Identification and characterisation of murine metastable epialleles conferred by endogenous retroviruses

Anastasiya Kazachenka

Department of Genetics

Darwin College

University of Cambridge

September 2017

This dissertation is submitted for the degree of Doctor of Philosophy

The research in this dissertation was carried out in the Department of Genetics, University of Cambridge, under the supervision of Professor Anne Ferguson-Smith. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except specified in the text

It does not exceed the prescribed word limit of 60,000 words.

Summary

Anastasiya Kazachenka

Identification and characterisation of murine metastable epialleles conferred by endogenous retroviruses

Repetitive sequences, including transposable elements, represent approximately half of the mammalian genome. Epigenetic mechanisms evolved to repress these potentially deleterious mobile elements. However, such elements can be variably silenced between individuals – so called 'metastable epialleles'. The best known example is the A^{vy} locus where an endogenous retrovirus (ERV) of the intracisternal A-particle (IAP) class was spontaneously inserted upstream of the agouti coat colour gene, resulting in variable IAP promoter DNA methylation, variable expressivity of coat phenotype, and environmentally modulated transgenerational epigenetic inheritance within genetically identical individuals. It is not known whether the behaviour exhibited by the ERV at A^{vy} represents a common occurrence throughout the genome or is unusual. Taking a genetic approach in purified cell populations, I have conducted a systematic genome-wide screen of murine metastable epialleles. I have identified over 100 murine IAPs with properties of metastable epialleles. Like A^{vy}, each exhibits a stable epigenetic state within an individual but epigenetic variability between individuals. Methylation levels are locus-specific within an individual, suggesting cis-acting control. The same screening strategy was applied for identification of metastable epialleles associated with other types of LTR-retroelements. However, many of identified candidates showed no inter-individual methylation variation upon experimental validation. These results suggest that IAPs are the dominant class of ERVs capable of acquiring epigenetic states that are variable between genetically identical individuals. I have conducted an analysis of IAP induced initiation and termination of transcription events using *de novo* assembled transcriptomes generated for B and T cells. 142 IAPs have been identified to overlap de novo assembled transcripts. 32 IAPs are metastable epialleles. Several of them show an inverse correlation between LTR promoter methylation and adjacent gene expression. In addition, I have shown that metastable epialleles have a characteristic pattern of histone modification and are flanked by the

methylation sensitive binding factor CTCF, providing testable hypotheses concerning the establishment and/or maintenance of the variable methylation state.

My findings indicate that metastability is, in general, specific to the IAP class of ERVs, that only around 1% of these elements have this unusual epigenetic property and that the ability to impact transcription, such as at agouti in A^{vy}, is not a ubiquitous feature of these loci.

Contents

List of figures

List of tables

Abbreviations

5mC - 5-methylcytosine

5hmC - 5-hydroxymethylcytosine

5fC - 5-formylcytosine

5caC - 5-carboxylcytosine

BER - base excision repair

BPA - bisphenol A

CGI -CpG island

CLP - common lymphoid progenitor

DE - differentially expressed

DMI - differentially methylated IAP

DMR - differentially methylated region

E - embryonic day

ERV - endogenous retrovirus

ESC - embryonic stem cell

EST - expressed sequence tag

ETn - early transposon

ETP - early thymic progenitor

FGO – fully-grown oocytes

FPKM - fragments per kilobase of exon per million

GMP - G/M progenitor

HSC - hematopoietic stem cell

IAP - intracisternal A-particle

ICM - inner cell mass

ICR - imprinting-control region

KO – knockout

LINE - long-interspersed nucleotide element

LMPP - lymphoid-primed multipotent progenitor

LMR - low methylated region

LTR - long terminal repeat

MaLR - mammalian apparent LTR retrotransposon

ME - metastable epiallele

MuERV-L - murine ERV-L element

MZ - marginal zone

NGO – nongrowing oocytes

NSC - neural stem cell

PGCs - primordial germ cells

p/mUPD - paternal/maternal uniparental disomy

SINE - short-interspersed nucleotide element

T1 B cells - transitional type I B cells

TAD - topologically associating domain

TCR complex - T-cell receptor complex

TDG - thymine-DNA glycosylase

TE - transposable element

TPM - transcript per million

TSS - transcription start site

UMR - unmethylated region

WGBS-seq - whole-genome bisulfite sequencing

WGoxBS-seq - whole-genome oxidative bisulfite sequencing

Chapter 1

Introduction

The term “epigenetics” was firstly proposed in 1940s by the developmental biologist Conrad H. Waddington and was mostly referred to as the interactions between genes and their surroundings within the fertilized egg (Hurd, 2010). These interactions were suggested to direct epigenesis and result in multicellular organism development. Since that time the meaning of “epigenetics” has changed significantly. The discovery of DNA structure in 1953 by J. Watson and F. Crick started new era in biology where DNA was considered to be the molecule of heredity and gene was the unit of heredity. However, it took time to understand that while inheriting the same DNA sequence, different cells acquire absolutely different morphologies and functions.

Today epigenetics can be defined as the study of heritable gene expression changes that do not involve changes in the underlying DNA sequence (Hurd, 2010). There are two types of epigenetic modifications impacting chromosome organisation: histone modifications and DNA methylation. The influence of a particular epigenetic mark on gene expression is context dependent. The cell-specific interplay between the DNA sequence and epigenetic modifications results in acquisition of different chromatin states and as a result different cell fates.

Histone modifications. Histones are proteins that interact with DNA and allow its organisation into a nucleoprotein complex known as chromatin. The histone “tails” that are not involved in direct interactions with DNA are targets for different modifications: acetylation, methylation, phosphorylation, and ubiquitination. These post-translational histone modifications act as substrates for protein complexes that can influence gene expression and chromatin structure. Interplay between histone modifications and DNA methylation creates distinctive epigenetic signatures marking regulatory elements, heterochromatin, transcriptionally active regions, silenced transposons etc. For example, active enhancers generally correlated with low methylation, H3K27ac and H3K4me1 (Zentner *et al*., 2011). Promoter regions of active genes are unmethylated and marked by H3K4me3. The function and localization of some of histone marks are represented in **Table 1**.

Table 1. Function and localization of histone marks

Histone Mark	Function	Localization	Reference
H3K4me3	Active	Poised and active promoters	Mikkelsen et al., 2007; Schuettengruber at al., 2007
H3K4me1	Active	Active and poised enhancers	Heintzman et al., 2007; Creyghton et al., 2010
H3K27ac	Active	Active enhancers	Creyghton et al., 2010
H3K27me3	Repressive	Poised and silenced promoters; inactive X chromosome	Plath et al., 2003; Mikkelsen et al., 2007; Schuettengruber at al., 2007
H3K9me3	Repressive	Heterochromatin; silenced promoters; retrotransposons	Bannister et al., 2001; Barski et al., 2007; Matsui et al., 2010
H4K20me3	Repressive	Heterochromatin, LTR retrotransposons	Schotta et al., 2004; Matsui et al., 2010; Rangasamy, 2013
H3K36me3	Active	Gene bodies of actively transcribed genes	Barski et al., 2007; Mikkelsen et al., 2007
H3K9ac	Active	Active promoters	Karmodiya et al., 2012

DNA methylation. One of the longest studied epigenetic modifications contributing to regulation of gene expression is methylation of the carbon atom at the 5th position in cytosine residue (5mC) (Jones, 2012). 5mC can be found in a variety of contexts (CG, CNG, CNN) in plants, but it is mostly restricted to CpG dinucleotides in mammals (Suzuki & Bird,2008). There are two DNA methyltransferases that are responsible for de novo establishment of DNA methylation: DNMT3a and DNMT3b. *Dnmt3b*$^{-/-}$ KO mice are embryonic lethal and exhibit growth and development arrest after E9.5. In contrast, *Dnmt3a*$^{-/-}$ mutants can survive up to 4 weeks after birth. [*Dnmt3a*$^{-/-}$, *Dnmt3b*$^{-/-}$] double KO embryos die before E11.5 and fail to initiate de novo

DNA methylation after implantation (Okano et al., 1999). DNMT3L, a member of DNMT family that lacks enzymatic activity, can cooperate with DNMT3a/3b and is required for establishment of maternal imprints and spermatogenesis (Hata et al., 2002; Bourc'his & Bestor, 2004). During replication DNA methylation is maintained by DNMT1. *Dnmt1*$^{-/-}$ mutants are embryonic lethal and have decreased global DNA methylation levels (Li et al, 1992). DNMT1 is tethered to hemimethylated DNA by NP95 to promote DNA methylation maintenance after replication (Bostick et al., 2007).

The erasure of DNA methylation can happen through active and passive mechanisms. Active DNA demethylation is catalyzed by Tet methylcytosine dioxygenase enzymes that oxidize 5mC to 5-hydroxymethylcytosine (5hmC). 5hmC can be further oxidized to 5-formylcytosine (5fC) and to 5-carboxylcytosine (5caC) that can be replaced by unmethylated cytosine through thymine-DNA glycosylase (TDG)-mediated base excision repair (BER). Passive demethylation happens gradually through cell divisions in the absence of DNA methyltransferase activity when newly synthesized DNA strands cannot get methylated (Bhutani et al., 2011).

Mammalian genomes are generally globally methylated in somatic tissues excluding unmethylated CpG islands (CGIs) and active regulatory elements (Suzuki & Bird, 2008). The function of DNA methylation is context dependent. Methylation of CGIs correlates with long-term transcriptional silencing such as observed at the inactive X chromosome or at imprinting genes. At the same time, methylation of gene bodies does not cause gene silencing and, according to some evidence, positively correlates with transcriptional elongation (Aran et al., 2011). Apart from these examples, DNA methylation is inversely correlated with enhancer activity, CTCF and transcription factor binding (Jones, 2012), although some factors have been shown to prefer a methylated substrate, such as some Methyl Binding Proteins (Hendrich & Bird, 1998) and the KRAB-zinc finger protein ZFP57 that regulates imprinting control regions (Li et al., 2008). Another role for DNA methylation is in the silencing of transposable elements (Walsh et al., 1998; Suzuki & Bird, 2008).

Genome-wide epigenetic reprogramming

Epigenetic marks are erased and re-established genome-wide twice during mammalian development (Reik et al., 2001; Cantone and Fisher, 2013; Messerschmidt et al., 2014). A first wave of epigenetic reprogramming happens during the development of primordial germ cells (PGCs) in the embryo. Reprogramming in PGCs starts from nearly complete loss of DNA methylation by E13-E14 days of embryonic development. Global DNA demethylation is accompanied by transient loss of histone marks including H3K9me2, H3K9me3, H3K27me3, and H3K9ac. De novo remethylation of the genome starts during the prospermatogonia stage in male germline and after birth in female germline. This reprogramming allows reestablishment of the imprints and derepression of inactive X chromosome in female germ cells. Furthermore, germline reprogramming is important as prevents the inheritance of parental epigenetic marks by the offspring.

The next reprogramming wave happens during early embryonic development. After fertilization, the paternally inherited genome goes through replacement of protamines with maternally-inherited histones and an almost immediate loss of DNA methylation through an active demethylation process. Active demethylation is mediated by TET enzymes that catalyze conversion of 5mC to 5hmC. There are three TET proteins in mouse, however only TET3 is highly expressed in the zygote. TET3 is believed to mediate rapid demethylation of the paternal genome after fertilization. Demethylation of the maternal genome happens gradually during cleavage divisions (passive demethylation) due to nuclear exclusion of DNMT1 although some active demethylation has recently been described (Guo et al., 2014). This wave of reprogramming on the two parentally inherited genomes is required to eliminate germline specific marks and establish pluripotency in the preimplantation embryo. In contrast to reprogramming in PGCs, many more genomic regions are resistant to global DNA demethylation during early embryo development. For example, methylation of imprints and some transposons is maintained.

Epigenetics of repetitive genome

Transposable elements (TEs) are mobile DNA sequences that comprise around 37% of the murine genome and 45% of the human genome (Muñoz-López & García-Pérez, 2010). TEs can be divided into two classes based on their mechanism of transposition: DNA transposons and retrotransposons. DNA transposons excise themselves and integrate into another region ("cut-and-paste" transposition). Retrotransposons are transposed by "copy-and-paste" mechanism that involves a reverse transcription of an RNA intermediate (Muñoz-López & García-Pérez, 2010).

There are three classes of retrotransposons found in mammals: long-interspersed nucleotide elements (LINEs), short-interspersed nucleotide elements (SINEs), and long-terminal direct repeat-containing retrotransposons (LTR retrotransposons). Around 10% of the murine genome is represented by LTR retrotransposons, all of which belong to the endogenous retrovirus (ERV) superfamily (Stocking & Kozak, 2008). ERVs are descendants of exogenous retroviruses that colonised the host genome. Three classes of ERVs can be distinguished in the murine genome. Class III ERVs are the most ancient ERVs and have been identified in all placental mammals. This class includes murine ERV-L elements (MuERV-L) and the most common murine ERV type - mammalian apparent LTR retrotransposons (MaLRs). Some of these are still active and can be responsible for new insertions in the genome. Class II ERVs are responsible for the most polymorphic mutations and are driven by the transposition of two ERV families: early transposons (MusD/ETn) and intracisternal A-type particles (IAPs) (Zhang et al., 2008). Class I is the smallest class and some class I ERVs are still active.

A full length intact ERV consists of 5' and 3' long terminal repeats flanking "internal" retroviral sequence (Falzon & Kuff, 1988; Mietz et al., 1987). The 5'LTR provides the promoter region for transcription of viral RNA that is an intermediate substrate during retroviral transposition. In addition to its promoter, the LTR contains enhancer sequences, transcription factor binding sites and a poly A site (functional at the 3'LTR). The "internal" retroviral sequence encodes viral genes (gag, pol and mutated non-functional env) and the primer binding site required for initiation of reverse transcription during transposition. ERVs that are capable of self-replication are called autonomous. Other ERVs that do not encode functional viral genes and require

proteins expressed from autonomous relatives for transposition are referred to as non-autonomous.

Active ERVs are capable of insertional mutagenesis, which can be detrimental to the organism. One of the main mechanisms ensuring ERV silencing is through the epigenetic control of retroviral insertions. DNA methylation is considered to have specifically evolved to regulate the activity of transposons. This hypothesis is supported by the fact that transposable elements are the main targets for methylation in fungi and some plants. In mouse, Dnmt1 is responsible for transcriptional silencing of at least some IAPs in somatic cells (Walsh et al., 1998). Dnmt3L knockout caused demethylation and transcriptional activation of ERVs in male germ cells (Bourc'his & Bestor, 2004). However, [*Dnmt1*$^{-/-}$, *Dnmt3a*$^{-/-}$, *Dnmt3b*$^{-/-}$] triple knockout ESCs do not show elevation of ERV transcription even though ERVs are largely demethylated (Tsumura et al., 2006; Matsui et al., 2010). Indeed, it has been shown that ERV silencing in cells undergoing reprogramming is maintained through DNA-methylation-independent recruitment of H3K9me3 to transposons (Matsui et al., 2010). H4K20me3 is another histone mark associated with LTR transposon silencing (Rangasamy, 2013). However no transcriptional activation of ERVs have been observed in cells depleted for H4K20 methyltransferases Suv420h1 and Suv420h2 suggesting that ERV silencing does not depend on the presence of this mark (Matsui et al., 2010). Apart from this, some IAPs have been shown to be resistant to the global demethylation that occurs after fertilization and they appear to be only partially demethylated during reprogramming in PGCs (Lane et al., 2003; Seisenberger et al., 2012; Arand et al., 2015). De novo methylation of ERVs relies on the activity of both Dnmt3a and Dnmt3b (Okano et al., 1999; Kato et al., 2007). The relative extent to which DNA methylation and H3K9me3 contribute to retroelement silencing and prevention of active transcription is still not clear for two reasons. First, the activation of repetitive elements in cells with mutations in the epigenetic mechanisms is generally assessed through analysis of bulk ERV family members hence those resistant to activation cannot be quantified. Second, mutation in the H3K9 methylation machinery results in severely unhealthy cells (Yuan et al., 2009) making direct effects on transposons, or indeed any other loci, difficult to assess and results hard to interpret.

Impact of retroelements on transcriptome

Recruitment of silencing epigenetic marks to transposable elements was shown to have an impact on heterochromatin formation at flanking genomic regions (Rebollo et al., 2011). Polymorphic insertion of active LTR transposons - IAPs and ETns - can induce spreading of H3K9me3 and H4K20me3 to the host genome. Sometimes this ERV-induced heterochromatin formation might influence gene expression through epigenetic silencing of promoter regions. At least one case has been reported when IAP insertion upstream of the B3galtl gene caused gene silencing in ESCs through spreading of H3K9me3 from IAP to TSS of this gene. Similarly, neighbouring genes can affect activity of ETn elements inserted in the proximity of TSSs through spreading of active epigenetic marks from gene promoter to retroelement (Rebollo at all., 2012). Demethylation of ETns caused ectopic expression of neighboring gene from the ETn promoter. Interestingly this spreading of active promoter marks (H3K4me3 and DNA demethylation) was observed only when the gene was expressed. No demethylation of ETn was observed when the neighbouring gene was not expressed in the tissue suggesting that transcription contributed to the ability of the adjacent ETn to acquire the active state. The interplay between the epigenetic marks of the host and the ERV was therefore implicated in the fine tuning of gene transcriptional activity.

Multiple cases have been described where ERVs have been found to be involved in regulation of neighbouring genes (Maksakova et al., 2006). Ectopic in vitro induced expression of erythroid transcription factor Pu.1 was initiated from the LTR promoter of ORR1A0 retrotransposon (Mak et al., 2014). In vivo this LTR transposon activity is regulated by antagonistic activities of erythroid specific transcription factors KLF1 and KLF3. The LTR driven chimeric transcript promotes erythroid differentiation and antagonizes normal PU.1 activity. Furthermore, a number of ORR1A0 driven transcripts have been found during EST screens suggesting in vivo activity of these retroelements.

Retrotransposon-derived genes were identified in the placenta where global DNA methylation levels are low (Sharif *et al*., 2013). For example, placenta-specific expression of MIPP was found to be induced from an IAP promoter that is silenced in other tissues (Chang-Yeh et al., 1991). Recruitment of ERVs for placenta-specific

expression of the prolactin gene (PRL gene) has also been described. The PRL gene is expressed from two different types of ERVs: MER39 and MER77 in human and mouse respectively (Emera et al., 2012). ERVs are a source of active enhancers in placental tissue (Chuong et al., 2013).

Apart from the placenta, ERVs are widely utilized as regulatory element in ESCs and testis, where global DNA methylation is low enough to create permissive epigenetic environment for ERV activity (Chuong et al., 2013). Intronic ERV insertions can cause premature termination of transcription from gene promoters either by providing alternative poly-A termination sites or through transcriptional interference (Li *et al*., 2012; Jern & Coffin, 2008). ERVs also can provide alternative splice sites that lead to transcription of different isoforms (Kapitonov & Jurka, 1999; Hughes, 2001). Interestingly, the usage of the mechanisms of aberrant transcription might be different between ERV types (Maksakova et al., 2006). There are more cases described with premature termination within ETns rather than IAPs. At the same time, IAP-driven ectopic gene expression is more common than ETn-driven transcription. Transposable element sequences were also reported to harbour binding sites for transcription factors including POU5F1-SOX2, CTCF, and ESR1 (Bourque et al., 2008).

Epialleles

Epigenetic mechanisms are widely suggested to link environmental exposures to phenotypic outcomes without triggering changes in the underlying genetic sequence (Feil & Fraga, 2012). Factors, such as temperature and diet, are associated with long-term epigenetic alterations genome-wide and in a locus-specific manner in plants, insects, and mammals. Genomic loci with interindividual epigenetic variation that is stably maintained during mitosis are called “epialleles” (Finer et al., 2011).

According to the classification proposed in 2006, epialleles can come in three different types depending on their relation to genetic variation: obligatory, facilitated, and pure epialleles (Richards, 2006). Obligatory epialleles are the direct outcomes of genetic polymorphisms and inherited in a Mendelian manner. Studies of epigenetic variation at the IGF2/H19 locus in monozygotic and dizygotic twins showed

significant correlation between DNA methylation levels and the presence of SNPs that sometimes cause an abolishment of CpG sites (Heijmans et al., 2007). Cis acting SNPs caused heritable epigenetic variation at this locus. Pure epialleles are the primary or secondary results of environmental influences and genetically independent. Monozygotic twins were found to have some epigenetic differences that impact gene expression. These differences accumulate with age and might be enhanced by diet, lifestyle and exercise (Fraga et al., 2005). The last group of epialleles is the class of facilitated epialleles. In this case genetic variation creates an environment for stochastic establishment of epigenetic marks but does not determine them. Facilitated epialleles do not follow Mendelian inheritance. An example of facilitated epialleles are the so-called "metastable epialleles" where epigenetic variation in genetically identical individuals is linked to the insertion of a transposable element.

Metastable epialleles

The term "metastable epiallele" is used to describe altered gene expression that is regulated by variable DNA methylation in genetically identical individuals. Methylation state at these regions is established stochastically during early development and not genetically determined. (Rakyan *et al*., 2002). So far a relatively small number of "metastable epialleles" have been reported with several being alleles at the Agouti locus (**Table 2**).

The most established example of a metastable epiallele is the Agouti viable yellow (A^{vy}) epiallele. It was first discovered as a spontaneous mutation in the colony of C3H/HeJ mice in 1960 (Dickie, 1962). Wild type Agouti encodes a paracrine signalling molecule and is expressed in skin, testes, and during fetal development (Yen et al., 1994). The Avy epiallele resulted from an IAP insertion upstream of Agouti gene promoter. The IAP provides a cryptic promoter that can drive an expression of Avy in nearly all tissues (Yen et al., 1994). Activity of an IAP promoter depends on its methylation state (Michaud et al., 1994). Partially or lowly methylated IAP promoter drives ubiquitous expression of agouti gene in multiple tissues that results in yellow or mottled coat color (Duhl et al., 1994, Yen et al., 1994; Michaud et

al., 1994; Waterland & Jirtle, 2003; Dolinoy, 2008). Moreover ectopic expression of agouti gene is associated with obesity, diabetes and cancer susceptibility (Wolff et al., 1986; Yen et al., 1994; Wolff, 1996). Alternatively, mice with highly methylated IAP look normal and have pseudoagouti coat color. The absence of methylation variation at this IAP locus between different tissues of the same individual suggests that IAP methylation happens early in development and is stably maintained through mitotic divisions (Waterland & Jirtle, 2003). Another well described metastable epiallele is $Axin^{fused}$. Similarly to A^{vy}, it arose from the insertion of an IAP element into the 6th intron of the Axin gene (Vasicek et al., 1997). The Axin gene produces a protein important for axis formation during development and is also expressed in adults. The intragenic IAP drives the expression of downstream Axin exons producing a truncated Axin transcript that misses upstream exons. The expression of this aberrant transcript correlates with IAP methylation and results in a “kinky tail” phenotype in hypomethylated individuals (Rakyan et al., 2003; Waterland et al., 2006). Again, IAP methylation levels are consistent between tissues of the same individual suggesting early developmental establishment of the incomplete methylation state (Waterland et al., 2006).

Table 2. Murine metastable epialleles

Allele	Founder strain	Retroelement insertion	Phenotype	Reference
A^{vy}	C3H/HeJ	IAP insertion in the 5'UTR exon 1A	Coat color, obesity	Duhl et al., 1994
A^{iapy}	C57BL/6J	IAP insertion upstream of the 2nd exon	Coat color	Michaud et al., 1994
A^{hvy}	C3H/HeJ	IAP insertion in the 5'UTR exon 1C	Coat color, obesity	Argeson et al., 1996
$Axin^{Fused}$	129/Rr	IAP insertion into intron 6	Kinky tail	Rakyan et al., 2003
$CABP^{IAP}$	C57BL/6J	IAP insertion into intron 6	no phenotype	Druker et al., 2004

The two described cases share four main features: 1) both epialleles result from insertion of an IAP element that provides an alternative promoter; 2) interindividual DNA methylation variation at the 5' LTR region correlates with variation of IAP-driven gene expression; 3) the methylation state at the IAP locus is constant between different tissues within the same individual; and 4) in both cases there is a visible phenotype that correlates with IAP methylation levels.

Metastable epialleles are sensitive to environmental exposures

Metastable epialleles have been used to study the influences of nutritional and environmental exposures on the establishment of DNA methylation marks during gestational periods. A number of studies found a phenotypic shift in the offspring towards pseudoagouti upon maternal methyl donor supplementation of A^{vy} mice (Wolff et al.,1998; Waterland & Jirtle, 2003; Cropley et al., 2006). The shift was even stronger when the methyl-donor supplemented diet included L-methionine and zinc - a DNMT cofactor (Wolff et al., 1998). However, the sensitivity of A^{vy} methylation to a maternal methyl supplemented dieting was observed only in the cases when the A^{vy} allele was paternally inherited. Methylation levels of maternally inherited A^{vy} metastable epialleles were not different between pups that were born from mothers fed a methyl donor diet and a control group of mothers (Cropley et al., 2006). Different strains have different susceptibility to methyl donor supplementation showing different distribution of phenotypes in the offspring (Wolff et al., 1998). A maternal diet enriched in methyl donors affects the establishment of methylation marks at the A^{vy} epiallele not only in the F1 offspring (fed normal diet) but also in the germline of the F1 pups. Furthermore, the spectrum of phenotypes in the F2 generation that were also fed a normal diet was shifted to pseudoagouti (Cropley et al., 2006). The $Axin^{fused}$ metastable epiallele is prone to be highly methylated in the offspring from methyl supplemented dams (Waterland et al., 2006). Similarly to the methyl donor supplementation experiments, genistein supplementation also causes a shift towards the pseudoagouti phenotype in the F1 generation (Dolinoy et al., 2006). In both cases, the methylation levels of metastable epialleles induced by maternal nutrition were consistent in the tissues derived from the same individual. This

suggests that maternal diet induces DNA methylation changes early in development prior germ layer differentiation.

In the A^{vy} models opposite effects on the distribution of phenotypes in the offspring was observed when maternal diet contained bisphenol A (BPA) (Dolinoy et al., 2007). Maternal BPA exposure caused an increase of yellow mice in the offspring. Interestingly, the BPA diet, accompanied by methyl donor or genistein supplementation, negated the shift of progeny phenotypes observed during BPA-only maternal exposure. Methylation levels at the A^{vy} metastable epiallele were also shown to be sensitive to maternal ethanol consumption (Kaminen-Ahola et al., 2010). Both preconceptional and gestational ethanol consumption caused a shift towards pseudoagouti mice in the A^{vy} progeny. The vulnerability of the methylation at A^{vy} during preimplantation development was also observed during the in vitro culture of zygotes (Morgan et al., 2008). Hypomethylation of A^{vy} occurred at a higher frequency among embryos from in vitro cultured zygotes rather than from transferred blastocysts or pregnancies without any interventions. It should be noted that most of these studies involved the analysis of coat colour phenotypes and did not directly assess methylation levels at A^{vy} individuals, which might lead to some bias in the obtained results.

Metastable epialleles as a model for transgenerational epigenetic inheritance

Two waves of genome-wide epigenetic reprogramming leave almost no chances for the inheritance of epigenetic marks by the next generations. However there is growing evidence that environmentally induced epigenetic changes might be transmitted to the next generation (Youngson & Whitelaw, 2008; Miska & Ferguson-Smith, 2016). Transmission of epigenetic traits though maternal germline quite often might be attributed to the in utero environment influences on embryo and primordial germ cells or behavioural differences and have to be persistent through at least three generations to be considered transgenerationally inherited (Blake & Watson, 2016). True transgenerational epigenetic inheritance via paternal transmission can be confirmed when epigenetic traits are maintained beyond two generations (Jimenez-Chillaron et al., 2009; Carone et al., 2010; Blake & Watson,

2016). Examples of paternal and maternal inter- and transgenerational effects in rodents are presented in Table 3. These studies imply the existence of a non-genetic memory of parental exposure to environmental factors that can be maintained during the early development reprogramming wave.

Imprinted genes, transgenes and retroelements provide examples where epigenetic marks can be resistant to reprogramming events. Imprinted genes are expressed according to their parent-of-origin (Ferguson-Smith, 2011). Parental-specific methylation that regulates the expression of imprinted genes is established during gametogenesis and is resistant to reprogramming during preimplantation embryo development in order to maintain the germline-derived methylation memory of parental origin. Mechanisms involved in DNA methylation maintenance at imprinting control regions during embryonic development are emerging. One of them is through recruitment of PGC7 that protects paternal (H19 and Rasgrf1) and maternal (Peg1, Peg3, Peg10) imprints from demethylation (Nakamura et al., 2007). Another involves binding of Zfp57 to imprinting-control regions (ICRs) that, in complex with the Trim28/KAP1 scaffolding protein, recruits a number of epigenetic modifiers required for imprint maintenance after fertilization (Li et al., 2008; Zuo et al., 2012). The IG-DMR, Rasgrf1, Snrpn, Nnat, Zsr1, Peg1, Peg3, and Zac1/Plagl1 imprinting control region methylation is lost in maternal and zygotic Zfp57 mutants (Takahashi et al., 2015).

Imprinting of several transgenes have been reported (Swain et al., 1987; Surani et al., 1988; Chaillet et al., 1991; Sasaki et al., 1991). However, in some cases transgenes were expressed in a parent-of-origin manner different from classical genomic imprinting (Allen et al., 1990; Weichman & Chaillet, 1997; Kearns et al., 2000). For example, the RSVIgmyc transgene is always silenced and highly methylated when it is inherited from mother (Weichman & Chaillet, 1997). However, when the offspring harbours paternally transmitted RSVIgmyc transgene, the degree of methylation and expression of this transgene varies. Interestingly, some of transgenes were shown to exhibit inter-individual epigenetic variation in a genetic-independent manner similar to metastable epialleles (Weichman & Chaillet, 1997; Sutherland et al., 2000; Kearns et al., 2000).

Tabel 3. Non-genetic inter-/transgenerational effects of parental exposures in rodents

Transmission	Environmental exposure (exposed generation)	Inter- and transgenerational effect (generation)	Model	Reference
Paternal transmission	Undernutrition in utero (F1)	Glucose intolerance, obesity, altered expression of metabolic genes (F2)	Mouse	Jimenez-Chillaron et al., 2009; Radford et al., 2014
Paternal transmission to female offspring	Paternal high fat diet (F0)	Impaired glucose–insulin homeostasis (F1)	Rat	Ng et al., 2010
Paternal transmission	Paternal low protein diet (F0)	Elevated hepatic expression lipid and cholesterol biosynthesis (F1)	Mouse	Carone et al., 2010
Maternal transmission to grand-offspring	Maternal low protein diet during pregnancy (F0)	Altered glucose homeostasis (F3)	Rat	Benyshek et al., 2006
Paternal transmission	Paternal food deprivation (F0)	Decrease in average serum glucose (F1)	Mouse	Anderson et al., 2006
Paternal transmission	In utero exposure to viscozilin (F1)	Increased spermatogenic cell apoptosis (F2-F4)	Rat	Anway et al., 2005; Guerrero-Bosagna et al., 2010

All known metastable epialleles are associated with retroviral elements. Endogenous retroviral elements are targets for DNA methylation and silencing histone modifications that ensure their repression in somatic tissues and during development (Walsh et al., 1998; Matsui et al., 2010). Epigenetic marks associated with retrotransposons were shown to be partially resistant to genome-wide reprogramming events creating a possibility for transgenerational epigenetic memory associated with these elements (Lane et al., 2003; Seisenberger et al., 2012; Arand et al., 2015).

Transgenerational effects at metastable epialleles

The inheritance of epigenetic states at metastable epialleles at the two classic models has been studied. Methylation levels of maternal A^{vy} influences the distribution of phenotypes in the progeny (Morgan et al., 1999). Yellow mothers have more yellow pups then pseudoagouti mothers. This maternal effect is not due to an altered intrauterine environment as there was no difference in the phenotype distribution between offspring of a yellow A^{vy} mother and offspring obtained after fertilized oocyte transfer from a yellow A^{vy} mouse to a recipient black mouse (Morgan et al., 1999). No paternal inheritance of phenotype or methylation pattern was observed for the A^{vy} epiallele in a C57BL/6J background. Inheritance of $Axin^{fused}$ phenotype was observed after both maternal and paternal transmission in 129P4Rr/Rk background (Rakyan et al., 2003). Interestingly, A^{vy} also showed paternal inheritance in 129P4Rr/Rk mice, suggesting a strain-specific effect on metastable epiallele-mediated non-genetic inheritance. Intriguingly, studies of DNA methylation dynamics at the A^{vy} locus during preimplantation development suggested complete methylation reprogramming during maternal and paternal transmission (Blewitt et al., 2006). However, A^{vy} methylation levels in sperm reflected the methylation levels in somatic tissues of the yellow and pseudoagouti mice and similar data was described for $Axin^{fused}$ methylation in sperm (Rakyan at al., 2003). Upon paternal transmission there is a dramatic decrease of methylation during the zygote stage. Low methylation levels of A^{vy} epiallele are maintained in blastocysts independent of the methylation levels observed in sperm. Similarly to sperm, oocytes derived from yellow and pseudoagouti mice showed different methylation levels, being low in yellow mice and high in pseudoagouti. However, after fertilization demethylation of the maternally-inherited allele occured more gradually (compared to the paternally inherited epiallele) with complete erasure of methylation being evident at the blastocyst stage. To summarize these findings, metastable epialleles might avoid reprogramming during primordial germ cells development but, unlike the recently described methylomes of IAPs, become reprogrammed during preimplantation development. Complete erasure of DNA methylation by the blastocyst stage suggests that DNA methylation is not the source of epigenetic inheritance observed for A^{vy} and $Axin^{fused}$.

Identification of novel metastable epialleles

The aforementioned unique features of metastable epialleles made them a useful model to study environment-epigenetic-phenotype relationships and transgenerational epigenetic inheritance. To date, it remains unclear how common metastable epialleles in murine genome are. A^{vy} and $Axin^{fused}$ were discovered due to pronounced phenotypes that were caused by IAP-driven alterations in gene expression levels between individual mice. However, the discovery of the $CABP^{IAP}$ metastable epiallele indicated that metastable epialleles might not necessarily have a clear impact on phenotype (Druker *et al*., 2004). $CABP^{IAP}$ was found in an analysis of C57BL/6J cDNAs that contained IAP sequences. This metastable epiallele is a result of a C57BL/6J-specific IAP insertion in the 6th intron of the Cdk5rap1 gene. Consistent with the two previously known metastable epialleles, this IAP can induce expression of downstream exons of the host gene when it is lowly methylated. The expression levels of this aberrant IAP driven transcript correlated with IAP methylation and showed interindividual variation. This finding indicates that metastable epialleles might be extensive in the mammalian genome. However, the total number of murine metastable epialleles is still unclear.

One of the first attempts to identify metastable epialleles genome wide used Affymetrix expression array data to screen for genes with high interindividual and low inter-tissue variation (Weinhouse et al., 2011). This screen identified six candidate metastable epialleles. Despite two candidates showing some interindividual methylation variation, the link between this interindividual variation and gene expression was not established. Moreover, these two candidates showed inconsistent intertissue variation and hardly any methylation change upon bisphenol-A exposure. One further approach that aimed to identify novel metastable epialleles screened for IAPs marked by the active histone modification H3K4me3 in the proximity of the transcriptional start sites of mRNAs (Ekram et al., 2012). 143 ERVs were identified as potential candidates. Only 13 of them were analysed for interindividual methylation variation. Three candidates were successfully validated.

Both of these screens assumed that variable epigenetic states at metastable epialleles have a direct impact on transcriptional events. However, the H3K4me3-based screen might not be optimal since low methylation at the A^{vy} locus

was associated with histone acetylation, but no significant differences in H3K4me3 enrichment was observed between yellow and pseudoagouti mice (Dolinoy et al., 2010).

The most extensive screen, performed by Oey et al. using comparative whole genome bisulfite sequencing (WGBS) data in five individuals, identified around 300 non-repetitive and 55 ERV regions exhibiting inter-individual differential methylation (Oey et al., 2015). This study proved that naturally occurring germline mutations and inter-individual genetic differences do not underlie the epigenetic variation observed at the identified regions. The majority of identified ERVs belong to IAP and MuLV classes. However, only a small number of the identified regions were experimentally analysed for interindividual methylation variation and the outcome was poor. The methylation variation at five validated non-repetitive regions was found to be tissue specific differential methylation and therefore, these regions cannot be referred to as metastable epialleles. The authors speculated that their reported number of candidate metastable ERVs might be twice what they had estimated due to technical limitations of the designed screen. However, this statement was not supported by experimental validation.

To summarise, screens previously conducted to identify murine metastable epialleles found different numbers of metastable epialleles present in the murine genome. Their results varied from six to over a hundred candidates. However, only four common loci were identified in the datasets by Ekram et al and Oey et al. Both studies had very limited experimental validation analysing less than 10% of identified candidates.

Blueprint project

In 2011, a large European epigenome collaboration called BLUEPRINT was funded. The scientists from this project aimed to produce around 100 ex vivo reference epigenomes of purified hematopoietic cells from healthy and diseased individuals. This extensive amount of data promised some understanding how much of epigenetic variation exists between individuals, cell types, health and disease. The

better understanding of disease development and possible treatments are also expected to be outcomes of the project (Abbott, 2011; Adams *et al*., 2012).

One work package on the BLUEPRINT project utilised murine models to explore and quantify genomic/epigenomic variation comparatively. In our laboratory two inbred mouse strains (C57BL/6J and CAST/Eij) are used to address questions about genotype-epigenotype-transcriptome-phenotype interactions in the absence of genetic heterogeneity. More than 17 million SNPs and 86 000 structural variant differences exist between CAST/Eij and published C57BL/6J reference genomes (Keane et al., 2011). The correlation between epigenetic variation and the extensive genetic variation in these two strains can be studied using these animal models. For our aims, two quiescent cell populations were purified: CD4+ CD62L+ CD44low CD25- naïve T cells and CD19+ CD43- B cells. For technical reasons, the purified B cells are a mixture of transitional, follicular and marginal zone B cells. According to our FACS sorting data, the majority of purified cells (~70%) are the follicular type I B cells. Cell cycle analysis confirmed that the purified B cells represent the resting population. Isolated populations from multiple animals were pooled and used to prepare total RNA for sequencing (RNA-seq), DNA for WGBS-seq and oxidative bisulfite sequencing (WGoxBS-seq). Biological and technical replicates were generated. Whole-genome sequencing of bisulfite converted DNA is a widely used method to identify genomic location of 5mC+5hmC at a single-base resolution. DNA treatment with sodium bisulfite results in deamination of cytosine to uracil that is read as thymine during sequencing. However, 5mC and 5hmC are resistant to bisulfite conversion and cannot be discriminated after sequencing both being read as cytosine. Oxidative-bisulfite sequencing involves oxidation of 5hmC prior bisulfite treatment to produces 5fC. 5fC is deaminated during bisulfite treatment and read as thymine upon sequencing. Thus, WGoxBS-seq allows genome-wide mapping of 5mC only (Booth et al., 2013).

B and T cell models

Hematopoietic stem cells (HSCs) are the common progenitors of all hematopoietic lineages including B and T lymphocytes. HSCs retain a self-renewing capacity

through the life of the organism. The pool of HSCs is maintained in the bone marrow of the adult where the majority of HSCs are resting in G0 stage of cell cycle. However ~5-10% of HSCs keep cycling and serve as a source of all hematopoietic cells. HSCs, through intermediate stages, differentiate to lymphoid-primed multipotent progenitor (LMPP) and from them to G/M progenitors (GMP), common lymphoid progenitors (CLP) and early thymic progenitors (ETP) (Dias *et al*., 2008; Matthias & Rolink, 2005).

T cell development happens in thymus. Early thymic progenitors (double negative stage) migrate to thymus where they become gradually restricted to the T-cell differentiation program. Double negative progenitors have the potential to differentiate into other cell types (dendritic cells, NK cells, macrophages) until rearrangements of γ, δ, β chains of the T-cell receptor complex occurs. After final rearrangement of the α chain, a functional TCR complex is expressed. This stage of T-cell differentiation is called the double positive stage. Then, cells undergo two rounds of selection and finally differentiate into CD4+ or CD8+ single positive naïve T cells (Koch & Radtke, 2011; Rothenberg, 2012).

CLPs are differentiated from LMPPs and can be characterized by expression of IL-7Rα, EBF1 transcription factor and E2A. They retain the potential to develop into Natural Killer and T cells. CLPs finally become finally primed to the B cell lineage when the expression levels of Pax5 increase. This happens during differentiation of CLPs into pre-pro-B cells. A complex transcriptional network that consists of EBF1; PU.1; E2A; Pax5; Stat5 and other transcription factors acts to activate the B-cell lineage-specific program and finally prime cells towards B-cell development (Dias *et al*., 2008). The further differentiation involves changes in gene expression and gradual rearrangement of light and heavy chains of B-cell receptor complex. During these rearrangements cells are checking for autoreactivity and finally differentiate into IgM+ IgD- naïve immature B cells (Rolink & Melchers, 1996; Osmond *et al*., 1998; Matthias & Rolink, 2005; Vaughan *et al*., 2011).

Two cellular pathways that lead to the generation of naive mature B cells exist. Around 25% of newly formed immature B cells do not leave the bone marrow and differentiate into CD23+, IgD-high T2-like cells. These cells can further differentiate into follicular type I and type II B cells that comprise the “perisinusoidal niche”

(Cariappa *et al*., 2007; Allman & Pillai, 2008). The majority of immature B cells after passing selection against autoreactivity, differentiate into transitional type I (T1) B cells. T1 B cells enter spleen and differentiate further to T2 and T3 B cells. Unlike mature populations of B cells transitional B cells keep expressing AA4 marker and have low expression of CD21. Transitional B cells give rise to two splenic subsets of mature naïve B cells: follicular type I B cells and Marginal zone (MZ) B cells (Pillai & Cariappa, 2009; Allman & Pillai, 2008).

Pure naïve B and T cell populations can be isolated in usable numbers ex vivo using cell specific markers and hence can be used to study genotype-epigenotype relations. Representing pure non-cycling populations, these cell models avoid the main sources of intrinsic epigenetic variation.

Research questions and experimental aims

My thesis work focused on two goals – the first involved validation of the BLUEPRINT datasets to be utilised for the second part. The second goal involved a genome-wide screen for alleles that were subjected to variable methylation between individuals – so-called 'metastable epiallelles'.

Part 1. Validation of BLUEPRINT datasets and manual curation of cell type specific differentially methylated regions (DMRs).

My project relied on WGBS-seq and RNA-seq datasets generated for the BLUEPRINT project by my postdoctoral colleagues, Dr. Sjoberg and Dr. Walker. These datasets were aligned to the mm10 mouse genome to visualize DNA methylation and transcription in the WashU Epigenome Browser. As a starting point for my project and that of others in the group, it is important to conduct experimental validation of these datasets. 5 imprinted regions and 20 genomic regions were selected for pyrosequencing analysis to confirm the bioinformatics on the WGBS data. Expression of 25 genes was experimentally checked to validate RNA-seq datasets.

Identification of DMRs between B and T cells is a key step in WGBS-seq data analysis as it provides insight into differential roles for DNA methylation in B and T

cell development and transcriptional regulation within these cell types. Because of the large amount of data it is appropriate that comparison and analysis of the results is done bioinformatically. At the same time most of bioinformatics tools are based on generalising and simplifying real biological models. The output of such analysis can quite often be compromised and contain false positive results. To optimize and improve bioinformatics tools, I conducted manual curation of a partial dataset to check the bioinformatics output. While optimization of bioinformatic analysis takes time, the characterization of manually identified DMRs provides preliminary data to underpin future experiments.

Part 2. Genome-wide identification of novel metastable epialleles.

$Axin^{fused}$ and A^{vy} metastable epialleles are two classic examples of alleles subject to variable DNA methylation between individuals. They have been identified due to their clear phenotypic outcomes subsequently explained by variable IAP methylation. Both are results of spontaneous mutation caused by IAP transposition events. The more recent discovery of $CABP^{IAP}$ suggests that naturally existing polymorphic IAP insertions can be a subject to interindividual DNA methylation and have impact on transcription without distinctive phenotypic variation as an outcome. Hence, the extent to which metastable epialleles are present in the murine genome is still not clear. Genome-wide identification of metastable epialleles should help to understand the mechanisms of establishment of variable methylation at these regions, their impact on mouse phenotype, and mechanisms of ERV regulation and silencing in mouse.

The project strategy is illustrated on figure 1. Our aim was to determine methylome parameters that would identify a metastable epiallele from WGBS datasets making the assumption that these alleles were associated with variable methylation at endogenous retroviruses. Our experimental design consisted of two steps. Step 1 represented a limited screen of polymorphic IAPs between the two strains. A list of polymorphic retroelements between 18 inbred mouse strains has been published by Nellaker et al. We hypothesised that overlap between C57BL/6J specific ERVs and genes that are differentially expressed between C57BL/6J and CAST/Eij might be expected to identify a set of ERVs with features similar to metastable epialleles.

Step 2 involved the generation of a model using methylation and/or transcription features of metastable ERVs identified during Step 1 to conduct a genome-wide unbiased screen for ERV-derived metastable epialleles.

The final set of metastable epiallele candidates provided a dataset to explore the mechanisms and features that might underlie inter-individual methylation variation between genetically identical individuals and the global impact of metastable epialleles on phenotype.

Part 3. Characterization of IAP impact on transcription

Transposable elements are often considered to be “junk” DNA. However, evidence is accumulating to suggest that these elements might provide alternative promoters, poly-A sites, splicing sites, enhancers and transcription factor binding sites to regulate the expression of host genes.

IAP elements are one of the most active ERV class in mouse and some impact on host gene expression has been attributed to them. Most IAP-driven transcription has been reported in tissues with low global methylation levels, however, the extent to which IAP regulatory elements are involved in the regulation of transcription in somatic tissues is not clear. We used RNA-seq data generated for the BLUEPRINT project to screen for IAPs that can initiate and/or terminate transcription in B and T cells. The available methylation data allows an exploration of the role of DNA methylation in these processes. While IAPs have been mostly considered to provide alternative promoters, we have also analysed the use of alternative splicing sites provided by this group of retroelements.

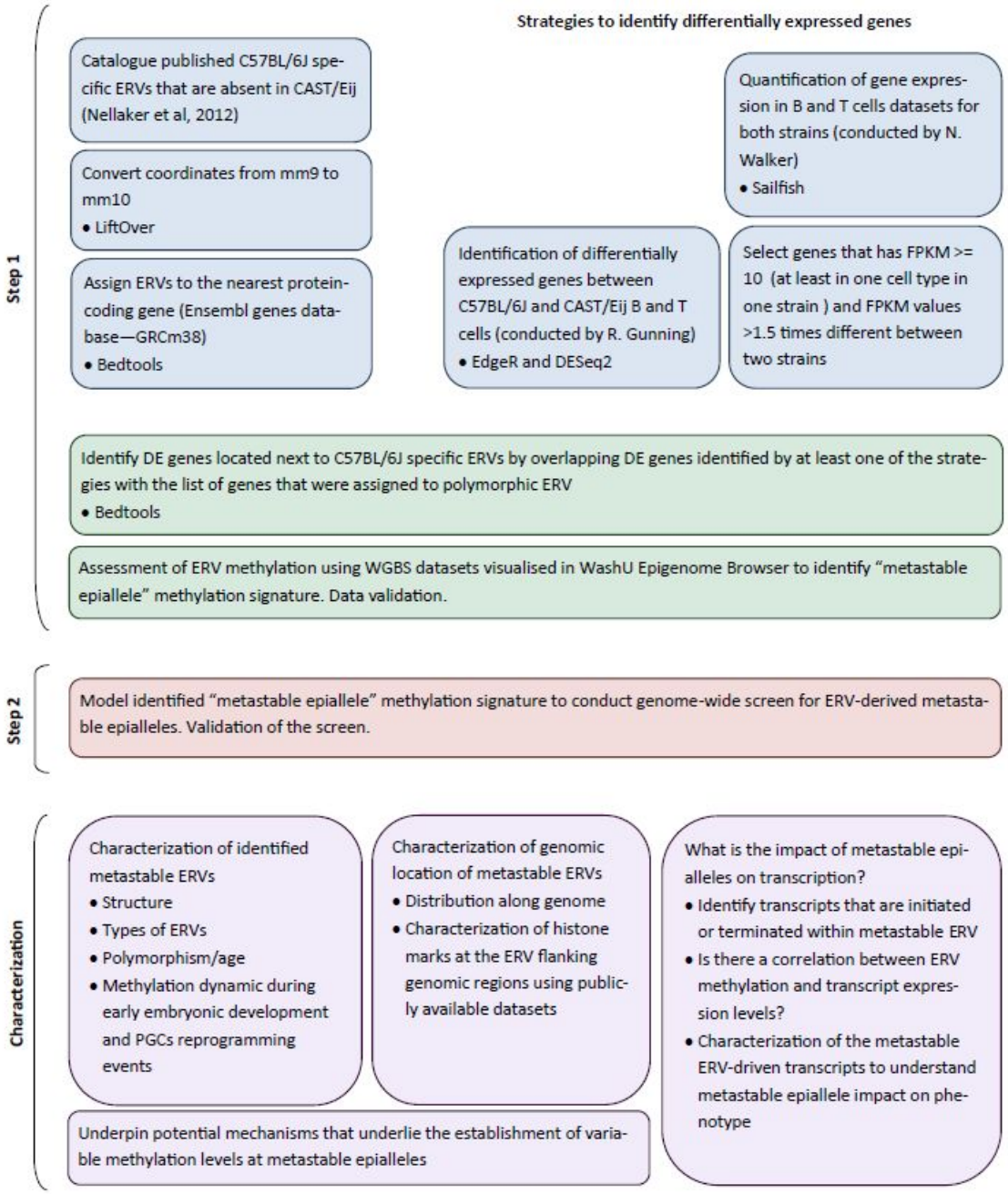


Figure 1.1 Project outline. *Explanation of genome-wide screen strategy for identification of metastable epialleles and summary of the addressed questions.*

Chapter 2

Materials and Methods

Tissues and cells

Following dissection, somatic C57BL/6J ana CAST/Eij tissues were snap frozen in liquid nitrogen and manually pulverized. B cells were isolated from fresh splenic tissues using the B Cell Isolation Kit (Miltenyi Biotec). All mouse work was carried out in accordance with UK government Home Office licensing procedures (HO project licence number: PC9886123).

DNA/RNA purification and conversion

Tissues were dissected from inbred C57BL/6J and CAST/Eij mice. 30-35 ug of tissue sample was used for simultaneous purification of genomic DNA and total RNA using the AllPrep DNA/RNA Mini Kit (QIAGEN). During purification, RNA was treated with DNaseI using the RNase-Free DNase Set (QIAGEN). B cell DNA was purified using standard phenol-chloroform extraction protocols (Maniatis et al., 1982)

Bisulfite conversion

Bisulfite conversion and post-modification clean-up of DNA samples was performed with the Imprint DNA Modification Kit (Sigma). The samples were treated according two-step modification protocol provided by the manufacturer.

