpap1 (-38) | **CTGTGGCTCGCCCGGGA** |
| chr11 | 94937826 | 94938026 | 51 | Itga3 (+102) | **AGGGGGCGCGGCCGGGA** |
| chr12 | 117519525 | 117519725 | 51 | Esyt2 (-70) | **GAGGGGCTTGCTCGGGC** |
| chr3 | 94846279 | 94846479 | 51 | Psmd4 (+88) | **GGGGGGCGCGTCTGGGT (+1)** |
| chr4 | 43667441 | 43667641 | 51 | Spag8 (-1117); Hint2 (+1776) | **TGGGGGCTTGTCTGGGA (+1)** |
| chr9 | 61220678 | 61220878 | 51 | Tle3 (+605) | **GGGGGGCGCGTCTGGGC** |

Shown are the 200 bp core regions of peaks that were at least 50 fold enriched over the input control and are located not more than 3 kb from a Refseq TSS. All genes, whose TSS is located within this range are shown with their relative distance of the peak to the TSS indicated in parentheses. The presence of one or two additional target motifs that have been identified in one peak is indicated in parentheses next to the top-scored target motif.

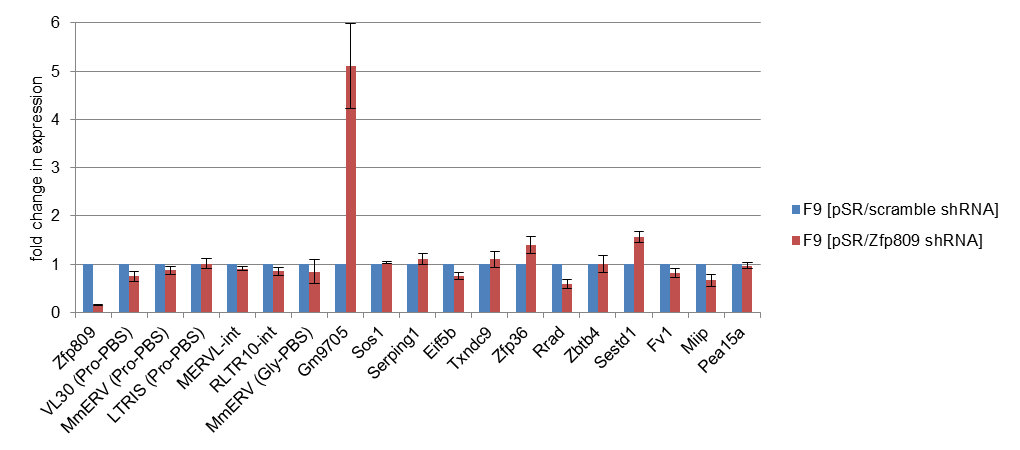
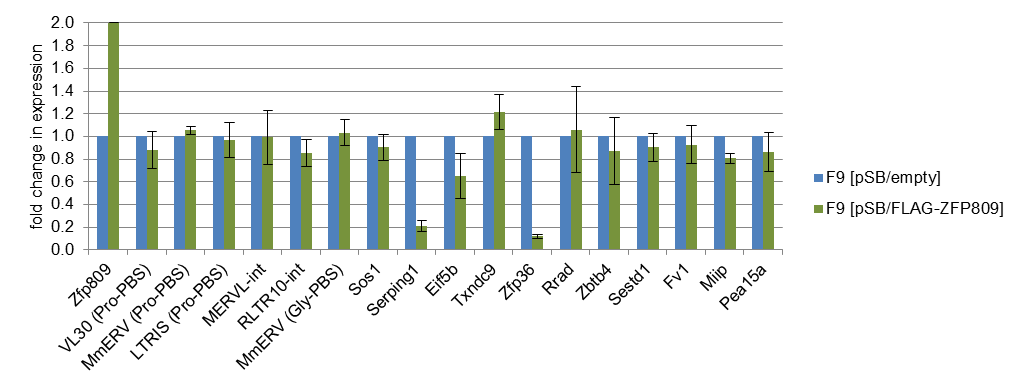
These DNA samples have been isolated from the same pooled F9 [pSB/FLAG-ZFP809] cell line that was used for ChIP-seq but using a different ChIP protocol (see chapter 2). We designed primer pairs annealing up- and downstream of some of the putative ZFP809 target motifs that have been identified by ChIP-seq. As shown in Fig. 3.9, ZFP809 binding to promoter regions that have been identified as strong ZFP809 targets by ChIP-seq could be readily detected by ChIP-qPCR. ZFP809 binding to the identified target motifs in MERV-L and RLTR10 elements was strongly enriched as compared to the *Hprt* promoter region but only slightly enriched over other ERV regions in which no ChIP-seq peaks have been identified. We also observed relative ZFP809 enrichment at the heterochromatin associated *Mest* promoter over the transcriptionally active *Hprt* promoter. We manually analysed the read density around the *Mest* promoter and found that ChIP-seq reads were mapped to this region at an about three times higher frequency than in the *Hprt* promoter whereas reads from the input sample were found at an equal frequency at both regions (data not shown). Altogether these data suggest that ZFP809 binds specifically to genomic targets but may also have a slight affinity to heterochromatin regions, possibly mediated by interactions with heterochromatin associated proteins or modified histones.



**Figure 3.9: ChIP-qPCR analysis to confirm FLAG-ZFP809 binding to putative target sequences**. Chromatin was isolated from FLAG-ZFP809 overexpressing F9 cells and immunoprecipitated against an anti-FLAG antibody or an unspecific IgG antibody, as described in Fig. 2.7C. Primers amplifying a short region that included the identified ZFP809 target motif or control regions were used for qPCR.

### Regulation of gene expression by ZFP809

To determine whether ZFP809 knock-down has an effect on the expression of the identified genes with strong ChIP-seq peaks in the promoter region, RNA was isolated from F9 cells, stably transfected with a retroviral vector expressing an short hairpin (sh) RNA targeting the Zfp809 mRNA or an unspecific scrambled shRNA. Since we have shown that ERVs with a Pro-PBS are the main targets of ZFP809 we used primers for several ERV families with Pro-PBS sequences. Additionally, we analysed expression levels of RLTR10 and MERV-L elements, which were identified as potential secondary ZFP809 targets by ChIP-seq. Primers amplifying endogenous MLV elements with Pro-PBS sequences were also designed but we could not detect expression of these elements in F9 cells. We also could not detect expression of *Cacna1g* and *Npas4* mRNAs in these cells. Surprisingly, neither any of the analysed Pro-PBS containing ERVs nor RLTR10 or MERV-L ERVs was upregulated upon ZFP809 knock-down (Fig. 3.10A). Also the majority of the analysed ZFP809 target genes were not or only slightly up- or downregulated in ZFP809 knock-down cells. However, *Gm9705* expression levels were five times increased upon ZFP809 knock-down. Notably, the identified ZFP809 target motif in the promoter region of *Gm9705* differs from the Pro-PBS in only two positions which, according to the determined target motif, did not seem to be crucial for efficient ZFP809 binding (Table 3.2 and Fig. 3.5).



24.7 (± 3.1)

**A**

**B**

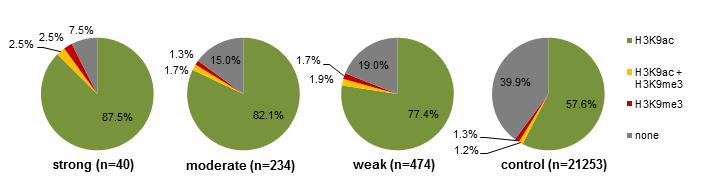
**Figure 3.10: Effect of ZFP809 knock-down and overexpression in F9 EC cells.** (A) F9 cells were stably transduced with retroviral vectors expressing a shRNAs directed against the Zfp809 mRNA or random scrambled shRNA without a matching mRNA target. Expression levels were determined by RT-qPCR. Values are normalised to Act expression and shown relative to the expression in F9 cells transduced with a scramble shRNA expressing vector. (B) F9 cells were stably transfected with a transposon based FLAG-ZFP809 expressing vector or an empty transposon. Values are normalised to Act expression and shown relative to the expression in F9 cells transfected with the empty transposon vector.

To test whether ZFP809 overexpression can silence transcription of ZFP809 targeted genes, RNA was isolated from F9 cells stably transduced with a transposon based FLAG-ZFP809 expressing vector or an empty control vector. As show in Fig. 3.10B, two of the eleven tested genes, *Serping1* and *Zfp36*, were strongly downregulated in FLAG-ZFP809 overexpressing cells whereas ERV expression remained unaffected upon ZFP809 overexpression.

### Correlation of ZFP809 peaks with available histone modification ChIP-seq data

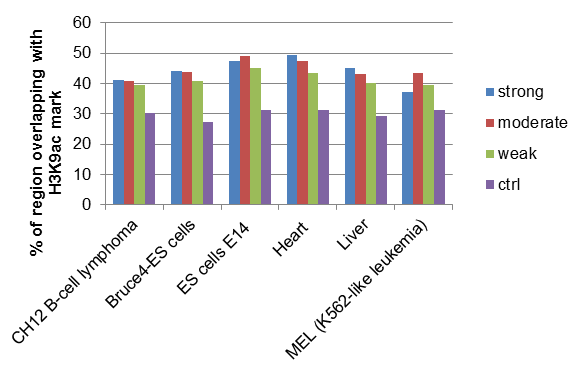
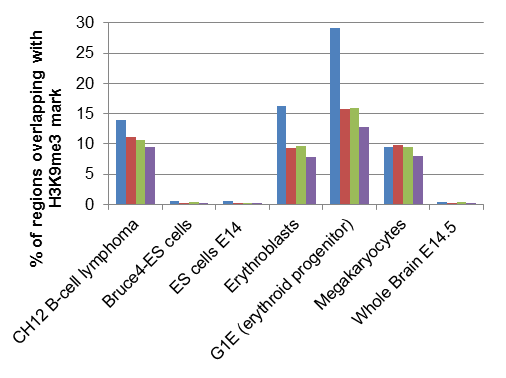
We next wanted to investigate whether the promoter regions targeted by ZFP809 are associated with transcription activating or repressing histone modifications. For this, we retrieved all genomic regions ranging from -2kb to +2kb relative to the TSS of all annotated Refseq genes in the UCSC table browser, hereafter referred to as core promoter regions. Of the 30,656 received core promoter regions 748 regions were overlapping with ZFP809 peaks. We then divided these core promoter regions into three categories according to the enrichment values of the matching ZFP809 peaks (strong: enrichment >50; moderate: enrichment 20-50; weak: enrichment 10-20) and determined how many of these regions were overlapping with non-repetitive H3K9me3 or H3K9ac peaks that had been identified in the Bruce4 ES cell line previously [[142](#_ENREF_142),[143](#_ENREF_143)]. As control regions, we used all core promoter regions that did not overlap with any of the ZFP809 peaks.

As shown in Fig. 3.11, promoter regions of ZFP809 targeted genes were generally not marked by H3K9 trimethylation in ES cells. Strikingly, the only strong ZFP809 peak in a gene promoter region that was associated with H3K9me3 in ES cells was located upstream of the *Gm9705* gene, which was upregulated upon ZFP809 knock-down (Fig. 3.9A). Interestingly, gene promoters with strong ZFP809 peaks were rather associated with H3K9 acetylation in ES cells than promoters with weaker or no peaks.



**Figure 3.11: H3K9 trimethylation and acetylation at promoter regions of ZFP809 targeted genes in ES cells**. Core promoter regions (-2kb to +2 kb of TSS) of Reseq genes were categorised according to overlapping strong, moderate or weak FLAG-ZFP809 ChIP-seq peaks. Promoter regions without ZFP809 peaks were included as control. The peaks were categorised according to their enrichment values as described above. Promoter regions were correlated with the peak coordinates identified for H3K9me3 and H3K9ac ChIP-seq data in Bruce4 ES cells available in the UCSC Genome Browser. Charts show the percentage of promoter regions (number of regions) overlapping with H3K9me3 or H3K9ac peaks by at least one bp.

Since, according to expression profiles on the BioGPS database and our own observations (data not shown), ZFP809 expression is not restricted to ES cells and it is therefore possible that ZFP809 regulates gene expression in other cell types or tissues, we correlated the core promoter region with all H3K9me3 tracks currently available on the UCSC Genome Browser [[144](#_ENREF_144)]*.* As shown in Fig. 3.12A, the 40 genes that were associated with strong ZFP809 peaks at the promoter region were more frequently marked by H3K9me3 in erythroblasts and G1E erythroid progenitor cells than promoter regions with weaker ZFP809 peaks or the included control regions. In several other cell types and tissues, we found that core promoter regions with ZFP809 peaks were rather associated with H3K9 acetylation than promoters without ZFP809 peaks (Fig. 3.12B).



**A**

**B**

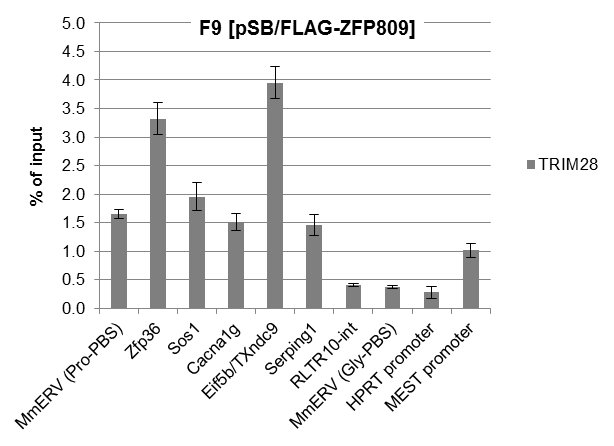
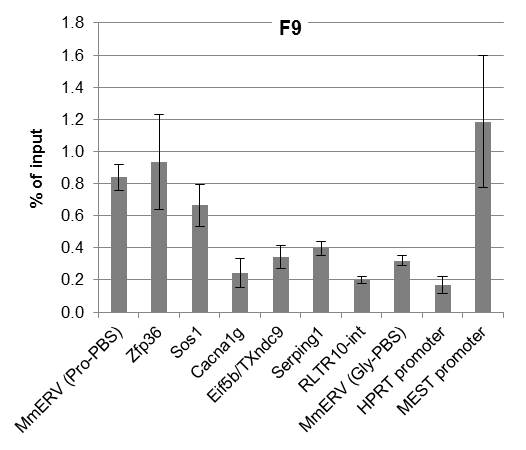
**Figure 3.12: Correlation of ZFP809-targeted promoter regions with H3K9me3 and H3K9ac ChIP-seq data**. 4 kb Refseq promoter regions were categorised as in Fig. 3.11 and correlated with H3K9me3 and H3K9ac peak coordinates that are available on the UCSC Genome Browser*.* Histograms show the proportion of the total bp in promoter regions that overlap with (A) H3K9me3 or (B) H3K9ac peaks.

### Histone methylations and TRIM28 binding at non-repetitive ZFP809 targets

We performed our ChIP-seq analysis with F9 cells that were expressing ZFP809 at more than 20 times higher levels than native F9 cells (Fig. 3.10B). The identified target regions may therefore not represent *bona fide* ZFP809 targets in cells that express ZFP809 at normal levels. This would explain why ZFP809 peaks in core promoter regions did not co-localise with H3K9me3 peaks in ES cells. To investigate whether ZFP809 binding at the identified target sites recruits the TRIM28 corepressor, we re-examined F9 ChIP DNA samples that have been prepared in the previous chapter (Fig. 2.5) using the new primer pairs that have been designed to confirm ZFP809 binding (Fig. 3.9). We also performed ChIP with chromatin that has been isolated from FLAG-ZFP809 overexpressing F9 cells that have been described above (Fig. 2.7C and Fig. 3.10B) to monitor eventual changes in TRIM28 binding upon ZFP809 overexpression. Of the five analysed ZFP809 targets in promoter regions, the regions in the promoter regions of *Zfp36* and *Sos1* were clearly enriched in TRIM28 binding, as compared to the included control regions, indicating that these genes are also bound by ZFP809 in native F9 cells (Fig. 3.13A). Also TRIM28 binding at the ZFP809 targets near the TSS of *Eif5b*/*Txndc9* and *Serping1* was about two fold enriched as compared to the *Hprt* promoter, suggesting that also these genes might be *bona fide* targets of ZFP809. In FLAG-ZFP809 overexpressing F9 cells, strong TRIM28 enrichment was detected at all identified ZFP809 target regions (Fig. 3.13B). We also observed increased TRIM28 binding to the Pro-PBS of MmERV whereas TRIM28 enrichment at the *Mest* promoter, which is not targeted by ZFP809, remained largely unchanged. RLTR10 ERVs were not significantly enriched in TRIM28 in native F9 cells and only a very slight increase was observed in ZFP809 overexpressing F9 cells.

**A**

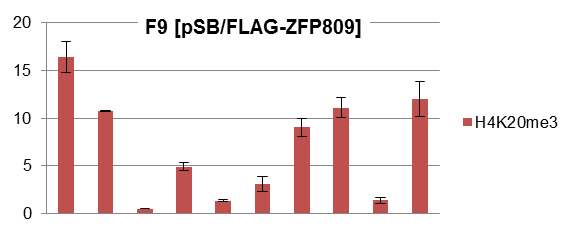
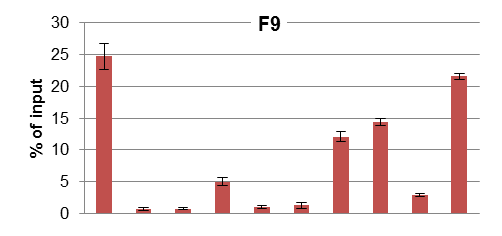
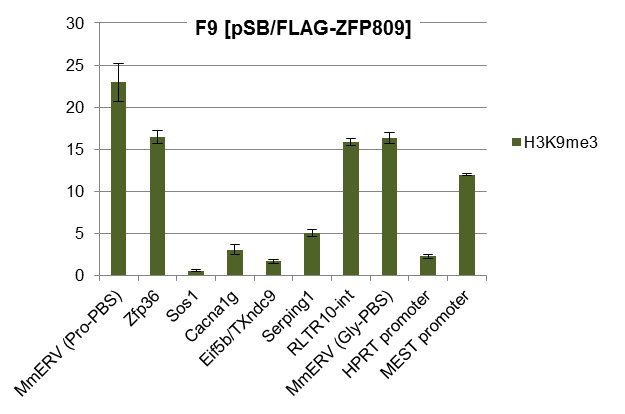
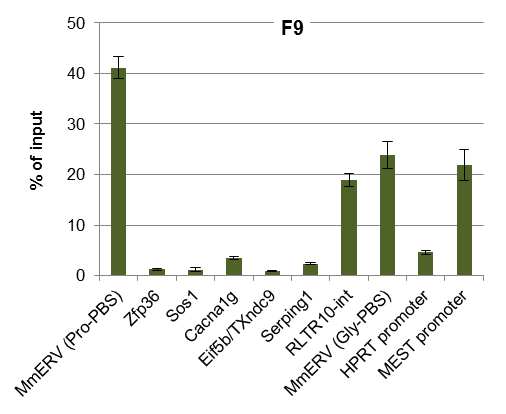
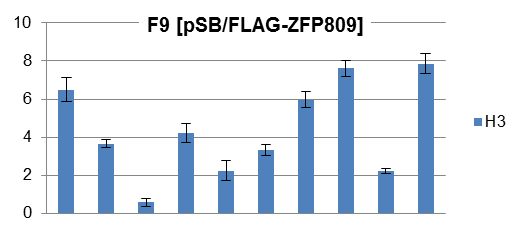
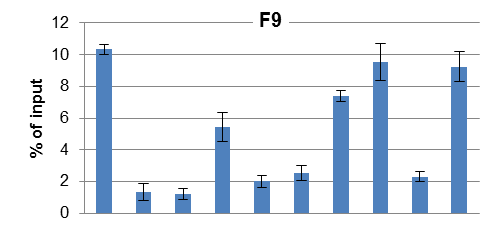
**B**



**Figure 3.13: TRIM28 binding to newly identified ZFP809 binding sites**. Cross-linked chromatin isolated from F9 and FLAG-ZFP809 overexpressing F9 cells was immunoprecipitated with antibodies against TRIM28 or an unspecific IgG control antibody. Enrichment of target sequences was determined by qPCR as described in Fig. 2.4. Enrichment of target regions in IgG immunoprecipitated chromatin was undetectable or less than 0.02% of input and is therefore not shown.

Although we have found that the identified ZFP809 binding sites near TSS are generally not marked by H3K9 trimethylation in ES cells (Fig. 3.11), it is possible that F9 EC cells may show a different H3K9m3 distribution pattern from ES cells. We therefore re-examined ChIP DNA samples that have been isolated from F9 cells in chapter 2 (Fig. 2.5) using primers for the newly identified ZFP809 targets. We also performed ChIP with Chromatin from FLAG-ZFP809 overexpressing F9 cells that have been described above. None of the analysed ZFP809 targets in gene promoter regions was clearly marked by H3K9 and H4K20 trimethylation in F9 cells (Fig. 3.14A). However, ZFP809 overexpression seemed to induce H3K9 and H4K20 trimethylation at the ZFP809 binding site in the *Zfp36* promoter region (Fig. 3.14B). Also the target in the *Serping1* promoter showed moderate levels of the H3K9me3 and H4K20me3 marks in ZFP809 overexpressing cells. Interestingly, the ZFP809 target region in the Sos1 promoter region, although marked by higher levels of TRIM28 than the *Mest* promoter in FLAG-ZFP809 overexpressing cells (Fig. 3.13B), was completely devoid of repressive histone methylations in these cells (Fig. 3.14B).

