Transposon regulation upon dynamic loss of DNA methylation

Marius Walter

Université Pierre et Marie Curie
École doctorale Complexité du Vivant

Génétique et Biologie du developpement
Institut Curie`
CNRS U934 - INSERM UMR 3215

Transposon regulation upon dynamic loss of DNA methylation

Thèse de Doctorat de Biologie présentée par

Marius WALTER

et dirigée par

Déborah BOURC'HIS

Présentée et soutenue le **10 Décembre 2015** devant un jury composé de:

Dr. **Antonin Morillon**	President du jury
Pr. **Wolf Reik**	Rapporteur
Pr. **Donal O'Carroll**	Rapporteur
Dr. **Vincent Colot**	Examinateur
Dr. **Michael Weber**	Examinateur
Dr. **Déborah Bourc'his**	Directrice de thèse

"There is large amount of evidence which suggests, but does not prove, that much DNA in higher organisms is little better than junk. […] We therefore need to explain how such DNA arose in the first place and why it is not speedily eliminated, since, by definition, it contributes little or nothing to the fitness of the organism."

Orgel and Crick, 1980

"In the future attention undoubtedly will be centered on the genome, and with greater appreciation of its significance as a highly sensitive organ of the cell, monitoring genomic activities and correcting common errors, sensing the unusual and unexpected events, and responding to them, often by restructuring the genome."

McClintock, 1983

"They didn't know it was impossible so they did it."

Mark Twain

Acknowledgements

I am grateful to the members of my jury who have agreed to evaluate my work: Pr. Wolf Reik and Pr. Donal O'Carroll, *rapporteurs;* Pr. Vincent Colot and Dr. Michael Weber, *examinateur*, and Dr. Antonin Morillon, *president du jury*. Being evaluated by such inspiring scientists is a rare privilege and I thank them in advance for their time and their help. This is a great honor.

I started to write this PhD manuscript being convinced that it was a massive waste of time. Indeed, there was so little time left before I had to leave the lab and so many cool ideas I wanted to test… I had to reconsider this statement after a few weeks, being forced to realize that it was exactly the opposite: an incredible gift of time. I could finally read and learn. Writing became an incredibly inefficient process and I spent days lost in the literature, often chasing exotic transposons in obscure species. The frustration never went away, and in fact grew continuously as new ideas accumulated and time to test them inexorably disappeared. Nonetheless, in the end I really enjoyed it. I therefore want to thank very warmly the members of my jury who have to read this manuscript. I really hope they will enjoy it.

To anybody else that would ever read this thesis, even a small part, or would just look at the figures and find them pretty, please send an email at marius.walter@m4x.org. I would add your name in the acknowledgment. Nothing could make me happier than being read, especially if you randomly found this thesis somewhere in the Internet. Please, really, send an email!

I joined the lab of Deborah Bourc'his almost five years ago. My interest for biology was three months old and I had only a vague idea of what was DNA. Deborah nonetheless gave me a pipette, put me in the arms of two brave postdocs, and it was the start of an incredible human and scientific adventure. I have been incredibly lucky to work in this lab and with Deborah. I could not have been freer to do everything I wanted, nor receive better mentoring. I always had the help, the support and the trust of Deborah, which is invaluable. After five years spent in her lab, I feel that I think and act like a scientist. I am incredibly grateful for that.

But what made my PhD such a wonderful experience is also the general atmosphere of the team and the department, a powerful mixing of scientific passion and true friendship. I especially want to thank Max and Joan, postdocs in the lab. Their curiosity, their apparently infinite knowledge and dedication for science constantly pushed me forward. They will never actually believe that I wrote that, but they represent the models of the young scientist I wanted to become. Beside, my debt to them (expressed in liters of beers they paid for me) is so staggering that it could never be repaid.

Natasha also occupies a very special place. She introduced me to the fantastic world of transposons and with her strange Australian accent, she is probably the best English teacher I ever had. Having my first (and for the moment only) paper published with her means a lot to me. She is one of the closest friends I made here, and was present at very important time. I never believed she would actually leave, but the lab and the whole BDD are now missing their rightful Queen. That being said, we will all probably become half deaf very young because of her, but well ☺.

I cannot write a paragraph for everybody, but I really enjoyed working with all the members of the team. Rachel that left too early, Raquel that is the nicest person I ever met, Julian that patiently taught me how to dissect tiny embryos, Tomek with its language impossible to master, Sophie that brought the world "crazy" to a new dimension. In a few

months I will transmit my crown of senior PhD to Juliane and I know she will make a good usage of it. I wish her all the failure she is impatiently looking for.

Finally, I am incredibly grateful to Aurelie. I had the chance to work with her on a daily-basis on the bioinformatic analysis for more than a year and I consider myself incredibly lucky. I could never have done half of what I needed without her help and I could not be more thankful. I am conscious that working with me was not always easy and that I often asked too much, but I think that at the end it was worth it. I am fully aware of what I owe her and I think that we can be proud.

I also want to acknowledge all the people in the department that contributed to my PhD, and especially all the team leaders. Their continual guidance during the course of this project, in unit seminars, thesis committees or elsewhere was very valuable and helpful. I especially want to thank Edith Heard, Raphael Margueron and their respective teams. I really learned a lot working in close contact with them. In particular, Michel and Elphège have been incredibly helpful, both in terms of technical help and scientific exchange.

I cannot think about a better place to work than BDD, our department. And I should really write "live" and not "work" here, since I barely left the lab on this last year. At the end of my second year of PhD almost one year ago, I decided that I would finish it in three years instead of four. This manuscript shows it was possible, but at that time nobody believed I could do it. As consequence I seldom left the lab. The life of PhD students and postdocs sometimes involves insane amount of work. I survived because I could share that with close friends from all over the world. This department is full of amazing people, making almost unnecessary to maintain a social life out of it. Some might find that it is too much sometimes, but I had an amazing year. "Work hard, party hard" could definitely be the motto for PhDs and postdocs of our department. This is not the place to elaborate on that, but I want nonetheless to mention some of the people that contributed to the Friday night usual craziness. They made me happy during tough time, and I thank them for that.

Neuza because she is the best, for real; Joke, my climbing partner that once did the biggest baby step ever; Diana, because she is worth a hundred bubbles and is a very good 👮‍♀️; Eskeww, the black devil that we should never listen to; Anahi because she is Anahi; Tim, because we are the proof that European Union works, Natasha, undisputed Queen of BDD; Ellis and Eve for making sure that Max would stay in France forever; Raquel, because she is the voice of reason; Juliane after midnight, because this is another person; Joan, for being always there with me on the late night shifts; Simao, of course; Sophie, for pushing the limit another step further every time; Ines, one of my numerous Portuguese teacher; The WOS, best bar of Paris, therefore the world, why would anybody want to go anywhere else ?; MacDonald, for 1€ hamburgers; My bike, that always knew the way home; Jennifer Doudna and Emmanuelle Charpentier, for changing the word, the French state, for giving money to have fun in the lab; Adobe Illustrator, my one true love; The Point Ephemere, the only reason not to go to the WOS, and finally the girls of BDD, my sun and stars. ☺

J'ai aussi une pensée spéciale pour mes meilleurs amis, de l'X ou d'avant: Robi, Fouch, Bellouch, Qt, Tom et Gabi en particulier. Vous êtes des mecs en or, ne changez rien. Je voudrais aussi remercier George, pour son soutien sans failles depuis tant d'années.

Delphine, aussi.

Finalement je tiens à remercier ma famille. Leur amour inconditionnel est la fondation sur laquelle je suis construit. Les avoir avec moi permet tout le reste.

Résumé

Les transposons sont des séquences d'ADN qui ont la capacité de se dupliquer de façon autonome, posant une menace pour l'intégrité et la stabilité du génome. De nombreux mécanismes existent pour contrôler l'expression des transposons, parmi lesquels la méthylation de l'ADN joue un rôle particulièrement important. Chez les mammifères, les profils de méthylation sont stables tout au long de la vie de l'individu, mis-à-part pendant deux moments clés du développement embryonnaire. Pendant ces deux périodes, la méthylation de l'ADN est globalement effacée, ce qui corrèle avec l'acquisition d'un état cellulaire pluripotent, puis rétablie. En utilisant un système cellulaire de reprogrammation de méthylation induite, ce travail s'est attaché à comprendre comment le génome parvient à maintenir le contrôle des transposons en l'absence de cette protection d'ordinaire essentielle, J'ai pu démontrer que divers mécanismes chromatiniens compensent progressivement la disparition de la méthylation de l'ADN pour le maintien de la répression des transposons. En particulier, la machinerie Polycomb prend en partie le relai et acquiert un rôle primordial, spécifiquement en l'absence de méthylation de l'ADN. Dans un second temps, la contribution du cofacteur d'ADN méthyltransférase DNMT3l lors de la méthylation *de novo* a été étudiée. Dans sa globalité, ces découvertes offrent des perspectives nouvelles sur la façon dont le génome se réorganise lors de moments clés du développement embryonnaire.

Mots clés : transposons – méthylation de l'ADN – chromatine – reprogrammation

Abstract

Transposons are DNA sequences that can duplicate autonomously in the genome, posing a threat for genome stability and integrity. To prevent their potentially harmful mobilization, eukaryotes have developed numerous mechanisms that control transposon expression, among which DNA methylation plays a particularly important role. In mammals, DNA methylation patterns are stable for life, at the exception of two key moments during embryonic development, gametogenesis and early embryogenesis. After a phase a global loss of genomic methylation accompanying the acquisition of pluripotent states, DNA methylation patterns are re- established *de novo* during differentiation. This work attempted to elucidate how the genome copes with the rapid loss of DNA methylation, in particular regarding the control of transposons in absence of this essential protective mark. Using an embryonic cellular model of induced methylation reprogramming, I showed that various chromatin-based mechanisms can compensate for the progressive loss of DNA methylation. In particular, my results suggest that the Polycomb machinery acquires a critical role in transposon silencing, providing a mechanistic relay specifically when DNA methylation patterns are erased. In a second phase, this work analyzed the contribution of the DNA methyltransferase cofactor DNMT3l during events of embryonic *de novo* methylation. Overall, these findings shed light onto the processes by which genome regulation adapts during DNA methylation reprogramming.

Key words: transposons – DNA methylation – chromatin – reprogramming

Résumé des travaux

Les éléments transposables, ou transposons, sont des séquences d'ADN qui ont la capacité de se dupliquer de façon autonome dans le génome. Ils sont présents chez tous les eucaryotes et représentent une force évolutive importante, contribuant au fonctionnement normal du génome. Néanmoins, par leur potentiel mutagénique, ils constituent une menace certaine et immédiate pour la stabilité du génome. En conséquence, les génomes eucaryotes sont dotés de divers mécanismes de protection contre l'activité des transposons. En particulier, la méthylation de l'ADN est l'un des mécanismes de défense les plus conservés, que l'on trouve chez les plantes et les animaux. Chez les mammifères en particulier, les transposons se sont multipliés dans des proportions impressionnantes, et représentent environ la moitié de la masse génomique. La majorité de ces éléments ont accumulé au cours du temps des mutations qui les rendent inactifs. Cependant, une petite minorité a conservé une activité de mobilisation, avec des conséquences potentiellement délétères, comme en témoigne leur implication dans de nombreuses pathologies, congénitales et acquises.

La méthylation de l'ADN exerce différentes fonctions chez les mammifères. Elle est l'un des principaux mécanismes qui assure la répression des transposons, et est aussi associée au contrôle de l'expression des gènes, en particulier ceux associés à l'empreinte génomique, ceux soumis à l'inactivation du chromosome X chez les femelles enfin, les gènes impliqués dans la pluripotence et les fonctions de la lignée germinale. Les profils de méthylation sont remarquablement stables tout au long de la vie de l'individu, mis-à-part pendant deux moments clés du développement, l'embryogénèse précoce et la gamétogénèse. Pendant ces deux périodes, la méthylation de l'ADN est globalement effacée, ce qui corrèle avec l'acquisition d'un état cellulaire pluripotent. Dans un second temps, les profils de méthylation sont rétablis *de novo* lors de la différenciation de l'embryon ou de la spécification des gamètes. La méthylation *de novo* dépend de l'activité catalytique des ADN méthyltransférases DNMT3a et DNMT3b, assistées du cofacteur DNMT3l.

Mon travail de thèse s'est attaché à répondre à deux questions spécifiques :

- Comment le génome s'adapte-il à la perte rapide de méthylation de l'ADN, et en particulier comment le génome parvient-il à maintenir le contrôle des transposons en l'absence de cette protection d'ordinaire essentielle ?
- Quel est le rôle particulier de DNMT3l lors de la méthylation *de novo* du génome de l'embryon ?

Pour ce faire, j'ai eu recours à des approches d'ingénierie génétique par les outils CRISPR/Cas9 sur des cellules embryonnaires murines (ES), combinées à des méthodes de cartographie à grande échelle des profils transcriptionnels (RNA-seq), de méthylation génomique (par Whole Genome Bisulfite Sequencing, WGBS), et de modifications d'histones (ChIP-seq) suivies d'analyses bioinformatiques pertinentes.

Afin de reproduire les vagues de déméthylation qui se produisent pendant le développement embryonnaire et gamétique, j'ai utilisé un système de culture différentiel de cellules ES murines. Le changement de ces cellules d'un milieu de culture contant du sérum à un milieu contant des inhibiteurs chimiques (milieu « 2i) ») permet de convertir leur génome d'un état globalement hyperméthylé à un état pratiquement dénué de méthylation, sans modifier leur état de pluripotence. Pendant cette transition, j'ai observé que l'expression des transposons suivaient deux phases distinctes. Dans un premier temps, pratiquement toutes les familles de transposons montrent une réactivation. Puis dans un second temps, les transposons sont remis sous silence. Cela indique que le contrôle des transposons peut être compensé par des mécanismes alternatifs lorsque la méthylation de l'ADN disparait. A cet égard, j'ai pu observer que la déméthylation du génome s'accompagnait d'une reconfiguration des profils chromatiniens de type répressif : alors que la tri-méthylation de l'histone H3 sur la lysine 9 (H3K9me3) reste stable, H3K9me2 disparait totalement alors que les profils de H3K27me3, une marque répressive associée à la machinerie Polycomb, se réorganisent et s'accumulent sur les transposons. En utilisant des cellules mutantes pour des gènes du complexe Polycomb, j'ai pu confirmé génétiquement que H3K27me3/Polycomb devenaient des importants régulateurs des transposons en absence de méthylation de l'ADN.

De façon intéressante, j'ai de plus observé que H3K9me3 et H3K27me3 occupent des familles de transposons distinctes, et ou des territoires séparés à l'intérieur de la séquence d'un même transposon. Cela nous à permis de séparer les familles de transposons en trois classes fonctionnelles, qui déterminent comment ces familles s'adaptent à la perte de méthylation de l'ADN. Ces résultats n'étaient pas soupçonnés auparavant et mettent en lumière les mécanismes possibles impliqués dans la répression des transposons pendant le développement embryonnaire. En particulier, ce travail montre que des voies de répression différentes agissent de concert spécifiquement en absence de méthylation de l'ADN pour sécuriser le contrôle d'une large gamme de transposons, permettant ainsi le maintien de la stabilité du génome lors de périodes développementales critiques.

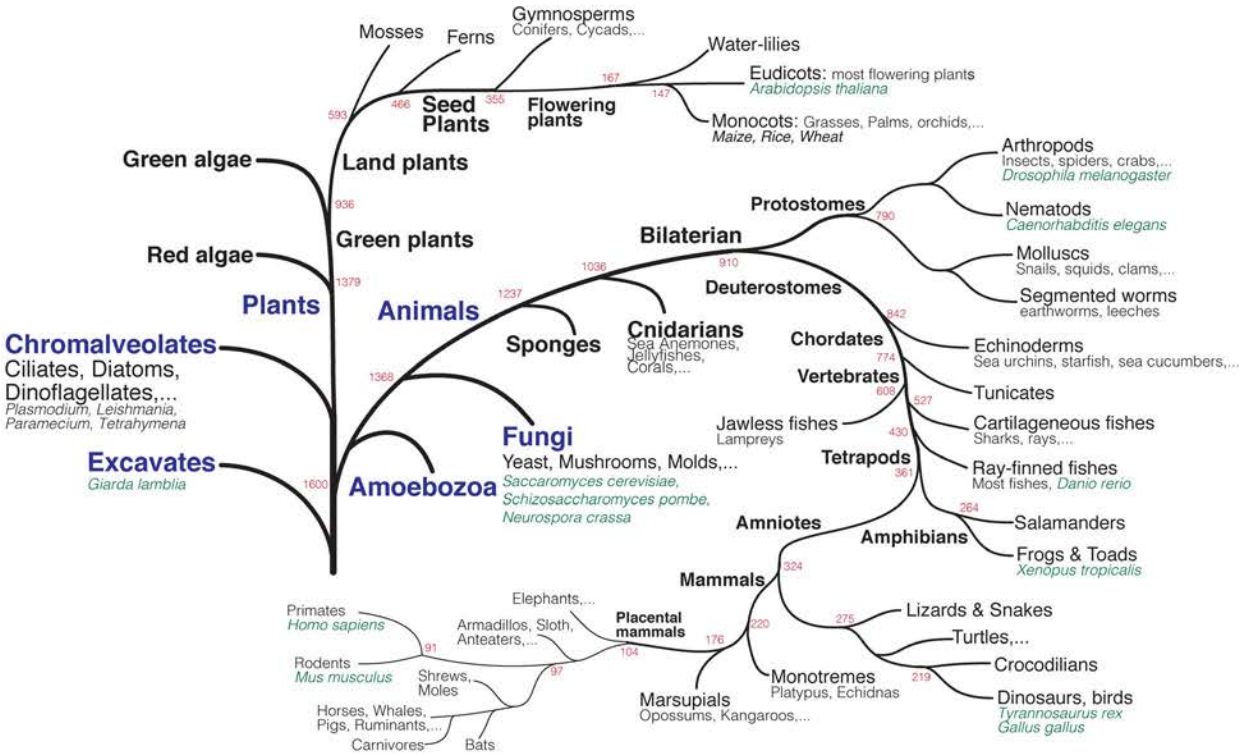
La seconde partie de ma thèse a concerné l'étude du rôle de DNMT3l pendant la méthylation *de novo* des cellules embryonnaires. DNMT3l est un cofacteur des ADN méthyltransférases DNMT3a and DNMT3b. DNMT3l n'a pas d'activité enzymatique en soi, mais stimule l'activité de DNMT3a et DNMT3b en stabilisant leur conformation tridimensionnelle. Dans la lignée germinale, sa présence est absolument requise pour la méthylation *de novo*. En revanche, sa fonction dans l'embryon précoce reste incertaine. Afin d'analyser précisément la contribution de DNMT3l dans les évènements de méthylation global du génome embryonnaire, j'ai utilisé des souris et des cellules ES mutantes pour DNMT3l. J'ai ensuite cartographié à la base près la dynamique de la reméthylation de l'ADN dans les embryons post-implantatoires et pendant la différenciation des cellules souches, en comparaison avec des cellules de la lignée germinale.

Ces résultats montrent que pendant la différenciation des cellules souches mais pas dans l'embryon post-implantatoire, DNMT3l accélère globalement la mise en place de la méthylation de l'ADN. Ce retard de méthylation en l'absence de DNMT3l n'a néanmoins que des conséquences minimes sur l'expression des gènes et ne semble pas affecter la dynamique de différenciation cellulaire. De façon intéressante, nous avons remarqué que le défaut de méthylation en l'absence de DNMT3l était pratiquement inexistant dans le corps des gènes fortement exprimés, que ce soit en contexte embryonnaire ou germinal. Cela indique que même si DNMT3l est nécessaire à l'acquisition rapide de la méthylation dans l'ensemble du génome, sa présence est superflue au niveau des régions fortement transcrites.

Dans sa globalité, ce travail de thèse met en lumière la façon dont le génome s'adapte aux changements globaux de méthylation de l'ADN. L'observation que la machinerie Polycomb entre en jeu dans le contrôle des transposons spécifiquement en l'absence de méthylation de l'ADN est particulièrement intéressant et novateur. Ce résultat illustre comment des mécanismes différents se passent le relais pendant le développement embryonnaire afin de préserver en continu l'intégrité du génome.

Table of content

INTRODUCTION



Introductory Figure 1. Tree of life

Phylogenetic tree of eukaryotes. Data compiled from Hedges et al., 2015; Maddison et al., 2007 (http://www.timetree.org/book and http://www.tolweb.org/).

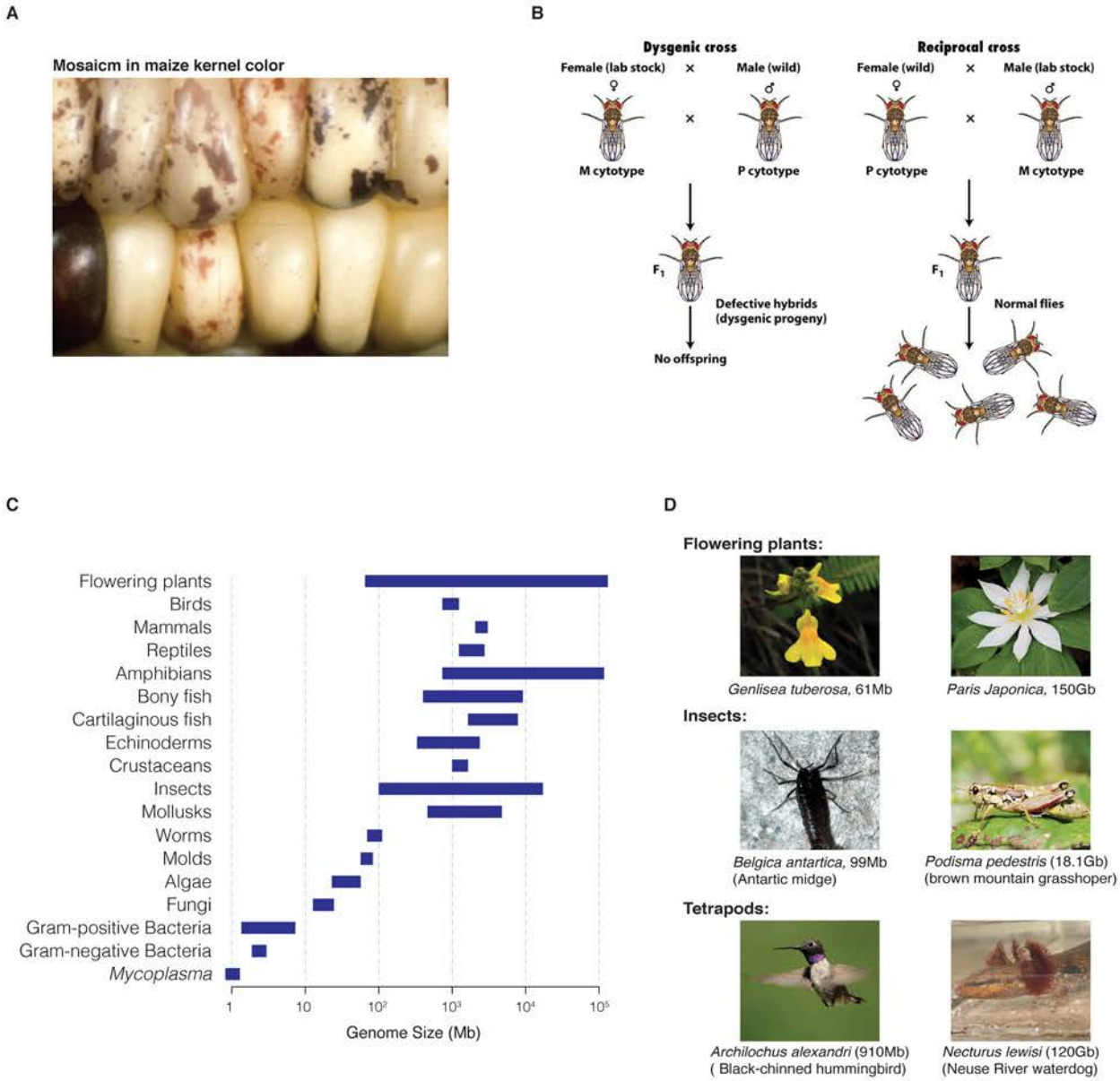
1 TRANSPOSONS

Long believed to be "junk" DNA without any function, Transposable Elements (TEs) are now considered as major players for genome regulation and contributed to shape the evolution of virtually every eukaryote (**Introductory Figure 1**).

1.1 Historical perspective

In the 1910s, Thomas Hunt Morgan discovered that chromosomes were the carriers of genetic information and the material basis of Mendelian heredity (Morgan, 1915). After he established the first genetic maps of *Drosophila,* genes started to be represented as fixed units stably ordered in a linear pattern on chromosomes, like "beads-on-a-string". Mutations were thought to be permanent and irreversible, and considered to be the main driving force of Darwinian evolution. This model would prevail during most of the 20th century. When Barbara McClintock observed in 1950 that some genes could move along chromosomes, her discovery received a very cold reception from the scientific community (McClintock, 1950). McClintock was working on the mechanism of chromosome breakage and fusion in maize. She had identified a locus on chromosome 9 where breakage was always occurring. She named it "*Ds*" for "Dissociation". She noticed that *Ds* could change position within the chromosome and switch on and off the expression of pigment genes, resulting in mosaicism in the maize kernel colors (**Introductory Figure 2A**). Even if this discovery would eventually earn her a Nobel Price in 1983, the concept of mobile elements that could reversibly be inserted elsewhere in the genome and alter the expression of other genes did not fit within the framework of genetics at that time.

The presence of mobile DNA in bacteria was acknowledged at the end of the 1960s (Shapiro, 1969) but McClintock work would need another decade and the discovery of P elements in fruit flies before starting to receive recognition from the scientific community. When crossing *Drosophila Melanogaster* strains used in laboratory with strains found in nature, researchers observed an increase of the rate of mutations, recombination defects, chromosomal rearrangements and sterility (**Introductory Figure 2B** and Kidwell et al., 1977). This phenomenon of "hybrid dysgenesis" was explained by the presence of a family of TEs in wild flies that did not exist in lab strains (Rubin et al., 1982). Researchers understood that P elements had the ability to move in the genome and had invaded all known populations of *D. Melanogaster* worldwide in less than 50 years (Anxolabehere et al., 1988). Lab strains



Introductory Figure 2. Historical perspectives and C-value paradox

A. Mosaicism in maize kernel color due to the interplay between a transposable element and a gene involved in pigmentation. From Feschotte et al., 2002.

B. Hybrid dysgenesis phenomenon caused by P elements. Mating of female from lab stock (M cytotype) with wild male (P cytotype) causes an increase of the rate of mutations, recombination defects, chromosomal rearrangements and sterility, while mating of a P male with a P female has no effect. From Griffiths et al., 2008.

C. Illustration of the C-value paradox. Data represents variation of genome size among various groups of species. No correlation between genome size and complexity is observed. Adapted form Federoff, 2012.

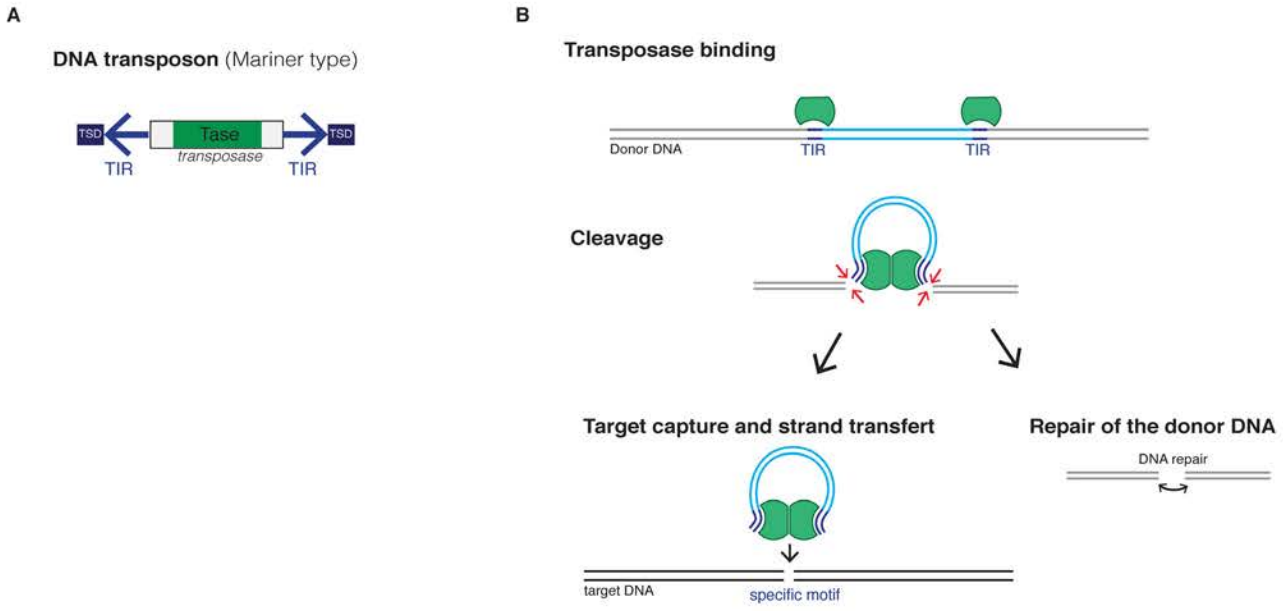
D. Examples of extreme genome size variation. The smallest and largest known genomes from various groups of species are represented. (Fleischmann et al., 2014; Gregory, 2015; Gregory et al., 2009; Kelley et al., 2014; Pellicer et al., 2010)

isolated before the invasion burst had been protected. In the following years, it became evident that TEs were present in almost every eukaryote species and represented a significant proportion of genomes.

During the 1960s came also the realization that cells from different species could contain very different amount of DNA. Whereas prokaryotes tend to have small genomes, there is no correlation between genome size and evolutionary complexity for eukaryotes. This observation is referred to as the "C-value Paradox" (**Introductory Figure 2C, 2D** and Thomas, 1971). Some species of amphibians or fishes have for example 50 times as much DNA per nucleus as humans. Even species of similar biology and complexity can harbor striking differences in genome size. Some flowering plants, like the plant model system *Arabidopsis thaliana,* harbor very small genomes (~100Mb, similar to *Caenorhabditis elegans*), while other flowering plants contain 2,000 times as much DNA (Bennet and Leicht, 2012). These observations were even more intriguing considering that the estimated number of genes (defined as discrete, locatable and protein-coding units of DNA) does not vary in such proportions among species. Indeed, gene number and genome size range from 2,000 genes in 2.3Mb for *Encephalitozoon intestinalis*, an intracellular fungal parasite (Corradi et al., 2010), to around 20-30,000 genes in complex eukaryotes and an enormous size of 150Gb for the flowering plant *Paris japonica* (Pellicer et al., 2010). To summarize, gene number varies in eukaryotes in a 1 to 15 ratio, while genome sizes vary between 1 and 65,000.

In 1972, Susumu Ohno popularized the idea that the majority of DNA in the genome had no function and could be considered as "junk" (Ohno, 1972). Furthermore, in order to explain how such an amount of useless DNA could accumulate in genomes, two papers in *Nature* in 1980 proposed that much of eukaryotic genomes was in fact composed of "selfish DNA" (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). These notions shifted from the prevalent view that the whole nuclear DNA was functional and under heavy selective pressure. On the contrary, and as long as it remained without consequences, DNA capable of proliferating could theoretically accumulate in tremendous amount. Being neutral from an evolutionary point of view, the concept of "junk" and "selfish" DNA solved the C-value paradox and gave a reasonable explanation for the accumulation of TEs in most eukaryote genomes. As Orgel and Crick stated it, "The spread of selfish DNA can be compared to the spread of a not-too-harmful parasite within its host".

This view of TEs as genomic parasites remained dominant for several decades. The mutagenic potential of TEs turned them into powerful and popular genetic tools in many model systems, but they otherwise received little attention. The interest changed with the era



Introductory Figure 3. DNA Transposons

A.Structure of DNA transposons (Mariner type). Two inverted tandem repeats (TIR) flank the transposase gene. Two short tandem site duplications (TSD) are present on both sides of the insert.

B. Mechanism of transposition: Two transposases recognize and bind to TIR sequences, join together and promote DNA double-strand cleavage. The DNA-transposase complex then inserts its DNA cargo at specific DNA motifs elsewhere in the genome, creating short TSDs upon integration.

of genome sequencing and the realization that TEs occupy much more genomic space than anticipated. They constitute around 50% of the mouse or human genome and an impressive 90% of the corn genome (Lander et al., 2001; Schnable et al., 2009; Waterston et al., 2002). In the recent years, TEs became an active topic for many fields of research, and are now seen as major contributors of genome evolution and regulation in virtually every eukaryotes (Fedoroff, 2012). Ironically, McClintock was initially presenting TEs as "controlling elements" because of their ability to modify gene expression. Almost 50 years were necessary for the rest of the scientific community to adopt this idea.