Pyrosequencing

Methylation quantification was carried out by pyrosequencing. The assays and primers were designed using PyroMark Assay Design Software and provided in Supplementary Table 1. The annealing temperature for PCR primers was optimized by gradient PCR. The PCR conditions were the following: 1) 95°C – 5 min; 2) 94°C – 30 sec, optimized t°C – 30 sec, 72°C – 55sec , 40 cycles; 3) 72°C – 5 min. The PCR product was shaken with streptavidin sepharose high performance beads (GE healthcare) dissolved in binding buffer for 20 min.Purification of the biotinylated strand was done using the PyroMark vacuum workstation. Sequencing primers were

dissolved in annealing buffer and combined with the purified product in a PSQ plate, followed by incubation at 85°C for 3 min. The plates were centrifuged for 4 minutes at 2500 rpm and loaded onto the PyroMark™ MD pyrosequencer (Biotage) Sequencing was carried out using PyroMark Gold Q96 SQA Reagents (QIAGEN) and methylation was quantified using Pyro Q-CpG 1.0.9 software (Biotage)

cDNA synthesis

DNAseI treated RNA was used for cDNA synthesis. Thermo Scientific RevertAid H Minus First Strand cDNA Synthesis kit was used. The reactions were performed according to the manufacturer's protocol.

Q-PCR

Q-PCR primers were designed using Primer3 software (Untergasser et al., 2012) and are provided in Supplementary Table 2. cDNA was amplified using the LightCycler 480 SYBR Green I Master mix (Roche). Samples were run on LightCycler 480 Instrument (Roche). The PCR conditions were the following: 1) 95°C – 5 min; 2) 95°C – 10 sec, 60°C – 10 sec, 72°C – 10sec , 45 cycles; 3) 95°C - 5 sec, 65°C - 1 min, 97°C - continuous ; 4) 40°C – 30 sec. Relative cDNA abundance was calculated using the ΔCT method and normalized to the expression of housekeeping gene Actin. Expression of Tcf7 gene was normalized to the expression of 18S RNA.

PCR to identify Dlk1 isoforms

B and T cell cDNA was used as a PCR template. Forward 5'-CTGCACACCTGGGTTCTCTG-3' and reverse 5'-ATCACCAGCCTCCTTGTTGA-3' Dlk1 primers were used. Depending on the expressed isoforms these primers would implify fragment of a different size:

ENSMUST00000056110	Dlk1-001	805bp	385aa	secreted
ENSMUST00000109844	Dlk1-002	805bp	385 aa	secreted
ENSMUST00000109842	Dlk1-008	586bp	312aa	membrane-bound
ENSMUST00000109843	Dlk1-003	520bp	290aa	membrane-bound
ENSMUST00000109846	Dlk1-005	652bp	334aa	secreted

The PCR conditions were the following: 1) 95°C – 5 min; 2) 94°C – 30 sec, 60°C – 30 sec, 72°C – 55sec ,_40 cycles; 3) 72°C – 5 min. 1.5% agarose gel was used to assess product size.

Manual curation of DMRs

WGBS-seq datasets were visualized in the WashU Epigenome Browser (http://epigenomegateway.wustl.edu/) systematically visually comparing WGBS data for B and T cells from C57BL/6J to identify differentially methylated regions. The majority of CpGs in these non-cycling cells were 100% methylated. Chromosome ideograms were built using Idiographica web server (http://www.ncrna.org/idiographica). The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used for gene ontology analysis (https://david.ncifcrf.gov/)

Biased screen of polymorphic ERVs and identification "metastable epiallele" methylation signature

Genomic coordinates of C57BL/6J specific ERVs that are absent from the CAST/Eij genome were extracted from the published list of polymorphic ERVs (Nellaker et al., 2012). ERV coordinates were converted from mm9 to mm10 mouse genome assembly using LiftOver and assigned to the nearest protein-coding gene from the Ensembl gene database (GRCm38) using Bedtools (Quinlan & Hall, 2010).

Genes differentially expressed between C57BL/6J and CAST/Eij were identified using two strategies. The first one relied on bioinformatic analysis. DESeq2 and edgeR were used to identify DE genes for B and T cell samples (conducted by R. Gunning). Due to the absence of strain-specific annotation of ncRNAs, they were removed from the analysis. Significant hits from both programs were used to compile the list of differentially expressed genes between the two strains. The second strategy used FPKM gene expression levels calculated by Sailfish software (conducted by N. Walker). Genes that had FPKM >= 10 in at least one cell type of one strain were selected. Out of these lists, I have extracted genes that had >1.5 difference of expression between two strains within the same cell type.

DE genes identified by at least on of the described methods were overlapped with genes containing polymorphic ERV insertion nearby or within. The overlapping genes

provided a list of metastable ERV candidates for visual assessment of their methylation levels. Validation of interindividual methylation variation of identified candidates was carried out by pyrosequencing.

Assembly of ERV coordinates

The RepeatMasker database was downloaded through the UCSC Table Browser to determine genomic coordinates of ERV fragments (Karolchik et al., 2004). ERV fragments were separated into 4 groups according to RepeatMasker annotation: ERV1, ERVL, ERVK and IAPs. The fragments were assembled using Bedtools.

For the IAP group, IAP fragments were considered to be part of an insert if they were no longer than 150 bp apart and located on the same strand:

bedtools merge -s -d 150 -i <iap_fragments_coordinates.bed> > <iap_coordinates.bed>

For assembly of ERV1, ERVK and ERVL groups 100 bp distance between fragments was allowed, fragments should have same orientation:

bedtools merge -s -d 100 -i <fragments_coordinates.bed> > <ERV_coordinates.bed>

Structure of ERV insertions was determined based on the annotation of the fragments belonged to the insertion.

Genome-wide screen for metastable epialleles

For the purpose of the genome-wide screen WGBS-seq and WGoxBS-seq generated for B and T cells (16 datasets in total) were treated as biological replicates. Bedgraph files were used to extract methylation levels of ERV CpGs:

bedtools map -a <ERV_coordinates.sorted.bed> -b <bedgraph file WGBS or WGoxBS> -c 4 -o collapse

Methylation of a CpG site is represented by two numbers reflecting methylation levels at sense and antisense strands. Average methylation for 16 methylation values representing 8 distal CpG from the 5' and 3' ends of an ERV was calculated for each biological replicate to estimate average ERV methylation in this replicate. To determine methylation variation at an ERV across 16 biological replicates, the average methylation levels were sorted and the difference between second highest and second lowest values was used as a computational score for methylation

variation in our datasets (**Figure 2.1**). Analysis for methylation variation at 5' and 3' ERV ends was done separately and the results were overlapped. The cut-off for a final list of metastable epiallele candidates was determined by experimental assessment of the interindividual methylation variation at a selection of regions with different computational methylation variation scores (**Supplementary table 3**). Regions that were differentially methylated between B and T cells were excluded from the final list of variably methylated regions since these are categorised as cell type specific DMRs and hence do not fulfil the criteria of methylation consistency between tissues. The region was considered to be differentially methylated between B and T cells if 8 highest and/or 8 lowest average methylation levels were coming from same cell type replicates and the computational methylation score at this region was higher than the selected cut-off.

Characterization of metastable epiallele candidates

Catalogue of structural variants across 18 inbred mouse strains generated for Mouse Genomes Project (http://www.sanger.ac.uk/sanger/Mouse_SnpViewer/rel-1505) was used to quantify polymorphism of identified metastable epiallele candidates (Keane et al., 2011; Yalcin et al., 2011).

A neighbour-joining tree for IAPLTR1_Mm containing IAPs was built using Geneious 9.0.5 software using default parameters (http://www.geneious.com/; Faulk et al., 2013). IAP sequences were downloaded from UCSC Table Browser. "+" strand sequence was used for antisense oriented IAPs.

Publicly available ENCODE datasets were used for histone modification and CTCF binding analysis. Signal p-value bigwig files were downloaded and analysed using Galaxy deepTools (https://usegalaxy.org/; Ramírez et al., 2014).

Identification of IAPs capable of transcription initiation and termination

De novo transcriptomes were assembled using StringTie 1.3.3 software (conducted by J. Gardner). 12 RNA-seq datasets were used for de novo transcriptome assembly: 3 B cell female replicates, 3 B cell male replicates, 3 T cell female replicates, 3 T cell male replicates. Using Bedtools software, coordinates of identified transcripts were overlapped with IAP coordinates to identify transcripts initiated, terminated or spliced within IAPs (**Figure 2.2**). Only IAPs that consistently overlap transcripts within at

least 3 biological replicates representing the same cell type and sex were further analysed.

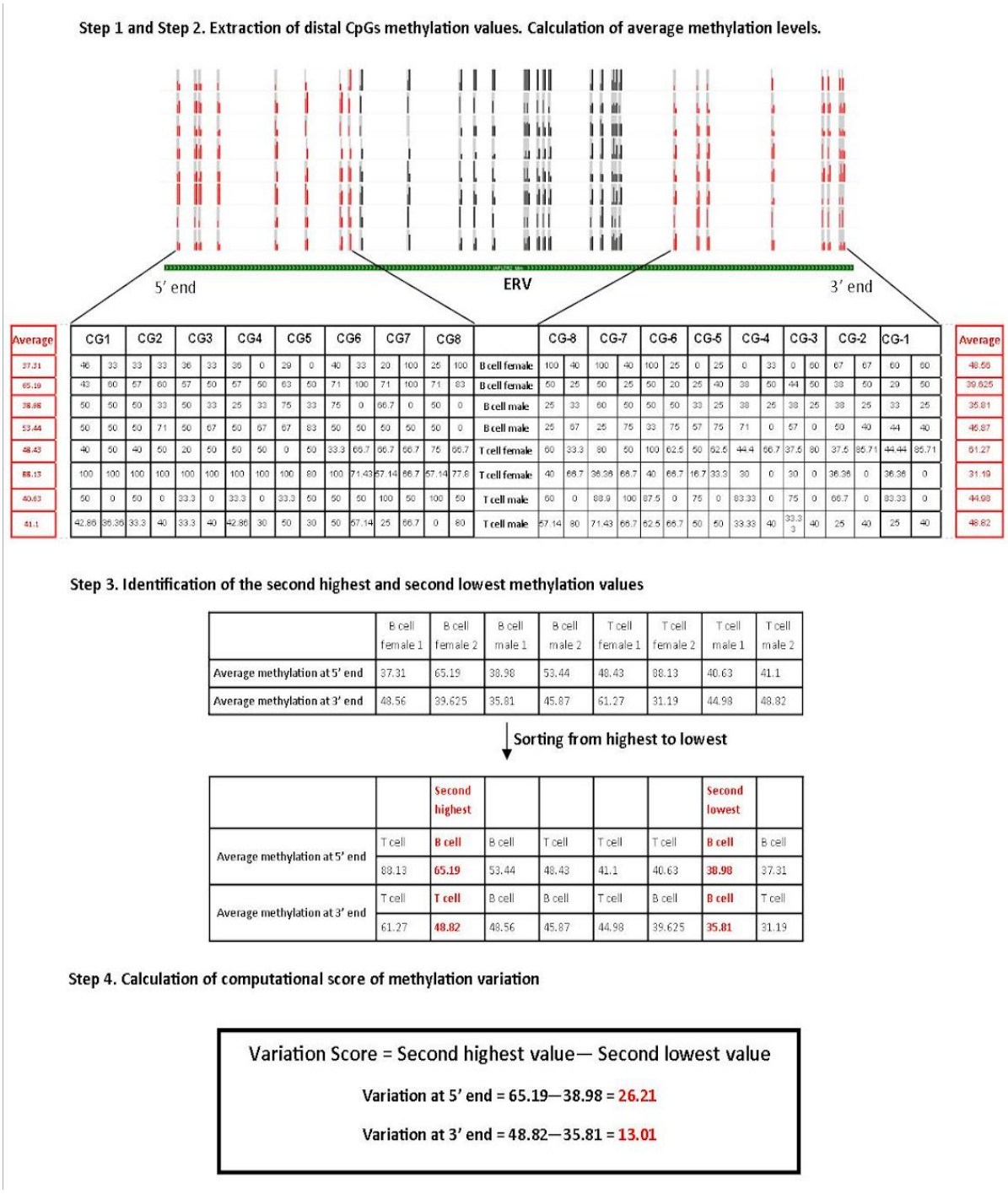


Figure 2.1 Simplified schematic representation example of the genome-wide screening strategy to identify variably methylated ERVs. *Methylation values of 8 distal CpGs from 5' and 3' IAP ends were extracted from WGBS-seq datasets. Average methylation of 8 distal CpGs in each biological replicate was quantified. Average methylation levels from all replicates were sorted from highest to lowest. Second highest and second lowest values were used to quantify variation between biological replicates. Analysis of methylation variation was done separately for at 5' and 3' IAP ends*

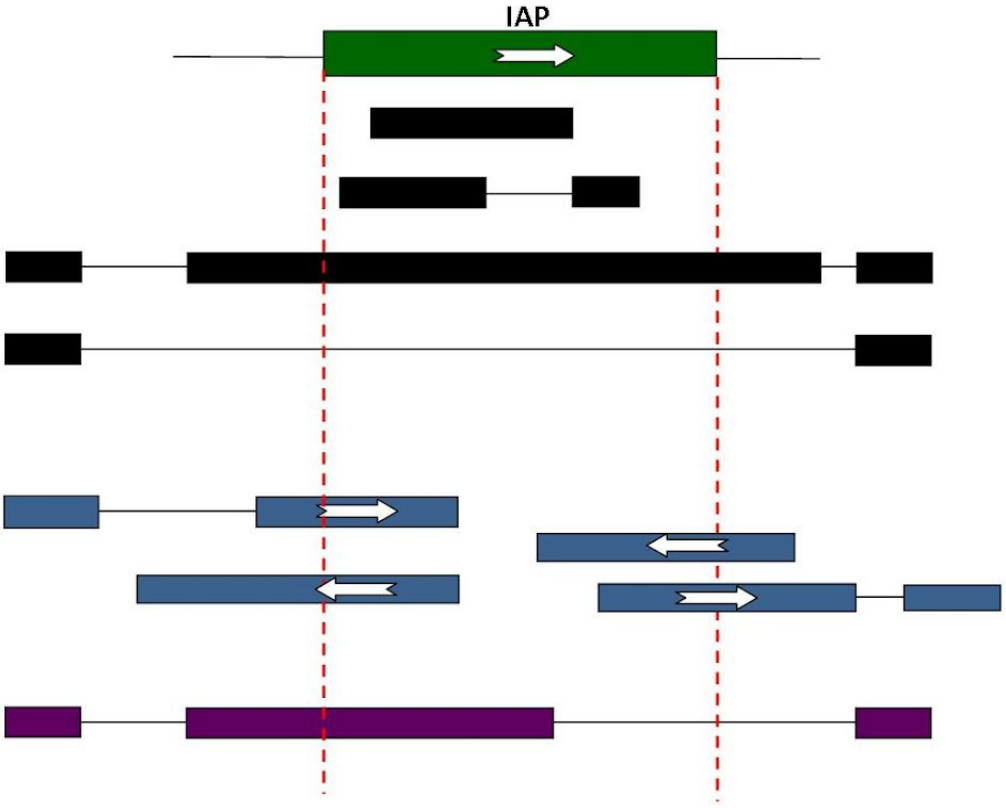


Figure 2.2 Schematic representation of the identification of transcripts initiated or terminated within IAP (colored in blue) and transcripts spliced within IAP (colored in purple). *Transcripts that might overlap IAPs representing either IAP transcription or IAP exonic inclusion are colored in black. The screen was performed by intersecting IAP coordinates with transcript coordinates using multiple biological replicates of RNA-seq data*

Chapter 3

Validation of BLUEPRINT datasets

Validation of BLUEPRINT datasets for WGBS and RNA-seq

My project is based on the use of WGBS-seq and RNA-seq data that have been previously generated in our lab. Sequencing data was aligned to the mm10 mouse genome and visualized in the WashU Epigenome browser (http://epigenomegateway.wustl.edu/). The obtained bioinformatics outcomes were experimentally validated. For WGBS-seq data validation, 7 imprinted regions and 20 genomic regions that are differentially methylated between B and T cells and include both intragenic and intergenic regions from 11 different chromosomes were selected. Their methylation levels were analysed by pyrosequencing and compared to WGBS-seq data. DNA samples used in the experiments were purified by M.Sjoberg from B and T cells pooled from 4-6 mice. Pyrosequencing results correlate with the methylation levels seen in the browser (**Figure 3.1, 3.2**) with imprinted regions generally showing around 50% methylation as expected, and the selected hypo and hypermethylated regions validating.

Interestingly, the IG-DMR imprinting control locus on mouse chromosome 12 that is expected to have around 50% methylation and was shown to be hypermethylated (75-95%) in C57BL/6J by pyrosequencing. High methylation levels of the IG-DMR were observed bioinformatically as well. To see whether such hypermethylation is a common feature of B and T cells, and/or is dependent on genetic background 6 imprinted regions were pyrosequenced using B and T samples from CAST/Eij. All of them showed the expected 40-60% methylation apart from IG-DMR (**Figure 3.3**). However CAST/Eij IG-DMR methylation was lower than in C57BL/6J. The data indicate that the methylation imprint at the IG-DMR is lost from these cell types.

The normalization of RNA-seq data was done using the Sailfish bioinformatics tool (conducted by N.Walker; Patro et al., 2014). For each transcript, its Transcript per Million (TPM) value was calculated. TPM values represent normalized RNA-seq data (Wagner et al., 2012). Q-PCR was used to validate the relative expression of 25

genes between B and T cells of C57BL/6J. The results were compared to normalized RNA-seq data (**Table 4**) and confirmed the accuracy of conducted RNA-seq analysis.

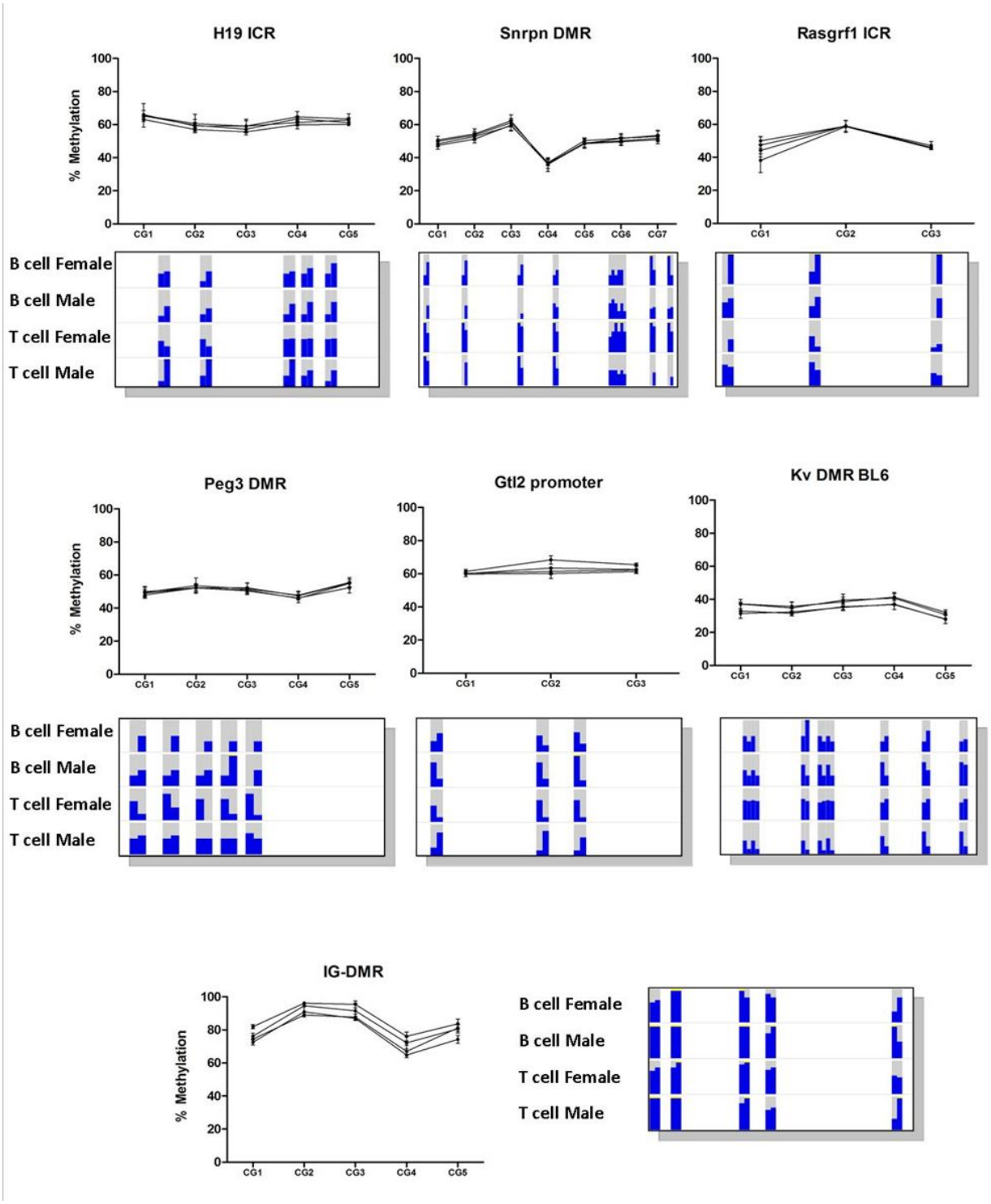


Figure 3.1 Validation of imprinting regions in C57BL/6J. *Methylation at imprinting regions was analysed by pyrosequencing. Four samples were used: B cell male, B cell female, T cell male and T cell female. The screenshots represent methylation levels of corresponding CpGs in WGBS-seq datasets*

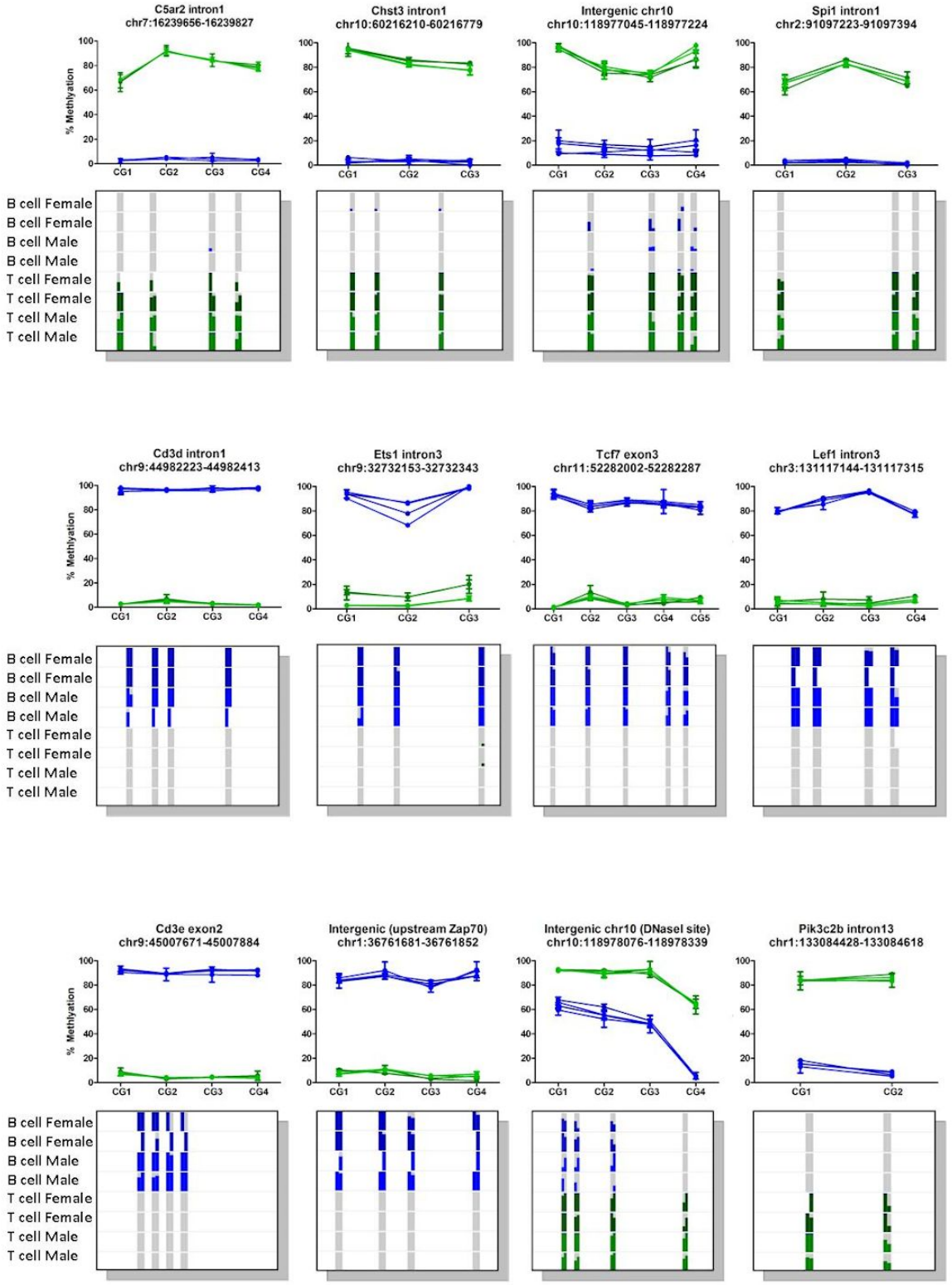


Figure 3.2 Continued next page

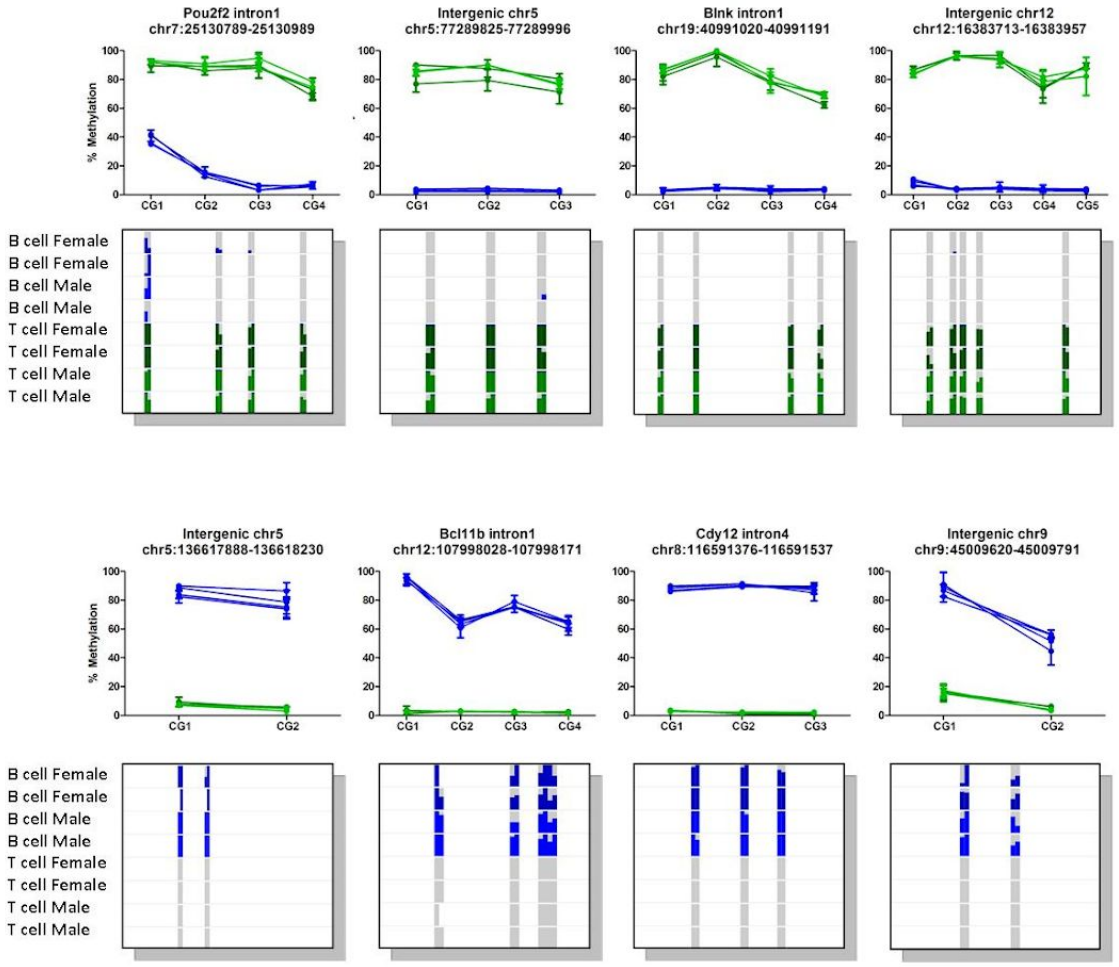


Figure 3.2. Validation of genomic regions differentially methylated between B and T cell. *Two B cell male and two B cell female samples are highlighted in blue. Two T cell male and two T cell female samples are highlighted in green. The screenshots represent methylation levels of corresponding CpGs in WGBS-seq datasets*

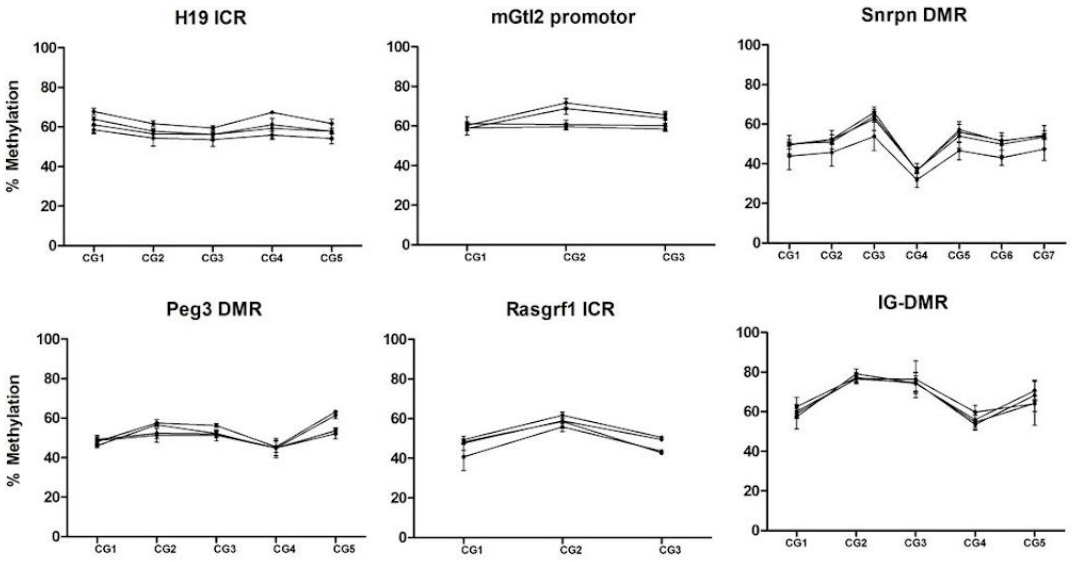


Figure 3.3. Validation of imprinting regions in CAST/EiJ. *Four samples were used: B cell male, B cell female, T cell male and T cell female*

Table 4. RNA-seq validation. *Average TPM value was calculated for six biological replicates (both male and female). dCt values are the average values for two biological replicates (each has three technical repeats). Tcf7 expression was normalized to 18S RNA gene expression. Expression of other genes was normilized to Actin gene expression.*

Gene	TPM B cells	TPM T cells	TPM(B) vs. TPM(T)	dCt values (B cell/T cell)	Fold change (B vs. T)
Ets1	179.17	230.33	0.78	3.65 / 3.29	0.78
Spi1	32.83	0.13	252.54	7.26 / -	NA
Bcl11b	0	19.67	NA	- / 6.48	NA
Cd2	91.68	134.86	0.68	4.91 / 4.43	0.72
Cd3d	0.83	93.24	0.0089	- / 5.8	NA
Cflar	12.35	17.5	0.71	8.32 / 7.65	0.63
Chst3	65.37	0.14	466.93	5.88 / -	NA
Elf1	56.75	46.14	1.23	6.26 / 6.5	1.18
Iscu	13.2	11.33	1.17	9.04 / 9.06	1.02
Itk	0.31	69.11	0.005	- / 4.46	NA
Pik3c2b	17.46	0.37	45.21	7.29 / 12.89	48.56
Rap1b	97.68	99.43	0.98	5.07 / 5.21	1.01
Rhoh	23.02	23.15	0.99	7.11 / 7.04	0.95
Ski	5.98	13.7	0.44	12.94 / 10.94	0.25
Stat5b	14.03	21.97	0.64	6.93 / 6.36	0.67
Tcf7*	0.43	182.43	0.002	- / 12.33	NA
Stk17b	176.34	106.03	1.66	7.72 / 8.78	2.09
Ugcg	20.22	39.64	0.51	8.35 / 7.48	0.55
Ptprc	227.09	224.9	1.01	5.62 / 6.35	1.65
Nfkbia	79.24	44.14	1.79	6.24 / 7.01	1.7
Smim14	15.53	3.43	4.53	6.11 / 8.48	5.16
Dgka	36.27	163.33	0.22	5.98 / 3.75	0.21
Fam105a	3.35	27.21	0.12	10.71 / 7.42	0.1
Arpc2	36.65	26.1	1.41	5.48/5.76	1.21
Hvcn1	26456.9	466.95	56.66	4.79/10.62	56.89
Selplg	1606.24	10979.59	0.15	8.45/5.79	0.16
Laptm5	134.47	122.82	1.09	2.72/2.55	0.89

Is there selective absence of imprinting in hematopoietic cell types?

During pyrosequencing validation, hypermethylation of the IG-DMR imprinting control region was found. The IG-DMR locus is located 70 kb downstream of the Dlk1 gene. This region is normally paternally methylated. Deletion of the IG-DMR from the unmethylated maternal copy causes the biallelic expression of Dlk1 and the other two main protein coding genes in the cluster Rtl1 and Dio3 that are normally maternally repressed (Georgiades et al., 2000). Furthermore, paternal uniparental disomy for chromosome 12 exhibits biallelic methylation at the locus and similarly results in biallelic expression of the protein coding genes (Georgiades et al., 2000). A role for Dlk1 in lymphocyte development has been suggested (Sakajiri et al., 2005; Raghunandan et al., 2008). It has been shown previously that in $Dlk1^{-/-}$ mice, the ratios between different splenic B cell subsets were altered. Marginal zone B cells numbers were increased in $Dlk1^{-/-}$ mutants in comparison with wild-type. At the same time, a decrease in numbers of follicular B cells was observed in Dlk1-deficient mice. This suggests that Dlk1 is important for proper B cell differentiation and cell fate decisions. Compared to C57BL/6J, the methylation levels of the IG-DMR in CAST/Eij T and B cells appeared to be ~20% less (**Figures 3.1, 3.3**). The analysis of IG-DMR methylation in B cells purified from C57BL/6JxCAST/Eij reciprocal hybrids showed the inheritance of this hypermethylation in a parent-of-origin depending manner (**Figure 3.4A**). Hybrids from CAST/Eij mothers x C57BL/6J fathers showed lower hypermethylation levels comparing to hybrids with C57BL/6J mothers and CAST/Eij fathers. DNA from mice with paternal or maternal uniparental disomy (pUPD and mUPD) was used for controls (Georgiades et al., 2000). 50% methylation control was made by mixing equal amounts of pUPD and mUPD DNA. Controls were used to confirm that the observed IG-DMR hypermethylation is not a technical artefact and methylation quantification was accurate. These results confirm the loss of DNA methylation imprinting in B and T cells.

Selective functionally important absence of imprinting of Dlk1 was previously reported for postnatal neural stem cells and niche astrocytes (Ferron et al., 2011). In these cells, hypermethylation of the IG-DMR correlates with biallelic expression of Dlk1 and may contribute to the regulation of this Dlk1 dosage increase. However, hypermethylation might also be a secondary consequence of biallelic Dlk1. I first

analyzed the expression of Dlk1 in B and T cells in C57BL/6J background. Dlk1 expression was found in B cells but no Dlk1 expression in T cells was detected (**Figure 3.4B**). Dlk1 expression was detected in B cells purified from C57BL/6J and CAST/Eij hybrids (**Figure 3.4C**). To assess whether Dlk1 is biallelically expressed in B cells, pyrosequencing analysis using reciprocal hybrids cDNA was conducted. The maternally and paternally expressed Dlk1 transcripts were distinguished based on SNPs between C57BL/6J and CAST/Eij using established assays (Ferron et al, 2011). Biallelic expression of Dlk1 was observed in B cells in reciprocal hybrids (**Figure 3.4D**).

Dlk1 protein (Delta Like Non-Canonical Notch Ligand 1) is a member of epidermal growth factor-like family and is a vertebrate specific atypical Notch ligand. It has been suggested to act as Notch-signaling antagonist (Nueda et al., 2007, Bray et al., 2008). Notch signaling is involved in regulation of lymphopoiesis at multiple stages. Notch activity is required for common lymphoid progenitor commitment to T cell development over B cell. Apart from this, it directs development of CD8+ T cells from CD4+CD8+ T cells and supports differentiation of mature B cells into marginal zone B cells (Robey, 1999; He and Pear, 2003). There are several Dlk1 transcript isoforms that produce either membrane-bound or secreted Dlk1 proteins. Studies in Drosophila showed that membrane-bound Dlk1 has a more profound effect on Notch signalling than the secreted form (Bray et al., 2008) and in the mouse postnatal neurogenic niche, the secreted form functions in a Notch independent manner (Ferron et al., 2011). To determine what isoforms are expressed in naive B cells, a PCR assay was performed. Dlk1 primers specifically amplified fragments of different size depending on the isoform were used (Ferron et al., 2011). Two secreted isoforms would produce bands 805 and 652 bp long; two membrane-bound isoforms, 586 and 520 bp. The assay confirmed the presence of two isoforms of Dlk1 (secreted isoform - 805bp and membrane-bound isoform - 586 bp) in B cells in C57BL/6J and hybrids (**Figure 3.4E**).

These results suggest that Dlk1 is expressed in B cell but not in T cells and is represented in both secreted and membrane bound forms in B cells.

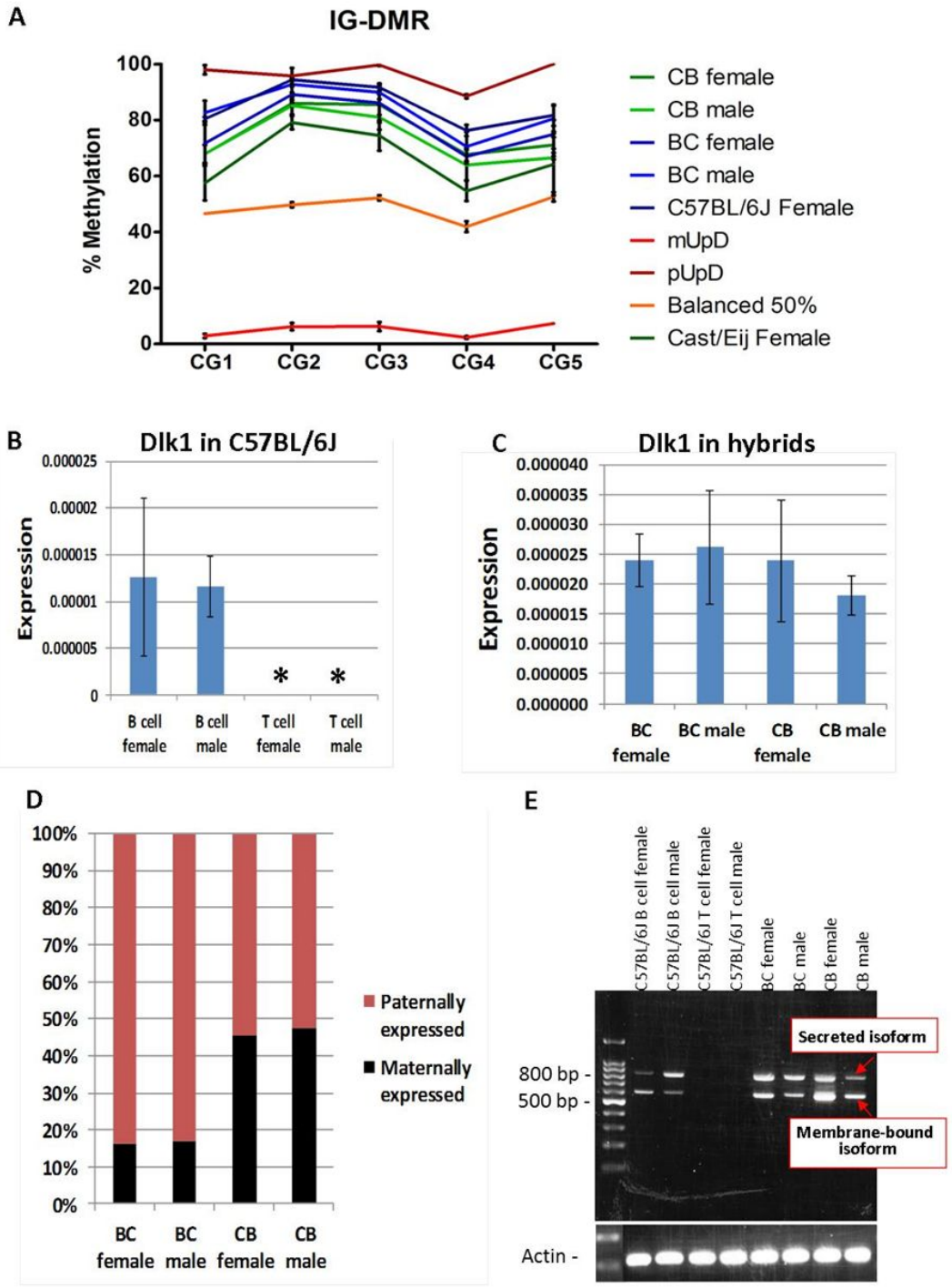


Figure 3.4. Loss of Dlk1 imprinting in B and T cells. *A) IG-DMR methylation in CAST/EijxC57BL/6J (CB) and C57BL/6JxCAST/Eij (BC) hybrid B cells. mUPD, pUPD and balanced 50% methylation samples are used as controls; B) Dlk1 expression in B and T cells in C57BL/6J mice. T cells do not express Dlk1 despite also exhibiting hypermethylation at the IG-DMR; C) Dlk1 expression in hybrid B cells.; D) Biallelic expression of Dlk1 in hybrid B cells; E) Dlk1 membrane bound and secreted isoforms are expressed in B cells in hybrids and C57BL/6J*

Identification of cell type specific differentially methylated regions (DMRs)

Validated WGBS data was used to find DMRs between T and B cells of C57BL/6J. To date, a number of software tools are available for computational identification of DMRs (Chen et al., 2016). Comprehensive identification of DMRs often relies on the implementation of different methods with subsequent overlap of the results. The majority of the methods rely on prior knowledge of certain DMR criteria including CpG density, region length, coverage, etc (Chen et al., 2016). Computational identification of DMRs between C57BL/6J B and T cells was conducted by N. Walker. He applied a combination of methods to analyse the data including MethylSeekr, MOABS and Methylpipe (Burger et al., 2013; Sun et al., 2014). In addition, he designed his own pipeline "GeneTiles" for DMR identification. MethylSeekr, MOABS and Methylpipe use Hidden-Markov Model to identify low methylated regions (LMRs) since these are the most prone to functional variation in the mammalian genome (Stadler et al., 2011). By comparing identified LMRs in T and B cells, DMRs could then be identified in an unbiased fashion. GeneTiles software breaks the genome into tiles. Methylation values are calculated for each tile. All tiles together with methylation values are used to create a "profile" to visualise what fraction of the gene, feature or region is hypermethylated or hypomethylated ultimately to generate a heatmap of gene categories with different methylation patterns that can be compared between the two cell types.

To reaffirm the results of the bioinformatic screen, I independently identified DMRs by systematically curating the mapped methylation data for individual chromosomes in the browser. Since I had manually scored 10 chromosomes for intragenic DMRs (**Figure 3.5**), I was able to identify 500 DMRs. Twenty of the identified DMRs were experimentally validated (**Figure 3.2**). To assess DMR conservation between mouse strains, 8 selected potential DMRs were analysed in CAST/Eij B and T cells by pyrosequencing (**Figure 3.6A**). Only one region (Cdyl2) was shown not to be DMR in CAST/Eij hence not conserved with C57BL/6J. All other DMRs were present in the CAST/Eij genome. Such conservation of the DMRs between two strains suggests functional importance.

I anticipated that bioinformatic tools would be more successful in the identification of small DMRs than my manual curation however the ability to test modified informatics

tools on my curated dataset contributed to the optimisation of the computational CpG methylation tools. The differences between single CpGs methylation are harder to pick up during visual screen. At the same time, a manual screen is a reliable method for identification of large DMRs. Computational methods define large DMRs by merging identified small DMRs located close to each other. Identification of large DMRs by bioinformatic tools is complicated as it can misinterpret the presence of single highly methylated CpGs, low CpG density and low coverage during small DMR merging. My independent manual screen identified 150 intragenic DMRs that are >1500 kb long between T and B cells. Most of these DMRs are unmethylated in T cells (**Figure 3.6B**) and are randomly distributed across analysed chromosomes. More than half of DMRs tend to locate at the 5' half of genes (5' end) and only a few large DMRs were found at the 3' end (**Figure 3.6C**). It is possible that the functional relevance and the nature of these large DMRs can differ depending on properties, such as size and position. Gene ontology analysis of the genes differentially methylated between B and T cells at 5' end showed an enrichment of genes associated with lymphopoiesis and B and T cell differentiation (**Table 5**).

The function of gene body methylation is not fully understood hence the identification of intragenic DMRs provide an opportunity to address this. While computational analysis is still in progress, experimental data suggests that gene body methylation is not necessary associated with high transcriptional levels. For example, Tcf7, which shows gene body hypomethylation, is specifically expressed in the T cell lineage and is involved in regulation of Wnt-signalling. The DMR is lowly methylated in T cells and highly methylated in B cells (**Figure 3.7A**). Similarly, the DMR in Chst3 is lowly methylated in B cells and highly methylated T cells which inhibits Chst3 expression (**Figure 3.7B**). In contrast, Ets1 has similar expression levels in B and T cells but is intragenically differentially methylated in each cell type (**Figure 3.7C**). Expression of the genes associated with large DMRs are summarised on **Figure 3.7D**. Some of these genes were reported to have an enrichment of 5hmc at their gene bodies in T cells (Tsagaratou et al., 2014). The presence of 5hmc was correlated with high transcription levels of the host gene. Characterization of 5hmC and histone marks associated with large intragenic DMRs will contribute to our understanding of gene body methylation roles and is in progress.

My DMR identification for the ten chromosomes that I scored has been compared with bioinformatically generated data (Dr. N.Walker - Methpipe). For those hypermethylated in T cells, of the 245 DMRs >1.5kb identified informatically over the whole genome, there is 96% (108/113) concordance with those on the ten chromosomes that I scored. 3% of the discrepancy was due to the differences in the estimate of DMR length during manual screen and computational analysis. In summary, independent manual curation of DMRs and comparison of its results with the results of bioinformatic screen convinced us that our computational approach was robust and accurate.

