

RESEARCH

Open Access

Genomic properties of variably methylated retrotransposons in mouse



Jessica L. Elmer[†], Amir D. Hay[†], Noah J. Kessler[†], Tessa M. Bertozzi, Eve A. C. Ainscough and Anne C. Ferguson-Smith^{*}

Abstract

Background: Transposable elements (TEs) are enriched in cytosine methylation, preventing their mobility within the genome. We previously identified a genome-wide repertoire of candidate intracisternal A particle (IAP) TEs in mice that exhibit inter-individual variability in this methylation (VM-IAPs) with implications for genome function.

Results: Here we validate these metastable epialleles and discover a novel class that exhibit tissue specificity (tsVM-IAPs) in addition to those with uniform methylation in all tissues (constitutive- or cVM-IAPs); both types have the potential to regulate genes in *cis*. Screening for variable methylation at other TEs shows that this phenomenon is largely limited to IAPs, which are amongst the youngest and most active endogenous retroviruses. We identify sequences enriched within cVM-IAPs, but determine that these are not sufficient to confer epigenetic variability. CTCF is enriched at VM-IAPs with binding inversely correlated with DNA methylation. We uncover dynamic physical interactions between cVM-IAPs with low methylation ranges and other genomic loci, suggesting that VM-IAPs have the potential for long-range regulation.

Conclusion: Our findings indicate that a recently evolved interplay between genetic sequence, CTCF binding, and DNA methylation at young TEs can result in inter-individual variability in transcriptional outcomes with implications for phenotypic variation.

Keywords: Retrotransposon, Endogenous retrovirus, Intracisternal A particle, DNA methylation, Metastable epiallele, CTCF, Chromatin conformation

Background

Transposable elements (TEs) are DNA sequences that account for about 40% of the mouse genome [1]. The vast majority (96%) of TEs are retrotransposons, which mobilise via an RNA intermediate prior to re-integration into the genome [1, 2]. There are three classes of retrotransposons in mammals: long-interspersed nuclear elements (LINEs), short-interspersed nuclear elements (SINEs), and long-terminal repeat elements (LTRs; which include endogenous retroviruses (ERVs)). Whilst ERVs mostly exist as solo LTRs, and the majority of full-length elements have lost coding potential due

to an accumulation of mutations, some mouse ERVs retain the ability to retrotranspose and account for up to 12% of all germline mutations [3]. Due to the risk of insertional mutation and the potential activity of internal regulatory sequences, retrotransposons are commonly targeted for silencing by ncRNAs, repressive histone modifications, and DNA methylation [4].

Of all the types of ERVs in the mouse genome, intracisternal A-particle elements (IAPs) are amongst the most active [5, 6] and evolutionarily young [2, 3, 7, 8]. Unlike 90% of the genome, IAPs have been reported to resist the epigenetic reprogramming that occurs during early embryonic development and remain methylated [9, 10]. It is therefore rare for IAPs to be partially or completely unmethylated. *Agouti viable yellow* (A^{vy}) and

* Correspondence: afsmith@gen.cam.ac.uk

[†]Jessica L. Elmer, Amir D. Hay and Noah J. Kessler contributed equally to this work.

Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Axin Fused (*Axin^{Fu}*) represent two-well studied examples of IAPs that are not fully methylated. These alleles were first identified due to observable differences between littermates in coat colour and tail morphology, respectively [11, 12] and were termed ‘metastable epialleles’ [13–15]. It was determined that these phenotypic differences result from an IAP insertion and that the methylation level of the element is inversely correlated with the expression of the affected gene [16, 17]. These two IAP insertions have inter-individual variation in DNA methylation that gives rise to individual mice exhibiting coat colours ranging from yellow to pseudoagouti (*A^{vy}*), and tail morphology ranging from highly kinked to straight (*Axin^{Fu}*). Furthermore, both models transmit a memory of the parental methylation state to the offspring, providing a paradigm for transgenerational epigenetic inheritance [14, 16].

Previously, our group performed a systematic genome-wide screen in C57BL/6J mice to identify other variably methylated IAPs (VM-IAPs) [18]. Unlike what was observed at the *A^{vy}* and *Axin^{Fu}* loci, the majority of VM-IAPs identified in that screen did not have any detectable effect on the expression of adjacent genes. In offspring, VM-IAPs do not possess a memory of the parental methylation state; instead, the DNA methylation at VM-IAPs is predictably reprogrammed following fertilization and re-established stochastically in the next generation [13]. The functional implications of DNA methylation at these elements and the mechanisms behind the establishment of variable methylation are not fully understood.

In this study, we validate candidate VM-IAPs that show consistent levels of methylation between tested tissues, which we term constitutive variably methylated IAPs (cVM-IAPs). In addition, we identify and characterise IAP elements that only have variable methylation in some, but not all, tested tissues and term these tissue-specific variably methylated IAPs (tsVM-IAPs). We find that cVM-IAPs are enriched for specific sequences that may play a role in the acquisition of variable methylation and test whether they are sufficient to confer variable methylation. We show an inverse correlation between binding levels of the multifunctional transcription factor CTCF and DNA methylation at cVM-IAPs, and identify distinct patterns of chromatin interactions at several of these loci with levels of DNA methylation at these loci correlating with levels of H3K9me3. We expand our screen beyond IAP elements and find that variable methylation is uncommon in other types of retrotransposons. Overall, our findings suggest that variable methylation at IAP elements occurs via complex interactions between IAP sequence, CTCF binding, histone modifications and DNA methylation machinery, and may represent a transient evolutionary state with the

potential to cause inter-individual variability in transcription and phenotype.

Results

Individual VM-IAPs possess constitutive or tissue-specific methylation variability

Our previous screen for VM-IAPs was performed using whole genome bisulphite sequencing (WGBS) datasets from B and T cells of C57BL/6J mice generated as part of the BLUEPRINT consortium [18]). This resulted in the identification of 104 candidate VM-IAP elements. The IAP annotations used in the screen were created by RepeatMasker, a program that identifies repetitive portions of the genome (including TEs). Upon closer inspection we noted that many of the annotated elements were ‘fragmented’, i.e. neither solo LTRs nor fully-structured IAPs with tandem LTRs. Fragmented elements may arise naturally in the genome, through later insertions of other elements, accumulation of polymorphisms over evolutionary time, or through interchromosomal recombination. RepeatMasker often separates the components of an intact IAP into several distinct annotations, resulting in an inflated proportion of fragmented elements in the mouse genome and an inaccurate understanding of IAP boundaries. We therefore developed a method to piece together elements that were artificially fragmented in the annotation. Applying this method to the existing annotation decreased the total count of IAP elements in the mouse genome from 13,065 to 10,678, and reduced the proportion of fragmented elements from 36 to 19% (https://github.com/knowah/vm-retrotransposons/blob/master/data/repeat_annotations/mm10.IAP.mended.tsv).

This improved IAP annotation, along with an updated WGBS screen algorithm (Methods), identified twelve new candidate VM-IAPs, bringing the total to 116. We were able to assay 103 of these IAPs (Supplemental Table S1) using bisulphite pyrosequencing in ear – a distinct tissue from the cell types used in the screen. Of the tested candidates, 51 elements showed variable methylation in ear, passing a threshold of $\geq 10\%$ inter-individual methylation variation. We termed these ‘constitutive VM-IAPs’ (cVM-IAPs) (Fig. 1a, left panel and Supplemental Data).

We next assessed whether the 52 candidates that did not validate in ear samples represent IAPs which are variably methylated in other tissues. Using the same tissues as in the WGBS VM-IAP screen, we sorted B and T cells from C57BL/6J mice (detailed in [18]), and performed bisulphite pyrosequencing for each candidate using these samples. Half (26) of the candidates that did not validate in ear were variably methylated in B cells; we termed these ‘tissue-specific VM-IAPs’ (tsVM-IAPs) (Fig. 1a, middle panel and Supplemental Data). The candidates not

