

A genome-wide transcriptional network controlled
by the human SETMAR protein

Michael Tellier

**Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy**

July 2015

Abstract

Transposable elements are discrete segments of DNA, which can be mobilized and amplified within a host genome. They are found in almost every living organism where their activities can contribute to gene expression and phenotypic variability. Hsmar1, a DNA transposon from the ITm superfamily, has entered the anthropoid primate lineage 55 to 65 Mya and remained active for ~ 15 My. The human genome contains 8,527 Hsmar1 remnants, which are essentially distributed between Made1 elements, an 80 bp deletion-derivative of the Hsmar1 transposon containing two transposon ends, and solo ITRs, which are deletion-derivative of Made1. Amidst these remnants, a fusion event between a pre-existing histone methyltransferase gene and one copy of the Hsmar1 transposase gave birth to the SETMAR gene.

Except for a possible role in non-homologous end joining, the functions of SETMAR in human cells are still poorly understood. Since the DNA-binding domain of SETMAR is under purifying selection and several thousands transposon ends are located within genes, I investigated whether SETMAR binds to the genic transposon ends to regulate gene expression by dimethylating the K36 of histone H3. By using RNA-Seq, I established that a change in SETMAR expression modifies the expression of a third of the 1,523 protein-coding genes containing an ITR, with a bias towards up-regulation. I determined SETMAR binding sites with ChIP-Seq and found that SETMAR binds essentially to genic and intergenic Made1 elements and to a group of putative miRNAs derived from Made1. Even though only 2% of the differentially expressed genes with an ITR are found in the ChIP-Seq data, the genic ChIP peaks are enriched in the genes with an ITR.

SETMAR may fine-tune by itself or by interacting with other proteins the expression of the subset of genes, which are differentially expressed and bound by SETMAR. Members of this group of genes encode proteins involved in signalling pathways such as FAM83B and ARSG, which could in turn modify gene expression. SETMAR could also potentially regulate the expression of the miRNAs-like derived from Made1.

Acknowledgements

My first acknowledgements go to Ronald Chalmers for giving me the opportunity and the freedom to work on different projects. Thank you for the discussions and advices during the last four years. You were a great supervisor.

I am grateful to the past and present members of the Chalmers lab. Neil for his assistance, Danxu for technical advises, Corentin, Mar, George, and Shiv for discussions and sharing views.

I am also grateful to David Onion for its help for the flow cytometry.

I also thank Vlad, Marcus, and Abdul for the interesting discussions on science and biohacking.

I also thank the people I lived with over the last four years.

I particularly thank my family for their support.

I also gratefully acknowledge the financial support of the University of Nottingham.

Table of Contents

Abstract.....	2
Acknowledgements	3
Table of Contents.....	4
Table of Tables.....	9
Table of Figures.....	10
Abbreviations	10
1. INTRODUCTION	14
1.1 The classification of Transposable Elements.....	15
1.2 Hsmar1 and the ITm superfamily	16
1.2.1 The mariner family	21
1.2.2 Hsmar1	23
1.2.3 Mechanism of Hsmar1 transposition	24
1.2.4 Life-cycle of a transposon and regulation of transposition	26
1.2.5 Structural insights into SETMAR and the related Mos1 transposase.....	28
1.2.6 Miniature inverted repeat transposable elements (MITEs).....	32
1.3 Exaptation	36
1.3.1 Co-option of regulatory sequences derived from transposable elements	36
1.3.2 Exaptation of MITEs	38
1.3.3 Domestication of the transposase gene.....	41
1.3.4 SETMAR.....	43
1.4 Aims of the research	48
2. MATERIALS AND METHODS	52
2.1 Strains and cell lines	53
2.2 Oligonucleotides and plasmids.....	54
2.3 Chemicals	59
2.4 General DNA manipulation	59
2.4.1 Polymerase Chain Reaction (PCR)	59
2.4.2 Plasmid construction	60
2.5 Cell culture.....	60
2.5.1 Media and growth conditions	60
2.5.2 Transient transfection of the T-Rex-U2OS cell line	61
2.5.3 Stable transfection of T-Rex-U2OS cells	61

2.5.4 Proteins extraction	62
2.5.5 Western blotting.....	63
2.5.6 RNA extraction	64
2.5.7 RNA-Seq.....	64
2.5.8 Preparation of the sheared chromatin for ChIP-exo.....	65
2.5.9 ChIP-exo	66
2.5.10 Non-homologous end-joining assay and FACS analyses	67
2.5.11 Illegitimate DNA integration assay	68
2.6 Bioinformatics analysis.....	69
2.6.1 Hsmar1 remnants and ITRs in the human genome.....	69
2.6.2 RNA-Seq analysis.....	69
2.6.3 ChIP-exo analysis	72
2.7 Bacterial assays.....	73
2.7.1 Constitutive promoters.....	73
2.7.2 Western blotting.....	75
2.7.3 Papillation assays.....	76
2.7.4 Mating-out assay.....	76
3. BIOINFORMATICS ANALYSIS OF THE HSMAR1 TRANSPOSON IN THE HUMAN GENOME.....	78
3.1 Introduction	79
3.2 Results.....	79
3.2.1 Distribution of the Hsmar1 remnants in the human genome.....	79
3.2.2 Distribution of the Hsmar1 transposon ends in the human genome.	84
3.2.2.1 Origins of the transposon ends	84
3.2.2.2 Distribution of the ITRs in the human genome.....	88
3.2.2.3 Gene Ontology enrichment of genes with an ITR.....	95
3.3 Discussion	97
3.3.1 Hsmar1 remnants in the human genome	98
3.3.2 Are a fraction of the ITRs under purifying selection?.....	99
3.3.3 A class of miRNA-like genes derived from Made1	100
3.3.4 Hsmar1 remnants in human lncRNAs.....	101
4. SETMAR IS AN EPIGENETIC REGULATOR OF A GENE NETWORK ..	103
4.1 Introduction	104
4.2 Results.....	105
4.2.1 Stable cell lines used for RNA-Seq and ChIP-exo.....	105

4.2.2 Between 20 and 35% of the differentially expressed genes are unique to each cell line	108
4.2.3 SETMAR overexpression is associated with an enrichment in differentially expressed genes containing an ITR.....	112
4.2.4 SETMAR expression level correlates with a decrease in the strength of genes down-regulation.....	117
4.2.5 A change in gene expression is only observed in active genes.....	119
4.2.6 The differential expression of genes does not depend on whether they have a solo ITR or a Made1 element.....	123
4.2.7 Gene Ontology (GO) enrichment in the different RNA-Seq data.....	124
4.2.8 No significant enrichment in the ITR position compared to the start of the gene or the nearest splicing site in the differentially expressed genes.....	128
4.2.8 Characterization of the SETMAR binding sites <i>in vivo</i>	133
4.2.9 SETMAR core binding sites in the human genome.....	140
4.2.10 SETMAR binds mostly to Made1 elements rather than solo ITRs.....	142
4.2.11 The intragenic ChIP peaks are enriched in genes with an ITR	144
4.2.12 ChIP peaks are found in both expressed and silenced genes	147
4.2.13 Only a fifth of the genes with a ChIP peak are also differentially expressed.....	148
4.3 Discussion	150
4.3.1 Overlapping of the three sets of differentially expressed genes.....	150
4.3.2 Increasing SETMAR expression skews differentially expressed genes towards up-regulation.....	151
4.3.3 Depth of sequencing coverage for RNA-Seq and ChIP-Seq	152
4.3.4 Regulation of the expression of genes containing an ITR by SETMAR	153
4.3.5 SETMAR and Made1 elements.....	154
4.3.6 Do the mir-548 like genes regulate the differentially expressed genes in the RNA-Seq experiments?	155
4.3.7 ChIP peaks enrichment in the satellite regions.....	156
5. SETMAR FUNCTIONS IN NON-HOMOLOGOUS END JOINING REPAIR AND ILLEGITIMATE DNA RECOMBINATION.....	158
5.1 Introduction	159
5.2 Results.....	159
5.2.1 SETMAR overexpression does not promote cell proliferation in the U2OS genetic background.	161

5.2.2 The SET and MAR domains have opposite influence on DNA repair...	163
5.2.3 The SET and MAR domains but not SETMAR promote DNA integration	
.....	171
5.3 Discussion	176
5.3.1 SETMAR overexpression and cell proliferation.....	176
5.3.2 The relationship between DNA repair and integration.....	178
5.3.3 Functions of the SET and MAR domains in DNA repair and integration	
.....	179
5.3.3.1 Function of the SET domain in DNA repair and integration	180
5.3.3.2 Function of the MAR domain in DNA repair and integration.....	181
5.3.4 SETMAR variants and DNA repair and integration	182
6. USE OF CONSTITUTIVE PROMOTERS TO FIND HYPERACTIVE	
TRANSPOSASES IN A BACTERIAL GENETIC SCREEN.....	185
6.1 Introduction	186
6.2 Results.....	189
6.2.1 Characterization of the papillation assay using a strong inducible	
promoter.....	189
6.2.2 Constitutive promoters provide expression levels below and above the	
OPI threshold	195
6.2.3 Characterization of several Hsmar1 mutants	204
6.2.4 An Hsmar1, PiggyBac and Sleeping Beauty triple-transposon reporter	
strain.....	208
6.2.5 Two types of colonies are produced with the single copy expression	
vectors	210
6.3 Discussion	214
6.3.1 Alteration of the transposition rate by the MBP tag in <i>E. coli</i>	214
6.3.2 In <i>E. coli</i> , the transposition rate peaks with a stochastic transcription	
and translation of the transposase	215
6.3.3 Mutations in the DNA binding domain and OPI-resistance	216
6.3.4 Decreasing the transposase dimer stability produces bacterial	
hyperactive mutants.....	217
6.3.5 Comparison of the single and triple transposon ends constructs.....	218
6.3.6 Permissive colonies with the single-copy expression vectors	220
6.3.7 Variant of the papillation assay.....	222
7. CONCLUSION AND PERSPECTIVES.....	224

7.1 Summary of the research.....	225
7.2 Outlook	227
8. APPENDICES	231
Supplementary Figures	232
Bibliography.....	240

Table of Tables

Table 2-1: Escherichia coli strains	53
Table 2-2: Mammalian cell lines	54
Table 2-3: Oligonucleotides used in this study	56
Table 2-4: Plasmids used in this study	58
Table 2-5: Constitutive promoters	74
Table 3-1: GO enrichment for the genes containing an Hsmar1 remnant or an ITR and for the largest genes in the human genome	97
Table 3-2: List of Made1 elements under purifying selection.....	100
Table 4-1: Number of differentially expressed genes in each cell line	109
Table 4-2: No enrichment in differentially expressed genes near an ITR	114
Table 4-3: Strong enrichment in differentially expressed genes with an ITR	114
Table 4-4: Strong enrichment in genes with an ITR and at least a 2-fold change in expression	115
Table 4-5: Enrichment in down- and up-regulated genes with an ITR	116
Table 4-6: Enrichment in down- and up-regulated genes with an ITR and at least a 2-fold change	116
Table 4-7: GO enrichment of all differentially expressed genes in the RNA-Seq samples	126
Table 4-8: GO enrichment of the differentially expressed genes with an ITR in the RNA-Seq samples	128
Table 5-1: T-Rex-U2OS cell lines used in the DNA repair assay	166
Table 6-1: Transposition frequencies of wild type and mutant Hsmar1 transposases	195
Table 6-2: Promoters and vectors used in this Chapter.....	198
Table 6-3: Transposition frequencies of different concentrations of wild type Hsmar1 transposase.....	203
Table 6-4: Transposition frequencies of different concentrations of Hsmar1 transposase mutants	206
Table 6-5: Transposition frequencies of different wild type transposases ...	210

Table of Figures

Figure 1-1: Taxonomic distributions of DNA transposons across the eukaryotic tree of life.	18
Figure 1-2: Organization of ITm transposons and mariner transposases.	20
Figure 1-3: Phylogeny of the ITm transposons.	21
Figure 1-4: Mechanism of Hsmar1 transposition.	25
Figure 1-5: Autoregulation of mariner transposition by overproduction inhibition (OPI).	28
Figure 1-6: Structural features of SETMAR catalytic domain and Mos1 transposase.	31
Figure 1-7: Model for the generation of some MITE families.	34
Figure 1-8: Alignment between a Made1 element and four miRNAs derived from Made1.	40
Figure 3-1: Hsmar1 remnants in the human genome.	81
Figure 3-2: Distribution of each group of Hsmar1 remnants in the human genome.	83
Figure 3-3: Source of ITRs in the human genome.	87
Figure 3-4: Distribution of the Hsmar1 ITRs in the human genome.	89
Figure 3-5: The ITRs are particularly found in the largest human genes.	92
Figure 3-6: Distribution of the intragenic ITRs compared to the start of the genes and the nearest splicing site.	95
Figure 4- 1: Cell lines overexpressing SETMAR used for the RNA-Seq and ChIP-exo experiments.	107
Figure 4-2: The SM3 cell line and the genes with at least a 2-fold change contain a higher proportion of unique genes.	110
Figure 4-3: SETMAR overexpression correlates with an upregulation of the genes with an ITR expressed at a low or medium level.	111
Figure 4-4: The highest fold changes are observed for the lowest expressed genes.	112
Figure 4-5: Inverse relationship between SETMAR expression and gene down-regulation.	119
Figure 4-6: The differentially expressed genes are on average more expressed than the non-differentially expressed genes.	122
Figure 4-7: The differentially expressed genes with an ITR are not enriched in solo ITRs nor Made1 elements.	124
Figure 4-8: The position of the ITR compared to the start of the gene is not correlated to transcriptional changes.	130
Figure 4-9: The position of the ITR compared to the nearest splicing site is not correlated to transcriptional changes.	132
Figure 4-10: Negative and positive control of a ChIP experiment on the control cell line TOF1.	134

Figure 4-11: ChIP-exo profiles over a genomic region.	135
Figure 4-12: Both ChIP-exo replicates contain an enrichment for the ZNF354C binding site.	136
Figure 4-13: The ChIP-exo replicates overlap partially and the common peaks are located in the chromosomes centromere.	138
Figure 4-14: The non-centromeric ChIP peaks are enriched for the ITR sequence.	140
Figure 4-15: Determination of the consensus SETMAR binding site in vivo and putative transcription factor binding sites.	142
Figure 4-16: SETMAR binds essentially to Made1 elements and miRNAs-like.	144
Figure 4-17: The ChIP peaks are enriched in genes with an ITR.	145
Figure 4-18: Only the genes with an ITR bound by SETMAR are enriched in Made1 elements.	146
Figure 4-19: SETMAR does not bind specifically to expressed genes.	148
Figure 4-20: The genes with a ChIP peak overlap only partially with the three sets of differentially expressed genes.	149
Figure 5-1: The overexpression of SET or SETMAR does not promote cell proliferation in an U2OS genetic background.	163
Figure 5-2: Reporter substrate for the in vivo analysis of NHEJ.	165
Figure 5-3: SETMAR constructs used in the NHEJ repair assay.	167
Figure 5-4: The SET and the MAR domains of SETMAR have an antagonist activity on DNA repair.	170
Figure 5-5: Preliminary experiments on DNA integration.	173
Figure 5-6: The SET and MAR domains increase the frequency of DNA integration in the genome of the U2OS cell line.	175
Figure 5-7: The transcriptomes of the HEK293 and U2OS cell lines are similar.	177
Figure 5-8: SETMAR second isoform is specific to the species containing the MAR domain.	183
Figure 6-1: Papillation and mating-out assays.	187
Figure 6-2: The number of papillae per colony varies with the concentration of lactose and IPTG.	192
Figure 6-3: Hsmar1 transposase expression from different promoters.	199
Figure 6-4: Characterization of the papillation assay with the set of constitutive promoters expressing the Hsmar1 transposase.	202
Figure 6-5: Transposition rate of different Hsmar1 transposase mutants.	205
Figure 6-6: Characterization of the triple transposons papillation strain with Hsmar1, PiggyBac, and Sleeping Beauty transposases.	209
Figure 6-7: Characterization of the normal and permissive colonies.	211
Figure 6-8: The permissive phenotype is also present with the Hsmar1 transposase mutants and is encoded by the transposase expression plasmid.	212

Supplementary Figure 1: Full pictures of the Western blot presented in Figure 4.1.	232
Supplementary Figure 2: SETMAR encodes two isoforms in the U2OS cell line.	233
Supplementary Figure 3: Papillation assay on minimal media in the RC5096 strain.	234
Supplementary Figure 4: Papillation assay with the wild type Hsmar1 transposase in the RC5096 strain.	235
Supplementary Figure 5: Papillation assay with Hsmar1 transposase mutants in the RC5096 strain.	236
Supplementary Figure 6: Papillation assay with the wild type Hsmar1 transposase in the RC5118 strain.	237
Supplementary Figure 7: Papillation assay with the wild type PiggyBac and Sleeping Beauty transposase in the RC5118 strain.	238
Supplementary Figure 8: Characterization of normal and permissive colonies.	239

Abbreviations

a.a.	amino acid(s)	M&M	materials and methods
Amp	ampicillin	MBP	maltose binding protein
bp	base pair	mb	megabase pairs
cDNA	complementary DNA	MITE	miniature inverted-repeat transposable element
ChIP	chromatin immunoprecipitation	MLE	<i>mariner</i> -like element
Cm	chloramphenicol	My	million years
DNA	deoxyribonucleic acid	Mya	million years ago
DMEM	Dulbecco's modified Eagle's medium	NEB	New England BioLabs
DMSO	dimethyl sulfoxide	NHEJ	non-homologous end joining
dNTPs	deoxynucleotides triphosphates	nt	nucleotide
DTT	dithiothreitol	OPI	overproduction inhibition
e.g.	for example	ORF	open reading frame
EDTA	ethylenediaminetetraacetic acid	ori	origin of replication
EGFP	enhanced green fluorescent protein	pBS	pBluescript vector
EtOH	ethanol	PBS	phosphate buffered saline
FACS	fluorescence-activated flow cytometry	PCR	polymerase chain reaction
FBS	fetal bovine serum	PVDF	polyvinylidene difluoride
HR	homologous recombination	RFP	red fluorescent protein
HTH	helix-turn-helix	RNA	ribonucleic acid
i.e.	that is	rt	room temperature
IPTG	isopropyl-beta-D-thiogalactopyranoside	SINE	short interspersed nuclear elements
ITR	inverted terminal repeat	TF	transcription factor
Kan	kanamycin	TFBS	transcription factor binding site
kb	kilobase pairs	TEs	transposable elements
LINE	long interspersed nuclear elements	UTR	untranslated region
LTR	long terminal repeat	W.T.	wild type

1. INTRODUCTION

Transposable elements (TEs) are discrete segments of mobile DNA. They are almost ubiquitous and their transposases are the most abundant genes in nature (Aziz et al., 2010). Because the genetic information encoded by TEs is usually used only for their own survival, they have been considered as 'parasitic' or 'selfish' genes (Orgel and Crick, 1980). This view is supported by the mutagenic nature of transposition. TE insertions not only disrupt open reading frames (ORFs) but can modify their expression (Fedoroff, 2001). Furthermore, the presence of several transposon copies in the genome may also initiate chromosomal rearrangements. Transposons are therefore a source of genetic novelty. In this work, I focus on the human SETMAR protein. It is a primate-specific protein that arose some 40 to 55 Mya from the fusion of a histone methyltransferase and a mariner family transposon.

1.1 The classification of Transposable Elements

Transposable elements are divided into two classes. Class I elements, or retrotransposons, transpose via an RNA intermediate whereas class II elements, or DNA transposons, transpose directly via a DNA intermediate. There is still no consensus for a further classification of eukaryotic TEs (Wicker et al., 2007) (Kapitonov and Jurka, 2008; Seberg and Petersen, 2009). However, the retrotransposons can be further divided into 5 orders: the long terminal repeat (LTR), Dictyostelium intermediate repeat sequences (DRIS), Penelope-like elements (PLE), long interspersed nuclear elements (LINE), and short interspersed nuclear elements (SINE) (Wicker et al., 2007). Except for the SINE that are non-autonomous elements, all the groups

encode several proteins, which include the reverse transcriptase, required for their mobility.

The DNA transposons are divided into two groups according to whether they move via a non-replicative (subclass I, cut-and-paste) or replicative (subclass II, copy-and-paste) mechanism. Subclass 1 is the most numerous group and is further divided into 17 superfamilies based on the origin of the catalytic domain of the transposases (Yuan and Wessler, 2011). This group includes the prokaryotic insertion sequences, which encompass a dozen families of transposons (Siguier et al., 2006). Subclass 2 includes the Helitrons and Mavericks in eukaryotes, and phage Mu and Tn3 in prokaryotes (Wicker et al., 2007).

1.2 Hsmar1 and the ITm superfamily

A domain of the SETMAR protein is derived from the Hsmar1 transposon, which is a subclass 1 cut-and-paste DNA-transposon belonging to the ITm superfamily of transposable elements. The ITm acronym is derived from the IS630, Tc1 and mariner transposons, which represent three major divisions within the grouping. The first member of the superfamily to be documented was Tc1 in *Caenorhabditis elegans* in 1983 (Emmons et al., 1983). A few years later, Mos1 and IS630 were identified in *Drosophila mauritiana* and the bacterium *Shigella sonnei*, respectively (Jacobson et al., 1986) (Matsutani et al., 1987). A steady stream of ITm elements entered the literature in subsequent years. However, these were the tip of an iceberg and the depth and breadth of their phylogenetic distribution did not start to become apparent

until 1993 when PCR experiments using mariner-specific primers revealed their presence in seven orders of insects (Robertson, 1993). We now know that ITm is probably the most widespread superfamily of transposons in nature and that they are present in all branches of the tree of life (Figure 1.1) (Yuan and Wessler, 2011).

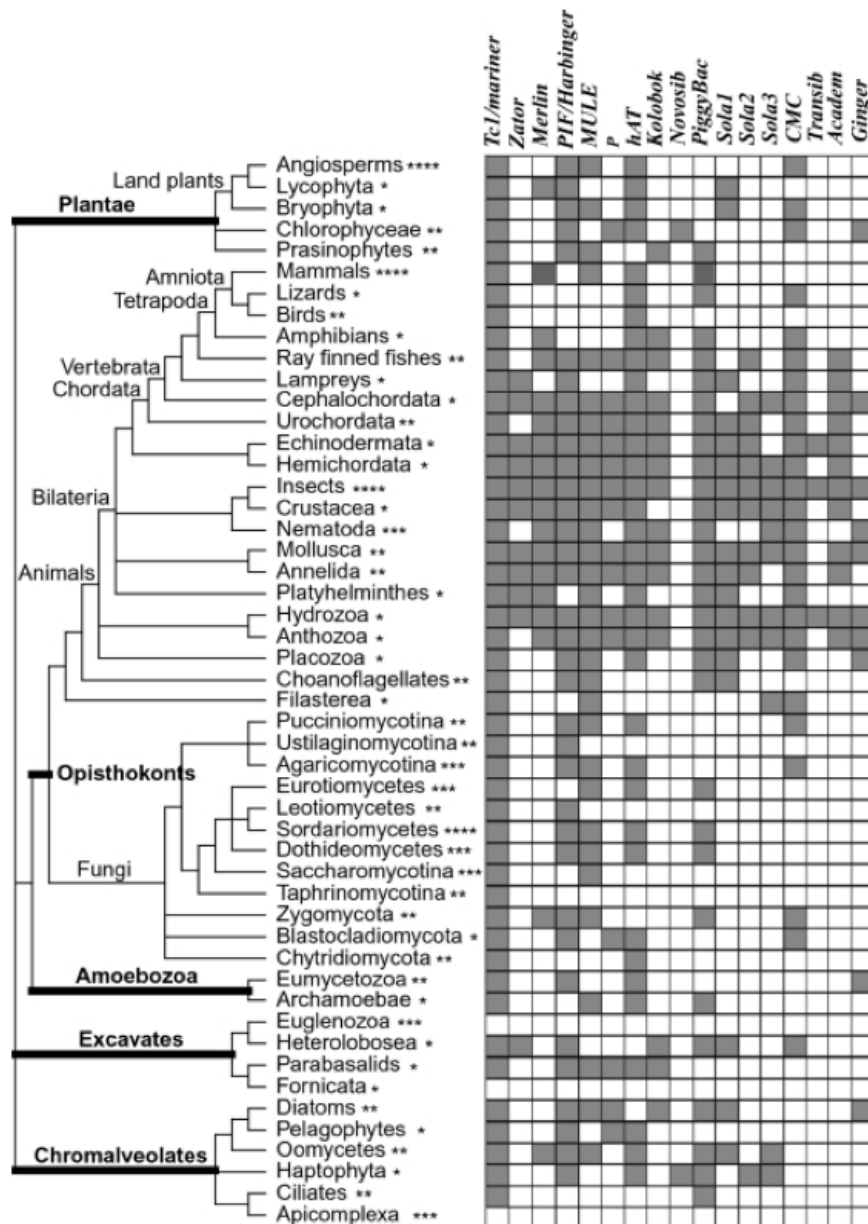


Figure 1-1: Taxonomic distributions of DNA transposons across the eukaryotic tree of life.

Five eukaryotic super-groups are indicated with thickened lines. Asterisks indicate the numbers of genomes representing the branch: *, 1 genome; **, 2 to 5 genomes; ***, 6 to 10 genomes; ****, over 10 genomes. Figure reproduced from (Yuan and Wessler, 2011).

The ITm transposons generally carry a single open reading frame (ORF) encoding the transposase (Figure 1.2A). The transposases share a common

catalytic domain containing a triad of conserved aspartate and/or glutamate residues (Figure 1.2B). These coordinate a divalent metal ion in the active site, which resides within an RNase H-like structural fold. The ITm transposases also have an N-terminal domain for site-specific DNA-binding.

The transposase gene is flanked by a pair of inverted terminal-repeats (ITR) (Figure 1.2A). These can be as short as 20 to 30 bp, encoding a single transposase binding site. The longer ITRs may have additional binding sites and spacer regions, which differ between opposite ends of the transposon. Across the group as a whole, the ITRs are so divergent that no consensus can be discerned. However, the ITRs are always flanked by a TA dinucleotide, which is derived from the target site and duplicated during the integration step of the reaction. Apart from the evolutionary origin of the catalytic domain, this is the one feature that unites all members of the ITm family. The mechanistic significance of the invariant target site duplication will be considered in more detail below.

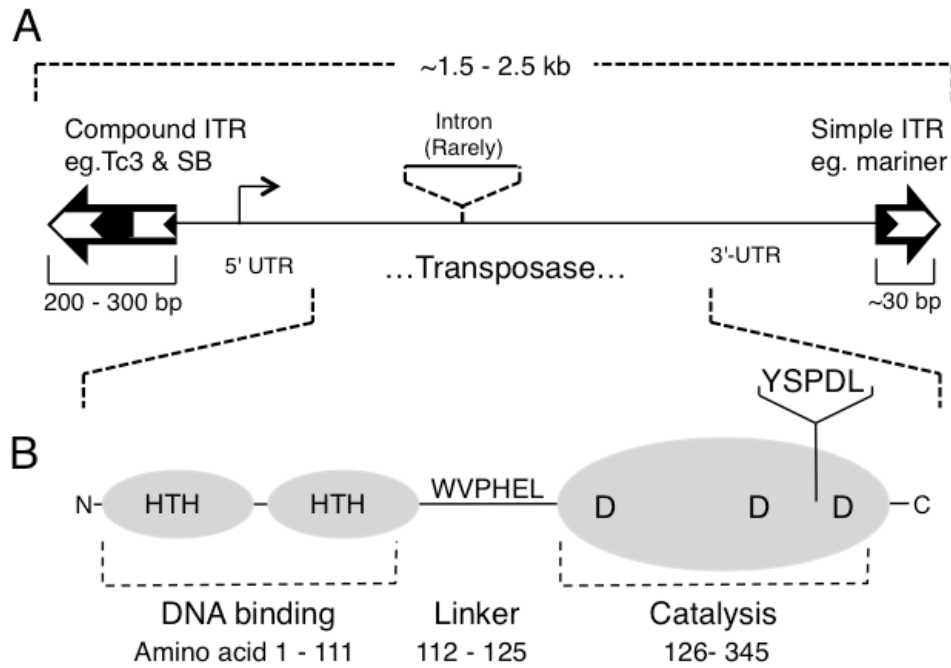


Figure 1-2: Organization of ITm transposons and mariner transposases.

(A) ITm transposons are ~1.5-2.5 kb in length. In Tc3 and Sleeping Beauty (SB), inverted terminal repeats (ITRs) have two transposase binding sites. In mariner, ITRs are ~30 bp and carry a single transposase binding site. ITm transposons have a single ORF, which codes for the transposase and is rarely interrupted by an intron.

(B) Mariner transposases have an N-terminal DNA-binding domain with two helix-turn-helix (HTH) motifs and a C-terminal catalytic domain with three aspartate (D) residues that coordinate the catalytic metal ions. The DNA-binding and catalytic domains are joined by a linker region that includes highly conserved stretch of amino acids, WVPHEL. YSPDL is another highly conserved motif that occurs just before the third aspartate. Both conserved motifs are important in communication between the active sites.

Phylogenetic analysis of the ITm catalytic domain reveals seven distinct lineages (Figure 1.3). The most mechanistically-relevant nomenclature uses the identities of the residues in the DDE/D catalytic triad and the distances between them (Shao and Tu, 2001). The spacing between the first and second aspartate residues is variable. However, the distance between the

second aspartate residue and the third aspartate or glutamate residue is sufficiently well conserved to distinguish the various lineages.

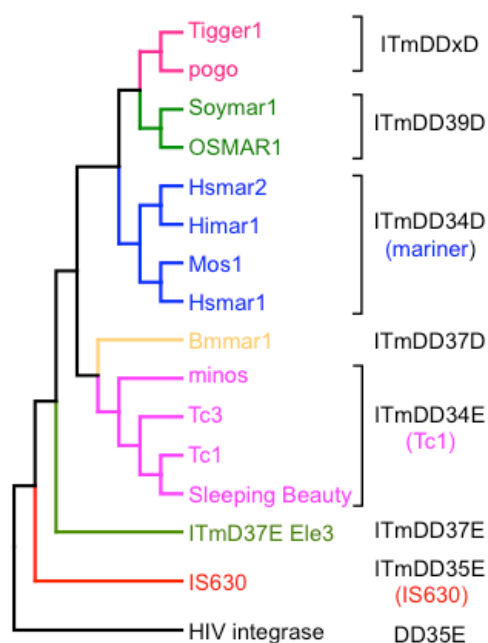


Figure 1-3: Phylogeny of the ITm transposons.

ITm families are classified according to their triad of catalytic residues and the number of amino acids separating the second aspartate (D) and the third aspartate or glutamate (E). Sequences of the catalytic domains for 15 ITm transposases were extracted from Repbase, aligned with MUSCLE and an unrooted tree was constructed with Mega 6.06 using HIV integrase as an outgroup.

1.2.1 The mariner family

The mariner (ITmD34D) transposons are generally about 1.5 kb long and encode an intron-less transposase of about 340 amino acids which is flanked by simple inverted repeats of about 30 bp (Figure 1.2). Mariner are particularly widespread in higher animals where they are capable of horizontal transfer between species. This is evident from the incongruent phylogenies of

the transposons and their hosts (Robertson and Lampe, 1995). Thus, closely related mariners may be found in distantly related hosts and vice versa. Furthermore, their distribution is patchy and they may be present or absent in closely related species. The mechanisms and vectors underlying horizontal transfer between multicellular organisms are still poorly known. However, virus with their ability to inject DNA or RNA into host cells and endosymbiotic bacteria such as *Wolbachia*, which are known to contain eukaryotic genes, are considered to be the most likely candidates {Routh, 2012 #532} {Wijayawardena, 2013 #533} {Dupeyron, 2014 #530}.

The first mariner element to be studied *in vitro* was Himar1 (Lampe et al., 1996). The transposase is a consensus sequence derived from six defunct elements cloned by PCR of the Horn Fly genome. This work established that the transposase was the only protein required for transposition, which is consistent with their wide phylogenetic distribution. In general, most transposons that have been tested are active in heterologous hosts, indicating that they are independent of specific host factors. Exceptions are Sleeping Beauty (ITmD34E), which is essentially restricted to vertebrate hosts, and the P element, which is restricted to Drosophilids. The second mariner element to be studied *in vitro* was the naturally active Mos1 (Tosi and Beverley, 2000). This was followed by *in vitro* analysis of Mboumar-9, Mcmar1 and Hsmar1 (Miskey et al., 2007) (Munoz-Lopez et al., 2008) (Claeys Bouuaert and Chalmers, 2010) (Renault et al., 2010) (Trubitsyna et al., 2014). However, most of the information set out below has been derived from Mos1 and Hsmar1.

1.2.2 Hsmar1

In common with other *mariner* elements, Hsmar1 (Homo sapiens mariner 1) has an incongruent phylogeny, indicating several independent horizontal transfers between species {Pace, 2008 #68}. It entered the tree shrew, a specie closely related to the primates, the bushbaby (a primate) and the anthropoid primates lineages about 55 to 65 Mya {Pace, 2008 #68}. It also invaded the cow and armadillo lineages ~45 to 50 Mya {Pace, 2008 #68}. Another mariner transposon, MMAR1, is found in rodents and is 95% identical to Hsmar1. The invasion of the rodent lineages is the oldest one with an estimated age comprised between 72 and 84 Mya {Pace, 2008 #68}. To calculate the age of a transposon, its degree of divergence is first determined by calculating the number of nucleotide substitutions per site in all individual copies in comparison to the consensus transposon sequence given by RepeatMasker. The age of a transposon is then estimated by comparing the divergence degree and the neutral mutation rates, which correspond to the estimated number of substitutions per site/per million years, for rodents, primates, and anthropoid primates {Pace, 2008 #121} {Pace, 2008 #68}.

More than half of the human genome is composed of repetitive sequences, mostly consisting of defunct copies of transposable elements (Lander et al., 2001). About 40% of the genome is derived from retro-transposons such as the LINE-1 elements, which are the only transposons currently active in modern humans (Beck et al., 2010). The DNA transposons all became extinct 40-50 Mya and now represent ~3% of the genome with a third of it constituted by the ITm superfamily (Pace and Feschotte, 2007).

Mariner elements are the second most abundant representatives of the superfamily after the Tigger transposons, which belong to the Tc1 family (Smit and Riggs, 1996). The mariner copies are distributed between Hsmar1 and 2 (Pace and Feschotte, 2007). Hsmar1 in association with Made1, an 80 bp deletion-derivative, is the only mariner transposon to have achieved a high copy number with 8,527 remnants in the human genome while Hsmar2 left 1,829 remnants (Robertson and Martos, 1997) (Pace and Feschotte, 2007). The reason behind the differential success of transposable elements is unclear but could be connected to differences in the transposon auto-regulatory mechanisms or in the host defences against transposable elements.

In other mammalian species, Hsmar1 activity differed significantly. Indeed, only 331 and 475 copies of Hsmar1 are found in mouse and bushbaby, respectively. The number of copies increases to 1900, 2347, and 2648 in tree shrew, armadillo, and cow, respectively {Pace, 2008 #68}.

1.2.3 Mechanism of Hsmar1 transposition

The transposition reaction only requires a single Hsmar1 transposase homodimer (Figure 1.4A). Like other helix-turn-helix (HTH) proteins, the transposase binds rapidly to the first transposon end to form the single-end complex 2 (SEC2), which is composed of a transposase dimer bound to one transposon end (Claeys Bouuaert et al., 2013) (Claeys Bouuaert et al., 2011). The recruitment of the second transposon end results in the formation of the paired-end complex (PEC), which is composed of a transposase dimer bound

to two transposon ends (Claeys Bouuaert et al., 2011). The formation of the PEC is a crucial auto-regulatory step which specifically initiates the transposase catalytic activity. Catalysis is initiated by the 5'-nicking of the transposon ends followed by their 3'-cleavages (Figure 1.4) (Claeys Bouuaert and Chalmers, 2010) (Claeys Bouuaert et al., 2013). The capture of a target site by the transposase results in the integration of the transposon in a new genomic location (Figure 1.4) (Claeys Bouuaert and Chalmers, 2010).

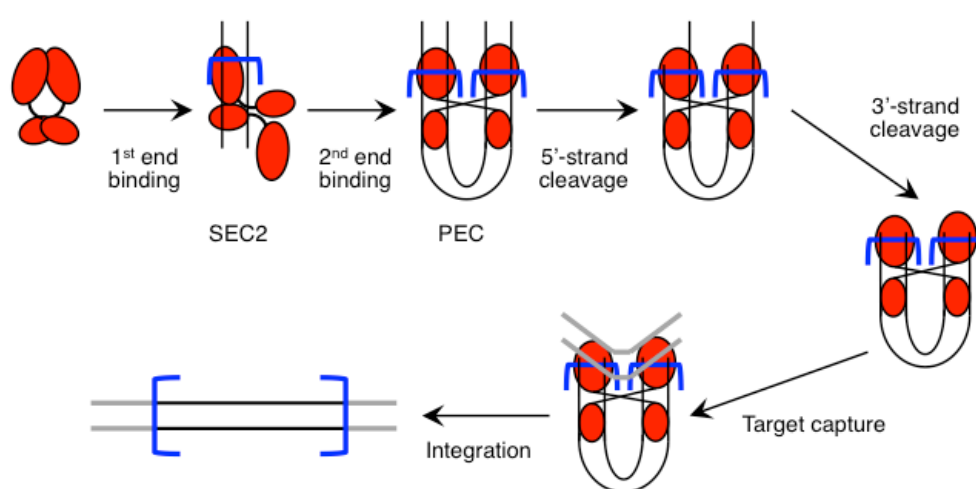


Figure 1-4: Mechanism of Hsmar1 transposition.

The binding of a transposase dimer to the first transposon end forms the single-end complex (SEC2). The recruitment of the second end within SEC2 forms the paired-end complex (PEC). After synapsis, the 5' strands of the transposon ends are cleaved first, followed by cleavage of the 3'-ends and transposon integration. Red blob, transposase; blue bracket, ITR.

The transposon excision leaves behind a double-strand DNA break (DSB), which is repaired by the host proteins. In eukaryotes, two major pathways repair the DSBs. If a sister chromatid or a homologous chromosome is present, homologous recombination (HR) can restore the initial transposon copy, which increases the transposon copy number by one if

the excised transposon integrates successfully. In absence of homologous DNA sequence, the DSB will be repaired by the non-homologous end-joining (NHEJ) pathway. The two DNA ends will be ligated together, which generates a transposon footprint. The transposon copy number will be preserved if the transposon integration is successful.

1.2.4 Life-cycle of a transposon and regulation of transposition

As alluded to above, DNA transposons do not persist for long periods in a given eukaryotic genome and their survival requires frequent horizontal transfer into new hosts. The life-cycle of a transposon might therefore be considered as the period between its first appearance in a virgin genome and its eventual extinction. Most of the knowledge on transposons life-cycle is derived from the analysis of both active and extinct transposons and from mathematical models derived from these experimental analyses. Immediately following horizontal transfer, the transpositional activity of the element is under positive selection; an element with a fully active transposase, and perfect cognate transposon ends, will have a high rate of transposition. A transposon with a hypoactive transposase or imperfect transposon ends will be less likely to start a successful invasion because of its lower transpositional activity which might not sustain a sufficient number of active transposons to compensate the removal or the silencing of the transposons by the host. As time passes some copies will acquire mutations. However, the presence of multiple copies relaxes selection because a pool of freely diffusing

transposase acts on all copies of the element, including those with defective transposase genes. Eventually, genetic drift poisons the pool of active transposase by dominant-negative complementation leading to the extinction of the transposon (Hartl et al., 1997; Le Rouzic et al., 2007; Le Rouzic and Capy, 2005; Lohe et al., 1995). Eukaryotic transposases are multimeric proteins. Thus, the presence of wild type and mutant monomers in a cell will result in the formation of chimeric multimers which are unable to perform transposition.

Notwithstanding their inevitable demise in a given genome, the transposon's life-history has many of the hallmarks of a classical host-parasite relationship. The host has adaptations, such as RNAi, to suppress the parasite, while the parasite has adaptations to spare the fitness of the host. Indeed, Hsmar1 and other mariner transposons have a set of auto-regulatory features controlling the timing and the rate of transposition (Claeys Bouuaert et al., 2011) (Claeys Bouuaert et al., 2013). The most important auto-regulation is an emergent feature of the transposition mechanism itself. The mechanism of transpososome assembly forces transposase multimers to compete for binding sites at the transposon ends (Figure 1.5) (Claeys Bouuaert et al., 2013). Consequently, above a critical threshold, doubling of the transposase concentration halves the rate of transposition. This is known as overproduction inhibition (OPI) (Lohe and Hartl, 1996).

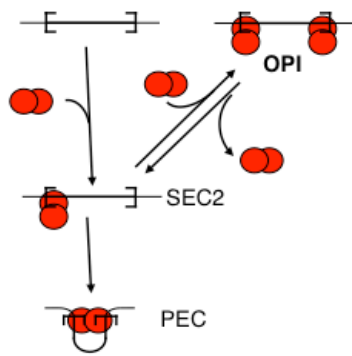


Figure 1-5: Autoregulation of mariner transposition by overproduction inhibition (OPI).

Model of the mechanism that underlies overproduction-inhibition (OPI) in mariner transposition. PEC assembly is accomplished by recruitment of a naked transposon end. As the concentration of transposase rises, naked ends are sequestered leading to inhibition of the reaction. Red blob, transposase; bracket, ITR.

1.2.5 Structural insights into SETMAR and the related Mos1 transposase

The domain structure of the mariner transposases is illustrated in Figure 1.2A. It has a bipartite N-terminal DNA binding domain connected to the catalytic domain by a proteolitically sensitive linker. Apart from the catalytic triad of D residues, the most highly conserved sequence motifs are WVPHEL and YSPDL (Robertson, 1993). Although these motifs are less well conserved in other ITm transposons, the general arrangement is probably common. For example, the ITmD34E elements Tc3 and Sleeping Beauty are similar in size to mariner and both have bipartite DNA binding domains (Watkins et al., 2004) (Carpentier et al., 2014). Other ITm transposase, such as Osmar, are significantly larger but appear to have a similar arrangement (Feschotte et al., 2005).

There are currently two published crystal structures available for SETMAR (Goodwin et al., 2010). Both structures are very similar but correspond only to the catalytic domain of the transposase. Since the transposase domain of SETMAR provides the site-specific binding activity and the dimer interface, it is therefore pertinent to examine transposases related to Hsmar1. Several crystal structures are available for DDE(D) transposases. The most relevant to SETMAR is that of the related Mos1 (Davies et al., 2000) (Richardson et al., 2009). The primary sequences of the catalytic domains from SETMAR and Hsmar1 transposase are 91% identical. Hsmar1 and Mos1 transposase are 37% identical with a single residue indel. Their structures are thus largely superimposable.

The crystal structure for the Mos1 post-cleavage intermediate presents a classical trans-architecture, in which the transposase subunit bound to one transposon end contributes its catalytic domain to the opposite end (Figure 6A) (Richardson et al., 2009). It also helped to clarify a long-standing question about why the WVPHEL motif is conserved given its location in an inter-domainal linker, which might be expected to have an unconstrained sequence. The structure revealed that the linker contributes a second level of complexity to the trans-architecture of the complex (Figure 1.6A and B). An extended “clamp-loop” feature emerges from the catalytic core of one transposase subunit and interacts with the transposon end and the WVPHEL motif on the other side of the complex. These interactions account for more than 70% of the dimer interface. On the opposite side, WVPHEL is in contact with the YSPDL motif, which forms part of the active site. This network of

interactions connects the active sites of both subunits and provides a potential conduit for signals.

The relative spatial orientation of the catalytic cores in the Mos1 PEC is shown in Figure 1.6D. The orientation is significantly different in the apo-structure in the absence of DNA (Figure 1.6E). In this structure the dimer interface has a small area and it may be a crystal lattice interaction with no biological significance. The orientation of the catalytic cores in SETMAR is quite different from either of the Mos1 structures (Figure 1.6F). In this case the dimer interface is extensive and one element is shown in Figure 1.6G. The highlighted residue is R469, which is homologous to R142 in Mos1 and R141 in Hsmar1. In a random screen for hyperactive Hsmar1 mutations, R141L was identified as increasing the rate of transposition by 140-fold in an *in vivo* assay (D. Liu and RC, to be presented elsewhere). This suggests that the SETMAR subunit interface may also be present in one of the transposition intermediates.

One way to interpret the hyperactivity of the R141L transposase is that the SETMAR interface is present in SEC2 and that the mutation promotes the conformational change that accompanies second end recruitment and activation for catalysis (see Figure 1.4A). However, the presence of the SETMAR dimer interface in the catalytic domain is inconsistent with a neutron scattering study of Mos1 (Cuypers et al., 2013). This study suggested that the transposase subunits have head-to-head interactions, with an extended structure that becomes even more extended in SEC2 (Figure 1.6H).

1.2.6 Miniature inverted repeat transposable elements (MITEs)

Because a freely diffusing pool of transposase acts on all copies of the transposon, autonomous transposons may be parasitized by non-autonomous transposons. These are elements that have acquired inactivating point mutations in the transposase or sizable internal deletions, which yield miniature elements. Miniature elements are often very numerous and therefore appear to have a higher rate of transposition. In the case of the deletion derivatives, the shorter distance between transposon ends probably promotes synapsis and provides a lower probability of non-productive autointegration. Although some of the miniature elements are simple deletion derivatives of their parental element, others have a more complex genesis. However, the more complex miniature elements may have sequence-features that enhance transposition (Yang et al., 2009).

The higher rate of transposition of miniature elements results in MITEs outnumbering their transposase-encoding elements in plants, nematodes, and primates (Pace and Feschotte, 2007). Therefore, MITEs provide most of the transposon binding sites in these organisms.

Recent results have shed some light on one of the mechanisms leading to the birth of MITEs (Yang et al., 2013). The presence of short microhomologies around the break points, i.e. the locations on a MITE where the DNA sequence stops being homologous to the autonomous transposon, and the presence of mismatches in some of the microhomologies are hallmark features of microhomology-mediated end joining (MMEJ) (McVey and Lee, 2008). The MMEJ repairs double-strand DNA breaks but unlike the

non-homologous end joining (NHEJ), it requires limited resection (Symington and Gautier, 2011). Thus, the break points are always flanked by deletions (McVey and Lee, 2008). The MMEJ is active throughout the cell cycle but predominates in G2 and S phases (Symington and Gautier, 2011) (Decottignies, 2007). Consequently, a double-strand DNA break left by an excised transposon during the S phase may be initially repaired by homologous recombination (Figure 1.6). After the synthesis of the left and right ITRs, the newly synthesized nascent strands may detach from the recombination complex. If microhomologous sequences are present in both synthesized strands, the MMEJ pathway can anneal both DNA strands at the microhomologies. The trimming of DNA flaps followed by DNA synthesis and ligation will produce a MITE (McVey and Lee, 2008). However, this mechanism explains the formation of only a fraction of the MITEs because several non-autonomous elements do not contain microhomologous sequences.

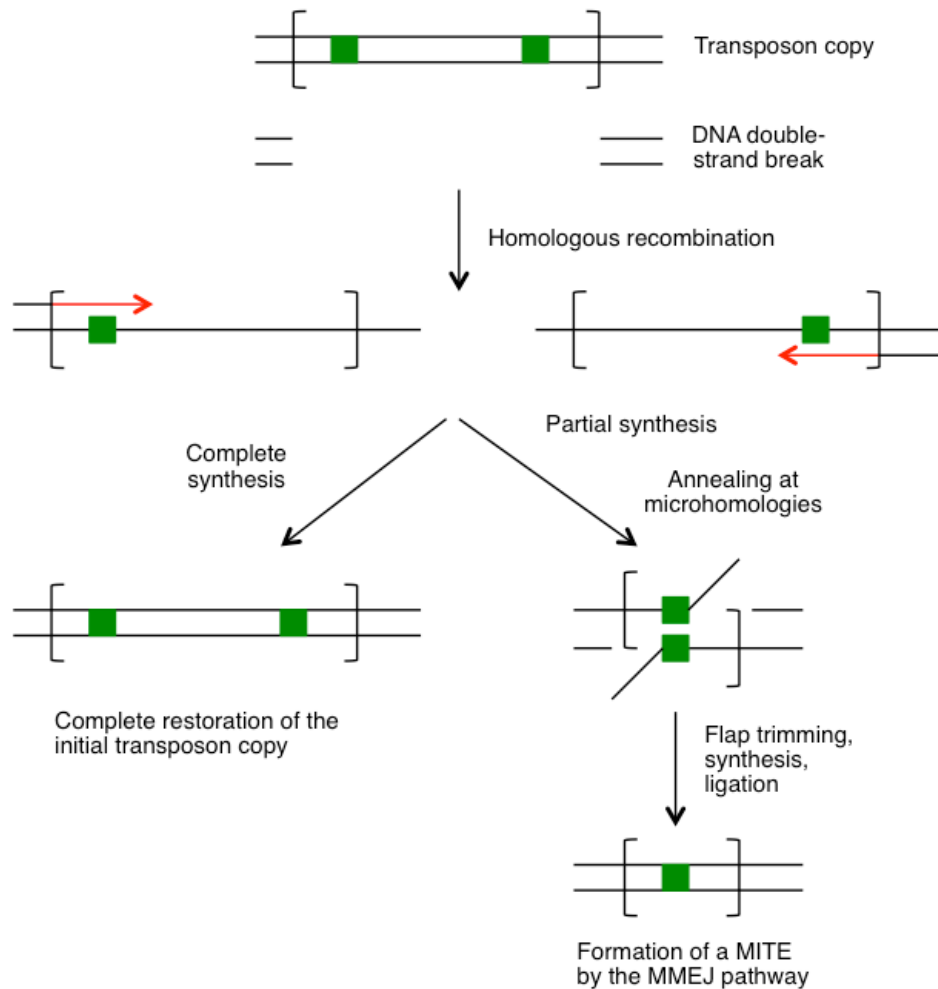


Figure 1-7: Model for the generation of some MITE families.

After the excision of a transposon copy, the double-strand DNA break is repaired by homologous recombination using another chromatid as a template. If the DNA synthesis is completed, the initial transposon copy is restored. However, if the DNA synthesis stops before finishing, the MMEJ pathway can anneal both synthesized strands at the microhomology region. The DNA flaps need to be trimmed before the gaps can be filled and ligated. The repair by the MMEJ pathway generates a deletion-derivative of the initial transposon. Open brackets, transposon ends; red arrows, DNA synthesis; green boxes, microhomology regions.

Non-autonomous and miniature elements are much rarer in bacteria than eukaryotes. One reason for this is that the cis-action of some bacterial transposases restricts their activity to the encoding element eg. (Bouuaert et al., 2014) (Duval-Valentin and Chandler, 2011). This contrasts with the

situation in eukaryotes where the separation of transcription and translation mandates a freely diffusing transposase, which facilitates the amplification of non-autonomous elements.

Transposons have been intensively studied in bacteria, where they have been shown to affect physiology and drive evolution (reviewed in (Chalmers R, 1999)). In experiments where strong selection is applied, strains possessing a *Tn10* element outcompete isogenic strains lacking the element (Chao and McBroom, 1985) (Chao et al., 1983) (Stoebel and Dorman, 2010). Similar results were obtained with the *Ty* element in yeast (Wilke and Adams, 1992). In these experiments the transposons are acting as mutators and affect the global regulatory network and the mutation rates by inserting inside or close to genes and by promoting non-homologous recombination {Gaffe, 2011 #534} {Song, 2007 #537}. The competitive advantage provided by transposons is therefore linked to their mobile nature. Genetic changes introduced by transposons are only kept in a bacterial population if they improve the host's fitness. Transposons also provide a competitive advantage in environments where antibiotics or heavy metals are present by facilitating the transfer of resistance genes in the population {Davies, 1994 #536} {Baker-Austin, 2006 #535}. It is therefore unsurprising that eukaryotic transposons, which generally constitute a much larger fraction of the genome than their bacterial counterparts, have also contributed to evolution in higher eukaryotes.

1.3 Exaptation

With the large number of whole genome sequences available it has become clear that transposons sometimes experience exaptation, or domestication. The key transitions are when a transposon-derived sequence begins to perform a 'useful' function and therefore comes under purifying selection, just like any other bona fide host gene or regulatory region eg. (Alzohairy et al., 2013) (Feschotte, 2008). A transposon may come under purifying selection simply because it has inserted near a gene and changed its expression either directly or by altering or adding regulatory elements. The transposase may also contribute functional domains to an expressed protein.

1.3.1 Co-option of regulatory sequences derived from transposable elements

The ability of transposable elements to move and replicate within the genome can facilitate the spreading of regulatory elements such as transcription factor binding sites (TFBSs) in the host genome. Indeed, numerous transposable elements contain TFBSs, or other regulatory elements, which can be directly co-opted by the host (Romanish et al., 2007; Wessler et al., 1995). A comparison of the occupancy profiles of OCT4, NANOG, and CTCF in mouse and human embryonic stem cells found that the CTCF profile was similar between both species but only ~5% of OCT4 and NANOG sites were homologously occupied (Kunarso et al., 2010). Interestingly, up to 20% of the OCT4 and NANOG binding regions overlap with species-specific transposable

elements. However, only a fraction of these binding regions have been shown to directly influence gene expression (Kunarso et al., 2010). Furthermore, it has been reported that new regulatory elements can emerge *de novo* by mutations from the inactive TE copies present in the genome (Medstrand et al., 2005; Zhou et al., 2002). In both cases, the dispersal of the TE copies can result in the recruitment at several chromosomal locations of the same motif(s) by the host. This has the potential to promote the emergence of new regulatory networks by putting genes under the same set of regulatory controls (Johnson et al., 2006; Wang et al., 2007).

In addition to TFBSs, transposable elements are also a source of promoter and alternative promoter sequences (Jordan et al., 2003; van de Lagemaat et al., 2003). Indeed, a quarter of the human promoters contain a TE-derived sequence. The SINE and LINE elements are predominant in the promoter regions and represents 87% of the TEs (Jordan et al., 2003). The TE-derived sequences are essentially located a few hundred base pairs away from the transcription start sites (Jordan et al., 2003).

