THESE DE DOCTORAT

De l'Université Sorbonne Paris Cité
Préparée à l'Université Paris Diderot
École Doctorale 562. BIO SORBONNE PARIS CITE
Unité Épigénétique et Destin Cellulaire UMR7216

et

De la Universidad Colegio Mayor de Nuestra Señora del Rosario
Faculté des Sciences Naturelles et Mathématiques et
École de Médecine et des Sciences de la Santé
Doctorat en Sciences Biomédicales et Biologiques

Un mécanisme de trans-méthylation entre les deux principales méthyltransférases de H3K9 SETDB1 et SUV39H1, régule l'établissement de l'hétérochromatine

A trans-methylation mechanism between the two major H3K9 methyltransferases SETDB1 and SUV39H1 regulates heterochromatin establishment

Par **Paola Andrea CRUZ TAPIAS**

Thèse de doctorat de Biologie Moléculaire

Dirigée par le Dr. Slimane AIT SI ALI et le Dr. Sandra RAMIREZ CLAVIJO

Présentée et soutenue publiquement à Paris, le 21 Septembre 2018

Président du jury : Dr Sandra DUHARCOURT / CNRS - IJM - Université Paris Diderot

Rapporteurs : Dr Robert FEIL / CNRS - IGMM, Montpellier
Dr Reini FERNANDEZ DE LUCO / CNRS - IGH, Montpellier
Dr José-Manuel LOZANO / Universidad Nacional de Colombia

Examinateurs : Dr Fabien LE GRAND / CNRS - INSERM - Institut de Myologie
Dr Hernando CURTIDOR / Universidad del Rosario

Directeurs de thèse : Dr Slimane AIT SI ALI / CNRS - Université Paris Diderot
Dr Sandra RAMIREZ CLAVIJO / Universidad del Rosario

ACKNOWLEDGEMENTS

Many people provided guidance and support during my thesis work and I am ever in their debt. First, I thank my thesis co-director Dr. Slimane Ait Si Ali for his guidance along the course of my program. He constantly encouraged and challenged me with insightful ideas and always provided a stimulating intellectual environment, giving me the tools I needed to become an independent researcher.

I am also very grateful to Lauriane Fritsch, she taught me a lot about techniques and work organization. I also thank Slimane's lab members, past and present, Véronique Joliot, Laurence del Maestro, Vlada Zakharova, Sonia Albini, Roberta Rapone, Gilles Tellier, Cecile Philippon, Katia Boyarchuk and Philippe Robin for their help and encouragement. I also thank my colleagues in the Epigenetics and Cell Fate Unit who have helped to shape my ideas, given me the opportunity to present my work and have provided technical assistance and support. I cannot be more grateful than I am for the excellent work environment, the friendship and scientific feedback.

I also would like to thank Dr. Sandra Ramirez for her help and support.

I am grateful to the Agence française pour la promotion de l'enseignement supérieur, l'accueil et la mobilité -Bourse Eiffel as well as the Fondation ARC pour la recherche sur le cancer and the Colombian Government -COLCIENCIAS for believing in me and my project, allowing me to work for these years. Many thanks to all the collaborators, who worked with us and share with us projects, techniques, skills and results.

This work would not have been possible without the support and encouragement of my family. I am grateful to my grandparents, Luis and Beatriz and my “aunt-moms” Flor, Ines, Cecilia and Rosita for their unparalleled love and support.

It is my proud privilege to express my sincere thanks and deep sense of gratitude to my father and my sisters, Elizabeth and Angélica. They have never stopped believing in me.

I offer my heartfelt thanks to my fiancé, Oscar, who has patiently journeyed with me over these years, providing me with encouragement and steady support.

All my work is dedicated to my mother, the most courageous person I know…

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

5mC – 5-methyl-cytosine
AGO – Argonaute
CGI – CpG Island
ChIP – Chromatin Immuno-Precipitation
CTCF – CCCTC-binding factor
DNMT – DNA methyltransferase
Dpc – days post coitum
EC – Ectoplacental Cone
EED – Embryonic ectoderm development
ERV – Endogenous Retroviral Elements
ESC – Embryonic Stem Cell
HAT – Histone Acetyltransferase
HDAC – Histone Deacetylase
HMT – Histone Methyltransferase
HOTAIR – lncRNA HOX transcript antisense RNA
HP1 – Heterochromatin Protein 1
IAP – Intracisternal A-Particle
ICM – Inner Cell Mass
ICR – Imprinting Control Region
JMJD – Jumonji domain containing protein
Kb – Kilobase
KMD – Lysine Methyldemethylase
KMT – Lysine Methyltransferase
KRAB – Kruppel-Associated box
LIF – Leukemia Inhibitory Factor
LINE – Long Interspersed Nuclear Elements
LTR – Long Terminal Repeat
MBD – Methyl-Binding Protein
MEF – Mouse Embryonic Fibroblast
MLV – Moloney Murine Leukemia Virus
mRNA – messenger RNA
ncRNA – non-coding RNA
NuRD – Nuclear Remodelling and Deacetylation Complex
ORF – Open Reading Frame
PAD – Peptidyl arginine deiminase
PRC – Polycomb Repressive Complex
PTMs – Post-translational modifications
PRMTs – Protein arginine methyltransferases
PWS – Prader-Willi Syndrome
RBAP – Retinoblastoma-associated protein
REST – RE1-silencing transcription factor
RNA Pol – RNA Polymerase
RNP – ribonucleoprotein
SAM – S-Adenosyl Methionine
SINE – Short Interspersed Nuclear Elements
SUV39HDN – SUV39H1/2 Double knockout
SUMO – Small Ubiquitin-related MOdifier protein
TE – Transposable Elements
XIST – X-Inactive Specific Transcript
SUZ – Suppressor-of-Zeste

ABSTRACT

Histone H3 lysine 9 (H3K9) methylation, which is established by the lysine methyltransferases (KMTs) SETDB1, SUV39H1, G9A and GLP, is a central epigenetic mechanism involved in cell fate regulation. In particular, H3K9 methylation is directly involved in heterochromatin formation and gene silencing. Our lab showed that the main H3K9 KMTs (SETDB1, G9A, GLP and SUV39H1) form a functional megacomplex involved in transcriptional silencing (1), probably *via* the cooperative establishment of the different H3K9 methylation levels. However, up to now, the regulation of the H3K9 KMT complexes is not fully understood. Interestingly, post-translational modifications (PTMs) have been implicated in the regulation of H3K9 KMT functions (2). In this, my PhD thesis aimed to decipher how methylation of SETDB1, regulates its activity (complex formation, interaction with partners, recruitment to chromatin), which ultimately could impact on heterochromatin formation, gene expression and cell fate regulation.

SETDB1 is crucial during development and cellular differentiation. Moreover, SETDB1 is essential in mouse embryonic stem cells (mESCs) pluripotency and self-renewal (3,4), *Setbd1* KO is lethal at the peri-implantation stage at 7.5 days postcoitum (dpc) (5). Beside histones, SETDB1 is also able to methylate other proteins (e.g. UBF, ING2, p53) (6–8). Notably, my current data show that SETDB1 undergoes (auto)methylation on the lysines K1170 and K1178 located inside its catalytic SET domain.

SETDB1 and SUV39H1 coordinate the establishment and maintenance of H3K9me3 at constitutive pericentromeric heterochromatin (9,10) and co-regulate many genomic targets

within heterochromatin, including transposable elements, such as long interspersed nuclear elements (LINEs) and endogenous retroviruses (ERVs) (11). Since SUV39H1 is a H3K9 tri-methyltransferase that uses H3K9me1 or H3K9me2 as a primary substrate, SETDB1 could probably provide mono- or di-methylated H3K9 (9,10). Interestingly, my results point to a model in which SETDB1 auto-methylation paves the path to a subsequent trans-methylation by SUV39H1. This mechanism could regulate not only the SETDB1/SUV39H1 physical interaction (*via* the SUV39H1 chromodomain), but also cooperation in the establishment and maintenance of both heterochromatin blocks (large domains) and transposable elements (TEs) silencing, at least in ES cells. Thus, we would like to better understand how the crosstalk between these two key H3K9 KMTs, SETDB1 and SUV39H1, occurs in terms of interaction and recruitment to target loci.

Keywords*: SETDB1, SUV39H1, post-translational modification, lysine methylation*

RÉSUMÉ

La méthylation de la lysine 9 de l'histone 3 (H3K9), établie par les lysine méthyltransférases (KMTs) SETDB1, SUV39H1, G9A et GLP, représente un mécanisme épigénétique central dans la régulation du destin cellulaire. En particulier, la methylation d'H3K9 est directement impliquée dans la formation de l'hétérochromatine et l'extinction des gènes. Notre laboratoire a montré que les principales KMTs (SETDB1, G9A, GLP et SUV39H1) spécifiques de H3K9 forment un méga-complexe impliqué dans la répression transcriptionnelle (1), probablement *via* une coopération pour établir les différents niveaux de méthylation. Néanmoins, la régulation des complexes de H3K9 KMT n'est jusqu'à présent pas bien comprise. Il est à noter que des modifications post-traductionnelles (PTM) ont été impliquées dans la régulation des fonctions des KMTs (2). Dans ce contexte, mon projet visait à comprendre comment la méthylation de SETDB1 régulerait son activité (incorporation dans des complexes, interaction avec ses partenaires, recrutement à la chromatine). Le but étant d'établir quel impact auraient ces modifications de SETDB1 sur la formation de l'hétérochromatine, l'expression des gènes et la régulation du destin cellulaire.

SETDB1 est cruciale lors du développement et de la différençiation cellulaire. De plus, SETDB1 est essentielle pour la pluripotence et le renouvellement des cellules souches embryonnaires murines (mESC) (3,4). L'inactivation génique de ou KO de *Setdb1* est létal au stade préimplantatoire à 7,5 jours post-coïtum (dpc) (5). En plus des histones, SETDB1 méthyle d'autres protéines comme UBF, ING2 et p53 (6-8). Mes résultats montrent notamment, que SETDB1 s'autométhyle sur les lysines K1170 et K1178 localisées dans le domaine catalytique SET.

SETDB1 et SUV39H1 coordonnent l'établissement et la maintenance de H3K9me3 dans l'hétérochromatine péricentromérique constitutive (9,10) et co-régulent de nombreuses cibles génomiques dans l'hétérochromatine, dont les éléments transposables comme les *Long Interspersed Nuclear Elements* (LINEs) et les rétrovirus endogènes (ERVs) (11).

Comme SUV39H1 est une triméthyltransférase qui utilise H3K9me1 ou H3K9me2 comme substrat primaire, SETDB1 pourrait probablement fournir les mono- ou di-méthyl H3K9 (9,10). Mes résultats suggèrent un modèle dans lequel l'auto-méthylation de SETDB1 est pré-requise à la trans-méthylation subséquente par SUV39H1. Ce mécanisme pourrait réguler non seulement l'interaction physique entre SETDB1 et SUV39H1, *via* le chromodomaine de SUV39H1, mais aussi leur coopération dans l'établissement et la maintenance des blocs (grands domaines) d'hétérochromatine et l'extinction des éléments transposables, au moins dans les cellules souches. Ainsi, nous souhaitons mieux comprendre comment le « dialogue » entre ces deux H3K9 KMT majeures, SETDB1 et SUV39H1, est impliqué dans leurs interactions et leurs recrutements aux *loci* cibles.

***Mots clés**: SETDB1, SUV39H1, modifications post-traductionnelles, méthylation de lysines*

RESUMEN

La metilación de la lisina 9 de la histona 3 (H3K9) es establecida por la familia de las lisina metiltransferasas (KMTs) SETDB1, SUV39H1, G9A y GLP. Esta metilación representa un mecanismo epigenético central en la regulación del destino celular. En particular, la metilación de la H3K9 esta directamente implicada en la formación de la heterocromatina y la represión de transcripción génica. Nuestro laboratorio demostró que las principales KMTs (SETDB1, SUV39H1, G9A y GLP) especificas de H3K9 forman un complejo multimérico implicado en la represión transcripcional (1), probablemente mediante una cooperación para establecer diferentes niveles de metilación. Sin embargo, aun se desconoce como se regula este complejo de H3K9 KMTs. Es importante mencionar que las modificaciones post-traduccionales (PTMs) han sido implicadas en la regulación de las KMTs (2). En este contexto, mi proyecto busca comprender como la metilación de SETDB1 regula su actividad (incorporación dentro del complejo, interacción con otras proteínas y reclutamiento a la cromatina). El objetivo es establecer el impacto que tendrán las PTMs de SETDB1 con respecto a la formación de heterocromatina, la expresión génica y la regulación del destino celular.

SETDB1 es crucial durante el desarrollo y la diferenciación celular. Es mas, SETDB1 es esencial para la pluripotencia y la renovación de las células madre embrionarias murinas (mESCs) (3,4). La inactivación génica o « Knockout » de *Setdb1* es letal en un estado de pre-implantación de 7,5 dias post-coito (dpc) (5). Además de las histonas, SETDB1 también puede metilar otras proteínas como UBF, ING y p53 (6-8). Mis resultados muestran que SETDB1 puede auto-metilarse en las lisinas K1170 y K1178, las cuales se encuentran ubicadas dentro su dominio catalítico SET.

SETDB1 y SUV39H1 coordinan el establecimiento y el mantenimiento de H3K9me3 en la heterocromatina pericentromérica constitutiva (9,10) y co-regulan numerosos blancos genómicos en la heterocromatina, dentro de los cuales se encuentran los elementos transposables como los *Long Interspersed Nuclear Elements* (LINEs) y los retrovirus endógenos (ERVs) (11).

Debido a que SUV39H1 es una trimetiltransferasa que utiliza H3K9me1 o H3K9me2 como sustrato primario, SETDB1 podría probablemente aportar los mono- o di-metil H3K9 (9,10). Mis resultados sugieren un modelo en el que la auto-metilacion de SETDB1 es un pre-requisito para la trans-metilación por SUV39H1. Este mecanismo podría regular no solamente la interacción física entre SETDB1 y SUV39H1 (*vía* el cromodominio de SUV39H1), sino también la cooperación para el establecimiento y el mantenimiento de la heterocromatina y el silenciamiento de los elementos transposables, al menos en células madre. En general, el proyecto busca comprender mejor cómo ocurre el "*crosstalk*" entre estas dos H3K9 KMT, SETDB1 y SUV39H1 en términos de su interacción y el reclutamiento a sus genes blanco.

***Palabras clave**: SETDB1, SUV39H1, modificaciones post-traduccionales, metilación de lisinas*

STATE OF ART AND OBJECTIVES

There are distinct epigenetic mechanisms (DNA methylation, covalent histone modifications, non-covalent mechanisms such as incorporation of histone variants and ATP-dependent nucleosome remodeling and non-coding RNAs), which cooperate in the nucleus to modify the conformation of chromatin, impacting on all nuclear processes that need access to the DNA, including gene transcription. These epigenetic mechanisms are essential for normal embryonic development and maintenance of tissue-specific gene expression patterns in mammals and their disruption can lead to altered gene function and participate in pathogenesis. For example, modulation of histone PTMs, including methylation of H3K9, has been tightly linked to tumorigenesis (12). H3K9 methylation, which is mainly involved in gene repression and plays key roles in cell fate regulation, is established by the KMTs SETDB1, SUV39H1, G9A and GLP. H3K9 methylation has gained paramount importance. Indeed, aberrant changes in H3K9 methylation were linked with increased recurrence and poor survival of several malignancies and the main H3K9 KMTs are deregulated in many cancer types (13). For example, G9A has an oncogenic function by conferring survival and growth advantages to tumors (14); and pharmacological inhibition of G9A triggers death of cancer cells of different tissue origins (14), providing proof-of-principle for a H3K9 KMT inhibition as a possible pharmacological approach for cancer.