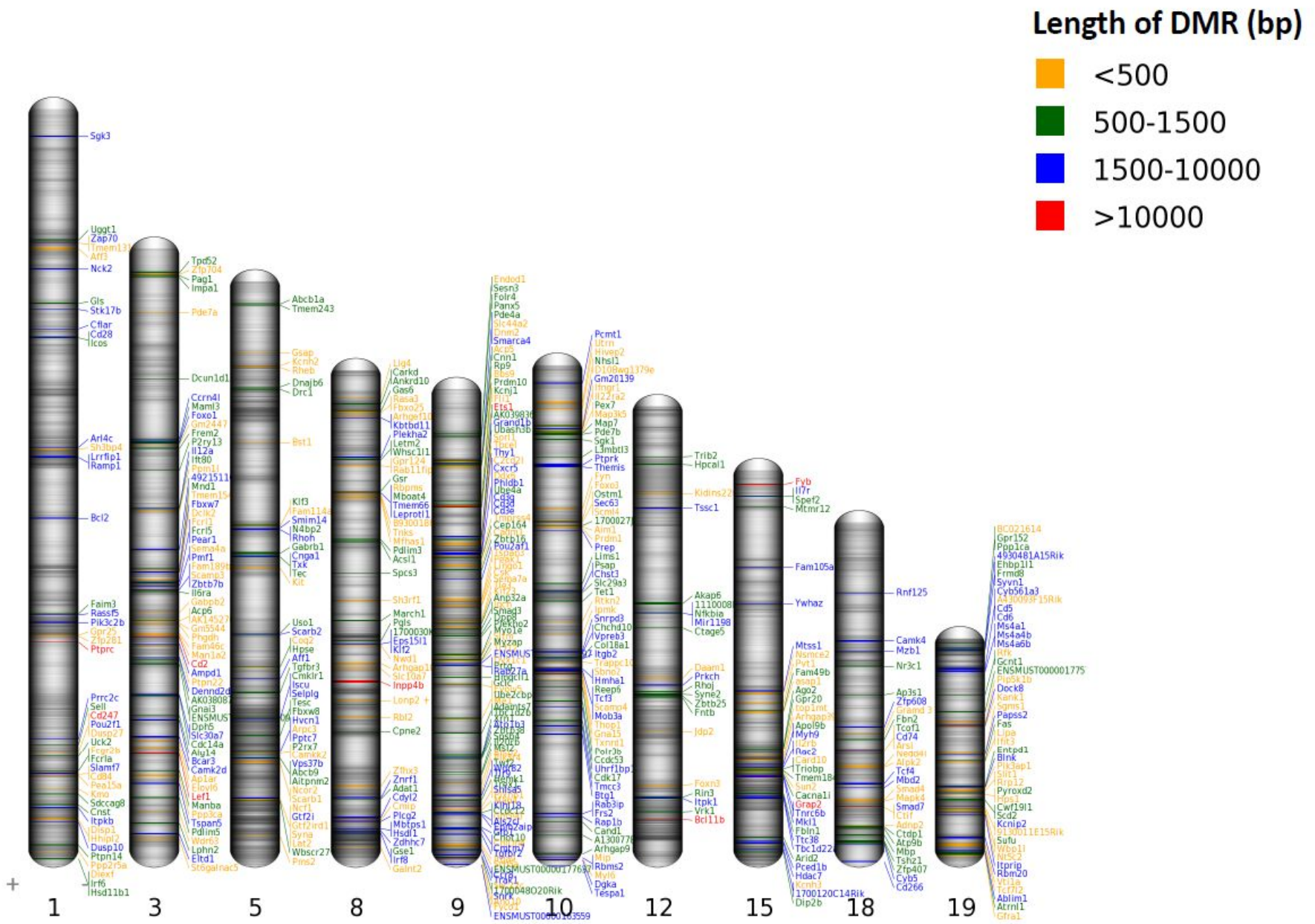


Figure 3.5. Distribution of DMRs

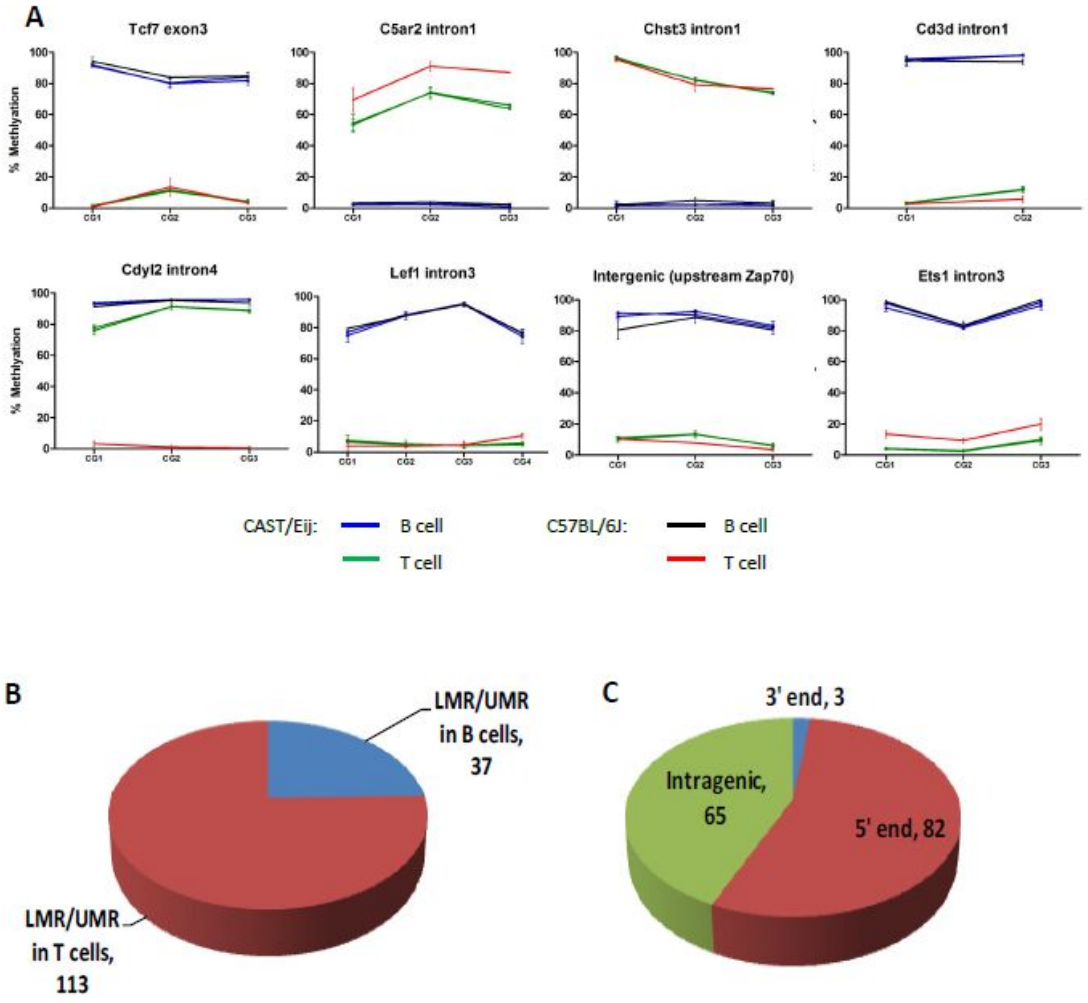


Figure 3.6. Analysis of manually curated DMRs in B and T cells. *A) Validation of DMRs in CAST/Eij. 2 CAST/Eij B cell biological replicates are highlighted in blue, 2 CAST/Eij T cell replicates are in green, 1 C57BL/6J B cell replicate – black, 1 C57BL/6J T cell replicate – red; B) Proportions of LMR/UMR in T and B cell DMRs. C) Distribution of DMRs across gene bodies.*

Table 5. Gene ontology of large 5'end DMRs

Term	Count	P-value
leukocyte activation	21	1.97E-21
cell activation	21	2.05E-20
lymphocyte activation	19	1.53E-19
hemopoietic or lymphoid organ development	18	3.33E-15
immune system development	18	7.53E-15
hemopoiesis	17	1.11E-14
cell surface receptor linked signal transduction	17	0.081233
leukocyte differentiation	16	4.15E-17
regulation of transcription	16	0.06456
lymphocyte differentiation	14	2.42E-15
T cell activation	14	3.03E-15
regulation of apoptosis	14	7.95E-07
regulation of programmed cell death	14	9.16E-07
regulation of cell death	14	9.73E-07
positive regulation of macromolecule metabolic process	14	3.58E-06
transcription	14	0.048321
positive regulation of immune system process	13	9.71E-11
positive regulation of nitrogen compound metabolic process	13	3.08E-06
positive regulation of cellular biosynthetic process	13	5.05E-06
positive regulation of biosynthetic process	13	5.54E-06
T cell differentiation	12	2.64E-14
positive regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	12	1.43E-05
positive regulation of macromolecule biosynthetic process	12	2.04E-05
regulation of cell proliferation	12	2.34E-05
regulation of transcription, DNA-dependent	12	0.059573
regulation of RNA metabolic process	12	0.065237

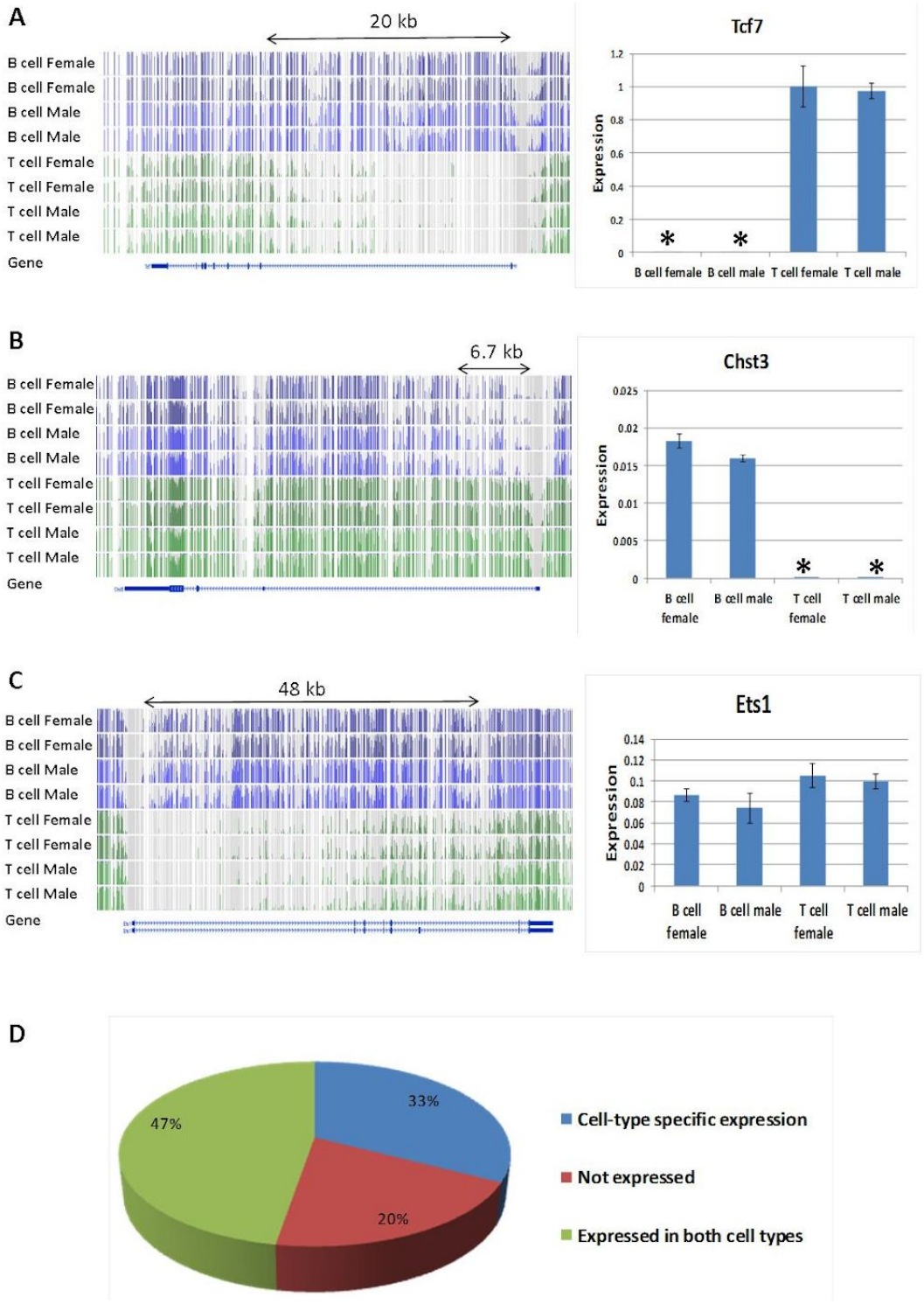


Figure 3.7. Relation between DMRs and expression of host genes. *A,B,C) Left panels represent WGBS-seq screenshots of large DMRs. B cell replicates colored in blue, T cell – in green. Expression of genes that are differentially methylated between B and T cells is on the right panels; D) Correlation between DMRs and expression of host genes*

Summary and discussion

Experimental DNA methylation and expression validation confirmed the accuracy of WGBS and RNA-seq data generated for pure non-cycling populations of B and T cells from the C57BL/6J background. Validation of imprinted loci showed hypermethylation of the IG-DMR imprinting control region at the Dlk1-Dio3 imprinted region. The other imprinted regions showed the expected 40-60% methylation in both cell types. Interestingly, the IG-DMR has a much higher methylation levels in C57BL/6J comparing to CAST/Eij and these level were parent-of-origin specific in reciprocal hybrids. Despite hypermethylation of IG-DMR in B and T cells, only B cells showed detectable Dlk1 expression. However the expression was quite low. The expression of membrane-bound and secreted Dlk1 isoforms was observed. Loss of IG-DMR methylation imprinting correlates with biallelic Dlk1 expression in neural stem cells and astrocytes (Ferron et al., 2011). Similarly, biallelic expression of Dlk1 in C57BL/6JxCAST/Eij reciprocal hybrids was confirmed. Interestingly, the relative amounts of maternally and paternally expressed Dlk1 were different between C57BL/6JxCAST/Eij and CAST/EijxC57BL/6J hybrids suggesting strain background effects on Dlk1 expression.

Cell-type specific loss of IG-DMR methylation imprinting might be important for Dlk1 dosage regulation. Alternatively, it may be an indirect effect of little or no functional relevance but perhaps associated with the fact that the cells are quiescent and might acquire methylation as functionally irrelevant unmethylated loci – the maternally unmethylated allele of the IG-DMR in these cells might be one such functionally irrelevant locus. Taking into account low Dlk1 expression in B cells and the absence of expression in T cells, it is unlikely that the observed hypermethylation of IG-DMR is involved in the regulation of the naive B and T cell state. However, analysis of the expression of other genes from Dlk1-Dio3 locus would be valuable to draw a complete conclusion.

It is not clear when this loss of IG-DMR imprinting happens. The epigenome of hematopoietic stem cells was published by Jeong *et al*. in 2014. The IG-DMR is hypermethylated in these HSCs also, possibly suggesting that imprinting was lost at this stage or earlier, before HSC differentiation. This also suggests that other hematopoietic cell lineages should be hypermethylated at the IG-DMR region.

Interestingly, the Gtl2 promoter region in NSCs is also imprinted and has around 60% methylation independent of mouse strain or IG-DMR methylation levels (Ferron et al., 2011). This suggests that the hematopoietic hypermethylation may only affect Dlk1.

The extent to which biallelic expression of Dlk1 might impact B and T cell development is not known. The relative number of splenic B cell populations in $Dlk1^{-/-}$ mutants might suggest specific roles of membrane-bound and secreted Dlk1 isoforms on B cell differentiation (Raghunandan et al., 2008). Heterozygous $Dlk1^{-/+}$ and $Dlk1^{+/-}$ mutants might provide a further highlight the importance of the Dlk1 dosage control during lymphopoiesis.

Identification of DMRs between T and B cells raises a question about the relationship between the genome and the epigenome in these cell types. The identified set of regions is very diverse in terms of location, size, and host genes. A number of studies reported that hypomethylated regions often coincide with enhancers and transcription factor binding sites (Stadler et al., 2011; Xie et al., 2013). Therefore, it is likely that among our set of DMRs, the ones that are smaller (<1500 kb) could be tissue-specific enhancers or binding sites. Additional knowledge of histone marks and binding proteins that are associated with them are required. Enhancers are specifically marked by H3K4me1. The co-presence of H3K27ac mark is associated with active enhancers and its absence with poised enhancers. Additionally, binding of CHD7 or P300 proteins to hypomethylated regions would also signal an enhancer function for the region. Some of the identified DMRs might impact alternative splicing through regulating the rate of transcription. Previous association of alternatively spliced genes with presence of tissue specific DMRs was reported for mouse retina and brain tissue (Wan *et al*., 2013). This idea is also supported by the fact that a lot of host genes that contain large DMRs have splice isoforms. However, this hypothesis still needs to be tested. Furthermore, some of the DMRs may be linked to the expression of tissue-specific non-coding RNAs initiated intragenically. Differential expression analysis of RNA-seq data including coding and non-coding transcripts for T and B cells would allow us to design validation assays to test these hypotheses.

Previously, correlation of gene body methylation with actively transcribed genes was reported (Aran *et al*., 2011). However, our observations suggest that genes with low methylation are expressed at the same level or even higher than the same genes

with high body methylation levels. Furthermore, it was reported that quite often slow proliferating tissues have similar levels of gene body methylation for active and inactive genes. In general, T cells appear to have less methylation at these regions than B cells. This observation is supported by the fact that most of large DMRs are hypomethylated in T cells and hypermethylated in B cells. However, an interesting picture previously observed in brain tissues showed actively transcribed genes were less methylated than inactive genes (Aran *et al*., 2011). Hence it remains unclear what the actual relationship between gene body methylation and gene expression is. One possibility is that gene body methylation allows tighter control of gene expression and reduces transcriptional noise (Huh *et al*., 2013). To test this hypothesis, single cell targeted gene expression analysis can be conducted to determine whether gene body methylated genes show more or less variability between cells. Another possibility is that the absence or presence of methylation serves as a specific mark of a developmental program in B and T cells or is somehow correlated with quiescent state of cells. It is worth looking at methylation dynamics at large DMR regions during differentiation and after activation.

Chapter 4

Genome-wide identification and characterization of murine metastable epialleles

Introduction

Nearly 10% of the murine genome consists of endogenous retroviral elements (Muñoz-López & García-Pérez, 2010). ERVs are globally silenced by various epigenetic mechanisms. The discovery of metastable epialleles suggested that other ERVs might be methylated to a different extent in different individuals (Rakyan *et al*., 2002). Two well described cases of metastable epialleles are results of spontaneous mutations via insertions of IAP elements (Yen et al., 1994; Vasicek et al., 1997). Unlike the majority of IAPs (from gross analysis), these IAPs can be variably methylated in individuals and, when hypomethylated, can impact the expression of the neighbouring host genes. The expression of neighbouring genes inversely correlates with IAP methylation levels. Thus metastable epialleles represent so called facilitated epialleles when a particular genetic region provides a substrate for the establishment of variable epigenetic state at this region (Richards, 2006). A phylogenetic analysis of IAP clades using semi-quantitative methods indicated that individual IAPs are capable of acquiring inter-individually variable methylation levels (Faulk et al., 2013). However, the attempts to quantify metastable epialleles genome-wide have shown inconsistent results. Previous strategies to conduct genome-wide screens for metastable epialleles relied on different assumptions: the presence of active histone marks at metastable ERVs, variation of gene expression potentially linked to epigenetic variation at neighbouring loci, and inter-individual methylation variation at ERVs (Weinhouse et al., 2011; Ekram et al., 2012; Oey et al., 2015).

To understand the importance and role of metastable elements for mammalian evolution and development, it is important to know how common they are in the genome and to what extent they impact neighbouring genes. The mechanism

underlying the acquisition of the epigenetic and transcriptional properties of metastable epialleles is particularly interesting since, to date, there is no evidence for a genetic cause of variable ERV methylation. It is not clear how certain ERVs can escape full silencing and why this event is stochastic in some cases. Apart from these, naturally existing metastable epialleles might become a useful model to study transgenerational epigenetic inheritance and influences of environment on phenotype and development.

Genome-wide identification of IAP-derived metastable epialleles

I designed a two-step approach for genome wide identification of metastable epialleles (see Chapter 2). The first step relies on the hypothesis that newly inserted polymorphic ERVs will more likely behave like metastable epialleles and have an impact on neighbouring gene expression. I used previously published data about polymorphic ERVs in 18 different strains (Nellaker et al., 2012) and cataloged C57BL/6J specific ERVs that are absent in CAST/Eij (**Table 6**). ETns and IAPs are considered to be two of the most polymorphic ERV families. However, the total number of polymorphic ETns was much smaller than the number of IAPs, MaLR and RLTR10 in the published screen.

Table 6. Summary of ERVs present in C57BL/6J and CAST/Eij strains published in Nellaker et al., 2012

	CAST/Eij specific	C57BL/6J specific
ETn	350	269
IAP-I	2949	1994
IS2	154	89
MaLR	1882	1466
MuLV	108	54
RLTR1B	105	112
RLTR10	1155	964
RLTR45	86	71
VL30	24	157

ERV coordinates were converted from mm9 to mm10 mouse genome assembly. Then, the ERVs were assigned to the nearest protein-coding gene (Ensemble GRCm38). For further analysis, B and T cell RNA-seq datasets generated for BLUEPRINT projects were used to identify a set of genes that are differentially expressed between two strains. This was done using two strategies (**Figure 4.1A**). The first strategy relied on computational analysis conducted by R. Gunning, whereby quantification of differentially expressed genes was done separately for B and T cell datasets using DESeq2 and edgeR. Due to the absence of strain specific annotation of ncRNAs, they were removed from the analysis. The overlap of data between the two programs was used to generate a final set of differentially expressed genes between two mouse strains. Differential expression analysis identified 7613 transcripts that were differentially expressed in B cells between two strains and 4315 transcripts were similarly identified in T cells. These genes were overlapped with genes that contain ERV insertion nearby. The overlap was done separately for B cell differentially expressed genes and T cell differentially expressed genes. However, the conducted analysis did not identify the Cdk5rap1 gene to be differentially expressed between strains, which is a previously described metastable epiallele. This gene contains an intragenic insertion of C57BL/6J specific IAP that is variably methylated between individuals and can drive expression of downstream exons of Cdk5rap1 gene (Druker et al., 2004). To insure the validity of my screen, I have separately generated a list of differentially expressed genes that have been previously assigned to polymorphic ERVs based on their FPKM values. First, genes that had FPKM>=10 at least in one cell type of one strain were selected reassuring that they are expressed. The threshold for FPKM values was selected based on previous RNA-seq validation. Out of these genes, I selected genes with 1.4 times FPKM expression difference. Cdk5rap1 FPKM expression values were used to define differential expression. This way I ended up with 477 genes that are differentially expressed between two strain in at least one cell type and have an insertion of polymorphic ERV nearby (**Figure 4.1A**). The ERVs identified by each strategy are summarized in **Table 7**. The overlap between different strategies is illustrated on **figure 4.1B,C**. As a result, 1672 individual ERVs including 552 IAPs were found to be located next to differentially expressed genes.

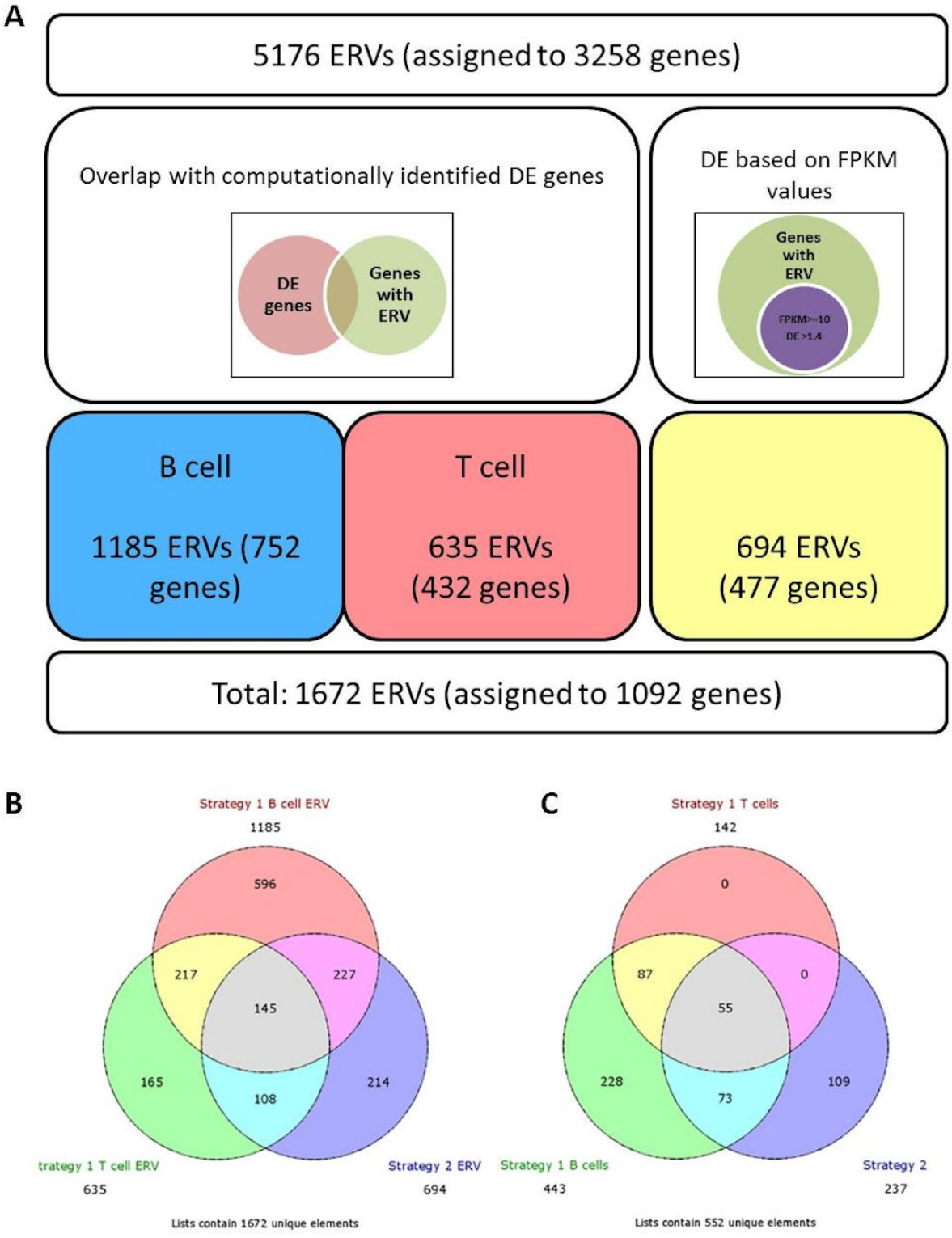


Figure 4.1. Summary of the biased screen for metastable epialleles. *A) Summary of ERVs and nearby genes identified during biased screen; B) Overlap between ERVs identified by different screen strategies; C) Overlap between IAPs identified by different screen strategies.*

Table 7. Summary of identified ERVs

	Overlap with DE genes in B cells	Overlap with DE genes in T cells	Genes with 1.4x FPKM difference
ETn	67	30	46
IAP	443	257	239
IS2	14	11	14
MaLR	364	179	198
MuLV	16	9	15
RLTR10	196	107	132
RLTR1B	30	11	18
RLTR45	24	8	8
VL30	31	23	24

So far all described cases of metastable epialleles are associated with the insertions of IAP elements. Furthermore, the majority of the identified ERVs that potentially impact expression of nearby genes were IAPs as well. So I first attempted to conduct a manual analysis of qualitative differences in methylation of the identified IAPs. Metastable epialleles have been previously described to have constant methylation levels across different tissues within the individual. This allows us to treat WGBS and WGoxBS datasets generated for B and T cells as biological replicates to assess methylation variation at our metastable epiallele candidates. WGBS and WGoxBS data was visualized on the WashU Epigenome Browser. Methylation levels of 552 IAPs were visually analysed. The majority of candidates (521 IAPs) were highly methylated in all 16 replicates. However, 31 IAPs were found to have variable methylation between replicates. We describe this pattern as “ragged” methylation. I selected a set of 8 highly methylated and 10 variably methylated IAPs to experimentally validate interindividual methylation variation at these loci in multiple individuals (n=10) by bisulfite pyrosequencing. The primers were designed to assess the methylation of 3-10 distal CpGs at the 5’ LTR of each IAP. All IAPs that were highly methylated according to visual assessment of BLUEPRINT datasets were highly methylated across 10 individuals with no interindividual variation (**Figure 4.2A,** included in **Supplementary figure 1**). In contrast, the 10 IAPs with “ragged”

methylation had different methylation levels across the 10 individuals (**Figure 4.2B,** included in **Supplementary figure 1**) consistent with the bioinformatics analysis that identified them. The degree of variation was specific for each IAP. I have additionally checked the consistency of the methylation levels across different tissues within the same individuals for 3 of these IAPs (**Figure 4.2C**). Similar to the known metastable epialleles, the methylation levels of these IAPs were consistent in liver, spleen, kidney and brain purified from the same individual. In summary, these results confirm that IAPs with "ragged" methylation validate as metastable epialleles and have individual and tissue methylation properties similar to A^{vy} and $Axin^{fused}$.

Manual screen of methylation of other ERVs identified during this biased screen showed the majority of them were highly methylated as well. MaLR regions were mostly CpG poor, sometimes having no CpGs at all. 49 ERVs showed methylation variation similar to IAPs across replicates. They mostly belonged to 3 classes: ETns (19 candidates); MuLVs (11 candidates) and VL30 (13 candidates). I selected three ETns, two MuLVs and one VL30 regions for experimental validation. However, only 1 MuLV ($MuLV_{Cep85}$) region was differentially methylated between mice (**Figure 4.3;** further discussed in section "Genome-wide identification of non-IAP-derived metastable epialleles").

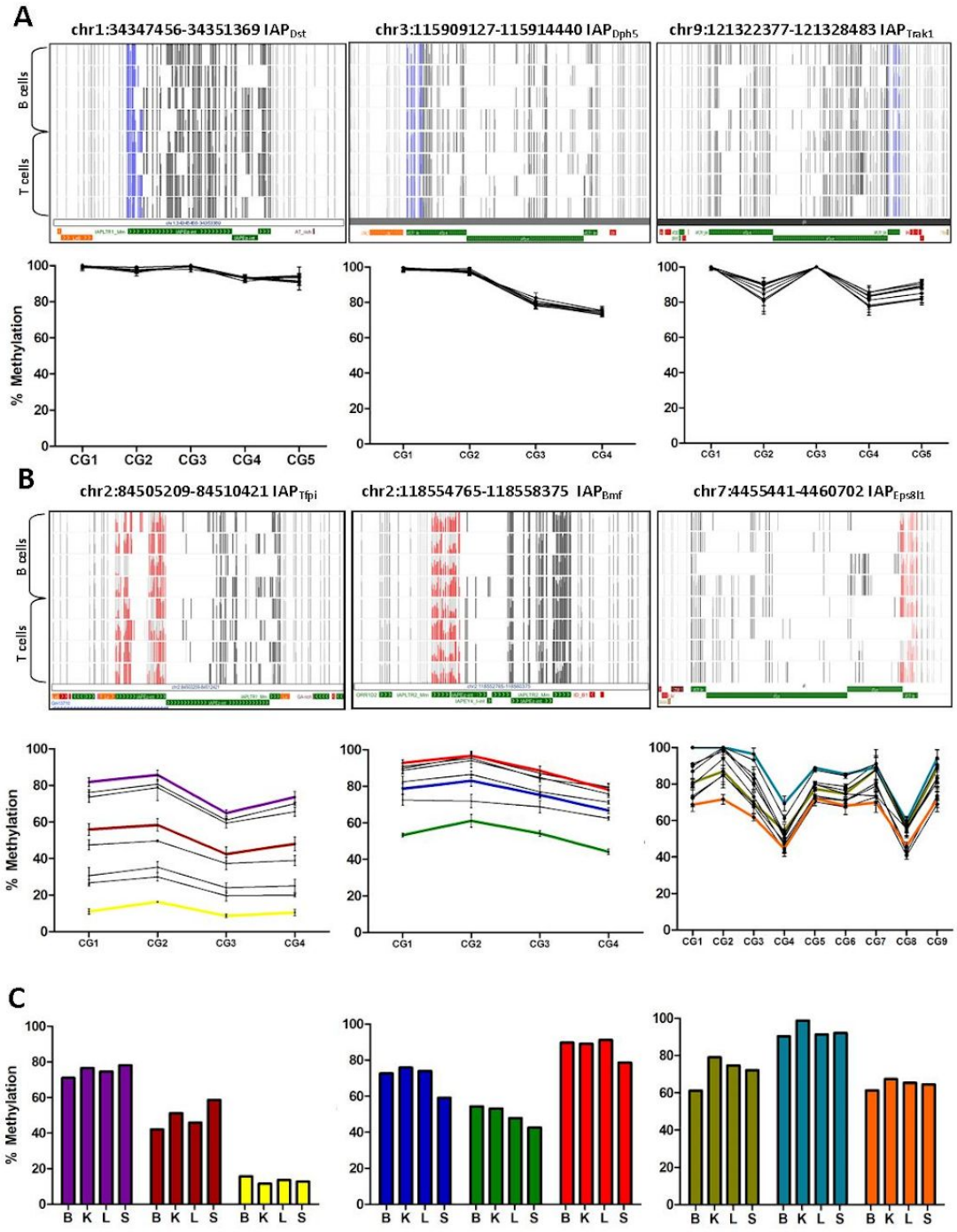


Figure 4.2. Validation of IAPs identified during biased screen. *A) Validation of highly methylated IAPs. IAPs are highlighted in black on the screenshots. Regions that were analysed by pyrosequencing are highlighted in blue. Each line represents an individual on pyrosequencing plots (10 individuals); B) Validation of IAPs (black on screenshots) with "ragged" methylation (red on screenshots). Each line represents an individual on pyrosequencing plots; C) Intra-individual methylation consistency across brain (B), liver (L), kidney (K), and spleen (S) tissues. The colors represent different individuals and correspond to the colors in panel B.*

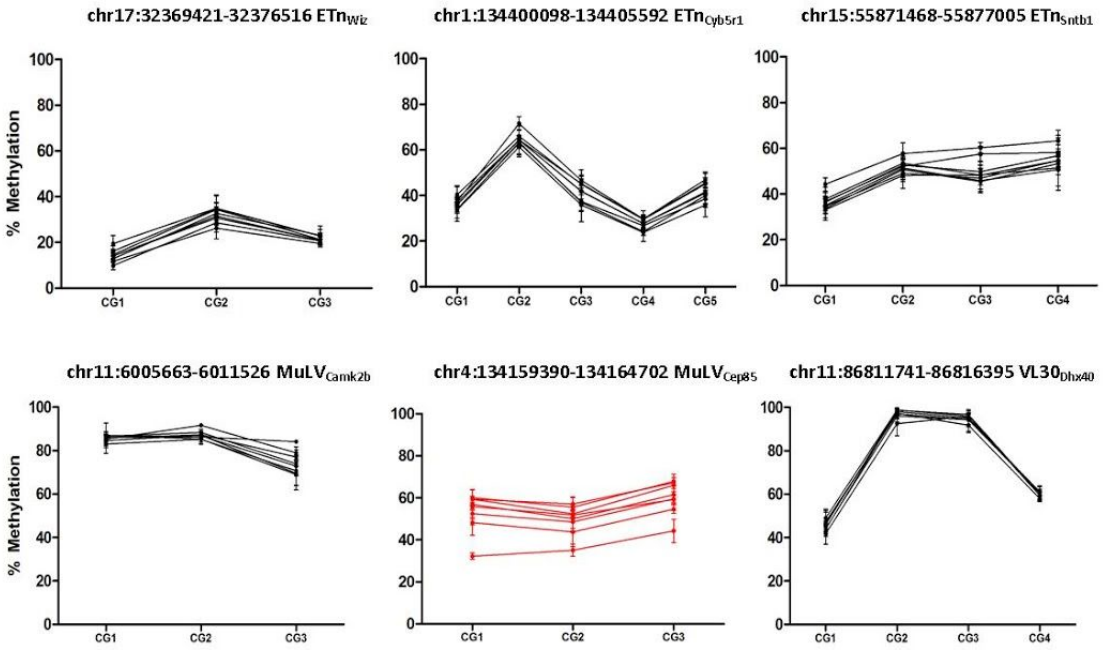


Figure 4.3. Validation of ERVs identified during biased screen. *Each line represents an individual on the pyrosequencing plots. Only one MuLV had the properties of a metastable epiallele (colored in red).*

The biased screen of polymorphic ERVs located next to differentially expressed genes identified a specific methylation pattern that is associated with inter-individual methylation variation at least at IAP regions. Manually identified and validated candidates were used to model the identified methylation pattern that can be applied to a genome-wide unbiased identification of potential metastable epialleles. Next, I applied the model to screen IAPs genome wide and, after validation, determined that this was a reliable approach. I then extended the screen to all other types of ERVs.

The repetitive genome is challenging to map. Most of IAP alignments miss specifically aligned reads to the internal IAP parts due to their repetitive nature and distance from unique flanking sequence. For the genome wide screen, my analysis assessed the 8 distal-most CpGs in an IAP. These CpGs were experimentally validated by pyrosequencing. Distal IAP sequences have a reliable alignment due to their location next to the unique genomic regions. This way it was possible to decrease the false discovery rate caused by low alignment efficiency.

RepeatMasker was used to identify the structure and location of IAPs in the C57BL/6J genome. RepeatMasker contains information about IAPs fragments (LTRs and retroviral genes) rather than the full IAP insertion. Bedtools software was used to assemble these fragments. Fragments were considered to belong to one IAP insertion if they were within 150 bp of each other and located on the same strand. This way 11559 IAPs were identified in the C57BL/6J genome (**Figure 4.4A**). Most of the identified insertions were solo LTRs with no retroviral gene sequences next to them (~5000 IAPs). A substantial number of IAPs were truncated or were missing 5' and/or 3' LTRs or contained large insertions of other sequences within them (~3300 IAPs). About 3200 IAPs contain 5' and 3' LTRs flanking coding IAP fragments resembling the full classical IAP structure.

To model the observed "ragged"methylation pattern in our datasets, I calculated the average methylation levels for 8 distal CpGs for each biological replicate (**Figure 2.1**). The difference between the second highest and second lowest values across average methylation levels of all biological replicates was used as a score for methylation variation in our data (computational score of variation). The screen was run separately for 5' and 3' end distal CpGs. To confirm that this way of scoring methylation variation across our datasets represents real interindividual methylation variation, 68 IAPs with different methylation variation scores were experimentally analysed for interindividual methylation variation (**Figure 4.4B; Supplementary table 3**). The actual experimental variation was defined as the difference between average methylation levels of the individuals with highest and lowest methylation levels. A T-test found a statistically significant correlation between the experimentally identified variation and the bioinformatically determined one ($p < 0.0001$). Consistent with this, that the higher the computational variation score that was assigned to an IAP, the higher likelihood that this IAP would have different methylation between individuals. Experimental validation was used to identify false positives/false negatives and determine a threshold for the final list of metastable epiallele candidates. The computational variation score of 25% was used as the threshold based on the following criteria. ~75% of experimentally analysed IAPs with a computational variation score of >25% showed >10% variation of methylation between individuals (**Supplementary figure 1**).

Because two different cell types were used in the screen, there was a possibility of differentially methylated IAPs (DMI) between B and T cells. These regions might be picked up as variably methylated and hence would represent false positive matches. To avoid this, I identified 35 IAPs where all high or all low average methylation came from the same cell type since this was most likely to represent DMIs. All informatically identified DMIs were visually assessed to confirm this assumption. Three of these were experimentally analysed for interindividual variation. All 3 regions had similar methylation levels between individuals (**Figure 4.4C**). Overall the screen for the 5' IAP end identified 100 IAPs with a computational score >25%, 26 of them were identified as differentially methylated between B and T cells. Similarly, the 3' end IAP screen identified 71 IAPs including 9 DMIs (**Figure 4.5A**). 31 IAPs were identified by both 5' and 3' IAP end screens. 7 of them are full structure IAP elements suggesting that both 5' and 3' LTRs are variably methylated. The rest were solo LTRs. In total 105 IAPs were identified having variable methylation at one end at least (**Supplementary table 4**). The final list of metastable epiallele IAP candidates was used for further analysis and characterisation of their metastable epiallele properties.

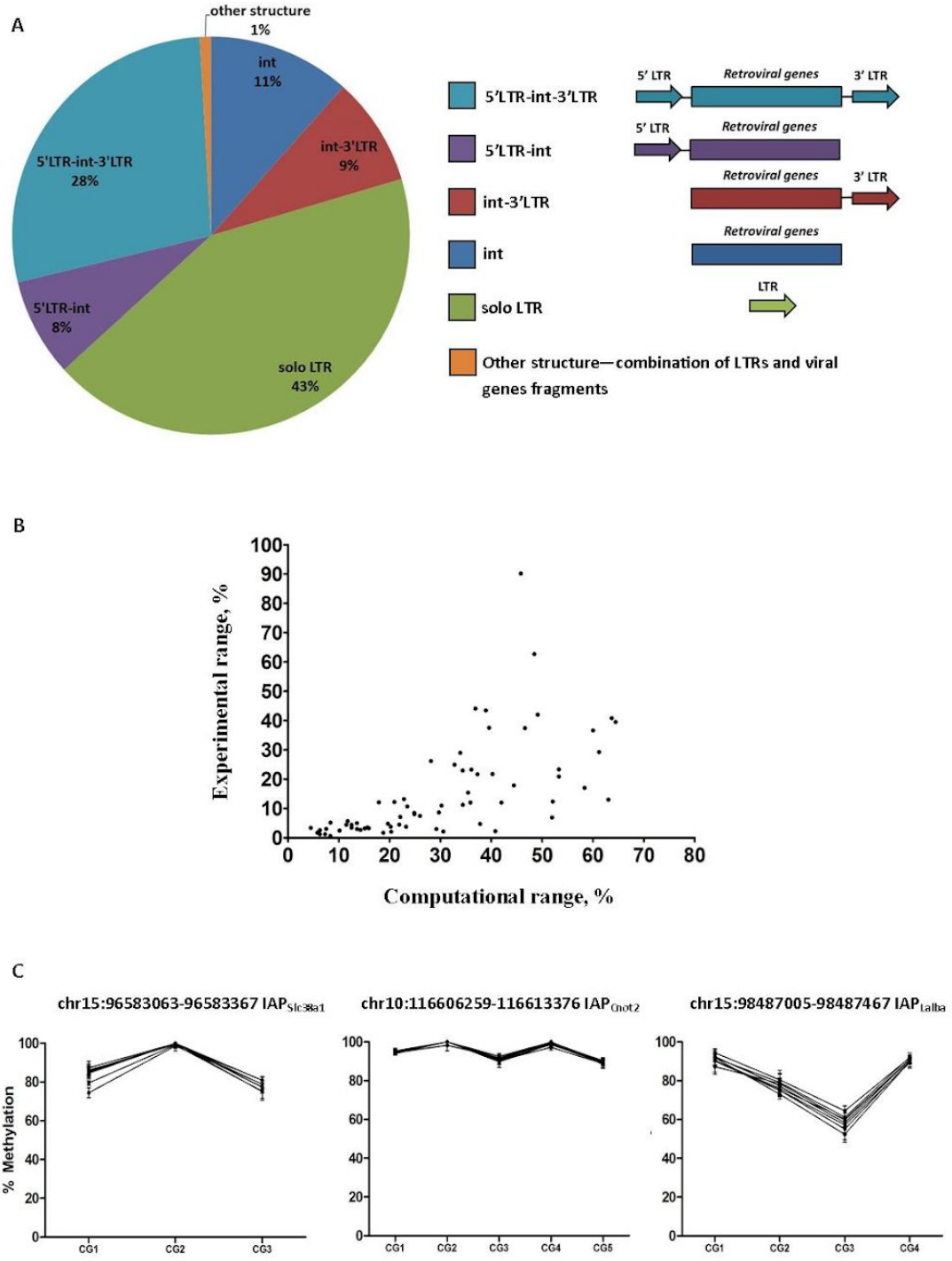


Figure 4.4. Genome-wide screen for metastable IAPs. *A) Pie chart distribution of all IAPs in the C57Bl/6J genome based on IAP structure; B) Validation of genome-wide screen for metastable epialleles. Each dot represents an IAP. Experimental range represents the difference between average methylation levels of the most highly and the most lowly methylated individuals identified via bisulfite pyrosequencing; C) Validation of IAPs that have been identified as differentially methylated between B and T cells.*

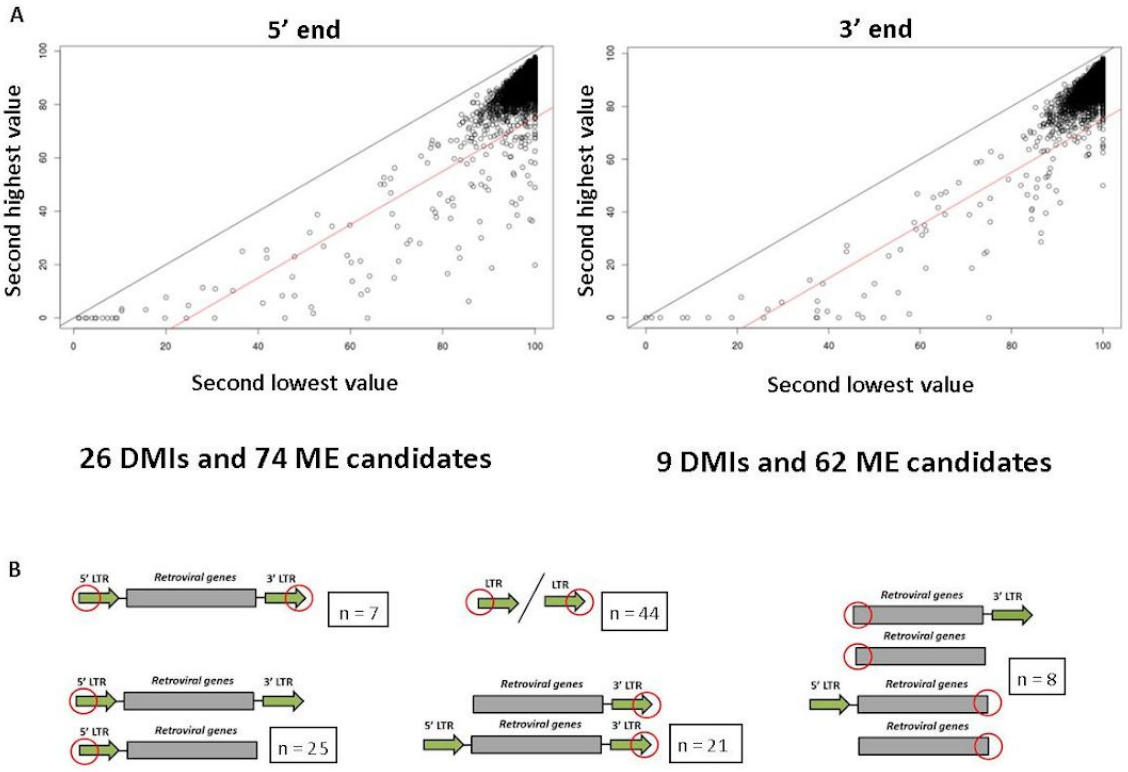


Figure 4.5. *A) Summary of genome wide screen at 5' and 3' IAP ends. Red line represents the threshold of 25% difference between second highest and second lowest average methylation levels quantified for an individual IAP. Each dot represents an IAP; B) Types of metastable IAP structures. Regions with "ragged" methylation profiles are marked by red circles*

Characterization of metastable epiallele candidates

Based on IAP structure and the location of the "ragged" methylation pattern, the identified set of IAPs can be divided in the following manner (**Figure 4.5B**):

1) solo LTRs identified during screen including both 5' and/or 3' end of IAP. In the majority of the cases "ragged" methylation covers the full LTR.
2) IAPs with a ragged 5' LTR. This group contains mostly full structured IAPs (5'LTR-int-3'LTR).
3) IAPs with a ragged 3' LTR. 5'LTR-int-3'LTR is also a common structure for this group.
4) IAPs with "ragged" 5' and 3' LTRs - also with the 5'LTR-int-3'LTR structure.
5) Truncated IAPs where the internal IAP fragment was identified as a variably methylated.

Inter-individual methylation at the 3' LTR of 9 metastable and 3 highly methylated IAPs was analysed. Three highly methylated IAPs showed no inter-individual methylation variation at 3' LTRs (**Figure 4.6A**). Similarly, five metastable IAPs showed no inter-individual methylation variation at 3' LTRs (**Figures 4.6B, 4.7A**). We have also compared methylation at 5' and 3' LTRs for 5 metastable IAPs. No correlation between 5' and 3' LTRs methylation levels within the same individual was found for four analysed IAPs. Moreover, the range of methylation variation at 5' and 3' LTRs was different (**Figures 4.6C, 4.7A,B**). Hence, though 5' and 3' LTR within the same IAP have identical sequences, they are functionally different and have different methylation levels within an individual.

However, one IAP among the validated candidates did not follow this rule. Quite strikingly, IAP_{Pgm1} showed methylation variation at both LTRs that was correlated within an individual. Unlike other IAP candidates, IAP_{Pgm1} showed a bimodal distribution of methylation amongst the individuals tested, being either highly or lowly methylated at both LTRs within the same individual (**Figure 4.7C**). We have now analysed IAP_{Pgm1} methylation in 18 individual mice. Only two individuals showed around 50% methylation at this IAP. The data for IAP_{Pgm1} suggests that this epiallele is inherited in a Mendelian manner hence potentially is a genetically determined epigenetic level. This hypothesis needs to be tested in future experiments since the putative genetic mechanism determining whether the methylation status in animals is

'homozygous high', 'homozygous low' or 'heterozygous high/low=intermediate' might contribute novel insight into genotype-epigenotype interactions.