Although TEs sequences are seldom present in the exons of protein-coding genes, they are particularly enriched in long non-coding RNAs (lncRNAs) (Gotea and Makalowski, 2006; Kapusta et al., 2013). Indeed, in humans they account for ~30% of the total lncRNAs sequences and are found in more than two thirds of the mature lncRNA transcripts (Kapusta et al., 2013). In total TEs may provide up to 30,000 regulatory signals, which are distributed between transcriptional initiation sites, splicing sites, and polyadenylation signals, necessary for the biogenesis of lncRNAs in humans (Kapusta et al., 2013).

Another notable example of transposon exaptation is the internally eliminated sequences (IES), which interrupt thousands of genes in ciliated protozoans. The IESs are highly degenerate, but they are all flanked by a TA dinucleotide, and many appear to be derived from an IT_m transposon (Arnaiz et al., 2012). Interestingly, IESs are excised by a domesticated PiggyBac transposon (Baudry et al., 2009). This is important because PiggyBac excision yields complementary overhangs, which can be ligated without leaving a footprint. This means that exonic IES insertions are not counterselected in ciliates.

1.3.2 Exaptation of MITEs

The MITEs are non-autonomous elements, which are composed of essentially two ITRs separated by a central region. Their small sizes compared to their autonomous counterparts allow them to form stable hairpin loops when expressed as RNA. The RNA interference enzymatic machinery can recognize and process transcribed MITEs to produce microRNAs (miRNAs) (Piriyapongsa and Jordan, 2008). Consequently, transcribed MITEs are recurrently exapted as miRNA genes (Piriyapongsa and Jordan, 2007). Indeed, in an analysis of the human genome it was found that 15% of the miRNA genes originate from a transposable element (Borchert et al., 2011).

miRNAs are genome-encoded small RNAs, which regulate gene expression at the post-transcriptional level. It is estimated that more than half of the human mRNAs are targeted by miRNAs, with each miRNA regulating up to a hundred transcripts (Gurtan and Sharp, 2013). In human, half of the

miRNAs are located in genes while the other half is found in intergenic regions (Hsu et al., 2006).

The intragenic miRNAs are essentially intronic and are directly processed from the intron of the host transcript (Gurtan and Sharp, 2013). Thus, the miRNAs are coordinately expressed with their host genes. The intergenic miRNAs are transcribed from their own promoter and can be found as a single gene or as a cluster of genes.

The miRNAs are transcribed by the RNA polymerase II into long primary transcripts (pri-miRNAs) that are processed by the RNase III enzyme Drosha in precursor miRNAs (pre-miRNAs) (Lee et al., 2003). The pre-miRNAs are then transported to the cytoplasm where the RNase III enzyme Dicer processes them to yield a mature miRNA (Hutvagner et al., 2001). Within the RNA-silencing complex (RISC), the miRNA is loaded into an Argonaute protein to guide the complex to target mRNAs through a specific interaction with its “seed” sequence, which spans positions 2-7 of the miRNA (Pasquinelli et al., 2000).

The miRNAs are regrouped into families based on the mature miRNA sequence and the structure of pre-miRNAs (Kozomara and Griffiths-Jones, 2011). Members of the same family have a similar “seed” sequence therefore they regulate the expression of mRNAs in a synergistic manner. Only a partial complementarity sequence is required between the miRNA “seed” sequence and the target mRNA, thus explaining the large number of mRNA targets for each miRNA (Bagga et al., 2005).

The Made1 elements are 80 bp long MITEs, which are deletion-derivative of Hsmar1 composed of two 37 bp ITRs separated by a 6 bp central

spacer (Morgan, 1995). In the human genome, the mir-548 family currently contains 69 miRNAs derived from Made1 elements but 116 miRNAs have been defined in primates (Figure 1.7) (Kamanu et al., 2013). Since the different miRNAs emerged from the same type of non-autonomous transposon, their sequences differ only in a few positions from each other. Several members of the mir-548 family act in synergy and have been characterized as regulators of a tumour-suppressor gene, FHIT, and the type III interferon gene, IFN- λ 1 (Hu et al., 2014) (Li et al., 2013b). A bioinformatics analysis of the predicted mir-548 targets found a subset of mRNAs containing a Made1 element within or close to their 3'UTR (Piriyapongsa and Jordan, 2007). This finding, if experimentally demonstrated, suggests that the exaptation of MITEs could lead to the establishment of specific post-transcriptional regulatory networks.

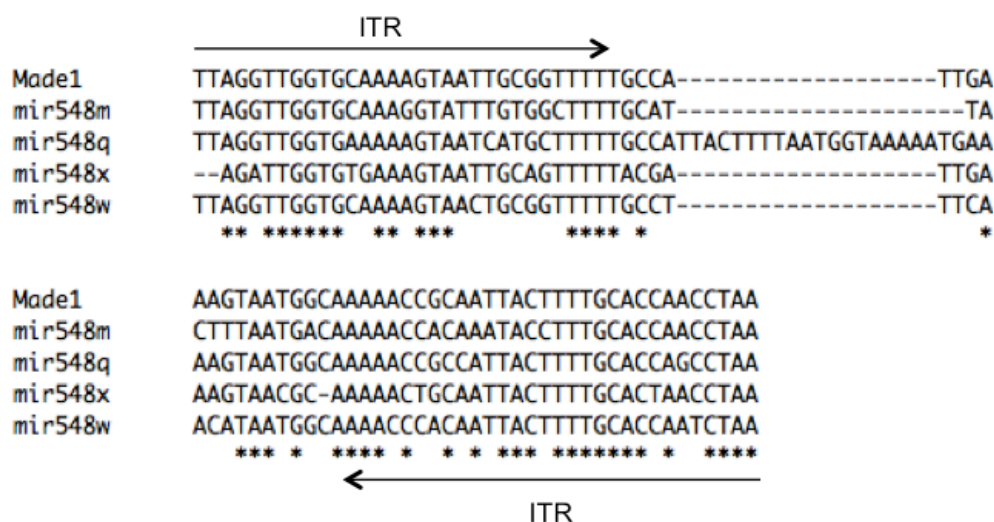


Figure 1-8: Alignment between a Made1 element and four miRNAs derived from Made1.

Multiple sequence alignment with Clustal Omega between Made1 and four miRNAs from the mir548 family. Each miRNA contain substitutions in the ITRs, which should decrease the affinity of SETMAR for the binding site. The consensus ITR sequence is shown in the Made1 element.

1.3.3 Domestication of the transposase gene

Transposable elements are also a source of new protein-coding genes (Feschotte, 2008). Indeed, new eukaryotic host genes have been recruited from a variety of TEs such as transposases, integrases, reverse transcriptases, Gag, and envelope proteins (Volff, 2006). It appears that the DNA transposases are the major source of exapted genes (Volff, 2006). The trend is particularly clear in human where 43 of the 47 domesticated genes are derived from DNA transposons despite the fact that the retrotransposons are many fold more numerous (Feschotte, 2008; Lander et al., 2001). The recurrent domestication of the DNA-binding domain of DNA transposases, which can provide a specific DNA binding activity and/or a dimer interface, might explain this overrepresentation {Quesneville, 2005 #529}.

The domesticated transposases can be classified in three categories (Sinzelle et al., 2009). The first category is when the transposase performs a new function such as the centromeric protein B (CENP-B) or the recombination-activating gene products RAG1/2 (Masumoto et al., 2004) (Schatz, 1999). The second category includes the chimeric genes, which result from a fusion between the transposase gene and a pre-existent gene. The major representatives of this group are SETMAR and the Cockayne syndrome Group B gene (CSB) (Cordaux et al., 2006) (Newman et al., 2008). The latter category comprises the chimeric genes containing only a transposase domain. The double HTH motif of the PAX proteins is, for

example, derived from the DNA-binding domain of an ancestral Tc1-like transposon (Breitling and Gerber, 2000).

Domesticated transposases originate from several DNA transposon families (Alzohairy et al., 2013). The functions of the majority of the exapted transposases are currently unknown but they are generally assumed to interact with DNA. The RAG1/2 genes, which are involved in V(D)J recombination in jawed vertebrates are derived from a Transib superfamily member (Kapitonov and Jurka, 2005). The *Arabidopsis thaliana* Daysleeper protein is a domesticated hAT transposase and acts as an essential transcription factor in plant development (Bundock and Hooykaas, 2005). Another important domesticated protein is CENP-B from the Pogo family, which specifically binds to a 17-bp sequence, called CENP-B box, to facilitate the assembly of nucleosomes at the centromeres (Yoda et al., 1998) (Smit and Riggs, 1996).

Transposases from the PiggyBac superfamily have also been domesticated several times (Pavelitz et al., 2013). In ciliates, the domesticated piggyMac transposases are involved in programmed genome rearrangements by excising the internally eliminated sequences (IES) (section 1.3.1) (Arnaiz et al., 2012) (Baudry et al., 2009). In humans, there are five domesticated PiggyBac transposases but only two have been characterized. The PGBD3 piggyBac transposon is inserted in an intron of the CSB gene (Newman et al., 2008). Alternative splicing of the CSB gene produces three proteins: CSB, CSB-PGBD3, and PGBD3. The PGBD3 domain of CSB-PGBD3 is able to bind its transposon ends *in vivo* (Gray et al., 2012). However, the fusion protein is also tethered to c-Jun motifs where the CSB

domain will specifically regulate gene expression by remodeling the chromatin (Gray et al., 2012). The domesticated PGBD3 domain is speculated to hinder the activity of the CSB gene, which results in the induction of an interferon response to help fight a viral infection (Weiner and Gray, 2013). The second characterized PiggyBac transposase is PGBD5, a highly conserved vertebrate-specific gene (Pavelitz et al., 2013). The function of the protein is currently unknown but its expression is located in specific areas of the brain and central nervous system. However, the protein is speculated to be inactive and fails to associate with both DNA and chromatin *in vivo* (Pavelitz et al., 2013).

1.3.4 SETMAR

The best-known example of mariner exaptation is the SETMAR protein, which is specific to anthropoid primates. It is a fusion between the Hsmar1 transposase and SET, which is a histone methyltransferase gene (Robertson and Zumpano, 1997) (Cordaux et al., 2006). The mechanism of the fusion involved several events which took place 40 to 58 Mya (Cordaux et al., 2006). Subsequent to the integration of an Hsmar1 transposon downstream the SET gene, is the insertion of an AluSx retrotransposon in the 5' transposon end resulting in the immobilization of the Hsmar1 copy. The ancestral stop codon of the SET gene has then been lost due a 27 bp genomic deletion, which allowed an extension of the second exon of the SET gene to a cryptic donor splicing site located 77 bp downstream. The presence of a consensus acceptor-splicing site 3 bp upstream of the start codon of the Hsmar1

transposase allowed the exonization of the transposase gene resulting in the formation of SETMAR.

The transposase domain is 94% identical to the Hsmar1 consensus sequence. However, the third active site D residue is substituted with an N residue, which abolishes the catalytic activity of the transposase. Although some activity, particularly 5'-end nicking, is recovered *in vitro* in the presence of DMSO and Mn^{2+} , this likely has no physiological significance (Liu et al., 2007). Nevertheless, SETMAR does retain robust transposon end binding and synapsis activity, which may be important to its function (Liu et al., 2007). The synapsis activity implies that the protein is a multimer, which is supported by the crystal structure of the transposase domain (Goodwin et al., 2010).

The SET domain is 33 to 35% identical to SUV39H1 and EHMT2, two members of the SUV39 family. Except for SETMAR, all the members of this family methylate the lysine 9 of histone H3 (H3K9me) (Schneider et al., 2002). This epigenetic mark is known to generate repressive heterochromatin (Towbin et al., 2012). However, SETMAR dimethylates the histone H3 lysine 36 (H3K36me₂) *in vitro* and *in vivo* (Lee et al., 2005) (Fnu et al., 2011).

The human genome is currently thought to encode eight proteins that methylate H3K36 (Wagner and Carpenter, 2012). Except for SETD2, which adds the third methyl group to a dimethylated K36, all the enzymes mono- or dimethylate H3K36 (Wagner and Carpenter, 2012). H3K36 methylation is associated with active chromatin and involved in several cellular processes such as transcriptional repression, DNA replication and repair, alternative splicing, and dosage compensation (Wagner and Carpenter, 2012). The output of the H3K36 methylation is therefore defined by its location, its

temporality, and which reader proteins bind to the epigenetic mark.

Some early experiments with SETMAR revealed that it was involved with DNA repair (Lee et al., 2005). Subsequent papers from this group attempted to dissect the contribution of SETMAR's methylase and transposase domains but failed to provide a coherent model for the activity of the protein. One of the difficulties in understanding the function of the protein is that it produced a response in a number of different assays, suggesting that it was involved in many different aspects of DNA metabolism. For example, its overexpression promotes classical non-homologous end joining (NHEJ), the random integration of transfected plasmid DNA and the restart of stalled replication forks (Lee et al., 2005) (Hromas et al., 2012). This group's *in vitro* analysis also contradicted Liu et al by showing that purified SETMAR is an endonuclease, at least *in vitro*, like Artemis (Beck et al., 2011a; Mohapatra et al., 2013). The MAR domain, which corresponds to the SETMAR transposase domain, trims damaged and undamaged DNA overhangs to form DNA ends that can be process by DNA ligase IV, another interacting partner of SETMAR (Hromas et al., 2008). SETMAR endonuclease activity has only been established *in vitro* and recent papers question its relevance *in vivo* (Mohapatra et al., 2013) (Rath et al., 2014). In contrast to Artemis, which promotes both trimming of DNA overhangs and DNA repair in cell extract assays, SETMAR did not stimulate DNA repair and only promotes trimming in one assay.

Hromas' group also discovered SETMAR interacting-partners. The SET methylase-domain was shown to interact with PRPF19, also known as PSO4, which is a protein involved in the classical NHEJ and the spliceosome (Beck

et al., 2008). This interaction was predicted to target SETMAR to double-strand DNA breaks where the SET domain could dimethylate the histone H3 lysine 36 of neighboring nucleosomes (Fnu et al., 2011). This epigenetic mark recruits and stabilizes the anchoring of Ku70 and NBS1, two early acting NHEJ factors, to the DNA ends (Fnu et al., 2011). Two papers recently established a role for H3K36me2 and H3K36me3 in double-strand DNA breaks by the NHEJ in *S. pombe*, supporting an evolutionary conserved function for these epigenetic marks in DNA repair (Jha and Strahl, 2014; Pai et al., 2014).

Hromas' group also showed that the DNA repair and replication restart activities of SETMAR are oppositely regulated by phosphorylation (Hromas et al., 2012). The balance between the functions is dependent on whether Chk1 phosphorylates serine 508 (S508), with the phosphorylated form favoring DNA repair (Williamson et al., 2014). In both papers, the residue is numerated S495 rather than S508 because the authors worked with a SETMAR variant missing the first 13 amino acids. Phosphorylation of the S508 has also been detected in three large-scale phosphoproteomic analyses (Mayya et al., 2009) (Olsen et al., 2010) (Rigbolt et al., 2011). The S508 is located in the MAR domain of SETMAR and this position is conserved in mariner transposases (Bouchet et al., 2014). Intriguingly, the mariner transposases, except for the *irritans* subfamily, and SETMAR contain a PKA and not a Chk1 consensus site in the vicinity of the S170 (or S508 for SETMAR) (Bouchet et al., 2014). For Mos1, the phosphorylation of the S170 residue inhibits the formation of the PEC and traps the protein in the cytoplasm (Bouchet et al., 2014). The phosphorylation of S508 should not hinder the nuclear importation of

SETMAR because a nuclear localization signal is also located in the SET domain. However, it could regulate SETMAR site-specific binding activity.

The role of SETMAR in the restart of stalled replication forks is unclear (De Haro et al., 2010). The DDN motif of the transposase domain and the SET domain appears to be essential but their specific functions have not been defined (De Haro et al., 2010; Kim et al., 2014). The requirement of the DDN motif could suggest a role for the MAR domain nicking activity. This model states that SETMAR could promote the re-start of stalled replication forks by nicking and decatenating the positively supercoiled DNA in front of the fork to release the topological tension (De Haro et al., 2010; Kim et al., 2014). It is supported by the promotion of the decatenation activity of Topoisomerase II α when it interacts with SETMAR (Williamson et al., 2008b). Topoisomerase II α is a highly conserved enzyme (Ramamoorthy et al., 2012). It relieves topological stress by catalysing a transient double-strand DNA break, then passing another double-stranded DNA through the break before re-joining the two DNA ends (Wang, 2002). SETMAR promotion of Topoisomerase II α decatenating activity is regulated by the auto-methylation of the residue K498 (Williamson et al., 2008b). The methylation of this lysine by the SET domain inhibits the promotion of chromosome decatenation (Williamson et al., 2008b). The mechanism underlying the auto-regulation is unknown but could indicate a regulation of the ancestral Hsmar1 transposase activity by the ancestral SET gene.

Two papers reported a correlation between SETMAR expression and the random integration efficiency of plasmid and lentivirus DNAs in the genome (Lee et al., 2005; Williamson et al., 2008a). Illegitimate DNA

recombination and lentivirus cDNA integration are dependent of the NHEJ pathway but the mechanism responsible for plasmid integration, which cannot rely on an integrase, remains uncertain (Wurtele et al., 2003) (Li et al., 2001). The current model states that the circular plasmid needs to be linearized by a DNA double-strand break for recruiting DNA repair proteins on the plasmid ends. For genomic integration to happen, one plasmid end needs to be in the vicinity of a genomic lesion for the NHEJ proteins to repair the genomic double-strand break by integrating the plasmid DNA (Wurtele et al., 2003).

The New Mexico group has also reported that SETMAR promotes cell proliferation in the HEK293 cell line and the suppression of chromosomal translocations (De Haro et al., 2010; Williamson et al., 2008b; Wray et al., 2010). However, they do not suggest a mechanism by which SETMAR might mediate these effects.

1.4 Aims of the research

The primary aim of my research is to illuminate the function of SETMAR in human cells. My hypothesis is based on the observation that the residues 342 to 512 of SETMAR are under purifying selection in the anthropoid primate lineage (Cordaux et al., 2006). This was established by comparing for the MAR domain the rate of non-synonymous (K_A) versus synonymous (K_S) nucleotide substitutions per site in each anthropoid primate. Only the DNA binding domain displayed a K_A/K_S ratio significantly inferior to 1, which signifies a strong signal of purifying selection. The pattern of amino acid substitutions in the C-terminal catalytic-domain of the transposase suggests

that it is experiencing genetic drift and that its evolution is therefore neutral. I therefore wanted to test the hypothesis that SETMAR acts as a transcriptional regulator by targeting the methylase activity of the SET domain to specific sites in the genome, which correspond to the location of Hsmar1 transposon ends. This is a significant challenge because there are about 6,000 transposon ends in the human genome. Although about half of them are located in the non-coding regions of genes, it is probably that many lack significance. My aim was therefore to identify sites at which SETMAR has the strongest effects.

A subsidiary aim was to try to clarify some of the confusing results from Hromas' group by addressing whether SETMAR or its domains are involved in NHEJ and plasmid integration (Lee et al., 2005). The SET domain is an ancient gene specific to bilaterian animals. Therefore, it should have one or more cellular functions, which could have been modified upon its fusion to the Hsmar1 transposase. It has been recently established that the H3K36me3 mark is involved in DNA repair (Pai et al., 2014). Interestingly, H3K36me3 has an opposite effect in yeast and in human where it promotes NHEJ and homologous recombination (HR), respectively {Jha, 2014 #538}. Thus, I investigated whether the SET gene was involved in DNA repair and whether its fusion to the MAR domain affects its activity.

The long turnaround time for my cell culture and deep sequencing experiments gave me the opportunity to pursue a side project on Hsmar1 transposition per se. I used a bacterial genetic screen and a range of constitutive promoters to separate the negative effects of OPI from the underlying rate of transposition. These two aspects are intimately linked by

the molecular mechanism of the transposition reaction (Claeys Bouuaert et al., 2013). The investigation of the transposition rate of a mutant transposase is therefore a combination of the consequence of the mutation on the transposase activity and of its effect on the sensibility of the transposase to OPI. Thus, my hypothesis was that library screening for hyperactive mutants would be improved if the transposase expression could be reduced below the level at which OPI is the dominant factor.

The work is structured as follows:

Chapter 3 presents a bioinformatics analysis of Hsmar1 remnants in the human genome, which is a prerequisite for interpretation of the subsequent RNAseq experiments. I report the distributions of Hsmar1, Made1 elements, solo ITRs, and putative SETMAR binding sites. Made1 and solo ITRs are the major source of transposon ends in the human genome. The ITRs are mostly located in genes and a fraction of the binding sites may be under purifying selection. I discovered that in addition to the 69 human miRNAs already reported to be derived from Made1, there are 456 additional elements that could also act as miRNAs.

Chapter 4 explores the role of SETMAR as a genome-wide epigenetic regulator. I used RNA-Seq to establish that a significant number of genes containing a transposon end are differentially expressed when SETMAR expression is altered. I also used ChIP-Seq to assess SETMAR binding directly. Although the SETMAR binding sites were significantly enriched for sequences encoding transposon ends, they did not strongly overlap with the differentially expressed genes in the RNAseq experiments. The ChIP-Seq

revealed that the strongest SETMAR binding sites are Made1 elements and the putative miRNAs discovered in the Chapter 3. I therefore propose a model in which SETMAR may act by two different mechanisms. SETMAR might directly regulate the expression of the putative miRNAs. It could also regulate the expression of the genes with an ITR by interacting with other regulatory proteins or recruiting them by dimethylating the H3K36.

Chapter 5 explores the role of SETMAR and its domains in NHEJ and plasmid integration. From a series of *in vivo* experiments, I established that expression of the SET domain alone promotes random integration of plasmid DNA, but not DNA repair. In contradiction to the results published by Hromas' group, I found that this activity of the SET domain is lost upon its fusion to the MAR domain. I suggest that this loss of function is owing to the dimerization mediated by the transposase domain.

Chapter 6 presents the refinement of a bacterial genetic screen to discover hyperactive transposase variants. From the analysis of a set of constitutive promoters, I select three expression points and verify that they are sufficient for finding auto-regulated hyperactive transposases by testing several Hsmar1 mutants.

Finally, Chapter 7 summarizes the findings and suggests directions of future researches. An Appendix section (Chapter 8) contains supplementary data and a formatted version of the article published during this work.

2. MATERIALS AND METHODS

2.1 Strains and cell lines

NEB5α	endA1 hsdR17 (rk-mk-) supE44 thi-1 recA1 gyrA (Nal ^r) relA1d(lacIZYA-argF) U169deoR (phi80dlacd(lacZ)M15)
RC5094	F- araD139 Δ(argF-lac)U169 rspL150 relA1 flbB5301 fruA25 deoC1 ptsF25 rpoS359::Tn10.
RC5096	F- fhuA2 Δ(lacZ)r1 glnV44 e14-(McrA-) trp-31 his-1 rpsL104 xyl-7 mtl-2 metB1 Δ(mcrC-mrr)114::/S10 argE::Hsmar1-lacZ'-kanR
RC5097	= RC5096 pOX38::miniTn10-CAT
RC5118	F- fhuA2 Δ(lacZ)r1 glnV44 (= amber sup) e14-(McrA-) trp-31 his-1 rpsL104 xyl-7 mtl-2 metB1 Δ(mcrC-mrr)114::/S10 argE::Sleeping Beauty-Piggy Bac-Hsmar1-lacZ'-kanR
RC5122	= RC5118 pOX38::miniTn10-CAT

Table 2-1: *Escherichia coli* strains.

T-Rex-U2OS	Human osteosarcoma cell line stably expressing the tetracycline repressor protein.
T-Rex-U2OS-TO	T-Rex-U2OS cell line stably transfected with an empty pcDNA4TO.
T-Rex-U2OS-SETMAR	T-Rex-U2OS cell line stably expressing SETMAR.
T-Rex-U2OS-TO-FLAG	T-Rex-U2OS cell line stably transfected with an empty pcDNA4TO-FLAG.
T-Rex-U2OS-SET-FLAG	T-Rex-U2OS cell line stably expressing the FLAG-tagged exons 1 and 2 of SETMAR (= SET domain).
T-Rex-U2OS-MAR-FLAG	T-Rex-U2OS cell line stably expressing the FLAG-tagged exon 3 of SETMAR (= MAR domain).
T-Rex-U2OS-SETMAR-FLAG	T-Rex-U2OS cell line stably expressing a FLAG-tagged SETMAR.
T-Rex-U2OS-SETMAR R432A-FLAG	T-Rex-U2OS cell line stably expressing a FLAG-tagged SETMAR with the mutation R432A, which decreases the affinity of SETMAR for the transposon end.
T-Rex-U2OS-SETMAR D483A-FLAG	T-Rex-U2OS cell line stably expressing a FLAG-tagged SETMAR with the D483A mutation, which abolishes the catalytic activity of transposase domain of SETMAR.

Table 2-2: Mammalian cell lines.

2.2 Oligonucleotides and plasmids

Name	Sequence	Description / Use
MT01	GCGGAATTCATGGCCGAGTTC	Add EcoRI site on the 5' of mSETMAR for cloning in pcDNA4 TO
MT02	GATAATTGCGGCCGCTCAGTCG	Add NotI site on the 3' of mSETMAR for cloning in pcDNA4 TO
MT03	gacctGATATCaagactacaaggacgacgat gacaagtaaTCTAGAtagc	Construction of 3x-FLAG to
MT04	gctaTCTAGAttactgtcatcgctgcctttagt cttGATATCaggtc	

MT05	gacctGATATCAAgactacaaggacgacgat gacaagGAATTCgatc	integrate into pcDNA4TO
MT06	gatcGAATTCctgtcatcgtcgtcctttagtcT TGATATCaggtc	
MT07	gatcCAATTGgattataaagatgatgacgataa aAGATATCtagc	
MT08	gctaGATATCTttatcgtcatcatctttataatcC AATTGgatc	
MT09	GATGaagcttATGGCCGAGTTCAAAG AG	Oligonucleotides to clone the mouse codon optimized SETMAR, SET, and MAR in pcDNA4TO:3xFLAG
MT10	GCATgaattcGTCGAAGTAGCTGCCG	
MT11	GCGGaagcttATGAAGATGATGCTGG	
MT12	GCATgaattcGGTTTCCAGGGTCAGC C	
MT13	gccCACCTGAAGCAGATCGGC	Oligonucleotides to insert the R432A mutation in SETMAR
MT14	GACGACGGTGGAGTGGTTCACG	
MT15	GCAGGTCACGATCCGGTCC	Oligonucleotides to insert the D483A mutation in SETMAR
MT16	gccGAGAAGTGGATTCTGTACGAC	
MT17	CAGAATTCGCATGAGCTCCAG	Add EcoRI site on the 5' of the constitutive promoter-Hsmar1 construction, for cloning in pBS
MT18	CTTCTAGACAGGGTACCCAGC	Add XbaI site on the 3' of the Constitutive promoter-Hsmar1 construction, for cloning in pBS
MT19	CATCTAGAGCATGAGCTCCAG	Add XbaI site on the 5' of the Constitutive promoter-Hsmar1 construction, for cloning in pGHM491
MT20	CTGAATTCCAGGGTACCCAGC	Add EcoRI site on the 3' of the Constitutive promoter-Hsmar1 construction, for cloning in pGHM491
MT21	CAACTAGTGCATGAGCTCCAG	Add SpeI site on the 5' of p1- Hsmar1 construction, for cloning in pGHM491
MT22	GGTGGCGAACCACAGTAGAG	Oligonucleotides to sequence the

		promoter p1 to p6 in pGHM491
MT23	CGAGTcaattgGCATGAGCTCCAG	Cloning of p1 to p6-Hsmar1 in pBACe3.6
MT24	CGAGTcaattgCAGGGTACCCAGC	
MT25	GTGGTTCcAAAAGTTTCGCAAAGG CG	Oligonucleotides to insert the K50Q mutation in Hsmar1
MT26	CATTGCACGGTACGCTCGTTTCG	
MT27	CACATTCAGTTCTTCAGCAACTTC	Oligonucleotides to insert the D98N mutation in Hsmar1
MT28	aacCACTCTACTGTGGTTTCGCC	
MT29	ACGACGGTTTTTCTGGTTTTTCAG	Oligonucleotides to insert the F132A mutation in Hsmar1
MT30	gctGAAGTTAGCTCCTCTCTG	
MT31	CGAGGGAAGGATTTCAGaattcCATA TG	Oligonucleotides for constructing a mutant library for Hsmar1 and PiggyBac transposases
MT32	GCAGGTCTGACTCTAGAGGATCCTT A	

Table 2-3: Oligonucleotides used in this study.

	R.C. number	Other name	Description / Use
1	pRC1701		pcDNA4TO
2	pRC1702		Mouse codon optimized <i>SETMAR</i> cloned in pcDNA4TO between EcoRI and NotI
3	pRC1703		3xFLAG cloned in pRC1701 between EcoRI and XbaI
4	pRC1704		Mouse codon optimized <i>SETMAR</i> cloned in pRC1703 between HindIII and EcoRI
5	pRC1705		Mouse codon optimized <i>SET</i> cloned in pRC1703 between HindIII and EcoRI
6	pRC1706		Mouse codon optimized <i>MAR</i> cloned in pRC1703 between HindIII and EcoRI
7	pRC1707		R432A mutation in pRC1704
8	pRC1708		D483A mutation in pRC1704
9	pRC1709	pEGFP-Pem1- Ad2	Vector containing an inactivated <i>EGFP</i> gene used for following NHEJ repair in human cells
10	pRC1710	pRFP	Vector containing a <i>RFP</i> gene used as a transfection control.

11	pRC1711		pBluescript SKII+
12	pRC1712	pMM2	Neomycin resistance gene cloned in pRC1711 in BamHI
13	pRC1713		pMAL-c2x
14	pRC1714		pGHM491
15	pRC1715		pcDNA3.1+
16	pRC1716		pBACe3.6
17	pRC1717	pRC880	<i>E. coli</i> codon optimized <i>Hsma1</i> cloned in pRC1713 between NdeI and BamHI (MBP-tag +)
18	pRC1718	pRC1231	W118R mutation in pRC1717
19	pRC1719		pRC1717 without the MBP-tag
20	pRC1720		pRC1718 without the MBP-tag
21	pRC1721		Vector containing the constitutive promoters p1 and p2
22	pRC1722		Vector containing the constitutive promoters p3 to p6
23	pRC1723		pRC1711 with p1-Hsma1 construct
24	pRC1724		pRC1711 with p2-Hsma1 construct
25	pRC1725		pRC1711 with p3-Hsma1 construct
26	pRC1726		pRC1711 with p4-Hsma1 construct
27	pRC1727		pRC1711 with p5-Hsma1 construct
28	pRC1728		pRC1711 with p6-Hsma1 construct
29	pRC1729		pRC1714 with p1-Hsma1 construct (Ip1)
30	pRC1730		pRC1714 with p2-Hsma1 construct (Ip2)
31	pRC1731		pRC1714 with p3-Hsma1 construct (Ip3)
32	pRC1732		pRC1714 with p4-Hsma1 construct (Ip4)
33	pRC1733		pRC1714 with p5-Hsma1 construct (Ip5)
34	pRC1734		pRC1714 with p6-Hsma1 construct (Ip6)
35	pRC1735		pRC1714 with p1-Sleeping Beauty construct (Ip1)
36	pRC1736		pRC1714 with p1-PiggyBac construct (Ip1)
37	pRC1737		pRC1714 with p2-PiggyBac construct (Ip2)
38	pRC1738		pRC1715 with p1-Hsma1 construct (pcDNA3.1-p1)

39	pRC1739		pRC1715 with p2-Hsmar1 construct
40	pRC1740		pRC1715 with p3-Hsmar1 construct
41	pRC1741		pRC1715 with p4-Hsmar1 construct
42	pRC1742		pRC1715 with p5-Hsmar1 construct
43	pRC1743		pRC1715 with p6-Hsmar1 construct (pcDNA3.1-p6)
44	pRC1744		pRC1716 with p1-Hsmar1 construct (Bp1)
45	pRC1745		pRC1716 with p2-Hsmar1 construct (Bp2)
46	pRC1746		pRC1716 with p3-Hsmar1 construct (Bp3)
47	pRC1747		pRC1716 with p4-Hsmar1 construct (Bp4)
48	pRC1748		pRC1716 with p5-Hsmar1 construct (Bp5)
49	pRC1749		pRC1716 with p6-Hsmar1 construct (Bp6)
50	pRC1750		K50Q mutation in pRC1744
51	pRC1751		D98N mutation in pRC1744
52	pRC1752		R104A mutation in pRC1744
53	pRC1753		F132A mutation in pRC1744
54	pRC1754		R141L mutation in pRC1744
55	pRC1755		K50Q / R141L mutations in pRC1744
56	pRC1756		D155A mutation in pRC1744
57	pRC1757		K50Q mutation in pRC1729
58	pRC1758		D98N mutation in pRC1729
59	pRC1759		R104A mutation in pRC1729
60	pRC1760		F132A mutation in pRC1729
61	pRC1761		R141L mutation in pRC1729
62	pRC1762		K50Q / R141L mutations in pRC1729
63	pRC1763		D155A mutation in pRC1729
64	pRC1764		K50Q mutation in pRC1734
65	pRC1765		D98N mutation in pRC1734
66	pRC1766		R104A mutation in pRC1734
67	pRC1767		F132A mutation in pRC1734
68	pRC1768		R141L mutation in pRC1734
69	pRC1769		K50Q / R141L mutations in pRC1734
70	pRC1770		R141L mutation in pRC1743

Table 2-4: Plasmids used in this study.

2.3 Chemicals

Chemicals were generally of the highest quality available. Most chemicals were purchased from Sigma, unless stated otherwise. Restriction enzymes, T4 DNA ligase, T4 DNA polymerase, Vent polymerase, and molecular weight marker (DNA and protein) were from New England BioLabs (NEB). DNA preparation and purification kits were from QIAGEN. Cell culture media were from Sigma and selective reagents from Invitrogen.

2.4 General DNA manipulation

2.4.1 Polymerase Chain Reaction (PCR)

For cloning, DNA was amplified with the Vent polymerase (NEB) in 1x ThermoPol reaction buffer with 200 nM of each primer (Table 2.2), 200 μ M dNTPs, 10 ng of template, and 1 unit of enzyme per 50 μ l. After an initial 5 min denaturation step at 95°C, the DNA was denatured 30 s at 95°C, annealed 30 s at the T_m calculated from the NEB website, and elongated for 1 min per kb of product at 72°C, for 30 cycles, followed by a final 5 min elongation step. The PCR products were separated on an agarose gel and purified with the QIAquick gel extraction kit (Qiagen).

2.4.2 Plasmid construction

Plasmids DNA and purified PCR products were digested with restriction enzymes from NEB in the chosen restriction buffer for 1 h at the recommended temperature. The restriction fragments were purified with the QIAquick PCR purification kit (Qiagen) or separated on agarose gel and purified with the QIAquick gel extraction kit (Qiagen). If blunt-end DNA was required, 5' overhangs were filled and 3' overhangs removed with T4 DNA polymerase (NEB) according to manufacturer's instructions. The vector backbone was dephosphorylated with calf intestinal phosphatase (CIP, NEB) for 1 h at 37°C. Ligations were performed with T4 DNA ligase (NEB) in the supplied buffer at 16°C overnight prior to transformation in chemically competent NEB5 α cells. Plasmids DNA were purified with the Qiagen miniprep kit. Plasmids constructed with a PCR step were confirmed by sequencing.

2.5 Cell culture

2.5.1 Media and growth conditions

The T-Rex-U2OS cell lines were maintained in complete Dulbecco's modified Eagle's medium (DMEM, Sigma) supplemented with 10% heat inactivated Fetal Bovine Serum (FBS), 100 u/ml of streptomycin, 100 μ g/ml of penicillin, and 5 μ g/ml of blasticidin at 37°C with 5% CO₂.

The medium of T-Rex-U2OS cell lines stably expressing a gene of interest from an integrated pcDNA4TO plasmid was supplemented with 400 µg/ml of zeocin. Cell lines were passaged at 90% confluency by a trypsin/EDTA treatment and ~ 10% of the cells seeded to a new 75 cm² flask. Cell lines were kept for 10 passages before a new culture was set up from a frozen aliquot. Conservation of the cell lines was done by resuspending near confluent 100 mm dishes in 2 ml of freezing medium (50% complete medium, 40% FBS, 10% DMSO), freezing the tubes 24 hours at -80°C, and keeping them in the vapor phase of liquid nitrogen for long-term storage.

2.5.2 Transient transfection of the T-Rex-U2OS cell line

For each transfection, cells were seeded in a 6-well plate and grown overnight in DMEM supplemented with 10% FBS. The plasmids were transfected using Lipofectamine 2000 (Invitrogen), following manufacturer's instruction. Cells were retrieved 24 hours after transfection and processed as needed.

2.5.3 Stable transfection of T-Rex-U2OS cells

For each transfection, 2.5×10^5 of cells were seeded in a 6-well plate and grown overnight in DMEM supplemented with 10% FBS. The plasmids were transfected using Lipofectamine 2000 (Invitrogen), following manufacturer's instruction. After 24 hours, a quarter of the cells were transferred to 100 mm dishes and the medium supplemented with 400 µg/ml of zeocin. After 2 weeks

of selection, single foci were picked and grown in a 24-well plate. The expression of the gene of interest was verified in each cell line by inducing the P_{CMV} promoter with doxycycline at a final concentration of 1 $\mu\text{g/ml}$ for 24 hours.

2.5.4 Proteins extraction

Whole cell extracts were harvested from cultures at 90-95% confluency in 6-well plates or 60 mm dishes. The cells were put on ice for 5 min before being washed two times with ice-cold PBS then resuspended in 750 μl of ice-cold PBS. The cells were pelleted for 5 min at 3,000x g at 4°C. The samples were resuspended in 100 μl of Radio ImmunoPrecipitation Assay (RIPA) buffer (50 mM Tris pH 7.5, 150 mM NaCl, 1 mM Na_2EDTA , 1% Triton X-100, 0.5% sodium deoxycholate and 0.1% SDS) with freshly added protease inhibitor cocktail (Roche Applied Science) then incubated on ice for 30 min with a vortexing step every 10 minutes. The cell lysates were centrifuged for 15 min at 14,000x g at 4°C. The supernatants were transferred in a new tube, quantified by the Bradford method then aliquoted, fast frozen and stored at -80°C. The presence of detergents in the RIPA buffer results in an overestimation of ~10% of the final protein concentration.

2.5.5 Western blotting

For each western blot, 15 or 20 µg of proteins were mixed with 2X SDS loading buffer (100 mM Tris, pH 6.8, 4% SDS, 0.2% bromophenol blue, 20% glycerol, 200 mM β-mercaptoethanol) and denatured by heating at 95°C for 5 min. Proteins were separated on 10% gels by SDS-PAGE gels using the Mini-Protean 3 Cell (BioRad) in a Tris/glycine/SDS buffer (3.03 g/l Tris base, 14.4 g/l glycine, 1% SDS). Proteins were transferred to a polyvinylidene difluoride (PVDF) membrane in 25 mM Tris, 192 mM glycine, and 20% ethanol buffer using a Mini Trans-Blot Cell (BioRad). After transfer, the membranes were stained with ponceau red to verify the transfer efficiency then incubated overnight at 4°C in TBST (50 mM Tris, pH 7.6, 150 mM NaCl, 0.1% Tween 20) plus 5% nonfat dry milk. The membrane was then incubated at room temperature for one hour in TBST plus 5% nonfat dry milk with a dilution of the primary antibody, washed three times for 10 min each with TBST, incubated for one hour in TBST plus 5% nonfat dry milk with a dilution of the secondary antibody, and finally washed three times for 10 min each in TBST alone. Chemiluminescent detection was performed with the ECL system (Promega) and Fuji medical X-ray film (Fujifilm) or LAS-3000 imaging system (Fujifilm). For reprobing, membranes were washed at room temperature twice for 10 min each in stripping buffer (200 mM Glycine, 0.01% SDS, 1% Tween 20, pH 2.2), twice for 10 min each in PBS and finally twice for 5 min each in TBST before incubating the membranes in TBST plus 5% nonfat dry milk for one hour thirty at room temperature.

The following antibodies have been used: anti-Hsma1 antibody (goat polyclonal, 1:500 dilution, ab3823, Abcam), anti-Hsma1 (rabbit polyclonal, 1:500 dilution, Eurogentec), anti-beta Tubulin (rabbit polyclonal IgG, 1:500 dilution, ab6046, Abcam), and anti-FLAG (rabbit, 1:500 dilution, F7425, Sigma). The secondary antibodies were horseradish peroxidase-conjugated anti-goat (rabbit polyclonal, 1:5000 dilution, ab6741, Abcam) and anti-rabbit (goat polyclonal, 1:5000, ab6721, Abcam).

2.5.6 RNA extraction

Total RNAs for RNA-Seq were isolated from cells grown at 90% confluency in 6-well plate with the High Pure RNA isolation kit (Roche Applied Science), following manufacturer's instructions. The samples were quantified with a Nanodrop and their quality verified with a Bioanalyzer (Agilent), only samples with a RIN number > 9 were used.

2.5.7 RNA-Seq

The Deep Seq unit at the University of Nottingham processed the total RNAs. Approximately 10µg of Total RNA was used to enrich for mRNA using two rounds of enrichment with Dynalbeads Oligo(dT)25 (Life tech, 61005). Solid whole transcriptome libraries were made according to the Solid Total RNA-Seq protocol (Life tech, 4445374). Briefly, enriched mRNA was fragmented using chemical hydrolysis followed by phosphorylation with T4 PNK and

column purification using Purelink RNA Micro Kit (Life tech, 12183-016). cDNA synthesis was carried out after adaptor ligation. cDNA libraries were amplified following size selection and adaptor removal using Ampure XP beads. Equimolar pools of RNA-Seq libraries were made following qPCR quantitation using Kapa Library Quantification kit (Kapa Biosystems, KK4823). ePCR and templated bead enrichment was carried out with Solid EZ bead system according to manufacturer's preparation guide. Enriched beads were sequenced on an ABI SOLiD 5500xl analyzer according to the manufacturer's instructions to generate 50bp reads in colour space.

2.5.8 Preparation of the sheared chromatin for ChIP-exo

The T-Rex-U2OS-SMF2 cell line was grown in four 150 mm dishes until they reach ~90% confluency. The cells were fixed with 1% formaldehyde for 10 min at room temperature and quenched with 0.125 M glycine for 5 min at room temperature. The dishes were put on ice and washed two times with ice-cold PBS. Cells from two dishes were pooled together in an eppendorf tube. Cell pellets were resuspended in 0.5 ml of Cell Lysis Buffer (10 mM Tris pH 8.0, 10 mM NaCl, 0.5% NP40) supplemented with fresh Complete Protease Inhibitor (CPI) cocktail (Roche Applied Science) and incubated for 10 min on ice. The nuclei were pelleted for 5 min at 660x g at 4°C and washed once with 1 ml of ice-cold PBS. Nuclei pellets were resuspended in 1 ml of room temperature Nuclei Lysis Buffer (50 mM Tris pH 8.0, 10 mM EDTA, 0.32% SDS) supplemented with fresh CPI cocktail and incubated for 10 min on ice. The nuclei lysates were transferred to a fresh, ice-cold 15 Falcon tube and the

ependorf tube washed with 0.6 ml of cold IP Dilution Buffer (20 mM Tris pH 8.0, 2 mM EDTA, 150 mM NaCl, 1% Triton X-100) complemented with fresh CPI cocktail and combined with the nuclei lysates. For obtaining fragments of ~ 300 bp, solubilized chromatin was sonicated in ice-cold condition with a Bioruptor (Diagenode) using the following conditions: power setting: high, time setting: 15 cycles of 30 sec “on” / 30 sec “off” during two sessions of 15 min. The sonicated extracts were transferred to fresh, ice-cold ependorf tubes and centrifuged for 10 min at 20,800x g at 4°C. The supernatants were transferred to fresh, ice-cold ependorf tubes, fast frozen and stored at -80°C. For determining the size of the chromatin fragments, 45 µl of IP dilution Buffer and 1 µl of RNase A (10 mg/ml) were added to 5 µl of sonicated extracts and incubated for 30 min at 37°C. The samples were supplemented with 1 µl of Proteinase K and incubated for 2 hours at 62°C before being loaded on a 2% agarose gel and stained with ethidium bromide.

2.5.9 ChIP-exo

Peconic LLC further processed the sheared genomic DNAs as described in (Rhee and Pugh, 2011). Briefly, sonicated chromatin was immunoprecipitated overnight at 4°C with an anti-FLAG M2 antibody (mouse monoclonal, F3165, Sigma) and followed by protein A sepharose precipitation at 4°C for 1.5 hour. Immunoprecipitates were washed several times then subjected to enzymatic reactions to ligate P2 adaptors to both ends of sonicated DNA fragments and to remove uncrosslinked fragments by λ and RecJ_f exonucleases, which digest respectively one strand of the double-stranded DNA and 5'-3' single-

strand DNA. DNA fragments not protected by a bound protein will be digested by the exonucleases diminishing the background. The λ exonuclease treatment produces single-stranded DNA outside the protein bound DNA fragment. The protected fragment remains double stranded and will provide the final overlapping region representing the protein-binding site in the genome. The resin with the ChIP samples were washed several times before the elution of the ChIP-exo treated DNA. The crosslinks were reversed and the proteins degraded with a Proteinase K treatment. DNA were extracted with phenol:chlorophorm:isoamyl alcohol and precipitated with ethanol before being denaturated at 95°C to make single-stranded DNA. Each single-stranded DNA was made double-stranded by a TruSeq Adaptors Index 1 extension with the Φ 29 polymerase. TruSeq Universal Adaptors were then ligated to the exonuclease-treated ends. The DNA fragments were amplified by PCR and only the 120-160 bp PCR products were gel-purified and subject to single-end deep sequencing with Illumina genome sequencer.

2.5.10 Non-homologous end-joining assay and FACS analyses

Prior to transfection, the pEGFP-Pem1-Ad2 plasmid was digested overnight with *HindIII* or *I-SceI*. The digested plasmids were heat-inactivated and column-purified before being co-transfected with a pRFP plasmid for controlling the transfection efficiency. A day before transfection, 8×10^5 cells were seeded in 60 mm dishes for obtaining a ~70 % confluency on the transfection day. Transfections were performed with 3 μ g of linear pEGFP-Pem1-Ad2, 3 μ g of pRFP and 14 μ l of Lipofectamine 2000 (Invitrogen),

according to manufacturer's instructions. After 24 hours, green (GFP) and red (RFP) fluorescence was measured by fluorescence-activated flow cytometry (FACS). For FACS analysis cells were harvested with Accutase (Sigma), washed once in 1X PBS and fixed in 2% formaldehyde (Sigma). FACS analysis was performed on a Coulter FC500 (Beckman Coulter). The numbers of repaired events are reported as the ratio of green and red positive cells over the total number of red positive cells. This ratio normalizes the numbers of repaired events to the transfection efficiency. The values for all the cell lines are reported as a percent of the control cell lines.

2.5.11 Illegitimate DNA integration assay

For integration assays in the T-Rex-U2OS cell lines, 8×10^5 cells were seeded onto 6-well plates with 2.5 μg of pRC1712, a neomycin-resistant derivative of pBS SKII+ (Agilent), and 5 μl of Lipofectamine 2000 (Invitrogen). Twenty-four hours later, cells were trypsinized and 5×10^4 cells of each transfection were seeded onto 10 cm dishes in medium containing 800 $\mu\text{g/ml}$ of G418 (Sigma). After two weeks of selection, surviving foci were fixed for 15 min with 10% formaldehyde in PBS, stained for 30 min with methylene blue buffer (1% methylene blue, 70% ethanol), washed with water, air dried, and photographed.

2.6 Bioinformatics analysis

2.6.1 Hsmar1 remnants and ITRs in the human genome

Hsmar1 remnants locations were extracted from the output of the RepeatMasker (RM) Genomic Datasets produced for the human genome, hg19 assembly, RM 3.3.0 track, repbase libraries 20120124. The dataset has been generated by screening the human genome against the Repbase curated library of repeats with the RM software, which compares two DNA sequences to determine similar regions. The fasta file generated by the RM software provides the locations of the repeats which match their Repbase consensus sequence. Hsmar1 transposon ends were obtained using the BLAST software against the human genome, hg19 assembly, and the outputs parsed for transposon ends retaining at least 80% of the length and identity of the ancestral Hsmar1 transposon end. Human genome annotation for the hg19 assembly was retrieved with BioMart from the Ensembl website. Bioinformatics analyses were done with custom Python scripts using the Biopython library.

2.6.2 RNA-Seq analysis

Read Mapping

The RNA-seq colour-space reads were initially processed using The LifeTechnologies LifeScope (v2.5.1) Whole Transcriptome Pipeline. Reads

were first filtered against a set of sequences to remove those reads derived from undesired sequences, such as rRNA, tRNA, and sequencing contaminants. The reads that pass the filter are first mapped against the reference genome sequence (Homo sapiens GRCh37 hg19 from Ensembl). Reads are also mapped to a library of exon junction sequences derived from the gene exon coordinates defined in the genome sequences companion genome annotation file (GTF file), thereby allowing reads that map across exon boundaries to be recorded.

BAM files and Mapping Types

The read alignments are recorded in a BAM formatted alignment file (named *.bam), and companion BAM index file (named *.bam.bai). The BAM (or SAM) format is the industry standard format for recording and presenting NGS sequencing alignment results. There are 3 types of recording the mapping of reads against a reference sequence:

'Uniquely' mapped reads are reads that map to only a single location, and there are no significant hits to any other reference genome location.

'Primary' mapped reads record the best mapping position of each read, regardless of whether they map uniquely.

'All' mapped reads are when every single significant mapping position is recorded against the reference genome, therefore a read may have multiple entries within the BAM file.

Primary aligned reads were recorded in the result BAM files, with those that aligned uniquely marked as so. A minimum of 30 millions primary reads were

mapped in each sample replicate and at least 78% of these reads were uniquely mapped.

Mapped Read Counts per Gene

For each sample replicate, read counts per gene were calculated using the program htseq-count using the corresponding BAM file and genome annotation information. Uniquely aligned reads with a minimum mapping quality score of at least 10 (MAPQ10) were used in calculating the counts data.

RPKM Expression Values

The RPKM is a normalized read count for a given gene. The read count of a gene is normalized against the total number of (primary) mapped reads in that particular sample (i.e. the BAM file) and against the total length of the gene's exon-space (Mortazavi et al., 2008).

Differential Gene Expression

Differential genes expression analysis was performed using Baggerly's test in CLC Genomics Workbench software package. The false discovery rate (FDR) corrected p-values of < 0.001 and p-value < 0.0001 were examined (CLCbio, Qiagen).

Reads are also aligned separately to the human SETMAR gene (ENSG00000170364) and to the SETMAR mouse codon optimized sequence

using primary aligned reads in order to assess the different isoforms of the SETMAR gene.

2.6.3 ChIP-exo analysis

After sequencing 10 millions reads, a total of 5 millions of reads for each biological replicate were mapped to human genome version GRCh37 (hg19) using BWA version 0.7.5a, and BAM-formatted files were created using samtools version 0.1.18 (Li and Durbin, 2009) (Li et al., 2009b). Peak calling has been done with GeneTrack using a fine or a course grain (Albert et al., 2008). For each genomic coordinate, GeneTrack approximates a measurement by replacing each read with a probability function. When summed over all reads, it produces a smooth and continuous curve where the peaks indicate the most likely positions of the measured genomic feature. Peak detection is done in an iterative manner. For each iteration, the algorithm selects the highest peak along the chromosome and establishes an exclusion zone of 40 bp on each side of the peak summit coordinate. Any other peak in an exclusion zone is therefore removed from the analysis. The peak calling process is repeated until no new peak can be placed. The singleton peaks, which contain only one type of read, were discarded.

	TATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCACCTC GAGTCCCTATCAGTGATAGAGATTGACCTCCCTATCAGTGAT AGAGATACTGAGCACATCAGCAGGACGCACTGACCACTTTAA GAAGGAGATATACATATG	0.159
p4 (K)	CAATTCCGACGTCTAAGAAACCATTATTATCATGACATTAACC TATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCACCTC GAGTCCCTATCAGTGATAGGGATTGACATCCCTATCAGTGAT AGAGACACTGGGCACATCAGCAGGACGCACTGACCACTTTA AGAAGGAGATATACATATG	50,376 / 0.299
p5 (E)	CAATTCCGACGCCTAAGAAACCATTATTATCATGACATTAGCC TATAAAAATAGGCGTACCACGAGGCCCTTTCGTCTTCACCTC GAGTCCCTATCAGTGATAGAGATTGACACCCCTATCAGTGAT AGAGATACTGAGCACATCAGCAGGACGCACTGACCACTTTAA GAAGGAGATATACATATG	183,694 / 0.743
p6 (W / pltetO)	CAATTCCGACGTCTAAGAAACCATTATTATCATGACATTAACC TATAAAAATAGGCGTATCACGAGGCCCTTTCGTCTTCACCTC GAGTCCCTATCAGTGATAGAGATTGACATCCCTATCAGTGAT AGAGATACTGAGCACATCAGCAGGACGCACTGACCACTTTAA GAAGGAGATATACATATG	145,807 / 1

Table 2-5: Constitutive promoters.

Nomenclature (the letters indicated between brackets are from (Alper et al., 2005)), sequence, and strength of the constitutive promoters used in this study.

Transposase expression plasmids were built by cloning the respective transposase genes in pBACe3.6, pGHM491, and pMAL-c2X (New England Biolabs) between NdeI and BamHI restriction endonuclease sites. Plasmids pRC1717 and pRC1719 encode the wild-type transposase in pMAL-c2X in presence and absence of the MBP tag respectively. Plasmids pRC1744 to 1749 and pRC1729 to 1734 encode the wild-type transposase downstream the p1 to p6 promoters in pBACe3.6 and pGHM491. Plasmids pMT1750 to 1754 encode K50Q, D98N, R104A, F132A, and R141L mutants' downstream p1 in pBACe3.6. Plasmids pRC1757 to 1761 encode K50Q, D98N, R104A,

F132A, and R141L mutants' downstream p1 in pGHM491 whereas plasmids pRC1764 to 1768 encode the same mutants downstream p6 in pGHM491. Plasmids pMT1735 and 1736 encode respectively wild-type Sleeping Beauty and PiggyBac transposases downstream p1 in pGHM491. pRC1737 encode the PiggyBac transposase downstream p2 in pGHM491.