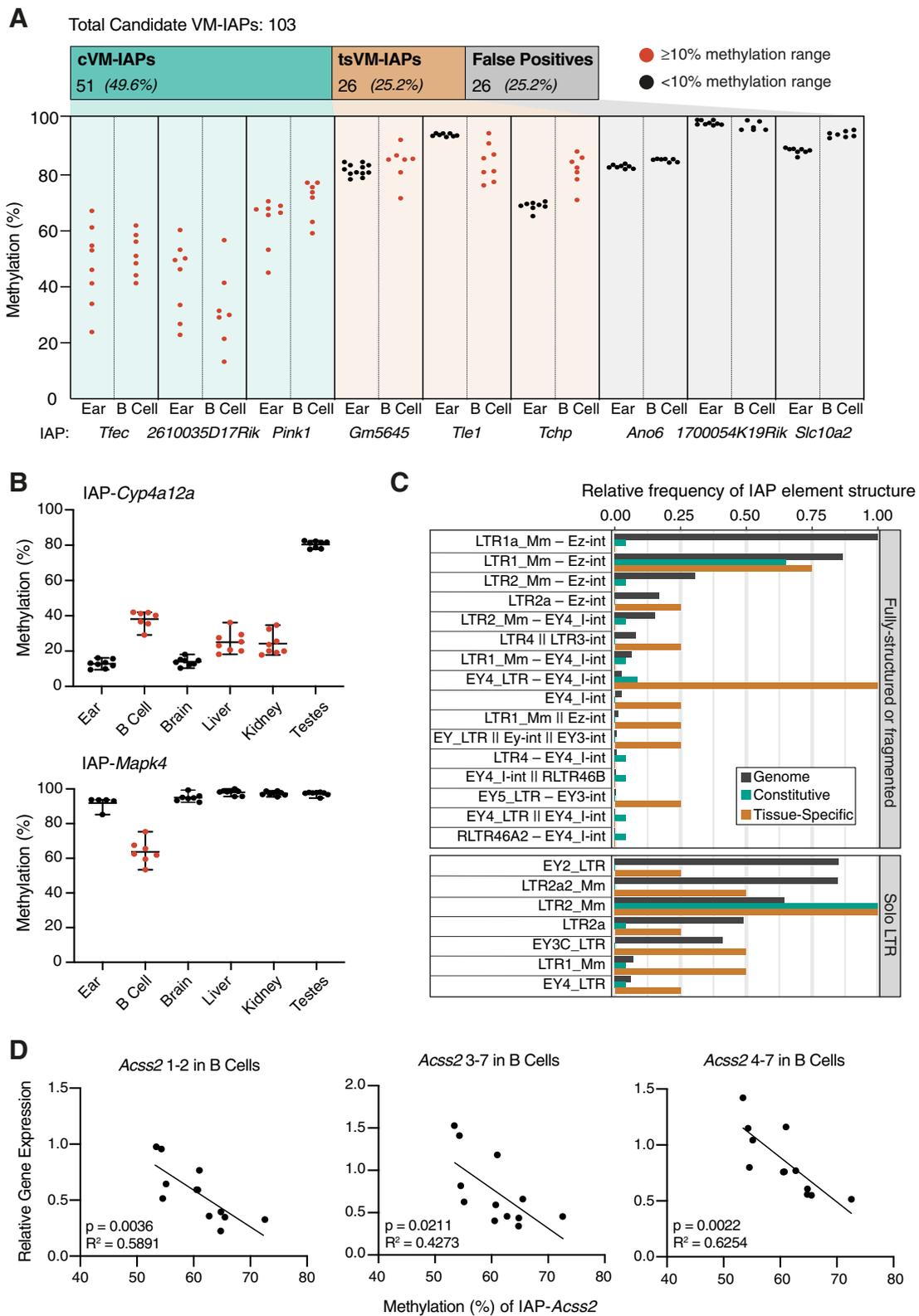


Fig. 1 (See legend on next page.)

(See figure on previous page.)

Fig. 1 Genome-wide screens and site-specific validation of candidate variably methylated IAP elements (VM-IAPs) identify 51 loci with constitutive variable methylation and 26 loci with tissue-specific variability. **a** Bisulphite pyrosequencing of 103 candidate VM-IAPs in ear tissue and B cells led to classification of constitutive VM-IAPs (cVM-IAPs), tissue-specific VM-IAPs (tsVM-IAPs), or false positives – three representative examples are shown for each. The threshold for validation as a VM-IAP is a 10% methylation range amongst individuals. IAP elements with ≥ 10 and $< 10\%$ range in methylation are coloured red and black, respectively. Each point represents an individual and is the average methylation of four distal CpGs in the LTR. $n = 8$ for ear samples and 7 for B cell samples. **b** IAP-*Cyp4a12a* (top) is a representative example of tsVM-IAPs with variable methylation in multiple tissues. IAP-*Mapk4* (bottom) is a representative example of tsVM-IAPs whose variable methylation is restricted to B cells. **c** IAP elements of the LTR1_Mm-Ez-int (fully-structured) and the LTR2_Mm (solo LTR) types are over-represented in cVM-IAPs. The frequencies of cVM-IAP and tsVM-IAP structures are compared with the relative frequency of genome-wide IAP structures. In the IAP type labels, fully-structured IAPs with flanking LTRs are indicated with “–”, and incomplete IAPs missing one or both flanking LTRs are indicated with “||”. Only IAP types with at least one cVM-IAP or tsVM-IAP are shown. **d** In B cells, expression of *Acss2* is inversely correlated with the methylation level of the nearby tsVM-IAP, IAP-*Acss2* (two-tailed Pearson). Expression was quantified by qPCR, normalised to housekeeping genes, *Pgk1* and *Gapdh*, and analysed across multiple exon-exon junctions: 1–2, 3–7 and 4–7

variably methylated in ear or B cells were also not variably methylated in T cells; these IAPs were termed ‘false positives’, representing 25% (26 out of 103) of the tested loci and 22% of all identified candidates (Fig. 1a, right panel and Supplemental Data).

To determine whether tsVM-IAPs possess consistent intra-individual methylation variation, we assayed multiple tissues, including brain, liver, kidney, and testes. These represent a diverse set of tissues derived from all three embryonic germ layers. Ten tsVM-IAPs are variably methylated in multiple tissues (Fig. 1b, upper panel and Supplemental Fig. S1A). These ten tsVM-IAPs display inter-tissue methylation consistency – i.e., an individual which has low methylation in one tissue tends to be lowly methylated in the other variable tissues (Supplemental Fig. S1B). In contrast, the majority of tsVM-IAPs (16 out of 26) are variably methylated in B cells but not in the other tested tissues (Fig. 1b, lower panel and Supplemental Fig. S1C). Compared to cVM-IAPs, the tsVM-IAPs have less consistent IAP structure (Fig. 1c) and lower methylation variability on average (Supplemental Fig. S1D). The multi-tissue assays used to identify the tsVM-IAPs also confirmed the lack of methylation variability in five false positive IAPs (Supplemental Fig. S1E). We cannot rule out the presence of additional tsVM-IAPs occurring in cell types that we have not assayed.

A role for tsVM-IAPs in transcriptional regulation

We previously reported examples of cVM-IAP-initiated transcripts which overlap with annotated genes and for some of these, the expression level correlated inversely with the methylation level of the cVM-IAP [18]. Around 10% of cVM-IAPs show this effect (Supplemental Table 1). However, in all cases this correlation was tissue-specific. We therefore probed our set of tsVM-IAPs to characterise their effect on transcription of nearby genes. As no tsVM-IAP was located in the vicinity of the transcriptional start site of a gene, we focussed on four tsVM-IAPs located within

introns of genes. We investigated whether expression of the surrounding gene correlates with the methylation level of the tsVM-IAP; both expression and methylation were assayed in the tissues in which the tsVM-IAP is variably methylated (Supplemental Fig. S1A). Using three primer sets targeted to different exons, we found a statistically significant inverse correlation between *Acss2* gene expression and IAP-*Acss2* methylation level in B cells (Fig. 1d), but not in brain or kidney where the IAP was also variably methylated. Because expression levels of exons on either side of IAP-*Acss2* were inversely correlated with the methylation level of the element, it is unlikely that this tsVM-IAP is acting as an alternative promoter. There was no transcriptional effect in the other three tsVM-IAPs investigated (Supplemental Fig. S2).

Repeat-associated variable methylation is mainly a feature of IAPs

To determine whether other families of retrotransposons exhibit the properties of VM-IAPs, we carried out a genome-wide screen for variably methylated LINES, SINEs, ERVs and non-ERV LTRs in the same WGBS datasets used for the IAP screen. The methylation ranges for candidate variably methylated LINES, SINEs, ERVs, and non-ERV LTRs were lower compared to IAPs (LINES, SINEs, non-ERV LTRs, Fig. 2a; IAPs, Supplemental Fig. S3A; ERVs, Supplemental Fig. S3B). We experimentally validated a total of 34 elements representing the top candidates in each repeat family and found two new variably methylated ERVs (VM-ERVs) in addition to 13 previously validated [18] (Supplemental Fig. S3C). We also checked whether the ERVs which did not pass the validation threshold in ear samples were variably methylated in B cells, as this was the hallmark of tissue-specific variability in IAPs. None of the tested ERVs were variably methylated in B cells, suggesting that tissue-specific variable methylation occurs exclusively at IAP elements (Supplemental Fig. S3D).

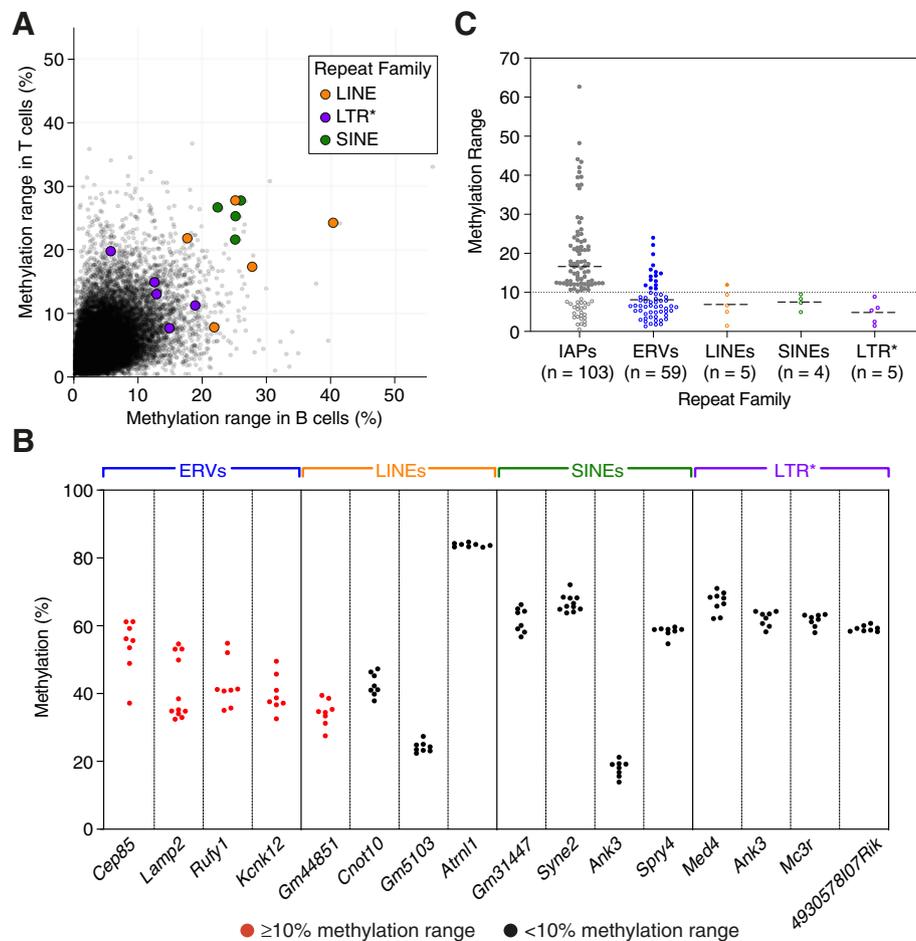


Fig. 2 Variable methylation in TEs is unique to evolutionarily young ERV insertions. **a** Methylation ranges at the edges of LINEs, SINEs and non-ERV LTRs (denoted LTR*) in B and T cell samples from the WGBS dataset. Each point represents an individual element. Coloured points represent the top candidates in each family, which were selected for bisulphite pyrosequencing validation. **b** Bisulphite pyrosequencing validation in ear tissue of the top candidates from the WGBS screen in each repeat family. Elements with ≥ 10 and $< 10\%$ range in methylation between individuals are coloured in red and black, respectively. Each point represents an individual and is the average methylation of 2–4 distal CpGs in the element; $n \geq 8$ for all elements shown. **c** Methylation ranges of validated elements grouped by repeat family suggests that variable methylation is uncommon in LINEs, SINEs and non-ERV LTRs, while validated IAP elements have the highest range. Each point represents an individual element. Elements with ≥ 10 and $< 10\%$ range in methylation between individuals are denoted by filled and blank circles, respectively. The thin black dotted line shows the 10% threshold. The thick black dashed lines represent the average methylation range per repeat family. Methylation was assayed in B cells for tsVM-IAPs and in ear tissue for all other types of elements