### Repression activity of ZFP809 target motifs incorporated in retroviral vectors



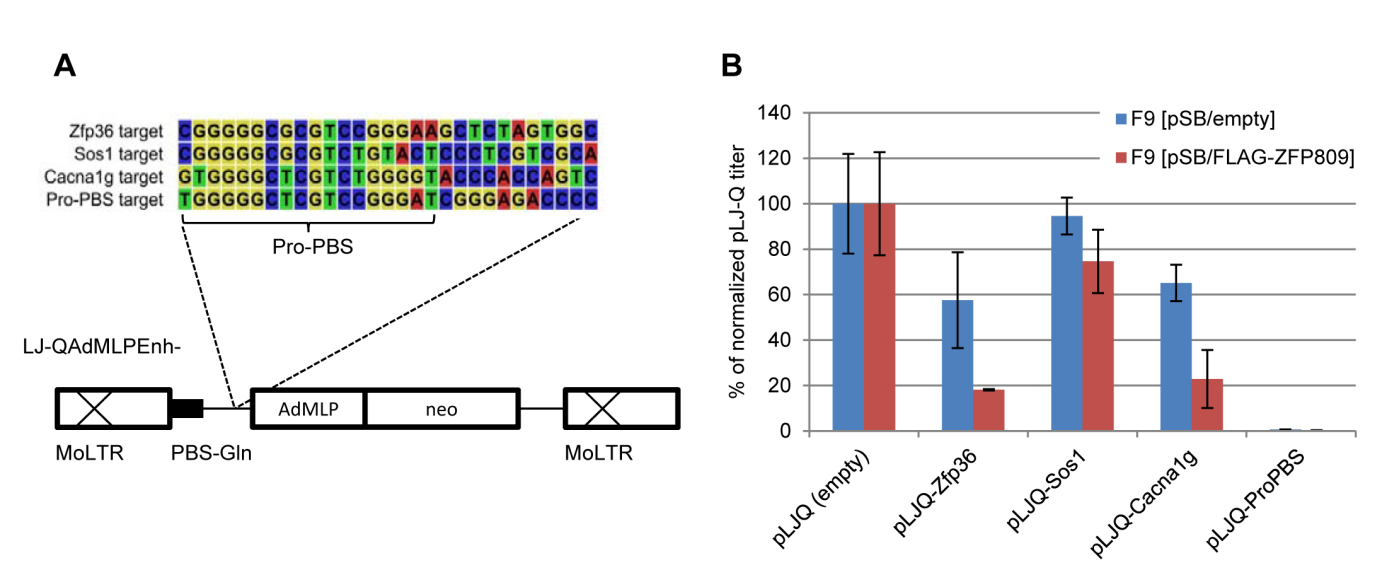
**A**

**B**

**Figure 3.14: Histone modifications at ZFP809 target genes**. Cross-linked chromatin, isolated from (A) F9 cells or (B) FLAG-ZFP809 overexpressing F9 cells, was immunoprecipitated with antibodies against histone 3 (H3), H3K9me3 or H4K20me3. Enrichment of genomic regions in immunoprecipitated DNA was determined as described above.

The observation that even high levels of recruited TRIM28 did not lead to repressive histone methylations in the majority of the analysed target regions led us speculate that certain flanking *cis*-acting elements may prevent heterochromatin formation at these sites. To analyse repression activity of ZFP809 target sequences independent of the surrounding genomic DNA, we generated retroviral vectors that contained the ZFP809 target motifs identified in the promoter regions of *Zfp36*, *Sos1* and *Cacna1g*. The 11 bp regions downstream of the target sites were included in the vector since these regions might be required for repression activity [[84](#_ENREF_84)]. As a positive control for repression, a vector with the 18 bp of the Pro-PBS including the 11 bp downstream region of the Pro-PBS of Moloney MLV was created and included in the assay (Fig. 3.15A). As shown in Fig. 3.15B, the titer of the LJQ-ProPBS vector was strongly reduced in F9 cells compared to the empty control vector. Vectors with genomic ZFP809 target sequences transduced F9 cells with only slightly lower efficiencies than the empty pLJQ vector. Repression of the vectors containing the *Zfp36* and *Cacna1g* ZFP809 target motifs was more pronounced in ZFP809 overexpressing cells but still clearly less than repression of the LJQ-ProPBS vector.

## Discussion



**Figure 3.15: Repression activity of ZFP809 target motifs.** (A) Schematic representation of the LJ-Q AdMLPEnh- vector (modified from ref. [[1](#_ENREF_1)]). The LJQAdMLPEnh- vector contains enhancer deleted Moloney MLV LTRs (MoLTRs), a glutamine PBS (PBS-Gln) and an adenovirus major late promoter (AdMLP) driving expression of a neomycin marker (neo). Four 29 bp oligos, containing the ZFP809 target motifs identified in ChIP-seq peaks close to gene promoters and downstream flanking regions, or the Pro-PBS sequence with the Moloney MLV downstream 11 bp were inserted in sense orientation upstream of the AdMLP. (B) F9 cells, stably transfected with a FLAG-ZFP809 overexpressing vector or an empty control vector and NIH/3T3 fibroblasts were transduced in parallel with serial dilutions of retroviral particles containing the indicated ZFP809 target motifs. Viral titers (CFU/ml) for each vector were determined in all cell lines by counting resistant colonies after 10 to 14 days of selection with G418. Viral titers in F9 cells were normalised to titers in NIH/3T3 cells and are shown relative to the normalised titer of a vector containing the empty LJQAdMLPEnh- vector.

#### Identification of protein binding sites in repetitive elements by ChIP-seq

The ChIP-seq analysis presented in this chapter was not only conducted to identify the genome-wide binding pattern of ZFP809 but was also planned to test whether the binding sites of a transcription factor targeting repetitive elements can be accurately identified by this approach. The genome-wide identification of binding sites of DNA-binding proteins that target repetitive elements is challenging since ChIP-seq reads matching to several genomic positions may result in enriched peaks in regions that are similar to target regions up- and/or downstream of a transcription factor binding site but lack the target site itself. Therefore, the majority of ChIP-seq studies investigating genome-wide binding sites of transcription factors and genomic distribution patterns of histone modifications have ignored ChIP-seq reads that align to multiple genomic loci [[72](#_ENREF_72)]. However, by ignoring reads within repetitive elements, transcription factor binding sites in these regions may remain undetected. To distinguish between *bona fide* binding sites in repetitive elements and peaks that are merely the result of reads randomly mapped to elements that are homologous to the actual targeted region, it is necessary to carefully analyse identified target regions and combine information on the peak sequence with the determined peak enrichment values. Repetitive ChIP-seq reads are expected to be mapped to repetitive regions containing true binding sites at a higher frequency since repetitive elements are often not perfectly conserved. Furthermore, reads overlapping with the target motif itself would be correctly mapped to repeats that are targeted by the transcription factor.

However, it is also possible that peak regions with a degenerated target motif that are less enriched over input than homologous regions with intact target motifs are bound by the analysed transcription factor with a lower efficiency. If a suspected target motif is contained in several unrelated repeat types, it is possible to test whether this target recruits the investigated DNA-binding protein independent of the flanking sequence. It should be also noted that enrichment values of peaks in repetitive elements do not reliably reflect the binding affinity of a DNA-binding protein to this element. If a large number of homologous elements lacking the actual target site are present in the genome, a major fraction of the reads originating from true binding sites will be mapped to these false target regions. The peak enrichment at repetitive elements therefore does not only depend on the affinity of the binding protein to this target but also on the number and sequence similarity of related elements that do not contain the target motif. If possible, binding sites in repetitive elements identified by ChIP-seq should always be confirmed by ChIP qPCR with primers that anneal to the target motif itself and therefore selectively amplify target regions that contain identical target motifs. Considering the limitations mentioned above, we believe that ChIP-seq is a powerful tool to study binding sites of ERV targeting KRAB-ZFPs.

#### Analysis of the identified ZFP809 ChIP-seq target motif

It has been previously shown that most of the mutations that had been introduced to the ZFP809/TRIM28 RBS strongly reduced repression activity [[145](#_ENREF_145)] and repressor complex binding [[94](#_ENREF_94)]. Concordantly, the only nucleotide that was different in the ZFP809 consensus target motif derived from non-repetitive ChIP-seq peaks and the Pro-PBS sequence, T instead of C at position 13 of the PBS, was found at a position that had been changed from C to T without a loss of the repression activity in transduction assays [[145](#_ENREF_145)]. We found that the individual target motifs were not well conserved at two positions (C9 and T11), indicating that these nucleotides may be less important for ZFP809 binding. However, the vast majority of the individual motifs contained either a C or T at these positions whereas A and G nucleotides were rarely found. In transduction assays, these nucleotides were substituted by an A nucleotide which resulted in a complete loss of repression [[145](#_ENREF_145)]. Also in *in vitro* binding assays, the C9, T11 and C13 nucleotides were replaced by G or A nucleotides which resulted in a total loss of repressor complex binding [[94](#_ENREF_94)]. Although this should be confirmed by *in vitro* binding assays and transduction experiments, our results strongly indicate that three positions of the Pro-PBS/RBS tolerate both pyrimidine but not purine bases without losing binding specificity to ZFP809.

#### ZFP809 binding sites in repetitive elements

As expected, virtually all strong ZFP809 ChIP-seq peaks in the repetitive fraction of the genome were located in ERV elements containing a Pro-PBS, indicating that these ERVs are primary endogenous ZFP809 targets. It should be noted that, as discussed above, random mapping of reads that align to multiple loci is expected to lead to an overestimation of the number of ZFP809 binding sites in ERVs. Furthermore, we could not identify a target motif in some of the observed ZFP809 binding sites in ERVS (e.g. IAP elements). Moreover, we detected a general slight enrichment of ZFP809 at repetitive and non-repetitive heterochromatin regions by ChIP-qPCR, suggesting that ZFP809 has some affinity to heterochromatin associated proteins or histone modifications. Importantly, IAP elements are strongly marked by repressive histone methylation [[71](#_ENREF_71)] which might explain the enrichment of ZFP809 at these elements without a detectable target motif.

Apart from the Pro-PBS in various ERV groups, we identified two potential ZFP809 target motif in RLTR10 and MERV-L elements. However, ZFP809 enrichment at these motifs was clearly weaker than at Pro-PBS sequences. Furthermore, we could not detect substantial TRIM28 binding at these regions in native or FLAG-ZFP809 overexpressing F9 cells, indicating that these motifs do not represent functional RBS sequences and RLTR10 and MERV-L ERVs are therefore most likely not repressed by ZFP809.

Nevertheless, our results indicated that ZFP809 has some binding affinity to retroviral sequences that differ in several positions from the Pro-PBS. Mutations in the ZFP809 DNA binding region may increase the affinity to these targets leading to efficient transcriptional silencing of these elements. However, we have shown in the previous chapter that the ZFP809 target specificity is evolutionary conserved, indicating that the need to silence integrated Pro-PBS containing ERVs may have prevented a ZFP809 specificity shift to other targets. Surprisingly, we did not detect an upregulation of ERVs containing a Pro-PBS in F9 cells upon ZFP809 knock-down. It has been shown that KRAB-ZFP/TRIM28 complexes can induce DNA methylation in ES cells [[146](#_ENREF_146)] and may therefore lead to inheritable stable silencing that does not require ZFP809 binding after DNA methylation was established at the target region. ZFP809 knock-down might therefore not be sufficient to release these ERVs from repression. Alternatively, additional repression mechanisms that act independent of a Pro-PBS target sequence might repress these elements in F9 cells. Indeed, we have shown in chapter 2 that also ERVs without a Pro-PBS are marked by repressive histone methylation in EC and ES cells. Another possible explanation for the lack of ERV upregulation upon ZFP809 knock-down might be that the analysed ZFP809 targeted ERVs require certain enhancer binding transcription factors for efficient expression. Those transcription factors might not or not sufficiently be expressed in F9 cells.

#### Zfp809 binding to non-repetitive genomic regions

As we have speculated in chapter 2, specific ZFP809 binding to targets other than ERVs might be required for cellular functions, such as gene regulation. Indeed, we found that ZFP809 target sites in the non-repetitive fraction of the genome are preferably located around the TSS of genes, suggesting that this protein may have a role in the regulation of gene expression. Although ZFP809 recruitment to these targets seemed to be dependent on the presence of target motifs, other factors might bias ZFP809 binding pattern to these regions. First, certain sequence features, such as GC content, may be preferably found at promoter regions and therefore lead to a higher chance that a sequence similar to the GC rich Pro-PBS is present at these regions. We also found that preferably promoter regions that are marked by H3K9 acetylation in ES cells are bound by ZFP809. An open chromatin structure generally favours transcription factor binding. However, we found that ZFP809 efficiently binds to heterochromatin targets and is rather enriched at heterochromatin associated regions such as the *Mest* promoter than at the open *Hprt* promoter. Nevertheless, we cannot exclude that ZFP809 has an increased sequence-independent affinity to both acetylated and methylated histones.

In a previous study, it has been shown that genes expressed in B-enriched splenocytes were significantly closer to TRIM28 peaks in these cells than transcriptionally inactive genes [[147](#_ENREF_147)]. The authors suggested that this may indicate that TRIM28 is rather associated with genes that are expressed and therefore need to be regulated [[147](#_ENREF_147)]. This might possibly also explain our observation that ZFP809 targeted genes are marked by H3K9ac to a higher degree than genes without ZFP809 binding sites in ES cells. However, we were unable to show that ZFP809 binding to these promoter regions generally affects gene expression. Only one of the analysed genes, in which a strong ZFP809 binding site was detected at the promoter region, was strongly upregulated upon ZFP809 knock-down indicating that possibly only a very small subset of the identified ZFP809 targeted genes is repressed by this protein in EC cells. Concordantly, except the *Gm9705* promoter, none of the promoter regions with ZFP809 binding sites have been identified as H3K9me3 targets in ES cells in a previous ChIP-seq study and we confirmed the absence of repressive histone methylations at some of these promoters in F9 cells. Due to time constraints we were not able to perform ChIP-qPCR to confirm ZFP809 and TRIM28 binding at the *Gm9705* promoter. Furthermore, it should be tested whether the ZFP809 target motif in this region induces transcriptional silencing in retroviral transduction assays. The *Gm9705* locus and the possible consequences of ZFP809-dependent repression of *Gm9705* transcription will be discussed in more detail at the end of chapter 4.

#### Functionality of ZFP809 and TRIM28 binding at gene promoters

Although we have identified strong ZFP809 ChIP-seq peaks in gene promoters, these regions were not associated with repressive heterochromatin. It is therefore possible that the ZFP809 binding affinity to these regions is simply too weak to induce epigenetic silencing. It should also be stressed that our ChIP-seq analysis has been performed in F9 cells, in which ZFP809 mRNA was expressed more than 20 times higher than in native F9 cells. Some of the identified targets might therefore merely be the result of a saturation of true ZFP809 binding sites with abundant ZFP809 that led to ZFP809 binding to weaker targets, which would otherwise remain unbound. However, we detected substantial TRIM28 binding at two of the five tested promoter regions in native F9 cells, strongly indicating that at least some of the identified targets are indeed bound by ZFP809 under normal expression levels. Nevertheless, these regions were free of repressive histone methylations indicating that TRIM28 binding does not necessarily result in heterochromatin formation. Also in ZFP809 overexpressing cells, where strong TRIM28 enrichment could be detected at all analysed ZFP809 targeted promoter regions, the majority of these regions remained free of heterochromatin markers. The RBCC domain of TRIM28 was shown to bind as a heterotrimeric complex to the KRAB domains of ZFPs [[148](#_ENREF_148)]. ZFP809 binding might be strong enough to recruit single TRIM28 proteins to these sites but too weak to allow the formation a functional repressor complex.

As shown in a previous study, only about 25% of the identified TRIM28 binding sites near gene promoters were also marked by H3K9me3 in human EC cells [[134](#_ENREF_134)]. Also in murine B-cells, only 30% of the identified TRIM28 binding sites were associated with the H3K9me3 mark [[147](#_ENREF_147)]. Insufficient binding affinity of KRAB-ZFPs to these regions might therefore explain these observations. Alternatively, counteracting transcription factors, such as histone demethylases, may prevent H3K9 trimethylation [[149](#_ENREF_149)] at TRIM28 binding sites. Nevertheless, we have also shown that ZFP809 overexpression can lead to repressive histone methylation and gene silencing at regions that are not silenced when ZFP809 is expressed at normal levels.

Since none of the analysed target sites induced strong transcriptional silencing of retroviral vectors in our transduction assays, it seems unlikely that the absence of heterochromatin at these loci is due to *cis*-acting elements in the promoter regions. However, we cannot exclude that such elements were included in the 11 bp downstream of the target motifs that were included in our target oligos or possibly even overlap with the ZFP809 target motifs themselves. Indeed, the ZFP809 target motif in the *Zfp36* promoter overlaps with binding site for the transcription factor EGR-1 and a putative TPE1 promoter element which both have been identified as cis-acting elements acting to confer full inducible *Zfp36* transcription [[150](#_ENREF_150)]. ZFP36, also known as tristetraprolin (TTP), does not contain a KRAB domain and was shown to bind AU-rich elements within mRNAs, such as the tumour necrosis factor-a (TNF-a) mRNA. ZFP36 binding to these targets leads to a destabilisation and decreased secretion of TNF-a and other proteins [[151](#_ENREF_151)]. ZFP36 mediated mRNA decay was shown to regulate expression of a number of critical genes that are frequently overexpressed in inflammation and cancer [[152](#_ENREF_152)]. One might speculate that ZFP809 binding to transcription factor binding sites in the *Zfp36* promoter may prevent binding of these transcription factors that would otherwise induce strong ZFP36 expression. Since ZFP36 expression is induced by growth factors and mitogens under certain circumstances [[152](#_ENREF_152)], the observed lack of *Zfp36* upregulation in F9 cells upon ZFP809 knock-down does therefore not exclude that ZFP809 affects *Zfp36* transcription under certain circumstances.

The observation that the identified gene promoter regions which were overlapping with strong ZFP809 peaks had a tendency to higher levels of H3K9 trimethylation in erythroblasts and G1E erythroid progenitor cells does not necessarily imply that ZFP809 is responsible for this histone modification in these cells. However, it remains an interesting possibility that should be addressed in future studies.

#### Conclusion

We have confirmed that ZFP809 is efficiently recruited to Pro-PBS sequences of various ERV groups, yet did not detect an upregulation of these elements upon ZFP809 knock/down in F9 cells. Although we have also identified a large number of ZFP809 binding sites in gene promoters, the biological significance of these sites remains largely unknown. However, our results indicate that transcriptional repression of genes by TRIM28 depends on a high affinity of TRIM28 recruiting KRAB-ZFPs to the genomic target site. Considering the little understood role of KRAB-ZFP and TRIM28 in transcriptional control of gene expression it should not be excluded that the identified ZFP809 binding sites within promoter regions that are not marked by heterochromatin in EC and ES cells are biologically relevant in some cell types or tissues. To further investigate the biological functions ZFP809 we initiated the generation of a ZFP809 knock-out mouse, which will be presented in the next chapter.