SETDB1 is also deregulated in several cancer types. In collaboration with Zon L group, our lab participated in showing that *SETDB1* gene is amplified in human melanoma and accelerates melanoma onset *in vivo* (15). Moreover, SETDB1 has been described as a tumorigenic factor in other tumor types such as liver (7), lung (16) and breast cancer (17). Interestingly, it has been shown that SETDB1 modulates p53 stability *via* direct methylation, which in turn impacts on liver cancer cell growth (7). It is, therefore, likely that the activity of

SETDB1 fuels the transformation process and could be considered as a target in cancer therapies.

SETDB1 has been shown to be part of a functional megacomplex with other H3K9 KMTs (G9A, GLP and SUV39H1), involved in transcriptional silencing (1). Interestingly, SETDB1 and SUV39H1 not only share common targets, including transposable elements such as LINEs and ERVs (11), but also coordinate the propagation of H3K9me3 in pericentromeric heterochromatin (9,10). Remarkably, the role of SETDB1 in transposon repression is highly relevant since retrotransposition of LINEs and ERVs has been described to induce cancer-associated processes, including genomic instability, disruption of regulatory elements and generation of mutations (18). Thus, we would like to better understand how the crosstalk between these two key H3K9 KMTs occurs in terms of interaction and recruitment to target loci.

Indeed, up to date, the regulation of the KMT complexes is not fully understood. Interestingly, PTMs have been implicated in the regulation of H3K9 KMT functions (2). G9A and GLP are known to methylate themselves (14) and our results point to SETDB1 being able to undergo methylation at K1770 and K1778 inside its catalytic domain (15). Thus, we aimed to better understand how methylation of SETDB1 regulate its activity (auto-methylation, tans-methylation within the KMT complexes), interaction with partners and recruitment to chromatin, which ultimately could impact gene expression and could have a link to cellular transformation.

My thesis project was focused on two axes:

1- Understand how SETDB1 auto-methylation regulates its activity, interaction with partners and recruitment to chromatin, which ultimately could impact on gene expression.

2- Identify the SETDB1 and SUV39H1 co-regulated genomic targets and investigate the role of SETDB1 methylation in this regulation.

INTRODUCTION

CHAPTER 1. EPIGENETICS AND CHROMATIN STRUCTURE

This chapter gives an overview on the structure and compaction of chromatin and the most studied histone post-translational modifications. I focus on describing the main lysine methyltransferases that are involved in gene silencing: namely, those that methylate histone H3 at lysines 9 and 27 (H3K9 and H3K27, respectively). Then, I summarize their biochemical properties and biological functions, and discuss their recruitment to chromatin and how they mediate transcriptional silencing.

1. CHROMATIN STRUCTURE

Waddington originally defined the term "epigenetics" as the "branch of biology which studies the causal interactions between genes and their products which bring the phenotype into being" (19). This definition implies the understanding of chromatin structure and its impact on gene function. Nowadays, the definition of epigenetics has evolved as "the study of heritable changes in gene expression that occur independently of changes in the primary DNA sequence" (20). This heritability of gene expression patterns is mediated by epigenetic modifications, which include DNA methylation, post-translational modification of histones, incorporation of histone variants as well as the positioning of nucleosomes along chromatin. Remarkably, failure in the proper maintenance of heritable epigenetic landscape can result in inappropriate activation or inhibition of various signaling pathways leading to diseases, such as cancer (21).

In mammals, the genomic information is organized into chromatin (22). Based on the level of compaction, chromatin is divided into euchromatin and heterochromatin (**Illustration 1**).

Euchromatin is generally gene-rich, less condensed, and associated with active gene transcription, whereas heterochromatin is generally highly condensed and mostly contains inactive genes. Two distinct subtypes of heterochromatin have been defined: facultative and constitutive heterochromatin. Facultative heterochromatin regions can alternate between different transcriptional states in a development-specific manner, and usually contain genes that must be kept silent upon developmental cues. The most prominent example of facultative heterochromatin is the inactive X chromosome in females (23). On the other hand, constitutive heterochromatin encompasses the regions containing a high density of repetitive DNA elements, such as clusters of satellite sequences and transposable elements at centromeres, pericentric foci, and telomeres (24) For instance, the silencing of repetitive or repeat-derived sequences has an important role in maintaining chromosome stability by regulating proper chromosome segregation and preventing recombination or transposition activity (25). Constitutive heterochromatin has long been considered to be a highly stable state, however, recent studies show that constitutive heterochromatin is also dynamically regulated and responsive to stimuli. For example, gene silencing at constitutive heterochromatin in fission yeast is less stable at elevated temperatures. Although these changes could potentially help organisms to adapt to new environments, in some cases they have also been associated with disease process (26).

DNA is tightly packaged into nucleosomes that form the structural and functional unit of chromatin (**Illustration 2**). Each nucleosome consists of an octamer of four core histone proteins organized as an $(H3/H4)_2$ tetramer and two H2A/H2B dimers positioned on both sides of the tertramer. Around the histone octamer core are wrapped 147 bp of DNA (27). Besides, the linker histone H1 binds the DNA entering and exiting the core nucleosome and protects the linker DNA, further compacting chromatin. The core histones (H2A, H2B, H3

and H4) have a similar structure, with a N-terminal domain, a globular domain organized by the histone fold, and a C-terminal tail. The N-terminal tails of histones are subject to a large number of post-translational modifications such as acetylation, methylation, phosphorylation, ubiquitination and sumoylation (28). These modifications regulate the ability of chromatin and DNA regulation machineries to access the underlying DNA, representing a fundamental regulatory mechanism for transcription, replication and genome stability (27).

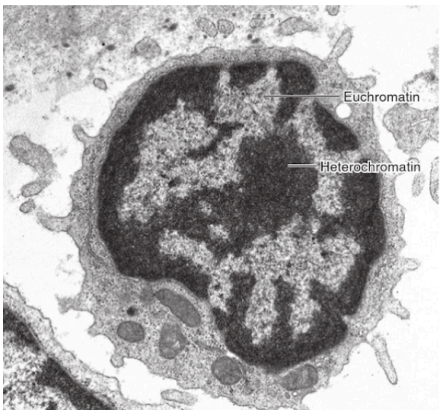

Illustration 1. Euchromatin and heterochromatin in the nucleus of a lymphocyte a seen by transmission electron microscopy. Chromatin presents different layers of compaction: euchromatin and heterochromatin. *From: Ronner P. 2016* (29)

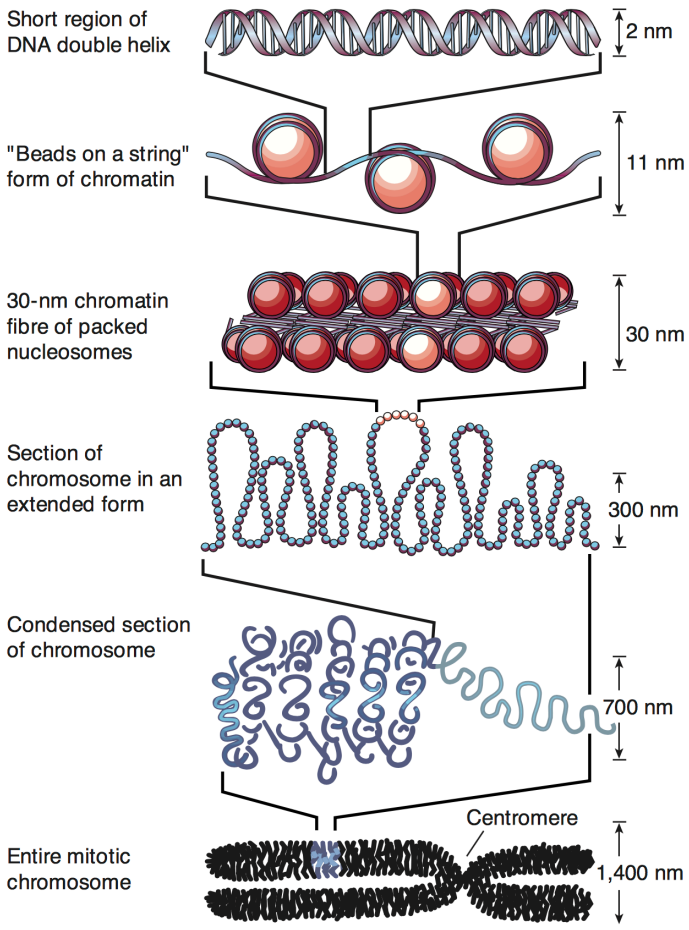

Illustration 2. Multiple levels of chromatin structure. DNA is wrapped around a histone octamer to form nucleosomes. Nucleosomes are connected by stretches of linker DNA. This basic nucleosome structure is folded into a fiber-like structure of about 30 nm in diameter. These 30-nm fibers are further compacted into higher-order structures. *From: Felsenfeld G, et al. 2003* (30).

2. POST-TRANSLATIONAL MODIFICATIONS OF HISTONES

In general, two mechanisms are thought to govern the function of PTMs. First, these different marks could affect the nucleosome-nucleosome or DNA-nucleosome interactions through the addition of molecules or by changing histone charges. Second, different marks could represent docking sites for the recruitment of specific proteins which could result in different functional outcomes. All these modifications define the proposed "histone code" that can alter higher-order chromatin structure and/or mediate subsequent targeting of transcriptional complexes (22) (**Illustration 3** and **Table 1**).

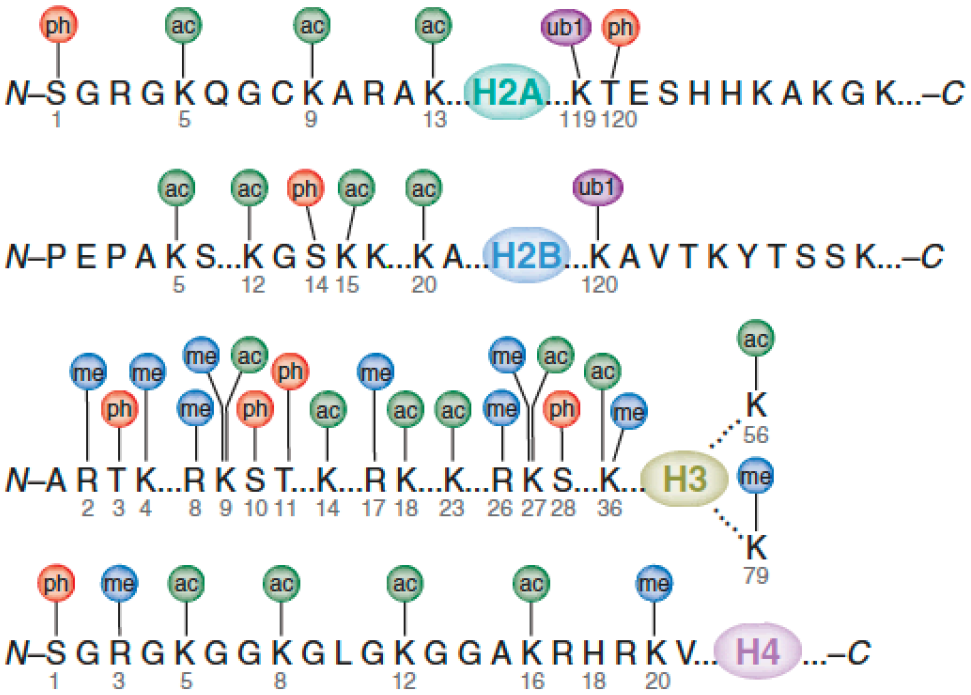

Illustration 3. Post-translational modifications of human nucleosomal histones. The modifications include acetylation (ac), methylation (me), phosphorylation (ph) and ubiquitination (ub). *From: Bhaumik SR, et al. 2007* (31).

Post-translational histone modifications associated with transcriptional repression		
Histone modification	**Residues modified**	**Enzymes**
Lysine deacetylation	H2AK5, H2AK9, H2BK5, H2BK12, H2BK15, H2BK20, H2BK120, H3K4, H3K9, H3K14, H3K18, H3K23, H3K27, H3K36, H3K56, H4K5, H4K8, H4K12, H4K16, H4K91	HDAC1, HDAC2, HDAC3, HDAC4, HDAC5, HDAC7, HDAC8, HDAC9, HDAC11, SIRT1, SIRT2, SIRT6, SIRT7
Lysine methylation	H3K9, K3K27, H4K20	EHMT1, EHMT2, EZH2, PRDM2, SETD8, SETDB1, SUV39H1, SUV39H2, SUV420H1, SUV420H2
Lysine demethylation	H3K4, H3K36	KDM1A, KDM1B, KDM2A, KDM2B, KDM4A, KDM4B, KDM4C, KDM4D, KDM5A, KDM5B, KDM5C, KDM5D, KDM8
Arginine methylation	H2AR3, H3R2, H3R8, H4R3	PRMT5, PRMT6, PRMT7
Arginine citrullination	H3R2, H3R8, H3R17, H4R3	PADI4
Serine phosphorylation	H2AS1	RPS6KA5
Lysine ubiquitylation	H2AK119	RNF2

Post-translational histone modifications associated with transcriptional activation		
Histone modification	**Residues modified**	**Enzymes**
Lysine acetylation	H2AK5, H2AK9, H2BK5, H2BK12, H2BK15, H2BK20, H2BK120, H3K4, H3K9, H3K14, H3K18, H3K23, H3K27, H3K36, H3K56, H4K5, H4K8, H4K12, H4K16, H4K91	ATF2, CREBBP, ELP3, EP300, HAT1, KAT2A, KAT2B, KAT5, KAT6A, KAT6B, KAT7, KAT8, NCOA1
Lysine methylation	H3K4, H3K36, H3K79	ASH1L, DOT1L, MLL, MLL2, MLL3, MLL5, NSD1, SETD1A, SETD1B, SETD2, SETD7, SMYD2, SMYD3
Lysine demethylation	H3K9, K3K27, H4K20	JHDM1D, KDM1A, KDM3A, KDM3B, KDM4A, KDM4B, KDM4C, KDM4D, KDM6A, KDM6B, PHF2, PHF8
Arginine methylation	H2AR3, H3R2, H3R8, H3R17, H3R26, H4R3	PRMT1, PRMT2, CARM1, PRMT5, PRMT6, PRMT7
Arginine citrullination	H3R26	PADI2
Serine phosphorylation	H2BS36, H3S10, H3S28	CHUK, PRKAA2, RPS6KA4, RPS6KA5, SLK
Threonine phosphorylation	H3T6, H3T11	PKN1, PRKCB
Lysine ubiquitylation	H2BK120	UBE2A, UBE2B, UBE2E1

Table 1. Major post-translational modifications on histone tails. *Adapted from: Huynh JL, et al. 2013* (32).

2.1. Histone methylation

Methylation is a covalent modification that represents the addition of a methyl group from the donor S-adenosylnethyonine (SAM) on the side-chain nitrogen atoms of lysine and arginine residues (**Illustration 4**). The enzymes that catalyze histone methylation have been grouped into three different classes: 1) the lysine-specific SET domain-containing histone methyltransferases involved in methylation of lysines 4, 9, 27, and 36 of histone H3 and lysine 20 of histone H4; 2) non-SET domain-containing lysine methyltransferases involved in methylation of lysine 79 of histone H3; and 3) arginine methyltransferases involved in methylating arginines 2, 8, 17, and 26 of histone H3 as well as arginine 3 of histone H4 (33).