Aside from VM-ERVs, LINE-*Gm44851* was the only variably methylated retrotransposon that passed the validation threshold of $\geq 10\%$ methylation range (Fig. 2b and Supplemental Fig. S3E). Two SINEs located on the X chromosome with high methylation ranges in the WGBS dataset were found to be bistable epialleles – loci which have two methylation states within a population – with the two methylation states segregating by sex; these elements were excluded from further analysis (Supplemental Fig. S3F). In total we have identified 77 VM-IAPs (51 constitutive and 26 tissue-specific), 15 VM-ERVs and one VM-LINE. We compared the validated methylation ranges of individual TEs across the different repeat families and found that VM-IAPs are more variable than

the other variably methylated TEs (Fig. 2c). These findings indicate that variable methylation most commonly occurs at IAPs and is not a universal property of TEs.

Sequence specificity of variable methylation at cVM-IAPs
IAPs can be classified into different types based on the sequence of the LTRs (15 types) and internal portions (10 types) [19]. The previously reported VM-IAPs were shown to be enriched in the LTR2_Mm, LTR1_Mm, and EY4_LTR types of IAP LTRs ([18]; ‘IAP’ prefixes omitted for brevity). With the methylation ranges of all candidate IAPs confirmed by pyrosequencing and the improved annotation of IAP elements, we were able to reassess this enrichment by incorporating the internal portions of

non-solo LTR elements into the analysis. Genome-wide, the most common types of IAPs are, in order, LTR1a_Mm – Ez-int and LTR1_Mm – Ez-int (both fully-structured IAPs), followed by EY2_LTR and LTR2a2_Mm (both solo LTR IAPs). The black bars in Fig. 1c show the frequency of each type of IAP relative to the most common type in the genome, LTR1a_Mm – Ez-int. The majority (74%) of cVM-IAPs are made up of just two types of IAPs, LTR2_Mm and LTR1_Mm – Ez-int (Fig. 1c, green bars), despite these being only the second and fifth most common types of IAP genome-wide. The majority of solo LTR VM-IAPs are of the LTR2_Mm type and the majority of full-length VM-IAPs are of the LTR1_Mm – Ez-int type. In contrast to cVM-IAPs, we did not find that tsVM-IAPs are enriched in any one type of element (Fig. 1c, orange bars). These differences – in particular, the underlying sequence difference between the IAP types – may underpin divergent mechanisms responsible for variable methylation at these elements.

It is known that CpG density is generally positively correlated with DNA methylation, but at high CpG densities there is an inverse correlation with DNA methylation [20, 21]. Previously, we found that IAPs have higher CpG density than other ERV retrotransposons in the mouse genome [18]. When comparing CpG density between different types of LTRs, we found that the IAP types for which cVM-IAPs are enriched, namely LTR1_Mm – Ez-int and LTR2_Mm, have higher CpG density than other IAP type LTRs (Supplemental Fig. S4A). Furthermore, cVM-IAPs of the LTR2_Mm type have significantly higher CpG density than non-variable LTR2_Mm elements. On the other hand, cVM-IAPs of the LTR1_Mm – Ez-int type have slightly lower CpG density in their LTRs than non-variable LTR1_Mm – Ez-int elements. Given these contrasts, if CpG density is involved in establishing variable methylation states, the mechanism by which that occurs is dependent on IAP type. Similar to IAPs, VM-ERVs have higher CpG density compared to ERV elements in general, but they have lower CpG density than both cVM-IAPs and tsVM-IAPs (Supplemental Fig. S4B).

Since IAP type and CpG density are both potential sequence-based determinants of methylation variability at cVM-IAPs, we sought to ascertain whether specific sequences within the IAP LTRs may be conferring the variable methylation. We performed a *k*-mer analysis to identify enriched sequences amongst the LTRs of cVM-IAPs (Supplemental Table S2). We identified 14 sequences which are each present in multiple cVM-IAP LTRs and which are present in no more than 2% of all IAP LTRs in the genome. A total of 37 out of the 51 cVM-IAPs contained at least one of these sequences, as well as 6 out of the 26 tsVM-IAPs. Each sequence is

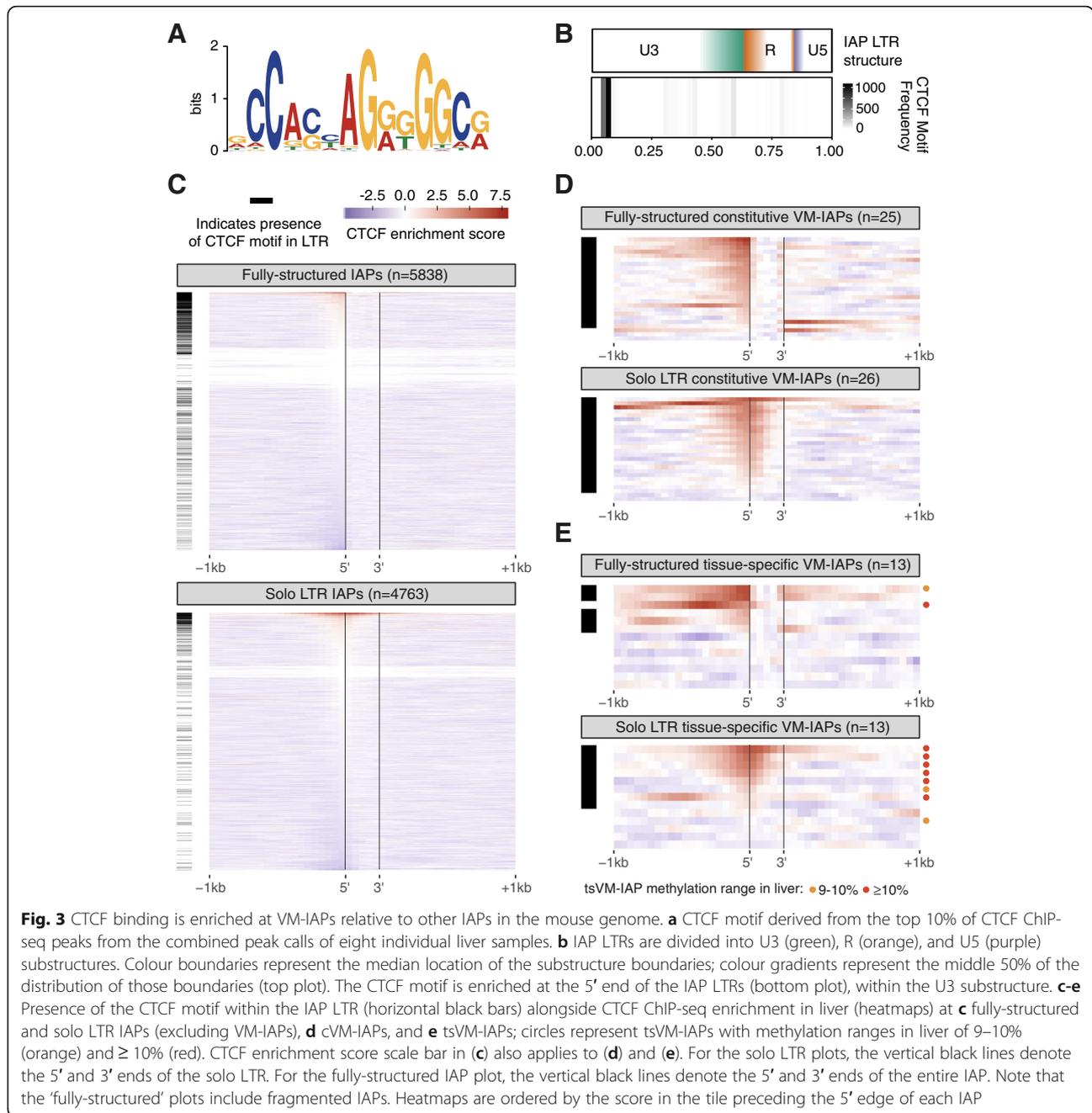
mostly present in only one type of LTR, either LTR2_Mm or LTR1_Mm. We experimentally assessed DNA methylation at 18 IAP elements containing multiple sequences enriched in cVM-IAPs and found that they were all hypermethylated with little inter-individual variability (Supplemental Fig. S5). This proves that sequence is not the sole determinant of variable methylation. However, the existence of sequences enriched amongst cVM-IAPs suggests that variable methylation may be driven, at least in part, by underlying genetic features.

CTCF and its motif are enriched at VM-IAPs

Using CTCF ChIP-seq data from ENCODE, we previously reported that VM-IAPs appear closer to CTCF binding sites compared to non-variable IAPs [18]. To expand and refine our previous findings, we generated CTCF ChIP-seq datasets from livers of eight individuals (Additional File 1). Although the presence of a binding site can be easily determined for a given element, it is difficult to discern whether CTCF binds within the IAP itself or its flanking regions. This is because the ChIP-seq fragments from IAPs often do not contain unique sequence, meaning they cannot be mapped confidently to a specific IAP element. To address this, we mapped the CTCF ChIP-seq datasets to consensus sequences of all named IAP LTR types and found that CTCF is enriched in four LTR types (Supplemental Fig. S6A), including those that are specifically enriched among cVM-IAPs (LTR2_Mm and LTR1_Mm) (Fig. 1c).