## Supplementary data for chapter 3

Supplementary Table S3.1: LTR-class repeat enrichment in ZFP809 ChIP-seq peaks.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Repeat** | **Class** | **Family** | **Bp in genomee** | **Bp in peaks** | **Repeat enrichment in peaks** | **No. of overlapping peaks and summits** | **Mean peak height** | **No. of motifs** | **Mean motif score** |
| RLTR6\_Mm | LTR | ERV1 | 4.6E+05 | 6.8E+04 | 145.1 | 279 (117) | 37.8 | 115 | 6.6 |
| RLTR10-int | LTR | ERVK | 2.0E+06 | 2.0E+05 | 100.1 | 539 (534) | 19.4 | 499 | 3.4 |
| MER50B | LTR | ERV1 | 1.1E+03 | 9.1E+01 | 77.2 | 1 (0) | - | - | - |
| MERVL-int | LTR | ERVL | 4.3E+06 | 2.6E+05 | 59.9 | 653 (652) | 20.6 | 618 | 3.4 |
| LTRIS\_Mm | LTR | ERV1 | 2.5E+05 | 1.5E+04 | 58.9 | 68 (16) | 17.2 | 50 | 3.1 |
| RLTR1D | LTR | ERV1 | 2.4E+05 | 8.6E+03 | 34.5 | 29 (13) | 14.9 | 4 | 3.1 |
| RLTR6-int | LTR | ERV1 | 2.3E+06 | 7.5E+04 | 31.5 | 196 (153) | 57.6 | 125 | 6.8 |
| RLTR1 | LTR | ERV1 | 1.2E+05 | 3.4E+03 | 29.0 | 15 (3) | 19.0 | 1 | 3.6 |
| RLTRETN\_Mm | LTR | ERVK | 9.0E+05 | 2.6E+04 | 28.6 | 102 (95) | 18.1 | 1 | 2.8 |
| MuRRS-int | LTR | ERV1 | 1.0E+06 | 2.8E+04 | 26.5 | 76 (61) | 50.9 | 95 | 6.0 |
| LTRIS\_Mus | LTR | ERV1 | 4.5E+05 | 1.0E+04 | 21.6 | 36 (12) | 68.6 | 14 | 3.5 |
| RLTR17 | LTR | ERVK | 1.1E+06 | 2.4E+04 | 20.5 | 59 (55) | 27.4 | 69 | 4.4 |
| RLTR47\_MM | LTR | ERV1 | 2.1E+05 | 4.4E+03 | 20.3 | 19 (3) | 36.7 | 1 | 2.6 |
| LTRIS4 | LTR | ERV1 | 9.0E+04 | 1.8E+03 | 19.2 | 8 (3) | 15.2 | - | - |
| MMVL30-int | LTR | ERV1 | 6.7E+05 | 1.2E+04 | 17.8 | 34 (21) | 56.0 | 16 | 7.0 |
| RMER21A | LTR | ERV1 | 2.2E+05 | 3.9E+03 | 17.0 | 12 (11) | 17.7 | 2 | 4.3 |
| RLTR1B | LTR | ERV1 | 6.2E+05 | 1.0E+04 | 16.5 | 36 (30) | 15.1 | 19 | 2.9 |
| RLTR27 | LTR | ERVK | 2.7E+05 | 4.0E+03 | 14.8 | 48 (4) | 18.8 | 1 | 2.1 |
| MMERGLN-int | LTR | ERV1 | 1.2E+06 | 1.8E+04 | 14.3 | 53 (31) | 48.1 | 28 | 5.8 |
| LTRIS5 | LTR | ERV1 | 6.8E+04 | 9.8E+02 | 14.0 | 4 (1) | 18.5 | 1 | 8.0 |
| RLTR19C | LTR | ERVK | 9.5E+04 | 8.9E+02 | 9.1 | 3 (2) | 26.5 | - | - |
| RLTR5\_Mm | LTR | ERV1 | 2.5E+05 | 2.3E+03 | 9.0 | 10 (3) | 38.8 | 2 | 7.3 |
| MER90 | LTR | ERV1 | 5.6E+04 | 5.1E+02 | 8.9 | 2 (1) | 14.1 | 1 | 4.8 |
| IAPEz-int | LTR | ERVK | 1.2E+07 | 1.1E+05 | 8.7 | 298 (270) | 13.4 | 11 | 2.2 |
| RLTR10A | LTR | ERVK | 3.6E+05 | 3.2E+03 | 8.5 | 23 (5) | 17.9 | - | - |
| IAPLTR3-int | LTR | ERVK | 1.3E+06 | 1.1E+04 | 8.3 | 30 (27) | 20.6 | 2 | 3.5 |
| IAPLTR1a\_Mm | LTR | ERVK | 6.6E+05 | 5.1E+03 | 7.5 | 25 (10) | 13.5 | - | - |
| RLTR10B | LTR | ERVK | 9.6E+04 | 7.4E+02 | 7.5 | 4 (0) | - | - | - |
| MERVL\_2A-int | LTR | ERVL | 2.1E+06 | 1.5E+04 | 7.1 | 38 (38) | 24.0 | 39 | 3.4 |
| RLTR9E | LTR | ERVK | 5.6E+05 | 3.8E+03 | 6.6 | 16 (5) | 15.4 | 6 | 3.6 |
| RLTR1C | LTR | ERV1 | 3.2E+05 | 2.1E+03 | 6.6 | 8 (2) | 14.0 | 2 | 5.7 |
| RLTR10 | LTR | ERVK | 1.1E+06 | 7.1E+03 | 6.2 | 75 (0) | - | - | - |
| IAPLTR2b | LTR | ERVK | 3.4E+05 | 2.0E+03 | 5.8 | 15 (1) | 12.6 | - | - |
| RLTR23 | LTR | ERV1 | 8.0E+05 | 4.8E+03 | 5.8 | 17 (2) | 56.9 | 1 | 4.5 |
| RMER16-int | LTR | ERVK | 2.0E+06 | 1.1E+04 | 5.5 | 33 (30) | 16.4 | 2 | 3.8 |
| IAPEY2\_LTR | LTR | ERVK | 3.4E+05 | 1.9E+03 | 5.4 | 9 (7) | 12.3 | - | - |
| RLTR10D | LTR | ERVK | 2.9E+05 | 1.6E+03 | 5.3 | 10 (2) | 10.4 | - | - |
| MER110 | LTR | ERV1 | 2.9E+04 | 1.6E+02 | 5.2 | 1 (1) | 32.7 | 1 | 3.6 |
| RLTR9B | LTR | ERVK | 7.6E+04 | 4.0E+02 | 5.1 | 1 (1) | 12.9 | 1 | 2.2 |
| RLTR44C | LTR | ERVK | 8.0E+04 | 4.1E+02 | 5.0 | 2 (0) | - | - | - |

The table shows the coverage of Repbase LTR elements in the mouse genome and in determined 6499 ChIP-seq peak regions. The enrichment in ChIP-seq peaks was calculated as the ratio of the peak proportion that is overlapping with the repeat to the percentage of the mouse genome that is annotated with the repeat. The number of ChIP-seq peaks that have at least one bp overlap with elements of a given LTR repeat type are shown, the number of peak summits that are located within repeats is shown in parentheses. The genomic coordinates of all ZFP809 target motifs that have been identified in ChIP-seq peaks by *CisFinder* were correlated with the coordinates of Repbase repeats to identify motifs that overlap by at least one bp with a repeat type. The mean motif score indicates the similarity of repeat-specific motifs to the ZFP809 target motif identified in the 100 top-scored non-repetitive ChIP/seq peaks. Only LTR repeats that were at least five fold enriched in peak regions are shown.

Supplementary Table S3.2: Non-LTR repeat enrichment in ZFP809 ChIP-seq peaks

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Repeat** | **Class** | **Bp in genome** | **Bp in peaks** | **Repeat enrichment in peaks** | **No. of overlapping peaks (and summits)** | **Mean peak height** | **No. of motifs** | **Mean motif score** |
| tRNA-Arg-AGG | tRNA | 7.3E+02 | 2.1E+02 | 277.1 | 7(0) | - | - | - |
| (CCTTG)n | Simple\_repeat | 6.1E+03 | 6.5E+02 | 104.6 | 1(0) | - | - | - |
| (TTAGG)n | Simple\_repeat | 1.4E+03 | 9.0E+01 | 61.3 | 1(0) | - | - | - |
| (CGGGA)n | Simple\_repeat | 1.2E+03 | 6.3E+01 | 53.0 | 1(1) | 25.1 | 2 | 2.8 |
| (TCTTG)n | Simple\_repeat | 7.7E+03 | 3.9E+02 | 48.8 | 2(1) | 13.9 | - | - |
| U3 | snRNA | 9.1E+03 | 4.1E+02 | 43.6 | 2(2) | 15.4 | - | - |
| (TAGGG)n | Simple\_repeat | 4.3E+03 | 1.6E+02 | 35.9 | 1(0) | - | - | - |
| tRNA-Ile-ATT | tRNA | 2.6E+03 | 7.4E+01 | 27.2 | 1(0) | - | - | - |
| (CTGGGG)n | Simple\_repeat | 4.4E+04 | 1.2E+03 | 26.8 | 14(9) | 19.6 | 31 | 3.3 |
| (CCCTG)n | Simple\_repeat | 2.7E+04 | 7.0E+02 | 25.0 | 5(2) | 20.6 | 3 | 3.1 |
| (CAGCC)n | Simple\_repeat | 1.7E+04 | 4.0E+02 | 22.7 | 5(3) | 18.2 | 21 | 3.0 |
| (CCGG)n | Simple\_repeat | 3.7E+02 | 8.0E+00 | 21.0 | 1(0) | - | - | - |
| (CCGGG)n | Simple\_repeat | 3.9E+03 | 8.4E+01 | 20.8 | 2(1) | 29.6 | 2 | 2.1 |
| (CCCCAA)n | Simple\_repeat | 5.7E+03 | 1.2E+02 | 19.8 | 2(1) | 12.0 | 2 | 2.8 |
| (CCCCAG)n | Simple\_repeat | 5.3E+04 | 1.0E+03 | 18.3 | 11(4) | 30.3 | 32 | 3.1 |
| (CAGGG)n | Simple\_repeat | 2.4E+04 | 4.3E+02 | 17.4 | 3(2) | 20.0 | 3 | 4.0 |
| (TTAGGG)n | Simple\_repeat | 6.5E+04 | 1.1E+03 | 17.3 | 5(5) | 14.8 | - | - |
| (CCCCG)n | Simple\_repeat | 3.3E+04 | 4.7E+02 | 13.8 | 8(0) | - | 3 | 3.3 |
| (CTATG)n | Simple\_repeat | 6.6E+03 | 8.8E+01 | 13.0 | 1(0) | - | 1 | 2.5 |
| (TTGGGG)n | Simple\_repeat | 8.2E+03 | 1.0E+02 | 12.2 | 2(0) | - | - | - |
| (CCCCCG)n | Simple\_repeat | 5.3E+03 | 6.6E+01 | 12.1 | 1(0) | - | - | - |
| tRNA-Lys-AAG | tRNA | 1.2E+04 | 1.4E+02 | 11.8 | 2(0) | - | - | - |
| (GGCTG)n | Simple\_repeat | 1.4E+04 | 1.7E+02 | 11.7 | 2(2) | 11.3 | 10 | 4.0 |
| (CCCTAA)n | Simple\_repeat | 6.1E+04 | 7.1E+02 | 11.3 | 6(3) | 17.1 | 2 | 2.4 |
| LSU-rRNA\_Hsa | rRNA | 7.2E+04 | 7.6E+02 | 10.2 | 3(2) | 13.9 | 2 | 2.7 |
| (GGTTG)n | Simple\_repeat | 8.2E+03 | 7.8E+01 | 9.3 | 2(0) | - | - | - |
| (GGGTG)n | Simple\_repeat | 5.1E+04 | 4.8E+02 | 9.2 | 4(1) | 11.6 | 4 | 2.5 |
| (CACAC)n | Simple\_repeat | 2.8E+04 | 2.2E+02 | 7.7 | 2(1) | 86.4 | 5 | 5.0 |
| (CAAG)n | Simple\_repeat | 4.4E+04 | 3.3E+02 | 7.3 | 1(1) | 10.8 | 6 | 3.0 |
| GC\_rich | Low\_complexity | 3.0E+05 | 2.0E+03 | 6.6 | 59(6) | 15.1 | 27 | 3.5 |
| (CTCG)n | Simple\_repeat | 7.6E+03 | 4.7E+01 | 6.0 | 1(0) | - | 1 | 3.8 |
| L5 | LINE | 2.2E+04 | 1.3E+02 | 5.5 | 1(1) | 24.3 | - | - |
| (GGGAA)n | Simple\_repeat | 1.0E+05 | 5.7E+02 | 5.3 | 3(2) | 15.1 | - | - |
| (CGAG)n | Simple\_repeat | 3.9E+03 | 2.1E+01 | 5.3 | 1(0) | - | - | - |

Table 1

Table 2

Non-LTR repeat enrichment in ZFP809 ChIP-seq peaks. See legend of Supplementary Table S2.2 for explanation

Supplementary Table S3.3: Primer sequences

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Primer target |  | Sequence | Application | Amplicon size (bp) |
| RLTR10-int (71 copies) | FW | TGGTGGCCCGTACGGGGA | ChIP / RT-qPCR | 108 |
| RV | CTCCTGGGGTTGCAAGTTTC |
| MERV-L-int (613 copies) | FW | GGAGGTTATGCTTGGGCTCAG | ChIP / RT-qPCR | 116 |
| RV | TATCTGGGGCTCAGTGTTGGT |
| LTRIS (Pro-PBS) (16 copies) | FW | TGGGGGCTCGTCCGGGAT | RT-qPCR | 115 |
| RV | CAAATTTCAGCACAGAACCTAGCC |
| MLV (Pro-PBS) (5 copies) | FW | TGGGGGCTCGTCCGGGAT | RT-qPCR / qPCR | 195 |
| RV | TCCACGGAACCGCCAGAT |
| MMETn (Lys1,2 PBS) (31 copies) | FW | CAACAATTGGCGCCCAACGT | ChIP | 127 |
| RV | CAGCTTAGCCCATACCTTCTC |
| Zfp36 peak | FW | GGAGCTTTCTCCTTTTATGGG | ChIP | 119 |
| RV | TAACCAGCAGATAGGCGGAA |
| Sos1 peak | FW | CAGCCTCACTGCCAGAACG | ChIP | 97 |
| RV | GAACTCCGCTACAATGCCT |
| Cacna1g peak | FW | GGAAGGCTGAGGTGTGTGTA | ChIP | 102 |
| RV | GTGGGGGACTGTTAAGGAGG |
| Eif5b/TXndc9 peak | FW | GGACCCGATTTCGCACTAAC | ChIP | 122 |
| RV | GCGTCCCAGTGTTGATGAAG |
| Serping1 peak | FW | TCCTACGAAGCCAGTTAGCC | ChIP | 132 |
| RV | AGAGGGAGGAGGAGAGAACACA |
| Zfp809 exon 4-5 | FW | AGACACCAAGTGAAGACACG | RT-qPCR | 107 |
| RV | GCGGCACAAAGATTTTACGG |
| Gm9705 exon2-3 | FW | TTGGAGCGGCATCTTGGATT | RT-qPCR | 152 |
| RV | CATGCCTTCCTCATTGTGCT |
| Sos1 exon 9-10 | FW | GTGAAGATCAAGAAGACAAGGAG | RT-qPCR | 162 |
| RV | CTAGCTGTTTCCCCTTCATC |
| Serping1 exon 1-3 | FW | TGATTCTCTGCGACTGTCTG | RT-qPCR | 171 |
| RV | TGGTAGCTTCGGGATCTGAG |
| Eif5b exon 1-3 | FW | CAGAAAAACAAGAGCGAAGACA | RT-qPCR | 183 |
| RV | CAACTCCTCCAGCTCTCTCA |
| Txndc9 exon 1-2 | FW | CTCTCTCCGTAGGCCTCATC | RT-qPCR | 185 |
| RV | CTCCAGTTCATCTTCACCAATC |
| Zfp36 exon 1-2 | FW | TCTCTTCACCAAGGCCATTC | RT-qPCR | 136 |
| RV | GTCCGAGTTTATGTTCCAAAGT |
| Rrad exon 2-3 | FW | TCCTCCGACTCGCTCAGTTC | RT-qPCR | 190 |
| RV | ACGCTTCTTCTCCATCCACA |
| Zbtb4 exon 1-2 | FW | GAGCTGGAGGAGAGCGAAAC | RT-qPCR | 143 |
| RV | TAGCCAACATGGAGAAGCGT |
| Sestd1 exon 2-3 | FW | GGCAAGGACGTGAAATGGAG | RT-qPCR | 112 |
| RV | AGCGGAATTGTCAAGATGAG |
| Fv1 mRNA | FW | CTCTTCTTTCCTACCGCCAC | RT-qPCR | 187 |
| RV | TCCTCCCACACATCTCCATC |
| Miip exon1 | FW | CCCAGCCTTAGCCAATTAGGA | RT-qPCR | 117 |
| RV | CATGGATAAAAGGACGCCAAC |
| Pea15a exon 2-3 | FW | AGTGAGAAGAGTGAGGAGAT | RT-qPCR | 138 |
| RV | CATAGTGAGGAGGTCAGGAC |
| Npas4 exon1-2 | FW | TCTTGCCTGCATCTACACTC | RT-qPCR | 104 |
| RV | CCACAATGTCTTCAAGCTCT |
| Cacna1g exon1-2 | FW | CCCCGGTGGTTTTCTTCTAC | RT-qPCR | 146 |
| RV | CTCACACGGCCTGAACATAC |

# Chapter 4: Generation of a ZFP809 knock-out mouse model

## Introduction and summary

ZFP809 has been shown to restrict expression of *de novo* integrated retroviruses and retroviral vectors in murine ES and EC cells [[94](#_ENREF_94)]. However, the functions of this protein in live mice have not been investigated yet. One might speculate that ZFP809 acts as a restriction factor that prevents expression and amplification of exogenous retroviruses. This may be of special importance in cells of the germ line and ES cells, where retroviral integrations can become ERVs. However, we have shown in chapter 2 that two closely related MmERV subgroups have colonised the mouse genome at similar rates, suggesting that ZFP809 may not primarily function to protect the germ line from ERV amplification. Alternatively, ZFP809 might restrict infecting exogenous retroviruses in somatic cells and tissues. It has been shown that Pro-PBS targeted retroviral repression also takes place in murine hematopoietic stem cells [[85](#_ENREF_85)]. Since retroviruses often infect hematopoietic cells, ZFP809 might restrict retroviral replication in infected mice. However, replicating exogenous retroviruses evolve rapidly and could easily escape repression by utilising new PBS sequences that cannot be targeted by ZFP809. Furthermore, it has been shown that replication-competent MLV variants using different PBS sequences replicate with similar efficiencies in infected mice [[105](#_ENREF_105)]. In chapter 3, we have shown that ZFP809 binds to ERVs and also to non-repetitive regions with a bias to promoter regions around the TSS of genes. We have also shown that expression of one of the genes targeted by ZFP809 in the promoter region, *Gm9705*, is upregulated upon ZFP809 knock-down in F9 cells. To investigate whether expression of ZFP809 targeted ERVs and genes is affected by ZFP809 in other cell types than EC and ES cells, we decided to generate a ZFP809 knock-out mouse which should serve as a rich source for ZFP809 knock-out cell lines and tissues that can be analysed to investigated the *in vivo* functions of ZFP809.

We here show that homozygous ZFP809 knock-out mice are viable and fertile and therefore represent a promising model to study ZFP809 functions. Our preliminary results indicate that out of several tested ERV groups with Pro-PBS sequences, only VL30 elements are strongly upregulated in ZFP809 knock-out MEFs but not ES cells. We also detected a strong VL30-Pro upregulation in hematopoietic cells isolated from ZFP809 knock-out mice. Furthermore, expression of the *Gm9705* gene, which was identified as ZFP809 target in the previous chapter, was strongly induced in ZFP809 knock-out blood cells.

## Materials and methods

#### Generation of ZFP809 knock-out mice

ZFP809 knock-out mice were generated in collaboration with Annette Füchtbauer and Ernst-Martin Füchtbauer. All experimental procedures involved in blastocyst injection and isolation of MEFs and ES cells have been performed by Annette Füchtbauer. The F065A07 ES cell line, derived from the 129S2/SvPas mouse strain, was purchased from the German Gene Trap Consortium (GGTC). This cell line contains a pT1ATGbetageo gene-trap vector in the intron between exon1 and exon2 of the *Zfp809* gene. F065A07 ES cell were injected into blastocysts harvested from B6D2 mice. The resulting chimeric animals were paired with C57BL/6JBom mice to generate the F1 generation of heterozygous *Zfp809*+/- animals. F1 mice were genotyped and two pairs of heterozygous *Zfp809*+/- animals were chosen for inbreeding to generate the F2 generation consisting of wild-type, heterozygous *Zfp809*+/- and homozygous *Zfp809*-/- animals.

#### Genotyping of transgenic animals

According to information available on the GGTC database, the 5’ flanking RNA region of the pT1ATGbetageo gene trap vector has been mapped to exon1 of the *Zfp809* mRNA, indicating that the vector integration site is located between exon1 and exon2, and leads to a splicing event in the transcribed mRNA between the splice donor at the 5’ end of intron1 and the En-2 splice acceptor site of the gene-trap vector. To determine the genomic vector integration site, a series of antisense primers annealing to regions within the first intron of the *Zfp809* gene were designed and used together with a sense primer annealing within the gene trap vector for PCR on genomic DNA isolated from F065A07 ES cells. PCR products were sequenced to determine the precise integration site. New primers annealing up- and downstream of the integration site or at the 3’ end of the integrated vector were designed to allow detection of the *Zfp809* allele containing the gene-trap vector and the wild-type *Zfp809* allele (see Fig. 4.1). The following primers were used for genotyping: P1 (GGACTTCCATTCACAGACTG); P2 (GCATTGGCTGGTTTGCATAG); P3 (GGAATAAGGGCGACACGGAA).

#### Preparation of heterozygous and homozygous MEFs from individual embryos

E8.5 to E14.5 embryos were harvested from F2 heterozygous females that had been paired with F2 homozygous *Zfp809*-/- males and transferred from the yolk sac to a 6-well dish. Brain and red organs were removed from the embryo before mincing the embryo. The carcass was transferred into a 96-well dish containing Trypsin-EDTA and further disaggregated by pipetting up and down. After stopping trypsinisation by adding FBS and centrifugation, cells were seeded on a 12-well dish and grown until passage 2 before analysis. MEFS were cultivated in DMEM supplemented with 10% FBS and penicillin/streptomycin.

#### Isolation of wild-type, heterozygous and homozygous Zfp809-/- ES cells

Wild-type *Zfp809*+/+ and heterozygous *Zfp809*+/- ES cells were derived from blastocysts harvested from F2 *Zfp809*+/- females that had been mated with F2 *Zfp809*+/- males. Homozygous *Zfp809*-/- ES cells were derived from blastocysts that had been isolated from F2 Zfp809-/- females paired with F2 *Zfp809*-/- males. Briefly, plug checked females were sacrificed 3.5 days after mating and blastocysts were harvested and seeded on 96-well dishes coated with inactivated MEF feeder cells. Blastocyst outgrowths were disaggregated by trypsinisation at day 7 after blastocysts harvesting and plated into feeder coated 96-well plates. At day 13, ES colonies were trypsinised and seeded in new feeder coated 96- or 48-well dishes. ES cells were further cultivated until passage 4 to 6 before analysis. For the last two passages before analysis, ES cells were seeded on gelatine coated plates without feeder cells to decrease the proportion of MEF feeder cells in the ES cell culture. ES cells were grown in DMEM supplemented with 15% ES cell tested FBS, glutamine, penicillin/streptomycin, non-essential amino acids, 0.1 mM β-mercaptoethanol, nucleosides and 103 U/ml LIF (Murine Leukemia Inhibitory Factor ESGROTM).

#### Inhibitor treatments

Heterozygous and homozygous ZFP809 knock-out MEFs at passage 3 were treated with 1µM Aza-dC, 0.5 µM TSA or both inhibitors at these concentrations combined, as described in chapter 2.