Histone lysine methylation is one of the most extensively studied epigenetic marks. The majority of histone lysine methylation in mammals occurs on the histone H3 and H4 N-terminal tails. Histone lysine methylation effects depend on the amino-acid residue that is methylated, its degree of methylation and the positioning of the methylated nucleosome within the gene and the genome. For instance, lysine residues can be mono-, di- or tri-methylated (me1, me2 and me3, respectively) (34). In general, methylation of histone H3 lysine 4 (H3K4), H3K36 and H3K79 have been linked to active gene expression, whereas di- and tri-methylation on H3K9, H3K27 and H4K20 have been associated with gene silencing and/or heterochromatin formation (33).

Lysine methylation is reversible, with demethylation being carried out by two families of enzymes, amine oxidases such as LSD1 and hydroxylases of the JmjC family, with specific activities towards histone residues. As for example, LSD1, an amine oxidase that associates with histone deacetylase complexes can demethylate H3K4me2 and H3K4me1. The JMD2

family includes JMJD2A, which can both demethylate H3K9 and H3K36, JMJD2B and JMJD2D, which demethylate H3K9, JHDM1 that demethylates H3K36 (35).

Histone lysine methylation has been also associated with regulation of splicing (36). For instance, H3K36me3 marks highly transcribed exons and are more enriched in constitutive exons than in alternatively spliced ones (37). Moreover, local increases in H3K9me2 and H3K9me3 enhance exon inclusion, whereas H3K9 demethylation correlates with exon skipping (38).

Arginine methylation is catalyzed by a family of enzymes called protein arginine methyltransferases (PRMTs). There are three main forms of methylated arginines identified in mammalian cells: mono-methylarginines, asymmetric dimethylarginines and symmetric dimethylarginines. PRMTs are generally classified as type I, type II, or type III enzymes. Type I and type II catalyze the formation of a mono-methylarginine intermediate, then type I PRMTs (PRMT1, 2, 3, 4, 6 and 8) further catalyze the production of asymmetric dimethylarginines, while type II PRMTs (PRMT5 and 7) catalyze the formation of symmetric dimethylarginines (39). To date, two proteins have been reported to have a potential arginine demethylation activity: the Jumonji domain containing protein 6 (Jmjd6) and the peptidyl arginine deiminase 4 (PAD4) (40). Notably, due to the methylation state of individual arginine residues, the adjacent chromatin region is either transcriptional active or repressed (40).

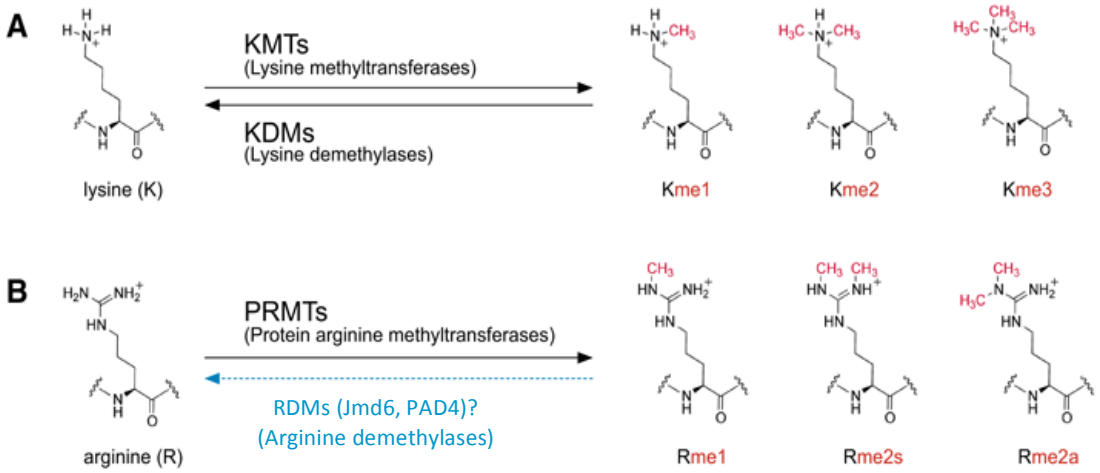

Illustration 4. Chemistry of lysine and arginine methylation. **A)** Mono-, di-, or tri-methylation (Kme1, Kme2, and Kme3) of lysine at the ε-amino group is catalyzed by lysine methyltransferases (KMTs). This modification is reversible and can be erased by lysine demethylases (KDMs). **B)** The protein arginine methyltransferases (PRMTs) can modify the guanidinium group of arginine residues and result in either mono-methylated (Rme1) or symmetric (Rme2s) or asymmetric (Rme2a) di-methylated arginine. Currently, arginine demethylation activity has been described for Jumonji domain containing protein 6 (Jmjd6) and peptidyl arginine deiminase 4 (PAD4). *Adapted from: Wesche J, et al. 2017* (40).

2.2. Histone acetylation

Lysine acetylation is a highly abundant mark that is known to regulate many cellular processes including transcription. Acetylation of histones reduces the positive charge of the histone tails, thereby loosening the interaction between histones and DNA, allowing access of transcription factors for active transcription (22). Several lysine residues located in histone tails are acetylated by different families of histone acetyltransferases (HATs) including the GNAT family (Gcn5, PCAF, Hat1, Elp3, and Hpa2); the p300/CBP family (p300, and CBP); the MYST family (Esa1, MOF, Sas2, Sas3, MORF, Tip60, and Hbo1) and Rtt109 (22). Likewise, acetylation is reversed by histone deacetylases (HDAC). There are three distinct families of HDACs: the class I and class II histone deacetylases and the class III NAD-dependant enzymes of the Sir family. They are involved in multiple signaling pathways and they are present in numerous repressive chromatin complexes. Remarkably, acetylation of H3K9 and H3K27, have been found primarily at active enhancers, regulatory DNA regions that dictate the spatial and temporal patterns of gene expression during development (41).

2.3. Histone phosphorylation

Phosphorylation occurs on serine, threonine and tyrosine residues predominantly, but not exclusively, in the N-terminal histone tails. Histone kinases transfer a phosphate group from ATP to the hydroxyl group of the target amino-acid side chain (22). Regarding histone phosphorylation associated with transcriptional regulation, phosphorylation of H3S10 and H3T11 was shown to enhance its interaction with Gcn5 at promoters of Gcn5-dependent genes such as the cell- cycle regulators *cyclin B* and *cdk1*, leading to increased H3K9 and K14 acetylation and stimulation of transcription (42). Another important event is phosphorylation

of the variant histone H2A(X) at serine 139 that is a key event during cell cycle and DNA damage response. This modification accumulates around the DNA double strand breaks to recruit DNA damage repair and signaling factors, including the critical mediator protein MDC1 (42).

2.4. Histone ubiquitination

Ubiquitin is a 76-amino acid polypeptide protein highly conserved in eukaryotes. Ubiquitination refers to the post-translational modification of the lysine residue by the covalent attachment of one (monoubiquitination) or more (polyubiquitination) ubiquitin monomers. Typically, polyubiquitination marks a protein to be degraded *via* the 26S proteasome, whereas monoubiquitination modifies protein functions. The dominant form of ubiquitinated histones are monoubiquitinated H2A (H2Aub) and H2B (H2Bub). A single molecule of ubiquitin is added to the highly conserved lysine residues: K119 for H2A, and K120 for H2B in eukaryotes (43). Monoubiquitination of H2A and H2B have been clearly implicated in transcriptional regulation. H2Aub occupation is more frequently correlated with gene silencing, while H2Bub is mostly associated with transcription activation, where the polycomb group protein RING1B-mediated H2Aub is required for gene silencing (44). In contrast, H2Bub occupation is strongly correlated with active gene expression in most cases, likely through multiple mechanisms, including promoting other active histone modifications and RNA pol II elongation. In a ChIP-on-Chip experiment, H2B monoubiquitination was found in the transcribed regions of highly expressed genes (45).

Many studies have demonstrated that histone ubiquitination crosstalk with other histone modifications and they act in combination or sequentially to regulate transcription. For example, H2B monoubiquitination is required for both H3K4 and H3K79 methylation (46).

2.5. Histone sumoylation

Sumoylation consists in the addition of a "Small Ubiquitin-related MOdifier protein" (SUMO) of ~100 amino acids to a target protein. It is one of the most essential post-translational modifications and plays a key role in various physiological and pathological processes. SUMO exists in four isoforms named SUMO-1, -2, -3, which are 12 kDa in size, and share a structural similarity with ubiquitin (47). Protein sumoylation does not lead to proteosomal degradation; in contrast, it can control protein-protein interactions, cellular localization, enzymatic activity and protein stability (47). On the other hand, desumoylation is mediated by specific cysteine-protease Sentrin-specific proteases (SENPs), which exhibit specific proteolytic activity for the SUMO-lysine isopeptide bond (48). Emerging evidence indicates that SUMOylation of chromatin-bound proteins plays an important role in transcription and may impact KMT-dependent regulation of heterochromatin structure. For example, in fission yeast, the E2 conjugating enzyme Ubc9 binds chromatin and promotes SUMOylation of Clr4 (Suv39 orthologue) to promote its activity and maintain gene silencing at heterochromatic regions (49). Likewise, SUMO2/3 directly interact with human MCAF1/SETDB1 complex suggesting a requirement for SUMOs for heterochromatin assembly (50). Remarkably, sumoylation is also known to regulate the enzymatic activity of histone-modifying enzymes such as histone deacetylases and EZH2 (51).

3. EPIGENETIC REGULATION BY WRITERS, ERASERS AND READERS

Histone modifications are reversible and are dynamically regulated through the dual action of enzymes that have become known as “chromatin writers” and “chromatin erasers” (**Illustration 5**). The group “writers” includes proteins that establish histone modifications such as the family of lysine methyltransferases. Conversely, chromatin erasers, such as lysine demethylases, catalyze the removal of histone methylation marks (**Illustration 6**).

Interestingly, most of the KMTs have at least one domain, such as PHD, bromo, chromo and Tudor, which confers the ability to recognize different PTMs. Such domains facilitate the coupling of “writer” with “reader” properties, which allows KMTs to identify specific histone PTMs and catalyze methylation, indicating the existence of protein crosstalk mechanisms for the establishment and/or propagation of different histone marks (33). For instance, HP1 proteins contain a chromo-domain which allows the interaction with H3K9me3 and the formation of heterochromatin (52).

Crosstalk involving histone modifications have been described. For instance, competition between different PTMs for the same amino acid. This is the case for a lysine that can be either acetylated or methylated, and these two modifications are mutually exclusive. Moreover, the presence of one modification may inhibit the interaction of a chromatin reader and the establishment of marks at neighboring residues. Phosphorylation of H3S10 disrupts the interaction between H3K9me3 and its reader HP1 during mitosis (53). Finally, the establishment of a modification may depend on the presence of another modification. For example, ubiquitylation of H2B is required for H3K4 and H3K79 methylation (46).

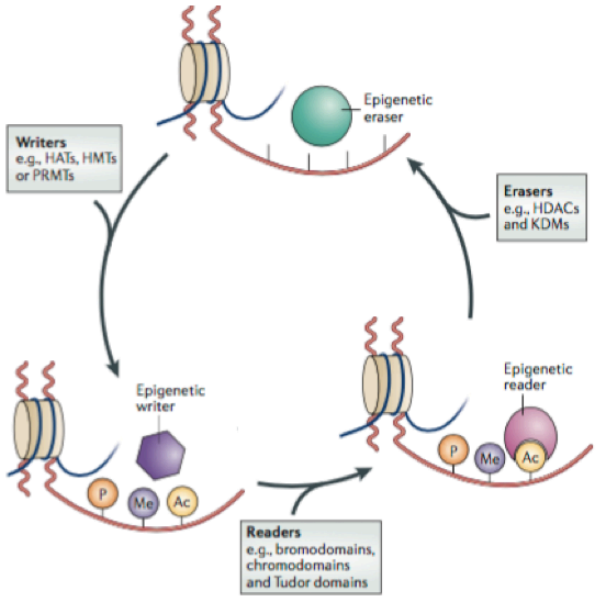

Illustration 5. Epigenetic writers, readers and erasers. Epigenetic regulation is a dynamic process. Epigenetic writers such as histone methyltransferases (HMTs) deposit epigenetic marks on residues on histone tails. Epigenetic readers such as proteins containing bromodomains, chromodomains and Tudor domains bind to these epigenetic marks. Epigenetic erasers such as lysine demethylases (KDMs) catalyze the removal of epigenetic marks. *Adapted from: Falkenberg KJ, et al. 2014* (54).

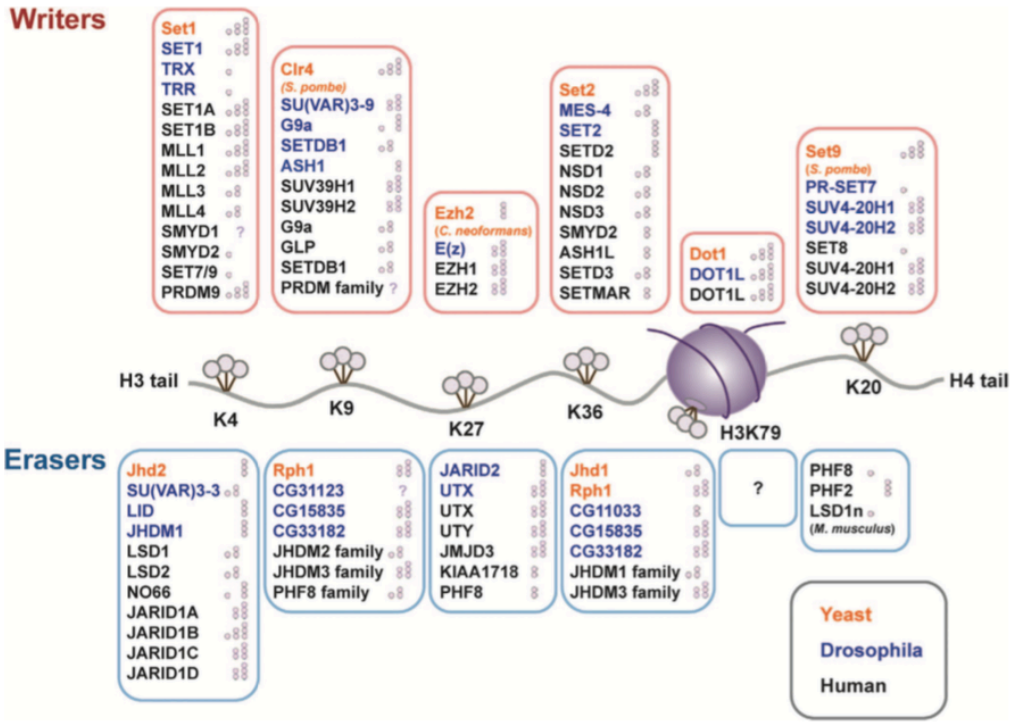

Illustration 6. Principal lysine methylation sites on histones H3 and H4. The reported writers (methyltransferases) and erasers (demethylases) for each lysine methylation are also depicted with their methylation state specificities: single circle: me1; double circle: me2; triple circle: me3. *From: Hyun K, et al. 2017* (55).