To generate a robust CTCF binding motif that correlates with strong CTCF binding, we identified a 14 nucleotide sequence matrix (i.e., motif) derived from the top 10% of ChIP-seq peaks across the pooled eight datasets (Fig. 3a). Binding sites matching this motif are present in around 10% of IAP LTRs at the 5' end of the U3 region (Fig. 3b-c). When analysing CTCF enrichment at specific IAP elements, we found that most IAPs are not bound by CTCF, including many of those which contain the motif (Fig. 3c). In contrast, almost all the cVM-IAPs are enriched for CTCF binding and only five elements do not have the motif (Fig. 3d). Although tsVM-IAPs are not as ubiquitously bound by CTCF as cVM-IAPs, those that are enriched for CTCF binding tend to be variably methylated in liver, the tissue used to generate the ChIP-seq datasets (Fig. 3e). This validates and refines our previously reported finding that CTCF is specifically enriched at VM-IAPs compared to other IAPs in the genome and suggests that there may also be a relationship between CTCF and tsVM-IAPs.

To ask whether the CTCF enrichment at cVM-IAPs is due to the sequence of the CTCF binding site within those elements, we generated motifs using only the binding sites within IAPs, cVM-IAPs, and tsVM-IAPs (Supplemental Fig. S6B). The similarity of these motifs both



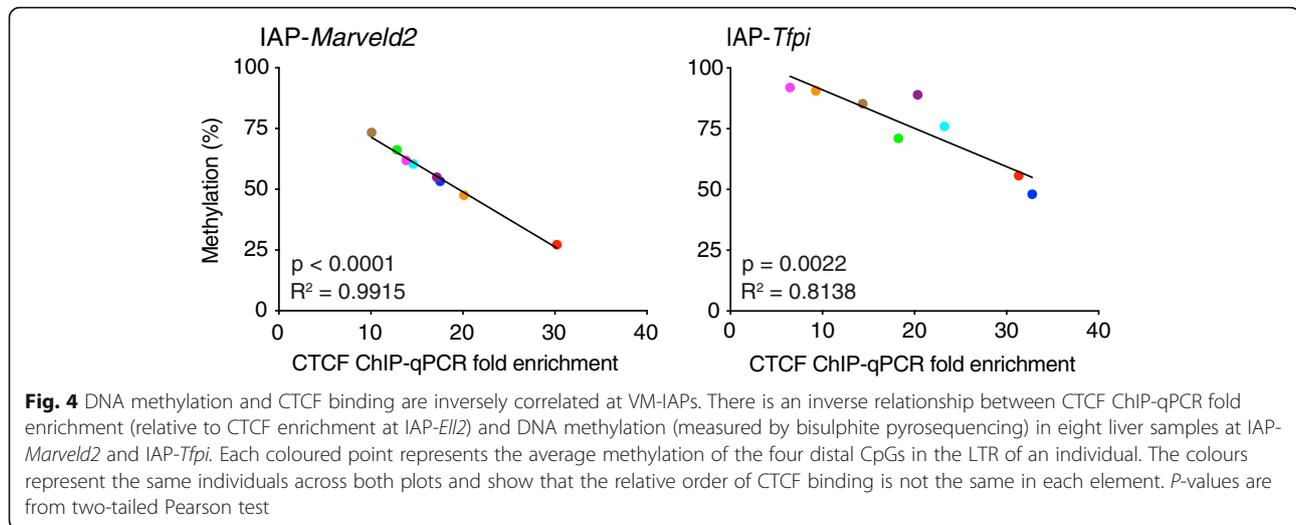
to each other and to the more general genome-wide motif suggests that the specific sequence of the binding site is unlikely to be the cause of the specific CTCF enrichment at cVM-IAPs.

CTCF binding and DNA methylation have an inverse relationship at cVM-IAPs

There is a known inverse relationship between DNA methylation and reduced CTCF binding at many regions in the genome, including imprinted genes and some differentially methylated regions [22–31]. In addition, many

CTCF binding sites in the mouse and human genomes are located within TEs [32–35]. We used our ChIP-seq datasets to ask if CTCF binding is variable between individuals at cVM-IAPs, and if so, whether there is a relationship between CTCF binding and DNA methylation. We found that at six out of the seven analysed cVM-IAPs, there is a significant inverse correlation between DNA methylation at the cVM-IAP and CTCF binding (Supplemental Fig. S7A).

This inverse relationship was validated by performing ChIP-qPCR at IAP-*Marveld2* and IAP-*Tfpi* across the



same eight individuals on which we performed the ChIP-seq experiments (Fig. 4 and Supplemental Fig. S7B), calculating fold enrichment relative to IAP-*Eil2* and to IAP-*Dst* (both non-variable hypermethylated IAPs, Supplemental Fig. S7C). We found that CTCF binding is also variable at some non-variably methylated IAPs (Supplemental Fig. S7D), which is consistent with reports that CTCF binding can be methylation sensitive at some loci and not at others [28, 36]. Our data indicate that methylation is associated with the level of CTCF binding at cVM-IAPs. In contrast to CTCF binding, H3K9me3 levels correlated positively with DNA methylation at these loci (Supplemental Fig. S8).

Chromatin interactions with cVM-IAPs

CTCF is important for establishing chromatin architecture [37–40] and recent findings suggest that TEs bound by CTCF can contribute to chromatin looping, which in turn can influence gene regulation [41–43]. Therefore, we hypothesized that inter-individual variation in CTCF binding at cVM-IAPs could contribute to variation in genome topology. We investigated this using circularised chromatin conformation capture sequencing (4C-seq), a technique used to reveal chromatin interactions between a locus of interest and other parts of the genome. We performed 4C-seq on five individual mice at four cVM-IAPs (IAP-*Marveld2*, IAP-*Tfec*, IAP-*Pink1*, and IAP-*Mbnl1*) and one non-variable IAP (IAP-*Dst*) (Supplemental Fig. S9). In a 400 kb window surrounding the elements, we found that the two cVM-IAPs that were lowly methylated within individuals (IAP-*Marveld2* and IAP-*Tfec*) have more long-range and variable interactions compared to the two cVM-IAPs that were highly methylated within individuals, which show fewer interactions (IAP-*Pink1* and IAP-*Mbnl1*). The control non-variable IAP has a more uniform interaction pattern. Due to

limited sample size we were unable to confidently quantify differential interactions between individuals.

Methylation variability is not confined to the LTR boundaries of cVM-IAPs

To determine the extent to which methylation variability exists outside of the LTR, we first assessed the DNA methylation within 1 kb of cVM-IAPs by bisulphite pyrosequencing. This revealed that variable methylation between individuals extends outside the TE and into the adjacent unique sequence (Fig. 5a and Supplemental Fig. S10A). The relative order of methylation levels across individuals is maintained until inter-individual variability is lost, at a position around 500–1000 bp from the end of the LTR. Although methylation levels beyond this point can show some variability between the individuals (Fig. 5b and c, and Supplemental Fig. S10B and C), this occurs independently of the VM-IAP since the relative order of methylation levels across individuals is no longer retained.

We next considered whether the methylation variability at cVM-IAPs and beyond is associated with a distinct methylation pattern near the elements. By examining the WGBS datasets, we found that the distribution of DNA methylation in the 5 kb flanking each element is not uniform amongst all 51 cVM-IAPs (Supplemental Fig. S10B). We observed a variety of methylation patterns, including fully hypermethylated flanks, hypomethylated regions, intermediately-methylated regions, and tissue-specific methylation; these patterns are not mutually exclusive. We confirmed this by performing bisulphite pyrosequencing on the flanking regions of six cVM-IAPs representative of these patterns (Fig. 5b and c, and Supplemental Fig. S10C). For example, at the regions flanking IAP-*Marveld2*, we found hyper- and intermediate methylation (Fig. 5b), and at the regions flanking IAP-