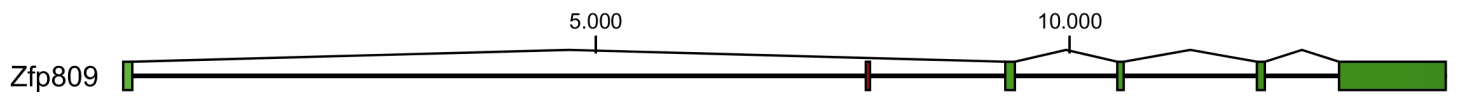
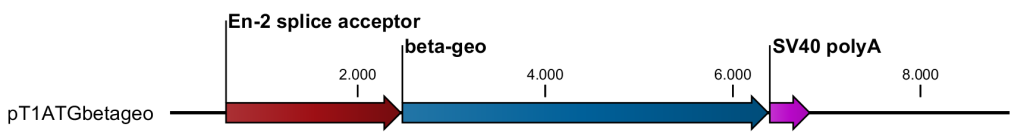
#### RNA purification and RT-qPCR

RNA was isolated from MEFs and ES cells using the TRIzol reagent as described in chapter 2. RNA from whole blood samples, isolated from the eye vein of adult F2 mice, was prepared using the GeneJET Whole Blood RNA Purification Mini Kit (Fermentas). RNA was purified from possible DNA contaminants and reverse-transcribed as described in chapter 2. RT-qPCR was performed as described in chapter 2. Primer sequences are given in Supplementary Table S3.3.

## Results

### Generation of homozygous *Zfp890*-/- mice

Heterozygous *Zfp809*+/- mice with a mixed strain background (129S2/SvPas × C57BL/6J) were generated by blastocysts injection of a *Zfp809*+/- ES cell line as described in the methods section. The *Zfp809* knock-out allele contains a gene-trap vector integration between exon 1 and exon 2, introducing a SV40 polyadenylation (SV40 polyA) and an En-2 splice acceptor site (Fig. 4.1). Two mating pairs of F1 heterozygous animals were set up to generate homozygous offspring. As shown in Table 4.1, homozygous *Zfp809*-/- mice were born at a frequency below expectation. However, the relatively small number of genotyped animals does not allow statistically accurate conclusions about a potentially increased prenatal lethality of *Zfp809*-/- mice. Heterozygous and homozygous ZFP809 knock-out animals showed no obvious phenotype, disease or growth deficit, as determined by routine checks at the animal facility (Annette Füchtbauer, personal communication). Both male and female *Zfp809*-/- mice produced offspring when paired with heterozygous or homozygous animals (Table 4.1), indicating that ZFP809 depletion does not severely affect vitality or fertility of these mice when held in husbandry conditions.



P1

P2

P3

**Figure 4.1: Map of the ZFP809 knock-out allele with the integrated gene trap vector.** Precise integration site was determined by genomic PCRs and sequencing. Zfp809 exons are shown as green conjoined rectangles. The approximate annealing sites of the used genotyping primers P1-3 are indicated as arrows.

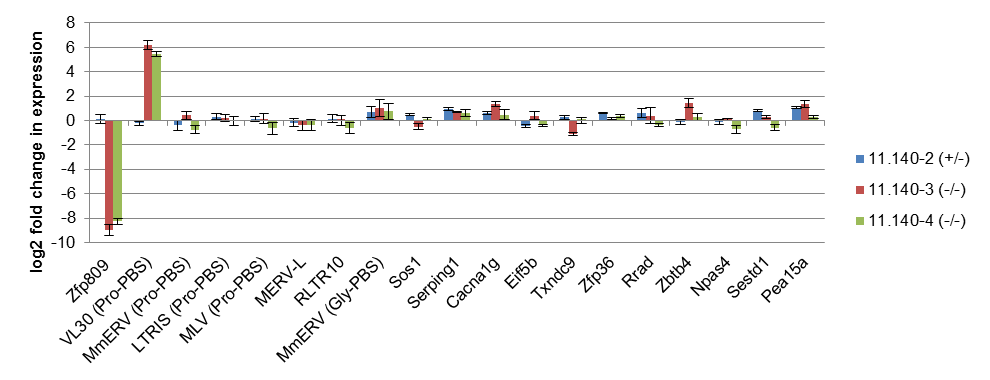
**Table 4.1: Birth frequency statistics of *Zfp809* knock-out mice**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Genotype count | | |
| Breeding pair | Genotype of parents | No. of litters | No. of offspring | +/+ | +/- | -/- |
| 10.838 × 10.830 | M(+/-) × F(+/-) | 4 | 31 | 7 (90%) | 19 (123%) | 5 (65%) |
| 10.836 × 10.841 | M(+/-) × F(+/-) | 3 | 26 | 10 (154%) | 13 (100%) | 3 (46%) |
| 11.134 × 10.956 | M(-/-) × F(+/-) | 2 | 19 | - | 13 (137%) | 6 (63%) |
| 11.528 × 11.380 | M(+/-) × F(-/-) | 1 | 8 | - | 4 (100%) | 4 (100%) |
| 11.312 × 11.298 | M(-/-) × F(-/-) | 1 | 7 | - | - | 7 (100%) |

The total number of offspring with indicated genotypes is shown. The percentage of the expected number of offspring (according to Mendels’s law) with the indicated genotype is shown in parentheses.

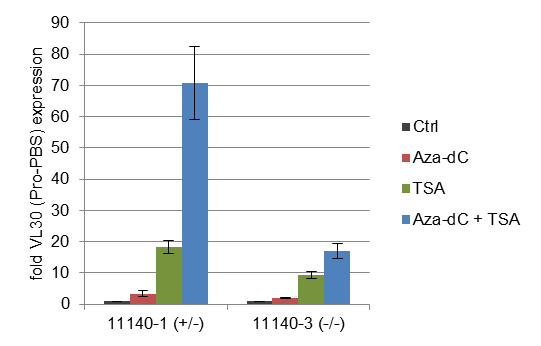
### Endogenous VL30-Pro elements are derepressed in *Zfp809*-/- MEFs

To confirm efficient Zfp809 knock-out and investigate expression of ZFP809 targeted ERVs and genes that have been identified by ChIP-seq (chapter 3), MEFs were prepared from single embryos and genotyped. Since these MEFs were isolated from F2 mice that had a mixed strain background, ERV and gene expression levels may vary from individual to individual. This is a major concern for the ERV expression analysis since many active murine ERVs are polymorphic in different mouse strains. To increase the fidelity of our approach and monitor variations in expression levels between individuals of the same genotype, we chose two heterozygous and two homozygous MEF lines for analysis. RT-qPCR was performed using primers that have been described in chapter 3. As shown in Fig. 4.2, ZFP809 expression in *Zfp809*-/- MEFs was reduced to about 0.2 to 0.3% of the expression levels in *Zfp809*+/- MEFs, confirming successful ZFP809 knock-out by the inserted gene-trap vector.



**Figure 4.2: ERV and ZFP809 target gene expression in ZFP809 knock-out MEFs**. Expression levels were determined by RT-qPCR. Values are normalised to ActB expression and are shown relative to the expression in a heterozygous MEF (11.140-1) prepared from an embryo of the same parents as the analysed MEFs.

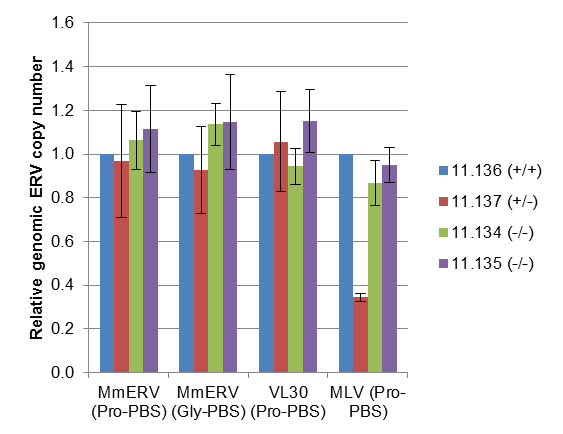
Of all analysed ERVs, only VL30 elements containing a Pro-PBS were transcriptionally upregulated in both *Zfp809*-/- MEF cell lines (Fig. 4.2). No difference in VL30-Pro expression was observed between the two analysed heterozygous MEF cell lines, indicating that the mixed strain background is unlikely to be responsible for the observed VL30-Pro upregulation in *Zfp809*-/- MEFs. None of the analysed ZFP809 target genes was significantly upregulated in both *Zfp809*-/- MEF lines. Furthermore, we were unable to detect expression of the *Gm9705* gene, which was upregulated in F9 cells upon ZFP809 knock/down by shRNAs (chapter 3), in the analysed MEFs. To test whether VL30 expression in *Zfp809*-/- MEFs can be further enhanced by DNA demethylation and inhibition of histone deacetylation, heterozygous and homozygous MEFs were treated with Aza-dC, TSA or both inhibitors combined. As shown in Fig. 4.3, VL30-Pro elements were strongly upregulated in inhibitor treated *Zfp809*+/- MEFs. VL30-Pro upregulation was also observed in *Zfp809*-/- MEFs, yet to a lesser extent.



**Figure 4.3: Effect of TSA and Aza-dC treatment on VL30-Pro expression in MEFs.** Heterozygous *Zfp809*+/- (11.140-1) and homozygous *Zfp809*-/- (11.140-3) MEFs were treated with Aza-dC, TSA or both inhibitors combined. Expression levels were normalised to ActB expression and are shown relative to the expression in untreated MEFs.

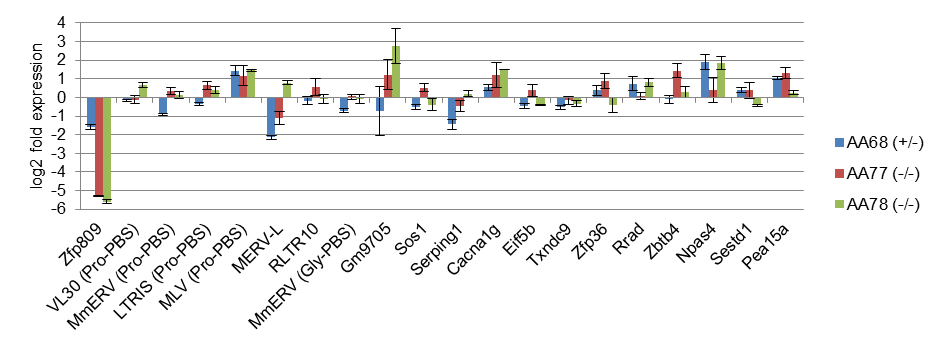
To investigate whether the observed VL30-Pro upregulation also results in genomic amplification of this ERV group, four genomic DNA samples that had been isolated from the tails of ZFP809 knock-out mice for genotyping purpose were analysed by qPCR to determine relative ERV copy numbers. As shown in Fig. 4.4, MmERV-Pro and VL30-Pro copy numbers were constant in wild-type and ZFP809 knock-out mice. The MLV-Pro copy numbers was constant in wild-type and homozygous *Zfp809*-/- mice but significantly lower in the heterozygous animal. To further analyse strain differences in MLV-Pro copy numbers, we screened DNA isolated from the F065A07 parental ES cell line by genomic qPCR. We found that these cells contain the same amount of MLV-Pro copies in the genome as the 11.137 mouse (data not shown). This indicates that MLV-Pro is present at different copy numbers in the two mouse strains that had been used to generate ZFP809 knock-out mice.

### ERV and ZFP809 targeted gene expression in *Zfp809*-/- ES cells



**Figure 4.4: Genomic ERV copy number variation in ZFP809 knock-out mice**. Genomic DNA, isolated from tails was analysed by qPCR. Values were normalised using primers amplifying a genomic region on exon 5 of *Zfp809* and shown as relative change in ERV content relative to the ERV content in the genome of a wild-type mouse (11.136).

Next, ES cells were generated from blastocysts that had been harvested from pregnant F2 mice. In a first attempt, several heterozygous females were paired with heterozygous males to produce wild-type, heterozygous and homozygous ES cells. Although seven individual ES cell lines could be established, genotyping revealed that none of these ES cells was homozygous for the gene-trapped *Zfp809* allele. For the second attempt, homozygous animals were paired and gave rise to two *Zfp809*-/- ES cell lines. These cells were analysed together with one *Zfp809*+/- and one wild-type *Zfp809*+/+ ES cell line that had been generated in the first attempt. As shown in Fig. 4.5, ZFP809 expression in *Zfp809*-/- ES cells was reduced to about 2 to 3 % compared to wild-type ES cells. The relatively high remaining levels of ZFP809 expression in these cells is likely to be due to the presence of contaminating feeder cells. Expression levels of some of the analysed ERVs and ZFP809 targeted genes varied considerably in these cells, however no clear correlation of ERV or gene expression levels with the mouse genotype could be observed. Slight differences in the state of the ES clones, partial differentiation and the mixed strain background may be responsible for the observed differences in expression. It should also be noted that some of the analysed genes were only expressed at levels at the border of detection, which made the assay technically challenging and prone for variations. Also expression of *Gm9705* was nearly undetectable in these cells, however, a trend to higher *Gm9705* expression levels in *Zfp809*-/- ES cells could be observed (Fig. 4.5).



**Figure 4.5: ERV and ZFP809 target gene expression in heterozygous and homozygous *Zfp809*-/- ES cells**. Expression levels were determined by RT-qPCR. Values are normalised to ActB expression and are shown relative to the expression in *Zfp809*+/+ ES cells of the same mixed strain background as the analysed cells.

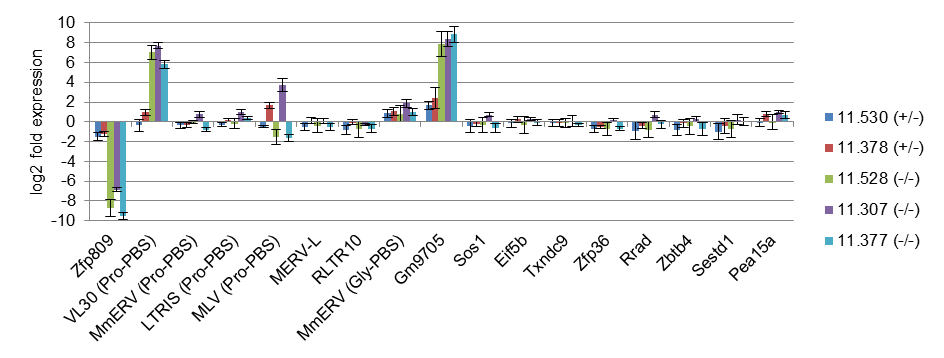
### Derepression of ZFP809 targets in hematopoietic cells of *Zfp809*-/- mice

To investigate ERV and gene expression live mice, we took blood samples from a total of six adult mice and isolated RNA for RT-qPCR analysis (Table 5.2). We found highly increased VL30-Pro expression levels in hematopoietic cells of all three *ZFP809*-/- animals whereas expression of these elements in the heterozygous animal was relatively constant compared to the wild-type control (Fig. 4.6). Expression of endogenous MLV-Pro elements varied strongly between different animals but was not clearly correlated with their genotype.

**Table 4.2: Animals used for blood RT-qPCR analysis**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Animal | Sex | *Zfp809* genotype | Parents (F1 x F1) | Age (weeks) |
| 11.307 | m | -/- | 10.836 x 10.841 | 12 |
| 11.377 | m | -/- | 10.838 x 10.830 | 11 |
| 11.378 | m | +/- |
| 11.528 | f | -/- | 10.836 x 10.841 | 6 |
| 11.529 | f | +/+ |
| 11.530 | f | +/- |

Although we did not analyse the genomic DNA of these animals for MLV-Pro copy number variations, it seems likely that the observed differences are the result of polymorphic MLV-Pro elements. Interestingly, also *Gm9705* expression was strongly induced in blood cells of homozygous mice, confirming the observed upregulation of this gene upon ZFP809 knock-down in F9 cells (chapter 3).

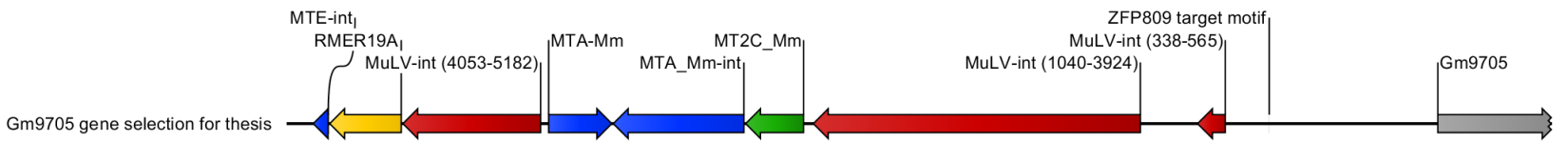


**Figure 4.6: ERV and ZFP809 target gene expression in hematopoietic cells of heterozygous and homozygous *Zfp809*-/- mice**. Expression levels were determined by RT-qPCR. Values are normalised to *ActB* expression and are shown relative to the expression in whole blood of a wild-type mouse (11.529 f (+/+)) of the same mixed strain background.

### Characterisation of the *Gm9705* gene and its ZFP809 target site

The *Gm9705* gene consists of 13 short exons which contain an ORF for a 546 amino acid long predicted protein. The ZFP809 target site in the *Gm9705* promoter region, as identified by ChIP-seq, is located 1,350 bp upstream of the TSS in antisense direction. As noted in chapter 3, the target motif differs in only two nucleotides from the Pro-PBS sequence. Since several ERV elements are annotated upstream of the target motif in the UCSC Genome Browser, we investigated this locus in more detail. The closest annotated internal ERV repeat, MuLV-int, is located 355 bp upstream of the determined Zfp809 target motif which, considering the relative position of the MuLV-int element to the MuLV-int consensus sequence, strongly suggests that the ZFP809 target motif represents a proviral PBS (Fig. 4.7). The provirus upstream of *Gm9705* seems to be strongly degenerated as indicated by the absence of potential MLV LTR elements in this region. Furthermore, the provirus is disrupted by an approximately 2 kb long insertion that consists mainly of MaLR and ERVL elements (Fig. 4.7). According to information available on the UCSC Genome Browser, the MuLV-int elements in this locus are about 34% divergent from the MuLV-int consensus sequence, indicating that this provirus is only remotely related to MLV. Furthermore, a genome-wide BLAT search revealed that no sequences that are more than 75% identical to the provirus upstream of the *Gm9705* gene can be found in the mouse genome. We attempted to reconstruct the ORFs of this ERV by alignment to the corresponding coding region of an intact MLV element. Apart from a large number of insertions and deletion, we counted more than 20 premature stop codons in the 4.6 kb long potential Gag-Pol coding region of the degenerated MuLV-like provirus. Altogether, this indicates that the MuLV-like ERV upstream of the *Gm9705* gene introduced the ZFP809 binding site in the promoter region of this gene but did not recently integrate in the mouse genome.

**Figure 4.7: Genomic locus upstream of the Gm9705 gene.** The depicted region on chromosome 17 is spanning from bp 32748970 to bp 32759191 on the NCBI37/mm9 genome assembly. Repetitive elements as annotated on the UCSC *Genome Browser* are shown as coloured arrows, colours indicate the ERV family: red: ERV1; green: ERVL; yellow: ERVK; blue: MaLR. The relative positions of MuLV-int elements to the MuLV-int consensus sequence, as annotated on the UCSC *Genome Browser*, are shown in parentheses.



1kb 2kb 3kb 4kb 5kb 6kb 7kb 8kb 9kb 10kb

*Gm9705* is a predicted gene that belongs to the cytochrome P450 (*CYP*) gene family and is located within a cluster of *CYP* genes on chromosome 17qB1. Interestingly, this gene has been recently duplicated as indicated by the presence of a homologous gene (*Cyp4f16*), which is located about 85 kb upstream of *Gm9705*. The homology between these genes spans the entire gene including introns and a 745 bp region upstream of the TSS. The MuLV-like provirus and its putative PBS that is targeted by ZFP809 are absent upstream of the *Cyp4f16* gene. Unlike *Gm9705*, *Cyp4f16* has an annotated gene orthologue in the rat genome (*Cyp4f5*). The protein encoded by *Gm9705* is 22 amino acids shorter than CYP4f16 and CYP4f5 than but otherwise 94% identical to CYP4f16. While it seems clear that the *Gm9705*/*Cyp4f16* gene duplication has occurred after the mouse-rat spit (as shown in the Ensembl database), the exact time point of this event is unknown. We therefore attempted to estimate the time of the gene duplication by determining the substitution rates between *Gm9705* and *Cyp4f16* coding regions. We determined dN and dS rates of 0.031 and 0.080, respectively. Assuming that the dS rate between duplicated genes is comparable to the dS rate between gene orthologues in different species we compared this value with the dS values that were determined for the DNA binding region of *Zfp809* in various *Mus* species (Table 2.2). According to this estimation, the *Gm9705*/*Cyp4f16* gene duplication may have occurred at approximately the same time as the *M. musculus* / *M. caroli* split and therefore relatively recently in history.

## Discussion

In chapter 3, we have shown that several ERV groups with Pro-PBS sequences are not upregulated in F9 cells upon ZFP809 knock-down by shRNA. As we have speculated in the previous chapter, the lack of ERV reactivation in F9 cells upon ZFP809 knock-down might be due to inheritable silencing that is initiated by ZFP809 and does not require the presence of the protein once established. This does not seem to apply for these ERVs since we observed that also a complete knock-out of ZFP809 in ES cells was not sufficient to induce expression of any of the tested ERVs. Unlike in *Zfp809-/-* ES cells, VL30-Pro elements were strongly upregulated in *Zfp809-/-* MEFs and hematopoietic cells of *Zfp809-/-* mice. This indicates that either ZFP809-independent silencing mechanism that are active in ES cells but not MEFs or hematopoietic cells repress transcription of these elements, or that transcription factors required for full VL30-Pro expression are not present in ES cells. It remains unknown whether the lack of transcriptional upregulation of the other tested ERV groups is due to degenerated retroviral enhancer elements, missing transcription factors or ZFP809-independent repression mechanisms.