1.2 Biology of Transposable Elements

There are two major groups of TEs. Class II or DNA transposons move by a "cut-and-paste" mechanism. Both *Ds* transposons discovered by McClintock in maize and P elements in *Drosophila* are DNA transposons. Class I or retrotransposons move by a "copy-and-paste" mechanism. They require an RNA intermediate and the insertion of its cDNA complement at a new site in the genome (Finnegan, 1989).

DNA transposons and retrotransposons can further be subdivided into subclasses, orders, superfamilies, families and subfamilies and an unified nomenclature has been proposed (Wicker et al., 2007). In this classification, retrotransposons are subdivided into five orders, the three main ones being LTR-retrotransposons, and Long and Short Interspersed Nuclear Elements (LINEs and SINEs). In addition to these classical groups, two other orders have been described and would only be named here for the sake of being exhaustive: DIRS-like and Penelope-like retrotransposons. Their mechanism of retrotransposition differs from the other orders and members have been detected in plants, fungi and animals (Evgen'ev and Arkhipova, 2005; Poulter and Goodwin, 2005).

1.2.1 DNA-transposons

DNA transposons are found in almost all eukaryote genomes. They move by excising themselves from the genome and integrating at a new location (Muñoz-López and García-Pérez, 2010). Typical DNA transposons are 1.5-5kb in length and encode a transposase gene flanked by two Tandem Inverted Repeats (TIRs) (**Introductory Figure 3A**). TIRs length usually varies between 20pb to 1kb. Two transposase proteins bind to the TIRs and cleave the 5'-end of both repeats, forming a DNA-transposase dimer. Most DNA transposons then recognize a specific target sequence elsewhere in the genome, specific for every family of

transposons. The cleaved DNA is then integrated into the new location. The transposition process creates short Target Site Duplications (TSDs) of typically 4-8pb at both ends of the insert (**Introductory Figure 3B**). At least nine superfamilies of DNA transposons were reported to move according to this mechanism (Wicker et al., 2007).

However, some DNA transposons mobilize in a completely different way. For example, *Helitrons,* which are present in protists, plants and animals, do not contain TIRs and move through a complex "rolling-circle" mechanism (Kapitonov and Jurka, 2007). *Mavericks*, which are found in diverse eukaryotes (except in plants), contain TIRs and encode between five and eleven genes, including an integrase and a DNA polymerase (Pritham et al., 2007). Contrary to most DNA transposons, *Helitrons* and *Mavericks* probably move trough a copy-and-paste mechanism that involves the replication of a single-stranded DNA intermediate (Feschotte and Pritham, 2007).

Except for *Helitrons* and *Mavericks,* mobilization of DNA transposons occurs in a non-replicative manner. Multiplication of transposon copies can therefore only occur through indirect mechanisms. For example, during DNA replication, transposition of a DNA transposon from a newly replicated chromatid to an unreplicated one would lead to the gain of one transposon copy in one of the daughter cell. Moreover, excision of a transposon from its donor site creates DNA double-strand breaks that need to be repaired. One possible pathway of repair is homologous recombination that uses the homologous chromosome (or the sister chromatid) as a template. This scenario results in the regeneration of the transposon at its site of origin and therefore in an increase in transposon copy number in the genome. (Feschotte and Pritham, 2007). DNA repair can occur otherwise by Non-Homologous End-Joining (NHEJ), leading to the formation of transposon "footprints" formed by the remaining TSDs. Using these various mechanisms, DNA transposons have been able to accumulate in large amounts in certain organisms.

Most DNA transposons belong to a handful of superfamilies that have been categorized based on homology of their transposase gene (e.g. *Tc1/mariner, hAT, piggyBac*). Some of the most widespread superfamilies, like *Tc1/mariner,* are found in almost every eukaryotic kingdom, indicating a very ancient evolutionary origin (Capy et al., 1996; Plasterk et al., 1999). Imperfect DNA repair can give rise to various internal deletions and create degenerated transposon copies. However, because transposition only requires the terminal repeats, degraded and non-autonomous copies can still be mobilized by intact transposases encoded elsewhere by intact elements. For example, internally deleted Miniature Inverted-

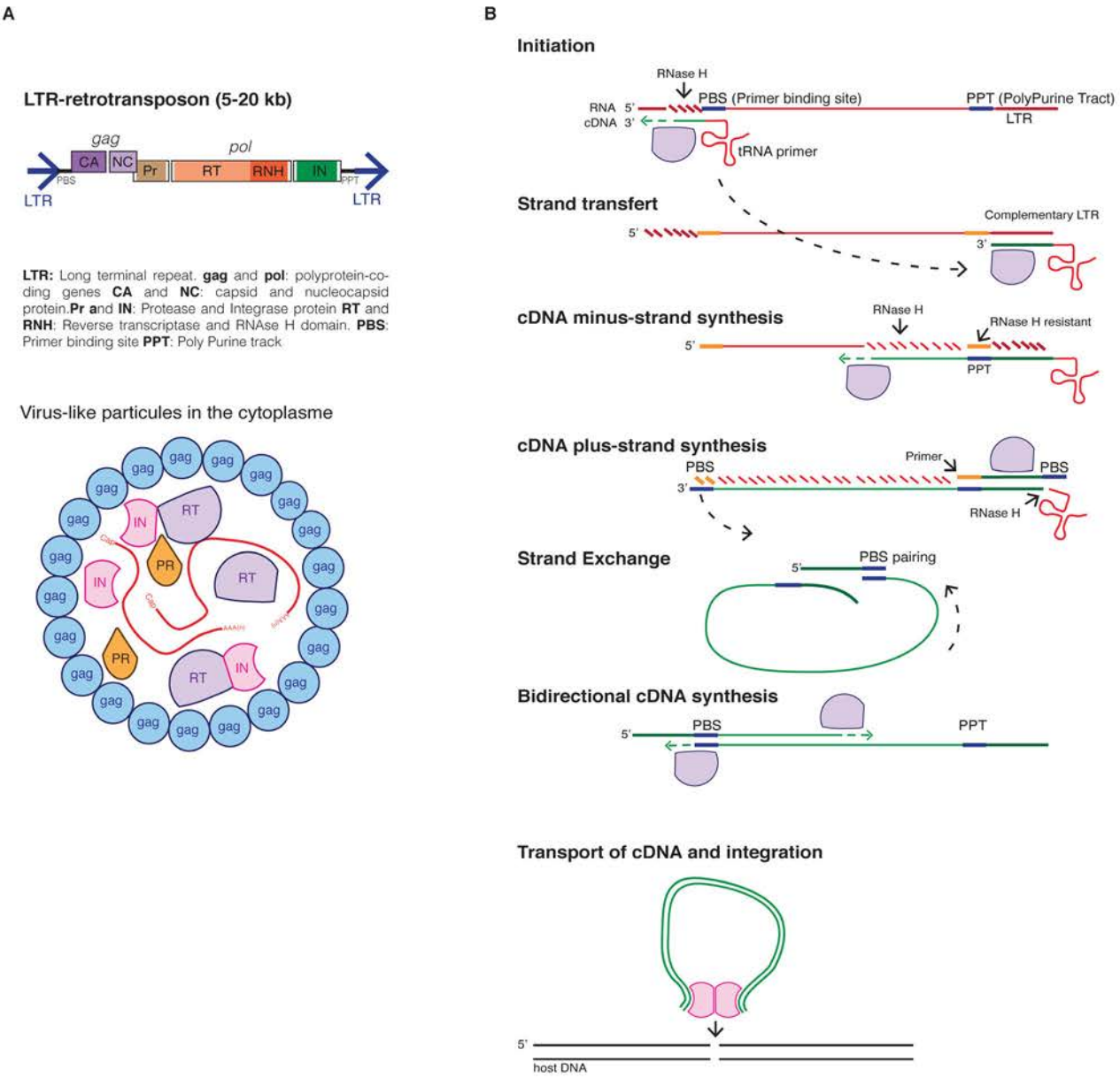
repeat Transposable Elements (MITEs) have accumulated to large numbers in many genomes and are often present in many more copies that intact elements (Feschotte et al., 2002).

1.2.2 LTR-retrotransposons

LTR-retrotransposons are characterized by the presence of Long Terminal Repeats (LTR) directly flanking an internal coding region. As retrotransposons, they mobilize through reverse-transcription of their mRNA and integration of the newly created cDNA into another location. Their mechanism of retrotransposition is shared with retroviruses, with the difference that most LTR-retrotransposons do not form infectious particles that leave the cells and therefore only replicate inside their genome of origin. Their size ranges from a few hundred base pairs to 25kb, like the Ogre retrotransposon in the Pea genome (Neumann et al., 2003). All functional LTR-retrotransposons encode a minimum of two genes, *gag* and *pol,* that are sufficient for their replication (**Introductory Figure 4A**). *Gag* encodes a polyprotein with a capsid and a nucleocapsid domain (Sandmeyer and Clemens, 2010). Gag proteins form virus-like particles in the cytoplasm inside which reverse-transcription occurs. The *Pol* gene produces three proteins: a protease (PR), a reverse-transcriptase endowed with an RT (reverse-transcriptase) and an RNAse H domains, and an integrase (IN) (Wicker et al., 2007).

Typically, LTR-retrotransposon mRNAs are produced by the host RNA pol II acting on a promoter located in their 5' LTR. The *Gag* and *Pol* genes are encoded in the same mRNA. Depending on the host species, two different strategies can be used to express the two polyproteins: a fusion into a single open reading frame (ORF) that is then cleaved or the introduction of a frameshift between the two ORFs (Gao et al., 2003). Occasional ribosomal frameshifting allows the production of both proteins, while ensuring that much more Gag protein is produced to form virus-like particles. Reverse-transcription usually initiates at a short sequence located immediately downstream of the 5'-LTR and termed the primer binding site (PBS). Specific host tRNAs bind to the PBS and act as primers for reverse-transcription, which occurs in a complex and multi-step process, ultimately producing a double-stranded cDNA molecule (Craig et al., 2002). The cDNA is finally integrated into a new location, creating short TSDs and adding a new copy in the host genome (**Introductory Figure 4B**).

Based on phylogenic analyses of the conserved RT domain and on the order of the RT and the IN domains in the *Pol* gene, LTR-retrotransposons can be classified into two main superfamilies, *copia*-like and *gypsy*-like (named after copia/gyspy transposons in *Drosophila* (Havecker et al., 2004; Xiong and Eickbush, 1990). *Copia* and *gypsy* LTR-retrotransposons can



Introductory Figure 4. LTR-retrotransposons.

A. Structure of LTR-retrotransposons (*gypsy*-type).

B. Mechanism of retrotransposition, occurring inside viral-like particles in the cytoplasm. Reverse-transcription initiates at a host tRNA primer binding site (PBS) located immediately downstream of the 5'LTR. The newly synthesized minus-strand cDNA copy of the 5'LTR is then transferred to the 3'LTR and used as a primer for reverse-transcription of the entire minus-strand sequence. An RNase H-resistant polypurine tract then serves as a primer for plus-strand synthesis of the 3'LTR and complementary PBS. The newly-synthesized plus-strand PBS then associates with the already-synthesized minus-strand PBS, and double-stranded cDNA is finally produced. Double-stranded cDNA is then transferred to the nucleus by integrase proteins, and a new copy is integrated into the genome.

be found in all eukaryote kingdoms, indicating a very ancient origin. Except in vertebrates, these two superfamilies account for the vast majority of LTR-retrotransposons. Both *copia* and *gypsy* elements sometime encode an additional envelope (*env*) gene with surface and transmembrane units, giving to some retrotransposons the ability to infect other cells or other organisms. In *Drosophila*, gypsy elements have for example been shown to be potentially infectious (Kim et al., 1994).

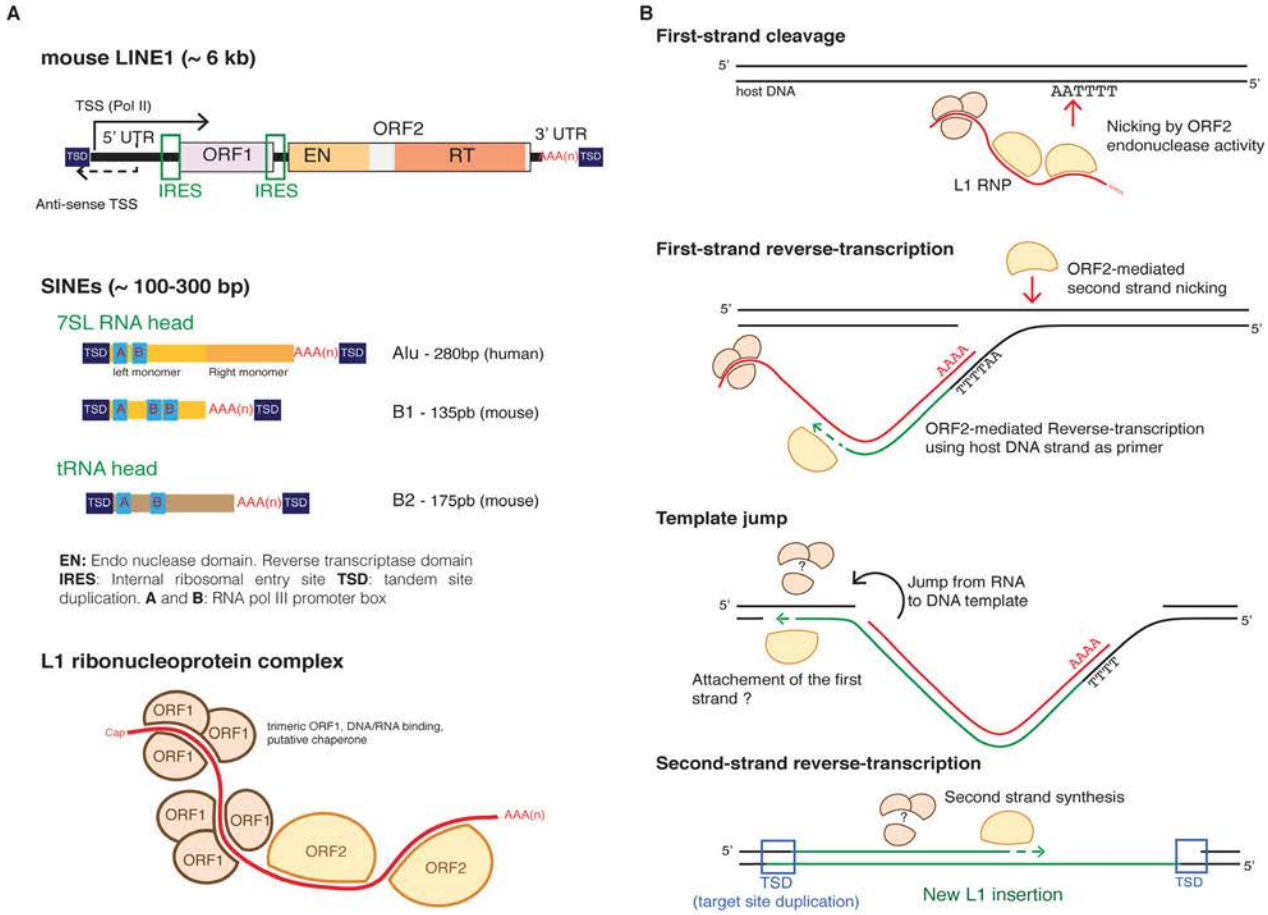
A lack of a functional *env* gene is what distinguishes LTR-retrotransposons from *bona fide* retroviruses. In vertebrates and especially in mammals, the vast majority of LTR-retrotransposon is in fact thought to have resulted from the endogenization of retroviruses, through inactivation or deletion of the domains that enable extracellular mobility (Boeke and Stoye, 1997). In mammals, endogenous retroviruses (ERVs) are usually further categorized into three classes, I, II and III, depending on the family of exogenous retroviruses (XRVs) they are related to. The respective evolutionary origins of LTR-retrotransposons and ERVs are complex and will be discussed later.

Many LTR-retrotransposons lack functional ORFs and cannot be replicated autonomously. They can nonetheless be mobilized *in trans* by functional elements, the minimal requirement being the presence of flanking LTRs and of the priming sequences in the PBS necessary for the initiation of reverse-transcription. Many non-autonomous LTR-retrotransposons can be identified in plants or animals, sometimes representing distinct families like the VL30 or malR elements in mice, and have sometimes accumulated to significant amounts (Stocking and Kozak, 2008). Additionally, in a majority of species, LTR sequences can be found alone, without internal coding sequences. In mammalian genomes, these "solo-LTRs" are ten-time more numerous than full-length copies. They are thought to result from homologous recombination between the two LTRs of a retrotransposon, leaving a solitary LTR after excision of most of the transposon sequence (Mager and Goodchild, 1989; Sverdlov, 1998).

1.2.3 LINEs

LINEs are autonomous retrotransposons found in all eukaryotic kingdoms. Five superfamilies have been described – R2, RTE, Jockey[1], L1 and I – based on RT domain phylogeny (Kapitonov et al., 2009; Wicker et al., 2007). All LINEs encode a least one protein, ORF2, which contains an RT and an endonuclease (EN) domains. Except for the evolutionary ancient R2 and RTE superfamilies, LINEs usually encode for another protein

[1] L2 elements are for example member of the Jockey superfamily



Introductory Figure 5. LINEs

A. Structure of murine LINE1 and SINEs. Bottom: proposed structure of L1 RNA-protein (RNP) complexes. ORF1 proteins form trimers, exhibiting RNA binding and nucleic acid chaperone activity.

B. Mechanism of target-primed reverse transcription (TPRT), directly at the site of integration. L1 RNP recognize AAAATT hexanucleotides and ORF2 endonuclease activity cleaves the DNA first-strand. L1 polyA tail associate with TTTT overhang and the host DNA is used as a primer to initiate reverse-transcription. ORF2 probably also mediate second-strand cleavage and attachment of newly synthesized cDNA to the DNA template, using again host DNA as a primer for second-strand synthesis.

named ORF1. LINE elements are relatively rare compared to LTR-retrotransposons in plants, fungi or insects, but are dominant in vertebrates and especially in mammals, where they represent around 20% of the genome. Most of the knowledge about LINE biology comes from the study of L1 elements in mice and humans, even though the retrotransposition mechanism was first elucidated from the study of an R2 element in the silkmoth *Bombix mori* (Luan et al., 1993).

Full-length mammalian L1s are typically around 6kb long. They are composed of a 5' untranslated region (UTR), which acts as an internal promoter, two ORFs and a 3'UTR containing a polyadenylation signal and a polyA tail (**Introductory Figure 5A** and Babushok et al., 2007). The 5'UTRs of mouse L1s contain a variable number of GC-rich tandemly repeated monomers of around 200pb, followed by a short non-monomeric region. Interestingly, the promoter activity of mouse 5'UTRs was shown to be proportional to the number of monomers (DeBerardinis and Kazazian, 1999). Human 5'UTRs on the other hand are ~900pb in length and do not contain such repeated motifs. All families of human L1s harbor in their most 5' extremity a binding motif for the transcription factor YY1 (Becker et al., 1993). Younger families have also two binding sites for SOX-family transcription factors, and both YY1 and SOX sites were shown to be required for human L1 transcription initiation and activation (Athanikar et al., 2004; Tchenio, 2000). Both mouse and human 5'UTRs contain as well a week antisense promoter of unknown function (Li et al., 2014; Speek, 2001).

L1s produce a single bicystronic mRNA encoding ORF1 and ORF2 proteins. In human and mice, the two ORFs are non-overlapping (although not necessarily in the same reading frame), but this is not the case in rats. In human, ORF2 is thought to be translated by an unconventional termination/reinitiation mechanism (Alisch et al., 2006), while mouse L1s contain an internal ribosome entry site (IRES) upstream of each ORF (Li et al., 2006b). ORF1 is a 40kDa protein that lacks homology with any protein of known function. In vertebrates, it contains a conserved C-terminus domain and a highly variable coiled-coil N-terminus that mediates the formation of ORF1 trimetric complexes (Martin et al., 2003). ORF1 trimmers have RNA-binding and nucleic acid chaperone activity that are necessary for retrotransposition (Martin and Bushman, 2001; Martin et al., 2003, 2005).

ORF2 is a 150kDa protein with an endonuclease domain (EN), an RT domain and sometimes, an RNAse H domain. The RNAse H domain is absent in mammalian L1s but is present in *Drosophila "I* factors". Both ORF1 and ORF2 proteins primarily associate *in cis* with their encoding mRNA, forming a ribonucleoprotein (RNP) complex, likely composed of two ORF2s and an unknown number of ORF1 trimers (Babushok et al., 2007; Kulpa and Moran,

2006). The complex is transported back into the nucleus, where the L1 endonuclease opens the DNA at TTAAAA hexanucleotide motifs (Jurka, 1997). Reverse-transcription then occurs directly at the site of integration trough a mechanism named target-primed reverse transcription (TPRT) (**Introductory Figure 5B**), which was originally described for an LINE R2 in silkworms (Cost et al., 2002; Luan et al., 1993). New insertions create short TSDs, and the majority of new inserts are severely 5'-truncated (average insert size of 900pb in humans) and often inverted (Szak et al., 2002). Because they lack their 5'UTR, most of new inserts are non functional. LINE proteins sometimes fail to associate *in cis* with their encoding mRNA and were shown to mobilize *in trans* various genic mRNAs, creating most of the time "dead-on-arrival" pseudogenes lacking introns and promoters (Esnault et al., 2000).

1.2.4 SINEs

SINEs are the only TEs that are non-autonomous by nature, meaning that they did not evolve from autonomous elements. They are small (80-500pb) and rely in *trans* on functional LINEs for their replication, but their evolutionary origin is very distinct. SINEs can be found in very diverse eukaryotes, but they have only accumulated to impressive amount in mammals, where they represent between 5 and 15% of the genome with millions of copies. SINEs typically possess a "head" with an RNA pol III promoter that enables autonomous transcription, and a body of various composition (**Introductory Figure 5A** and Kramerov and Vassetzky, 2005). SINEs replication mechanism was shown to rely on LINEs, either in human or in fish (Dewannieux et al., 2003; Kajikawa and Okada, 2002). SINE RNAs form a complex with LINE ORF2 proteins and are inserted into the genome by TPRT, creating short TSDs upon insertion. Some SINE families are thought to rely on specific LINEs for their replication, while others seem to be more generalist.

SINEs are postulated to originate from the accidental retrotransposition of various pol III transcripts, and have appeared separately numerous times in evolution history (Kramerov and Vassetzky, 2011). The type of pol III promoter defines the different superfamilies and reveal their origin: tRNA, ribosomal 5S RNA or signal recognition particle 7SL RNA. *Alu* and B1 elements, with their 1.1 million and 650,000 copies in the human and mouse genomes, respectively, harbor a 7SL promoter. The 350,000 copies of B2 SINEs in the mouse are on the other hand tRNA-related (Vassetzky and Kramerov, 2013). The origin of the 3' region is more obscure, and is thought to contain sequence elements allowing recognition by the LINE proteins. Some SINE 3'-regions are indeed in some cases very similar to the 3'-end of a LINE of the same genome (Kajikawa and Okada, 2002; Okada et

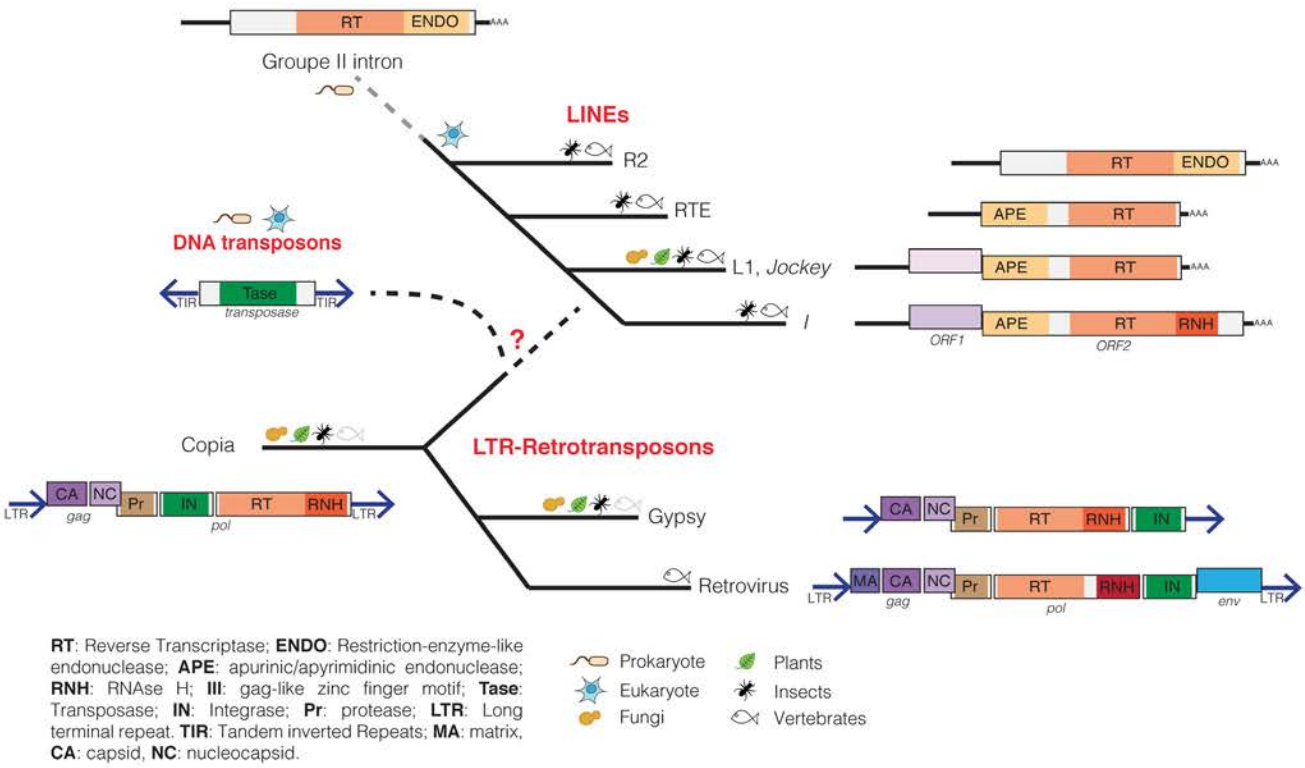
al., 1997). In other cases, the 3' end of SINEs can be either A- or AT-rich, with short tandem repeat or a poly-T tail (the pol III termination signal). The tail of *Alu* was, for example, shown to be essential for their mobility (Dewannieux et al., 2003).

LINE proteins preferentially assemble in *cis* with their encoding mRNA, directly after translation. The mechanism by which SINE RNAs are transported to the cytoplasm and incorporated into LINE RNPs in place of LINE RNAs remains poorly understood. However, the majority of SINEs are transcribed from promoters of RNAs usually involved in the translation machinery. Some elements, like *Alu* and B1, can still form complexes with protein associated with ribosomes (Weichenrieder et al., 2000). It has been proposed that most SINEs maintain the ability to associate with the translation machinery, giving them the opportunity to present their 3'-end to newly translated LINE proteins (Kramerov and Vassetzky, 2005).

Hominid genomes contain original elements termed SVA. They are composite transposons formed by the fusion of a SINE-R and an *Alu*, separated by a variable number of tandems repeats (Ostertag et al., 2003). Less than 3kb in length and apparently mobilized using L1 machinery, they are around 2500-3000 copies in human or gorilla genomes, and less than 1000 in orangutan (Wang et al., 2005). SVA are one of the youngest TE in great apes genome and among the most active and polymorphic in the human population.

1.2.5 Evolutionary origin of Transposable Elements

DNA transposons and retrotransposons can be found in virtually every eukaryotes, including very primitive unicellular protozoans, like the intestinal parasite *Giarda Lambia* (Arkhipova and Meselson, 2000). The conservation of the structure and mode of replication of TEs indicate that they appeared very early in evolution, and their presence seems to be a constitutive feature of eukaryotic genomes. Even if numerous TEs have the capacity to horizontally infect other cells or organisms, the transmission of TEs is thought to primarily occur vertically from one cell to its daughter cells (Malik et al., 1999). Because TEs are mobile by nature, it is easy to imagine that the integration of a transposon into a host gene or into other transposons would allow the acquisition of increasingly complex new functions (Lerat et al., 1999; Malik and Eickbush, 2001). The vertical mode of transmission enables the reconstruction of an evolutionary history of TEs, from prokaryotic precursors to highly complex retroviruses (**Introductory Figure 6**). However, some crucial steps remain obscure and speculative.



Introductory Figure 6. Phylogeny of transposable elements.

Primitive LINEs (R2 elements in B. mori for example) possess a single ORF composed of the RT domain and a sequence-specific endonuclease. LINEs exchanged this restriction-enzyme-like endonuclease for a less stringent apurinic/apyrimidinic endonuclease and also acquired a second protein, ORF1, with RNA binding and chaperone activity. Some of the more complex LINEs have an additional RNAse H domain, like *I* factors in Drosophila.

LTR retrotransposon could have evolved from the fusion between an advanced form of LINE and a DNA transposon. ORF2 would have brought the RT and RNH domain, ORF1 could have evolved into gag, and a DNA transposon could have contributed to the IN gene.

Copia and *Gypsy* (including endogenous and exogenous retroviruses) differ by the ordering of protein domains: IN is upstream of RT-RNH in *Copia* and downstream in *Gypsy*. Retroviruses have evolved away from other *gypsy* elements: their RT domain have diverged, and their gag polyproteins encode an additional matrix domain used to direct the virus particle to the plasma membrane. Retroviruses have also acquired a new RNAse H domain and the ancestral one has degenerated into a tether (connection) domain. The env genes of retroviruses have multiple sources, and probably originate from sequences captured from the host or other virus genomes

DNA transposons are thought to originate directly from the prokaryotic world. "Insertion Sequences" in bacteria move by a mechanism that is nearly identical to DNA transposons in eukaryotes (Cerveau et al., 2011). Some superfamilies of transposases even appear to be shared between animals and bacteria, indicating that their emergence preceded the evolution of eukaryotes (Feschotte, 2004).

The defining feature of autonomous retrotransposons is the presence of a reverse transcriptase (RT) domain. Sequence similarity of the RT region has been used to establish phylogenetic analyses of retroelements (Xiong and Eickbush, 1988, 1990). LINEs appear to be the most ancient retrotransposons and are probably of prokaryotic origin. Their RT domain and TPRT mode of replication is similar to group II introns, a class of mobile genetic elements found in bacteria and mitochondria (Lambowitz and Zimmerly, 2011; Malik et al., 1999). LINEs are also the only retrotransposons found in the very primitive eukaryote *Giarda lambia*, (Arkhipova and Meselson, 2000). Primitive LINEs, like the well-studied R2 elements of *B. mori*, possess a single ORF composed of the RT domain and of a sequence-specific endonuclease similar to bacterial restriction enzymes (Luan et al., 1993). LINEs exchanged this restriction-enzyme-like endonuclease for a less stringent apurinic/apyrimidinic endonuclease, probably coming from the DNA repair machinery of the host cell (Malik et al., 1999). They also acquired a second protein, ORF1, with RNA binding and chaperone activity. Some of the more complex LINEs have an additional RNAse H domain, like the *I* factors in *Drosophila* (Malik, 2005; Wicker et al., 2007). RNAse H is thought to be necessary to remove the template RNA from the newly synthetized cDNA. This domain is absent from mammalian L1s, which appears as a more primitive lineage that probably use host RNAse H to carry out this function.

Based on RT domain phylogeny, LTR-retrotransposons (including ERVs and XRVs) form a monophyletic group. Vertebrate retroviruses cluster together within the *gypsy* superfamily, while *copia* is separated (Malik et al., 1999; Xiong and Eickbush, 1988, 1990). This suggests that retroviruses evolved from *gypsy* elements by acquisition of *env* genes and additional regulatory sequences, giving them the opportunity to invade other cells. Invertebrate retroviruses are structurally similar to *gypsy* retrotransposons, except for the presence of *env* genes that were probably acquired through recombination with other dsDNA or ssRNA viruses (Malik, 2000; Pearson and Rohrmann, 2002). Vertebrate retroviruses on the other hand have evolved separately from the rest of LTR-retrotransposons. Their RT domain has diverged away from other *gypsy* retrotransposons and their gag polyproteins encode an additional matrix domain used to direct the virus particle to the plasma membrane

(Sandmeyer and Clemens, 2010). Vertebrate retroviruses have also acquired a new RNAse H domain and the ancestral one has degenerated into a tether (connection) domain (Malik and Eickbush, 2001). The *env* gene of retroviruses has multiple sources, likely originating from sequences captured from the host or from other viruses (Kim et al., 2004). These observations converge towards the possibility that complex retroviruses have evolved from simpler LTR-retrotransposons rather than the opposite. In fact, almost all mammalian LTR-retrotransposons seem to result from the endogenization of retroviruses that have lost the ability to infect other cells (Boeke and Stoye, 1997). Ironically, mammalian LTR-retrotransposons are therefore the descendant of elements that initially managed to escape their host genome, but were then trapped into another genome.