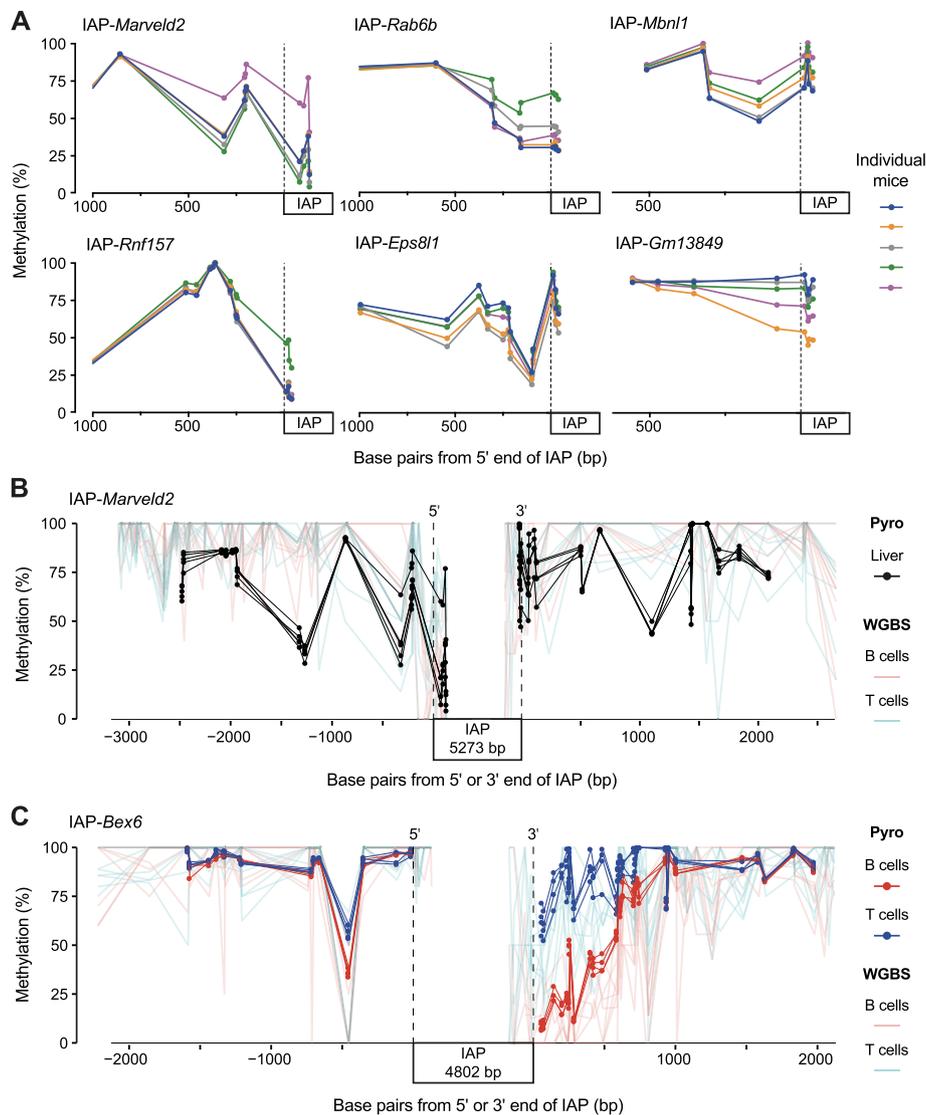


Fig. 5 Methylation variability is not confined to the LTRs of VM-IAPs. **a** Bisulphite pyrosequencing of liver samples from five individuals shows inter-individual methylation variation within 500–1000 bp beyond the edge of six cVM-IAP LTRs. **b–c** WGBS data from B and T cells (light red and blue) in the region flanking cVM-IAPs with bisulphite pyrosequencing (Pyro) validation in **b** five liver samples (black) at IAP-*Marveld2* and **c** B cells and T cells (red and blue) from four individuals at IAP-*Bex6*. IAP length is not to scale. For the pyrosequencing data, each dot represents a CpG. The dashed vertical lines represent the 5' and 3' ends of the IAPs

Bex6, we found a tissue-specific methylation pattern that differed between the B and T cells (Fig. 5c), as expected from the WGBS data. These findings suggest that a particular flanking DNA methylation landscape is not required for variable methylation, nor do VM-IAPs influence the flanking DNA methylation landscape in a specific way.

Discussion

Using WGBS data combined with bisulphite pyrosequencing of independent samples, we classified 51 VM-IAPs as constitutive, meaning variably methylated in all

tested tissues, and 26 as tissue-specific, of which 16 were only variably methylated in B cells. Furthermore, fifteen VM-ERVs were also identified. This provides a validated resource of murine (C57BL/6J) metastable epialleles for further studies.

Given the hypothesis that recently active retrotransposons may have facilitated the rapid adaptive evolution undergone by protein-coding immune genes [44–47], it is noteworthy that tsVM-IAPs are enriched for variable methylation in B cells, an immune cell population. We also found an inverse correlation between the methylation level of a tsVM-IAP and gene expression; the

transcriptional effect was specific to B cells, despite the presence of variable methylation in other tissue types.

DNA methylation variability between individuals is not confined to the boundaries of cVM-IAPs, but also exists in the immediate flanking regions. A few rare examples of DNA methylation spreading from fully methylated retrotransposons have been reported [48–55]. In all tested cVM-IAPs, we found variable methylation within 500 bp of the element boundaries, and sometimes beyond, suggesting that any effects of variable TE methylation may extend into the adjacent sequence. This, along with our finding that the methylation of a tsVM-IAP can correlate with the expression of a nearby gene, indicates that VM-IAPs have potential *cis*-regulatory effects. The absence of a common pattern at these boundaries indicates that VM-IAPs do not confer, or be influenced by, a flanking genomic landscape in a specific way.

We have shown that CTCF is highly enriched at cVM-IAPs and tsVM-IAPs compared to other IAPs in the genome. Unlike VM-IAPs, many non-variable IAP LTRs contain a CTCF motif yet are not bound by the protein; this may be because the majority of IAP LTRs are hypermethylated and CTCF binding is sensitive to methylation. Although it is already known that TEs harbour and spread CTCF binding sites throughout mammalian genomes through mobilisation [33, 34], it is not clear to what extent these CTCF binding sites affect gene regulatory function. Some studies have indicated that the methylation sensitivity of CTCF is not a major contributor to its function [28, 36], although this is not the case at imprinted domains [22–24, 56]. CTCF is known for its role in establishing genomic interactions; we found that these interactions vary depending on the methylation range of the element whereby lowly-methylated cVM-IAPs appear to have more interactions than highly-methylated cVM-IAPs. However, whether the CTCF bound to cVM-IAPs contributes to interactions with specific loci is yet to be determined. It has been shown in both mouse and human that TE insertions enriched for CTCF induce differential loop formation that is significantly associated with effects on gene expression [42]. Our results indicate that epigenetic differences at TEs might also influence CTCF-mediated conformational states.

Results presented here point to mechanisms for when and how variable methylation at TEs is established. The identification of multiple tsVM-IAPs shows that variable methylation can occur in a tissue-specific manner, which indicates that VM-IAPs can likely arise at different developmental time points. As originally observed at the *A^{vy}* and *Axin^{Fu}* loci, cVM-IAPs have inter-individual methylation variation but consistent methylation levels within an individual across tissues, indicating that variable methylation establishment at these loci likely occurs

prior to germ layer specification [18, 57, 58]. Since the extent of methylation variation and the absolute methylation levels differ between tissues at tsVM-IAPs, the mechanisms underlying establishment of variable methylation at these loci are likely to be different from those at cVM-IAPs, and probably occur later in development. These findings also suggest that, rather than being resistant to the early developmental erasure and re-establishment of methylation as has been proposed for IAP elements in general [9, 59], the VM-IAPs (which represent around 1% of the ~10,000 IAP elements) have increased susceptibility to this developmental process [18].

Transcription factor binding at cVM-IAPs and the genetic sequence of cVM-IAP LTRs are likely contributors to the mechanism underlying establishment of variable methylation. We rule out the possibility that genomic methylation context is an important factor in establishing variable methylation by showing that there is no discernible common pattern of DNA methylation flanking cVM-IAPs. Moreover, a correlation between methylation and H3K9me3 was observed along with an inverse correlation between the enrichment of transcription factor CTCF and DNA methylation at cVM-IAPs. During early development, CTCF may compete with DNA methylation machinery for access to VM-IAP LTRs. This hypothesis is consistent with research showing that CTCF can influence the presence or absence of DNA methylation during development with functional consequences [22–25, 28, 36, 56, 60–63]. Therefore, the molecular antagonism between CTCF and DNA methylation machinery could contribute to the formation of variable methylation levels between genetically identical individuals. CTCF may also facilitate interactions with genomic regions that contribute to the regulation of VM-IAP methylation.

The updated categorisation of VM-IAPs revealed that the majority of cVM-IAPs are solo LTR2_Mm or full length LTR1_Mm – Ez-int elements, while tsVM-IAPs are not enriched for any specific type of IAP. This is further evidence that cVM-IAPs and tsVM-IAPs likely arise via separate mechanisms. We have also identified sequence correlates and CpG density profiles of variable methylation at IAPs. Subsets of cVM-IAPs are highly enriched for specific sequences compared to other IAPs in the genome, however these are unable to predict variable methylation. At this point it is unclear exactly how the enriched sequences contribute to establishing variable methylation. A plausible explanation is that they contain binding sites for transcription factors such as CTCF or KRAB zinc finger proteins (KZFPs) [34, 64]. KZFPs are the largest family of transcription factors in mammals and are known to coevolve with and regulate TEs by recruiting heterochromatic machinery to distinct

loci in a sequence specific manner [65–68]. KZFPs are rapidly evolving and highly polymorphic between vertebrates and even among different mouse strains [69]. Variable methylation could arise if there were changes in KZFP binding sites within the VM-IAPs or the binding domains of specific KZFPs that target VM-IAPs. Ultimately, the enrichment of specific sequences at VM-IAPs could contribute to variable methylation via novel or disrupted protein interactions.

The genome-wide screens that we have conducted at TEs in mouse reveal that variable methylation is a rare occurrence that is mostly restricted to IAPs, which are amongst the evolutionarily youngest ERV types [70]. Whether variable methylation is a phenomenon exclusive to young TEs is an open question. Genome-wide screens for metastable epialleles at non-repetitive regions have been performed on the human genome, and variably methylated non-repetitive regions appear to exist [71, 72]. The extent to which these regions are driven by inter-individual genetic differences has not yet been fully determined as it is difficult to eliminate the confounding effect of human genetic variation. Due to the difference in how TEs and non-repetitive regions are regulated by DNA methylation [73], it is likely that the overarching mechanisms of variable methylation establishment and its potential function also vary between these two types of genomic loci.

Conclusions

We have shown that VM-IAPs have the capacity to affect gene regulation by either *cis* or long-range mechanisms and in a tissue-specific manner. Addressing the question of what defines a VM-IAP is made more difficult by the fact that there is no unifying characteristic which distinguishes them from the other 99% of IAPs. Instead, we have shown that there are correlations of varying specificity and confidence by which subpopulations of VM-IAPs differ from the main IAP population with regards to tissue-specific methylation states, CTCF binding, histone modifications, genomic sequence, and CpG density. This shows that there are multiple factors contributing to variable methylation between individuals. The functional implications of TE epigenetic variability and the extent to which this can influence phenotypic outcome remain to be determined.

