Two repeated motifs enriched within some enhancers and origins of replication are bound by SETMAR isoforms in human colon cells

Aymeric Antoine-Lorquin, Peter Arensburger, Ahmed Arnaoty, Sassan Asgari, Martine Batailler, Linda Beauclair, Catherine Belleannée, Nicolas Buisine, Vincent Coustham, Serge Guyetant, et al.

Two repeated motifs enriched within some enhancers and origins of replication are bound by SETMAR isoforms in human colon cells.

Aymeric Antoine-Lorquin[1,11], Peter Arensburger[2,11], Ahmed Arnaoty[3,11], Sassan Asgari[4,11], Martine Batailler[5,11], Linda Beauclair[5,11], Catherine Belleannée[1,11], Nicolas Buisine[6,11], Vincent Coustham[7,11], Serge Guyetant[8,11], Laura Helou[5,11], Thierry Lecomte[3,11], Bruno Pitard[9,11], Isabelle Stévant[10,11] & Yves Bigot[5]

[1] IRISA, 263 avenue du Général Leclerc, 35042 Rennes, France

[2] Biological Sciences Department, California State Polytechnic University, Pomona, CA 91768 - United States of America

[3] EA GICC 7501, CHRU de Tours, 37044 TOURS Cedex 09

[4] School of Biological Sciences, The University of Queensland, Brisbane QLD 4072, Australia

[5] PRC, UMR INRA 0085 and CNRS 7247, Centre INRA Val de Loire, 37380 Nouzilly, France

[6] UMR CNRS 7221, Muséum National d'Histoire Naturelle, 75005 Paris, France

[7] BOA, INRAE, Université de Tours, 37380 Nouzilly, France

[8] Tumorothèque du CHRU de Tours, 37044 Tours cedex, France

[9] Université de Nantes, CNRS ERL6001, Inserm 1232, CRCINA, F-44000 Nantes, France.

[10] Institut de Génomique Fonctionnelle de Lyon, Univ Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, 46 allée d'Italie, 69364 Lyon, France

[11] Author names were ranked in alphabetic order because all contributions were essential to the project managed by Y. Bigot.

Corresponding author: yves.bigot@inrae.fr

Abstract

Setmar is a gene specific to simian genomes. The function(s) of its isoforms are poorly understood and their existence in healthy tissues remains to be validated. Here we profiled SETMAR expression and its genome-wide binding landscape in colon tissue. We found isoforms V3 and V6 in healthy and tumour colon tissues as well as incell lines. In two colorectal cell lines SETMAR binds to several thousand *Hsmar1* and MADE1 terminal ends, transposons mostly located in non-genic regions of active chromatin including in enhancers. It also binds to a 12-bp motifs similar to an inner motif in *Hsmar1* and MADE1 terminal ends. This motif is interspersed throughout the genome and is enriched in GC-rich regions as well as in CpG islands that contain constitutive replication origins. It is also found in enhancers other than those associated with *Hsmar1* and MADE1. The role of SETMAR in the expression of genes, DNA replication and in DNA repair are discussed.

Keywords: transposon, MITE, domestication, coevolution, DNA binding, expression enhancer, DNA replication

1. Introduction

Setmar is one of 54 genes in the human genome that is derived from DNA transposable elements (TEs) by exaptation of an open reading frame (ORF) encoding a transposase (Table S1). Transposases are enzymes that mediate all the DNA cleavage and strand transfer reactions required for transposition of DNA fragments and are a hallmark of TEs [1]. The *setmar* gene is approximately 14 kbp long and is located on human chromosome 3 (positions 4,292,212 to 4,328,658 in GRCh38/hg38). This gene arose 44 ± 4.4 million years ago in the simian lineage (anthropoid lineage at the origin of the hominoids and the old and new world monkeys) [2,3] from the functional fusion of two neighbouring genes. The first gene, *set* (so-called Etet2 in mouse), is composed of two exons coding a lysine methyltransferase, while the second gene, *Hsmar1*, encodes a transposase (HSMAR1) belonging to the *mariner* family (Fig. 1a). *Mariners* are structurally simple TEs composed of one transposase ORF flanked by two inverted terminal repeats (ITRs). These flanking repeats are specifically bound by the transposase that excises the TE from a 'donor' site before catalysing its insertion at another 'target' locus. During evolution of the anthropoid lineage the accumulation of at least three mutations modified the expression properties of the *set* and *Hsmar1* genes. They acquired the ability to be transcribed together within the same mRNA and thus form the *setmar* gene. Alternative splicing events of most *setmar* transcripts are concentrated in the SET domain (Fig. 1) that is fused to the HSMAR1 domain and produced 4 to 8 SETMAR isoform [2], except for the V4 isoform which is the original SET protein.

The *setmar* gene was shown to have evolved under strong purifying selection [2], suggesting it is subject to functional constraints. Significant efforts have been carried out to elucidate the enzymatic activities of the SET and HSMAR1 moieties, to identify cellular protein partners, and to integrate this knowledge into biological pathways. As a fusion gene, *setmar* accumulates the activities of its two moieties. The SET moiety is associated with the methylation of lysine 130 of the snRNP70 splicing factors [4] and the dimethylation of histone 3 lysine 36 (H3K36me2) [5]. H3K36me2 is involved in double-strand break repair by recruiting early repair factors such as NBS1 and Ku70 [5], as well as in the recruitment of the DNA methyltransferase DNMT3A and the maintenance of DNA methylation at intergenic regions [6]. The N-terminal domain of the HSMAR1 moiety has kept its binding specificity to the ITRs of the *Hsmar1*/MADE1 TE *in vitro* [7]. Its C-terminal domain carries out most of the cleavage and strand transfer activities of the *mariner* transposases, except that the 3' second strand cleavage at the outer ITR ends is severely affected by a mutation of the

catalytic triad (DDN instead of DDD) [7,8]. In addition to the respective activities of its moieties, SETMAR is also able to form a stable complex with the human psoralen 4 protein (hPso4 a.k.a. PRP19) [9,10] that binds *in vitro* to non-ITR double-stranded DNA targets. The biological activities of SETMAR have been documented by R. Hromas' team and his collaborators since 2005 through intensive study of its largest isoform (V1, 684 amino acid residues; Fig. 1a) in the context of at least three housekeeping mechanisms using cancerous cell lines. First, SETMAR V1 has been shown to enhance chromosomal decatenation [11,12]. Second, SETMAR V1 improves the efficiency and the accuracy of the DNA repair by non-homologous end-joining (NHEJ) [13-16] which is the primary repair pathway throughout the cell cycle. Third, SETMAR V1 positively impacts the restart of stalled replication forks after DNA damage repair [17]. In addition, SETMAR V1 appears to enhance the chromosomal integration of both transfected DNA fragments [18] and lentiviral DNA into host cell genomes [19]. In agreement with these properties, SETMAR acts negatively on the onset of apoptosis and mediates resistance to DNA damaging cancer chemotherapy [20,21]. In 2019, the Chalmers team showed that the presence of SETMAR V1 could be correlated with changes in the transcription rate of expressed genes in the human osteosarcoma cell line U2-OS. Although it is not clearly established whether this is an absolute requirement, this transcriptional effect would be strongest at genes containing at least one *Hsmar1* or MADE ITR [22]. The same team found afterwards that SETMAR V1 is not involved in illegitimate DNA recombination and non-homologous end joining repair in U2-OS cells [23]. In light of the fast-growing set of literature and the regular updates of public databases, it is clear that the current understanding of the roles and functions of SETMAR, first assessed in these pioneering studies, relies on postulates based on limited knowledge and suboptimal experimental designs ultimately leading to biased conclusions. Below, we outline several points regarding these issues.

First, functional studies have, so far, focused on the largest SETMAR isoform (V1) while there are several other isoforms. At the NCBI portal nine protein-coding SETMAR transcripts (Fig. 1b, Table S2 and File S1) are described from cancerous tissues and from cell lines [18,24-28]. The Ensembl and GTEx portals describe a total of eight transcripts found in healthy tissues (Fig. 1c) and their expression levels (Fig. 1d), only three of them being shared with cancerous tissues and cell lines (V1, V2 and V4 (a.k.a. the mammal SET)). To our knowledge, the diversity of SETMAR proteins in both healthy and cancerous patient biopsies is not known. There is no published data indicating whether the V1 protein isoform is expressed either alone or together with other protein isoforms in healthy tissues. The V1 isoform has only been found in six cancer cell lines [25,27].

The second issue is related to the evolution of the SETMAR sequence. In a previous study [2], the ratio of non-synonymous (dN) to synonymous (dS) nucleotide substitutions was used to support the fact that the SET domain and the Tpase DNA-binding moiety of the HSMAR1 domain are under purifying selection while the catalytic Tpase domain is drifting [22,23]. Because improved SETMAR sequence descriptions are now available (both in terms of diversity and accuracy), we have updated this important point and confirmed that SETMAR is under purifying selection (dN/dS = 0.21789 (<<1)). However, we also found that three domains are under strong purifying selection: the SET and catalytic Tpase domains display similar dN/dS ratios (0.30394 and 0.27193 (<<1), respectively) and the Tpase DNA binding domain has an extremely low ratio (0.01788 (<<1)). As a result, the three SETMAR domains are certainly biologically active and should display functional properties.

The third issue is related to the fact that if a transposase has been co-opted and is currently under selection it is likely that the corresponding DNA binding sites (BS) are also under selection and may acquire novel cellular function(s). In the case of *setmar*, its emergence after the split of prosimian and anthropoid lineages occurred concurrently with a *Hsmar1*-derived miniature TE (MADE1; 80 bp), and subsequently amplified by transposition to several thousands of copies in the ancestral genome of the anthropoid lineage. These MADE1 sequences have been maintained in the genomes of all current anthropoid species, including humans. Therefore, investigating the function(s) of SETMAR in the human genome also requires that binding target sequences be described, including their sequence conservation and their distribution relative to genes. The sequence properties of MADE1 bound by individual SETMAR isoforms are still an unresolved point. For all these reasons, profiling SETMAR protein content in healthy cells, tumour cells and cancer cell lines, is a clear prerequisite to further functional characterisation.

To address these points, we first profiled the expression of SETMAR isoforms at the protein level in various colon materials. We unambiguously showed that the largest SETMAR isoform (V1) is not expressed, but shorter isoforms with a spliced out SET domain are. Second, we revisited the annotation of the human genome for the *Hsmar1* element and its associated 80 bp miniature inverted-repeat transposable elements (MITEs), MADE1, using BLAST+ and logol [29]. We discovered 2604 novel MADE1 copies and improved the human genome annotation, which proved crucial for defining the sequence properties of SETMAR BS at high resolution and for showing that MADE1 copies are statistically depleted in gene-rich regions. Third, our chromatin immunoprecipitation sequencing (ChIP-Seq) results revealed that SETMAR isoforms V2, V3 and HSMAR1 bind to several thousand BS and

occur mostly at MADE1 sequences. However, a significant fraction of them (up to ~50%) were also found in regions that did not contain *Hsmar1*/MADE1 ITR, but an inner ITR motif. Together our results provide robust evidence of the functional interactions between SETMAR isoforms and *Hsmar1*/MADE1 ITRs. Although the role of these isoforms in healthy colon cells and in cancerous cells remains elusive, our work clearly shows that it does not involve the histone methylase activity of the SET domain.

2. Materials and Methods

2.1. Culture of cell lines

HeLa cells, human colorectal cancer cell lineages (SW48 and HT29) were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% foetal bovine serum, at 37 °C and with 5% CO_2 as described [27]. The sources of the cell lineages were those of the EA GICC 7501 (CHRU de Tours, 37044 TOURS Cedex 09) which acquired in 2009 them from the ATCC (HeLa (ATCC® CCL-2); SW48 (ATCC® CCL-231™; HT-29 (ATCC® HTB-38™)). They were HeLa cell transfection with 1 µg per 100,000 cells of pVAX-HSMAR1 or pVAX-SETMAR isoforms DNA was carried out as described [27]. All used pVAX plasmids was presenting the same Kozak' s box around the translation initiation codon and the nucleic acid sequence of each isoform was that available in hg38.

2.2. Non-tumour and tumour colon tissue samples

Two samples of colon tissues, tumours and adjacent non-tumour tissues, were recovered from patients after surgery for a colorectal cancer in 2007 or 2008. Samples were stored at -80 °C at the tumour bank of the Tours Centre Hospitalier Régional Universitaire (CHRU). The tissue collection was declared to the French Ministry of Research and High Education (n°DC2008-308). Patients were informed of the possibility that samples could be used for research and their agreement to participate in this research was collected. Tumour tissues were phenotyped by a pathologist member of the tumour bank of the Tours CHRU. Samples were selected to contain more than 50% tumour cells (Table S3).

2.3. Protein extraction from non-tumour and tumour colon tissues

Tissue sections (10 µm) were generated with a cryomicrotome on frozen samples of tumour and non-tumour tissues, and stored at -80 °C. For each patient, protein lysates were made by suspending a tenth of the sections in iced RIPA buffer (20 mM Tris-HCl pH7.2, 150 mM

NaCl, 1 mM EDTA, 1% Glycerol, 1% Triton X100, 0.5% doxycholate, 0.1% SDS, 1X-Complete Protease Inhibitor Cocktail (Roche Applied Sciences, Meylan, France)), vortexed for 1 min, incubated on ice for 15 min, and centrifuged at 15,000 g for 15 min at 4 °C. After recovery of supernatants, proteins were quantified with a Quick Start™ Bradford Protein Assay (Bio-Rad, Richmond, USA) and conserved at -20 °C.

2.4. Antibodies

As previously reported [27], the quality of antibodies is crucial to study the expression of domesticated transposases, including SETMAR. The commercial anti-SETMAR polyclonal antibodies (pA) ab3823 (Abcam) was used for Western blot analysis as described [27]. This pA was directed against the RWQKCVDCNGSYFD peptide that is located at the C-terminal end of the HSMAR1 moiety of SETMAR. Because there is no suitable anti-SETMAR or anti-HSMAR1 commercial antibody for ChIP analyses, a murine pA was produced using DNA vaccination technology ICANtibodies™ (In Cell Art, Nantes, France). This pA was obtained using a pVAX mammal expression plasmid coding the complete HSMAR1 moiety of SETMAR [27]. Such pA has been reported to be effective for Western blots, ELISA, cells spread on slide [27], and here for ChIP. For this study, new batches of murine pre-immune and anti-HSMAR1 polyclonal sera were produced. These were validated as described [27] and displayed the same properties as previously reported batches [27]. pA contained in pre-immune and anti-HSMAR1 murine sera were purified using Protein A/G MagBeads (GenScript, Piscataway, USA). Their quality was verified by polyacrylamide gel electrophoresis (PAGE) after staining with Coomassie blue and their concentration was defined using the BCA Protein quantification Kit (Interchim, Montluçon, France).

2.5. PAGE, immunoblotting, and hybridisation of antisera

Protein extracts from cultured cells were separated on PAGE and transferred onto a polyvinylidene difluoride (PVDF) membrane, and antibody incubations and imaging with a FUJI LAS4000 imager were carried out as described [27].