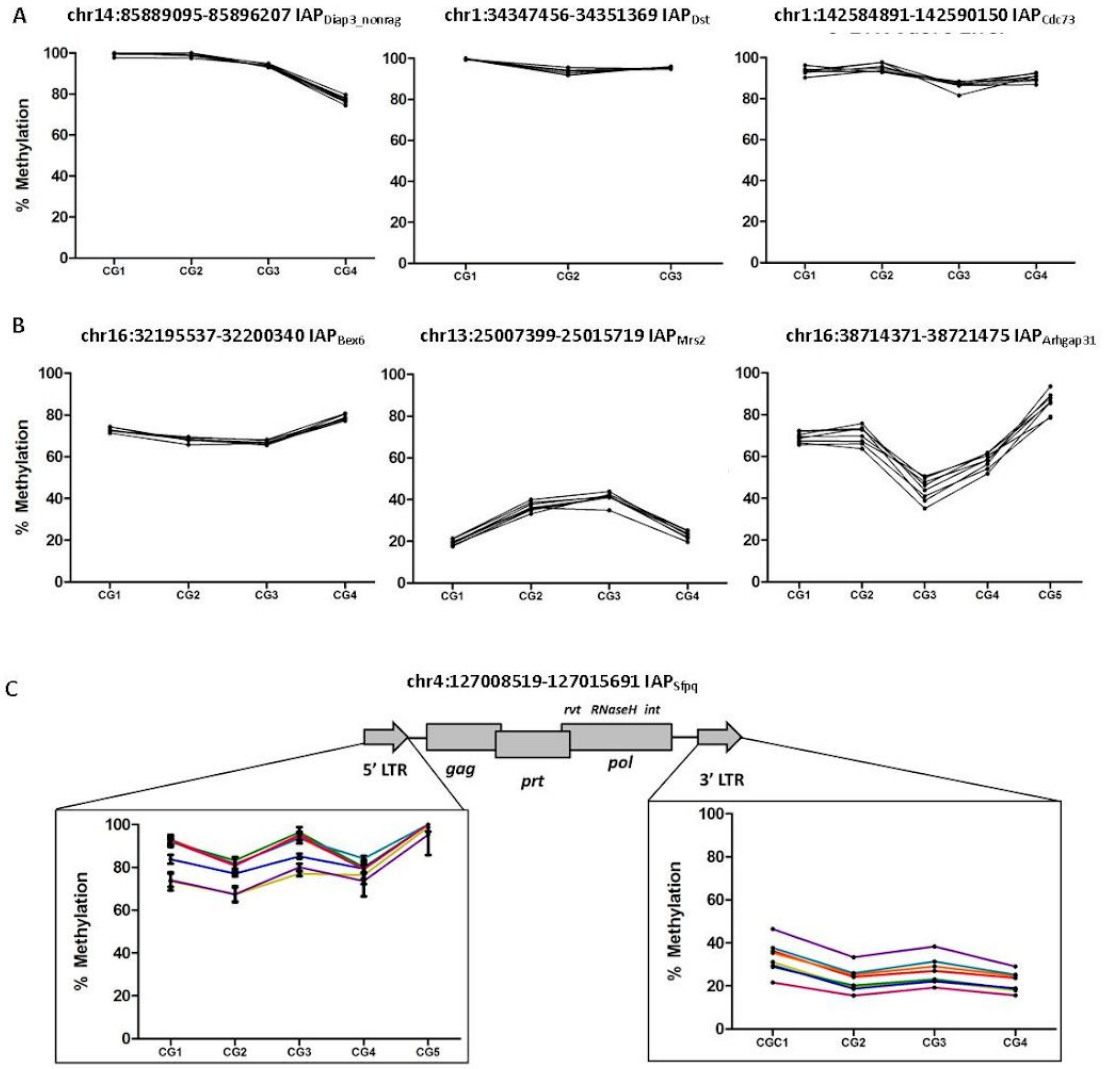


Figure 4.6. Validation of interindividual methylation variation at 3' LTR. *A) Interindividual methylation variation at 3' LTRs of highly methylated IAPs. Each line represents an individual (n=8 individuals); B) Validation of "ragged" methylation at 3' LTR of 3 metastable epiallele candidates; C) Interindividual methylation variation at 5' and 3' LTRs of metastable IAPs. It is noteworthy that the same individuals (colour-coded) have both different methylation levels and different relative methylation levels at 5' and 3' LTRs of IAP$_{Sfpq}$*

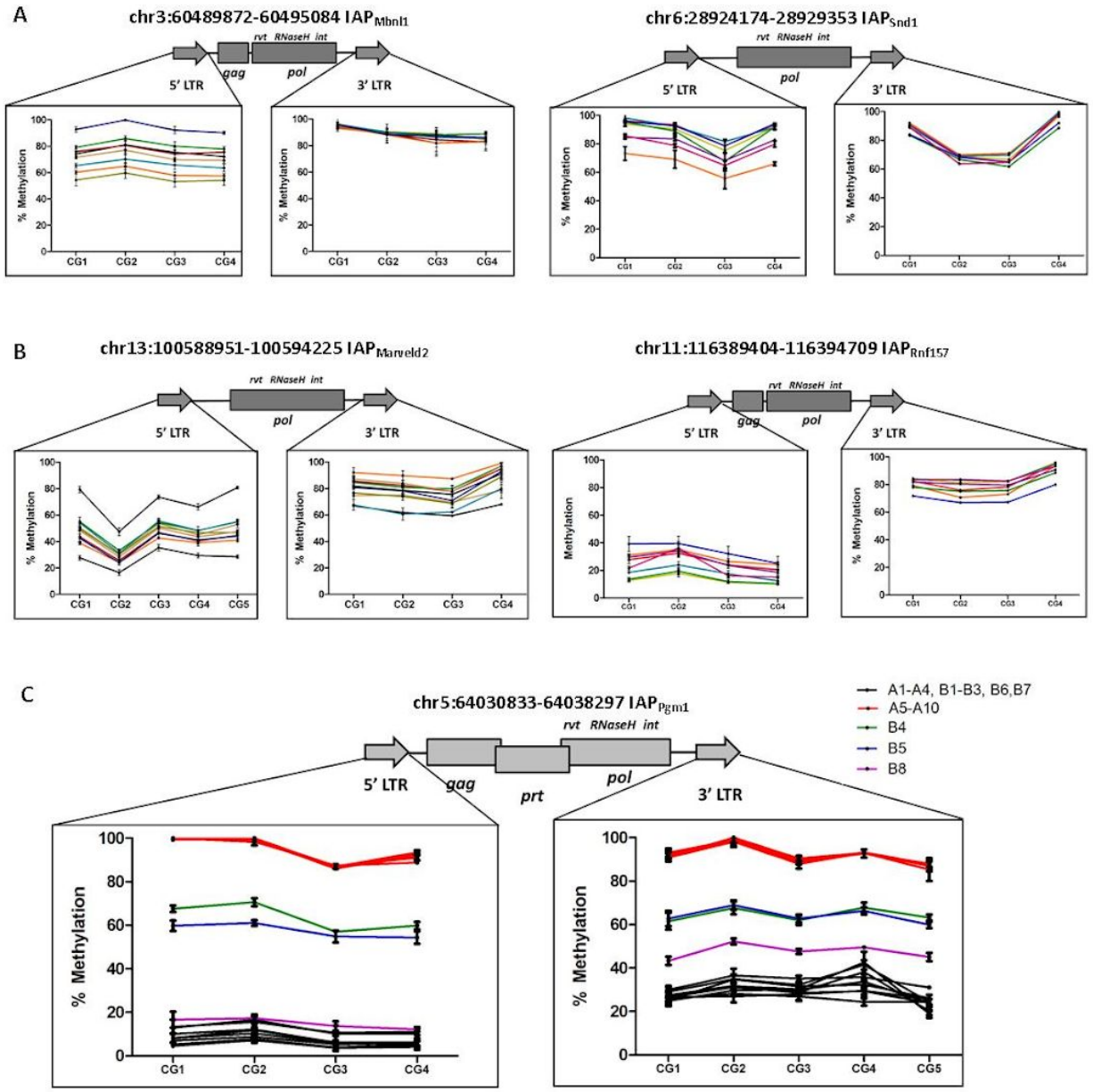


Figure 4.7. Relation between 5' and 3' LTR methylation. *A) Unlike 5' LTRs, 3' LTRs of IAP_{Mbnl1} and IAP_{Snd1} are not variably methylated between individuals; B) Inter-individual methylation variation at 5' and 3' LTRs of $IAP_{Marveld2}$ and IAP_{Rnf157}. The same individuals were tested for both LTRs, shown in corresponding colors. The methylation variation at 5'LTR and 3' LTR are different; C) Inter-individual methylation variation at 5' and 3' LTRs of IAP_{Pgm1}. Here, methylation at 5' LTR correlates with methylation at 3' LTR reflecting the special properties of IAP_{Pgm1}. Highly methylated individuals are coloured in red. Lowly methylated individuals are coloured in black. Individuals with intermediate methylation levels at both 5' and 3' LTRs are the same colour on both pyrosequencing plots*

To ensure that the observed interindividual methylation variation at these loci is not a result of tissue heterogeneity, we analysed interindividual methylation variation for some metastable epiallele candidates using DNA from naive B cells. Naive B cells were purified from the spleens of different individuals. These cells represent pure single cell type populations unlike multicellular tissues and represent the cells used to generate the BLUEPRINT datasets used in the screen. Moreover, these cells are non-cycling, eliminating cell-cycle related methylation variation. All analysed IAPs were found to have interindividual methylation variation in pure B cell populations (**Figure 4.8A**) reflecting the data observed in the dissected tissues of individual animals. This indicates that the variation in methylation seen in tissues unlikely reflects cell type heterogeneity but rather different methylation states within the same cell type.

CpG dinucleotides are the main targets of DNA methylation in mammals, however there is increasing evidence for the existence and functional importance of non-CpG methylation in developing gametes, stem cells, and brain (Pinney, 2014). Non-CpG methylation is important for the efficient silencing of transposable elements in Arabidopsis (Stroud et al., 2014). Therefore, we tested whether similar mechanisms might be involved in the regulation of activity of the metastable epialleles. Four metastable IAPs were assessed for non-CpG methylation by bisulfite pyrosequencing. No non-CpG methylation was found at metastable epiallele candidates (**Figure 4.8B**). Indeed, cytosines in all non-CpG contexts were completely unmethylated.

In addition to their differences in structure (presence of both LTRs and retroviral genes - **figure 4.5B**), the metastable epialleles belong to different IAP subtypes (**Figure 4.8C**). Interestingly, an enrichment of 3 subtypes was found: IAPLTR1, IAPLTR2 and IAPEY4. The majority of the full structure metastable epiallele IAPs had IAPLTR1_Mm flanking sequences, while most solo LTRs belong to the IAPLTR2_Mm subtype. The IAPLTR1 subtype has been previously described in the literature as the youngest IAP subtype and responsible for the majority of recent insertions (Qin et al., 2010).

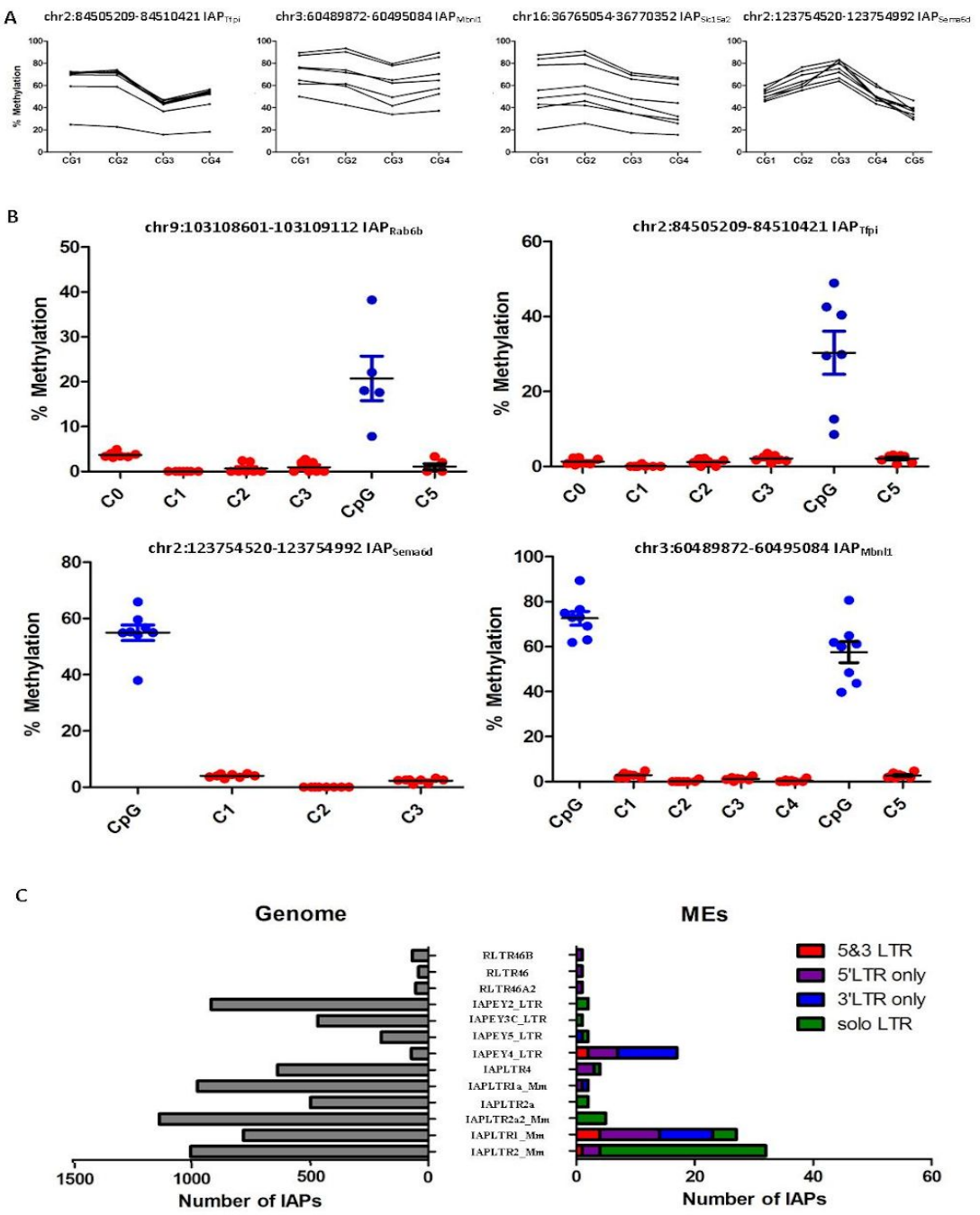


Figure 4.8. Characterization of metastable IAPs. *A) Interindividual methylation variation in pure naive B cell samples at 4 metastable IAPs validating a cell type utilised for the BLUEPRINT project; B) Non-CG methylation at 4 metastable IAPs, assessed by pyrosequencing. Each dot represents one individual. Distal non-CpG cytosines are shown as red dots and CpGs are shown as blue dots – variable methylation only occurs at CpGs; there is no non-CG methylation at these IAPs; C) Enrichment of IAP subtypes in the identified set of metastable IAPs. The left side represents the total number of IAPs in the genome of a particular LTR subtype. The right side shows the number of metastable IAPs of a particular LTR subtype, colour-coded accordingly to what part of the IAP is variably methylated between individuals*

The mouse genomes project conducted at the Sanger Institute provides an extensive database of structural variants across 18 mouse strains (Keane et al., 2011). 6774 out of 11559 IAPs overlap with structural variants reported in this database. We used this data to assess the genetic polymorphism of our metastable IAPs (**Figure 4.9**). Only 17 of our 105 C57BL/6J IAP insertions were absent amongst the structural variants existing between the 18 strains (on the figure reported as being present in 18 strains). 46 IAPs (44%) were present in less than 11 strains. Together this indicates an enrichment of polymorphic insertions in our set of candidates. An enrichment of polymorphic IAP insertions that belong to evolutionary young IAP subtypes amongst our variably methylated IAPs suggests that the more recent insertion are more prone to interindividual epigenetic variation.

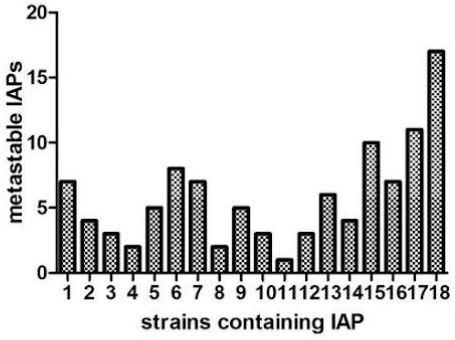


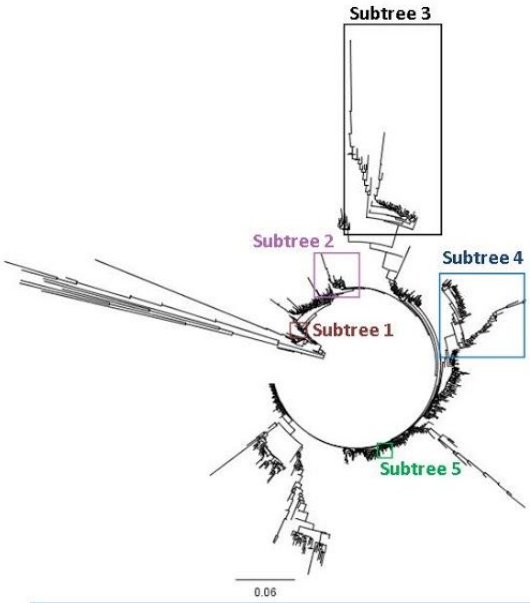
Figure 4.9. Distribution of metastable IAPs across 18 mouse strains (as determined from The Mouse Genomes Project http://www.sanger.ac.uk/science/data/mouse-genomes-project)

Relationship between metastability and IAP sequence, chromatin, and genetic background.

Given the variety of structures and subtypes at the identified set of metastable IAPs, it is very unlikely that the establishment of variable DNA methylation at these regions is dictated by the IAP sequence itself. Furthermore, the validation showed that the ranges of interindividual methylation variation at metastable IAPs are different for individual IAP regions. To further investigate the relation between IAP sequence and its methylation, I have built a neighbor-joining tree for the IAPLTR1_Mm subtype. Sequences of all IAPLTR1_Mm subtype IAPs (780 IAPs) were downloaded to build a neighbour-joining tree. 27 of these IAPs have been identified as metastable epialleles

in our screen. Neighbour-joining tree analysis helps to identify closely related IAP sequences. If IAP metastability is determined by IAP sequence, our 27 metastable IAPs are expected to form a separate clade on the tree.

Using the neighbour-joining tree, I found 5 distinct subtrees that contain metastble IAPs. Most of our IAPLTR1 metastable retroelements were found within Subtree 4 (**Figure 4.10**). This likely reflects their recent integration into the C57BL/6J genome. Interestingly, metastable $IAP_{Slc15a2}$ had about 99% identity to two IAPs that were not identified as metastable. These two IAPs (IAP_{Gpsm1} and IAP_{Zak}) were highly methylated in all 10 analysed individuals (**Figure 4.11A**). Metastable IAP_{Mbnl1} and IAP_{Tfpi} appeared to have 100% identical sequences. If IAP sequence was determining the property of IAP methylation variation within an individual it would be expected that both of these IAPs would have similar methylation levels within the individual and possibly have similar interindividual methylation variation ranges. However, highly methylated individuals at IAP_{Tfpi} were lowly methylated at IAP_{Mbnl1} and vice versa. While IAP_{Tfpi} showed about 80% methylation variation between individuals, IAP_{Mbnl1} methylation varied within a 40% range (**Figure 4.11B**). Once more, this suggests that IAP sequence is unlikely to be a deterministic factor for inter-individual methylation variation at metastable IAPs. Moreover, the establishment of methylation at metastable IAPs is locus specific. Methylation within an individual is established independently at each single IAP locus, even at closely related or identical ones. This suggests that the mechanism governing this process acts in cis, confirming the absence of an overarching trans-mediated mechanism targeting all metastable epialleles within a single individual in the same way.



Subtree 1	IAP_{Rftn1}
Subtree 2	IAP_{Gm4262}, IAP_{Aire}
Subtree 3	IAP_{Acss2}, IAP_{Pifo}, IAP_{Bex6}, IAP_{Fam32a}
Subtree 4	$IAP_{Mir30c-2}$, IAP_{Snd1}, $IAP_{3110079015Rik}$, IAP_{Txn2}, IAP_{Sfpq}, IAP_{Lrfn3}, IAP_{Ube2e2}, $IAP_{E030019B13Rik}$, IAP_{Rims2}, $IAP_{Slc15a2}$, IAP_{Tfpi}, IAP_{Mbnl1}, $IAP_{Gm13849}$, IAP_{P4ha1}, IAP_{Eps8l1}, IAP_{Rnf157}, IAP_{Nrros}, $IAP_{Marveld2}$, IAP_{Kcnh6}
Subtree 5	IAP_{Ccr9}

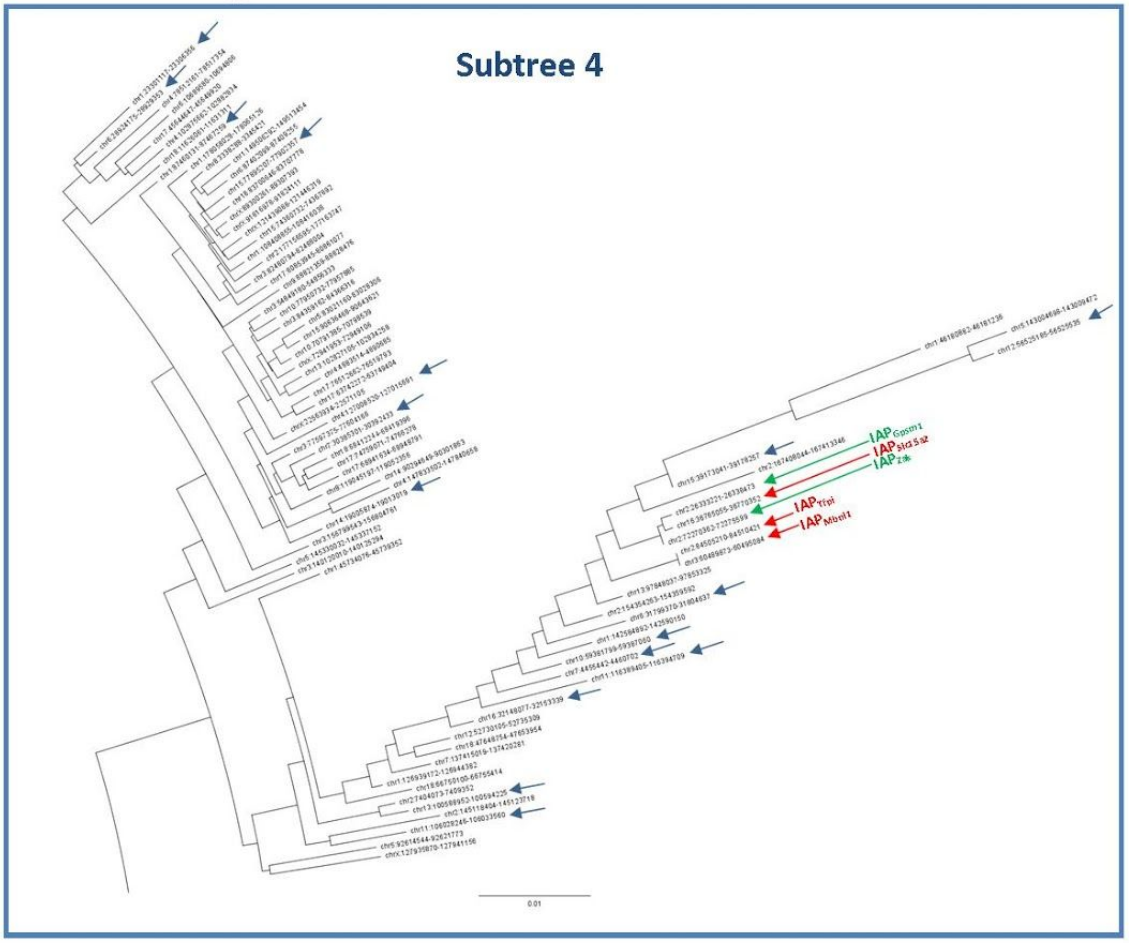


Figure 4.10. Neighbour-joining tree for IAPLTR1 subtype. *Metastable IAPLTR1 retroelements are distributed across 5 subtrees highlighted on the tree. Their location at subtrees is summarized in the table. Subtree 4 contains the majority of metastable IAPs. Metastable IAPs are pointed out by arrows on subtree 4 graph. Location of metastable $IAP_{Slc15a2}$, IAP_{Tfpi} and IAP_{Mbnl1} are highlighted by red arrows on the tree. IAP_{Zak} and IAP_{Gpsm1} are two the most closely related IAPs to $IAP_{Slc15a2}$ (green arrows)*

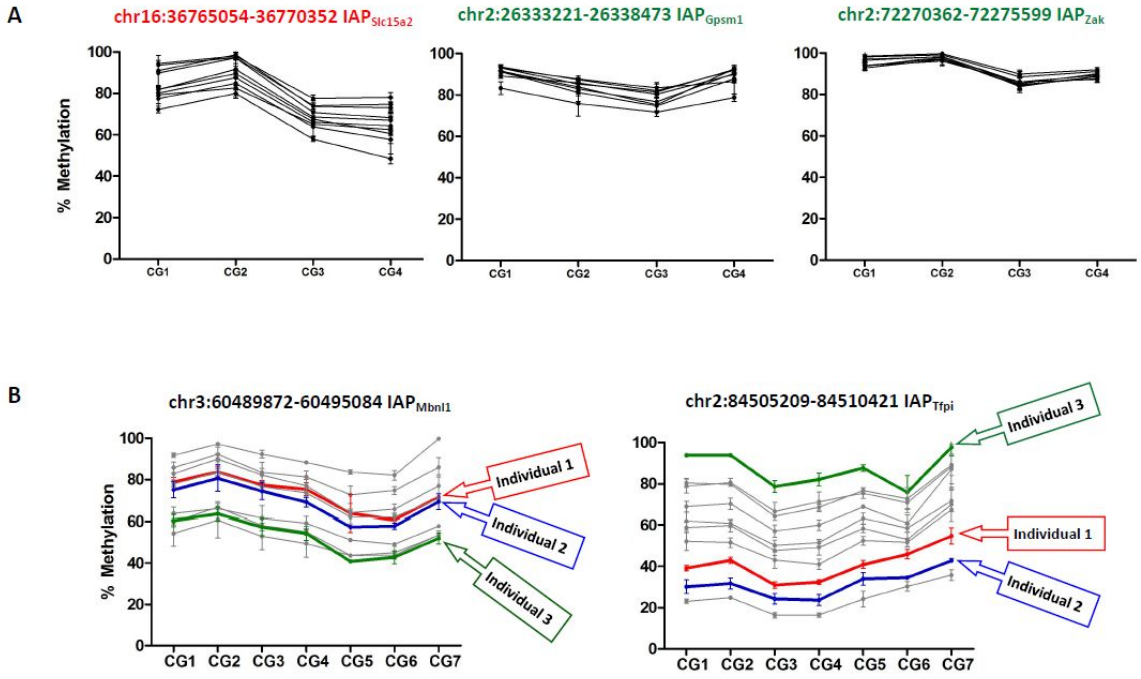


Figure 4.11. Methylation variation at closely related IAPs. *A) IAP_{Zak} and IAP_{Gpsm1} that closely related to metastable $IAP_{Slc15a2}$ are highly methylated in individual mice; B) Methylation levels at metastable IAPs are established in a locus-specific manner and are sequence-independent. IAP_{Tfpi} and IAP_{Mbnl1} have identical sequences. The same individuals (color-coded) have different methylation levels at these two loci.*

We next analysed to what extent specific features of metastable IAPs might be determined by their genomic location. The fact that transgenes were previously shown to be variably methylated between individuals suggest that genomic location might trigger the establishment of a variable methylation state at certain regions (Weichman & Chaillet, 1997; Sutherland et al., 2000; Kearns et al., 2000). Unlike imprinted regions, metastable IAPs are not organized in clusters. Furthermore, metastable IAP insertions appear randomly distributed along the genome. Nearly a quarter of them are intragenic while the majority are intergenic (**Supplementary table 4**). No correlation between metastable IAP location and topologically associated domains (TADs) was found.

Unfortunately, available ENCODE histone ChIP-seq datasets do not provide enough power to analyse histone marks at individual IAPs. However, it has previously been described that polymorphic IAP insertions are capable of spreading epigenetic marks to flanking genomic DNA and vice versa (Rebollo et al., 2011). I have analysed available ENCODE histone CHIP-seq datasets to explore the epigenetic profiles of the regions flanking metastable IAP insertions because they potentially can reflect the epigenetic state of individual IAPs. H3K9me3 is largely associated with IAP silencing (Matsui et al., 2010). However, no clear difference in H3K9me3 distribution was found between regions flanking metastable and control non-metastable IAPs (**Figure 4.12**).

CTCF has been suggested to play a role in the regulation of IAP-induced heterochromatin spreading on flanking genomic DNA (Rebollo et al., 2011). Analysis of CTCF ChIP-seq data showed a substantial CTCF enrichment at the genomic regions flanking metastable IAPs (**Figure 4.13**). Interestingly, CTCF peaks were often located next to the variable end of the IAP only (**Figure 4.14**). This enrichment was observed in different tissues (liver, kidney brain) as well as in ESCs, suggesting stable maintenance of CTCF binding at these locations throughout development. CTCF is crucial for preimplantation and postimplantation development (Phillips & Corces, 2009). In fact, CTCF-deficient oocytes cannot progress to the blastocyst stage following fertilization and CTCF KO embryos die before implantation (Wan et al., 2008; Moore et al., 2012). Of note, it has been suggested that the establishment of methylation levels at A^{vy} occurs during preimplantation development (Waterland et al., 2006; Blewitt et al., 2006). In addition, CTCF binding is known to be sensitive to DNA methylation, preferring unmethylated binding sites, and has been shown to inhibit Dnmt1 activity to prevent methylation of its binding domain (Bell & Felsenfeld, 2000; Zampieri et al., 2012). This raises the hypothesis that an interplay between IAP methylation and CTCF binding site demethylation/hypomethylation may be involved in the establishment and/or maintenance of variable methylation levels observed at metastable epialleles.

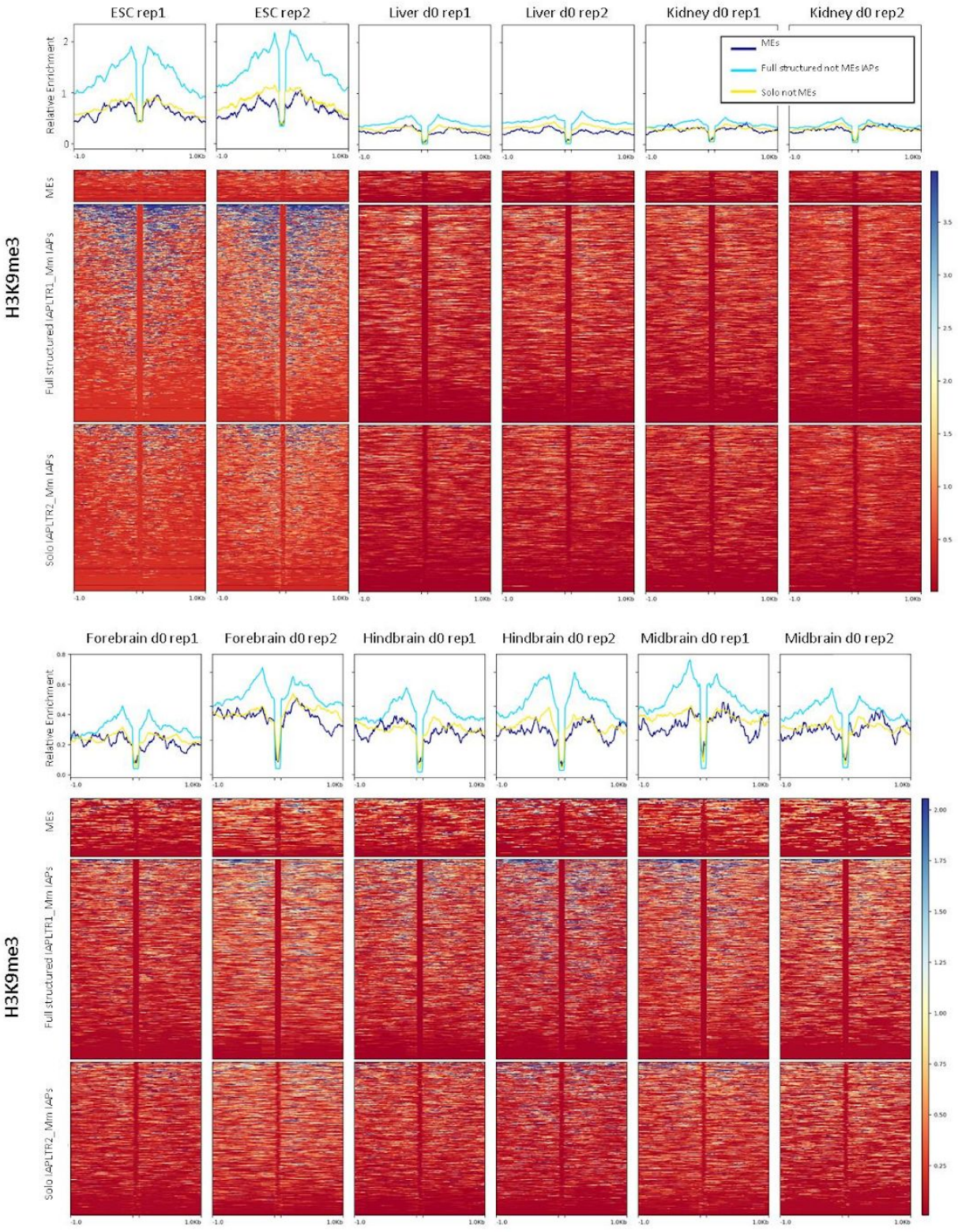


Figure 4.12. Relative H3K9me3 enrichment profiles of metastable IAP flanking regions. *Highly methylated full structure IAPs of the IAPLTR1_Mm and solo IAPLTR2_Mm subclasses serve as controls.*

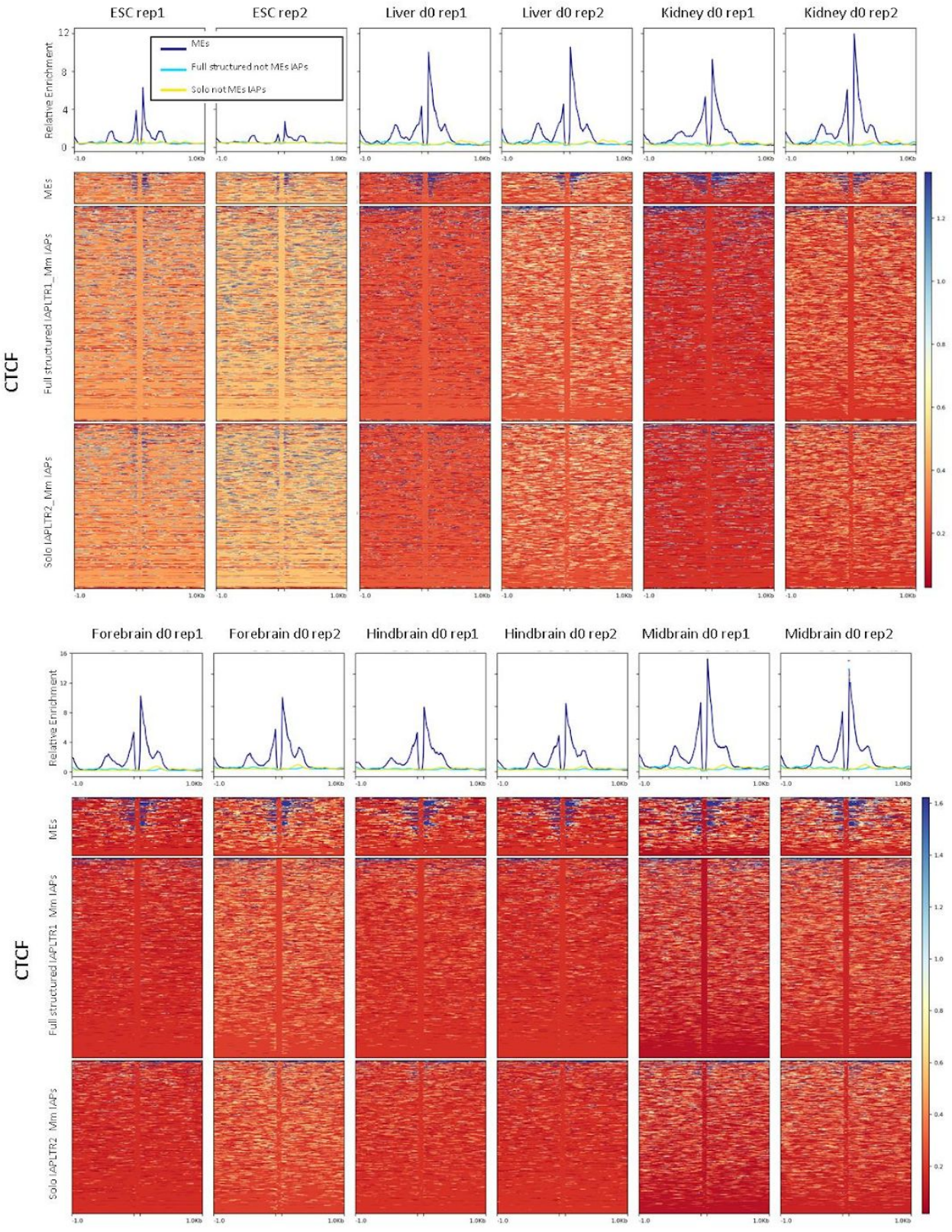


Figure 4.13. Relative CTCF enrichment profiles of metastable IAP flanking regions. *Highly methylated full structure IAPs of the IAPLTR1_Mm and solo IAPLTR2_Mm subclasses serve as controls.*

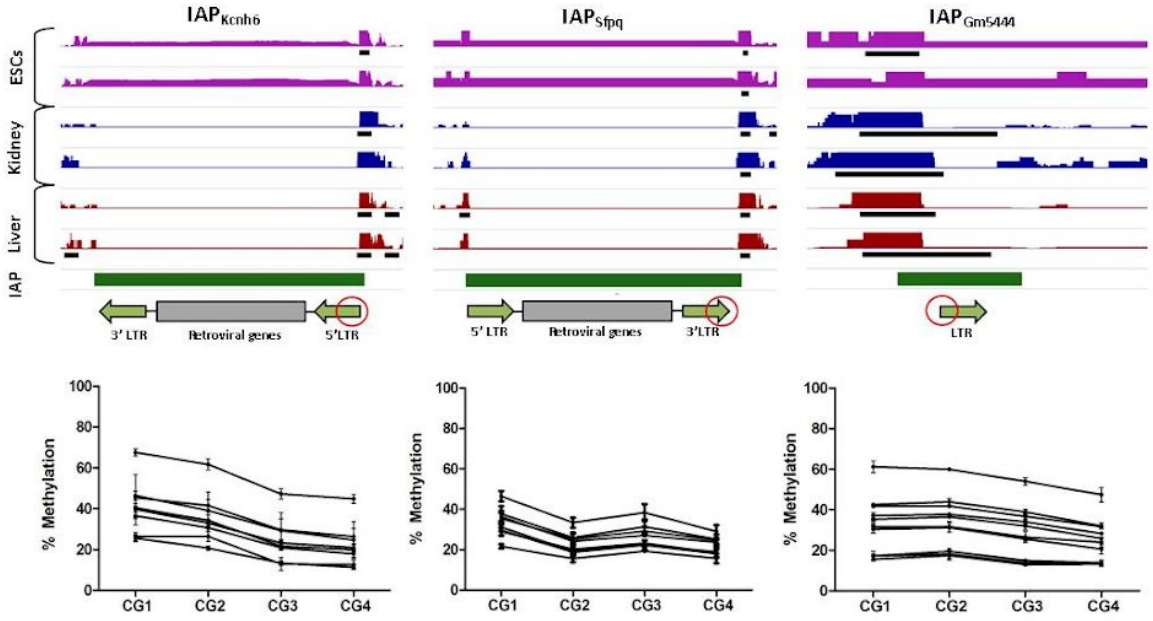


Figure 4.14. CTCF ChIP-seq profiles of 3 metastable IAPs. *A schematic representation of the IAP's orientation and structure is shown under the corresponding profile. CTCF peaks are found near the variably methylated end of the IAP (red circles). Experimental validation via pyrosequencing of these regions is shown.*

Strain-specific behaviour of conserved integrations.

A^{vy} was originally found in a C3H/Hej inbred mouse strain and $Axin^{fused}$ in a 129/Rr background suggesting metastable epialleles might exist in different strain backgrounds. While our set of metastable IAPs contain a lot of highly polymorphic insertions, some of them might be present in strains closely related to C57BL/6J. Identification and characterization of IAPs that are present in more than one strain would provide a useful model to study interindividual methylation variation, its origin, and its potential impact on transcriptome and phenotype.

To do this, I experimentally verified the presence of a subset of C57BL/6J metastable IAPs in the 129Sv background. Because 129Sv mice came from two different labs, I have checked the presence of the C57BL/6J IAPs in both 129Sv colonies. And used all of the available 129Sv individuals to analyse interindividual methylation variation at shared IAPs. Out of 20 C57BL/6J metastable IAPs assessed, 5 regions were amplified in both 129Sv colonies:$IAP_{Gm20110}$, IAP_{Nrros}, IAP_{Tfpi}, IAP_{Rab6b}, and IAP_{Diap3} (**Figure 4.15A**). Interestingly, $IAP_{Slc15a2}$, IAP_{Wdr1}, IAP_{Fam78b}, IAP_{Pgm1} showed a clear band in the 129Sv mouse coming from only one of the colonies. Interindividual

methylation variation of 3 IAPs was analysed in 129Sv individuals. One of them (IAP_{Tfpi}) was differentially methylated between 129Sv mice but the range of methylation variation was smaller than in C57Bl/6J mice (**Figure 4.15B**). However, the two other IAPs appeared to be highly methylated in 129Sv background (**Figure 4.15B**) and hence are not metastable epialleles. These results suggest that strain background might have an impact on the establishment of methylation at these loci and that the existence of metastability at the same IAP is not an essential function between strains of mice. To what extent this difference in behaviour might be explained by genetic polymorphism(s) existing between the two strains is not clear.

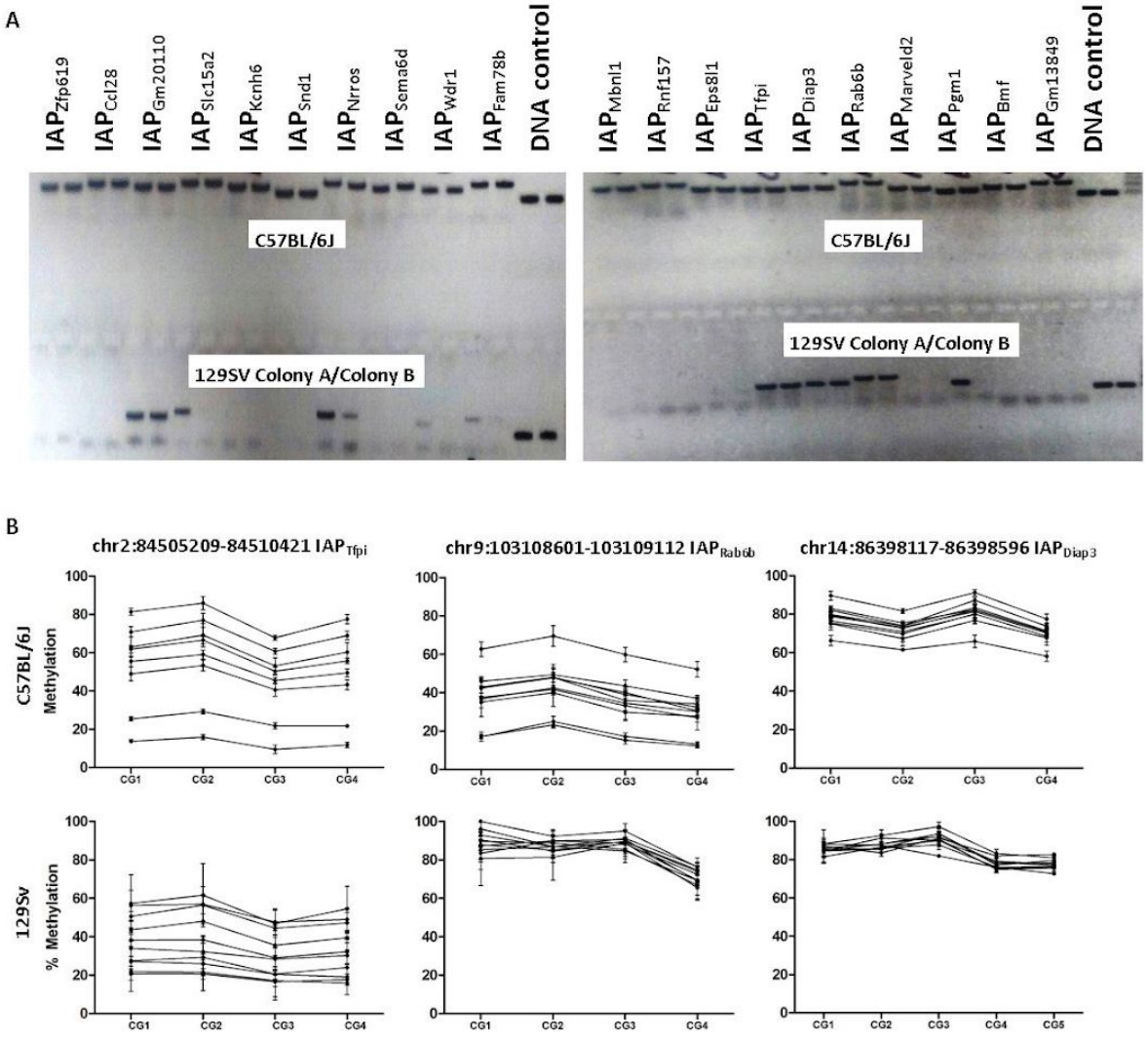


Figure 4.15. Strain-specific behaviour of conserved integrations. *A) Presence of C57BL/6J metastable IAPs in the 129Sv background.Each IAP was analysed in two 129Sv mice that were coming from two different colonies used for the analysis; B) Interindividual methylation variation at 3 IAPs that are present in C57BL/6J and 129Sv background*

Genome-wide identification of non-IAP-derived metastable epialleles

The biased screen of ERVs for metastable epialleles was reasonably successful. $MuLV_{Cep85}$ showed clear interindividual methylation variation at this step (**Figure 4.3**). Overall this suggests that the "ragged" methylation pattern might be specific for metastable IAPs and a good enrichment for metastability at ERVs more generally. However, the biased screen was based on visual assessment of methylation variation, which is often subjective and targeted, and hence not appropriate for a genome-wide screen of metastability at all ERVs in the C57BL/6J genome.

Therefore, the methylation model that was used for the unbiased genome-wide screen of metastable IAPs was used to conduct a similar screen of all ERVs. This method is robust, unbiased and avoids subjective estimation. The assembly of ERV coordinates was carried out separately for each ERV class including ERV1, ERVK and ERVL. The ERV segments were considered to be a part of a single element if they belonged to the same class, were on the same strand and located equal to or less than 100bp from each other. Then these elements were screened using the same genome-wide strategy that was used for the IAP screen. 208 ERV1, 760 ERVK and 174 ERVL candidates were identified using the threshold levels developed in the model. 11 ERV candidates were selected for experimental validation (in addition to those previously validated in **figure 4.3**). However, most candidates showed a modest range of methylation variation between individuals. Out of these 11 ERVs, only 1 additional MuLV candidate ($MuLV_{Pik3c3}$) showed interindividual methylation variation of about 11%. The other ERVs showed no interindividual methylation variation (**Figure 4.16**). Hence, while the identified ERV candidates show variable methylation levels between our WGBS-seq datasets, hardly any of them showed interindividual methylation variation during experimental validation. I suggest three possible explanations for the observed discordance between the genome-wide screen and its validation. First, less than 1% of those identified have been analysed experimentally and this may not be enough to draw a final conclusion. Second, the quantified methylation variation in WGBS-seq data might be a technical artefact related to the different CpG density across different ERVs. Finally, this data might reflect some other biological variation in pure non-cycling cell populations. However, our experiments confirm that this variation is unlikely related to metastable epiallele

properties. More experimental validation of these regions in pure B and T cells and in additional samples might confirm or disprove these possibilities.

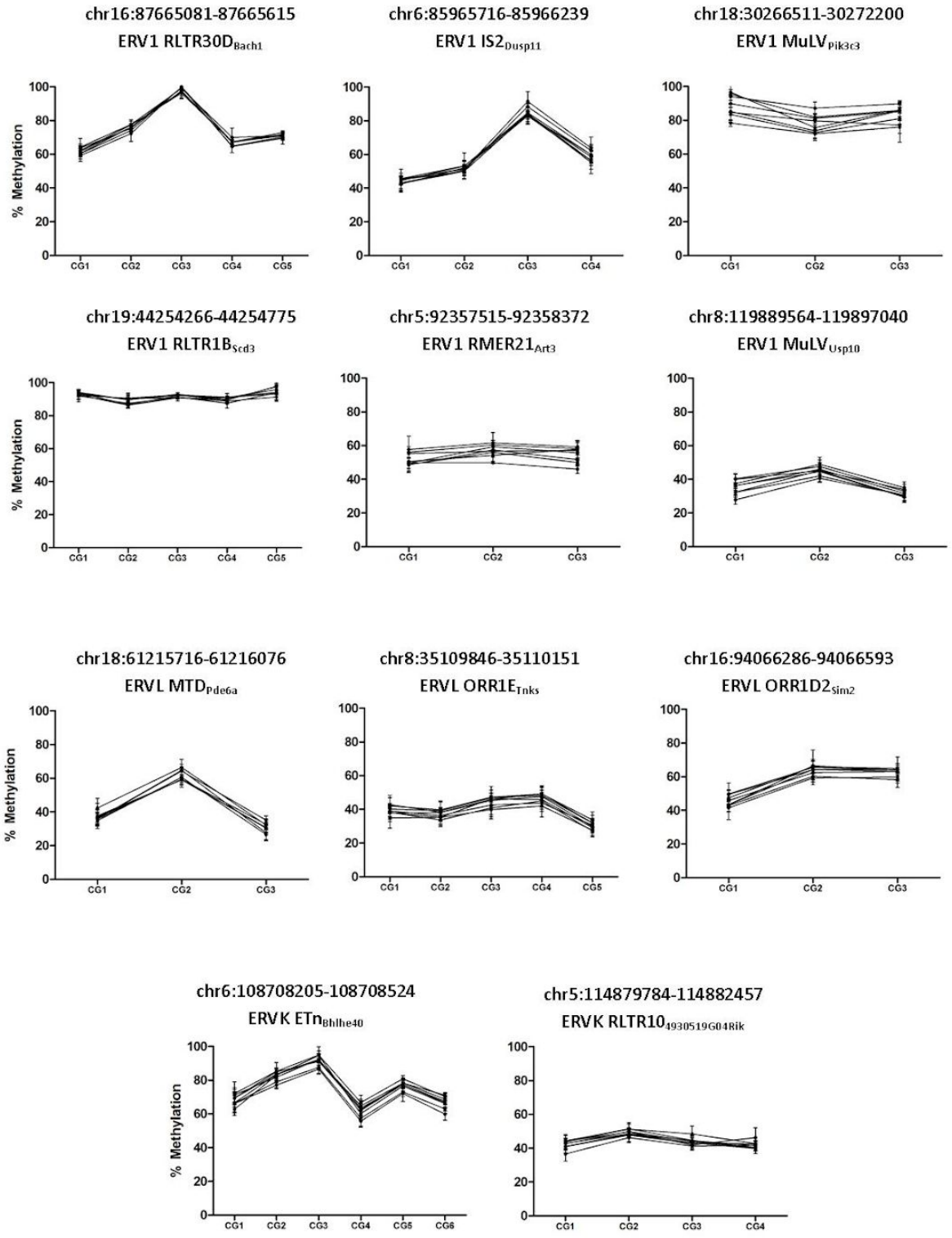


Figure 4.16. Validation of ERV metastable epiallele candidates.

Metastable IAP methylation dynamic during male and female germline development

The A^{vy} metastable epiallele is a commonly used model for transgenerational epigenetic inheritance in mammals. The exact mechanism that could explain transgenerational effects in A^{vy} mice remain unclear. It was shown that this metastable epiallele undergoes postfertilisation methylation reprogramming upon maternal transmission despite heritability of coat colour penetrance, while the coat color phenotype is not transmitted in a heritable fashion upon paternal transmission at least on the C57BL/6J background (Morgan et al., 1999; Blewitt et al., 2006). In contrast, offspring of both male and female of $Axin^{fused}$ animals transmit a memory of the penetrance of the tail kink phenotype on a C57BL/6J background. The methylation state in the sperm of $Axin^{fused}$ males is reported to reflect the methylation levels of the somatic tissues of each male, suggesting heritable maintenance of methylation variability levels (Rakyan et al., 2003).

Our metastable IAPs share many of similar characteristics with the A^{vy} and $Axin^{fused}$ metastable epialleles. While transgenerational studies are in progress, we decided to analyse recent published genome-wide methylation data to assess methylation dynamics at our metastable epiallele candidates during germline development.

This work was done in collaboration with Hiroyuki Sasaki and Kenjiro Shirane (Fukuoka, Japan). They blindly quantified average methylation levels for a set of IAPs (both metastable and control) during different developmental time points using published methylation datasets. These datasets were generated for samples in which multiple individuals were pooled together and hence can only provide a general picture of the behaviour of each allele. In particular, this does not allow us to estimate interindividual methylation variation at individual metastable IAPs during germline development. However, it has the potential to identify general trends that can be compared to known data about A^{vy} reprogramming during development (Blewitt et al., 2006; Rakyan at al., 2003). The IAP set for the analysis included 20 metastable IAPs and 14 non-metastable control IAPs. The obtained methylation data was further analysed by me.

Metastable IAPs are generally more lowly methylated than control IAPs in the inner cell mass (ICM). However at this time point, both metastable IAPs and controls show quite a lot of variation in their average methylation levels between individual IAPs (**Figure 4.17A**). While maternally inherited A^{vy} was shown to be hypomethylated at the blastocyst stage (0-20%) (Blewitt et al., 2006), and the methylation levels of our metastable alleles are consistent with this with a range from 0-60% and significantly different to the control levels of 40-80%. In the epiblast, all control IAPs were highly methylated varying within a 74-94% methylation range (**Figure 4.17B**). Metastable IAP methylation levels are significantly lower than controls with a methylation range of around 20-80% (**Figure 4.17B**). The lowest level of methylation (22.4%) in the epiblast was observed for IAP_{Rab6b}. IAP_{Snd1} had the highest average methylation (84%) in the epiblast.