Methods

Improving the catalogue of IAPs

The GRCm38/mm10 RepeatMasker (henceforth RM) annotation (Dfam v4.0.7) was downloaded from the UCSC table browser [19, 74]. This annotation set contains entries for each transposable element, which may be categorized into one or more ‘subelements’, which are united by an ‘element ID’. All transposable elements

that were considered IAP elements contained at least one IAP subelement of the following types: IAPLTR1a_Mm, IAPEY2_LTR, IAPEy-int, IAPEY3_LTR, IAPLTR2b, IAPLTR2_Mm, IAPEz-int, IAPEY4_I-int, IAP-d-int, IAPEY3-int, RLTR10B2, IAP1-MM_I-int, IAP1-MM_LTR, IAPLTR2a2_Mm, IAPLTR4, IAPLTR1_Mm, IAPEY3C_LTR, IAPLTR4_I, IAPLTR3, IAPLTR3-int, IAPEY5_LTR, IAPEY5_I-int, IAPEY_LTR, IAPEY4_LTR, IAPA_MM-int.

Elements in the RM annotation that overlap 500 kb boundaries were erroneously categorized as two entries with separate element IDs. Each element at one of these boundaries was patched to unify the two entries under the same element ID.

Furthermore, many of the elements are ‘fragmented’ for the following reasons: subsequent insertion of other transposable elements; sequence divergence from the Dfam models; general poor performance of RepeatMasker at ERV elements; or the retrotransposition of an already fragmented IAP element. An element is ‘fragmented’ if it is neither fully-structured (containing an internal portion comprised of ERV genes flanked by tandem LTRs) nor a solo LTR (a single LTR with no ERV genes, formed from intra- or inter-element recombination of two LTRs [75]).

For each fragmented element annotated as missing a 5′ LTR, the following heuristic was used in an attempt to ‘mend’ it, forming a fully-structured IAP: (1) merge the element with an adjacent fragmented element missing a 3′ LTR; (2) merge the element with an adjacent solo LTR; (3) merge the element with an adjacent fully-structured IAP. Adjacent elements must have been within 2000 bp of the 5′ end of annotation to form a match. The same algorithm was used for fragmented elements missing a 3′ LTR, or vice versa. Note that sometimes an element was annotated as containing just an internal portion, so the attempt at mending was performed on both edges. Step 3 of the heuristic could result in the formation of a double or higher-order fully-structured IAP.

Screen for variably methylated transposable elements

A screen for VM-IAPs, similar to that described in Kazachenka et al. 2018, was performed using the improved catalogue of IAPs. The 16 C57BL/6J whole-genome bisulphite sequencing (WGBS) datasets (8 T cell samples, 8 B cell samples) from the BLUEPRINT Epigenome project [76] were mapped to the mm10 reference using Bismark v0.20.0 with default options [77]. This dataset contains 8 ‘standard’ WGBS and 8 oxidative WGBS samples, which were not distinguished in the analysis since hydroxymethylation levels are very low in these samples. Methylation calls were obtained from the aligned reads with a MAPQ ≥ 10 . LTR methylation states

in each sample were calculated at the 5' and 3' edge of each IAP element by determining the average methylation level of the 8 CpGs nearest each LTR edge (only the two outward-facing edges were considered for each element; the internal-facing LTR edges of fully-structured IAPs were excluded). At each LTR edge, a sample with fewer than 20 methylation calls across the 8 CpGs, or fewer than 4 CpGs with coverage, was considered uninformative. LTR edges with fewer than 5 informative samples in either cell type were then excluded from further analysis. The methylation range at each LTR edge surviving these filtering steps was then calculated in each cell type. Unlike in the previous screen, inter-replicate methylation ranges were calculated without excluding the highest and lowest methylation values in each cell type, as this conservative measure was deemed unnecessary in light of the improved IAP reference and stricter filtering on both read quality and region-level coverage.

The screen in LINE, SINE, and non-ERV LTR elements was performed in a similar manner with some modifications: the RepeatMasker annotation was only fixed by combining adjacent elements of the same class within 100 bp of each other; and only the first 8 CpGs within 200 bp of each element edge were considered.

k-mer and CpG density analysis

All sequences of length 15 nt (*k*-mers with $k=15$) present in two or more of the 5' LTRs of the cVM-IAPs ($N=51$) were identified using Jellyfish v2.3.0 [78]. Enrichment of each *k*-mer ($N=2363$) was then calculated relative to a background set of IAP 5' LTRs ($N=9650$). *k*-mers present in at least 5 cVM-IAPs and with an enrichment of at least 20-fold ($N=229$) were then grouped by sequence, such that all *k*-mers in a group overlap with another *k*-mer in the group by $k-1$ nucleotides. Each group was then merged into a single extended sequence using *abyss-align* from the tool ABySS v2.2.3 [79]. The extended sequences were then trimmed to the sub-sequence maximising the enrichment in cVM-IAPs versus background IAPs. Identified sequences are listed in Supplemental Table S2; pyrosequencing primers are listed in Supplemental Table S1.

CpG density was calculated for IAP LTRs by normalising the number of CpGs in the LTR by the LTR length in base pairs.

Tissue collection, DNA/RNA extraction and bisulphite pyrosequencing

Immediately following dissection, C57BL/6 J tissues were snap frozen in liquid nitrogen and manually pulverised. 30µg of tissue (brain, liver, kidney, testes, B and T cells) was used for simultaneous purification of genomic DNA and total RNA with the AllPrep DNA/RNA Mini Kit -

Quick-Start Protocol (QIAGEN, cat. no. 80204). Ear notch samples were lysed (Lysis Buffer: 10 mM EDTA, 150 mM NaCl, 10 mM Tris-HCl pH 8, 0.1% SDS) and DNA was purified using a standard phenol chloroform extraction protocol. 0.5-1 µg of DNA per sample was bisulphite converted using the two-step protocol of the Sigma Imprint® DNA Modification Kit according to the manufacturer's instructions. Following PCR amplification, CpG site-specific methylation was quantified using the PyroMark™ Q96 MD pyrosequencer (Biotage) as previously described [18]. Primers are listed in Supplemental Table S1 (for IAPs, ERVs, LINES, SINEs, non-ERV and spreading).

B and T cell sorts

Splenic tissue was ground through a 70 µm cell strainer to obtain a single cell suspension in PBS with 2% heat-inactivated fetal calf serum (FCS). Red blood cell lysis was performed on ice using cold ammonium chloride (Stem Cell Technologies, cat. No 07800). Cells were subsequently stained as described in the panel (Table 1) for 30 min at 4 °C. Cells were washed twice with 2% FCS in 2 ml PBS. 7-aminoactinomycin D (7-AAD) (1:100; Biolegend, 420,404) was added as a viability stain. Using an Influx Cell Sorter (BD Biosciences), cells were sequentially gated based on cell size, presence of singlets, and live cells to sort CD3+, CD4+, CD25-, CD44lo, CD62L+ T cells and CD19+ CD43- B cells. The sorted populations include naïve CD4+ T cells and T1, T2, marginal zone and follicular B cells. Gates were confirmed using 'fluorescence minus one' controls for CD44 and CD43 markers.

RT-qPCR

RNA was treated with RNase-free DNase I (ThermoScientific, EN0521) prior to cDNA synthesis using RevertAid H Minus First Strand cDNA Synthesis Kit with oligo dT and random hexamer primers (ThermoScientific). qPCR primers were designed using Primer-BLAST and are listed in Supplemental Table S3. qPCR was performed with Brilliant II SYBR® Green QPCR Master Mix (Agilent Technologies) in a LightCycler® 480

Table 1 Fluorophores for B and T cell panel

Antigen	Fluorochrome	Clone	Company	Catalogue Number
CD19	BV711	6D5	Biolegend	115,555
CD43	APC	S11	Biolegend	143,207
CD3	FITC	17A2	Biolegend	100,203
CD4	APC/Cy7	RM4-5	Biolegend	100,525
CD25	BV421	PC61	Biolegend	102,033
CD44	PE/Cy7	IM7	Biolegend	103,029
CD62L	PE	MEL-14	Biolegend	104,407

Instrument II (Roche). Relative gene expression was calculated using the standard curve method and cDNA input was normalised using housekeeping genes *Pgk1* and *Gapdh*. The significance of correlations between gene expression and VM-IAP methylation levels was assessed by computing Pearson correlation coefficients followed by two-tailed *p* values in GraphPad Prism.

Chromatin immunoprecipitation (ChIP) - qPCR and sequencing

Chromatin immunoprecipitation (ChIP) was performed as previously described with some modifications [65]. 100 mg of powdered frozen mouse liver was crosslinked in 1% formaldehyde for 10 min and subsequently quenched with Tris pH 8.0 (250 mM final) for 10 min. The quenched cells were washed twice with ice-cold PBS supplemented with a protease inhibitor cocktail (EDTA-free cOmplete™, Sigma Aldrich), flash frozen in liquid nitrogen, and stored at -80°C . Crosslinked cells were thawed on ice and then lysed sequentially on ice and for 10 min at each step in each of the following buffers: LB1 (50 mM HEPES- KOH pH 7.4, 140 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 10% Glycerol, 0.5% NP-40, 0.25% Triton-X-100 and EDTA-free cOmplete™), LB2 (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA and EDTA-free cOmplete™), and SDS shearing lysis buffer (10 mM Tris-HCl pH 8, 1 mM EDTA, 0.15% SDS and EDTA-free cOmplete™). The lysates were sonicated (Bioruptor) at 4°C to generate DNA fragments of 100–500 bp (3 repetitions of 5 sonication cycles of 30 s on and 30 s off for the CTCF ChIP; 1 repetition of 8 cycles of 30 s on and 30 s off for the H3K9me3 ChIP) and the sonicated lysates were subsequently clarified by centrifugation (15,000 rpm for 15 min at 4°C). The sonicated lysate was divided into 1.5-mL Eppendorf tubes based on the number of ChIPs performed and topped up to 1 mL with Lysis Buffer 500NaCl (20 mM Tris-HCl, pH 7.5, 500 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 1% Triton-X-100, 0.1% Sodium deoxycholate, 0.1% SDS and EDTA-free cOmplete™). This mixture was incubated overnight at 4°C with beads (Protein A Dynabeads, Invitrogen, for CTCF ChIP; Protein G Dynabeads, Invitrogen, for H3K9me3 ChIP) that had been pre-blocked with 0.5% BSA and mixed with polyclonal CTCF antibody (C15410210–50, Diagenode) or polyclonal H3K9me3 antibody (AB_2532132, Active Motif). To remove non-specifically bound proteins from the CTCF ChIP, beads were washed five times with RIPA Buffer (50 mM HEPE S-KOH, pH 7.4, 500 mM LiCl, 1 mM EDTA, 1% NP-40 and 0.7% Sodium deoxycholate) and once with TE Buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA) at 4°C . To remove non-specifically bound proteins from the H3K9me3 ChIP, beads were washed twice with low salt

buffer (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.15% SDS, 1 mM PMSF) followed by single successive washes with high salt buffer (10 mM Tris-HCl pH 8.0, 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.15% SDS, 1 mM PMSF), LiCl buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA, 0.5 mM EGTA, 250 mM LiCl, 1% NP40, 1% Na- deoxycholate, 1 mM PMSF), and 10 mM Tris pH 8.0 at 4°C . The DNA-protein complex was eluted from the beads in Elution Buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA and 1% SDS) for 20 min at 65°C , and reverse-crosslinked overnight at 65°C . The eluted samples were then treated with RNase A (Wako) and Proteinase K (Roche), and purified using a PCR purification kit (NEB).