2.6. ChIP-Seq experiments and analyses

Chromatin samples were prepared from non-synchronous and exponentially growing SW48 and HT29 cells. Chromatin shearing was performed with a Bioruptor ultrasonicator (Diagenode, Ougrée, Belgium). Chromatin immunoprecipitation was performed with 10 μg

of purified pre-immune or HSMAR1 pA, and purification of immunoprecipitated DNA was done using the iDeal ChIP-Seq kit following the supplier's recommendations (Diagenode). Libraries for Illumina sequencing were made using iDeal ChIP-Seq & Library Preparation Kit (Diagenode). DNA quantities were monitored at various steps in the procedure with the Qubit® dsDNA HS Assay Kit (Molecular Probes, Eugene, USA). Fragment size selection, library quality control and Illumina sequencing (HiSeq 51 nucleotides, TruSeq SBS Kit v3) were achieved at the Plateforme de Séquençage Haut Débit I2BC (Gif-sur-Yvette, France), following published quality recommendations [30]. Data published in this paper are based on three biological replicates.

ChIP-Seq sequence reads were mapped to the human genome assembly hg38 (December 2013; available at http://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26) with the Bowtie short read aligner [31]. Peak calling was done with two tools from bam files and normalised over input (three biological replicates), the peak-calling prioritisation pipeline PePr1.1.16 [32] and Callpeak from package MACS2 [33]. PePr was used to call peaks in order to benefit from the statistical power linked to the use of replicates, while MASC2 was used to call peaks with files resulting from the fusion of replicates for exploiting the power of sequencing depth. For both tools, a q-value threshold of 5.10^{-2} was used. Other parameters were those per default for PePr1.1.16 while for MACS2, we used a mfold lower bound = 10, mfold upper bound = 30. In final results, we fused those obtained with each of both peak-calling tools. Intersections between ChIP-Seq peaks and the *Hsmar1* or MADE1 copies were calculated using bedtools.

2.7. Computational analyses

Annotation of *Hsmar1* and MADE1 copies in hg38 is described in File S2. Searches for conserved motifs were carried out with the MEME suite (MEME-ChIP and GLAM2) [34] and the RSAT pipeline [35,36]. Searches for specific conserved motifs using a model were done using FIMO [37]. Ontology analyses were done using ClueGo within the Cytoscape package [38]. Annotations files recovered from the literature in an hg18 or hg19 version were transformed in hg38 versions using liftover facilities at https://genome.ucsc.edu/cgi-bin/hgLiftOver.

Data analyses, t-tests and most graphic representation were done using bedtools and Prism 6 package (GraphPad Software Inc). Before t-tests, a F test was done to compare variances of samples and a Welch's correction was used when they were significantly

different (p>0.05). P-values for hypergemotric distribution and chi2 test were computed with Microsoft Excel. Permutation tests (10,000 per test) were computed using shuffleBed (with options -noOverlapping and -chromFirst) to produce samples containing non-overlapping features and intersectBed to count overlaps between two feature files. Probabilities were calculated from Z score at https://www.fourmilab.ch/rpkp/experiments/analysis/zCalc.html.

Annotation files used to analyse the *Hsmar1*, MADE1 and ChIP-seq peak distribution were downloaded from the websites described in the literature and updated in hg38 using liftover at the UCSC website. For that concerning the regions containing at least one replication origin (ORI), we used recent data that split them in two categories in hg38: core origins (annotations) and stochastic ORI [39].

For the substitution rate analyses, an amino acid alignment was first performed using MAFFT [40] Version 7.407. Nucleotide sequences of protein-coding genes were extracted and aligned by using PAL2NAL [41] Version 14 according to the corresponding amino acid alignment. A phylogenetic tree was constructed using the protein sequence alignments with RaxML [42]. To estimate the selective pressure on SETMAR domains, we estimated the ratio (ω) of the rate of nonsynonymous substitutions to the rate of synonymous substitutions using PAML [43]. We used different parameters in the CODEML control file: CodonFreq = 2 for estimating the codon frequencies using F3X4 model, runmode = 0 for evaluation of the tree topologies and model = 0 for a single dN/dS value across all branches, model = 0 is for one omega ratio for all sites. The accession numbers of the used sequences were: XM_012468511.1, XM_008982057.2, XM_017512643.1, XM_012059942.1, XM_007985050.1, XM_011937247.1, XM_004033522.1, NM_006515.3, XM_005547658.2, XM_001099426.4, XM_011734163.2, XM_011965455.1, XM_003831222.4, XM_526121.5, XM_003894164.4, XM_023218587.1, XM_002813499.3, XM_017870728.1, XM_010365759.1, XM_003927130.1, XM_025375280.1.

For RNA-Seq analyses, datasets were downloaded from NCBI using the SRA toolkit (2.8.2-1). After filtering the data using FastQC and Trimmotatics, each set of reads was mapped to hg38 using HISAT2.2.1.0. Read alignments were visualised with IGV and quantified with featureCounts.

3. Results

3.1 Expression of SETMAR isoforms in colon samples

The protein profile of SETMAR isoforms in tissues is largely unknown and highly variable

from one cell line to another. Western blot analyses showed that V1, V2, V3 and HSMAR1 isoforms each display apparent molecular weights consistent with their expected value (Fig. 2a). We observed SETMAR isoform profiles in three tissue types: non-tumour and tumour colon biopsies from 26 patients, and two colorectal cell lines (SW48 and HT29). In healthy and tumour colon tissues (Fig. 2b) we found that only two SETMAR isoforms were present and displayed apparent molecular weights expected for the V3 and V6 isoforms (55 and 49,7 kDa respectively; Fig. 1), with V6 being the most abundant. The V3 isoform was only expressed in a few samples, and always at low rates. This protein profile is markedly different from the *setmar* transcript profiles available at Ensembl and GTEx portals, where transcripts coding the V1 were expected in this tissue (Fig. 1c,d; V4 could not be detected with the antibody used). This suggested that two inner translation initiation sites were used in V1 transcripts to produce V3 and V6, or that they are not translated.

Three protein isoforms with apparent molecular weights corresponding to V2, V3 and HSMAR1 were found in SW48 cells, while only one (with an apparent molecular weight matching that of HSMAR1) was found in HT29 cells (Fig. 2c). An analysis of transcripts diversity, by using the combined RNA-Seq datasets SRR6368612 to SRR6368614 for HT29 cells, and the combined datasets SRR7108292 to SRR7108339 for SW48 cells, revealed that transcripts corresponding to V1, V2, V3, V6 and setmar-003 isoform models might be expressed in HT29 cells, and two transcripts corresponding to V1, V2, V3, and V6 were found in SW48 cells (Fig. S1). On a side note, it is difficult to match protein levels to transcript levels, but such lack of stoichiometric correlation between the amount of mRNA and the protein level is a commonl feature, found for ~50-60% of transcribed genes [44-51]. Overall, the stark contrast between protein profiles in colon tissues and cell lines shows that SETMAR expression is highly cell-specific and very sensitive to the cellular context.

3.2. Features of Hsmar1 and MADE1 in the human genome

An in-depth description of SETMAR V1 binding to the *Hsmar1*/MADE1 ITRs *in vitro* [10] and to certain chromosomal ITRs in U2OS cells [22] critically relies on accurate and thorough TE annotations. Indeed, functional examination of genomes relies heavily on high quality annotations, which are often far from optimum in the case of TEs. Therefore, we undertook to generate an improved annotation of *Hsmar1,* MADE1 elements and their ITRs in the hg38 genome. We characterised their conservation and distribution in different structural and functional components of the genome.

3.2.1. Improved Hsmar1 and MADE1 annotation.

Annotation of TEs is notoriously difficult to carry out, especially when these are very small such as MADE1 (80 bp) that contain two short (24-bp long) ITRs. We verified and improved the quality of *Hsmar1* and MADE1 annotations generated by RepeatMasker using logol and BLAST+. Methods and results are summarised in File S2 and our revised annotation is provided in File S3a. In summary, the revised annotation revealed the presence of 519 *Hsmar1* copies (composed of 615 contiguous and split annotations and 609 ITRs that were already annotated by the RepeatMasker annotation), along with 10374 MADE1's displaying a total of 9806 ITRs (lacking less than 5 nucleotides) and composed of 2312 full-length elements, 5181 elements truncated at one end (displaying a single ITR, i.e. a single SETMAR binding site of 24-bp), and 2881 elements with damaged ITRs (outer ends lacking ≥ 4 nucleotides at each ITR).

Based on this updated annotation, we first measured the sequence conservation of *Hsmar1* and MADE1 copies to their consensus sequence to estimate their degree of divergence. We found that the *Hsmar1* copies had more sequence identity to their consensus than MADE1 copies. Among MADE1 copies, those that possess one or two ITRs were more conserved than copies with no ITRs (identity mean 90.52% ± 0.04338 vs 87.17% ± 0.1660, respectively; t-test with Welch's correction, $p<0.0001$). This suggested there were two populations of MADE1; one highly conserved and displaying ITRs, and one less conserved, devoid of ITRs and accumulating more mutations. Interestingly, *Hsmar1* copies are even more conserved (identity mean of *Hsmar1* fragments = 91.39% ± 0.1557) than MADE1 copies (t-test with Welch's correction, $p<0.0001$). Sequence conservation was further investigated at the level of their ITR sequence. The average conservation of *Hsmar1* ITRs (91.41 ± 0.180) was lower than that of MADE1 ITRs (92.23 ± 0.040; t-test with Welch's correction, $p<0.0001$). For MADE1, our results suggested that at least one of the two ITRs in MADE1 accumulated fewer mutations than the rest of the sequence or that MADE1 ITRs were under selection. It should be noted that there was a little subpopulation of *Hsmar1* and MADE1 with at least one ITR which displays 99-100% of identity to their consensus sequence. Our downstream analysis focussed on *Hsmar1* and MADE1 copies harbouring at least one ITR.

3.2.2. Distribution of Hsmar1 and MADE1.

Description of the landscape of *Hsmar1* and MADE1 elements was performed at different

scales: chromosome, intra and intergenic regions, lamina associated domains (LAD) [52], topological associated domains (TAD) [53-55], and at origins of replication [39,56-58] depending on the local mutation rate [59] (File S2b). Results showed that 1) *Hsmar1* and MADE1 elements with at least one putative SETMAR binding site in TIR sequences were not distributed at random in the human genome, and 2) they accumulated mutations with rates depending on the evolutionary age of the chromosomal segments in which they were located. These results also showed that *Hsmar1* and MADE1 are enriched in intergenic regions contained in TADs and those in constitutive LAD (cLAD) and constitutive inter-LADs (ciLADs) regions. No link was found between the location of *Hsmar1* copies and MADE1 location and origins of replication. Finally, we speculate that gene expression interference by SETMAR using *Hsmar1* and MADE1 ITRs was not from promoter regions but rather from enhancers contained within introns and intergenic regions. The enrichment in Hsmar1 and MADE1 copies within enhancers was investigated using two annotation sources (File S3g,h). The first is composed of consensus enhancers (259801 in hg38) predicted based on multiple high throughput experimental datasets (e.g. histone modification, CAGE, GRO-Seq, transcription factor binding and DHS) and is available at http://www.enhanceratlas.org/indexv2.php [60]. The second, Fantom5, is available at http://slidebase.binf.ku.dk/human_enhancers/ and is a sub-database of enhancers (32689 in hg38) predicted from CAGE data, i.e. with an activity reflected by the presence of non-coding enhancer RNA (eRNA) [61]. Permutation tests revealed that there was no enrichment or depletion of Hsmar1 or MADE1 with at least one ITR within Fantom5 enhancers and those common to both annotations ($p = 0.435$, 0.293, 0.064, and 0.084, respectively). In contrast, intact MADE1 elements or MADE1 elements with at least one ITR, were significantly enriched among non-Fantom5 enhancers ($p = 0.000126$ and 0.0185, respectively). Interestingly, we observed that MADE1 elements were neither enriched nor depleted among Fantom5 enhancers ($p = 0.293508$), but those with at least one ITR were significantly depleted ($p=0.001015$). Altogether, we can thus estimate that 383 MADE1 elements located at non-Fantom5 enhancers (383 = 6632 annotated MADE1 – 5949 expected by chance; see permutation tests in methods) might be involved in enhancers when they were bound by all or some SETMAR isoforms.

3.3. SETMAR binding sites (BS) along human chromosomes in two colon cell lines

SETMAR chromosomal targets were identified by ChIP-Seq in HT29 and SW48 cell lines, using a custom-made polyclonal antibody directed against the HSMAR1 domain of SETMAR.

The specificity of this antibody has been addressed previously and is characterised by a good specificity/sensitivity ratio [27]. Peak calling, independently carried out with PePr and MACS2, revealed 6280 and 5059 peaks in HT29 cells, and 8097 and 11059 peaks in SW48 cells (with a false discovery rate (FDR) below 5%, File S3i-l). In HT29 cells, 4028 peaks (60%) overlapped with 4360 *Hsmar1* and MADE1 annotations while 4375 peaks (37%) overlapped with 4786 annotations in SW48 cells. In HT29, 3860/4028 *Hsmar1* and MADE1 were annotated as having at least one ITR and 4191/4786 in SW48. This was partly due to the stringency of our definition to annotate the presence of one ITR with respect to that of a SETMAR binding site. If more than 4 nucleotides were missing at the outer end, an ITR was considered as being absent. In terms of overlap with *Hsmar1* and MADE1 annotations ('HM peaks'), each peak-caller provided similar results (Fig. 3a). However, these programs strongly diverged calling peaks that did not co-localise with *Hsmar1* and MADE1 annotations ('noHM peaks', Fig. 3b). PePr and MACS2 files were combined into a final non-redundant dataset containing 6699 and 11724 peaks in HT29 and SW48 cells, respectively (examples of peaks differentially detected by both methods are supplied in Fig. S2). HM peaks displayed significantly higher width when compared to noHM peaks (Fig. 3c; t-test with Welch's correction, $p<0.0001$ for both cells), as well as higher normalised read enrichments (NRE, Fig. 3d; t-test with Welch's correction, $p<0.0001$ for both cells) and lower FDR (Fig. 3e; t-test with Welch's correction, $p<0.0001$ for both cells). This analysis clearly identifies two SETMAR binding peaks populations, 1) peaks overlapping with *Hsmar1* and MADE1 elements (HM peaks) associated with a strong and robust binding, and 2) other peaks (noHM peaks) associated with much weaker binding.