Although it has been reported that some IAPs are resistant to reprogramming during germline development (Seisenberger et al., 2012), during male primordial germ cell development, all IAPs that we analysed go through a demethylation process (**Figure 4.17C,D,E**). with lowest levels at E13.5 when the PGCs have predominantly erased their epigenetic marks. Many more metastable epialleles have 0% methylation than the control IAPs. All become hypermethylated at the prospermatogonia stage (**Figure 4.17F**). Therefore, hypermethylation of all IAPs is retained throughout sperm maturation. This hypermethylation of IAPs in the male germline differs from what has been reported for A^{vy} (Rakyan et al., 2003) and has been experimentally confirmed in our lab (Tessa Bertozzi, unpublished results). Overall, the obtained results suggest that both metastable IAPs and non-metastable IAPs get reprogrammed during male primordial germ cells development.

Alternatively, during female primordial germ cell development, metastable IAPs lose methylation much quicker than control IAPs (**Figure 4.18**). Compared to male PGCs female PGCs appear slower both to demethylate and remethylate between E10.5 and E16.5. Metastable IAPs are less resistant to reprogramming than controls at these stages. While the majority of control IAPs still maintain about 40-50% methylation, most of metastable IAPs go below 20% methylation in female primordial germ cells at E13.5. This might reflect initially lower methylation levels at metastable IAPs at the early post implantation stages when the few cells destined to become

PGCs become set aside. In the case of A^{vy}, oocytes from a pseudoagouti mouse had higher methylation levels than oocytes from a yellow mouse (Blewitt et al., 2006). Our analysis does not allow us a direct comparison with A^{vy} since we do not know the extent to which methylation at the metastable IAPs in fully-grown oocytes is related to metastable IAP methylation in the donor female mouse.

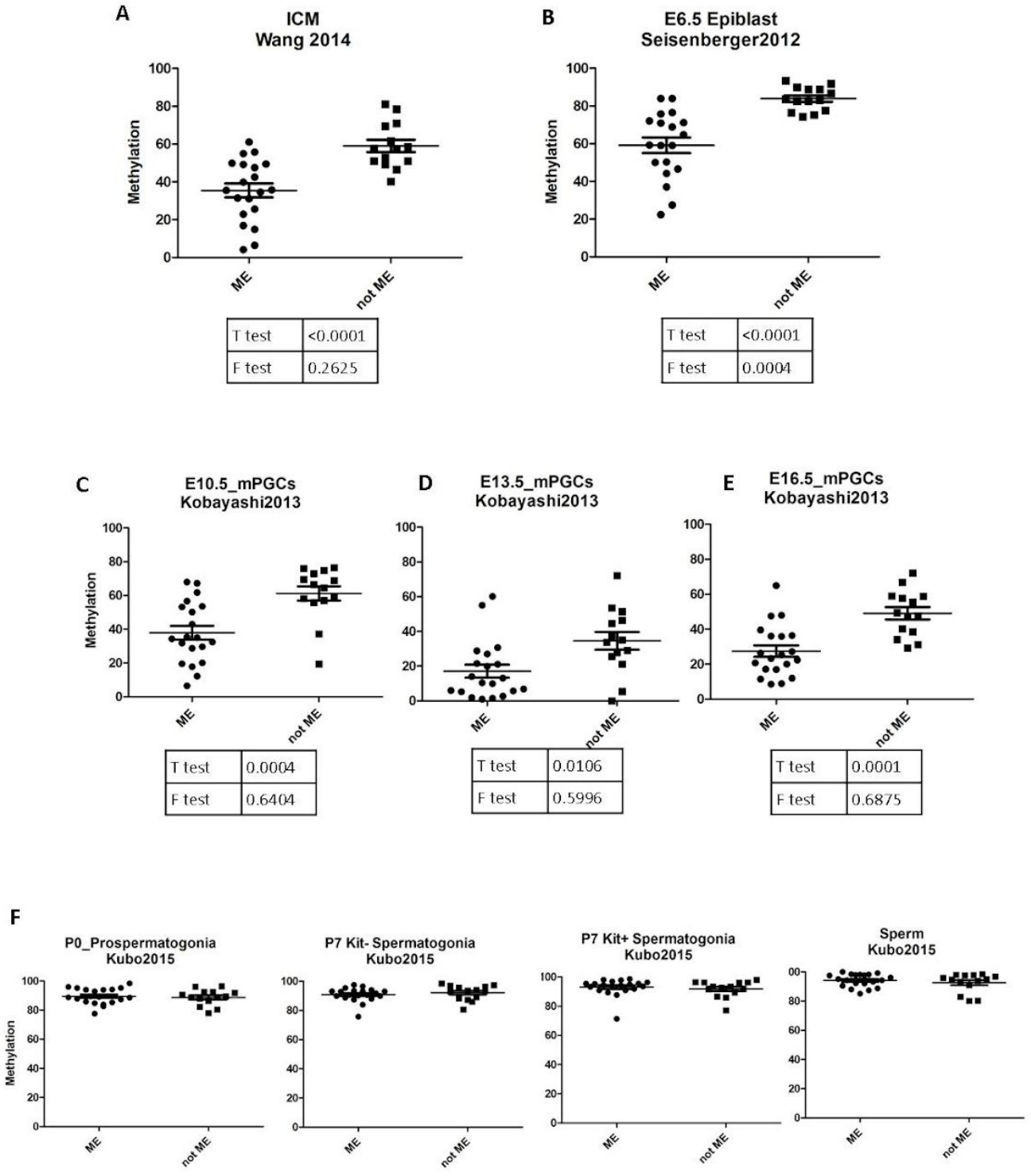


Figure 4.17. IAP methylation dynamics during male germline development. *A,B) Average methylation levels at metastable (ME) and non-metastable (not ME) IAPs in the ICM and epiblast; C-E) Methylation dynamic at metastable IAPs during male germline development. ICM – inner cell mass, mPGCs - male primordial germ cells; F) Methylation dynamic at metastable IAPs during spermatogenesis*

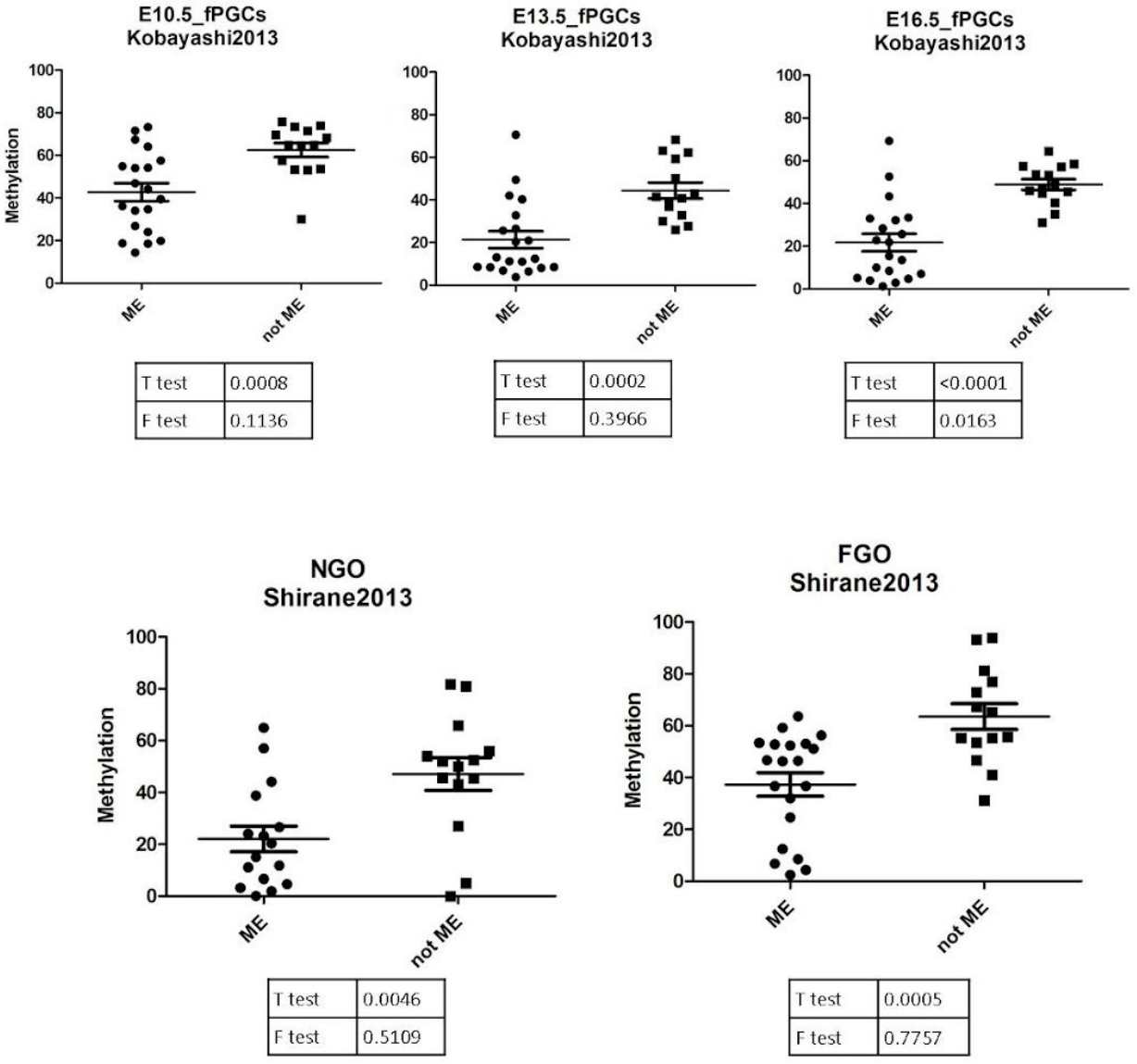


Figure 4.18. Methylation dynamic at metastable (ME) and non-metastable (not ME) IAPs during female germline development. *fPGCs – female primordial germ cells, NGO – nongrowing oocytes, FGO – fully-grown oocytes*

Summary and discussion

The proposed genome-wide screen for metastable epialleles identified over 100 IAP-derived metastable epialleles. The extensive experimental evaluation of these candidates was carried out and confirmed the screen strategy validity. An identical screen was conducted for other types of ERVs, however experimental analysis of randomly selected candidates showed modest or no interindividual methylation variation at analysed ERVs with the exception of 2 MuLVs. The optimisation of the IAP screen largely relied on extensive validation. Experimental validation of larger number of non-IAP ERVs might help to confirm or deny the relative absence of metastability at other ERVs. The outcome of biased screen suggests that the observed ERV methylation variation across the WGBS-seq biological replicates might reflect other biological phenomenon, such as cell-type specific differential methylation or other processes not related to methylation metastability.

There have been three screens aimed to identify novel metastable epialleles. The screen conducted by Weinhouse et al. (2011) identified 6 novel metastable epiallele candidates. These candidates were variably expressed between individuals with relatively low expression variation between tissues within individuals. Two candidates were analysed for interindividual methylation variation at their promoter or adjacent retroelement. Both of them showed a little methylation variation between individuals but with no clear consistency between tissues. This screen did not manage to confirm any relation between their variably expressed candidate genes and ERVs. None of these genes were identified in our screen or in that of two other groups (Ekram et al., 2012; Oey et al., 2015).

The metastable epiallele screen conducted by Ekram et al. (2012) was based on the presence of H3K4me3 at retroelement promoter regions and identified 143 epiallele candidates (Ekram et al., 2012). Out of them 99 candidates were ERVs: 40 ERVLs, 47 ERVKs including 6 IAPs, and 12 ERV1s. The screen conducted by Oey et al. (2015) identified 51 metastable ERVs, including 26 IAPs. The overlap between IAPs separately identified by these two screens and ours is illustrated on **figure 4.19A**. Validation in both studies was limited.

Only 3 ERVs (all of them are IAPs) have been identified by all three screens. The screen of Ekram et al. was based on the analysis of H3K4me3 data.The logic of this approach is unclear. H3K4me3 is a mark of active promoter regions and no H3K4me3 enrichment was found at lowly methylated A^{vy} individuals, limiting the possibility of discovering novel metastable epialleles based on the presence of this mark (Dolinoy et al., 2010). There is a possibility that the majority of ERVs identified by Ekram et al. are tissue-specific promoters rather than epialleles with interindividual epigenetic variation.

Oey et al. screened only CpGs that met 6 read coverage requirements reducing the total number of ERVs that were analysed. We identified 22 IAPs that were also identified by Oey et al. The differences in the screen designs might partially explain the absence of total overlap between the identified IAPs. The screen of Oey et al. used a different method to quantify methylation variation across their datasets and used a less stringent threshold for candidate selection (20%).

Out of 5 IAPs that were identified by the other screens and not in ours, 3 IAPs had computational variation score 20-25% in our data (**Table 8**). One of them is the previously reported $IAP_{Cdk5rap1}$. Validation of this IAP showed considerably smaller interindividual methylation variation in our datasets than that previously been reported (Druker et al., 2004; **Figure 4.19B**). IAP_{Phf14} showed no interindividual methylation variation upon validation (**Figure 4.19C**). The other three IAPs need to be tested.

We did not identify any specific IAP sequence features that could explain the acquisition of inter-individual methylation variation. In fact, even closely related IAPs were found to have different ranges of methylation variation. In some cases, near-perfect sequence identity was observed between a metastable and a non-metastable IAP. That said, we did find an enrichment in our set of candidates for young classes of IAPs (IAPLTR1_Mm and IAPLTR2_Mm). It is therefore possible that metastable epialleles represent evolutionarily young IAPs in the process of becoming epigenetically silenced. The majority of these young IAPs show absence/presence polymorphisms between mouse strains, which further supports for recent integration into the C57BL/6J genome.

Table 8. Summary of potentially metastable IAPs that were not identified in our screen

IAP coordinates	Nearest gene	Pre-validation computational variation score in our datasets	Screen where was identified
chr6:71955426-71960760	Polr1a	15.625	Ekram et al., 2012
chr6:10689579-10694806	Phf14	20.85417	Oey et al., 2015
chr2:154354262-154359592	Cdk5rap1	20.9375	Oey et al., 2015
chr19:5406362-5406822	4930481A15Rik	24.4375	Ekram et al., 2012; Oey et al., 2015
chr17:33981268-33981698	AA388235	2.5	Oey et al., 2015

The enrichment of CTCF binding that we observed in the flanking regions of metastable IAPs is the most consistent, and indeed the only parameter to date, that separates metastable epialleles from their hypermethylated counterparts. This introduces the possibility that insertion near de-repressed CTCF binding sites could provide retroviruses with a defensive mechanism against repression mechanisms. Overall, our findings suggest that the metastability observed at this small and epigenetically unusual subset of ERVs is subject to complex genetic and non-genetic factors.

We have conducted an analysis of methylation dynamics at the set of metastable IAPs during germline development using published methylation datasets. Unlike what has been reported for A^{vy}, many of metastable IAPs show intermediate methylation levels in the ICM but it is not clear whether this reflects a constant level, delayed loss of methylation or efficient acquisition after post-fertilisation reprogramming. The data shows that they get reprogrammed during spermatogenesis and are highly methylated in sperm hence some element of reprogramming must be taking place after fertilisation. Metastable IAPs seem to be partially resistant to demethylation during oogenesis but to a lesser extent than non-metastable IAPs. This analysis was limited since interindividual methylation variation could not be tested due to the unavailability of methylation data for individual mice. However, the data suggests

different behaviours of metastable IAPs during PGCs reprogramming compared to what has been reported for A^{vy}.

Analysis of the A^{vy} epiallele has also suggested that methylation levels in the male germline recapitulate that observed in the somatic tissues of a male despite this not being reflected in the penetrance of the coat colour phenotype upon paternal transmission. We have initiated experiments to re-assess the A^{vy} data using the more contemporary and quantitative pyrosequencing approaches that we have applied to our novel alleles and also to conduct heritability studies using a blind methylation readout rather than coat colour phenotypes, which introduce researcher bias into the scoring process.

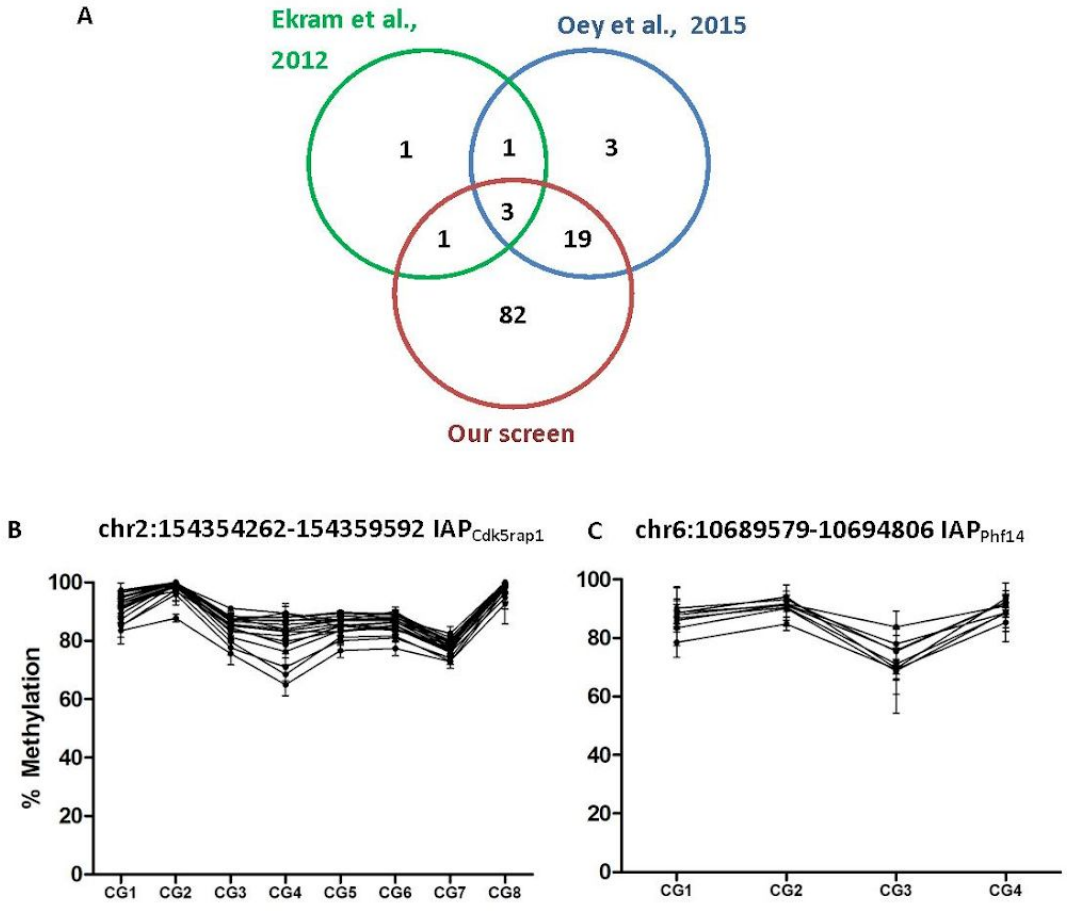


Figure 4.19.Comparison of genome-wide screens for metastable epialleles. *A) Overlap between IAPs identified by Ekram et al., Oey et al. screens and ours; B) Interindividual methylation variation at $IAP_{Cdk5rap1}$. Each line represents an individual; C) Interindividual methylation variation at IAP_{Phf14}. Each line represents an individual.*

Chapter 5

Impact of IAPs on transcription

Introduction

A^{vy} and $Axin^{fused}$ metastable epialleles have been shown to impact the transcription of nearby gene that contributed to their identification, to the correlation between methylation and expression, and has allowed the phenotypic characterisation of penetrance upon maternal and paternal inheritance. In both cases, the LTR of the IAP drives ectopic expression of neighbouring exons when it is hypomethylated. Intronic IAP insertions have been reported to be able to cause premature termination of gene transcription and some alternative splicing (Maksakova et al., 2006; Li et al., 2012). For example, the IAP insertion within Slc15a2 gene causes a premature transcriptional termination of this gene that leads to a nearly 40 fold decrease in the expression of the normal transcript (Li et al., 2012). Interestingly, we have identified this IAP as a metastable epiallele. Two alternative isoforms of Adamts13 gene were found in the C57BL/6J background (Zhou et al., 2007). Both of these isoforms are terminated within an intronic IAP. IAP-driven termination of Adamts13 expression is associated with a reduced activity of Adamts13 protein in this strain.

The exon skipping caused by an IAP insertion in the Reeler gene was previously described (Royaux et al., 1997). However, the examples of IAP-driven alternative splicing that are not associated with a premature termination within IAP are limited (Maksakova et al., 2006). While IAP activity is mostly reported in tissues with globally low methylation levels, it is still not clear to what extent IAPs can impact transcription in somatic cells. Metastable epialleles are particularly interesting models to study these effects as they maintain low methylation levels in somatic tissues in some individuals and hence have the potential to impact inter-individual phenotypic diversity.

BLUEPRINT RNA-seq data was used to explore the extent to which IAPs impact the transcriptome in pure populations of T and B cells. First of all I was interested in exploring the functional impact of metastable IAPs and the available data allows

exploration of the global impact of IAPs on transcription in T and B cells and on the cell-type specific activity of IAPs. B and T cells retain high global methylation levels similar to the majority of somatic tissues. IAP activity is largely determined by its methylation in somatic tissues. Expression levels of A^{vy} and $Axin^{fused}$ are inversely correlated with IAP methylation. Similar correlations between LTR methylation and activity at a promoter have been observed for LTR-LacZ transgenes (Dupressoir and Heidmann, 1996). Transgene expression was limited to the male germline where global DNA methylation was low. Available BLUEPRINT methylation data and ENCODE datasets were analysed to further understand epigenetic regulation of IAPs and their ability to impact transcription in somatic tissues.

Strategy design

RNA-seq datasets generated for the BLUEPRINT project and validated by me were used to conduct de novo transcriptome assembly that would contain previously unannotated transcripts in addition to those in existing database. Joseph Gardener used StringTie software to reconstruct de novo transcriptomes for six B cell RNA-seq biological replicates and six T cell RNA-seq biological replicates. The identified transcripts were intersected with the IAP coordinates generated from the metastable epiallele screen (**Figure 2.2**). There are 5 possible outcomes designed into this "overlap" strategy:

1) IAPs that cover the full transcript. These transcripts would be indicative of IAP transcriptional activity. IAPs are highly repetitive regions and IAP derived RNA-seq reads might be aligned to multiple regions hence multiply aligned reads were excluded from the transcriptome reconstruction. Therefore, IAPs that would be picked up in this group would not represent the full list of active IAPs in the genome. For this reason, this group was not further analysed.

2) IAPs that fall entirely within exonic regions. This group might include strain specific gene isoforms if IAPs are strain specific. However, the absence of comparative analysis of gene expression between different strains does not

allow the assessment of the functional impact of this group of IAPs on gene expression.

3) IAPs that fall entirely within intronic regions. The majority of these IAPs are not functional and indeed are silenced in somatic tissues. There is a possibility that they might affect gene transcription in cis and trans by providing enhancer and transcription factor binding site sequences or have an impact on alternative splicing. However, this impact was extremely hard to assess or confirm within my PhD project timeline and hence has been excluded from further analysis.

4) IAPs that overlap transcription initiation or termination sites. This group would contain IAPs that can initiate and/or terminate transcription of adjacent genes. Due to strategy design, this group might pick up IAPs that drive expression of small RNAs as well. If any of our metastable IAPs can drive ectopic expression of nearby genes similarly to A^{vy} or $Axin^{fused}$, they will be picked up in this group.

5) IAPs that partially overlap exonic sequences of transcripts that were initiated and terminated outside of the overlapping IAP. This group would contain IAPs that provide alternative splice sites in B and T cells. So far, there is no evidence that metastable epialleles provide alternative splice sites for gene expression. There is, however, growing evidence that DNA methylation is involved in splicing regulation (Maor et al., 2015). This group helps to assess the impact of DNA methylation on IAP driven splicing alterations.

Due to time constraints, my further analysis was concentrated on the results of the 4th and 5th groups of IAPs and IAPs identified from these groups were further annotated and analysed in more details. The Ensembl gene database was used to annotate overlapping transcripts. A transcript would be considered to be a gene isoform if it contains a full or partial exon sequence of an annotated gene and is transcribed in the same direction.

Characterization of IAP-driven initiation and termination events

This strategy as applied to the BLUEPRINT datasets and included all IAPs in C57BL/6J identified 142 IAPs that are involved in the initiation or termination of de novo assembled transcripts. 32 of these IAPs have been identified as metastable epialleles and will be discussed later (**Supplementary table 4**).

For the remaining 110 IAPs, the total number of terminating and initiating transcripts across all 12 transcriptomes were calculated (**Supplementary table 5**). 100 IAPs were associated with both initiation and termination (**Figure 5.1A**). 7 IAPs were overlapping only termination sites of transcripts (**Figure 5.1B**). 3 IAPs were associated only with initiation (**Figure 5.1C**).

Characterization of identified IAPs. 91 identified IAPs are full structure IAPs that contain 5' and 3' LTRs that flank internal IAP sequence (**Figure 5.2A**). Only 6 of the 110 IAPs were solo LTRs suggesting that these smaller elements have a lesser impact on transcription. The other IAPs were truncated from their 5' and/or 3' end.

An enrichment of young IAP type (IAPLTR1_Mm) was observed among IAPs that are involved in the initiation and termination of transcription (**Figure 5.2B**). Moreover, the majority of these insertions are show absence/presence polymorphism between strains (**Figure 5.2C**). Only 15 IAPs did not contain any structural variants between 18 inbred strains. This data suggests that IAPs that provide alternative promoters or termination sites are evolutionary young intact IAP insertions. Moreover, the enrichment of polymorphic insertions suggest that these IAPs might contribute to the phenotypic divergence of inbred mouse strains.

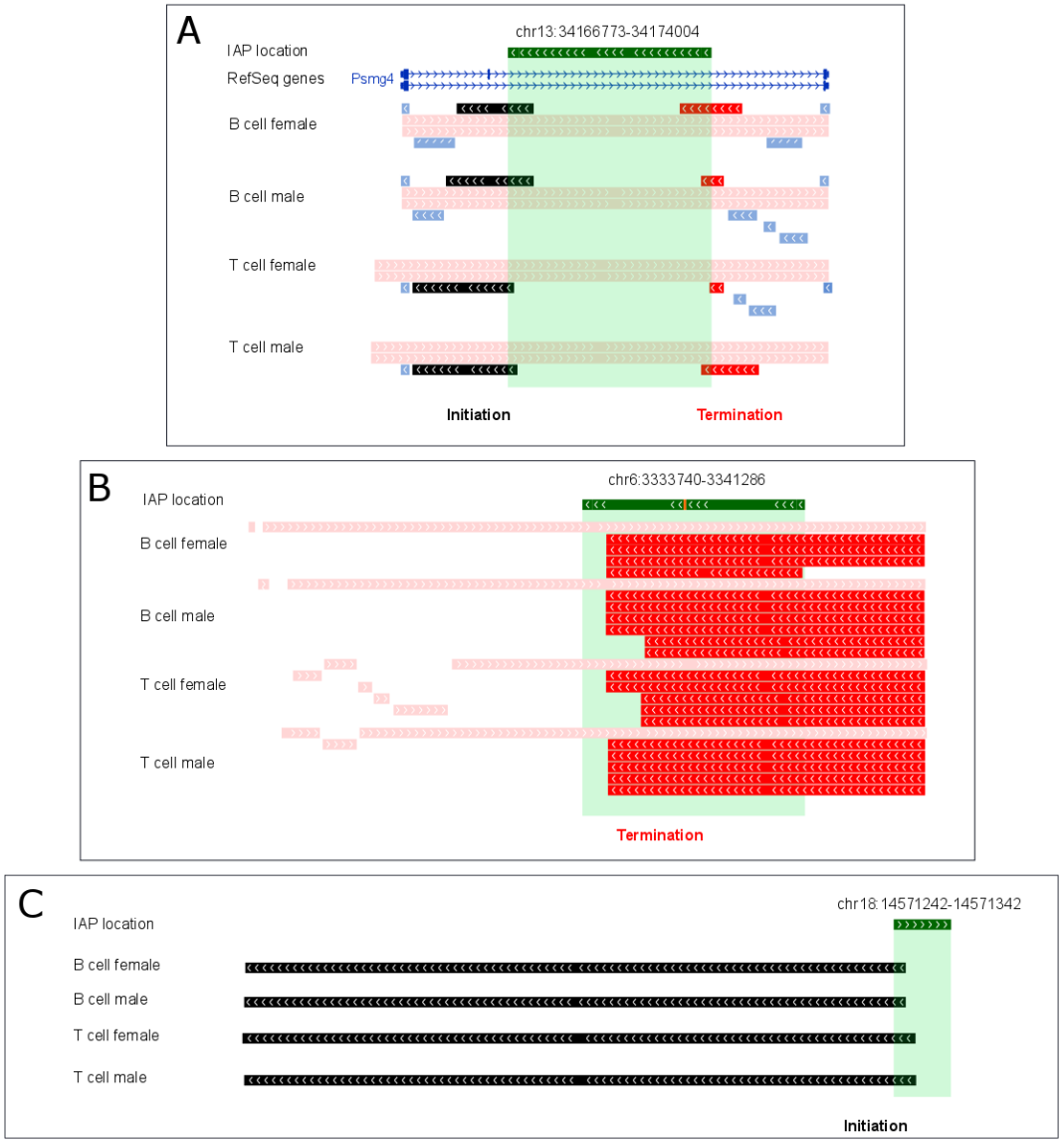


Figure 5.1. Screenshots of IAPs impacting transcription. *A) IAP (chr13:34166773-34174004) is highlighted in green. Transcripts that are highlighted in red are terminated within the IAP and transcripts that are coloured in black are initiated within the IAP. Psmg4 transcripts and other transcripts that do not overlap the IAP are coloured in pale pink and pale blue respectively; B) IAP (chr6:3333740-3341286) that is highlighted in green is involved in termination of transcripts coloured in red. Other transcripts are coloured in pale pink; C) Transcripts coloured in black are initiated within an IAP (chr18:14571242-14571342) that is coloured in green.*

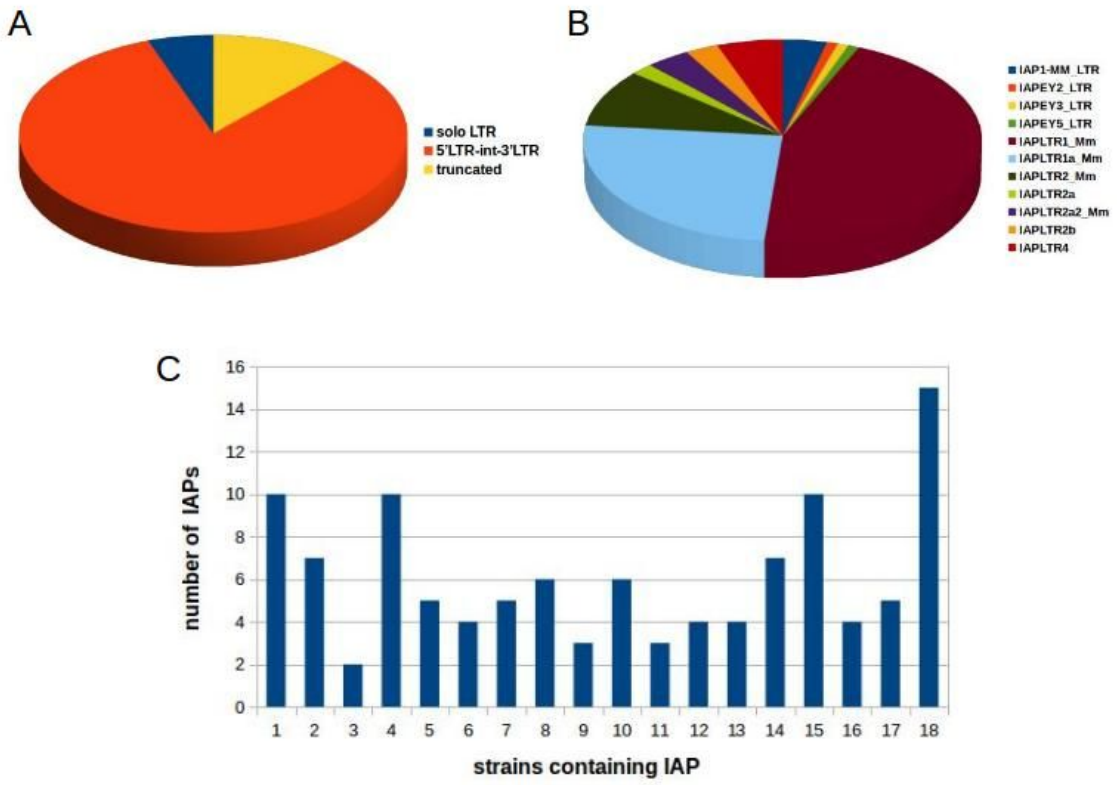


Figure 5.2. Characterization of IAPs involved in transcription termination and/or initiation. *A) Pie chart distribution of identified IAPs based on IAP structure; B) Enrichment of IAP subtypes in the identified set of IAPs; C) Distribution of IAPs involved in transcription termination and/or initiation across 18 mouse strains*

Characterization of identified transcripts. The majority of identified transcripts that were initiated or terminated within IAPs are unannotated sequences of 100-500 bp long (**Figure 5.1A**). The presence and length of some transcripts are variable between RNA-seq biological replicates. However, some of them were constantly present in all de novo transcriptomes. While the function of these transcripts is not clear, some of them might be noncoding RNAs or pseudogenes.

19 IAPs were found to overlap with cell-type specific transcripts: 9 IAPs associated with B cell-specific transcripts and 10 IAPs with those from T cells. Transcripts for only 27 IAPs contained exonic sequences of annotated genes. Moreover, 11 IAPs were initiating the transcription of annotated transcripts. The rest were involved in premature termination of transcription that was initiated at a known-gene promoter. However, the majority of IAP initiation or termination events associated with these genes were not consistent between replicates. Sometimes transcripts were present in

1 of the transcriptomes only. Only 3 genes were found to be initiated within an IAP in at least 9 transcriptomes: Pde3b, Poc1b, Bckdhb. Similarly termination of 3 genes within an IAP (Gm20559, Senp2, Gpbp1) was consistent between replicates.

To date, DNA methylation is considered by some as the major epigenetic mechanism for IAP silencing in somatic tissues (Bestor, 1998). However, according to our WGBS-seq data, the majority of IAPs that are involved in transcription initiation and/or termination are highly methylated in B and T cells hence methylation is not necessarily limiting their activity. Out of 110 identified IAPs, 7 IAPs were differentially methylated between B and T cells being low in T cell methylomes. However only 4 of them overlapped with cell-type specific transcripts (**Figure 5.3A**). 3 differentially methylated IAPs were overlapping transcripts in both cell types (**Figure 5.3B**). 15 IAPs that were associated with cell-type specific transcription were highly methylated in both cell types (**Figure 5.3C**).

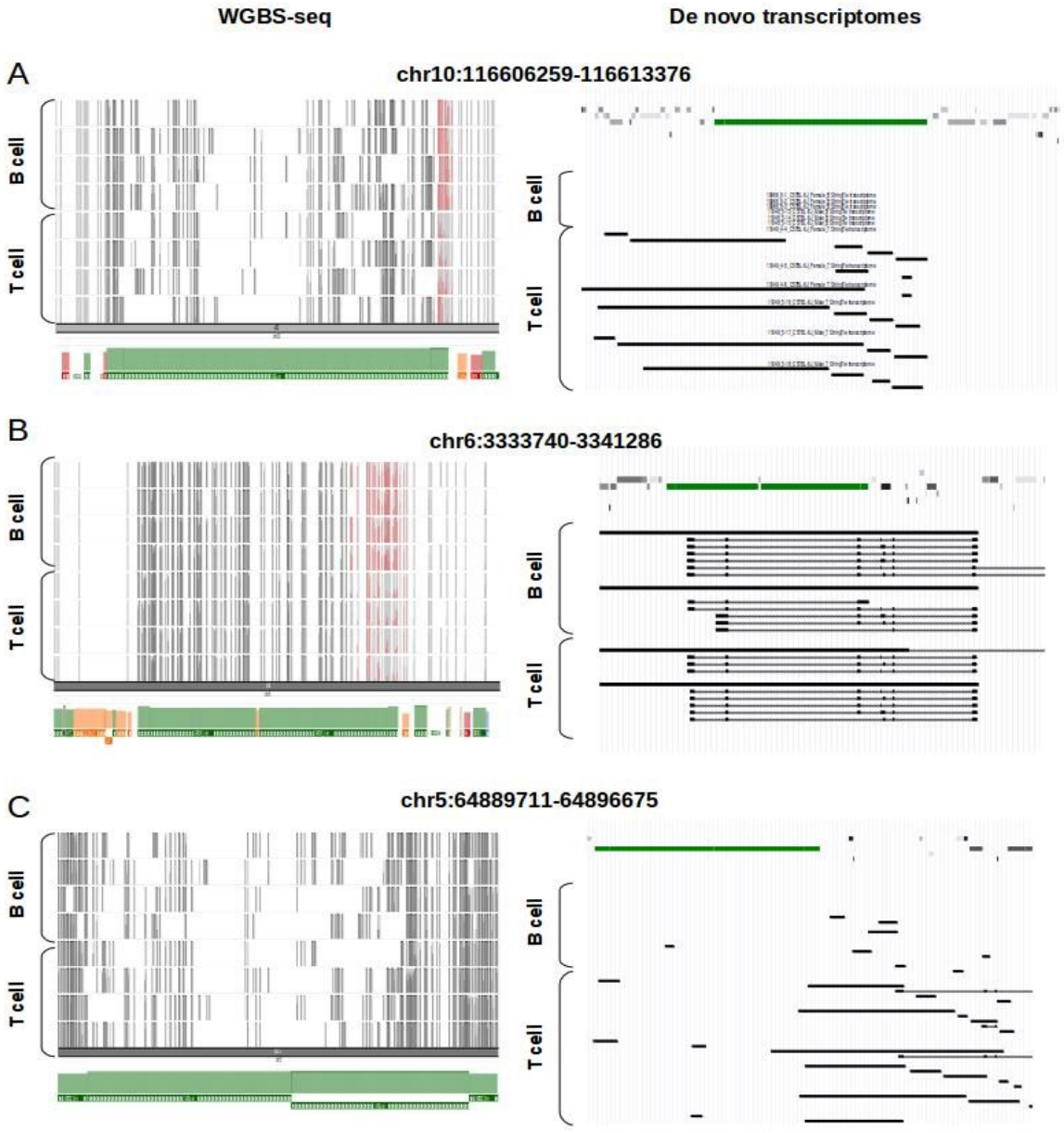


Figure 5.3. Relation between IAP methylation and its ability to impact transcription initiation and/or termination. *Left screenshots represent IAP methylation in WGBS-seq datasets. IAP methylation is highlighted in black and red (if the regions is differentially methylated between B and T cells). Right screenshots represent transcripts initiated or terminated within a corresponding IAP (green). IAP coordinates are on the top of the screenshots. A) Differentially methylated IAP (lowly methylated in T cells, highly methylated in B cells) overlaps with T cell specific transcripts; B) Similar transcripts are present in both cell types despite differential methylation of an overlapping IAP lowly methylated in T cells, highly methylated in B cells); C) IAP that is highly methylated in both cell types overlaps with T cell specific transcripts*

Characterization of metastable epiallele impact on transcription. Of the 142 IAPs with an influence on transcription initiation or termination, 32 IAPs associated with transcription termination and initiation were metastable epialleles generated from my screen. Half of the identified IAPs have a full structure with IAPLTR1_Mm flanking sequences and hence these 32 IAPs are enriched for the full structure IAPs.

Nine metastable IAPs were associated with initiation or termination of previously annotated genes. Five of these genes were initiated within an IAP in at least one biological replicate: Bmf, Slc15a2, Eps8l1, 2610035D17Rik, AK134158. We experimentally validated the expression levels of four of these genes and compared their expression levels with IAP methylation within individuals (expression of AK134158 needs to be analysed). Eps8l1 IAP-driven transcripts were found in all 12 RNA-seq replicates. Its expression was variable between individuals in the 4 analysed tissues (brain, liver, kidney, and spleen) and IAP methylation was inversely correlated with Eps8l1 expression in all tissues (**Figure 5.4**). The long non-coding RNA, 2610035D17Rik, was initiated within an IAP in six B cell and five T cell transcriptomes. Similar to Eps8l1, it showed higher expression in individuals with lower IAP methylation (**Figure 5.5**). Expression of downstream exons of Slc15a2 from the metastable IAP promoter was observed in two B cell and five T cell replicates (**Figure 5.6**). No transcripts actually initiating from the Slc15a2 promoter were found in our datasets but we found that expression of downstream exons was inversely correlated with IAP methylation in splenic tissues where Slc15a2 gene is normally not expressed. Similarly, the expression of only Slc15a2 downstream exons was observed in liver (**Figure 5.6B**). However no statistically significant correlation between gene expression and IAP methylation was found (**Figure 5.6C**). Slc15a2 is highly expressed from its own promoter in kidney and brain. However, we did not find a statistically significant correlation between expression of downstream and upstream exons in these tissues. The expression of downstream exons was lower comparing to upstream exons in these tissues probably due to premature termination within the IAP that has previously been described for this gene (Li et al., 2012; **Figure 5.6**). At the Bmf/Bub1b locus, no correlation between IAP methylation and Bmf expression was found (**Figure 5.7A**). However, correlation between IAP methylation and

expression of Bub1b gene located downstream of this IAP was found in brain tissues (**Figure 5.7B**).

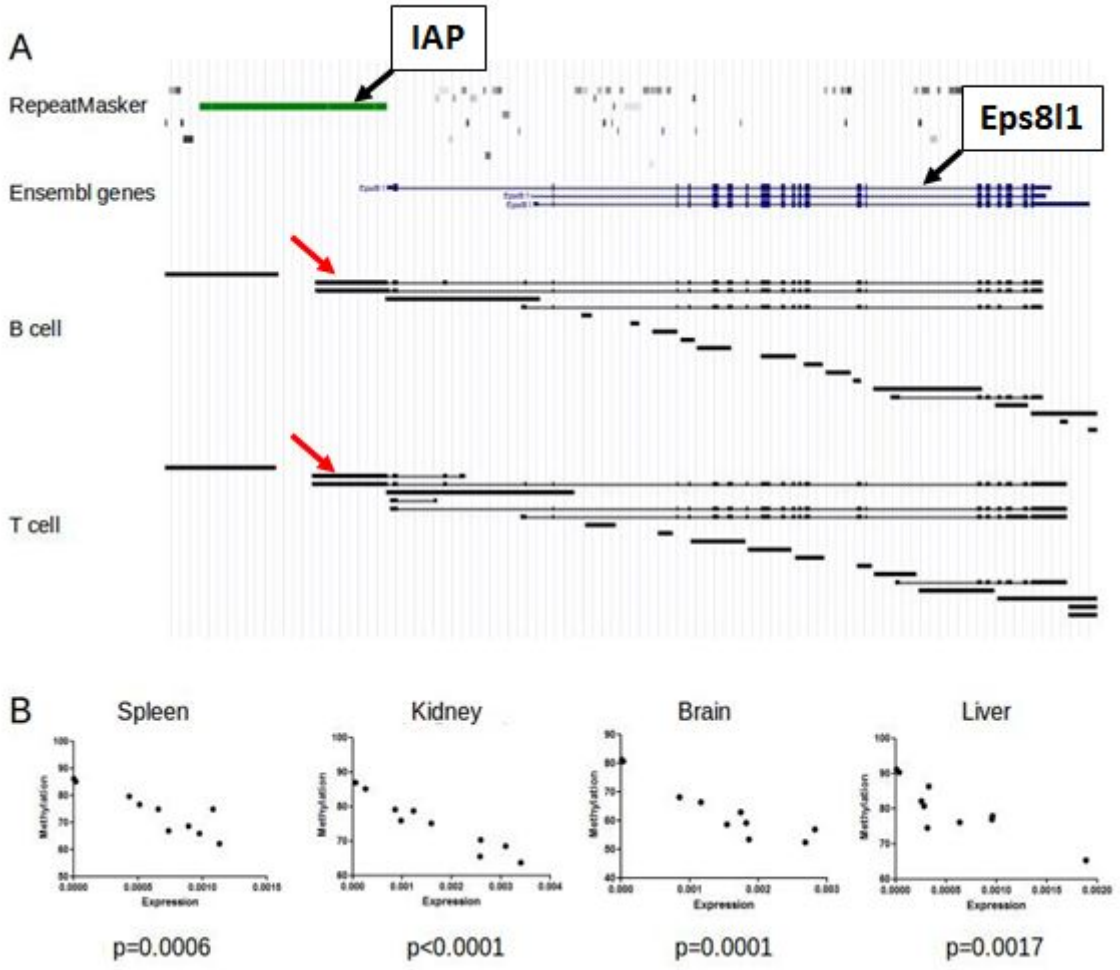


Figure 5.4. Validation of Eps8l1 expression. *A) Eps8l1 transcripts (pointed by red arrows) initiated within an IAP (highlighted in green) in B and T cell de novo transcriptomes. Eps8l1 annotation is shown at "Ensembl genes" track; B) Correlation between Eps8l1 expression and IAP methylation in 4 analysed tissues (two-tailed Pearson).*

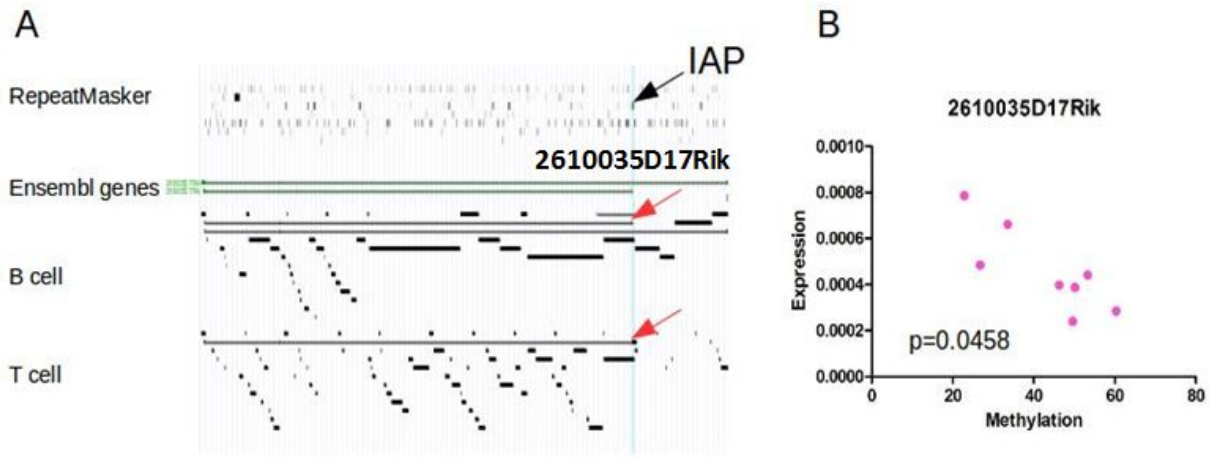


Figure 5.5. Validation of 2610035D17Rik expression. *A) 2610035D17Rik transcripts initiated within an IAP (black arrow) in B and T cell transcriptomes are pointed by red arrows; B) Correlation between IAP methylation and 2610035D17Rik expression in liver (two-tailed Pearson). Each dot represents an individual.*

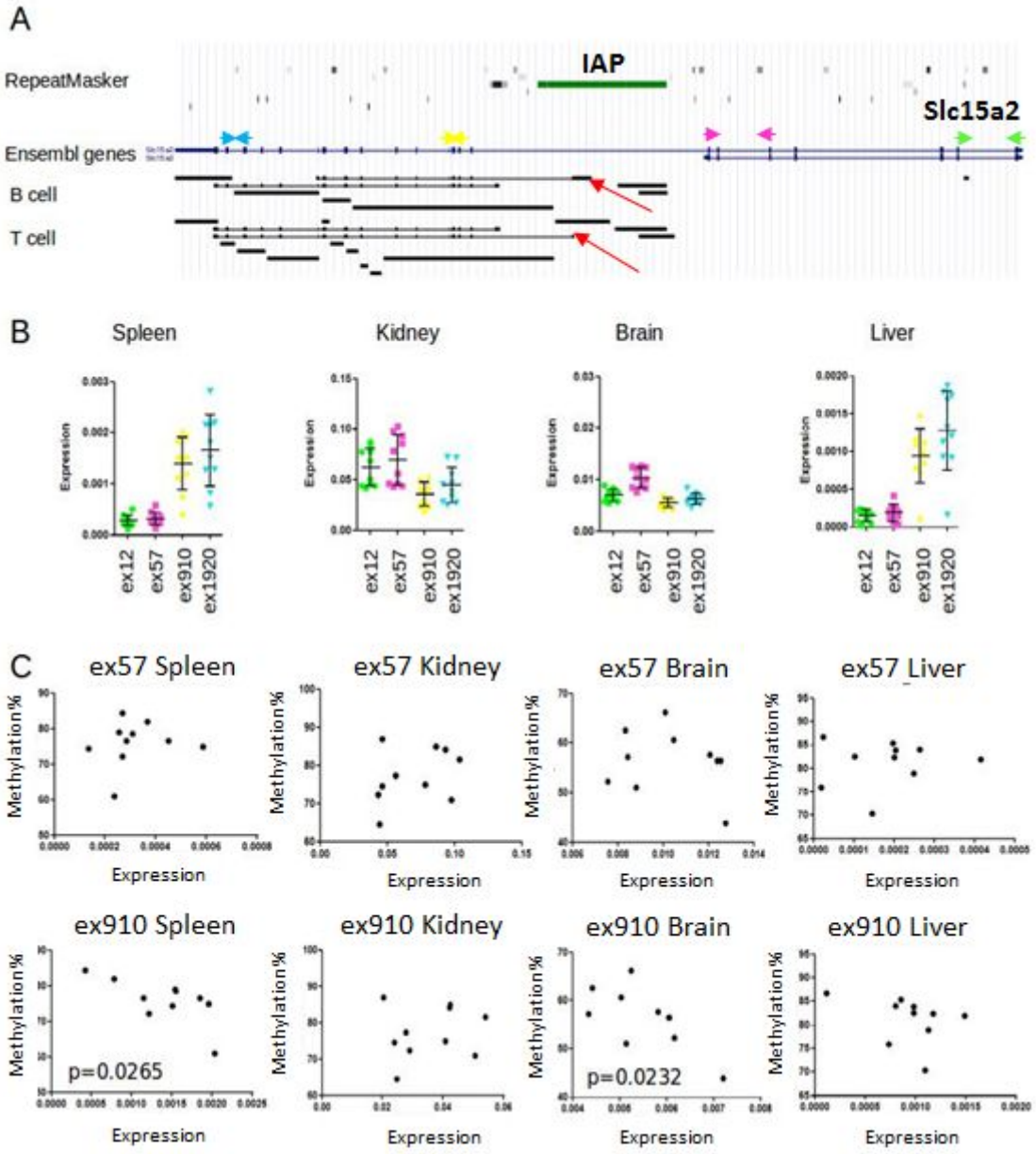


Figure 5.6. Validation of Slc15a2 expression. A) Slc15a2 transcripts (pointed by red arrows) that are initiated with an IAP (highlighted by green). Slc15a2 annotation is *is shown at "Ensembl genes" track. Primers location is indicated by arrows. Primer colors correspond to the colors in panel B; B) Expression of Slc15a2 upstream (ex12 and ex57) and downstream (Ex910 and ex1920) exons in analysed tissues; C) Expression of Slc15a2's downstream exons is inversely correlated with IAP methylation in spleen and brain tissues (two-tailed Pearson). Each dot represents an individual.*

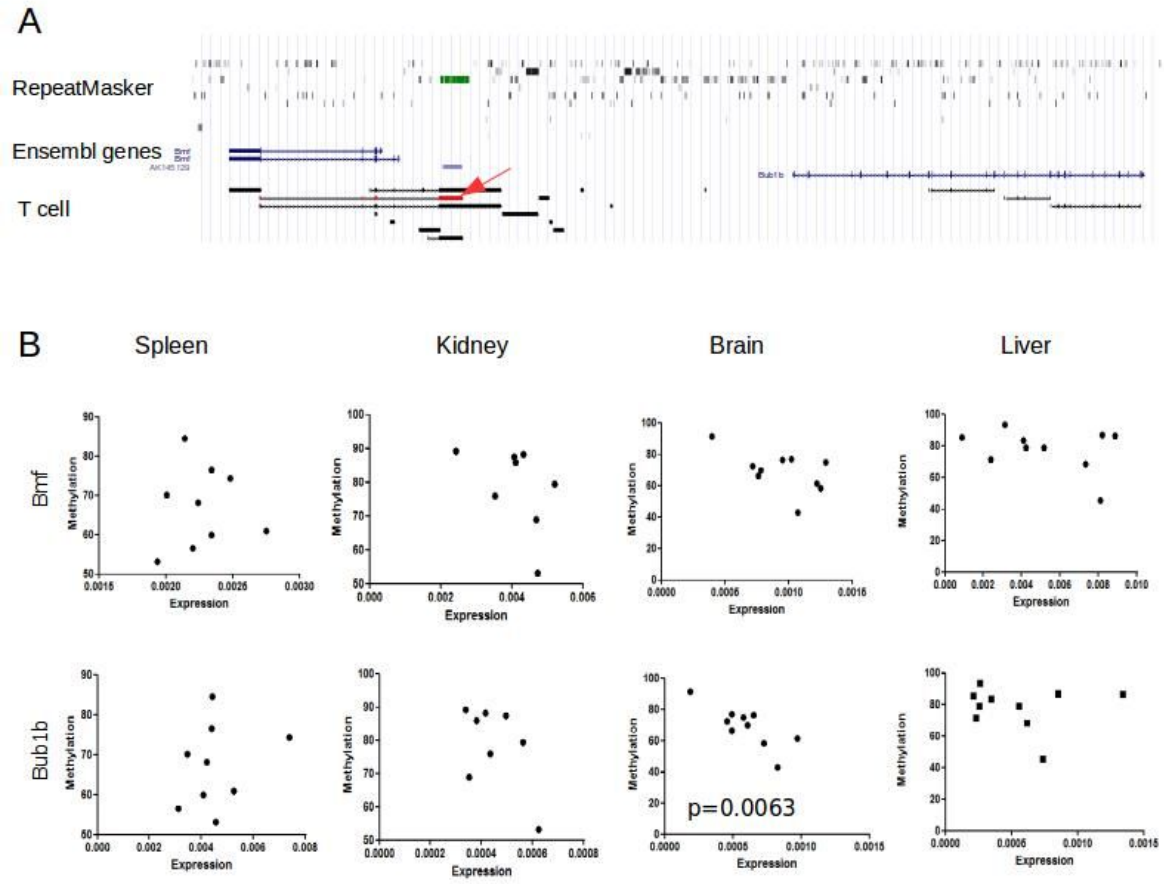


Figure 5.7. Validation of Bmf and Bub1b expression. *A) Bmf transcript (pointed by red arrow) is initiated at IAP (colored in green at RepeatMasker track); B) No correlation between Bmf expression and IAP methylation was found in analysed tissues. Expression of Bub1b is inversely correlated with IAP methylation in brain tissues (two-tailed Pearson). Each dot represents an individual.*

To date, no tissue-specific activity of metastable epialleles has been described in the literature. Tissues-specific activity of a metastable IAP was experimentally confirmed for IAP_{Tfpi}, which is adjacent to the Gm13710 transcript (**Figure 5.8**). While there was no IAP-driven transcription of Gm13710 found in the B and T cell transcriptomes, a correlation between Gm13710 expression and IAP_{Tfpi} methylation was observed in liver (**Figure 5.8**). This suggests the presence of tissue specific transcription factors that act in concert with the a tissue-wide epigenetic state at IAP_{Tfpi} to regulate the transcription of Gm13710. No expression of this gene was observed in CAST/Eij individuals that lack this IAP insertion further supporting IAP-driven expression of Gm13710 in C57BL/6J individuals. We found termination of 4 genes within metastable IAPs. In 3 cases the terminating transcripts were found only in the T cell

samples. Therefore, once more, the cell-type specific presence of terminating transcripts suggests tissue-specific functional roles of metastable IAPs and the involvement of cell-type specific regulators. The relationship between IAP methylation and transcription termination requires experimental analysis.