VL30-Pro repression in MEFS might be mediated via ZFP809/TRIM28 binding and subsequent histone methylations or be merely the consequence of inheritable silencing earlier in development. Although H3K9me3 marks have been reported at ERVs in MEFs [[72](#_ENREF_72)], this histone methylation and TRIM28 binding seemed to be absent at ERVs in MEFS in a recent study [[77](#_ENREF_77)]. However, it is possible that a small subset of ERVs containing a strong repressor binding site, such as the Pro-PBS, retains repressive histone methylations in MEFs that are lost on other ERVs upon differentiation. The observation that VL30-Pro elements are more strongly upregulated upon DNA demethylation in *Zfp80+/-* MEFs than in *Zfp809-/-* MEFs suggests that ZFP809 mediated silencing may also induce DNA methylation in these cells. Further analysis of histone modifications and the DNA methylation status of VL30-elements in wild-type and *Zfp809-/-* MEFs will be required to clarify the mechanism behind the observed ERV silencing in differentiated cells.

IAPs and several other ERV families have been shown to be upregulated in TRIM28 knock-out ES cells but no IAP upregulation was observed in TRIM28 knock-out MEFs [[95](#_ENREF_95)]. Although ERV upregulation in TRIM28 knock-out MEFs has not been investigated by large-scale RNA sequencing, the authors concluded that TRIM28 dependent ERV silencing acts primarily in ES cells [[95](#_ENREF_95)]. Our results strongly indicate that TRIM/KRAB-ZFP mediated ERV repression is not restricted to ES cells and may also be a common mechanism in differentiated cells. Interestingly, we also observed a strong VL30-Pro reactivation in blood cells of adult mice. This is not unexpected since it has been shown that Pro-PBS targeted silencing of retroviral vectors takes place in murine hematopoietic stem cells [[85](#_ENREF_85)]. Since we analysed VL30-Pro expression in a mixed population of blood cells, it remains unknown whether the observed VL30-Pro upregulation is limited to a certain cell type, such as hematopoietic stem cells, or affects the majority of hematopoietic cell types. However, these results indicate that VL30-Pro upregulation is not restricted to ES and MEFs and it will be therefore interesting to analyse ERV expression in a larger number of tissues and organs of ZFP809 knock-out mice. VL30-Pro elements only represent a small subgroup of VL30 ERVs, which usually use a Gly-PBS for reverse transcription and are therefore most likely not derepressed upon ZFP809 knock-out. As VL30 ERVs in general, the VL30-Pro subgroup does not code for retroviral proteins and can therefore only amplify if these proteins are provided by *in trans* complementation, or if the VL30 mRNA is co-packaged into viral particles derived from replication-competent retroviruses [[15](#_ENREF_15)]. It is therefore not surprising that we did not observe germ line amplification of these elements in ZFP809 knock-out mice. VL30 elements have been shown to regulate steroidogenesis and possibly other processes by forming a complex with a cellular protein [[22](#_ENREF_22)]. However, we found that the identified VL30 sequences interacting with PSF are not present in the VL30 subgroup that uses a Pro-PBS.

The strong upregulation of *Gm9705* mRNAs in hematopoietic cells of ZFP809 knock-out mice strongly substantiates the observed upregulation of this gene in F9 cells upon ZFP809 knock-down. Recently, it has been shown that a polymorphic IAP LTR that is located 368 bp upstream of the *B3galtl* TSS induces transcriptional silencing of this gene via heterochromatin spreading into the *B3galtl* promoter [[153](#_ENREF_153)]. Although further analysis will be required, our results indicate that a MLV-like proviral insertion induces heterochromatin formation at the *Gm9705* promoter, leading to transcriptional silencing. Importantly, the DNA-binding repressor that recruits the ERV silencing machinery to this provirus, ZFP809, is known. *Gm9705* therefore represents the first example of a gene that is repressed by TE induced heterochromatin spreading induced by a known target-specific repressor protein. The IAP LTR upstream of the *B3galtl* was shown to induce CpG methylation at the CpG island located within the *B3galtl* promoter region in ES cells [[153](#_ENREF_153)]. Although the *Gm9705* promoter region does not contain a CpG island, it will be interesting to analyse DNA methylation at this region in wild-type and ZFP809 knock-out cells and tissues.

Cytochrome P450 (CYP) enzymes are a superfamily of heme-containing monooxygenases that have been found in species across all kingdoms of life. In human and mice, 57 and 102, respectively, putatively functional full length *CYP* genes have been identified [[154](#_ENREF_154)]. CYP enzymes are involved in a large number of metabolic processes such as detoxification of foreign chemicals and pollutants, hormone synthesis and breakdown, and cholesterol synthesis [[154](#_ENREF_154)]. The functions of the proteins encoded by the *Gm9705* gene and its duplicated counterpart *Cyp4f16* have not been investigated so far. However, a comparative analysis of more than 600 *CYP* genes in different species revealed that these genes can be roughly divided into phylogenetically stable and unstable genes [[155](#_ENREF_155)]. Stable *CYP* genes, which are normally present in only one or few copies per species, often encode enzymes that are involved in the metabolism of endogenous substrates such as hormones. In contrast, unstable *CYP* genes frequently undergo gene duplications and losses and are often associated with xenobiotic detoxification [[155](#_ENREF_155)]. *Gm9705*, which is located in a cluster of seven *CYP450* genes has evidently undergone recent gene duplications and can therefore be considered as a rather unstable gene. Although experimental proof is missing, it is tempting to speculate that *Gm9705* encodes for a detoxifying enzyme. According to the BioGPS database, *Gm9705* is mainly expressed in the small and large intestine, which appears to be the right tissue for a protein that is involved in the detoxification of substances that are taken up via ingestion.

At this point, we can only speculate that ZFP809 dependent repression of the *Gm9705* gene is biologically relevant. However, the *Gm9705* ORF remained intact after the *Gm9705/Cyp4f16* gene duplication event, suggesting that purifying selection prevented the loss of the *Gm9705* gene function in mice. It remains unknown whether the MLV-like proviral insertion that introduced a ZFP809 binding site upstream of the *Gm9705* promoter occurred before or after the gene duplication event. If the retroviral insertion predated the gene duplication, the gene duplication might have been a mechanism that released this gene from being repressed by translocating it to a locus more distant from the ZFP809 binding site. Alternatively, the provirus might have been present in the mouse genome at the time of the gene duplication but distant from the *Gm9705/Cyp4f16* progenitor gene. The duplication event may have brought *Gm9705* in proximity to the ZFP809 binding site. One might speculate that these two genes have adapted to different biological functions that require a differential expression activity of these genes. In this case ZFP809 might indeed have a beneficial role in repressing *Gm9705* in certain cell types. Further studies should therefore address the biological function of *Gm9705* and the relevance of its transcriptional repression by ZFP809.

# Chapter 5: Epigenetic marking and de-repression of human endogenous retroviruses

## Introduction and summary

Approximately 10% of the human genome was estimated to be of retroviral origin [[5](#_ENREF_5)]. Unlike in mice and pigs, no replication-competent human endogenous retroviruses (HERVs) have been identified so far. HERVs are usually classified based on the tRNA specificity of their PBS and named by adding the one-letter code for the corresponding amino acid to HERV (e.g HERV-H, HERV-K, HERV-W, etc.). With the remarkable exception of several human-specific HERV-K (HML-2) elements, the majority of all HERVs are believed to entered colonised the human genome between 25 and 70 million years ago [[7](#_ENREF_7),[156](#_ENREF_156)]. Some members of the HERV-K family have integrated into the human genome after the hominid-chimpanzee split approximately 6 million years ago. Two of these human-specific HERV-K proviruses (HERV-K113 and HERV-K115) are insertionally polymorphic and maintained intact ORFs, indicating that the HERV-K group has been active until very recent evolutionary times [[6](#_ENREF_6)]. The integration time of HERV-K113 and HERV-K115 has been estimated to be at least 800,000 and 1.1 million years ago [[157](#_ENREF_157)]. In a recent study, it has been suggested that one HERV-K insertion, HERV-K106, that is fixed in humans integrated approximately 150,000 years ago, after the emergence of anatomically modern humans [[7](#_ENREF_7)]. Furthermore, the putative ancestral progenitor element of the HERV-K family has been artificially reconstructed and it has been shown that functional HERV-K elements can be generated by *in vitro* recombination of human-specific HERV-K proviruses [[158](#_ENREF_158)].

Although no fully replication-competent HERV-K element has been identified so far, it has been shown that HERV-K transcripts, mainly transcribed from one single provirus (HERV-K101), are selectively packaged into retroviral particles in a human EC cell line [[159](#_ENREF_159)]. HERV-K101 has also been identified as one of the most actively transcribed HERV-K elements in tumour samples from patients with testicular cancer but HERV-K101 expression in normal testes was not detected [[160](#_ENREF_160)]. It remains unclear whether HERV-K101 expression in testicular tumours and EC cells is merely a secondary effect or may actively contribute to germ cell tumour formation. In addition to Gag, Pol and Env proteins, type 2 HERV-K elements express a Rec protein which is functionally related to the regulatory Rev/Rex proteins of HIV and HTLV. Rec is translated from alternatively spliced *env* transcripts and has been associated with aberrant germ cell development and carcinoma formation [[161](#_ENREF_161)]. It has been shown that Rec interacts with the promyelocytic leukemia zinc finger (PLZF) tumour suppressor, which negatively regulates transcription of the *c-myc* proto-oncogene and, thus, indirectly alters expression of c-Myc regulated genes [[162](#_ENREF_162)]. A similar mechanism has been shown for Np9, a protein translated from alternatively spliced env transcripts in type 1 HERV-K elements such as HERV-K101 [[162](#_ENREF_162)].

Increased HERV-K, HERV-H and especially HERV-W expression has also been associated with multiple sclerosis (MS) and other neurological diseases [[163](#_ENREF_163),[164](#_ENREF_164),[165](#_ENREF_165)]. Furthermore, a recent study reported a strong association between several single nucleotide polymorphisms (SNPs) in a HERV-Fc1 insertion and MS [[166](#_ENREF_166)]. The HERV-Fc1 provirus is a member of the low-copy HERV-H/F family and was estimated to have integrated into the human genome 10 to 15 million years ago [[166](#_ENREF_166)]. Expression of HERV-Fc1 derived mRNAs and Gag proteins have been found to be significantly increased in peripheral blood mononuclear cells from MS patients with active disease [[167](#_ENREF_167)]. However, the mechanism by which HERV-Fc1 may influence the onset of MS remains unknown.

Recently, an endogenous THE1B solo LTR of the MaLR ERV family was shown to initiate transcription of the non-B, myeloid-specific proto-oncogene *colony-stimulating factor 1 receptor* (*CSF1R*) in B-cell derived Hodgkin´s lymphoma cells [[25](#_ENREF_25)]. Hodgkin´s lymphoma is known to originate from mature B-cells but in malignant Hodgkin/Reed-Sternberg (HRS) cells expression of B-cell specific genes, which are required for B-cell survival, is largely lost [[168](#_ENREF_168)]. LTR driven expression of *CSF1R*, which is normally transcriptionally silent in B-cells, provides an alternative survival pathway for HRS cells [[25](#_ENREF_25)]. The non-canonical *CSF1R* transcript was detected in several HRS cell lines such as KM-H2 but not in myeloid HL60 cells or in several non-Hodgkin’s lymphoma cell lines. It has been suggested that the loss of epigenetic repression of the THE1B LTR is caused by an HRS cell-specific loss of CBFA2T3, a transcriptional repressor that interacts with HDACs and other corepressors [[25](#_ENREF_25)].

Several studies have investigated the effect of DNA methylation on HERV-K expression. HERV-K Gag production was shown to be upregulated in Tera1 EC cells treated with 5-Aza [[169](#_ENREF_169)]. Also in cells which normally do not express HERV-K, DNA demethylation induced HERV-K expression and the activity of *in vitro* methylated HERV-K LTRs was strongly reduced in teratocarcinoma and melanoma cell lines [[170](#_ENREF_170)]. Furthermore, many of the CpG sites in older HERV-K elements have been found to be mutated, indicating that DNA methylation at these sites induced conversion of methylated cytosines to thymines [[171](#_ENREF_171)].

The role of epigenetic silencing by repressive histone modifications in the control of potentially pathogenic HERVs is little understood. Although the genome-wide distribution of a series of histone modifications has been investigated in human CD4+ T-cells [[142](#_ENREF_142)], this study was conducted using an approach that ignores histone modifications in repetitive elements. A follow-up study reinvestigated the distribution of these histone modifications in repetitive elements and found that LTR elements were enriched in H3K9me2 but not in H3K9me3 or H4K20me3 marks [[172](#_ENREF_172)]. However, the same data were later analysed using improved computational methods and it was reported that both H3K9me2 and H3K9me3 marks are enriched at several human ERV1 families, including HERV-K [[72](#_ENREF_72)]. Although human ES and EC cells have been intensively analysed by ChIP-seq to identify genome-wide distribution patterns of a large number of histone modifications, these studied focused on the non-repetitive fraction of the genome and did not investigate histone modifications at ERVs [[173](#_ENREF_173)].

In this chapter, we analyse histone modifications at several disease-associated HERV elements in human EC and ES cells and investigate differences in the chromatin status at the CSF1R THE1B LTR in T-cell lymphomas. Although we found that HERVs are generally marked by H3K9 and H4K20 trimethylation in ES and EC cells, we could not detect significant enrichment of TRIM28 at the investigated HERV loci. Furthermore, our results indicate that the HERV-K101 provirus escapes epigenetic repression mediated by histone methylations in EC cells. Finally, we show that the CSF1R THE1B LTR is marked by H3K9 acetylation in HRS and T-cell lymphomas but retains a rather compact chromatin structure and low levels of repressive histone methylations in these cells.

## Materials and methods

#### Cell culture

Human Ntera-2 EC cells (CRL-1973) were purchased from the American Type Culture Collection (ATCC) and cultivated in DMEM supplemented with 10% FBS and penicillin/streptomycin. The following lymphoma cell lines were kindly provided by Steffen Junker: Reh6 [[174](#_ENREF_174)], HL-60 [[175](#_ENREF_175)], KM-H2 [[176](#_ENREF_176)], MyLa 2054 and 2059 [[177](#_ENREF_177),[178](#_ENREF_178)], SeAx 1621 and 4542 [[179](#_ENREF_179)], Daudi [[180](#_ENREF_180)], DG.75 [[181](#_ENREF_181)], BJA-B [[182](#_ENREF_182)] and Namalwa [[183](#_ENREF_183)]. Cells were received in culture flasks containing the appropriate media and processed for ChIP analysis at the same day.

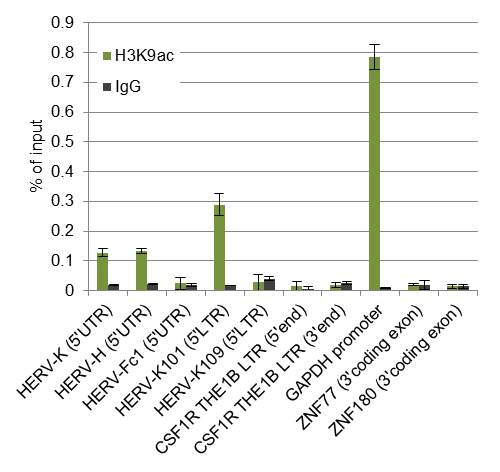
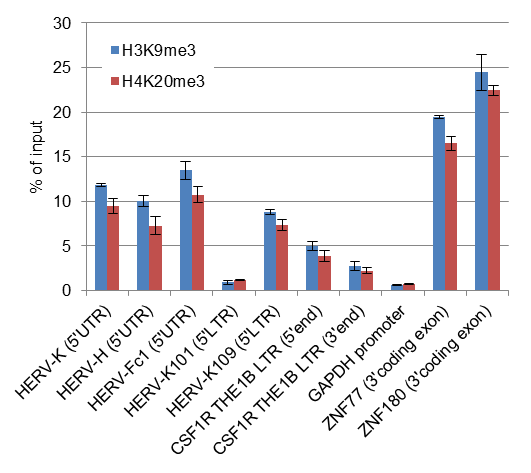
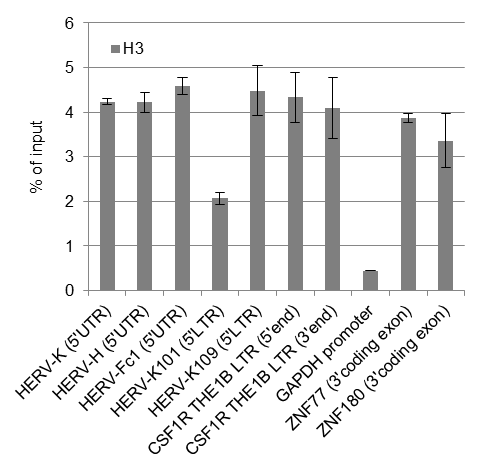
#### ChIP-qPCR

Cross-linked human MEL-1 ES cells (passage 25) [[184](#_ENREF_184)] were kindly provided by Jun Liu and shipped to our laboratory on dry ice. Cross-linked MEL-1 and NTERA-2 cells were sonicated as described in chapter 2. ChIP was performed using the MAGnify Chromatin Immunoprecipitation kit (Invitrogen) as described in chapter 2. Fragment length of reverse-cross-linked sheared chromatin isolated from NTERA-2 cells was analysed by electrophoresis and found to be in the range of 500 to 1500 bp. The size of sonicated MEL-1 chromatin fragments was not analysed. Lymphoma cells were cross-linked and lysed as described previously [[139](#_ENREF_139)]. Chromatin was sonicated for 25 minutes on a Covaris S2 sonicator under the settings described in chapter 2. Fragment length of reverse-cross-linked sheared chromatin was analysed by electrophoresis and found to be in the range of 100 to 300 bp. Sheared chromatin of 1 × 107 cells was used per ChIP according to the protocol established by *O’Geen et al.* [[139](#_ENREF_139)]. Per ChIP, 5 µg of H3, 2 µg of H3K9me3, H4K20me3 and H3K9ac or 1 µg of unspecific IgG antibodies was used. All antibodies were purchased from Abcam as described in chapter 2. Enrichment of specific ERV sequences was determined by qPCR as described in chapter 2. Primer sequences are shown in Supplementary Table S5.1

## Results

### Epigenetic marking of HERVs in human EC and ES cells

Since ERV silencing by histone methylation was proposed to be a mechanism that is predominantly active in pluripotent cells [[71](#_ENREF_71)], we decided to perform ChIP analysis of a human EC cell line (Ntera-2). This cell line was shown to express a large number of pluripotency markers and can be differentiated into neuronal, mesoderm and endoderm lineages [[185](#_ENREF_185)]. Furthermore, HERV-K elements have been shown to be selectively upregulated in human EC cells [[159](#_ENREF_159)]. Additionally, we received cross-linked human MEL-1 ES cells which were included in the analysis. As we have shown in chapter 2, the Pro-PBS of murine ERVs is bound by TRIM28 which is recruited by ZFP809 binding. We speculated that, if KRAB-ZFPs targeting PBS sequences are expressed in human EC and ES cells, these PBS sequences should be also bound by TRIM28 in these cells. Since we only identified three intact Pro-PBS sequences in the human genome by BLAST search, we further speculated that it is unlikely that a human KRAB-ZFP has evolved to target this sequence. It seemed more likely that PBS sequences used by HERVs that have been active until relatively recently in history, such as HERV-K, would be targeted for repression in human cells. Although the HERV-H family is believed to have lost its last replication competent member long before the human-chimpanzee split, we identified a large number of intact histidine PBS sequences in the human genome, which may therefore be another promising target for target-specific repression in human cells. We therefore designed forward primers annealing to the lysine or histidine PBS of HERV-K and HERV-H, respectively, and reverse primers annealing to the 5’UTR of these elements to ensure that only regions containing an intact PBS will be amplified. Since primers annealing within the *CSF1R* THE1B LTR element that has been shown to initiate aberrant *CSF1R* transcription in HRS cells were predicted to amplify also several other THE1B LTRs in the human genome, we took advantage of the non-repetitive flanking regions up- and downstream of this solo THE1B LTR and designed two primer pairs where one of the primers anneals to the unique flanking regions and one in the LTR region (Fig. 5.3A). We also designed primers specifically amplifying the 5’LTR of two individual HERV-K insertions (HERV-K101 and HERV-K109) using the same principle, and primers amplifying a region in the PBS/5’UTR of the HERF-Fc1 provirus. As controls, we designed primers amplifying a sequence in the *GAPDH* promoter region and sequences in the 3’ coding exons of two zinc finger protein encoding genes, *ZNF77* and *ZNF180*, which have been shown to be targeted by TRIM28 in human EC and 293 cells [[134](#_ENREF_134),[135](#_ENREF_135)]. As shown in Fig. 5.1B, HERV-K, HERV-H and HERV-Fc1 ERVs were marked by high levels of the H3K9me3 and H4K20me4 marks in human EC cells whereas histone methylations at the solo THE1B LTR upstream of the *CSF1R* gene were less pronounced. Interestingly, the two analysed individual HERV-K insertions differed greatly in their chromatin status. Unlike HERV-K109, which was marked by repressive chromatin at the 5’LTR, repressive histone methylations were entirely absent at the HERV-K101 5’LTR. Moreover, we observed relatively high levels of H3K9 acetylation at the HERV-K101 but not HERVK-109 5’LTR or in the HERV-Fc1 5’UTR (Fig. 5.1C). Some degree of H3K9 acetylation was also detectable at the 5’UTRs of HERV-K and HERV-H ERVS using primer pairs amplifying a larger number of these elements (see Supplementary Table S5.1), indicating that possibly several other HERV-K and HERV-H insertions are marked by active chromatin markers, as observed for HERV-K 101. However, apart from the two control regions in the *ZNF77* and *ZNF180* genes, we did not detect significant TRIM28 enrichment in any the analysed HERV loci as compared to the *GAPDH* promoter (Fig. 5.1D). Also in human MEL-1 ES cells, we observed high levels of H3K9me3 and H4K20me3 marks at the 5’UTR of HERV-K, HERV-H and HERV-Fc1 ERVs. However, in these cells the 5’LTRs of both HERV-K101 and HERV-K109 seemed relatively devoid of repressive histone methylations. This might be due to possible differences in the size of the sonicated chromatin fragments between. Although both cell types were sonicated under the same settings, it is possible that sonication efficiency varies between cell types. If histone methylations are predominantly located on ERVs, the primers amplifying a region at the border between the retroviral 5’LTR and the upstream non-repetitive flanking region might detect lower levels of these marks if the sonicated chromatin size is smaller.