Similarly, we found that almost all *Hsmar1 and* MADE1 copies displaying canonical ITR (99–100% identical to the consensus; 24 out of 25, and 90 out of 91, respectively) were bound by SETMAR in both cell types (black ellipse Fig. 4a). This is consistent with a model of SETMAR binding to canonical DNA sequences where chromatin accessibility is the limiting factor, irrespective of the genomic location. Sequence conservation of SETMAR bound to *Hsmar1* ITR is 4.5% higher than that of unbound ITRs (t-test, $p<0.0001$, 94.49 ± 0.249, $n=171$ versus 89.84 ± 0.191, $n=335$, respectively). This is also true for bound MADE1 ITRs, although the conservation level is slightly lower (t-test with Welch's correction, $p<0.0001$, 93.66 ± 0.047, $n=4210$ versus 90.39 ± 0.057, $n=3283$ respectively). There is virtually no SETMAR binding at ITRs with ~85% or lower sequence conservation. We also found no significant correlation ($R^2 <0.05$) between peak NRE or peak q-value with the conservation level of bound ITR, further strengthening the view that chromatin configuration might be the limiting step for SETMAR binding.

3.4. Features of conserved motifs covered by HM peaks

A significant fraction of HM peaks (595 peaks) overlapped with *Hsmar1* and MADE1 annotations, although they did not display any predicted SETMAR binding sequences. This raises the question of the diversity of SETMAR binding sequences. We ran the *de novo* motif discovery programs RSAT, GLAM2 and MEME-ChIP on three datasets: 1) *Hsmar1* and MADE1 copies with at least one canonical ITR (this dataset acted as a positive control), 2) *Hsmar1* and MADE1 copies with no known predicted SETMAR binding sequence, and 3) unbound *Hsmar1* and MADE1 copies. We found a single overrepresented motif (e-value $<1.10^{-100}$, with RSAT, GLAM2 and MEME-ChIP), corresponding to *Hsmar1* and MADE1 ITR. This analysis also showed that bound ITRs were more conserved than unbound ITRs (Fig. 4b), and that the sub-motif corresponding to the SETMAR binding site [2] (Fig. 4b, boxed with a black line) was the most conserved but only in the ITR from nucleotides 6 to 24. This can be further restricted to a 12 bp motif showing the highest conservation level overall (Fig. 4b, boxed with a purple line). The second half of this motif (CTTTTG) is located exactly at the center of the ITR. In addition, *Hsmar1* and MADE1 copies devoid of predicted canonical SETMAR BS displayed a mutated sequence, with a C nucleotide insertion at position 11. Unbound ITR sequences harboured a CT dinucleotide at positions 11 and 12 which may prevent SETMAR binding.

3.5. Motifs and sequence features of noHM peaks

We carried out similar motifs searches with sequences overlapping noHM peaks and found a statistically significant overrepresented short motif, CTTTTG, present in numerous sequences. Since this 6 bp motif is part of the most conserved core motif (12 bp) of ITRs (see above), we undertook a systematic survey in order to verify its presence in noHM peak sequences by deriving a position weight matrix (PWM) of the 12 bp motif based on MEME-ChIP results [34], and using FIMO [37]. At a p-value threshold below 0.01, this core motif could be found in the vast majority of binding peaks: 2152 (15312 core motifs; 7.1 motifs per peak) out of 2671 peaks (80.6%) in HT29 cells, and 6400 (29600 core motifs; 4.6 motifs per peak) out of 7348 peaks (87.1%) in SW28 cells. At a p-value threshold below 0.001, we found a conserved ITR core motif in 1023 (1978 core motifs; 1.9 motifs per peak) out of 2671 peaks (38,3%) in HT29 cells and 1808 (2498 core motifs; 1.4 motifs per peak) out of 7348 peaks (34.0%) in SW28 cells (consensus motifs are shown in Fig. 5a). These results show

that SETMAR binds to DNA not only at *Hsmar1* and MADE1 ITRs with high affinity, but also at other genomic sequences that contain the conserved core sub-motif CTTTTG (found in *Hsmar1* and MADE1 ITR), albeit with lower affinity.

We also found that noHM peaks are preferentially distributed within GC-rich regions, in contrast to HM peaks (0.445 ± 0.001, n=2671, in HT29 cells, and 0.538 ± 0.001, n=7348, in SW48 cells for noHM peaks versus 0.391 ± 0.001, n=4029, in HT29 cells and 0.404 ± 0.001, n=4376, in SW48 cells for HM peaks). This distribution bias is highly significant (t-test with Welch's correction, $p<0.0001$, Fig. 5b,c). Active replication origins are often located within GC-rich regions, or near CpG islands and G-quadruplex (G4) sequence motifs [56-58]. This raises issues regarding SETMAR binding within the noHM regions that would be located within or near replication origins.

3.6. Distribution of SETMAR binding sites in the human genome

The number of HM peaks found in both cell types had a linear relationship with chromosome size (R^2 = 0.9136 and 0.9145, respectively) and the number of annotated *Hsmar1* and MADE1 copies (R^2 = 0.9862 and 0.9892, respectively) (Fig. 6a,b). This is in striking contrast with no HM-peaks (Fig. 6c; R^2 = 0.1013 and 0.0052), which are overrepresented (two-fold) in chromosomes 7, 8, 17, and 20 in HT29 cells and in chromosomes 7, 11, 12 16, 17, and 20 in SW48 cells. They are also underrepresented (two-fold) in chromosomes 4, 6, 14, 18, 22 and X in HT29 cells and in chromosomes 2, 3, 4, 5, 6, 9, 14, 15, 18, X and Y in SW48 cells.

On a smaller genomic scale, and as expected from our previous results, intergenic regions are consistently enriched in HM peaks when compared to intragenic regions (Fig. 6d,e; R^2 values ranging from 0.8485 to 0.9627, respectively). HM peaks were significantly underrepresented in chromosomes 4, 6, 9, 14, 18 and X in both cells (Fig. 6f, green bars), while they were overrepresented in chromosomes 7, 8, 17 and 20 (Fig. 6f, grey bars). They also revealed important differences in intragenic and intergenic regions between HT29 and SW48 cells in chromosomes 10, 12 and 16 (Fig. 6f, yellow bars).

These observations show that the genomic distribution of SETMAR binding peaks belongs to two sub-populations of peaks: I) high affinity peaks closely matching the location of well conserved *Hsmar1* and MADE1 ITR and mostly located in intergenic regions, and ii) lower affinity peaks matching only a core sub-motif of ITRs, with a more contrasted distribution between intra and intergenic regions, depending on the chromosome and the cell type.

3.6.1. HM peaks.

With regard to the various gene components, including promoter sequences and 3′ UTR (over a 3 kb genome span), peaks were significantly underrepresented (by ~10-30%) only in long non-coding RNAs (lncRNAs), introns of protein coding genes, microRNAs and uncharacterised genes ($p<0.05$; hypergeometric distribution). We found that the occupancy rates of HM peaks as a function of available ITRs were more elevated in inter-LADs (iLAD) (Fig. 6g, grey whisker boxes) and in TAD boundary regions (Fig. 6h, grey whisker boxes) than in LAD and nbTAD (Fig. 6g,h, whisker boxes in black). This indicates that SETMAR binding at ITRs occurs mostly at chromatin regions surrounding genes and with open chromatin configuration (i.e. defined as corresponding to both iLAD and TAD boundaries). Investigations of gene ontology containing HM peaks did not yield significant results in either cell lines, but permutation tests revealed, in both cell lines, that HM peaks were enriched in enhancers ($p < 10^{-100}$).

3.6.2. noHM peaks.

Low affinity binding peaks stood in stark contrast to the preceding results, they were more abundant in intragenic regions, with 75% of them located in protein-coding genes. Ontology analyses showed no clear trend in biological processes, pathways or molecular functions (1058 genes in HT29 cells, and 3372 genes in SW48 cells, Fig. S3b). Analysis of their distribution in LAD and TAD regions (Fig. 6i to l) revealed a profile similar to that of HM peaks, i.e. more noHM peaks than expected in iLAD regions and TAD boundaries in SW48 cells and to a somewhat lower extent in HT29 cells. noHM peaks were located in GC-rich regions located in iLAD and TAD boundaries. Because such regions have been previously described as corresponding to replication origins, we verified whether or not they displayed any enrichment in noHM peaks [39,56-58,62,63]. We found that regions containing both kinds of replication origin, constitutive and stochastic, were enriched in noHM peaks: 10 and 3 folds for constitutive and stochastic replication origins in SW48 (Fig. 7a), respectively, and 3 and 2 folds in HT29 cells, respectively (Fig. 7b). We also found that their distribution was not uniform between chromosomes. In both cell lines, they were under-represented in chromosomes 2, 3, 6, 9, 14, 18, 22 and X and over-represented in chromosomes 7, 20 and Y (Fig. 7c and 7d). Interestingly, the distribution in chromosomes 4, 5, 8, 15 and 19 is cell type specific. Given that ~76% of CpG islands (CGI) are located at constitutive origins of

replication origins (i.e. core replication origins) [39,56-58], we verified that they were enriched in noHM peaks (File S3k). We found 310 noHM peaks colocalizing with 310 CGI in HT29 cells and 2959 peaks colocalizing with 2959 CGI in SW48 cells. These results were highly significant for SW48 cells (permutation test, $p < 10^{-100}$), and significant in HT29 cells at a 5% threshold ($p = 0.033$).

Consistent with results above regarding the distribution of *Hsmar1* and MADE1 copies at the vicinity of replication origins, HM peaks were found to be depleted in these regions. Given that GC-rich regions and CpG islands can also be found near or upstream of promoters and contained enhancers, permutation tests were done to verify whether noHM peaks were enriched in these elements, considering that 25651/27708 CGI colocalized with enhancers ($p < 10^{-100}$). In HT29, 1976 peaks colocalized with an enhancer while there were 5516 in SW28 ($p < 10^{-100}$). These peak numbers could not be compared with those obtained above because the CGI selection for these analyses was very stringent and therefore did not allow us to define to which element SETMAR was binding to.

4. Discussion

4.1. Relevance of SETMAR V1 isoform in healthy and cancer cells

A major bottleneck for dissecting the function of SETMAR in terms of human physiology and physio-pathology is the lack of biological models. Currently, there is no simian model in which gene invalidation can be performed. Furthermore, the complexity of the SETMAR/*Hsmar1*/MADE1 system prevents its synthetic reconstruction in mouse models. Finally, no structural and sequence polymorphisms at *setmar* have been so far linked to human diseases.

Our initial results extended previous observations [44-51] where mRNAs synthesised *in vivo* from the *setmar* gene were not indicative of the protein isoforms that are actually expressed in healthy and tumour tissues, and in cell lines. Our first key result is that the longest SETMAR isoform, V1, which has received the most of attention so far, is not expressed in healthy and tumour colon tissues. This result, based on a cohort of 26 patients with colorectal cancer. So far, V1 has only been detected in a few cell lines derived from various tumours, most of which do not express V1 [25]. This suggests that V1 expression might result from phenotypic drift, a well-known fact where cancer cells (and cell lines derived from them) hijack the regulatory mechanisms governing transcription, splicing and translation [47,64]. The question of how general V1 expression is among cancer cells and healthy

tissues, and whether it is idiosyncratic of some cancer and healthy cells is still an open question that will clearly deserve a more systematic survey. In fact, an absence of V1 in healthy tissues and the existence of shorter isoforms with a damaged SET domain unable to display histone methyltransferase activity is something rather expected. Indeed, the human *set* gene (from which the V4 mRNA is synthesised) and its murine *Etet2* ortholog display similar expression profiles (Fig. 1 versus data available at https://www.mousephenotype.org/data/genes/MGI:1921979#phenotypesTab) and are very likely functional analogues. The invalidation of *Etet2* in mice results in diverse effects, including fat tissue content and glycerol metabolism, vertebral column development, and behavioural control at the central nervous system. In contrast, the functional output of SETMAR transcripts is much less clear. Artificial overexpression of SETMAR V1 in U2OS cells has a massive transcriptional impact (~8890 genes with fold change 2<or>2 [22]), although this effect is not direct and does not involve changes in H3K36me2 levels at target genes. This casts further doubt on the putative involvement of the SETMAR histone methyltransferase activity. Therefore, the function, if any, of the extra-human specific SETMAR transcripts remains a matter of debate.

The in-depth description of alternative transcripts and protein isoforms we have provided here will be a key resource to understand the biological activity associated with SETMAR. With regard to the mode of action, it is clear that if SETMAR alters the expression profile of several genes in colorectal cells [21,22], this effect cannot be mediated by the V1 isoform but rather with the SET-less isoforms V2, V3, V6 and/or HSMAR1. This fact has two important consequences: 1) the protein domain containing biological activity involved in the modulation of gene expression is almost certainly located at the C-terminal of SETMAR, and 2) an artificial construct overexpressing the *setmar* gene coding a mutated version of the V1 isoform (as carried out in [22] with the N210A mutation, which suppresses the histone methylase activity of the SET domain) certainly does not indicate that the N-terminal domain carries the biological activity involved in changes of gene expression, and the resulting phenotype does not reflect the physiological/physio-pathological mechanisms. Our interpretation is that the N210A mutation may result in a misfolded protein, thereby preventing any biological activity associated with the C-terminal domain of V1.

Future investigations aimed at dissecting the role(s) of *setmar* in human physiology and physiopathology will require two essential observations. First, it is absolutely necessary to precisely describe the repertoire of HSMAR1 and-or SETMAR isoforms beside V1. Indeed, information available in public databases on translation initiation sites [65] and mass

spectrometry ([66] and https://www.proteomicsdb.org/) suggest that the diversity of SETMAR isoforms might be even higher than what can be inferred from transcriptomic analyses. The second point will be to profile the expression of SETMAR isoforms in diverse healthy and tumours tissues, and to address whether the SET moiety is a catalytically active histone methylase in SETMAR, or just a chimera of SET subdomains facilitating interactions with other proteins. Our current understanding of the SETMAR functional properties is very limited as it is only based on the correlation between the observed versus expected (from transcriptomic data) molecular weight, and has not been addressed experimentally.

4.2. Two kinds of SETMAR binding sites

The structural and functional characterisation of the repertoire of SETMAR BS is critically dependent on accurate genome annotation. Unfortunately, as opposed to genes, annotation of transposable elements, which is technically complex and time consuming, often receives much less attention and is frequently incomplete. It is therefore not uncommon for experts in a family of TEs to curate and incrementally improve existing annotations. Here, we describe 4080 new *Hsmar1* and MADE1 ITR annotations, extending the official RepeatMasker annotation set of 6334 copies by 64.4%, leading to a final dataset of 10414 ITR. This proved critical as numerous SETMAR binding peaks colocalise with this extended dataset (see below).

The first set of BS is located at MADE1 and *Hsmar1* ITRs. Within these BS, sequence variability is such that we could identify a minimal 20 bp sub-motif, which can be further split into two components: 1) a highly conserved 12 bp core (CAATTACTTTTG), surrounded by 2) somewhat more variable residues. The second set of BS is restricted to the core 12 bp motifs, and displays overall more sequence variability. Given their differences in sequence conservation level and ChIP-Seq features (peak height), we propose that these two sets of BS correspond to high and low affinity SETMAR BS (although without additional biochemical evidence, they could equally be labelled as ‘stable’ and ‘less stable’ BS). This is reminiscent of the *Sleeping Beauty* (*SB*) transposase, which belongs to the same ITR transposon family of *mariner*-like elements (i.e. *Hsmar1*) [1], and which also binds two kinds of sequences: *SB* ITR and a 12 bp ITR sub-motif (GTTTACATACAC), although the latter tolerates relatively high sequence variability [67]. Therefore, SETMAR and *SB* binding features most certainly reflect generic properties of the *mariner*-like elements transposases. Conventional biochemistry of DNA-protein interactions (EMSA, footprint, etc.) failed to identify these two families of BS, which emphasizes the importance of taking into account the molecular

context (and especially chromatin state/modification and/or accessory factors) when defining the transposase binding properties *in vivo*.