The rest of identified transcripts that were terminating or initiating within metastable IAPs are previously unannotated. Most of them are 100-500 bp sequences that might be non-coding RNAs or pseudogenes. However, low expression levels of the majority of these IAP-derived transcripts makes experimental validation challenging. The function of unannotated transcripts is not clear. However, the functions and roles of the annotated transcripts are summarized in **table 9**.

I have compared the enrichment of CTCF and histone marks at flanking regions of metastable IAPs that do not interfere with any transcriptional events and those that do. CTCF binding in the IAP proximity was found in both groups suggesting that the presence of CTCF does not correlate with any transcription initiation or termination properties of the IAP and further implicating CTCF in the establishment and/or maintenance of methylation metastability rather than a particular role in initiation or termination (**Figure 5.9**).

No clear enrichment of H3K36me3 was observed in transcription initiation or termination cases. However, metastable IAPs involved in transcription termination and initiation were flanked by H3K27ac, H3K9ac and H3K4me3 enriched genomic regions compared to those that are not (**Figure 5.9**). These histone marks are associated with active promoter and enhancer regions. While it is unclear whether the observed chromatin state at these boundary regions is IAP-induced or was present before IAP insertion, no enrichment of similar marks was found at the flanking regions of metastable IAPs that do not interfere with transcriptional events.

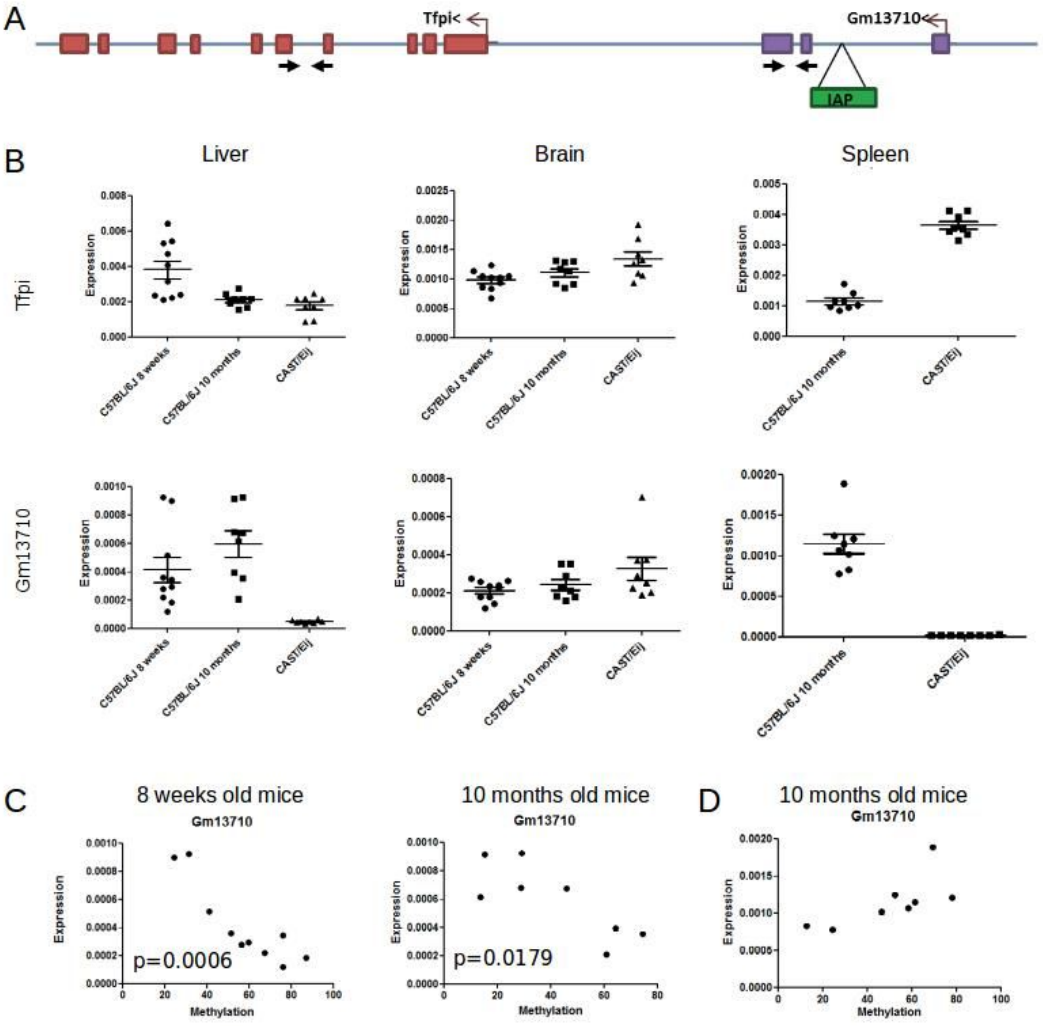


Figure 5.8. Validation of Tfpi and Gm13710 expression. *A) Schematic representation of the genomic region. IAP is inserted in the 1st intron of Gm13710 in C57BL/6J. There is no IAP insertion in this region in CAST/Eij; B) Expression of Tfpi and Gm13710 in C57BL/6J 8 weeks and 10 months old mice and CAST/Eij mice. Each dot represents an individual; C) Gm13710 expression is inversely correlated with IAP methylation in liver tissues. Each dot represents a C57BL/6J individual mouse; D) No correlation between Gm13710 expression and IAP methylation was found in splenic tissues.*

Table 9. Summary of genes potentially regulated by metastable IAPs

Gene name	Type	GO: Molecular function and biological process	Associated phenotype
Fam32a	Protein coding	RNA binding; apoptotic process	unknown
Bmf	Protein coding	Apoptotic process	B cell-restricted lymphadenopathy (Labi et al., 2008); defects in uterovaginal development (Hübner et al., 2010)
Bub1b	Protein coding	protein kinase activity; cell cycle; mitotic spindle assembly checkpoint	embryonic lethal at E8.5 (Wang et al., 2004)
Rnf157	Protein coding	metal ion binding	no phenotype (Matz et al., 2015)
Gm15800	Protein coding	ubiquitin-protein transferase activity; glucose metabolic process	unknown
2610035D17Rik	LincRNA	-	unknown
Slc15a2	Protein coding	oligopeptide transport	no obvious phenotype, abnormal renal reabsorbtion (Rubio-Aliaga et al., 2003)
Eps8l1	Protein coding	Rho protein signal transduction	unknown
Mbnl1	Protein coding	RNA binding; regulation of alternative mRNA splicing	muscle, eye, and RNA splicing abnormalities (Kanadia et al., 2003)
AK134158	Mus musculus adult male thymus cDNA	-	unknown
Gm13710	LincRNA	-	unknown

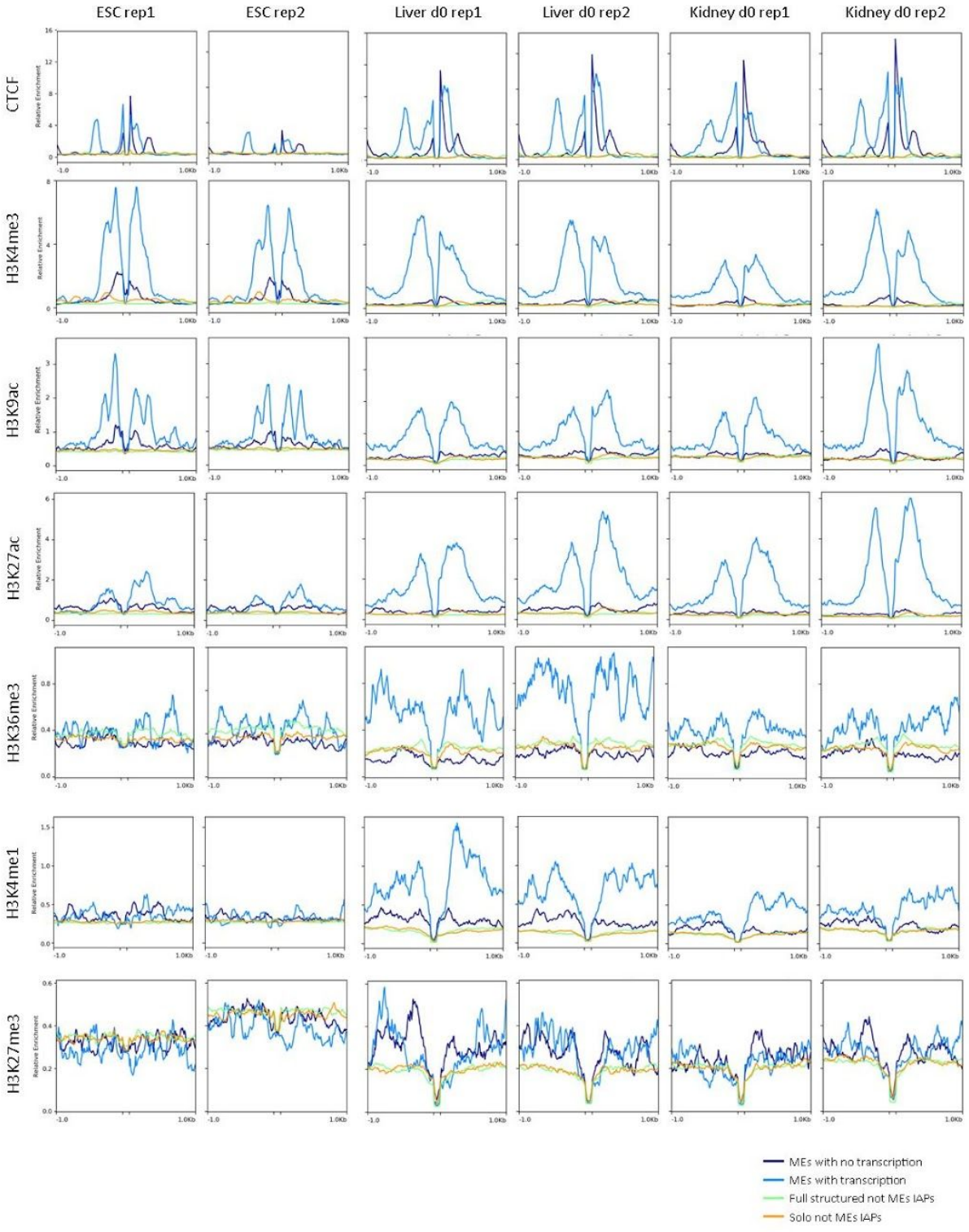


Figure 5.9. ChIP-seq profiles of metastable IAP flanking regions that either overlap (MEs with transcription) or do not overlap (MEs with no transcription) with *de novo* assembled transcripts. *Highly methylated full structure IAPs of the IAPLTR1_Mm and solo IAPLTR2_Mm subclasses serve as controls.*

Characterization of splicing events within IAPs

We used the suggested strategy to analyze splicing events happening within IAPs (**Figure 2.2**). There were two conditions used to separate these IAPs from other groups:

- An IAP should overlap an exon boundary consistently between 3 biological replicates of one cell type and sex.
- The identified transcript that contained a splicing site within an IAP should be initiated and terminated outside of this IAP.

Only 3 IAPs were identified that satisfied these conditions: chr3:60489872-60495084, chr1:173572497-173580638; chr9:40314952-40315123. One of them is the metastable IAP_{Mbnl1}. This intronic IAP was spliced out from the unannotated transcript transcribed in antisense direction to the Mbnl1 gene. In two other cases, transcripts contained full exons within IAP boundaries. This data suggests that IAPs are rarely involved in alternative splicing events in B and T cells.

Summary and discussion

De novo assembled transcriptomes for six B cell and six T cell RNA-seq datasets were used to explore the impact of all IAPs and then the metastable IAPs on transcription. Only IAPs that were consistently involved in transcription initiation, termination and/or splicing were further analysed. We identified 142 IAPs overlapping transcripts' starts or ends. 32 of these IAPs are metastable epialleles. In a few cases we managed to confirm the inverse correlation between metastable IAP methylation and expression of the neighbouring gene. Expression of IAP-driven transcripts was often low, perhaps partially explaining the absence of a clear transcriptional phenotype between differentially methylated individuals. Most of identified transcripts were not previously annotated and probably represent non-coding RNAs or pseudogenes whose functional importance is not clear. Apart from transcription initiation, metastable IAPs were found to cause transcriptional termination. However, no correlation between IAP methylation and the expression of terminating genes was found and perhaps a deeper or more focused analysis of these alleles is required.

Interestingly, our data suggests that metastable IAPs might impact transcription in a tissue specific manner despite being similarly variably methylated in all tissues. At least two cases of the cell type-specific presence of transcripts terminating within metastable IAPs were found. In particular, we found a tissue specific correlation between metastable IAP methylation levels and expression of Gm13710, which might have phenotypic implications given that this appears to be a lncRNA of unknown function and hence is worthy of further study.

The enrichment of active histone marks at metastable IAP flanking regions is indicative of open chromatin next to IAP insertions. This might be a secondary effect of a metastable IAP insertion or might reflect CTCF binding. However, this is unlikely since the absence of similar marks next to CTCF-associated metastable IAPs that do not interfere with transcriptional events. This suggests that these marks, which are associated with active promoters and enhancers, are functionally related to the behaviour of the specific IAP-associated gene and may have been there prior to insertion. One cannot rule out the possibility that the specific IAP insertion has caused the recruitment of the active marks subsequent to insertion.

Just under 1% of IAP elements are metastable. Interestingly, less than 1% of IAPs can cause transcription initiation and termination in B and T cells. Most of these IAPs are evolutionary young full-structured insertions. Despite the general opinion that DNA methylation is sufficient to silence IAP elements, 88% of the identified functionally relevant IAPs were highly methylated in both cell types. Only 4 cases (out of 19) were IAPs associated with cell-type specific transcription initiation and/or termination that are differentially methylated between B and T cells.

Alternative splicing within IAP elements is extremely rare. We manage to identify only 3 IAPs that were consistently providing splice sites within their boundaries.

To sum up, IAP involvement in the regulation of transcription in B and T cells is extremely rare. However such IAP activity is unlikely associated with IAP methylation indicating that other epigenetic mechanisms might be involved in the silencing of young IAP insertions in somatic tissues.

Chapter 6

General conclusion and perspectives

The aim of my PhD was to conduct a genome wide screen for metastable epialleles using sequencing datasets generated in our lab. Multiple biological replicates of WGBS-seq and RNA-seq data were generated for pure non-cycling populations of B and T cells. Each biological replicate represents a sample where B or T cells purified from 4-6 individual mice were pooled together. The datasets were generated for C57BL/6J and CAST/Eij inbred mouse strains. I have conducted an experimental validation of these datasets to confirm the accuracy of the sequencing alignment and analysis with all of the functional analysis focusing on C57BL/6J. This work provides a framework for future comparative analysis with variably methylated alleles in other strains of mice using their presence/absence to consider functional and mechanistic properties.

The screen for metastable epialleles consisted of two steps. The first step used an assumption that metastable strain-specific retroelements might impact expression of nearby genes. This assumption was based on a few major findings about metastable epialleles:

1) The few metastable epialleles are associated with a retroelement insertion (IAP elements).

2) Metastable IAPs provide cryptic promoters to drive an expression of nearby genes.

3) The previously reported $CABP^{IAP}$ metastable epiallele results from an insertion of a polymorphic C57BL/6J specific IAP element absent in other strains.

I have catalogued published C57BL/6J specific ERVs that were absent in the CAST/Eij strain (Nellaker et al., 2011). Our RNA-seq datasets were analysed to identify genes that are differentially expressed between two strains. Polymorphic retroelements were assigned to the nearest protein coding genes. ERVs that were assigned to the differentially expressed genes were further considered. I manually analyzed the methylation status of the identified retroelements across visualized

WGBS-seq and WGoxBS-seq replicates. The absolute majority of ERVs were highly methylated consistently between biological replicates. A small subset of ERVs showed some methylation variation between analysed datasets and the majority of these ERVs were IAP elements. I hypothesised that this inconsistent methylation pattern, that we termed “ragged” methylation, might reflect interindividual methylation variation at these regions. To confirm this hypothesis I purified DNA and RNA from multiple tissues (liver, kidney, spleen and brain) dissected from individual mice. Due to the repetitive nature of ERVs, only methylation of the distal CpGs of retroelements can be experimentally assessed for an individual ERV. Pyrosequencing was used to quantify interindividual methylation variation at the identified ERVs. All analysed IAPs with “ragged” methylation showed different methylation levels between individuals and, consistent with the known metastable epialleles, constant methylation levels at IAP elements were found between different tissues purified from the same individual. Control IAPs that were highly methylated in our sequencing datasets showed no interindividual methylation variation. These results confirmed that IAPs with “ragged” methylation profiles are regions that are variably methylated between individuals.

The second step of the screen was to model the identified “ragged” methylation patterns - associated with interindividual methylation variation at IAPs. Unlike first step, the second step of the screen was purely based on the analysis of the methylation status of ERVs and hence was agnostic to the class or type of element. Methylation variation between WGBS-seq and WGoxBS-seq biological replicates was quantified for 5’ and 3’ ends of IAPs. Because of low mapping quality to internal IAP regions, only distal CpGs can be reliably analysed. Experimental validation of quantified methylation variation score was used to define a threshold for IAPs with interindividual methylation variation. This step identified around 100 metastable IAPs including all IAPs that were identified during the first step of the screen. The same model was used to screen other types of ERVs. Unlike IAPs, other analysed ERVs did not show significant interindividual methylation variation despite having similar methylation profiles in our datasets. However, the numbers selected for validation were very low and more need to be assessed. Whether this apparent discordance between ERVs is due to technical issues associated with mapping quality of sequencing reads to different ERV types with different CpG contents, or is a true reflection of a different type of methylation variation that might exist in B and T cells,

requires further research. Only two MuLV regions showed clear interindividual methylation variation.

Unlike previous screens conducted to identify novel metastable epialleles, my screen benefits from utilising much bigger number of WGBS and WGoxBS biological replicates generated for ex vivo B and T cells (16 replicates in total). This experimental design allows to minimize the discovery of regions that are variably methylated due to tissue heterogeneity. Moreover, the screen relies on the extensive experimental validation of identified regions. However, the necessity to pool B and T cells from different individuals might minimize the sensitivity of the screen and IAPs with low range of interindividual methylation variation might be missed.

Majority of identified metastable IAPs are polymorphic insertions that belong to young IAP subtypes. The identified set of IAPs was used to further explore the mechanisms underlying the establishment of different methylation levels between individuals at these regions. The range of methylation variation between individuals was different between metastable IAPs suggesting a stochastic establishment of methylation during development. The establishment of methylation at metastable IAPs is locus specific. Metastable IAPs with identical sequences are methylated to a different extent within the same individual. Indeed, some metastable IAPs are much more closely related to the highly methylated insertion than to each other. Together, this suggests that IAP sequence is not deterministic factor for the acquisition of methylation metastability at these regions but rather that genomic context plays a role. The mechanism governing the establishment of methylation at metastable IAPs therefore acts in *cis*. It is unlikely that a trans-mediated mechanism that targets all metastable epialleles within a single individual in the same way.

We have identified an enrichment of CTCF occupancy in the proximity of the variably methylated ends of metastable IAPs. CTCF is a methylation sensitive DNA binding protein that is extremely important during early development (Bell & Felsenfeld, 2000; Zampieri et al., 2012). CTCF knockouts are embryonic lethal (Wan et al., 2008; Moore et al., 2012). Methylation at metastable IAPs is established during early embryonic development (Waterlan et al., 2006; Blewitt et al., 2006). This observation gives rise to the hypothesis that an interplay between CTCF binding and IAP methylation in the early embryo might cause the establishment and/or maintenance

of interindividual methylation variation at metastable IAPs. This analysis was based on publicly available ChIP-seq datasets. However, the observed CTCF enrichment needs to be validated and experimentally manipulated in future work. CTCF enrichment was consistently observed in different tissues and in ESCs. It will be important to investigate CTCF binding next to highly and lowly methylated metastable IAPs in individual mice. We identified two C57BL/6J metastable IAPs that are present in but highly methylated in 129Sv background. More extensive comparative analysis of metastable epialleles in different inbred mouse strains will help further explore CTCF role in the establishment of IAP methylation. If CTCF binding can drive metastability of ERV insertions, it is expected that loss of methylation variation at metastable IAPs in 129Sv strain correlates with the absence of CTCF binding in IAP proximity. Datasets for CTCF binding in 129Sv ES cells are available (Nora et al., 2017) and will be analysed in the near future.

Published WGBS-seq data was used to track methylation dynamics at metatable IAPs during male and female germ cell development (Wang et al., 2014; Seisenberger et al., 2012; Kobayashi et al., 2013; Kubo et al., 2015; Shirane et al., 2013). While the available data can not be used to assess the interindividual methylation variation ranges at analysed stages, some major differences between metastable IAPs and A^{vy} methylation was found:

1) While A^{vy} is hypomethylated in blastocyst, some metastable IAPs retain some methylation in ICMs - these represent both paternally and maternally-inherited genomes and hence the homozygosity of the alleles makes it unclear what the methylation status on the two parentally-inherited alleles is, at this time and earlier.

2) Metastable IAPs are fully methylated in sperm. A^{vy} and $Axin^{fused}$ methylation at sperm apparently reflects methylation levels in somatic tissues within the individual – however the technology for quantifying methylation in those papers was less sophisticated and prone to error hence our lab plans to repeat those experiments using pyrosequencing on A^{vy} scoring the methylation blind without knowing the coat colour of the animals in advance.

3) Though containing lower in methylation than control alleles, some metastable IAPs may be partially resistant to demethylation during oocyte development yet they are fully programmed during sperm development.

This data suggests that reprogramming of our metastable IAPs is different from the A^{vy} IAP (Blewitt et al., 2006). More complete analysis of our own alleles as well as A^{vy} is required in order to conduct a detailed comparison. A^{vy} and $Axin^{fused}$ showed parent-of-origin effects and strain effects on the heritable establishment of methylation at these regions in the offspring (Morgan et al., 1999; Rakyan et al., 2003). While transgenerational inheritance of the methylation state of our metastable IAPs is in progress and most alleles are variably methylated in offspring, detailed analysis of methylation dynamics at these regions during development would help to determine the key processes responsible for transmission or de novo establishment of epigenetic marks at these IAPs.

Dnmt1 activity is required for the maintenance of A^{vy} IAP methylation in somatic tissues (Gaudet et al., 2004). During early embryonic development the Dnmt1 short maternal isoform (Dnmt1o) is expressed in the stage cleavage embryo (Ratnam et al., 2002). The presence of this protein at a particular stages might therefore be important for the establishment and maintenance of metastable IAP methylation. Analysis of methylation dynamic at our metatable IAPs in Dnmt1o mutants is required to test this and we have obtained these mice (Howell et al, 2001). CTCF has been shown to be able to inhibit Dnmt1 activity (Zampieri et al., 2012). Therefore, there might be specific interactions between CTCF and Dnmt1o activities during early embryonic development when the establishment of methylation at metastable IAPs is happening.

Thirty three metastable IAP were found to be associated with transcription initiation or termination. Most of the identified affected transcripts were previously unannotated and from our BLUEPRINT datasets. Their functions, as yet, are not known. However we hypothesise that they represent non-coding RNAs or pseudogenes that might be involved in the regulation of gene transcription in cis or trans. We managed to confirm an inverse correlation between IAP methylation and expression at some of the genes. Interestingly, we found that some metastable epialleles are expressed in a tissue specific manner despite a consistent methylation state across tissue types.

Whether tissue-specific IAP activity is associated with availability of tissue-specific transcription factors or chromatin conformation is not clear and needs to be further explored. Moreover, these results suggest that metastable epiallele screens that rely solely on the identification of transcripts variably expressed between individuals would not be fully comprehensive. Therefore, our two-pronged approach provided a more comprehensive screen.

The enrichment of active histone marks next to metastable IAPs that could impact transcription indicates that these IAPs may have inserted into more active regulatory regions that are more open and conducive to retrotransposition events. However, these active regulatory regions were not functionally critical since their disruption did not result in a negative selection against IAP insertions at these regions. This hypothesis might be further tested through comparative analysis of polymorphic metastable IAPs between different inbred strains. While publicly available ChIP-seq datasets can be used to analyse epigenetic states at flanking genomic regions, it would be interesting to experimentally analyse histone marks associated with metastable IAPs that have different effects on transcriptional events.

Less than 1% of IAPs genome-wide were found to be involved in transcription termination or initiation. Most of identified IAPs were providing both transcription termination and initiation sites. We found an enrichment of evolutionary young subtypes of IAPs among these insertions. Interestingly, despite their impact on the transcriptome, most identified IAPs were highly methylated in B and T cells. This suggests that DNA methylation is not sufficient for complete silencing of young potentially active IAP insertions in somatic tissues. It would be interesting to further explore what histone marks are associated with these elements. The observed expression levels of IAP-derived transcripts were quite low and therefore it would be of value to see if there is an increase in transcription associated with these elements in tissues with lower global methylation levels. One cannot assume that the elements themselves would reflect the lower methylation state of the host tissue since methylation at an hypermethylated IAP is quite stable in situations where global methylation levels are reduced (Bertozzi, unpublished results).

Our analysis showed that IAPs rarely provide alternative splicing events. However one of the identified examples was a metastable IAP. So far it is the first example of a

metastable epiallele involved in alternative splicing of a transcript that is initiated and terminated outside of the metastable IAP hence this locus can be exploited to further explore the relationship between expression and splicing of the identified transcript and IAP methylation. Interestingly, this transcript is expressed in an antisense direction to the host Mbnl1 gene. Whether this transcript is involved in Mbnl1 regulation and what the role of methylation metastability at this region is an interesting question for further experimentation.

Acknowledgments

I would like to thank my supervisor Prof. Anne Ferguson-Smith for welcoming me in her lab and for valuable discussions about my project and guidance. Dr. Marcela Sjoberg selected and purified the cellular models from each one of the strains and reciprocal hybrids used in this project, prepared the libraries for WGBS and RNA-seq and provided me with RNA and DNA samples from each cell type which I used in the validation experiments described in this report. I am thankful to Dr. Nic Walker who performed the bioinformatics analysis of the WGBS/oxWGBS data sets. Thanks to Richard Gunning who performed the differential expression analysis of the RNA-seq datasets and Joseph Gardner who conducted de novo transcriptome assembly for the RNA-seq datasets. I am grateful to students whose projects I supervised: Part III student Eleni Pahita who worked on IAP-driven transcription; Part II student Sarah Adams and summer students James Lee and Antara Majumdar who helped with metastable IAP validation. I am thankful to my collaborators Hiroyuki Sasaki and Kenjiro Shirane (Fukuoka, Japan) who helped with the analysis of metastable IAP methylation dynamics during germline development. I would like to thank my family and friends for care and support during my PhD.

My project was funded by Darwin Trust of The University of Edinburgh and BLUEPRINT.

References

Abbott A. (2011) Europe to map the human epigenome. Nature, 477:518

Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A, Dahl F, Dermitzakis ET, Enver T, Esteller M, Estivill X, Ferguson-Smith A, Fitzgibbon J, Flicek P, Giehl C, Graf T, Grosveld F, Guigo R, Gut I, Helin K, Jarvius J, Küppers R, Lehrach H, Lengauer T, Lernmark Å, Leslie D, Loeffler M, Macintyre E, Mai A, Martens JH, Minucci S, Ouwehand WH, Pelicci PG, Pendeville H, Porse B, Rakyan V, Reik W, Schrappe M, Schübeler D, Seifert M, Siebert R, Simmons D, Soranzo N, Spicuglia S, Stratton M, Stunnenberg HG, Tanay A, Torrents D, Valencia A, Vellenga E, Vingron M, Walter J, Willcocks S. (2012) BLUEPRINT to decode the epigenetic signature written in blood. Nature Biotechnology, 30:224-226

Allen ND, Norris ML, Surani MA. (1990) Epigenetic control of transgene expression and imprinting by genotype-specific modifiers. Cell, 61(5):853-61

Allman D, Pillai S. (2008) Peripheral B cell subsets. Current Opinion in Immunology, 20: 149-157

Anderson LM, Riffle L, Wilson R, Travlos GS, Lubomirski MS, Alvord WG. (2006) Preconceptional fasting of fathers alters serum glucose in offspring of mice. Nutrition, 22(3):327-31

Anway MD, Cupp AS, Uzumcu M, Skinner MK. (2005) Epigenetic transgenerational actions of endocrine disruptors and male fertility. Science, 308(5727):1466-9

Aran D, Toperoff G, Rosenberg M, Hellman A. (2011) Replication timing-related and gene body-specific methylation of active human genes. Human Molecular Genetics, 20(4):670-680

Arand J, Wossidlo M, Lepikhov K, Peat JR, Reik W, Walter J. (2015) Selective impairment of methylation maintenance is the major cause of DNA methylation reprogramming in the early embryo. Epigenetics Chromatin, 8(1):1

Argeson AC, Nelson KK, Siracusa LD. (1996) Molecular basis of the pleiotropic phenotype of mice carrying the hypervariable yellow (Ahvy) mutation at the agouti locus. Genetics, 142(2):557-67

Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T. (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature, 410(6824):120-4

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. (2007) High-resolution profiling of histone methylations in the human genome. Cell, 129(4):823-37

Bell AC, Felsenfeld G. (2000) Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature, 405(6785):482-5

Benyshek DC, Johnston CS, Martin JF. (2006) Glucose metabolism is altered in the adequately-nourished grand-offspring (F3 generation) of rats malnourished during gestation and perinatal life. Diabetologia, 49(5):1117-9

Bestor TH. (1998) The host defence function of genomic methylation patterns. Novartis Found Symp., 214:187-95

Bhutani N, Burns DM, Blau HM. (2011) DNA demethylation dynamics. Cell, 146(6):866-72

Blake GET, Watson ED (2016) Unravelling the complex mechanisms of transgenerational epigenetic inheritance. Curr Opin Chem Biol., 33:101-7.

Blewitt ME, Vickaryous NK, Paldi A, Koseki H, Whitelaw E. (2006) Dynamic reprogramming of DNA methylation at an epigenetically sensitive allele in mice. PLoS Genet., 2(4):e49

Booth MJ, Ost TW, Beraldi D, Bell NM, Branco MR, Reik W, Balasubramanian S. (2013) Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc., 8(10):1841-51

Bostick M, Kim JK, Estève PO, Clark A, Pradhan S, Jacobsen SE. (2007) UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science, 317(5845):1760-4

Bourc'his D, Bestor TH. (2004) Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature, 431(7004):96-9

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, Chew JL, Ruan Y, Wei CL, Ng HH, Liu ET. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res., 18(11):1752-62

Bray SJ, Takada S, Harrison E, Shen SC, Ferguson-Smith AC. (2008) The atypical mammalian ligand Delta-like homologue 1 (Dlk1) can regulate Notch signalling in Drosophila. BMC Dev Biol., 8:11

Burger L, Gaidatzis D, Schübeler D, Stadler MB. (2013) Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res., 41(16): e155

Cantone I, Fisher AG. (2013) Epigenetic programming and reprogramming during development. Nat Struct Mol Biol., 20(3):282-9

Cariappa A, Chase C, Liu H, Russell P, Pillai S. (2007) Naïve recirculating B cells mature simultaneously in the spleen and bone marrow. Blood, 109:2339-2345

Carone BR, Fauquier L, Habib N, Shea JM, Hart CE, Li R, Bock C, Li C, Gu H, Zamore PD, Meissner A, Weng Z, Hofmann HA, Friedman N, Rando OJ. (2010) Paternally induced transgenerational environmental reprogramming of metabolic gene expression in mammals. Cell, 143(7):1084-96

Chaillet JR, Vogt TF, Beier DR, Leder P. (1991) Parental-specific methylation of an imprinted transgene is established during gametogenesis and progressively changes during embryogenesis. Cell, 66(1):77-83

Chang-Yeh A, Mold DE, Huang RC. (1991) Identification of a novel murine IAP-promoted placenta-expressed gene. Nucleic Acids Res., 19(13):3667-72.

Chen DP, Lin YC, Fann CS. (2016) Methods for identifying differentially methylated regions for sequence- and array-based data. Brief Funct Genomics, 15(6):485-490.

Chuong EB, Rumi MA, Soares MJ, Baker JC. (2013) Endogenous retroviruses function as species-specific enhancer elements in the placenta. Nat Genet., 45(3):325-9

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, Boyer LA, Young RA, Jaenisch R. (2010) Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A., 107(50):21931-6

Cropley JE, Suter CM, Beckman KB, Martin DI. (2006) Germ-line epigenetic modification of the murine A vy allele by nutritional supplementation. Proc Natl Acad Sci U S A., 103(46):17308-12

Dias S, Xu W, McGregor S, Kee B. (2008)Transcriptional regulation of lymphocyte development. Opinion in Genetics & Development, 18:441-448

Dickies MM. (1962) A new viable yellow mutation in the house mouse. J. Hered. 53(2), 84–86

Dolinoy D.C. (2008) The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. Nutr Rev., 66: S7–11

Dolinoy DC, Huang D, Jirtle RL. (2007) Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. Proc Natl Acad Sci U S A., 104(32):13056-61

Dolinoy DC, Weidman JR, Waterland RA, Jirtle RL. (2006) Maternal genistein alters coat color and protects Avy mouse offspring from obesity by modifying the fetal epigenome. Environ Health Perspect., 114(4):567-72

Dolinoy DC, Weinhouse C, Jones TR, Rozek LS, Jirtle RL. (2010) Variable histone modifications at the A(vy) metastable epiallele. Epigenetics, 5(7):637-44

Druker R, Bruxner TJ, Lehrbach NJ, Whitelaw E. (2004) Complex patterns of transcription at the insertion site of a retrotransposon in the mouse. Nucleic Acids Research, 32(19):5800–5808

Duhl DM, Vrieling H, Miller KA, Wolff GL, Barsh GS. (1994) Neomorphic agouti mutations in obese yellow mice. Nat Genet., 8(1):59-65.

Dupressoir A, Heidmann T. (1996) Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice. Mol Cell Biol., 16(8):4495-503

Ekram MB, Kang K, Kim H, Kim J. (2012) Retrotransposons as a major source of epigenetic variations in the mammalian genome. Epigenetics, 7(4):370-82

Emera D, Casola C, Lynch VJ, Wildman DE, Agnew D, Wagner GP. (2012) Convergent evolution of endometrial prolactin expression in primates, mice, and

elephants through the independent recruitment of transposable elements. Mol Biol Evol, 29(1):239-47

Falzon M, Kuff EL. (1988) Multiple protein-binding sites in an intracisternal A particle long terminal repeat. J Virol., 62(11):4070–4077

Faulk C, Barks A, Dolinoy DC. (2013) Phylogenetic and DNA methylation analysis reveal novel regions of variable methylation in the mouse IAP class of transposons. Genomics, 14:48

Feil R, Fraga MF. (2012) Epigenetics and the environment: emerging patterns and implications. Nature Reviews Genetics, 13(2):97-109

Ferguson-Smith AC. (2011) Genomic imprinting: the emergence of an epigenetic paradigm. Nat Rev Genet., 12(8):565-75

Ferrón SR, Charalambous M, Radford E, McEwen K, Wildner H, Hind E, Morante-Redolat JM, Laborda J, Guillemot F, Bauer SR, Fariñas I, Ferguson-Smith AC. (2011) Postnatal loss of Dlk1 imprinting in stem cells and niche astrocytes regulates neurogenesis. Nature, 475(7356):381-5

Finer S, Holland ML, Nanty L, Rakyan VK. (2011) The hunt for the epiallele. Environ Mol Mutagen., 52(1):1-11

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. RNAS 102(30): 10604 – 10609

Gaudet F., Rideout WM, Meissner A, Dausman J, Leonhardt H, Jaenisch R. (2004) Dnmt1 Expression in Pre- and Postimplantation Embryogenesis and the Maintenance of IAP Silencing. Mol Cell Biol., 24(4): 1640–1648

Georgiades P, Watkins M, Surani MA, Ferguson-Smith AC. (2000) Parental origin-specific developmental defects in mice with uniparental disomy for chromosome 12. Development, 127(21):4719-28

Guerrero-Bosagna C, Settles M, Lucker B, Skinner MK. (2010) Epigenetic transgenerational actions of vinclozolin on promoter regions of the sperm epigenome. PLoS One, 5(9)

Guo F, Li X, Liang D, Li T, Zhu P, Guo H, Wu X, Wen L, Gu TP, Hu B, Walsh CP, Li J, Tang F, Xu GL. (2014) Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. Cell Stem Cell., 15(4):447-459

Hata K, Okano M, Lei H, Li E. (2002) Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development, 129(8):1983-93

He Y, Pear WS. (2003) Notch signalling in B cells. Semin Cell Dev Biol., 14(2):135-42.

Heijmans BT, Kremer D, Tobi EW, Boomsma DI, Slagboom PE. (2007) Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. Hum Mol Genet., 16(5):547-54

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. , 39(3):311-8

Hendrich B, Bird A. (1998) Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol Cell Biol., 18(11):6538-47

Howell CY, Bestor TH, Ding F, Latham KE, Mertineit C, Trasler JM, Chaillet JR. (2001) Genomic imprinting disrupted by a maternal effect mutation in the Dnmt1 gene. Cell, 104(6):829-38

Hughes D.C. (2001) Alternative Splicing of the Human VEGFGR-3/FLT4 Gene as a Consequence of an Integrated Human Endogenous Retrovirus. J. Mol. Evol., 53:77-79

Huh I, Zeng J, Park T, Yi SV. (2013) DNA methylation and transcriptional noise. Epigenetics & Chromatin, 6:9

Hurd PJ. (2010) The era of epigenetics. Briefings in functional genomics, 9(5-6):425-428

Hübner A, Cavanagh-Kyros J, Rincon M, Flavell RA, Davis RJ. (2010) Functional cooperation of the proapoptotic Bcl2 family proteins Bmf and Bim in vivo. Mol Cell Biol., 30(1):98-105

Jern P, Coffin JM. (2008) Effects of retroviruses on host genome function. Annu Rev Genet., 42:709-32

Jimenez-Chillaron JC, Isganaitis E, Charalambous M, Gesta S, Pentinat-Pelegrin T, Faucette RR, Otis JP, Chow A, Diaz R, Ferguson-Smith A, Patti ME. (2009) Intergenerational transmission of glucose intolerance and obesity by in utero undernutrition in mice. Diabetes, 58(2):460-8

Jones PA. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet., 13(7):484-92

Kaminen-Ahola N, Ahola A, Maga M, Mallitt KA, Fahey P, Cox TC, Whitelaw E, Chong S. (2010) Maternal ethanol consumption alters the epigenotype and the phenotype of offspring in a mouse model. PLoS Genet., 6(1):e1000811

Kanadia RN, Johnstone KA, Mankodi A, Lungu C, Thornton CA, Esson D, Timmers AM, Hauswirth WW, Swanson MS. (2003) A muscleblind knockout model for myotonic dystrophy. Science, 302(5652):1978-80

Kapitonov VV, Jurka J. (1999) The long terminal repeat of an endogenous retrovirus induces alternative splicing and encodes an additional carboxy-terminal sequence in the human leptin receptor. J Mol Evol., 48(2):248-51

Karmodiya K, Krebs AR, Oulad-Abdelghani M, Kimura H, Tora L. (2012) H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. BMC Genomics., 13:424

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. (2004) The UCSC Table Browser data retrieval tool. Nucleic Acids Res., 32(Database issue):D493-6.

Kato Y, Kaneda M, Hata K, Kumaki K, Hisano M, Kohara Y, Okano M, Li E, Nozaki M, Sasaki H. (2007) Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Hum Mol Genet., 16(19):2272-80

Keane TM, Goodstadt L, Danecek P, White MA, Wong K, Yalcin B, Heger A, Agam A, Slater G, Goodson M, Furlotte NA, Eskin E, Nellåker C, Whitley H, Cleak J, Janowitz D, Hernandez-Pliego P, Edwards A, Belgard TG, Oliver PL, McIntyre RE, Bhomra A, Nicod J, Gan X, Yuan W, van der Weyden L, Steward CA, Bala S, Stalker J, Mott R, Durbin R, Jackson IJ, Czechanski A, Guerra-Assunção JA, Donahue LR, Reinholdt LG, Payseur BA, Ponting CP, Birney E, Flint J, Adams DJ. (2011) Mouse genomic variation and its effect on phenotypes and gene regulation. Nature, 477(7364):289-94

Kearns M, Preis J, McDonald M, Morris C, Whitelaw E. (2000) Complex patterns of inheritance of an imprinted murine transgene suggest incomplete germline erasure. Nucleic Acids Res., 28(17):3301-9

Kobayashi H, Sakurai T, Miura F, Imai M, Mochiduki K, Yanagisawa E, Sakashita A, Wakai T, Suzuki Y, Ito T, Matsui Y, Kono T. (2013) High-resolution DNA methylome analysis of primordial germ cells identifies gender-specific reprogramming in mice. Genome Res., 23(4): 616–627.

Koch U, Radtke F. (2011) Mechanisms of T cell development and transformation. Annu. Rev. Cell Dev. Biol., 27:539-562

Kubo N, Toh H, Shirane K, Shirakawa T, Kobayashi H, Sato T, Sone H, Sato Y, Tomizawa S, Tsurusaki Y, Shibata H, Saitsu H, Suzuki Y, Matsumoto N, Suyama M, Kono T, Ohbo K, Sasaki H. (2015) DNA methylation and gene expression dynamics during spermatogonial stem cell differentiation in the early postnatal mouse testis. BMC Genomics, 20;16:624

Labi V, Erlacher M, Kiessling S, Manzl C, Frenzel A, O'Reilly L, Strasser A, Villunger A. (2008) Loss of the BH3-only protein Bmf impairs B cell homeostasis and accelerates gamma irradiation-induced thymic lymphoma development. J Exp Med., 205(3):641-55

Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J, Reik W. (2003) Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. Genesis, 35(2):88-93

Li E, Bestor TH, Jaenisch R. (1992) Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell, 69(6):915-26

Li J, Akagi K, Hu Y, Trivett AL, Hlynialuk CJ, Swing DA, Volfovsky N, Morgan TC, Golubeva Y, Stephens RM, Smith DE, Symer DE. (2012) Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. Genome Res., 22(5):870-84

Li X, Ito M, Zhou F, Youngson N, Zuo X, Leder P, Ferguson-Smith AC. (2008) A maternal-zygotic effect gene, Zfp57, maintains both maternal and paternal imprints. Dev Cell., 15(4):547-57

Mak KS, Burdach J, Norton LJ, Pearson RC, Crossley M, Funnell AP. (2014) Repression of chimeric transcripts emanating from endogenous retrotransposons by a sequence-specific transcription factor. Genome Biol., 15(4):R58

Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL. (2006) Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet., 2(1):e2

Maniatis T, Fritsch EF, Sambrook J. (1982) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory

Maor GL, Yearim A, Ast G. (2015) The alternative role of DNA methylation in splicing regulation. Trends Genet., 31:274–280

Matsui T, Leung D, Miyashita H, Maksakova IA, Miyachi H, Kimura H, Tachibana M, Lorincz MC, Shinkai Y. (2010) Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature, 464(7290):927-31

Matthias P, Rolink A.G. (2005) Transcriptional networks in developing and mature B cells. Nature Reviews Immunology, 5(6):497-508

Matthias P, Rolink AG. (2005) Transcriptional networks in developing and mature B cells. Nature Reviews Immunology, 5(6):497-508

Matz A, Lee S-J, Schwedhelm-Domeyer N, Zanini D, Holubowska A, Kannan M, Farnworth M, Jahn O, Göpfert MC, Stegmüller J. (2015) Regulation of neuronal survival and morphology by the E3 ubiquitin ligase RNF157. Cell Death Differ., 22(4): 626–642

Messerschmidt DM, Knowles BB, Solter D. (2014) DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes Dev., 28(8):812-28

Michaud EJ, van Vugt MJ, Bultman SJ, Sweet HO, Davisson MT, Woychik RP. (1994) Differential expression of a new dominant agouti allele (Aiapy) is correlated with methylation state and is influenced by parental lineage. Genes Dev., 8(12):1463-72.

Mietz JA, Grossman Z, Lueders KK, Kuff EL. (1987) Nucleotide sequence of a complete mouse intracisternal A-particle genome: relationship to known aspects of particle assembly and function. J Virol., 61(10):3020–3029

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES,

Bernstein BE. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, 448(7153):553-60

Miska EA, Ferguson-Smith AC. (2016) Transgenerational inheritance: Models and mechanisms of non-DNA sequence-based inheritance. Science, 354(6308):59-63

Moore JM, Rabaia NA, Smith LE, Fagerlie S, Gurley K, Loukinov D, Disteche CM, Collins SJ, Kemp CJ, Lobanenkov VV, Filippova GN. (2012) Loss of Maternal CTCF Is Associated with Peri-Implantation Lethality of Ctcf Null Embryos. PLoS One., 7(4): e34915

Morgan HD, Jin XL, Li A, Whitelaw E, O'Neill C. (2008) The culture of zygotes to the blastocyst stage changes the postnatal expression of an epigentically labile allele, agouti viable yellow, in mice. Biol Reprod., 79(4):618-23

Morgan HD, Sutherland HG, Martin DI, Whitelaw E. (1999) Epigenetic inheritance at the agouti locus in the mouse. Nat Genet., 23(3):314-8

Muñoz-López M, García-Pérez JL. (2010) DNA transposons: nature and applications in genomics. Curr Genomics., 11(2):115-28

Nakamura T, Arai Y, Umehara H, Masuhara M, Kimura T, Taniguchi H, Sekimoto T, Ikawa M, Yoneda Y, Okabe M, Tanaka S, Shiota K, Nakano T. (2007) PGC7/Stella protects against DNA demethylation in early embryogenesis. Nat Cell Biol., 9(1):64-71

Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, Flint J, Adams DJ, Frankel WN, Ponting CP. (2012) The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol., 13(6):R45

Ng SF, Lin RC, Laybutt DR, Barres R, Owens JA, Morris MJ. (2010) Chronic high-fat diet in fathers programs ß-cell dysfunction in female rat offspring. Nature, 467(7318):963-6

Nora EP, Goloborodko A, Valton AL, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG. (2017) Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell, 169(5):930-944

Nueda ML, Baladrón V, Sánchez-Solana B, Ballesteros MA, Laborda J. (2007) The EGF-like protein dlk1 inhibits notch signaling and potentiates adipogenesis of mesenchymal cells. J Mol Biol., 367(5):1281-93

Oey H, Isbel L, Hickey P, Ebaid B, Whitelaw E. (2015) Genetic and epigenetic variation among inbred mouse littermates: identification of inter-individual differentially methylated regions. Epigenetics Chromatin, 8:54

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(3):247-57

Osmond D, Rolink A, Melchers F. (1998) Murine B lymphopoiesis: towards a unified model. Immunology today, 19:65-68

Patro R, Mount SM, Kingsford C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nat Biotechnol., 32(5):462-4.