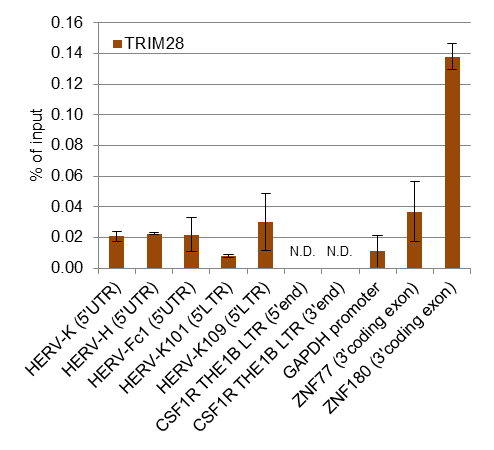
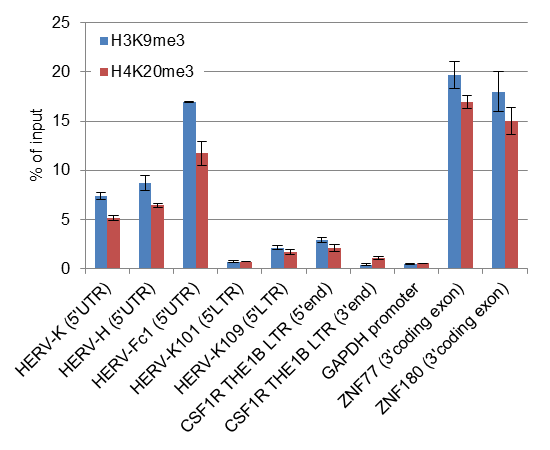
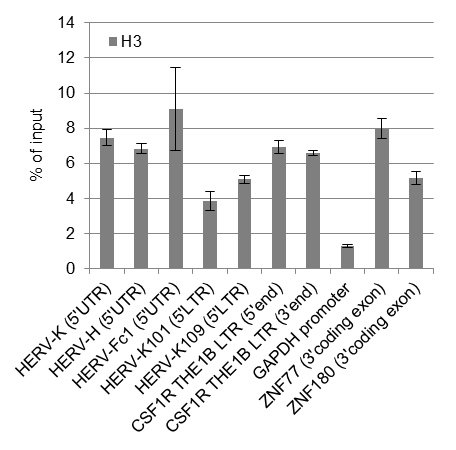
**A**

**B**

**C**

**D**

**Figure 5.1: Histone modifications and TRIM28 binding at HERV loci in human EC cells**. Chromatin was immunoprecipitated with antibodies against modified histones, TRIM28 or an unspecific IgG antibody. Histogram shows mean ChIP enrichment values as percentage of the input control with standard deviations (n = 3, technical replicates).



**B**

**A**

**C**

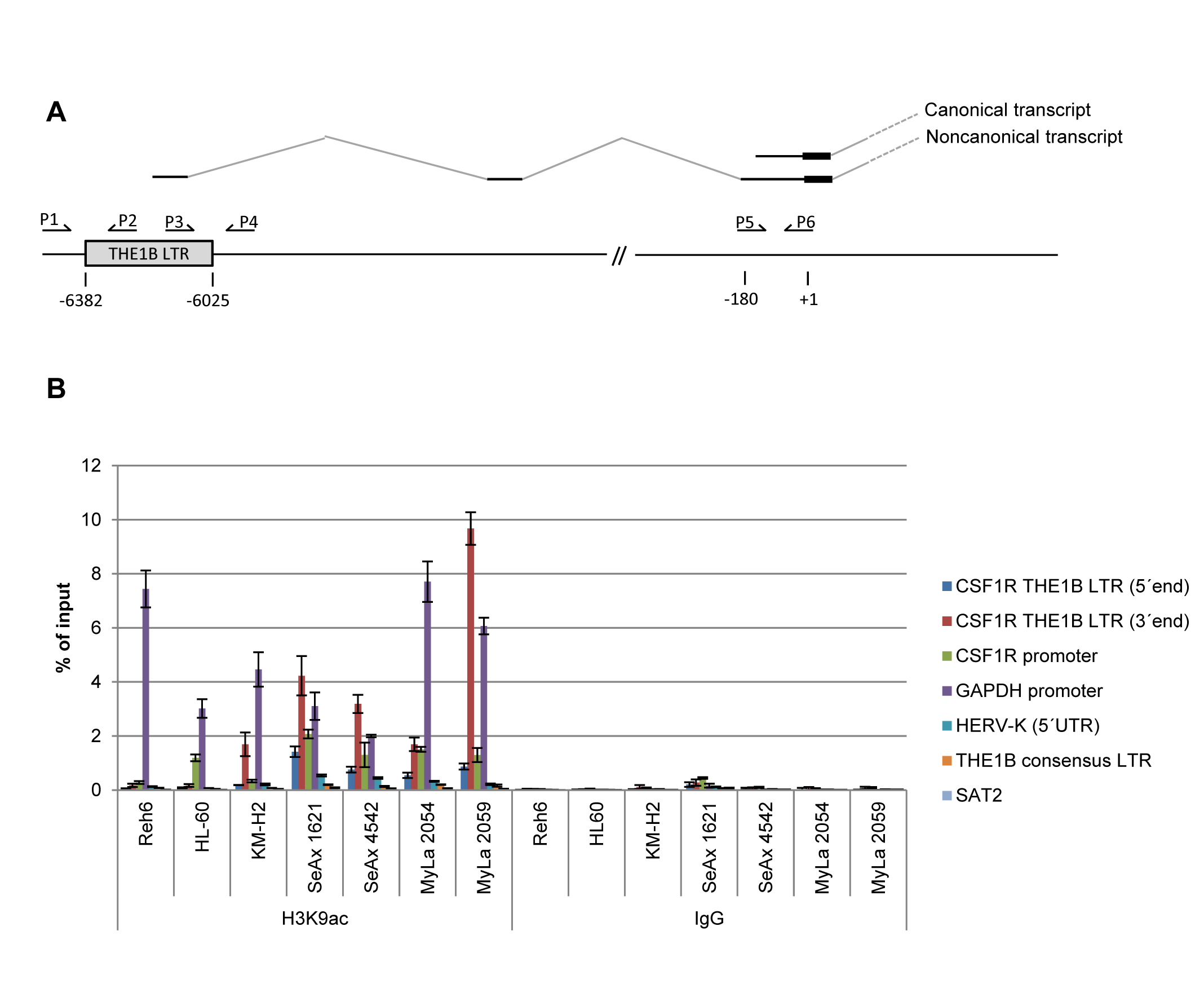
**Figure 5.2: Histone modifications and TRIM28 binding in human MEL-1 ES cells**. ChIP was performed as in Fig. 5.1. N.D.: no PCR product was detected by qPCR.

Nevertheless, we observed more than seven fold higher levels of H3K9 trimethylation at the 5’ end of the THE1B LTR upstream of the *CSF1R* gene than on the 3’ end of the LTR (Fig. 5.2B), whereas the nucleosome density at both ends of the LTR was nearly identical (Fig. 5.2A), indicating that the epigenetic repression of this LTR is asymmetrical. In NTERA-2 cell, the difference in H3K9 trimethylation at these regions was less than two fold (Fig. 5.1B), possibly due to the relatively large chromatin fragments which are less suitable to detect epigenetic differences between regions that are located close to each other. We did not detect significant TRIM28 enrichment at any of the analysed ERV loci but at the *ZNF180* 3’ coding exon in MEL-1 cells. Unfortunately, the ChIP with an H3K9ac specific antibody seemed to have failed in MEL-1 cells since we could not detect significant enrichment of this mark at any of the analysed regions, including the *GAPDH* promoter (data not shown).

### Increased H3K9 acetylation at the *CSF1R* THE1B LTR in T-cell lymphomas

*CSF1R* THE1B LTR derepression has been described in HRS cells [[25](#_ENREF_25)] and was recently also observed in several T-cell lymphomas (S. Junker, unpublished results). To investigate potential cell type-specific changes in the histone modifications at this region, we performed ChIP on several lymphoma cell lines using antibodies against various histone modifications and THE1B LTR specific primers (Fig. 5.3A). As shown in Fig. 5.3B, the THE1B LTR was marked by H3K9 acetylation in the HRS cell line (KM-H2) and all of the analysed T-cell lymphomas but not in Reh6 or HL-60 cells, in which the aberrant THE1B LTR derived transcript has not been detected previously [[25](#_ENREF_25)]. Although the primer pairs amplified products which were only about 130 bp apart, H3K9 acetylation levels were several times higher at the 3´end of the THE1B LTR than at its 5´end in KM-H2 and T-cell lymphomas. In accordance with a previous study [[25](#_ENREF_25)], we observed H3K9 acetylation at the TSS of the canonical *CSF1R* transcript in HL-60 but not KM-H2 cells. Interestingly, H3K9ac was also detectable at the *CSF1R* promoter in all analysed T-cell lymphomas (Fig. 5.3B).

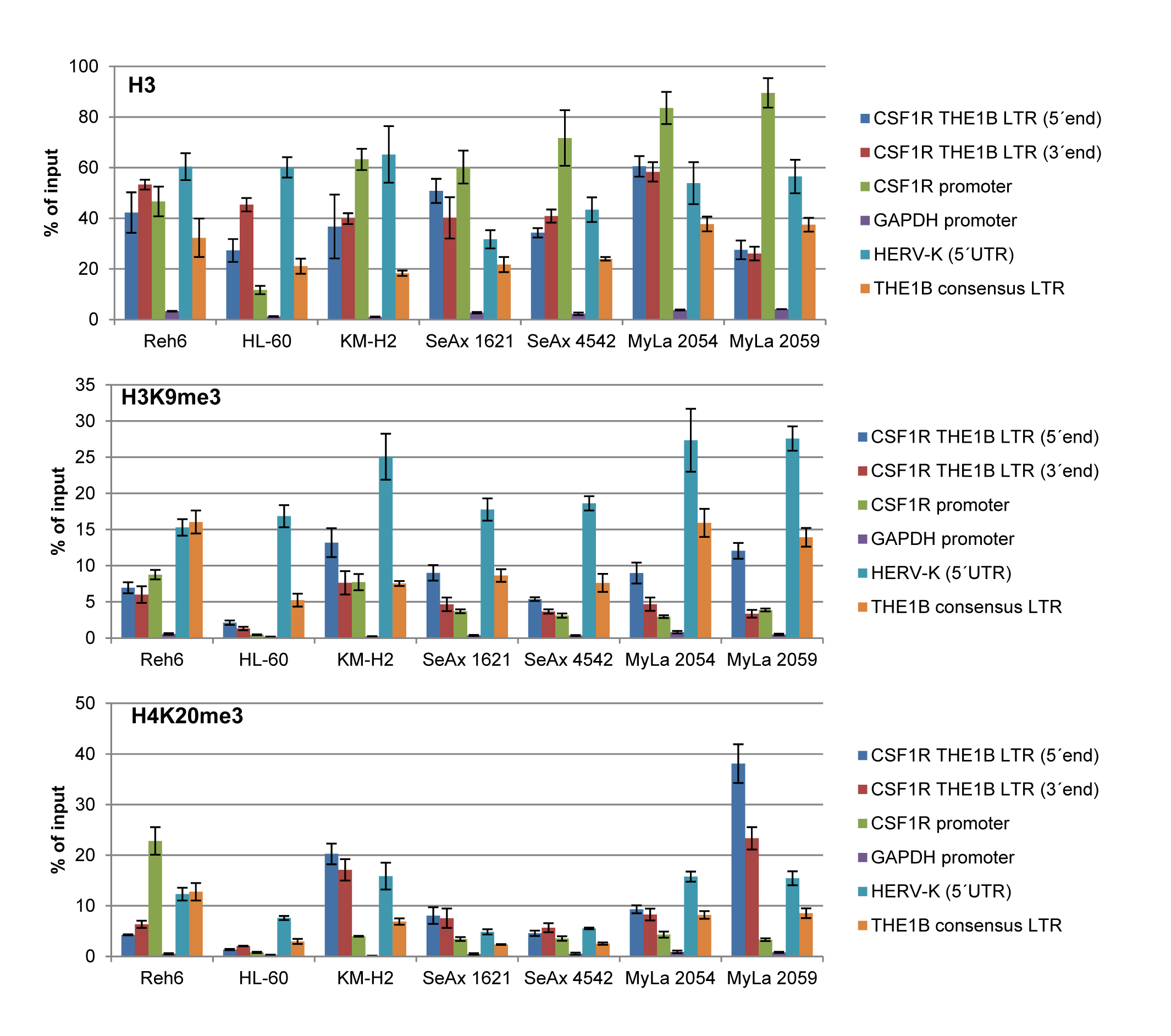
Since THE1 LTR reactivation was reported to be a widespread phenomenon in HRS cells [[25](#_ENREF_25)], we designed primers predicted to amplify the 3’end of approximately 80 THE1B LTRs to test whether histone acetylation is a general feature at these LTRs in HRS and T-cell lymphomas. Furthermore, we included primers specific for HERV-K elements to investigate whether also these ERVs are marked by increased levels of H3K9 acetylation in some of the investigated cells. As shown in Fig. 5.3B, HERV-K sequences were only slightly enriched over the SAT2 control region in the H3K9ac immunoprecipitated DNA of some of the T-cell lymphomas and no H3K9 acetylation could be detected with primers amplifying a large number of THE1B LTRs.



**Figure 5.3: Histone acetylation at the *CSF1R* THE1B LTR in T-cell lymphomas.** (A) Schematic representation of the *CSF1R* locus including the upstream located THE1B LTR, from which aberrant transcription has been reported. The numbers below the sequence representation indicate the relative positions to the translation start site of the canonical *CSF1R* mRNA. The approximate position of the primer pairs is indicated on top of the sequence. The canonical and non-canonical transcripts are shown above the sequence. (B) ChIP analysis of H3K9 acetylation at the *CSF1R* THE1B LTR in lymphoid (Reh6), HRS (KM-H2), myeloid (HL60) and T-cell lymphomas (SeAX and MyLa). Chromatin was immunoprecipitated against an H3K9ac specific antibody or an unspecific control IgG antibody. Enrichment of target sequences was determined by qPCR using a 5% input control for normalisation. Primers amplifying the promoter region of the constitutively expressed *GAPDH* gene were included as controls. The human SAT2 locus is known to be transcriptionally inactive and was included as negative control.

### Repressive histone methylations at the *CSF1R* THE1B LTR in lymphomas

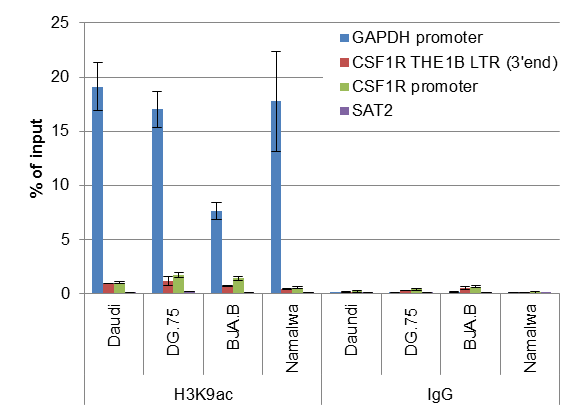
We next wanted to investigate whether the loss of DNA methylation and increased histone acetylation at the *CSF1R* THE1B LTR are accompanied by a loss of repressive histone methylations and performed ChIP with antibodies against H3 or the H3K9me3 and H4K20me3 marks. Both the THE1B LTR and the *CSF1R* promoter seemed to be located in tightly packed chromatin structures in all tested cell types, as assessed by ChIP with H3 antibody (Fig. 5.4). Furthermore, enrichment of the H3K9me3 and H4K20me marks was detectable at the THE1B LTR upstream of the *CSF1R* gene in all of the investigated T-cell lymphomas (Fig. 5.4), indicating that this LTR is not fully derepressed in these cells. Interestingly, repressive histone methylations were nearly absent at the THE1B LTR in HL-60 cells, in which we did not detect H3K9 acetylation at this elements. Unlike in all other tested cell lines, the *CSF1R* promoter was free of H3K9me3 and H4K20me3 marks in HL60-cells. Furthermore, we observed that the 3’ end of the THE1B LTR was less enriched in H3K9me3 than the 5’ end in KM-H2 cells and T-cell lymphomas. Altogether, these data suggest that the transcriptional THE LTR activity is associated with H3K9 acetylation but not necessarily with chromatin relaxation or histone demethylation in HRS cells and T-cell lymphomas.



**Figure 5.4: Repressive histone methylations in T-cell lymphomas.** Chromatin was immunoprecipitated with H3, H3K9me3 or H4K20me3 specific antibodies. Enrichment of target sequences was determined by qPCR using a 5% input control for normalisation.

### Lack of H3K9 acetylation at the *CSF1R* THE1B LTR in Burkitt’s lymphomas

The THE1B LTR initiated *CSF1R* transcript has also been detected in two Burkitt´s lymphoma cell lines, Daudi and Namalwa, that harbour the Epstein-Barr virus (EBV) but not in the DG.75 and BJA.B Burkitt´s lymphoma cell lines in which EBV is absent (S. Junker, personal communication). To confirm transcriptional THE1B LTR activity in these cells, we performed ChIP as described above. However, we could not detect high levels of H3K9 acetylation at the THE1B LTR or *CSF1R* promoter in any of the four Burkitt’s lymphomas (Fig. 5.5).



**Figure 5.5: H3K9 acetylation in Burkitt’s lymphomas.** ChIP was performed as in Fig. 5.3B.

## Discussion

Several recent studies have indicated a possible role of HERVs in disease. Nevertheless, epigenetic ERV repression by histone modifications has been almost exclusively studied in mice and murine cell models. Although the results presented in this chapter are preliminary and require confirmation by functional studies, our data suggest that HERV repression by histone modifications may represent an important aspect of HERV-associated human diseases. We found that one HERV-K insertion, HERV-K101 which was shown to be expressed at high levels in human EC cell lines and testicular cancers [[159](#_ENREF_159),[160](#_ENREF_160)] escapes epigenetic repression by histone methylations in EC cells. In contrast, we found high levels of repressive histone methylations at the HERV-K109 provirus which, concordantly, has been shown to be transcriptionally inactive or expressed at lower levels in these cells and tissues [[159](#_ENREF_159),[160](#_ENREF_160)]. The discrepancy in the transcriptional activity between these HERV-K elements as well as the observed difference in the chromatin status might be due to differences in the LTR regions of these elements. Transcription factors may bind specifically to the HERV-K101 5’LTR and counteract the cellular silencing mechanisms that normally lead to repressive histone methylation at HERV-K elements. To our knowledge, the transcriptional activity of these two HERV-K 5’LTRs has not been investigated in reporter assays so far. However, the LTR of HERV-K108 has been positively tested for transcriptional activity [[170](#_ENREF_170)] and we found that the HERV-K109 LTR has a deletion of 8 consecutive nucleotides as compared to HERV-K108 and HERVK-101. Apart from this deletion the HERV-K101 and HERV-K109 are 98.4% identical. Recently, it has been shown that transcription from HERV-K LTRs mainly depends on the transcription factors Sp1 and Sp3 [[186](#_ENREF_186)] and all of the identified Sp1 and Sp3 binding site are perfectly conserved in the LTRs of HERV-K108, HERV-K101 and HERV-K109 [[186](#_ENREF_186)]. Although it is possible that also other elements may be required for expression, it seems that the epigenetic differences between HERV-K101 and HERV-K109 are not due to transcription factor binding. The HERV-K101 provirus is located in antisense direction approximately 2 kb downstream of the TSS of the *proline dehydrogenase (oxidase) 1* (*PRODH*) gene whereas HERV-K109 is located more than 260 kb downstream of the closest annotated RefSeq gene. It is therefore tempting to speculate that the proximity of HERV-K101 to the *PRODH* promoter is the cause for the observed lack of repressive histone methylations and the presence of H3K9 acetylation mark at this provirus. In chapter 3 and 4, we have presented results indicating that a retroviral insertion close to a gene promoter leads to gene silencing by heterochromatin spreading in murine cells. In the case of HERV-K101 and *PRODH*, the opposite might be true. *Cis*-acting enhancer elements within the *PRODH* promoter may disturb heterochromatin formation of the nearby HERV-K101 LTR and ultimately lead to expression of this provirus. Indeed, a recent study found that gene promoters often prevent epigenetic silencing of nearby LTRs whereas heterochromatin spreading from ERVs into gene promoters is a rare phenomenon [[187](#_ENREF_187)]. Nevertheless, the HERV-K101 provirus may also influence expression of the neighbouring *PRODH* gene. In a recent study, it was shown that about half of the human-specific HERVK LTRs promote transcription of flanking non-repetitive DNA and active LTRs were preferably located in gene-rich regions [[188](#_ENREF_188)]. However, the applied method in this study was not suitable to map transcripts initiated from the 5’LTR of full length proviruses and no clear correlation was observed between the transcriptional activities of genes and their neighbouring genes [[188](#_ENREF_188)]. Future studies may aim to determine whether a potential influence of the HERV-K101 provirus on *PRODH* expression is mediated by antisense transcription initiated from the LTR or by long distance effects of transcription factors recruited to the proviral LTR.

We detected H3K9 acetylation of HERV-K and HERV-H elements using primers that amplify a larger number of these elements and therefore represent an average of all amplified regions. This may indicate that several members of these ERV groups may escape epigenetic repression in a similar way as HERV-K101. Further ChIP studies with primer pairs specific for a larger number of single HERV-K and HERV-H proviruses and possibly the re-examination of existing ChIP-seq data obtained for H3K9me3 and H3K9ac in human ES and EC cells may be useful to generate an epigenetic map of single HERV insertions.