4.3. Artificial systems affect genome wide SETMAR binding profiles

A previous study in U2OS cells overexpressing an N-terminal flag-tagged version of the V1 SETMAR isoform [22] also showed binding at the high and low affinity BS, although the total number of BS is much lower (875 total; 7.65 and 13.40 times less than in HT29 and SW48 cells). They also identified a third motif corresponding to the binding sequence of the centromeric protein CENP-B. Of note, the vast majority of BS (64.5%) in this study did not match with these three sequence motifs, leaving the majority of them uncharacterised. Overall, the results in this study [22] differ from our work in several ways. First, BS profiling was carried out by ChIP-exo-Seq, arguably a gold standard to precisely identify sequence motifs at protein BS [68]. Unfortunately, the exonuclease treatment can lower the signal intensity of numerous binding sites, thus restricting the overall dataset [69]. Second, U2OS cells are derived from osteosarcoma (bone lineage) and for which, to our knowledge, there are no RNA-Seq data available in the public database on simian bone cells. As a result, the expression level of the endogenous *setmar* is not characterized in this kind of biological material. This is unfortunate because, based on the known binding properties of the RAG1/2 transposase with respect to their RSS and RSS-like binding motifs in lymphoid and non lymphoid cells [70-78], one would predict that expression of co-opted transposases and accessibility to their BS would be tightly regulated. If *setmar* is not expressed in bone cells, only a few of its BS would be accessible, thus preventing a description of the full binding site repertoire. The third point is that the construct expressing the SETMAR V1 isoform used to genetically engineer U2OS cells was calibrated from transcript rates observed in non-modified U2OS cells. Because the expression level of the exogenous SETMAR was not measured at the protein level (with anti-*Hsmar1* antibody), it is not possible that the experimental setup reflects a non-artificial system. In agreement with this, the fact that SETMAR binding peaks colocalise with CENP-B BS is a clear indication that its expression level is too high. Another strong line of evidence is that 124 out of 875 ChIP-seq peaks obtained in [22] are located within centromeres (14.17%). Since the centromere coverage in the human genome is only ~3%, this indicates that there is a ChIP-seq peak enrichment of 4.72 fold at the centromeres. It is well known that overexpressed nuclear proteins tend to accumulate and bind to DNA at centromeres and pericentromeres. This is well illustrated in the case of the negative elongation factor complex (NELF), a regulator of transcriptional

elongation [79,80]. In the first of these two studies, the authors monitored NELF binding with an anti-NELF-A monoclonal antibody in unmodified human cells lines. The genome wide binding profiles revealed discrete peaks mainly located at transcriptionally poised promoters (Omnibus project GSE53008). In the second study, the authors used an anti-NELF-E polyclonal antibody and genetically modified HeLa cells expressing a recombinant NELF-E. In this case, the dataset was dominated by non-specific binding, with a strong enrichment at centromeres (Omnibus project GSE60586). An even stronger heterogeneity of BS has been described in the case of a human transcription factor, the nuclear receptor subfamily 3 member 1 (NR3A1, so-called oestrogen receptor, ESR1), with various studies based on healthy tissues, cancer lines and artificially reconstructed systems [81]. Qualitatively, these studies consistently identified the consensus sequence of NR3A1 binding site, but quantitatively, the number of BS varied over an order of magnitude between experiments and the overlap between datasets was modest at best. We therefore argue that using artificially reconstructed systems correctly identifies some binding site properties but at the cost of increasing the risk of artefacts, which preclude their use in the context of physiology and physiopathology (i.e. tumours). Therefore, the use of biopsies originating from healthy and tumour tissues, together with suitable anti-*Hsmar1* antibodies, should be favoured instead.

4.4. Binding site distribution and prediction of SETMAR molecular function

Virtually all (>95%) of the SETMAR BS found in HT29 and SW48 cells are located at instances of *Hsmar1*/MADE1 ITR or the core 12 bp sub-motif CAATTACTTTTG. Their genomic distribution supplies indirect but important information to further examine the roles of SETMAR in the general physiology of the cell. The distribution of HM peaks reveals that SETMAR isoforms bind to ITRs displaying a well-conserved sequence (identity>85%). These peaks were located preferentially in non-genic regions of active chromatin including in some enhancers. This suggests that SETMAR effect on gene transcription might be mediated through these enhancers. The effect of SETMAR on gene transcription has already been shown using at least two different approaches [21,22]. Indeed, Tellier's works [22] describes a massive and deep SETMAR dependent transcriptional reprogramming, where 53% (~8890 of the 16776 genes investigated) of genes are labelled as differentially expressed, with 1500 of them displaying |fold change| ≥2. Unfortunately, such dramatic figures often result from technical (unfiltered or uncorrected sequencing bias, normalization issues…) or biological (*e.g.* a few genes capturing millions of reads and distorting the shape

of the reads distribution among genes) issues that need to be properly addressed [82]. Also, differential analysis carried out within the ANOVA framework is only valid when the total number of DE genes is low, typically less than 10% of all genes (*e.g.* in [83]). As such, ANOVA-based differential analysis performs poorly when detecting extensive alterations of the transcriptome, and even failed to detect the massive genome wide amplification of transcription induced by *c-myc* [83], thus making the extent of the SETMAR-dependent transcriptional reprogramming reported by [22] suspicious. Therefore, this is still an open question that requires additional scrutiny.

NoHM peaks displayed a genomic distribution biased toward the two kind of replication origins, with a marked statistical trend in favor of core replication origins compared to stochastic ones. Because replication initiates at different genomic locations depending on the cell type and the replication timing, we could not further investigate the extent of this statistically significant association. Despite being statistically significant, the extent of the proximity between NoHM peaks and replication origins is likely underestimated due to our experimental setup and some intrinsic features of stochastic replication origin. Non-synchronized SW48 and HT29 cells were used in our experiments, together with a somewhat limited sequencing depth. The weakness of this experimental approach is that the DNA replication phase only occurs in a limited fraction of the cell population examined, which reduces *de facto* the sensitivity of our test. In this regard, it is remarkable that despite a sub-optimal experimental setup, we were still able to detect such strong association. The resulting lack of sensitivity is expected to be much stronger for stochastic replication origins, which are variable within and between cell populations [39]. Future improvements in the experimental plan are clearly needed to further evaluate the connection between SETMAR biological activity and DNA replication, but also to define whether SETMAR is implicated in a limited number of replication origins or is involved in the functioning of most of them [84,85].

4.5. Does SETMAR position the DNA repair machinery at replication origins?

The fact that a co-opted eukaryotic transposase such as SETMAR interplays with NHEJ is expected. Indeed, well-studied models of naturally occurring (*P* [86], *piggyBac* [87] and *SB* [88]) and co-opted (*RAG1* [89], *piggyMac* [90]) transposases have been found to interact with Ku70-Ku80 complex, which are early factors triggering and channelling DNA repair through the NHEJ pathway. Given that *SB* (which belongs to *Tc1-mariner* family) interacts with Ku80, it is very much expected that HSMAR1 and SETMAR would also do so and in a similar manner, most likely through their C-terminal domain. Surprisingly, and despite recent

literature describing the functional connections between SETMAR and NHEJ [5,13-16], one group opposes this view and proposes instead that SETMAR has little to no effect on the rate of cell division, exogenous DNA integration in the genome or even is involved in the NHEJ machinery [22]. The experimental setup used by this group is based on U2OS cells expressing SET, MAR or SETMAR proteins containing, or not, the N201A, D432A or D483A mutations in order to suppress methytransferase activity, ITR binding activity and the remaining nuclease activities found in MAR and SETMAR. The first drawback of this study is that it has been clearly established that even small variations in DSB repair capacity between cell lines (and clones) often results in large differences in cell resistance to DNA breaks [91-93]. It is therefore difficult to conclude that there is 'no effect' without proper estimation of the cell resistance to DSB. At most, even if true in U2OS cells, this result certainly does not reflect a general property of SETMAR. The second drawback is that one of the SET-less isoforms (V3 or V6) mediate a robust increase of DNA repair efficiency by NHEJ [28], further supporting our view that the SETMAR V1 isoform is not physiologically relevant in the SETMAR/*Hsmar1*/MADE1 system, and that any conclusion drawn from this isoform is artificial in regards to the role of *setmar* in healthy tissue. Finally, this was supported in a recent article about the SETMAR-H3K36me2-NHEJ repair axis in glioblastoma (Kaur et al., 2020) [94].

Another way to address the existence of V1 in healthy tissue is to ask what the functional and evolutionary advantages of the *set* + *mar* = *setmar* gene fusion would be? As with all gene fusions, the biological activity(ies) encoded by *mar* falls under the transcriptional control of elements located within and upstream of *set*, a process which is reminiscent of molecular specialization (sub-functionalization). Alternative transcripts and translation initiation codons generate a variety of fusion proteins, structurally and functionally connecting the MAR domain with the pre-SET (V2 isoform), post-SET (V3 isoform) or the end of the post-SET domain (V6 isoform). As such, the various SET subdomains might act as docking platforms that differentially bind host factors (such as hPso4 [9,10]) and provide coordinated recruitment of SETMAR activity(ies) to the DNA repair machinery [95-97] at dedicated (SETMAR BS) genomic locations. This model is well supported by the fact that the C-terminal domain in the HSMAR1 moiety, which very likely contains an active interface for the binding with Ku70/80, is well conserved since the divergence of the 21 primate species with sequenced genomes (~44 million years). It should also be noted that the SET domain in the V1 isoform was found to be enzymatically inactive in experiments done in cell culture, likely because the substrate binding pocket is blocked by the protein dimerisation and its binding to DNA [22]. Indeed, the two SET domains in a dimer are close with each

other since the HSMAR1 has kept ability to bind to ITRs and to homodimerize even when its N-terminal domain is fused to another protein [7,98-100].

5. Conclusions

We have provided a repertoire of expressed SETMAR isoforms in healthy and cancer colon tissues, the genomic distribution of predicted BS relative to the different genome components, as well as a genome-wide profile of SETMAR binding. To our knowledge, this is the first comprehensive report of SETMAR expression and binding properties in a non-artificial setup. We also confirmed that there are two kinds of BS for SETMAR isoforms, *Hsmar1* and MADE1 ITRs and an inner 12-bp motif that displays sequence similarities with the inner core of *Hsmar1* and MADE1 ITRs. These two kinds of BS might, respectively, be committed to two SETMAR isoform functions: 1) the expression of certain genes such as MADE1 as previously proposed [21,22] because ITRs are overrepresented in some enhancers, and 2) chromosomal DNA replication as proposed by the Hromas' group [11,12,16,17,21] because NoHM peaks containing 12-bp motifs are enriched in at least some origins of replication.

List of abbreviations

BS, binding sites

ChIP-Seq, chromatin immunoprecipitation sequencing

CHRU, Centre Hospitalier Régional Universitaire

ciLADs, constitutive inter-LADs

cLAD, constitutive LAD

CRC, colorectal cancer

dN, rate of non-synonymous nucleotide substitutions (dN)

dS, rate of synonymous nucleotide substitutions

DMEM, Dulbecco's modified Eagle's medium

eRNA, enhancer RNA

FDR, false discovery rate

HM peaks, peaks that co-localise with *Hsmar1* and MADE1 annotations

hPso4, human psoralen 4 protein

ITR, inverted terminal repeat

LAD, lamina associated domain

MADE1, *Hsmar1*-derived miniature TE

MSI, microsatellite instable

MSS, microsatellite stable

NELF, negative elongation factor complex

NHEJ, non-homologous end joining

noHM peaks, peaks that did not co-localise with *Hsmar1* and MADE1 annotations

NR3A1, nuclear receptor subfamily 3 member 1

Obs, observed

ORF, open reading frame

pA, polyclonal antibodies

PAGE, polyacrylamide gel electrophoresis

PVDF, polyvinylidene difluoride

SB , *Sleeping Beauty* transposase

TAD, topological associated domains

TE, transposable element

Th, theoretical

Declarations

Ethics approval and consent to participate

Tumor samples came from a cohort of colorectal cancers (CRCs) collected in the pathology department of the University Hospital of Tours. The constitution of the collection was approved by the ethics committees of the “Centre Hospitalier Régional Universitaire” (CHRU) of Tours (France). The collection contains frozen and formalin-fixed, paraffin-embedded tumor material, as well as paired normal tissue; written informed consent was obtained from all patients. According to French laws and recommendations, the collection has been declared to the French Ministry of Scientific Research and is registered under N DC-2008-308.

Consent for publication

Not applicable

Availability of data and materials

All raw and processed data are available through the European Nucleotide Archive under accession number PRJEB19196. The files describing the updated annotation of *Hsmar1* and MADE1 copies in the hg38 release and the ChIP-Seq peaks are supplied as File S2.

Competing interests

The authors declare that they have no competing interests

Authors' contributions

AAL, AA, MB, LB, SG, LH, BP and YB did the data acquisition and analyses, CB, NB, IS and YB did interpretation of data, YB designed the project, PA, SA, NB, VC, TL, IS and YB contributed to the theoretical developments, discussion and writing. AAL, PA, AA, SA, MB, LB, CB, NB, VC, SG, LH, TL, BP, IS and YB read and approved the final manuscript.

Funding

This work was funded by the C.N.R.S., the I.N.R.A., and the GDR CNRS 2157. It also received funds from a research program grant from the Cancéropôle Grand-Ouest, the Ligue Nationale Contre le Cancer and grants from Amgen and the French National Society of Gastroenterology. Laura Helou holds a PhD fellowship from the Région Centre Val de Loire. The funders have had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Acknowledgements

We thank Dr Olivier Hyrien for kindly supplying annotation files about replication origins obtained by OK-seq. We acknowledge the High-throughput sequencing facility of I2BC for its sequencing and bioinformatics expertise.