2.7.2 Western blotting

Western blotting were done in similar a condition as the one described in section 2.3.8 with the following differences. Cells containing a derivative of pMAL-c2x were grown in LB supplemented with 100 µg/ml of ampicillin at 37°C until an OD₆₀₀ of ~ 0.5 and were then induced with the required concentration of IPTG for 2 hours at 37°C. Cells containing pGHM491 or pBACe3.6 derivatives were grown in LB supplemented with respectively 100 µg/ml of spectinomycin or 50 µg/ml of chloramphenicol at 37°C for the same amount of time as the induced cells. Promoters expression were analyzed by pelleting $\sim 1.5 \times 10^9$ cells. The samples were resuspended in SDS sample buffer, boiled for 5 min, and loaded on 10% SDS-PAGE gels. Proteins were transferred to PVDF membrane, probed with an anti-Hsma1 antibody (goat polyclonal, 1:500 dilution, ab3823, Abcam) followed by a horseradish peroxidase-conjugated anti-goat secondary antibody (rabbit polyclonal, 1:5000 dilution, ab6741, Abcam). Proteins were visualized by using the ECL system (Promega) and Fuji medical X-ray film (Fujifilm).

2.7.3 Papillation assays

Transposase expression plasmids were transformed into the reporter strain *E. coli* RC5096 or RC5118. Both strains encode a transposon flanked with respectively Hsmar1 ends or Sleeping Beauty, PiggyBac, and Hsmar1 ends. RC5096 transformants were plated on LB-agar medium supplemented with 0.01% lactose, 40 µg/ml of X-gal and either 200 µg/ml of ampicillin, 50 µg/ml of chloramphenicol, or 100 µg/ml of spectinomycin. RC5118 transformants were plated on similar conditions as RC5096 except that 0.1% lactose was added. Plates were incubated 5 days at 37°C and photographed. The transposition rate is determined by the number of papillae per colony or by the rate of appearance of papillae on the colonies.

2.7.4 Mating-out assay

A chloramphenicol resistant derivative of the conjugative plasmid pOX38 has been introduced in the RC5096 or RC5118 papillation strains to create the donor strains RC5097 and RC5122, respectively. Briefly, RC5097, or RC5122, transformants and the recipient strain, RC5094, were grown overnight in LB supplemented with antibiotics at 37°C. The next day, respectively one and three volumes of RC5097, or RC5122, and RC5094 were centrifuged for 5 min at 6,000x g. Each pellet was resuspended in 3 ml of fresh LB, pool together, and incubated in a shaking water bath for 3 hours at 37°C. After the mating, the transposition events were detected by plating

200 µl of each culture on LB-agar medium supplemented with tetracycline and kanamycin. The number of transconjugants was obtained by plating a 10^{-5} -fold dilution of each culture on LB-agar medium supplemented with tetracycline and chloramphenicol. The plates were incubated overnight at 37°C and the transposition rate determined the next day by dividing the number of kanamycin-resistant cells by the number of chloramphenicol-resistant cells.

3. BIOINFORMATICS ANALYSIS OF THE HSMAR1 TRANSPOSON IN THE HUMAN GENOME

3.1 Introduction

This chapter focuses on a bioinformatics analysis of the Hsmar1 transposon in the human genome. Prior to the analysis of the distribution of Hsmar1 transposon ends, which is necessary for the study of SETMAR function in gene regulation, I will focus on the distribution of the different groups of remnants. Hsmar1 remnants, and by extension the ITRs, are mostly located in genes. This is presumably explained by the targeting of transposon insertions to open chromatin. The exaptation of several Made1 elements in miRNAs was previously established but the current bioinformatics analysis identifies 456 additional examples that could also act as miRNAs. These miRNAs-like have retained their ITRs, which makes them possible targets of SETMAR binding and regulation. Furthermore, preliminary results have established a possible selection of a fraction of ITRs in non-coding genes and protein-coding genes, supporting a putative epigenetic regulation of these genes by SETMAR.

3.2 Results

3.2.1 Distribution of the Hsmar1 remnants in the human genome

Hsmar1 remnants can be classified in four groups (Figure 3.1A). The first group is the full-length Hsmar1 transposons, which retains at least 80% of the ancestral transposon length. The second group contains the truncated Hsmar1 transposons, which retains less than 80% of the ancestral transposon

length. The third group is the Made1 elements, which are composed of 37 bp inverted repeats with 6 bp of unique DNA in the middle. The fourth group contains the solo ITRs, which are deletion-derivatives of Made1.

According to RepeatMasker (V4.0.5), the human genome (version hg19) contains 8,527 Hsmar1 remnants (Figure 3.1B). The full-length and truncated Hsmar1 transposons represent less than 9% of the number of remnants while the solo ITRs and Made1 represent each ~46% of the Hsmar1 remnants. To provide a better representation of the link between solo ITRs and Made1, I plotted the size distribution of Hsmar1 non-autonomous elements and their deletion-derivatives (Figure 3.1C). The distribution displays a broad and a narrow peak, which corresponds to the solo ITRs and Made1 elements, respectively, and a series of smaller peaks joining each type of remnants. The solo ITRs peak at 46 nucleotides but every length between 39 and 55 nucleotides is represented by more than a hundred remnants. The remnants between the solo ITRs and Made1 peaks correspond to partial deletion of Made1 where one of the ITRs is incompletely deleted. Made1 elements peak at a length of 80 nucleotides with more than a hundred remnants found for each length between 71 and 81 nucleotides. Henceforth, I will define Made1 as any element with at least 80% identity to the full length 80 bp element plus the small fraction of remnants longer than 80 bp, which correspond to elements containing a DNA insertion.

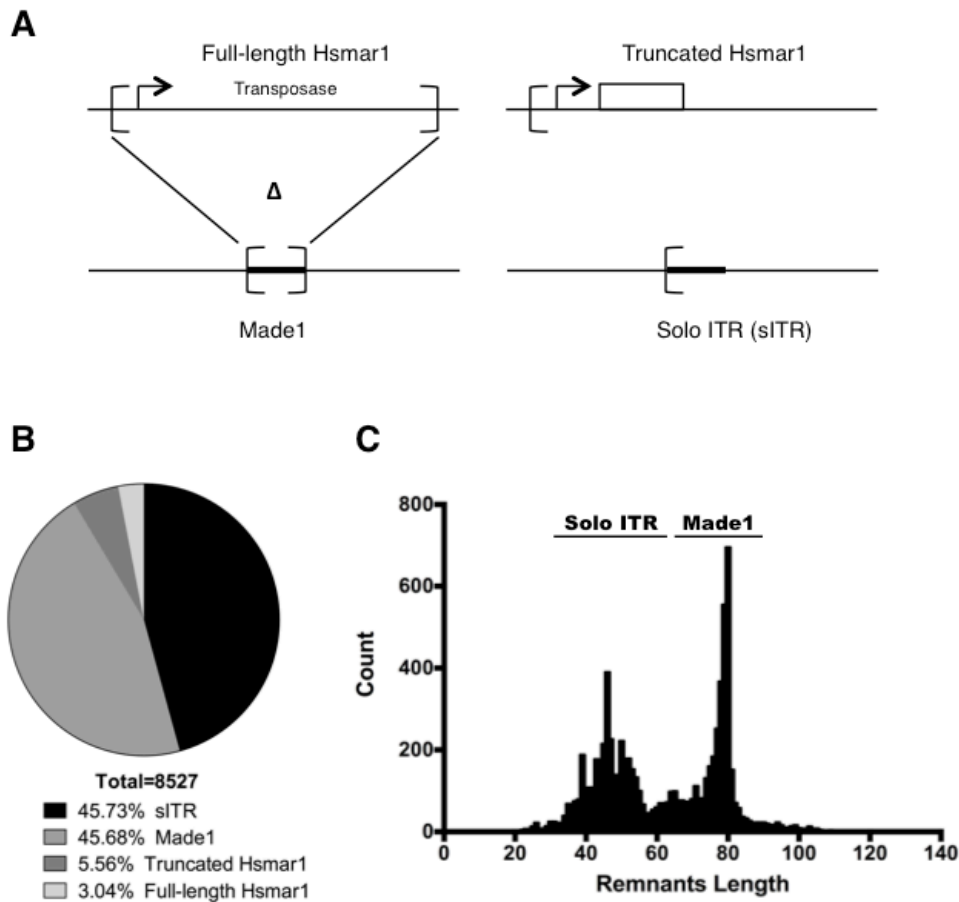


Figure 3-1: Hsmar1 remnants in the human genome.

(A) Hsmar1 remnants can be divided in four types: full length and truncated Hsmar1, Made1 elements composed of two ITRs and a short central region, and solo ITR which are deletion-derivative of Made1 and composed of a single ITR. Open brackets, ITR. (B) Distribution of the four groups of Hsmar1 remnants in the human genome. (C) Size distribution of the non-autonomous elements. Most remnants are found in the two major peaks corresponding to solo ITR and Made1.

I next determined the distribution of Hsmar1 remnants in the human genome at the chromosomal and the whole-genome level. For the chromosomal analysis, I merged the truncated and full-length Hsmar1 transposons to obtain a sufficient number of remnants to investigate. The initial analysis was to determine whether Hsmar1 remnants were enriched in some chromosomes. For this I compared the number of remnants per

chromosome and the size of the chromosome (Figure 3.2A). There is a positive correlation ($r^2 \geq 0.8$) between the number of remnants and the chromosome size. This is the distribution expected if transposon integrations were randomly distributed.

To determine whether there is a difference in the genomic distribution between the four groups of remnants, I compared the location of each remnant to the position of non-coding and protein-coding genes in the human genome (Figure 3.2B). Respectively, 54 and 63% of the truncated and full-length Hsmar1 transposons are located outside genes. Approximately 13% of the remnants of both types are found in non-coding genes whereas the remaining truncated (33%) and full-length (24%) transposons are located in protein-coding genes. The genomic distributions of solo ITRs and Made1 are roughly alike with 49% and 42% of intergenic elements, respectively. Approximately a third of the solo ITRs and Made1 are located in protein-coding genes whereas the remaining 15% and 20% are found in non-coding genes.

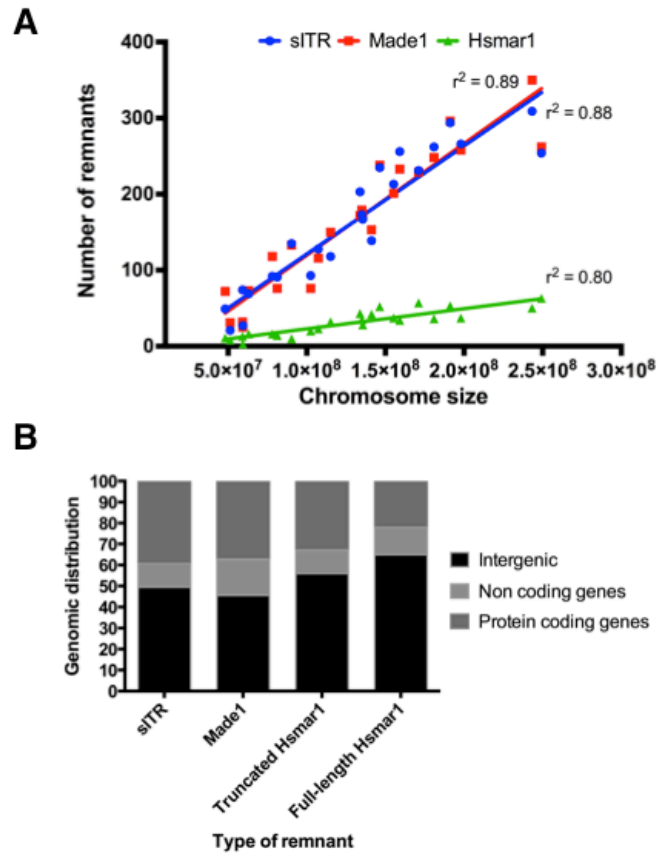


Figure 3-2: Distribution of each group of Hsmar1 remnants in the human genome.

(A) The distribution of Hsmar1 remnants in the human genome is random. A strong correlation is observed between the size of the chromosome and the number of remnants. The truncated and full-length Hsmar1 were merged for getting a sufficient number of remnants to analyze. **(B)** Genomic distribution of the solo ITRs, Made1, truncated and full-length Hsmar1. Except for the full-length Hsmar1, roughly half of the remnants are located inside a gene. The intragenic remnants are mostly found in protein-coding genes.

3.2.2 Distribution of the Hsmar1 transposon ends in the human genome.

3.2.2.1 Origins of the transposon ends

The Hsmar1 transposon has been inactive for the last 40 My (Pace and Feschotte, 2007). As described in the previous section, 8,527 remnants are currently annotated in the human genome. However, the vast majority of the remnants should have evolved under neutral selection so the accumulation of new mutations should have also affected the transposon binding sites located in the Hsmar1 remnants. To determine which putative binding sites could still be recognized by SETMAR, and therefore potentially subject to purifying selection, I used BLAST to retrieve the list of every ITR present in the human genome. I then parsed the list to keep only the ITRs retaining at least 80% of the length and identity of the ancestral ITR. This list was then compared to the list of Hsmar1 remnants from RepeatMasker to confirm the transposon origin of the ITRs.

Out of the 8,527 Hsmar1 remnants in the human genome, 5,467 contain 6,334 ITRs. To characterize the source and the distribution of the ITRs, I first compared the location of the Hsmar1 remnants and the ITRs to obtain the list of Hsmar1 remnants with an ITR. This list was then compared to the position of human non-coding and protein-coding genes (Figure 3.3A). In all four distributions, more than 90% of the ITRs are present as solo ITRs and Made1 elements, while the remaining binding sites are distributed between the truncated and full-length Hsmar1. At the genomic level (Figure 3.3A, first

column), 44 and 49% of the ITRs are in solo ITRs and Made1, respectively, with the remaining 7% distributed between truncated (4%) and full-length (3%) transposons. The distribution of intergenic ITRs is similar to its genomic counterpart, 47 and 44% of the ITRs are located in solo ITRs and Made1, respectively, and the remaining 8% is evenly distributed between the truncated and full-length Hsmar1 (Figure 3.3A, second column). The intragenic distribution is slightly different: 38% and 44% of the solo ITRs are in non-coding and protein-coding genes, respectively. For Made1, the proportion increases to 57 and 50% in non-coding and protein coding genes, respectively. For truncated and full-length Hsmar1, they respectively give 3 and 2% of the ITRs for both types of genes (Figure 3.3A, third and fourth columns).

Since this seems to indicate a bias towards Made1 elements as the source of intragenic ITRs, I determined the proportions of conserved ITRs at the genomic level by comparing the number of remnants with an ITR to the total number of remnants (Figure 3.3B). The proportion of solo ITRs and full-length Hsmar1 elements with a binding site is similar at 62%, whereas this percentage increases to 68% in the case of Made1. Only 47% of the truncated transposons still contain one ITR.

I next determined whether the difference between the numbers of conserved ITRs observed in the remnants was also found for the intragenic and intergenic remnants. I compared the number of remnants with an ITR to the total number of remnants for the following conditions: outside a gene, in a non-coding gene, or in a protein-coding gene (Figure 3.3C). For intergenic remnants, ~ 60% of the solo ITRs, Made1 elements, and full-length Hsmar1

elements contain at least one ITR but this fraction falls to 50% for truncated Hsmar1. These proportions are similar to the genomic ones except for the slight decrease of Made1. For intragenic remnants, the results are more variable. In non-coding genes, ~80% of the solo ITRs and Made1 retain an ITR whereas for both types of Hsmar1 transposons the proportion remain at ~60%. In protein-coding genes, ~70% of the Made1 and full-length Hsmar1 transposons still have an ITR whereas the solo ITRs and truncated Hsmar1 transposons have a similar proportion as their intergenic counterparts, respectively, 60 and 50%.

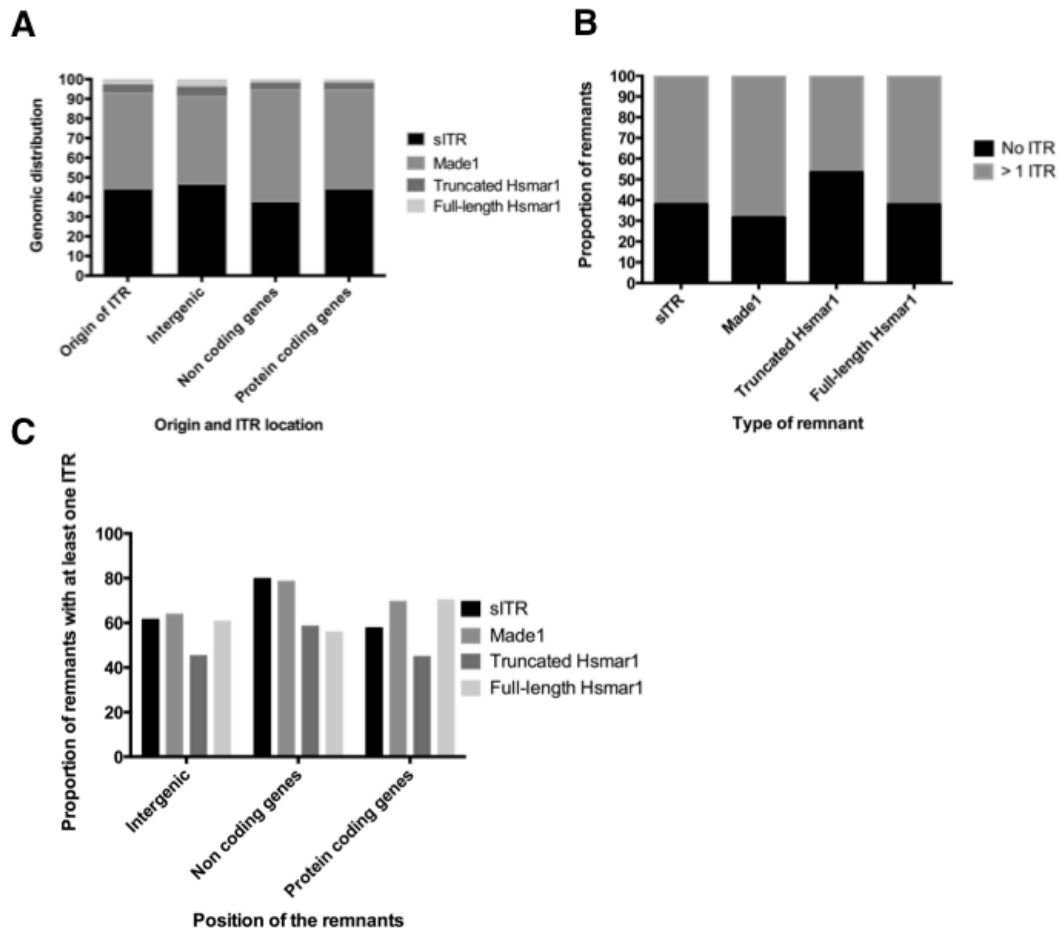


Figure 3-3: Source of ITRs in the human genome.

(A) Source of ITRs at the genomic level and origin of intergenic and intragenic ITRs. The solo ITRs and Made1 provides more than 90% of the intergenic and intragenic ITRs. **(B)** Proportion in each type of remnants of elements containing at least one ITR. Except for the truncated Hsmar1, ~60% of the remnants still retain one Hsmar1 binding site. **(C)** Proportion of intergenic and intragenic remnants containing one ITR. Approximately 80% of the solo ITRs and Made1 in non-coding genes and 70% of Made1 and full-length Hsmar1 in protein-coding genes have retained one ITR, a higher fraction than the ones observed at the genomic and intergenic levels.

In summary, the results of this section suggest that I should focus on solo ITRs and Made1 because they represent more than 90% of the remnants and are the source of more than 90% of the ITRs. The intergenic ITRs originate evenly from solo ITRs and Made1 (respectively 51 and 49%), a

proportion equal to the distribution of solo ITRs and Made1 in intergenic remnants (respectively 52 and 48%). This is explained by a similar proportion of conserved ITRs for the intergenic solo-ITRs and Made1 (respectively 61 and 64%). In the case of non-coding genes, 40% of the ITRs are from solo ITRs and the last 60% come from Made1 elements. A similar proportion of solo ITRs and Made1 in non-coding genes (respectively, 40% and 60%) and a comparable number of conserved ITRs for both types of remnants (respectively 80% and 79%) explain the observation that Made1 is the main source of ITR in non-coding genes. For protein-coding genes, 53% of the ITRs come from Made1 and the last 47% from solo ITRs. However, the remnants distribution in protein-coding genes establishes that 51% of the remnants are solo ITRs and the last 49% Made1 elements. This discrepancy between the remnants distribution and the source of ITRs is explained by an ITR conservation rate of 70% in Made1 compared to 58% for solo ITRs.

3.2.2.2 Distribution of the ITRs in the human genome

In this section I will focus on the genomic distribution of the ITRs. I first compared the location of the ITRs to the position of non-coding and protein-coding genes (Figure 3.4A). Out of the 6,334 ITRs found in the human genome, 4,132 ITRs are located in 2,697 genes. Of these 2,697 genes, 56% encode proteins whereas the remaining 44% encode non-coding RNAs (Figure 3.4B). The fraction of protein-coding genes with an ITR is noticeably higher than the fraction of protein-coding genes in the human genome, which represents 35% of the genes (Figure 3.4B).

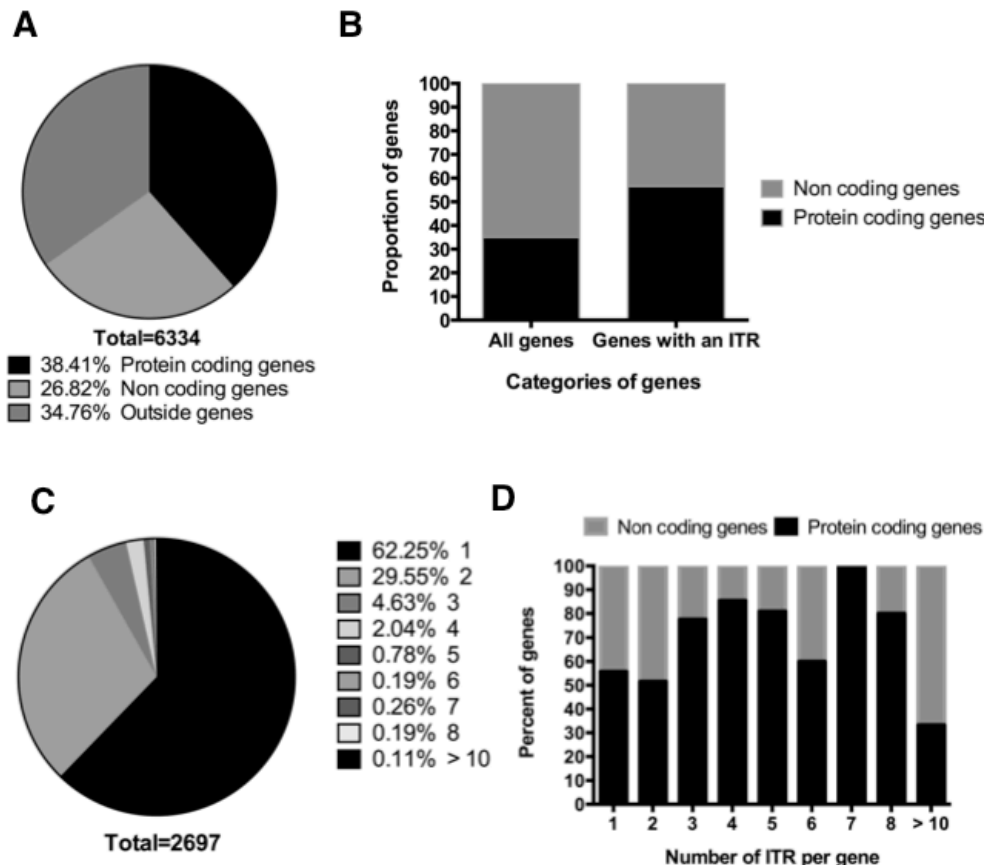


Figure 3-4: Distribution of the Hsmar1 ITRs in the human genome.

(A) Distribution of the 6,334 ITRs in the human genome. Similar to the Hsmar1 remnants, 65% of the ITRs are found in genes. (B) The 4,132 intragenic ITRs are located in 2,697 genes. Nearly 60% of the genes with an ITR encode for proteins whereas the human genome contain 35% of protein-coding genes. (C) Distribution of the 2,697 genes by the number of ITRs they contain. Approximately 92% of the genes contain one or two ITRs while the remaining 8% is inversely distributed to the number of ITRs per gene. (D) The genes with one or two ITRs are evenly distributed between non-coding and protein-coding genes. Most of the genes with more than 2 ITRs encode for proteins except for the genes with more than 10 ITRs.

The fact that the 4,132 ITRs are only found in 2,697 genes means that a fraction of these genes contains more than one ITR. I first investigated the proportion of genes with multiple ITRs amongst the 2,697 genes (Figure 3.4C). The number of ITRs per gene varies between 1 and 15 (the three

genes with more than 10 ITRs are regrouped in the >10 category). The vast majority of genes, 62% and 30%, contain, respectively, one or two ITRs. Amongst the 797 genes with 2 ITRs, 81% (or 648 genes) contain a Made1 element, which has two ITRs. The remaining genes possess two independent remnants. The Made1 with two ITRs are more frequent in non-coding genes (87% of the non-coding genes with 2 ITRs) than protein-coding genes (76%).

I next determined if there was any difference in the distribution of genes with multiple ITRs between non-coding and protein-coding genes (Figure 3.4D). It is worth noting that categories with 5 or more ITRs have so few members that the distinction between non-coding and protein-coding genes becomes uninformative. The genes with one or two ITRs, totalling 92% of all the genes with an ITR, are evenly distributed between non-coding and protein-coding genes. Except for the genes with 6 or more than 10 ITRs, the remaining categories are mostly filled with protein-coding genes while the non-coding genes represents less than a quarter of each category.

From the previous sections, we know that the Hsmar1 remnants, and by extension the ITRs, are particularly present in genes, a feature likely explained by the targeting of transposon insertions to open chromatin (Li et al., 2013a) (Jursch et al., 2013). This predisposition towards open chromatin should result in the integration of Hsmar1 transposons and Made1 elements in the largest human genes. To verify this assumption, I compared the size of the genes with an ITR against the remaining human genes (Figure 3.5A). In agreement with previously published genomic data, the protein-coding genes are on the average larger than the non-coding genes (Lander et al., 2001). Thus, the genes with ITRs have a median length 27-fold larger than the

average human gene. Most of this increase is due to the over-representation of non-coding genes containing an ITR, which are 16-fold longer than average. If I consider only the ITR-containing coding genes, they are 6-fold longer than average.

I next determined whether the number of ITRs is related to the size of the gene (Figure 3.5B). For the analysis, I merged the non-coding and protein-coding genes to obtain a more meaningful result. The genes with 1 or 2 ITRs have a similar median size because more than 80% of the genes with 2 ITRs contain a complete *MaeI* element, which is equal to a single integration event. Except for the genes with 6 ITRs, I observe a positive correlation between the gene median size and the number of ITRs. The low number of genes for the categories with 6 or more ITRs per genes makes the size distribution sensitive to outliers. However, the median size of genes with 5 ITRs is 10 times larger than the genes with one or two ITRs and a similar trend is observed for the genes with 3 and 4 ITRs with a 4- and 5-fold increase, respectively.

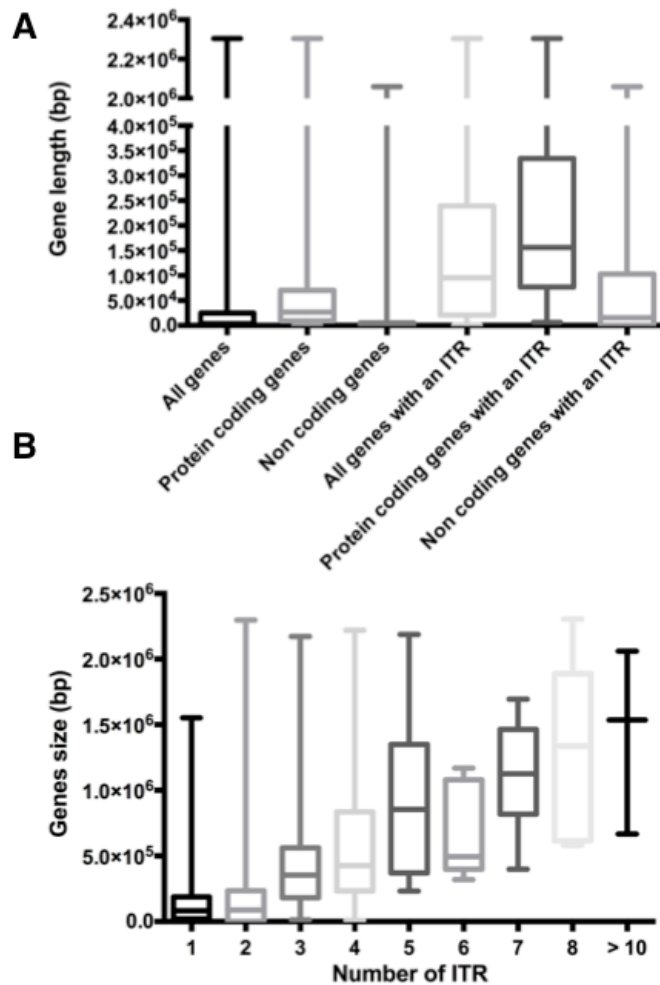


Figure 3-5: The ITRs are particularly found in the largest human genes.

(A) Box and whisker plot representing the size distribution of human genes with and without an ITR. Human protein-coding genes are on average larger than non-coding genes. The ITRs are distributed in the largest genes for both non-coding and protein-coding genes. **(B)** Box and whisker plot representing the size distribution of human genes versus the number of ITRs. Except for the genes with 6 ITRs, a positive relation is observed between the number of ITRs per gene and the median size of the gene.

To determine whether the ITRs are particularly enriched at a specific location in the genes, I first analyzed the distribution of ITRs relative to the start of the gene (Figure 3.6A). The choice of the start of the gene is motivated by the idea that if SETMAR binds to the intragenic ITRs and

regulates the gene expression by modifying the neighbouring nucleosomes, the location of the transposon ends in the gene could be important. The distance between the ITR and the start of the gene has been normalized by the size of the gene and categorized by decile with 0 and 1 being respectively the start and the end of the gene. In protein-coding genes, the location of the 2,432 ITRs is roughly evenly distributed within the gene with each decile comprising 9 to 12% of the ITRs. The presence of an ITR at the start of several non-coding genes is due to the miRNAs derived from *Made1*. Contrary to protein-coding genes, the distribution of ITRs in non-coding genes is not uniform with a proportion of ITRs in each decile varying from 5 to 21%. The first and the seventh deciles account for 40% (21 and 19%, respectively) of the ITRs followed by the second and the sixth deciles representing 22% (11% each).

A series of papers over the last few years have demonstrated a role for H3K36me3 in alternative splicing (Kolasinska-Zwierz et al., 2009) (Kim et al., 2011). SETMAR dimethylates the histone H3 K36 and also interacts with PRPF19 and Spf27, two proteins involved in the spliceosome (Lee et al., 2005) (Beck et al., 2008). SETMAR could target the splicing complex to the transposon ends near splicing sites where it may regulate alternative splicing.

To select only the ITRs present in genes with at least one intron, I first characterized the distribution of the intragenic binding sites in exons and introns (Figure 3.6B). From the 4,132 ITRs analyzed, ~19% are located in exons while the remaining 81% are found in introns. To determine whether the exonic ITRs are located in single or multiple exons genes, I compared their location with the type and the length of the gene (Figure 3.6C).

Approximately 94% of the 792 exonic binding sites are found in non-coding genes. These genes are essentially divided between miRNAs (9%) and miRNAs-like (83%). These miRNAs-like are annotated as non-coding genes with a size comprised between 61 and 129 nucleotides. These genes are essentially composed of a Made1 element and are annotated by the GENCODE project as miRNAs. The last 8% is divided between non-coding genes with at least one intron (2%) and protein-coding genes (6%). The non-coding genes with at least one intron encode long non-coding RNAs (lncRNAs). For the protein-coding genes, the ITRs are found in exonic 5' and 3' untranslated regions and so are not encoded in the protein sequence.

To determine whether the ITRs are enriched at specific positions compared to a splicing site, I calculated the relative distance between the ITRs and the splicing site for intronic and exonic ITRs present in genes with at least one intron (Figure 3.6D). I first evaluated the absolute distance between the ITR and the nearest splicing site before normalizing the distance by the size of the intron or exon where the ITR is located. This normalization step is necessary for avoiding skewing the distribution owing to ITRs present in small introns. The distances are categorized by decile with 0 and 0.5 being respectively the start and the middle of the intron or exon. Since the analysis is based on the smallest distance between the ITR and the splicing site, there cannot be any distance superior to half the size of the intron or exon. The distances are approximately evenly distributed with each decile containing between 8 and 11% of the ITRs for protein-coding genes and between 7 and 12% for non-coding genes.

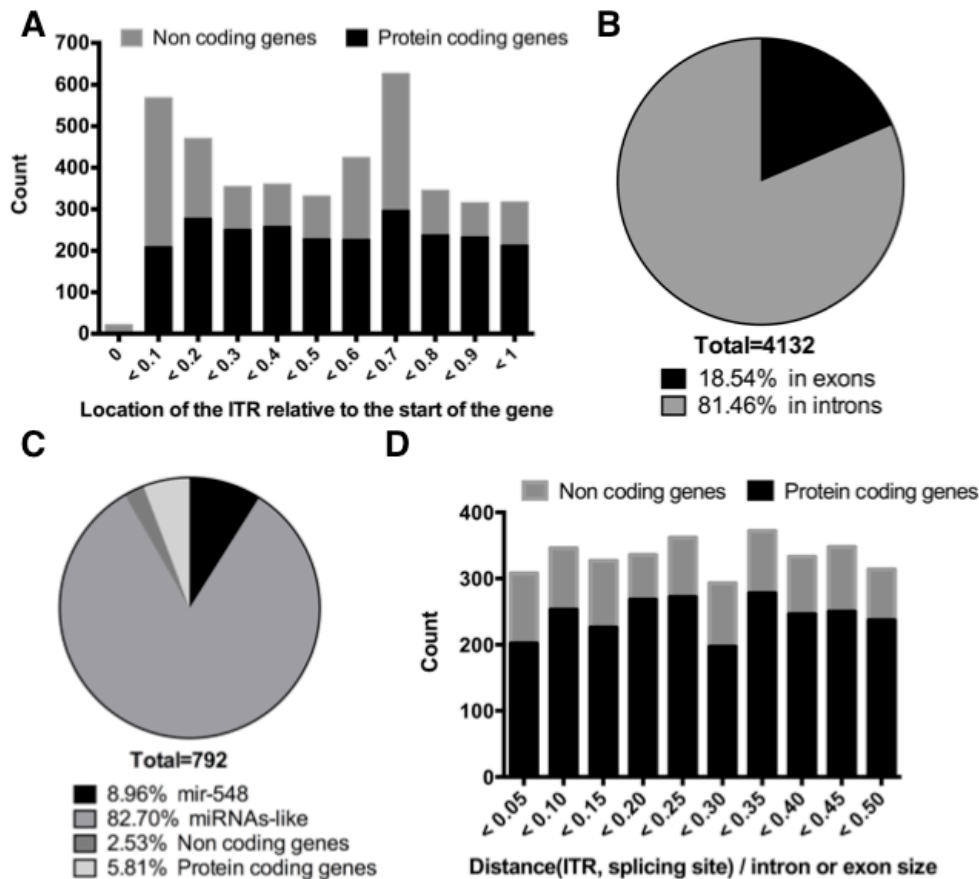


Figure 3-6: Distribution of the intragenic ITRs compared to the start of the genes and the nearest splicing site.

(A) Distribution of the ITRs compared to the start of their gene. In protein-coding genes, the ITRs are evenly distributed throughout the entire gene length. In non-coding genes, the ITRs are enriched at the start and at the three-quarter of the gene.

(B) Distribution of the intragenic ITRs in exons and introns. (C) Gene distribution of the exonic ITRs. These ITRs are mostly found in non-coding genes (94%) and more particularly in miRNAs-like, which are essentially composed of a Made1 element. The exonic ITRs present in protein-coding genes are located in the 5' and 3' UTRs.

(D) Distribution of the ITRs compared to the nearest splicing site. For both non-coding and protein-coding genes, the ITRs are evenly distributed throughout the size of the intron or exon.

3.2.2.3 Gene Ontology enrichment of genes with an ITR

SETMAR could act as an epigenetic regulator of gene expression over a network of 2,697 genes from which 1,523 are protein-coding genes. All these

genes contain a binding site in their introns. To get a better understanding of the putative cellular functions regulated by SETMAR, I looked for possible enrichments of genes in specific functions (Table 3.1). I first determined the overlap in enriched Gene Ontology terms (GO terms) between Hsmar1 remnants and the ITRs. The number of genes used for each category is lower than the total number of genes with Hsmar1 remnants (3,432 genes) or ITRs (2,697 genes). This is explained by the unknown functions of most non-coding genes. The genes with Hsmar1 remnants are significantly enriched ($p < 1.0E^{-10}$) in three GO terms: “cell adhesion”, “nervous system development”, and “signal transduction”. If I keep only the genes with an ITR, I observe a significant enrichment ($p < 1.0E^{-8}$) for five GO terms with two of them, “nervous system development” and “cell adhesion”, also found for the genes with Hsmar1 remnants. Of the three GO terms unique to the genes with ITR, two of them are also involved in the nervous system, “generation of neurons” and “transmission of nerve impulse”, while the last unique GO term is “cell projection organization”.

In the previous sections, I determined that the ITRs were particularly found in the largest genes so the enriched GO terms observed for the genes with an ITR could be explained by this localization. Indeed, the biggest genes, longer than 1 or 5 Mb, are enriched for the same GO terms as the genes with an ITR except for one, “generation of neurons” (Table 3.1).

	Gene Ontology (GO) enrichment				
Genes with an Hsmar1 remnant(s) (1971 genes)	Cell adhesion, 184 genes, adjP = 1.46E-12	Nervous system development, 280 genes, adjP = 5.02E-11			Signal transduction, 568 genes, adjP = 4.60E-10
Genes with an ITR(s) (1517 genes)	Nervous system development, 224 genes, adjP = 3.23E-09	Generation of neurons, 149 genes, adjP = 3.78E-08	Transmission of nerve impulse, 114 genes, adjP = 1.92E-08	Cell adhesion, 137 genes, adjP = 3.30E-08	Cell projection organization, 135 genes, adjP = 3.30E-08
Gene size > 100,000 nucleotides (3423 genes)	Nervous system development, 548 genes, adjP = 2.76E-40	Cell projection organization, 328 genes, adjP = 3.33E-31	Cellular compartment morphogenesis, 333 genes, adjP = 1.99 E-27	Single organism signalling, 1126 genes, adjP = 1.99E-27	
Gene size > 500,000 nucleotides (294 genes)	Nervous system development, 82 genes, adjP = 2.64E-16	Cell adhesion, 55 genes, adjP = 1.39E-13			Transmission of nerve impulse, 43 genes, adjP = 8.21E-11

Table 3-1: GO enrichment for the genes containing an Hsmar1 remnant or an ITR and for the largest genes in the human genome.

The number of genes used for each category is indicated in the left column. The GO terms enrichment has been obtained with Webgelstat (Wang et al., 2013).

3.3 Discussion

In common with other transposable elements, Hsmar1 has left numerous remnants in the human genome (Pace and Feschotte, 2007). The 8,527 Hsmar1 remnants can be divided in four groups: solo ITRs, Made1, truncated

Hsmar1, and full-length Hsmar1. Made1 are specific to anthropoid primates as is SETMAR, the fusion between a histone methyltransferase and an *Hsmar1* transposase gene (Cordaux et al., 2006). A previous study has established that the DNA binding domain but not the catalytic domain of the transposase domain is under purifying selection (Cordaux et al., 2006). The domestication of the transposase could have affected the evolution of the transposon binding sites in the genome if SETMAR requires them for performing its functions.

3.3.1 Hsmar1 remnants in the human genome

The Hsmar1 remnants, and by extension the ITRs, are principally found in the largest human genes. This phenomenon is also observed with other transposable elements such as PiggyBac or the L1 retrotransposon (Yant et al., 2005) (Prufer et al., 2012). This is explained by a combination of integration randomness and natural selection. Some Tc1/*mariner* transposases require open chromatin for integration, so the probability of intragenic integration is positively correlated to the size of the gene. Furthermore, integrations in the largest genes are less likely to affect the gene function because of the inverse correlation between the gene size and the fraction of coding DNA. The smaller probability of reducing host fitness also decreases the prospect of transposon copies to be removed by natural selection.

The Gene Ontology terms (GO) enrichment observed in the genes with Hsmar1 remnants or ITRs is explained by the size of the genes where the

remnants or the ITRs are located rather than an integration bias of Hsmar1 transposons or Made1 elements (Table 3.1). However, if SETMAR regulates the expression of genes containing an ITR we may expect these binding sites to be under purifying selection. This evolutionary conservation could in fact also explain the GO terms enrichment.

3.3.2 Are a fraction of the ITRs under purifying selection?

The analysis of the number of intergenic and intragenic remnants with an ITR revealed a higher fraction of solo ITR, Made1, and truncated Hsmar1 with an ITR in non-coding genes (80% versus 60% for intergenic remnants). A similar but smaller increase is also observed for Made1 and full-length Hsmar1 in protein-coding genes (70% of remnants with an ITR) (Figure 3.3C). This increased conservation of the ITRs in genes hints at a possible purifying selection acting on a fraction of the binding sites. This hypothesis is supported by a preliminary analysis of Made1 elements in several anthropoid primates (Table 3.2, Cedric Feschotte, personal communication). The human Made1 locations and sequences were compared to the Made1 existing in other primates to determine which elements were common to all species and under purifying selection. The analysis found 16 Made1 under purifying selection with half of the elements being intergenic. With the sequencing of new anthropoid primates genomes, this analysis could be done on the ITRs rather than Made1 elements to discover which binding sites are under purifying selection and if they are bound by SETMAR.

Locus (human assembly hg19, Feb. 2009)	Genomic environment
chr1:184795525-184795570	Intron, FAM129A
chr1:39362167-39362245	Intron, RHBDL2
chr2:8305864-8305932	LINC00299
chr3:151679155-151679231	Gene desert*
chr4:101051108-101051187	RP11-15B17.1
chr4:31757539-31757612	Gene desert*
chr5:106767014-106767064	Intron, EFNA5
chr5:163382624-163382696	Gene desert*
chr5:71315991-71316070	Gene desert*
chr6:111849356-111849409	TRAG3IP2-AS1
chr6:160341024-160341103	12 kb downstream of MAS1
chr6:169871229-169871304	Intron, WDR27
chr8:66356903-66356979	Gene desert*
chr18:69494214-69494292	10 kb upstream of RP11-723G8.1
chr21:39277765-39277801	Intron, KCNJ6
chrX:135222985-135223014	5.8 kb upstream of FHL1

Table 3-2: List of Made1 elements under purifying selection.

The list was generated by comparing the location and the sequence of the Made1 elements in different anthropoid primates (Cedric Feschotte, personal communication). Gene desert corresponds to region where no known gene is located at less than 50 kb away from the Made1 element.

3.3.3 A class of miRNAs-like genes derived from Made1

From the most recent mirBase update, there are 69 miRNAs, regrouped in the mir-548 family, derived from Made1 in the human genome (Piriyapongsa and Jordan, 2007). Several members of the mir-548 family have been characterized as regulators of a tumour-suppressor gene, FHIT, and the type III interferon gene, IFN- λ 1 (Hu et al., 2014) (Li et al., 2013b).

The analysis of the ITRs position relative to the start of non-coding genes revealed a skewed distribution towards the start and the three-quarter

position within the genes (Figure 3.6A). The mir-548 family alone cannot explain this bias because there are 456 non-coding genes containing an ITR at these positions (Figure 3.6C). These non-coding genes are derived from Made1 and they are between 61 and 129 nucleotides long with 95% of them longer than 80 bp, the size of the original Made1 elements. These longer sizes explain why the distribution of ITRs does not peak at the start and at the end of the non-coding genes but rather at the start and at three-quarter position.

Since the Made1 elements are not transcriptionally active, the intergenic non-coding genes are transcribed from an upstream promoter whereas the intragenic non-coding genes are transcribed with their mRNA. The functions of these genes are currently unknown. However, their small size and the presence of a Made1 element suggest they may be putative members of the mir-548 family, making it one of the largest families of miRNAs in the human genome. Since these miRNAs-like still contain their Hsmar1 binding sites, they are also good candidates for being regulated by SETMAR.

3.3.4 Hsmar1 remnants in human lncRNAs

The majority of the intragenic ITRs are located in introns (Figure 3.6B). The exonic ITRs are mainly found in miRNAs and miRNAs-like (see Section 3.3.2). Only 7% of the exonic ITRs are located in non-coding and protein-coding genes with at least one intron. These non-coding genes encode for long non-coding RNAs, RNA molecules involved in the regulation of cellular

processes such as gene regulation, pluripotency, or cancer (Rinn and Chang, 2012).

Transposable elements are major contributors of lncRNAs with a few classes of DNA transposons including Hsmar1 being over represented in lncRNAs exons (Kapusta et al., 2013). *PCAT-1*, a long intergenic ncRNA (lincRNAs), is currently the best-described lncRNA containing an Hsmar1 transposon (Prensner et al., 2011). *PTAC-1* is expressed specifically in the prostate and promotes cell proliferation by repressing the expression of several genes (Prensner et al., 2011). The full-length Hsmar1 transposon is inactivated by the insertion of an AluY element in the transposase gene but the second *PTAC-1* exon and the polyA signal is entirely provided by the Hsmar1 transposon. The role of the second exon and by extension of the Hsmar1 transposon in *PTAC-1* function is currently unknown. However, around 30 Hsmar1 transposons and Made1 elements are currently found in lncRNAs (Kapusta et al., 2013). These Hsmar1 remnants contribute their sequences to exons and provide polyA signal and splicing sites. Further investigation is needed to determine whether these exonic Hsmar1 and Made1 are necessary for the function of the lncRNAs.

4. SETMAR IS AN EPIGENETIC REGULATOR OF A GENE NETWORK

4.1 Introduction

The objective of the work described in this Chapter is to determine whether SETMAR acts as an epigenetic regulator of a gene network. To accomplish this, I used RNA-Seq and ChIP-exo to investigate which genes are differentially regulated when SETMAR expression is modified and whether these candidate genes have a transposon end bound by SETMAR. Modest overexpression of SETMAR modified the expression of thousands of genes in a human cell line. Amongst these differentially expressed genes, a strong enrichment is observed in genes containing an ITR. However, only 20% of the genes with an ITR bound by SETMAR are also differentially expressed, which represents ~2% of the genes with an ITR found in the RNA-Seq data. SETMAR binds essentially to Made1 elements rather than solo ITRs. SETMAR could fine-tune the expression of the genes with an ITR by interacting with other regulatory proteins or recruiting them by dimethylating the H3K36. However, SETMAR binding sites are also enriched in the miRNAs-like described in Chapter 3. This suggests a possible regulation by SETMAR of the expression of these small RNAs, which could in turn regulate the genes differentially expressed in the RNA-Seq experiments.

4.2 Results

4.2.1 Stable cell lines used for RNA-Seq and ChIP-exo

To determine whether SETMAR can regulate the genes containing an ITR, I first investigated whether altering SETMAR expression results in a modification of gene expression in human cells. In principle, SETMAR expression could be changed either by repression using an shRNA or by overexpression from an exogenous copy of the gene. I decided against the first approach because the natural level of SETMAR expression is below the sensitivity of a Western blot and there is therefore no easy way to quantify a reduction in its expression. Thus, although SETMAR mRNA can be detected in most tissues by PCR, the protein level is low. This is supported by a quantitative mass spectroscopy experiment on the U2OS cell line, which suggested the presence of less than 500 SETMAR molecules per cell (Beck et al., 2011b). I therefore decided to engineer a various increased levels of SETMAR expression. I select a U2OS cell line where the T-Rex system repressor was already integrated in the genome. This system is based on the Tetracycline operator-repressor where the addition of tetracycline to the medium induces the expression of the gene of interest. To differentiate in the RNA-Seq data between the endogenous and overexpressed mRNA, I used an artificial SETMAR gene that had been codon optimized and therefore has quite a distinct DNA sequence from the endogenous gene. A second version was tagged with a C-terminal FLAG epitope to aid detection and quantification.

Since the transfection process itself can stress cells and change their gene expression profile, I wished to create stable cell lines in which the SETMAR overexpression cassette was integrated in the genome. After drug selection, I expanded several foci, which were grown on and assessed for SETMAR expression level. I used western blotting with an antibody directed against the last nine amino acids of SETMAR, or directed against the FLAG epitope, depending on whether the protein was tagged or not. Surprisingly, many of the cell lines expressed significant levels of SETMAR in the absence of tetracycline induction. I therefore decided to compare these cell lines directly and thus avoid any artifactual changes in gene regulation that might be caused by tetracycline treatment. I selected cell lines with a range of SETMAR concentrations. SETMAR expression level is determined by the number of vector integrated in the genome and their genomic locations. These are designated as SM1, SM2 and SM3 for the native protein and SMF2 for the FLAG tagged proteins, respectively (Figure 4.1A). As a control, I also created the cell line SETF2, expressing the FLAG tagged set domain of SETMAR. Further controls were provided by the parental cell line TO and TOF, which contains an integrated copy of the T-Rex expression plasmid (Figure 4.1A).

Western blotting with the SETMAR antibody was unable to detect SETMAR expression in TO, TOF, or SM1 (Figure 4.1A). However, cell lines SM2 and SM3 produced detectable levels of the protein. The FLAG antibody detected the tagged SETMAR and SET-domain proteins in cell lines SMF2 and SETF2, respectively. Further, SETMAR was detected in the SMF2 cell line with the FLAG antibody but not with the SETMAR antibody. This shows

that the FLAG antibody is more sensitive (Figure 4.1A). Cell lines expressing the untagged SETMAR were subsequently used in the RNA-Seq experiments, while SMF2 was used for ChIP-exo analysis (below). The SETF2 cell line is not analyzed further in this Chapter.

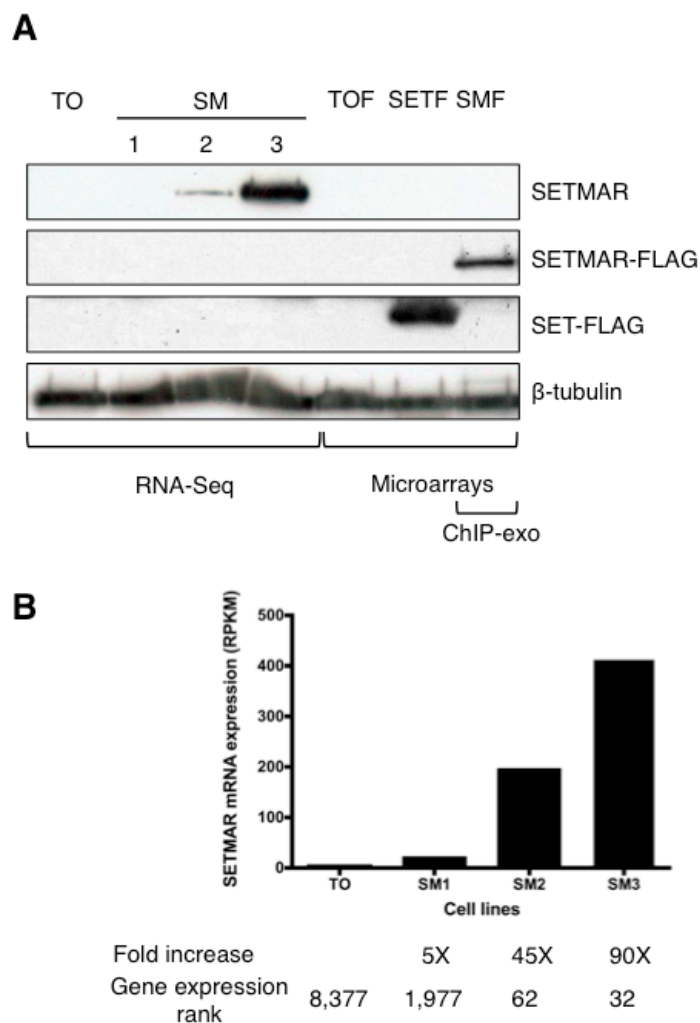


Figure 4- 1: Cell lines overexpressing SETMAR used for the RNA-Seq and ChIP-exo experiments.

(A) Expression of SETMAR in the different T-Rex-U2OS cell lines used for RNA-Seq and ChIP-exo. The SETMAR protein was detected with an antibody raised against the last 9 amino acids of the native protein or a monoclonal antibody against the FLAG tag. Full pictures of the Western blot are presented in Supplementary Figure 1

(B) mRNA expression of SETMAR in the cell lines used for RNA-Seq. The overexpression fold and the gene expression rank are indicated below the figure.

RNA was extracted from cell lines TO, SM1, 2, and 3 and prepared for RNA-Seq using the SOLiD system (M&M section 2.3.7). The analysis was done on biological duplicates for each cell lines to assess the reproducibility of the transcriptomic analysis. The data sets also provided quantification of the amount of SETMAR mRNA in the different cell lines (Figure 4.1B). In the control cell line, TO, SETMAR is ranked as the 8,377th out of the 29,355 expressed genes. Compared to this level, cell lines SM1, SM2 and SM3 overexpress SETMAR by 5-, 45-, and 90-fold. In these cell lines SETMAR is the 1,977th, 62th, and 32nd most expressed gene. Even though there is only a 2-fold difference at the mRNA level between SM2 and 3, the difference appears much greater at the protein level (Figure 4.1A). However, this may reflect the dose-response curve of the western blot rather than the actual level of the protein present.