The emergence of the first LTR retrotransposon is obscure, and it has been proposed to have occurred through the fusion between an advanced form of LINE and a DNA transposon (Malik, 2005; Malik and Eickbush, 2001). Like DNA transposons, LTR-retrotransposons integrate double-stranded DNA molecules into the genome and require tandem repeat flanking the insert. A DNA transposon could have brought the integrase activity and the necessity to harbor flanking repeats to early LTR-retrotransposons (Capy et al., 1997). The LINEs *I* in *Drosophila* possess an RNase H domain and its ORF1 contains some zinc finger motifs that are reminiscent of gag proteins (Martin, 2006; Wicker et al., 2007). Based on RNAse H domain phylogeny, it has been proposed that LTR-retrotransposon *pol* gene could have evolved from an *I*-like LINE element and the *gag* from ORF1 (Malik, 2005). In line with this hypothesis, the entire lineage of LTR-retrotransposons appears to be no older than the youngest lineage of LINEs (Malik and Eickbush, 2001). Even if this model is speculative, and can not explain how LTR-retrotransposons developed their complex retrotransposition mechanism, the only missing domain after a fusion between a LINE and a DNA transposon would be the protease domain, which could have originated from an ancestral form of the host pepsin gene (Lin et al., 1992).

1.3 Birth, life, death and afterlife of transposons

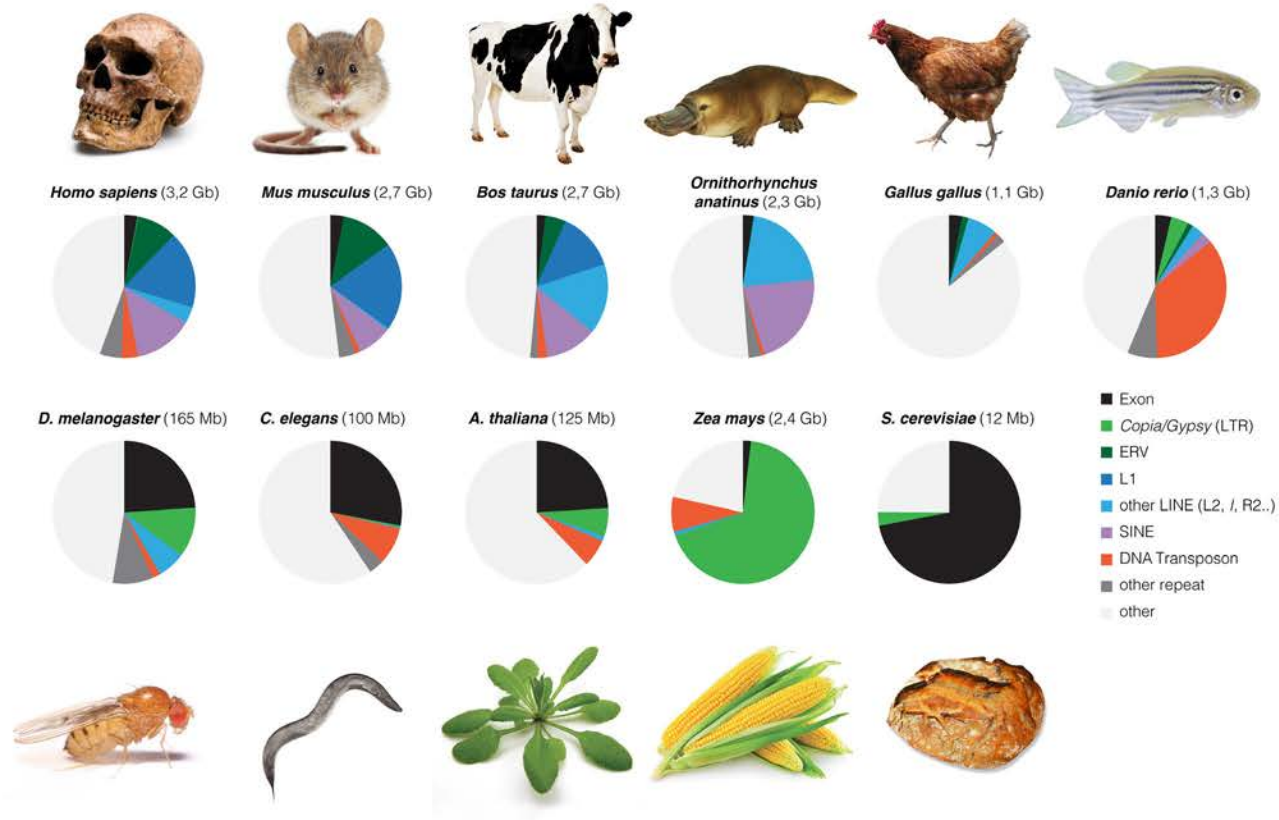
TEs are a constitutive feature of every eukaryote genomes. However, there is an important diversity of TE distribution and genome composition between species. In order to understand how TEs contribute to the genomic structure, it is necessary to analyze in details the forces that facilitate or restrict their expansion.

1.3.1 Transposons: an heavy, diverse and ancient genomic load

Because of their replicative nature, TEs have accumulated to significant proportion in the majority of eukaryotes, and many different families of TE usually coexist inside the same genome (**Introductory Figure 7**). However, the vast majority of them are usually not functional anymore. Transposition itself often introduces errors and many new inserts carry important mutations. For example, the majority of new LINE insertions in mammals are severely 5' truncated (Szak et al., 2002). Beside this, TEs are rarely under positive selection pressure and rather accumulate mutations at a neutral rate over evolutionary time. Most TEs are therefore merely molecular fossils, truncated, mutated and unable of further mobilization. Out of the millions of copies that can populate a genome, only a subset is thought to be potentially active. Among the 500,000 L1 copies that populate the human genome, it has been estimated that between 5-7,000 are full-length (Khan et al., 2006; Lander et al., 2001), that 80-100 are retrotransposition-competent, and that only six "hot L1s" contributed to the majority of retrotransposition in the human population (Brouha et al., 2003).

TEs are not subject to the same selective constraints as protein coding genes. While the coding part of the genome can be relatively similar even between distantly related species, the transposon-derived fraction can evolve rapidly into a great diversity. One good example is the comparison between the two distantly related frogs *Nanorana parkeri* and *Xenopus tropicalis* (Sun et al., 2015). Despite having diverged 266 Myrs ago, the two frogs have conserved a considerable genome synteny, but greatly differed in their TE content. The genome of *N. parekeri* is significantly larger than the one of *X. tropicalis* (2.3 vs 1.5Gb): TEs account for most of this difference, both in term of genomic mass (970 *vs* 318Mb of TEs) and composition (mostly *gypsy* LTR retrotransposons in *N. parkeri* and DNA transposons *X. tropicalis*).

The comparison of TE composition between species reveals important differences in the histories of expansion, contraction and activity of their genomes (**Introductory Figure 7**). *Saccharomyces cerevisiae* harbors a very small genome (12Mb) and most of the genomic space is occupied by exonic sequences. The budding yeast genome is probably under strong selective pressure to maintain a small size and TEs only represent 3% of the DNA. The intestinal and intracellular fungal parasite *Encephalitozoon intestinalis* represents an extreme case: it has the smallest eukaryotic genome ever sequenced (2.3Mb) and is the only eukaryote that apparently lacks TEs. Its genome is in fact so compact that non-coding sequences appear more conserved that genes themselves (Corradi et al., 2010). *D. melanogaster, C. elegans* or *A. thaliana* have also relatively dense genomes (100-150Mb). Exons occupy around one quarter of the genetic space and TEs less than 15% (Arabidopsis Genome Initiative, 2000; Smit et al., 2013). Interestingly,



Introductory Figure 7. Genome composition
Genome composition of different species. Data for animals was obtained from the Repeatmasker database (Smit et al., 2013)

most of the TEs in these three genomes are relatively young, less than 20 millions year old, and often active (Kapitonov and Jurka, 1999, 2003).

These features are in strong contrast with vertebrate genomes, which are usually much bigger (>1 Gb). Exons represent only 2-3%, while old fossils of TEs are the main components of the vertebrate DNA. It is commonly assumed that around 50% of the human genome is composed of TEs, but this number is likely underestimated because very ancient TEs are almost impossible to recognize and annotate. The TE-derived fraction of the human genome might in fact be closer to two-third (de Koning et al., 2011). The diversity of TE composition in vertebrate genomes is astonishing and reflects the diverse successful invasions of different TE families in separated lineages. The zebrafish genome is composed of 35% of DNA transposons; mammalian genomes are dominated by LINEs and SINEs; and *gypsy*-like LTR-retrotransposons can make up to 30% of the Salamander genome (Smit et al., 2013; Sun et al., 2012a). Birds on the other hand have the smallest genome among tetrapods (~1Gb), which contains only 10% of TEs.

Except for a set of "hot L1s", most human TEs are inactive. LINEs are in fact the most successful transposons in all mammalian species, where they usually represent around 20% of the genome. Mouse and human genomes are dominated by L1s, but other LINE superfamilies have been successful in other mammals: Bov-B, an RTE-type LINe, represents 15% of the cow genome. L1s have been continuously amplifying for the last 160 Myrs in mammals, and bursts of amplification have alternated with periods of low activity (Khan et al., 2006; Pascale et al., 1990). However, it is estimated that the rate of human TE amplification, including L1s, has been decreasing for the past 50 Myrs (Lander et al., 2001), except for a peak of Alu (SINEs) expansion that occurred around 40 Myrs ago. The remnants of past periods of amplification are, however, numerous. For example, the human genome carries 3% of DNA transposons, which were active until 40-50Myrs ago in the primate lineage (Pace and Feschotte, 2007). Similarly, L2 elements (2% of human DNA) have been immobile for a long time, but they probably played an important role in the past. In contrast, they represent 20% of the platypus genome, and domesticated L2-derived sequences are now involved in T cell-specific gene regulation in humans (Donnelly et al., 1999). The most prolific mammalian LTR retrotransposon elements (ERVL and MaLRs, 5.8% of the genome) greatly multiplied 100Myrs ago but have been extinct for the last 40 Myrs (Lander et al., 2001). With the potential exception of HERVK-HML2 (Subramanian et al., 2011), a recently acquired ERV that is expressed in cancers and diseases, LTR retrotransposons appear to be on the edge of extinction in humans.

The reasons for this decline in transposons expansion in the human lineage are obscure and no good explanation has been offered so far. The mostly quiescent human genome stands in opposition with the mouse genome, which harbors many active LINEs and ERVs. The composition of the murine genome will be discussed in details later in this manuscript.

1.3.2 A subtle equilibrium between opposing forces

The evolution of the genome and, especially, of its repeated fraction is not a linear process. When a new transposon copy inserts into the germline of an individual, this copy can be transmitted to the next generation and then spreads into the population. Interestingly, the multiplication of TEs seems to often occur by bursts of activity, followed by periods of decay. Theoretical models have been proposed to reconstitute the initial invasion (Le Rouzic and Capy, 2005) and long time evolution of TEs (Le Rouzic et al., 2007a). Mutations introduced by TEs (inactivation of genes, chromosomal breaks, translocations, etc.) can be deleterious for the host and its descendants. TEs were indeed shown to cause 50% of deleterious mutations in *Drosophila* (Finnegan, 1992), and 10-15% in laboratory mice (Kazazian, 1998; Maksakova et al., 2006). The spreading of TEs in a population is therefore essentially controlled by natural selection (Charlesworth and Charlesworth, 1983). Moreover, eukaryotes have developed an important and diverse range of defense mechanisms to protect themselves against the invasion of TEs (Slotkin and Martienssen, 2007). It is often assumed that the genome is the theater of a constant arms race between the TEs and their hosts (Lisch and Slotkin, 2011). Newly invading elements are first not recognized by the host and can proliferate into the genome until protective mechanisms evolve and slow down the multiplication process. Unable of further mobilization, TEs then accumulate genetic mutations that definitely inactivate them. On the long range, the accumulation of TEs is a subtle equilibrium between their own activity, host defense mechanisms and natural selection.

1.3.3 Appearance of new transposons

New transposons can appear in a population either vertically, by modification of an existing one, or horizontally by endogenization of sequences originating from other species or from viruses.

LINEs and SINEs are transmitted essentially vertically (Burke et al., 1998; Pascale et al., 1990). In mammals (at least), L1s constantly change their regulatory units by modifying their 5'UTR, while ORF1 and ORF2 sequences remain relatively conserved (Adey et al., 1994; Khan et al., 2006). The 5'UTRs of L1s are often completely unrelated, especially between

different species. Only the 5'UTR of closely related L1 families seem to originate from the modification of a common ancestor. On the contrary, L1s are thought to often acquire completely novel regulatory units by inaccurately switching template during the retrotransposition reaction (Hayward et al., 1997). The human L1 5'UTR was modified at least eight time in the 70Myrs of primate evolution (Khan et al., 2006). Analysis of the mouse genome reveals also that different L1 families (active or not) often recombine together, exchanging regulatory or coding sequences to form new mosaic elements with renewed activity (Saxton and Martin, 1998; Sookdeo et al., 2013). Interestingly, in most mammals analyzed, L1 evolves as a single lineage: a new family emerges from the modification of an existing one, amplifies to thousand of copies and becomes extinct after being replaced by younger elements. This is currently the case in humans, where all L1s have derived from the single dominant L1PA lineage over the last 40Myrs (Khan et al., 2006). Moreover, it appears that concurrent L1 lineages only coexist when they harbor different promoter types. It is the case in the mouse, where two lineages with different promoter types (A and F) are currently active; it occurred as well in humans past before the extinction of the L1PB lineage 40Myrs ago (Goodier et al., 2001; Sookdeo et al., 2013). As a good example of the arm race between TEs and their hosts, the apparition of a new L1 family with modified regulatory or coding units was often followed by a period of massive amplification of this new family. Moreover, the coiled-coil domain of the human ORF1 protein appears to be rapidly evolving and under high selective pressure (Boissinot and Furano, 2001; Khan et al., 2006).

Similar cycles of activity/quiescence can be observed for LTR-retrotransposons. For example, ERV-L is a very ancient LTR retrotransposon that is present in all placental mammals. The mouse genome contains the remains of several boosts of activity, and mouse ERV-L (MERVL) is still one of the most active murine TEs (Bénit et al., 1999). The ability to episodically modify their regulatory and coding sequences likely explains why some TEs, especially LINEs, originate from the beginning of eukaryote existence and are still active nowadays. Like perfect parasites, they are constantly adapting to their host.

Alternatively, the appearance of new TEs in a population can result from the horizontal transfer of foreign DNA. The endogenization of retroviruses in vertebrates is probably the most documented example. Exogenous retroviruses (XRVs) are similar to LTR retrotransposons in terms of structure and mode of replication, with the difference that virus particles leave (and often kill) the infected cells after retrotransposition. However, if a retrovirus manages to invade the germline of an individual and is not too harmful for its host, the retrovirus genome can be transmitted to the next generations. After the initial infection,

the *env* gene allowing cell-to-cell mobility is often lost or mutated, forcing the new transposons to adopt an entirely cell-autonomous life cycle. Virtually every LTR retrotransposons in mammals are the result of the endogenization of an ancient exogenous retrovirus (Boeke and Stoye, 1997). For example, MLV (Mouse Leukemia Virus) elements have integrated the mouse genome recently (<1.5Myr ago) and are still very similar to their exogenous counterparts, with some members having maintained infectious properties (Stocking and Kozak, 2008). Similarly, in Koalas, KoRVs have all the characteristics of a functional retrovirus and endogenization into the Koala population is an ongoing process (Tarlinton et al., 2006).

Examples of horizontal transfer are not restricted to ERVs but involve all types of TEs in plants, animals and fungi (Wallau et al., 2012). *Mariner* and P elements (DNA transposons), *I factors* (LINEs) and *Gypsy* (LTR retrotransposons) were shown to move between *Drosophila* or other insect species (Abad et al., 1989; Daniels et al., 1990; Robertson, 1993). In vertebrates as well, surprising cases of horizontal transfers were observed. Bats are the only mammals known to harbor many active DNA transposons. Multiple waves of *Mariner*, *hAT* or *Helitron* amplification have invaded the bat genome in the last 30Myrs and likely result from horizontal transfer (Ray et al., 2007, 2008). Bov-B LINEs are present in all ruminants. They represent 15% of the cow genome but are absent in related species (Jobse et al., 1995). However, closely related LINEs were observed in the genome of different snakes and lizards (Kordis and Gubensek, 1997). It was in fact established that Bov-B LINEs most probably originated from the horizontal transfer 40-50 Myrs ago of elements from scaled reptiles to the ancestor of ruminants (Kordis and Gubensek, 1998).

How exogenous TEs were horizontally transferred from one species to the other is often mysterious. Retroviruses (and some *gypsy* LTR-retrotransposons) are naturally infectious but the majority of TEs are normally unable to leave their host cell. It seems highly surrealistic that a TE could jump directly from a snake to a cow: some sort of vector is necessary to mediate the physical transfer of DNA between the donor cell and the recipient germline. This vector needs to have access both to the intracellular and extracellular environment. Viruses, parasites (insects, mites…) or intracellular parasites make suitable vectors for transfer between natural populations, and examples of their involvement are numerous and increasing (Silva et al., 2004).

Bats are important reservoirs of viruses (Calisher et al., 2006) and it has been postulated that bat DNA transposons could have used dsDNA viruses as carriers to enter the bat genome (Pace et al., 2008; Ray et al., 2008). By contrast with retroviruses that have ssRNA genomes

and need to integrate cDNA into their host genome, dsDNA viruses never use RNA intermediates nor integrate into the host. Virus specialty is precisely to introduce new genetic material into foreign cells. By jumping into the DNA of invading virus instead of their host genome, TEs can find a way to be transported out of their genome of origin and potentially toward completely unrelated species (depending on the infectious range of the virus). Accordingly, the insertion of TEs into dsDNA virus has been observed in insects, with the case of a *gypsy*-like retrotransposon transposing into the circular genome of a Baculovirus (Friesen and Nissen, 1990), or in vertebrates by the presence of a snake-specific SINE in the genome of a poxvirus infecting west African rodents (Piskurek and Okada, 2007).

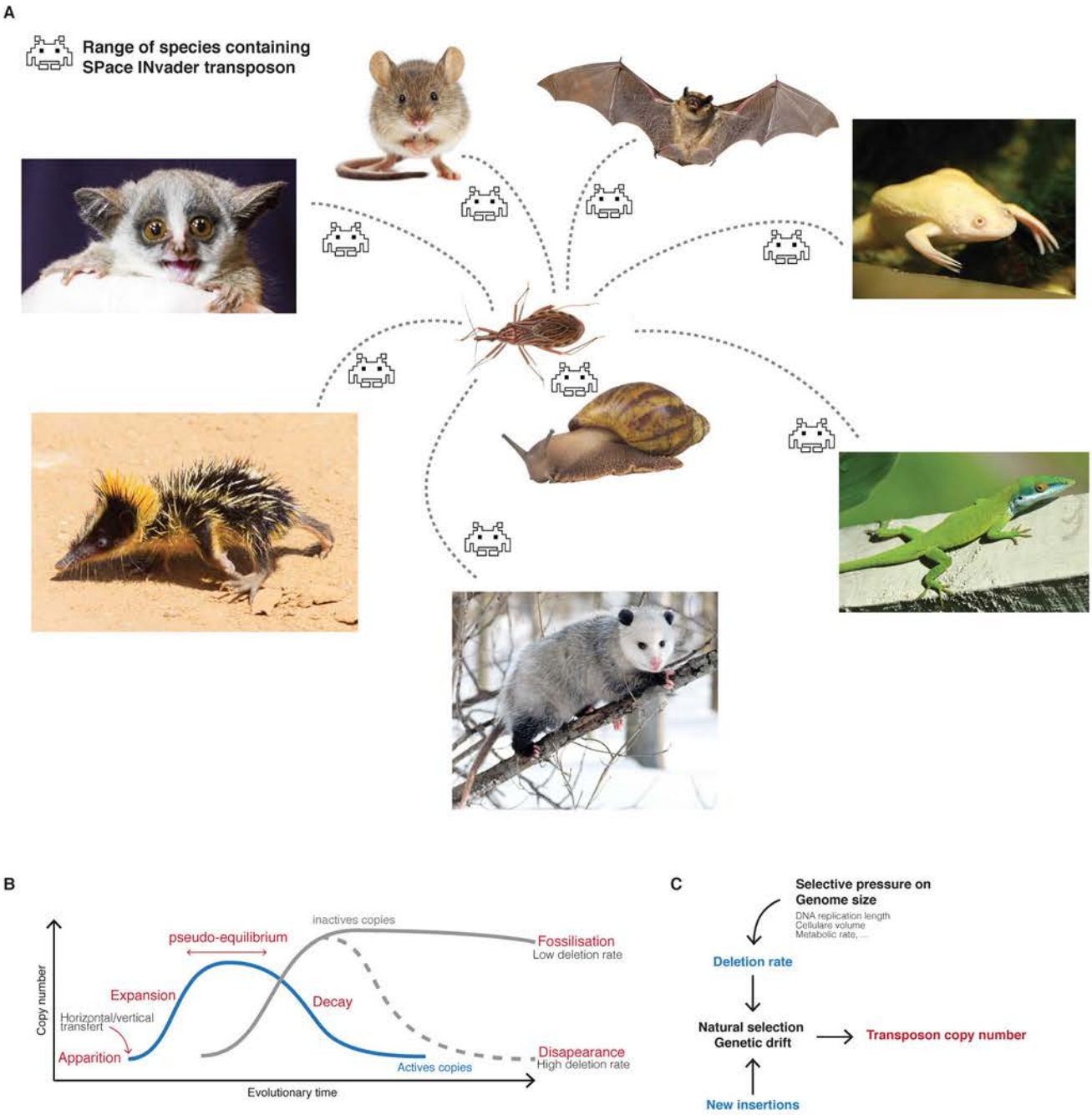
Other lines of evidences suggest that TEs can use different parasites as carriers between species. It was for example postulated that *P* elements were transferred between *Drosophila* species by a semi-parasitic mite (Houck et al., 1991). Another DNA transposon was also observed moving from a moth to its parasitoid wasp (Yoshiyama et al., 2001). In vertebrates, a family of *hAT* DNA transposon, SPINs[2], was shown to be highly conserved between a lizard, a frog and five mammalian species (rodents, a primate, a bat, a tenrec and an opossum), but could not be detected in closely related species (Pace et al., 2008). SPINs were further identified in the genome of a triatomine bug, which feeds on the blood of various tetrapods, and in a freshwater snail, which is a known intermediary for the parasite (Gilbert et al., 2010). The combination snail/triatomine bug probably allowed the transfer of SPINs between these very distant tetrapod species (**Introductory Figure 8A**).

The favorite mode of appearance of new TEs, either by vertical modification or horizontal transfer, probably varies across species and, as a consequence, affects the diversity of TEs in various genomes. Species where horizontal transfer is prevalent (like probably *Drosophila*) are expected to harbor many different and diverse TE families (Silva et al., 2004). On the other hand, mammals seem to have comparatively few horizontal transfers, maybe because their well-developed immune system efficiently guards against the transfer of DNA by infectious vectors (Lander et al., 2001). Mammals tend in general to have few TE families, but composed of many members. Indeed, half of TEs in the mouse and human genome come from the different variants of a single family of LINEs (Khan et al., 2006).

1.3.4 Initial amplification and long-term maintenance

TEs are present as hundreds or thousands of copies in eukaryotic genomes. Every successful TE families originally derive from a unique element, which integrated into the

2 For SPace INvaders…



Introductory Figure 8. Dynamic life of transposable elements

A. Example of horizontal transfer of the DNA transposon SPIN (SPace INvaders) between seven tetrapod species: counterclockwise from top, little brown bat (laurasiatherian), rat/mouse (murine rodents), bushbaby (prosimian primate), tenrec (afrotherian), opossum (marsupial), and two non-mammalian tetrapods (anole lizard and African clawed frog). The triatomine bug Rhodnius prolixus and the pond snail Lymnaea stagnalis are potential vectors for the spreading of SPIN transposons (Gilbert et al., 2010; Pace et al., 2008).

B. Model for the expansion of transposable elements into the genome. After the initial invasion, the number of active copies increases rapidly until it reach a pseudo-equilibrium state where the gain of new copies is compensated by the disappearance or the mutation of existing one. After the genome evolved efficient counter-measures, the number of active copies decreases inexorably. In species with low deletion rates, inactive copies fossilize and remain in the genome for extended period of time, while genomes with high deletion rates eliminate rapidly any inactive copies.

C. Evolutionary forces acting on transposon copy number.

germline of a single individual, started to be expressed and increased his copy number throughout the population (**Introductory Figure 8B**). However, invasion by a new TE is not always possible, especially if the founder element originates from the horizontal transfer from a distant species; TEs are often adapted to their host and transposition into a different environment might not be successful, especially if the transposition process requires specific host factors. *P* elements use the *Drosophila* protein IRBP (inverted repeat binding protein) to be excised from their locus (Beall and Rio, 1996; Beall et al., 1994). In relationship with the polymorphic presence of this factor, *P* elements can only mobilize in species of the *Drosophilidae* family, and not even in non-drosophilids fruit flies (*Tephritidae*) (O'brochta and Handler, 1988). On the contrary, *Mariner* DNA transposons only require their own transposase for mobilization (Vos et al., 1996), which probably explains why they are so widely distributed in eukaryotes (Plasterk et al., 1999). Along with the necessity to escape build-in defense mechanisms, adaptation of TEs to their former host probably restrains their ability to successfully spread into new organisms.

Once the first element is integrated into the genome, the spread of a new TE family in the population is a balance between the rate at which new copies arise and the rate at which they are lost (Charlesworth and Charlesworth, 1983). Loss of a functional TE can happen trough various mechanisms, including random genetic drift (the chromosome carrying a transposon copy is not passed to the next generation by chance), inactivating mutations, and natural selection against deleterious consequences. Mutations by small or large insertions/deletions or nucleotide substitutions occur at a rate of around 0.1-100 per genome per generation (Drake et al., 1998; Lynch, 2010); this is several orders of magnitude lower that the transposition rate of an invading transposon, which can potentially adds several new copies per generation. Immediately after the initial invasion, spontaneous mutations are therefore only playing a negligible role and the spreading of a TE in a population is principally governed by natural selection and genetic drift (Charlesworth et al., 1994). In the first generations after the initial invasion, theoretical models show that low rates of transposition likely lead to the rapid disappearance of the new transposon (Le Rouzic and Capy, 2005; Le Rouzic et al., 2007a). The probability of fixation of any mutation depends on the population size and the selective advantage or disadvantage brought by the mutation. With too few copies in too few individuals, a new transposon would very unlikely be able to get fixed in the population. On the other hand, a continuous high rate of transposition would cause a massive amplification and natural selection against the deleterious consequences would lead to the rapid elimination of the transposon from the population. These theoretical

models suggest in fact that a successful invasion need to be biphasic, with an initial burst of amplification characterized by a (moderately) high transposition rate, followed by a longer period where the transposition rate decreases (Le Rouzic and Capy, 2009). The initial burst of TE amplification was observed several times in natural and laboratory populations, and the most documented example is the rapid spread of *P* elements in all populations of *D. melanogaster* worldwide, in less than 50 years (Anxolabehere et al., 1988).

The reason for the decrease of transposition rate is not well understood. However, it is easy to speculate that once a TE managed to get fixed in a population, natural selection would favor individuals that either developed defense mechanisms against the invader, or have accumulated inactivating mutations of the invader. The appearance of non-autonomous elements (SINEs, MITEs or degenerated copies) that hijack the activity of autonomous ones was also proposed to slow down the accumulation of functional TEs (Feschotte and Pritham, 2007; Le Rouzic et al., 2007b). Indeed, autonomous and non-autonomous elements would compete for the same proteins, decreasing the probability to insert new functional copies.

Some authors have speculated that transposon copy number can reach an equilibrium, when the appearance of new elements is compensated by their loss (Biémont, 1994; Charlesworth and Charlesworth, 1983). However, such an equilibrium would likely be unstable and very transient (Le Rouzic and Capy, 2009). Indeed, the decay observed after the initial invasion is mainly irreversible. Once the host has developed mechanisms to prevent further transposition, active elements cannot be replaced and would either accumulate mutations or disappear from the population because of natural selection. Depending on the selective pressure, the mutation rate and the ability of a species to eliminate degenerated sequences, the general dynamics of the invasion could take different forms. Strong selective pressure and high mutation rate would rapidly eliminate remaining functional elements once the peak of amplification is passed. On the other hand, a low mutation rate and a weak selective pressure would allow for a very long decay phase, which may look like a pseudo-equilibrium situation (**Introductory Figure 8B**).

1.3.5 Death and fossilization

Once their activity starts decreasing, TEs begin to accumulate mutations and, given enough time, become molecular fossils. Human and mouse genomes share around 165Mb of common ancestral repeats that date back to more than 100Myrs (Waterston et al., 2002). In fact in the human genome, the majority of resident repeats precede the radiation of placental mammals (Lander et al., 2001). This very ancient origin of mammalian TEs is in strong

opposition with the observed age of transposons in many other species. For example, *D. melanogaster* contain only TEs younger than 20Myrs (Kapitonov and Jurka, 2003). All the LTR retrotransposons present in the euchromatic part of the *Drosophila* genome appear to be younger that their host species, having integrated well after the split with its closest relative (*D. simulans*) 3Myrs ago (Bowen and McDonald, 2001). Similarly, *S. cerevesiae* LTR retrotransposons are all younger than 100,000 years (Promislow et al., 1999) and the majority of *A. thaliana* LTR retrotransposons have less than 4Myrs of age (Devos et al., 2002). The maize has a genome size similar to mammals (2.4Gb) composed of 90% of TEs, but all these elements appear nonetheless to be not older than 6Myrs (SanMiguel et al., 1998).

These differences in the age distribution of TEs can be explained by the differential ability of species to eliminate nonessential DNA. Insertion of new TEs tend to increase genome size, and the maize genome has apparently gained 1.2Gb of TE sequences in the last 3Myrs (SanMiguel et al., 1998). However, despite continuous TE activity, the majority of individuals apparently managed to keep their genome size relatively constant, and some species like *A. thaliana* are even actively shrinking (Hu et al., 2011). This indicates that genomes have the ability to lose DNA over evolutionary time, counterbalancing the inflating pressure of TE mobilization.

It was shown in *Arabidopsis* and *Drosophila* that an important proportion of global DNA loss occurred by spontaneous deletion (Devos et al., 2002; Petrov et al., 1996). The mechanisms that generate insertions and deletions are generally not well understood. They could originate from mistakes during DNA replication (Levinson and Gutman, 1987), and it was shown that DNA repair after a DNA double-strand break often generates insertions and deletions ranging from a few base pairs to several kilobases (Kirik et al., 2000). In mammals, plants and insects, deletions systematically outnumber insertions, both in terms of average size and frequency, indicating that these genomes tend to naturally contract (Graur et al., 1989; Kirik et al., 2000; Petrov, 2002). By analyzing pseudogenes ("dead" sequences that evolve at a neutral rate), it was estimated that *Drosophila* is losing DNA 20 times faster than mammals (Petrov and Hartl, 1998), while the nucleotide substitution rate is similar (Petrov and Hartl, 1999). In *Drosophila*, deletions are both more frequent and larger in average than in mammals and the same appears true for *C. elegans* (Robertson, 2000). Nonessential sequences are expected to lose half of their DNA in 14Myrs in *Drosophila*, compared to 884Myrs in mammals (a period anyway so long that point substitutions would have made the sequence unrecognizable long before that) (Petrov and Hartl, 1998).

DNA loss can also occur by unequal recombination between non-allelic but identical regions. During meiosis, successful crossing-over involves the pairing of identical regions located on matching chromosomes. Unequal homologous recombination occurs when crossing-over happens between two non-allelic regions and can generate deletions, duplications, inversions and other alterations (Sasaki et al., 2010). Unequal recombination can occur as well during homology-dependent DNA double-strand break repair in somatic cells (Hurles, 2005). Because TEs of the same family are nearly identical, they represent a very good platform for unequal recombination. Solo-LTRs are precisely the consequence of unequal recombination between the two flanking sequences of LTR-retrotransposons (Mager and Goodchild, 1989). However, unequal recombination between two LTRs has probably a limited contribution to the overall DNA loss, at least in *Arabidopsis* (Devos et al., 2002). Similarly, recombination between two different TE copies located on the same chromosome could lead to the deletion of one of the copy and of the sequence located between them. In *Arabidopsis,* these events are even more rare that recombination between two-LTRs, probably because large structural variations would be often counter-selected. A clear estimation of the overall contribution of unequal recombination to total DNA loss is currently missing.

The differences in the rate of spontaneous deletions probably account for the important variation observed in the age of transposons between species. In species with a high deletion rate, non-expanding TEs are rapidly eliminated from the genome. Conversely, they can remain almost forever in species with a slow deletion rate (**Introductory Figure 8B**). Even if the rate is comparatively slow, DNA loss probably played an important role in mammals. Estimation of ancestral genome sizes from fossils indicates that the common ancestor of mouse and human may have had a genome of similar size (Organ et al., 2007). Given that the human genome contain around 700Mb of lineage-specific repeats, it indicates that the human genome lost an equal amount of DNA since its divergence with the mouse (Waterston et al., 2002).