References

1. Piégu B, Bire S, Arensburger P, Bigot Y, A survey of transposable element classification systems - a call for a fundamental update to meet the challenge of their diversity and complexity. Mol. Phylogenet. Evol. 86 (2015) 90-109. https://doi.org/10.1016/j.ympev.2015.03.009.
2. Cordaux R, Udit S, Batzer MA, Feschotte C, Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci USA. 2006;103:8101-8106. https://doi.org/10.1073/pnas.0601161103.
3. Finstermeier K, Zinner D, Brameier M, et al. A mitogenomic phylogeny of living primates. PLoS. One. 8 (2013) e69504. https://doi.org/10.1371/journal.pone.0069504.
4. Carlson SM, Moore KE, Sankaran SM, et al., A proteomic strategy identifies lysine methylation of splicing factor snRNP70 by the SETMAR Enzyme. J. Biol. Chem. 290 (2015) 12040-12047. https://doi.org/10.1074/jbc.M115.641530.
5. Fnu S, Williamson EA, De Haro LP, et al., Methylation of histone H3 lysine 36 enhances DNA repair by nonhomologous end-joining. Proc. Natl. Acad. Sci. USA. 108 (2011) 540-545. https://doi.org/10.1073/pnas.1013571108.
6. Weinberg DN, Papillon-Cavanagh S, Chen H, et al., The histone mark H3K36me2 recruits DNMT3A and shapes the intergenic DNA methylation landscape. Nature. 573 (2019) 281-286. https://doi.org/10.1038/s41586-019-1534-3.
7. Liu D, Bischerour J, Siddique A, et al., The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase. Mol. Cell. Biol. 27 (2007) 1125-1132. https://doi.org/10.1128/MCB.01899-06.
8. Miskey C, Papp B, Mátés L, et al., The ancient mariner sails again: transposition of the human Hsmar1 element by a reconstructed transposase and activities of the SETMAR protein on transposon ends. Mol. Cell. Biol. 27 (2007) 4589-4600. https://doi.org/10.1128/MCB.02027-06.
9. Beck BD, Park SJ, Lee YJ, et al., Human Pso4 is a metnase (SETMAR)-binding partner that regulates metnase function in DNA repair. J. Biol. Chem.283 (2008) 9023-9030. https://doi.org/10.1074/jbc.M800150200.
10. Beck BD, Lee SS, Hromas R, et al., Regulation of Metnase's TIR binding activity by its binding partner, Pso4. Arch. Biochem. Biophys. 498 (2010) 89-94. https://doi.org/10.1016/j.abb.2010.04.011.
11. Williamson EA, Rasila KK, Corwin LK, et al., The SET and transposase domain protein Metnase enhances chromosome decatenation: regulation by automethylation. Nucleic. Acids. Res. 36 (2008) 5822-5831. https://doi.org/10.1093/nar/gkn560.
12. Wray J, Williamson EA, Sheema S, et al., Metnase mediates chromosome decatenation in acute leukemia cells. Blood. 114 (2009)1852-1858. https://doi.org/10.1182/blood-2008-08-175760.
13. Hromas R, Wray J, Lee SH, et al., The human set and transposase domain protein Metnase interacts with DNA Ligase IV and enhances the efficiency and accuracy of non-homologous end-joining. DNA Repair. 7 (2008) 1927-1937. https://doi.org/10.1016/j.dnarep.2008.08.002.

14. Wray J, Williamson EA, Chester S, et al., The transposase domain protein Metnase/SETMAR suppresses chromosomal translocations. Cancer. Genet. Cytogenet. 200 (2010) 184-190. https://doi.org/10.1016/j.cancergencyto.2010.04.011.
15. Beck BD, Lee SS, Williamson E, et al., Biochemical characterization of metnase's endonuclease activity and its role in NHEJ repair. Biochemistry. 50 (2011) 4360-4370. https://doi.org/10.1021/bi200333k.
16. Kim HS, Chen Q, Kim SK, et al., The DDN catalytic motif is required for Metnase functions in non-homologous end joining (NHEJ) repair and replication restart. J. Biol. Chem. 289 (2014) 10930-10938. https://doi.org/10.1074/jbc.M113.533216.
17. Kim HS, Kim SK, Hromas R, et al., The SET Domain Is Essential for Metnase Functions in Replication Restart and the 5′ End of SS-Overhang Cleavage. PLoS. One. 10 (2015) e0139418. https://doi.org/10.1371/journal.pone.0139418.
18. Lee SH, Oshige M, Durant ST, et al., The SET domain protein Metnase mediates foreign DNA integration and links integration to nonhomologous end-joining repair. Proc. Natl. Acad. Sci. USA. 102 (2005) 18075-18080. https://doi.org/10.1073/pnas.0503676102.
19. Williamson EA, Farrington J, Martinez L, et al., Expression levels of the human DNA repair protein metnase influence lentiviral genomic integration. Biochimie. 90 (2008) 1422-1426. https://doi.org/10.1016/j.biochi.2008.05.010.
20. Williamson EA, Damiani L, Leitao A, et al., Targeting the transposase domain of the DNA repair component Metnase to enhance chemotherapy. Cancer Res. 72 (2012) 6200-6208. https://doi.org/10.1158/0008-5472.CAN-12-0313.
21. Apostolou P, Toloudi M, Kourtidou E, et al., Potential role for the Metnase transposase fusion gene in colon cancer through the regulation of key genes. PLoS. One. 9 (2014) e109741. https://doi.org/10.1371/journal.pone.0109741.
22. Tellier M, Chalmers R., Human SETMAR is a DNA sequence-specific histone-methylase with a broad effect on the transcriptome. Nucleic Acids Res. 47 (2019) 122-133. https://doi.org/10.1093/nar/gky937.
23. Tellier M, Chalmers R, The roles of the human SETMAR (Metnase) protein in illegitimate DNA recombination and non-homologous end joining repair. DNA. Repair. (Amst). 80 (2019) 26-35. https://doi.org/10.1016/j.dnarep.2019.06.006.
24. Wray J, Williamson EA, Chester S, et al., The transposase domain protein Metnase/SETMAR suppresses chromosomal translocations. Cancer. Genet. Cytogenet. 200 (2010) 184-190. https://doi.org/10.1016/j.cancergencyto.2010.04.011.
25. Wray J, Williamson EA, Royce M, et al., Metnase mediates resistance to topoisomerase II inhibitors in breast cancer cells. PLoS. One. 4 (2009) e5323. https://doi.org/10.1371/journal.pone.0005323.
26. Jeyaratnam DC, Baduin BS, Hansen MC, et al., Delineation of known and new transcript variants of the SETMAR (Metnase) gene and the expression profile in hematologic neoplasms. Exp.

Hematol. 42 (2014) 448-456. https://doi.org/10.1016/j.exphem.2014.02.005.

27. Arnaoty A, Gouilleux-Gruart V, Casteret S, et al., Reliability of the nanopheres-DNA immunization technology to produce polyclonal antibodies directed against human neogenic proteins. Mol. Genet. Genomics. 288 (2013) 347-363. https://doi.org/10.1007/s00438-013-0754-8.
28. Dussaussois-Montagne A, Jaillet J, Babin L, et al., SETMAR isoforms in glioblastoma: A matter of protein stability. Oncotarget. 8 (2017) 9835-9848. https://doi.org/10.18632/oncotarget.14218.
29. Belleannée C, Sallou O, Nicolas J, Logol: Expressive Pattern Matching in sequences. Application to Ribosomal Frameshift Modeling. In Comin,M., Kall,L., Marchiori,E., Ngom, A., Rajapakse,J. (eds.), PRIB2014 - Pattern Recognition in Bioinformatics, 9th IAPR International Conference (2014) Springer International Publishing, Stockholm, Vol. 8626, pp.34-47. https://doi.org/10.1007/978-3-319-09192-1.
30. Bailey T, Krajewski P, Ladunga I, et al., Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS. Comput. Biol. 9 (2013) e1003326. https://doi.org/10.1371/journal.pcbi.1003326.
31. Langmead B, Salzberg SL, Fast gapped-read alignment with Bowtie 2. Nat. Methods. 9 (2012) 357-359. https://doi.org/10.1038/nmeth.1923.
32. Zhang Y, Lin YH, Johnson TD, et al., PePr: a peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. Bioinformatics.30 (2014) 2568-2575. https://doi.org/10.1093/bioinformatics/btu372.
33. Feng J, Liu T, Qin B, et al., Identifying ChIP-seq enrichment using MACS. Nat. Protoc. 7 (2012) 1728-1740. https://doi.org/10.1038/nprot.2012.101.
34. Bailey TL, Johnson J, Grant CE, et al., The MEME Suite. Nucleic. Acids. Res. 43 (2015) W39-W49. https://doi.org/10.1093/nar/gkv416.
35. Thomas-Chollier M, Herrmann C, Defrance M, et al., RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic. Acids. Res. 40 (2011) e31. https://doi.org/10.1093/nar/gkr1104.
36. Thomas-Chollier M, Darbo E, Herrmann C, et al., A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat. Protoc. 7 (2012) 1551-1568. https://doi.org/10.1038/nprot.2012.088.
37. Grant CE, Bailey TL, Noble WS, FIMO: scanning for occurrences of a given motif. Bioinformatics. 27 (2011) 1017-1018. https://doi.org/10.1093/bioinformatics/btr064.
38. Bindea G, Mlecnik B, Hackl H, et al., ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25:1091-1093. https://doi.org/10.1093/bioinformatics/btp101.
39. Akerman I, Kasaai B, Bazarova A, et al., A predictable conserved DNA base composition signature defines human core DNA replication origins. Nat. Commun. . 2020;11:4826. https://doi.org/10.1038/s41467-020-18527-0.
40. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30 (2013) 772-780.

https://doi.org/10.1093/molbev/mst010.

41. Suyama M, Torrents D, Bork P, PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic. Acids. Res. 34 (2006) W609-W612. https://doi.org/10.1093/nar/gkl315.
42. Stamatakis A, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 30 (2014) 1312-1313. https://doi.org/10.1093/bioinformatics/btu033.
43. Yang Z, PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24 (2007) 1586-1591. https://doi.org/10.1093/molbev/msm088.
44. Uhlén M, Fagerberg L, Hallström BM, et al., Proteomics. Tissue-based map of the human proteome. Science. 347 (2015) 1260419. https://doi.org/10.1126/science.1260419.
45. Havelock JC, Rainey WE, Carr BR, Ovarian granulosa cell lines. Mol. Cell. Endocrinol. 228 (2004) 67-78. https://doi.org/10.1016/j.mce.2004.04.018.
46. Song XC, Fu G, Yang X, et al., Protein expression profiling of breast cancer cells by dissociable antibody microarray (DAMA) staining. Mol. Cell. Proteomics. 7 (2008) 163-169. https://doi.org/10.1074/mcp.M700115-MCP200.
47. Pan C, Kumar C, Bohl S, et al., Comparative proteomic phenotyping of cell lines and primary cells to assess preservation of cell type-specific functions. Mol. Cell. Proteomics. 8 (2009) 443-540. https://doi.org/10.1074/mcp.M800258-MCP200.
48. Ghazalpour A, Bennett B, Petyuk VA, et al., Comparative analysis of proteome and transcriptome variation in mouse. PLoS. Genet. 7 (2011) e1001393. https://doi.org/10.1371/journal.pgen.1001393.
49. Vogel C, Marcotte EM, Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. 13 (2012) 227-232. https://doi.org/10.1038/nrg3185.
50. Payne SH, The utility of protein and mRNA correlation. Trends. Biochem. Sci. 40 (2015) 1-3. https://doi.org/10.1016/j.tibs.2014.10.010
51. Liu Y, Beyer A, Aebersold R, On the dependency of cellular protein levels on mRNA abundance. Cell. 165 (2016) 535-550. https://doi.org/10.1016/j.cell.2016.03.014.
52. Kind J, Pagie L, de Vries SS, et al., Genome-wide maps of nuclear lamina interactions in single human cells. Cell. 163 (2015) 134-147. https://doi.org/10.1016/j.cell.2015.08.040.
53. Dixon JR, Selvaraj S, Yue F, et al., Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 485 (2012) 376-380. https://doi.org/10.1038/nature11082.
54. McCole RB, Erceg J, Saylor W, et al., Ultraconserved Elements Occupy Specific Arenas of Three-Dimensional Mammalian Genome Organization. Cell. Rep. 24 (2018) 479-488. https://doi.org/10.1016/j.celrep.2018.06.031.
55. Hong S, Kim D, Computational characterization of chromatin domain boundary-associated genomic element. Nucleic. Acids. Res. 45 (2017) 10403-10414. https://doi.org/10.1093/nar/gkx738.

56. Picard F, Cadoret JC, Audit B, et al., The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells. PLoS. Genet. 10 (2014) e1004282. https://doi.org/10.1371/journal.pgen.1004282.

57. Langley AR, Gräf S, Smith JC, et al., Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic. Acids. Res. 44 (2016) 10230-10247. https://doi.org/10.1093/nar/gkw760.

58. Petryk N, Kahli M, d'Aubenton-Carafa Y, et al., Replication landscape of the human genome. Nat. Commun. 7 (2016) 10208. https://doi.org/10.1038/ncomms10208.

59. Kuruppumullage Don P, Ananda G, Chiaromonte F, et al., Segmenting the human genome based on states of neutral genetic divergence. Proc. Natl. Acad. Sci. USA. 110 (2013) 14699-14704. https://doi.org/10.1073/pnas.1221792110.

60. Gao T, Qian J, EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic. Acids. Res. 48 (2020) D58-D64. https://doi.org/10.1093/nar/gkz980.

61. The FANTOM 5 Consortium and the RIKEN PMI and CLST (DGT), A promoter-level mammalian expression atlas. Nature. 507 (2014) 462-470. https://doi.org/10.1038/nature13182.

62. Miotto B, How genomic approaches help the understanding of the initiation of DNA replication. Med. Sci. (Paris). 33 (2017) 143-150. https://doi.org/medsci/20173302009.

63. Hyrien O, Peaks cloaked in the mist: the landscape of mammalian replication origins. J. Cell. Biol. 208 (2015) 147-160. https://doi.org/10.1083/jcb.201407004.

64. Vaklavas C, Blume SW, Grizzle WE, Translational Dysregulation in Cancer: Molecular Insights and Potential Clinical Applications in Biomarker Development. Front. Oncol. 7 (2017) 158. https://doi.org/10.3389/fonc.2017.00158.

65. Ingolia NT., Ribosome Footprint Profiling of Translation throughout the Genome. Cell. (2016) 165 (2016) 22-33. https://doi.org/10.1016/j.cell.2016.02.066.

66. Na CH, Barbhuiya MA, Kim MS, et al., Discovery of noncanonical translation initiation sites through mass spectrometric analysis of protein N termini. Genome Res. 28 (2018) 25-36. https://doi.org/10.1101/gr.226050.117.

67. Gogol-Döring A, Ammar I, Gupta S, et al., Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4(+) T Cells. Mol. Ther. 24 (2016) 592-606. https://doi.org/10.1038/mt.2016.11.

68. Rhee HS, Pugh BF, ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. Curr. Protoc. Mol. Biol. 21 (2012) Unit 21.24. https://doi.org/10.1002/0471142727.mb2124s100.

69. Rossi MJ, Lai WKM, Pugh BF, Simplified ChIP-exo assays. Nat. Commun. 9 (2018) 2842. https://doi.org/10.1038/s41467-018-05265-7.