4.2.2 Between 20 and 35% of the differentially expressed genes are unique to each cell line

To determine which genes are differentially expressed, each cell line overexpressing SETMAR was compared to the control cell line and only the genes with a false discovery rate corrected p-value < 0.001 were retained (Table 4.1). Approximately 5,000 genes were differentially expressed in each cell line, with about four-fifths of them with a fold change between -2 and 2. The change in gene expression is only observed between cell lines. No significant variation is observed between replicates.

	Control vs. SM1	Control vs. SM2	Control vs. SM3
Number of differentially expressed genes	5,171	4,477	5,644
Number of differentially expressed genes > 2 fold	1,091	856	954

Table 4-1: Number of differentially expressed genes in each cell line.

Number of differentially expressed genes with a p-value < 0.001 in SM1, 2, and 3 compared to the control cell line.

I next determined the overlap between the sets of differentially expressed genes in the three cell lines (Figure 4.2A). If all significant (p < 0.001) genes are counted, approximately 35% of the differentially expressed genes are common to the three cell lines. The number of genes found in two cell lines represents 40% of SM1 and 2 but only 30% of SM3. The remaining genes, 20% for SM1 and 2 and 35% for SM3, are unique to each cell line.

If genes with at least a 2-fold change are counted the number of genes common to the three cell lines decreases to 25% of SM2 and 20% of SM1 and 3. This decrease is balanced by an increase in the number of genes unique to each cell line. The remaining genes are common to two cell lines and account for 40% of SM1 and 2 and 25% of SM3.

I next investigated the overlap between the sets of differentially expressed genes that encode an ITR (Figure 4.2B). The results were broadly similar to the complete set of differentially expressed genes in Figure 4.2A.

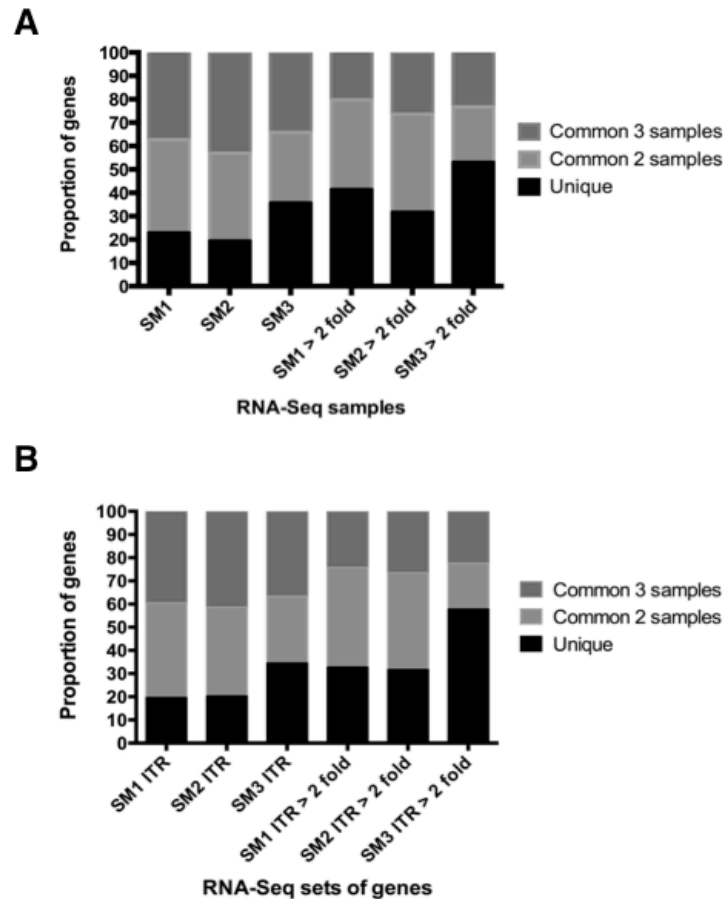


Figure 4-2: The SM3 cell line and the genes with at least a 2-fold change contain a higher proportion of unique genes.

(A) Overlap between the three sets of differentially expressed genes. **(B)** Overlap between the three sets of differentially expressed genes with an ITR.

To determine whether the genes differentially expressed behave in a similar way, I compared for each gene its expression in the control cell line and in the SM1, 2, and 3 cell lines (Figure 4.3A, B, and C). The genes with an ITR down- or up-regulated are indicated in orange and blue, respectively. The majority of the differentially expressed genes with an ITR have a low to medium expression, i.e. a $\log_2(\text{RPKM})$ between 0 and 5, in the control cell line (Figure 4.3A, B, and C).

To investigate whether the genes with an ITR behave differently to genes without an ITR, I compared the proportion of down- and up-regulated

genes without an ITR, in grey and black at the top left of each graph, respectively, to the genes with an ITR, in the bottom right of each graph (Figure 4.3A, B, and C). For the three cell lines overexpressing SETMAR, I observed a greater up-regulation of the genes with an ITR than for the genes without an ITR. The effect is the most visible for the SM3 cell line where ~70% of the genes with an ITR are up-regulated compared to half of the genes without an ITR (Figure 4.3C).

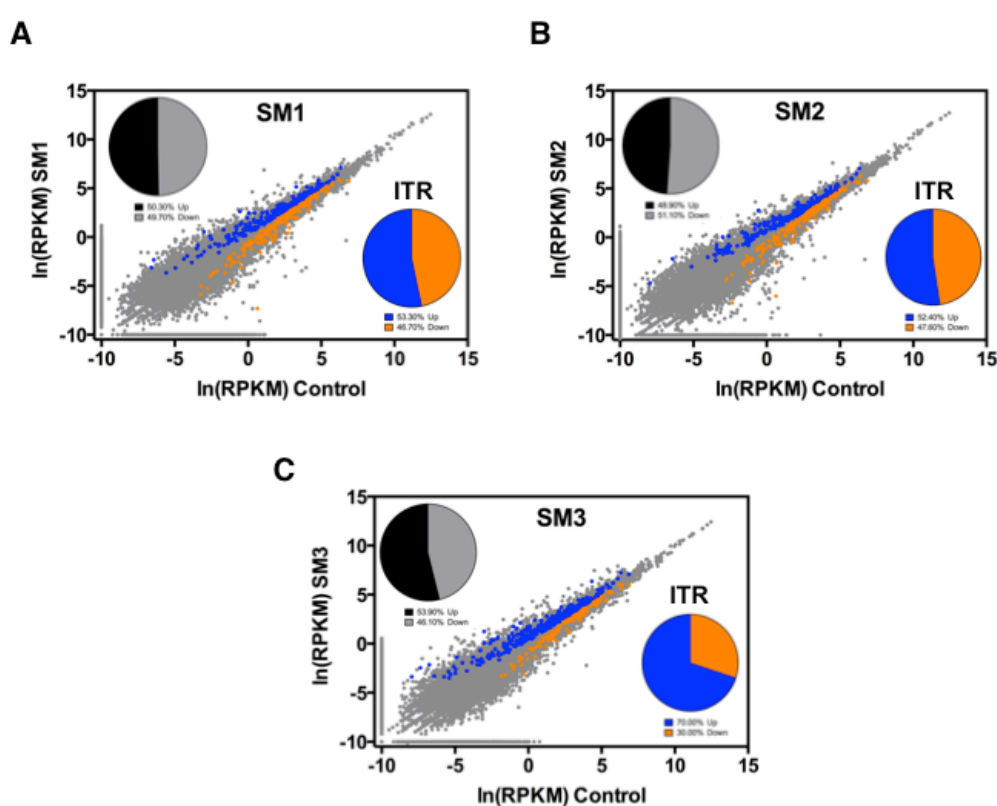


Figure 4-3: SETMAR overexpression correlates with an upregulation of the genes with an ITR expressed at a low or medium level.

Comparison of the expression of every gene in the control cell line and SM1 (A), SM2 (B), and SM3 (C). The up-regulated genes with an ITR are indicated in blue whereas the down-regulated genes with an ITR are indicated in orange. The proportion of down- and up-regulated gene with an ITR is indicated on the bottom right of each graph. The proportion of down- and up-regulated genes without an ITR is represented in grey and black, respectively, on the top left of each graph.

I next investigated whether the expression and fold change of genes with or without an ITR were different (Figure 4.4A, B, and C). However, for both type of genes the less expressed genes are the most misregulated. Furthermore, there is no visible difference between the genes with or without an ITR.

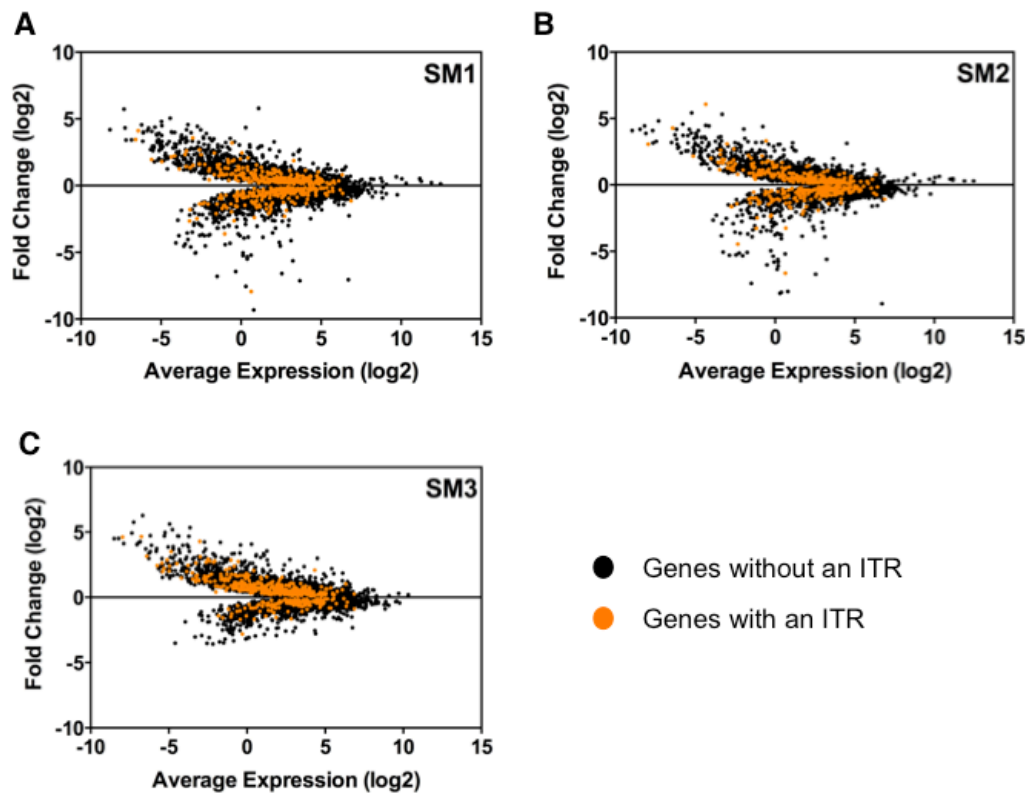


Figure 4-4: The highest fold changes are observed for the lowest expressed genes.

Comparison for each differentially expressed gene of its expression in the control cell line versus its fold change in SM1 (A), 2 (B), and 3 (C). The genes without or with ITR are indicated in black and orange, respectively.

4.2.3 SETMAR overexpression is associated with an enrichment in differentially expressed genes containing an ITR

It seems likely that any changes in gene expression caused directly by SETMAR will be accompanied by secondary changes. For example, if SETMAR acts directly on a transcription factor, it could start a cascade of regulatory changes, which spread through the cellular network. The most likely candidates for direct action are the genes near or containing an ITR. I first analyzed whether the differentially expressed genes are enriched in genes near an ITR (Table 4.2). I defined as 'near' an ITR all genes (ORFs) that do not contain an ITR but are less than 10 kb from an ITR. The 10 kb threshold was chosen because it should encompass any short and medium range regulatory effects of SETMAR on gene regulation (Maston et al., 2006). To detect any enrichment, I compared the proportion of genes near an ITR in each cell line to the proportion of expressed genes near an ITR in the U2OS cell line. I select the expressed genes and not the total number of genes because all the differentially expressed genes in the RNA-Seq data are expressed in the control cell line. The list of expressed genes is defined as any gene with an RPKM value greater than zero in the control cell line. This analysis revealed that there was no significant enrichment in differentially expressed genes near an ITR ($p < 0.5$). SETMAR therefore does not appear to regulate this category of genes.

	Human genome	SM1	SM2	SM3
Number of genes near an ITR (< 10 kb)	814	153	136	159
Total number of genes	29,355	5,171	4,477	5,644
p-value		0.024	0.019	0.035

Table 4-2: No enrichment in differentially expressed genes near an ITR.

No specific enrichment in genes near an ITR (<10 kb) is observed in the RNA-Seq samples irrespectively of the fold difference.

I next investigated whether the genes containing an ITR are enriched amongst the differentially expressed genes (Table 4.3). Statistical analysis found a strong enrichment in differentially expressed genes with an ITR in each cell line ($p < 10^{-22}$). Even if I only keep the genes common to the three cell lines, a significant enrichment ($p < 10^{-11}$) is still obtain supporting the direct targeting of SETMAR to the genes with an ITR.

	Human genome	SM1	SM2	SM3	Common (SM1&2&3)
Number of genes with an ITR	1,735	460	440	496	182
Total number of genes	29,355	5,171	4,477	5,644	1,926
p-value		4.55E-22	5.41E-30	6.77E-23	6.23E-11

Table 4-3: Strong enrichment in differentially expressed genes with an ITR.

An enrichment in genes with an intragenic ITR is observed in the RNA-Seq samples irrespectively of the fold difference.

The first analysis was done on all the differentially expressed genes irrespectively of the fold difference. I next determined the extent of enrichment in the genes with at least a 2-fold change in expression (Table 4.4). The

enrichment in this set was weaker in each cell line ($p < 10^{-8}$). If I kept only the common genes between the three cell lines, the enrichment dropped ($p < 10^{-4}$).

	Human genome	SM1 > 2 fold	SM2 > 2 fold	SM3 > 2 fold	Common > 2 fold (SM1&2&3)
Number of genes with an ITR	1,735	107	98	116	26
Total number of genes	29,355	1,076	828	939	218
p-value		4.20E-08	2.39E-11	2.07E-14	3.03E-04

Table 4-4: Strong enrichment in genes with an ITR and at least a 2-fold change in expression.

An enrichment in genes with an intragenic ITR is observed in the genes with at least a 2-fold difference in expression.

I investigated whether the enrichment differs between down- and up-regulated genes (Table 4.5). Except for the SM3 cell line, an enrichment in differentially expressed genes with an ITR is observed for both down- and up-regulated genes in the SM1 and 2 cell lines ($p < 10^{-8}$). For the SM3 cell line, only the up-regulated genes are significantly enriched in genes with an ITR ($p < 10^{-33}$).

	Human genome	SM1 down	SM1 up	SM2 down	SM2 up	SM3 down	SM3 up
Number of genes with an ITR	1,735	215	245	209	230	149	347
Total number of genes	29,355	2,551	2,606	2,258	2,191	2,515	3,114
p-value		1.80E-08	4.09E-14	1.44E-11	1.52E-18	0.04	3.28E-33

Table 4-5: Enrichment in down- and up-regulated genes with an ITR.

An enrichment in down- or up-regulated genes with an intragenic ITR is observed in the RNA-Seq samples irrespectively of the fold difference.

I next focused on the genes with more than a 2-fold change (Table 4.6).

The enrichment in this set was weaker for each cell line ($p < 10^{-5}$).

	Human genome	SM1 down > 2 fold	SM1 up > 2 fold	SM2 down > 2 fold	SM2 up > 2 fold	SM3 down > 2 fold	SM3 up > 2 fold
Number of genes with an ITR	1,735	48	59	40	58	18	98
Total number of genes	29,355	456	620	299	529	251	688
p-value		4.08E-05	9.38E-05	1.38E-07	2.41E-06	0.07	3.34E-16

Table 4-6: Enrichment in down- and up-regulated genes with an ITR and at least a 2-fold change.

An enrichment in down- or up-regulated genes with an intragenic ITR is observed in the genes with at least a 2-fold difference in expression.

In the set of all differentially expressed genes, about 9% contain an ITR.

The proportion remains similar in the three cell lines. However, if I only keep

the genes with more than a 2-fold change, the proportion containing an ITR increases with increasing SETMAR expression. Thus, 9.9%, 11.8%, and 12.4% of the differentially expressed genes have an ITR in the SM1, 2, and 3 cell lines, respectively.

4.2.4 SETMAR expression level correlates with a decrease in the strength of genes down-regulation

The differentially expressed genes are either down- or up-regulated by SETMAR overexpression. To establish whether SETMAR overexpression is associated with an inhibition or activation of gene expression, I placed the differentially expressed genes into sets according to whether they are underexpressed or overexpressed and whether the fold change was less or greater than 2-fold (Figure 4.4A). The proportion of under and overexpressed genes is evenly distributed in the three cell lines. About 80% of the genes changed expression by less than two fold up or down. For the genes with more than a 2-fold change, the proportion of overexpressed genes remains stable at 12% irrespectively of the level of SETMAR overexpression. However, for the underexpressed genes the proportion falls from 10% to 7% and 5% for SM1, 2, and 3, respectively.

I next investigated whether the genes with an ITR are also distributed in a similar way (Figure 4.4B). There is no difference between genes with and without an ITR for the SM1 and 2 cell lines. However, the proportion of up-regulated genes with an ITR increases to 70% for the SM3 cell line. The increase is essentially due to the genes with less than a 2-fold change.

This general analysis suggested an inverse correlation between SETMAR expression level and the strength of gene down-regulation. To determine whether the direction and strength of response is affected by an increased SETMAR expression, I divided the genes into 12 categories depending on their fold change (Figure 4.4C). Compared to SM1, SM2 has the same proportion of genes with a fold change inferior to -5. However, a 25% decrease is observed in the number the genes with a fold difference between -5 and -2. In SM3, the number of genes with more than a 2-fold decrease in expression is reduced. There are virtually no genes with a fold change inferior to -10 and there is a 50% decrease in the genes with a fold-change between -2 and -10. Except for a slight increase of the proportion of genes with an increase below 3-fold in SM3, the proportion of up regulated genes is distributed equally between the three cell lines.

I next determined whether a similar effect is also observed for the subset of genes with an ITR (Figure 4.4D). Compared to SM1 there is only a slight decrease in the proportion of down regulated genes in SM2. In contrast, in SM3 the number of down regulated genes is greatly reduced, while the number of up-regulated genes is greatly increased, compared to SM1.

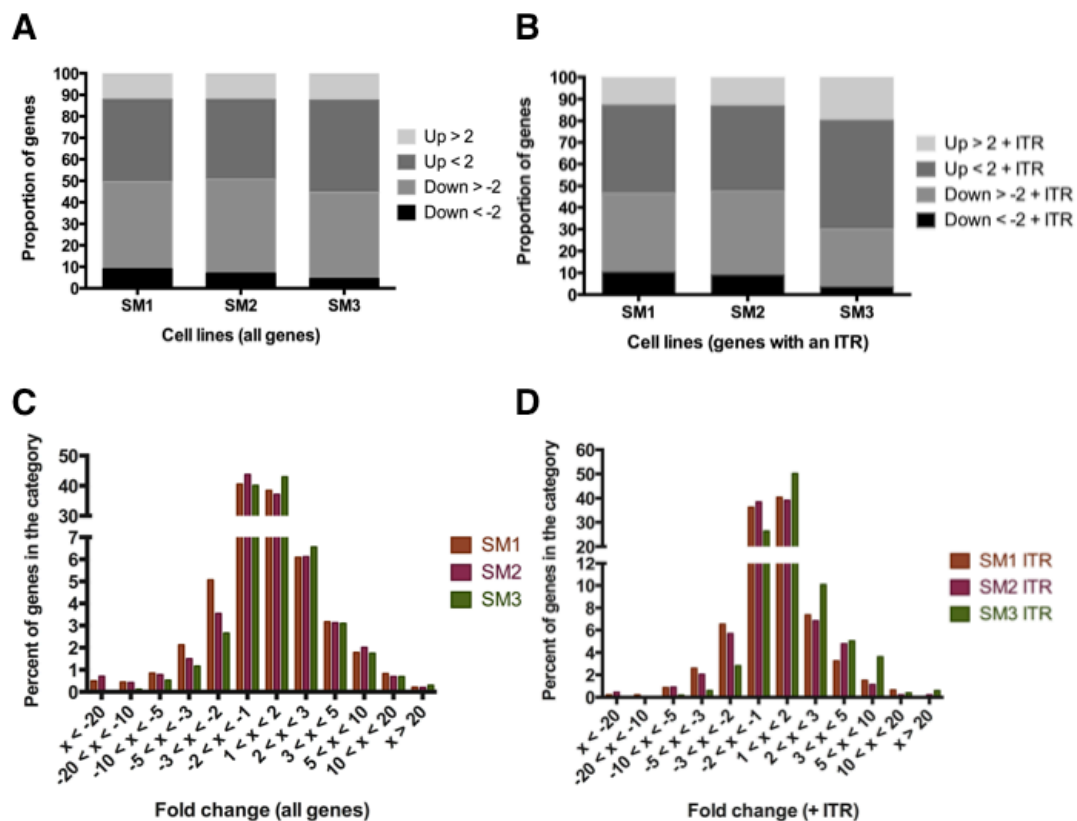


Figure 4-5: Inverse relationship between SETMAR expression and gene down-regulation.

(A) Distribution of all differentially expressed genes depending of their down- or up-regulation and their fold change. (B) Distribution of differential genes with an ITR by their up down- or up-regulation and their fold change. (C) Distribution of the differentially expressed genes depending of their fold change. (D) Distribution of the differentially expressed genes with an ITR by their fold change. An inverse relation is observed between the level of SETMAR expression and the number of genes down-regulated more than 2-fold.

4.2.5 A change in gene expression is only observed in active genes

A regulatory protein can either turn a gene on, turn it off or modulate the level of its expression. To define how SETMAR might act, I first investigated the number of genes with more than a 20-fold difference in expression, which can

be considered as being turned on or off (Figure 4.4C). For the three cell lines, less than 4% of the genes with more than a 2-fold change are being turned on or off. For the SM1 and SM2 cell lines, two-thirds of the genes with more than a 20-fold change are being turned off. However, for the SM3 cell line all the genes with more than a 20-fold change are being turned on.

To investigate whether SETMAR might modulate gene expression, I retrieved the relative expression level of every gene in the control cell line and classify them in the following categories and sub-categories: differentially expressed genes (DE), non-differentially expressed genes (NDE), DE with or without an ITR (DE ITR and DE NITR), NDE with or without an ITR (NDE ITR and NDE NITR) (Figure 4.5A). The RPKM value, which corresponds to the number of reads per kilobase of exon per million of mapped reads, does not represent the absolute expression level but the relative expression of a gene compared to all the other genes. The RPKM measure allows the comparison of transcript levels both within and between samples (Mortazavi et al., 2008). I merged the list of differentially expressed genes found in the three cell lines to retrieve the list of genes differentially expressed in at least one cell line. The NDE list corresponds to the list of human genes minus the DE list. On average, the DE genes are 18-times more expressed than NDE genes and a similar increase is observed between DE and the NDE genes without an ITR. For the genes with an ITR, the DE genes are on average 9-times more expressed than their NDE counterparts.

Because SETMAR may directly regulate the expression of the DE genes with an ITR, I investigated whether the genes found in one, two, or three cell lines differ in their average expression levels (Figure 4.5A, last three bars).

From the global results, I know that DE genes are more expressed than NDE genes. The DE genes found in more than one cell line could therefore, on average, be more highly expressed than the unique genes. This is indeed the case. For the DE genes with an ITR there is more than a 2-fold increase in the average expression for the genes found in 2 or 3 cell lines compared to genes found in only one.

To get another view of the distribution of the expression level in the different sets of genes, I divided them into the following categories according to their initial expression level in the parental cell line: null to low (0-1), medium (1-10), high (10-100), and very high (> 100) (Figure 4.5B). For each set of genes, only up to 2% of the members were initially highly expressed. For the genes without an ITR, 90% of the DE genes have a medium and high expression whereas for the NDE genes, these two categories represent less than 40% of the total. For the genes with an ITR, the difference is quite similar with 80 and 40% of respectively the DE and NDE genes having a medium or a high initial expression level. The decrease in highly expressed genes observed between the DE genes without and with an ITR is owing to the genes with an ITR found in only one cell line (Figure 4.5B, last three columns). Within this group there are 2-times more genes with a null or low expression and half the number of highly expressed genes. Overall, these results suggest that SETMAR may regulate only the genes located in open chromatin.

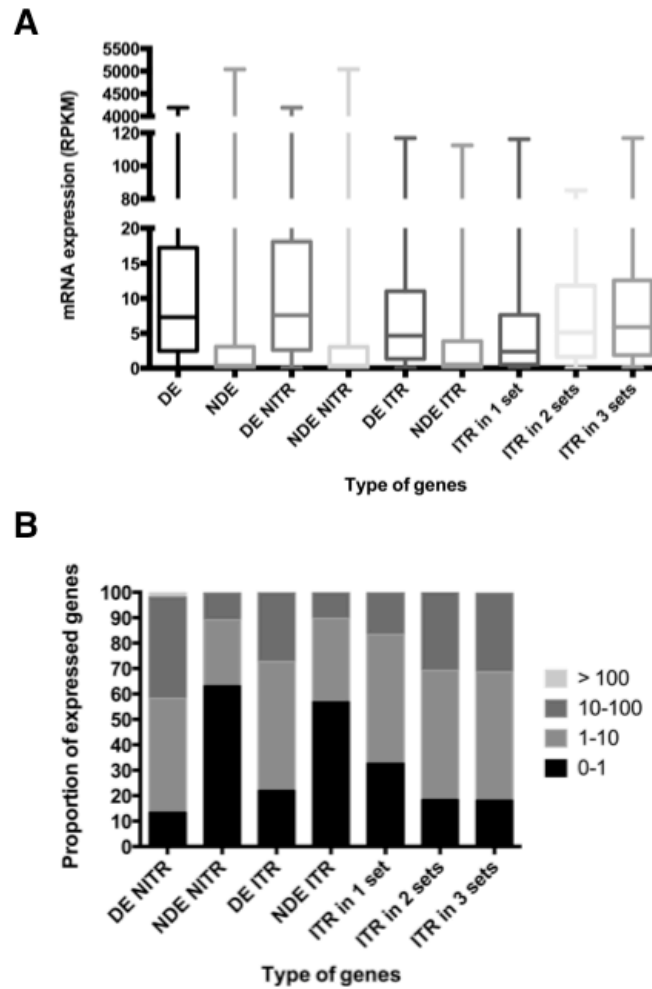


Figure 4-6: The differentially expressed genes are on average more expressed than the non-differentially expressed genes.

(A) Box and whisker plot representing the distribution of gene expression in different categories of genes. DE: Differentially expressed genes; NDE: Non-differentially expressed genes; NITR: genes without an ITR; ITR: genes with an ITR; 1, 2, 3 samples: genes found to be differentially expressed in 1, 2, or 3 RNA-Seq sets, respectively. **(B)** Distribution of gene expression in the different categories of genes. 0-1: no or low expression; 1-10: medium expression; 10-100: high expression; > 100: very high expression.

4.2.6 The differential expression of genes does not depend on whether they have a solo ITR or a Made1 element

Most of the ITRs in the human genome are solo ITRs and Made1 elements. As noted in Chapter 3 these are transposon remnants and miniature transposons, respectively. I determined whether the differentially expressed genes with an ITR are more likely to contain one or more of these types of elements (Figure 4.6A). In the human genome, 60% of the genes with an ITR have their binding sites from a Made1 element. A similar proportion is observed for the differentially expressed genes with an ITR in the three cell lines. Focusing on the genes with a 2-fold change does not modify the proportion of genes with an ITR from a Made1 element.

I next investigated whether the DE genes with an ITR present in one, two, or three cell lines were enriched in solo ITRs or Made1 elements (Figure 4.6B). Similar to the results obtained with individual cell lines, no significant enrichment is observed for the differentially expressed genes with an ITR found in 1, 2, or 3 RNA-Seq sets. The analysis of genes with at least a 2-fold change did not modify significantly the distribution. The variability for the genes common to two or three cell lines is due to the small number of genes.

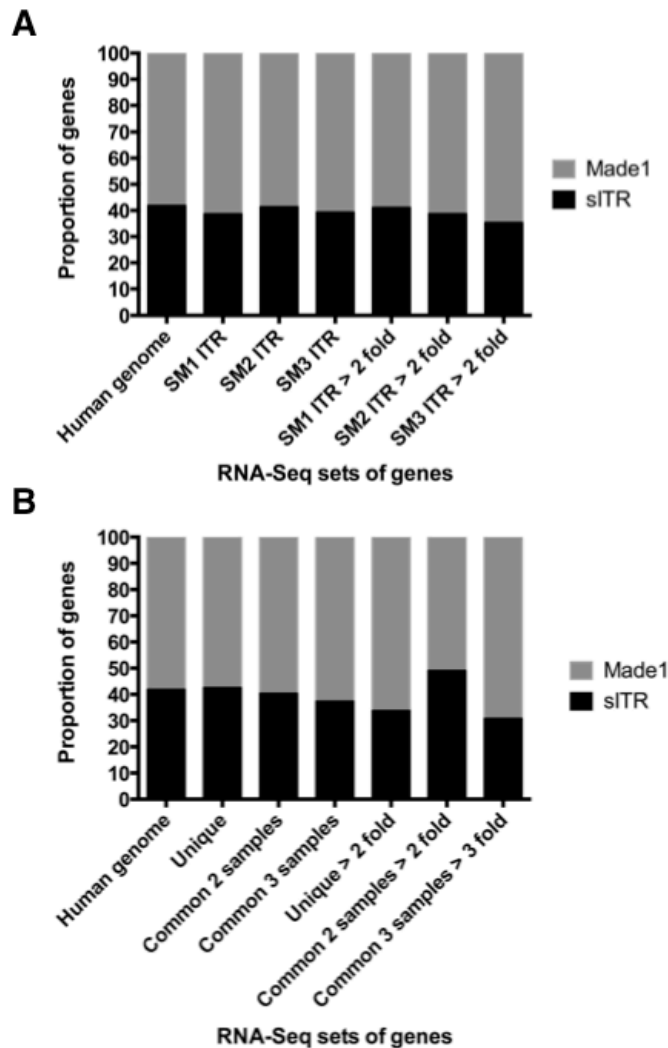


Figure 4-7: The differentially expressed genes with an ITR are not enriched in solo ITRs nor Made1 elements.

(A) Distribution of the solo ITRs and Made1 elements in the human genes with an ITR or in the three sets of differentially expressed genes with an ITR. **(B)** Distribution of the solo ITRs and Made1 elements in the human genes with an ITR or in the differentially expressed genes with an ITR found in 1, 2, or 3 RNA-Seq sets.

4.2.7 Gene Ontology (GO) enrichment in the different RNA-Seq data

To get a better idea of the functions of the DE genes, I first determined whether specific Gene Ontology (GO) terms are enriched in each cell line

irrespectively of the presence of an ITR (Table 4.7). The analysis was restricted to genes with at least a 2-fold change in gene expression. In all three cell lines there was an enrichment in 'System development' and 'cardiovascular or blood vessel development' genes. In addition, SM1 and SM3 are enriched in 'Cell adhesion', SM1 for 'Response to wounding', SM2 for 'Immune system process' and SM3 for 'Response to external stimulus'.

I next established whether there is an enrichment in the genes common to the three cell lines (Table 4.7). If I do not take into account the fold change, the common genes are enriched in the 'Organelle organization' function and in terms related to proliferation such as 'Cell cycle' and 'Nuclear division'. However, these enrichments disappear if I keep only the common genes with at least a 2-fold change.

	Gene Ontology (GO) enrichment			
SM1 genes > 2 fold (866 genes)	Response to wounding, 113 genes, adjP = 6.46E-12	Cell adhesion, 95 genes, adjP = 1.15E-09	System development, 242 genes, adjP = 9.67E-09	Cardiovascular system development, 75 genes, adjP = 1.95E-07
SM2 genes > 2 fold (659 genes)	Immune system process, 107 genes, adjP = 2.52E-05		System development, 182 genes, adjP = 1.85E-05	Blood vessel morphogenesis, 40 genes, adjP = 2.52E-05
SM3 genes > 2 fold (741 genes)	Cell adhesion, 86 genes, adjP = 2.01E-09	System development, 213 genes, adjP = 5.55E- 08	Response to external stimulus, 102 genes, adjP = 9.50E-08	Blood vessel development, 40 genes, adjP = 2.52E-05
Common genes (1805 genes)	Cell cycle, 291 genes, adjP = 2.10E-30		Organelle organization, 367 genes, adjP = 1.26E-22	Nuclear division, 107 genes, adjP = 8.40E-21
Common genes > 2 fold (178 genes)	Cardiovascular system development, 22 genes, adjP = 2.50E- 03		Extracellular matrix organization, 11 genes, adjP = 2.50E-03	Regulation of multicellular organismal process, 34 genes, adjP = 1.52E-02

Table 4-7: GO enrichment of all differentially expressed genes in the RNA-Seq samples.

The lists of differentially expressed genes were compared to every human gene to detect any enrichment in GO terms.

I next focused on GO enrichments in the differentially expressed genes containing an ITR (as determined in Section 3.2.2.3). When I considered the genes with at least a 2-fold change (Table 4.8) no enrichment is observed. However, when the total set of DE genes is considered several weak enrichments are seen: SM2, 'Chromosome organization', 'Cellular protein

modification process', and 'Regulation of Jun kinase activity' terms; SM3, only the 'Cellular component organization at cellular level' term.

Since there is no enrichment found for the common genes with an ITR, I merged the list of genes with an ITR present in each cell line and analyzed the resulting list of 743 genes against the human genes (Table 4.8). I observed an enrichment in 'Nervous system development', 'Neuron differentiation', and 'Cell projection morphogenesis'. These are the same GO terms as found for the set of all genes containing an ITR (see Chapter 3). A comparison of the 743 genes against the list of human genes with an ITR led to only one weakly enriched GO term not found in the SM2 and 3 cell lines, 'Organelle organization'. If I kept only the genes with an ITR with a 2-fold change in expression, a weak enrichment in GO terms is only found when compared to all human genes. Two terms are related to synapse, 'Synapse organization' and 'Synaptic transmission', and a third one in 'Cell adhesion'.

	Gene Ontology (GO) enrichment		
SM2 genes with an ITR vs. Genes with an ITR (418 genes)	Chromosome organization, 32 genes, adjP = 2.97E-02	Cellular protein modification process, 90 genes, adjP = 4.78E-02	Regulation of Jun kinase activity, 7 genes, adjP = 4.78E-02
SM3 genes with an ITR vs. Genes with an ITR (472 genes)	Cellular component organization at cellular level, 142 genes, adjP = 4.81E-02		
Differentially expressed genes with an ITR (743 genes)	Nervous system development, 123 genes, adjP = 7.01E-07	Neuron differentiation, 80 genes, adjP = 2.44E-06	Cell projection morphogenesis, 62 genes, adjP = 1.80E-06
Differentially expressed genes with an ITR vs. Genes with ITR (743 genes)	Organelle organization, 127 genes, adjP = 1.07E-02		
Genes with an ITR and > 2 fold (194 genes)	Synapse organization, 9 genes, adjP = 5.60E-03	Synaptic transmission, 21 genes, adjP = 5.60E-03	Cell adhesion, 24 genes, adjP = 1.64E-02

Table 4-8: GO enrichment of the differentially expressed genes with an ITR in the RNA-Seq samples.

The list of differentially expressed genes with an ITR was compared to the list of every gene with an ITR to detect any specific enrichment in the list of overrepresented GO terms defined in the Chapter 3.

4.2.8 No significant enrichment in the ITR position compared to the start of the gene or the nearest splicing site in the differentially expressed genes

In the previous Chapter I mapped the distribution of the ITRs compared to the start of genes and/or the nearest splice site. Since the overexpression of SETMAR increased the differential expression of genes containing an ITR, I

investigated whether the location of the ITR within these genes was biased or random. Because the DE genes with an ITR are at 95% protein-coding genes, I compared the DE protein-coding genes with an ITR from the SM1, SM2, and SM3 cell lines to the distribution of all the protein-coding genes with an ITR.

I first examined the location of the ITR compared to the start of the gene (Figure 4.7A). I present for each decile the distribution established in the Figure 3.6A and the distribution obtained from the DE genes with an ITR from the SM1, SM2, and SM3 cell lines. Except for the enrichment in the 7th decile for SM3 ($p < 0.01$), the ITRs are similarly distributed between the protein-coding genes and the DE protein-coding genes in the three cell lines.

Since there is no general ITR location enriched in the DE genes, I investigated whether the expression fold change and the ITR location were correlated (Figure 4.7B). It is possible that the expression fold change could be higher for the genes with an ITR near the start. The fold change values were log2 transformed to aid presentation. The horizontal lines at $y = -0.3$ and 0.3 represents the -2-fold and +2-fold change cutoffs. There is no visible correlation between the fold change and the position of the ITR in the gene for the three cell lines.

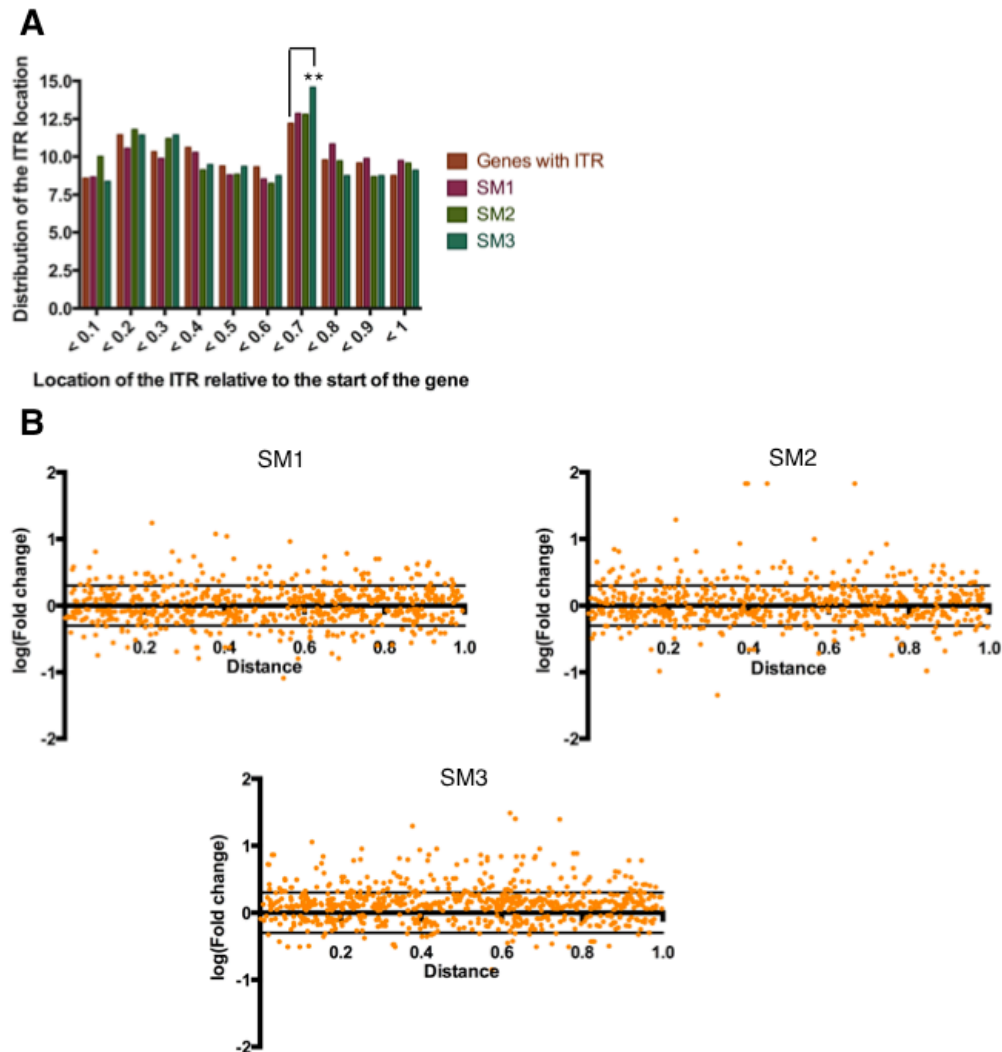


Figure 4-8: The position of the ITR compared to the start of the gene is not correlated to transcriptional changes.

(A) Distribution of the location of the ITR relative to the start of the gene in protein-coding genes. **: $p < 0.01$, hypergeometric test. **(B)** Comparison of the position of the ITR relative to the start of the gene and the log₂ transformed fold changes for the three RNA-Seq sets. The lines at $y = -0.3$ and $+0.3$ represent the 2-fold difference in gene expression threshold.

I next investigated whether there is an enrichment in differentially expressed genes for an ITR position compared to the nearest splicing site (Figure 4.8A). Except for the enrichment of the ITR found in the middle of the intron or exon in the SM2 cell line ($p < 0.01$), there is no significant difference

between the distribution of ITRs in the protein-coding genes and in the DE protein-coding genes in the SM1, SM2, and SM3 cell lines.

I also examined whether the distance between the ITR and a splicing site was correlated to the fold change (Figure 4.8B). If SETMAR can influence the mRNA splicing of genes with an ITR, this could trigger a change in the number of mRNA molecules. If there is an effect, the genes with an ITR near a splicing site may be enriched for a higher fold change. However, there is no visible correlation between the fold change and the location of the ITR in the intron or exon.

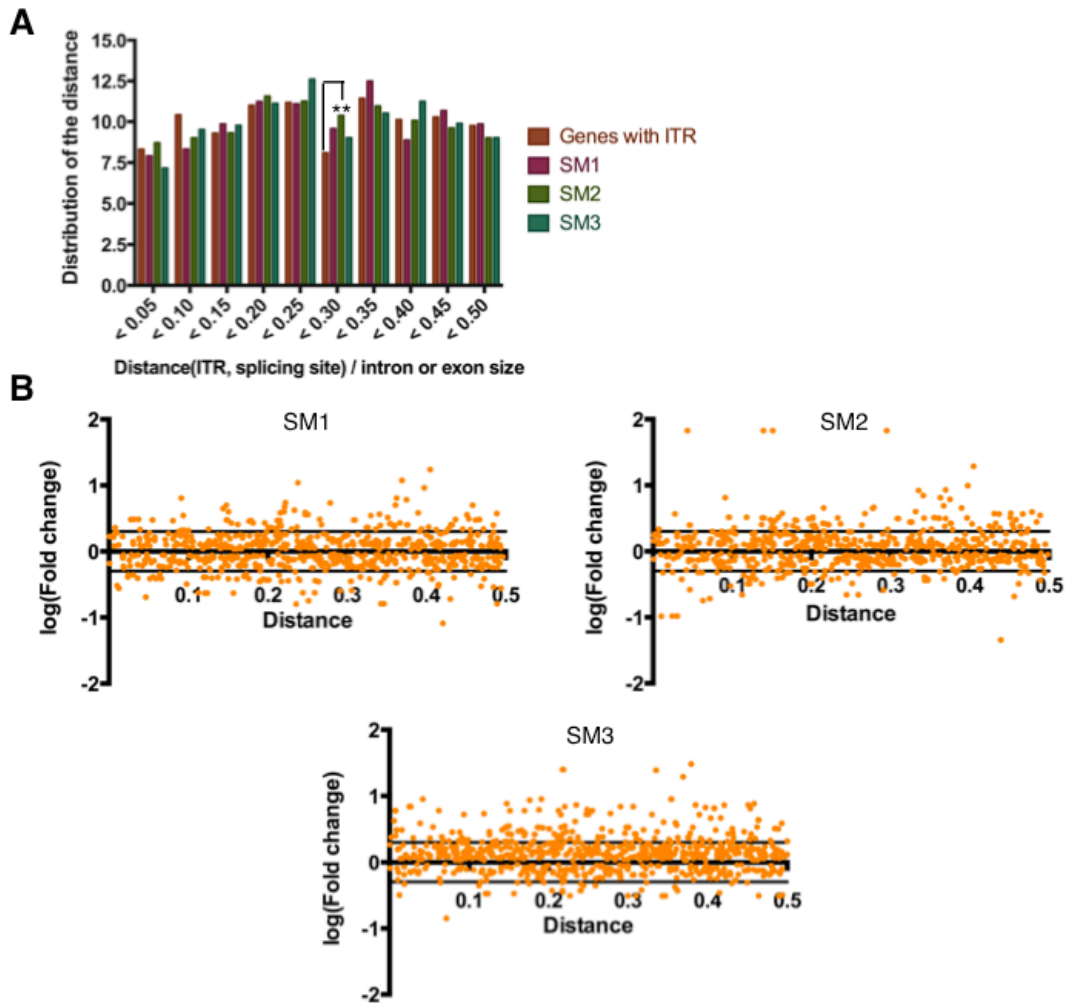


Figure 4-9: The position of the ITR compared to the nearest splicing site is not correlated to transcriptional changes.

(A) Distribution of the ITR relative to the nearest splicing site in protein-coding genes. **: $p < 0.01$, hypergeometric test. **(B)** Comparison of the position of the ITR relative to the nearest splicing site and the log2 transformed fold changes for the three RNA-Seq sets. The lines at $y = -0.3$ and $+0.3$ represent the 2-fold change in gene expression threshold.

To summarize this Chapter part, I have determined that modifying SETMAR expression has a widespread but weak effect on the cell transcriptome, which is similar to the effect observed for other chromatin modifiers. Amidst the genes differentially expressed, I observed an enrichment in the genes containing an ITR and a trend towards up-regulation, which is not observed in the genes without an ITR. Interestingly, only the

genes with an ITR expressed at a low or medium level are misregulated upon SETMAR overexpression, which indicates that SETMAR does not activate gene expression and its effect on expression is not strong enough to affect significantly the genes with an ITR expressed at a high level. The mechanism by which SETMAR acts on gene expression is still unclear since the location of the ITR in the gene seems to be non important.

4.2.9 Characterization of the SETMAR binding sites *in vivo*

The RNA-Seq experiments produce a set of candidate genes containing an ITR that are regulated by SETMAR expression. To investigate whether the regulation is direct or indirect, I wished to ask whether the sites are bound by SETMAR by using ChIP-exo. This is a variant of ChIP-seq with an additional exonuclease digestion step to improve the mapping accuracy and reduce the background (Rhee and Pugh, 2011) (Rhee and Pugh, 2012).

I first established a human cell line expressing at a low level of FLAG tagged SETMAR (see Figure 4.1A above). Biological duplicates, named R1 and R2 were prepared and sent to Peconic LLC for ChIP-exo analysis and single-end Illumina sequencing. Out of the 10 millions reads, between 4 to 5 millions reads have been mapped to the reference human genome. GeneTrack was used to called the ChIP-exo peaks (Albert et al., 2008). A peak is defined as a cluster of tags in a fixed interval of 80 bp centred over the peak summit coordinate. R1 and R2 produced 682 and 405 peaks, respectively. A peak is defined as any region of the genome where an enrichment in sequencing reads is observed on both strands. The sequencing

reads, which are mapped against the genome, are 81 bp long. A ChIP-PCR assay on the TOF1 cell line with an antibody against the RNA polymerase II is shown in Figure 4.10.

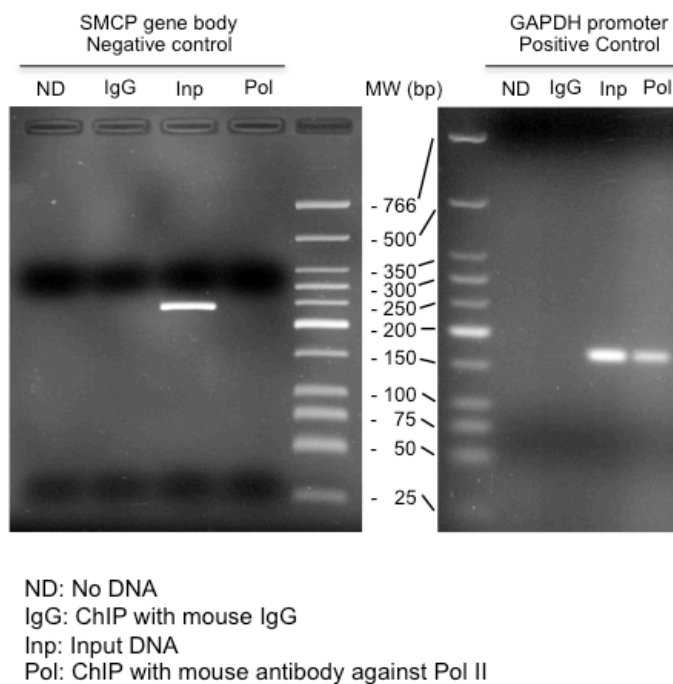


Figure 4-10: Negative and positive control of a ChIP experiment on the control cell line TOF1.

The ChIP experiments were done with a mouse IgG or a mouse antibody against RNA Polymerase II on 30 ug of chromatin extracted from the TOF1 cell line. The ChIP experiments were then tested by PCR on two genomic loci. The SMCP gene is not active in the U2OS cell line and was therefore used as a negative control. For the positive control, primers specific to the GAPDH promoter were used.

To determine the quality of the ChIP-exo, I uploaded on the UCSC genome browser the reads of the first biological replicate (Figure 4.11). Most of the ChIP-exo peaks are found in satellite repeats in centromeric regions (Figure 4.11A) but a minority of peaks are found overlapping intragenic Made1 elements (Figure 4.11B).

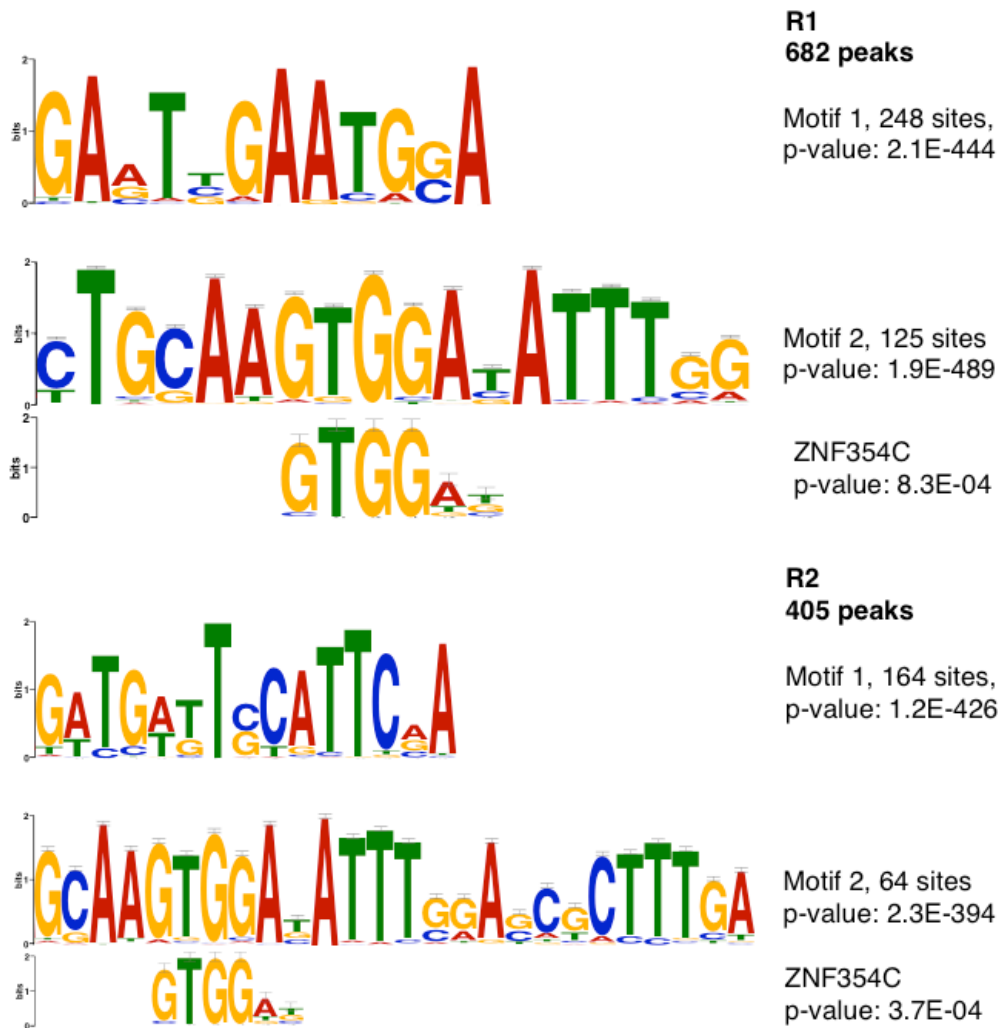


Figure 4-12: Both ChIP-exo replicates contain an enrichment for the ZNF354C binding site.

Motifs and transcription factor binding sites enriched in each replicate ChIP peaks. The analysis was done by using Multiple Em for Motif Elicitation (MEME), which determines enriched motifs in a set of DNA sequences, followed by the Tomtom motif comparison tool, which determines which transcription factor binding sites are present in a DNA sequence. The most significant matches of each replicates are shown.

Only 257 peaks overlap between both replicates, representing 38 and 63% of the R1 and R2 peaks, respectively (Figure 4.10A). The MEME software identified two motifs enriched in the common peaks (Figure 4.10B). However, the two motifs do not correspond to the ITR sequence. Together,

the two motifs represent 39% of the common peaks. Within each motif there is a sub-section, which is related to a transcription factor. Motif 1 and Motif 2 have binding sites for ZNF354C and SOX10, respectively. However, no other TFBSs are found in each motif while SOX10 is known to interact with Pax3 and to co-localize with EGR2 (Lang and Epstein, 2003) (Srinivasan et al., 2012). Furthermore, the common ChIP peaks are not located in human genes but in satellite repeats in the chromosomes centromere. It therefore seems likely that the enrichment for the two TFBSs occurs by chance rather than by an interaction between SETMAR and the transcription factors.