In general, active and repressive histone marks are not found in the same genomic region. However, in mESCs many key developmental genes are marked by the dual presence of activating H3K4me3 and repressive H3K27me3. Such bivalent domains poise differentiation genes for activation or stable repression following the removal, or a further gain, of H3K27me3, respectively (56). Bivalent domains gain or lose H3K27me3 following differentiation of ES cells, and this correlates with the progressive restriction of differentiation potential (57) (**Illustration 7**). Mechanistically, the presence of PRC2 at bivalent promoters interferes with transcription by keeping Pol II in a paused state (58).

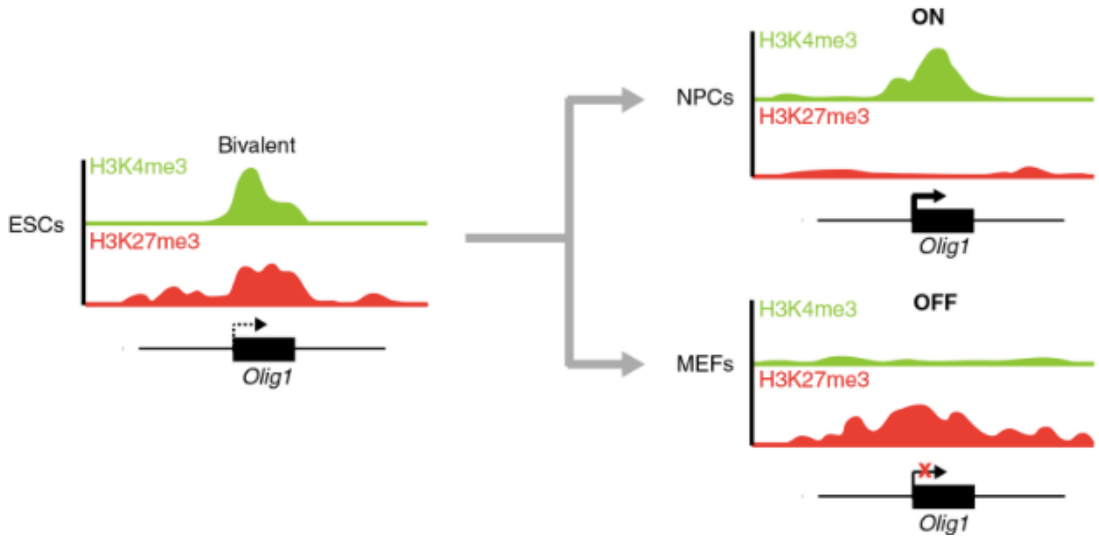

Illustration 7. Bivalent chromatin. Bivalent chromatin domains are characterized by the coexistence of active H3K4me3 and repressive H3K27me3 histone marks and are often found at promoter regions of developmentally regulated genes in pluripotent cells. Upon differentiation, this poised chromatin state is altered, which results in either gene activity or gene repression. This concept is illustrated for the "bivalent" *Olig1* gene. ECSs (Embryonic stem cells), NPCs (Neural Precursor Cells), MEFs (mouse embryonic fibroblasts). *From: Allis D, et al. 2015* (59).

4. LYSINE METHYLTRANSFERASES (KMTs)

KMTs can transfer one, two or three methyl groups from S-adenosyl-L-methionine to the ε-amino group of a lysine residue. Most KMTs contain a SET domain, which is a highly conserved peptide chain containing 130 amino acids. It was named after the initials of the three proteins first discovered in *Drosophila melanogaster*: Suppressor of variegation 3-9 (Su(var) 3-9), Enhancer of zeste (E(z)) and Trithorax (Trx) (60). In general, the SET domain possesses catalytic activity towards the ε-amino group of lysines. Moreover, there are KMTs characterized by the presence of the pre-SET domain that function to maintain the structural stability of the protein and the post-SET domain, which is involved in binding to histone tails (or other substrates) and interaction with cofactors (61). SET domain-containing proteins have now been found in all studied eukaryotes. Currently, six families of SET domain-proteins are distinguished: the SUV39 family of Su(var)3-9 homologues; the E(z) homologue (EZH); the SET-domain containing 1 (SET1) and SET2 family; the PR domain-containing (PRDM); the SET and MYND domain-containing (SMYD) families (33).

Interestingly, many KMTs have been shown to display catalytic activity towards particular lysine residues on histones, suggesting that there are recognition motifs for catalysis. Although many non-histone substrates have been identified for many KMTs, (detailed in below), the effect of methylation on most non-histone proteins are still being investigated. However, in some cases, methylation of non-histone proteins can regulate signaling pathways, modify the chromatin/DNA binding of the target protein, or affect the interaction with chromatin modifiers (33,62).

Hereafter, I chose to detail only a few families of KMTs to illustrate their different roles in gene silencing, since my thesis project aimed to better understand the crosstalk between two H3K9 KMTs (SETDB1 and SUV39H1).

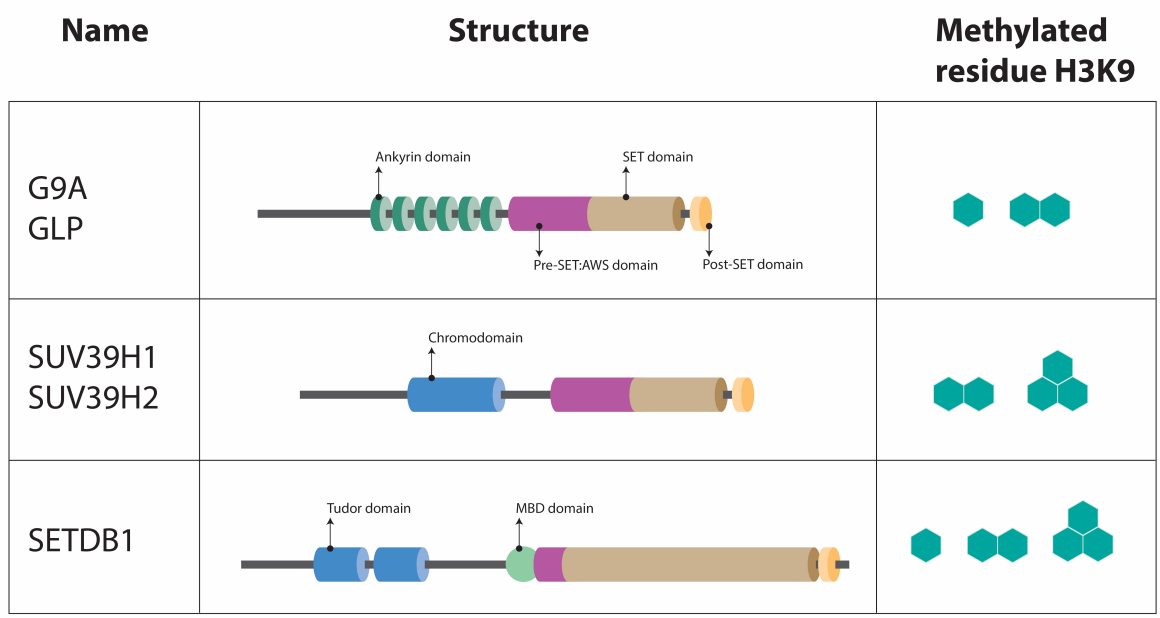

Illustration 8. The domain composition of the SUV39 and EZH families of lysine methyltransferases (KMTs). For each KMT, alternative names and corresponding domain structures are shown. The histone substrate specificities of the different KMTs, with the number of icons corresponding to the degree of lysine methylation are also shown are shown. *Adapted from Mozzetta C, et al. 2015* (33).

4.1. H3K27 Methyltransferases

Polycomb repressive complex 2 (PRC2) has generally been recognized to be responsible for the maintenance rather than the establishment of transcriptional silencing by of H3K27me2 and H3K27me3 (33). H3K27me3 mark correlates with gene silencing, and the distinct genomic locations of H3K27me3 and the active mark H3K36me3 support this idea. While H3K27me3 has been mostly associated to CpG-rich promoters and involved in the cell-specific maintenance of epigenetic silencing (57), H3K27me2 has been instead suggested to exert a protective function by inhibiting the firing of non-cell type-specific enhancers.

Two main H3K27 methyltransferases have been described so far in mammals: EZH1 and EZH2 (**Illustration 8**). These are the catalytic subunits of PRC2, which are responsible for H3K27 di- and trimethylation. Other subunits which compose the core PRC2 complex include suppressor-of-Zeste 12 (SUZ12), embryonic ectoderm development (EED) and retinoblastoma-associated protein 46 (RBAP46) (63).

SUZ12 and RBAP46 have been shown to anchor the complex to chromatin via their binding to histone H3. This allows the catalytic components, EZH1 and EZH2, to catalyze the di- and tri-methylation of H3K27 from a mono-methylated substrate (64). Furthermore, PRC2 can also be recruited to H3K27me3 through the C-terminal domain of EED, reading H3K27me3 through its WD-40 domain. The molecular structure of PRC2 suggested that EED binds to one nucleosome and places EZH2 close to the following nucleosome, suggesting a mechanism for H3K27me3 spreading (65). An additional co-factor necessary for the stabilization of PRC2 complex is Jumonji/ARID domain-containing protein 2 (JARID2), a DNA-binding protein involved in the recruitment of PRC2 complex to chromatin (65).

Notably, the recruitment of PRC2 and subsequent establishment of the H3K27me3 mark on the inactive X chromosome is mediated by the lncRNA X inactive-specific transcript (*XIST*), which coats the chromosome during inactivation of the X chromosome in female cells (66). The lncRNA HOX transcript antisense RNA (*HOTAIR*) mediates PRC2 recruitment in *trans* to the *HOXD* gene cluster, thus ensuring the proper temporal and spatial expression of HOX genes during development (67).

4.2. H3K9 Methyltransferases

Two families of H3K9 KMTs have been characterized: six members of the SUV39 subfamily and four members of the PRDM subfamily. SUV39 subfamily includes SUV39H1, SUV39H2, SET domain bifurcated 1 (SETDB1), SETDB2, G9A (also known as EHMT2) and G9A-like protein 1 (GLP; also known as EHMT1) (**Illustration 8**). Likewise, PRDM subfamily includes PRDM2, PRDM3, PRDM8, and PRDM16 (33). Importantly, our lab showed that four members of the SUV39 subfamily (G9A, GLP, SETDB1, and SUV39H1) form a functional megacomplex which cooperates in *de novo* gene silencing probably by sequential methylation of H3K9 (1).

4.2.1. G9A and GLP

In addition to the SET domain, G9A and GLP harbor ankyrin repeats, which are involved in protein-protein interactions. The SET domain is responsible for the transfer of methyl groups on histone H3, whereas the ankyrin repeats of G9A and GLP bind H3K9me1 and H3K9me2 marks (68). For instance, G9A and/or GLP deficient cells lacked nearly all the H3K9me2 and showed a significant decrease in H3K9 methylation in euchromatin (69,70). Overall, G9A and

GLP are responsible for H3K9 methylation at the euchromatin and facultative heterochromatin and are mainly involved in transcriptional repression (33).

G9A and GLP have been suggested to be chromatin compaction gatekeepers due to their capacity to interact with and recruit the H1 linker histone and the chromodomain-containing HP1 proteins. H1 positioning affects chromatin compaction, as it prevents nucleosome sliding. Methylation of H1.4K26 by G9A was suggested to create a binding site for HP1, promoting chromatin condensation (71). Furthermore, G9A and GLP are capable of automethylation, which can be recognized by HP1 and mediates its recruitment to chromatin and binding to G9A- and GLP-catalyzed H3K9me2 marks (72).

G9A and GLP are essential during development. In mouse embryos, loss of either G9A or GLP protein results in severe growth retardation and early lethality (73). Also, silencing of pluripotent markers such as *Oct4* and *Nanog* is mediated by G9A-dependent deposition of H3K9me2 and *de novo* DNA methylation. Consequently, G9A-deficient ESCs showed normal self-renewal but exhibit prolonged expression of *Oct4* and *Nanog* (74).

4.2.2. SUV39H1

Position-effect variegation (PEV) results when a gene generally positioned in euchromatin is juxtaposed with heterochromatin by rearrangement or transposition. Studies of PEV in *Drosophila melanogaster* have resulted in the identification of some PEV suppressors including Su(var)3–9 and its human ortholog SUV39H1 (27,75). SUV39H1/2 have two functionally distinct domains, the N-terminal chromodomain and the C-terminal SET domain. The chromodomain functions as a binding module that targets H3K9me3 marks, and the SET domain is responsible for SUV39H1/2 enzymatic activity.

SUV39H1 or SUV39H2 single knockout mice display normal viability and fertility, but double mutants are born at sub-Mendelian ratios with growth retardation, male sterility and present impaired formation of heterochromatin associated with genomic instability (76).

SUV39H1/2 were shown to be responsible for the establishment of H3K9me3 at pericentromeric repeats as well as at telomeric, subtelomeric regions and LINE elements (11,77). At pericentromeric repeats, SUV39H1/2 can recognize H3K9me1 through their chromodomain and to specifically catalyze H3K9 trimethylation (76,78). In fact, SETDB1 (see below) can monomethylate H3K9 and thereby provides the substrate for SUV39H1 to establish H3K9me3 (see below). Thereafter, HP1α which is a reader of H3K9me2 and H3K9me3 forms a stable complex with SUV39H1 and leads to the spreading and maintenance of H3K9me3 at heterochromatin. This propagation would involve a 'self-sustaining' loop, in which methylated H3K9 histones bind to HP1, which in turn recruits more H3K9 KMTs (52).

Interestingly, in fission yeast, H3K9 methylation is catalyzed by the SUV39H1 ortholog Clr4. Clr4-mediated H3K9 methylation serves to recruit Swi6, the fission yeast ortholog protein of mammalian HP1. It has been suggested that Swi6 mediates the recruitment of additional proteins that are required for heterochromatin formation (79).

It has also known that SUV39H1 and SUV39H2 are linked to the silencing of specific genes located in euchromatin. For instance, SUV39H1 and SUV39H2 can be recruited by retinoblastoma (Rb) protein to the promoters of several S-phase genes and repress their expression (80).

4.2.3. SETDB1

In general, *SETDB1* is a 36-kb single copy gene, and full-length SETDB1 transcript consists of 22 exons. However, there is a splicing variant which consists of only the first 12 exons and it lacks sequences encoding the MBD domain and the catalytic SET domain (81). Besides the catalytic SET domain, SETDB1 possesses two Tudor domains, a putative methyl CpG-binding domain (MBD) and a bifurcated domain, which is an insertion of 347 amino acids long in the SET domain. The two adjacent Tudor domains participate in binding to methylated Lys, whereas the bifurcated domain might be involved in regulation of the catalytic activity of the enzyme (33). Although MBD-containing proteins have been characterized as transcriptional repressors, the function of MBD domain in SETDB1 is still unknown (82).

SETDB1 activity is essential during development and cellular differentiation. SETDB1 is essential in mouse ESC pluripotency and self-renewal (3,4,83), and *Setdb1* KO is lethal at the peri-implantation stage at 7.5 days postcoitum (dpc) (5). ESCs cannot be derived from *Setdb1*-null blastocysts *in vitro* (5), and SETDB1 knockdown in mouse ESCs results in loss of Oct4 expression and abnormal expression of various differentiation markers (3,84). Consistently, SETDB1 has been found to occupy and silence trophoblastic and developmental genes in ESCs, in cooperation with Oct3/4 (4,83,84). SETDB1 is also essential for pluripotency and differentiation of many progenitor cell types. SETDB1 is required for spermatogenesis (85) and early neurogenesis in mice by promoting proliferation and/or cell survival of progenitor cells (86) (**Illustration 9**).