70. Oettinger MA, How to keep V(D)J recombination under control. Immunol. Rev.200 (2004) 165-

181. https://doi.org/10.1111/j.0105-2896.2004.00172.x

71. Spicuglia S, Franchini DM, Ferrier P, Regulation of V(D)J recombination. Curr. Opin. Immunol. 18 (2006) 158-163. https://doi.org/10.1016/j.coi.2006.01.003.

72. Hillion S, Rochas C, Youinou P, et al., Signaling pathways regulating RAG expression in B lymphocytes. Autoimmun. Rev. 8 (2019) 599-604. https://doi.org/10.1016/j.autrev.2009.02.004.

73. Spicuglia S, Pekowska A, Zacarias-Cabeza J, et al., Epigenetic control of Tcrb gene rearrangement. Semin. Immunol. 22 (2010) 330-336. https://doi.org/10.1016/j.smim.2010.07.002.

74. Majumder K, Bassing CH, Oltz EM, Regulation of Tcrb Gene Assembly by Genetic, Epigenetic, and Topological Mechanisms. Adv. Immunol. 128 (2015) 273-306. https://doi.org/10.1016/bs.ai.2015.07.001.

75. Zhang Y, Cheng TC, Huang G, et al., Transposon molecular domestication and the evolution of the RAG recombinase. Nature. 569 (2019) 79-84. https://doi.org/

76. Navarro JM, Touzart A, Pradel LC, et al., Site- and allele-specific polycomb dysregulation in T-cell leukaemia. Nat. Commun. 6 (2015) 6094. https://doi.org/10.1038/s41586-019-1093-7.

77. Papaemmanuil E, Rapado I, Li Y, et al., RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. Nat. Genet. 46 (2014) 116-125. https://doi.org/10.1038/ng.2874.

78. Halper-Stromberg E, Steranka J, Giraldo-Castillo N, et al., Fine mapping of V(D)J recombinase mediated rearrangements in human lymphoid malignancies. BMC. Genomics. 14 (2013) 565. https://doi.org/10.1186/1471-2164-14-565.

79. Liu P, Xiang Y, Fujinaga K, et al., Release of positive transcription elongation factor b (P-TEFb) from 7SK small nuclear ribonucleoprotein (snRNP) activates hexamethylene bisacetamide-inducible protein (HEXIM1) transcription. J. Biol. Chem. 289 (2014) 9918-9925. https://doi.org/10.1074/jbc.M113.539015.

80. Stadelmayer B, Micas G, Gamot A, et al., Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. Nat. Commun. 5 (2014) 5531. https://doi.org/10.1038/ncomms6531.

81. Welboren WJ, Sweep FC, Span PN, et al., Genomic actions of estrogen receptor alpha: what are the targets and how are they regulated? Endocr. Relat. Cancer. 16 (2009) 1073-1089. https://doi.org/10.1677/ERC-09-0086.

82. Francesca Finotello F, Camillo BD, Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. Brief Funct. Genomics. 14 (2015) 130-142. https://doi.org/10.1093/bfgp/elu035.

83. Anders S, Huber W, Differential expression analysis for sequence count data. Genome Biol. 11 (2010) R106. https://doi.org/10.1186/gb-2010-11-10-r106.

84. Nie Z, Hu G, Wei G, et al., c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell. 151 (2012) 68-79. https://doi.org/10.1016/j.cell.2012.08.033.

85. Lin CY, Lovén J, Rahl PB, et al., Transcriptional amplification in tumor cells with elevated c-Myc. Cell. 151 (2012) 56-67. https://doi.org/10.1016/j.cell.2012.08.026.

86. Staveley BE, Heslip TR, Hodgetts RB, et al., Protected *P*-element termini suggest a role for inverted-repeat-binding protein in transposase-induced gap repair in *Drosophila melanogaster*. Genetics. 139 (1995) 1321-1329. PMID: 7768441.

87. Jin Y, Chen Y, Zhao S, et al., DNA-PK facilitates piggyBac transposition by promoting paired-end complex formation. Proc. Natl. Acad. Sci. USA. 114 (2017) 7408-7413. https://doi.org/10.1073/pnas.1612980114.

88. Izsvák Z, Stüwe EE, Fiedler D, et al., Healing the wounds inflicted by sleeping beauty transposition by double-strand break repair in mammalian somatic cells. Mol. Cell. 13 (2004) 279-290. https://doi.org/10.1016/s1097-2765(03)00524-0.

89. Raval P, Kriatchko AN, Kumar S, et al., Evidence for Ku70/Ku80 association with full-length RAG1. Nucleic. Acids. Res. 36 (2008) 2060-2072. https://doi.org/10.1093/nar/gkn049.

90. Marmignon A, Bischerour J, Silve A, et al., Ku-mediated coupling of DNA cleavage and repair during programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. PLoS. Genet. 10 (2014) e1004552. https://doi.org/10.1371/journal.pgen.1004552.

91. Kolacsek O, Pergel E, Varga N, et al., Ct shift: A novel and accurate real-time PCR quantification model for direct comparison of different nucleic acid sequences and its application for transposon quantifications. Gene. 598 (2017) 43-49. https://doi.org/10.1016/j.gene.2016.10.035.

92. Kolacsek O, Orbán TI, Transcription activity of transposon sequence limits Sleeping Beauty transposition. Gene. 676 (2018) 184-188. https://doi.org/10.1016/j.gene.2018.07.045.

93. Kasten-Pisula U, Tastan H, Dikomey E, Huge differences in cellular radiosensitivity due to only very small variations in double-strand break repair capacity. Int. J Radiat. Biol. 81 (2005) 409-419. https://doi.org/10.1080/09553000500140498.

94. Kaur, E., Nair, J., Ghorai, A., et al., Inhibition of SETMAR-H3K36me2-NHEJ repair axis in residual disease cells prevent glioblastoma recurrence. Neuro. Oncol. (2020) in press. https://doi.org/10.1093/neuonc/noaa128.

95. Maréchal A, Li JM, Ji XY, et al., PRP19 transforms into a sensor of RPA-ssDNA after DNA damage and drives ATR activation via a ubiquitin-mediated circuitry. Mol. Cell. 53 (2014) 235-246. https://doi.org/10.1016/j.molcel.2013.11.002.

96. Mahajan K, hPso4/hPrp19: a critical component of DNA repair and DNA damage checkpoint complexes. Oncogene. 35 (2016) 2279-2286. https://doi.org/10.1038/onc.2015.321.

97. Augé-Gouillou C, Hamelin MH, Demattei MV, et al., The ITR binding domain of the Mariner Mos-1 transposase. Mol. Genet. Genomics. 265 (2001) 58-65. https://doi.org/10.1007/s004380000386

98. Zhang L, Dawson A, Finnegan DJ, DNA-binding activity and subunit interaction of the mariner transposase. Nucleic Acids Res. 29 (2001) 3566-3575. https://doi.org/10.1093/nar/29.17.3566.

99. Augé-Gouillou C, Brillet B, Germon S, et al., Mariner Mos1 transposase dimerizes prior to ITR binding. J. Mol. Biol. 351 (2005) 117-130. https://doi.org/10.1016/j.jmb.2005.05.019.

100. Dematteï MV, Hedhili S, Sinzelle L, et al., Nuclear importation of Mariner transposases among eukaryotes: motif requirements and homo-protein interactions. PLoS. One. 6 (2011) e23693. https://doi.org/10.1371/journal.pone.0023693.

Fig. Legends

Fig. 1. Organisation of the *setmar* gene and the various mRNA transcripts coding SETMAR isoforms. (**a**) Exon - intron organisation of the *setmar* gene. Encoded protein domains and subdomains are coloured and indicated below the graphic. (**b**) Domain organisation of SETMAR transcripts expressed in cancer cell lines and tissue biopsies. The dark blue box between positions 10458 and 12908 in a, and in X1 in b correspond to an alternative exon unrelated to the SET domain. (**c**) Domain organisation of SETMAR transcripts that were characterised from healthy tissues. In (**b**) and (**c**), regions coding for protein domains and subdomains were located by coloured box when they were in frame with the rest of the transcript. Transcripts of non-coding regions are in dark grey. (**d**) Relative expression of SETMAR transcripts in healthy tissues. Normalised expression levels expressed with TPM (Transcripts Per kilobase Million) scale. Details about the data extracted using the NCBI, GTEx and Ensembl portals are supplied in File S1 and Table S2.

Fig. 2. Western blot profiling of SETMAR isoforms. (**a**) Profiling of HeLa transfected with pVAX plasmids expressing HSMAR1 (1), V1 (2), V2 (3), V3 (4) and in two biopsies of healthy colorectal tissues from patients (5 ,6). Lanes M corresponds to PageRuler™ Prestained Protein Ladder, 10 to 180 kDa (ThermoFisher Scientific, Illkirch, France). Four peptides were detected in lane 4, V3, HSMAR1 and two smaller polypeptides (~37 and 28 kDa). HSMAR1 and the two polypeptides a priori resulted from translation initiations occurring within the HSMAR1 coding region of the V3 mRNA. V3 and V6 in lanes 4, 5 and 6 respectively migrated above and below 55 kDa protein of the protein ladders. (**b**) Profiling in healthy and tumorous colorectal tissues of patients. Protein extracts from healthy (H) and tumour (T) tissue biopsies of 26 patients (P1 to P26). Actin is shown as an internal loading reference. The sex of each patient and the phenotype of their colon cancer (microsatellite stable (MSS) and microsatellite instable (MSI)) are indicated. The isoforms (V3 and V6) that matched with the apparent molecular weights of the bands are indicated in the right margin. Protein molecular weights are indicated in the left margin. (**c**) Profiling in SW48 and HT29 cells lines. The isoforms (V2, V3 and HSMAR1) matching with the apparent molecular weights of the bands are indicated. Protein molecular weights are indicated in the left margin.

Fig. 3. Features of peaks co-localising (HM) or not (noHM) with *Hsmar1* and MADE1 annotations. (**a**) Venn diagram illustrating the distribution of ChIP-Seq peaks co-localising

with *Hsmar1* and MADE1 annotations and calculated with PePr and MACS2 from HT29 and SW48 datasets. (**b**) Venn diagram illustrating the distribution of ChIP-Seq peaks that did not co-localised with *Hsmar1* and MADE1 annotations and calculated with PePr and MACS2 from HT29 and SW48 datasets. (**c**) Widths of peaks co-localising (HM) or not (noHM) with *Hsmar1* and MADE1 annotations. The horizontal line in the middle, boxes, and whiskers respectively represent the median, the quartiles 1 and 3 and the data spread values. (**d**) Coverage in reads expressed in log2(fold change) in peaks that co-localised (HM) or not (noHM) with *Hsmar1* and MADE1 annotations in regard to the input controls. (**e**) peaks that co-localised (HM) or not (noHM) with *Hsmar1* and MADE1 annotations. In **c** to **f**, HT29 data are represented with blue symbols while those for SW48 are in green. Red bars represented the median and the quartiles 1 and 3 values.

Fig. 4. Features of *Hsmar1* and MADE1 ITRs bound by SETMAR isoform in HT29 and SW48 cells. (**a**) Distribution of identities between the consensus ITR and *Hsmar1* and MADE1 ITRs bound or unbound by SETMAR isoform in HT29 (blue symbols) and SW48 cells (green symbols). Red bars represent the median and the quartiles 1 and 3 values. (**b**) Conserved motifs located by GLAM2 in three categories of *Hsmar1* and MADE1 ITRs. On the top, the consensus sequence of *Hsmar1* and MADE1 ITR is shown, and below this sequence is supplied the BS (boxed with a line in black) defined *in vitro* [2,76]. Positions along the ITR are indicated below the horizontal axes of logos. The most conserved block of 12 nucleotides in bound canonical and non-canonical ITRs is boxed with a line in purple. The size of the letters in logos reflect their conservation that is scaled from 0 to 2 bits (0, 0.25, 0.50, 1.00, 1.50 and 2 bits, respectively, corresponding to a conservation of about 25, 35, 42, 50, 90 and 100% of the main letter [77]).

Fig. 5. Sequence features of noHM peaks. (**a**) Logo representation of conserved motifs located by FIMO within noHM peak features at two probability thresholds: 0.1% and 1%. (**b** and **c**) Distribution of GC contents in HM and noHM peaks obtained from HT29 and SW48 data.

Fig. 6. Features of HM and NoHM peaks in regard to some features of the human genome organisation. (**a**) Distribution of numbers of HM peaks in each chromosome taking into account the size of each chromosome. (**b**) Distribution of numbers of HM peaks in each

chromosome taking into account the number of HM annotations in each chromosome size. (**c**) Distribution of numbers of noHM peaks in each chromosome considering the size of each chromosome. (**d**) Distribution of numbers of HM peaks in intra and intergenic regions of each chromosome considering the size of each chromosome. (**e**) Distribution of numbers of HM peaks in intra and intergenic regions of each chromosome considering the number of HM annotations in each chromosome size. In (**a**) to (**e**), values are represented by blue symbols in HT29 and red symbols for SW48. Lines corresponding to linear regressions are drawn with the same colours. (**f**) Distribution of numbers of noHM peaks in intra and intergenic regions of each chromosome. Symbol colours were the same as in (**d**) and (**e**). Areas in grey locate chromosomes in which noHM peaks were enriched in both cell lines, in green those in which they were depleted, and in yellow those in which noHM peaks were enriched in one cell lines while it was depleted in the other one. (**g**) and (**h**), Occupancy rate of *Hsmar1* and MADE1 ITR by HM peaks considering their intra and intergenic distributions in LAD and iLAD (**g**) and in non-boundary TAD and boundaries (**h**). Data of both cell lines were gathered in these representations. (**i**) and (**j**), number of no HM peaks in LAD and iLAD regions in HT 29 (**i**) and SW48 (**j**) cell lines. (**k**) and (**l**), number of no HM peaks in non-boundary TAD and boundaries regions in HT 29 (**i**) and SW48 (**j**) cell lines. The colour legend for the bars in (**a**) to (**e**) is supplied in (**k**) and correspond to the number of noHM peaks observed (Obs) or theoretical (Th) in intra and intergenic regions. In c, blue and red arrows indicate chromosomes in which noHM peaks were overrepresented. In g and h, the horizontal line in the middle, boxes, and whiskers represent the median, the quartiles 1 and 3 and the data spread values respectively.

Fig. 7. Features of HM and noHM peaks in HT29 and SW48 cells in regard to their overlaps with regions containing origins of replication. (**a**) Ratio of observed and expected numbers of HM and noHM peaks that overlap regions containing core origins of replication in both cell lines. (**b**) Ratio of observed and expected numbers of HM and noHM peaks that overlap regions containing stochastic origins of replication in both cell lines. In (**a**) and (**b**), red arrows indicate a significant peak enrichment, while those in black depict a significant peak depletion in permutation tests ($p<0.01$). (**c** and **d**) Chromosomal distribution in percentage of noHM peaks displaying an overlap with a core (**c**) or a stochastic (**d**) origin of replication in HT29 (black bars) and SW48 (grey bars). Red bars represent the percentage of core (**c**) or stochastic (**d**) origins of replication in each chromosome. Chromosome numbers typed in blue or red respectively indicate statistically significant depletion or enrichments in noHM

peaks that overlap regions containing core (**c**) or stochastic (**d**) origins of replication in both cell lines ($p<0.05$; hypergeometric distribution).