Although a genome-wide binding analysis of TRIM28 binding has been previously performed in Ntera-2 cells, this analysis was based on a ChIP-chip approach using arrays representing the non-repetitive fraction of the human genome [[134](#_ENREF_134)]. TRIM28 binding sites in repetitive elements would therefore remain undetected. The authors of this study later confirmed their results by ChIP-seq, but binding sites in repetitive elements have not been investigated [[135](#_ENREF_135)]. We have shown that the investigated HERV loci are marked by repressive histone methylations but are not targeted by TRIM28 in human EC and ES cells. However, TRIM28 might be recruited to HERVs at other regions than the ones included in our analysis and it remains therefore unknown whether KRAB-ZFP and TRIM28 play a role in HERV silencing. Nevertheless, our results suggest that the lysine and histidine PBS sequences of HERV-K and HERV-H elements, respectively, as well as the 5’UTR of HERV-Fc1 are not TRIM28 targets in human EC and ES cells.

Our ChIP analysis of T-cell lymphomas confirmed preliminary results showing that the non-canonical *CSF1R* transcript is expressed in T-cell lymphomas and indicates that LTR derepression in these cells is accompanied by H3K9 acetylation but not necessarily chromatin relaxation and H3K9 demethylation. It remains unclear whether the observed H3K9 acetylation at the *CSF1R* THE1B LTR is a result of the loss of DNA methylation or vice versa. Notably, repressive histone methylations were entirely absent at the 3’end of the CSF1R THE1B LTR in human ES cells, indicating that this LTR may generally escape repression and therefore be prone to transcriptional activation in certain lymphomas. Interestingly, we detected H3K9 acetylation at the *CSF1R* promoter in T-cell lymphomas and HL-60 cells but not in KM-H2 cells, indicating that THE1B LTR derepression in T-cell lymphomas may also be correlated with transcriptional activation of the *CSF1R* promoter. *CSF1R* upregulation in HRS cells and T-cell lymphomas may therefore be mediated by different mechanisms.

Although preliminary results indicated that the non-canonical *CSF1R* transcript is also expressed in two Burkitt’s lymphoma cell lines in which EBV was detected (S. Junker, personal communication), we did not detect significant levels of H3K9 acetylation at the CSF1R THE1B LTR in any of the four tested Burkitt’s lymphomas. However, concerns about the assays that have been used to detect the non-canonical CSF1R transcript in the Burkitt’s lymphoma cell lines were raised after this study was conducted and these assays are currently being repeated (S. Junker, personal communication).

Research on epigenetic repression of HERVs has so far almost exclusively focused on DNA methylation. Since ERV repression in murine ES cells was shown to be primarily mediated by histone methylations, this mechanism could be also relevant for several HERV associated diseases. The preliminary data presented in this chapter may inspire future studies aiming to evaluate the role KRAB-ZFPs, TRIM28 and repressive histone methylations in the epigenetic repression of HERVs.

## Supplementary data for chapter 5

Supplementary Table S5.1: Primer sequences

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primer target |  | sequence (5´to 3´) | Amplicon size (bp) | Target copy number | Ref. |
| HERV-K (5'UTR) | FW | AACGTGGAGGCTTTTCTCTAGG | 134 | 16 |  |
| RV | TTATCACCCTAGCTTCTTCCGAG |
| HERV-H (5'UTR) | FW | GTGCCGTGACTCGGATCGGG | 115 | 79 |  |
| RV | TGGGCTGGTCGGTCTGAGGA |
| HERV-Fc1 (5'UTR) | FW | GAGCATCTCCTCTCCCTACCTG | 115 | 1 |  |
| RV | GATTACAGACAGAGGTTCCCGG |
| CSF1R THE1B LTR (5'end) | FW | TGATCTCTTACTTTGCCTAACTG | 150 | 1 |  |
| RV | GCCCCCATGATTCAATTATCT |
| HERVK-101 LTR (5’LTR) | FW | TTTTGTGGAGACAGGGTCTGG | 175 | 1 |  |
| RV | GGGGTTGGGGGTAAGGTTATAG |
| HERVK-109 LTR (5’LTR) | FW | TCTCAGGTGTTTGGATCTTCC | 162 | 1 |  |
| RV | TTGGGGGTAAGGTCACAGAAT |
| GAPDH promoter | FW | TACTAGCGGTTTTACGGGCG | 166 | 1 |  |
| RV | TCGAACAGGAGGAGCAGAGA |
| ZNF77 (3'coding exon) | FW | TGGGAAAACCTTCAGGTATCTCG | 106 | 1 |  |
| RV | TTTACGGTTGAACACTCTCCCAG |
| ZNF180 (3'coding exon) | FW | TGATGCACAATAAGTCGAGCA | 137 | 1 |  |
| RV | TGCAGTCAATGTGGGAAGTC |
| CSF1R THE1B LTR (3'end) | FW | GCCTTCCACTATGATTCTGA | 167 | 1 |  |
| RV | ACTAAACCCTGTATATGCCA |
| CSF1R promoter | FW | AGAAGAGGATCAGCCCAAGGA | 91 | 1 | [[25](#_ENREF_25)] |
| RV | AGGGATCGGGACACTGGAC |
| THE1B consensus LTR | FW | TGTGGAACTGTGAGTCCATTA | 92 | 71 |  |
| RV | TGTATTAGTCTGTTCTCACGC |
| SAT2 | FW | CATCGAATGGAAATGAAAGGAGTC | 160 | nd | [[189](#_ENREF_189)] |
| RV | ACCATTGGATGATTGCAGTCAA |

# Concluding remarks and further perspectives

The study of KRAB-ZFP/TRIM28 mediated ERV silencing in species other than mice is highly interesting yet challenging due to the limited availability of suitable non-murine ES cell. In consideration of the enormous number of potential KRAB-ZFP target sites in ERVs, the successful identification of such sequences is difficult, even if a proper ES cell line is available. Knock-down of evolutionarily conserved components of the ERV silencing machinery, such as TRIM28 or SETDB1, by RNAi in non-murine ES cells followed by RNA sequencing appears promising but is demanding since non-murine ES cells are difficult to maintain in culture without differentiation. Furthermore, such studies would fail to detect ERVs that are targeted by KRAB-ZFPs and other target-specific repressor proteins, but have already lost their transcriptional activity or are lacking the necessary transcription factors in the investigated cell types. Moreover, ERV upregulation upon knock-down of TRIM28 or SETDB1 might be due to secondary effects and does not easily lead to the identification of target-specific repressor proteins. Also knock-down of KRAB-ZFPs may not be a promising approach to determine whether these proteins are involved in ERV silencing. As we have shown, RNAi knock-down of ZFP809 in EC cells and even a complete ZFP809 knock-out in ES cells did not reveal that this protein is involved in silencing of VL30-Pro elements. Many ERV groups may be repressed simultaneously by multiple mechanisms or lack essential transcription factors in the investigated cell type.

We have shown that overexpression of a FLAG-tagged KRAB-ZFP in F9 cells allows the identification of its genome-wide binding sites in repetitive and non-repetitive genomic regions. We believe this could be a promising approach to identify novel ERV sequences that are targeted by KRAB-ZFP in mice and possibly other mammals. The approach has several advantages: In contrast to knock-down studies in combination with large-scale RNA sequencing, ChIP-seq does not depend on the transcriptional activity of targeted elements to identify repressor binding. Furthermore, some ERV targeting KRAB-ZFPs may not be expressed in a given cell type. Overexpression of FLAG-tagged KRAB-ZFPs can overcome this obstacle and also circumvent the production of KRAB-ZFP specific antibodies, which is costly, time consuming and technically challenging due to homologies that are shared by many KRAB-ZFPs. Furthermore, it is reasonable to speculate that the binding specificity of most KRAB-ZFPs is independent of the cell type. Therefore, our approach might also be successful in differentiated cell types that are available for a species of interest. However, although the costs for a single ChIP-seq, as it has been performed in this study, is relatively low and will probably further decrease in the near future, the ChIP-seq screen of all KRAB-ZFPs that are present in a given mammalian species would be too costly and demanding to be conducted by a single research group. Nevertheless, although possibly only very few of the several hundreds of KRAB-ZFPs in pigs, mice and humans may target ERVs, the results of such screens would be also useful in the general study of KRAB-ZFPs and their genomic target sites.

Host factors restricting microbial infection are known to undergo rapid evolution to match the high mutation rate of microbial genomes. In this study, we addressed if this principle applies to the repression of ERVs through sequence-specific recognition of integrated retroviral DNA. Surprisingly, our results indicate that the recognition specificity of ZFP809 is conserved in mice, hamsters and blind mole rats, which have separated about 25 million years ago and have since been colonised by distinct ERV families of which the majority does not contain the repressor target sequence. Moreover, two closely related subgroups of a relatively young family of ERVs were equally efficient in colonizing the mouse genome even though only one of them was targeted by the sequence-specific repressor. Our evolutionary analysis of ZFP809 and its genomic ERV targets therefore does not support a model of an on-going arms race in which the DNA-target specificity of a transcriptional repressor protein adapts to ERVs that, in turn, evolve to escape repression. It remains unknown whether also other ERV targeting KRAB-ZFPs are similarly conserved or whether ZFP809 represents an exception of an otherwise evolutionary dynamic and adaptive set of ERV repressing KRAB-ZFPs. Assuming that the observed transcriptional repression of the *Gm9705* gene is beneficial for mice, one might speculate that this, and possibly other ZFP809 functions that are not related to ERV silencing, prevented a specificity shift or degeneration of ZFP809, even during periods in which no selective pressure was exerted by potentially hazardous exogenous and/or endogenous retroviruses with Pro-PBS sequences. Expression of host protein encoding chimeric transcripts initiated from ERV LTRs and other TEs has been described before and it has also been speculated that heterochromatin spreading from TEs into cellular promoters may suppress gene expression. Although awaiting further studies, our results suggest that a KRAB-ZFP that had evolved to repress ERVs has been later utilised as transcriptional repressor by the introduction of a retrovirus derived binding site close to a cellular promoter. Since ZFP809 might have controlled gene expression before it adapted its binding specificity to a retroviral sequence, this may represent a functional cycle that could help to understand the evolutionary dynamic of KRAB-ZFPs.

However, further studies will be required to substantiate these speculations. Most importantly, expression levels of both the ZFP809 targeted *Gm9705* as well as its duplicated counterpart *Cyp4f16* should be analysed in organs that are commonly associated with CYP enzymes. It also needs to be determined whether *Gm9705* mRNA is indeed translated into a protein product in these organs and which reactions are catalysed by this protein. It will be also highly interesting to compare the substrate specificities and catalytic activities of the *Gm9705* encoded protein and its homologue CYP4f16 to determine whether these two proteins may have adapted to different biological functions.

Our ZFP809 knock-out mouse model should serve well to investigate the biological relevance of the transcriptional repression of *Gm9705* by ZFP809. Moreover, this model represents an interesting opportunity to study KRAB-ZFP/TRIM28 mediated ERV silencing. In contrast to TRIM28 or SET1DB knock-out, the depletion of ZFP809 does not lead to prenatal death of mouse embryos and therefore allows studying the function of KRAB-ZFP silencing in adult animals and virtually all tissues of interest. This may be of special importance since our results indicate that KRAB-ZFPs is not restricted to ES and EC cells, which were the main focus of previous studies related to TRIM28-dependent ERV silencing.

In summary, we have conducted the first study investigating epigenetic repression of PERVs by histone modifications and provide preliminary data indicating that repressive histone methylations of HERVs may be a key factor in HERV associated diseases. The genome-wide binding analysis of ZFP809 in combination with the generation a ZFP809 knock-out mouse model confirmed a speculated function of ZFP809 in ERV repression but also identified ZFP809 binding sites at gene promoter regions of which at least one seems to serve as a control element for gene expression. Altogether, this study brings new insights into the recently emerged field of target-specific ERV silencing and highlights the evolutionary and epigenetic interplay between ERVs, retroviral repressors and host genes.

# References

1. Modin C, Pedersen FS, Duch M (2000) Lack of shielding of primer binding site silencer-mediated repression of an internal promoter in a retrovirus vector by the putative insulators scs, BEAD-1, and HS4. J Virol 74: 11697-11707.

2. Weiss RA (2006) The discovery of endogenous retroviruses. Retrovirology 3: 67.

3. Katzourakis A, Tristem M, Pybus OG, Gifford RJ (2007) Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A 104: 6261-6265.

4. Gifford R, Tristem M (2003) The evolution, distribution and diversity of endogenous retroviruses. Virus Genes 26: 291-315.

5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.

6. Turner G, Barbulescu M, Su M, Jensen-Seaman MI, Kidd KK, et al. (2001) Insertional polymorphisms of full-length endogenous retroviruses in humans. Curr Biol 11: 1531-1535.

7. Jha AR, Nixon DF, Rosenberg MG, Martin JN, Deeks SG, et al. (2011) Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. PLoS One 6: e20234.

8. Stocking C, Kozak CA (2008) Murine endogenous retroviruses. Cell Mol Life Sci 65: 3383-3398.

9. Le Tissier P, Stoye JP, Takeuchi Y, Patience C, Weiss RA (1997) Two sets of human-tropic pig retrovirus. Nature 389: 681-682.

10. Patience C, Takeuchi Y, Weiss RA (1997) Infection of human cells by an endogenous retrovirus of pigs. Nat Med 3: 282-286.

11. Tarlinton R, Meers J, Young P (2008) Biology and evolution of the endogenous koala retrovirus. Cell Mol Life Sci 65: 3413-3421.

12. Katzourakis A, Rambaut A, Pybus OG (2005) The evolutionary dynamics of endogenous retroviruses. Trends Microbiol 13: 463-468.

13. Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T (2004) Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. Nat Genet 36: 534-539.

14. Ribet D, Dewannieux M, Heidmann T (2004) An active murine transposon family pair: retrotransposition of "master" MusD copies and ETn trans-mobilization. Genome Res 14: 2261-2267.

15. French NS, Norton JD (1997) Structure and functional properties of mouse VL30 retrotransposons. Biochim Biophys Acta 1352: 33-47.

16. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, et al. (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet 2: e2.

17. Mi S, Lee X, Li X, Veldman GM, Finnerty H, et al. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403: 785-789.

18. Blaise S, de Parseval N, Benit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. Proc Natl Acad Sci U S A 100: 13013-13018.

19. Dupressoir A, Marceau G, Vernochet C, Benit L, Kanellopoulos C, et al. (2005) Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc Natl Acad Sci U S A 102: 725-730.

20. Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. Annu Rev Genet 42: 709-732.

21. Beyer U, Moll-Rocek J, Moll UM, Dobbelstein M (2011) Endogenous retrovirus drives hitherto unknown proapoptotic p63 isoforms in the male germ line of humans and great apes. Proc Natl Acad Sci U S A 108: 3624-3629.

22. Song X, Sui A, Garen A (2004) Binding of mouse VL30 retrotransposon RNA to PSF protein induces genes repressed by PSF: effects on steroidogenesis and oncogenesis. Proc Natl Acad Sci U S A 101: 621-626.

23. Murchison EP, Stein P, Xuan Z, Pan H, Zhang MQ, et al. (2007) Critical roles for Dicer in the female germline. Genes Dev 21: 682-693.

24. Piriyapongsa J, Marino-Ramirez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. Genetics 176: 1323-1337.

25. Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, et al. (2010) Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat Med 16: 571-579, 571p following 579.

26. Young GR, Eksmond U, Salcedo R, Alexopoulou L, Stoye JP, et al. (2012) Resurrection of endogenous retroviruses in antibody-deficient mice. Nature.

27. Ikeda H, Sugimura H (1989) Fv-4 resistance gene: a truncated endogenous murine leukemia virus with ecotropic interference properties. J Virol 63: 5405-5412.

28. Benit L, De Parseval N, Casella JF, Callebaut I, Cordonnier A, et al. (1997) Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. J Virol 71: 5652-5657.

29. Pryciak PM, Varmus HE (1992) Fv-1 restriction and its effects on murine leukemia virus integration in vivo and in vitro. J Virol 66: 5959-5966.

30. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, et al. (2004) The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. Nature 427: 848-853.

31. Towers GJ (2007) The control of viral infection by tripartite motif proteins and cyclophilin A. Retrovirology 4: 40.

32. Kaiser SM, Malik HS, Emerman M (2007) Restriction of an extinct retrovirus by the human TRIM5alpha antiviral protein. Science 316: 1756-1758.

33. Harris RS, Bishop KN, Sheehy AM, Craig HM, Petersen-Mahrt SK, et al. (2003) DNA deamination mediates innate immunity to retroviral infection. Cell 113: 803-809.

34. Esnault C, Heidmann O, Delebecque F, Dewannieux M, Ribet D, et al. (2005) APOBEC3G cytidine deaminase inhibits retrotransposition of endogenous retroviruses. Nature 433: 430-433.

35. Okeoma CM, Lovsin N, Peterlin BM, Ross SR (2007) APOBEC3 inhibits mouse mammary tumour virus replication in vivo. Nature 445: 927-930.

36. Sanville B, Dolan MA, Wollenberg K, Yan Y, Martin C, et al. (2010) Adaptive evolution of Mus Apobec3 includes retroviral insertion and positive selection at two clusters of residues flanking the substrate groove. PLoS Pathog 6: e1000974.

37. Gao G, Guo X, Goff SP (2002) Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. Science 297: 1703-1706.

38. Guo X, Ma J, Sun J, Gao G (2007) The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. Proc Natl Acad Sci U S A 104: 151-156.

39. Neil SJ, Zang T, Bieniasz PD (2008) Tetherin inhibits retrovirus release and is antagonized by HIV-1 Vpu. Nature 451: 425-430.

40. Liu J, Chen K, Wang JH, Zhang C (2010) Molecular evolution of the primate antiviral restriction factor tetherin. PLoS One 5: e11904.

41. Iguchi-Ariga SM, Schaffner W (1989) CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev 3: 612-619.

42. Hendrich B, Bird A (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol Cell Biol 18: 6538-6547.

43. Leonhardt H, Page AW, Weier HU, Bestor TH (1992) A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. Cell 71: 865-873.

44. Ooi SK, Bestor TH (2008) Cytosine methylation: remaining faithful. Curr Biol 18: R174-176.

45. Gowher H, Liebert K, Hermann A, Xu G, Jeltsch A (2005) Mechanism of stimulation of catalytic activity of Dnmt3A and Dnmt3B DNA-(cytosine-C5)-methyltransferases by Dnmt3L. J Biol Chem 280: 13341-13348.

46. Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. Trends Genet 13: 335-340.

47. Jaenisch R, Schnieke A, Harbers K (1985) Treatment of mice with 5-azacytidine efficiently activates silent retroviral genomes in different tissues. Proc Natl Acad Sci U S A 82: 1451-1455.

48. Brunmeir R, Lagger S, Simboeck E, Sawicka A, Egger G, et al. (2010) Epigenetic regulation of a murine retrotransposon by a dual histone modification mark. PLoS Genet 6: e1000927.

49. Walsh CP, Chaillet JR, Bestor TH (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat Genet 20: 116-117.

50. Bourc'his D, Bestor TH (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature 431: 96-99.

51. Jahner D, Stuhlmann H, Stewart CL, Harbers K, Lohler J, et al. (1982) De novo methylation and expression of retroviral genomes during mouse embryogenesis. Nature 298: 623-628.

52. Okano M, Bell DW, Haber DA, Li E (1999) DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 99: 247-257.

53. Cooper DN, Krawczak M (1989) Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 83: 181-188.

54. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, et al. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128: 1089-1103.

55. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ (2007) Developmentally regulated piRNA clusters implicate MILI in transposon control. Science 316: 744-747.

56. Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, et al. (2008) DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev 22: 908-917.

57. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, et al. (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell 31: 785-799.

58. Maksakova IA, Mager DL, Reiss D (2008) Keeping active endogenous retroviral-like elements in check: the epigenetic perspective. Cell Mol Life Sci 65: 3329-3347.

59. Leung DC, Lorincz MC (2012) Silencing of endogenous retroviruses: when and why do histone marks predominate? Trends Biochem Sci 37: 127-133.

60. Williams K, Christensen J, Pedersen MT, Johansen JV, Cloos PA, et al. (2011) TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature 473: 343-348.

61. Hutnick LK, Huang X, Loo TC, Ma Z, Fan G (2010) Repression of retrotransposal elements in mouse embryonic stem cells is primarily mediated by a DNA methylation-independent mechanism. J Biol Chem 285: 21082-21091.

62. Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, et al. (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell Stem Cell 8: 676-687.

63. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature 389: 251-260.

64. Kouzarides T (2007) Chromatin modifications and their function. Cell 128: 693-705.

65. Tse C, Sera T, Wolffe AP, Hansen JC (1998) Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III. Mol Cell Biol 18: 4629-4638.

66. Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, et al. (2004) Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell 119: 941-953.

67. Metzger E, Wissmann M, Yin N, Muller JM, Schneider R, et al. (2005) LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. Nature 437: 436-439.

68. Macfarlan TS, Gifford WD, Agarwal S, Driscoll S, Lettieri K, et al. (2011) Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev 25: 594-607.

69. Grewal SI, Jia S (2007) Heterochromatin revisited. Nat Rev Genet 8: 35-46.

70. Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, et al. (2005) The profile of repeat-associated histone lysine methylation states in the mouse epigenome. EMBO J 24: 800-812.

71. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448: 553-560.

72. Day DS, Luquette LJ, Park PJ, Kharchenko PV (2010) Estimating enrichment of repetitive elements from high-throughput sequence data. Genome Biol 11: R69.

73. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, et al. (2006) Control of developmental regulators by Polycomb in human embryonic stem cells. Cell 125: 301-313.

74. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, et al. (2006) Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature 441: 349-353.