SETDB1 is responsible for H3K9me1/2/3 deposition in euchromatin and facultative heterochromatin, being mainly involved in transcriptional repression. Moreover, SETDB1 interacts with the chromatin-associated factor ATF7IP (MCAF) to allow the transition from

H3K9me2 to H3K9me3 state (62).

SETDB1-mediated H3K9me3 deposition seems to be crucial for the formation of heterochromatin at interspersed repeats, in particular at class I and II retrotransposons in ESCs (87), and recently has been reported also in MEFs (88). The mechanism of heterochromatin formation at retrotransposons is complex because it varies with the class of elements and the developmental stage. However, it was shown that SETDB1 could play a crucial role in depositing H3K9me3 at ERVs during the developmental stages characterized by the absence of DNA methylation, such as early embryonic and germline development (89).

SETDB1 is tightly associated with KAP1 protein (90). KAP1 is a multi-domain protein defined as a platform for the recruitment of repressive enzymes. It contains an N-terminal Ring finger and a coiled-coil domain (RBCC domain), a C-terminal PHD-bromodomain and a central HP1-binding domain (91). The RBCC domain is crucial for the interaction with the N-terminal part of KRAB-Zinc finger proteins (KRAB stands for Kruppel-Associated box), a family of proteins able to bind DNA through their C2H2 zinc finger domain. The interaction of KAP1 with KRAB-zinc finger proteins allows the recruitment of KAP1 at specific loci. KAP1 then recruits SETDB1 as well as the catalytic subunit of the Nuclear Remodelling and Deacetylation Complex (NuRD complex), promoting the compaction of chromatin (92,93).

SETDB1 is not exclusively targeted towards histones. So far, four non-histone substrates have been described for SETDB1: ING2 (6), HIV Tat (94), p53 (7) and UBF (8) proteins (see below for more details). Interestingly, SETDB1 is the only H3K9 KMT showing distinct subcellular distributions (95,96), however the mechanistic understanding of which remains elusive.

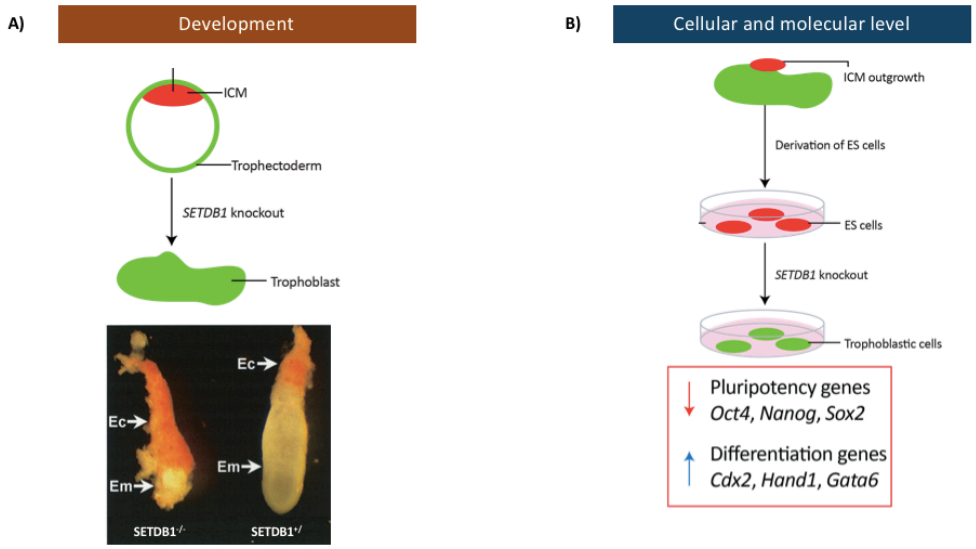

Illustration 9. SETDB1 is required for embryonic stem cells self-renewal and viability. A) In mice embryos, loss of SETDB1 results on failure in the development of the pluripotent inner cell mass (ICM), only the trophoblast is developed. Lower part: An SETDB1$^{-/-}$ embryo (left) and SETDB1$^{+/-}$ littermate (right) at 7.5dpc. SETDB1$^{-/-}$ embryo is severely malformed, with only the ectoplacental cone discernible (Ec). **B)** In ESCs derivated from normal inner cell mass, the loss of SETDB1 results on differentiation into trophoblastic cells. At molecular level, there is down-regulation of genes associated with pluripotency, and up-regulation of genes associated with differentiation. *Modified and adapted from: Dodge JE, et al. 2004* (5) *and Yeap, et al.. 2009* (3).

5. TARGETING MECHANISMS OF H3K9 KMTs

H3K9 methyltransferases potentially interact and are (co)-recruited to their genomic target loci by interacting with sequence-specific transcription factors, other chromatin binding proteins and non-coding RNAs (ncRNAs) (**Illustration 10**).

5.1. Transcription factor-mediated recruitment

H3K9 methyltransferases can be (co)-recruited to their genomic target loci by interacting with sequence-specific transcription factors. For instance, the zinc-finger proteins (ZFPs) ZNF518A and ZNF518B mediate the structural association between G9A, GLP and PRC2 and probably their genomic targeting (97). In addition, G9A and GLP can be preferentially targeted to promoters by the ZFPs widely interspaced zinc-finger-containing protein (WIZ) and RE1-silencing transcription factor (REST) (98). Furthermore, SETDB1 can interact with pluripotency transcription factors such as POU class 5 homeobox 1 (POU5F1) (84) and ZNF274 (99). Likewise, association of SUV39H1 with transcription factors such as RUNT-related transcription factor 1 (RUNX1) and RUNX3 provides a mechanism for gene silencing (100,101).

5.2. Recruitment by other chromatin-binding proteins

Chromatin-binding proteins that interact with H3K9 KMTs can regulate their genomic targeting. JARID2 interacts with G9A and GLP, and its binding overlaps to some extent with G9A-binding sites, suggesting that it could co-recruit PRC2, G9A and GLP (102). Similarly, CDYL, which has been described as a potential reader of both H3K9me2 and H3K27me3

(103), mediates the recruitment of both PRC2 and G9A, pointing to co-recruitment of PRC2 and H3K9 methyltransferases as a mechanism of gene silencing (104).

SUV39 subfamily members interact with the transcription co-repressors including KAP1 (92), methyl-CpG-binding domain protein 1 (MBD1) (105), ATF7IP (106) with DNA (cytosine-5)-methyltransferase 3A (DNMT3A) and DNMT3B (77,107) and with NAD-dependent protein deacetylase sirtuin (108). Binding of SUV39H1, SUV39H2 and HP1 to methylated H3K9 at pericentromeric heterochromatin is stabilized by their interactions with other heterochromatin components such as methyl-CpG-binding protein 2, MBD1 and sentrin-specific protease 7 highlighting the importance of keeping multiple interactions between various chromatin-binding proteins in the generation of heterochromatin nucleation sites. Remarkably, in fission yeast, H3K9 methylation is catalyzed by the SUV39H1 ortholog Clr4. Clr4-mediated H3K9 methylation serves to recruit Swi6, the fission yeast ortholog protein of mammalian HP1. The Swi6 oligomers provide a platform for the recruitment of other heterochromatin components, and this results in heterochromatin spreading (79)

5.3. Non-coding RNA-mediated recruitment

Long ncRNAs (lncRNAs) and short ncRNAs are important regulators of the recruitment of H3K9 KMTs to chromatin. For instance, lncRNAs can bind different KMTs to control genomic imprinting (for details see chapter imprinting). Furthermore, interaction of H3K9 KMTs with member of the Argonaute family has been also reported. Notably, Argonaute family has been shown to be essential components of RNA-induced silencing complexes, which, in conjugation with short interfering RNAs, target genes for silencing.

SETDB1 was found to interact with Argonaute protein 2 (AGO2) in the nucleus and to regulate AGO2-mediated transcriptional gene silencing. AGO2-mediated recruitment of SETDB1 depends on small antigene RNAs (agRNAs) that are complementary to the promoter of the androgen receptor gene. Furthermore, G9A and SUV39H1 were also identified as part of an AGO2 chromatin-associated complex that is involved in promoting pre-mRNA splicing by methylating H3K9 to slow down the transcription machinery and thus facilitate splicing (109). Moreover, AGO-dependent recruitment of G9A and GLP to mRNA 3′ untranslated regions was implicated in termination of transcription.

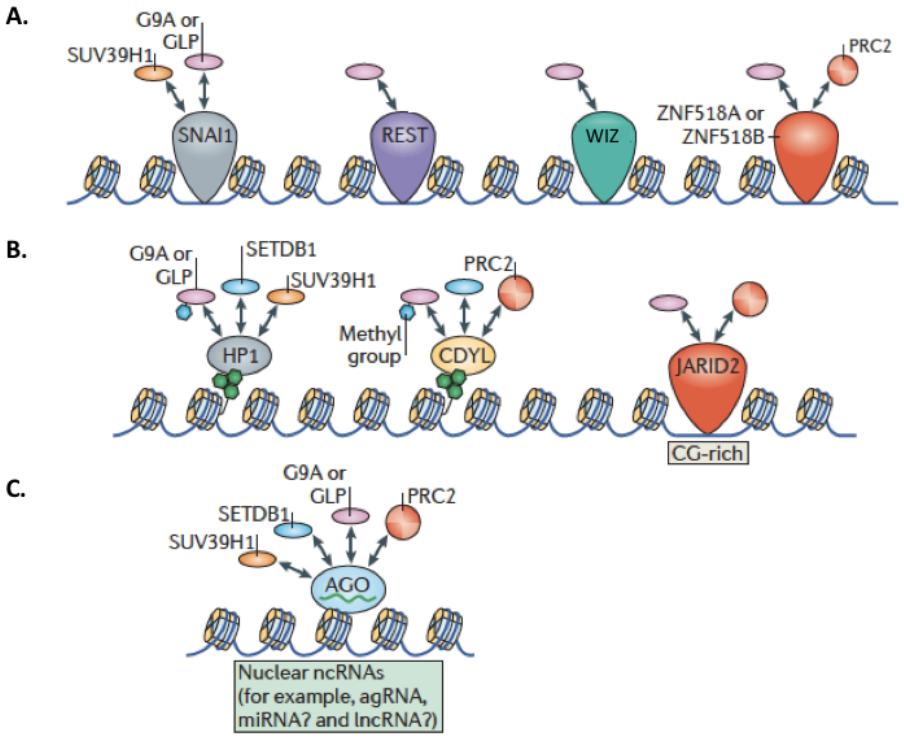

Illustration 10. Targeting mechanisms of H3K9 methyltransferases. A) Sequence-specific transcription factors, such as SNAI1, REST, WIZ, zinc-finger protein 518A (ZNF518A) and ZNF518B interact with and recruit H3K9 KMTs. **B)** The chromatin-binding proteins CDYL, HP1 and JARID2 interact with and probably mediate the recruitment of both PRC2 and H3K9 KMTs to their genomic targets. Auto-methylation of G9A and GLP is involved in mediating protein–protein interactions. **C)** Argonaute (AGO)-associated nuclear non-coding RNAs (ncRNAs) mediate the genomic targeting of PRC2 and/or H3K9 KMTs. agRNA: antigene RNAs; miRNA: microRNA. *Modified from: Mozzetta C, et al. 2015* (33).

CHAPTER 2. EMERGING ROLES OF NON-HISTONE PROTEIN LYSINE METHYLATION

As was described in the chapter 1, PTMs of histone proteins are key to the regulation of chromatin dynamics. Notably, many reports have revealed that KMTs and KDMs also regulate the methylation dynamics of non-histone proteins. However, the biological and physiological significance of non-histone lysine methylation/demethylation balance is still under investigation. In this chapter, I highlight the non-histone KMTs substrates with emphasis on H3K9 KMTs.

1. KMTs THAT ACT ON NON-HISTONE SUBSTRATES

Currently, there are three KMTs that have more known non-histone substrates than other enzymes; those are SMYD2, SETD7 and G9A. All of them have not only an important role in epigenetic regulation of gene expression but also in the regulation of cellular signal transduction pathways. Remarkably, some of the non-histone substrates have important roles in the development of different diseases including cancer.

1.1. SMYD2

The KMT "SET and MYND-domain containing 2" or SMYD2 was shown to methylate both H3K36 and H3K4 and act as a transcriptional coactivator. The reported non-histone substrates of SMYD2 include p53, ERα, RB and the chaperone protein HSP90. SMYD2 methylates p53 K370, and this methylation has been associated with repression of p53-mediated transcriptional regulation (110). Moreover, SMYD2-dependent p53K370 methylation impairs the expression of *CDKN1A*, a downstream target of p53, suggesting that SMYD2 represses

the function of p53 through K370 mono-methylation. Moreover, p53K372 methylation by SETD7 seems to inhibit SMYD2-depedent p53K370 methylation through blocking the interaction between SMYD2 and p53 (111).

Estrogen receptor-α (ERα), a ligand-activated transcription factor involved in human breast cancer, can be methylated by SMYD2 at K266. Upon estrogen stimulation, ERα-K266 methylation is diminished, thereby enabling p300/cAMP response element-binding protein-binding protein to acetylate ERα at K266, which is known to promote ERα transactivation activity (112).

The retinoblastoma tumor suppressor (RB) is a key cell cycle regulator and tumor suppressor and is dysfunctional in several cancer types. RB can be methylated by SMYD2 at lysine 860 and this modification permits direct binding of RB to the lysine methyl-binding protein L3MBTL1, which may alter the function of RB in cells (113).

The heat shock protein HSP90 is an evolutionarily conserved molecular chaperone that participates in the stabilization and activation of more than 200 proteins. Cancer cells use the HSP90 chaperone machinery to protect different mutated and overexpressed oncoproteins from misfolding and degradation. Therefore, HSP90 is recognized as a critical facilitator of oncogene addiction and cancer cell survival. Notably, SMYD2 methylates HSP90A on K531 and K574 and these methylation sites were shown to be important for the dimerization and chaperone complex formation (114).

1.2. SETD7

SETD7 (also known as KMT7 or SET7/9) was first identified as a histone H3K4-specific

KMT associated with transcriptional activation (115). However, later studies revealed that SETD7 knockdown in cells did not affect histone H3K4 methylation, indicating that SETD7 does not play a major role in maintaining this histone methylation (116). SETD7 methylates different non-histone proteins including p53, RB, SUV39H1 and the RelA subunit of NFκB transcription factor.

p53 is methylated by SETD7 at K372 in response to DNA damage, initiating cell cycle checkpoints and stimulating the acetylation of neighboring lysine residues in the regulatory domain of p53 by the acetyltransferase Tip60 (116,117). This acetylation prevents ubiquitin-mediated degradation of p53 by the proteasome, which can be associated with a longer half-life of p53.

SETD7 methylates K314 and K315 in the RelA subunit of NFκB, a transcription regulatory complex associated mainly with inflammatory response. This methylation has been associated with repression of the expression of NFκB target genes due to ubiquitination and proteasomal degradation of RelA (118). However, other study showed that methylation of K37 in the RelA subunit by SETD7 stimulates transactivation by NFκB through enhancing DNA binding by RelA (119). Although these results showed that SETD7 has opposing functions in NFκB regulation, they also suggest that methylation of RelA is tightly controlled to elicit the appropriate signaling output through the NFκB pathway.

SETD7 methylates the tumor suppressor RB and regulates RB-dependent cellular functions through modulating protein–protein interactions. For instance, SETD7-mediated methylation of RB at K873 promotes its interaction with HP1 (120). In addition, SETD7 methylates SUV39H1 at K105 and K123, and this methylation inhibits the H3K9 methyltransferase activity of SUV39H1 leading to heterochromatin relaxation and genome instability (121).