Additional files

Fig. S1. IGV graphic representation of aligned reads within the 20kbp chromosomal region containing the *setmar* gene.

Fig. S2. Differential peak detection in SW48 and HT29 cells for HM and NoHM peaks using MACS2 and PePr.

Fig. S3. (a) Ontologies of protein-coding genes containing *Hsmar1* and-or MADE1 ITR within their non-coding regions; (b) Ontologies of protein-coding genes contained in peaks with no overlap with a *Hsmar1* or MADE1 annotation.

File S1. Protein sequences alignment of the 3 SET subdomains and the 11 putative SETMAR isoforms.

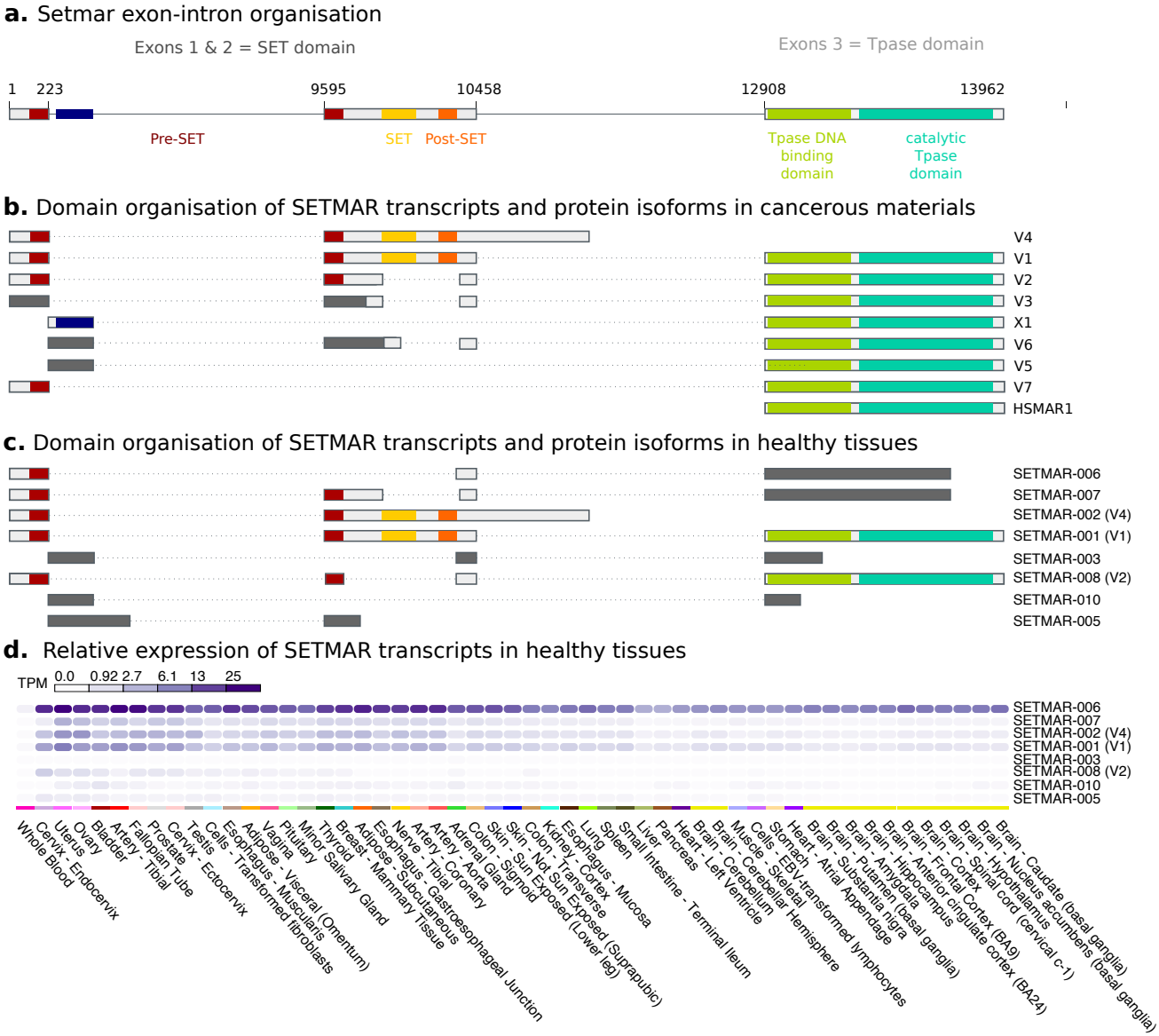
File S2. (a) Inventory of MADE1 elements in the hg38 version of the human genome. (b) Distribution in the different structural and functional components of the human genome.

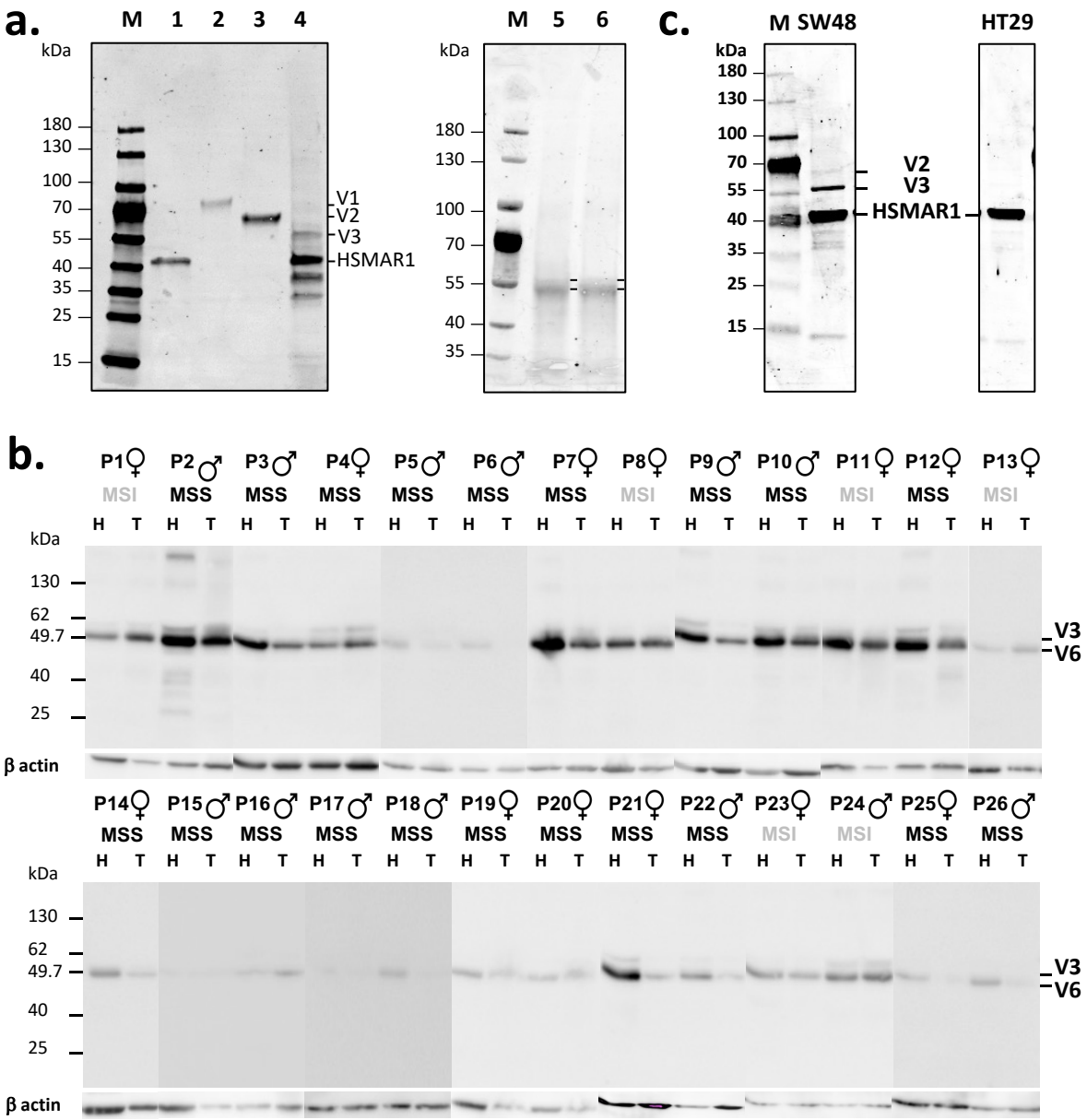
File S3. Annotations files.

Table S1. Characteristics of the 54 neogenes derived from DNA transposons in the human genome. Update of the list published in [27].

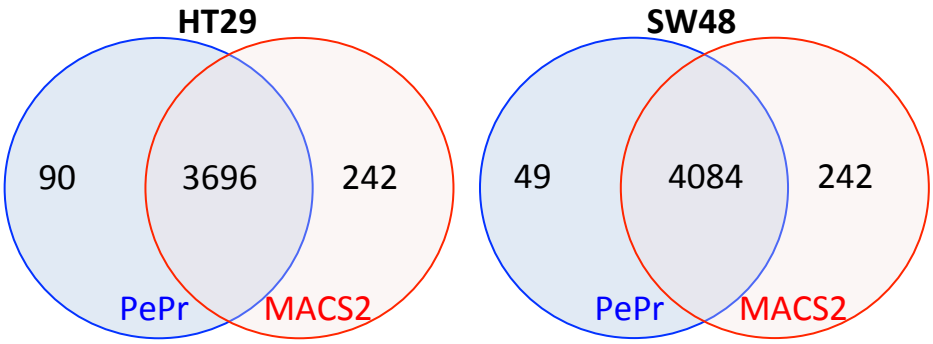
Table S2. Features of *Setmar* transcripts.

Table S3. Features of tumoral samples used.

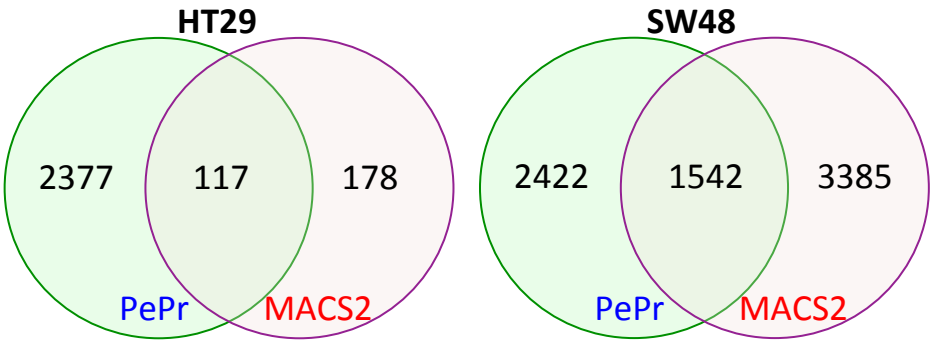


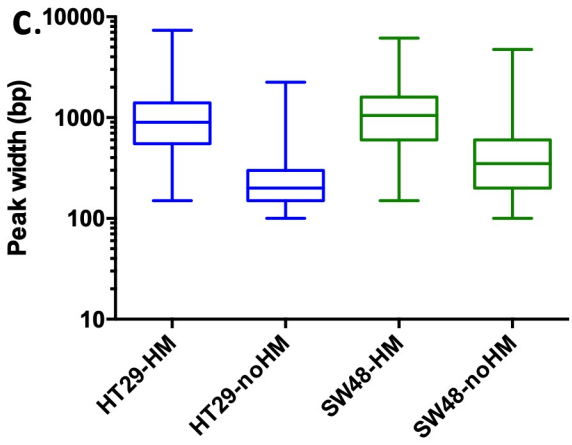


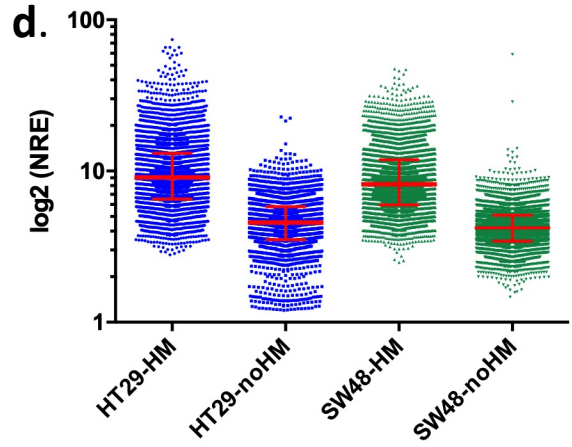
a. Peaks overlapping with a *Hsmar1* or MADE1 annotations