Phillips JE, Corces VG. (2009) CTCF: master weaver of the genome. Cell, 137(7):1194-211

Pillai S, Cariappa A. (2009) The follicular versus marginal zone B lymphocyte cell fate decision. Nature reviews. Immunology, 9(11):767 – 777

Pinney SE. (2014) Mammalian Non-CpG Methylation: Stem Cells and Beyond. Biology (Basel)., 3(4):739-51

Plath K, Fang J, Mlynarczyk-Evans SK, Cao R, Worringer KA, Wang H, de la Cruz CC, Otte AP, Panning B, Zhang Y. (2003) Role of histone H3 lysine 27 methylation in X inactivation. Science, 300(5616):131-5

Qin C, Wang Z, Shang J, Bekkari K, Liu R, Pacchione S, McNulty KA, Ng A, Barnum JE, Storer RD. (2010). Intracisternal A particle genes: Distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. Mol. Carcinog. 49, 54-67

Quinlan AR, Hall IM. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6):841-2

Radford EJ, Ito M, Shi H, Corish JA, Yamazawa K, Isganaitis E, Seisenberger S, Hore TA, Reik W, Erkek S, Peters AHFM, Patti ME, Ferguson-Smith AC. (2014) In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism. Science, 345(6198):1255903

Raghunandan R, Ruiz-Hidalgo M, Jia Y, Ettinger R, Rudikoff E, Riggins P, Farnsworth R, Tesfaye A, Laborda J, Bauer SR. (2008) Dlk1 influences differentiation and function of B lymphocytes. Stem Cells Dev., 17(3):495-507

Rakyan VK, Blewitt ME, Druker R, Preis JI, Whitelaw E. (2002) Metastable epialleles in mammals. TRENDS in Genetics, 18(7):348-351

Rakyan VK, Chong S, Champ ME, Cuthbert PC, Morgan HD, Luu KV, Whitelaw E. (2003) Transgenerational inheritance of epigenetic states at the murine Axin(Fu) allele occurs after maternal and paternal transmission. Proc Natl Acad Sci U S A., 100(5):2538-43

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res., 42(Web Server issue):W187-91

Rangasamy D. (2013) Distinctive patterns of epigenetic marks are associated with promoter regions of mouse LINE-1 and LTR retrotransposons. Mob DNA., 4(1):27

Ratnam S, Mertineit C, Ding F, Howell CY, Clarke HJ, Bestor TH, Chaillet JR, Trasler JM. (2002) Dynamics of Dnmt1 methyltransferase expression and intracellular localization during oogenesis and preimplantation development. Dev Biol., 245(2):304-14.

Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, Goyal P, Keane TM, Jones S, Hirst M, Lorincz MC, Mager DL. (2011) Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS genetics, 7(9): e1002301

Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, Mager D. (2012) Epigentic interplay between mouse endogenous retroviruses and host genes. Genome Biology, 13:R89

Reik W, Dean W, Walter J. (2001) Epigenetic reprogramming in mammalian development. Science, 293(5532):1089-93

Richards EJ. (2006) Inherited epigenetic variation - revisiting soft inheritance. Nat Rev Genet., 7(5):395-401

Robey E. (1999) Regulation of T cell fate by Notch. Annu Rev Immunol., 17:283-95

Rolink A, Melchers F. (1996) B-cell development in the mouse. Immunology Letters, 54:157-161

Rothenberg EV. (2012) Transcriptional drivers of the T-cell lineage program. Current Opinion in Immunology, 24:132–138

Royaux I, Bernier B, Montgomery JC, Flaherty L, Goffinet AM (1997) Relnrl-Alb2, an allele of reeler isolated from a chlorambucil screen, is due to an IAP insertion with exon skipping. Genomics, 42:479–482

Rubio-Aliaga I, Frey I, Boll M, Groneberg DA, Eichinger HM, Balling R, Daniel H. (2003) Targeted disruption of the peptide transporter Pept2 gene in mice defines its physiological role in the kidney. Mol Cell Biol., 23(9):3247–3252

Sakajiri S, O'kelly J, Yin D, Miller CW, Hofmann WK, Oshimi K, Shih LY, Kim KH, Sul HS, Jensen CH, Teisner B, Kawamata N, Koeffler HP. (2005) Dlk1 in normal and abnormal hematopoiesis. Leukemia, 19(8):1404-10

Sasaki H, Hamada T, Ueda T, Seki R, Higashinakagawa T, Sakaki Y. (1991) . Inherited type of allelic methylation variations in a mouse chromosome region where an integrated transgene shows methylation imprinting. Development, 111(2):573-81

Schotta G, Lachner M, Sarma K, Ebert A, Sengupta R, Reuter G, Reinberg D, Jenuwein T. (2004) A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes Dev., 18(11):1251-62

Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G. (2007) Genome regulation by polycomb and trithorax proteins. Cell, 128(4):735-45

Seisenberger S, Andrews S, Krueger F, Arand J, Walter J, Santos F, Popp C, Thienpont B, Dean W, Reik W. (2012) The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells. Mol Cell, 48(6):849-62

Sharif J, Shinkai Y, Koseki H. (2013) Is there a role for endogenous retroviruses to mediate long-term adaptive phenotypic response upon environmental inputs? Philos Trans R Soc Lond B Biol Sci., 368(1609):20110340

Shirane K, Toh H, Kobayashi H, Miura F, Chiba H, Ito T, Kono T, Sasaki H. (2013) Mouse oocyte methylomes at base resolution reveal genome-wide accumulation of non-CpG methylation and role of DNA methyltransferases. PLoS Genet., 9(4):e1003439

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, Schübeler D. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature, 480(7378):490-5.

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

Stroud H, Do T, Du J, Zhong X, Feng S, Johnson L, Patel DJ, Jacobsen SE. (2014) Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. Nat Struct Mol Biol., 21(1):64-72

Sun D, Xi Y, Rodriguez B, Park HJ, Tong P, Meong M, Goodell MA, Li W. (2014) MOABS: model based analysis of bisulfite sequencing data. Genome Biol., 15(2):R38

Surani MA, Reik W, Allen ND (1988) Transgenes as molecular probes for genomic imprinting. Trends Genet 4:59-62

Sutherland HG, Kearns M, Morgan HD, Headley AP, Morris C, Martin DI, Whitelaw E. (2000) Reactivation of heritably silenced gene expression in mice. Mamm Genome, 11(5):347-55

Suzuki MM, Bird A. (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet., 9(6):465-76

Swain JL, Stewart TA, Leder P. (1987) Parental legacy determines methylation and expression of an autosomal transgene: a molecular mechanism for parental imprinting. Cell, 50(5):719-27

Takahashi N, Gray D, Strogantsev R, Noon A, Delahaye C, Skarnes WC, Tate PH, Ferguson-Smith AC. (2015) ZFP57 and the Targeted Maintenance of Postfertilization Genomic Imprints. Cold Spring Harb Symp Quant Biol., 80:177-87

Tsagaratou A, Äijö T, Lio CW, Yue X, Huang Y, Jacobsen SE, Lähdesmäki H, Rao A. (2014) Dissecting the dynamic changes of 5-hydroxymethylcytosine in T-cell development and differentiation. Proc Natl Acad Sci U S A., 111(32):E3306-15

Tsumura A, Hayakawa T, Kumaki Y, Takebayashi S, Sakaue M, Matsuoka C, Shimotohno K, Ishikawa F, Li E, Ueda HR, Nakayama J, Okano M. (2006) Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes Cells, 11(7):805-14

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. (2012) Primer3 - new capabilities and interfaces. Nucleic Acids Res., 40(15):e115

Vasicek T, Zeng L, Guan X-J, Zhang T, Costantini F, Tilghman S. (1997) Two Domain Mutations in the Mouse Fused Gene Are the Results of Transposon Insertions. Genetics, 147:777-786

Vaughan A, Roghanian A, Cragg M. (2011) B cells – Masters of the immunoverse. The intenational Journal of Biochemistry and Cell Biology, 43:280-285

Wagner GP, Kin K, Lynch VJ. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci., 131(4):281-5.

Walsh C.P., Chaillet J.R., Bestor T.H. (1998) Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nature Genetics, 20:116-117

Wan J, Oliver VF, Zhu H, Zack DJ, Qian J, Merbs SL. (2013) Integrative analysis of tissue-specific methylation and alternative splicing identifies conserved transcription factor binding motifs. Nucleic Acids Research, 41(18): 8503 – 8514

Wan LB, Pan H, Hannenhalli S, Cheng Y, Ma J, Fedoriw A, Lobanenkov V, Latham KE, Schultz RM, Bartolomei MS. (2008) Maternal depletion of CTCF reveals multiple functions during oocyte and preimplantation embryo development. Development, 135(16):2729-38

Wang Q, Liu T, Fang Y, Xie S, Huang X, Mahmood R, Ramaswamy G, Sakamoto KM, Darzynkiewicz Z, Xu M, Dai W. (2004) BUBR1 deficiency results in abnormal megakaryopoiesis. Blood, 103(4):1278-85

Wang L, Zhang J, Duan J, Gao X, Zhu W, Lu X, Yang L, Zhang J, Li G, Ci W. (2014) Programming and inheritance of parental DNA methylomes in mammals. Cell, 157:979–991

Waterland RA, Dolinoy DC, Lin JR, Smith CA, Shi X, Tahiliani KG. (2006) Maternal methyl supplements increase offspring DNA methylation at Axin Fused. Genesis, 44(9):401-6

Waterland RA, Jirtle RL. (2003) Transposable Elements: Targets for Early Nutritional Effects on Epigenetic Gene Regulation. Mol. Cell. Biol., 23(15):5293-5300

Weichman K, Chaillet JR. (1997) Phenotypic variation in a genetically identical population of mice. Mol Cell Biol., 17(9):5269-74

Weinhouse C, Anderson OS, Jones TR, Kim J, Liberman SA, Nahar MS, Rozek LS, Jirtle RL, Dolinoy DC. (2011) An expression microarray approach for the identification of metastable epialleles in the mouse genome. Epigenetics, 6(9):1105-13.

Wolff GL, Kodell RL, Moore SR, Cooney CA. (1998) Maternal epigenetics and methyl supplements affect agouti gene expression in Avy/a mice. FASEB J., 12(11):949-57

Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, Zhou X, Lee HJ, Maire CL, Ligon KL, Gascard P, Sigaroudinia M, Tlsty TD, Kadlecek T, Weiss A, O'Geen H, Farnham PJ, Madden PA, Mungall AJ, Tam A, Kamoh B, Cho S, Moore R, Hirst M, Marra MA, Costello JF, Wang T. (2013) DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nature Genetics, 45:836-847

Yalcin B, Wong K, Agam A, Goodson M, Keane TM, Gan X, Nellåker C, Goodstadt L, Nicod J, Bhomra A, Hernandez-Pliego P, Whitley H, Cleak J, Dutton R, Janowitz D,

Mott R, Adams DJ, Flint J. (2011) Sequence-based characterization of structural variation in the mouse genome. Nature, 477(7364):326-9

Yen TT, Gill AM, Frigeri LG, Barsh GS, Wolff GL. (1994) Obesity, diabetes, and neoplasia in yellow A(vy)/- mice: ectopic expression of the agouti gene. FASEB J., 8(8):479-88

Youngson NA, Whitelaw E. (2008) Transgenerational epigenetic effects. Annu Rev Genomics Hum Genet., 9:233-57

Yuan P, Han J, Guo G, Orlov YL, Huss M, Loh YH, Yaw LP, Robson P, Lim B, Ng HH. (2009) Eset partners with Oct4 to restrict extraembryonic trophoblast lineage potential in embryonic stem cells. Genes Dev., 23(21):2507-20

Zampieri M, Guastafierro T, Calabrese R, Ciccarone F, Bacalini MG, Reale A, Perilli M, Passananti C, Caiafa P. (2012) ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites. Biochem. J., 441:645–652

Zentner GE, Tesar PJ, Scacheri PC. (2011) Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res, 21:1273–1283

Zhang Y, Maksakova IA, Gagnier L, van de Lagemaat LN, Mager DL. (2008) Genome-Wide Assessments Reveal Extremely High Levels of Polymorphism of Two Active Families of Mouse Endogenous Retroviral Elements. PLoS Genetics, 4(2): e1000007

Zhou W, Bouhassira EE, Tsai HM. (2007) An IAP retrotransposon in the mouse ADAMTS13 gene creates ADAMTS13 variant proteins that are less effective in cleaving von Willebrand factor multimers. Blood, 110(3):886-93

Zuo X, Sheng J, Lau HT, McDonald CM, Andrade M, Cullen DE, Bell FT, Iacovino M, Kyba M, Xu G, Li X. (2012) Zinc finger protein ZFP57 requires its co-factor to recruit DNA methyltransferases and maintains DNA methylation imprint in embryonic stem cells via its transcriptional repression domain. J Biol Chem., 287(3):2107-18

Supplementary table 1. Primers for pyrosequencing

Primers for BLUEPRINT WGBS-seq validation			
Gene	Forward primer	Reverse primer	Sequencing primer
C5ar2 1st intron	[Btn]GATGTGAGAAATTGGGAAGTTAG	CCCACCAAACTAATAAACTAAATCC	CCTAAAAAAAAATATACACAAACA
Pik3c2b 13th intron	[Btn]GTGAGAAAGGAAATAGGTTAGGTTAAGAT	AATCTCCTTACCCTAAACTCAAA	CTAAACTCAAAAAATATAACTAAAT
Cd3e 2nd intron	TTGATTTGGGTTTGTTATAGTGGTTAG	[Btn]CTCACCCTACACTAAAACTATTCATAA	GGTTTGTTATAGTGGTTAGT
Intergenic	AAATGTTGTGTGGAAGGAAAATTAT	[Btn]TACAATCCCCACCTTCCTCTTATA	GGAAGGAAAATTATAGGTTATATG
Lef1 3rd intron	[Btn]TTTGGGTAAGGAATAGTTGAAGG	CCCTATAACCTAAACCACCATTCAATACAC	ATTCAATACACACTATACTCA
Blnk 1st intron	TGTATAAGTATAGGTGGGGAAGTT	[Btn]ACCCAAATTCTCAATTCTTCAAC	ATGTAAAGTTTTATTTTTAAGTTAG
Cdyl2 4th intron	AAGAGAGGAAGTAATAAGTAGTAAGTGTAA	[Btn]ACATCTATCTATTCTCAACCTATATCAA	AGTAATAAGTAGTAAGTGTAAG
Ets1 3rd intron	TGTTTTTAGGAGGTTAGTAGGTATTTG	[Btn]CCACACTACCTAAATTCCAACTTCATTAA	GTAAAGAAATGTAGTTAGGTT
Chst3 1st intron	AAGTTGTTTGATTTTATGGGATATAAGAGA	[Btn]CCTACTCCTCCAAAAATAACTCTAAT	TGATTTTATGGGATATAAGAGAA
Bcl11b 1st intron	TTGATGGTGGGGTTTTTAGATT	[Btn]ACTTTCCTATATCAAAACAAACTCTTC	GTTTGTGAGATGTATAGATAG
Pou2f2 1st intron	ATGGGTTTTTATAGTGTAGTATTGGTAAGA	[Btn]AACCACCAAAACCTCTAAAACTCTCC	GTGTAGTATTGGTAAGAATTA
Intergenic	AGGGTAAAAGTGAGTAGTAAAATTTGT	[Btn]CTTAACCACTACCTACCCTACTAA	AGTAGTAAAATTTGTAATTAAGGA
Intergenic	AGTATAGGGGTTTGTGGTAGA	[Btn]CCACATCTAATTTCTATTCCTCTCTAACT	GTTTGTGGTAGAAGTGTA
Intergenic	TGAATAGTGAAAGATGGTTTGAAAGATAG	[Btn]TAATTTCCCCCCTTCTTTTTTTCTTTACTT	ATGGTTTGAAAGATAGTTTATAT
Tcf7 3rd exon	GGTAGTTAGTTTTATTATAGGGTTTGATT	[Btn]CTACTCTACCTTCAATCTACTCAT	GTTTTATTATAGGGTTTGATTGT
Intergenic	GAATTGGGAGGTGTATTTTGATTTAA	[Btn]TTCATAATAAAAAACTCAAACCCCATAC	AGATTAGAGGAAAGATATAATTG

Spi1 1st intron	TGTGATAGGTGGGGTTTGAATAG	[Btn]CTCCAAATCTCCTCATCCATAAC	GTTATTTATAGATGGGTTTAGT
Cd3d 1st intron	GGAGTAGGATTTTTGGTTATGTGAT	[Btn]CCTCTACCTATCCTTCCCATTCAA	GGATTTTTGGTTATGTGATTAT
Intergenic	GTAATTAGGGATGGTAATAAAATTAAGTGG	[Btn]AAAAAAAAACATAAATCTCTCAACCTTAA	GAGATTTTGTTAAAGGATGAA
Intergenic	GTGTTTGAATTTTATATAGGAAGT	[Btn]TACTAAAACCCCCCTTATTCATC	GAAGTGTAGAGGGGA

ERV primers

	Forward	Reverse	Sequencing
IAP_5LTR_Slc15a2_1	GGAAGTATAGAGAGAGTTATGGGGTTTA	[Btn]CCTTCTATTTTAAAAAAAACAACCATTACC	TTTTGGAGTGTGGGA
IAP_5LTR_Slc15a2_2	GGAAGTATAGAGAGAGTTATGGGGTTTA	[Btn]CCTTCTATTTTAAAAAAAACAACCATTACC	ATTTGAGTATGAGTTAAGGGTAT
IAP_5LTR_Eps8l1_1	AAGGGGGATTGTGGTTGGTTATTAG	[Btn]TCACTCCCTAATTAACTACAACCCATAAC	GGGTGGGTGTGGGAA
IAP_5LTR_Eps8l1_2	AAGGGGGATTGTGGTTGGTTATTAG	[Btn]TCACTCCCTAATTAACTACAACCCATAAC	TGTTTTAAGTTGGTAAATAAATAAT
IAP_5LTR_Bmf_1	TTTTGGTTGAAGGGATTTTTATAGT	[Btn]CTCCCTAATTAACTACAACCCATAAC	TTTTATAGTTAATAATTGTTGGGA
IAP_5LTR_Bmf_2	AGTAGGATTAGATGGGTAGAGT	[Btn]CCCCTCCCCTTTAAAAAAAAATAACCCT	GTGTTTTAAGTGGTAAATAAATAAT
IAP_Ylpm1	GGGAAAGGTAGAGTATAAGTAGT	[Btn]TCCCTAAACCTTTTCTTCCTTCTC	ATTATTTAGAATATAGGATGTTAG
IAP_Rab6b	AGTGGGTAGGGTAGGTTAGATAAA	[Btn]CCCTAATTAACTACAACCCATAACC	ATGTAAATAGTTGTTGGGA
IAP_5LTR_Rnf157	GTTATTTGGGATAGGATGAGAGTTAGTA	[Btn]ATCACTCCCTAATTAACTACAACC	AGGAGGGTGTGGGAA
IAP_3LTR_Rnf157	TAGAAAGTTGAGTGGGTTGTATTG	[Btn]CTTACTTCTTACTCTCTTTTCCTAAA	AGGTTGATATTTTTTTTTGTTAT
IAP_5LTR_Marveld2	AGGAAGGAAGGTGATATAGAGA	[Btn]CCCTAATTAACTACAACCCATAAC	TTTAAGTTGGTAAATAAATAATTTG
IAP_3LTR_Marveld2	AGGGTTTTAGGATTTAGTTATGTTAGTATT	[Btn]ACCATAAACTACAACCAATCAAAAAATA	TTAGTTATGTTAGTATTTTATTTGT

IAP_5LTR_Mbnl1_1	AAGTTTGTAATGGTGGGAGATTAA	[Btn]CCCTCTTTAAAAAAAACAACCATTACCT	AAAATAAATTGTGGGAAGT
IAP_5LTR_Mbnl1_2	AAGTTTGTAATGGTGGGAGATTAA	[Btn]CCCTCTTTAAAAAAAACAACCATTACCT	TTTGTGTTTTAAGTTGGTAAAT
IAP_3LTR_Mbnl1	TTGTTTTAGGGTTAATTTTGAGTGT	[Btn]TTCCTTTTCTCTCTCTTACTTCTTACTCTC	AAAAGGTTATATTATATAGTTTGT
IAP_5LTR_Tfpi	GGAGGAAGTAATAGTGTTAATAGGT	[Btn]CTCCCTAATTAACTACAACCCATAAC	AGTAATAGTGTTAATAGGTTAGA
IAP_5LTR_Pgm1_1	GTTTGGTTTTTATATAGAAGGAAAGAAGAA	[Btn]ACAAATAATCATAAAATACCCTTAACTCAT	ATGTTTGTTTGTTGGG
IAP_5LTR_Pgm1_2	GTTTGGTTTTTATATAGAAGGAAAGAAGAA	[Btn]ATCACTCCCTAATTAACTACAACC	GTGTTTTAAGTGGTAAATAAATAAT
IAP_5LTR_Sfpq	AGATATTGATTAGATTTGTTAGGAAGTAT	[Btn]CCCTTTCCTCTTTAAAAAAAACAACAATT	TGTTTTAAGTTGGTAAATAAATAAT
IAP_3LTR_Sfpq	GGTTGTAGTTAATTAGGGAGTGA	[Btn]TCAACTAACTATACACACTCCATTACTC	TTGGTTTGTGGTGTT
IAP_Gm5444	GGTTTGGTTTTGTTTATGAAGAGTT	[Btn]ATCACTCCCTAATTAACTACAACC	ATATGTATATGTTGTTGGGA
IAP_5LTR_Rims2	TTGTAAGAGTTTTTTGAGGGTTTTATGTAA	[Btn]CATCACTCCCTAATTAACTACAACC	ATTTTGTTTGTGGGAAG
IAP_Sema6d	AGGATTAAGTGAAATGATGTATTAGAGT	[Btn]TCCCTAATTAACTACAACCCATAACC	GTATTAGAGTTTTTTTTTTTGTAAG
IAP_5LTR_Gm13849	AGGGAGTATTATTTTTTGATTGGTTGTAGT	[Btn]ACACAATTCTATTTCTAATCCATTATATCT	TTTATTAATTTAGAATATAGGATGT
IAP_Wdr1	ATGGGGTAGAGAAATATTTTTGATATATG	[Btn]ATACTTAAATAAACCCAAAACTATCCC	GTTTATTATTTAAAATATAGGATGT
IAP_Ect2l	GGGAAAGGTAGAGTATAAGTAGT	[Btn]CCACTACCCAACTAAATCATAACA	TGTTTATTATTTAGAATATAGGATG
IAP_5LTR_Vmn2r97	GGAGGAGGAGATTTAGAGTATAGG	[Btn]ATAACCCTAAAAAAAAACCTTCATACCC	AGGAGATTTAGAGTATAGGA
IAP_2610035D17Rik	TTTTGGATAGGTATAGTTTGTTGAGTTT	[Btn]AATAATCATAAAATACCCTTAACACATAC	GGTGATTGTTGGGAG
IAP_Tfec	GGGAAAGGTAGAGTATAAGTAGTTATAAA	[Btn]CTATCTAAAAACCAATCAAAACTTTAAAA	TGTTTATTATTTAGAATATAGGATG
IAP_Pifo	[Btn]GGGAAAGGTAGAGTATAAGTAGT	CCACACTTCCATCTTCCTTCTTCTTAA	CTTCTTAAATTTCATATAATTTATT
IAP_5LTR_Tle1	GGAATTTTAGAGTTGTTTTGATTTGTATTT	[Btn]ATTTTTCACCCTAACTAATTACTCAC	AGGTGTTTGTGAGGA

IAP_Tmprss11d	GGGAAAGGTAGAGTATAAGTAGT	[Btn]AAAAATATTCCTCTTCTCCACATATTCA	TTTATTATTTAGAATATAGGATGTT
IAP_5LTR_Kcnh6	GTGGAAGTTAGTGATTAGTTTTAGATG	[Btn]TCCCTAATTAACTACAACCCATAACC	TTTGTTTTTGTGTGTGG
IAP_5LTR_Bmf	TTTTGGTTGAAGGGATTTTTATAGT	[Btn]CTCCCTAATTAACTACAACCCATAAC	TTTTATAGTTAATAATTGTTGGGA
IAP_5LTR_Vmn1r3	AGGTTAAAGTTTTTAGGGGTTGG	[Btn]CTTTACCACCTAAATACAAAATCTAACA	TTGGTTGGTTGTGTT
IAP_5LTR_Ccl28	GTGTGGGTGTTTTAGTGGT	[Btn]AACAACACCCATAATCCTCATAC	GTGGAATTGTAAAGGG
IAP_Fam78b	TGTTTAAGTTGAGATGTGGATAGT	[Btn]ATCACTCCCTAATTAACTACAACC	TTTATTTAGTATAATTTTGTTGGG
IAP_Gm5936	ATGGATAGATAGAAGGATTGGGAATA	[Btn]ACAACACCCATAATCCTCATACA	GGAATATTTAAGATTGTAAAGGG
IAP_Zfp619	AGTTTTATTAAAGGTGAAGAAGTGTAGA	[Btn]CTTCACTTAAAACCCATCACTCCCTAAT	TGTAGATAGTTGTTGGGA
IAP_Mmp16	AATATTTTATTGTGGTAAGGATGTATAGAA	[Btn]CATCACTCCCTAATTAACTACAACC	GGATGTATAGAATGTTGGG
IAP_Diap3	AGATGTAAGATAGAAGGGGTTTT	[Btn]ACAAACTAAAAATTTCCCACTCTCC	AGTTTTTGAAGATGTAAGTAATAAA
IAP_5LTR_Snd1	GGGAAAAATAGAGTATAAGTGGT	[Btn]CCAAACTAAAATCCAAAAAACATAACC	GTTTATTAATTTAGAATATAGGATG
IAP_3LTR_Snd1	GTGGTTAATGGTTTTAAGGGATAGA	[Btn]CAACTCCACCATAAACTACAACCAATCAAA	AGAGGGATAGGGAGG
IAP_Creb3l2	TGTAGATGGAGTTTGTGTGTAGTA	[Btn]ACAAATAATCATAAAATACCCTTAACTCAT	GGTTTGTGTTGGGAG
IAP_5LTR_P4ha1	AGGTTGGAGAGATAAAGGTATTGT	[Btn]TCACTCCCTAATTAACTACAACCC	GTGTTTGTTTGTGGGA
IAP_Cdh19	GGGATTTTTTGGTTTTAATATAGAGGT	[Btn]ATCACTCCCTAATTAACTACAACC	ATGTGTAAATTGTTGGGA
IAP_5LTR_Rftn1	[Btn]GGGGAAGGTAGAGTATAAGTAGT	AAAACCAACATAAAATATCTTCCTATATC	CCTATATCACTCTCTACCTATT
IAP_5LTR_Nrros_1	GGGAGTATTATTTTTTGATTGGTTGTAGT	[Btn]TCCTATACTTTACCCCTCTACAATCTTCCT	GTTTATTAATTTAGAATATAGGATG
IAP_5LTR_Nrros_2	TAGGGTTGAGGTTTTAGAGTG	[Btn]CCATTACCTAAAACCCATCACTCCCTAATT	TGTTTTAAGTTGGTAAATAAATAAT
IAP_Psd3	GAAAGGTAGAGTATATGTAGTGGTAAAA	[Btn]AAAAAAAATCCAAAAACTTATATAACCT	ATTTGTTTATTATTTAGAATATAGG

IAP_2010015L04Rik	TAGTTGGAAATGGGTGTAGGA	[Btn]ACAAATAATCATAAAATACCCTTAACTCA	GGAAATGGGTGTAGGAA
IAP_Mrps30	TGGGAAAGGTAGAGTATAAGTAGT	[Btn]ACATATATAACCAAACTTCAAACACAA	TGTTTATTATTTAGAATATAGGATG
IAP_Kmt2e	AGATGTTTTTTAATTGAGGAATGGATATAG	[Btn]CTAAAACCCATCACTCCCTAATTAACTACA	ATGAAATTTTTAGGTTGTTGG
IAP_Zfp820	AGATGAGGTTATGGTTAGTTGTGA	[Btn]CACTAAAATTAACTACCCACTATCAATC	GGTTAGTTGTGAGGAG
IAP_Letm1	AGGGTTGAGGATATGTTGG	[Btn]AAACCCATCACTCCCTAATTAACTACA	TTGAGGATATGTTGGG
IAP_Cyp4f18	AGTGTTTTTTTGAAATGGATATAGGT	[Btn]TAATAACAACACCCATAATCCTCATACA	GGTGTAAAGGGTTGTAATAT
IAP_5LTR_Cdk5rap1_1	GGATTTGGGGAAGAAAGTATTAGTA	[Btn]ACTCCCTAATTAACTACAACCCATAA	GGTGTTTTAGTGTGGG
IAP_5LTR_Cdk5rap1_2	GGATTTGGGGAAGAAAGTATTAGTA	[Btn]ACTCCCTAATTAACTACAACCCATAA	AAATAATTTGAGTATGAGTTAAGG
IAP_Smg1	GGGGAAGGTAGAGTATAAGTAGTTATAAGA	[Btn]TTATTTCCCTAAACATAAAATCCCTCATCA	TGTTTATTATTTAGAATATAGGATG
IAP_E030019B13Rik	GGGAGTATTATTTTTTGATTGGTTGTAGT	[Btn]ATAATATATAAACTCTAAACCATAACACAA	GTTTATTAATTTAGAATATAGGATG
IAP_Bai1	GTTTAAGGGAAAAATAGAGTATAAGTAGT	[Btn]ACCCAACTAAAAACTCTTCCA	GTTTATTAATTTAGAATATAGGATG
IAP_Dph5	TGGTTATAGTGATAGGGAAATGAGG	[Btn]AAACACAAATAATCATAAAATACCCTTAAC	AGGGTGATGTTGGGA
IAP_5LTR_Cdc73	TGTGATAAAAGAAAATAGAGGGAATTAAG	[Btn]CCCTAATTAACTACAACCCATAACC	TTTTATTTATAGTTTATATGTGGGA
IAP_3LTR_Cdc73	TGGAGGAGTTAGAGAAGGTATTAAAAG	[Btn]CTTCTTACACTCTAACCCCATAAAAATATA	GGATTTTGTTTGTATGTGAA
IAP_Zscan4c	GGGAAGGTAGAGTATAAGTAGTTATAAGA	[Btn]AATAACCCTTCTCCTTAACCTCTTT	AGATTATTTGTTTATTATTTAGAAT
IAP_H2T24_1	GGTAGAGTATATGGAGTGAAGAATTAT	[Btn]ATTCCCTAAAAAACAAACTCACTT	TTGTTTATTATTTAGAATATAGTT
IAP_H2T24_2	GGAAGTGTTATTTTTTGATTGGTTGTAGT	[Btn]ATTCCCTAAAAAACAAACTCACTT	ATTTTTTGATTGGTTGTAGTT
IAP_Slc24a3	GGGAAAAATAGAGTATAAGTGGT	[Btn]TAATCCTACTTCAACAACTAAACTATCTC	GTTTATTAATTTAGAATATAGGATG
IAP_Celf2	TTTTAGTTGAAGTTGGGTGGGTAGGAA	[Btn]CACTCCCTAATTAACTACAACCCATAAC	GGGTAGGAAGGTGTG

IAP_Trak1	TGAGTTAGTTATAGTTGGTTTGTATGAA	[Btn]ACACAAATAATCATAAAATACCCTTAACA	GGGTAGTTGTTGGGA
IAP_Cr2	GGGAAAGGTAGAGTATAAGTAGT	[Btn]TCTAAAACTAAAAACTCCATCTACTTAC	ATTTGTTTATTATTTAGAATATAGG
IAP_5LTR_Diap3_nonrag	GGGGTTTGATGTTTTATTTTATTTTGTTTA	[Btn]AACACATACCCAAATTATTTATTTACCACT	GTGATGTTTTGTTGGGA
IAP_3LTR_Diap3_nonrag	TTTGGTTTTTGAAGATGTAAGAATAAAGT	[Btn]ACTCACACTTCCAAAAAATTCAATTAATA	AAGTTTTGTAGTAGAAGATT
IAP_Adamts13	GGGAATTGAAGGTATATGTGAGT	[Btn]CAAATTATTTATTTACCAACTTAAAACACA	GGGTATTAGTGTGGGAA
IAP_Zak	GTTGTTTTTTTGTTTTTTGGATGATTT	[Btn]TCCCTAATTAACTACAACCCATAACC	TGTTTTTAGATATTGTGGGA
IAP_Scn7a	GGGAAAGGTAGAGTATAAGTAGT	[Btn]CCCATTTTAAATTTATTAACTAACACCACT	ATTTGTTTATTATTTAGAATATAGG
IAP_Steap1	GGAAGGTAGAGTATAAGTAGTTATAAGAT	[Btn]CAAACCTACAAATAAACATTTAATTTTAAC	TTATTATTTAGAATATAGTTGTTAG
IAP_Ell2	GGTTATGGAAGAAAATGTAAATGATAAAAG	[Btn]ACACAAATAATCATAAAATACCCTTAACA	AAATTATTTTTATGTGTTGGGA
IAP_Cdh20	GGGGAAGGTAGAGTATAAGTAGTTA	[Btn]ATCTTCTACAACCACCCTTAATC	GTTTATTATTTAGAATATAGGATGT
IAP_5LTR_Dst	GATTAATGAAAGGATATAGGGAGAAAG	[Btn]AACACATACCCAAATTATTTATTTACCACT	ATTTGTTGATTGTTGGGA
IAP_3LTR_Dst	GTATTTTGGTTTTTGAAGATGTAAGAAA	[Btn]AAACCTTAATAAACTCTACATCTCTATCA	AAGTTTTGTAGTAGAAGATT
IAP_Sqcz	TGGGAGTGATTAGGAGTGAGTT	[Btn]TTCCCATATTCTATTCTAAAAACTTCAC	GGAGGTTAGAGTTATTTTATAATGG
IAP_D19Bwg1357e	AGATGGAAAAGTGTGTAGGAGAAGTA	[Btn]CATCACTCCCTAATTAACTACAACC	AGGTTTGGGTTTTATAATG
IAP_Cdk15	GAGTAAAGGTAGAGTATATGTAGTGG	[Btn]CACCAAAAAACAAAACCAAACTATACTT	TGTTTATTATTTAGAATATAGGATG
IAP_Lmcd1	TGTTGAGATATATGGGTTGTAATGTG	[Btn]TTTCCCAATTCCCAATCTTCCCTCCTAATA	GTTGTAGTGGGATTGT
IAP_3LTR_Bex6	GGGGTTGTGTAGGAAGGAAATG	[Btn]CCTACACCCTAACTCCTAAAAATATA	ATTTGAGGGTTTTTGTTAT
IAP_3LTR_Mrs2	TGTGAGTAGTAGGGTTTGTTTTAAT	[Btn]CCACACACTTTCTAAACTTTCTTTATCT	GTTTTTAAAGTAGTAATTTAGTATT
IAP_3LTR_Arhgap31	AGATAAGTTTTTTAGTTTGGAGGTATTT	[Btn]TTACTTCTTACACTCTAACCCCAAAA	AGGTATTTTTGTTTGTTTGTTA

RLTR10_4930519G04Rik	GTGTTTTATATGAGGAAGTTAGGTGTAAG	[Btn]AAATTTCTCTTTCTAAACCCACCTTAT	AAGTTATATGATTTAGTTGTAGT
ETn_Sntb1	GGGTAAAAGAAGAAATGTAGTTAAGATAG	[Btn]TCTCATTAAAAAAAAAATAACCCACACTC	TGGTTTAGTAGAAAGGTGTA
ETn_Bhlhe40	GTTGGTTTTGTGAAAAATAGAGTGT	[Btn]CCCCCTCTAAATTCCTCTCTTACAACT	AAGTTTTTTAGTTTGAGTTATTTTT
ETn_Wiz	GGGTAAAAGAAGAAATGTAGTTAAGATAG	[Btn]ATAACCTTCACCTCCTCTACA	TGGTTTAGTAGAAAGGTGTA
RLTR30D_Bach1	AGATTGGGTAGATTTTAGTGATTATAAGTT	[Btn]AACTCCACTTCTCCTCCCTAAAT	GGAGAATAATAATTTTTAATTTTTT
LTRIS2_Dusp11	GATTATTGGTAAGATTGGAATAGGGTTAAA	[Btn]CCCTTCCCACCTCCTTAAATTAT	TTAGTGTAGATGATGAAAGA
RLTR1B_Scd3	TTTTGTTTTTTGAGGGAGATATGTGT	[Btn]CCAACTTAAACACACCTACATTATC	ATTTTATGTTTTTAAATTAAGATG
RMER21_Art3	TTTGAGGGGTAGTTTTTATTGAGT	[Btn]TTCCTCTTATCTTAATCACTAATTCTT	GGGGAAGGTAGTGTATTA
VL30_RLTR6_Dhx40	AAGGTGGTAAGTTGGGTTAATT	[Btn]ACTACCTTATTAAATCTTACCTTCTACA	GTTAATTTGAGAAATTGAAAGAT
MuLV_RLTR4_Pik3c3	AGTATAATTGTAAATGGAGATTAAATTGT	[Btn]ACCAAAACATTATACACCTCTATTTTTCT	ATATTGGGATAGGGG
MuLV_RLTR4_Cep85	GTTGGAAGTTAAATTTTGGGTAAGTA	[Btn]AATACCATCTATTCTTAACCCTAAAC	ATTAGGATAAGGGTTAAATA
MuLV_RLTR4_Usp10	ATGGATGGTGATATGGTTATAGG	[Btn]AACCAAAACACTATACACCTCTA	ATATTGGGATAGGGG
MTD_Pde6a	AGATGGGTTTGTTATTGTTTGAT	[Btn]CCTACTTAACTCCTTTCTACATCAC	TGTTTAGATTGATAATTGATGTG
ORR1E_Tnks	GGGAATAGGGAGGTAGAGAT	[Btn]ACAAACTAACCCTTTCCTAAACCATTAACC	AGGGAGGTAGAGATT
ORR1D2_Sim2	GAAGTAGATGGAATGAATTGTAATGT	[Btn]CAACTAACCAAACCAATAACTAATCTTTC	AATGTTGAGTTTTTTAGATTTTT

Supplementary table 2. Primers for Q-PCR

Q-PCR primers for BLUEPRINT RNA-seq validation		
Genes and isoforms	Forward primer	Reverse primer
Rap1b-201	CAGGAACGGAGCAATTCACA	ATGTCGACTGTGCTGTGATG
Ugcg-001	TGGCTTCAAATGTGTGACGG	TGAAAACCTCCAACCTCGGT
Ugcg-001-002	TGTCCATCATCTACACCCGG	CCCCTTCAGTGGCTTCAGAA
Ski-001-003	ACTCAGCCCAGATTGAGGAC	ACCACCTTCTCCAGATGCTC
Itk-001-002	AGGGCTCCATTGAACTCTCC	CATGCACGACCTGAAAAGGG
Itk-001-004	TGGAGGAGGTTTGGTCACTC	GCACGAACGTTAGCTCTGAG
Itk-001-002-004	ACAACAGGCGGTCATTTCAG	TTGTCTTGAACCCTCCACCA
Chst3-201-001	CCTCACAACTTTCCCACAGC	GGGGACACTCTGATCCTGTA
Chst3-201-001-002	TTTCCGGGACCTTGTACACA	TCGGAGACCCTGGATATGATT
Nfkbia-201	GAAAGCTGGCTGTGATCCTG	ACGTGTGGCCATTGTAGTTG
Ptprc-all	TTTCTTAGGGGCACAGCTGA	GTGTAGGTGTTTGCCCTGTG
Smim14-001	GCTGTTTTGGGCGGGTTTAA	GCCTCCTCATAGCATGTTCG
Smim14-001-004	GTCTGCTCTCACGAACATGC	AGAAGGAAGAGGAGCATGGC
Arpc2-002	GCAGATTTTGATGGCGTCCT	ATCAGCTCCATGTGCCTGAA
Dgka-201	TAGTCAAGTTGCGGCTGAGA	TTCCGTGCTATCCAGGATCC
Fam105a-201	CAGCTGAAATGGTGGATCGG	CTCCCTTCCACTCTCTTGCA
Laptm5-001-006-007-009-010	GTACCTCAGGATGGCTGACC	GTACGCTGGCAATTCGATGT
Selplg-201	CTCATCCCGGTGAAGCAATG	CACTGGGTACATGTGGGTCT
Hvcn1-all	CTACTCGAGACAGCAGGGTG	CCTTTGGGTCATGGGAAGTC
Cflar-002-003-004-005	TGCTACCTTGGCTGAATTGC	CTCCATCAGCAGGACCCTAT
Iscu-001	CGTCATGAAACTGCAGATCCA	ACCGTTTTCCCCTTTACCCA
Pik3c2b-001	TCCCTAGCCGGTTTGTGATT	AGATCACACTCAGCCACCTC
Rhoh-201	GCATTCTACCCCAGTTCCCA	GTGAACAGCTCTTGACTGGC
Stk17b-001	ACTGTAAGAAGCGTCGGTCA	GGCAGACTTTTACCTCGCAC
Spi1-001	GGAGACAGGCAGCAAGAAAA	CCTTGTGCTTGGACGAGAAC

Spi1-001-002	GGATGTTACAGGCGTGCAAA	CCAGTAGTGATCGCTATGGC
Cd3d-201 (1)	TACCGAATGTGCCAGAACTG	CGGTCTCATGTCCTGCAAAG
Cd3d-201 (2)	CTTCTGGGGCTGCTGAGG	CAAGACGGCTGTACTGGGTA
Tcf7-001-002(1)	CGCGGGATAACTACGGAAAG	AGCACTGTCATCGGAAGGAA
Tcf7-001-002(2)	CCCCAGCTTTCTCCACTCTA	GCCTGTGAACTCCTTGCTTC
Bcl11b-003	GCAGGAGAACATTGCAGGTAA	TCCAGGTAGATTCGGAAGCC
Cd2-001	CGATAAACCCGGTCAGCAAG	AGCTCTTCATCTTTTCTCCTCCT
Cd2-001-002	CGTGAGGATTCTGGAGAGGG	TCGCCTCACACTTGAATGGT
Stat5b-002	ATTCGGGCTCTGTAGGTTGT	TGGCTTTCGATCCACTGTGA
Stat5b-001-002	TTCAACCGGGAGAATTTGCC	CCGTCTGGCTTGTTGATGAG
Ets1-001-003-004	TCCTGCAGAAAGAGGATGTGA	TCCGAGCTGATAGGATGCAG
Elf1-202	TTAACAGCACAGAACTTGCCG	AGGTCTTCTTGTCCCTTTAAACT
Elf1-201-202	AGATGGGAAGGGAAACACAAT	CCCCATGGTCTCGTAGTTCA

Q-PCR primers for IAP-driven expression analysis

	Forward	Reverse
Eps8l1_ex12	GAGTCCTCAGGCACCTCC	ACTTTGCACTTGGTTTGGGT
Eps8l1_ex45	TGTCAATCACCTGGTCACCT	TCAGTAGCATCTCCTGTGCC
Slc15a2_ex12	TGAGTCCAAGGAAACGCTCT	GAAGAGCTTCGGAGTTGACT
Slc15a2_ex57	TGAAGAGGAACATGCAGAGG	GCATAGCAGTCTTCGCCAAA
Slc15a2_ex910	GCTTCAGGAACCGTTCTGAG	CGTCCATAATGAGGTGCTTTGG
Slc15a2_ex1920	GCAGTGCATTGTGAAACGGA	CCCCGGTTGGTGATATTAGTG
2610035D17Rik	CCCAACAGTCACCCATCCAT	GTCTCCTCGGCCTTTCTCTT
Bmf	CCAGAGACTCTTTTACGGCAAC	TGTTGCGTATGAAGCCGATG
Bub1b	GTCTCTGGATCAAATTGGGACA	CCTTGCGTTCAATCCCTTCC
Tfpi	TGAAGGCAGATGATGGTCCA	AGCTGTCTTCTCATAACCTGGT
Gm13710	TTGCTGTCCTGGAACTCGAT	GCCATGTGTGCTGTATTCATCTT
Cdk5rap1_ex56	CTATCAGGGAGAAGGCCGAG	TCAGCCATGCAGCCTAGAAT
Cdk5rap1_ex89	AACCCAAACAAGGAGGGCTT	CGAATCAGCTGAAGAACCTCA

Supplementary table 3. Validation of the threshold for genome-wide identification of metastable epialleles

IAP coordinates	Computational Variation Score	Nearest gene	Tissue	N of individuals	Experimental variation
chr3:60489872-60495084	64.4375	Mbnl1	liver	10	39.47619
chr13:4942652-4943122	63.69791667	Gm5444	spleen	10	40.835
chr15:39173040-39178257	63.02083333	Rims2	spleen	10	13.03583
chr2:123754520-123754992	61.1875	Sema6d	liver	8	29.2225
chr6:31799369-31804637	60	Gm13849	brain	8	36.62979
chr5:38598377-38598869	58.33333333	Wdr1	liver	10	17.07375
chr16:36765054-36770352	53.27380952	Slc15a2	spleen	10	23.34
chr11:116389404-116394709	53.27083333	Rnf157	spleen	8	20.93938
chr10:18139120-18139597	52.08333333	Ect2l	brain	8	12.35646
chr17:18967877-18969281	51.95436508	AK134158	spleen	8	6.936667
chr13:100588951-100594225	49.10416667	Marveld2	brain	10	42.012
chr2:84505209-84510421	48.44998751	Tfpi	liver	10	62.68666
chr11:113173016-113173489	46.61011905	2610035D17Rik	spleen	8	37.44111
chr5:64030833-64038297	45.8125	Pgm1	liver	18	90.16625
chr6:16359772-16360249	44.4375	Tfec	spleen	8	17.9025
chr3:106003601-106003967	41.97916667	Pifo	brain	8	12.02688
chr4:71291215-71293236	40.77083333	Tle1	liver	8	2.238125
chr5:86318272-86318748	40.24404762	Tmprss11d	brain	8	21.75584
chr11:106028245-106033560	39.54166667	Kcnh6	liver	8	37.59125
chr2:118554765-118558375	38.90029762	Bmf	liver	8	43.44938
chr4:3192211-3200530	37.7827381	Vmn1r3	brain	8	4.708335
chr13:119661605-119669925	37.2485119	Ccl28	spleen	8	21.69
chr9:103108601-103109112	36.875	Rab6b	liver	10	44.12667
chr1:166940021-166940493	36.08333333	Fam78b	liver	10	23.288
chrX:74872946-74880330	35.88541667	Gm5936	brain	8	12.03375
chr7:39655241-39655724	35.41964286	Zfp619	spleen	8	15.40917

chr4:17826925-17827396	34.39583333	Mmp16	liver	8	11.27313
chr14:86398117-86398596	34.375	Diap3	liver	10	22.95222
chr7:4455441-4460702	33.875	Eps8l1	spleen	10	28.99167
chr6:28924174-28929353	32.77083333	Snd1	liver	8	24.9575
chr6:37464145-37464607	30.5625	Creb3l2	spleen	10	2.13
chr10:59381798-59387060	30.20833333	P4ha1	brain	8	10.99375
chr1:110612994-110613452	29.66666667	Cdh19	brain	8	8.63375
chr17:50074462-50078819	29.16666667	Rftn1	spleen	10	2.986668
chr16:32148076-32153339	28.125	Nrros	kidney	8	26.15292
chr12:85003517-85003995	25.95833333	Ylpm1	liver	10	7.511663
chr8:68160165-68160570	24.85119048	Psd3	spleen	8	8.098002
chr4:155421681-155422048	24.82954545	2010015L04Rik	spleen	10	8.48067
chr13:118269009-118269481	23.4375	Mrps30	spleen	10	10.7075
chr5:23379564-23380035	23.22916667	Kmt2e	spleen	8	3.79778
chr17:21801081-21808185	22.8125	Zfp820	spleen	10	13.19867
chr5:33759665-33760140	22.08333333	Letm1	spleen	10	7.1275
chr8:72024782-72033082	21.875	Cyp4f18	kidney	8	4.505833
chr2:154354262-154359592	20.9375	Cdk5rap1	spleen	18	12.19667
chr7:118268458-118273558	20.3125	Smg1	spleen	8	2.030835
chr12:56525184-56525535	20.1875	E030019B13Rik	spleen	8	3.784167
chr15:74360731-74367892	19.63728632	Bai1	liver	8	4.83375
chr3:115909127-115914440	18.75	Dph5	brain	8	1.7
chr4:127008519-127015691	17.85714286	Sfpq	liver	10	12.11084
chr1:142584891-142590150	15.875	Cdc73	kidney	8	3.264373
chr7:11054280-11058647	15.625	Zscan4c	liver	8	3.5065
chr17:36007143-36014240	15.10416667	H2T24	liver	8	3.188645
chr2:145118403-145123718	14.23611111	Slc24a3	spleen	8	2.673125
chr2:7404072-7409352	13.5625	Celf2	liver	8	3.0125
chr9:121322377-121328483	13.54166667	Trak1	brain	8	5.003

chr1:195320289-195327446	12.5	Cr2	lung	10	4.44333
chr14:85889095-85896207	12.5	Diap3_nonrag	kidney	8	3.491
chr2:26998297-27004317	11.71875	Adamts13	kidney	8	5.645625
chr2:72270361-72275599	11.45833333	Zak	kidney	8	4.4275
chr2:66845102-66852262	10.10416667	Scn7a	brain	8	2.4425
chr5:6184072-6191390	8.333333333	Steap1	kidney	8	5.167168
chr13:75677017-75682366	8.3125	Ell2	spleen	8	0.671667
chr1:104289577-104294815	7.5	Cdh20	liver	8	3.045553
chr1:34347456-34351369	7.291666667	Dst	kidney	8	1.221
chr19:27492836-27497769	6.25	D19Bwg1357e	kidney	8	1.219167
chr8:37382201-37390379	6.25	Sqcz	brain	8	2.587143
chr1:59345146-59350179	5.75	Cdk15	spleen	8	1.74
chr6:112099994-112100334	4.479166667	Lmcd1	brain	8	3.38361