2. MOLECULAR FUNCTIONS OF LYSINE METHYLATION BY H3K9 KMTs

There are different biological processes affected by lysine methylation of non-histone substrates by H3K9 KMTs. Some examples include regulation of protein–protein interactions through binding to chromodomains, regulation of promoter binding affinity of transcription factors, thereby changing transcription levels of target genes, and regulation of nucleolar transcription (**Illustration 10**).

Concerning protein-protein interactions, methylated-lysine-specific binding proteins have been reported, and these proteins contain motifs that specifically recognize methylated lysine residues such as ankyrin domains, chromodomains and Tudor domains. For example, auto-methylation of G9A enables the binding of the chromodomain-containig protein, CDYL at the methylated site (72). Interestingly, a role for CDYL and its interaction with G9A in mediating REST tumor suppression function has been shown (122). Regarding modulation of transcription factors activity, Lysine methylation also regulates the binding affinity of transcription factors for promoters, which changes the transcription levels of target genes. For example, G9A methylates p53 at K373 and during DNA damage, the overall level of p53 modified at p53K373me2 does not increase, despite the dramatic increase in total p53, indicating that this methylation correlates with inactive p53 (123). In addition, lysine methylation by H3K9 KMTs also regulates nuclear chromatin condensation, which could result in the repression of rDNA transcription. For example, SETDB1 methylates the upstream binding factor (UBF) at the K232/254 residues. Accordingly, mutations at K232/254 blocked UBF trimethylation and derepressed the transcriptional suppression of rDNA by methylated UBF in response to SETDB1. In fact, methylated UBF has been correlated with deregulation of ribosomal transcription in Huntington disease (8).

Furthermore, there is evidence that lysine methylation by H3K9 KMTs can affect other PTMs. For example, an operational binary 'methyl/phospho switch' has been reported for K9/S10 region of histone 3. HP1 associates with the H3K9me3 mark, recruiting HP1 to heterochromatin regions and mediating the formation and maintenance of heterochromatin (124). However, during mitosis, phosphorylation of H3S10 by Aurora B disrupts the HP1–H3K9me3 interaction and, as a consequence, HP1 is released from heterochromatin, even though the H3K9me3 mark persists (53).

3. NON-HISTONE SUBSTRATES OF H3K9 KMTs

A summary of all non-histone targets of H3K9 KMTs described to date can be found in **Table 2.**

3.1. G9A

G9A is able to methylate itself, facilitating the binding of HP1 at the methylated site and creating a platform for recruitment of the transcriptional repressor CDYL (72,125). Furthermore, based on results from peptide arrays, additional non-histone substrates of G9A have been identified including WIZ, CDYL, CSB, ACINUS, HDAC1, DNMT1 and KLF12 (126). Notably, methylation of CDYL was found to alter its chromodomain binding to H3K9me3, perhaps by blocking its binding pocket for trimethylated lysine (126). Many transcription factors, such as C/EBPβ, p53, MEF2 and MyoD, are methylated by G9A, which appears to negatively impact their transcriptional activity. For instance, G9A methylates MyoD at K104 and blocks its transcriptional activity and expression of muscle differentiation genes (127). Similarly, methylation of the myocyte enhancer factor 2 (MEF2) at K267 by

G9A is involved in modulation of skeletal muscle cell differentiation (128). Furthermore, C/EBPβ methylation at K39 interferes with activation of myeloid genes (129,130). Additionally, p53 rendered transcriptionally inactive after it is dimethylated by G9A/GLP complex (123).

3.2. SUV39H1/2

SUV39H1 methylates CBX4 at K191. This methylation event is required for the binding of CBX4 to the ncRNA TUG1, which then targets methylated CBX4 to Polycomb bodies to repress transcription (131). A very recent study on Crl4 (SUV39H ortholog in the fission yeast *Schizosaccharomyces pombe*) showed that this protein undergoes auto-methylation at K455. Moreover, this auto-methylation promotes a conformational switch that enhances the enzymatic activity of Clr4 (132). On the other hand, Piao and collaborators recently reported that SUV39H2 is able to undergo auto-methylation at K392. Interestingly, SUV39H2 auto-methylation led to the impairment of its binding affinity to substrate proteins such as histone H3 and LSD1, suggesting that SUV39H2 may regulate interactions with, and dissociation from, substrates through its auto-methylation (133).

3.3. SETDB1

SETDB1 was also shown to exert enzymatic activity towards non-histone substrates. The UBF has been demonstrated to interact with and gets methylated by SETDB1, thus leading to nucleolar chromatin condensation and decreased rDNA transcriptional activity (8). Interestingly, SETDB1 has been found to be overexpressed in the neurons of patients diagnosed with Huntington disease (134). Intriguingly, this overexpression results not only in

elevated levels of H3K9me3 but also in high levels of UBF methylation by SETDB1 (8). This increased UBF methylation causes alterations in nucleolar function (8). Furthermore, ING2 (inhibitor of growth protein-2), a component of a HDAC complex has also been reported as a methylation target of SETDB1 (6). Moreover, methylation of the transcription factor p53 at K370 by SETDB1 was associated with cancer cell growth (7). Finally, SETDB1 has been reported to target the HIV-1 protein Tat at K50 and K51, leading to transcriptional repression of viral mRNA (94).

Given the fact that other H3K9 KMTs such as G9A, GLP and SUV39H1/2 are subject to auto-methylation and this process affects different biological pathways and processes, the present project not only studied the capability of SETDB1 to undergo auto-methylation, but also what is the impact on its activity, interaction with partners and recruitment to chromatin.

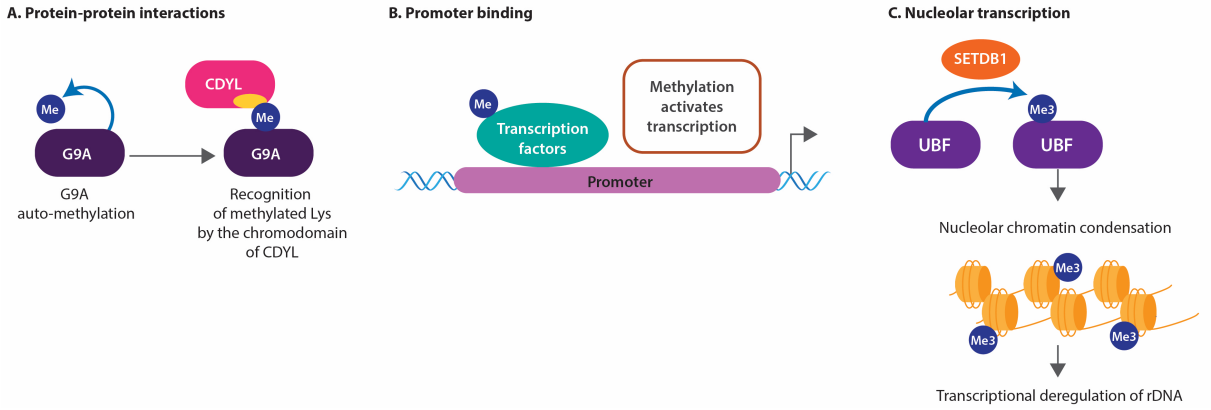

Illustration 11. Examples of molecular functions of non-histone lysine methylation by H3K9 KMTs. Three examples are shown. **A)** Lysine methylation regulates protein–protein interactions through binding to chromodomains or Tudor domains. This is the case of interaction of CDYL with the methylated form of G9A **B)** Lysine methylation regulates promoter binding affinity of transcription factors, thereby changing transcription levels of target genes. This is the case of p53, that can be methylated by G9A and SETDB1. **C)** Lysine methylation regulates nuclear chromatin condensation, which could result in the repression of rDNA transcription. This is the case of methylation of the upstream binding factor (UBF) by SETDB1. *Adapted and modified from: Hamamoto R, et al. 2015* (111) *and Lee J, et al. 2014* (135).

Function	Substrate	KMT	Biological outcome
Transcription factors	C/EBPβ	G9A	Inhibition of transactivation activity
	MyoD	G9A	Inhibition of transactivation activity
	MEF2D	G9A	Inhibition of transactivation activity
	p53	G9A/GLP	Inhibition of p53
	p53	SETDB1	Decrease on the recognition and degradation of p53 by MDM2
	UBF	SETDB1	Nucleolar chromatin condensation, decreased rDNA transcription
Chromatin modifiers	DNMT1	G9A	Unknown
	DNMT3	G9A/GLP	Binding of MPP8
	GLP	GLP	Binding of MPP8
	G9A	G9A	Binding of HP1 and CDYL
	SUV39H1	SUV39H1	Loss of heterochromatin domains and inhibition of growth
	SUV39H2	SUV39H2	Regulation of interaction with substrates
	SIRT1	G9A	Unknown
Chromatin- or DNA-binding proteins	CBX4/PC2	SUV39H1	TUG1 ncRNA-dependent recruitment to Polycomb bodies
	HP1α	SUV39H1	Unknown
	CDYL	G9A	Decreased interaction with H3K9me3
	WIZ	G9A	Unknown
	Acin1	G9A	Unknown
	RUVBL2	G9A	Negative regulation of hypoxia-inducible genes
Cell cycle regulators	ING2	SETDB1	Unknown
	MAGEH1	G9A	Unknown
Others	KIFC	G9A	Unknown
	UBE2D2	G9A	Unknown
	RDH11	G9A	Unknown
	CALN1	G9A	Unknown
	Tat	SETDB1	Inhibition of HIV transcription

Table 2. Known non-histone H3K9 KMTs substrates. *Modified and adapted from: Mozzetta C, et al. 2015* (33).

CHAPTER 3. EPIGENETIC REGULATION OF HETEROCHROMATIN ESTABLISMENT BY H3K9 KMTs

This chapter is focused on describing how H3K9 KMTs contribute to the chromatin repression. I provide two model systems including maintenance of genomic imprinting and repression of transposable elements. As I will present in the Results section, these two systems are affected by SETDB1 auto-methylation and are relevant to better understand the crosstalk between SETDB1 and SUV39H1.

1. DNA METHYLATION

DNA methylation is a covalent modification occurring mainly on cytosines. The methylation reaction is catalyzed by DNA methyltransferases (DNMTs), enzymes that transfer a methyl group from *S*-adenosyl-l-methionine (SAM) to the C5 of a cytosine (136). In mice, approximately 40% of promoters contain or are proximal to regions containing high frequencies of CpGs known as CpG islands. Methylation patterns are maintained during DNA replication and cell divisions and are vital in the regulation of gene expression, as well as other important processes such as X chromosome inactivation and genomic imprinting (137). In general, DNA methylation is established on unmethylated DNA through the synergistic action of DNMT3A and DNMT3B enzymes, together with their enzymatically inactive co-factor DNMT3L. The patterns of methylation are then maintained by DNMT1 which is able to copy these patterns from a hemi-methylated template during DNA replication (138).

In mice, the genome undergoes two waves of DNA methylation reprogramming (**Illustration 12**). These take place during gametogenesis and early embryogenesis. The parental-specific imprints must be ‘erased’ and new imprints reflecting the sex of the embryo are ‘established’

during germ cell development. Following fertilization, these imprint marks are 'maintained' as the cell propagates, except in the germ cells where the imprints are erased and re-established for the next generation (139).

Demethylation is in charge of the Tet family of proteins (Tet1, Tet2 and Tet3). These proteins have been implicated in the DNA methylation reprogramming process. Notably, Tet1 and Tet2 are expressed at high levels in mESCs (140) and in blastocysts (141), where 5hmC is relatively abundant. Furthermore, these enzymes play an important role in the maintenance of the pluripotent state (140). It has been shown that Tet proteins function in part to respond to the repressive function of the de novo DNMTs, which are also highly expressed in mESCs. In fact, genome-wide studies revealed that Tet1 and 5hmC are enriched in the promoter regions and gene bodies respectively, of many actively transcribing genes (142).

2. CROSSTALK BETWEEN DNA METHYLATION AND H3K9 METHYLATION

The interplay between DNA methylation and histone methylation is complex. For instance, Suv39h-mediated H3K9 methylation has been reported to direct DNMT3B through a HP1α-DNMT3B interaction to methylate major satellite repeats at pericentric heterochromatin (77). In addition, the absence of G9A and GLP results in loss of DNA methylation at pluripotency genes in ESCs (143,144). Notably, this loss of DNA methylation was associated with the absence of the recruitment of DNMT3A and DNMT3B by G9A and GLP (145). Moreover, Epsztejn-Litman and colleagues showed that DNMT3A and DNMT3B can be recruited by G9A, independently of its methyltransferase activity, based on its ankyrin repeat domain (145).

There is evidence showing that DNMTs recruitment is mediated by H3K9me3. For instance, one of the main recruiter of DNMT1, UHRF1, co-crystalizes with H3K9me3 through its Tandem Tudor Domain (TTD)-PHD domain (146). Subsequently, it was shown that the domain of UHRF1 protein called "Linker", localized between the TTD and PHD domain, was also crucial for H3K9me3 recognition and essential for the maintenance of DNA methylation at the level of chromatin (147).

DNA methylation maintenance is particularly important for repeated sequences that resist to DNA methylation reprograming, like transposable elements, as well as for genomic imprinting as I will detail it below.

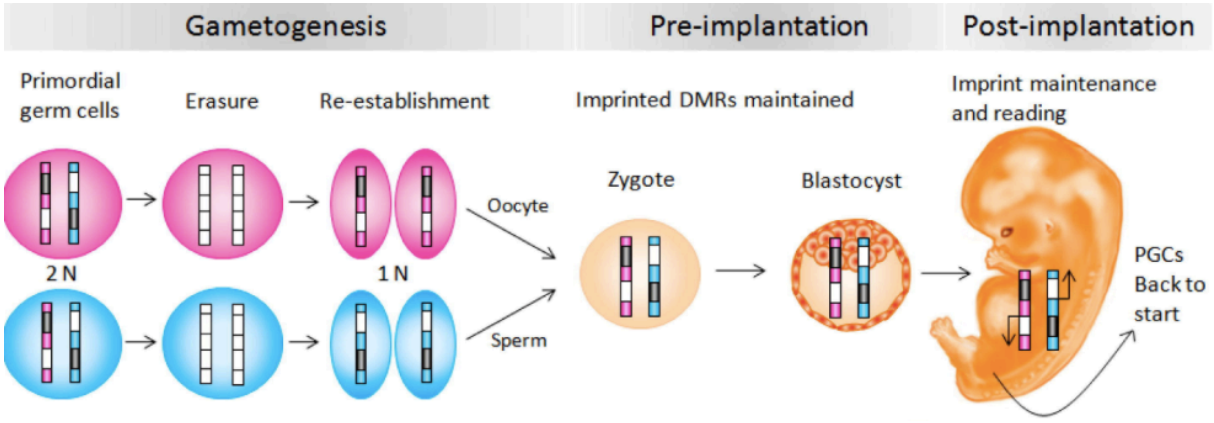

Illustration 12. DNA methylation reprograming during development. Methylation imprint erasure, re-establishment and maintenance at the germline differentially methylated regions (gDMRs) during gametogenesis and early embryonic development. Maternal and paternal chromosomes are represented in pink and blue bars, respectively. The black and white filled boxes on the chromosomes indicate the presence or absence of allelic modifications, respectively. *From: Ishida M, et al. 2013* (139).