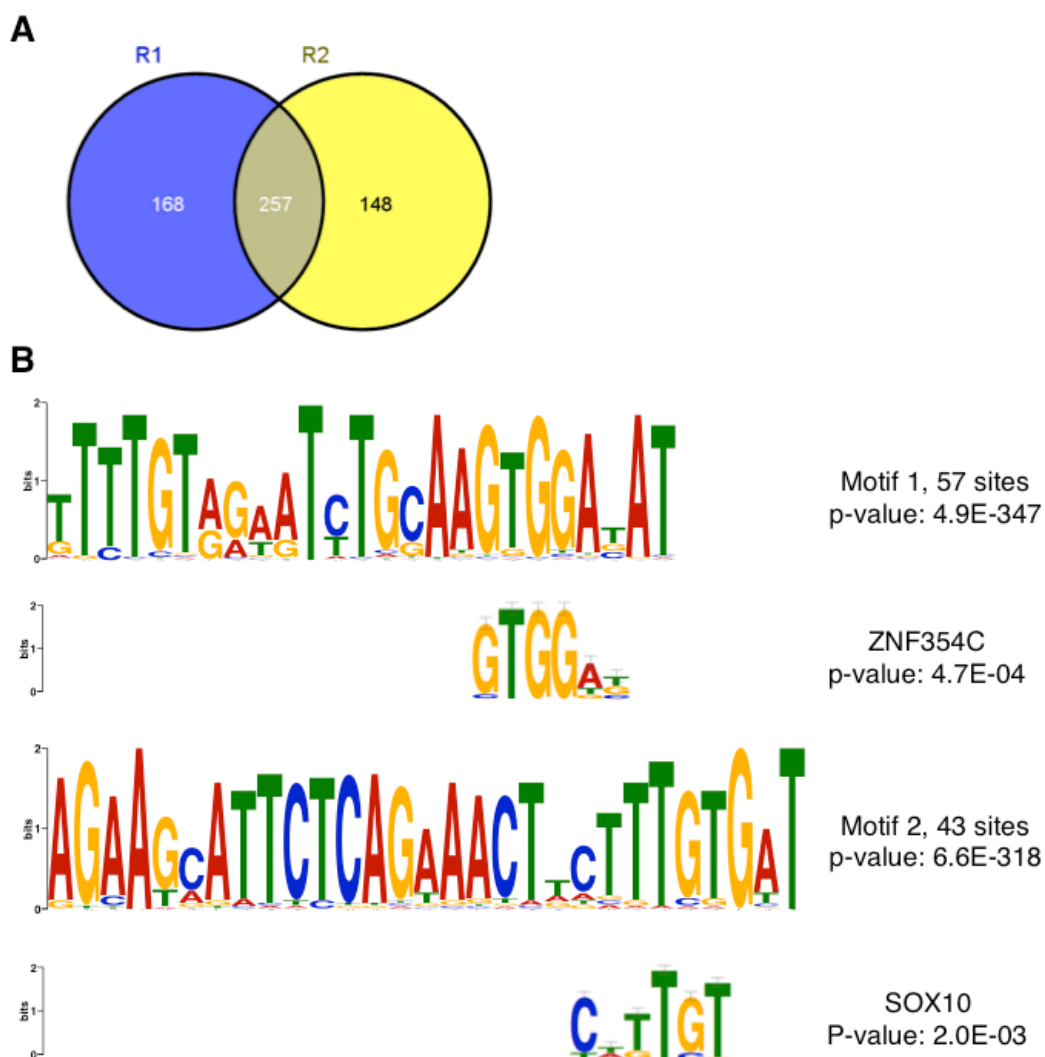


Figure 4-13: The ChIP-exo replicates overlap partially and the common peaks are located in the chromosomes centromere.

(A) Venn diagram representing the overlap of the ChIP peaks between both ChIP-exo replicates. (B) Motifs and transcription factor binding sites enriched in the 257 common peaks. The analysis was done by using Multiple Em for Motif Elicitation (MEME) followed by the Tomtom motif comparison tool. The most significant matches are shown.

Since the common ChIP peaks are not located in human genes and may be experimental artifacts, I parsed the list of ChIP peaks of both replicates and kept only the peaks outside the centromeres. There are now 186 and 99 peaks in R1 and R2, respectively. Analysis of this set also identified two motifs enriched in both replicates (Figure 4.11). For both replicates, the most

enriched motif corresponds to the ITR sequence and represents 8 and 16% of the R1 and R2 peaks respectively. The difference in the ITR length is due to the MEME software which trims the motif to eliminate any positions where there is a gap in any of the occurrences. Thus, the smaller size of the ITR sequence in R2 is due to either the presence of deletions in one or more occurrences of the ITR or to the presence of one or more truncated ITRs due to the ChIP-exo exonuclease step. The second enriched motif differs between the replicates and represents respectively 5 and 14% of R1 and 2. Each also contains a TFBS, respectively ZNF354C and ARID3A. ARID3A, also known as Bright, promotes transcription from IgH promoters by interacting with Btk and the transcription initiation factor-II (Schmidt et al., 2009).

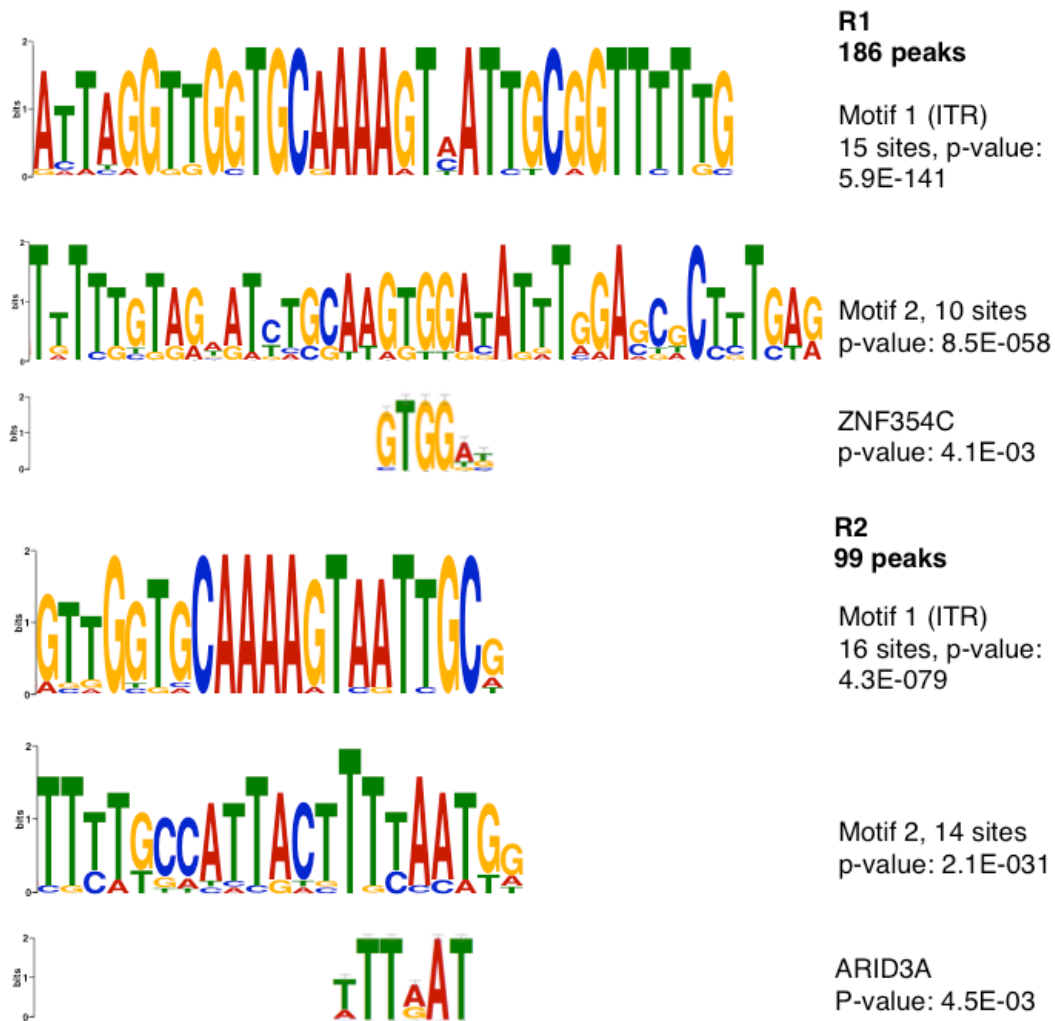


Figure 4-14: The non-centromeric ChIP peaks are enriched for the ITR sequence.

For both replicates, the ChIP peaks outside the centromere regions are enriched for the ITR sequence. The second highest enriched motifs contain putative binding sites for ZNF354C and ARID3A. The analysis was done by using MEME followed by the Tomtom motif comparison tool. The most significant matches of each replicate are shown.

4.2.10 SETMAR core binding sites in the human genome

SETMAR is a fusion between a histone methyltransferase and an Hsmar1 transposase, which binds specifically to the Hsmar1 transposon ends (Cordaux et al., 2006) (Liu et al., 2007). *In vitro* assays with the purified DNA-

binding domain showed that mutations located between bp +8 and +26 of the transposon interfered with binding (Cordaux et al., 2006). A crystal structure for the closely related Mos1 transposase bound to the cognate transposon end provided further details (Richardson et al., 2009). The DNA binding domain contains two HTH motifs, which bind at positions centered about bp +9 and +27 of the transposon end. The most important interactions are thought to be with the most distal sequences, closest to the end of the transposon.

The ChIP-Seq results give me the opportunity to analyze the consensus sequence bound by SETMAR *in vivo*. I retrieved for both replicates the sequences of the ITRs located within or at less than 200 nucleotides from a ChIP peak and aligned them to determine the consensus sequence (Figure 4.12). I selected the 200 nucleotides threshold for retrieving the ITR sequences because SETMAR may be part of a higher order complex. If this is the case, the exonuclease step might have been stopped before it reached the core SETMAR binding site. The consensus sequence logo was a very good match to the Hsma1 transposon end.

Since the consensus binding site is present in several thousands copies in the human genome, I investigated whether the ITR sequence could contain binding sites for transcription factors (Figure 4.12). Indeed, two homeobox transcription factors, Barx1 and Prrx2, may also bind the ITR sequence. Barx1 regulates the canonical Wnt signalling to specify gastric epithelial identity during gut organogenesis (Kim et al., 2005). Prrx2 is involved in pituitary organogenesis and scarless wound healing (Stelnicki et al., 1998) (Higuchi et al., 2014). In the Made1 elements, the transcription factor binding sites

located in each ITR are separated by 36 nucleotides. Therefore, if Barx1 or Prrx2 function as multimer, they may stably bind to the Made1 elements. However, there is no RNA-Seq or ChIP-Seq data available for either protein to determine whether they could regulate the expression of a subset of the genes containing an ITR.

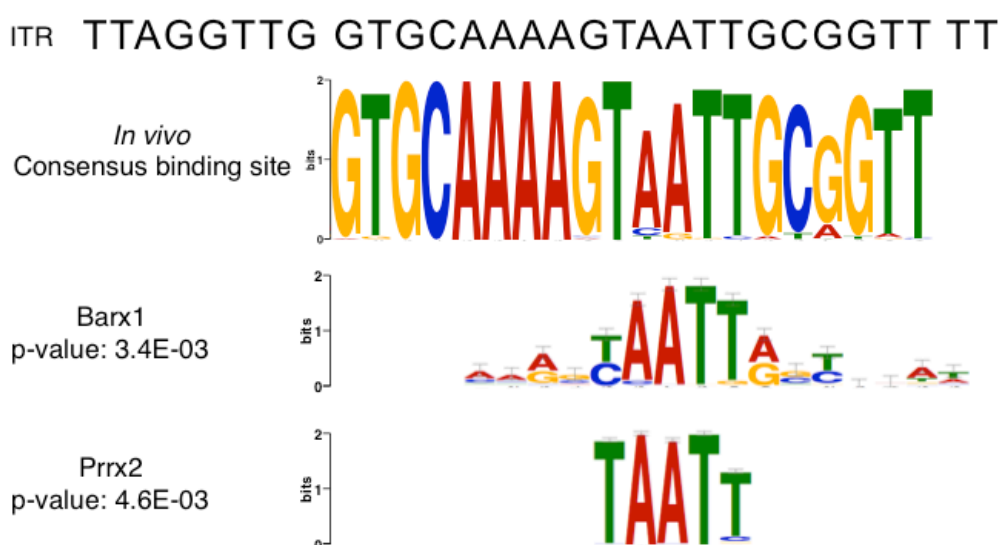


Figure 4-15: Determination of the consensus SETMAR binding site in vivo and putative transcription factor binding sites.

The consensus binding site was determined by aligning all the ITR sequences located within and at less than 200 nucleotides of a ChIP peak. The analysis was done by using MEME followed by the Tomtom motif comparison tool. The most significant matches are shown.

4.2.11 SETMAR binds mostly to Made1 elements rather than solo ITRs

SETMAR binding sites are located in Hsmar1 remnants in the human genome. To determine whether SETMAR binds preferentially to a specific type of remnant, I first investigated the proportion of ChIP-exo peaks located in an ITR (Figure 4.13A). For both replicates, 24% of the peaks are found in

an ITR, which represents 45 peaks in 68 ITRs for R1 and 23 peaks in 28 ITRs for R2. There are 10 common ITRs between R1 and R2. The fact that there are fewer peaks than ITRs is explained by the Made1 elements, which contains two ITRs. Approximately 80% of the ITRs are located in Made1 elements while the remaining 20% are found in solo ITRs (Figure 4.13B). In comparison, the solo ITRs and Made1 elements are evenly distributed in the human genome (Figure 4.13B). Therefore, SETMAR preferentially targets the Made1 elements rather than the solo ITRs ($p < 0.001$ for R1 and < 0.05 for R2).

Several Made1 elements have been co-opted as miRNAs and miRNAs-like (Chapter 3). If expression of these genes is regulated by SETMAR, it may help to explain the large number of DE genes observed in the RNA-Seq data. I therefore investigated whether SETMAR binds specifically to these domesticated Made1 elements (Figure 4.13C). No ChIP peaks are located in one of the defined mir-548 members. However, a significant binding enrichment in miRNAs-like is observed for R1 only (15 hits, $p < 10^{-5}$). For R2, a non-significant 2-fold increase is observed ($p = 0.08$).

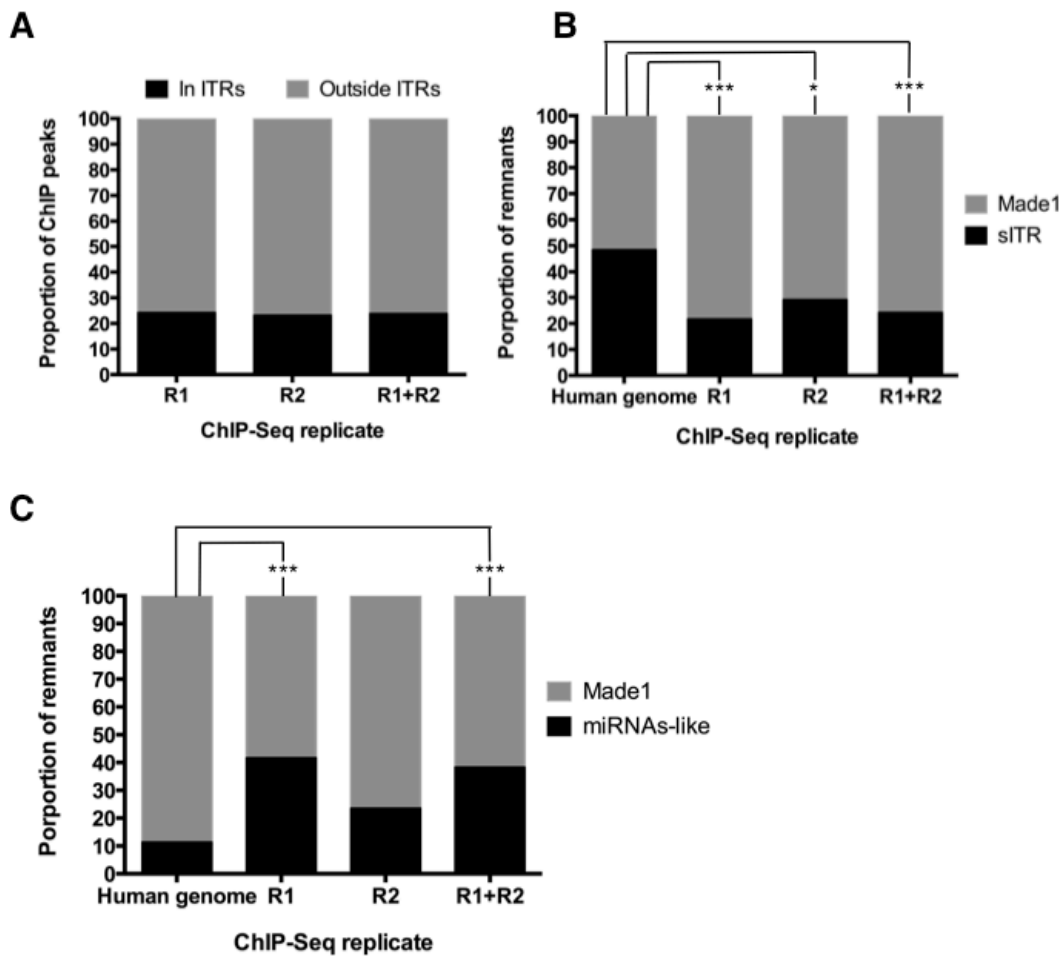


Figure 4-16: SETMAR binds essentially to Made1 elements and miRNAs-like.

(A) Proportion of ChIP peaks located in an ITR for both replicates. (B) SETMAR binds essentially to Made1 elements rather than solo ITRs. (C) Amidst the Made1 elements, SETMAR binding sites are enriched in miRNAs-like. *: $p < 0.05$, ***: $p < 0.001$, hypergeometric test.

4.2.12 The intragenic ChIP peaks are enriched in genes with an ITR

I next established the distribution of the ChIP peaks between genic and inter-genic regions of the genome (Figure 4.14A). Most peaks are genic with only 30% located outside genes. The genic peaks are mainly found in protein-coding genes with 55% and 45% of the total R1 and 2 peaks, respectively.

Since the ChIP peaks are enriched for ITRs (Figure 4.11), I investigated whether there is a reciprocal enrichment of ITRs in genes containing a ChIP peak (Figure 4.14B). Indeed, this is the case. Although only 8% of the protein-coding genes contain an ITR, 40-45% of the ChIP peaks of both replicates are located in a gene with an ITR. This represents 55 and 25 genes for R1 ($p < 10^{-35}$) and R2 ($p < 10^{-37}$), respectively. Approximately 40% of the genes with an ITR and a ChIP peak have their ITR bound by SETMAR. For the remaining 60%, the distance between the ChIP peak and the ITR varies between 7 kb and 5.8 Mb with a median distance of ~ 1 Mb.

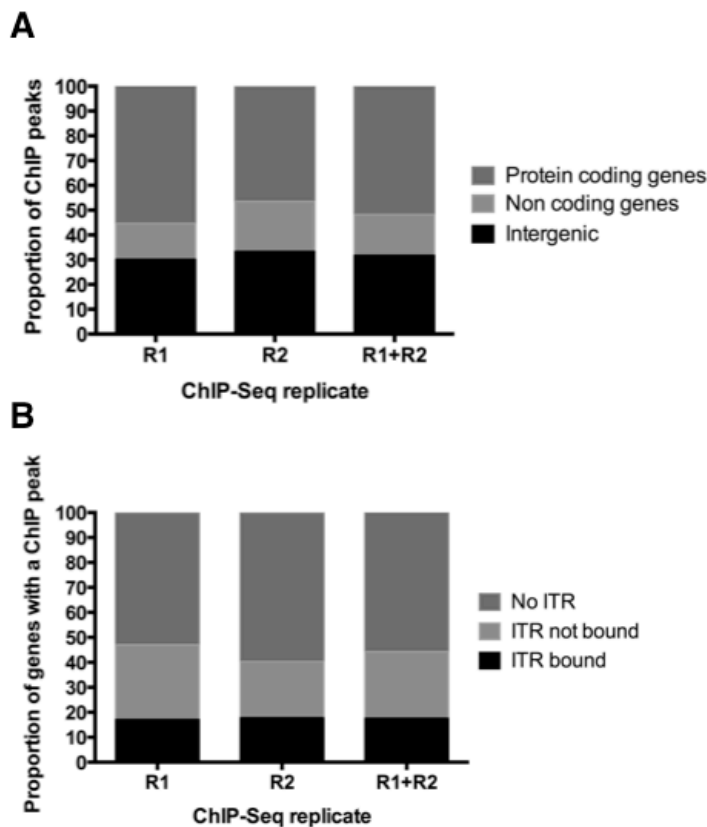


Figure 4-17: The ChIP peaks are enriched in genes with an ITR.

(A) Proportion of intragenic and genic ChIP peaks for both replicates. **(B)** Distribution of intragenic ChIP peaks between genes without an ITR, with an unbound ITR, and with a bound ITR in each replicate.

Since most of the ITRs bound by SETMAR are located in Made1 elements (Figure 4.13B), I examined whether the genes with an ITR and a ChIP peak were enriched in Made1 elements (Figure 4.15A). Indeed, an enrichment for Made1 in the genes with an ITR is observed for R1 and R1+R2 ($p < 0.01$). I next investigated whether the enrichment is also present in the genes with an unbound ITR or with a bound ITR (Figure 4.15B and C, respectively). Interestingly, the Made1 enrichment is only observed for the genes with an ITR bound by SETMAR ($p < 0.001$ for R1+R2).

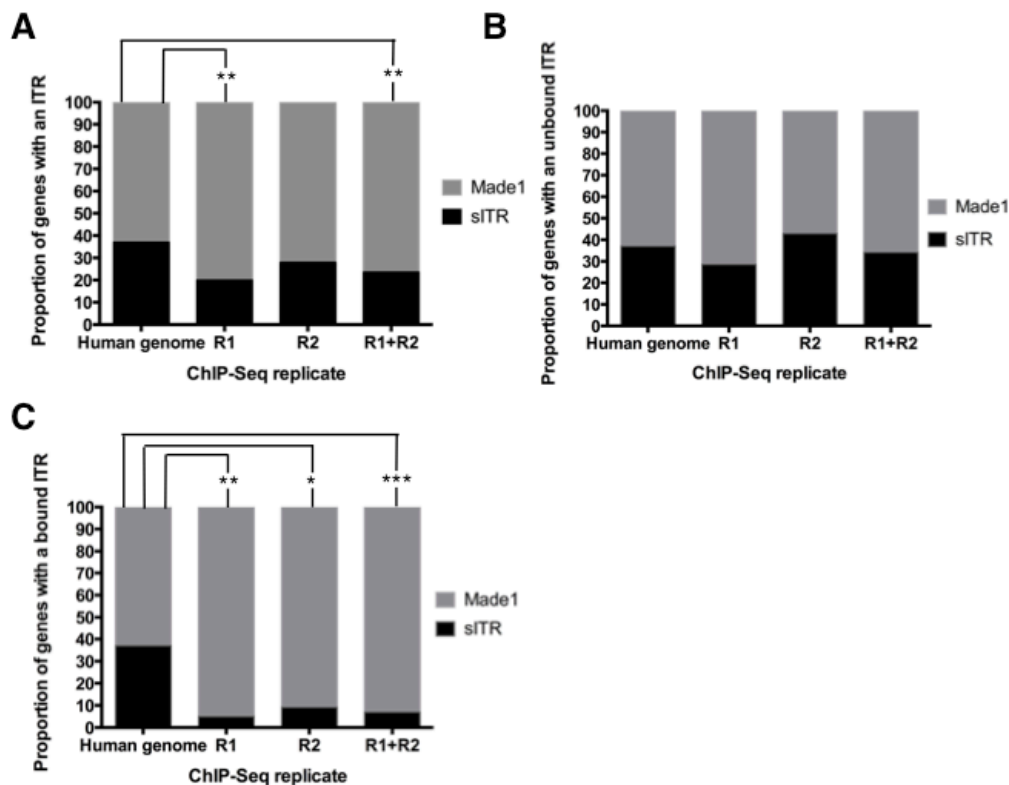


Figure 4-18: Only the genes with an ITR bound by SETMAR are enriched in Made1 elements.

(A) Source of the ITR in the genes with a ChIP peak in each replicate. (B) Source of the ITR in the genes with a ChIP peak but an unbound ITR. (C) Source of the ITR in the genes with a ChIP peak and a bound ITR. *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, hypergeometric test.

4.2.13 ChIP peaks are found in both expressed and silenced genes

To determine whether SETMAR binds specifically to open chromatin, I first investigated the expression level of the genes with a ChIP peak (Figure 4.16A). I estimated the gene expression level of each gene by using its number of reads, the RPKM value, in the control cell line. A gene was considered to not be expressed if there is no read mapping to it (0 RPKM). A gene with a RPKM value between 0 and 1, 1 and 10, 10 and 100, and more than 100 was considered to have a low, medium, high, and very high expression, respectively. In both replicates, more than half of the genes are either lowly expressed or not expressed at all. The remaining genes are essentially expressed at a medium or high level.

I next examined whether the genes with their ITR bound by SETMAR were more expressed than the genes that are not bound (Figure 4.16B). The genes with an ITR are on average less expressed than the genes without an ITR. However, the SETMAR bound or unbound genes have a similar median expression value. A 2-fold increase in average expression is only observed for the half most expressed genes with an ITR bound compared to their unbound counterparts.

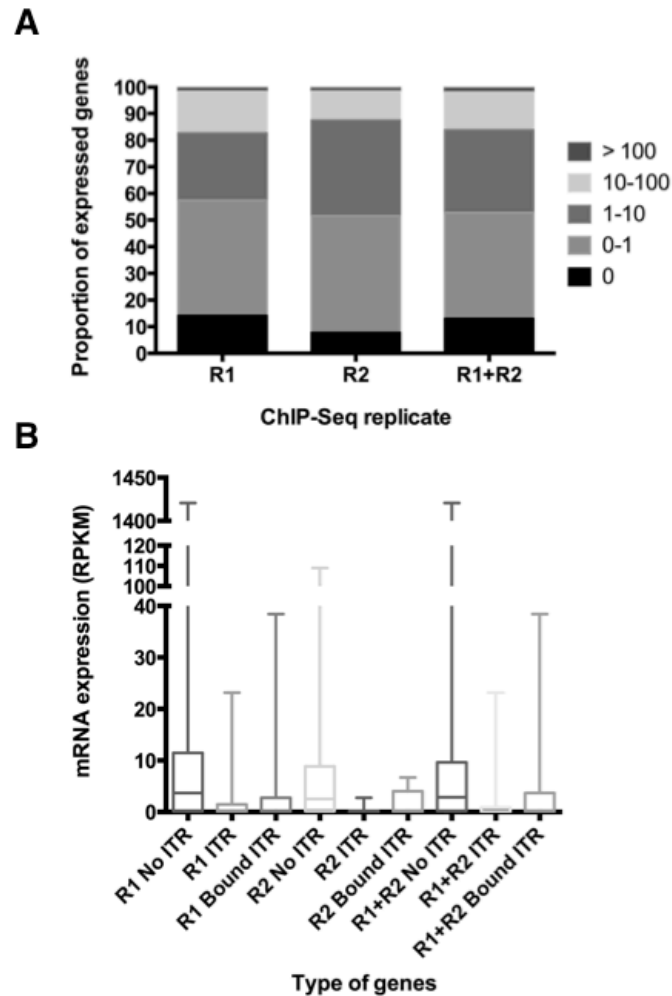


Figure 4-19: SETMAR does not bind specifically to expressed genes.

(A) Distribution of gene expression for the genes containing a ChIP peak. 0: no expression; 0-1: low expression; 1-10: medium expression; 10-100: high expression; > 100: very high expression. **(B)** Box and whisker plot representing the distribution of gene expression for the three categories of genes. No ITR: genes without an ITR; ITR: genes with an unbound ITR; Bound ITR: genes with a bound ITR.

4.2.14 Only a fifth of the genes with a ChIP peak are also differentially expressed

I investigated whether the genes with a ChIP peak overlap with the sets of differentially expressed genes determined from the RNA-Seq data (Figure 4.17). For the genes with a ChIP peak, a ChIP peak and an ITR, and a bound

ITR, only ~ 20% of the genes of each category are also found in each set of differentially expressed genes (Figure 4.17, left, middle, and right column, respectively).

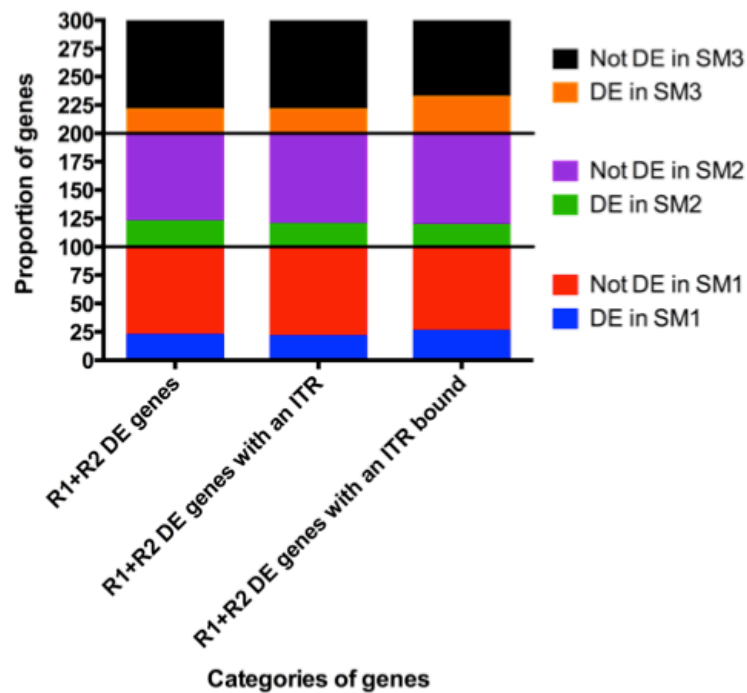


Figure 4-20: The genes with a ChIP peak overlap only partially with the three sets of differentially expressed genes.

Proportion of genes with a ChIP peak also differentially expressed (DE) in each cell line for the following categories: proportion of genes with a ChIP peak also DE (R1+R2 DE genes), proportion of genes with an ITR and a ChIP-peak also DE (R1+R2 DE genes with an ITR), and proportion of genes with a bound ITR also DE (R1+R2 DE genes with an ITR bound).

Approximately 1% of the DE genes contain a ChIP peak with more than half being up-regulated. For the DE genes with an ITR, ~ 2% have their ITR bound by SETMAR in ChIP-Seq with half of them being up-regulated. FAM83B, which encodes an oncogene activating the MAPK and mTOR signalling pathways, is the most down-regulated gene bound by SETMAR

with a 3-fold decrease in expression (Cipriano et al., 2012) (Cipriano et al., 2013). ARSG, which encodes for a sulphatase involved in neuronal homeostasis, is the most up-regulated gene with a 3.5-fold increase in expression (Abitbol et al., 2010).

4.3 Discussion

4.3.1 Overlapping of the three sets of differentially expressed genes

A quarter of the differentially expressed genes are unique to each of the three overexpressing cell lines (Figure 4.2A). This increases to 41% for the genes with at least a 2-fold change in expression level (Figure 4.2A). However, this increase is explained by the difference in fold change for the genes present in two or three cell lines. Thus, a gene with a fold change inferior to 2 in one cell line but superior to 2 in another will be counted as unique to the second cell line. Therefore, the fold changes of the unique genes are evenly distributed below and above the 2-fold change threshold.

Between 20 and 30% of the differentially expressed genes with an ITR are unique to each cell line (Figure 4.2B). The patchy distribution of genes with an ITR between the three sets is explained by the presence of 4,132 binding sites in 2,697 genes. Thus, a difference in the regulated genes is expected because of the probabilistic nature of site-specific binding. The genes with an ITR found in two or three cell lines may correspond to the core intragenic binding sites while the unique genes contain suboptimal binding sites.

The variability in the differentially expressed genes without an ITR could be due to the unique genes with an ITR. If some of the unique genes encode for regulatory proteins, a modification of their expression will affect the regulation of their target genes.

4.3.2 Increasing SETMAR expression skews differentially expressed genes towards up-regulation

The inverse relationship between SETMAR expression level and the proportion of down-regulated genes suggests a de-repressing role for SETMAR (Figure 4.4C). The effect is even more significant for the genes with an ITR in the SM3 cell line where an increase in the proportion of up-regulated genes is also observed (Figure 4.4B).

The H3K36me2 mark is widespread in the genome with 30 to 50% of the histone H3 containing this mark (Rose and Klose, 2014). In *Saccharomyces cerevisiae*, H3K36me2 repress spurious transcription initiation (Li et al., 2009a). However, in human its role is less defined and seems to be involved in both transcriptional initiation and repression (Wagner and Carpenter, 2012). The response to this epigenetic mark is dependent of where and when it is located in the gene body and on which reader proteins bind to the modification (Wagner and Carpenter, 2012). For the genes with an ITR, SETMAR dimethylation of K36 at the histone H3 may be associated to a promotion of gene expression because more than half of these genes in the SM1 and 2 cell lines are up-regulated (Figure 4.4B). For the SM3 cell line, the proportion of up-regulated genes increases to 70% (Figure 4.4B).

For the genes without an ITR, the bias towards up-regulation may be explained by an indirect targeting of SETMAR to these genes by interacting partners. A more likely hypothesis may be a regulation of the genes without an ITR by the differentially expressed genes with an ITR. The shift towards up-regulation of the genes with an ITR could promote a similar modification of their target genes. If SETMAR directly up-regulates another TF, it is most likely to up-regulate other genes because most TFs in eukaryotic are positive regulators (Hurst et al., 2014).

4.3.3 Depth of sequencing coverage for RNA-Seq and ChIP-Seq

For the RNA-Seq experiments, at least 30 millions mappable reads were generated for each sample, in agreement with the ENCODE standards for the experiments evaluating the difference between transcriptional profiles of poly(A)+ samples.

For the ChIP-exo experiment, the number of mappable reads is comprised between 4 and 5 millions for each sample, which is equal to a sequencing depth coverage of 12.5%. This is four times less than the 20 millions mappable reads recommended by the ENCODE guidelines for ChIP-Seq on point-source transcription factors, which corresponds to proteins binding specific DNA sequences such as SETMAR (Landt et al., 2012). However, the ChIP-exo requires less depth of sequencing coverage than a standard ChIP-Seq because of the reduced background provided by the exonuclease step (Rhee and Pugh, 2011) (Serandour et al., 2013).

The low overlap between both the peaks of both ChIP-exo replicates is explained by a less successful immunoprecipitation of SETMAR in the second replicate. The lower quality of the ChIP of the second replicate led to a two-fold decrease in the final number of peaks.

4.3.4 Regulation of the expression of genes containing an ITR by SETMAR

The enrichment of ITRs in the set of genes with a DE of at least 2-fold in the RNA-Seq experiments (section 4.2.3) suggests that SETMAR binding may regulate these genes directly. These genes are essentially up-regulated and represent 55%, 59%, and 85% in the SM1, 2, and 3 cell lines, respectively (Table 4.6). In contrast to the ChIP-Seq data, where the non-centromeric SETMAR binding sites and the bound genic ITRs were mostly Made1 elements, there was no enrichment of these elements in the set of DE genes encoding an ITR (Figure 4.6A). Since Made1 elements encode two ITRs, which have the potential for a higher order complex (see below, (Claeys Bouuaert et al., 2011)) it seems possible that ChIP-Seq is only identifying the strongest binding sites or those that are stabilized by interacting partners. Indeed, site-specific binding by DNA binding proteins is generally about 10^5 fold higher than their affinity for non-specific DNA. In a genome of $\sim 10^9$ bp with a single binding site the protein would spend only 1/10000 of its time bound to its specific site.

Alternatively, regulation by SETMAR may depend on additional factors working together with SETMAR. This is supported by the overrepresentation

of genes with an ITR within the genes with a ChIP peak even though less than half of them have their ITR directly bound by SETMAR (Figure 4.14B).

Another non-exclusive hypothesis is SETMAR may dimethylate the K36 of histone H3 near the ITR, which may in turn recruit other proteins that could regulate gene expression, depending on the surrounding epigenetic landscape.

4.3.5 SETMAR and Made1 elements

The preferential binding of SETMAR to Made1 elements can have several non-exclusive explanations. Made1 elements possess either two functional binding sites or a combination of functional and mutated sites. The presence of mutations may decrease the affinity of SETMAR for the ITR but it could still stabilize the SETMAR dimer on the Made1 element.

Hsmar1 ITRs and Made1 elements have a GC-content of 35%. Both types have several poly(dA:dT) sequences which are known to strongly inhibit nucleosomes formation (Struhl and Segal, 2013). The solo ITRs are too small to affect nucleosomes positioning but the Made1 elements are long enough to hinder the nucleosomes formation. Consequently, the presence of Made1 elements in nucleosomes-free region could favour their binding by SETMAR compared to the solo ITRs.

4.3.6 Do the miRNAs-like regulate the differentially expressed genes in the RNA-Seq experiments?

All the differentially expressed genes are expressed in the parental cell line. In contrast, the genes bound by SETMAR are a combination of silenced and expressed genes. This disparity suggests a possible post-transcriptional regulation of the differentially expressed genes by small RNAs. This could be mediated in many ways. For example, if SETMAR alters the expression of another master regulator it could initiate a cascade of changes. However, some of the Made1 elements, which are 80 bp palindromic sequences, have been recruited into miRNAs and miRNAs-like (Chapter 3) (Piriyapongsa and Jordan, 2007). These are good candidates for being regulated by SETMAR.

Out of the 456 miRNAs-like, 18 appear in the ChIP-Seq data set (Figure 4.13C). However, a further 47 are also bound by SETMAR but I cannot determine yet if they correspond to a signal or to a background noise. The low number of miRNAs-like bound might correspond to the members with the strongest binding sites. There is less than 500 SETMAR molecules per cell in the U2OS cell line and SETMAR has several interacting partners not implicated in gene expression (Introduction section 1.3.4) (Beck et al., 2011b). Thus, the endogenous SETMAR should bind to only a subset of genic ITRs in the cell.

Since the current RNA-Seq experiments were performed on poly(A)+ RNAs, the expression of the miRNAs-like in the U2OS cell line is unknown. An increased binding is also observed in the second replicate but it is not statistically significant (Figure 4.13C). Oppositely, SETMAR does not bind to

any of the 69 members of the mir-548 family. This is explained by the degeneration of the transposon binding sites in the miRNAs (Chapter 3).

If these miRNAs-like are functional and their expressions regulated by SETMAR, it could explain the vast number of genes differentially expressed in each cell line because each miRNA can regulate the expression of up to a hundred targets (Bartel, 2009). Because SETMAR expression skews the differentially expressed genes towards up-regulation, the miRNAs-like might be down-regulated if they inhibit gene expression. However, they could also down-regulate only a subset of the differentially expressed genes.

The miRNAs post-transcriptional regulation should affect the expression of genes with and without an ITR. Thus, it cannot explain the enrichment observed in differentially expressed genes containing an ITR. If there is a specific regulation of the genes with an ITR by the miRNAs-like, it should not be by a targeting of their ITR sequence by the putative miRNAs. Indeed, the ITRs are located in introns whereas miRNAs target only mature mRNAs (Lee et al., 2002). I cannot exclude a different mechanism for the miRNAs-like, which could result in a direct regulation of the pre-mRNAs containing an intronic ITR. However, introns are co-transcriptionally spliced in the nucleus within 5 to 10 minutes after synthesis questioning the feasibility of a small RNAs based regulation at this step (Singh and Padgett, 2009).

4.3.7 ChIP peaks enrichment in the satellite regions

The majority of the ChIP-exo peaks are located in satellite DNA located in the chromosomes pericentromeric regions (section 4.2.8). I did not consider these

peaks for the analysis because they did not resemble ITRs, which are the only known site-specific binding determinants for SETMAR. I cannot therefore determine whether they represent a true signal or an artefact of the ChIP technique.

Some types of human satellite DNA are known to maintain open chromatin structure, which could result in a misleading localization of SETMAR to these regions (Kim et al., 2009). Indeed, it has been recently established in *S. cerevisiae* that 238 euchromatic loci exhibit enrichment for several unrelated proteins (Teytelman et al., 2013). Furthermore, the ChIP-Seq experiments performed by the ENCODE consortium established a blacklist of regions in the human genome corresponding to signal artifacts (Bernstein et al., 2012). Half of the 226 regions are located in either centromeric repeat or satellite repeat. The regions defined by the ENCODE project do not overlap with the ones found for SETMAR but this discrepancy may be explained by the technical differences between ChIP-exo and ChIP-Seq.

**5. SETMAR FUNCTIONS IN NON-
HOMOLOGOUS END JOINING REPAIR
AND ILLEGITIMATE DNA
RECOMBINATION**

5.1 Introduction

This chapter focuses on gaining a better understanding of the functions of SETMAR and its domains in DNA repair and integration. I used two different plasmid-based assays to test several human cell lines overexpressing SETMAR or its constituent domains. SETMAR overexpression did not promote DNA repair or integration and its endonuclease activity was also not essential *in vivo*, in agreement with recent published papers (Mohapatra et al., 2013) (Rath et al., 2014). Surprisingly, the overexpression of the SET domain decreases DNA repair efficiency but increases DNA integration whereas the MAR domain promotes DNA repair and at a lesser level DNA integration. Since the SET domain should work as a monomer, I speculate that the promotion of plasmid integration by the MAR domain is due in fact by the increase in SETMAR-MAR dimer containing only one SET domain. Further work is required for understanding the underlying mechanisms of the SET and MAR domains in the NHEJ pathway.

5.2 Results

I constructed a series of T-Rex-U2OS cell lines stably expressing the SET or the MAR domains alone, full length SETMAR, SETMAR R432A, and SETMAR D483A. The two mutants encoding respectively a protein with a lower affinity for Hsma1 transposon ends and a catalytically defective protein unable to trim DNA overhangs (Roman et al., 2007) (Mohapatra et al., 2013).

The control cell line has been stably transfected with an empty expression vector.

SETMAR probably functions as a dimer in the cell with the transposase domain providing the whole of the dimer interface (Goodwin et al., 2010). The endogenous concentration of SETMAR in the U2OS cell line is low (less than 500 molecules per cell (Beck et al., 2011b)) so the overexpression of SETMAR mutants should produce essentially dimers of two mutant monomers and a small proportion of dimers containing wild type and mutant monomers. Dimers with two wild type monomers are the less likely so the activity of SETMAR at the cell level should be hindered by the overexpression of the mutants.

The R432A mutation decreases the affinity of SETMAR for the transposon ends (Roman et al., 2007) (Claeys Bouuaert et al., 2013). Therefore, this mutant will allow me to investigate the relative contribution of ITR binding in DNA repair and integration. Presumably, the decrease in ITR-bound SETMAR will increase the concentration of free and non-specifically bound SETMAR in the nucleus. This shift in the equilibrium will help to determine whether the fraction of wild type ITR-bound SETMAR can be recruited for DNA repair and integration. Indeed, any difference observed between the overexpression of SETMAR and the R432A mutant will be due to the change in the equilibrium of bound/unbound SETMAR.

The D483A mutant is catalytically defective because of the mutation of the first D of the DDD triad, which is necessary for the incorporation of one the Mg^{2+} ion (Roman et al., 2007) (Mohapatra et al., 2013). SETMAR catalytic domain can trim damaged and undamaged DNA overhangs to form DNA

ends that can be processed by DNA ligase IV (Hromas et al., 2008). However, SETMAR endonuclease activity has only been established *in vitro* and recent papers question its relevance *in vivo* (Mohapatra et al., 2013) (Rath et al., 2014). Therefore, this mutant will allow me to investigate whether SETMAR endonuclease activity is required in DNA repair and integration *in vivo*.

Both assays used in this Chapter rely on plasmid transfection. The number of plasmids received per cell is variable. The cells with the most plasmids are more likely to have a successful DNA repair or integration event than the cells with few plasmids. However, this bias should be found in each cell line, which makes the comparison of the results valid.

5.2.1 SETMAR overexpression does not promote cell proliferation in the U2OS genetic background.

It was previously observed that SETMAR overexpression increases the growth rate of the HEK293 and HEK293T cell lines (De Haro et al., 2010). To determine whether a similar effect is observed on another cell line, and whether this effect is connected to an unregulated activity of the histone methyltransferase domain, I used stable T-Rex-U2OS cell lines overexpressing flag-tagged version of SETMAR or of the SET domain. The expression level of the SET domain or SETMAR was determined by western blotting using an anti-FLAG antibody to allow the comparison between the cell lines (Figure 5.1A). The SETF1, SMF2, and SMF3 cell lines overexpress at a medium level SET or SETMAR. I have selected SMF2 and 3 to account for any positional effect of the random genomic integration of the expression

vector. SMF1 overexpresses SETMAR at a lower level and has been chosen to determine whether the effect on cell proliferation is concentration-dependent.

The growth rate was determined by counting the number of cells everyday for a period of 8 days (Figure 5.1B). A slight but significant decrease in cell proliferation is observed for the cell lines with the stronger overexpression. SETF1, SMF2, and SMF3 grew 1.3 to 1.5 times slower than the control cell line between the days 6 to 8. The SMF3 cell line starts to grow significantly slower from day 5. The SMF1 cell line has a similar growth rate than the control cell line with a slight decrease on the last day (Figure 5.1B).

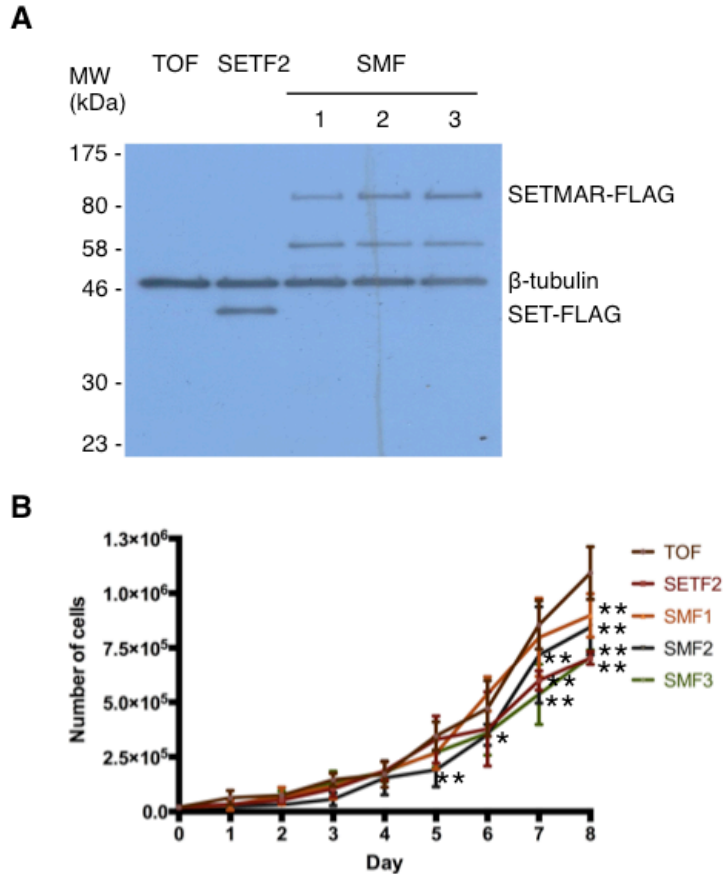


Figure 5-1: The overexpression of SET or SETMAR does not promote cell proliferation in an U2OS genetic background.

(A) SET and SETMAR overexpression level in T-Rex-U2OS cell lines. The western blot was done with anti-FLAG and anti- β tubulin antibodies. **(B)** Growth rate of T-Rex-U2OS cell lines overexpressing SET or SETMAR. At day 0, 2.0×10^4 cells were seeded in eight dishes for each cell line and one dish was count every day for eight days. Average $\pm$ standard deviation of 3 to 5 biological replicates. Statistical test: 2way ANOVA, * p-value < 0.05, ** p-value < 0.01.

5.2.2 The SET and MAR domains have opposite influence on DNA repair

To improve our understanding of SETMAR functions in the NHEJ pathway, I used a previously described *in vivo* DNA repair assay (Fattah et al., 2010). This assay is based on a set of two plasmids, one encoding the

reporter gene (pEGFP) and the other serving as a transfection control (pRFP). The reporter plasmid encodes an EGFP gene interrupted by a 2.4 kb intron derived from the rat *Pem1* gene. An exon from the adenovirus (Ad2) has been integrated in the intron abolishing the GFP activity (Figure 5.2A). The Ad2 exon is flanked by HindIII and I-SceI restriction sites. Cleavage with HindIII or I-SceI yields compatible or incompatible ends, respectively (Figure 5.2B). These two types of ends require different steps for repair. Compatible ends can be directly ligated while incompatible ends need to be trimmed before the ligation step can occur. The repair of the linearized plasmid by the NHEJ pathway restores the GFP ORF making the cell green (Figure 5.2C). The repair events were detected by flow cytometry measuring at least 10,000 cells per assay. The repair efficiency was calculated as the ratio of green and red cells over the total number of red cells, thus normalizing the transfection efficiency between cell lines.

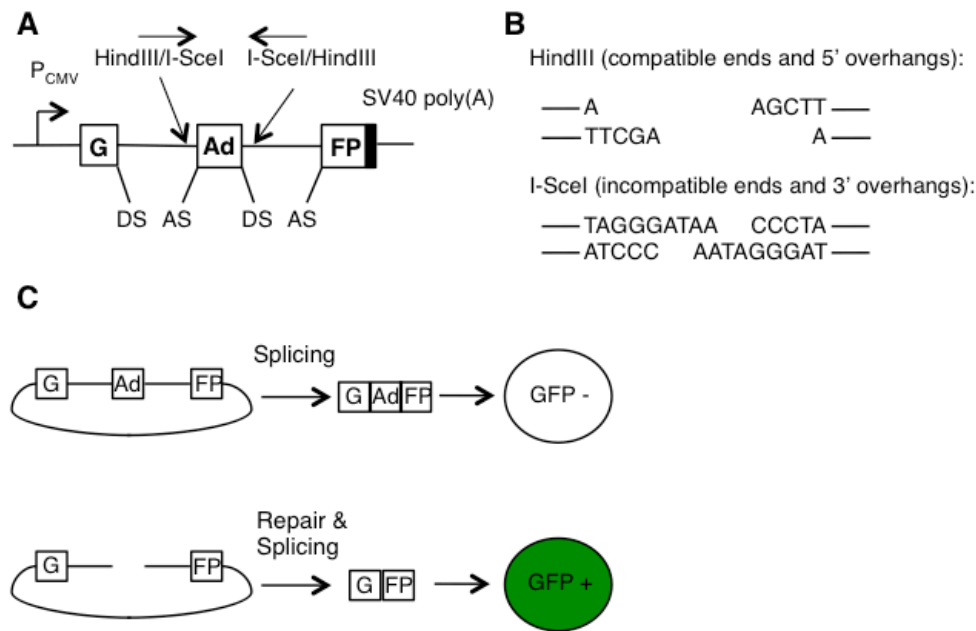


Figure 5-2: Reporter substrate for the in vivo analysis of NHEJ.

(A) The reporter construct (pEGFP-Pem1-Ad2) is composed of a GFP cassette flanked by a P_{CMV} promoter and a SV40 poly(A) sequence. The GFP coding sequence is interrupted by a 2.4 kb intron containing an adenovirus exon (Ad). The Ad exon is flanked by HindIII and I-SceI restriction sites. The donor (DS) and acceptor (AS) splicing sites are shown. **(B)** HindIII and I-SceI restriction sites are respectively composed of a palindromic 6-bp and a nonpalindromic 18-bp sequence. Digestion of the reporter construct by HindIII or I-SceI generates respectively compatible and incompatible ends. **(C)** The presence of the Ad exon in the GFP ORF inactivates the GFP activity thus making the cell GFP negative. Removal of the Ad exon by HindIII or I-SceI followed by a successful intracellular repair will restore the GFP expression that can be quantified by flow cytometry. Adapted from (Fattah et al., 2010).

To analyze the role of SETMAR and its domains in DNA repair, I constructed stable cell lines overexpressing different SETMAR variants (Figure 5.3A, Table 5.1). The expression levels of each cell line were determined by western blotting using anti-SETMAR and anti-FLAG antibodies (Figure 5.3B). An anti-FLAG antibody was used for the cell lines containing an F (for FLAG-tag) in their names. SM2 and 3 overexpress a version of

SETMAR without any FLAG-tag so an antibody against the last nine amino acids of SETMAR was used to determine their expression level. The antibody also recognizes SMF2 and 3 therefore the absence of bands with this antibody indicates that SM2 and 3 overexpress SETMAR at a higher level than SMF2 and 3.

Full name	Abbreviation	Expression level
T-Rex-U2OS-TO	TO	Null
T-Rex-U2OS-SETMAR	SM2	High
	SM3	Very high
T-Rex-U2OS-TO-FLAG	TOF	Null
T-Rex-U2OS-SET-FLAG	SETF1	Low
	SETF2	Medium
T-Rex-U2OS-MAR-FLAG	MARF	Medium
T-Rex-U2OS-SETMAR-FLAG	SMF2	Medium
	SMF3	Medium
T-Rex-U2OS-SETMAR R432A-FLAG	SMF R432A	Medium
T-Rex-U2OS-SETMAR D483A-FLAG	SMF D483A	Medium

Table 5-1: T-Rex-U2OS cell lines used in the DNA repair assay.

The expression level of the stably integrated construct in each cell line has been determined by western blotting (see main text and Figure 5.3A).