1.3.6 Solving the C-value paradox

The amplification of TEs is one of the main forces that increases genome size and TEs often contribute to genomic gigantism, like in salamanders where genome size ranges from 14 to 120Gb (Sun et al., 2012a). On the other hand, DNA loss (most likely by spontaneous deletion) has been proposed to be the main counterbalancing force (Petrov, 2001; Petrov et al., 2000), even if other mechanisms are likely involved (unequal recombination, genes and chromosome duplication/deletion, simple repeats expansion/shrinkage, etc.). To present it in

a simple manner, genomes with a high transposition rate increase in size, while genomes with a high DNA loss rate shrink. For example, *Drosophila*, which loses DNA 20-time faster than humans, has 20-fold smaller genome [3]. Genome size is therefore a balance between these two counteractive forces (**Introductory Figure 8C**). In order to further test that hypothesis, the rate of DNA loss at pseudogenes was compared between *Drosophila* (genome size of 165Mb) and two other insects: *Laupala* Hawaiian crickets (1.9Gb) and the grasshopper *Podisma pedestris* (18.1Gb) (Bensasson et al., 2001; Petrov et al., 2000). In good correlation with its 11-time bigger genome, *Laupala* tends to lose DNA 40 times slower than *Drosophila*. *Podisma* loses DNA at an even slower rate, which probably explains its enormous genome. The deletion rate is in fact not significantly different than the insertion rate in *Podisma*, meaning that genome size cannot decrease and that new TE insertions would remain virtually forever, until nucleotide substitution render them unrecognizable. Similarly, Salamanders have both the biggest tetrapod genome and the slowest DNA loss rate (Sun et al., 2012b).

Why species have small or big genomes is probably not random. Genome size itself, independently of the sequence content, is a cellular characteristic with many physiological consequences upon which natural selection can act. Bigger genomes for example need more time to replicate during mitosis and meiosis (Bennett, 1971). Genome size therefore puts a limit on cell-cycle length and on the speed at which an organism can develop. Species that need to develop fast are expected to require smaller genomes: in herbaceous plants, ephemeral species have indeed significantly less DNA than species that live several years (Bennett, 1972). Similarly, among salamanders, species with the biggest genome appear to have also the longest embryonic development time (Jockusch, 1997).

Genome size also strongly correlates with cellular size (Cavalier-Smith, 1982; Szarski, 1970), which has a direct influence on a very wide range of mechanisms. In plants, genome size was shown to positively correlate with seed and pollen size (Grotkopp et al., 2004; Knight et al., 2010), especially for species that rely on the wind for their dispersion. In vertebrates, cell size has a direct impact on the cell metabolic level (the rate at which a cell consume energy) (Szarski, 1983). For example, an increase in cell size diminishes the energy necessary to maintain the osmotic pressure between the interior and the exterior of the cells. An expanding cell size can represent a good strategy for species leaving in energy-restricted environments. Indeed, some salamanders or lungfishes leave in oxygen-poor fresh waters and have a very low basal metabolic rate with very large cells (Licht and Lowcock, 1991; Szarski, 1970). Conversely, small cells have bigger surface/volume ratios that facilitate gas exchanges and

[3] Sometimes, numbers are magical.

they are probably a necessary condition to maintain a high metabolism (temperature regulation, high mobility, complex nervous system), which is especially critical for red blood cells that need to travel through small blood vessels (Gregory, 2001; Szarski, 1983). Probably because of the high-energy cost required by flight, birds have the highest metabolic rate of vertebrates and smaller cells than mammals (Szarski, 1983). Interestingly, they have also the smallest vertebrate genomes (around 1Gb), with a very low TE content (around 10% in chicken) (International Chicken Genome Sequencing Consortium, 2004). It was proposed that small genome size was a necessary feature to evolve the high metabolism required by flight (Hughes and Hughes, 1995), and indeed this hypothesis is reinforced by the observation that bats, the only flying mammals, have also the smallest mammalian genomes (Van den Bussche et al., 1995). Interestingly, estimation of ancestral genome size in dinosaur fossils indicate that genome size reduction in the avian lineage long predated the acquisition of flight (Organ et al., 2007).

These examples show that, in certain conditions, genome size can be a highly selectable character that influences organism development and metabolism. This selective pressure on genome size constrains the balance between amplification and deletion mechanisms (**Introductory Figure 8C**). Species that favors big cell size like salamanders, or where shrinking constraints are suddenly relaxed like in maize, would accumulate a lot of TEs. On the other hand, organisms like *Drosophila* or *Arabidopsis* have small genomes and high deletion rates, and maintain their genome small by actively eliminating DNA. Interestingly, it was shown that natural selection is more likely to act on the mechanisms that create the deletions (fidelity of DNA replication or of DNA repair for example), than on the individual deletions themselves (Petrov and Hartl, 2000). Bird genomes are also small for vertebrates, but birds probably use different mechanisms to control TE expansion. Indeed, birds TEs are very old (International Chicken Genome Sequencing Consortium, 2004), which indicates that birds probably maintain their genome small by somehow preventing efficiently the fixation of new TE sequences.

Distribution of TEs can therefore be viewed as an indirect consequence of the selective pressure acting on genome size. Mammals have moderately big genomes that do not appear to be under strong decreasing or increasing pressure. As a consequence, TEs are let free to accumulate and remain indefinitely as molecular fossils.

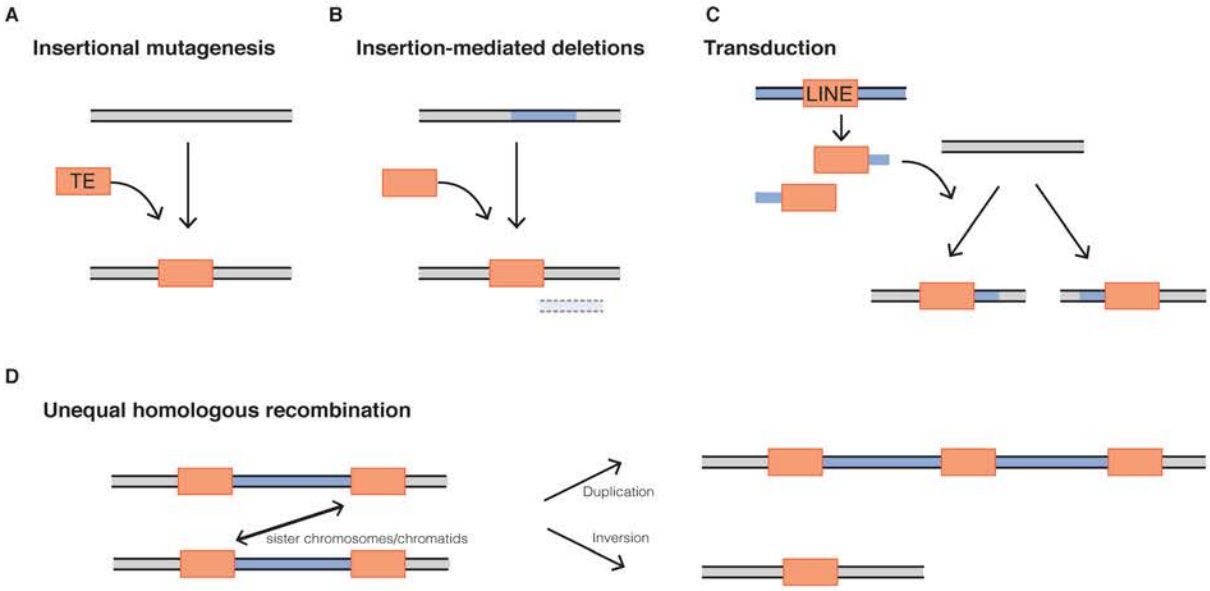
1.4 Distribution and contribution of TEs

We have reviewed in details the dynamic forces that control the expansion of TEs. Parasitic by nature, TEs play no systematic role beyond their own selfish replication. However, their self-centered proliferation has important consequences for the host genome. Before analyzing in depth TEs in the mouse model, it is important to understand how TEs are spatially distributed and how they sometimes directly contribute to genome function.

1.4.1 TEs create genetic variation, the substrate of natural selection

Evolution acts by selecting individuals that present advantageous variations compared to the rest of the population (Darwin, 1859). By naturally creating genetic variation, TEs are widely recognized as major drivers of evolution (Fedoroff, 2012). TEs cause 50% of spontaneous mutation in *Drosophila* (Finnegan, 1992). In the mouse, ERV insertions make up about 10-12% of mutations (Maksakova et al., 2006), and L1s account for another 2-3% (Druker and Whitelaw, 2004). TEs are mostly silent in humans: they are responsible for only ~0.3% of human mutations, and all the reported insertions concern L1, *Alu* and SVA elements (Callinan and Batzer, 2006). Nevertheless, based on the frequency of disease-causing insertions and on comparisons between the human and chimpanzee genomes, it is estimated that a new *Alu* element inserts into the genome every 20 births (Cordaux et al., 2006), and an new L1 every 20 or 200 births, depending on the study (Kazazian, 1999; Xing et al., 2009).

The mutations created by TEs are of various types (**Introductory Figure 9**). The most straightforward is the direct insertion into a new locus, potentially disrupting regulatory or coding sequences. The LINE protein machinery is also able to insert *in trans* other types of RNAs, leading to the insertion of SINEs or even cellular RNAs. Typical TE insertions have short TSDs on both sides, but around 0.5-0.7% of human L1 and *Alu* insertions are non-canonical and lack TSDs (Sen et al., 2007; Srikanta et al., 2009). These insertions were shown to be independent of L1 endonuclease activity and linked to DNA double-strand break repair (Morrish et al., 2002). It appears in fact that L1s and *Alus* can insert into an exiting DNA break and repair it. Moreover, around 20% of L1 canonical insertions (and up to 90% of endonuclease-independent ones) generate additional deletions at the site of integration, ranging from 1pb to more than 20kb (Gilbert et al., 2002; Symer et al., 2002). By opposition, L1s also sometimes insert additional DNA originating from the flanking genomic sequences of the donor element (Moran et al., 1999). Named 5'- or 3' transduction, these events probably occur when L1 transcription uses alternative upstream promoters or downstream



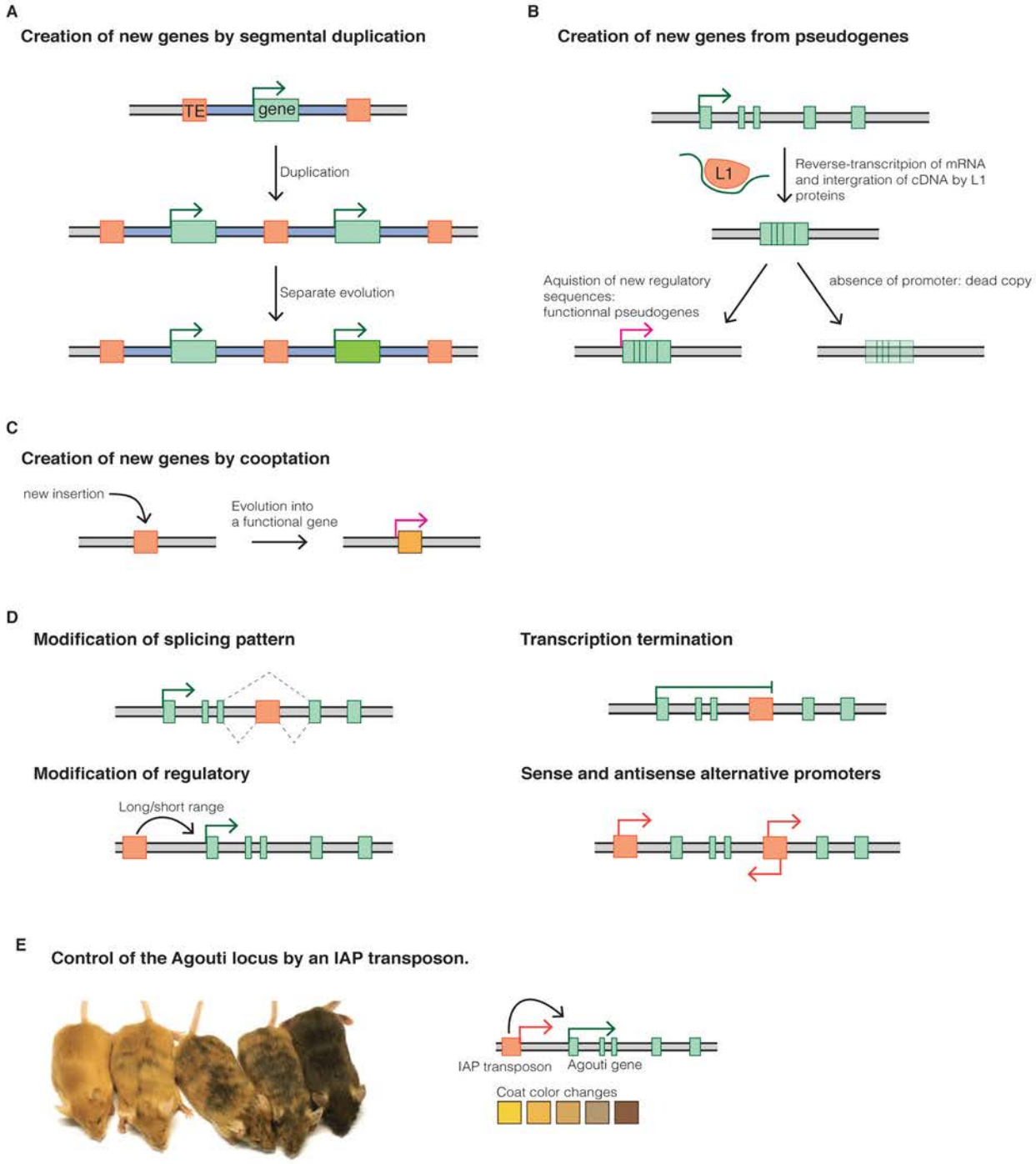
Introductory Figure 9. Mutagenic action of Transposable elements.
A. Insertional mutagenesis
B. Insertion-mediated deletion: TE insertion sometimes associates with the deletion of genomic DNA around the site of integration.
C. 5' or 3' transduction: LINE retrotransposition sometime carry over genomic DNA from the donor site.
D. Unequal homologous recombination between non-allelic but identical copies result either in deletion, inversion, or duplication of the sequence located between the two copies.

polyadenylation signals. 3' transduction takes place in around 15-20% of new L1 insertions and adds on average 200pb (Goodier et al., 2000; Pickeral et al., 2000); 5'-transduction appears less frequently, probably because most L1 insertions are 5'-truncated, but can nonetheless be observed in cell-based assays (Symer et al., 2002).

Deletion of genomic DNA upon insertion or transduction of additional sequences from the donor site is concomitant to the *de novo* integration. Another type of chromosomal rearrangement occurs when TEs are already integrated into the genome, and principally involves unequal homologous recombination between non-allelic but identical TEs. Recombination between TEs located on different chromosomes would lead to big chromosome translocations. If located on the same chromosome, unequal recombination would cause either the deletion, the duplication or the inversion of the sequence lying between the two copies (Sasaki et al., 2010). Recombination events between ERVs were reported (Hughes and Coffin, 2001), but because of their overrepresentation in human genomes L1s and *Alus* seem to be involved in the majority of cases. Numerous events of *Alu* recombination-mediated deletions are implicated in diseases and cancers (Deininger and Batzer, 1999). Since the divergence with the chimpanzee, recombination between *Alu* or L1 elements caused respectively 492 and 73 deletions events in the human genome, with a size ranging from 100pb to 64kb (Han et al., 2008; Sen et al., 2006). Similarly, at least 27% of segmental duplication and 44% of inversion events in human evolution can be confidently linked to unequal recombination between L1 or *Alu* elements (Bailey et al., 2003; Lee et al., 2008).

1.4.2 Advantageous and disadvantageous variations

L1, *Alu* and SVA elements have been directly involved in at least 50 genetic diseases in humans, including immunodeficiencies, vision abnormalities, renal anomalies, muscle and blood disorders (Kaer and Speek, 2013). TE-mediated mutations are also linked to cancer. Mutations in tumor-suppressor genes increase the probability to develop cancer: for example, several *Alu* insertions in the Breast Cancer 1 and 2 genes (*BRCA1-2)* were reported, and other insertions were linked to leukemia, skin and colon cancer (Chénais, 2013). Moreover, environment of a tumor cell itself might favor TE mobilization by disrupting mechanisms that normally prevent TE expression (Ross et al., 2010). Numerous *de novo* insertions of L1 elements were observed in different types of cancer, often in the proximity of cancer-related genes (Lee et al., 2012). Whether transposition in healthy somatic cells is the cause of cancers, or whether insertions are mere consequences of tumor development is an interesting question.



Introductory Figure 10. Contribution of of Transposable elements to genome evolution.
Transposable elements can contribute to the formation of new genes by:
A. Duplication of an existing gene by unequal recombination-mediated segmental duplication.
B. Creation of functional retrogenes by L1-mediated retrotransposition of genic mRNA.
C. Evolution of a transposon into a functional gene.
D. Transposable elements can also affect the expression of existing genes by altering splicing, transcription initiation and termination signals, or modifying short and long range regulatory interactions.
E. Example of the control of gene expression by the insertion of an IAP transposon upstream of the Agouti gene in mouse. Expression of IAP transposon affects the expression of the Agouti gene controlling coat color. Difference in the timing of IAP regulation during embryonic development leads to significant coat color variation in genetically identical mouse littermates. From Jirtle and Skinner, 2007.

In the mouse, MLV LTR-retrotransposons have long been shown to be able to induce leukemia (Jolicoeur et al., 1978)

Disease-causing mutations tend to disappear from the population, because of their deleterious effects. But TE-derived mutations had many positive consequences during evolution (**Introductory Figure 10**). The probability of evolving new protein-coding genes *de novo* by random association of amino acids is very low. Since François Jacob introduced the concept of "evolutionary tinkering", it is commonly considered that in the majority of cases, "The appearance of new molecular structures […] have rested on alteration of preexisting ones" (Jacob, 1977). There are numerous examples where TEs helped creating new genes, usually by duplicating existing ones. Around 40% of human or *Drosophila* genes are duplicated somewhere else in the genome (Zhang, 2003). Considering that unequal recombination between TEs is a major source of genomic duplication, unequal recombination probably account for the creation of many duplicated genes (**Introductory Figure 10A**).

New genes are also created from intronless retrogenes created by LINE *trans* mobilization (**Introductory Figure 10B**). Newly inserted retrogenes most often lack regulatory sequences and are inert, becoming therefore pseudogenes. However, some appear to have acquired new promoters or were transported with it. As a result, out of the 8,000 retrogenes identified in the mouse genome, around 1,000 are transcribed and a least 120 have evolved into *bona fide* new genes (Vinckenbosch et al., 2006; Zhang et al., 2003). It is estimated that at least one new functional retrogene per million year appeared during primate evolution (Marques et al., 2005). Finally, at least one gene family in primates has also been created by repeated 3'-transduction (Xing et al., 2006)

TEs did not only contribute to evolution by modifying the genome, but were also sometimes directly domesticated by their host to evolve new functions **Introductory Figure 10C** and Volff, 2006). The majority of coopted TEs were DNA transposons. The most famous case is the domestication of an ancient transposase 500Myrs ago to create the recombination-activating Rag1 protein (Kapitonov and Jurka, 2005; Thompson, 1995). Rag1 (and Rag2) are responsible for V(D)J recombination, the defining feature of the adaptive immune system of jawed vertebrates. Another vital mechanism that originates from the domestication of an ancient TE is the maintenance of chromosome telomeres. Telomerases are RNA-dependent DNA polymerases and probably evolved from a primitive LINE-like retrotransposon (Eickbush, 1997). In *Drosophila*, the ancestral telomerase was lost and telomere maintenance evolved a second time by using, again, two LINE retrotransposons (Levis et al., 1993). Many mammalian genes are also derived from the *gag* or *env* genes of LTR-retrotransposons (Volff,

2006). *Peg10* and *Rtl1,* two well-studied imprinted genes involved in placenta formation, derive from *gypsy*-like LTR retrotransposons (Youngson et al., 2005). *Syncytin* genes in human and mouse are also required for placenta formation and derive from the *env* genes of ERVs (Mi et al., 2000). Interestingly, the domestication of human and mouse *Syncytin* genes appears to have happened independently (Dupressoir et al., 2005).

In addition to creating new genes, TE insertions can also modify existing ones (**Introductory Figure 10D** and **10E**). TEs can introduce new coding sequences (Nekrutenko and Li, 2001), and modify splicing patterns (Kreahling and Graveley, 2004). TEs can also introduce new regulatory sequences. As examples, L1 and *Alu* polyadenylation signals were shown to cause premature mRNA truncation (Chen et al., 2009; Perepelitsa-Belancio and Deininger, 2003). The promoter sequence of TEs can also be used to drive expression of neighboring genes. In particular, solo-LTRs are often used as alternative promoters in mammals, and appear to play an important role during embryonic development (Cohen et al., 2009; Gifford et al., 2013).

1.4.3 Spatial distribution of TEs

TE distribution in the genome is not random. Some regions appear composed only of TEs, while other places are mostly depleted. For example, a 200kb segment in the human X chromosome is composed at 99% of TEs, while the four homeobox gene clusters contain less that 2% of TE sequences (Lander et al., 2001). As a general rule, TE-content is lower in gene-rich regions and higher elsewhere. *Drosophila* and *Arabidopsis* represent extreme cases. In their small genomes, TEs are highly compartmentalized. They are mostly absent in the vicinity of genes but strongly enriched in telomeres, centromeres and pericentric regions (Bartolome et al., 2002; Wright et al., 2003). In human as well, L1s appear to be particularly overrepresented in pericentric regions (Laurent et al., 1997; Schueler et al., 2001), an observation that was not however neither confirmed nor infirmed after sequencing of the human genome.

Estimating if TEs have preferential insertion bias for specific genomic contexts is difficult. Indeed, the current density of TEs probably does not reflect their initial integration sites, as selective pressure would act differently on gene-rich and gene-poor regions. It has long been observed that in human and mouse, L1s and ERVs are over-represented in AT-rich portions of the genome, while SINEs are preferentially in GC-rich regions (Smit, 1999; Soriano et al., 1983). This observation was puzzling because L1s and SINEs use the same protein machinery and should therefore not exhibit different preferences. It was in fact shown

that young SINEs have the same AT-rich insertion bias than LINEs, but that there is a strong (and unexplained) selective pressure to conserve older ones only in GC-rich portions (Lander et al., 2001; Waterston et al., 2002). The preference of L1 insertions for AT-rich regions is often explained by the requirement for L1 to insert at AATTTT hexanucleotides (Smit, 1999). However, gene-rich regions tend naturally to have a high GC-content, while gene-poor regions are conversely AT-rich. The observed preference for AT-rich regions could simply be the result of differential selective pressures.

The most unbiased method to estimate the real integration preferences of TEs is to rely on artificially integrated constructs. Several studies with human and mouse artificial L1 transgenes showed that L1 insertions tend in reality to be distributed randomly throughout the genome (Babushok et al., 2006; Gilbert et al., 2002; Symer et al., 2002). A modest preference for intergenic regions could only be noted for mouse L1s. In line with these observations, the most recent L1 insertions in the human genome are distributed randomly, and the observed preference for AT-rich, gene-poor regions is probably a selective bias (Ovchinnikov et al., 2001). Similarly, analysis of the integration sites of an artificially resurrected human ERV or of *bona fide* retroviruses showed that retroviral sequences insert preferentially into gene-rich regions, and especially inside genes (Brady et al., 2009; Mitchell et al., 2004). By contrast, remaining ERVs are mostly excluded from genes, and in general ancient ERVs are particularly underrepresented near genes compared to younger elements (Medstrand et al., 2002).

Gene-rich regions are often associated with a permissive chromatin environment that could make integration of new TEs easier. However, in the long term, the fixation and the maintenance of TEs is probably mainly controlled by natural selection. As a consequence, TEs are over-represented in gene-poor regions.

SINEs represent a very intriguing exception. They seem to integrate randomly but are strongly selected to remain in the proximity of genes. This suggests that they are beneficial for the organism (Lander et al., 2001), which fits with their incredible success in mammalian genomes. SINEs could potentially play a role during meiosis and contribute to synapsis formation between homologous chromosomes. During meiosis, sister chromatids adopt a complex structure: they align on a protein-made axis with chromatin extending out in loops. The protein axis of the two homologous chromosomes are brought together and joined via the zipper-like synaptomenal complex. The formation of this complex is critical for correct crossing-over, and recombination is thought to occur primarily within the loops and not in the DNA located near the axis (Keeney et al., 2014). Interestingly, it was suggested that SYCP3,

(one of the core protein of the synaptomenal complex) was preferentially binding SINE DNA in mouse (Johnson et al., 2013). SINEs could serve as a binding platform for the recruitment of the synaptomenal complex in gene-rich regions. Because recombination occurs preferentially in the chromatin loops that are not directly bound by the synaptomenal complex, it would also ensure that potentially harmful DNA breaks do not occur directly in genes, but only in their proximity. It was indeed shown in human that recombination "hotspots" were preferentially located near genes, but principally outside the transcribed domain (Myers et al., 2005). As synapsis formation is a genome-wide process, it would also explain why the localization of SINEs in gene-rich regions is such a global pattern. However, this hypothesis is highly speculative (Barau, 2015)[4] and the reason for the accumulation of SINEs near genes remains largely unknown.

1.4.4 TEs in the mouse genome[5]

The mouse genome is an interesting model to study TEs. It contains both ancestral repeats that predate the divergence with humans, and also very recently integrated ones. Rodent and primate lineages are estimated to have diverged between 65 and 100Myrs ago. The broad composition of human and mouse genomes is similar, but closer analysis reveals interesting differences (Waterston et al., 2002). The mouse genome (2.7Gb) is markedly smaller than the human one (3.2Gb), and is composed of at least 44% of TEs, with around 1Gb (85%) that is mouse-specific. The rate of spontaneous deletions and nucleotide substitutions is twice as high in mice than in humans. Fossilized repeats become then unrecognizable faster in mouse: they disappear after 100-120Myrs, compared to 150-200Myrs for humans. As a consequence, ancestral repeats comprise only 5% of the genome in mice, and 22% in humans. Mouse TEs are also much younger in average: 25% of mouse repeats have integrated in the last 25Myrs, while human TEs became almost extinct during that time. Whereas the majority of human TEs are now inert, the mouse genome contains thousands of active LINEs and ERVs.

L1s represent 20% of the mouse genome, with approximately 600,000 copies. Mouse L1s all derive from the evolution of a single ancestral lineage, the last common ancestor between humans and mice being L1MA6 (Smit et al., 1995; Waterston et al., 2002). Around 20,000 L1s are full-length and 3,000 copies are estimated to be retrotransposition-competent,

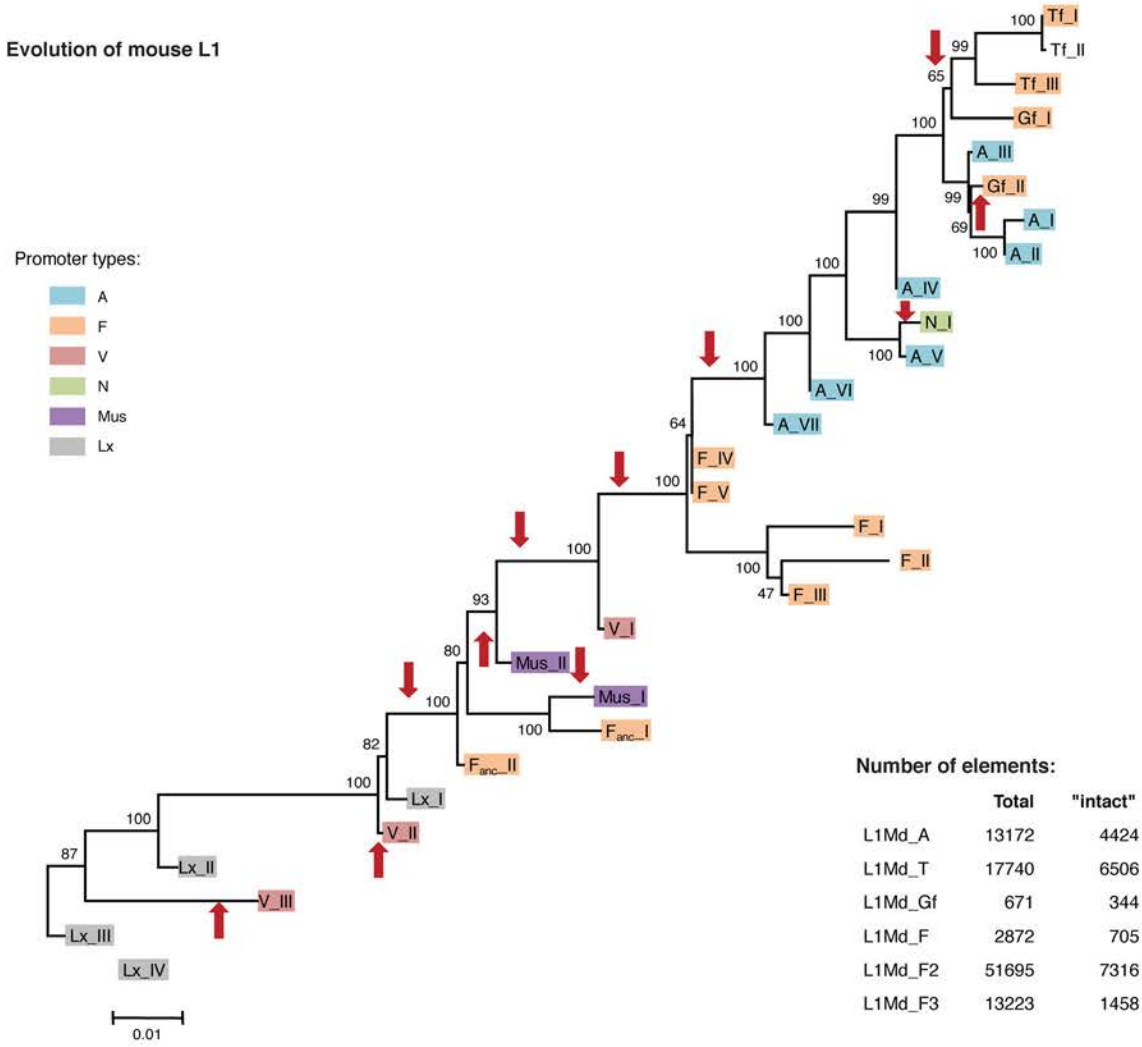
[4] And also very personal

[5] The names and the numbers given in this section are based on the last RepeatMasker annotation (Smit et al., 2013), using criteria to categorize full-length elements that will be developed in the result section of this thesis. They are in good concordance with numbers given in Stocking and Kozak, 2008.

namely 30 times more than in humans (Goodier et al., 2001). Since the divergence with the rat 13Myrs ago, the mouse L1 lineage has experienced 11 replacements of 5'UTR, and numerous other recombinations in the coding part (**Introductory Figure 11** and Sookdeo et al., 2013). Recent active families are classified in two groups based on their promoter type (A and F), while more older and extinct families have a V type promoter (Adey et al., 1994). The F promoter replaced a V promoter by resuscitating a more ancestral form 6.4Myrs ago. The A promoter further replaced this F-type 4.6Myrs ago. Respectively 2.2 and 1.2Myrs ago, A-elements exchanged again their promoter and gave birth to Tf and Gf elements (both F-type promoters) (Sookdeo et al., 2013). A-, Gf- and Tf-type account respectively for 900, 400 and 1,800 retrotransposition-competent elements (Goodier et al., 2001).

Whereas only one SINE lineage is active in the human lineage (*Alu),* mice have four distinct SINE families (B1, B2, ID, B4). There is more SINEs in the mouse than in the human genome (1.4 and 1.1 millions, respectively), but mouse SINEs occupy less genomic space (8 and 13%, respectively) because of their smaller size (Waterston et al., 2002). B1 and *Alu* element*s* are both derived from a 7SL RNA and have probably a common origin (Quentin, 1994). On the other hand, B2 and ID are related to tRNAs. ID closely matches an Ala-tRNA and is present with relatively few copies in the mouse genome, compared to its very successful amplification in the rat genome (Gibbs et al., 2004). B4 elements represent an interesting case of fusion between ID and B1 (Kramerov and Vassetzky, 2001). Both B4 and ID elements are inactive in the mouse genome, whereas B1 and B2 show continuous signs of multiplication (Gibbs et al., 2004).