3. GENOMIC IMPRINTING

Genomic imprinting is an epigenetic mechanism whereby gene expression is restricted to the allele inherited from either the maternal or paternal parent. In mammals, a small number of genes (97 confirmed in humans and 151 in mice) are imprinted (**Illustration 13**). More than hundred protein-coding genes are imprinted in humans and mice (Williamson et al., 2013; Morison et al., 2005). These play diverse roles in development and growth, and contribute to physiology and behavior (Peters, 2014). Emerging evidence also shows that many of regulatory non-coding RNAs (ncRNAs), including microRNAs, snoRNAs and long ncRNAs, are controlled by genomic imprinting (148). Although the biological functions of these imprinted ncRNAs are still not fully understood, recent reports have shown that imprinted small RNAs modulate specific functions in development and metabolism and also are frequently perturbed in cancer (148).

Imprinted genes are organized in clusters, forming chromosomal domains of tens of kilobases to several megabases in size. Parent-specific expression of multiple genes within a cluster is under the overall control of a *cis*-acting imprinting control region (ICR) (149). ICRs are essential regulatory DNA elements, which are rich in CpG dinucleotides, and are marked by a germline-derived DNA methylation region (DMR) on one of the two parental alleles. Parental allele-specific DNA methylation involves the 5-methyl-cytosine (5mC) methylation and covalent histone modifications (150). Notably, deletion of the DMR or loss of DNA methylation at the DMR will lead to abnormal expression of all imprinted genes in this cluster.

The discovery of clusters of imprinted genes was a strong indication that a common DNA

element may regulate imprinted expression of multiple genes in *cis*. The most significantly represented function among imprinted genes includes those that affect growth of the embryo, placenta, and neonate. This category contains paternally expressed imprinted genes that function as growth promoters (i.e., *Igf2, Peg1, Peg3, Rasgrf1, Dlk1*) and show growth retardation in embryos deficient for the gene. There are also maternally expressed imprinted genes that function as growth repressors (i.e., *Igf2r, Gnas, Cdkn1c, H19, Grb10*), as shown by a growth enhancement in embryos deficient for the gene. Another significant category includes genes with behavioral or neurological defects (e.g., *Nesp, Ube3a, Kcnq1*) (151).

3.1. Epigenetic control of mammalian genomic imprinting

DNA methylation has been shown to be essential for establishment and maintenance of imprinting. The acquisition of imprinting is mainly mediated by the de novo DNA methyltransferase DNMT3A, and that of a non-catalytically active partner protein, called DNMT3-like (DNMT3L) (152). Furthermore, the somatic maintenance of the DNA methylation imprints is mainly mediated by DNMT1. However, there are many other proteins contributing to this process including ZFP57, which binds to methylated DNA sequences through recognition of a specific sequence motif ('TGCCGC') and it creates a platform for binding of KAP1, which in turn, recruits histone-modifying enzymes, including SETDB1, that deposit H3K9me3 repressive mark (153).

Other proteins contributing to the maintenance of methylation imprints include the methyl-CpG-binding domain protein MBD3 and retinoblastoma-binding protein 1 (RBBP1). MBD3 is part of the 'nucleosome remodeling and deacetylase' (NuRD) complex. In mouse embryos, knock-down of *Mbd3* results in partial loss of DNA methylation at the paternally-methylated ICR of the *Igf2-H19* locus (154,155). On the other hand, RBBP1 is a nuclear protein involved

in the control of gene expression at the imprinted domain associated with the Prader-Willi Syndrome (PWS). In mouse, deletion of *Rbbp1* gene results in reduced levels of H3K9me3 and H4K20me3, and partial loss of DNA methylation at the ICR controlling the PWS domain (156).

Most imprinted domains express at least one lncRNA, transcription of which, or the RNA itself, controls adjacent genes (157). For instance, the *Igf2 receptor* (*Igf2r*) locus is controlled by a maternally methylated ICR that comprises the promoter of a 118-kb lncRNA expressed from the paternal allele, antisense to *Igf2r* (157).

Two lncRNAs can bind to both G9A and PRC2: KCNQ1 overlapping transcript 1 (*Kcnq1ot1*) and DHRS4 antisense RNA 1 (*DHRS4AS1*). *Kcnq1ot1* controls the imprinting and silencing of the potassium voltage-gated channel, subfamily Q (KCNQ) gene cluster by recruiting G9A and also PRC2 (158). Furthermore, *DHRS4AS1* silences the DHRS4 gene cluster by physically interacting with G9A and EZH2 and targeting them to the promoters of the downstream *DHRS4L2* and *DHRS4L1* genes, where G9A and EZH2 deposit H3K9me2 and H3K27me3 repressive marks respectively(159)**.** Interestingly, the imprinted lncRNA acute insulin response (*AIR*) regulates targeting of G9A to induce the allele-specific imprinting in *cis* of the solute carrier family 22 member 2 (*Slc22a2*), *Slc22a3* and *insulin-like growth factor 2 receptor* genes in the mouse placenta (160).

Mouse targeting studies on the *Dlk1-Dio3* domain have shown that its paternally methylated ICR activates imprinted gene expression on the maternal chromosome. The maternal ICR allele activates all the maternal ncRNAs of the locus, including *Gtl2* lncRNA. This enhancer-like action is linked to RNA polymerase II recruitment to the ICR, and to bidirectional transcription of multiple ncRNAs that are relatively short and non-poly-adenylated (161).

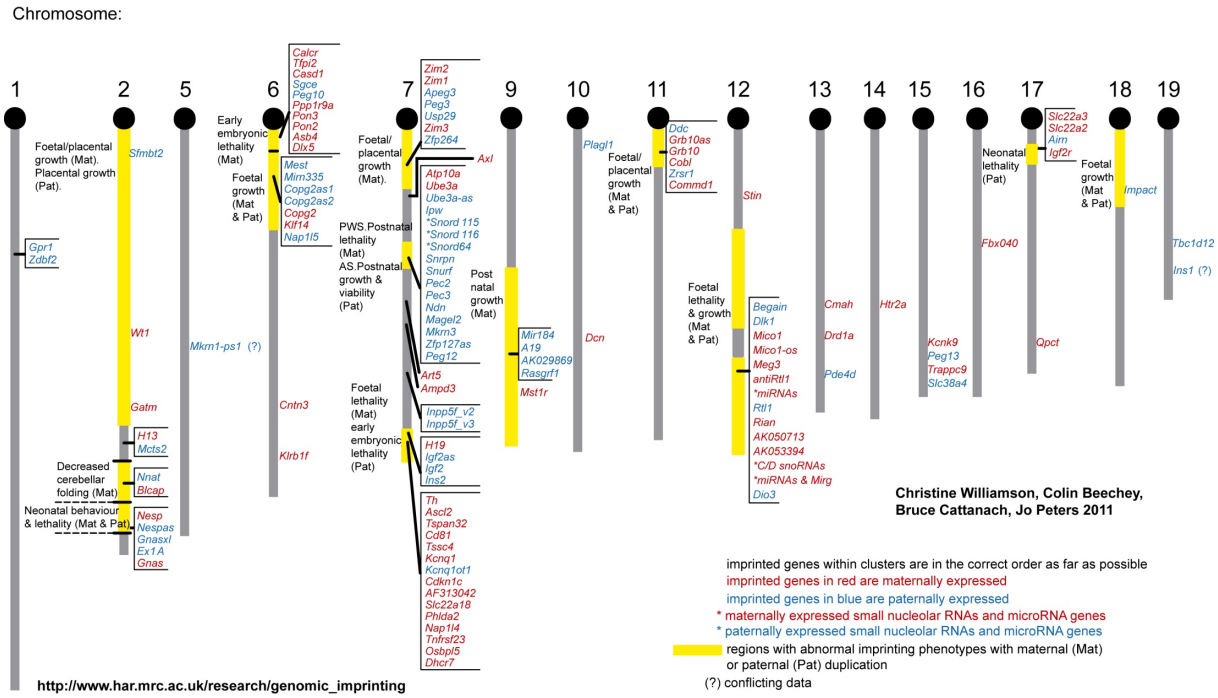

Illustration 13. Mouse imprinted genes, regions and phenotypes.
From: http://www.har.mrc.ac.uk/research/genomic_imprinting/

4. STRUCTURE OF RETROTRANSPOSONS IN THE MAMMALIAN GENOME

Genomic studies have revealed that only about 3% of the mammalian genome is composed of protein-coding sequences compared to 40 – 45% constituted by transposable elements (TEs) (162). In fact, despite often being classified as selfish or junk DNA, TEs have a significant impact on mammalian development and genome regulation.

TEs are mobile DNA sequences able to jump into new locations within genomes contributing to evolution through processes such as new gene formation, exonization, and formation of new transcriptional regulatory elements, like promoters and enhancers (163). Additionally, TEs can modulate gene expression and chromatin states, and large tandemly repeated arrays of DNA, which organize specialized domains such as centromeres and telomeres (164).

These repetitive elements are divided into two categories. Class II elements or DNA transposons, which move via a "cut-and-paste" mechanism and constitute a minor fraction of the mammalian genomes (**Illustration 14A**). Class I elements or retrotransposons, which replicate via a "copy and paste" mechanism, which involves transcription to generate an RNA intermediate followed by reverse transcription and subsequent integration into the host genome at a new location (165) (**Illustration 14B**). Retrotransposons are further classified into those that are flanked by Long Terminal Repeats (LTRs), also called endogenous retroviruses (ERVs), or those lacking LTRs. LTR elements/ERVs make up 8-10% of the mouse and human genomes. Non-LTR retrotransposons are the most abundant and active, comprising approximately 25-27% of the mouse and human genomes and include the long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (166) (**Illustration 15**).

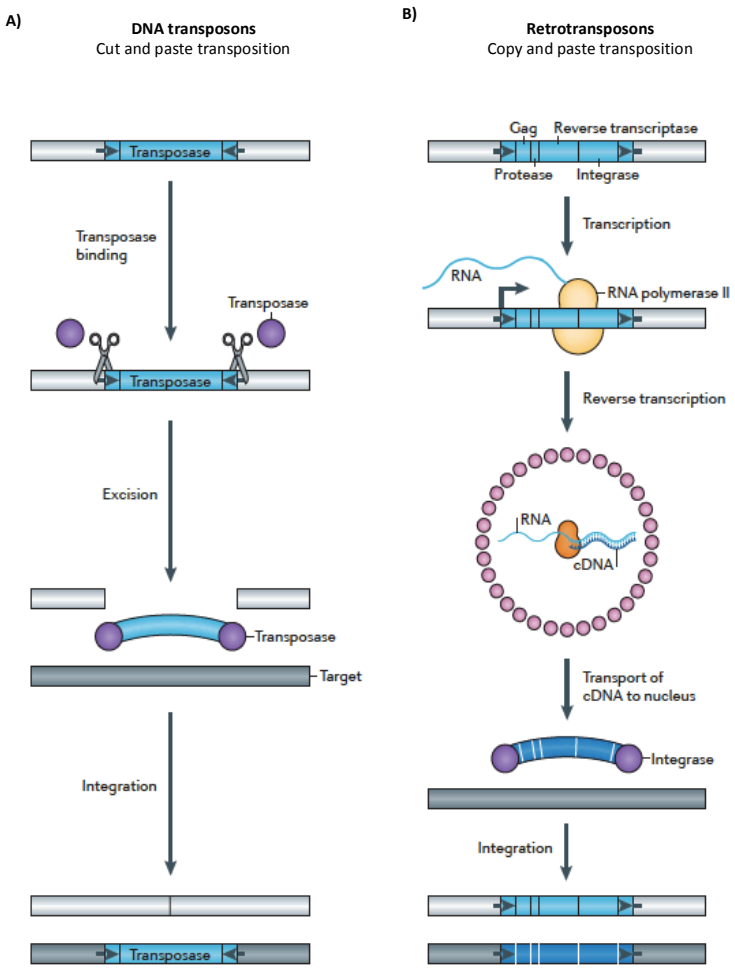

Illustration 14. The diverse mechanisms of transposon mobilization. A) DNA transposons. In general, DNA transposons mobilize by a 'cut and paste' mechanism. The transposase excises the transposon from its existing genomic location and pastes it into a new genomic location. The cleavages of the two strands at the target site are staggered, resulting in a target-site duplication. **B)** Retrotransposons mobilize by replicative mechanisms that require the reverse transcription of an RNA intermediate. In general, a sequence is recognized by the host RNA polymerase II and produces the mRNA of the retrotransposon. Then, the reverse transcriptase copies the retrotransposon mRNA into a full-length dsDNA. Finally, integrase inserts the cDNA into the new target site. *Adapted from: Levin HL, et al. 2011* (167).

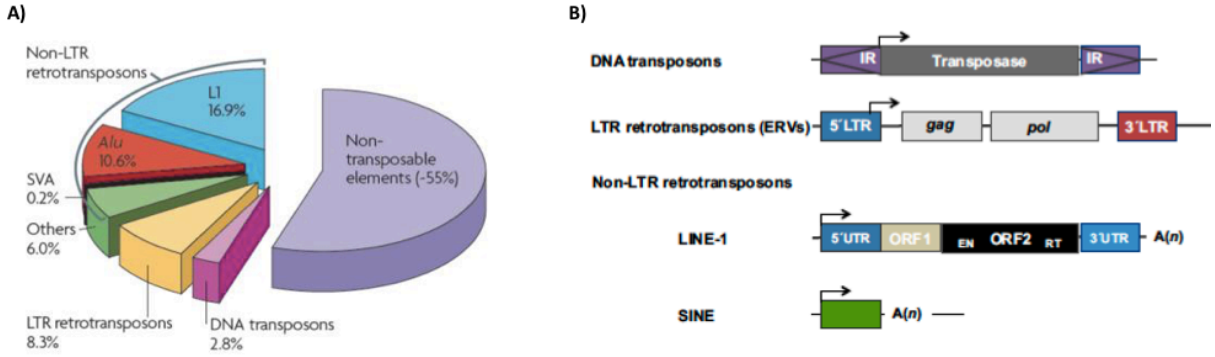

Illustration 15. Repetitive elements in the mammalian genome. A) About 45% of the human genome can currently be recognized as being derived from transposable elements. **B)** Structure of different types of TEs (DNA transposons; ERV, endogenous retrovirus; LINE-1, long interspersed element class 1; SINE, short interspersed element) found in mammalian genomes are represented (IR, inverted repeat; UTR, untranslated region; EN, endonuclease; RT, reverse transcriptase; LTR, long terminal repeat; ORF, open reading frame. *Adapted from: Garcia-Perez JL, et al. 2016* (168) *and Cordaux R, et al. 2009* (166).