qPCR was performed with Brilliant II SYBR® Green qPCR Master Mix (Agilent Technologies) in a LightCycler® 480 Instrument II (Roche). The ChIP-qPCR fold enrichments are calculated by $2^{\Delta\text{Ct}}$, where ΔCt is the difference in qPCR Ct value between the tested IAP and a control IAP (either *IAP-Dst*, *IAP-Ell2*, or *IAP-Asxl3*). qPCR primers were designed using Primer-BLAST and are listed in Supplemental Table S3.

ChIP-seq libraries were prepared using KAPA Adapters, KAPA HyperPrep Kit (KAPA Biosystems), and AMPure XP Beads (Beckman Coulter) and quality checked using Qubit, Bioanalyzer, and TapeStation. The libraries were sequenced as 150 bp paired-end reads on the Illumina HiSeq4000. The resulting ChIP-seq data was trimmed by Trim Galore and aligned using bwa 0.7.15. For heatmaps, bamCompare from deeptools 3.3.1 was used to generate CTCF binding scores in 50-bp tiles across the genome, using the combined reads of all eight individuals. Each IAP element is split into five equal-sized tiles. The score is the log₂ ratio of ChIP reads to input reads; Integrative Genomics Viewer (IGV) was used to visualise a bedGraph file of the log₂ ratios. Using only reads with MAPQ \geq 10, ChIP peaks and summits were called by MACS2 2.1.0 as described in [80] and were visualized using custom R scripts (see Data Access). Genomic context of CTCF peaks in Additional File 1 was analysed using the ChIPQC package from R/Bioconductor [81]. The MACS2 summits from all 8 CTCF ChIP-seq samples were combined, and the sequence of a 50 bp window centred on each summit ($N = 97,746$ summits) was used as input to MEME 5.0.4, using the strategy of motif finding from [33]. The FIMO tool from MEME was then used to identify genome-wide locations of the top motif.

4C-seq

4C-seq was performed as previously described with some modifications [82]. Tissues were fixed as outlined in the ChIP protocol above. DpnII (New England Biolabs) was used as the primary restriction enzyme and

NlaIII (New England Biolabs) as the secondary restriction enzyme. Prior to library preparation, samples were purified by the Monarch PCR & DNA Cleanup Kit (New England Biolabs). For the library preparation, 16 individual PCR reactions were performed for each sample per viewpoint with reverse primers containing indexes (see Supplemental Table S4). The 16 PCRs were combined and purified using 0.8x Agencourt AMPure XP beads (Beckman Coulter). Five libraries for each of the five tested viewpoints were multiplexed, quality checked using Qubit and Bioanalyzer, and sequenced as 150 bp paired-end reads on the Illumina HiSeq4000. The sequencing data for each viewpoint were processed using the *smoothCounts* method from the R Bioconductor package FourCSeq [83]. Read counts for each restriction fragment were then normalised and log transformed. The trend line and its 95% confidence interval were generated by the *loess.sd* method from the R package msir using span = 0.01 [84]. ChromHMM data is from mouse liver, and was downloaded from a public GitHub repository https://github.com/gireeshkbogu/chromatin_states_chromHMM_mm9/blob/master/liver_cStates_HMM.zip and lifted over to mm10 [85]. Gene tracks are from the R Bioconductor package EnsDb.Mmusculus.v79.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13100-021-00235-1>.

Additional file 1: CTCF ChIP-seq metadata and genomic features.

Additional file 2: Supplemental Figures.

Additional file 3: Supplemental Table 1.

Additional file 4: Supplemental Table 2.

Additional file 5: Supplemental Table 3.

Additional file 6: Supplemental Table 4.

Additional file 7: Supplemental Data.

Acknowledgements

We are grateful to Kay Harnish, Julie Ahringer and Alex Appert, Iwo Kucinski, Bertie Gottgens for access to facilities and technology optimisation. We thank Michael Imbeault, Felipe Karam Teixeira, Tyler Linderth, Carol Edwards and Fran Dearden for helpful discussions and Gloria Jansen for technical assistance. Thanks to members of the Ferguson-Smith lab for feedback and discussions including during manuscript preparation.

Consent for participation

Not applicable.

Authors' contributions

J.L.E., A.D.H., T.M.B. and E.A.C.A. conducted experimental work and analysis, N.J.K. and A.D.H. conducted the computational work, and A.C.F.S. conceived and supervised the project. J.L.E., A.D.H., N.J.K. and A.C.F.S. wrote the manuscript with input from T.M.B. All authors read and approved the final manuscript.

Funding

Research was funded by grants from the MRC (MR/R009791/1), BBSRC (BB/R009996/1), and a Wellcome Trust Investigator Award (210757/Z/18/Z) to A.C.F.S. We are grateful for PhD studentships from the BBSRC to J.L.E., from

the Cambridge Trust and Department of Genetics to A.D.H., and from the Cambridge Trust, Downing and Pomona Colleges to T.M.B.

Availability of data and materials

All raw and processed sequencing data generated in this study have been submitted to the NCBI Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>) under accession number GSE159110 (CTCF ChIP-seq and 4C-seq) and GSE159790 (BLUEPRINT methylation data). All code used to perform the analyses is available at GitHub (<https://github.com/AFS-lab/vm-retrotransposons>).

Ethics approval

Research in the Ferguson-Smith lab involving laboratory mice has been approved by the Animal Welfare and Ethical Review Body (AWERB) of the University of Cambridge and is governed by the UK Government Home Office licence PC213320E.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 26 November 2020 Accepted: 2 February 2021

Published online: 21 February 2021

References

1. Mouse Genome Sequencing Consortium, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 2002;420(6915):520–62.
2. Nellaker C, Keane TM, Yalcin B, Wong K, Agam A, Belgard TG, et al. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biol*. 2012;13(6):R45.
3. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS Genet*. 2006;2(1):e2.
4. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. *Genome Biol*. 2018;19(1):199.
5. Dewannieux M, Dupressoir A, Harper F, Pierron G, Heidmann T. Identification of autonomous IAP LTR retrotransposons mobile in mammalian cells. *Nat Genet*. 2004;36(5):534–9.
6. Heidmann O, Heidmann T. Retrotransposition of a mouse IAP sequence tagged with an indicator gene. *Cell*. 1991;64(1):159–70.
7. Gagnier L, Belancio VP, Mager DL. Mouse germ line mutations due to retrotransposon insertions. *Mob DNA*. 2019;10:15.
8. Zhang Y, Maksakova IA, Gagnier L, van de Lagemaat LN, Mager DL. Genome-wide assessments reveal extremely high levels of polymorphism of two active families of mouse endogenous retroviral elements. *PLoS Genet*. 2008;4(2):e1000007.
9. Lane N, Dean W, Erhardt S, Hajkova P, Surani A, Walter J, et al. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis*. 2003;35(2):88–93.
10. Seisenberger S, Peat JR, Hore TA, Santos F, Dean W, Reik W. Reprogramming DNA methylation in the mammalian life cycle: building and breaking epigenetic barriers. *Philos Trans R Soc Lond Ser B Biol Sci*. 2013;368(1609):20110330.
11. Dickies MM. A new viable yellow mutation in the house mouse. *J Hered*. 1962;53:84–6.
12. Reed SC. The inheritance and expression of fused, a new mutation in the house Mouse. *Genetics*. 1937;22(1):1–13.
13. Bertozzi TM, Ferguson-Smith AC. Metastable epialleles and their contribution to epigenetic inheritance in mammals. *Semin Cell Dev Biol*. 2020;97:93–105.
14. Rakyank VK, Chong S, Champ ME, Cuthbert PC, Morgan HD, Luu KV, et al. Transgenerational inheritance of epigenetic states at the murine Axin (Fu) allele occurs after maternal and paternal transmission. *Proc Natl Acad Sci U S A*. 2003;100(5):2538–43.
15. Rakyank VK, Blewitt ME, Druker R, Preis JI, Whitelaw E. Metastable epialleles in mammals. *Trends Genet*. 2002;18(7):348–51.