Bona fide LTR-retrotransposons of *gypsy* or *copia* superfamilies cannot be found in the mouse, but ERVs compose 12% of the genome. As a consequence, the two terms "LTR-retrotransposons" and "ERVs" are often used as synonyms. ERVs are usually further classified into three classes depending on the type of retrovirus they originate from **Introductory Figure 12A**). Class-I relates to gamma- and epsilonretroviruses, class-II to lentiviruses, alpha-, beta- and deltaretroviruses, while class-III is closer to spumaviruses (Gifford et al., 2005; Jern et al., 2005). Class-III ERVs are often referred as class-L ERVs, because in both mouse and human retrotransposition is initiated with a leucine tRNA (Bénit et al., 1997; Cordonnier et al., 1995). Similarly, class-II is also often referred as class-K, because the first identified elements (MMTV in mouse, HERV-K10 in human) use lysine tRNAs as primers (Ono et al., 1986; Peters and Glover, 1980). However, this classification is not correct since IAP, one of the most studied class-II family, appears to use a phenylalanine tRNA (Rowe et al., 2010). In general, the nomenclature of ERVs is incredibly confusing. The



	Total	"intact"
L1Md_A	13172	4424
L1Md_T	17740	6506
L1Md_Gf	671	344
L1Md_F	2872	705
L1Md_F2	51695	7316
L1Md_F3	13223	1458

Introductory Figure 11. Phylogenetic tree of mouse L1 families based on the longest non-recombining region of ORF2, including the reverse transcriptase domain. Red arrows indicate the acquisition of a new 5'UTR. From Sookdeo et al., 2013.
The number for total and intact elements are taken from Repeatmasker, which use a nomenclature slightly different from the one proposed in Sookdeo et al., 2013.

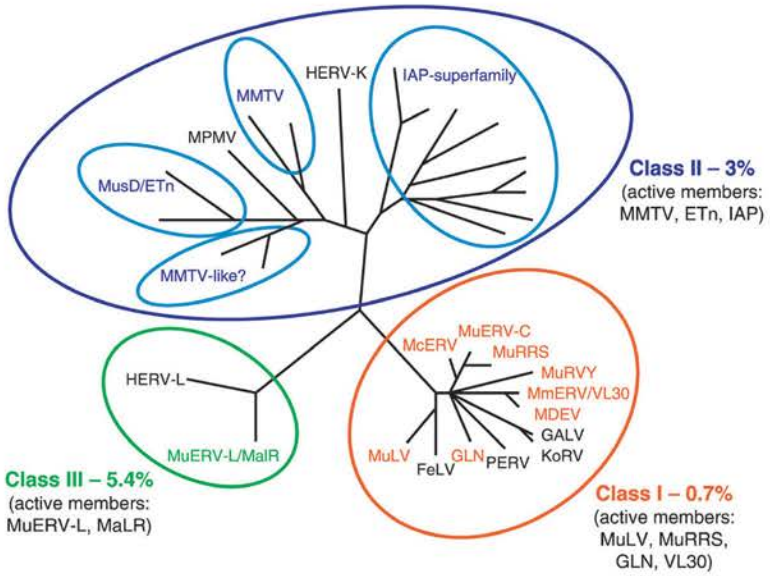
main families were discovered a long time ago and their current names (when they have one) often reflect their initial description, but in the form of now meaningless initials (IAP, MLV, MMTV, VL30, ETn, GLN, etc.). It is also difficult to have a clear estimation of the total number of families because this varies depending on the criteria used to classify them. Some studies proposed a limited number (around 20) (McCarthy et al., 2004), but the Repbase depository separates LTRs from internal sequences and currently annotate 484 different families (Jurka et al., 2005). Unfortunately, no coherent classification managed yet to replace this unintelligible imbroglio. In fact, because they were changing the name of known transposons, most of the attempts only added confusion (Baillie et al., 2004; McCarthy et al., 2004).

Class-I ERVs represent only 1.2% of the mouse genome, with a lot of different families composed of relatively few full-length copies (between 10 and 300 per genome). The best characterized is the MLV family (Mouse Leukemia Virus). MLVs entered the mouse genome recently (<1.5Myr ago) and intact elements remain potent retroviruses capable of infecting other cells. MLV particles has long been shown to be able to promote cancer (Friend, 1957). MLVs are in the process of endogenization and the number of element per genome varies between 10 and <100 depending on the mouse species (Stocking and Kozak, 2008). RLTR6 (also named MmERV) and VL30 are two other interesting class-I elements. They share the same LTRs, but RLTR6 is autonomous with predominantly intact *gag*, *env* and *pol* sequences, whereas VL30 lacks coding regions (Bromham et al., 2001; French and Norton, 1997). VL30 are likely derived from RLTR6 elements that lost their coding region, and continue to use *in trans* their protein machinery for their replication.

Class-II ERVs are very numerous and represent 4.9% of the mouse genome. The first discovered ERV-II was linked to breast cancer (Mouse Mammary Tumor Virus: MMTV) and was initially thought to be a *bona fide* retrovirus with a strong insertion preference for the same genomic region (Nusse and Varmus, 1982). However, there is only 2-3 full-length MMTV inserts in the mouse genome. Two related families are known as MusD[6] and ETn (Early Transposon), with around 300 intact copies each. ETns are very active and was first detected in embryonic carcinoma (Brulet et al., 1983). ETn elements are however internally recombined and lack coding sequences. They rely on retrotransposition-competent MusD proteins for their activity (Baust et al., 2003; Mager and Freeman, 2000). The most studied mouse ERVs are related to IAP (Intracisternal A-type Particles) that were initially observed in the endoplasmic reticulum of various cancer cells (**Introductory Figure 12B** and Dalton et

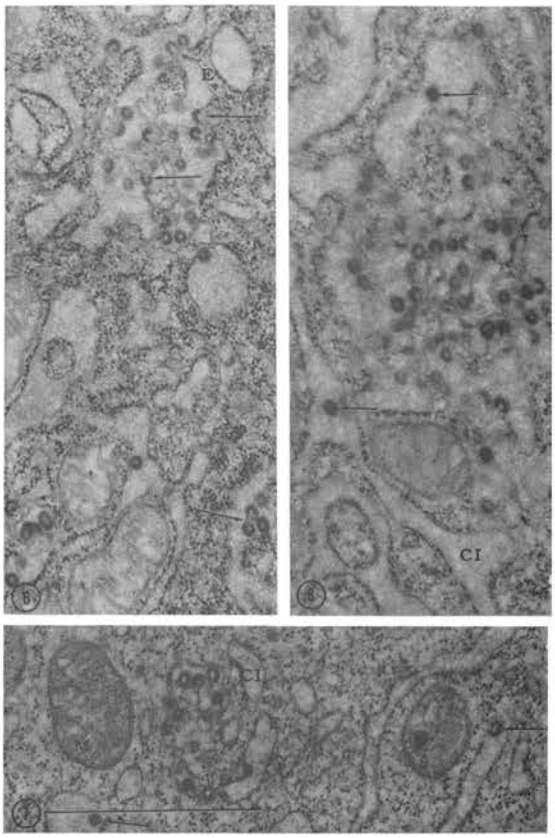
[6] In RepeatMasker nomenclature, MusD is named ETnERV or ERVB7_1 (sic)

A



	Total	Intact
MERV1 (McERV)	568	80
MLV	83	15
GLN	300	103
RLTR6 (MmERV)	516	242
vl30	343	113
MuRRS	875	267
MuRVY	319	252
IAPEz	2958	1936
IAPEy	2098	487
MMERVK10C	1077	312
RLTR10	1465	780
Etn/MusD	2541	549
MMTV	13	3
MERVL	3444	834
MalR-MT	10643	4380
MalR-ORR1	12179	3732

B



Introductory Figure 12. mouse Endogeneous retroviruses

A. RT sequences of ERVs from host species other than mouse are included for comparison and are in black letters. Four distinct clades or superfamilies are de- fined for the Class II ERVs, one of which (MMTV-like) is poorly characterized. Non-autonomous elements, such as the abundant VL30 s (Class I), ETns (Class II), and MaLRs (Class III) are listed with their presumed parental ERVs, as they do not contain RT domains. From Stocking and Kozack, 2008, numbers from this study.

B. Initial observation of IAP as virus-like particle in plasma-cell tumors. From Dalton et al., 1961.

al., 1961; De Harven and Friend, 1958). Eight families can be distinguished (including for example MMERVK10C or the severely internally deleted RLTR10) and they represent together around 2,000 intact copies (McCarthy et al., 2004). ETn and IAP elements are the most active transposons of the mouse genome and account for the vast majority of *de novo* insertions (Maksakova et al., 2006).

As in most mammals, Class-III are the most abundant ERVs in the mouse genome (5.9% of the total). They are also the most ancient, and they account for 80% of the ERVs predating the human-mouse speciation (Waterston et al., 2002). Probably for this reason, Class-III elements, and especially solo-LTRs, appear to have been coopted by mammalian evolution and they play an important role during gametogenesis and embryonic development (Gifford et al., 2013). Class-III is composed of only two types of elements: the retrotransposition-competent MERVLs and the non-autonomous MalRs (Mammalian apparent LTR-retrotransposons) (Bénit et al., 1997; Smit, 1993). MERVLs and MalRs share similar LTR sequences and probably have a common ancestor (McCarthy et al., 2004). MalRs are internally deleted and contain only non-coding repetitive DNA. The two main types, MT and ORR1, have an average length of 2 and 2.4kb, respectively. MalRs, and especially MT elements, have been incredibly successful, as they represent the vast majority of class-III elements and outnumber MERVLs by a 10 to 1 ratio.

Further reading:

The following reviews were extensively used as starting material to write this chapter.

Biemont, 2010: history of TEs discovery;
Feschotte and Pritham, 2007: DNA transposons;
Kramerov and Vassetzky, 2005: SINEs.
Babushok et al., 2007: LINE1;
Cordaux and Batzer, 2009: genome evolution;
Le Rouzic and Capy, 2009: transposon population genetics;
Silva et al., 2004: horizontal transfer;
Stocking and Kozak, 2008: murine LTR-retrotransposons.

The following studies are cited in figure legends:
Dalton et al., 1961; Fedoroff, 2012; Feschotte et al., 2002; Fleischmann et al., 2014; Gilbert et al., 2010; Gregory, 2015; Gregory et al., 2009; Griffiths et al., 2008; Hedges et al., 2015; Jirtle and Skinner, 2007; Kelley et al., 2014; Maddison et al., 2007; Pace et al., 2008; Pellicer et al., 2010; Smit et al., 2013; Sookdeo et al., 2013; Stocking and Kozak, 2008

2 TRANSCRIPTIONAL CONTROL OF TRANSPOSONS

Eukaryote cells have developed an important range of mechanisms that control the expression and mobility of TEs. Similar pathways are used in plants, fungi and animals and probably have a very ancient origin. Control of TEs is especially important for autonomous elements that retained the coding potential to mobilize themselves. However, mutated TEs can still perturb normal genome function by interfering with the transcription of neighboring genes or by modifying regulatory patterns. Considering that mutated TEs represent such an important proportion of the genome, tight control of non-autonomous elements is probably critical as well. Potentially for this reason, the most conserved silencing mechanisms do not target transposon proteins but act preferentially at the transcriptional and post-transcriptional levels. Transcriptional control ensures that transposon mRNAs are not produced and relies on a diverse range of chromatin modifications that alter the packing and condensation of DNA. On the other hand, post-transcriptional control principally acts by degrading transcribed mRNAs by a mechanism known as RNA interference.

RNA interference is widely used for transposon control in plants, fungi, insects or nematodes, and transcriptional and post-transcriptional mechanisms are often coupled together (Slotkin and Martienssen, 2007). However in mammals, at the notable exception of the male and female germlines, the bulk of transposon control is thought to rely on chromatin modifications. For this reason[7], the following sections of this work are centered on the mechanisms that ensure transcriptional silencing. How different pathways cooperate to specifically control TEs will be analyzed in details, but it is first necessary to review in depth how transcriptional silencing is achieved in general in eukaryotes, and in mammalian genomes in particular.

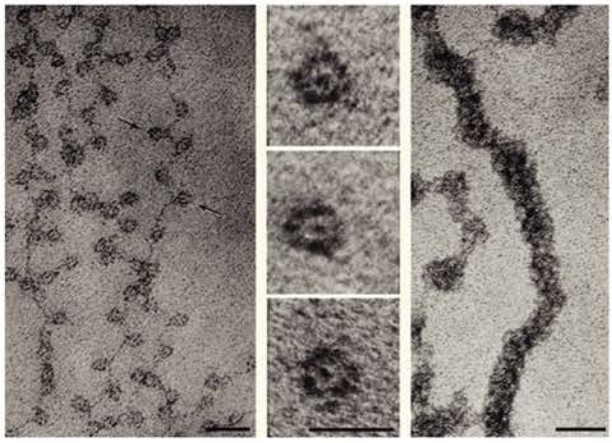
2.1 Chromatin

In eukaryotes, DNA does not float freely inside the nucleus but is tightly encapsulated with proteins to form highly organized macromolecules. The resulting DNA-proteins macro-complex is called chromatin and allows the compaction of a two meters-long DNA molecule (in human) into a nucleus of only 6µm in diameter.

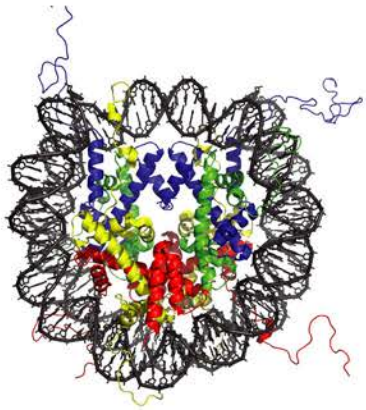
[7] And also because of the lack of time and space…

A

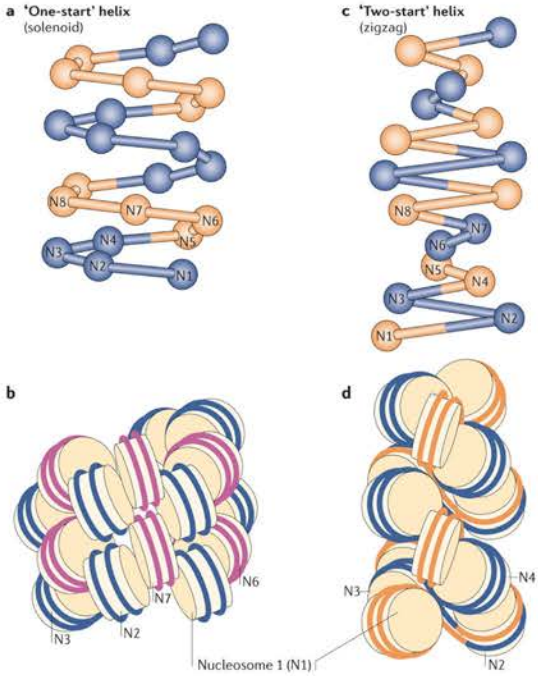
C



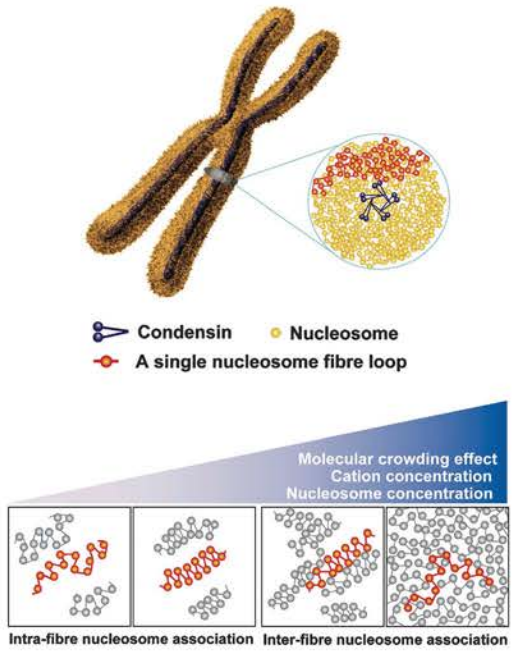
B



D



E



Introductory Figure 13. Chromatin core structure.

A. Histological section showing heterochromatc (dark) and euchromatic (light) regions of a a mammalian nucleus.

B. Crystal structure of the nucleosome core particle consisting of H2A , H2B , H3 and H4 core histones, and DNA. The view is from the top through the superhelical axis. From wikipedia.org.

C. Left: low ionic-strength chromatin spread; the 'beads on a string'. Size marker: 30 nm. Middle: Isolated mononucleosomes derived from nuclease-digested chromatin. Size marker: 10 nm. Right: Chromatin spread at a moderate ionic strength to maintain the 30-nm higher-order fiber. Size marker: 50 nm. From Olins and Olins, 2003.

D. Solenoid and zigzag model of 30nm chromatin fiber. From Luger et al., 2012.

E. Metaphase chromosome structure model. Chromosomes consist essentially of irregularly folded nucleosome (beads on a string) fibers. Condensins (blue) hold the nucleosome fibers (red) globally around the chromosome center. Locally, the nucleosome fiber is folded in an irregular or disordered manner, forming loop structures that are collapsed towards the chromosome center (blue). From Nishino et al., 2012.

2.1.1 Historical perspective

Walther Flemming first introduced the term chromatin in 1882 to refer to a fibrous scaffold in the cell nucleus that could be easily stained[8] (Flemming, 1882). A decade earlier, Friedrich Miescher had identified in the nucleus a strong phosphorus-rich acid (Miescher, 1871) and Albrecht Kossel proposed in 1884 that these "nucleic acids" might be bound to a novel class of proteins that he had termed histones (Kossel, 1884). Histones and DNA constitute indeed the core components of chromatin. For a long time, histones were even thought to be the carrier of genetic information, while DNA was viewed as a mere linker molecule (Schultz, 1941). This view completely shifted with the demonstration that DNA, and not proteins, could by itself transform a harmless strain of *Pneumococcus* bacteria into a very virulent one (Avery et al., 1944). The resolution of the double-helical structure of DNA then provided the molecular basis of genetic inheritance (Watson and Crick, 1953) that led to the enounciation of the central dogma of biology (Crick, 1956).

Chromatin was long established to exist in different forms. Chromosomes are at their most condensed form during the metaphase stage of mitosis, and chromosomes decondense during interphase. In 1928, Emil Heitz visualized chromosomal regions that do not undergo post-mitotic decondensation and termed these domains "heterochromatin" (Heitz, 1928). By opposition, "euchromatin" refers to chromosomal regions that decondense and spread diffusely in the interphase nucleus (**Introductory Figure 13A**). Heitz immediately proposed that the densely compacted heterochromatin could reflect a functionally inactive state of the genome (Heitz, 1929, 1932). This hypothesis was first validated in *Drosophila*, where the localization of the *white* gene next to an heterochromatic domain after an inversion was shown to produce stochastic changes in eye color (Schultz, 1936). This view is still prevalent today: heterochromatin is typically gene-poor and transcriptionally silent; whereas the less condensed euchromatin is gene-rich and transcriptionally permissive. Heterochromatin was further divided into constitutive and facultative heterochromatin to distinguish regions that are always condensed on both homologous chromosomes and in every cell types (like centromeres and telomeres), from regions that can alternate between different states in a development-specific manner (like the X chromosome in female) (Brown, 1966).

The molecular structure of chromatin in its various flavors has received incredible attention since then. While DNA is invariably composed of the same four nucleotides, the proteins it is packaged with come in a seemingly infinite diversity. In particular, histones can

[8] "chromo" stand for "color" in ancient Greek

harbor many different post-translational modifications and are further bound by numerous other proteins. Furthermore, DNA itself can be modified and methyl group are often added at cytosines. Both histone and DNA modifications directly influence chromatin condensation and gene expression. Following the work of Jacob and Monod, gene expression was initially believed to be regulated primarily by *trans*-acting factors that recognize specific DNA motifs and either block or activate transcription (Jacob and Monod, 1961). The discovery in yeast that histone themselves regulate gene expression (Kayne et al., 1988) was followed over the years by the description of apparently limitless combinations between various chromatin modifications. This development led to the "histone code" hypothesis[9], which states that the complex combination between various chromatin modifications "considerably extends the information potential of the genetic code" and "represents a fundamental regulatory mechanism that has an impact on most, if not all, chromatin-templated processes" (Jenuwein and Allis, 2001; Strahl and Allis, 2000).

Since 1988, Chromatin ImmunoPrecipitation (ChIP) has been widely used to analyze DNA-associated proteins and their modifications along chromosomes (Solomon et al., 1988). Over the last decade, tremendous energy has been spent to map the local structure of chromatin. These efforts, exemplified by the ENCODE project in mouse and human (Bernstein et al., 2012), allowed the linear segmentation of chromosomes into hundred of domains with different protein compositions. Moreover, the spatial organization of DNA in the nucleus is not random. Local and long-range interaction between sequences and the localization inside the nucleus appears highly controlled. For example in human or *Drosophila*, hundreds of domains are preferentially associated with the nuclear lamina, and genes inside these domains are transcriptional silent (Guelen et al., 2008; Pickersgill et al., 2006). The developments of chromosome conformation capture technology (3C, 4C, 5C and Hi-C) gave further insight into the global 3-dimensional structure of the genome and uncovered many unsuspected long range interactions between DNA sequences (Dekker et al., 2013). In *Drosophila*, mouse and human, chromosomes are composed of hundreds of discrete topological domains of typically 100kb-1Mb, inside which DNA preferentially interact (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012).

Hundreds of chromatin modifications have been identified and the potential combinations between them are infinite. However, most modifications preferentially associate together in a reduced set of combination, allowing the definition of only a small number of major chromatin types (Rando, 2012; van Steensel, 2011). One of the current challenges is to

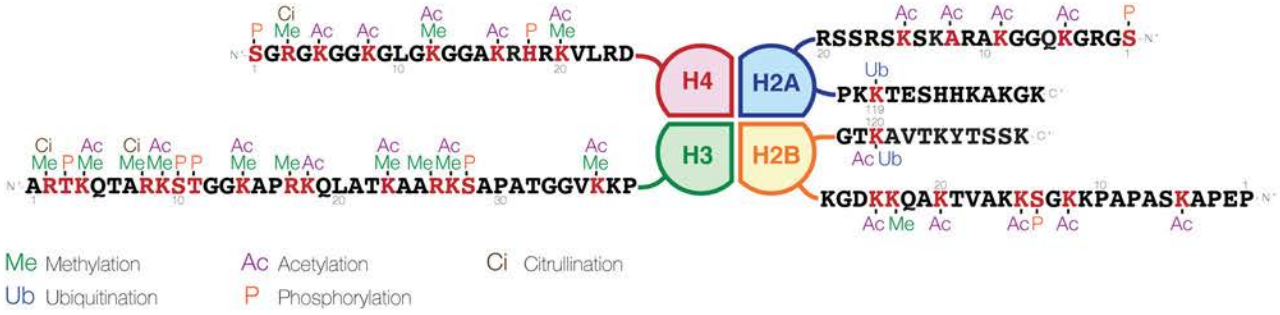
[9] that can be extended to incorporate DNA methylation and other non-histone proteins

integrate the enormous amount of linear and three-dimensional genomic data in order to build a global and compressive picture of chromatin organization (Bickmore and van Steensel, 2013). These new descriptions are going beyond the original heterochromatin/euchromatin dichotomy and allow for a better understanding of how local and large-scale structures affect gene regulation and other nuclear functions. It was for example proposed that chromatin complexity could be reduced to four "chromatin states" in *Arabidopsis* (Roudier et al., 2011), five "chromatin colors" in *Drosophila* (Filion et al., 2010) or six "chromatin compartments" in human (Rao et al., 2014).

2.1.2 Chromatin core structure and modifications

The first level of packaging is the nucleosome, which consists of 145-147pb of DNA wrapped 1.7 times around an octamer of core histones (Kornberg, 1974; Luger et al., 1997). Histone octamers are composed of two copies of the core histone proteins H2A, H2B, H3 and H4. Histones possess flexible tails that extend away from the disk-like structure of the nucleosome and that are involved in the interaction with other nucleosomes or various nuclear factors (**Introductory Figure 13B**). Nucleosomes are separated by short linker sequences of around 10-60bp that are often bound by the linker histone H1, with a global stoichiometry in mouse nucleus of 0.5 to 0.8 H1 per nucleosome (Woodcock et al., 2006). H1 localize at the entry and exit points of DNA and stabilize the interaction of an additional 20pb of DNA around the nucleosome, resulting in complete two-turn particle of ~166pb (Thoma et al., 1979). The repeated array of nucleosomes can be easily observed by electronic microscopy (Olins and Olins, 1974) and forms a "bead-on-a-string" primary structure known as the 10nm fiber (**Introductory Figure 13C**).

The packing of DNA into nucleosomes results in a reduction of the fiber length of about sevenfold. Chromatin can be further folded into a thicker secondary structure known as the 30nm fiber that was first observed by electron microscopy in salamander's red blood cells (**Introductory Figure 13C** and Gall, 1963). The 30nm fiber appears spontaneously *in vitro* (Finch and Klug, 1976) and relies on linker histone H1 for its stabilization (Robinson and Rhodes, 2006). Based on numerous *in vitro* studies, the 30nm fiber has long been thought to be the natural state of chromatin folding. The high density of the 30nm fiber makes, however, the analysis of its molecular conformation difficult. Two structures, the "zigzag" and "solenoid" models, are still in competition to explain how the repeated array of nucleosomes folds into the 30nm fiber (**Introductory Figure 13D** and Robinson et al., 2006; Schalch et al., 2005). The conformation of the 30nm fiber appears moreover to depend on many



Modification	histone	Residue	Known function
Acetylation	H2A	K5	Transcriptional activation
	H2B	K5, K12, K15, K20	Transcriptional activation
	H3	K4, K27,36	Transcriptional activation
		K9	transcriptional activation, Histone deposition
		K14, K23	transcriptional activation, Histone deposition, DNA repair
		K18	transcriptional activation, DNA repair, DNA replication
	H4	K5, K12	transcriptional activation, Histone deposition, DNA repair
		K8, K16	transcriptional activation, DNA repair
Methylation	H3	R2, K9, K27	Transcriptional repression
		R8	Transcriptional control ?
		R17, R26, K4, K79	Transcriptional activation
	H4	R3	Transcriptional control ?
		K20	Transcriptional repression
Phosphorilation	H2A	S1, T119	mitosis
	H2AX	S139	DNA repair
	H2B	S14	DNA repair, apoptosis
	H3	T3, S10, T11, S28	mitosis
Ubiquitination	H2A	K119	Transcriptional repression
	H2B	K120	Transcriptional activation, meiosis

Introductory Figure 14. Histone modifications. Known histone modification and putative functions. Adapted from Rodriguez-Paredes and Esteller, Nature, 2011, and Cell Signaling website.

different parameters like salt conditions, length of linker DNA, histone tails modifications, etc. (Hansen, 2002).

Chromatin can further fold into even denser tertiary structures, especially during metaphase. It was long assumed that the 30nm fiber was further twisted and coiled to form highly condensed domains. However, it was recently showed that 30nm fibers could not be observed in metaphase chromosomes and that high degree of compaction was achieved by a "fractal packaging" of irregularly arranged 10nm fibers (**Introductory Figure 13E** and Nishino et al., 2012). Even in interphase chromosome, the direct observation of *bona fide* 30nm fibers remains elusive (Tremethick, 2007) and some authors proposed that chromatin could in reality only consist of dynamic and disordered 10nm fibers (Maeshima et al., 2014). In fact, the 30nm fiber represents probably only one of many potential chromatin structures. *In vivo* chromatin is presumably not uniform and likely harbor different conformations depending on the context (transcriptional state, cell cycle stage, developmental stage, etc.).

Many "architectural" proteins have been described, including heterochromatin protein 1 (HP1), methyl-CpG-binding protein 2 (MeCP2), Polycomb group proteins and others. These proteins usually form complexes that bind to multiple nucleosomes and induce further chromatin compaction, sometime acting also as shields that block the access to the underlying DNA. Architectural proteins are critical to form and maintain high-order chromatin structures (McBryant et al., 2006).

Histone tails modifications are playing a very important role in chromatin compaction and have a direct role in transcriptional control. Histone tails represent 25-30% of the total mass of histones and provide an exposed surface for many protein-protein interactions. An extensive literature documents dozen of different post-translational modifications including acetylation, phosphorylation, methylation, ubiquitination and citrullination (**Introductory Figure 14**). Chromatin modifiers, *ie* proteins that add and remove ("writers") and/or interpret ("readers") histone modifications have also been described at length (Kouzarides, 2007). Histone acetylation is one of the most studied modification and correlate with open chromatin and active transcription (Grunstein, 1997). Acetylation neutralizes the positively charged lysine residue and thereby decreases the interaction of histone tails with the negatively charged DNA (Hong et al., 1993). This causes the nucleosomes to unfold (Norton et al., 1989) and facilitates the recruitment of the transcription machinery (Lee et al., 1993). Histone acetylation is primarily controlled by numerous histone acetyltransferase (HATs) and deacetylases (HDACs) (Marmorstein and Roth, 2001; Thiagalingam et al., 2003).

By contrast, lysine mono-, di- and tri-methylation do not modify the histone tail charge and can be linked either to active or inactive chromatin. In general, methylation of histone H3 lysine 4 (H3K4), H3K36 and H3K79 is associated with an active state, while H3K9, H3K27 and H4K20 correlate with repression and are prevalent in heterochromatin (Barski et al., 2007). Many lysine methyltransferases (KMTs) and lysine demethylases (KDMs) are described and a unified nomenclature based on the histone residue they target has been proposed (Allis et al., 2007). For example H3K9-KMTs are members of the KMT1 family, H3K4-KMTS of the KMT2 family, etc. At the notable exception of DOT1L[10]/KMT4 that deposit H3K79 methylation (Feng et al., 2002), histone lysine methyltransferases are all characterized by the presence of a SET domain that catalyzes the methylation reaction using S-Adenosyl methionine (SAM) as a methyl donor. SET stands for Su(var), Enhancer of zeste, and Tritorax, the names of the tree proteins in which it was first identified in *Drosophila* (Jenuwein et al., 1998).

The precise mechanisms that enable histone modifications to impact on chromatin compaction and gene regulation are rarely well understood. Some modifications, like in particular lysine acetylation, have a direct effect on chromatin compaction. The tail of histone H4 is known to mediate interaction between adjacent nucleosomes (Tse and Hansen, 1997) and the amino acids 16-25 of histone H4 are thought to interact with an acidic patch located on the surface of H2A-H2B of an adjacent nucleosome (Luger et al., 1997). This interaction has been shown to be required *in vitro* for the compaction into a 30nm fiber (Dorigo et al., 2003). Furthermore, it was shown that acetylation of H4K16, a well-studied mark associated with gene activation, was sufficient to prevent chromatin compaction (Shogren-Knaak et al., 2006) but that by contrast, tri-methylation of H4K20 resulted in a even more condensed chromatin (Lu et al., 2008). These examples show how histone modifications can directly modify the local chromatin structure and probably impact transcriptional regulation as a consequence.

Many chromatin modifications do not have a direct physical influence on chromatin structure but act indirectly, by attracting or preventing the binding of "histone readers". Over the years, numerous histone readers and specific protein interaction domains have been characterized. For example, methylation is recognized by PWWP, PHD or Chromo-like royal domains (Chromo, Tudor MBT), while bromodomains recognize acetylation (Kouzarides, 2007). The functions associated with histone reader domains are very diverse. Some proteins

[10] Because of the lack of significance it often represents, the origin and signification of many gene names will be most of the time deliberately ignored.

are directly linked to chromatin structure like architectural proteins, chromatin remodelers or chromatin modifiers, while others regulate various biological processes, including mRNA transcription and elongation, DNA repair, V(D)J recombination or DNA replication. An additional level of regulation is achieved by the use of histone variants. They are paralogs of the canonical histones and their integration can affect nucleosome stability and help to create functionally different environments. For example, H3 is replaced at centromeres by the variant CENP-A, or by H3.3 in actively transcribed genes, regulatory regions and in some cases in heterochromatin (Biterge and Schneider, 2014)

In plants, insects and mammals, two main type of repressive chromatin are identified and will be reviewed in details in the following sections. The first one is best characterized by H3K9 methylation, while the second is linked to H3K27 methylation. These two chromatin states are structurally different. They are accompanied by distinct histone modifications and rely on different protein complexes. H3K9 methylation mark essentially functional genomic structure such as telomeres, centromeres, or transposable elements, while H3K27me3 is primarily found in regulatory regions and is highly involved in gene regulation (Mikkelsen et al., 2007).

Methylation of cytosine nucleotides, usually simply referred as DNA methylation, is a really special chromatin modification, in the sense that it is the only one that actually has DNA as substrate. DNA methylation cannot be unambiguously linked to any chromatin compartment, as it is present throughout the genome in mammals. However, DNA methylation has a direct effect on transcriptional regulation and is involved in many critical biological processes.