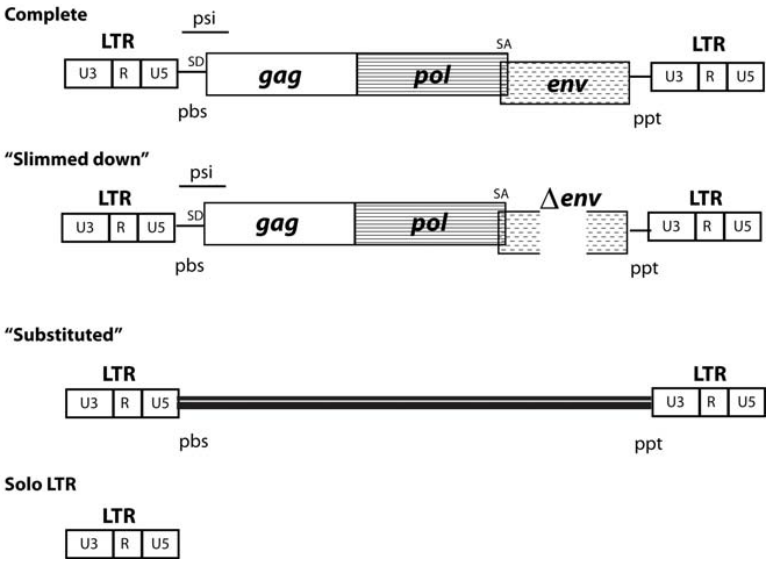

Illustration 16. Basic structures of ERVs. Complete ERVs are essentially identical to the integrated proviruses of simple exogenous retroviruses; they contain two LTRs made up of unique 3′ (U3), repeat (R), and unique 5′ (U5) regions, a primer binding site (pbs) and polypurine tract (ppt), as well as a full complement of coding sequences (gag, pol, and env). "Slimmed down" ERVs are elements lacking coding sequences compared to a complete ERV (e.g. deletion in env). "Substituted" ERVs are elements in which the ERV coding sequences have been replaced with non-viral sequences. "Solo LTRs" are single LTRs generated by homologous recombination between the two LTRs of a complete element. Adapted from: Mager DL, et al. 2015 (169).

4.1. Endogenous retroviruses (ERVs) family

ERVs are the descendants of exogenous retroviruses that integrated into the genome of germ cells but have lost the ability to exit the host cell. In fact, these ERVs are still competent for retrotransposition and expand in their host genome by vertical transmission (162). ERVs are generally 7-11 kb in length and typically contain 5' and 3' LTRs flanking the retroviral genes *Gag* (group-specific antigen), *Pro* (protease), *Pol* (polymerase) and *Env* (envelope). The *Gag* gene encodes for a group of specific antigen proteins, which include the viral matrix, capsid, and nucleocapsid. The *Pro* gene encodes a protease, which functions to cleave the Gag polyprotein precursor. The *Pol* gene encodes the reverse transcriptase and integrase proteins. The reverse transcriptase is an RNA-dependent DNA polymerase with RNase H activity required during retrotransposition. The integrase protein introduces breaks in the host chromosomal DNA to mediate integration of proviral DNA into the genome. *Env* gene encodes the virion surface glycoprotein and the transmembrane protein, which form a complex that interacts with cellular receptor proteins. However, most ERVs completely lack or have a non-functional *Env* gene. ERVs also harbor a tRNA primer-binding site (PBS) and the poly-purine tract (PPT), which are required for the process of reverse transcription. The flanking LTRs contain the U5 (unique to 5' end), R (repeated sequence) and U3 (unique to 3' end) sequences, which are necessary for the regulation of proviral transcription (18,170).

Based on its genetic structure, ERVs can be classified into four groups (169). The first group is composed of "complete ERVs" which are identical to exogenous retroviruses. The second group, "slimmed down" ERVs are those lacking coding sequences compared to a complete ERV (usually the Env gene). Examples include the intracisternal A-type particle (IAP) (171) and MusD (172). The third group, "substituted ERVs" contains elements with two LTRs,

PBS, and PPT, but no other recognizable homology with retroviral proteins, for example, the early transposons (ETns) of mice. Finally, the so-called "solo LTRs" includes single LTRs generated by recombination between the 5' and 3' LTRs of intact proviruses. Notably, this group is considered the largest of the annotated ERV sequences (173) (**Illustration 16**).

Mouse ERVs also can be classified as class I, class II, and class III based on the relative geological time scale of integration of these ERVs into the host genome.

Class I elements constitute about 0.7% of the mouse genome. Members of this ERV class are grouped based on their similarity to the Moloney Murine Leukemia Virus (MLV) (170). Class II elements include the HERVK (HML-2) family in humans (174) and the Intracisternal A Particles (IAPs), MusD and the closely related Early Transposon (ETn) elements in mice (173), which are retrotranspositionally competent subfamiles. Type I IAP elements encompass full-length members as well as deleted classes. Of these, the IΔ1 subclass, with a 1.9 kb deletion in the *gag-pol* region, is the most abundant deleted form in the mouse genome and is also responsible for the majority of IAP insertional mutations. Endogenous IAP expression has been detected in oocytes and preimplantation embryos. Furthermore, some adult mouse tissues, particularly thymus, also exhibit IAP transcription, which is abundant in tumors and tumor cell lines (175). ETns are flanked by LTRs and do not contain retroviral open reading frames. Two major subtypes of ETn elements, I and II, differ in the 3' portion of the LTR and a 5' internal segment. Although ETnI elements are more numerous than ETnII elements, the second subtype is more transcriptionally active. A related subfamily, MusD, which shares nearly identical LTRs with ETns, has been found to codify the proteins necessary for ETn retrotransposition (176). Finally, class III retroviral elements, are the most numerous of all three ERV classes, comprising of ~5.4% of the mouse genome. This class

includes the Murine ERV-L (MERV-L) and the non-autonomous MaLR elements, the latter being the most abundant ERV family in the mouse genome (169). Interestingly, it has been reported that Class III mouse MERV-L elements function as alternative promoters for many genes in oocytes and embryos, suggesting that such elements may participate in the reprogramming of the embryonic genome (177).

4.2. Non-LTR retrotransposons: LINE and SINE

Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements **(**SINEs) constitute ~30% of the human genome sequence and show a non-random genomic distribution. SINEs are generally localized in gene-rich regions, whereas LINEs are enriched in intergenic regions (178).

4.2.1. Long Interspersed Nuclear Elements (LINEs)

In mammals, the most widespread family of TEs capable of autonomous retrotransposition is the LINE-1 that represents approximately 20% of mammalian genomes. In humans, the largest majority of LINE-1 TEs have lost their ability to move and only about 80-100 copies of LINE-1 retain retrotransposition potential (179). In mice, LINE-1 elements comprise a similar fraction of the genome to humans, but only a thousand copies of LINE-1 elements may retain the capacity to retrotranspose (180).

Full-length LINE1 elements are 6 - 7 kb long and are transcribed from an internal promoter located in the 5’ UTR (179). Remarkably, this LINE-1 5’UTR varies significantly between species. LINE-1 encodes two open reading frames (ORFs) in mice and three ORFs in humans

(179). In mice, ORF1 encodes an RNA-binding protein that is implicated in the priming of LINE-1 reverse transcription. The ORF2 protein possesses endonuclease and reverse transcriptase activities as well as a Cys-rich domain, and all three domains are essential for retrotransposition (180). The LINE-1 integrates into a new genomic locus by a process termed Target-Primed Reverse Transcription (TPRT). During TPRT, the endonuclease recognizes and cleaves the DNA consensus target sequence 5'TTTT/AA-3' and then, the reverse transcriptase reverse transcribes the LINE-1 RNA (181).

4.2.2. Short Interspersed Nuclear Elements (SINEs)

SINEs are short sequences (80–500 bp) of non-autonomous retrotransposons that do not encode any protein but can retrotranspose by stealing the retrotrancriptase and endonuclease activities of LINE elements. SINEs are RNA polymerase III (RNA pol III)–transcribed elements constituted by a 5′ head, a body, and a 3′ tail. Head sequences harbor the internal RNA pol III promoter and have been used to categorize SINEs into three superfamilies. This classification is based on the SINE similarity to cellular RNA pol III-controlled genes encoding tRNA (such as mouse B2 or ID elements), 7SL RNA (such as mouse B1 and human *Alu* elements), or 5*S* rRNA (SINE3). Like LINEs, the majority of SINEs have also lost functional promoters or essential sequences required for retrotransposition (180).

5. TRANSCRIPTIONAL SILENCING OF RETROTRANSPOSONS BY H3K9 KMTs

Transcriptional silencing of retrotransposons is mediated by the establishment of a repressive chromatin structure that prevents the access and/or function of transcriptional activators. In mouse ESCs, proviral silencing mostly depends on histone lysine methylation (182), while in differentiated cells silencing relies mostly on DNA methylation (183). More specifically,

SETDB1/H3K9me3 ERV silencing pathway operates not only in cell types derived from embryonic stages, where DNA methylation is being reprogrammed (87,182), but it is also required for ERV silencing in differentiated cell types including neural progenitors (86), B lymphocytes (184) and somatic cells (88).

Comprehensive genomic analyses highlighted the importance of H3K9 KMTs in the silencing of several transposable elements. In fact, *SETDB1* conditional KO (cKO) mESCs show a dramatic up-regulation of several class I and class II ERVs with a concomitant decrease in H3K9me3 at these elements (87). Consistently, the depletion of the SETDB1-interacting protein KAP1 also results in up-regulation of the same ERVs in mESCs (87,183,185). KAP1 is targeted to intact ERVs by the KRAB zinc finger proteins (KRAB-ZFPs). In general, characterized KRAB-ZFPs binding sites recognize internal ERV sequences, including the PBS, 5'-UTR, Gag, and PTT regions (183). However, given the particular genomic structure of the "solo LTRs", these ERVs may escape the effect of the KRAB/KAP1/SETDB1 repression complex directed at full-length elements, facilitating their integration into existing regulatory networks as enhancers, promoters or as a source of novel non-coding and protein-coding genes (186).

The role of SETDB1 in ERV silencing in mESCs has been demonstrated as the deletion of *G9a*, *Glp*, *Suv39h1,* and *Suv39h2* did not revealed an increase in ERV expression (IAP, MusD and MLV) (87). Furthermore, complete loss of DNA methylation in undifferentiated mESCs due to deletion of DNMT1, DNMT3A and DNMT3B had only a modest effect on IAP silencing (87).

It has been shown that the SETDB1/KAP1 complex regulates *de novo* DNA methylation of proviral sequences (187) and that SETDB1 regulates DNA methylation turnover at ERV LTRs in mESCs (188). SETDB1 directs DNA methylation at H3K9me3-enriched ERV sequences, supporting the notion that the SETDB1/KAP1 silencing machinery can direct DNA methylation to maintain retroelement silencing (185). Furthermore, recent studies have also pointed to the role of KRAB-ZFP/KAP1/SETDB1 pathway in silencing of an evolutionarily distinct subset of ERV and LINE L1 elements in human ESCs (189,190). All in all, KRAB-ZFP/KAP1/SETDB1 complexes seem to be the essential epigenetic silencing machinery that acts independently of DNA methylation to silence newly integrated retroviral vectors and ERVs in specific cell types.

In addition to the role of SETDB1 in retrotransposon silencing, other H3K9-specific KMTs have also been shown to promote retroelement silencing in mESCs. For example, SUV39H1/2 are not necessary for silencing of class I and II ERVs, but are required for maintaining H3K9me3 and silencing of a distinct subset of LINE L1 elements in mESCs but not in differentiated somatic cell types (11). On the other hand, the catalytic activity of G9a/GLP complex is necessary for *de novo* DNA methylation to establish a silent state at newly integrated MSCV PBS proviruses (143). Besides, class III MERVL elements, which are not targeted by SETDB1 or H3K9me3 are enriched in H3K9me2 and are de-repressed in *G9a* KO and *Glp* KO mESCs (191).

CHAPTER 4. H3K9 LYSINE METHYLTRANSFERASES AND HUMAN DISEASE

Epimutations play an essential role in cancer initiation and progression. The cancer epigenome is characterized by changes in DNA methylation, histone tail modifications, histone variant incorporation patterns and nucleosomal remodeling, as well as altered expression profiles of chromatin-modifying enzymes. These epigenetic changes can lead to silencing of tumor suppressor genes and/or promote tumorigenesis by activating oncogenes. Remarkably, epigenetic alterations are mitotically heritable, can be selected for cancer cell population and confer a growth advantage to tumor cells resulting therefore in their uncontrolled growth (21).

Different studies have shown that alterations in H3K9 and H3K27 methylation patterns are associated with aberrant gene silencing in various forms of cancer. Dysregulation of H3K9 KMTs responsible for repressive marks lead to an altered distribution of these marks in cancer and leads to aberrant silencing of tumor suppressor genes. Thus, G9A, which has been identified as potential oncogene since G9A-mediated p53K373me2 suppresses its activity leading to cancer (123).

Likewise, different studies demonstrated that SUV39H1 plays an important role in tumorigenesis. For instance, SUV39H1 knockdown and the pharmacological inhibition of SUV39H1 reduced the tumorigenicity of hepatocellular carcinoma cells (192). In addition, SUV39H1 stability is negatively regulated by Prolyl isomerase Pin1. Pin1 interacts with SUV39H1 in a phosphorylation-dependent manner and promotes ubiquitination-mediated degradation of SUV39H1. Remarkably, overexpression of Pin1 leads to decrease SUV39h1 expression, which subsequently decreases H3K9me3, promoting tumorigenecity (193).

Importantly, the reversible nature of epigenetic aberrations has led to the emergence of the promising field of epigenetic therapy. The breast cancer treatments that are currently under evaluation are directed towards reversing aberrant DNA methylation, histone acetylation and histone methylation of tumor suppressor genes and genes involved in therapeutic response (194). For instance, re-establishing normal histone methylation patterns through treatment with KMT inhibitors have been shown to result in anti-tumorigenic effects. One inhibitor compound, DZNep, was shown to successfully induce apoptosis in cancer cells by selectively targeting PRC2 proteins (195). Furthermore, since G9A is overexpressed in numerous types of cancer, small molecules with the capacity to inhibit G9A catalytic activity have been tested in anti-tumor treatment. Indeed, cells pre-treated with the G9A-inhibitor BIX-01294 prior to injection into mice formed significantly smaller tumors when compared with untreated ones, suggesting that G9A inhibition could efficiently reduce tumor growth and metastatic potential (196).

Increasing evidence involve SETDB1 as a tumorigenic factor in many tumor types. It has been shown that *SETDB1* gene is amplified in human melanoma and accelerates melanoma onset *in vivo* in a *zebrafish* model. The same study also revealed that a copy number variation affecting *SETDB1* is associated with non-small-cell lung cancer, ovarian cancer, hepatocellular carcinoma and breast cancer (15).

SETDB1 has been proposed as a *bona fide* oncogene with potential for new therapeutic strategies. A genomic and transcriptomic analysis of 51 human KMTs in a panel of breast cancer cell lines and primary breast cancer samples showed *SETDB1* to be amplified and overexpressed in basal-like breast cancer, the most aggressive subtype (17). It has also been shown that epithelial-mesenchymal transition (EMT) process can be reversed when SETDB1

is inhibited by MiR-7 in breast cancer cell lines (197). Besides, *SETDB1* undergoes gene amplification in non-small and small lung cancer cell lines and primary tumors. Furthermore, depletion of SETDB1 expression in such cells reduces cancer growth in cell culture and *nude* mice models, suggesting that SETDB1 overexpression increases tumor invasiveness (16).

Interestingly, SETDB1 has been shown to directly interact with TP63, a transcription factor belonging to the p53 gene family. Specifically, SETDB1 interacts with the isoform ΔNp63, which is overexpressed in basal-like breast cancer and also contributes to the sustenance of tumor growth. Furthermore, this study identified 30 genes repressed by ΔNp63 in a SETDB1-dependent manner, whose expression is positively correlated to survival of breast cancer patients (198). Likewise, regulation of p53 stability by SETDB1 has been demonstrated. In hepatocellular carcinoma, SETDB1 forms a complex with p53 and catalyzes p53K370me2, thereby decreasing the efficiency of recognition and degradation of p53 by MDM2, promoting cell growth (7). Nonetheless, evidence suggests that p53 can regulate SETDB1 at transcriptional level.