16. Morgan HD, Sutherland HG, Martin DI, Whitelaw E. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet.* 1999;23(3):314–8.
17. Vasicek TJ, Zeng L, Guan XJ, Zhang T, Costantini F, Tilghman SM. Two dominant mutations in the mouse fused gene are the result of transposon insertions. *Genetics.* 1997;147(2):777–86.
18. Kazachenka A, Bertozzi TM, Sjoberg-Herrera MK, Walker N, Gardner J, Gunning R, et al. Identification, characterization, and heritability of murine metastable Epialleles: implications for non-genetic inheritance. *Cell.* 2018;175(6):1717.
19. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* 2016;44(D1):D81–9.
20. Edwards JR, O'Donnell AH, Rollins RA, Peckham HE, Lee C, Milekic MH, et al. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res.* 2010;20(7):972–80.
21. Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature.* 2008;454(7205):766–70.
22. Hark AT, Schoenherr CJ, Katz DJ, Ingram RS, Levorse JM, Tilghman SM. CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature.* 2000;405(6785):486–9.
23. Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature.* 2000;405(6785):482–5.
24. Engel N, Thorvaldsen JL, Bartolomei MS. CTCF binding sites promote transcription initiation and prevent DNA methylation on the maternal allele at the imprinted H19/Igf2 locus. *Hum Mol Genet.* 2006;15(19):2945–54.
25. Flavahan WA, Drier Y, Liao BB, Gillespie SM, Venteicher AS, Stemmer-Rachamimov AO, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature.* 2016;529(7584):110–4.
26. Han L, Lee DH, Szabo PE. CTCF is the master organizer of domain-wide allele-specific chromatin at the H19/Igf2 imprinted region. *Mol Cell Biol.* 2008;28(3):1124–35.
27. Lin S, Ferguson-Smith AC, Schultz RM, Bartolomei MS. Nonallelic transcriptional roles of CTCF and cohesins at imprinted loci. *Mol Cell Biol.* 2011;31(15):3094–104.
28. Maurano MT, Wang H, John S, Shafer A, Canfield T, Lee K, et al. Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep.* 2015;12(7):1184–95.
29. Merckenschlager M, Nora EP. CTCF and Cohesin in Genome folding and transcriptional gene regulation. *Annu Rev Genomics Hum Genet.* 2016;17:17–43.
30. Prickett AR, Barkas N, McCole RB, Hughes S, Amante SM, Schulz R, et al. Genome-wide and parental allele-specific analysis of CTCF and cohesin DNA binding in mouse brain reveals a tissue-specific binding pattern and an association with imprinted differentially methylated regions. *Genome Res.* 2013;23(10):1624–35.
31. Renda M, Baglivo I, Burgess-Beusse B, Esposito S, Fattorusso R, Felsenfeld G, et al. Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J Biol Chem.* 2007;282(46):33336–45.
32. Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.* 2008;18(11):1752–62.
33. Schmidt D, Schwali PC, Wilson MD, Ballester B, Goncalves A, Kutter C, et al. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell.* 2012;148(1–2):335–48.
34. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. *Genome Res.* 2014;24(12):1963–76.
35. Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Kaliskan M, et al. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res.* 2017;27(10):1623–33.
36. Wang H, Maurano MT, Qu H, Varley KE, Gertz J, Pauli F, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* 2012;22(9):1680–8.
37. Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, et al. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol.* 1996;16(6):2802–13.
38. Moon H, Filippova G, Loukinov D, Pugacheva E, Chen Q, Smith ST, et al. CTCF is conserved from *Drosophila* to humans and confers enhancer blocking of the Fab-8 insulator. *EMBO Rep.* 2005;6(2):165–70.
39. Phillips JE, Corces VG. CTCF: master weaver of the genome. *Cell.* 2009;137(7):1194–211.
40. Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, Qin ZS, et al. Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.* 2014;15(6):R82.
41. Choudhary MN, Friedman RZ, Wang JT, Jang HS, Zhuo X, Wang T. Co-opted transposons help perpetuate conserved higher-order chromosomal structures. *Genome Biol.* 2020;21(1):16.
42. Diehl AG, Ouyang N, Boyle AP. Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. *Nat Commun.* 2020;11(1):1796.
43. Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, et al. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet.* 2019;51(9):1380–8.
44. Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science.* 2016;351(6277):1083–7.
45. Ivancevic A, Chuong EB. Transposable elements teach T cells new tricks. *Proc Natl Acad Sci U S A.* 2020;117(17):9145–7.
46. Tie CH, Fernandes L, Conde L, Robbez-Masson L, Sumner RP, Peacock T, et al. KAP1 regulates endogenous retroviruses in adult human cells and contributes to innate immune control. *EMBO Rep.* 2018;19(10):e45000.
47. Ye M, Goudot C, Hoyler T, Lemoine B, Amigorena S, Zueva E. Specific subfamilies of transposable elements contribute to different domains of T lymphocyte enhancers. *Proc Natl Acad Sci U S A.* 2020;117(14):7905–16.
48. Graff JR, Herman JG, Myohanen S, Baylin SB, Vertino PM. Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions in de novo methylation. *J Biol Chem.* 1997;272(35):22322–9.
49. Magewu AN, Jones PA. Ubiquitous and tenacious methylation of the CpG site in codon 248 of the p53 gene may explain its frequent appearance as a mutational hot spot in human cancer. *Mol Cell Biol.* 1994;14(6):4225–32.
50. Mummaneni P, Bishop PL, Turker MS. A cis-acting element accounts for a conserved methylation pattern upstream of the mouse adenine phosphoribosyltransferase gene. *J Biol Chem.* 1993;268(1):552–8.
51. Mummaneni P, Walker KA, Bishop PL, Turker MS. Epigenetic gene inactivation induced by a cis-acting methylation center. *J Biol Chem.* 1995;270(2):788–92.
52. Oey H, Isbel L, Hickey P, Ebaid B, Whitelaw E. Genetic and epigenetic variation among inbred mouse littermates: identification of inter-individual differentially methylated regions. *Epigenetics Chromatin.* 2015;8:54.
53. Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, Zhang Y, et al. Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. *PLoS Genet.* 2011;7(9):e1002301.
54. Rebollo R, Miceli-Royer K, Zhang Y, Farivar S, Gagnier L, Mager DL. Epigenetic interplay between mouse endogenous retroviruses and host genes. *Genome Biol.* 2012;13(10):R89.
55. Yates PA, Burman R, Simpson J, Ponomoreva ON, Thayer MJ, Turker MS. Silencing of mouse *Apt* is a gradual process in differentiated cells. *Mol Cell Biol.* 2003;23(13):4461–70.
56. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell.* 1999;98(3):387–96.
57. Waterland RA, Jirtle RL. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol Cell Biol.* 2003;23(15):5293–300.
58. Waterland RA, Lin JR, Smith CA, Jirtle RL. Post-weaning diet affects genomic imprinting at the insulin-like growth factor 2 (*Igf2*) locus. *Hum Mol Genet.* 2006;15(5):705–16.
59. Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet.* 1998;20(2):116–7.
60. Feldmann A, Ivanek R, Murr R, Gaidatzis D, Burger L, Schubeler D. Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. *PLoS Genet.* 2013;9(12):e1003994.
61. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature.* 2011;480(7378):490–5.
62. Teif VB, Beshnova DA, Vainshtein Y, Marth C, Mallm JP, Hofer T, et al. Nucleosome repositioning links DNA (de) methylation and differential CTCF binding during stem cell development. *Genome Res.* 2014;24(8):1285–95.
63. Wiehle L, Thorn GJ, Raddatz G, Clarkson CT, Rippe K, Lyko F, et al. DNA (de) methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.* 2019;29(5):750–61.

64. Wolf D, Goff SP. Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature*. 2009;458(7242):1201–4.
65. Imbeault M, Helleboid PY, Trono D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*. 2017;543(7646):550–4.
66. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, et al. KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature*. 2010;463(7278):237–40.
67. Thomas JH, Schneider S. Coevolution of retroelements and tandem zinc finger genes. *Genome Res*. 2011;21(11):1800–12.
68. Wolf G, Yang P, Fuchtbauer AC, Fuchtbauer EM, Silva AM, Park C, et al. The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. *Genes Dev*. 2015;29(5):538–54.
69. Elmer JL, Ferguson-Smith AC. Strain-Specific Epigenetic Regulation of Endogenous Retroviruses: The Role of Trans-Acting Modifiers. *Viruses*. 2020; 12(8):810.
70. Qin C, Wang Z, Shang J, Bekkari K, Liu R, Pacchione S, et al. Intracisternal a particle genes: distribution in the mouse genome, active subtypes, and potential roles as species-specific mediators of susceptibility to cancer. *Mol Carcinog*. 2010;49(1):54–67.
71. Kessler NJ, Waterland RA, Prentice AM, Silver MJ. Establishment of environmentally sensitive DNA methylation states in the very early human embryo. *Sci Adv*. 2018;4(7):eaat2624.
72. Silver MJ, Kessler NJ, Hennig BJ, Dominguez-Salas P, Laritsky E, Baker MS, et al. Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptual environment. *Genome Biol*. 2015;16:118.
73. Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol*. 2019;20(10): 590–607.
74. Smit AFA, Hubley R & Green, P. *RepeatMasker Open-4.0*. 2013-2015. <http://repeatmasker.org/faq.html#faq3>.
75. Ji Y, DeWoody JA. Genomic landscape of long terminal repeat Retrotransposons (LTR-RTs) and solo LTRs as shaped by ectopic recombination in chicken and Zebra finch. *J Mol Evol*. 2016;82(6):251–63.
76. Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, et al. BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol*. 2012;30(3):224–6.
77. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. *Bioinformatics*. 2011;27(11):1571–2.
78. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. 2011;27(6):764–70.
79. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, et al. ABySS 2.0: resource-efficient assembly of large genomes using a bloom filter. *Genome Res*. 2017;27(5):768–77.
80. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178–92.
81. Carroll TS, Liang Z, Salama R, Stark R, de Santiago I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front Genet*. 2014;5:75.
82. Splinter E, de Wit E, van de Werken HJ, Klous P, de Laat W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: from fixation to computation. *Methods*. 2012;58(3):221–30.
83. Klein FA, Pakozdi T, Anders S, Ghavi-Helm Y, Furlong EE, Huber W. FourCSeq: analysis of 4C sequencing data. *Bioinformatics*. 2015;31(19):3085–91.
84. Scrucca L. Model-based SIR for dimension reduction. *Comput Stat Data Anal*. 2011;55(11):3010–26.
85. Bogu GK, Vizan P, Stanton LW, Beato M, Di Croce L, Marti-Renom MA. Chromatin and RNA maps reveal regulatory long noncoding RNAs in Mouse. *Mol Cell Biol*. 2015;36(5):809–19.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