2.2 DNA methylation

It has long been observed in a very diverse range of eukaryote species that a small subset of cytosine nucleotides carries an additional methyl group at the fifth position of the six-atoms aromatic ring (**Introductory Figure 15A** and Johnson and Coghill, 1925; Wyatt, 1950). In mammals, around 4-6% of cytosines are typically methylated in somatic cells. This proportion goes up to 25% in certain plant genomes, while other species like *C. elegans, D. menanogaster* (but not other insects such as ants and bees) and some fungi (*S. cerevisiae* and *S. pombe,* but not *N. crassa)* have no detectable levels (Zemach et al., 2010). In plants and other organisms, DNA methylation is found in three different sequence contexts: CG (or CpG), CHG or CHH (where H correspond to A,T or C). In mammals however, DNA methylation is almost

exclusively found in CpG dinucleotides, with the cytosines on both strand being usually methylated. The sequence symmetry of this context led very early to the groundbreaking hypothesis that DNA methylation could control transcriptional activity and serve as a form of non-coded cellular memory during DNA replication (Holliday and Pugh, 1975; Riggs, 1975). Indeed, DNA methylation has been shown to be stable trough multiple divisions or even from parents to offspring, and is functionally involved in many important processes such as genomic imprinting, X inactivation or silencing of transposable elements (Jones, 2012). In plants, "epialleles" carrying different transcriptional outcomes depending on their methylation status can be faithfully transmitted from one generation to the next. However, cases of transgenerationnal inheritance of non-coded information in mammals appear (at best) exceptional, especially because DNA methylation and other chromatin modifications are extensively reprogrammed during embryonic development (Heard and Martienssen, 2014). The question of these reprogramming events will be extensively addressed later, and the following paragraphs focus principally on the function of DNA methylation in steady-state cells.

The DNA methylation landscape of vertebrates is very particular compared to other organisms (**Introductory Figure 15B**). In vertebrate, around 80% of CpG are methylated in somatic cells and DNA methylation appears as a default state that has to be specifically excluded from defined locations (Lister et al., 2009; Stadler et al., 2011). By contrast, the genome of most plants, invertebrates, fungi or protists show "mosaic" methylation patterns, where only specific genomic elements are targeted, and they are characterized by the alternation of methylated and unmethylated domains (Suzuki and Bird, 2008; Zemach et al., 2010). Some mammalian cell types such as embryonic stem (ES) cells or neurons contain also a relatively high level of non-CpG methylation, but whether methylation in this context has a function remains unclear (Lister et al., 2009, 2013)

High CpG methylation in mammalian genomes has an evolutionary cost because it increases the frequency of spontaneous mutations. Loss of amino-groups occurs with a high frequency for cytosines, with different consequences depending on their methylation. Deamination of unmethylated cytosines results in uracils that are efficiently recognized and removed by base-excision repair mechanisms. Deamination of methylated cytosines on the other hand results in thymines, a proper genomic base that is not as efficiently repaired: this results in frequent C to T transitions over evolutionary time (Bird, 1980; Walsh and Xu, 2006). As a consequence, CpGs are globally depleted in mammalian genomes, occurring at only 20% of their excepted frequency (Lander et al., 2001; Swartz et al., 1962). The only

exception for this global CpG depletion resides in a specific category of GC- and CpG-rich sequences termed CpG islands that are generally unmethylated and therefore retained the expected CpG content (Bird, 1986; Bird et al., 1985).

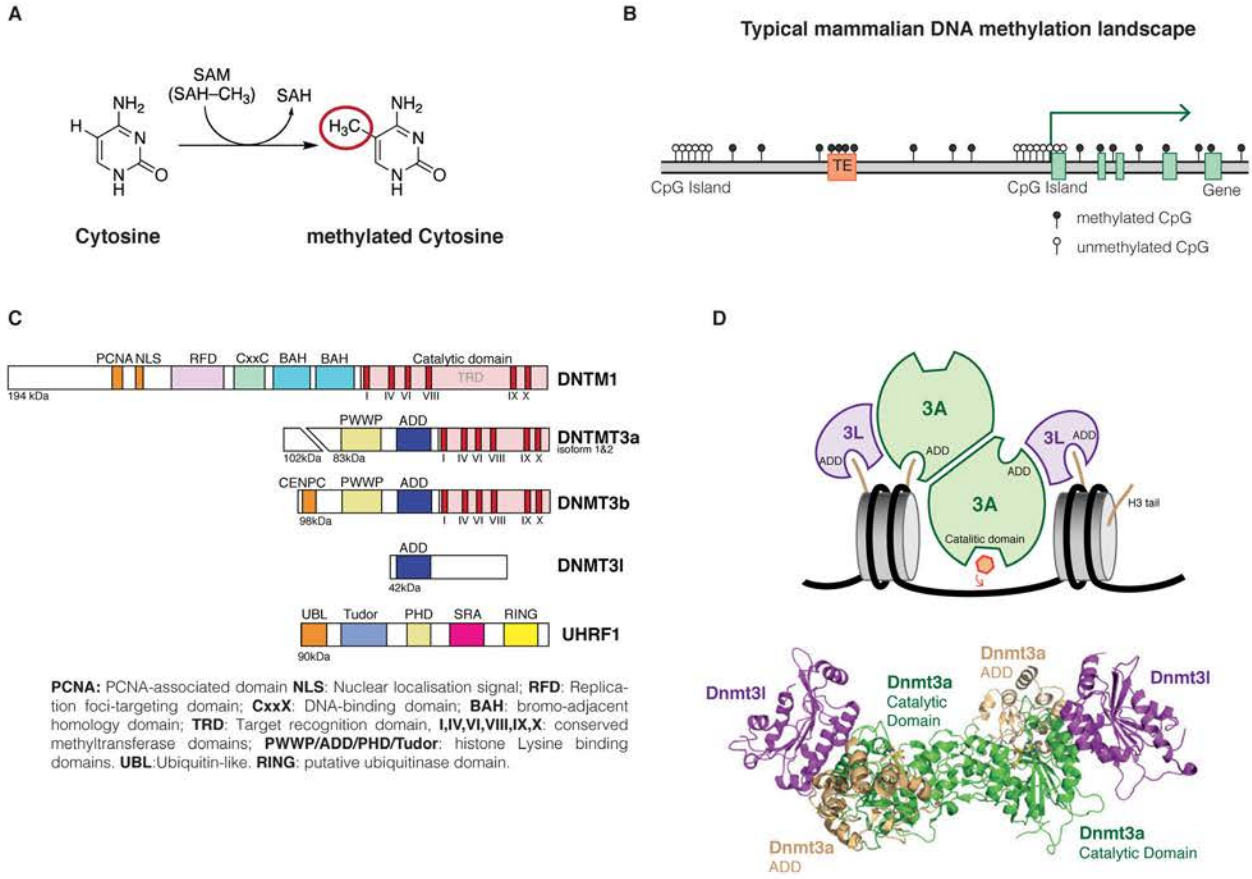
CpG islands are usually defined as regions with *1)* a length greater than 200pb, *2)* a G+C content greater than 50%, *3)* a ratio of observed to expected CpG greater than 0.6[11], although other definitions are sometimes used (Gardiner-Garden and Frommer, 1987). Excluding repeated sequences, there are around 25,000 CpG islands in the human genome, 75% of which being less than 850pb long (Lander et al., 2001). They are major regulatory units and around 50% of CpG islands are located in gene promoter regions, while another 25% lie in gene bodies, often serving as alternative promoters. Reciprocally, around 60-70% of genes have a CpG island in their promoter region (Illingworth et al., 2010; Saxonov et al., 2006). The majority of CpG islands are constitutively unmethylated and enriched for permissive chromatin modification such as H3K4 methylation. In somatic tissues, only 10% of CpG islands are methylated, the majority of them being located in intergenic and intragenic regions, and this number drop to 1% in the male germline (Smallwood et al., 2011).

Of note, ES cells are derived from the inner cell mass of the blastocyst, one of the only cell type during mouse development that is naturally depleted in DNA methylation. However, when cultured classically in presence of fetal bovine serum, mouse ES cells have a DNA methylation landscape similar to somatic cells: ~75% of CpGs are methylated and only CpG islands are protected (Stadler et al., 2011). Human ES cells are similarly highly methylated (Lister et al., 2009). Mouse ES cells can be also cultured without serum in presence of specific inhibitors. In these conditions, the DNA methylation landscape is closer to the blastocyst and cells are globally hypomethylated (Leitch et al., 2013). In almost all the studies referred to in this section, mouse ES cells were classically grown in presence of serum, and except if especially specified, it can therefore be considered that they are globally highly methylated.

2.2.1 DNA methyltransferases

DNA methylation is deposited by DNA (cytosine-5)-methyltransferase (DNMT) enzymes, which catalyze the transfer of a methyl group to a cytosine nucleotide from a SAM donor (**Introductory Figure 15C**). The first eukaryote DNMT to be identified and cloned was the mammalian DNMT1, which is principally involved in the maintenance of DNA methylation during cell divisions (Bestor and Ingram, 1983; Bestor et al., 1988). DNMT1

[11] Observed/Expected ratio= (number of CpG * length of the sequence)/(number of C * number of G)



Introductory Figure 15. DNA methylation and DNA methyltransferase.

A. Cytosine and methyl-Cytosine

B. Typical mammalian DNA methylation landscape. Most of the genome is CpG poor and highly-methylated, except for CpG-rich islands that are usually unmethylated.

C. Domain architecture of mammalian DNA methyltransferases and the cofactor, UHRF1.

D. Top: Model showing DNMT3l–DNMT3a–DNMT3a–DNMT3LI tetramer binds to unmodified H3 tail and then catalyze cytosine DNA methylation. Bottom: Structure of the DNMT3l–DNMT3a–DNMT3a–DNMT3l tetramer catalytic domains including the ADD (ATRX–DNMT3a–DNMT3l) domain of DNMT3a (in light brown). Adapted from Du et al., 2015.

contains a replication foci-targeting domain (RFD), a DNA-binding CxxC domain, two Bromo-adjacent homology domain (BAH) and a catalytic domain composed of core catalytic motifs and of a target recognition domain that recognizes the sequence to be methylated (Song et al., 2011; Takeshita et al., 2011). The overall structure is conserved with the plant homolog MET1, and the catalytic domain of both enzymes is also shared with prokaryotic DNMTs (Finnegan and Dennis, 1993). During mitosis, symmetrically methylated CpGs become hemi-methylated and full methylation must be reestablished in both daughter cells. DNMT1 is constitutively expressed in dividing cells and is most abundant during S phase. It is recruited to the replication fork by interacting *via* its RFD domain with the proteins PCNA and UHRF1. The latter binds specifically to hemimethylated CpG through its SRA domain and orient DNMT1 to the newly synthetized, unmethylated DNA molecule (Bostick et al., 2007; Chuang, 1997; Sharif et al., 2007). In addition, DNMT1 CxxC domain binds to unmethylated CpG and prevent the catalytic domain to access DNA (Song et al., 2011). Both UHRF1-mediated targeting to hemimethylated CpG and CxxC-mediated exclusion from unmethylated CpG ensure the faithful copy of DNA methylation patterns during mitosis. This maintenance activity is crucial in differentiated cells, and both UHRF1 and DNMT1-knockout mice show embryonic lethality around mid-gestation (Bostick et al., 2007; Li et al., 1992).

During embryonic development, DNA methylation is globally erased; it reaches a minimum at the blastocyst stage and in primordial germ cells and is reestablished around the time of implantation and during germ cell development (Kafri et al., 1992; Monk et al., 1987). Moreover, ES cells lacking DNMT1 still harbor significant DNA methylation and can *de novo* methylate foreign DNA sequences (Lei et al., 1996). These observations indicated that additional DNMTs were present in the genome, and led to the identification of the two *de novo* methyltransferases DNMT3a and DNMT3b (Okano et al., 1998). A third protein, DNM3l (DNMT3-like) was the last to be identified. It shares strong homology with DNMT3a and DNMT3b, but is severely truncated and lacks catalytic activity (Aapola et al., 2000). DNMT3 protein expression peaks during periods of *de novo* methylation in the peri-implantation embryo and in germ cells, and their expression become undetectable in most adult tissues. They are also highly expressed in ES cells, but downregulated after differentiation (Okano et al., 1998, 1999). *Dnmt3a*-knockout mice survive until birth but die in the next four weeks, while *Dnmt3b* knockout is lethal around mid-gestation. By contrast, mice depleted for *Dnmt3a* in germ cells specifically are sterile, whereas *Dnmt3b* knockout has not effect (Kaneda et al.,

2004). *Dnmt3l*-knockout mice develop to term and have no obvious phenotype, but males and females are infertile (Bourc'his et al., 2001).

DNMT3a and DNMT3b have non-redundant functions but share the same protein structure (**Introductory Figure 15C**). They contain a PWWP motif, a PHD-related domain termed ADD (Atrx-Dnmt3-Dnmt3l) and a catalytic domain similar to DNMT1. By contrast, DNMT3l lacks the PWWP domain and its catalytic sites are mutated, but it conserves a functional ADD domain (Chédin, 2011). DNMT3l does not bind DNA but its pseudo-catalytic site physically interacts with the catalytic domain of DNMT3a and DNMT3b to stimulate their activity (**Introductory Figure 15D** and Chen et al., 2005; Suetake et al., 2004). Crystallographic studies show that DNMT3l and DNMT3a form tetramers composed of DNMT3a and DNMT3l dimers, whereas in absence of DNMT3l, DNMT3a form long oligomers with reduced activity and progressivity (Holz-Schietinger and Reich, 2010; Jia et al., 2007). DNMT3a and DNMT3b preferentially methylate CpGs located in linker DNA between nucleosomes (Morselli et al., 2015; Takeshima et al., 2008), and their activity leads to the establishment of strand-specific patterns, with only one DNA strand being methylated (Lin et al., 2002). Furthermore, both proteins show significant (but modest) intrinsic sequence preference. For example, DNMT3a and DNMT3b appears to favor TNCGNC and TCGG sites, and disfavor ANCGN and NCGC sites[12], respectively (Wienholz et al., 2010). Interestingly, CpG dinucleotides inside CpG islands tend to be depleted for flanking sequences favored by DNMT3b, but enriched for disfavored sequences; this suggests that DNMT3b flanking site preference could have shaped the evolution of CpG islands by favoring sequences naturally resistant to *de novo* methylation (Wienholz et al., 2010). The presence of DNMT3l does not alter the sequence preference of DNMT3a and DNMT3b, but seems to increase the residence time of these enzymes on DNA, therefore increasing the chance to methylate anyway disfavored sites, and resulting in more uniform DNA methylation patterns.

PWWP and ADD domains regulate the interaction of Dmnt3 proteins with chromatin. The PWWP domain of both DNMT3a and DNMT3b binds to H3K36me3, a mark strongly associated with transcription elongation (Baubec et al., 2015; Dhayalan et al., 2010). Accordingly, it was shown that upon artificial expression in yeast, DNMT3b is recruited to H3K36me3-enriched chromatin (Morselli et al., 2015). By contrast, the ADD domain of DNMT3a, DNMT3b and DNMT3l binds to histone H3 tails, and these interactions are blocked by H3K4 methylation (Ooi et al., 2007; Otani et al., 2009). As a consequence, the *de*

[12] N represents any bases

novo machinery is excluded from regions marked by H3K4me3, such as promoters and enhancers. The anti-correlation between H3K4 and DNA methylation is observed genome-wide, in particular in the context of CpG islands (Weber et al., 2007). Finally, DNMT3a and DNMT3b appear to be specifically recruited to centromeric and pericentric regions. DNMT3a and DNMT3b physically interact, probably *via* their N-terminal domains, with CENP-C and Mis28α, two constitutive centromere proteins (Gopalakrishnan et al., 2009; Kim et al., 2012). The PWWP domain was also shown to be necessary for recruiting DNMT3a/b to pericentric heterochromatin, although the mechanism by which DNMT3a/b are targeted there is unclear since the PWWP domain does not bind efficiently to H3K9 and H4K20 methylation (Chen et al., 2004; Dhayalan et al., 2010).

While vertebrate genomes contain mainly CpG methylation in somatic cells, non-CpG methylation is widespread in other organisms, especially in plants. MET1, the plant homolog of Dnmt1, shares the same structure and maintains CpG methylation, while CRM2 is a plant-specific chromomethylase that especially maintains CHG methylation. By contrast with its mammalian homologs, the plant *de novo* methyltransferase DRM2 methylates cytosines in any sequence context (Law and Jacobsen, 2010). Mammalian and plant genomes contain another conserved DNMT-related enzyme, DNMT2. Despite its close analogy with authentic DNA methyltransferases, DNMT2 was in fact shown to be a tRNA methyltransferase with little affinity for DNA (Goll and Bestor, 2005). Sequence analysis of a wide range of eukaryotic genomes further revealed the existence of a total of six families of DNMT enzymes, only three of which exist in land plants and animals (Huff and Zilberman, 2014; Ponger and Li, 2005). The DNMT,5 and 6 families are poorly characterized, but have a very ancient origin. Dnmt5 was, for example, shown to be present in very diverse unicellular organisms (including diatoms, green algae or fungi) and to mediate CpG methylation maintenance of nucleosomal linker DNA.

Whereas DNMTs methylate DNA, there are at least two mechanisms by which DNA can be demethylated. Exclusion of DNMT1 from the replication fork leads to passive demethylation by dilution of the methylation mark over cell divisions. Active demethylation is mediated by the TET1/2/3 enzymes of the ten-eleven translocation (TET) family, which can oxidize methylated cytosines to convert them into hydroxymethylated cytosines (Tahiliani et al., 2009). Hydroxymethylated cytosines are then passively or actively reverted to unmethylated cytosines: *1)* UHRF1 correctly recognizes *in vitro* hemi-hydroxymethylated cytosines, but Dnmt1 appears to have very low affinity for this substrate (Frauer et al., 2011; Hashimoto et al., 2012). Hydroxymethylated cytosines would then be passively reverted to

unmethylated cytosines during DNA replication; *2)* TET enzymes catalyze further oxidation to formylcytosine and carboxylcytosine, which are specifically recognized, excised and replaced by unmethylated cytosines through base-excision repair mechanisms (He et al., 2011). *In vivo* observations indicate that both passive and active reversion of hydroxy-methylation probably occur during embryonic development (Kohli and Zhang, 2013).

2.2.2 DNA methylation represses transcription of CpG-dense promoters

DNA methylation was probably present at some extent in very early eukaryote ancestors. In virtually every organism analyzed, methylation in promoter regions correlates negatively with gene expression (Feng et al., 2010; Zemach et al., 2010). CpG-dense promoters of actively transcribed genes are never methylated, but reciprocally transcriptionally silent genes do not necessarily carry a methylated promoter. In mouse and human, around 60-70% of genes have a CpG island in their promoter region and most of these CpG islands remain unmethylated independently of the transcriptional activity of the gene, in both differentiated and undifferentiated cell types (Mohn et al., 2008; Weber et al., 2007). Of note, CpG-poor promoters appear highly methylated in somatic cells, which does not influence their activity, suggesting that DNA methylation can only repress transcription when found in a CpG-dense context.

Whereas DNA methylation is not necessary *per se* for transcriptional silencing, it is thought nonetheless to represent a "locked" state that definitely inactivates transcription. In particular, DNA methylation appears critical for the maintenance of mono-allelic silencing in the context of genomic imprinting and X chromosome inactivation (Beard et al., 1995; Li et al., 1993). In these cases, expressed and silent alleles differ by their methylation status, and loss of DNA methylation results in loss of imprinting and re-expression of *Xist* in somatic cells, indicating that DNA methylation is instrumental to maintain efficient silencing. During embryonic development, few genes change their methylation status, at the important exception of many genes specifically expressed in the germline (Borgel et al., 2010). Interestingly, germline genes are also among the few genes that have methylated promoters in ES cells, and loss of DNA methylation leads to their reactivation, both in ES cells and in the early embryo (Karimi et al., 2011b; Maatouk et al., 2006). These examples show that even if DNA methylation-mediated repression is limited in amplitude, it is nonetheless absolutely required for key biological processes. Of note, whereas DNA methylation of CpG islands is unambiguously linked with transcriptional repression, the function of DNA methylation at

CG-poor promoterss remains unclear; albeit there is little evidence that it could be functionally relevant (Schübeler, 2015).

DNA methylation appears absolutely required in differentiated cells, as knockout of any of the three competent DNMT results in embryonic or *post-partum* lethality. By contrast, DNA methylation is dispensable in undifferentiated cell types. The inner cell mass of the blastocyst and primordial germ cells are naturally hypomethylated and ES cells lacking one or several DNMTs can be isolated and do not present any obvious defect (Lei et al., 1996; Okano et al., 1999; Tsumura et al., 2006). However, they die upon differentiation, which further indicates that DNA methylation is a required feature of differentiated cell types. Since DNA methylation appears to directly regulate only a limited number of genes, how precisely DNA methylation absence causes the death of differentiated cells remain an open question.

2.2.3 Readers of DNA methylation

The early observation that methylated DNA sequences transfected into *Xenopus* oocytes or mammalian cells are transcriptionally inactive indicates that methylation by itself has an effect on transcription initiation (Stein et al., 1982; Vardimon et al., 1982). The link is thought to be primarily indirect, and DNA methylation most likely acts by modulating the recruitment of DNA binding factors, either preventing the recruitment of activators, or reversely, attracting transcriptional repressors. Of note, *in vitro* assays indicate that DNA methylation alone could modify the structure of chromatin by inducing a tighter wrapping of DNA around nucleosomes, but the relevance of this observation *in vivo* remains to be investigated (Lee and Lee, 2012)

Several methylation-sensitive transcriptional activators have been identified (Tate and Bird, 1993), and for example YY1 recruitment was shown to be abolished by the methylation of a single CpG inside its binding motif, playing a role in the regulation of the *Peg3* imprinted locus (Kim et al., 2003). By contrast, the KRAB-ZFP ZFP57 is known for binding only to methylated motifs, and the subsequent recruitment of KAP1[13] is functionally linked to the maintenance of genomic imprinting (Quenneville et al., 2012). However, many transcription factors are completely insensitive to DNA methylation. Additionally, potentially because transcriptional activators usually associate with H3K4 methylation, transcription factor-binding sites are often hypomethylated and several transcription factors were even shown to bind methylated sequences and subsequently promote their demethylation (Kress et al., 2006; Stadler et al., 2011). How transcription factors influence DNA methylation and whether a

[13] *Cf.* paragraph 2.3.2

given factor is sensitive or instructive to DNA methylation remains in most cases largely unexplored.

Transcription factors have usually recognition sequences of several base pairs, which can be sensitive or not to DNA methylation if the motif contains a CpG. Some proteins are by contrast CpG-specific readers that recognize methylated or unmethylated DNA. In the mouse genome, 12 proteins contain a CxxC domain that recognizes specifically unmethylated CpGs (Long et al., 2013). At the exception of DNMT1 that uses its CxxC domain for the faithful propagation of DNA methylation, CxxC-domains seem principally used to target various proteins to unmethylated CpG islands and favor a transcriptionally permissive chromatin environment. For example, the protein CFP1 is present in 80 % of CpG islands in the mouse genome. Cfp1 is part of the SETD1 H3K4 methylation complex and mediate H3K4 methylation, even at transcriptionally silent regions (Thomson et al., 2010). Similarly, the H3K4-KMT KMT2a and KMT2b also possess a CxxC-domain. The H3K36me2-KDMs KDM2a and KDM2b associate genome-wide with 90% of CpG islands, which leads to H3K36me2 depletion from CpG islands, whereas the rest of the genome is globally enriched for this mark (Blackledge et al., 2010, 2014). Finally, TET1 and TET3 enzymes also possess a CxxC domain. In mouse ES cells, TET1 is preferentially localized in CpG island-associated promoters and appears to help maintaining their hypomethylation (Wu et al., 2011a). As a whole, it appears that CxxC-proteins greatly contribute to defining the very particular chromatin environment of CpG islands. Interestingly, this permissive environment appears as the default state of many CpG islands, but is not necessarily coupled with active transcription (Weber et al., 2007). Many DNA unmethylated and H3K4 methylated CpG island promoters are silent, especially when Polycomb proteins are present as well.

While CxxX-domains recognizes unmethylated DNA, methylated CpGs are specifically bound by the Methyl-CpG-binding domain (MBD) proteins MBD1/2/4 and MeCP2, which are thought to contribute to transcriptional repression (Hendrich and Bird, 1998). MBD proteins are principally attracted to CpG-dense and highly methylated regions such as methylated CpG islands or pericentric major satellite sequences (Baubec et al., 2013). MeCP2 and Mbd2 promote transcriptional repression by recruiting histone deacetylase complexes (Nan et al., 1998; Ng et al., 1999), and MeCP2 was also shown to compact chromatin by assembling secondary structures (Georgel et al., 2003). MBD1 on the other hand associates with H3K9-KMTs and histone chaperones to ensure the faithful propagation of H3K9me3 during DNA replication (Sarraf and Stancheva, 2004).

2.2.4 Conserved functions of DNA methylation

The aforementioned observations highlight the intrinsic nature of DNA methylation as a powerful transcriptional repressor, at least in CpG dense contexts. Probably because most of the DNA methylation landscape is established early during embryonic development and in particular before organogenesis, and also because DNA methylation stays for life, transcriptional repression of protein-coding genes appears essentially limited to very specific classes of genes that need to be silent permanently and in all tissues. This situation is exacerbated in plants, which, by contrast to mammals, do not reprogram their DNA methylation during embryonic development. As a consequence, alterations of the DNA methylation landscape were shown to be transmitted over several generations and to affect several complex traits (Cortijo et al., 2014; Johannes et al., 2009). While DNA methylation does not have the flexibility required for the fine-tuning of gene regulation, its stability is perfect to ensure the permanent silencing of transposable elements. Transposon control is indeed one the most ancient function of DNA methylation that is shared by animals, plants and multiple protists (Huff and Zilberman, 2014). It is even suggested that DNA methylation evolved precisely for this purpose (Yoder et al., 1997).

In plants, loss of CpG methylation by genetic means results in the reactivation of TEs and their amplification in the genome for several generations, even if the DNA methylation machinery is reintroduced and fully functional (Marí-Ordóñez et al., 2013; Tsukahara et al., 2009). In mammals, it is canonically assumed that DNA methylation is required for transposon silencing in differentiated tissues. But because differentiated cells cannot survive in absence of DNA methylation, there are ironically few direct evidences for the role of DNA methylation in TE silencing *in vivo*, and the same two observations originating from the same laboratory are invariably cited: 1) *Dnmt1*-KO mouse embryos show a 1,000 fold reactivation of IAPs elements, and 2) both LINEs and ERVs are activated in the male germline in absence of Dnmt3l (Bourc'his and Bestor, 2004; Walsh et al., 1998; Zamudio et al., 2015). Similarly, conditional knockout of Dnmt1 in differentiated cells in culture or *in vivo* result in the strong reactivation of IAPs (Hutnick et al., 2010); it can further be added that in many cancer types, TEs are severely hypomethylated and highly expressed (Schulz et al., 2006). By contrast, DNA methylation depletion in undifferentiated ES cells does not result in major TE re-activation. ES cells are in fact thought to be able to compensate for its absence by relying on other mechanisms, in particular H3K9me-related pathways (Karimi et al., 2011b; Matsui et al., 2010). This observation often led to the conclusion that DNA methylation is dispensable

for TE silencing in ES cells. The work presented in this manuscript however shows that this statement needs to be revisited[14].

Surprisingly, a function that appears even more conserved than transposon silencing is positively correlated with gene expression. In almost all species where DNA methylation is present, DNA methylation is especially enriched in the body of highly transcribed genes (Feng et al., 2010). For example, some species such as the tunicate *Ciona intestinalis*, the anemone *Nematostella vectensis* or the honeybee *Apis mellifera* do not methylate transposons, but have high DNA methylation level in gene bodies (Suzuki et al., 2007; Zemach et al., 2010). The function of gene body methylation is not well understood. A body of evidence suggests that it could regulate splicing (Lev Maor et al., 2015) and suppress the activity of intragenic transcriptional units (cryptic promoters or transposable elements) (Maunakea et al., 2010). Gene-body methylation appears closely tied to H3K36 methylation. In yeast and mammals, H3K36 methylation is highly enriched in the body of highly transcribed genes. In yeast at least, H3K36me3 recruits enzymes such as histone deacetylases to condense chromatin and prevent the activation of cryptic start sites (Carrozza et al., 2005). In mammals, DNMT3a and DNMT3b PWWP domain binds to H3K36me3 and the two enzymes are recruited to the body of actively transcribed genes. In oocytes for example, DNA methylation is exclusively found in the body of actively transcribed genes and this deposition is dependent on DNMT3a and DNMT3l (Smallwood et al., 2011). By contrast in ES cells, DNMT3b, but not DNMT3a, localizes preferentially in the body of active genes (Baubec et al., 2015).

2.3 H3K9 methylation

H3K9 methylation is the hallmark of what is often referred as constitutive heterochromatin, especially around functional chromosome structures such as telomeres and centromeres. Centromeres are made of repeated DNA sequences known as "minor satellites" repeats. Pericentric heterochromatin refers to the large blocks of dense chromatin that surround centromeres and is principally composed in mouse of large array of AT-rich "major satellite" repeats interspersed with TEs (Guenatri et al., 2004). During interphase, pericentric regions from several chromosomes cluster together to form chromocenters, which are easily visualized as bright foci after DAPI staining or H3K9me3 immunofluorescence. H3K9 methylation is also present throughout the genome and is particularly enriched in TEs. H3K9 methylation exists at the mono-, di- or tri-methylated states, but only di- and tri- methylation

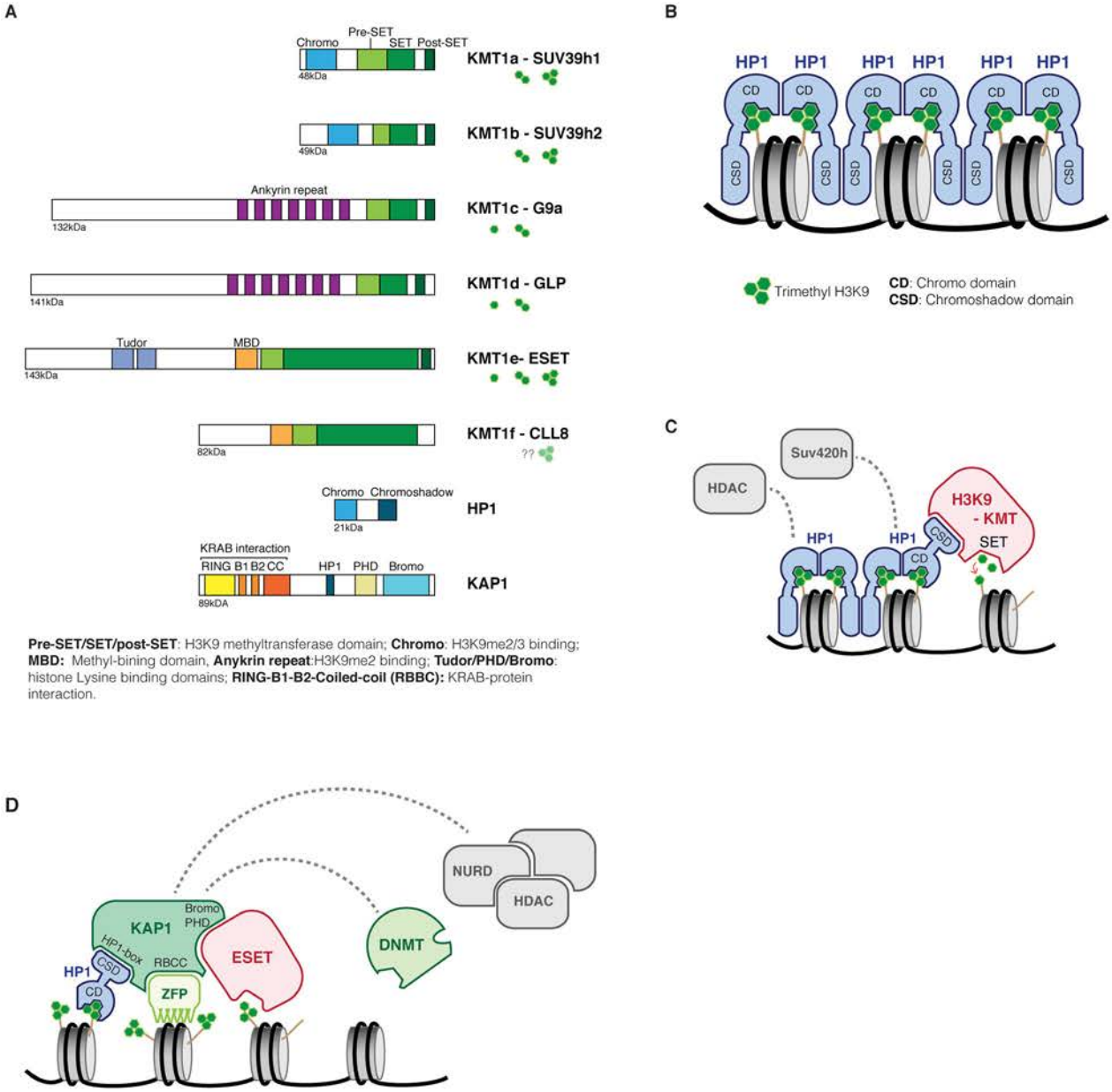
[14] *Cf.* Result part

are linked to chromatin compaction and gene silencing. At least 11 different KMTs, and other proteins such as HP1 and KRAB-associated protein 1 (KAP1), are implicated into H3K9-mediated repression. Di- and tri-methylation, as well as the numerous proteins involved in their deposition and reading, contribute to chromatin silencing in both overlapping and non-overlapping manners. In ES cells, mass spectrometry analysis indicates that 50% of histone H3 are unmodified at K9 residues, while 22% are mono-, 18% di- and 10% tri-methylated (Peters et al., 2003).