75. Leeb M, Pasini D, Novatchkova M, Jaritz M, Helin K, et al. (2010) Polycomb complexes act redundantly to repress genomic repeats and genes. Genes Dev 24: 265-276.

76. Lehnertz B, Ueda Y, Derijck AA, Braunschweig U, Perez-Burgos L, et al. (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol 13: 1192-1200.

77. Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, et al. (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature 464: 927-931.

78. Maksakova IA, Goyal P, Bullwinkel J, Brown JP, Bilenky M, et al. (2011) H3K9me3-binding proteins are dispensable for SETDB1/H3K9me3-dependent retroviral silencing. Epigenetics Chromatin 4: 12.

79. Dexter TM, Scott D, Teich NM (1977) Infection of bone marrow cells in vitro with FLV: effects on stem cell proliferation, differentiation and leukemogenic capacity. Cell 12: 355-364.

80. Linney E, Davis B, Overhauser J, Chao E, Fan H (1984) Non-function of a Moloney murine leukaemia virus regulatory sequence in F9 embryonal carcinoma cells. Nature 308: 470-472.

81. Barklis E, Mulligan RC, Jaenisch R (1986) Chromosomal position or virus mutation permits retrovirus expression in embryonal carcinoma cells. Cell 47: 391-399.

82. Loh TP, Sievert LL, Scott RW (1987) Proviral sequences that restrict retroviral expression in mouse embryonal carcinoma cells. Mol Cell Biol 7: 3775-3784.

83. Loh TP, Sievert LL, Scott RW (1990) Evidence for a stem cell-specific repressor of Moloney murine leukemia virus expression in embryonal carcinoma cells. Mol Cell Biol 10: 4045-4057.

84. Yamauchi M, Freitag B, Khan C, Berwin B, Barklis E (1995) Stem cell factor binding to retrovirus primer binding site silencers. J Virol 69: 1142-1149.

85. Haas DL, Lutzko C, Logan AC, Cho GJ, Skelton D, et al. (2003) The Moloney murine leukemia virus repressor binding site represses expression in murine and human hematopoietic stem cells. J Virol 77: 9439-9450.

86. Wolf D, Goff SP (2007) TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. Cell 131: 46-57.

87. Wolf D, Hug K, Goff SP (2008) TRIM28 mediates primer binding site-targeted silencing of Lys1,2 tRNA-utilizing retroviruses in embryonic cells. Proc Natl Acad Sci U S A 105: 12521-12526.

88. Schultz DC, Friedman JR, Rauscher FJ, 3rd (2001) 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-2alpha subunit of NuRD. Genes Dev 15: 428-443.

89. Schultz DC, Ayyanathan K, Negorev D, Maul GG, Rauscher FJ, 3rd (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 Dev 16: 919-932.

90. Lechner MS, Begg GE, Speicher DW, Rauscher FJ, 3rd (2000) Molecular determinants for targeting heterochromatin protein 1-mediated gene silencing: direct chromoshadow domain-KAP-1 corepressor interaction is essential. Mol Cell Biol 20: 6449-6465.

91. Wolf D, Cammas F, Losson R, Goff SP (2008) Primer binding site-dependent restriction of murine leukemia virus requires HP1 binding by TRIM28. J Virol 82: 4675-4679.

92. Urrutia R (2003) KRAB-containing zinc-finger repressor proteins. Genome Biol 4: 231.

93. Emerson RO, Thomas JH (2009) Adaptive evolution in zinc finger transcription factors. PLoS Genet 5: e1000325.

94. Wolf D, Goff SP (2009) Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature 458: 1201-1204.

95. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, et al. (2010) KAP1 controls endogenous retroviruses in embryonic stem cells. Nature 463: 237-240.

96. Modin C, Lund AH, Schmitz A, Duch M, Pedersen FS (2000) Alleviation of murine leukemia virus repression in embryonic carcinoma cells by genetically engineered primer binding sites and artificial tRNA primers. Virology 278: 368-379.

97. Groner AC, Meylan S, Ciuffi A, Zangger N, Ambrosini G, et al. (2010) KRAB-zinc finger proteins and KAP1 can mediate long-range transcriptional repression through heterochromatin spreading. PLoS Genet 6: e1000869.

98. Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature 410: 116-120.

99. Stewart MD, Li J, Wong J (2005) Relationship between histone H3 lysine 9 methylation, transcription repression, and heterochromatin protein 1 recruitment. Mol Cell Biol 25: 2525-2538.

100. Fuks F, Hurd PJ, Deplus R, Kouzarides T (2003) The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. Nucleic Acids Res 31: 2305-2312.

101. Honda S, Selker EU (2008) Direct interaction between DNA methyltransferase DIM-2 and HP1 is required for DNA methylation in Neurospora crassa. Mol Cell Biol 28: 6044-6055.

102. Lorincz MC, Schubeler D, Groudine M (2001) Methylation-mediated proviral silencing is associated with MeCP2 recruitment and localized histone H3 deacetylation. Mol Cell Biol 21: 7913-7922.

103. Schlesinger S, Goff SP (2012) Silencing of proviruses in embryonic cells: efficiency, stability and chromatin modifications. EMBO Rep.

104. Thomas JH, Schneider S (2011) Coevolution of retroelements and tandem zinc finger genes. Genome Res 21: 1800-1812.

105. Lund AH, Schmidt J, Luz A, Sorensen AB, Duch M, et al. (1999) Replication and pathogenicity of primer binding site mutants of SL3-3 murine leukemia viruses. J Virol 73: 6117-6122.

106. Lund AH, Duch M, Pedersen FS (2000) Selection of functional tRNA primers and primer binding site sequences from a retroviral combinatorial library: identification of new functional tRNA primers in murine leukemia virus replication. Nucleic Acids Res 28: 791-799.

107. Czauderna F, Fischer N, Boller K, Kurth R, Tonjes RR (2000) Establishment and characterization of molecular clones of porcine endogenous retroviruses replicating on human cells. J Virol 74: 4028-4038.

108. Deng YM, Tuch BE, Rawlinson WD (2000) Transmission of porcine endogenous retroviruses in severe combined immunodeficient mice xenotransplanted with fetal porcine pancreatic cells. Transplantation 70: 1010-1016.

109. Boneva RS, Folks TM (2004) Xenotransplantation and risks of zoonotic infections. Ann Med 36: 504-517.

110. Paradis K, Langford G, Long Z, Heneine W, Sandstrom P, et al. (1999) Search for cross-species transmission of porcine endogenous retrovirus in patients treated with living pig tissue. The XEN 111 Study Group. Science 285: 1236-1241.

111. Meije Y, Tonjes RR, Fishman JA (2010) Retroviral restriction factors and infectious risk in xenotransplantation. Am J Transplant 10: 1511-1516.

112. Petkov SG, Marks H, Klein T, Garcia RS, Gao Y, et al. (2011) In vitro culture and characterization of putative porcine embryonic germ cells derived from domestic breeds and Yucatan mini pig embryos at Days 20-24 of gestation. Stem Cell Res 6: 226-237.

113. Fujimura T, Miyagawa S, Takahagi Y, Shigehisa T, Murakami H (2008) Prevalence of porcine endogenous retroviruses in domestic, minature, and genetically modified pigs in Japan. Transplant Proc 40: 594-595.

114. Mang R, Maas J, Chen X, Goudsmit J, van Der Kuyl AC (2001) Identification of a novel type C porcine endogenous retrovirus: evidence that copy number of endogenous retroviruses increases during host inbreeding. J Gen Virol 82: 1829-1834.

115. Patience C, Switzer WM, Takeuchi Y, Griffiths DJ, Goward ME, et al. (2001) Multiple groups of novel retroviral genomes in pigs and related species. J Virol 75: 2771-2775.

116. Tonjes RR, Niebert M (2003) Relative age of proviral porcine endogenous retrovirus sequences in Sus scrofa based on the molecular clock hypothesis. J Virol 77: 12363-12368.

117. Morita S, Kojima T, Kitamura T (2000) Plat-E: an efficient and stable system for transient packaging of retroviruses. Gene Ther 7: 1063-1066.

118. Markert LD, Lovmand J, Foss M, Lauridsen RH, Lovmand M, et al. (2009) Identification of distinct topographical surface microstructures favoring either undifferentiated expansion or differentiation of murine embryonic stem cells. Stem Cells Dev 18: 1331-1342.

119. Dalsgaard T, Moldt B, Sharma N, Wolf G, Schmitz A, et al. (2009) Shielding of sleeping beauty DNA transposon-delivered transgene cassettes by heterologous insulators in early embryonal cells. Mol Ther 17: 121-130.

120. Xue X, Huang X, Nodland SE, Mates L, Ma L, et al. (2009) Stable gene transfer and expression in cord blood-derived CD34+ hematopoietic stem and progenitor cells by a hyperactive Sleeping Beauty transposon system. Blood 114: 1319-1330.

121. Nielsen AL, Ortiz JA, You J, Oulad-Abdelghani M, Khechumian R, et al. (1999) Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. EMBO J 18: 6385-6395.

122. Bromham L, Clark F, McKee JJ (2001) Discovery of a novel murine type C retrovirus by data mining. J Virol 75: 3053-3057.

123. Guenet JL, Bonhomme F (2003) Wild mice: an ever-increasing contribution to a popular mammalian model. Trends Genet 19: 24-31.

124. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M (2005) High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. Mol Biol Evol 22: 814-817.

125. Sripathy SP, Stevens J, Schultz DC (2006) The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. Mol Cell Biol 26: 8623-8638.

126. Talbert PB, Henikoff S (2006) Spreading of silent chromatin: inaction at a distance. Nat Rev Genet 7: 793-803.

127. Delson E (2006) Mammal species of the world: a taxonomic and geographic reference. Choice: Current Reviews for Academic Libraries 44: 456-456.

128. Steppan S, Adkins R, Anderson J (2004) Phylogeny and divergence-date estimates of rapid radiations in muroid rodents based on multiple nuclear genes. Syst Biol 53: 533-553.

129. Jansa SA, Weksler M (2004) Phylogeny of muroid rodents: relationships within and among major lineages as determined by IRBP gene sequences. Mol Phylogenet Evol 31: 256-276.

130. Maksakova IA, Zhang Y, Mager DL (2009) Preferential epigenetic suppression of the autonomous MusD over the nonautonomous ETn mouse retrotransposons. Mol Cell Biol 29: 2456-2468.

131. Rogers SL, Rouhi A, Takei F, Mager DL (2006) A role for DNA hypomethylation and histone acetylation in maintaining allele-specific expression of mouse NKG2A in developing and mature NK cells. J Immunol 177: 414-421.

132. Kosir R, Acimovic J, Golicnik M, Perse M, Majdic G, et al. (2010) Determination of reference genes for circadian studies in different tissues and mouse strains. BMC Mol Biol 11: 60.

133. Ayyanathan K, Lechner MS, Bell P, Maul GG, Schultz DC, et al. (2003) Regulated recruitment of HP1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. Genes Dev 17: 1855-1869.

134. O'Geen H, Squazzo SL, Iyengar S, Blahnik K, Rinn JL, et al. (2007) Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. PLoS Genet 3: e89.

135. Iyengar S, Ivanov AV, Jin VX, Rauscher FJ, 3rd, Farnham PJ (2011) Functional analysis of KAP1 genomic recruitment. Mol Cell Biol 31: 1833-1847.

136. Frietze S, O'Geen H, Blahnik KR, Jin VX, Farnham PJ (2010) ZNF274 recruits the histone methyltransferase SETDB1 to the 3' ends of ZNF genes. PLoS One 5: e15082.

137. Frietze S, Lan X, Jin VX, Farnham PJ (2010) Genomic targets of the KRAB and SCAN domain-containing zinc finger protein 263. J Biol Chem 285: 1393-1403.

138. Iyengar S, Farnham PJ (2011) KAP1 protein: an enigmatic master regulator of the genome. J Biol Chem 286: 26267-26276.

139. O'Geen H, Frietze S, Farnham PJ (2010) Using ChIP-seq technology to identify targets of zinc finger transcription factors. Methods Mol Biol 649: 437-455.

140. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9: R137.

141. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, et al. (2010) GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol 28: 495-501.

142. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, et al. (2007) High-resolution profiling of histone methylations in the human genome. Cell 129: 823-837.

143. Hawkins RD, Hon GC, Yang C, Antosiewicz-Bourget JE, Lee LK, et al. (2011) Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. Cell Res 21: 1393-1409.

144. Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, et al. (2012) ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res 40: D912-917.

145. Kempler G, Freitag B, Berwin B, Nanassy O, Barklis E (1993) Characterization of the Moloney murine leukemia virus stem cell-specific repressor binding site. Virology 193: 690-699.

146. Quenneville S, Turelli P, Bojkowska K, Raclot C, Offner S, et al. (2012) The KRAB-ZFP/KAP1 System Contributes to the Early Embryonic Establishment of Site-Specific DNA Methylation Patterns Maintained during Development. Cell Rep 2: 766-773.

147. Santoni de Sio FR, Massacand J, Barde I, Offner S, Corsinotti A, et al. (2012) KAP1 regulates gene networks controlling mouse B-lymphoid cell differentiation and function. Blood 119: 4675-4685.

148. Peng H, Begg GE, Schultz DC, Friedman JR, Jensen DE, et al. (2000) Reconstitution of the KRAB-KAP-1 repressor complex: a model system for defining the molecular anatomy of RING-B box-coiled-coil domain-mediated protein-protein interactions. J Mol Biol 295: 1139-1162.

149. Yokoyama A, Okuno Y, Chikanishi T, Hashiba W, Sekine H, et al. (2010) KIAA1718 is a histone demethylase that erases repressive histone methyl marks. Genes Cells 15: 867-873.

150. Lai WS, Thompson MJ, Taylor GA, Liu Y, Blackshear PJ (1995) Promoter analysis of Zfp-36, the mitogen-inducible gene encoding the zinc finger protein tristetraprolin. J Biol Chem 270: 25266-25272.

151. Blackshear PJ (2002) Tristetraprolin and other CCCH tandem zinc-finger proteins in the regulation of mRNA turnover. Biochem Soc Trans 30: 945-952.

152. Sanduja S, Blanco FF, Young LE, Kaza V, Dixon DA (2012) The role of tristetraprolin in cancer and inflammation. Front Biosci 17: 174-188.

153. Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, et al. (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet 7: e1002301.

154. Nelson DR, Zeldin DC, Hoffman SM, Maltais LJ, Wain HM, et al. (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. Pharmacogenetics 14: 1-18.

155. Thomas JH (2007) Rapid birth-death evolution specific to xenobiotic cytochrome P450 genes in vertebrates. PLoS Genet 3: e67.

156. Villesen P, Aagaard L, Wiuf C, Pedersen FS (2004) Identification of endogenous retroviral reading frames in the human genome. Retrovirology 1: 32.

157. Jha AR, Pillai SK, York VA, Sharp ER, Storm EC, et al. (2009) Cross-sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before Homo sapiens. Mol Biol Evol 26: 2617-2626.

158. Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, et al. (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. Genome Res 16: 1548-1556.

159. Ruprecht K, Ferreira H, Flockerzi A, Wahl S, Sauter M, et al. (2008) Human endogenous retrovirus family HERV-K(HML-2) RNA transcripts are selectively packaged into retroviral particles produced by the human germ cell tumor line Tera-1 and originate mainly from a provirus on chromosome 22q11.21. J Virol 82: 10008-10016.

160. Flockerzi A, Ruggieri A, Frank O, Sauter M, Maldener E, et al. (2008) Expression patterns of transcribed human endogenous retrovirus HERV-K(HML-2) loci in human tissues and the need for a HERV Transcriptome Project. BMC Genomics 9: 354.

161. Galli UM, Sauter M, Lecher B, Maurer S, Herbst H, et al. (2005) Human endogenous retrovirus rec interferes with germ cell development in mice and may cause carcinoma in situ, the predecessor lesion of germ cell tumors. Oncogene 24: 3223-3228.

162. Denne M, Sauter M, Armbruester V, Licht JD, Roemer K, et al. (2007) Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein. J Virol 81: 5607-5616.

163. Douville R, Liu J, Rothstein J, Nath A (2011) Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. Ann Neurol 69: 141-151.

164. Brudek T, Christensen T, Aagaard L, Petersen T, Hansen HJ, et al. (2009) B cells and monocytes from patients with active multiple sclerosis exhibit increased surface expression of both HERV-H Env and HERV-W Env, accompanied by increased seroreactivity. Retrovirology 6: 104.

165. Antony JM, Deslauriers AM, Bhat RK, Ellestad KK, Power C (2011) Human endogenous retroviruses and multiple sclerosis: innocent bystanders or disease determinants? Biochim Biophys Acta 1812: 162-176.

166. Nexo BA, Christensen T, Frederiksen J, Moller-Larsen A, Oturai AB, et al. (2011) The etiology of multiple sclerosis: genetic evidence for the involvement of the human endogenous retrovirus HERV-Fc1. PLoS One 6: e16652.

167. Laska MJ, Brudek T, Nissen KK, Christensen T, Moller-Larsen A, et al. (2012) Expression of HERV-Fc1, a human endogenous retrovirus, is increased in patients with active multiple sclerosis. J Virol 86: 3713-3722.

168. Kuppers R (2009) The biology of Hodgkin's lymphoma. Nat Rev Cancer 9: 15-27.

169. Gotzinger N, Sauter M, Roemer K, Mueller-Lantzsch N (1996) Regulation of human endogenous retrovirus-K Gag expression in teratocarcinoma cell lines and human tumours. J Gen Virol 77 ( Pt 12): 2983-2990.

170. Stengel S, Fiebig U, Kurth R, Denner J (2010) Regulation of human endogenous retrovirus-K expression in melanomas by CpG methylation. Genes Chromosomes Cancer 49: 401-411.

171. Lavie L, Kitova M, Maldener E, Meese E, Mayer J (2005) CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2). J Virol 79: 876-883.

172. Huda A, Marino-Ramirez L, Jordan IK (2010) Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mob DNA 1: 2.

173. Ram O, Goren A, Amit I, Shoresh N, Yosef N, et al. (2011) Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell 147: 1628-1639.

174. Kayibanda B, Rosenfeld C, Goutner A, Bornkamm GW (1978) A new lymphoid cell line, Reh 6, with characteristics of non-T and non-B cells, lacking the Epstein-Barr virus genome and derived from human acute lymphoblastic leukemia. Intervirology 9: 316-320.

175. Gallagher R, Collins S, Trujillo J, McCredie K, Ahearn M, et al. (1979) Characterization of the continuous, differentiating myeloid cell line (HL-60) from a patient with acute promyelocytic leukemia. Blood 54: 713-733.

176. Kamesaki H, Fukuhara S, Tatsumi E, Uchino H, Yamabe H, et al. (1986) Cytochemical, immunologic, chromosomal, and molecular genetic analysis of a novel cell line derived from Hodgkin's disease. Blood 68: 285-292.

177. Kaltoft K, Thestrup-Pedersen K, Jensen JR, Bisballe S, Zachariae H (1984) Establishment of T and B cell lines from patients with mycosis fungoides. Br J Dermatol 111: 303-308.

178. Kaltoft K, Bisballe S, Dyrberg T, Boel E, Rasmussen PB, et al. (1992) Establishment of two continuous T-cell strains from a single plaque of a patient with mycosis fungoides. In Vitro Cell Dev Biol 28A: 161-167.

179. Kaltoft K, Bisballe S, Rasmussen HF, Thestrup-Pedersen K, Thomsen K, et al. (1987) A continuous T-cell line from a patient with Sezary syndrome. Arch Dermatol Res 279: 293-298.

180. Klein E, Klein G, Nadkarni JS, Nadkarni JJ, Wigzell H, et al. (1968) Surface IgM-kappa specificity on a Burkitt lymphoma cell in vivo and in derived culture lines. Cancer Res 28: 1300-1310.

181. Ben-Bassat H, Goldblum N, Mitrani S, Goldblum T, Yoffey JM, et al. (1977) Establishment in continuous culture of a new type of lymphocyte from a "Burkitt like" malignant lymphoma (line D.G.-75). Int J Cancer 19: 27-33.

182. Menezes J, Leibold W, Klein G, Clements G (1975) Establishment and characterization of an Epstein-Barr virus (EBC)-negative lymphoblastoid B cell line (BJA-B) from an exceptional, EBV-genome-negative African Burkitt's lymphoma. Biomedicine 22: 276-284.

183. Lidin B, Adams A (1975) Spontaneous interferon production and Epstein-barr virus antigen expression in two human lymphoid cell lines and their somatic cell hybrids. Intervirology 5: 205-215.

184. Liu J, Jones KL, Sumer H, Verma PJ (2009) Stable transgene expression in human embryonic stem cells after simple chemical transfection. Mol Reprod Dev 76: 580-586.

185. Pal R, Ravindran G (2006) Assessment of pluripotency and multilineage differentiation potential of NTERA-2 cells as a model for studying human embryonic stem cells. Cell Prolif 39: 585-598.

186. Fuchs NV, Kraft M, Tondera C, Hanschmann KM, Lower J, et al. (2011) Expression of the human endogenous retrovirus (HERV) group HML-2/HERV-K does not depend on canonical promoter elements but is regulated by transcription factors Sp1 and Sp3. J Virol 85: 3436-3448.

187. Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, et al. (2012) Epigenetic interplay between mouse endogenous retroviruses and host genes. Genome Biol 13: R89.

188. Buzdin A, Kovalskaya-Alexandrova E, Gogvadze E, Sverdlov E (2006) At least 50% of human-specific HERV-K (HML-2) long terminal repeats serve in vivo as active promoters for host nonrepetitive DNA transcription. J Virol 80: 10752-10762.

189. Alexiadis V, Ballestas ME, Sanchez C, Winokur S, Vedanarayanan V, et al. (2007) RNAPol-ChIP analysis of transcription from FSHD-linked tandem repeats and satellite DNA. Biochim Biophys Acta 1769: 29-40.

# Appendix: manuscript under submission