Supplementary table 4. Metastable IAPs

Coordinates	Strand	Computational range	Nearest gene	Location relatively to gene	IAP structure	Region of variation	Type	Overlap with de novo transcripts
chr3:60489872-60495084	-	64.4375	Mbnl1	intragenic	5'LTR-int-3'LTR	5' and 3' LTRs	IAPLTR1_Mm	Mbnl1
chr13:4942652-4943122	+	63.69791667	Gm5444	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr15:39173040-39178257	+	63.02083333	Rims2	intergenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	not annotated
chr2:123754520-123754992	-	61.1875	Sema6d (4930517E11Rik)	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr6:31799369-31804637	-	60	Gm13849	intragenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	not annotated
chr5:38598377-38598869	-	58.33333333	Wdr1	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr5:121330118-121330536	-	57.58333333	Gm15800	intragenic	solo LTR	LTR	IAPLTR2a	Gm15800/not annotated
chr4:141207606-141210591	-	54.4760101	Rsg1	intergenic	5'LTR-int-3'LTR	5' LTR	IAPLTR2_Mm	not annotated
chrX:97141741-97142216	+	53.45833333	Pgr15l	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr16:36765054-36770352	-	53.27380952	Slc15a2	intragenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	Slc15a2/not annotated
chr11:116389404-116394709	+	53.27083333	Rnf157	intragenic	5'LTR-int-3'LTR	5' and 3' LTRs	IAPLTR1_Mm	Rnf157/not annotated
chr10:18139120-18139597	+	52.08333333	Ect2l	intragenic	solo LTR	LTR	IAPLTR2_Mm	not annotated
chr17:18967877-18969281	-	51.95436508	Vmn2r97	intergenic	5'LTR-int-3'LTR	5' LTR	IAPLTR4	AK134158/not annotated

chr2:155538450-155538811	+	51.66666667	Acss2	intragenic	solo LTR	LTR	IAPLTR1_Mm	-
chr16:11013422-11013790	-	49.79166667	Gm4262	3'UTR	solo LTR	LTR	IAPLTR1_Mm	-
chr13:100588951-100594225	-	49.10416667	Marveld2	intergenic	5'LTR-int-3'LTR	5' and 3' LTRs	IAPLTR1_Mm	not annotated
chr2:84505209-84510421	+	48.44998751	Tfpi(Gm13710)	intragenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	not annotated
chr19:9828847-9837164	-	47.91666667	Incenp	intergenic	5'LTR-int-3'LTR	3' LTR	IAPEY4_LTR	-
chr4:89875827-89876305	+	47.39583333	Dmrta1	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr7:30385300-30392433	+	47.08333333	Lrfn3	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	not annotated
chr14:41053401-41061665	+	47.0625	Dydc2	intragenic	5'LTR-int-3'LTR	5' LTR	IAPEY4_LTR	-
chr11:113173016-113173489	+	46.61011905	2610035D17Rik	intragenic	solo LTR	LTR	IAPLTR2_Mm	2610035D17Rik
chr5:64030833-64038297	+	45.8125	Pgm1	intergenic	5'LTR-int-3'LTR	5' and 3' LTRs	IAPLTR2_Mm	not annotated
chr6:16359772-16360249	-	44.4375	Tfec	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr10:99599558-99600037	+	44.38392857	Gm20110	intergenic	solo LTR	LTR	IAPLTR2_Mm	not annotated
chr1:87460130-87467259	+	43.75	3110079O15Rik	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	-
chr12:20845323-20852901	+	43.75	1700030C10Rik	intergenic	5'LTR-int-3'LTR	5' LTR	RLTR46A2	-
chr16:32195537-32200340	+	42.5	Bex6	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	not annotated
chr3:106003601-106003967	+	41.97916667	Pifo	intragenic	solo LTR	LTR	IAPLTR1_Mm	-
chr8:40609942-40610339	+	41.02083333	Mtmr7	intragenic	solo LTR	LTR	IAPLTR2a	-
chr4:138327385-138327850	+	40.83333333	Pink1	intergenic	solo LTR	LTR	IAPLTR2_Mm	-

chr4:71291215-71293236	+	40.77083333	Tle1	intergenic	5'LTR-int	5' LTR	IAPLTR4	not annotated
chr5:86318272-86318748	-	40.24404762	Tmprss11d	intragenic	solo LTR	LTR	IAPLTR2_Mm	-
chr11:106028245-106033560	-	39.54166667	Kcnh6	intragenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	not annotated
chr4:127008519-127015691	+	39.44642857	Sfpq	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	not annotated
chr2:118554765-118558375	+	38.90029762	Bmf	intergenic	5'LTR-int-3'LTR	5' LTR	IAPLTR2_Mm	Bmf/not annotated
chr15:77895206-77902357	-	38.39583333	Txn2	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	-
chr13:100447490-100447794	+	38.32589286	Naip1	intragenic	solo LTR	3' LTR	IAPEY3C_LTR	-
chr4:3192211-3200530	-	37.7827381	Vmn1r3	intergenic	5'LTR-int-3'LTR	5' and 3' LTRs	IAPEY4_LTR	-
chr12:22261651-22269953	-	37.5	Gm6838	intergenic	5'LTR-int-3'LTR	5' LTR	IAPEY4_LTR	-
chr8:72228228-72232950	-	37.5	Fam32a	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	not annotated
chr6:136801751-136802435	-	37.36111111	Hist4h4	5UTR	solo LTR	LTR	IAPLTR2a2_Mm	-
chr13:119661605-119669925	+	37.2485119	Ccl28	intergenic	5'LTR-int-3'LTR	5' LTR	IAPEY4_LTR	-
chr12:56525184-56525535	-	37.125	E030019B13Rik	intragenic	solo LTR	LTR	IAPLTR1_Mm	-
chr4:115281068-115281532	-	37	Cyp4a12a	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr9:103108601-103109112	+	36.875	Rab6b	intergenic	solo LTR	LTR	IAPLTR2_Mm	not annotated
chr1:166940021-166940493	-	36.08333333	Fam78b(Gm16701)	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chrX:74872946-74880330	-	35.88541667	Gm5936	intergenic	5'LTR-int	5' LTR	IAPEY4_LTR	-

chr1:8410687-8411149	+	35.86309524	Sntg1	intragenic	solo LTR	LTR	IAPLTR2_Mm	-
chr7:39655241-39655724	+	35.41964286	Zfp619	intergenic	solo LTR	LTR	IAPLTR2_Mm	not annotated
chr6:41563000-41563471	-	35.11904762	Trbv31	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr16:38714371-38721475	-	34.93303571	Arhgap31	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1a_Mm	not annotated
chr4:17826925-17827396	+	34.39583333	Mmp16	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr14:86398117-86398596	+	34.375	Diap3(4930529K09Rik)	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr13:94024731-94025225	+	34.21527778	Lhfpl2	intergenic	solo LTR	LTR	IAPLTR2_Mm	not annotated
chr7:4455441-4460702	-	33.875	Eps8l1	intragenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	Eps8l1/not annotated
chr18:73959218-73959588	+	33.30109127	Mapk4	intragenic	solo LTR	3' LTR	IAPEY2_LTR	-
chr3:38672092-38672522	-	32.87698413	Ankrd50	intergenic	solo LTR	LTR	IAPLTR2a2_Mm	-
chr6:28924174-28929353	-	32.77083333	Snd1	intragenic	5'LTR-int-3'LTR	5' and 3' LTRs	IAPLTR1_Mm	-
chr1:146805792-146806250	-	32.29166667	Fam5c	intragenic	solo LTR	LTR	IAPLTR2_Mm	-
chrX:165666796-165667264	-	31.98958333	Glra2	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr2:161511654-161512123	-	31.97172619	Ptprt	intergenic	solo LTR	LTR	IAPLTR2_Mm	not annotated
chr1:118012640-118019825	-	31.80059524	Tsn	intergenic	int-3'LTR	3' LTR	IAPEY4_LTR	-
chr7:7794177-7796900	+	31.25	Vmn2r35	intragenic	5'LTR-int	5' LTR	IAPLTR1a_Mm	-
chr7:27933745-27942065	-	31.25	1700049G17Rik	intergenic	5'LTR-int-3'LTR	3' LTR	IAPEY4_LTR	-
chr7:101562084-101562373	-	31.11053009	Art2a-ps	intergenic	solo LTR	LTR	IAPEY5_LTR	-

chr6:37464145-37464607	+	30.5625	Creb3l2	intergenic	solo LTR	LTR	IAPLTR2_Mm	not annotated
chr10:59381798-59387060	+	30.20833333	P4ha1	intergenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	not annotated
chr9:123742716-123749810	-	30.16666667	Ccr9	intergenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	-
chr8:72024782-72033082	-	30.10416667	Cyp4f18	intergenic	5'LTR-int-3'LTR	3' LTR	IAPEY4_LTR	-
chr2:93939727-93940154	+	30.08012821	Gm1335	intergenic	solo LTR	LTR	IAPLTR2a2_Mm	-
chr13:25007399-25015719	+	29.95833333	Mrs2	intragenic	5'LTR-int-3'LTR	3' LTR	IAPEY4_LTR	-
chr12:23414019-23422321	-	29.94791667	9030624G23Rik	intergenic	5'LTR-int-3'LTR	5' LTR	IAPEY4_LTR	-
chr10:23726340-23726813	+	29.89583333	Rps12	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr1:110612994-110613452	+	29.66666667	Cdh19	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr6:112099994-112100334	-	29.375	1700054K19Rik	intergenic	solo LTR	3' LTR	IAPEY2_LTR	-
chr17:50074462-50078819	-	29.16666667	Rftn1	intragenic	5'LTR-int	5' LTR	IAPLTR1_Mm	not annotated
chr10:78020225-78025621	+	29.16666667	Aire	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	-
chr12:115737421-115738757	+	28.95833333	abParts	intragenic	int-3'LTR	5' end	internal	-
chr17:20091770-20097820	+	28.94345238	Vmn2r-ps126	intragenic	int-3'LTR	3' LTR	IAPEY4_LTR	-
chr9:124081174-124082990	+	28.47948232	Ccr2	intergenic	int	5' end	internal	-
chr5:114699475-114699927	-	28.13762626	Tchp	intergenic	solo LTR	LTR	IAPLTR2a2_Mm	-
chr16:32148076-32153339	+	28.125	Nrros	intragenic	5'LTR-int-3'LTR	5' LTR	IAPLTR1_Mm	not annotated
chr7:47570135-47574698	-	28.125	Mrgpra3	intergenic	int-3'LTR	5' end	internal	-

chrX:154136452-154143405	+	27.83333333	Gm5645	intergenic	5'LTR-int-3'LTR	3' LTR	IAPEY5_LTR	-
chr14:19005873-19013019	-	27.4375	Ube2e2	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	-
chr12:18390683-18391250	+	27.29166667	5730507C01Rik	intergenic	solo LTR	LTR	IAPLTR2_Mm	-
chr12:23597689-23606009	-	27.14583333	9030624G23Rik	intergenic	5'LTR-int-3'LTR	5' and 3' LTRs	IAPEY4_LTR	-
chr5:33953277-33954255	-	27.125	Nelfa	intergenic	5'LTR-int-3'LTR	5' LTR	RLTR46	-
chrX:156318492-156323301	-	27.08333333	4930503H13Rik	intergenic	5'LTR-int-3'LTR	5' LTR	IAPLTR2_Mm	-
chr16:3625351-3632950	-	27.05952381	Olfr161	intergenic	5'LTR-int-3'LTR	5' LTR	RLTR46B	-
chr12:47417737-47418041	-	27.00595238	Nova1	intergenic	solo LTR	LTR	IAPLTR4	-
chr8:5217767-5218989	+	26.875	Slc10a2	intergenic	5'LTR-int	3' end	internal	-
chr15:95753315-95753705	+	26.70833333	Ano6	intergenic	solo LTR	LTR	IAPLTR2a2_Mm	-
chr7:9859857-9862628	+	26.5625	AK083498	intergenic	5'LTR-int	3' end	internal	-
chr14:43804373-43812688	-	26.375	Ear1	intergenic	5'LTR-int-3'LTR	3' LTR	IAPEY4_LTR	-
chr2:177236696-177241367	+	26.33333333	AK078446	intergenic	int	3' end	internal	-
chr1:100223822-100232140	-	26.0625	Cntnap5b	intragenic	5'LTR-int-3'LTR	3' LTR	IAPEY4_LTR	-
chr12:85003517-85003995	-	25.95833333	Ylpm1	intragenic	solo LTR	LTR	IAPLTR2_Mm	-
chr16:90341543-90349872	-	25.9375	Hunk	intergenic	5'LTR-int-3'LTR	3' LTR	IAPEY4_LTR	-
chr12:20526585-20527813	-	25.85416667	BC024416	intragenic	int-3'LTR	5' end	internal	-
chr7:48124123-48132441	-	25.67261905	Mrgprb5	intergenic	5'LTR-int-3'LTR	3' LTR	IAPEY4_LTR	-

chr1:23301116-23306356	+	25.52083333	Mir30c-2	intergenic	5'LTR-int-3'LTR	3' LTR	IAPLTR1_Mm	-
chrX:78041877-78044983	-	25.16666667	Obp1a	intergenic	int-3'LTR	5' end	internal	-
chr1:101897873-101903028	-	25.05492424	Cntnap5b	intergenic	5'LTR-int-3'LTR	5' LTR	IAPLTR4	not annotated

Supplementary table 5. IAP-driven transcription initiation and termination

Chr	Start	End	Methylation	Structure	LTR	Splicing to gene	Total number initiated transcripts	Total number terminated transcripts	Comments
1	127891984	127897254	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	11	11	
1	127558257	127565484	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	6	7	transcripts only in B
1	58930750	58935209	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	11	11	
1	159969770	159976851	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	12	6	transcripts in B
1	127875080	127882174	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	12	11	
1	58811690	58818858	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	10	9	
2	166812198	166817469	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	10	9	
2	3483513	3488773	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	5	NA	
2	154354262	154359592	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Cdk5rap1	10	10	initiation of Cdk5rap in 2B and 3T
2	26333220	26338473	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	9	13	
2	112548954	112553793	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	Aven	11	8	initiation of Aven in 1T
2	5684025	5691128	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	12	12	
3	96489247	96494539	low 3'LTR	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	9	12	
3	115909127	115914440	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Dph5	20	12	initiation of Dph5 in 4B and 2T
3	136173778	136180883	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	Bank1	6	7	transcripts only B, initiation of Bank1 1B and termination of Bank1 1B
4	98454409	98454761	high	LTR	IAPLTR1_Mm	NA	4	9	

4	46511086	46517639	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Trim14/ Nans	13	25	termination of Trim14 6B and 2T and Nans 5B and 2T; initiation Trim14 in T
4	147833501	147840658	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	11	11	
4	135455803	135462877	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Nipal3	12	19	termination of Nipal3 in 3B and 4T
4	135184494	135188779	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	10	1	
4	108284073	108291153	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	12	11	
5	45815198	45823190	high	5'LTR-int-3'LTR	IAPEY3_LTR	NA	10	8	
5	143004697	143009472	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	6	8	transcripts in B
5	105724153	105729395	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	9	5	
5	92476105	92482098	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	11	11	
5	23569988	23573397	high	5'LTR-int-3'LTR	IAPLTR2_Mm	Srpk2	11	9	initiation of Srpk2 in 1T
5	64889711	64896675	high	5'LTR-int-3'LTR	IAPLTR2_Mm	NA	6	1	transcripts only in T
5	110194893	110199850	high	5'LTR-int-3'LTR	IAPLTR2a	NA	11	7	
5	136155678	136160205	high	5'LTR-int-3'LTR	IAPLTR2a2_Mm	ALkbh4	5	20	termination of Alkbh4 in 2T and NA in 1B
5	38471467	38475987	high	5'LTR-int-3'LTR	IAPLTR2a2_Mm	NA	10	9	
5	30223044	30230157	high	5'LTR-int-3'LTR	IAPLTR2b	NA	12	11	
6	3333740	3341286	DMR -low T	int	-	Gm20559	NA	55	termination of Gm20559 in all
6	95475552	95479748	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	4	5	all 5 termination in B
6	71955426	71960760	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	12	11	
6	120812261	120817386	high	5'LTR-int-3'LTR	IAPLTR2_Mm	NA	12	11	

7	106561081	106563754	high	5'LTR-int	IAPEY5_LTR	NA	7	16	termination Gm1966 in 4B and 3T
7	118268458	118273558	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	6	6	transcripts only in T
7	79337065	79342596	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Abhd2	12	11	termination of Abhd2 in 1T
7	114404490	114411619	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Pde3b	17	13	initiation of Pde3b 4B and 6T
7	104266842	104273963	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Trim34a	7	13	termination of Trim34a in 4 B and 1T
7	89273141	89276993	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	4	8	
7	140014758	140021337	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	11	11	
7	55946767	55953945	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	12	9	
7	122051610	122052626	high	LTR	IAPLTR2_Mm	Ears2	NA	8	termination of Ears2 in 1B and 2T
8	94616511	94621872	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	10	11	
8	22551167	22558293	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Slc20a2	8	10	termination of Slc20a2 in 1T
9	123786635	123788707	high	int	-	NA	6	8	
9	121322377	121328483	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	9	7	transcripts in B
9	3461293	3466472	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Cwf19l2	10	18	termination Cwf19l2 in 2B and 4T
9	37331368	37338476	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Ccdc15	10	10	initiation Ccdc15 in 1B; termination Ccdc15 in 1T
9	83942547	83949868	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Bckdhb	22	2	initiation of Bckdhb in 4B and 6T
9	121406408	121413506	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	10	8	termination transcripts mostly in B
10	94693156	94695213	high	5'LTR-int-3'LTR	IAP1-MM_LTR	NA	9	6	
10	99098192	99103428	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Poc1b	17	3	initiation Poc1b in both cell types

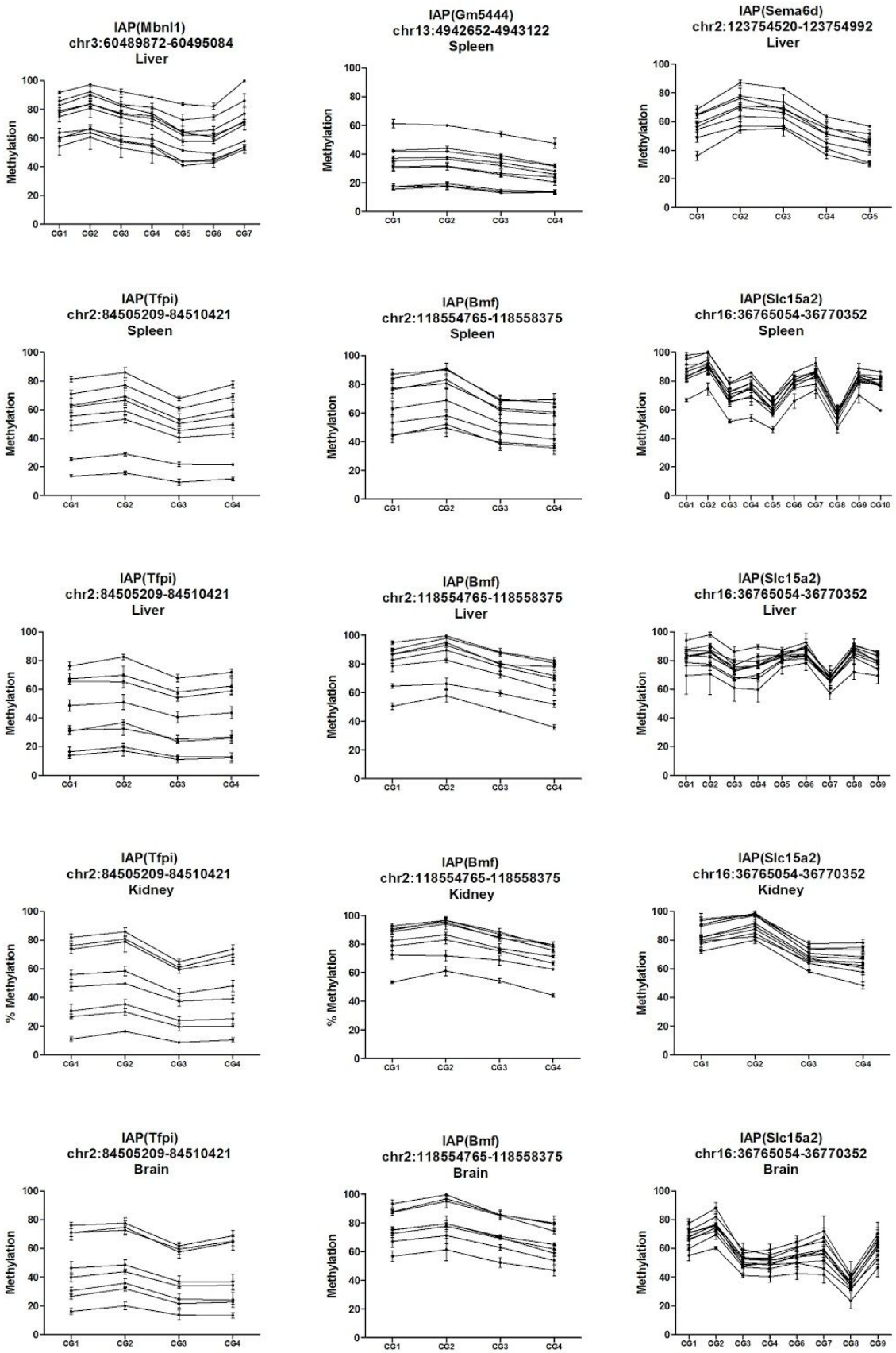
10	116606259	116613376	DMR - low T	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	2	5	transcripts only in T
10	94362587	94369250	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	6	6	transcripts only in B
10	122832872	122839915	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	4	3	transcripts only in T
10	41840531	41847635	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	6	8	
10	39760777	39764747	high	5'LTR-int-3'LTR	IAPLTR2_Mm	NA	11	8	
10	111000000	111002909	high	5'LTR-int-3'LTR	IAPLTR2_Mm	NA	9	9	
10	79412906	79418028	DMR - low T	5'LTR-int-3'LTR	IAPLTR4	NA	NA	11	transcripts only in T
11	26555445	26560784	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Vrk2	19	10	initiation 3B and 4T and termination 3B and 2T Vrk2
11	54576586	54583698	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	12	12	
11	6141511	6148630	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	12	12	
12	73310992	73315860	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Slc38a6	11	24	termination of Slc38a6/ initiation of Slc38a6 in 1T
12	81583094	81587495	high	5'LTR-int	IAPLTR1_Mm	NA	11	12	
12	41363543	41370627	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	6	7	transcripts in B
12	80417262	80424412	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	9	10	
12	91553506	91553915	high	LTR	IAPLTR2a2_Mm	NA	5	1	
13	111487909	111488349	*3CpG low	int	-	Gpbp1	NA	9	
13	63218000	63223147	DMR - low T	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	6	6	transcripts only in T
13	111694234	111700822	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	10	5	
13	56712880	56719977	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	12	11	
13	43277542	43284656	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	6	9	

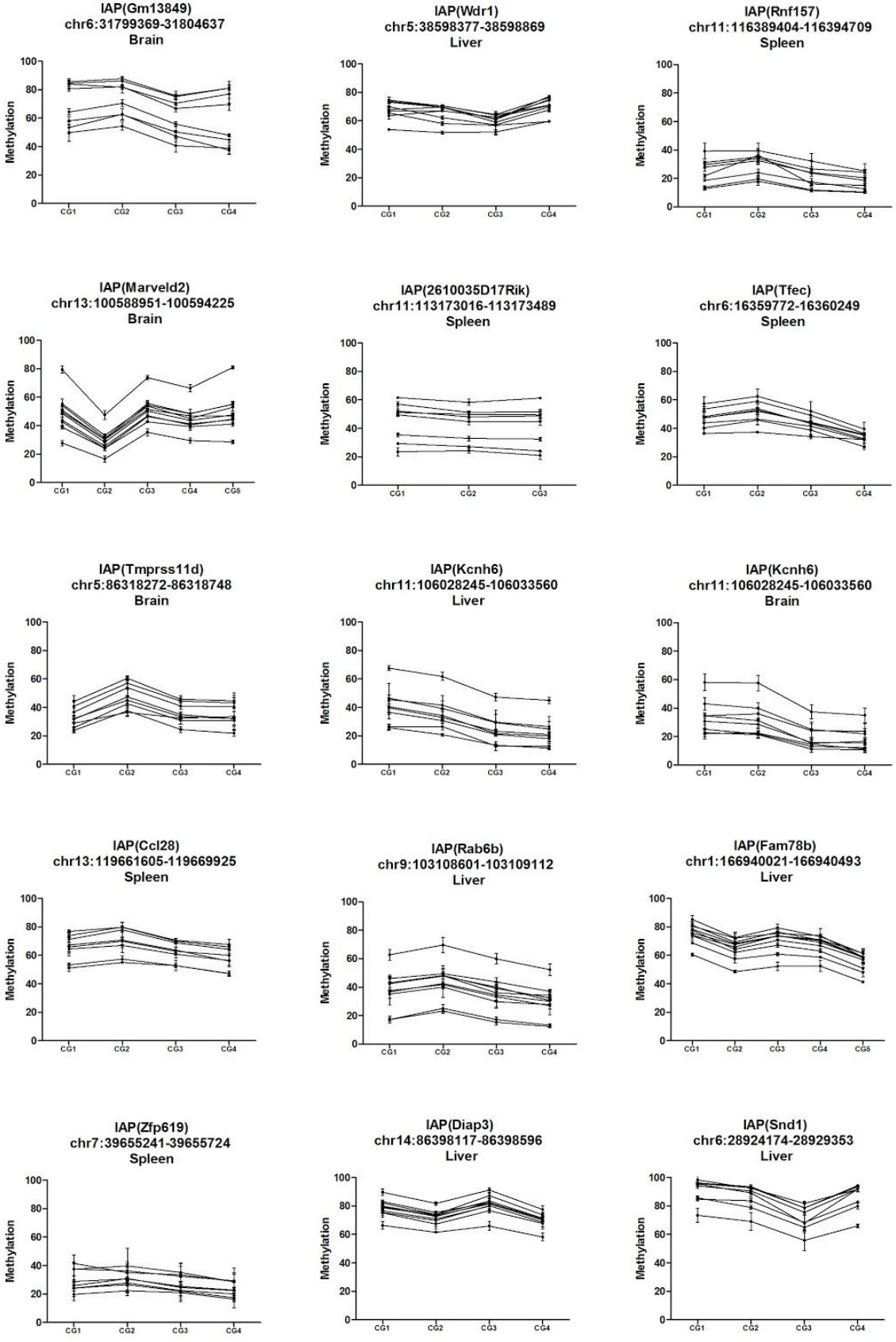
13	34166773	34174004	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	11	11	
13	43386793	43393876	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	9	5	
14	50880210	50884861	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	11	2	
15	100542521	100545705	high	5'LTR-int-3'LTR	IAPEY2_LTR	NA	7	10	
15	86312263	86317593	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	10	9	
16	98035575	98039240	high	int-3'LTR	IAPLTR2b	B230307C23Rik	12	4	termination of B230307C23Rik in 1B and 1T
16	11680013	11685308	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	8	5	transcripts only in B
16	77116957	77121840	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	12	6	
16	70542445	70547794	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	8	1	
16	4743410	4750484	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	10	5	
16	22033380	22035947	high	5'LTR-int-3'LTR	IAPLTR2a2_Mm	Senp2	1	29	termination Senp2 in all
17	34544222	34546535	high	int-3'LTR	IAP1-MM_LTR	NA	3	NA	only in male T
17	34543521	34544066	high	5'LTR-int	IAP1-MM_LTR	NA	1	6	transcripts only in T
17	51864437	51871261	high	5'LTR-int-3'LTR	IAPLTR1_Mm	Satb1	13	7	initiation Satb1 in 1B and 2T
17	55676216	55680378	high	5'LTR-int	IAPLTR1a_Mm	NA	10	7	
17	63125984	63133088	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	11	11	
17	36007143	36014240	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	H2-T24	10	10	termination of H2-T24 in 1B
17	31552408	31554076	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	3	6	transcripts only in T
17	33981268	33981698	low	LTR	IAPLTR2_Mm	AA388235	NA	10	
17	33738033	33740862	high	5'LTR-int-3'LTR	IAPLTR2_Mm	NA	8	1	
17	50660596	50665642	high	5'LTR-int-3'LTR	IAPLTR2a	NA	8	9	
17	18960873	18967329	high	5'LTR-int-3'LTR	IAPLTR2b	AK134158	5	8	termination of AK134158 in 4T

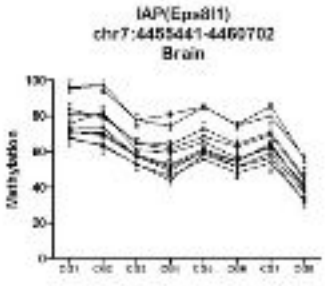
17	18558212	18566570	DMR - low T	5'LTR-int-3'LTR	IAPLTR4	NA	1	17	
18	14571242	14571342	no CpGs	LTR	IAPLTR1_Mm	NA	7	NA	
18	60528367	60533722	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	12	9	
18	5109642	5113531	high	5'LTR-int-3'LTR	IAPLTR2_Mm	NA	8	6	
19	8030707	8034135	high	int-3'LTR	IAPLTR4	NA	1	25	termination of NA in all
19	52985550	52990946	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	12	6	
19	37617435	37624485	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	12	11	
19	5406362	5406822	low	LTR	IAPLTR2_Mm	4930481A15Rik	9	23	initiation of 4930481A15Rik
19	8513116	8519212	DMR -low T	5'LTR-int	IAPLTR4	NA	NA	11	transcripts only in T
19	8313575	8319222	high	5'LTR-int-3'LTR	IAPLTR4	NA	NA	9	
19	8019332	8024504	low	5'LTR-int	IAPLTR4	NA	21	15	
X	94827561	94828105	DMR-low T	5'LTR-int	IAP1-MM_LTR	NA	10	6	
X	162978632	162985757	high	5'LTR-int-3'LTR	IAPLTR1_Mm	NA	12	11	
X	35915021	35919534	high	5'LTR-int-3'LTR	IAPLTR1a_Mm	NA	8	11	

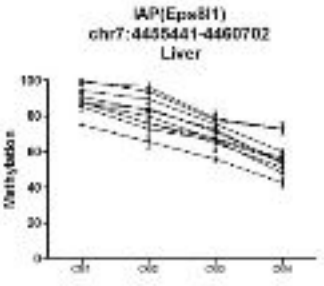
Supplementary figure 1. Validation of IAP inter-individual methylation variation

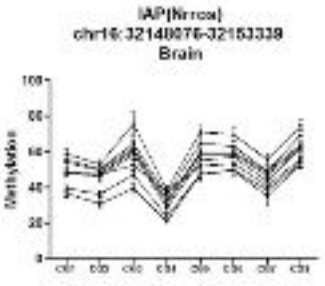
IAPs with computational score >25%

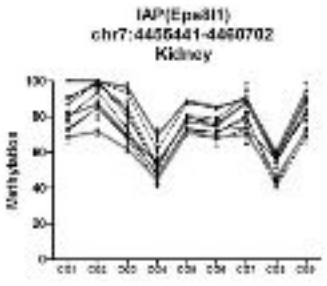


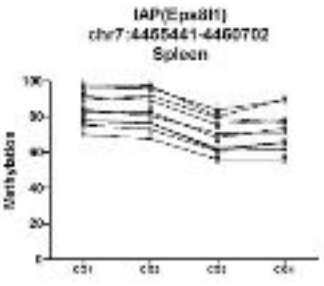

IAP(Gm13849)
chr6:31799369-31804637
Brain
IAP(Wdr1)
chr5:38598377-38598869
Liver
IAP(Rnf157)
chr11:116389404-116394709
Spleen
IAP(Marveld2)
chr13:100588951-100594225
Brain
IAP(2610035D17Rik)
chr11:113173016-113173489
Spleen
IAP(Tfec)
chr6:16359772-16360249
Spleen
IAP(Tmprss11d)
chr5:86318272-86318748
Brain
IAP(Kcnh6)
chr11:106028245-106033560
Liver
IAP(Kcnh6)
chr11:106028245-106033560
Brain
IAP(Ccl28)
chr13:119661605-119669925
Spleen
IAP(Rab6b)
chr9:103108601-103109112
Liver
IAP(Fam78b)
chr1:166940021-166940493
Liver
IAP(Zfp619)
chr7:39655241-39655724
Spleen
IAP(Diap3)
chr14:86398117-86398596
Liver
IAP(Snd1)
chr6:28924174-28929353
Liver
Methylation
CG1
CG2
CG3
CG4
CG5

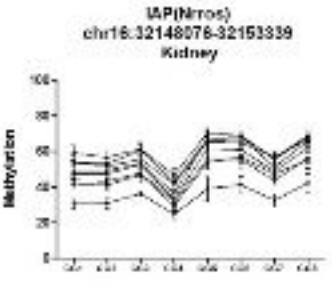

IAP(Epx8I1)
chr7:4455441-4460702
Brain

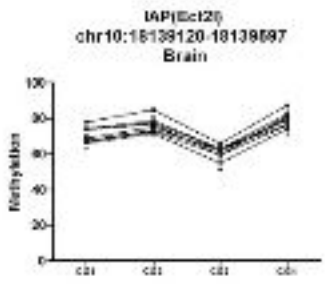

IAP(Epx8I1)
chr7:4455441-4460702
Liver

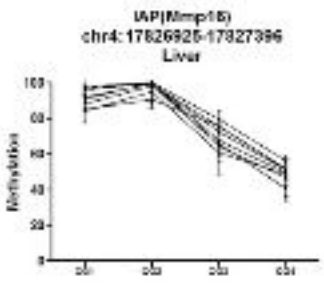

IAP(Nrros)
chr16:32148076-32153339
Brain

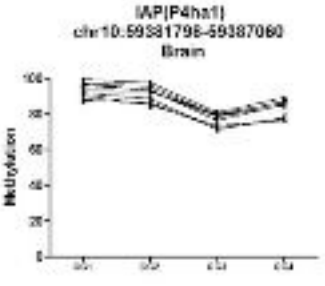

IAP(Epx8I1)
chr7:4455441-4460702
Kidney

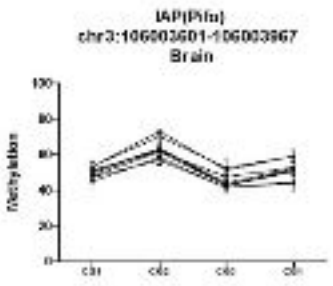

IAP(Epx8I1)
chr7:4455441-4460702
Spleen

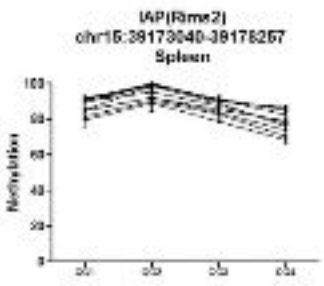

IAP(Nrros)
chr16:32148076-32153339
Kidney

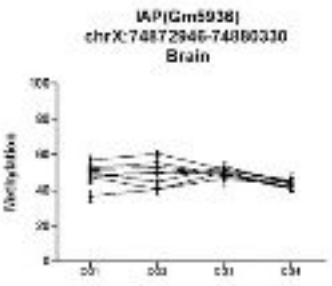

IAP(Ecl2l)
chr10:18139120-18139597
Brain

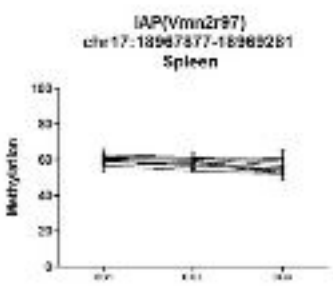

IAP(Mmp16)
chr4:17826925-17827396
Liver

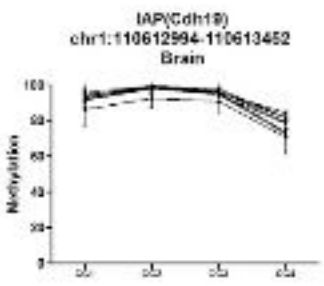

IAP(P4ha1)
chr10:59381798-59387060
Brain

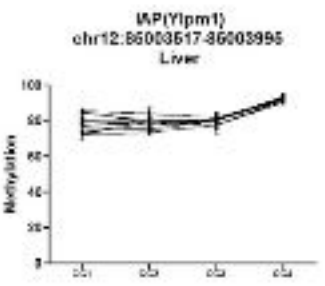

IAP(Pfn)
chr3:106003601-106003967
Brain

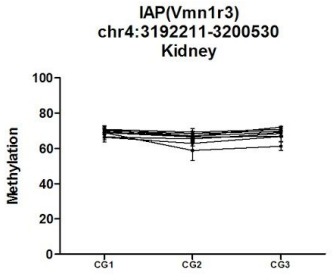

IAP(Rmx2)
chr15:39173040-39178257
Spleen

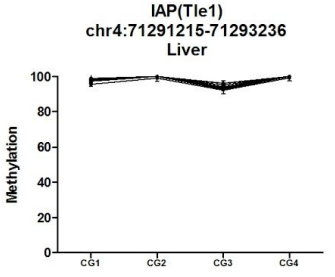

IAP(Gm5938)
chrX:74872946-74880330
Brain

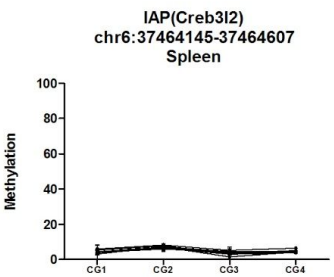

IAP(Vmn2r97)
chr17:18967877-18969281
Spleen

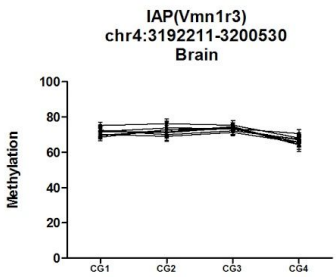

IAP(Cdh19)
chr1:110612994-110613452
Brain

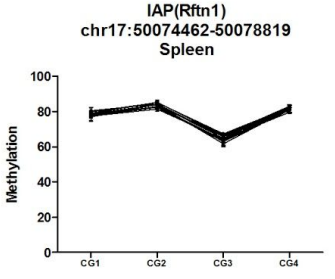

IAP(Ylpm1)
chr12:86003517-86003995
Liver



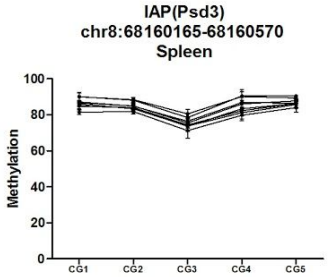


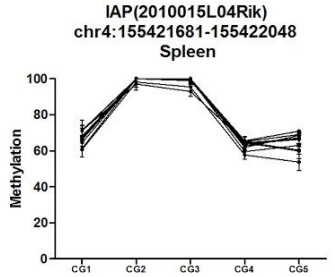


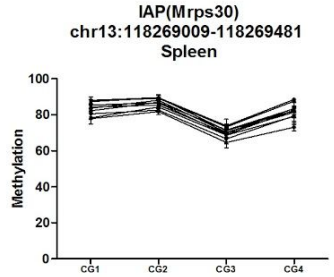


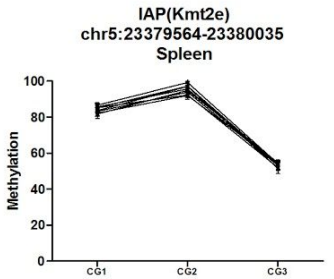


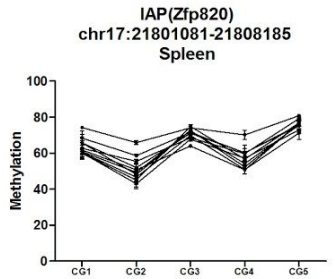
IAPs with computational score 20-25%

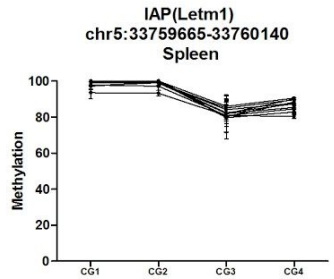


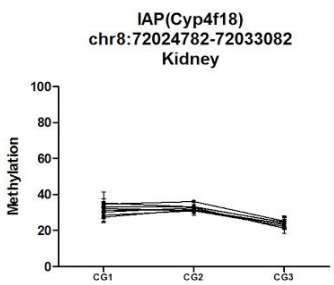


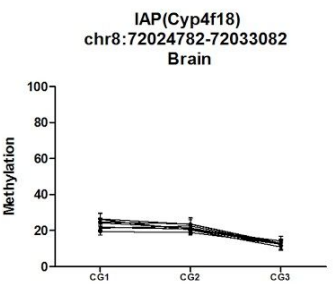


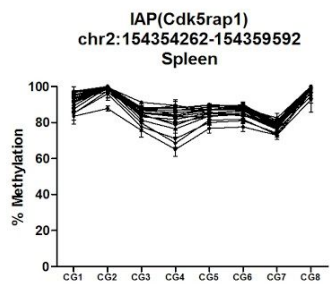


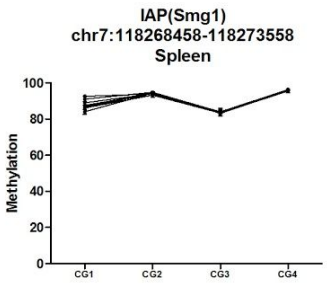


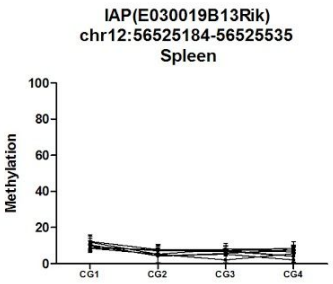




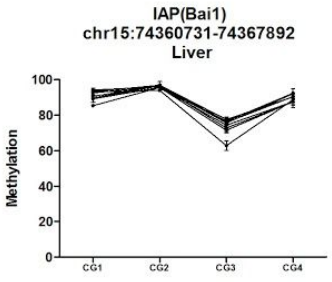


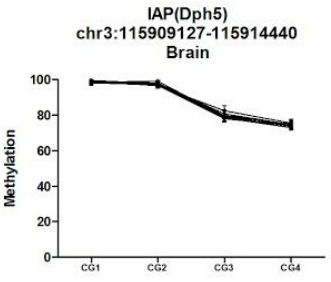


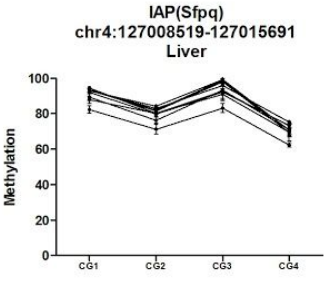


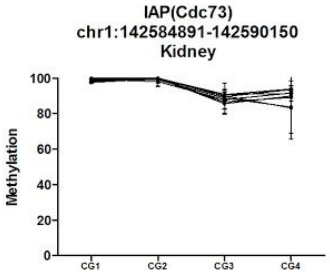


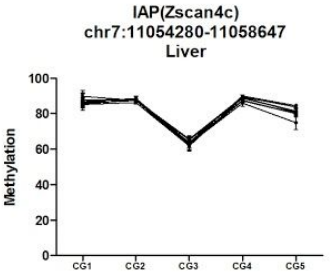
IAPs with computational score <20%

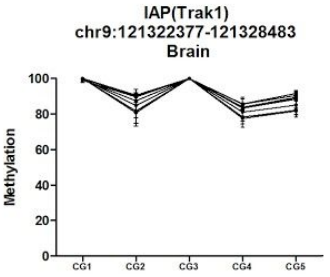


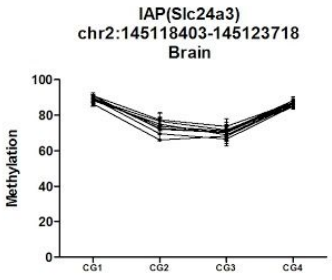


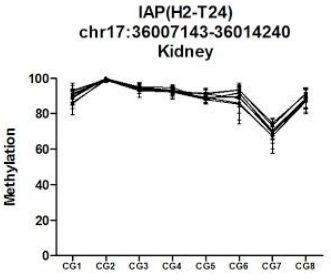


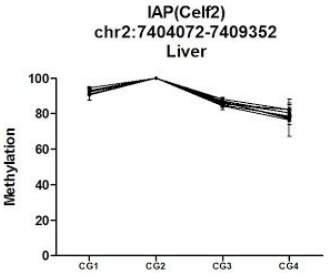


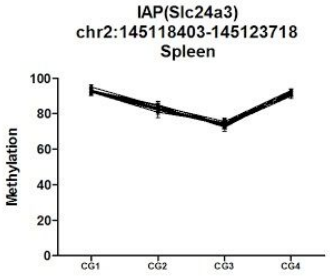


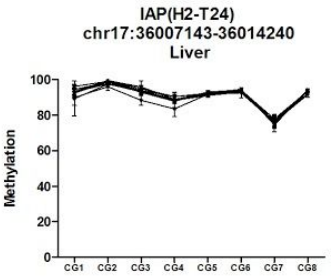


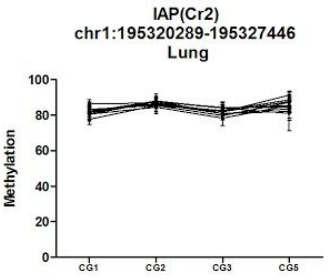


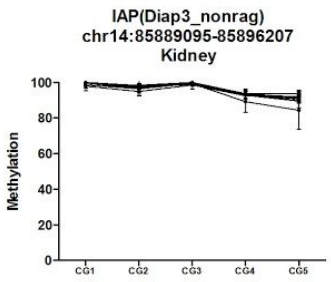


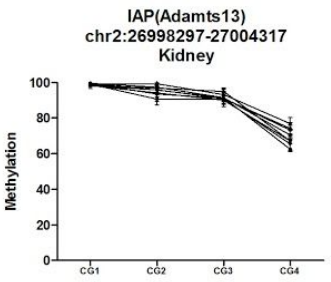


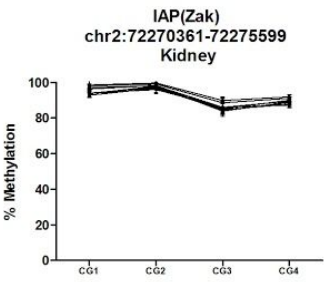


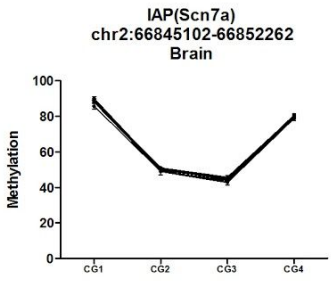


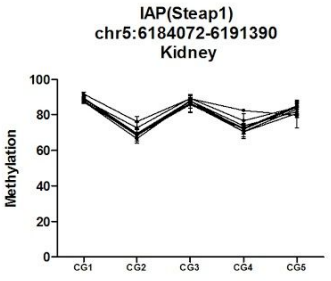


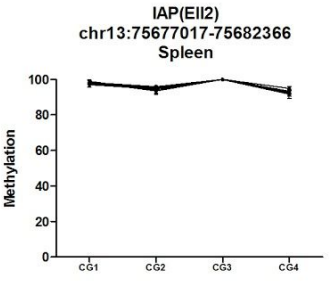

IAP(Diap3_nonrag)
chr14:85889095-85896207
Kidney
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4
CG5

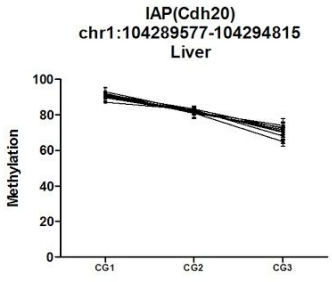

IAP(Adamts13)
chr2:26998297-27004317
Kidney
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4

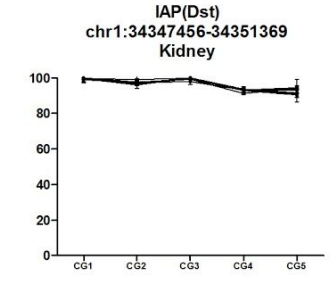

IAP(Zak)
chr2:72270361-72275599
Kidney
% Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4

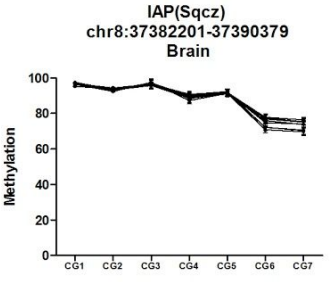

IAP(Scn7a)
chr2:66845102-66852262
Brain
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4

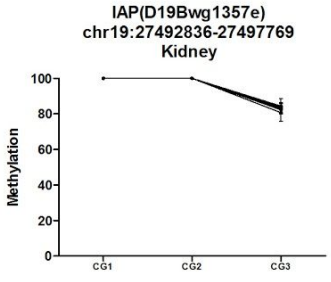

IAP(Steap1)
chr5:6184072-6191390
Kidney
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4
CG5

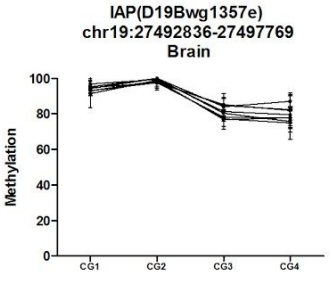

IAP(Ell2)
chr13:75677017-75682366
Spleen
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4

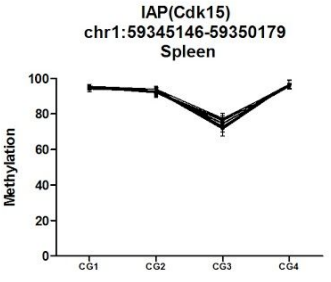

IAP(Cdh20)
chr1:104289577-104294815
Liver
Methylation
100
80
60
40
20
0
CG1
CG2
CG3

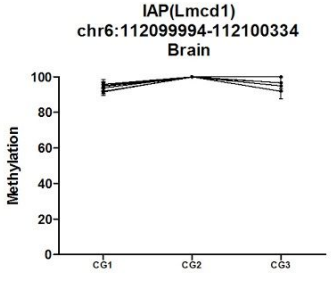

IAP(Dst)
chr1:34347456-34351369
Kidney
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4
CG5


IAP(Sqcz)
chr8:37382201-37390379
Brain
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4
CG5
CG6
CG7


IAP(D19Bwg1357e)
chr19:27492836-27497769
Kidney
Methylation
100
80
60
40
20
0
CG1
CG2
CG3


IAP(D19Bwg1357e)
chr19:27492836-27497769
Brain
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4


IAP(Cdk15)
chr1:59345146-59350179
Spleen
Methylation
100
80
60
40
20
0
CG1
CG2
CG3
CG4


IAP(Lmcd1)
chr6:112099994-112100334
Brain
Methylation
100
80
60
40
20
0
CG1
CG2
CG3