Importantly, the localization of H3K9me3 is globally unaffected by the absence of DNA methylation in *Dnmt*-tKO ES cells (Karimi et al., 2011b; Tsumura et al., 2006). Even if the two modifications are often found together, this indicates that the recruitment and maintenance of H3K9 methylation is essentially independent of DNA methylation, at least in mouse ES cells.

2.3.1 H3K9-mediated chromatin compaction trough HP1 recruitment

H3K9 methylation is thought to promote chromatin compaction principally by recruiting the architectural protein HP1, a highly conserved protein that was initially discovered in *Drosophila* (James and Elgin, 1986) and that count three isoforms in mammals, HP1α,β and γ (or CBX5, CBX1 and CBX3, respectively). HP1 is critical for the formation of high-order chromatin structure, even if the precise mechanisms by which HP1 mediates chromatin compaction remain unclear. HP1 proteins are small (~25kDA) and are characterized by conserved N-terminal chromo- and C-terminal chromoshadow-domains, separated by a central and more variable "hinge" region (**Introductory figure 16A**). HP1 is recruited to chromatin by its chromodomain and specifically binds to di- or tri- methylated H3K9 (Bannister et al., 2001; Lachner et al., 2001; Nielsen et al., 2002b). The chromoshadow-domain promotes the interaction with other proteins and allows in particular the formation of HP1 di- or multimeric complexes. In yeast, HP1 appears to bind nucleosomes as tetramers that bridge toward homologous HP1-structures on adjacent nucleosomes, which results in a highly compacted chromatin (**Introductory figure 16B** and Canzio et al., 2011). Because it can interact simultaneously with multiple proteins, HP1 oligomers probably act as binding platforms for the assembly of large macromolecular complexes in chromatin. HP1 also acts as a shield that blocks the interaction of the transcriptional machinery with the underlying chromatin (Smallwood et al., 2008). Interestingly, the presence of di- or tri-methylated H3K9 residues does not appear sufficient *in vivo* to recruit HP1, but requires direct protein-protein interactions between HP1



Introductory Figure 16. H3K9 methylation

A. Domain architecture of mammalian H3K9-KMT, KAP1 and HP1 proteins

B. Model showing multimerisation of HP1 proteins around nucleosomes. Adapted from Canzio et al. 2011.

C. Model for the maintenance and spreading of heterochromatin by HP1 and H3K9-KMT.

D. Model for KAP1-ESET mediated repression: KRAB-ZFPs recognize specific DNA sequences and recruit KAP1, that in turn recruits ESET, histone deacetylases, HP1 and DNMTs to create a repressive chromatin environment.

chromoshadow domains and the different KMTs that establish the mark (Chin et al., 2007; Sripathy et al., 2006; Stewart et al., 2005). As a consequence of the interaction between HP1 and H3K9-KMTs, artificial tethering of HP1α to the genome was shown to result in the deposition and spreading of H3K9me3 around the tethering site, forming large domains of several kbs. Of note, these H3K9me3 domains are heritably transmitted through multiple cell divisions, even after removal of the initiating stimulus, showing that once established, H3K9me3 domain can be stably and autonomously maintained (Hathaway et al., 2012).

HP1 isoforms are similar in structure, but differ in their localization: HP1α is restricted to pericentric regions, while HP1β and HP1γ can be found elsewhere in the genome (Minc et al., 2000; Nielsen et al., 2001). In line with these distinct localization, the three isoform appear to have also non-overlapping functions, as exemplified by the distinct phenotypes of each of the individual mouse knockout: HP1α absence has no detectable effect, HP1β knockout mice die around birth and HP1γ depletion leads to severe neonatal lethality and infertility (Aucott et al., 2008; Brown et al., 2010). Furthermore, whereas HP1α, β and γ physically block transcription initiation at promoters (Smallwood et al., 2008), in some cell types[15] HP1γ (but not H3K9me2/3) is also enriched in the body of highly transcribed genes and regulates splicing by interacting with elongating forms of RNA polymerase II (Smallwood et al., 2012; Vakoc et al., 2005).

Further compaction can probably be achieved through H4K20 methylation, which was shown *in vitro* to physically condense chromatin. H4K20 mono-methylation is catalyzed by SET8/KMT5a (Fang et al., 2002), while subsequent di- and tri-methylation rely successively on SUV4-20h1/KMT5b and SUV4-20h2/KMT5c (Schotta et al., 2004, 2008). H4K20me1 and H4K20me2 are associated with DNA replication and DNA repair and are broadly distributed in the genome, while H4K20me3 is found exclusively in heterochromatin and its absence is associated with severe defects in overall chromatin structure and genome integrity (Jørgensen et al., 2013). HP1, in particular α- and β-isoforms, was shown to bind directly to SUV4-20h2, therefore mediating the deposition of H4K20me3 in H3K9 methylated regions (Schotta et al., 2008). H4K20-dependent chromatin compaction is probably a two-step process: H3K9 methylation first recruits HP1, which in turn interact with SUV4-20h2 that deposits H4K20me3. In line with this model, H4K20me3 is abolished at pericentric regions in absence of H3K9me3, whereas H3K9 methylation and HP1 association are maintained in absence of SUV4-20h2. Potentially because SUV4-20h2 recruitment depends on specific HP1 isoforms, and because other factors are probably involved, H4K20me3 accompanies

[15] Of note, it was not described in ES cells so far.

principally H3K9me3 in pericentric regions and TEs, but is more rarely associated with H3K9me2.

Other factors than HP1 and H4K20 tri-methylation are probably involved in chromatin compaction and transcriptional repression. For example, H3K9 methylation and HP1 are thought to recruit HDACs (Schultz et al., 2001; Stewart et al., 2005), histone remodeling complexes (Nielsen et al., 2002a) and favor the integration of the H3.3 histones variant by the Daxx-Atrx chaperone complex (Eustermann et al., 2011; Lewis et al., 2010). Of note, H3K7 mono-methylation is also often found associated with H3K9me2/3 and H4K20me3, especially in pericentric heterochromatin, but its function remain obscure (Peters et al., 2003).

2.3.2 H3K9 methyltransferases

Eight distinct H3K9-KMTs have been described so far in mouse and human (**Introductory figure 16A**). The first discovered mammalian KMTs were SUV39h1/KMT1a and the testis-restricted SUV39h2/KMT1b (Aagaard et al., 1999; O'Carroll et al., 2000). SUV39h1/2 contain a conserved chromodomain and a SET domain. They use mono-methylated lysine residues as substrate and specifically govern H3K9 tri-methylation at pericentric heterochromatin. Absence of SUV39h enzymes results in severe chromosomal instabilities (Peters et al., 2003, 2001; Rea et al., 2000). Importantly, SUV39h1/2 bind to HP1 and HP1 is lost at pericentric regions in *Suv39h*-dKO; this suggests a self-enforcing loop where H3K9me3 recruits HP1, that in turn attracts more SUV39h1/2, allowing the spreading of H3K9me3 and HP1 along the chromatin fiber (**Introductory figure 16C** and Lachner et al., 2001). In yeast at least, the SUV39 homolog chromodomain binds as well to H3K9 methylation and this interaction is necessary for the spreading of heterochromatin. (Zhang et al., 2008a). In mammals, it was shown that the repressive property of SUV39h1/2 was located in the N-terminal part containing the chromodomain, and that the catalytic activity was not required for transcriptional silencing when the protein was artificially tethered to the genome. (Firestein et al., 2000). It is postulated that the chromodomain helps assembling multimeric repressive complexes on chromatin. Interestingly, while HP1 binding on chromatin is highly dynamic and in a permanent on-off flux, SUV39h remains immobile at pericentric heterochromatin and probably plays an important stabilizing role (Cheutin et al., 2003; Krouwels et al., 2005)

Pericentric H3K9 mono-methylation is mediated specifically by PRDM3 and PRDM16 (Pinheiro et al., 2012). The methylation reaction occurs on the cytoplasmic pool of histones; mono-methylated histones are then transported to pericentric heterochromatin to be

converted into H3K9me3 by SUV39h enzymes. Absence of PRDM3 and PRDM16 leads to the disappearance of H3K9me3, H4K20me3 and HP1 from pericentric regions. Interestingly, depletion of PRDM3 and PRDM16 further results in the disintegration of the DAPI-dense heterochromatin foci and in severe destabilization of the nuclear lamina, which is not the case in *Suv39h*-dKO cells (Pinheiro et al., 2012). This indicates a critical role for H3K9me1 in itself in the formation and maintenance of pericentric heterochromatin. How H3K9me1-marked histones and SUV39h-dependant H3K9 methylation are targeted to pericentric regions remains largely unanswered questions.

While pericentric heterochromatin relies extensively on PRDM3, PRMD16 and SUV39h1/2, H3K9 methylation is established by other KMTs in the rest of the genome. G9a/KMT1c and GLP/KMT1d are the primary enzymes responsible for H3K9 mono- and di-methylation in non-pericentric regions (Rice et al., 2003). G9a and GLP form heterodimers via their SET domains, and the G9a/GLP complexes are functional *in vivo*, as both G9a and GLP single KO contain only traces of H3K9me2 and severely reduced levels of H3K9me1 (Tachibana et al., 2002, 2005). G9a is able to methylate non-histone proteins, including itself, and its auto-methylation is necessary for the further recruitment of HP1 (Chin et al., 2007). G9a and GLP possess also a specific "ankyrin repeat domain" that binds to H3K9me1/2, meaning that the two enzymes are capable of both read and write the same mark (Collins et al., 2008). These self-enforcing loops are thought to favor the spreading of H3K9me2 into very large blocks (up to 5Mb) that cover between 4% and 50% of the genome (in mouse embryonic stem cells and in the liver, respectively) (Wen et al., 2009). In particular, large repressive chromatin domains associated with the nuclear lamina are highly enriched for H3K9me2, and the majority of genes upregulated in mouse *G9a*-KO ES cells are indeed localized at the nuclear periphery (Guelen et al., 2008; Yokochi et al., 2009). H3K9me2-mediated repression controls critical biological processes: in contrast with *Suv39h1/2*-knockout mice that are viable (Peters et al., 2001), *G9a* and *GLP*-knockout mice are embryonic lethal due to severe growth defects, and both male and female meiosis cannot be completed in absence of *G9a* (Tachibana et al., 2002, 2005, 2007). The recruitment of G9a/GLP complexes to chromatin likely involves multiple interactions with chromatin or sequence-specific DNA-binding molecules. In human and mouse, G9a/GLP complexes were shown to contain the zinc finger protein WIZ, and human complexes associate as well with ZNF644, another DNA binding protein. WIZ and ZNF644 mediate the recruitment of G9a to specific DNA sequences (Bian et al., 2015; Ueda et al., 2006). Other DNA-binding transcriptional

repressors such as E2F6 or CDP/cut were also shown to associate with G9a (Nishio and Walsh, 2004; Ogawa et al., 2002).

The last characterized H3K9-KMTs are the related proteins ESET/SetDB1/KMT1e and CLL8/SetDB2/KMT1f. Very little is known about CLL8. Depletion of the protein leads to globally reduced levels of H3K9me3, but not of H3K9me2 (Falandry et al., 2010). CLL8 appears important for chromosomal segregation during mitosis and was shown to repress specific genes during embryonic development in zebrafish and during anti-viral responses in mouse (Schliehe et al., 2015; Xu et al., 2010).

On the other hand, ESET has received considerable attention, especially because its biology is intimately linked with the KAP1 protein. ESET can successively mono-, di- and tri-methylate H3K9 residues and is thought to be responsible for most of non-pericentric H3K9me3 (Schultz et al., 2002; Wang et al., 2003). KAP1 is a multi-domain protein that acts as a binding platform for the establishment of a repressive environment around specific DNA sequences (**Introductory figure 16D**). It contains an N-terminal RBCC domain, a C-terminal PHD-bromo domain and a central HP1 binding domain. The RBCC domain binds specifically to the N-terminal KRAB (Krüppel-Associated box) domain of C2H2 zinc finger proteins (ZFPs)(Friedman et al., 1996), a superfamily of DNA binding proteins that represents around half of the annotated transcription factors. C2H2 zinc finger proteins recognize specific DNA sequences and 40% possess a KRAB domain, representing namely 423 genes encoding transcripts for 742 structurally distinct proteins in the human genome (Bellefroid et al., 1991; Huntley et al., 2006). After being targeted to specific genomic loci by KRAB-ZFPs, KAP1 further recruit ESET and the NuRD histone deacetylase complex through its PHD-bromodomain and establishes a repressive environment that is stabilized by the interaction of HP1 with both H3K9me3 and KAP1 (Ryan et al., 1999; Schultz et al., 2001, 2002).

Whereas KAP1 and ESET are ubiquitously expressed, cell-type specific expression of KRAB-ZFPs and the diversity of their recognition sequences allow the fine-tuning of numerous biological processes. In particular, both KAP1 and ESET are critical for embryonic development, as knockout embryos die after 4-5 days of gestation (Cammas et al., 2000; Dodge et al., 2004). KAP1 and ESET are also linked to pluripotency. KAP1 binds to numerous promoters in mouse ES cells and its depletion induces spontaneous differentiation (Hu et al., 2009). As another evidence for their critical role, ESET is the only H3K9-KMT for which knockout is known to be lethal in mouse ES cells, a phenotype shared by *Kap1* knockout (Dodge et al., 2004; Rowe et al., 2010). The binding motifs of most KRAB-ZFPs are still unknown and substantial efforts are currently undergoing to characterize the full repertoire of

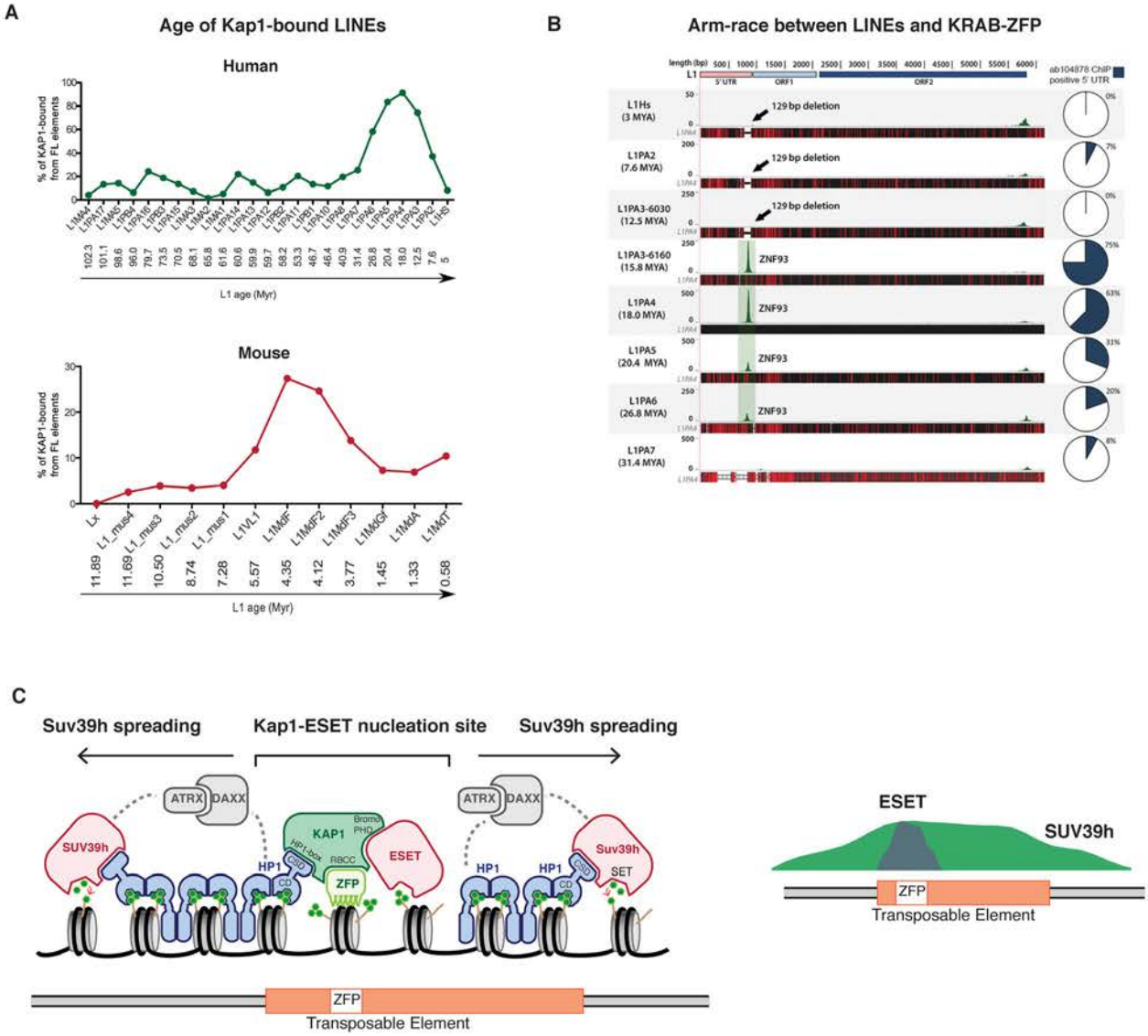
binding sequences. Nevertheless, it is clear from many studies that ESET and KAP1 play a critical role in the regulation of TEs, and indeed the majority of KRAB-ZFPs appears to bind specifically to TEs (Matsui et al., 2010; Najafabadi et al., 2015; Rowe et al., 2010)

Even if they have primarily non-overlapping functions, the different H3K9-KMTs seem nonetheless to collaborate. PRDM3 and PRDM16 are the main enzymes responsible for the mono-methylation of H3K9 in pericentric regions, but at least a small part of the mono-methylation appears to be provided by ESET. HP1α, KAP1 and ESET form with the histone chaperone CAF1 a complex that mono-methylates non-nucleosomal histones H3 and apparently delivers H3K9me1 to SUV39h enzymes in pericentric regions (Loyola et al., 2009). Furthermore, a subset of KMTs G9a, GLP, ESET and SUV39h1 was shown to form mega-complexes that are directly implicated in the repression of multiple G9a target genes and recruited at pericentric repeats (Fritsch et al., 2010). Importantly, deletion of SUV39h1 or G9a results in global destabilization of the other KMTs at the protein level, indicating that these interactions are functionally important.

2.3.3 Transposons repression by H3K9 methylation in ES cells

DNA methylation appears required for TE silencing in differentiated cells. However, in *Dnmts* triple knockout (*Dnmt*-tKO) ES cells, only two transposon families (IAPEz and MMERGLN) are modestly upregulated. By contrast, deletion of ESET result in the strong activation of many different ERVs from the class I and II, and more modestly of L1s (Karimi et al., 2011b; Matsui et al., 2010). Knockout of *Kap1* has similar effect on L1s and leads to the strong activation of the three classes of ERVs (Rowe and Trono, 2011). The ESET-KAP1 pathway is indeed thought to be the principal mechanism for TE control in ES cells: specific KRAB-ZFPs recognize specific transposon sequences and recruit KAP1, that in turn attracts ESET, histone deacetylase and HP1 to create a repressive environment. Of note, although HP1 is enriched in TEs marked by ESET, depletion of HP1α, β and γ alone or in combination results in only very modest TE reactivation, indicating that HP1 proteins are mainly dispensable for TE silencing (Maksakova et al., 2011).

Another layer of regulation is brought by the histone H3.3 chaperone complex containing the protein DAXX and the chromatin remodeler ATRX. Both proteins appear to be recruited by KAP1 and mediate the deposition of the histone variant H3.3, especially at IAP elements, reinforcing the efficient association of KAP1 with the chromatin and the deposition of H3K9 methylation (Elsässer et al., 2015; He et al., 2015; Sadic et al., 2015). Deletion of DAXX or ATRX result in reduced levels of H3K9 methylation and TE



Introductory Figure 17. H3K9me3-mediated repression of Transposable elements

A. Proportion of L1 elements bound by KAP1 depending on the family, ordered by age. From Castro-Diaz et al., 2014

B. Example of the arms race between a KRAB-ZFP and transposable elements. The figure represents ChIP-seq signal for ZFN93 mapped to human L1PA consensus sequences. ZFN93 recognize L1PA6 to L1PA3, but younger L1 elements carry a 129pb deletion containing exactly ZFN93 binding motif. From Jacobs et al., 2014

C. Model from the establishment of large repressive chromatin domain around transposable elements: after initial recruitment by KRAB-ZFP, KAP1 would further recruit HP1, ATRX-DAXX and ESET, leading to the deposition of H3K9me3 around the binding site. Subsequently, HP1, H3K9me3 and DAXX would recruit SUV39h, causing the spreading of H3K9me3 away from the KAP1-binding site. Inspired by Bulut-Karsioglu et al., 2014 and Elsässer et al., 2015.

upregulation, in particular of IAPs. ATRX contains an ADD domain that especially binds to H3K9me3 and interact with HP1, which could facilitate the recruitment of the complex to chromatin (Eustermann et al., 2011; Iwase et al., 2011).

KRAB-ZFPs represent the targeting module of the KAP1 complex. Most of the time, they bind near the promoter region of TEs, either close to the LTR of ERVs or in the 5'UTR of LINEs. As a result, H3K9 methylation is usually more strongly enriched in the 5' region of TEs, even if some elements like IAPs are enriched on their full sequence. For example, ZFP809 was shown to mediate the silencing of MLV elements by binding to the proline tRNA binding site (PBS), further recruiting KAP1 and ESET (Wolf and Goff, 2007, 2009). KRAB-ZFPs are the proteins that evolve the fastest in mammalian genomes, they are very divergent between species and under a strong selective pressure to change their DNA-binding specificity (Emerson and Thomas, 2009). KRAB-ZFPs are in fact one of the best example that illustrates the constant arm-race between TEs and host defense mechanisms. The family of KRAB-ZFPs is thought to evolve rapidly in order to constantly adapt genome defenses strategies to newly arriving TEs. It was demonstrated that the appearance of new families of ERVs in evolutionary time was a strong predictor of the appearance of new KRAB-ZFPs (Thomas and Schneider, 2011). However, the acquisition of new defenses is a slow process that takes millions of years. MLV elements have recently integrated into the mouse genome, and are not even present in every strains. Their recognition by ZFP809 is in fact a kind of lucky guess for the mouse genome, because ZFP809 binds also to the same PBS of the more ancient RLTR6 and VL30 elements (Wolf et al., 2015). Comparably, KAP1 does not bind to the youngest L1 elements, but represses only families of moderate age (10-25Myrs and 3-5Myrs in human and mouse, respectively) (Castro-Diaz et al., 2014). Saliently, KAP1 does not bind either to very old LINEs. These elements have become inactive a long time ago and the KRAB-ZFPs that used to target them were probably redirected to younger elements (**Introductory figure 17A**). Whereas KRAB-ZFPs evolved to silence TEs, some transposons were also caught trying to escape. In human, ZNF93 binds predominantly to a small region in the 5'UTR of L1PA3 and L1PA4 elements (active 15-20Myrs ago). The younger elements L1PA2 and L1PA1 harbor a 129pb deletion that precisely includes the ZNF93 binding motifs and probably allowed them to escape extinction (**Introductory figure 17B** and Jacobs et al., 2014). In good agreement with the observation that Kap1 does not bind to old elements, H3K9me3 appears restricted to relatively young and intact TEs in mouse ES cell (Bulut-Karslioglu et al., 2014).

SUV39h enzymes bring another layer of control of TEs, and especially of LINEs. In *Suv39h*-dKO, H3K9me3 diminishes at ERVs and almost totally disappears at L1s. As a result, some L1s of the A subtype (but not T or F) are strongly upregulated, while ERVs show only a vey modest activation (Bulut-Karslioglu et al., 2014). ESET and HP1 are preferentially enriched inside the transposon sequence, and particularly nucleate in the 5' region. SUV39h by contrast appears to be especially involved in the spreading of H3K9me3 inside and outside the transposon sequence and is found enriched in larger domains. Since little is known about the mechanisms that target SUV39h in general, it is tempting to propose a model where SUV39h would act downstream of KAP1, ESET and ATRX-DAXX, especially because SUV39h1 was shown to physically interact with DAXX (He et al., 2015). After an initial recruitment by KRAB-ZFPs, KAP1 would further recruit HP1, ATRX-DAXX and ESET, leading to the deposition of H3K9me3 around the binding site. Subsequently, HP1, H3K9me3 and DAXX would recruit SUV39h, causing the spreading of H3K9me3 distally from the Kap1-binding site (**Introductory figure 17C**).

Importantly, H3K9me3-mediated silencing appears restricted to undifferentiated cells. H3K9me3 marking at TEs is substantially lower in MEFs or upon differentiation of ES cells and deletion of SUV39h, ESET or KAP1 in differentiated cells does not affect transposon expression (Bulut-Karslioglu et al., 2014; Matsui et al., 2010; Rowe et al., 2013). This indicates that other silencing mechanisms, probably involving DNA methylation, take over during differentiation.

However, class III ERVs represent an exception: H3K9me3 is barely present in the sequence of MERVL in any cell type. Class III ERVs seem preferentially regulated by H3K9me2. In ES cells, MERVLs are highly enriched for H3K9me2 and are strongly activated following *G9a* knockout (Maksakova et al., 2013). Of note, MERVL LTRs contain the binding sequence of E2F6, a transcription factor that was shown to associate with G9a and which could therefore be used to target H3K9me2 to MERVL elements (Ogawa et al., 2002, personnal communication). But the picture is probably more complex. even though MERVL lacks H3K9me3 enrichment, it is strongly upregulated in *Kap1* and *Suv39* mutant ES cells, and modestly activated by *Eset* deletion. MERVL appears to be the only TE to be repressed simultaneously by the three major H3K9 pathways, even though it could be trough very indirect mechanism (Bulut-Karslioglu et al., 2014; Maksakova et al., 2013). Moreover, MERVL elements are also upregulated in absence of KDM1a (LSD1), an H3K4me1/2 demethylase that appears necessary to remove activating histone modification from MERVL promoters (Macfarlan et al., 2011).

2.4 H3K27 methylation and Polycomb

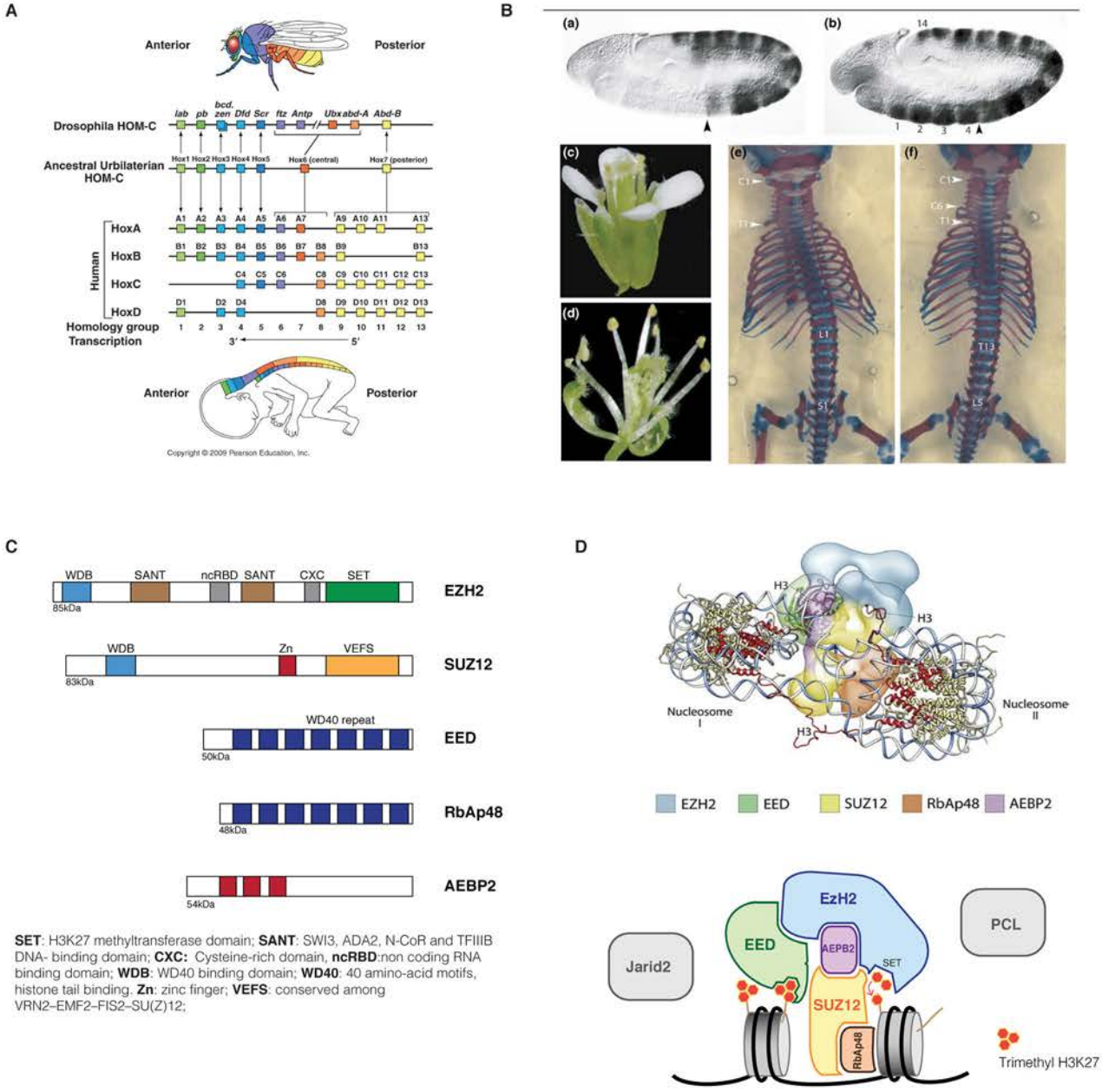
H3K27 methylation is associated with Polycomb group (PcG) proteins. The term Polycomb originates from a *Drosophila* mutant with improper body segmentation, and PcG proteins were first identified as repressors of homeotic (*Hox*) genes during early embryonic development (Lewis, 1978). PcG function is conserved between plants, animals and various unicellular eukaryotes, indicating a very primitive origin that probably dates back to the last unicellular common ancestor[16] (Shaver et al., 2014). Contrary to H3K9 methylation that marks essentially repetitive sequences such as pericentric repeats and transposons, Polycomb is principally associated with gene repression, especially of developmental regulators. The repressive function of Polycomb at *Hox* genes is conserved in mammals and PcG proteins are indeed essential for the regulation of developmental genes in multiple cell types (**Introductory figure 18A** and **B**). They appear critical for the transition or maintenance of cell fates and play particularly important roles in embryonic development, stem cells identity and cancer progression (Aloia et al., 2013; Sauvageau and Sauvageau, 2010). As an indication of their crucial role during development, deletion of the core PcG proteins results in early embryonic lethality.

In mammals and *Drosophila*, two major Polycomb complexes are conserved: Polycomb repressive complex 1 (PRC1) and 2 (PRC2). PRC1 compacts chromatin (Shao et al., 1999) and catalyzes the mono-ubiquitination of Lysine 119 on histone H2A (H2AK119ub or H2Aub) (Wang et al., 2004), while PRC2's main function is to deposit H3K27me3 (Cao et al., 2002).

2.4.1 PRC2

PRC2 is composed of four core components that are conserved between *Drosophila* and mammals: EZH1 or 2, SUZ12, EED and RbAp46 or 48, the first three ones being minimally required for its enzymatic activity *in vitro* (**Introductory figure 18C** and Cao and Zhang, 2004; Cao et al., 2002). The complex is further stabilized by AEBP2, which enhances its activity. Other cofactors such as JARID2 or PCL proteins are also involved in the activity and recruitment of PRC2 (Margueron and Reinberg, 2010). EZH1 and EZH2 both possess SET domains and preferentially di- and tri-methylate H3K27 from a mono-methylated residue. EZH1 and EZH2 do not appear to be interchangeable: EZH1 has a lower catalytic activity

[16] Polycomb has however been lost in fungi such as fission and budding yeasts, but is conserved in *Neurospora* (Jamieson et al., 2013).



Introductory Figure 18. Hox genes and PRC2 complex

A. Genetic map of hox gene clusters in *D. melanogaster* and human.

B. Homeotic transformations in PcG mutants. (a,b) Drosophila embryos, (c,d) Arabidopsis flowers and (e,f) mouse skeletons. Staining for the Drosophila Hox protein Ubx in (a) wild-type and (b) Su(z)12 mutant embryos. Arabidopsis flowers from (c) wild-type and (d) FIE (plant EED homolog)- deficient plants. Distal view of axial skeleton from (e) wild-type and (f) Cbx2-deficient mice. From Whitcomb et al. 2007.