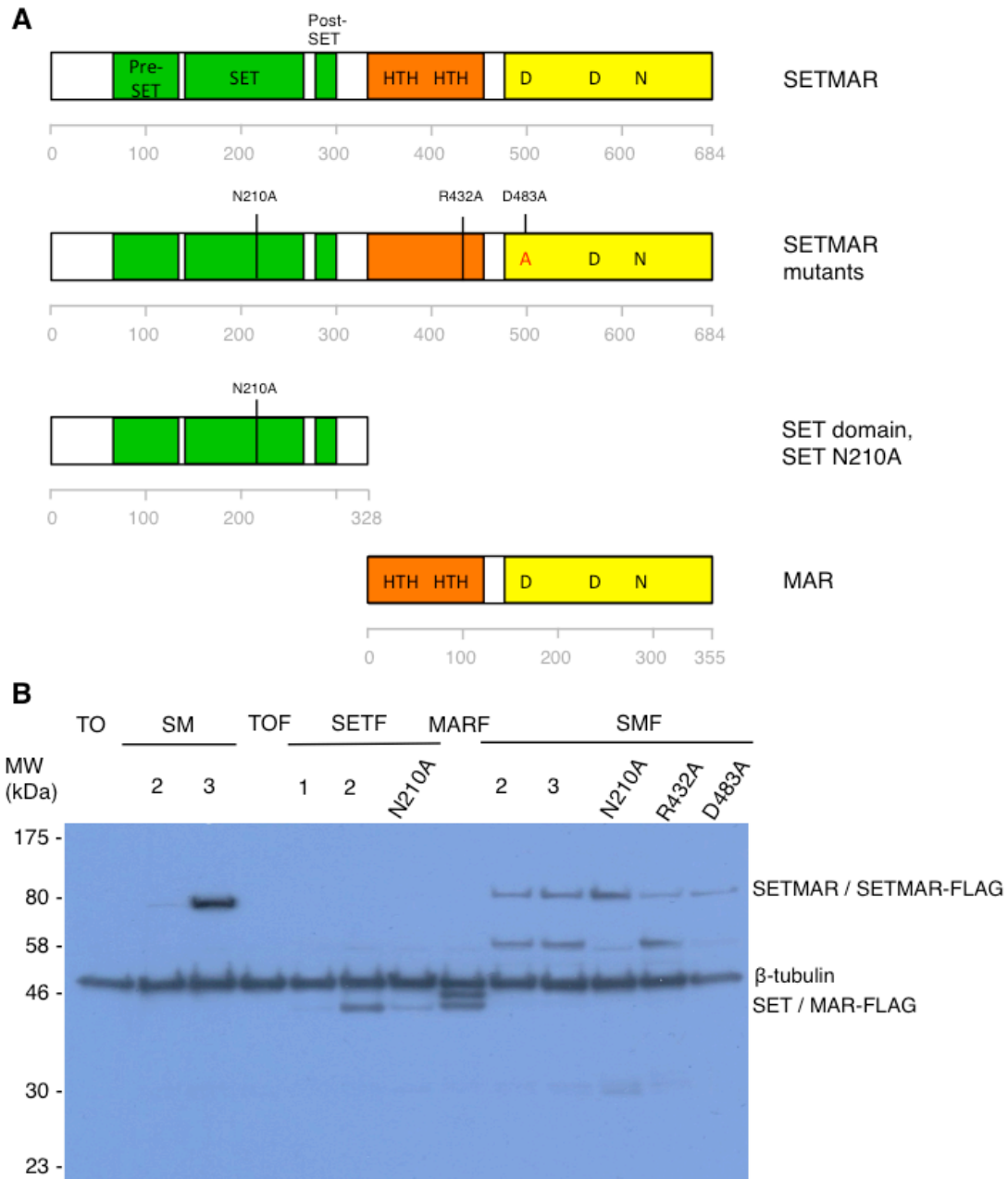


Figure 5-3: SETMAR constructs used in the NHEJ repair assay.

(A) Schematic representation of the SETMAR constructs used in this Chapter. **(B)** Expression of SET (SETF1&2, SETF N210A), MAR (MARF), SETMAR (SM2&3 and SMF2&3), and different SETMAR mutants (SMF N210A, R432A, and D483A) in T-Rex-U2OS cell lines. The western blot was probed with anti-FLAG and anti- β tubulin antibodies.

For normalizing between the biological replicates, I calculated for each replicate the ratio of the repair efficiency of each cell lines relative to their

respective control cell line. A FACS profile of each cell line is presented in Figure 5.4A. The average ratio from three independent experiments is presented in Figure 5.4B. The SM2 and 3 cell lines were compared to the TO control cell line whereas the remaining cell lines were compared to TOF to take into account any possible effect of the FLAG tag. Unsurprisingly, there was no significant difference in the repair efficiency of both types of ends between the two control cell lines (Figure 5.4A). In all cell lines, except SETF1 and 2, a ~10% decrease of the repair efficiency of incompatible ends compared to compatible ends was observed because of the necessary trimming of the DNA overhangs before the plasmid ends could be ligated (Figure 5.4A), in agreement with the previously published results for this assay (Fattah et al., 2010).

The overexpression of SETMAR, irrespective of the level of expression, does not promote the repair of either compatible or incompatible ends confirming the recent results obtained in the cell-extract assays (Figure 5.4B) (Mohapatra et al., 2013). A slight but non-significant decrease in the repair of compatible ends is observed when the SET domain is expressed at a low level whereas a medium overexpression significantly decreases by ~20% the repair efficiency of both compatible and incompatible ends, suggesting at a concentration-dependent effect (Figure 5.4B). In contrast, medium overexpression of the MAR domain increases the repair efficiency of both types of ends by ~20% compared to the control cell line (Figure 5.4B). The overexpression of SETMAR D483A increases by ~15% the repair efficiency of both types of ends but only the compatible ends change is statistically significant (Figure 5.4B). The results confirm that SETMAR endonuclease

activity is not required for DNA repair *in vivo* (Mohapatra et al., 2013). No difference is observed in the DNA repair efficiency of both types of ends by the SETMAR R432A mutant confirming that the ITR binding activity does not affect DNA repair (Figure 5.4B).

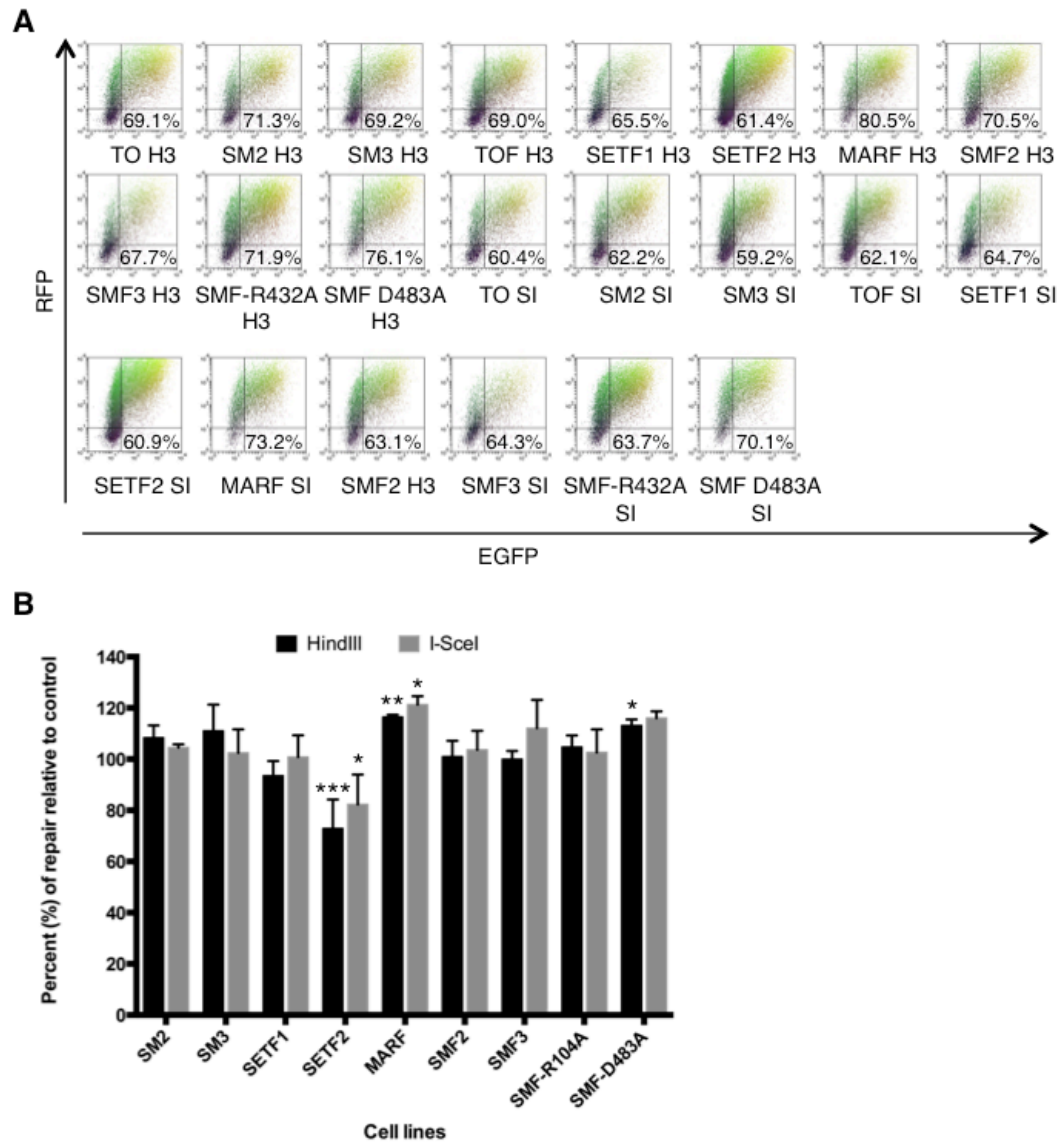


Figure 5-4: The SET and the MAR domains of SETMAR have an antagonist activity on DNA repair.

(A) FACS profiles of a pRFP plasmid used together with the pEGFP-Pem1-Ad2 reporter substrate are shown for each cell lines. The profiles were generated using HindIII- (H3) and I-SceI- (SI) linearized plasmids. The proportion of repaired substrate is indicated in the lower right-hand corner of each profile. The percent is calculated from the number of cells that were doubly EGFP (horizontal) and RFP (vertical) positive versus the number of RFP positive.

(B) Repair efficiency of the different cell lines relative to the control cell line (TO or TOF) of a linearized plasmid with compatible (HindIII) or incompatible (I-SceI) ends. Average $\pm$ standard deviation of 3 biological replicates. Statistical test: 1way ANOVA, * p-value < 0.05, ** p-value < 0.01, *** p-value < 0.001.

5.2.3 The SET and MAR domains but not SETMAR promote DNA integration

The observation that the overexpression of the SET and MAR domains but not SETMAR affect the DNA repair efficiency *in vivo* led me to test if each SETMAR domain and the full-length SETMAR were involved in illegitimate DNA integration.

For a preliminary integration assay, I used stable cell lines expressing the SET domain or SETMAR and determined the relative protein level by western blotting with an anti-FLAG antibody (Figure 5.5A). The SETF1 and 2 cell lines express at a low and medium level the SET domain whereas SMF1 and 2 express in the same manner SETMAR. The band for the SMF1 is extremely faint. However, the blot in Figure 5.1A is more sensitive and demonstrates the presence of the protein.

The plasmid integration efficiency was determined by transfecting in each cell line a circular plasmid encoding a neomycin gene resistance and by counting the number of surviving foci after a fourteen days selection with G418 (Figure 5.5B). A low overexpression of the SET domain does not change the number of plasmid integrated in the genome but a significant 3-fold increase is observed with a medium overexpression, confirming a role for the SET domain in NHEJ (Figure 5.5B). Similar to the DNA repair results, a low and medium overexpression of SETMAR did not promote DNA integration (Figure 5.5B).

Before testing the MAR domain and the different SETMAR mutants, I determined if the topology of the plasmid influences its integration frequency.

To compare the integration efficiency, I transfected a circular or linearized version of the same plasmid in the control cell line (Figure 5.5C). A ~ 3-fold decrease is observed for the linear form compared to its circular counterpart.

A representative selection of the integration plates of each cell line and their respective integration frequency are presented in Figure 5.5D. The results agree with published work showing that the efficiency of illegitimate recombination in most cell lines is less than 1% (Wurtele et al., 2003). An increase of the efficiency to 2% is only observed with the medium overexpression of the SET domain.

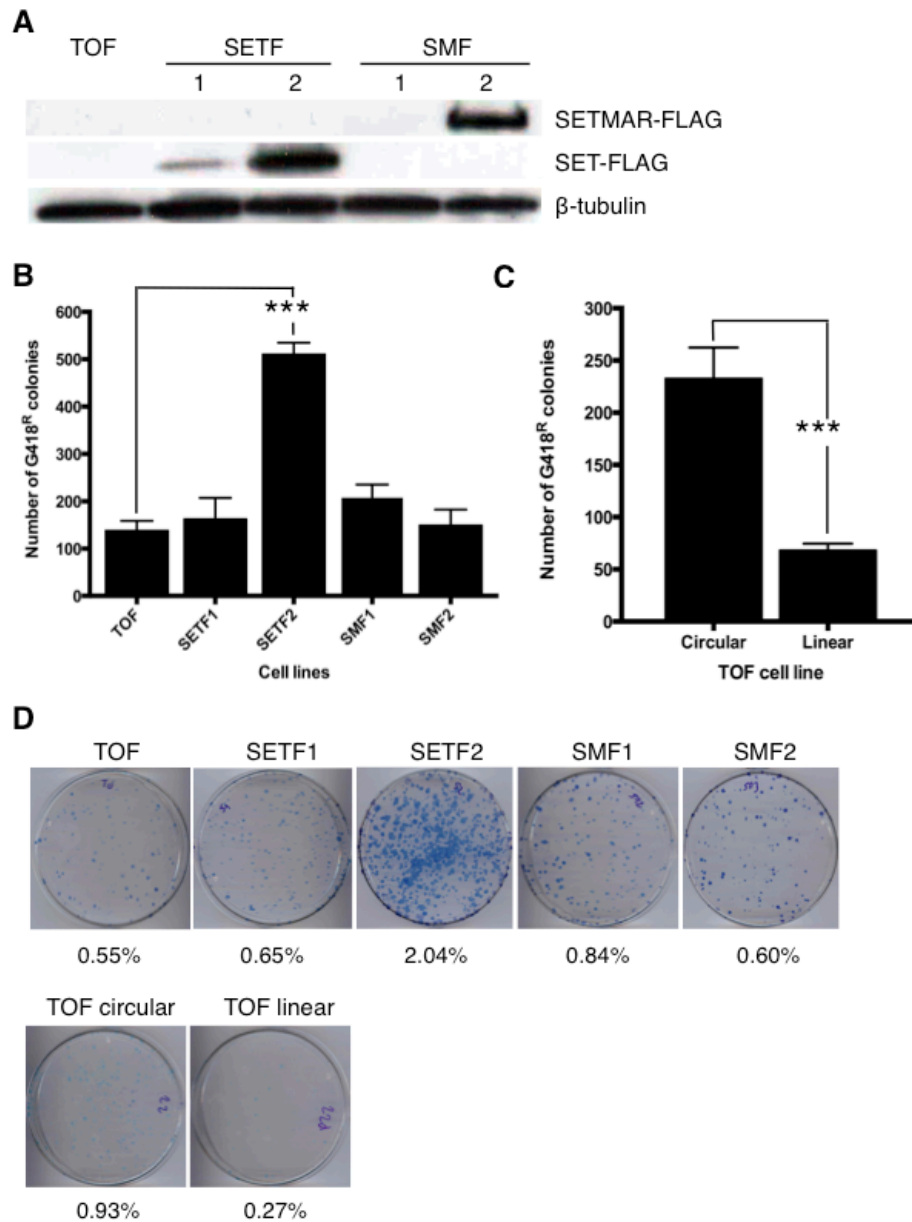


Figure 5-5: Preliminary experiments on DNA integration.

(A) SET and SETMAR overexpression level in T-Rex-U2OS cell lines. The western blot was probed with anti-FLAG and anti-β tubulin antibodies. **(B)** Comparison of the integration efficiency of a circular plasmid encoding a neomycin resistance gene. Average $\pm$ standard deviation of 3 biological replicates. Statistical test: 1way ANOVA, *** p-value < 0.001. Legend continues on the next page.

(C) Comparison of the integration efficiency in the control cell line TOF of a circular or linear plasmid encoding a neomycin resistance gene. Average $\pm$ standard deviation of 3 biological replicates. Statistical test: unpaired t-test, *** p-value < 0.001. (D) Representative pictures of integration plates. The integration rate of each cell line is indicated below each picture.

For testing whether the MAR domain and SETMAR ITR-binding and endonuclease activities are involved in DNA integration, I used a slightly different integration assay. The transfection efficiency of each cell line was determined using a pEGFP plasmid. The fraction of green-positive cells was determined by flow cytometry (Figure 5.6A). The transfection efficiencies of all cell lines except SETF2 were similar to the control cell line with ~90% of all cells transfected. SETF2 transfection efficiency is lower and corresponds to ~80 % of the control cell line or to a transfection of 70% of all cells. The reason for this decrease is unknown but it was also observed in the NHEJ assays. To account for any possible effect of the growth rate on the DNA integration assay, I transfect 5×10^5 cells and then after 24 hours took an aliquot of 5×10^4 cells and applied G418 selection. This accounts for any difference in cell proliferation during the previous 24 hours, which could bias the fraction of transfected cells counted.

Similar to the results presented in Figure 5.5B, the overexpression of the SET domain (SETF2) increases the number of plasmid integrated by 3-fold, whereas SETMAR overexpression (SMF2) does not affect DNA integration (Figure 5.6B). For the new cell lines, a 1.8-fold increase in the number of plasmid integrated is observed for the cell line overexpressing the MAR domain (Figure 5.6B). The overexpression of both SETMAR mutants did not change the frequency of DNA integration significantly (Figure 5.6B).

Therefore, SETMAR ITR-binding and endonuclease activities are not required for DNA integration. A representative selection of the integration plates of each cell line and their respective integration frequency are presented in Figure 5.6C.

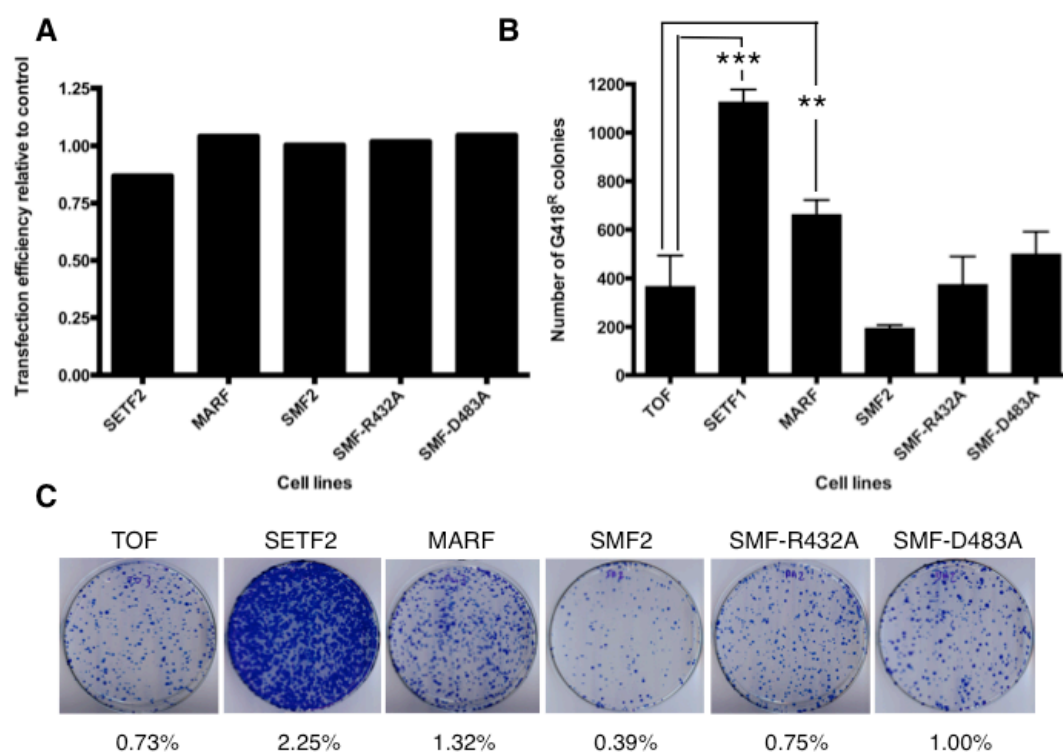


Figure 5-6: The SET and MAR domains increase the frequency of DNA integration in the genome of the U2OS cell line.

(A) Determination by flow cytometry of the transfection efficiency of a GFP plasmid in the different cell lines. The transfection efficiency is shown relative to the control cell line, which has an 88% transfection efficiency. **(B)** Comparison of the integration efficiency of a circular plasmid encoding a neomycin resistance gene. Average $\pm$ standard deviation of 3 biological replicates. Statistical test: 1way ANOVA, ** p-value < 0.01, *** p-value < 0.001. **(C)** Representative pictures of integration plates. The integration rate of each cell line is indicated below each picture.

5.3 Discussion

5.3.1 SETMAR overexpression and cell proliferation

Previous publications have associated SETMAR expression to cell proliferation in the HEK293 and the HEK293T cell lines (De Haro et al., 2010; Williamson et al., 2008b). This association was not observed in the U2OS cell line (Figure 5.1B). Indeed, a medium overexpression of the SET domain or SETMAR slightly reduces the growth rate. This is possibly a non-specific effect, as protein overexpression per se is known to reduce growth rate (Wurm, 2004).

Several methodological problems weaken the association observed in the HEK293 and HEK293T cell lines. The authors used an shRNA approach with only one siRNA to reduce SETMAR expression and showed a subsequent decrease in cell proliferation (De Haro et al., 2010). The main drawback of siRNAs to study the function of a gene comes from the multiple off-targets activity of each siRNA in the cell (Jackson and Linsley, 2010). Indeed, a siRNA can bind mRNAs in an imperfect sequence-specific manner increasing the number of undesirable target genes down regulated (Franceschini et al., 2014). To account for the off-targets effect, the phenotype must be observed with at least two different siRNAs repressing the same gene, an approach not done in the HEK293 and 293T cell lines.

Another criticism of the published work is in the use of an unregulated pCMV promoter. SETMAR expression from a very strong promoter can lead to phenotypic effects irrelevant to SETMAR functions. In contrast, the present

work uses stable cell lines expressing the various proteins at low or medium levels.

Another reason to doubt the published role of SETMAR in cell proliferation is that my gene ontology analysis revealed no enrichment in SETMAR binding sites amongst the cell cycle genes (Chapter 3). An indirect effect is also dubious because the transcriptomic analysis revealed no enrichment in cell cycle genes differentially expressed more than 2-fold (Chapter 4). The transcriptomes of the HEK293 and the T-Rex-U2OS cell lines correlate strongly so any effect on the cell cycle genes in the HEK293 should also be observed in the U2OS (Figure 5.7). It therefore seems likely that the apparent role of SETMAR in cell proliferation is an experimental artefact.

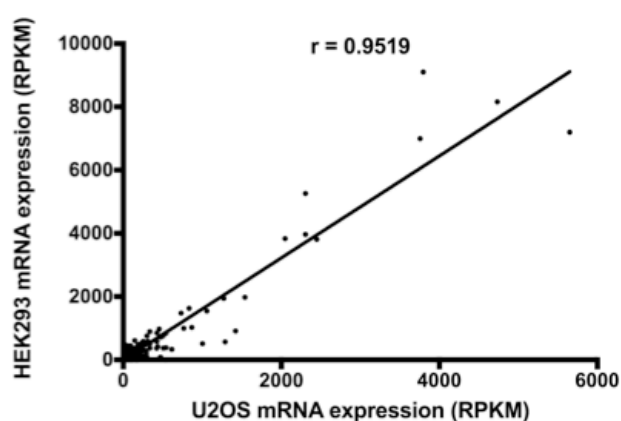


Figure 5-7: The transcriptomes of the HEK293 and U2OS cell lines are similar.

Comparison of the expression of 31,215 genes (protein-coding and non-coding genes) between the HEK293 and the U2OS cell lines. The RPKM (or reads per kilobase per million) values normalize for the sequencing depth and the gene length. The comparison of the expressed genes in both cell lines tell us how similar both transcriptomes are but not which genes are statistically differentially expressed.

5.3.2 The relationship between DNA repair and integration

The DNA repair and the DNA integration assays used in this Chapter rely on the NHEJ pathway. The current model of plasmid integration is based on the group of proteins from classical NHEJ (Wurtele et al., 2003). This is supported by the idea that for the cell, DNA repair or DNA integration of a linear plasmid produces the same result, i.e. the removal of free ends, which could induce apoptosis or cell cycle arrest. The choice between the outcomes is dependent on the presence or absence of a plasmid end in the vicinity of a genomic end. The connection between DNA repair and integration is also supported experimentally where cells deleted for a DNA repair protein became unable to integrate DNA (van Attikum et al., 2001) (Zhong et al., 2002).

Linearization of a plasmid is required before its integration into the genome (Wurtele et al., 2003). However, a linearized plasmid is more likely to be repaired than to be integrated. Indeed, the probability for the DNA repair complex to find the other plasmid end is higher than being near a genomic lesion due to physical continuity of the plasmid DNA, which ensures that the two ends can never be very far apart. Therefore, promotion of DNA repair should further reduce the amount of linear plasmid in the cell by re-circularizing them and by doing it decrease the frequency of integration.

The inefficiency of plasmid integration in mammalian cells comes from the combination of several limitations. The major ones are the low frequency of plasmid linearization, the re-circularization of the majority of linearized plasmids by the NHEJ pathway or their degradation, and the low probability of having a genomic and a plasmid end near each other. The proportion of

linearized plasmid degraded by the cell is unknown but might be quite high. Indeed, the transfection of a linearized plasmid decreases by 3-fold the integration frequency compared to its circular form (Figure 5.5C). From the data obtained in our NHEJ assay we can estimate the frequency of re-circularization as at least 65% (Figure 5.4A). Since the integration frequency is below 1%, we can estimate that about 35% of the linear plasmid should be degraded. A fraction of the transfected cells may also die from apoptosis if the free ends are not repaired in time. However, in the present experiment this is controlled by using a counted aliquot of the living cells for the G418 selection.

5.3.3 Functions of the SET and MAR domains in DNA repair and integration

The published work on the function of SETMAR in NHEJ claims specific roles for the SET and the MAR domains (Fnu et al., 2011). The SET domain dimethylates lysine 36 of histone H3 of nucleosomes near DNA double-strand break. This epigenetic mark recruits and stabilizes the binding of Ku70 and NBS1 to the DNA ends (Fnu et al., 2011). The MAR domain trims damaged and undamaged DNA overhangs before other NHEJ proteins ligate the ends (Beck et al., 2011a). They also claim that SETMAR activity is regulated by several interactions with other proteins involved in the NHEJ such as PRPF19 and DNA ligase IV (Beck et al., 2008) (Hromas et al., 2008). Only a direct interaction between the SET domain and PRPF19 has been confirmed and is supposed to promote the recruitment of SETMAR to DNA double-strand break sites (Beck et al., 2008).

5.3.3.1 Function of the SET domain in DNA repair and integration

The present results of the SET domain overexpression fit with a model wherein DNA repair and integration are inversely associated. In our assays, both DNA repair and integration are supposed to be dependent on the NHEJ pathway, supporting a role for the ancestral SET gene in this pathway. The mechanism by which the SET domain favors DNA integration over DNA repair is not clear. However, its histone methyltransferase activity and/or its interactions with other proteins such as PRPF19 are good candidates.

Almost all mammalian histone methyltransferases function as monomers (Kuzmichev et al., 2002) (Dou et al., 2006). The only known exception is vSET, a viral histone methyltransferase, which is active only as a dimer (Wei and Zhou, 2010). The crystal structure of SETMAR histone methyltransferase domain is a monomer strengthening the hypothesis that the pre-fusion SETMAR protein was operating as a monomer (Schapira, 2011). The MAR domain enforces the dimerization of SETMAR so even though the SET domain does not dimerize *sensu stricto*, the proximity between the two SET domains could affect its activity. This idea is supported by the failure of SETMAR overexpression to affect DNA repair and integration in the current experiments (Figure 5.4B and 5.6B).

How might the SET domain promote DNA integration? One way might be if it acted as a general DNA damaging agent, providing chromosomal breaks as substrates for NHEJ. However, this seems unlikely for a histone methyltransferase. Alternatively, H3K36me2 may delay the re-circularization of plasmids, increasing the window of opportunity for a plasmid end to be in

the vicinity of a genomic end. This is supported by a decrease in re-circularization with both compatible and incompatible ends (Figure 5.4B).

5.3.3.2 Function of the MAR domain in DNA repair and integration

The overexpression of the MAR domain stimulates both DNA repair and integration (Figure 5.4B and 5.6B). These results seem to conflict with the proposed inverse relationship between DNA repair and integration. However, the overexpression of the transposase domain should mostly produce MAR dimers and dimers between endogenous SETMAR and MAR monomers. This second type of dimer could be responsible for the increase in DNA integration. Indeed, the SETMAR-MAR dimers contain only one SET domain and could promote DNA integration as when the monomeric SET domain is expressed alone (Figure 5.6B). However, the low number of endogenous SETMAR molecules in the nucleus might not be sufficient to explain the increase in DNA integration.

It seems unlikely that the MAR domain promotes DNA repair directly, so any effect would have to be dependent on interactions with other NHEJ factors. Furthermore, the transposase domain has not been found to directly recognize γ -H2AX, which is associated with DNA double-strand breaks (Pilch et al., 2003). It is also unknown whether the MAR domain has specific interacting partners. However, direct interactions between the Hsmar1 transposase and DNA repair proteins cannot be excluded. A role for MAR's ITR-binding and endonuclease activities is not supported by the current data for the SETMAR R432A and D483A mutants.

Another idea is that deletion of the SET domain might relieve a structural hindrance blocking access of the transposase domain to the DNA ends where it could trim the DNA overhangs. However, this argument would only apply if one accepted the published claims for an *in vivo* nuclease activity, which seem tenuous at best. Nevertheless, this last hypothesis could also explain the increase in DNA repair for both compatible and incompatible ends (Figure 5.4B). However, if the MAR domain really does have a trimming activity, we would expect to observe a larger increase in the repair of incompatible ends versus compatible ends. In fact, the increase in the efficiency of repair is similar for both types of ends. Recovery and DNA sequencing of repaired junctions produced from compatible and incompatible ends would shed some light on these issues.

5.3.4 SETMAR variants and DNA repair and integration

The human SETMAR gene encodes three coding splicing variants but only two are observed in the U2OS cell line (Figure 5.8A and Supplementary Figure 2). The first isoform represents most of the transcripts and codes for the full-length protein. The second isoform represents only a few per cent of the transcripts but encodes for a defective histone methyltransferase protein because a splicing event removes the majority of the SET and post-SET domains in the second exon. Interestingly, the second variant is specific to the species where the SET gene is fused to the Hsmar1 transposase (Figure 5.8B). The 5' donor site is present in primates and other several mammals but the acceptor site is only found in anthropoid primates, except for the old-world

monkeys where a single mutation in their common ancestor abolishes the acceptor site. The marmosets also lost their 5' donor site but another less conserved site is present 20 nucleotides away.

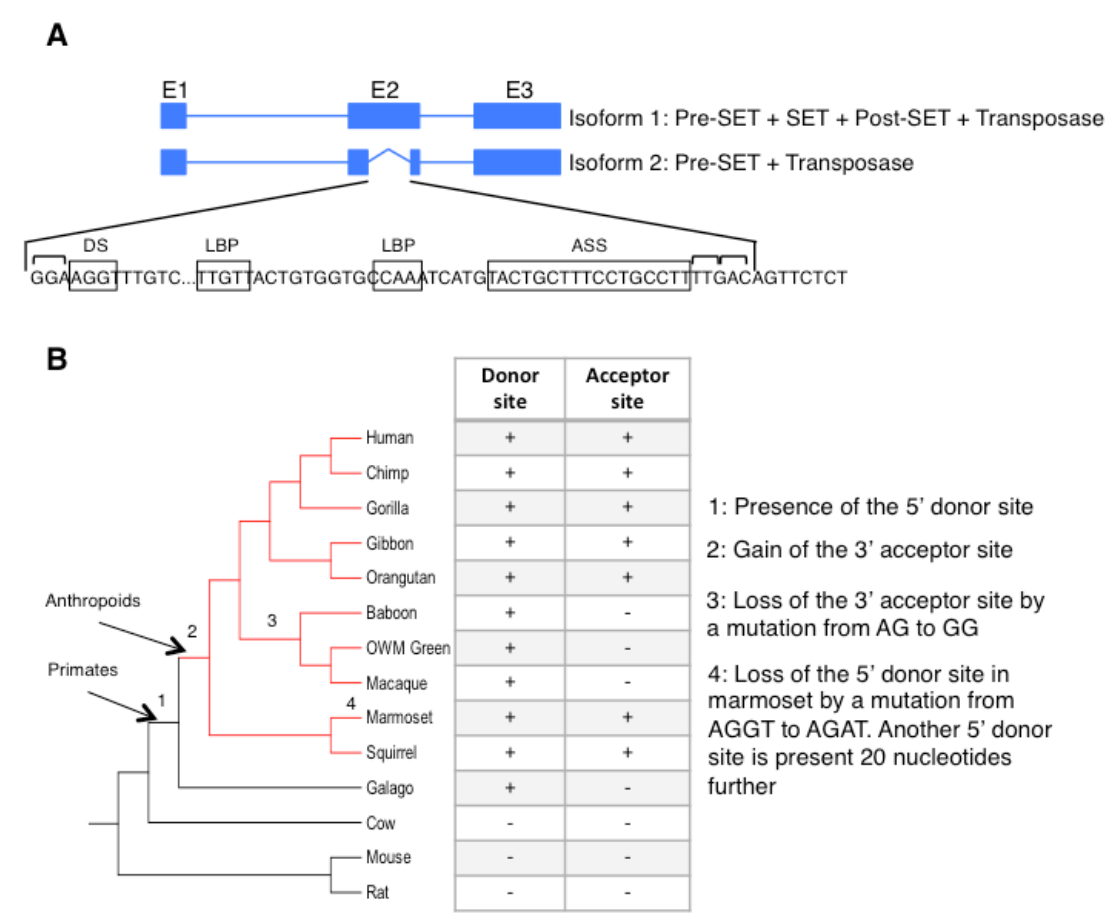


Figure 5-8: SETMAR second isoform is specific to the species containing the MAR domain.

(A) Human *SETMAR* gene encodes two isoforms, the full-length protein (isoform 1) and a truncated protein (isoform 2). The second isoform cannot perform its histone methyltransferase activity because of the deletion of the SET and post-SET domains. Canonical donor site (DS), lariat branch points (LBP), and acceptor splicing site (ASS) are present in the second exon of *SETMAR*. The top brackets represent the exon codons. **(B)** Phylogenetic tree of *SETMAR* second exon in several mammals. The 5' donor site is found in all primates except for the marmoset (see 4) but should have been gained before because it is present in several other mammals. The 3' acceptor site is specific to anthropoids' primates except for the old-world monkeys that lost it by a single point mutation.

The data presented in this Chapter suggests that there was an ancestrally active monomeric SET protein that lost its activity upon its fusion to the MAR domain. It is tempting to think that the increase in DNA integration observed with the MAR domain is due to the formation of SETMAR dimers with only one SET domain (SETMAR-MAR). These observations and the specificity of SETMAR second isomer to the species encoding the chimeric gene support the hypothesis that the dimerization enforced by the MAR domain might hinder the ancestral functions of the SET domain. The synthesis of SETMAR monomers with a deleted SET domain might have been a way for the first anthropoid primates to cope with the inactivation of the functions of the SET gene. The importance of the second SETMAR variant for the current anthropoid primates is unknown but the fact that old-world monkeys are unlikely to express it does not support a fundamental role.

**6. USE OF CONSTITUTIVE PROMOTERS
TO FIND HYPERACTIVE
TRANSPOSASES IN A BACTERIAL
GENETIC SCREEN**

6.1 Introduction

The discovery of hyperactive transposases in bacteria has mostly been accomplished by screening libraries of transposase mutants with the papillation assay (Figure 6.1) (Huisman and Kleckner, 1987) (Pajunen et al., 2010). This assay is based on a promoter-less *lacZ* gene flanked by transposon ends. This reporter is integrated in a silent region of the genome of *E. coli*. The transposase gene is provided in trans on a plasmid to simplify mutagenesis and library handling. Transposition events into an expressed ORF give rise to *lacZ* gene fusion proteins. When this happens within a colony growing on an X-gal indicator plate, it converts the cell to a *lac*⁺ phenotype, which allows the outgrowth of a blue microcolony (papillae) on a background of white cells. The transposition rate is given by the number of papillae per colony and by the rate of their appearance. The visual read-out of this assay simplifies the screening of mutant libraries for discovering hypo- or hyperactive transposases. The mating-out assay is a more quantitative assay using the same reporter strain. Here, the rate of transposition is measured by movement of the reporter from its chromosomal location into a conjugative plasmid.

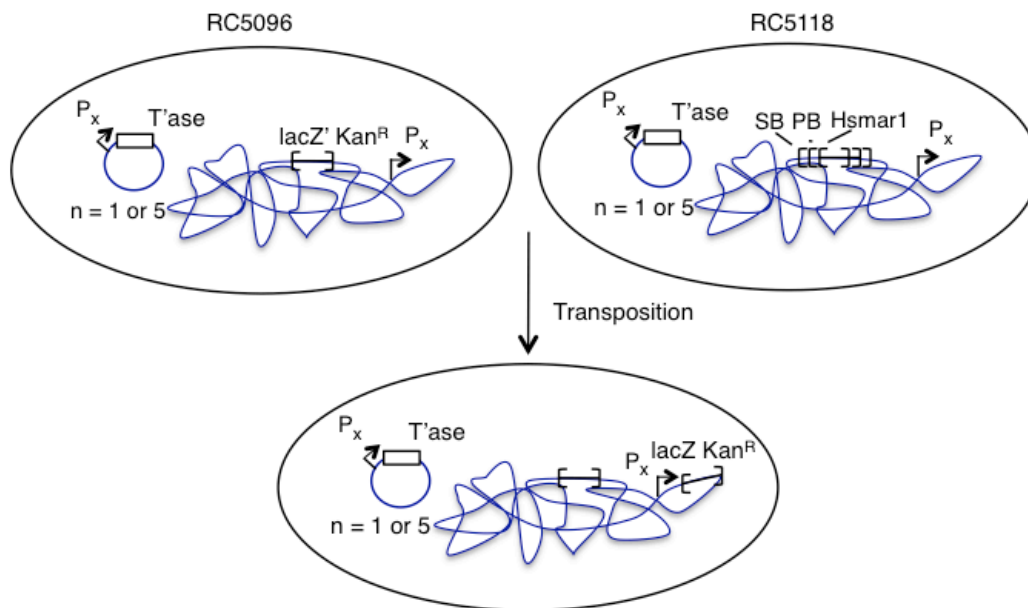


Figure 6-1: Papillation and mating-out assays.

The Hsmar1 (RC5096) or Sleeping Beauty (SB)-PiggyBac (PB)-Hsmar1 (RC5118) transposition reporters, which encode a promoter-less *lacZ* gene and a kanamycin resistance marker, has been integrated at a transcriptionally silent locus in a *lacZ*- *E. coli* strain. In absence of transposase, the *lacZ* gene is not expressed therefore the strain produces white colonies on X-gal reporter plates. When the transposase is supplied in trans from vector, if a transposon integrates into the ORF of a transcribed gene, a *lacZ* fusion protein will be produced. The cell's descendants will expressed LacZ and therefore will appear as blue papillae on X-gal reporter plates. Black arrow, promoter; open brackets, transposon ends; empty rectangle, transposase gene.

For the mating-out assay, a chloramphenicol resistant derivative of the conjugative plasmid pOX38 is introduced into the reporter strain. Transposition of the reporter into the plasmid is detected by selecting transconjugants after mating with a recipient strain on chloramphenicol and kanamycin.

My aim was to refine the papillation assay so that I could separate the negative effects of overproduction-inhibition (OPI) from the underlying rate of transposition. OPI is a phenomenon which arises when the increasing concentration of transposase in the cell leads to the sequestration of the transposon ends, inhibiting the transposition reaction (see 1.2.4 for more

details). These two aspects are intimately linked by the molecular mechanism of the transposition reaction. My hypothesis was that library screening for hyperactive mutants would be improved if the transposase expression could be reduced below the level at which OPI is the dominant factor.

In previous screening strategies the transposase was expressed from the strong P_{TAC} promoter. I therefore set out to use a set of constitutive promoters to obtain transposase concentrations below and above the OPI threshold. From the sets of constitutive promoters and vectors tested, I select three expression points corresponding to a low, optimal, and high transposase expression. I then compared the expected and experimental performance of several Hsmar1 mutants to demonstrate that these three expression points are sufficient for finding auto-regulated hyperactive transposases.

I also constructed a new bacterial strain in which the *lacZ* reporter gene is flanked by Hsmar1, PiggyBac, and Sleeping Beauty transposon ends. All three transposons can therefore be tested in a single reporter strain. In *E. coli*, the wild type Hsmar1 and PiggyBac transposases have a similar transposition rate whereas the Sleeping Beauty transposase was confirmed to be inactive, even at the lowest transposase concentrations where OPI is not the dominant factor.

6.2 Results

6.2.1 Characterization of the papillation assay using a strong inducible promoter

In the existing papillation assay, the transposase was provided by the protein expression vector pMAL-c2x (Liu and Chalmers, 2014). In this vector, the transposase gene is fused to a MBP tag under the control of the strong inducible P_{TAC} promoter. I began my investigation by analyzing the effect of the MBP tag on the transposition rate and by measuring the response of the assay to different concentrations of lactose and IPTG added to the LB agar plates (Figure 6.2A). The transposition rate is approximately proportional to the number of papillae per colony after 5 days of incubation (Huisman and Kleckner, 1987). In absence of transposase, the number of papillae per colony in all the conditions tested is either 0 or 1 (Figure 6.2A , $\emptyset$ lane). The single papillae are likely due to a stochastic recombination event resulting in the expression of the *lacZ* gene. In presence of the transposase or MBP-transposase (respectively middle and right lanes), the number of papillae per colony varies with the concentration of IPTG and lactose.

Independent of the type of transposase and the concentration of IPTG, the number of papillae increases with the concentration of lactose (Figure 6.2A). One explanation could be the induction of the P_{TAC} promoter by the lactose. However, since the strain is *lacZ*⁻ the lactose cannot be metabolized to allolactose, the inducer of the lac operator in the P_{TAC} promoter (Wheatley et al., 2013). Another explanation is that the *lac*⁺ cells form larger, more

clearly visible, papillae because they are able to continue growing after the lac^- cells have exhausted the carbon source in the LB agar. To determine whether this is the case, a papillation assay with the MBP-transposase was done normally except that after one day, the colonies were transferred to M9-agar plates supplemented with either 1% lactose or sucrose (Figure 6.2B, more colonies are presented in Supplementary Figure 3). Similar to the results obtained on LB-agar plates, the number of papillae is higher on M9-agar plates supplemented with lactose rather than sucrose. In absence of any carbon source except lactose, only the papillae continue to grow whereas in presence of sucrose, lac^- and lac^+ cells continue to divide until the depletion of the carbon source. Therefore any cells with a late transposition event, i.e. few divisions before the cells stop dividing, will not produce sufficient descendants to become visible to the naked eye. Lactose does not influence the transposition rate but does allow the late transposition events to become visible, explaining the positive correlation between the number of papillae and the lactose concentration.

At all lactose concentrations the transposition rate is highest at zero and 0.1 mM IPTG for the transposase and the MBP-transposase, respectively (Figure 6.2A). Any further increase in the IPTG concentration results in a decrease of the transposition rate, consistent with the effects of OPI (Claeys Bouuaert et al., 2013). Except for the MBP tag, the papillation assays for the transposase and MBP-transposase are isogenic. The difference in transposition rate between the transposases must therefore be due to an effect of the MBP tag on the transposase concentration or activity.

To distinguish between these alternatives, I investigated the expression level of the transposase and MBP-transposase (Figure 6.2C). On a Coomassie-stained gel, only the MBP-transposase expression is discernible and only when induced with at least 0.1 mM IPTG. A slight increase in protein concentration is visible between 0.1 and 0.5 mM IPTG.

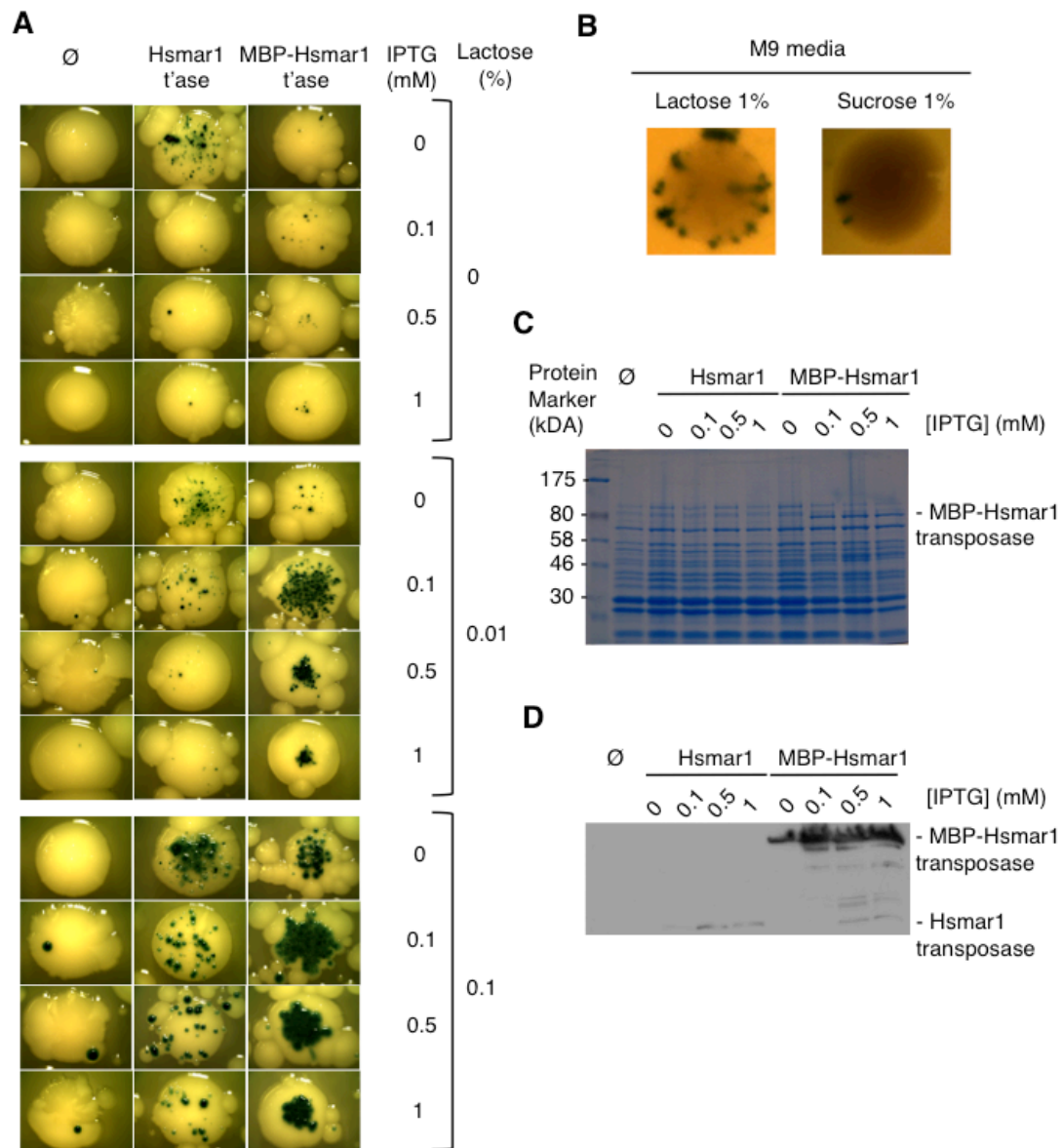


Figure 6-2: The number of papillae per colony varies with the concentration of lactose and IPTG.

(A) An expression vector encoding Hsmar1 or MBP-Hsmar1 transposase (t'ase) was transformed into the papillation strain and plated on different lactose and IPTG concentrations. Representative colonies of the papillation plates are shown. On some pictures, smaller colonies surrounding the main colony are visible. These satellite colonies appear only after several days of incubation when the ampicillin present on the plate has been degraded. They can be ignored because they do not contain any transposase expression plasmid. Legend continues on the next page.

(B) The papillation strain was transformed with an expression vector encoding the MBP-transposase, plated on an activated PVDF membrane on LB-agar supplemented with antibiotics and incubated overnight at 37 °C. The following day, the membrane was lifted off the LB agar and placed on M9-agar media supplemented with the necessary amino acids and 1% sucrose or lactose and incubated one week at 37 °C. Representative colonies of the papillation plates are shown. Pictures of the plates are presented in Supplementary Figure 3. **(C)** 10% acrylamide-SDS gel of Hsmar1 and MBP-Hsmar1 transposase expression induced with different concentration of IPTG. Only the MBP-Hsmar1 expression is visible with Coomassie staining. **D/** Western blot of the SDS-PAGE gel presented in B/ probed with an anti-Hsmar1 antibody.

Since the expression of the transposase is not detectable on a Coomassie-stained gel, I prepared a western blot of the same samples and probed with a polyclonal antibody against the last nine amino acids of the Hsmar1 transposase (Figure 6.2D). At 0.1 mM IPTG, the amount of transposase produced is at the limit of the detection. The amount of transposase increases sharply between 0.1 and 0.5 mM IPTG before reaching a plateau with little difference between 0.5 and 1 mM IPTG. The papillation assay is consistent with the amounts of protein detected in the western blot. The number of papillae decreases between zero and 0.1 mM, and then again from 0.1 to 0.5 mM (Figure 6.2A). The presence of the tag significantly increases the amount of protein induced (Figure 6.2D, compared 0 and 0.1 mM IPTG in both constructs). The signal in the western blot reaches saturation at 0.1 mM induction therefore there is no significant difference between 0.1, 0.5, and 1 mM IPTG induction. In the papillation assay, the number of papillae per colony peaks at 0.1 mM induction before decreasing at 0.5 and 1 mM confirming a higher transposase concentration at 1 mM than 0.1 and 0.5 mM. The concentration of the MBP-transposase far exceeds the

concentration of the untagged transposase. Even in absence of induction, the quantity of the MBP-transposase is higher than a fully induced transposase. Except for the MBP-tag, both conditions were similar therefore the presence of the tag must improve the protein stability.

The dose-response relationship of transposase expression was further investigated using a more quantitative mating-out assay. This revealed a 13-fold increase of the transposition rate of the transposase compared to the MBP-transposase in absence of IPTG, confirming the results of the papillation assay (Table 6.1). I next determined whether the hyperactivity of the previously described W118R mutant was also sensitive to the MBP-tag (Liu and Chalmers, 2014). The tagged mutant transposes 24-times more than the tagged wild type, in agreement with previously published data (Table 6.1). The hyperactivity decreases to a 16-fold increase when the untagged mutant is compared to the untagged wild type. For the following experiments, only untagged transposases were used.

Construct	Transposition frequency	Mutant/W.T.
P _{TAC} MBP-transposase	1.99 (±0.43) x 10 ⁻⁶	
P _{TAC} Transposase	2.58 (±0.02) x 10 ⁻⁵	
P _{TAC} MBP-transposase W118R	4.69 (±3.38) x 10 ⁻⁵	24
P _{TAC} Transposase W118R	4.11 (±0.97) x 10 ⁻⁴	16

Table 6-1: Transposition frequencies of wild type and mutant Hsmar1 transposases.

The bacterial mating-out assays have been done with the RC5097 strain and the pMAL-c2x vector. Transposition frequencies are the average of three independent experiments ± standard error of the mean.

The P_{TAC} promoter has many disadvantages for studying transposition. It is impossible to attain a level of protein expression below the one provided by the leaking of the promoter. Because of the unequal diffusion of the inducer in agar plate due to the colony morphology, the transposase expression varies between cells. The all-or-none induction, which results from the cooperativity between the three-lac repression binding sites in P_{TAC}, produces a very steep IPTG dose-response curve, which impedes an accurate analysis of the effect of transposase concentration (Figure 6.2D).

6.2.2 Constitutive promoters provide expression levels below and above the OPI threshold

To obtain a wider range of transposase expression, and for removing the inherent cell-to-cell variability of inducible promoters, we synthesized a set of five constitutive promoters based on Alper et al. (2005) (see M&M section

2.5.1 for more details). In addition we created a null-expression vector by replacing the promoter and ribosome binding site with a featureless, undistorted, DNA sequence consisting of a simple repeat of the tetranucleotide (GACT)_{n=44}. To increase the range of expression levels obtainable, the six promoters were cloned either into a one-copy vector or a five-copy vector (pBACe3.6 and pGHM491, respectively). Henceforth I will use the following nomenclature: Bp1 to Bp6 represents the six promoters cloned into the single copy vector, while Ip1 to Ip6 represents the six promoters cloned into the five copy vector.

To determine whether the set of constitutive promoters represents a better alternative than P_{TAC}, I used western blotting to compare the expression level of Ip1-6 with different IPTG inductions of the P_{TAC} promoter (Figure 6.3A). No band is visible for Ip1, which is consistent with the absence of both promoter sequence and RBS motif. An orderly and progressive increase in expression is observed from Ip2 to Ip4, followed by a steep increase to reach the saturation level of the western blot for Ip5 and Ip6. The results for the P_{TAC} promoter are similar to the ones obtained in Figure 6.2C with no visible band in absence of inducer and an increase of expression at 0.1 and 1 mM of IPTG. An induction of 0.1 mM of IPTG produces a transposase concentration between Ip3 and 4 whereas a 1 mM induction is located between Ip4 and 5. It should also be noted that the copy number of the P_{TAC} vector is about 20, four times more than the Ip1-6 plasmids.

Although the western blot in Figure 6.3A indicates a progressive increase in expression level with the different constitutive promoters, the dose-response curve of a western blot is not linear. To obtain an estimate of

the relative strengths of the promoters, a papillation cell extract containing the Ip6 expression vector was serially diluted and compared to a selection of undiluted extracts from some of the other promoters (Figure 6.3B). Bp6, the strongest promoter, when present in one copy per cell does not produce a detectible amount of transposase. According to the vector copy number, one would have expected Bp6 to produce one-fifth the amount of transposase than Ip6. However, it seems to produce much less, which is likely explained by the increase in stochasticity of the transposase expression from a single-copy vector rather than a five-copy vector. The dilution series also shows that Ip6 is ~1.5-fold stronger than Ip5, 2-fold stronger than Ip4, and ~ 6-fold stronger than Ip3. Ip3 is ~2-fold stronger than Ip2. The data are summarized in Table 6.2. These differences match more closely the mRNA production value of each promoter than their strength (see Table 2.5 in M&M section 2.5.1). It is also supported by the fact that Ip6 has an inferior promoter strength compared to Ip5 but a higher mRNA production value.