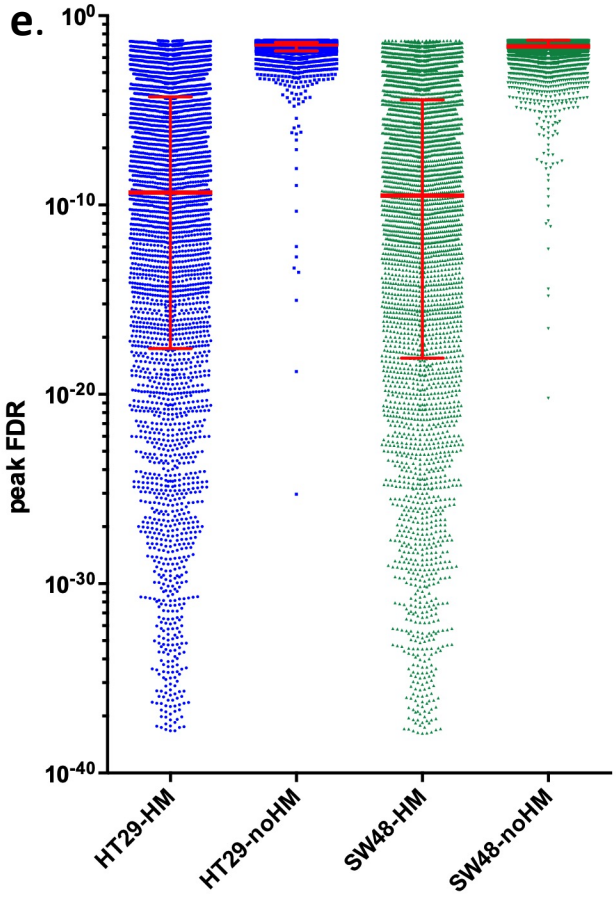


b. Peaks without a *Hsmar1* or MADE1 annotations

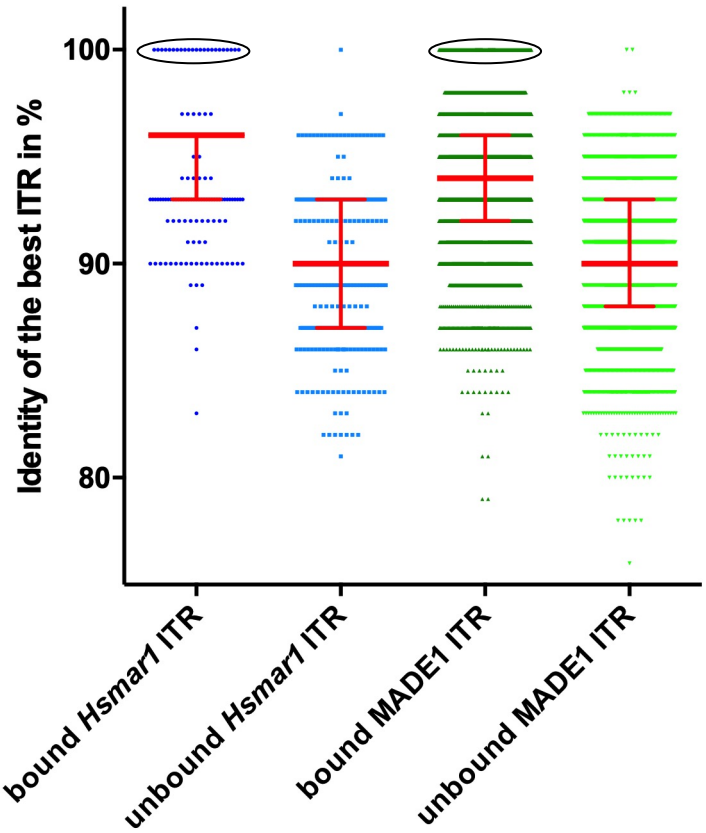




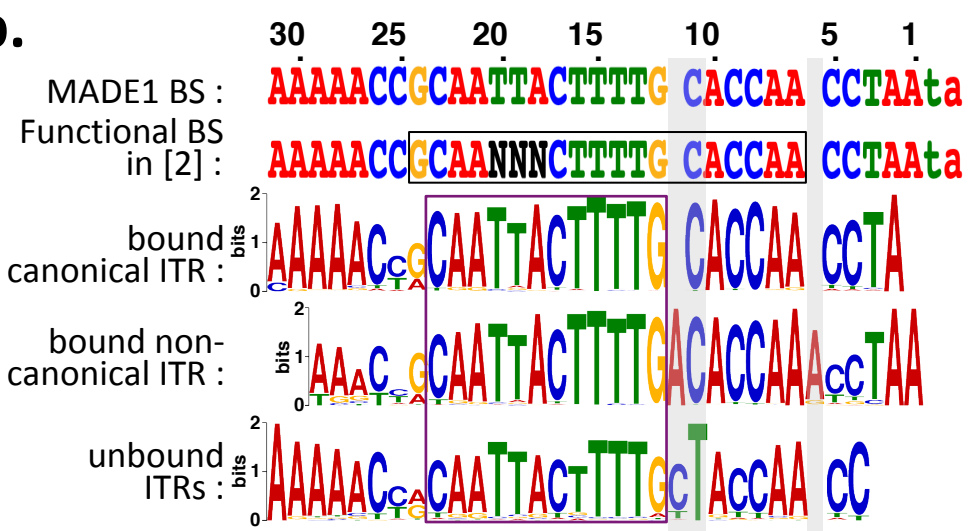


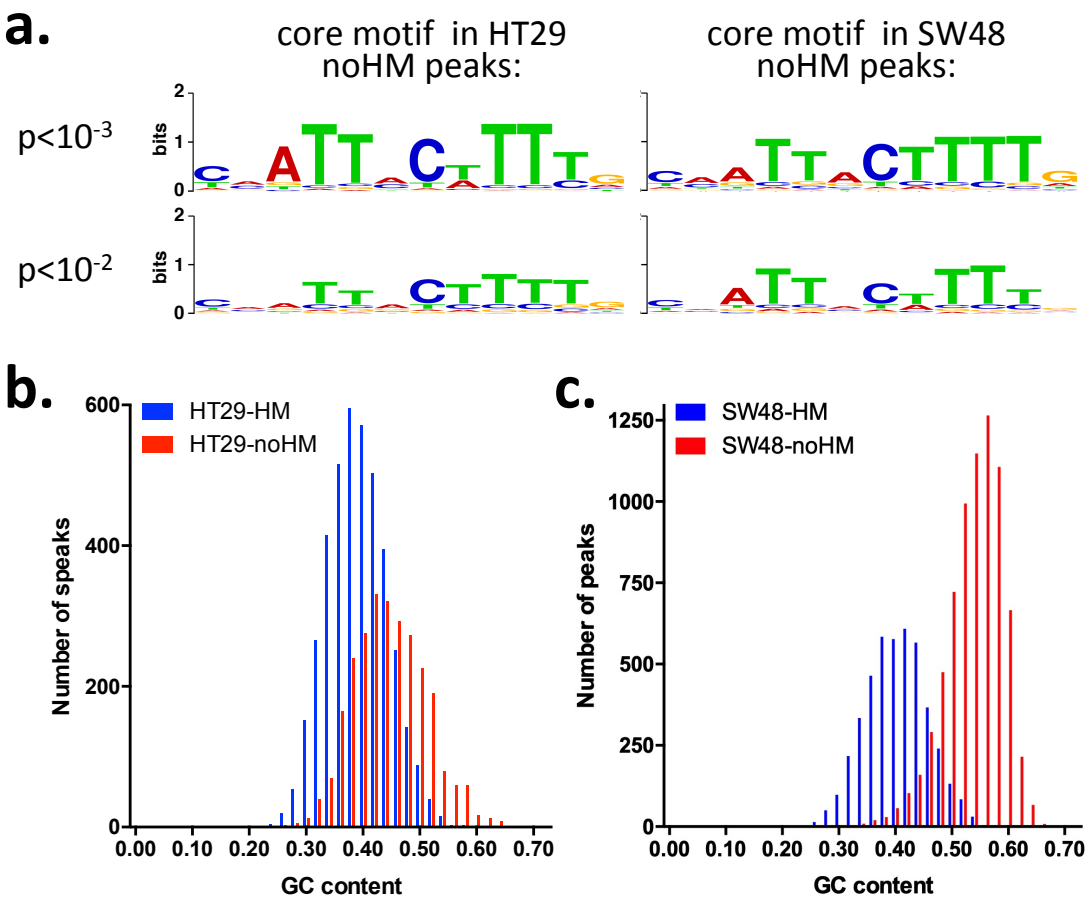


a.

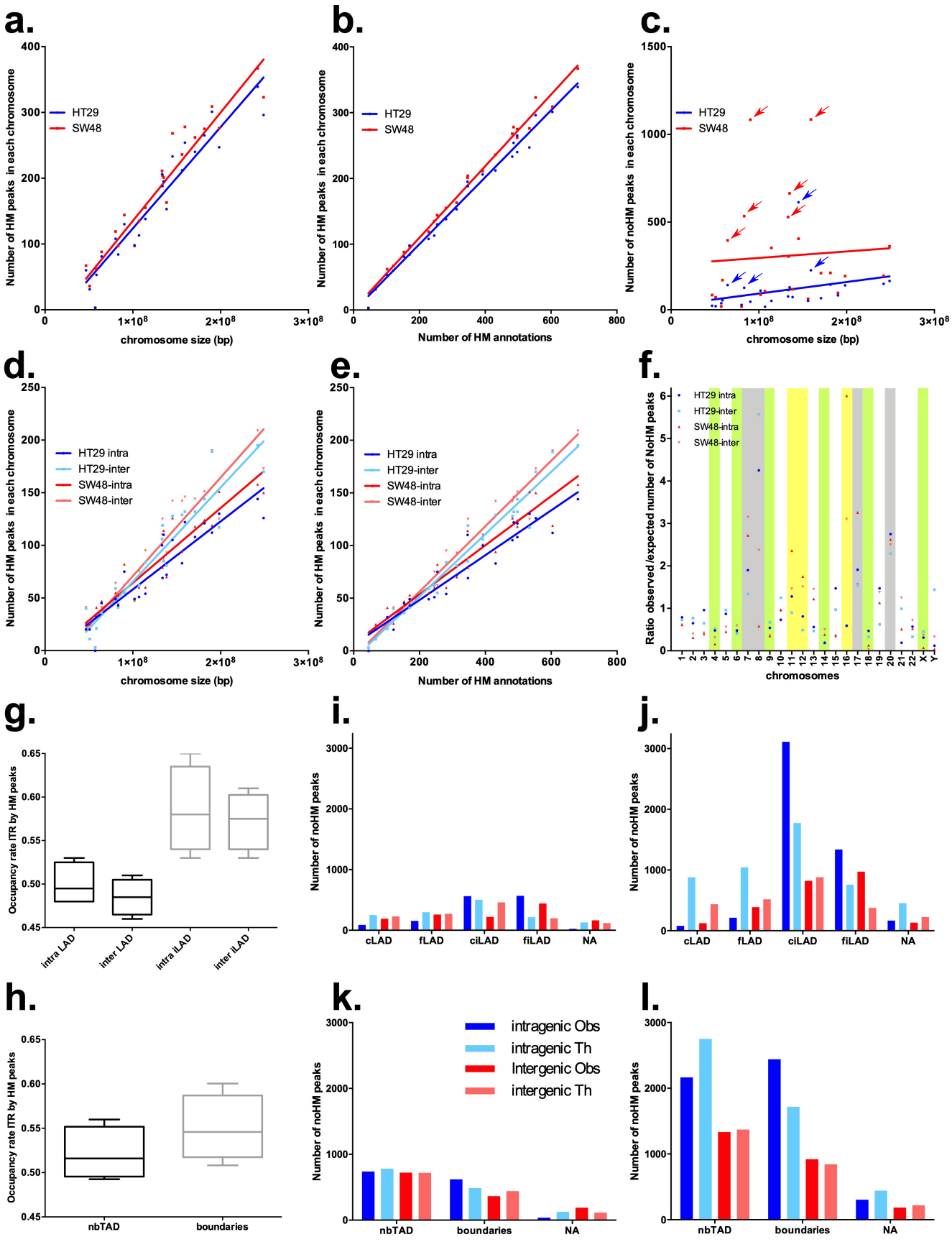

Identity of the best ITR in %
100
90
80
bound Hsmar1 ITR
unbound Hsmar1 ITR
bound MADE1 ITR
unbound MADE1 ITR

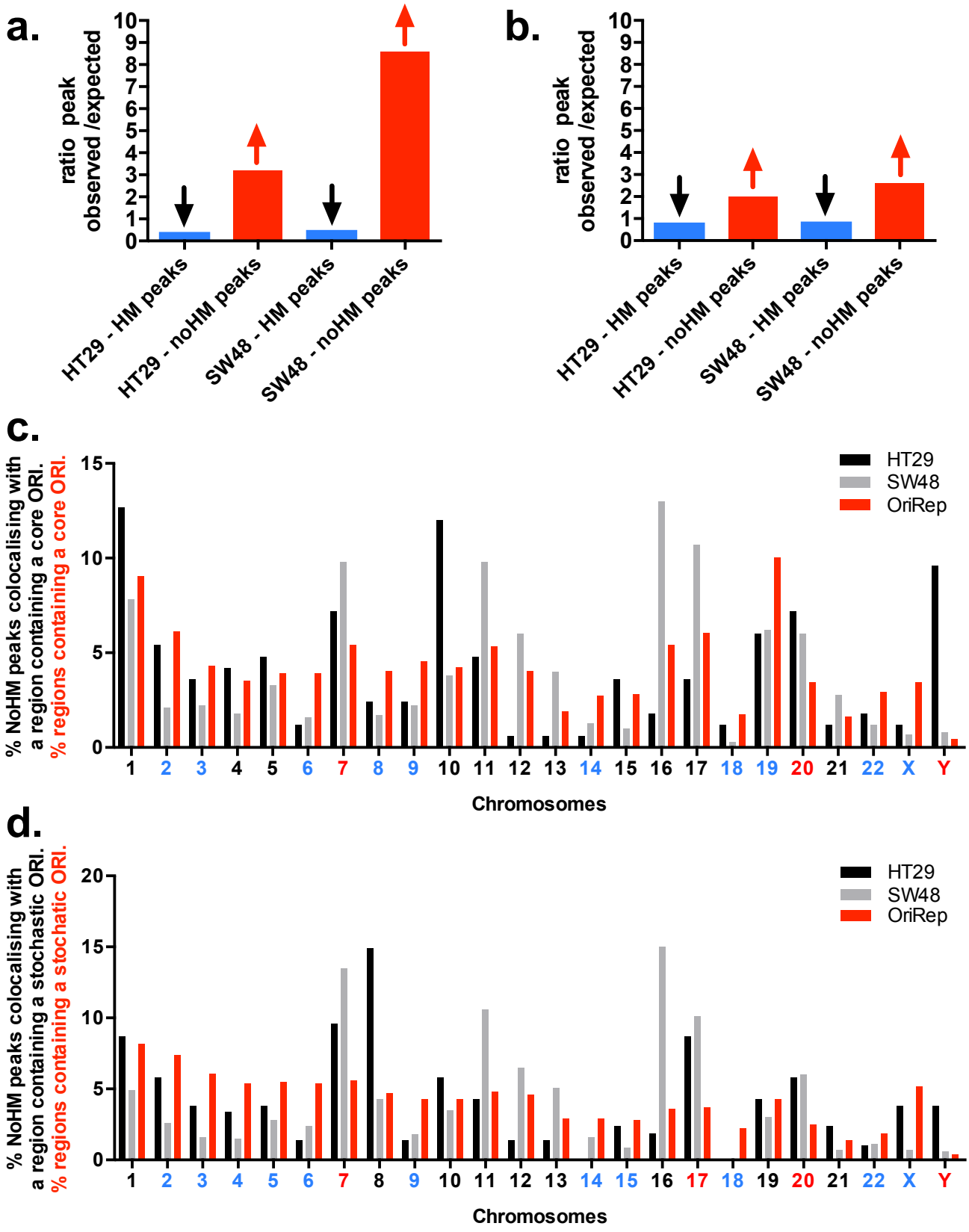
b.


30
25
20
15
10
5
1
MADE1 BS :
Functional BS in [2] :
bound canonical ITR :
bound non-canonical ITR :
unbound ITRs :
bits


a.
core motif in HT29 noHM peaks:
core motif in SW48 noHM peaks:
p<10-3
p<10-2
bits
b.
HT29-HM
HT29-noHM
Number of speaks
GC content
c.
SW48-HM
SW48-noHM
Number of peaks
GC content

Antoine-Lorquin_Fig.6





Author Statement

AAL, AA, MB, LB, SG, LH, BP and YB did the data acquisition and analyses, CB, NB, IS and YB did interpretation of data, YB designed the project, PA, SA, NB, VC, TL, IS and YB contributed to the theoretical developments, discussion and writing. AAL, PA, AA, SA, MB, LB, CB, NB, VC, SG, LH, TL, BP, IS and YB read and approved the final manuscript.