Name	Vector	Copy number	Promoter from (Alper et al., 2005).	Relative promoter strength (n =20). From (Alper et al., 2005).	Relative promoter strength (n =5)
Bp1	pBACe3.6	1	/	0	ND
Bp2	pBACe3.6	1	00	0.003	ND
Bp3	pBACe3.6	1	JJ	0.159	ND
Bp4	pBACe3.6	1	K	0.299	ND
Bp5	pBACe3.6	1	E	0.743	ND
Bp6	pBACe3.6	1	W	1	ND
Ip1	pGHM491	5	/	0	ND
Ip2	pGHM491	5	00	0.003	0.08
Ip3	pGHM491	5	JJ	0.159	.17
Ip4	pGHM491	5	K	0.299	.50
Ip5	pGHM491	5	E	0.743	.67
Ip6	pGHM491	5	W	1	1

Table 6-2: Promoters and vectors used in this Chapter.

The promoters from (Alper et al., 2005) are derivative of the constitutive bacteriophage $P_{L-\lambda}$ promoter. The original relative promoter strengths have been determined by using a 20-copy vector and correspond to the mRNA production value (Table 2.5 in M&M section 2.5.1) (Alper et al., 2005). ND, not defined.

To determine whether the effect of the MBP tag on the transposition rate is also observed with the constitutive promoters, I performed a papillation assay with Ip1 and Ip6 (Figure 6.3C). A decrease in the number of papillae per colony is also detected in presence of the MBP tag for both Ip1 and Ip6. To verify if the difference is also explained by a change in transposase concentration, I determined by western blotting the expression of the transposase and MBP-transposase from both vectors (Figure 6.3D). A similar increase in transposase concentration is observed in presence of the MBP tag for both Ip1 and Ip6 confirming the results obtained with the P_{TAC} promoter.

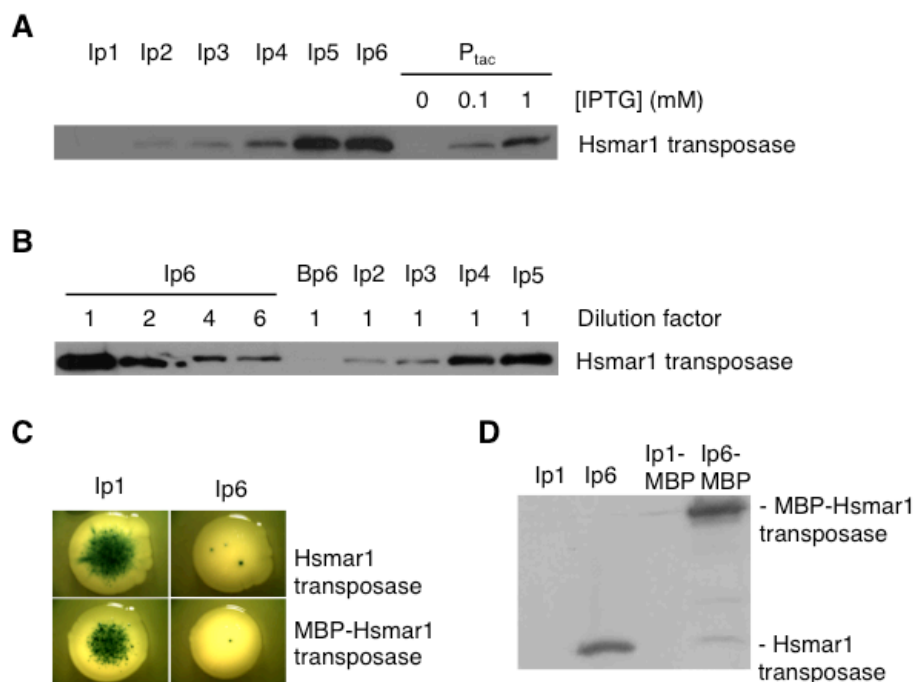


Figure 6-3: Hsmar1 transposase expression from different promoters.

(A) Western blot probed with an anti-Hsmar1 antibody. Hsmar1 transposase is expressed from the Ip1-6 expression vectors and from the P_{TAC} promoter induced with different concentrations of IPTG. (B) Western blot probed with an anti-Hsmar1 antibody. The expression of the Hsmar1 transposase from several constitutive promoters is compared to a serial dilution of the transposase expression from Ip6. To conserve the same number of bacterial cells in each lane, the Ip6 dilutions have been complemented with cells transformed with an empty expression vector. (C) Representative colonies of a papillation assay with Hsmar1 or MBP-Hsmar1 transposase expressed from Ip1 and Ip6. A decrease of the transposition rate is also observed in presence of the MBP tag. (D) Western blot probed with an anti-Hsmar1 antibody of Hsmar1 and MBP-Hsmar1 transposase expressed from Ip1 and Ip6. The presence of the MBP tag increases the transposase concentration.

Next, I performed the papillation assays with the complete set of expression vectors Bp1-6 and Ip1-6 (Figure 6.4A, whole plates are presented in Supplementary Figure 4). In absence of transposase, no papillae were observed for both empty vectors. Even though transposase expression from

Bp1-6 and lp1 are not detectable by western blotting, all of them produce papillae when transformed into the papillation strain. For the single copy vectors, the transposition rate is almost zero for Bp1 but peaks with Bp2. At this point it is worth remembering that the 'null' promoter p1 lacks a RBS, whereas all other constructs have a consensus RBS. The number of papillae per colony then decreases with the increasing strength of the promoter, consistent with the effects of OPI. For the five copy vectors, the transposition rate peaks with lp1 before decreasing sharply with lp2 then slightly up to lp6.

To obtain a more accurate estimation of the transposition rate, the number of papillae of four representative colonies from two independent experiments were counted and averaged for each expression vector. Based on our knowledge of the OPI mechanism, we know that above the OPI threshold, a doubling of the transposase concentration halves the transposition rate (Claeys Bouuaert et al., 2013). Consequently, below the OPI threshold, an increase in transposase concentration will result in an increase of the number of papillae per colony whereas above the threshold, the stronger the promoter is, the less papillae per colony it will produce. This allowed me to deduce the order of the promoter strengths based on their transposition rates (Figure 6.4B). The OPI threshold cannot be determined exactly but it should be somewhere between Bp1 and lp1 or lp1 and Bp2 because the transposition rate of lp1 is superior to both Bp1 and Bp2. Bp1 is the only indisputable expression vector below the OPI threshold and its low transposition rate is due to the limited amount of transposase produced. At first sight, the huge increase in transposition rate between Bp1 and lp1 seems inconsistent with the 5-fold difference in vector copy number. One explanation

is that the expression level from Bp1 is so low that few, if any, cells have two molecules of transposase, which is the lowest number that can support a transposition reaction.

Bp2, which corresponds to an inactivated promoter, is the only other expression vector with a transposition rate slightly lower than Ip1. Therefore, I can conclude that the OPI threshold in the papillation assay should be at a few dimers per cell. For the remaining expression vectors, the decrease of the transposition rate correlates with the increase of the promoter strength and the plasmid copy number. Thus, the higher the transposase expression level, the lower the rate of transposition. Theoretical consideration suggests that the inverse relationship between transposase concentration and transposition rate above the OPI threshold fits an inverse power law, i.e. a doubling of the transposase concentration halves the transposition rate (Claeys Bouuaert et al., 2013). I therefore plotted the relative promoter strength of Ip2-6 defined from the Figure 6.3B and summarized in Table 6.2 against the transposition rate established in Figure 6.4B (Figure 6.4C) and fitted an inverse power law curve. The correlation coefficient of 0.95 confirms the theoretical expectation and the results previously obtained *in vitro* and in HeLa cells (Claeys Bouuaert et al., 2013).

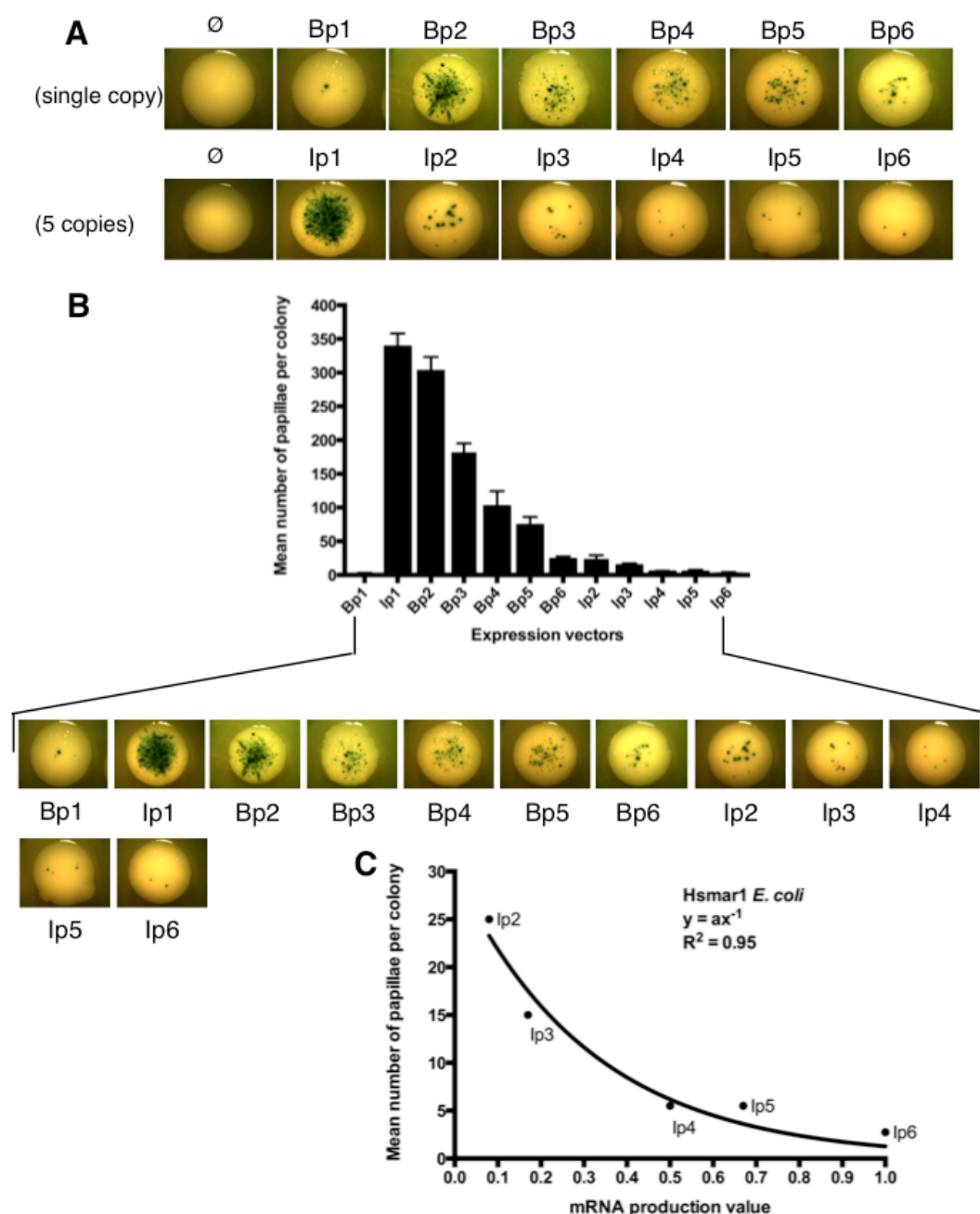


Figure 6-4: Characterization of the papillation assay with the set of constitutive promoters expressing the Hsmar1 transposase.

(A) Representative colony of each papillation plates. Whole plate pictures are presented in Supplementary Figure 4. (B) Quantification of the number of papillae per colony. The expression vectors have been ordered by increasing strength (see main text). Average $\pm$ standard deviation of four colonies from two independent experiments. Legend continues on the next page.

(C) Plot of the relative expression of Ip2 to 6, as determined in Table 6.2, against their average number of papillae, as quantified in Figure 6.4B. The curve fits well to an inverse power law.

To confirm the papillation assay results, I determined the transposition rate with Ip1-6 using the mating-out assay (Table 6.3). The Bp1-6 vectors were not tested because the pBAC expression vector and the pOX38, required for the assay, were incompatible because they both encoded a chloramphenicol-resistance gene. In agreement with the results of the papillation assay, the transposition rate in the mating-out assay peaks with Ip1 before decreasing with the increasing promoters' strength. There was a 10-fold decrease in the transposition rate between Ip1 and 2, which is similar to the 14-fold decrease observed in the papillation assay (Figure 6.4B). The difference increased to 50-fold for Ip3 and 100-fold for Ip4, 5, and 6. The transposition rate of Ip1 is similar to the non-induced P_{TAC} confirming a low leaking from the inducible promoter.

Construct	Transposition frequency
P_{TAC} Hsmar1	$2.58 (\pm 0.02) \times 10^{-5}$
Ip1 Hsmar1	$4.73 (\pm 1.02) \times 10^{-5}$
Ip2 Hsmar1	$3.42 (\pm 1.46) \times 10^{-6}$
Ip3 Hsmar1	$1.06 (\pm 0.87) \times 10^{-6}$
Ip4 Hsmar1	$5.64 (\pm 2.47) \times 10^{-7}$
Ip5 Hsmar1	$2.92 (\pm 1.26) \times 10^{-7}$
Ip6 Hsmar1	$3.22 (\pm 1.02) \times 10^{-7}$

Table 6-3: Transposition frequencies of different concentrations of wild type Hsmar1 transposase.

The bacterial mating-out assays have been done with the RC5097 strain and the Ip1-6 vectors. Transposition frequencies are the average of three independent experiments $\pm$ standard error of the mean.

6.2.3 Characterization of several Hsmar1 mutants

The transposition assay provides the opportunity to characterize *in vivo* the impact of mutations on the transposase activity. The weakest (Bp1) and the strongest (Ip6) expression vectors can be used to find hyperactive and OPI-resistant transposases, respectively. These two promoters offer complementary information on the effect of a mutation on the transposition rate and the auto-regulatory features. Furthermore, these two expression points can characterize the effect of a mutation on the ITR affinity. An increase in the affinity will improve the transposition rate below the OPI threshold but decrease it above the threshold compared to the wild type transposase (Claeys Bouuaert et al., 2013). Opposite results will be obtained for a mutation decreasing partially the ITR affinity.

To test whether this will be a fruitful experimental approach, I assayed several Hsmar1 transposase mutants using the papillation assay and/or the mating-out assay (Figure 6.5, Table 6.4). The results were broadly similar but most easily appreciated by the visual readout of the papillation assay (Figure 6.5, whole plates are presented in Supplementary Figure 5). The K50Q mutation is located in the first HTH DNA binding motif of the transposase. The mutation results in the loss of a positive charge that should decrease the affinity of the transposase for the transposon end. In HeLa cell assays, no difference has been observed between the wild type and the K50Q mutant (Danxu Liu, unpublished data). However, in the papillation assay, the mutant is slightly more active below the OPI threshold (Figure 6.5). At the optimal and high transposase concentration, the mutant transposase is similar or perhaps

slightly hypoactive compared to wild type. The mating out assay was broadly similar (Table 6.4). However, the mutant transposase is OPI-resistant compared to wild type with only a 36-fold decrease of the transposition rate between Ip1 and 6 compared to a ~100-fold for the wild type transposase, which is consistent with the OPI theory (Claeys Bouuaert et al., 2013).

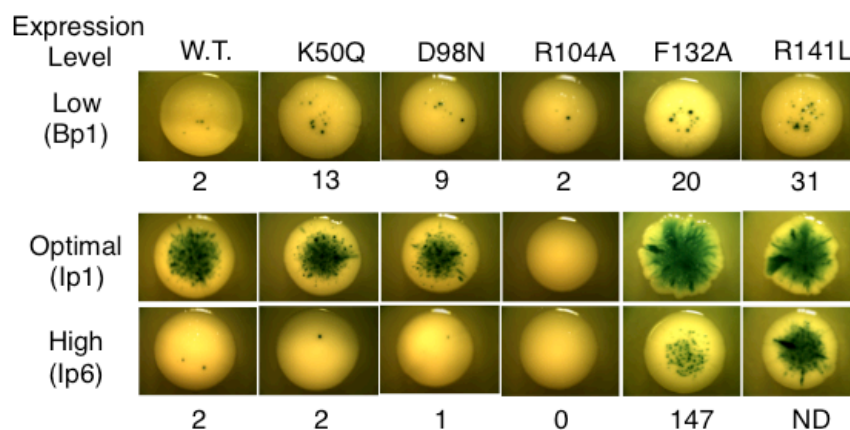


Figure 6-5: Transposition rate of different Hsmar1 transposase mutants.

Each mutant has been tested in low, optimal and high transposase expression level (Bp1, Ip1, and Ip6, respectively). Representative colony of each papillation plate is shown. The average number of papillae per colony is indicated below the pictures. Average $\pm$ standard deviation of four colonies from two independent experiments. ND: not determined. Pictures of the plates are presented in Supplementary Figure 5.

Construct	Transposition frequency	Mutant/W.T.
Ip1 W.T.	$4.73 (\pm 1.02) \times 10^{-5}$	
Ip6 W.T.	$3.22 (\pm 1.02) \times 10^{-7}$	
Ip1 K50Q	$3.01 (\pm 0.66) \times 10^{-5}$	0.64
Ip1 D98N	$4.43 (\pm 0.44) \times 10^{-5}$	0.94
Ip1 F132A	$9.73 (\pm 4.53) \times 10^{-4}$	21
Ip1 R141L	$2.42 (\pm 1.68) \times 10^{-4}$	5
Ip6 K50Q	$8.29 (\pm 3.61) \times 10^{-7}$	3
Ip6 D98N	$2.79 (\pm 1.32) \times 10^{-7}$	0.87
Ip6 F132A	$5.79 (\pm 2.63) \times 10^{-5}$	180
Ip6 R141L	$3.24 (\pm 1.43) \times 10^{-4}$	1006

Table 6-4: Transposition frequencies of different concentrations of Hsmar1 transposase mutants.

The bacterial mating-out assays have been done with the RC5097 strain and the Ip1 or Ip6 vectors. Transposition frequencies are the average of three independent experiments $\pm$ standard error of the mean.

The D98N mutation represents the only difference between the DNA binding domain of SETMAR and the ancestral Hsmar1 transposase sequence used in these experiments. The mutation substitutes a negatively charged amino acid to a neutral one. Therefore, an increased affinity of the mutant for the transposon ends might be expected because of the loss of the repulsive charge. However, the papillation and the mating-out assays show no significant difference to the wild type transposase. Thus, the mutation is presumably a neutral substitution (Figure 6.5 and Table 6.4).

The R104A is in the second DNA binding domain of the transposase (Roman et al., 2007) (Claeys Bouuaert et al., 2013). In the Mos1 structure this residue is seen to interact with a DNA base in the middle of the transposon

end. The loss of the positive charge reduces the affinity for the transposon end (Claeys Bouuaert et al., 2013). In HeLa cell assays, the mutant is generally hypoactive except in a severe OPI condition where it became hyperactive compared to the wild type transposase (Claeys Bouuaert et al., 2013). However, in *in vitro* transposition assay, this mutant is fully active and resistant to OPI (Claeys Bouuaert et al., 2013). In the papillation and mating-out assays, the R104A mutant is inactive except with Bp1 where a few papillae are observed (Figure 6.5 and Table 6.4).

In the structure of the catalytic domain of SETMAR, the F132A mutation is predicted to interact with the L258 residue of the second monomer to form a part of the dimer interface (Goodwin et al., 2010). In the papillation and mating out assay, the F132A mutant is hyperactive at all transposase concentrations (Figure 6.5 and Table 6.4). Interestingly, the F132A mutant is resistant to OPI and has only a 3-fold decrease of the transposition rate between Ip1 and 6 (Table 6.4).

R141L is also in the dimer interface, forming a salt bridge with E156 (Goodwin et al., 2010). The mutation was isolated in a papillation screen for hyperactive transposases. However, it has not been further characterized, except in a HeLa assay where its activity is 33-fold down (Danxu Liu, unpublished data). In the papillation and mating-out assays, the mutant is hyperactive at all transposase concentration (Figure 6.5 and Table 6.4). However, the two assays produce slightly different results. In the mating out assay, activity was undiminished at high transposase concentration, whereas it was significantly diminished in the papillation assay, but probably not as much as the wild type transposase. OPI resistance was confirmed by

experiments with a ~ 700-copy vector, which provides very high expression levels (result not shown).

6.2.4 An Hsmar1, PiggyBac and Sleeping Beauty triple-transposon reporter strain

Because of the success of the papillation assay in characterizing Hsmar1 mutations, I wished to expand the range of the assay. I therefore constructed a new papillation strain containing the same reporter construct but flanked by Hsmar1, PiggyBac, and Sleeping Beauty (SB) transposon ends. The three transposons were nested, with SB on the outside and Hsmar1 on the inside (Figure 6.1). To test the new reporter strain, I performed a papillation assay with the complete set of expression vectors for Hsmar1 (Figure 6.6, whole plates are presented in Supplementary Figure 6). Results were similar to the single reporter but the colonies presented a blue background instead of white. This made the results more difficult to interpret.

To determine whether the PiggyBac and Sleeping Beauty transposases are also active, I performed a papillation assay with the respective transposase expressed from the Ip1 and/or Ip2 promoters (Figure 6.6, whole plates are presented in Supplementary Figure 7). Similar to the Hsmar1 transposase, the transposition rate of the PiggyBac transposase is higher with Ip1 than Ip2, showing that it is sensitive to OPI in a bacterial assay. In contrast, Sleeping Beauty transposase does not significantly increase the rate of papillation, consistent with its known restriction to vertebrate hosts.

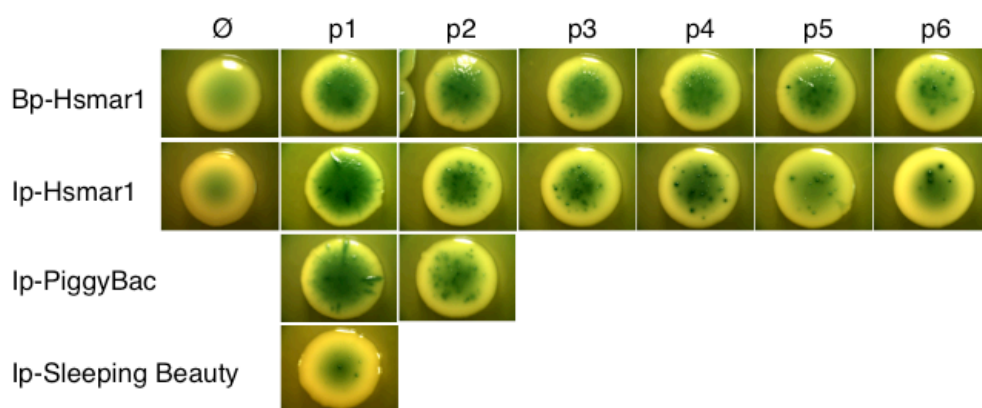


Figure 6-6: Characterization of the triple transposons papillation strain with Hsmar1, PiggyBac, and Sleeping Beauty transposases.

Representative colonies of the RC5118 strain transformed with plasmids encoding Hsmar1, PiggyBac, or Sleeping Beauty transposases. The Hsmar1 transposase has been tested with the 6 constitutive promoters cloned into pBACe3.6 and pGHM491. The PiggyBac and Sleeping Beauty transposases have only been tested with Ip1 plus Ip2 for PiggyBac. Pictures of the plates are presented in Supplementary Figures 6 and 7.

I also tested the reporter in the mating-out assay with the three transposases expressed from Ip1 (Table 6.5). At the optimal expression level, the Hsmar1 and PiggyBac transposases have a similar transposition rate, which is 10-fold higher than the one obtained for Ip1-Hsmar1 in the single-reporter strain RC5096 (Table 6.3). No transconjugants were retrieved with the Sleeping Beauty transposase, confirming that it is inactive in bacteria.

Construct	Transposition frequency
Ip1 Hsmar1	$4.45 (\pm 1.2) \times 10^{-4}$
Ip1 PiggyBac	$2.08 (\pm 0.4) \times 10^{-4}$
Ip1 Sleeping Beauty	0

Table 6-5: Transposition frequencies of different wild type transposases.

The bacterial mating-out assays have been done with the RC5122 strain and the Ip1 vector. Transposition frequencies are the average of three independent experiments $\pm$ standard error of the mean.

6.2.5 Two types of colonies are produced with the single copy expression vectors

An interesting phenomenon was observed in the papillation assays with the single-copy vectors Bp1 to Bp6. In contrast to the five-copy vectors, which produce a single type of colony, the same promoter in the single-copy vector produces two distinct populations of colonies, which I will call a 'normal' and a 'permissive' population (Figure 6.7A). The normal colonies are the ones presented in the Figure 6.4A and display the expected behaviour, where the transposition rate varies inversely with the promoter strength. The permissive colonies have significantly more papillae (Figure 6.7A). This is most clearly seen with Bp1 and Bp6, in which the normal colonies have the fewest number of papillae.

Since we have a mixture of two populations, I investigated whether the ratio of permissive colonies to the total number of colonies was constant and reproducible (Figure 6.7B). In four independent experiments, the ratio of permissive colonies is similar between biological replicates but varies with the

strength of the promoter. Except for Bp6, the ratio of permissive/normal colonies reflects their respective transposition rates.

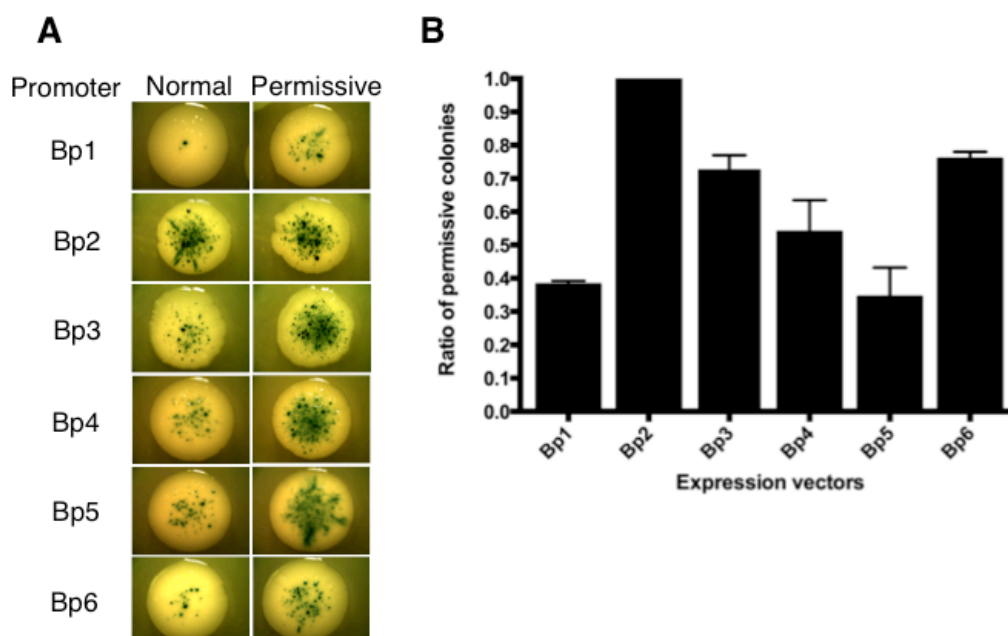
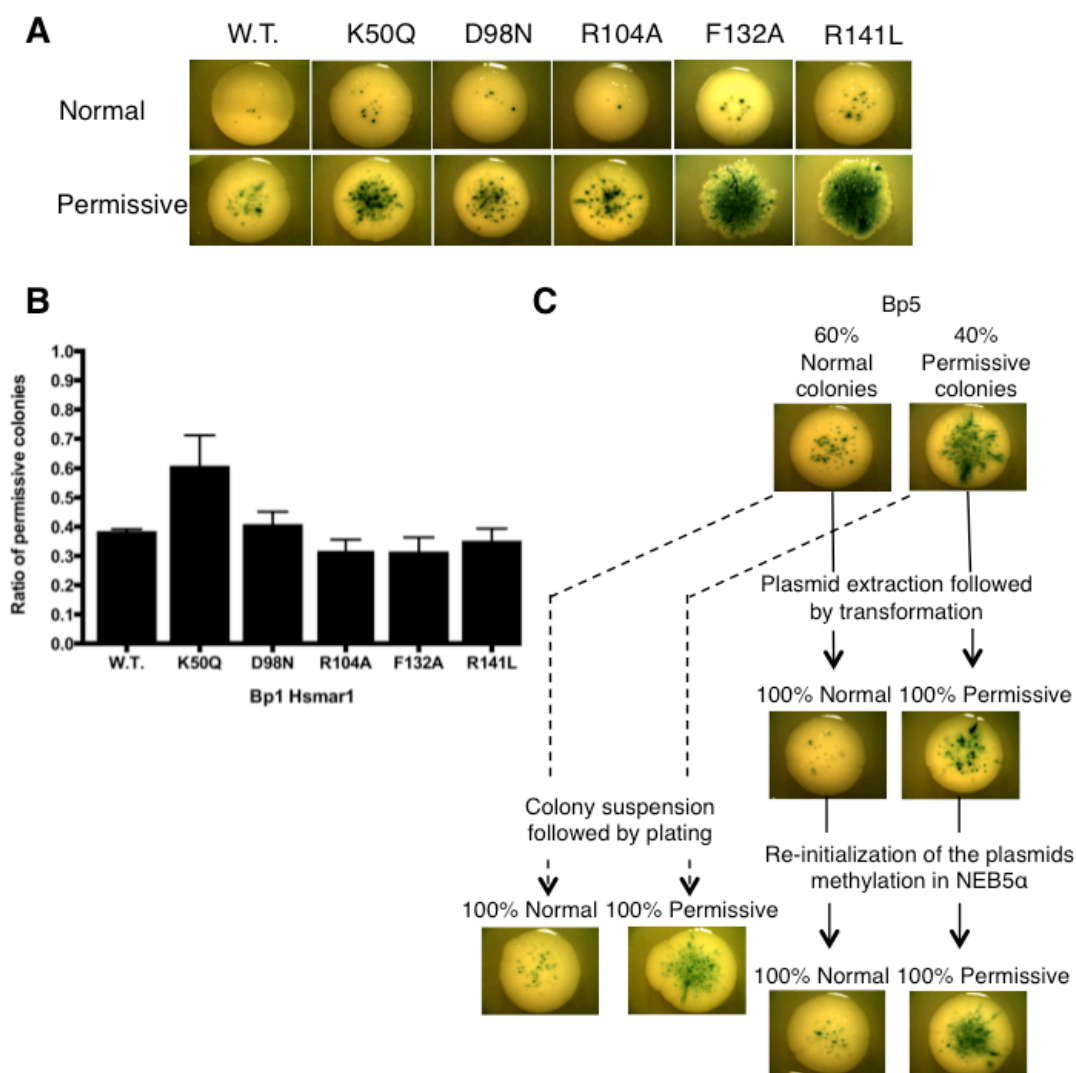


Figure 6-7: Characterization of the normal and permissive colonies.

(A) Comparison of representative normal and permissive colonies for the RC5096 strain expressing the Hsmar1 transposase from Bp1 to 6. Pictures of the plates are presented in Supplementary Figure 4. **(B)** The ratio of permissive colonies over the total number of colonies varies with the promoter strength. Average of four independent experiments $\pm$ standard deviation.

The two types of colonies are also observed with the Hsmar1 transposase mutants expressed from the Bp1 vector (Figure 6.8A). The number of papillae per permissive colony varies according to the hyperactivity of the mutation. The F132A and R141L mutants produce permissive colonies filled with papillae, whereas the D98N and R104A mutants transpose at an approximately similar rate as the wild type transposase. This variation in the

transposition rate confirms that the papillae observed in the permissive colonies are real transposition events.



(C) Normal and permissive colonies are respectively populations with a low or high transposition rate. The phenotypes are encoded by the plasmid and are independent of DNA methylation. Pictures of the plates are presented in Supplementary Figure 8.

I also investigated whether mutations in the transposase also affect the proportion of permissive colonies (Figure 6.8B). Except for K50Q there is no difference in the ratio of permissive colonies between the wild type and the mutant transposases.

To investigate the source of variability, I choose the Bp5 expression vector because it produces 60 and 40% of normal and permissive colonies, respectively, making it easier to recognize any differences in the distribution of the two populations (Figure 6.8C). The very large difference in the number of papillae for each type of colony makes them easier to score, with fewer ambiguities (see example of Bp5 in Figure 6.7A). To determine whether the phenotypes are vertically transmitted, I suspend a normal and a permissive colony in LB and plated it again on X-gal reporter plates (Figure 6.8C left panel). Some of the colonies were entirely blue. These were derived from the blue papillae in the original colony. The white colonies, which went on to papillate, were true to the phenotype of the parental colony. Normal colonies gave rise to normal colonies and permissive colonies gave rise to permissive colonies. The phenotype is therefore transmitted vertically.

To test whether the strain or the expression vector specify the normal and permissive phenotypes, I purified the plasmids from a normal and a permissive colony and performed a new papillation assay with both plasmids (Figure 6.8C, top right panel). This confirmed that the phenotype is encoded by the plasmids.

The two types of plasmids come from the same original stock. To determine whether the difference was due to a mutation in the transposase sequence or the promoter, I sequenced the transposase expression construct of both plasmids but did not find any mutation. To test whether differences were being maintained by passage through the papillation reporter strain, I transformed both plasmids in the NEB5 α strain. I then purified the respective plasmids and performed new papillation assays (Figure 6.8C, bottom right panel). However, the outcome remained unchanged and each plasmid produced only its own type of colony.

6.3 Discussion

6.3.1 Alteration of the transposition rate by the MBP tag in *E. coli*

It is quite surprising that a high number of papillae are visible with the MBP-transposase because the concentrations determined by western blotting are far above the OPI threshold (Figure 6.2A and B). It suggests that the MBP tag also affects the protein activity. From *in vitro* transposition assay, we know that two-third of the purified MBP-Hsmar1 transposase is active since three monomers per transposon are needed for an optimal transposition rate (Claeys Bouuaert et al., 2014). Even if a third of the cellular MBP-Hsmar1 transposase is inactive, it still does not account for all the difference observed with the untagged version. Several non-exclusive hypotheses can explain the effect of the MBP tag on the transposase activity. The presence of a 40 kDa

tag fused to each monomer could decrease the affinity of the transposase for its transposon ends, modify the synapsis rate, or hinder one the several structural changes required for transposition (Bouuaert et al., 2014). The MBP tag also influences the W118R mutant activity, which could partially explain the decrease of the mutants' transposition rate in HeLa cells where untagged transposases have been used (Table 6.1) (Liu and Chalmers, 2014).

6.3.2 In *E. coli*, the transposition rate peaks with a stochastic transcription and translation of the transposase

The p1 promoter is a feature-less track of GACT repeats and therefore contains no promoter or RBS. The number and timing of the mRNA production and the number of proteins produced per mRNA are therefore likely stochastic and different in each cell. However, the ProtParam tool estimates a half-life time of 10 hours for the Hsmar1 transposase in *E. coli* so the transposase concentration will increase over time. It is interesting to note that one copy of the p1-Hsmar1 construct per cell does not produce enough proteins for transposition to happen more than a few times in 10^9 cells whereas five copies of this construct is sufficient for obtaining nearly 350 successful transposition events in the same number of cells (Figure 6.4A).

Even a very weak promoter present in one copy per cell (Bp2) produces enough transposase for OPI to occur (Figure 6.4B). Even though the synthesis of an mRNA molecule is a rare event, the mRNA will produce enough transposases over time to observe a decrease of the transposition rate. Thus, OPI is even stronger when the p2-promoter is present in 5 copies

per cell and where a 12-fold decrease of the transposition rate is observed (Figure 6.4B).

6.3.3 Mutations in the DNA binding domain and OPI-resistance

According to the ASO model, the D98N mutation is expected to increase the affinity of the transposase domain for the transposon end because of the loss of a negative charge. However, no significant change has been observed with the papillation and mating-out assays (Figure 6.5 and Table 6.4). All the anthropoid primates with an available DNA sequence of SETMAR transposase domain possess this substitution therefore the copy of the transposase gene that integrates downstream of the SET gene already contained this mutation (Cordaux et al., 2006). However, the wild type Hsmar1 transposase used in this Chapter corresponds to the reconstructed ancestral (RA) sequence (Miskey et al., 2007). The RA sequence has been generated by aligning the Hsmar1 transposon remnants found in the human genome. Therefore, the D98 in the RA sequence could be the result of genetic drift and/or miscalculation. This is supported by the fact that Mos1 and Himar1, two other *mariner* transposases, have a serine, another polar residue, at this position.

The R104A mutation is in the second helix-turn-helix motif and an EMSA competition assay demonstrated that its affinity for the transposon end is reduced substantially (Claeys Bouuaert et al., 2013). Furthermore, the mutant was fully active in an *in vitro* reaction using a plasmid substrate and was also

completely resistant to OPI. It is therefore surprising that the mutant is barely active in the papillation assay while it is active in HeLa cells (Figure 6.5) (Claeys Bouuaert et al., 2013).

6.3.4 Decreasing the transposase dimer stability produces bacterial hyperactive mutants

Above the OPI threshold, the transposition rate decreases because of the probability of having a transposase dimer on each transposon end increases (Claeys Bouuaert et al., 2013). Destabilizing the dimer interface can therefore potentially generate hyperactive and OPI-resistant transposases. The F132 and R141 residues are located in the subunit interface and provide hyperactive transposases (see Introduction section 1.2.6) and may be important in one of the transposition intermediates. The hyperactivity of both mutants could be explained by the promotion of one of the conformational changes essential for transposition (Bouuaert et al., 2014). However, further investigation is needed to characterize the mechanism responsible for the hyperactivity of both mutants.

The OPI-resistance of the F132A and R141L mutations could be explained by two complementary mechanisms. Above the OPI threshold, a decrease in the dimer stability will shift the monomer-dimer equilibrium towards the monomeric form, decreasing the concentration of active transposases in the cell. Also, a less stable dimer bound to a transposon end may be more likely to fall apart allowing the recruitment of the previously bound end by another dimer. The strength of both mechanisms should be

proportional to the decrease of the dimer stability. Indeed, the R141 amino acid seems to be a key residue since its mutation abolishes the OPI sensitivity of the transposase in *E. coli* (Figure 6.5 and Table 6.4). This mutant conserved a similar transposition rate with the optimal (Ip1), strong (Ip6), and very strong (pcDNA3.1-p6) expression vectors. In comparison, the wild type transposase is nearly and completely inactive with respectively the strong and very strong vectors. Even though the F132A mutant is still sensitive to OPI as demonstrated by the 3-fold decrease in the transposition rate between Ip1 and Ip6, its greater OPI-resistance is clear when compared to the ~100-fold decrease of the wild-type transposase with the same expression points.

This type of mutant is more likely to be found hyperactive only in bacteria. Indeed, in mammalian cells the size of the nucleus and the bigger ratio of non-specific DNA to specific DNA, i.e. the transposon ends account for a smaller fraction of the genome, will increase the time necessary for a transposase to find a transposon end. Therefore, transposases with a weakened dimer interface are more likely to revert to an inactive monomeric state during this period of time resulting in hypoactive mutants. The R141L mutant illustrates this phenomenon by being hyperactive in *E. coli* but inactive in HeLa cells (Danxu Liu, unpublished data).

6.3.5 Comparison of the single and triple transposon ends constructs

The papillation assays presented in this Chapter have been done with two different strains, RC5096 and RC5118. Whereas the single reporter (RC5096)

produced a white colony on X-gal reporter plates, the triple reporter (RC5118) had a significant blue background (Figure 6.6). A likely explanation is adventitious transcription arising from the additional SB and PB binding sites at the 5' end of the promoter-less *lacZ* gene. This is supported by the fact that colonies expressing *lacZ* from my p1 construct are blue even though there is only one copy of the *lacZ* gene per cell and it is without any promoter sequence and RBS motif to aid expression (result not shown). Therefore, a few molecules of *lacZ* produced in each cell are sufficient for a colony to appear blue.

In HeLa cells, the PiggyBac transposase is 10 times more active than Hsmar1 whereas in the papillation and mating-out assays both transposases have a similar transposition rate (Table 6.3 and 6.5) (Claeys Bouuaert et al., 2013). The 10-fold increase in the transposition rate of the Hsmar1 transposase in the RC5118 strain is the most plausible explanation. Indeed, it is supported by the 10-fold increase in the transposition rate observed for Ip1-Hsmar1 in the triple reporter strain compared to its single counterpart. However, the mechanism responsible for this increased transposition rate is unknown. I also cannot exclude an effect on the transposition rate by the biological differences between bacterial and human cells such as the absence of the nucleus or the concurrent transcription and translation of genes.

Even though Sleeping Beauty is a member of the same superfamily of transposons as Hsmar1 (ITm), it is not active in *E. coli* (Figure 6.6 and Table 6.5). Sleeping Beauty is known to interact with several proteins in vertebrates therefore its inactivity could be explained by the absence of one or more cofactors in bacterial cells (Wang et al., 2014).

6.3.6 Permissive colonies with the single-copy expression vectors

From several experiments, I determined that the permissive phenotype is encoded by the transposase expression vectors (Figure 6.8C). Because the transposition rate is what differs between normal and permissive colonies, the mechanism responsible for both phenotypes must alter the transposase expression. I already know that the transposase gene and the promoter region are not mutated. Consequently, if a mutation or a DNA rearrangement is the source of the permissive phenotype, it should be located near the transposase construct or alter the vector copy number.

A comparison between the normal and permissive colonies showed that both types of colonies have a similar number of papillae with the Bp2 vector, which provides the maximum rate of transposition (Figure 6.7A). The p2 promoter has been deactivated by several point mutations and provides stochastic transcription, which should produce only a few transposases per cell (Alper et al., 2005). Consequently, the high transposition rate produces a saturating number of papillae preventing a visual distinction between normal and permissive colonies. I consider that all the colonies are permissive because it fits with the decreasing ratio of permissive colonies observed for Bp3 to 5 (Figure 6.7B). However, I cannot exclude that all the Bp2 colonies are normal or a mixture of both types.

The ratio of permissive colonies is conserved between biological replicates and most of the transposase mutants, but varies with the promoter

strength (Figure 6.7B and 6.8B). Except for Bp6, the trend in the ratio of permissive/normal colonies matches the transposition rate with a ratio inversely proportional to the promoter strength. The fact that Bp6 has a higher ratio could reflect the breakdown in the relationship at the very high transposase expression provided by this promoter. However, Bp1 and Bp6 permissive colonies produce less papillae than the Bp2 to Bp5 colonies. This difference may therefore be linked to the difference in the ratio of permissive colonies.

The decrease in the number of papillae per permissive colony for Bp1 and Bp6 may be explained by the respective weakness and strength of each promoter. For Bp1, the transposase concentration is too low to produce as many papillae as Bp2 whereas for Bp6, the concentration is too high and OPI is already affecting the transposition rate significantly. The major problem with this explanation, linking the phenomenon to OPI, is that it requires an increase in transposase expression from Bp1 but a decrease in the expression of Bp2 to Bp6. Furthermore, in the case of Bp6, it requires that the decrease must be insufficient to keep a high transposition rate. However, the amount of transposase produced by Bp2 to Bp5 increases with the promoter strength therefore the mechanism should also increase its inhibiting effect to keep a similar transposition rate for each expression vector. This seems unlikely because of the wide difference in promoter strength between p2 or p3 and p5 (Figure 6.3A).

Another outlier in the ratio of permissive colonies is the K50Q mutant. Compared to the wild type transposase, the K50Q mutation increases the ratio by 1.5-fold (Figure 6.8B). This observation hints to a possible role of the

transposase itself in addition to its expression in the emergence of the permissive phenotype. The hyperactivity of the K50Q mutation with Bp1 in the papillation assay does not explain the change in the ratio of permissive colonies because the hyperactive F132A and R141L mutants have a similar ratio as the wild type transposase. Similarly, it also cannot be explained by a mutation in the DNA binding domain since the D98N and R104A mutations did not affect the ratio. Thus, the K50Q mutation has a direct but unknown role in the permissive phenotype.

6.3.7 Variant of the papillation assay

The current transposition assays in mammalian cells use a strong promoter for expressing the transposase (Liu and Chalmers, 2014) (Wang et al., 2014). The high transposase concentration can decrease the transposition rate because of the effects of OPI (Claeys Bouuaert et al., 2013). We know that mutations in the DNA binding domain decreasing the affinity of the transposase towards the transposon ends increase the transposition rate above the OPI threshold (Claeys Bouuaert et al., 2013). Another way to obtain a similar effect without altering the transposase sequence would be to mutate the transposon ends to decrease the transposase affinity.

The current papillation assay can be easily modified to test a library of transposon end mutations to determine which nucleotides could be modified to partially decrease the affinity of the transposase for the transposon ends. The *lacZ* reporter gene flanked by one or both mutated transposon ends can be cloned into pBACe3.6 to provide a single transposon per cell. A

comparison of the transposition rates between the original and mutated transposon ends with the Ip6 expression vector encoding the wild type transposase would inform us on the strength of the mutation(s) on the transposase affinity towards the transposon ends. This approach is complementary to the search for hyperactive variants and offers the advantage to conserve the transposase auto-regulation.

7. CONCLUSION AND PERSPECTIVES

7.1 Summary of the research

The major part of my project was to determine how the functions of the SET gene have been transformed upon its fusion with the Hsmar1 transposase gene. The SET gene is specific to bilaterian animals and first appeared in the lineage 570-550 Mya (Shen et al., 2008). From the data presented in Chapter 5, one of the functions of the SET gene, at least in the ancestor of the anthropoid primates, might have been associated to NHEJ. However, its role seems to have promoted DNA integration rather than DNA repair (Chapter 5). The mechanism by which DNA integration is favoured is still unknown but should rely on histone methylation or protein-protein interactions.

The anthropoid and non-anthropoid primates have been invaded independently by the Hsmar1 transposon ~65-55 Mya {Pace, 2008 #68}. In the modern anthropoid primates several thousand Hsmar1 remnants are still recognizable. In addition to the full-length and partial transposons, which represent about 9% of the remnants, another ~8,000 are divided between Made1, an 80 bp miniature Hsmar1, and solo ITRs, many of which are a deletion-derivatives of Made1 (Chapter 3). A fraction of the Hsmar1 remnants have been co-opted over time by the host genome. Several Hsmar1 transposons and Made1 elements are now part of lncRNAs and provide exonic sequences, splicing sites, and polyadenylation signals (Kapusta et al., 2013). In addition to the 69 Made1 elements already reported to have been co-opted as putative miRNAs genes (Piriyapongsa and Jordan, 2007), I identified 456 additional candidate miRNAs (Chapter 3). The most significant

Hsmar1 exaptation is SETMAR, a chimeric gene resulting from the fusion of an Hsmar1 transposase gene on the C-terminus of the SET gene (Cordaux et al., 2006). The fusion event happened 40 to 58 Mya and is specific to the anthropoid primates.

The addition of the Hsmar1 transposase on the C-terminus of the SET domain presumably modified its function. Indeed, in contrast to the published data (Lee et al., 2005), I found that SETMAR lacks the DNA integration activity of the isolated SET domain (Chapter 5). Since characterized histone methyltransferase proteins function as monomers, this loss of function may be explained by the dimerization of SETMAR through the transposase domain (Chapter 5).

The main hypothesis addressed in this thesis is that the transposase DNA binding domain targets the histone methylase to specific sites in the genome. This is supported by the fact that the DNA binding domain is under purifying selection, while the transposase catalytic domain is drifting (Chapter 3) (Cordaux et al., 2006). I tested the hypothesis by using RNA-Seq to determine which genes are differentially expressed when SETMAR is overexpressed at different levels (Chapter 4). I observed an enrichment in the differentially expressed genes for the genes containing an ITR but not for the genes near an ITR. Half of the 1,523 protein-coding genes with an ITR are differentially expressed in at least one cell line. However, SETMAR binding sites, which I determined using ChIP-Seq, overlap only weakly with the list of differentially expressed genes (Chapter 4). Essentially, SETMAR binds to Made1 elements and to the miRNAs-like defined in Chapter 3 (Chapter 4).

The current results suggest that SETMAR could be involved in gene regulation by two different mechanisms. SETMAR may directly regulate the expression of the miRNAs-like, if they are active, which could in turn regulate their target genes. Since the genes containing an ITR are enriched in the differentially expressed genes and in the genes with a ChIP peak, SETMAR or the H3K36me2 mark could target interacting partners or H3K36me2 readers to the genes with an ITR to regulate their expression.

7.2 Outlook

A significant amount of work is still needed to obtain a clear picture of SETMAR functions in anthropoid primates and its connections with the other Hsmar1 remnants. The preliminary bioinformatics work presented in the Chapter 3 suggests that several ITRs are under purifying selection. As more anthropoid primate genomes enter the sequence database it will provide further information to determine which Hsmar1 remnants and ITRs are under purifying selection.

An important question is to determine whether SETMAR requires interacting partners to regulate the expression of the genes with an ITR and whether the H3K36me2 mark is sufficient in itself to recruit other proteins or change gene expression. In both cases, the mechanism leading to gene regulation needs to be characterized. If SETMAR interacts with other proteins, it will be interesting to determine which SETMAR domains are involved in the interactions. The outcome should suggest whether gene regulation is an emergent function of SETMAR provided by the transposase domain or an

ancestral function of the SET domain, which has been rewired upon its fusion to the Hsmar1 transposase.

SETMAR may regulate the expression of the miRNAs-like by binding to their ITRs and dimethylating the H3K36 of the nucleosomes in its vicinity. The RNA-Seq experiments presented in Chapter 4 were performed with polyadenylated mRNAs and therefore provide no information on the expression level of these genes. The miRNAs-like are annotated as miRNAs by the GENCODE project. However, their cell-type specific expressions or their target genes have not been characterized.

Several lncRNAs contain a full length Hsmar1 transposon. Although these all contain frame shift and point mutations in the transposase gene, they provide exonic sequence, splicing sites, and poly(A) signals (Kapusta et al., 2013). It is currently unknown if the Hsmar1 sequences recruited as lncRNAs exon are important for the lncRNAs functions in the cell.

When Hsmar1 transposase was fused to the 3'-end of the SET gene its 3'-UTR became the second intron. The fusion-gene then acquired the Hsmar1 3'-UTR (Cordaux et al., 2006). This could have several effects on SETMAR regulation. In addition to a possible change in the mRNA stability, the putative post-transcriptional regulation by miRNAs could also have been modified with the loss of the pre-fusion miRNAs target sites. There are only five substitutions and one nucleotide deletion between the Hsmar1 and SETMAR 3'-UTR. Furthermore, several Hsmar1 transposons are now part of long non-coding RNAs (lncRNAs) where they provide the poly(A) signal and therefore the 3'UTR. Thus, it could be interesting to investigate which regulations by miRNAs have been lost and gained upon the fusion and whether SETMAR

and the lncRNAs with an Hsmar1 transposon are co-regulated by similar small RNAs.

SETMAR encodes two isoforms in the U2OS cell line (Figure 5.7A and Supplementary Figure 2). The first isoform, which represents nearly all the transcripts in the RNA-Seq data, code the full-length protein whereas the second isoform, which accounts for a few per cent, is a variant containing a large deletion in the SET domain abolishing its histone methyltransferase activity (Figure 5.7A). The study of the second isoform of SETMAR should shed some light on the evolution of the gene following the fusion event. Although the junctions of this splice isoform are contained entirely within the SET domain, this variant is only detected in the anthropoid primates containing the fusion gene. It may be interesting to ask whether the two isoforms form mixed multimers and whether this affects their activity in the NHEJ pathway.

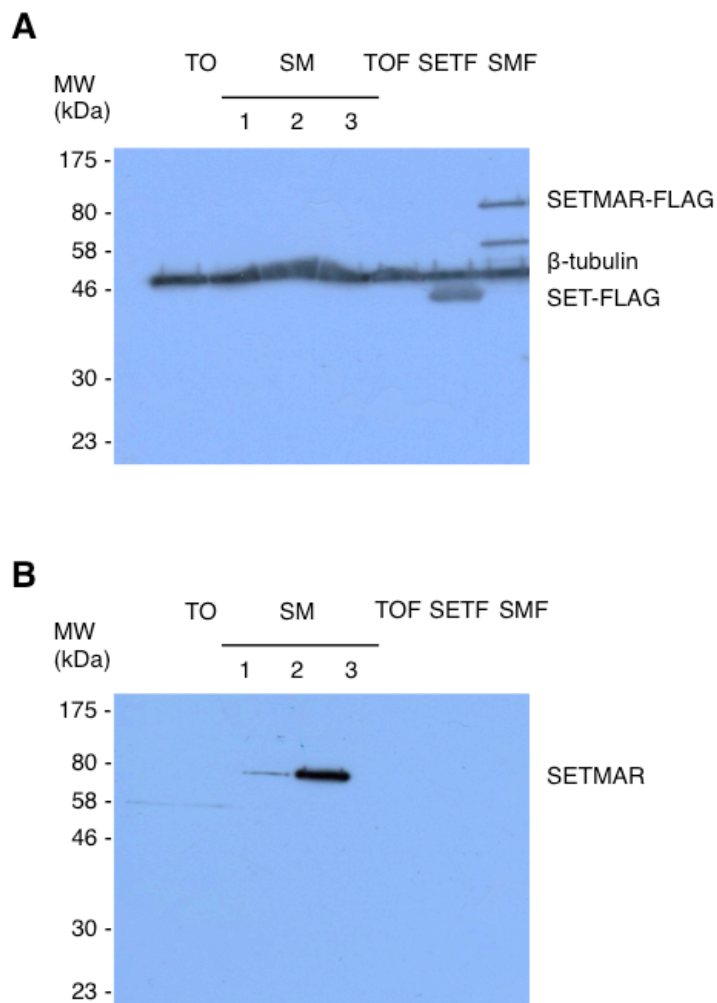
The crystal structure of the full-length SETMAR should provide some insight on the effect of the transposase domain on the location of the two SET domains towards each other. In addition, a crystal structure of SETMAR bound to an ITR should help to determine how the SET domains can dimethylate the H3K36 of the nucleosomes in its vicinity.

Future research is needed to gain a better understanding of the ancestral role of the SET gene in DNA repair and integration. The repair products of compatible and incompatible ends of each cell lines should be analyzed to determine whether SETMAR, or one of its domains, affects the accuracy of NHEJ. A recent paper concluded that a strong overexpression of SETMAR increases the inaccuracy of DNA repair (Rath et al., 2014).

The mechanism by which the SET domain promotes DNA integration over DNA repair is unclear. To determine whether DNA integration is dependent on histone methylation, a SET gene deficient for its methylation activity could be expressed and tested for DNA repair and integration.

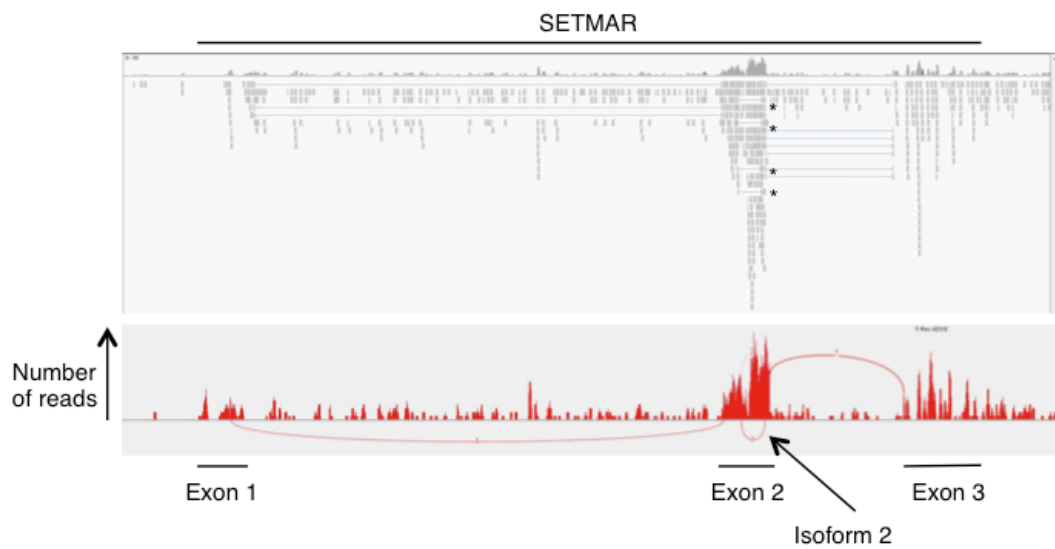
8. APPENDICES

Supplementary Figures



Supplementary Figure 1: Full pictures of the Western blot presented in Figure 4.1.

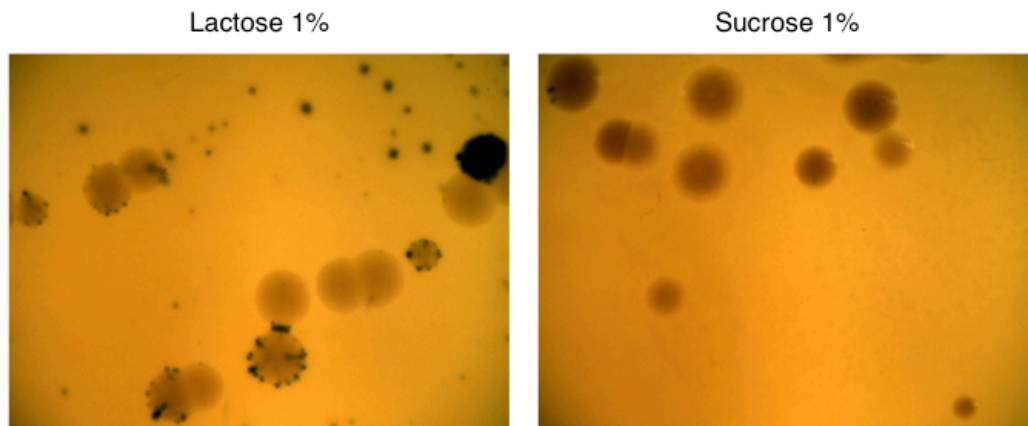
Expression of SETMAR in the different T-Rex-U2OS cell lines used for RNA-Seq and CHIP-exo. **(A)** The SET domain and SETMAR protein were detected with a monoclonal antibody against the FLAG tag. **(B)** The SETMAR protein was detected with an antibody raised against the last 9 amino acids of the native protein.



Supplementary Figure 2: SETMAR encodes two isoforms in the U2OS cell line.

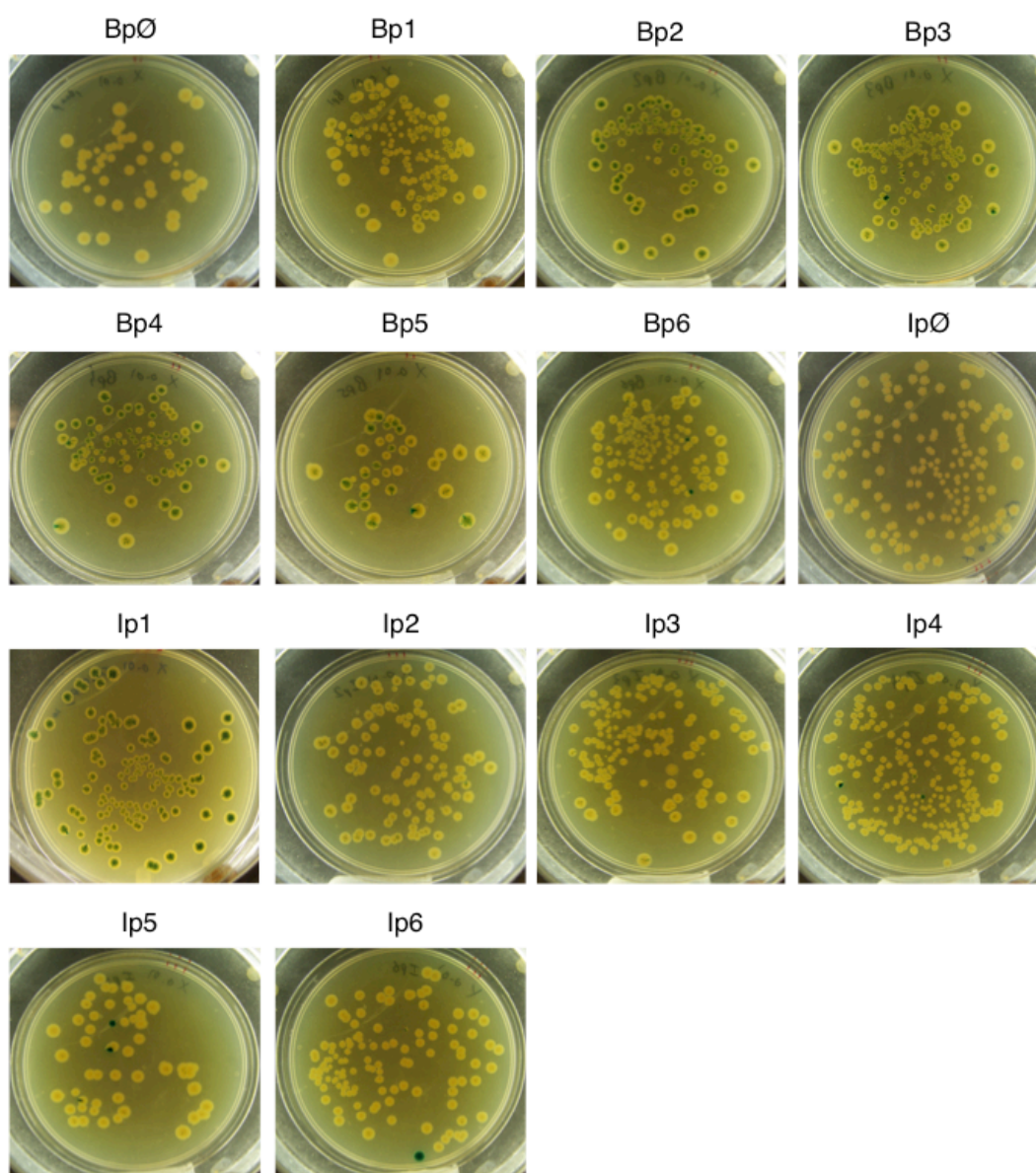
Reads mapping to SETMAR in one of the U2OS cell line used in the RNA-Seq analysis. The reads with exon-exon junction in the second exon of SETMAR correspond to the second isoform and are indicated with an asterisk (*).

M9 media



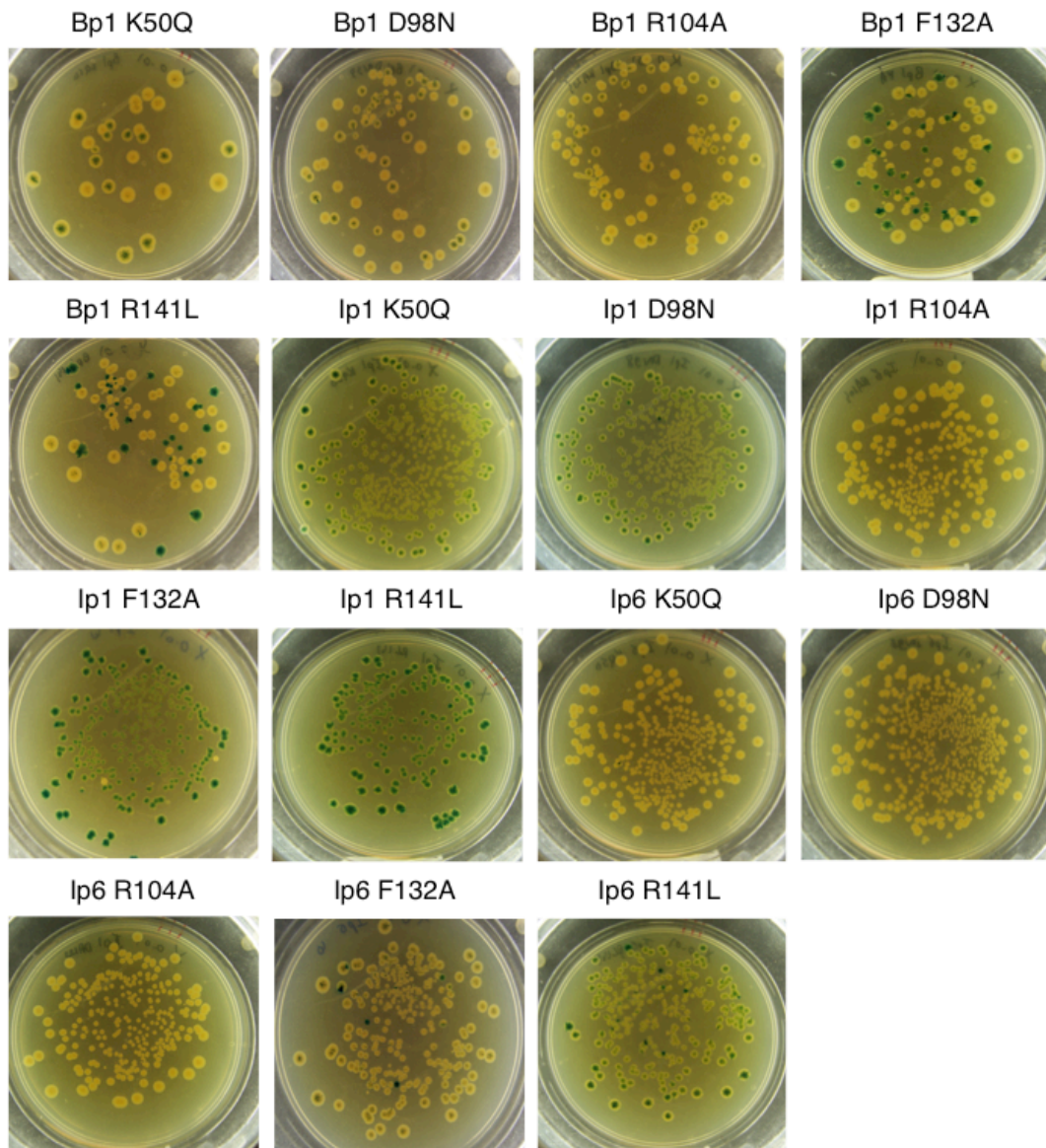
Supplementary Figure 3: Papillation assay on minimal media in the RC5096 strain.

Representatives field of view of the colonies presented in Figure 6.2B.



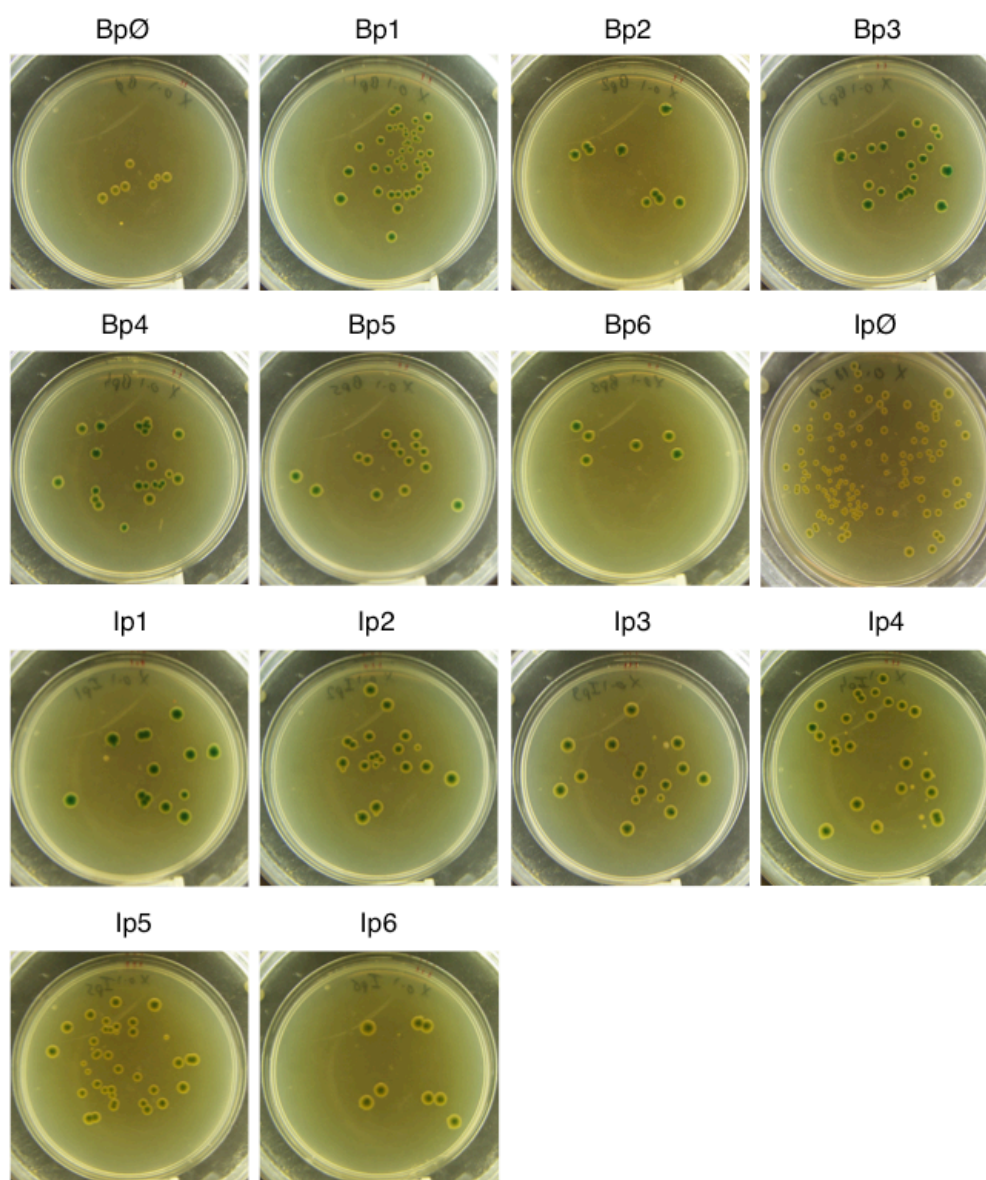
Supplementary Figure 4: Papillation assay with the wild type Hsmar1 transposase in the RC5096 strain.

The entirety of the plates from Figure 6.3A is shown.



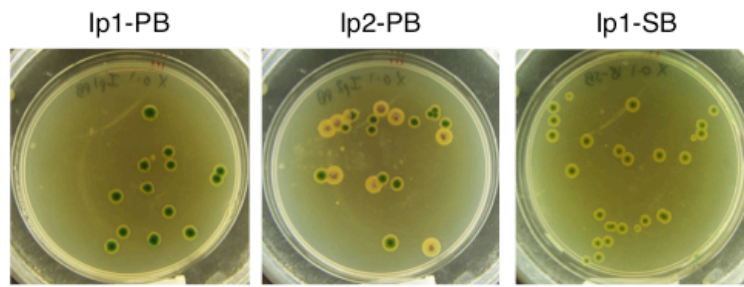
Supplementary Figure 5: Papillation assay with Hsmar1 transposase mutants in the RC5096 strain.

The entirety of the plates from Figure 6.5A is shown.



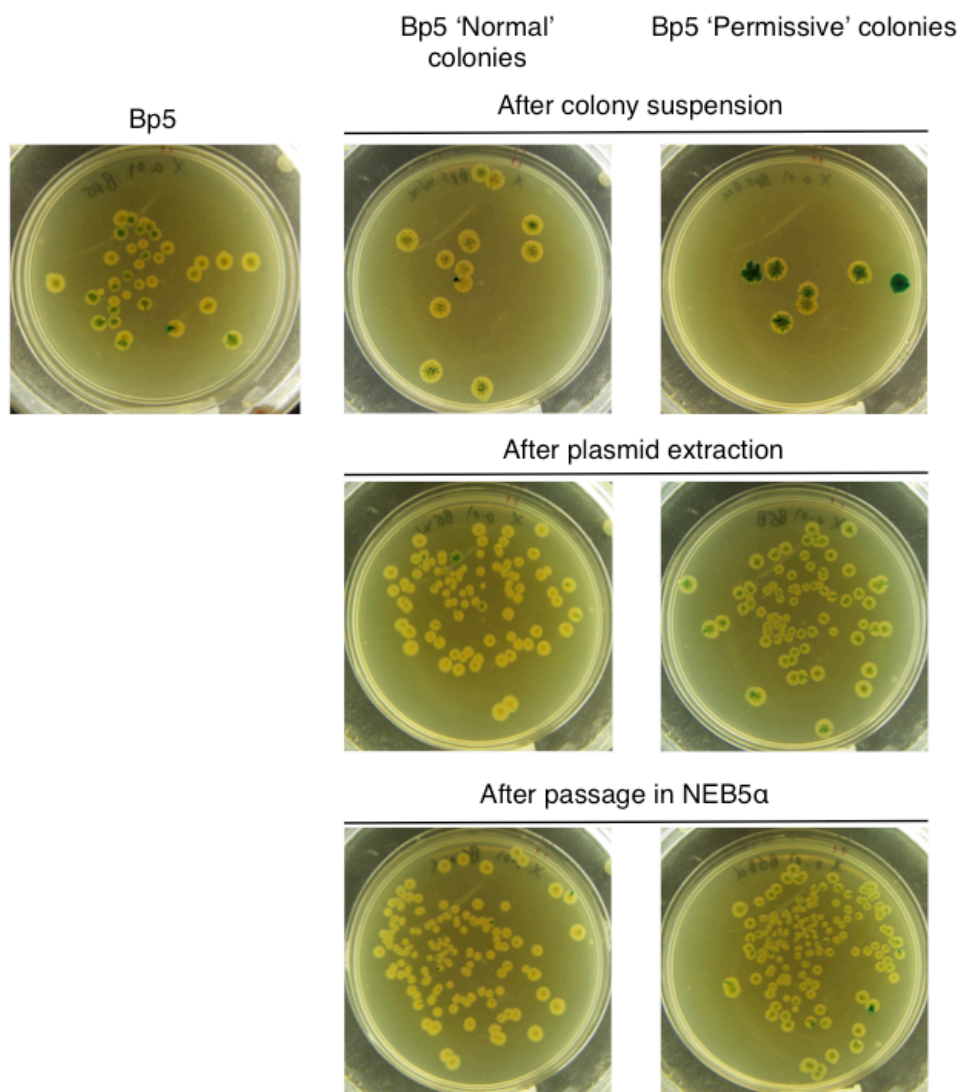
Supplementary Figure 6: Papillation assay with the wild type Hsmar1 transposase in the RC5118 strain.

The entirety of the Hsmar1 transposase plates from Figure 6.6 is shown.



Supplementary Figure 7: Papillation assay with the wild type PiggyBac and Sleeping Beauty transposase in the RC5118 strain.

The entirety of the plates for the PiggyBac (PB) and Sleeping Beauty (SB) transposases from Figure 6.6 is shown.



Supplementary Figure 8: Characterization of normal and permissive colonies.

The entirety of the plates for the normal and permissive colonies from Figure 6.8C is shown.

Bibliography

Abitbol, M., Thibaud, J.L., Olby, N.J., Hitte, C., Puech, J.P., Maurer, M., Pilot-Storck, F., Hedan, B., Dreano, S., Brahimi, S., *et al.* (2010). A canine Arylsulfatase G (ARSG) mutation leading to a sulfatase deficiency is associated with neuronal ceroid lipofuscinosis. *Proc Natl Acad Sci U S A* 107, 14775-14780.

Albert, I., Wachi, S., Jiang, C., and Pugh, B.F. (2008). GeneTrack--a genomic data processing and visualization framework. *Bioinformatics* 24, 1305-1306.

Alper, H., Fischer, C., Nevoigt, E., and Stephanopoulos, G. (2005). Tuning genetic control through promoter engineering. *Proc Natl Acad Sci U S A* 102, 12678-12683.

Alzohairy, A.M., Gyulai, G., Jansen, R.K., and Bahieldin, A. (2013). Transposable elements domesticated and neofunctionalized by eukaryotic genomes. *Plasmid* 69, 1-15.

Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.M., Wilkes, C.D., Garnier, O., Labadie, K., Lauderdale, B.E., Le Mouel, A., *et al.* (2012). The *Paramecium* germline genome provides a niche for intragenic parasitic DNA: evolutionary dynamics of internal eliminated sequences. *PLoS Genet* 8, e1002984.

Aziz, R.K., Breitbart, M., and Edwards, R.A. (2010). Transposases are the most abundant, most ubiquitous genes in nature. *Nucleic Acids Res* 38, 4207-4217.

Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A.E. (2005). Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* 122, 553-563.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215-233.

Baudry, C., Malinsky, S., Restituto, M., Kapusta, A., Rosa, S., Meyer, E., and Betermier, M. (2009). PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate *Paramecium tetraurelia*. *Genes Dev* 23, 2478-2483.

Beck, B.D., Lee, S.S., Williamson, E., Hromas, R.A., and Lee, S.H. (2011a). Biochemical characterization of metnase's endonuclease activity and its role in NHEJ repair. *Biochemistry* 50, 4360-4370.

Beck, B.D., Park, S.J., Lee, Y.J., Roman, Y., Hromas, R.A., and Lee, S.H. (2008). Human Pso4 is a metnase (SETMAR)-binding partner that regulates metnase function in DNA repair. *J Biol Chem* 283, 9023-9030.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. *Cell* 141, 1159-1170.

Beck, M., Schmidt, A., Malmstroem, J., Claassen, M., Ori, A., Szymborska, A., Herzog, F., Rinner, O., Ellenberg, J., and Aebersold, R. (2011b). The quantitative proteome of a human cell line. *Mol Syst Biol* 7, 549.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74.

Borchert, G.M., Holton, N.W., Williams, J.D., Hernan, W.L., Bishop, I.P., Dembosky, J.A., Elste, J.E., Gregoire, N.S., Kim, J.A., Koehler, W.W., *et al.* (2011). Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. *Mob Genet Elements* 1, 8-17.

Bouchet, N., Jaillet, J., Gabant, G., Brillet, B., Briseno-Roa, L., Cadene, M., and Auge-Gouillou, C. (2014). cAMP protein kinase phosphorylates the Mos1 transposase and regulates its activity: evidences from mass spectrometry and biochemical analyses. *Nucleic Acids Res* 42, 1117-1128.

Bouuaert, C.C., Tellier, M., and Chalmers, R. (2014). One to rule them all: A highly conserved motif in mariner transposase controls multiple steps of transposition. *Mob Genet Elements* 4, e28807.

Breitling, R., and Gerber, J.K. (2000). Origin of the paired domain. *Dev Genes Evol* 210, 644-650.

Bundock, P., and Hooykaas, P. (2005). An Arabidopsis hAT-like transposase is essential for plant development. *Nature* 436, 282-284.

Carpentier, C.E., Schreifels, J.M., Aronovich, E.L., Carlson, D.F., Hackett, P.B., and Nesmelova, I.V. (2014). NMR structural analysis of Sleeping Beauty transposase binding to DNA. *Protein Sci* 23, 23-33.

Chalmers R, B.M. (1999). Insertion Sequences and Transposons. In *Organization of the Prokaryotic Genome*, C.R. (ed.), ed. (Washington, D.C., American Society for Microbiology), pp. 151-169.

Chao, L., and McBroom, S.M. (1985). Evolution of transposable elements: an IS10 insertion increases fitness in *Escherichia coli*. *Mol Biol Evol* 2, 359-369.

Chao, L., Vargas, C., Spear, B.B., and Cox, E.C. (1983). Transposable elements as mutator genes in evolution. *Nature* 303, 633-635.

Cipriano, R., Graham, J., Miskimen, K.L., Bryson, B.L., Bruntz, R.C., Scott, S.A., Brown, H.A., Stark, G.R., and Jackson, M.W. (2012). FAM83B mediates EGFR- and RAS-driven oncogenic transformation. *J Clin Invest* 122, 3197-3210.

Cipriano, R., Miskimen, K.L., Bryson, B.L., Foy, C.R., Bartel, C.A., and Jackson, M.W. (2013). FAM83B-mediated activation of PI3K/AKT and MAPK signaling cooperates to promote epithelial cell transformation and resistance to targeted therapies. *Oncotarget* 4, 729-738.

Claeys Bouuaert, C., and Chalmers, R. (2010). Transposition of the human Hsmar1 transposon: rate-limiting steps and the importance of the flanking TA dinucleotide in second strand cleavage. *Nucleic Acids Res* 38, 190-202.

Claeys Bouuaert, C., Lipkow, K., Andrews, S.S., Liu, D., and Chalmers, R. (2013). The autoregulation of a eukaryotic DNA transposon. *Elife* 2, e00668.

Claeys Bouuaert, C., Liu, D., and Chalmers, R. (2011). A simple topological filter in a eukaryotic transposon as a mechanism to suppress genome instability. *Mol Cell Biol* 31, 317-327.

Claeys Bouuaert, C., Walker, N., Liu, D., and Chalmers, R. (2014). Crosstalk between transposase subunits during cleavage of the mariner transposon. *Nucleic Acids Res* 42, 5799-5808.

Cordaux, R., Udit, S., Batzer, M.A., and Feschotte, C. (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A* 103, 8101-8106.

Cuypers, M.G., Trubitsyna, M., Callow, P., Forsyth, V.T., and Richardson, J.M. (2013). Solution conformations of early intermediates in Mos1 transposition. *Nucleic Acids Res* 41, 2020-2033.

Davies, D.R., Goryshin, I.Y., Reznikoff, W.S., and Rayment, I. (2000). Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. *Science* 289, 77-85.

De Haro, L.P., Wray, J., Williamson, E.A., Durant, S.T., Corwin, L., Gentry, A.C., Osheroff, N., Lee, S.H., Hromas, R., and Nickoloff, J.A. (2010). Metnase promotes restart and repair of stalled and collapsed replication forks. *Nucleic Acids Res* 38, 5681-5691.

Decottignies, A. (2007). Microhomology-mediated end joining in fission yeast is repressed by pku70 and relies on genes involved in homologous recombination. *Genetics* 176, 1403-1415.

Dou, Y., Milne, T.A., Ruthenburg, A.J., Lee, S., Lee, J.W., Verdine, G.L., Allis, C.D., and Roeder, R.G. (2006). Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nat Struct Mol Biol* 13, 713-719.

Duval-Valentin, G., and Chandler, M. (2011). Cotranslational control of DNA transposition: a window of opportunity. *Mol Cell* 44, 989-996.

Emmons, S.W., Yesner, L., Ruan, K.S., and Katzenberg, D. (1983). Evidence for a transposon in *Caenorhabditis elegans*. *Cell* 32, 55-65.

Fattah, F., Lee, E.H., Weisensel, N., Wang, Y., Lichter, N., and Hendrickson, E.A. (2010). Ku regulates the non-homologous end joining pathway choice of DNA double-strand break repair in human somatic cells. *PLoS Genet* 6, e1000855.

Fedoroff, N. (2001). How jumping genes were discovered. *Nat Struct Biol* 8, 300-301.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9, 397-405.

Feschotte, C., Osterlund, M.T., Peeler, R., and Wessler, S.R. (2005). DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. *Nucleic Acids Res* 33, 2153-2165.

Fnu, S., Williamson, E.A., De Haro, L.P., Brenneman, M., Wray, J., Shaheen, M., Radhakrishnan, K., Lee, S.H., Nickoloff, J.A., and Hromas, R. (2011). Methylation of histone H3 lysine 36 enhances DNA repair by nonhomologous end-joining. *Proc Natl Acad Sci U S A* 108, 540-545.

Franceschini, A., Meier, R., Casanova, A., Kreibich, S., Daga, N., Andritschke, D., Dilling, S., Ramo, P., Emmenlauer, M., Kaufmann, A., *et al.* (2014). Specific inhibition of diverse pathogens in human cells by synthetic microRNA-like oligonucleotides inferred from RNAi screens. *Proc Natl Acad Sci U S A* 111, 4548-4553.

Gao, L., Sun, C., Qiu, H.L., Liu, H., Shao, H.J., Wang, J., and Li, W.X. (2004). Cloning and characterization of a novel human zinc finger gene, hKid3, from a C2H2-ZNF enriched human embryonic cDNA library. *Biochem Biophys Res Commun* 325, 1145-1152.

Goodwin, K.D., He, H., Imasaki, T., Lee, S.H., and Georgiadis, M.M. (2010). Crystal structure of the human Hsmar1-derived transposase domain in the DNA repair enzyme Metnase. *Biochemistry* 49, 5705-5713.

Gotea, V., and Makalowski, W. (2006). Do transposable elements really contribute to proteomes? *Trends Genet* 22, 260-267.

Gray, L.T., Fong, K.K., Pavelitz, T., and Weiner, A.M. (2012). Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. *PLoS Genet* 8, e1002972.

Gurtan, A.M., and Sharp, P.A. (2013). The role of miRNAs in regulating gene expression networks. *J Mol Biol* 425, 3582-3600.

Hartl, D.L., Lohe, A.R., and Lozovskaya, E.R. (1997). Regulation of the transposable element mariner. *Genetica* 100, 177-184.

Higuchi, M., Yoshida, S., Ueharu, H., Chen, M., Kato, T., and Kato, Y. (2014). PRRX1 and PRRX2 distinctively participate in pituitary organogenesis and a cell-supply system. *Cell Tissue Res* 357, 323-335.

Hromas, R., Williamson, E.A., Fnu, S., Lee, Y.J., Park, S.J., Beck, B.D., You, J.S., Leitao, A., Nickoloff, J.A., and Lee, S.H. (2012). Chk1 phosphorylation of Metnase enhances DNA repair but inhibits replication fork restart. *Oncogene* 31, 4245-4254.

Hromas, R., Wray, J., Lee, S.H., Martinez, L., Farrington, J., Corwin, L.K., Ramsey, H., Nickoloff, J.A., and Williamson, E.A. (2008). 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 (Amst)* 7, 1927-1937.

Hsu, P.W., Huang, H.D., Hsu, S.D., Lin, L.Z., Tsou, A.P., Tseng, C.P., Stadler, P.F., Washietl, S., and Hofacker, I.L. (2006). miRNAMap: genomic maps of microRNA genes and their target genes in mammalian genomes. *Nucleic Acids Res* 34, D135-139.

Hu, B., Ying, X., Wang, J., Piriyaopongsa, J., Jordan, I.K., Sheng, J., Yu, F., Zhao, P., Li, Y., Wang, H., *et al.* (2014). Identification of a tumor-suppressive human-specific microRNA within the FHIT tumor-suppressor gene. *Cancer Res* 74, 2283-2294.

Huisman, O., and Kleckner, N. (1987). A new generalizable test for detection of mutations affecting Tn10 transposition. *Genetics* 116, 185-189.

Hurst, L.D., Sachenkova, O., Daub, C., Forrest, A., and Huminiecki, L. (2014). A simple metric of promoter architecture robustly predicts expression breadth of human genes suggesting that most transcription factors are positive regulators. *Genome Biol* 15, 413.

Hutvagner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* 293, 834-838.

Jackson, A.L., and Linsley, P.S. (2010). Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nat Rev Drug Discov* 9, 57-67.

Jacobson, J.W., Medhora, M.M., and Hartl, D.L. (1986). Molecular structure of a somatically unstable transposable element in *Drosophila*. *Proc Natl Acad Sci U S A* 83, 8684-8688.

Jha, D.K., and Strahl, B.D. (2014). An RNA polymerase II-coupled function for histone H3K36 methylation in checkpoint activation and DSB repair. *Nat Commun* 5, 3965.

Johnson, R., Gamblin, R.J., Ooi, L., Bruce, A.W., Donaldson, I.J., Westhead, D.R., Wood, I.C., Jackson, R.M., and Buckley, N.J. (2006). Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res* 34, 3862-3877.

Jordan, I.K., Rogozin, I.B., Glazko, G.V., and Koonin, E.V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* 19, 68-72.

Jursch, T., Miskey, C., Izsvak, Z., and Ivics, Z. (2013). Regulation of DNA transposition by CpG methylation and chromatin structure in human cells. *Mob DNA* 4, 15.

Kamanu, T.K., Radovanovic, A., Archer, J.A., and Bajic, V.B. (2013). Exploration of miRNA families for hypotheses generation. *Sci Rep* 3, 2940.

Kapitonov, V.V., and Jurka, J. (2005). RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol* 3, e181.

Kapitonov, V.V., and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet* 9, 411-412; author reply 414.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M., and Feschotte, C. (2013). Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* 9, e1003470.

Kim, B.M., Buchner, G., Miletich, I., Sharpe, P.T., and Shivdasani, R.A. (2005). The stomach mesenchymal transcription factor Barx1 specifies gastric epithelial identity through inhibition of transient Wnt signaling. *Dev Cell* 8, 611-622.

Kim, H.S., Chen, Q., Kim, S.K., Nickoloff, J.A., Hromas, R., Georgiadis, M.M., and Lee, S.H. (2014). The DDN Catalytic Motif Is Required for Metnase Functions in Non-homologous End Joining (NHEJ) Repair and Replication Restart. *J Biol Chem* 289, 10930-10938.

Kim, J.H., Ebersole, T., Kouprina, N., Noskov, V.N., Ohzeki, J., Masumoto, H., Mravinac, B., Sullivan, B.A., Pavlicek, A., Dovat, S., *et al.* (2009). Human gamma-satellite DNA maintains open chromatin structure and protects a transgene from epigenetic silencing. *Genome Res* 19, 533-544.

Kim, S., Kim, H., Fong, N., Erickson, B., and Bentley, D.L. (2011). Pre-mRNA splicing is a determinant of histone H3K36 methylation. *Proc Natl Acad Sci U S A* 108, 13564-13569.

Kolasinska-Zwierz, P., Down, T., Latorre, I., Liu, T., Liu, X.S., and Ahringer, J. (2009). Differential chromatin marking of introns and expressed exons by H3K36me3. *Nat Genet* 41, 376-381.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39, D152-157.

Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.S., Ng, H.H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* 42, 631-634.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes Dev* 16, 2893-2905.

Lampe, D.J., Churchill, M.E., and Robertson, H.M. (1996). A purified mariner transposase is sufficient to mediate transposition in vitro. *Embo J* 15, 5470-5479.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., *et al.* (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* 22, 1813-1831.

Lang, D., and Epstein, J.A. (2003). Sox10 and Pax3 physically interact to mediate activation of a conserved c-RET enhancer. *Hum Mol Genet* 12, 937-945.

Le Rouzic, A., Boutin, T.S., and Capy, P. (2007). Long-term evolution of transposable elements. *Proc Natl Acad Sci U S A* *104*, 19375-19380.

Le Rouzic, A., and Capy, P. (2005). The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. *Genetics* *169*, 1033-1043.

Lee, S.H., Oshige, M., Durant, S.T., Rasila, K.K., Williamson, E.A., Ramsey, H., Kwan, L., Nickoloff, J.A., and Hromas, R. (2005). The SET domain protein Metnase mediates foreign DNA integration and links integration to nonhomologous end-joining repair. *Proc Natl Acad Sci U S A* *102*, 18075-18080.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., *et al.* (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature* *425*, 415-419.

Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *Embo J* *21*, 4663-4670.

Li, B., Jackson, J., Simon, M.D., Fleharty, B., Gogol, M., Seidel, C., Workman, J.L., and Shilatifard, A. (2009a). Histone H3 lysine 36 dimethylation (H3K36me2) is sufficient to recruit the Rpd3s histone deacetylase complex and to repress spurious transcription. *J Biol Chem* *284*, 7970-7976.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* *25*, 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009b). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* *25*, 2078-2079.

Li, L., Olvera, J.M., Yoder, K.E., Mitchell, R.S., Butler, S.L., Lieber, M., Martin, S.L., and Bushman, F.D. (2001). Role of the non-homologous DNA end joining pathway in the early steps of retroviral infection. *Embo J* *20*, 3272-3281.

Li, M.A., Pettitt, S.J., Eckert, S., Ning, Z., Rice, S., Cadinanos, J., Yusa, K., Conte, N., and Bradley, A. (2013a). The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Mol Cell Biol* 33, 1317-1330.

Li, Y., Xie, J., Xu, X., Wang, J., Ao, F., Wan, Y., and Zhu, Y. (2013b). MicroRNA-548 down-regulates host antiviral response via direct targeting of IFN-lambda1. *Protein Cell* 4, 130-141.

Liu, D., Bischerour, J., Siddique, A., Buisine, N., Bigot, Y., and Chalmers, R. (2007). The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase. *Mol Cell Biol* 27, 1125-1132.

Liu, D., and Chalmers, R. (2014). Hyperactive mariner transposons are created by mutations that disrupt allosterism and increase the rate of transposon end synapsis. *Nucleic Acids Res* 42, 2637-2645.

Lohe, A.R., and Hartl, D.L. (1996). Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Mol Biol Evol* 13, 549-555.

Lohe, A.R., Lidholm, D.A., and Hartl, D.L. (1995). Genotypic effects, maternal effects and grand-maternal effects of immobilized derivatives of the transposable element mariner. *Genetics* 140, 183-192.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* 7, 29-59.

Masumoto, H., Nakano, M., and Ohzeki, J. (2004). The role of CENP-B and alpha-satellite DNA: de novo assembly and epigenetic maintenance of human centromeres. *Chromosome Res* 12, 543-556.

Matsutani, S., Ohtsubo, H., Maeda, Y., and Ohtsubo, E. (1987). Isolation and characterization of IS elements repeated in the bacterial chromosome. *J Mol Biol* 196, 445-455.

Mayya, V., Lundgren, D.H., Hwang, S.I., Rezaul, K., Wu, L., Eng, J.K., Rodionov, V., and Han, D.K. (2009). Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Sci Signal* 2, ra46.

McVey, M., and Lee, S.E. (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. *Trends Genet* 24, 529-538.

Medstrand, P., van de Lagemaat, L.N., Dunn, C.A., Landry, J.R., Svenback, D., and Mager, D.L. (2005). Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet Genome Res* 110, 342-352.

Miskey, C., Papp, B., Mates, L., Sinzelle, L., Keller, H., Izsvak, Z., and Ivics, Z. (2007). 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, 4589-4600.

Mohapatra, S., Yannone, S.M., Lee, S.H., Hromas, R.A., Akopiants, K., Menon, V., Ramsden, D.A., and Povirk, L.F. (2013). Trimming of damaged 3' overhangs of DNA double-strand breaks by the Metnase and Artemis endonucleases. *DNA Repair (Amst)* 12, 422-432.

Morgan, G.T. (1995). Identification in the human genome of mobile elements spread by DNA-mediated transposition. *J Mol Biol* 254, 1-5.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5, 621-628.

Munoz-Lopez, M., Siddique, A., Bischerour, J., Lorite, P., Chalmers, R., and Palomeque, T. (2008). Transposition of Mboumar-9: identification of a new naturally active mariner-family transposon. *J Mol Biol* 382, 567-572.

Newman, J.C., Bailey, A.D., Fan, H.Y., Pavelitz, T., and Weiner, A.M. (2008). An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. *PLoS Genet* 4, e1000031.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A., *et al.* (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Sci Signal* 3, ra3.

Orgel, L.E., and Crick, F.H. (1980). Selfish DNA: the ultimate parasite. *Nature* 284, 604-607.

Pace, J.K., 2nd, and Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res* 17, 422-432.

Pai, C.C., Deegan, R.S., Subramanian, L., Gal, C., Sarkar, S., Blaikley, E.J., Walker, C., Hulme, L., Bernhard, E., Codlin, S., *et al.* (2014). A histone H3K36 chromatin switch coordinates DNA double-strand break repair pathway choice. *Nat Commun* 5, 4091.

Pajunen, M.I., Rasila, T.S., Happonen, L.J., Lamberg, A., Haapa-Paananen, S., Kiljunen, S., and Savilahti, H. (2010). Universal platform for quantitative analysis of DNA transposition. *Mob DNA* 1, 24.

Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M.I., Maller, B., Hayward, D.C., Ball, E.E., Degnan, B., Muller, P., *et al.* (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* 408, 86-89.

Pavelitz, T., Gray, L.T., Padilla, S.L., Bailey, A.D., and Weiner, A.M. (2013). PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. *Mob DNA* 4, 23.

Pilch, D.R., Sedelnikova, O.A., Redon, C., Celeste, A., Nussenzweig, A., and Bonner, W.M. (2003). Characteristics of gamma-H2AX foci at DNA double-strand breaks sites. *Biochem Cell Biol* 81, 123-129.

Piriyaongsa, J., and Jordan, I.K. (2007). A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One* 2, e203.

Piriyapongsa, J., and Jordan, I.K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. *Rna* 14, 814-821.

Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., *et al.* (2011). Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* 29, 742-749.

Prufer, K., Munch, K., Hellmann, I., Akagi, K., Miller, J.R., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., *et al.* (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486, 527-531.

Ramamoorthy, M., Tadokoro, T., Rybanska, I., Ghosh, A.K., Wersto, R., May, A., Kulikowicz, T., Sykora, P., Croteau, D.L., and Bohr, V.A. (2012). RECQL5 cooperates with Topoisomerase II alpha in DNA decatenation and cell cycle progression. *Nucleic Acids Res* 40, 1621-1635.

Rath, A., Hromas, R., and De Benedetti, A. (2014). Fidelity of end joining in mammalian episomes and the impact of Metnase on joint processing. *BMC Mol Biol* 15, 6.

Renault, S., Demattei, M.V., Lahouassa, H., Bigot, Y., and Auge-Gouillou, C. (2010). In vitro recombination and inverted terminal repeat binding activities of the Mcmar1 transposase. *Biochemistry* 49, 3534-3544.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147, 1408-1419.

Rhee, H.S., and Pugh, B.F. (2012). ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. *Curr Protoc Mol Biol Chapter 21*, Unit 21 24.

Richardson, J.M., Colloms, S.D., Finnegan, D.J., and Walkinshaw, M.D. (2009). Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell* 138, 1096-1108.

Rigbolt, K.T., Prokhorova, T.A., Akimov, V., Henningsen, J., Johansen, P.T., Kratchmarova, I., Kassem, M., Mann, M., Olsen, J.V., and Blagoev, B. (2011). System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. *Sci Signal* 4, rs3.

Rinn, J.L., and Chang, H.Y. (2012). Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81, 145-166.

Robertson, H.M. (1993). The mariner transposable element is widespread in insects. *Nature* 362, 241-245.

Robertson, H.M., and Lampe, D.J. (1995). Distribution of transposable elements in arthropods. *Annu Rev Entomol* 40, 333-357.

Robertson, H.M., and Martos, R. (1997). Molecular evolution of the second ancient human mariner transposon, Hsmar2, illustrates patterns of neutral evolution in the human genome lineage. *Gene* 205, 219-228.

Robertson, H.M., and Zumpano, K.L. (1997). Molecular evolution of an ancient mariner transposon, Hsmar1, in the human genome. *Gene* 205, 203-217.

Roman, Y., Oshige, M., Lee, Y.J., Goodwin, K., Georgiadis, M.M., Hromas, R.A., and Lee, S.H. (2007). Biochemical characterization of a SET and transposase fusion protein, Metnase: its DNA binding and DNA cleavage activity. *Biochemistry* 46, 11369-11376.

Romanish, M.T., Lock, W.M., van de Lagemaat, L.N., Dunn, C.A., and Mager, D.L. (2007). Repeated recruitment of LTR retrotransposons as promoters by the anti-apoptotic locus NAIP during mammalian evolution. *PLoS Genet* 3, e10.

Rose, N.R., and Klose, R.J. (2014). Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta*.

Schapira, M. (2011). Structural Chemistry of Human SET Domain Protein Methyltransferases. *Curr Chem Genomics* 5, 85-94.

Schatz, D.G. (1999). Transposition mediated by RAG1 and RAG2 and the evolution of the adaptive immune system. *Immunol Res* 19, 169-182.

Schmidt, C., Kim, D., Ippolito, G.C., Naqvi, H.R., Probst, L., Mathur, S., Rosas-Acosta, G., Wilson, V.G., Oldham, A.L., Poenie, M., *et al.* (2009). Signalling of the BCR is regulated by a lipid rafts-localised transcription factor, Bright. *Embo J* 28, 711-724.

Schneider, R., Bannister, A.J., and Kouzarides, T. (2002). Unsafe SETs: histone lysine methyltransferases and cancer. *Trends Biochem Sci* 27, 396-402.

Seberg, O., and Petersen, G. (2009). A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat Rev Genet* 10, 276.

Serandour, A.A., Brown, G.D., Cohen, J.D., and Carroll, J.S. (2013). Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biol* 14, R147.

Shao, H., and Tu, Z. (2001). Expanding the diversity of the IS630-Tc1-mariner superfamily: discovery of a unique DD37E transposon and reclassification of the DD37D and DD39D transposons. *Genetics* 159, 1103-1115.

Shen, B., Dong, L., Xiao, S., and Kowalewski, M. (2008). The Avalon explosion: evolution of Ediacara morphospace. *Science* 319, 81-84.

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., and Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res* 34, D32-36.

Singh, J., and Padgett, R.A. (2009). Rates of in situ transcription and splicing in large human genes. *Nat Struct Mol Biol* 16, 1128-1133.

Sinzelle, L., Izsvak, Z., and Ivics, Z. (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci* 66, 1073-1093.

Smit, A.F., and Riggs, A.D. (1996). Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* 93, 1443-1448.

Srinivasan, R., Sun, G., Keles, S., Jones, E.A., Jang, S.W., Krueger, C., Moran, J.J., and Svaren, J. (2012). Genome-wide analysis of EGR2/SOX10 binding in myelinating peripheral nerve. *Nucleic Acids Res* 40, 6449-6460.

Stelnicki, E.J., Arbeit, J., Cass, D.L., Saner, C., Harrison, M., and Largman, C. (1998). Modulation of the human homeobox genes PRX-2 and HOXB13 in scarless fetal wounds. *J Invest Dermatol* 111, 57-63.

Stoebel, D.M., and Dorman, C.J. (2010). The effect of mobile element IS10 on experimental regulatory evolution in *Escherichia coli*. *Mol Biol Evol* 27, 2105-2112.

Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. *Nat Struct Mol Biol* 20, 267-273.

Symington, L.S., and Gautier, J. (2011). Double-strand break end resection and repair pathway choice. *Annu Rev Genet* 45, 247-271.

Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc Natl Acad Sci U S A* 110, 18602-18607.

Tosi, L.R., and Beverley, S.M. (2000). cis and trans factors affecting Mos1 mariner evolution and transposition in vitro, and its potential for functional genomics. *Nucleic Acids Res* 28, 784-790.

Towbin, B.D., Gonzalez-Aguilera, C., Sack, R., Gaidatzis, D., Kalck, V., Meister, P., Askjaer, P., and Gasser, S.M. (2012). Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. *Cell* 150, 934-947.

- Trubitsyna, M., Morris, E.R., Finnegan, D.J., and Richardson, J.M. (2014). Biochemical characterization and comparison of two closely related active mariner transposases. *Biochemistry* **53**, 682-689.
- van Attikum, H., Bundock, P., and Hooykaas, P.J. (2001). Non-homologous end-joining proteins are required for *Agrobacterium* T-DNA integration. *Embo J* **20**, 6550-6558.
- van de Lagemaat, L.N., Landry, J.R., Mager, D.L., and Medstrand, P. (2003). Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19**, 530-536.
- Volff, J.N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**, 913-922.
- Wagner, E.J., and Carpenter, P.B. (2012). Understanding the language of Lys36 methylation at histone H3. *Nat Rev Mol Cell Biol* **13**, 115-126.
- Wang, J., Duncan, D., Shi, Z., and Zhang, B. (2013). WEB-based GENE SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* **41**, W77-83.
- Wang, J.C. (2002). Cellular roles of DNA topoisomerases: a molecular perspective. *Nat Rev Mol Cell Biol* **3**, 430-440.
- Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., and Haussler, D. (2007). Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* **104**, 18613-18618.
- Wang, Y., Wang, J., Devaraj, A., Singh, M., Jimenez Orgaz, A., Chen, J.X., Selbach, M., Ivics, Z., and Izsvak, Z. (2014). Suicidal autointegration of sleeping beauty and piggyBac transposons in eukaryotic cells. *PLoS Genet* **10**, e1004103.

Watkins, S., van Pouderoyen, G., and Sixma, T.K. (2004). Structural analysis of the bipartite DNA-binding domain of Tc3 transposase bound to transposon DNA. *Nucleic Acids Res* 32, 4306-4312.

Wei, H., and Zhou, M.M. (2010). Dimerization of a viral SET protein endows its function. *Proc Natl Acad Sci U S A* 107, 18433-18438.

Weiner, A.M., and Gray, L.T. (2013). What role (if any) does the highly conserved CSB-PGBD3 fusion protein play in Cockayne syndrome? *Mech Ageing Dev* 134, 225-233.

Wessler, S.R., Bureau, T.E., and White, S.E. (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. *Curr Opin Genet Dev* 5, 814-821.

Wheatley, R.W., Lo, S., Jancewicz, L.J., Dugdale, M.L., and Huber, R.E. (2013). Structural explanation for allolactose (lac operon inducer) synthesis by lacZ beta-galactosidase and the evolutionary relationship between allolactose synthesis and the lac repressor. *J Biol Chem* 288, 12993-13005.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8, 973-982.

Wilke, C.M., and Adams, J. (1992). Fitness effects of Ty transposition in *Saccharomyces cerevisiae*. *Genetics* 131, 31-42.

Williamson, E.A., Farrington, J., Martinez, L., Ness, S., O'Rourke, J., Lee, S.H., Nickoloff, J., and Hromas, R. (2008a). Expression levels of the human DNA repair protein metnase influence lentiviral genomic integration. *Biochimie* 90, 1422-1426.

Williamson, E.A., Rasila, K.K., Corwin, L.K., Wray, J., Beck, B.D., Severns, V., Mobarak, C., Lee, S.H., Nickoloff, J.A., and Hromas, R. (2008b). The SET and transposase domain protein Metnase enhances chromosome decatenation: regulation by automethylation. *Nucleic Acids Res* 36, 5822-5831.

Williamson, E.A., Wu, Y., Singh, S., Byrne, M., Wray, J., Lee, S.H., Nickoloff, J.A., and Hromas, R. (2014). The DNA repair component Metnase regulates Chk1 stability. *Cell Div* 9, 1.

Wray, J., Williamson, E.A., Chester, S., Farrington, J., Sterk, R., Weinstock, D.M., Jasin, M., Lee, S.H., Nickoloff, J.A., and Hromas, R. (2010). The transposase domain protein Metnase/SETMAR suppresses chromosomal translocations. *Cancer Genet Cytogenet* 200, 184-190.

Wurm, F.M. (2004). Production of recombinant protein therapeutics in cultivated mammalian cells. *Nat Biotechnol* 22, 1393-1398.

Wurtele, H., Little, K.C., and Chartrand, P. (2003). Illegitimate DNA integration in mammalian cells. *Gene Ther* 10, 1791-1799.

Yang, G., Fattash, I., Lee, C.N., Liu, K., and Cavinder, B. (2013). Birth of three stowaway-like MITE families via microhomology-mediated miniaturization of a Tc1/Mariner element in the yellow fever mosquito. *Genome Biol Evol* 5, 1937-1948.

Yang, G., Nagel, D.H., Feschotte, C., Hancock, C.N., and Wessler, S.R. (2009). Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. *Science* 325, 1391-1394.

Yant, S.R., Wu, X., Huang, Y., Garrison, B., Burgess, S.M., and Kay, M.A. (2005). High-resolution genome-wide mapping of transposon integration in mammals. *Mol Cell Biol* 25, 2085-2094.

Yoda, K., Ando, S., Okuda, A., Kikuchi, A., and Okazaki, T. (1998). In vitro assembly of the CENP-B/alpha-satellite DNA/core histone complex: CENP-B causes nucleosome positioning. *Genes Cells* 3, 533-548.

Yuan, Y.W., and Wessler, S.R. (2011). The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc Natl Acad Sci U S A* 108, 7884-7889.

Zhong, Q., Chen, C.F., Chen, P.L., and Lee, W.H. (2002). BRCA1 facilitates microhomology-mediated end joining of DNA double strand breaks. *J Biol Chem* 277, 28641-28647.

Zhou, Y.H., Zheng, J.B., Gu, X., Saunders, G.F., and Yung, W.K. (2002). Novel PAX6 binding sites in the human genome and the role of repetitive elements in the evolution of gene regulation. *Genome Res* 12, 1716-1722.