C. Domain architecture of mammalian PRC2 proteins.

D. Protein structure and model for PRC2 complex binding to nucleosomes. SUZ12 and RbAp48 serve as a nucleosome-binding module that anchors PRC2 to chromatin. EED binding to one nucleosome places the EZH2 SET domain in close proximity with the H3 tail of the adjacent nucleosome. From Ciferri et al. 2012.

and its depletion has little consequence on global H3K27me2/3 levels (Margueron et al., 2008). Importantly, global levels of H3K27me1 are unaffected by EZH1/2 depletion, indicating that H3K27me1 is deposited by an unrelated and still unknown process. It is sometimes pointed that G9a can methylate H3K27 *in vitro* (Tachibana et al., 2001) and H3K27me1 levels appear indeed reduced in *G9a*-KO (Wu et al., 2011b). However, H3K27me2/3 levels remain unaltered in *G9a*-KO and G9a cannot therefore be considered as a good candidate for the global deposition H3K27me1.

EED binds specifically to H3K27me3 through its C-terminal WD40 domain and H3K27me3 is abolished in absence of this interaction (Margueron et al., 2009). Additionally, resolution of the molecular structure of the PRC2 complex suggests that EED binding to one nucleosome places EZH2 SET domain in close proximity with the H3 tail of the next nucleosome (Ciferri et al., 2012). These positive-feedback loops probably allow for the maintenance and spreading of H3K27me3 along the chromatin fiber (**Introductory figure 18D**).

SUZ12 and RbAp46/48 are thought to serve as a nucleosome-binding modules that anchor PRC2 to chromatin (Nekrasov et al., 2005). In particular, the SUZ12-RbAp46/48 subunit binds to unmodified H3 tails and this interaction is lost in presence of H3K4me3 and H3K36me2-3, which subsequently prevents *in cis* PRC2 to methylate the same histone tail (Schmitges et al., 2011). As H3K4 and H3K36 methylation are associated with active transcription, this probably prevents the spreading of H3K27 methylation into active chromatin territories. Interestingly, whereas H3K4me3/H3K36me2-3 and H3K27me3 cannot be present in the same H3 molecule, they can nonetheless be present *in vivo* in the same nucleosome, each in a different H3 tail, forming "bivalent" structures marked at the same time by activating and repressing marks (Bernstein et al., 2006a; Voigt et al., 2012).

H3K27 methylation is very abundant and is found in approximately 80% of histone H3 in mouse ES cells, with around 15% of mono-, 50% of di- and 15% of tri-methylation (Peters et al., 2003). H3K27me2 and me3 deposition probably relies on slightly different PRC2 variants, as PCL cofactors are for example required for H3K27 tri- but not di-methylation (Sarma et al., 2008). H3K27me2 does not appear to be significantly associated with transcriptional repression, but likely represents an intermediate (or default) state that marks genomic regions to be repressed. H3K27me3 usually harbors two distinct patterns in mammalian cells, with either very large domains of more than 100kb, such as in *Hox* clusters or in the inactive X in females, or smaller domains of a few kilobases. In ES cells, around 50-60% of H3K27me3-enriched regions are localized in the proximity of promoters, marking

about 10% of the total number of genes, especially developmental regulators (Boyer et al., 2006; Marks et al., 2012; Zhao et al., 2007). Interestingly, in mouse or human ES cells, the majority (~80%) of H3K27me3-marked promoters are also enriched for H3K4me3, a mark associated with transcription activation, forming bivalent domains (Bernstein et al., 2006a). Bivalency is hypothesized to represent a specific state that poises key developmental genes for either activation or repression, and indeed many lineage-specific bivalent promoters are definitely activated or silenced following differentiation of ES cells (Mohn et al., 2008). A lot of controversy has followed the description of bivalency, which was criticized as being an *in vitro* artifact restricted to pluripotent ES cells in culture without an *in vivo* equivalent. However, bivalent promoters are also observed in cultured differentiated cells such as neurons, mouse embryonic fibroblasts (MEFs), hematopoietic progenitors and erythrocytes (Cui et al., 2009; Mikkelsen et al., 2007; Mohn et al., 2008). Furthermore, its observation *in vivo* in the brain, in primordial germ cells (PGCs) or in the sperm shows that bivalency likely represents an important mechanism for the regulation of many critical developmental processes (Cui et al., 2012; Hammoud et al., 2009; Sachs et al., 2013).

PRC2 and H3K27me3 have no obvious effect by themselves on chromatin compaction. H3K27me3 could indirectly regulate transcription by preventing other proteins to access chromatin. For example, levels of H3K27 acetylation increase upon loss of PRC2 and it was proposed that H3K27me3 could act by preventing the binding of activating HATs (Pasini et al., 2010). However, PRC2 and H3K27me3 are thought to control transcription mainly by recruiting other factors, principally PRC1.

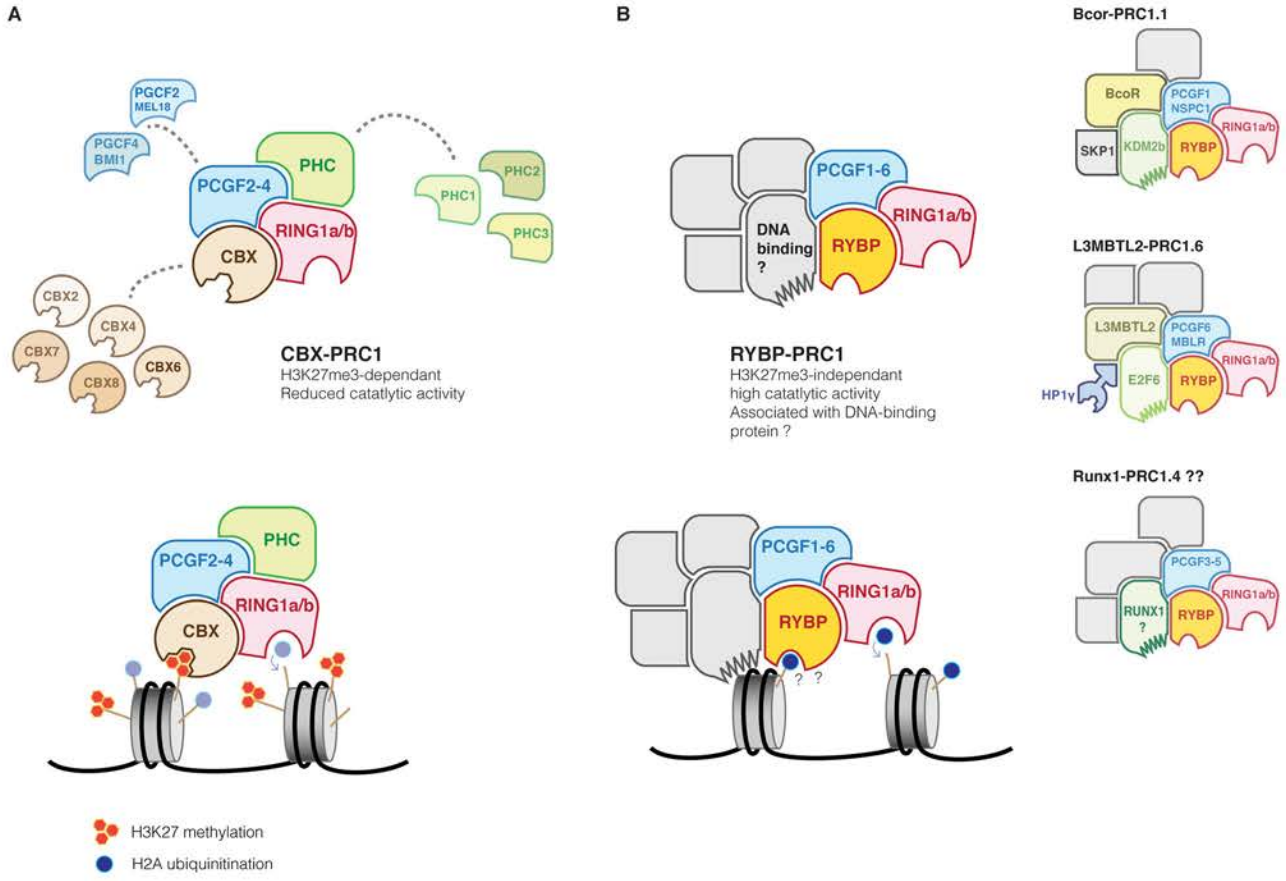
2.4.2 PRC1

In *Drosophila,* the canonical model for Polycomb-mediated silencing is relatively simple: PRC2 *1)* is recruited to chromatin by recognizing specific sequences; *2)* deposits H3K27me3, which *3)* further recruits PRC1 that *4)* compacts chromatin and prevents transcription (Levine et al., 2004). In mammals, the picture is much more complex. While the canonical PRC1 is composed of four proteins in *Drosophila,* there are a total of 16 homologues in mammals, and PRC1 is furthermore engaged into multiple non-canonical complexes (Simon and Kingston, 2013). Since the initial description of a *Drosophila*-like PRC1 in mammals (Levine et al., 2002), the variety of PCR1-related function has grown exponentially and a comprehensive (and still incomplete) picture of PRC1 complexity has only started to emerge recently (**Introductory figure 19**).

The defining structure of PCR1 complexes is the presence of the E3 ubiquitin ligases RING1a and RING1b. E3 ligases carry out the final step of a three-enzymes cascade, catalyzing the transfer of ubiquitin from an E2 donor enzyme (Berndsen and Wolberger, 2014). RING1a and RING1b catalyze mono-ubiquitination of Lysine 119 on histone H2A (H2Aub) and seem both present in PRC1 complexes (Gao et al., 2012; Wang et al., 2004). RING1b is the most active and broadly expressed of the two enzymes, its knockout is embryonic lethal while RING1a-depleted mice survive (del Mar Lorente et al., 2000; Voncken et al., 2003). However, even if *Ring1b*-KO ES cells have almost no H2Aub left, ES cells stop to proliferate and spontaneously differentiate only when RING1a is additionally depleted, indicating that RING1a can at least partially compensate for RING1b loss (Endoh et al., 2008). The second defining feature of PRC1 is the exclusive presence of one of six PCGF proteins (PCGF1 to 6). The association of RING1a/b with a specific PCGF protein defines six distinct PRC1 complexes that can be referred as PRC1.1 to PRC1.6 (Gao et al., 2012). The six PRC1 complexes have distinct protein partners and are functionally different. In particular, they target different genomic regions and regulate different sets of genes.

RING1a/b-PCGF modules further form two types of complexes by associating with either CBX proteins or with RYBP (or its homologue YAF2). CBX proteins and RYBP/YAF2 compete for the same binding pocket on RING1a/b C-terminal domain and are therefore mutually exclusive (Wang et al., 2010). CBX proteins are associated exclusively with PGCF2 (MEL18) and PGCF4 (BMI1), forming in addition with the Polyhomeotic proteins (PHC1/2/3) the canonical form of PRC1 (**Introductory figure 19A** and Gao et al., 2012). The five CBX proteins[17] (CBX2/4/6/7/8) are short proteins related to HP1 that bind H3K27me3 through their chromodomain. Of note, whereas the *Drosophila* homologue appears very specific, certain mammalian CBX proteins also recognize H3K9me3 *in vitro* (Bernstein et al., 2006b; Fischle et al., 2003). H3K27me3 binding by CBX proteins allows the recruitment of CBX-PRC1 to PRC2 targets, promoting chromatin compaction and gene repression (Introductory figure 19A). Indeed, in mouse ES cells, CBX7 is the only CBX protein associated with PRC1 and 90% of CBX7-binding sites in the genome are also occupied by PRC2 (Morey et al., 2012). Depletion of CBX7 leads to reactivation of many PRC2 targets, and loss of PRC2 greatly decreases the presence of PRC1 in gene promoters (Leeb et al., 2010), indicating that CBX-PRC1 act as a transcriptional repressor downstream of PRC2.

[17] Formally, HP1α,β and γ are also members of the Cbx family, but the term Cbx would only be used here to refer to the partners of PRC1.



Introductory Figure 19. PRC1

A. Canonical CBX-PRC1 complex.

B. Core RYBP-PRC1 complex is composed of RYBP, one of six PGCF proteins and of RING1a/b. Additional proteins define different RYBP-PRC1 complexes. On the right are three RYBP-complexes that have been described. From data in Gao et al., 2012; Trojer et al., 2011; Yu et al., 2012, among others.

Please note that the proposed placement of the different proteins is purely speculative.

Interestingly, the multiplicity of potential combination between PGCF2 and 4 and the five CBX proteins seems to provide an additional plasticity to PRC1 function. In pluripotent ES cells, CBX7 associates preferentially with PGCF2 and represses many developmental genes, including *Pgcf4*, *Cbx2* and *Cbx4*. Upon differentiation, CBX7-PGCF2 complexes switch for CBX2/CBX4-PGCF4 complexes that then regulate pluripotency genes, including *Cbx7* (Morey et al., 2012). How the different CBX proteins distinguish their targets remains mysterious. This fine-tuned regulation during cell differentiation highlights the complexity of PRC1-mediated repression and indicates without doubt that H3K27me3 is not the only factor implicated in PRC1 recruitment.

On the other hand, RYBP/YAF2 can be found with the six PGCF proteins and its presence further exclude PCH proteins (**Introductory figure 19B** and Gao et al., 2012). Importantly, RYBP-PRC1 does not contain a chromodomain subunit and is recruited to chromatin independently of H3K27me3 (Tavares et al., 2012). Indeed, while in ES cells 90% of RING1b binding sites are also occupied by H3K27me3, in other cells type like HEK-293T most of RING1B binding sites are free of H3K27me3, and these H3K27me3-free binding sites are also enriched for RYBP (Gao et al., 2012). Many examples show as well that PRC1 can be recruited to target sequences when PRC2 components are depleted, either at gene promoters or in large domains such as the inactive X in females (Pasini et al., 2007; Schoeftner et al., 2006). In several cases if not always, RYBP-PRC1 complexes contain DNA-binding factors that trigger their recruitment to specific locations: PRC1.4 and PRC1.6 associates with the transcription factors RUNX1 and E2F6, respectively (Trojer et al., 2011; Yu et al., 2012), whereas PRC1.1 contains the H3K36-KDM KDM2b that binds to unmethylated CpG-rich sequences trough it CxxC domain (Wu et al., 2013). Of note, RYBP-PRC1 appears to have a stronger catalytic activity than CBX-PRC1, as artificial tethering of CBX-PRC1, but not of RYBP-PRC1, fails to deposit H2Aub (**Introductory figure 20A** and Blackledge et al., 2014). Beside this, RYBP seems to have DNA and ubiquitin-binding activity *in vitro*, which could help stabilizing the complex to target regions (Arrigoni et al., 2006; Neira et al., 2009).

2.4.3 Polycomb recruitment and transcriptional repression

The interplay between PRC1 and PRC2 is complex and probably varies depending on the cell type. CBX-PRC1 is recruited to regions marked by H3K27me3, while RYBP-PRC1 is recruited independently of this mark. In *Drosophila*, PRC2 binds to specific DNA sequences, but such a mechanism does not seem to exist in mammals (Schuettengruber and Cavalli,

2009). At the exception of the large mega-domains found around *Hox* clusters or in the inactive X chromosome, genome-wide analysis of H3K27me3 distribution reveals in fact that PRC2 almost perfectly overlaps with CpG islands (Ku et al., 2008). Several studies showed indeed that PRC2 was systematically attracted to GC-rich sequences, at the condition that they are DNA unmethylated and depleted of activating transcription factors (Jermann et al., 2014; Mendenhall et al., 2010). The precise mechanism by which PRC2 is recruited to GC-rich sequences remains however unclear. JARID2, an important cofactor of PRC2, could be involved, as it possesses a DNA binding domain that binds preferentially to GC-rich sequences, especially short CCG tandem repeats (Li et al., 2010; Peng et al., 2009).

It appears clear as well that PRC2 can be recruited by long non-coding RNAs (ncRNAs), especially to form large H3K27me3 mega-domains. PCR2 is recruited *in cis* by the ncRNAs *Xist* and *Kcnq1ot1* at the inactive X in females and at the *Kcnq1* imprinted region, respectively (Pandey et al., 2008; Plath et al., 2003), or *in trans* at the *HoxD* locus by the ncRNA *HOTAIR* coming from the *HoxC* locus (Rinn et al., 2007). EZH2 strongly binds RNA, but this interaction has very low specificity and 5-25% of the mouse transcriptome is recovered after EZH2 pulldown assay (Zhao et al., 2010). JARID2 binds RNA as well and might be more specific, as it as been shown to been required for the recruitment of PRC2 by *Xist* to the inactive X chromosome (da Rocha et al., 2014).

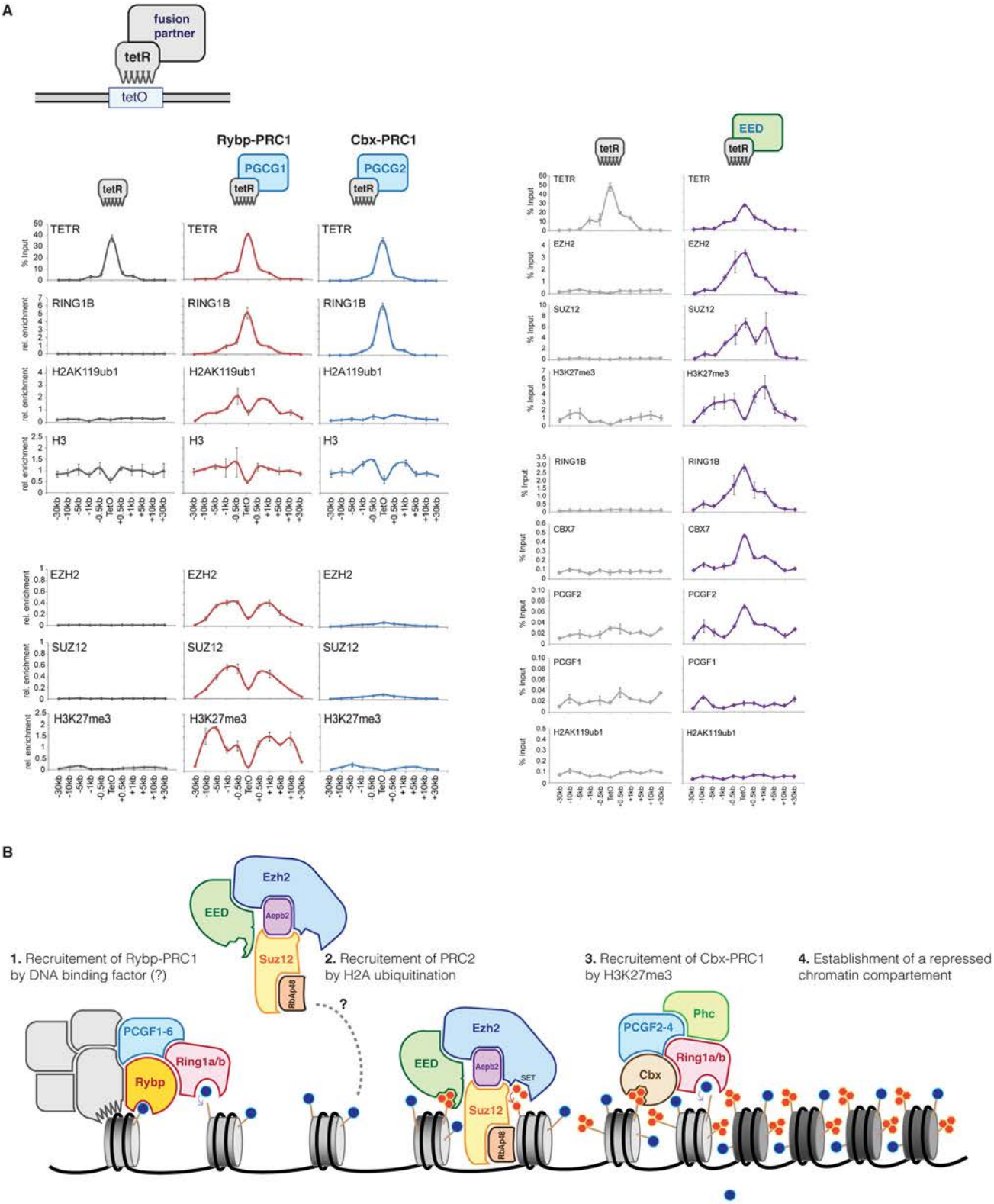
In mouse ES cells, 90% of RING1b binding-sites are also marked by H3K27me3, most of the time in the presence of CBX7, consistent with the canonical assumption that PRC1 is recruited by H3K27me3 (Morey et al., 2012). However, reciprocally, some studies reported that only half of H3K27me3-enriched regions are bound by Ring1b, indicating that CBX-PRC1 recruitment by H3K27me3 is not automatic and that other mechanisms control CBX-PRC1 recruitment (Ku et al., 2008).

In ES cells lacking PRC2, PRC1 enrichment decreases at sites normally occupied by H3K27me3, but interestingly it does not disappear completely and remains at around one quarter of its initial level. Moreover, H2A ubiquitination levels appear barely affected (Leeb et al., 2010; Tavares et al., 2012). This indicates that PRC1 recruitment and H2Aub at PRC2 target sites are partially independent of PRC2. In fact, it was shown in ES cells that *1)* artificial tethering of RYBP-PRC1, but not of CBX-PRC1, resulted in PRC2 recruitment, *2)* PRC2-binding and H3K27me3 were globally lost after depletion of PRC1, and finally that *3)* H2A ubiquitination was specifically responsible for the recruitment of PRC2 (**Introductory figure 20A** and Blackledge et al., 2014; Cooper et al., 2014). These results shift the usual paradigm of PRC1 and PRC2 recruitment and shows that PRC1 can efficiently attract PRC2

and lead to H3K27 methylation. Recruitment of PRC1 and PRC2 could occur in fact in a three-step process: RYBP-PRC1 would be first recruited to chromatin, probably by specific DNA binding factors, and deposit H2Aub. PRC2 would then recognize this mark by a yet unknown process and methylate H3 tails. Finally, CBX-PRC1 would bind to H3K27me3, leading to chromatin compaction and transcriptional repression (**Introductory figure 20B**). However in ES cells, RYBP-PRC1 presence does not necessary correlate with PRC2 enrichment. Additionally, CBX7 and RYBP co-occupy only 20-30% of their target promoters and regulate distinct classes of genes, indicating that the two PRC1-complexes have both common and exclusive targets (Morey et al., 2013).

KDM2b-PRC1.1 complexes are particularly interesting candidates for the recruitment of PRC2, because KDM2b CxxC domain recognizes GC-rich and DNA unmethylated sequences, which is precisely where H3K27me3 tends to be observed (Blackledge et al., 2014; Ku et al., 2008; Wu et al., 2013). Deletion of KDM2b has however little impact on the global repartition of Polycomb proteins, and only induces a modest reduction of PRC1 and PRC2 enrichment.

How PRC1 binding results in transcriptional repression is still not perfectly understood. Observation of nucleosome arrays *in vitro* by electronic microscopy shows clearly that PRC1 tightly condenses the chromatin (Francis et al., 2004; Grau et al., 2011). Dense Polycomb compartments can be observed both in *Drosophila* and mammalian nuclei (Alkema et al., 1997; Buchenau, 1998). PCH components of CBX-PRC1 were shown to polymerize, forming dense PRC1 clusters, and indeed CBX-containing complexes appear to compact chromatin more efficiently than in presence of RYBP (Gao et al., 2012; Isono et al., 2013); consequently genes targeted by CBX-PRC1 are more efficiently silenced that those bound only by RYBP-PRC1 (Morey et al., 2013). H2A ubiquitination appears to have a function on itself. RNA polymerase II is present in many bivalent promoters, being recruited by the H3K4me3 activating mark, but often stays bound in an inactive and poised state. Accordingly, H2Aub was shown to specifically block transcriptional elongation (Brookes et al., 2012; Stock et al., 2007). Indeed, ubiquitination seems necessary for the transcriptional repression of developmental genes, but interestingly it is dispensable for silencing the *Hox* clusters, while by contrast *Hox* genes are activated following CBX depletion (Endoh et al., 2012; Morey et al., 2012). As H2Aub is principally deposited by RYBP-PRC1, the efficiency of silencing in presence or absence of ubiquitination probably reflects the differential dependency on the various PRC1 complexes. Indeed, in ES cells, RYBP-PRC1 and CBX-PRC1 can be either found together or by themselves.



Introductory Figure 20. Model for PRC1 and PRC2 recruitement.

A. Targeting of different factors to a TetO array via the tetR binding domain. The data represent protein occupancy by ChIP-qPCR around tetO sequence. Targeting of RYBP-PRC1 deposit H2Aub and recruit PRC1 (left panel), while targeting of CBX-PRC1 fail to do so (mid panel). Targeting of PRC2 recruit CBX-PRC1 but not RYBP-PRC1 and do not result in H2A ubiquitination. From Blackledge et al., 2014

B. Model for a three-step establishment of a Polycomb repressed chromatin compartment.

Finally, Polycomb was never reported so far as a major transposon regulator, at least in cultured cells. At the exception of MLV elements, H3K27me3 is absent from transposon sequences in ES cells. Moreover, TEs do not get reactivated following deletion of PRC1 or PRC2, even though it has to be noted that the combined deletion of PRC1 and PRC2 leads to a strong upregulation of MLV and IAP elements (Leeb et al., 2010).

2.5 Methylation(s) crosstalk

H3K9 methylation and H3K27 methylation mark functionally distinct genomic compartments and are almost never observed together (Mikkelsen et al., 2007; Rao et al., 2014). Whether the two marks are structurally incompatible or whether their physical separation is merely the consequence of distinct recruitment mechanism is unclear. PRC2 catalytic activity is not affected by the presence of an H3K9 methylated histone tail (Schmitges et al., 2011), and G9a methylates H3K27 *in vitro*, indicating that its catalytic activity is not repelled by this mark (Tachibana et al., 2001). Whether the other H3K9-KMTs are affected by H3K27 methylation remains an unexplored question. In fact, a potential mechanist link between H3K9 and H3K27 methylation could be DNA methylation. Indeed, it is commonly admitted that DNA and H3K9 methylation are positively correlated, while DNA and H3K27 methylation are mutually exclusive.

2.5.1 H3K9 methylation recruits DNA methylation

Since DNA methylation is present by default in mammalian genomes, stating that DNA and H3K9 methylation are positively correlated is almost meaningless. In globally highly methylated cells, it simply indicates that H3K9me2/3 is never found in unmethylated loci such as CpG islands. Beside this, DNA methylation is clearly not a good predictor for the presence of H3K9 methylation: DNA methylation is widespread, whereas H3K9 methylation is localized to specific genomic places such as TEs, telomeres, centromeres and pericentric regions. In contrast, correlation between DNA and H3K9 methylation is much stronger in organisms that have mosaic DNA methylation pattern such as *Neuraspora* or *Arabidopsis*. In *Neuraspora*, the link is unidirectional: H3K9me3 directly recruits the DNA methylation machinery and all DNA methylation is lost in absence of H3K9me3 (Tamaru and Selker, 2001). In *Arabidopsis*, the pathways required to establish and maintain CHG and CHH methylation are highly dependent on H3K9me3 and the two marks interact in complex self-

reinforcing loops, whereas maintenance of CG methylation by MET1 is by contrast independent of H3K9me3 (Du et al., 2015).

In mammals, DNA methylation establishment and maintenance can occur independently of H3K9 methylation. Nonetheless, H3K9 methylation functionally attracts the DNA methylation machinery, which is especially relevant during developmental periods where global DNA methylation is constitutively low. In early development, the genome is globally hypomethylated at the blastocyst stage and IAP transposons and imprinting control regions are among the rare sequences that maintain high DNA methylation (Lane et al., 2003; Tremblay et al., 1997). Upon deletion of KAP1 in early embryo, DNA methylation is lost at imprinted loci and IAPs are strongly reactivated, showing that KAP1 is necessary for the DNA methylation of these genomic elements (Messerschmidt et al., 2012; Rowe et al., 2010). On the same line, in ES cells cultured in a medium that promotes global hypomethylation, residual DNA methylation is maintained specifically at IAPs, pericentric major satellite repeats and imprinted loci, where it usually correlates with the presence of H3K9 methylation (Ficz et al., 2013; Habibi et al., 2013). In ES cells and in the early embryo, maintenance of imprinted methylation was shown to rely specifically on the KRAB-ZFP ZFP57, which recognizes methylated binding-motif located in the imprinting control regions (Quenneville et al., 2011, 2012). Finally, transfection in ES cells of reporter genes coupled with known KRAB-ZFP binding motifs results in the rapid silencing of the reporter and *de novo* methylation of its promoter, undoubtedly demonstrating that KAP1 and ESET action are not limited to the protection of DNA methylation, but can also promote its initiation (Rowe et al., 2013).

By contrast with the ADD domain of ATRX, DNMT3a and 3b ADD domains have no particular affinity for H3K9me3 (Iwase et al., 2011). DNMT3a and b are therefore not directly recruited by H3K9me3, but by the different proteins associated with repressive chromatin. For example, DNMT3a and 3b co-immunoprecipitate with ESET (Li et al., 2006a) and KAP1 (Quenneville et al., 2011; Zuo et al., 2012). DNA methylation deposition at pericentric heterochromatin is on the other hand mediated by direct interaction of DNMT3a and 3b with SUV39h and HP1, and DNA methylation indeed diminishes in major satellite repeats in *Suv39h*-dKO ES cells (Fuks, 2003; Lehnertz et al., 2003). Reciprocally, the methylated DNA-binding protein MeCP2 associates with SUV39h, engaging DNA and H3K9 methylation in a self-reinforcing loop (Lunyak et al., 2002). Of note, DNA methylation of centromeric regions is not affected by SUV39h deletion, and DNMT3b is recruited directly to minor satellite repeats by centromeric proteins independently of H3K9 methylation (Gopalakrishnan et al., 2009).

In *Dnmt3a/b* double knockout (*Dnmt3*-dKO) ES cells, global DNA methylation reaches a very low level at late passages, but TEs, such as IAP elements, and imprinted loci still remain hypermethylated, indicating that DNMT1 as well can be recruited by H3K9 methylation (Leung et al., 2014). DNMT1 has no specific affinity for H3K9 methylation by itself, but was shown to bind to G9a (Estève et al., 2006). Moreover, UHRF1 recognizes H3K9me2/3 with its Tudor and PHD domains, allowing DNMT1 recruitment to H3K9 methylated regions (Arita et al., 2012; Rothbart et al., 2013). The recognition of H3K9me2/3 reinforces the maintenance of DNA methylation, and can even direct some *de novo* deposition independently of the presence of hemi-methylated CpGs (Arand et al., 2012; Liu et al., 2013).

G9a can also recruit DNMT3a and 3b, either directly or indirectly through the chromodomain-containing protein MPP8 (Epsztejn-Litman et al., 2008; Kokura et al., 2010); this correlates with a reduction of DNA methylation at transposons and germline genes in *G9a*-KO ES cells (Dong et al., 2008). Interestingly, the recruitment by G9a is independent of its catalytic activity. G9a is especially important for the inactivation and methylation of pluripotent genes (such as *Oct4*) during ES cells differentiation. Furthermore, the transcription factor E2F6 was shown to interact with G9a: as E2F6 mediates the proper methylation of germline genes during early embryogenesis, this indicates that G9a is probably responsible for methylating developmental gene *in vivo* as well (Ogawa et al., 2002; Velasco et al., 2010).

As a whole, the link between DNA methylation and H3K9 methylation appears mostly unidirectional in mammals. H3K9me2/3-mediated DNA methylation appears especially necessary during developmental periods where DNA methylation is either low or dynamic: H3K9me2/3 allows the maintenance of DNA methylation at TEs and imprinted loci during DNA methylation erasure, and promotes *de novo* methylation of developmental genes during differentiation. Importantly, H3K9 methylation is no longer anymore once DNA methylation has been established. Indeed, absence of KAP1, SUV39h or ESET in differentiated cells does not result in DNA methylation defect nor in transcriptional activation of underlying sequences (Bulut-Karslioglu et al., 2014; Matsui et al., 2010; Rowe et al., 2010).

2.5.2 DNA methylation excludes Polycomb from GC-rich regions

In ES cells, 80-90% of PRC1- or PRC2-enriched regions overlap with unmethylated CpG islands and these regions tend to be hyper conserved during evolution (Ku et al., 2008; Tanay et al., 2007). In fact, lack of DNA methylation and high CG content are the only characteristics that have been shown so far to promote H3K27me3 recruitment in mammals. Indeed, the introduction of GC-rich sequences in ES cells is sufficient to recruit PRC2, at the

condition that the region is protected against DNA methylation and devoid of activating motifs (Jermann et al., 2014; Lynch et al., 2011; Mendenhall et al., 2010). However, the restriction of PRC2 complexes to unmethylated CpG islands appears specific to ES cells. In mouse or human differentiated cells, H3K27me3 expands away from CpG islands and forms large domains spanning more than 10% of the genome and with a typical size of 15-40kb (Hawkins et al., 2010; Pauler et al., 2009). These large H3K27me3 domains remain highly methylated, indicating that H3K27me3 and DNA methylation are not always incompatible. Nonetheless, even though H3K27me3 and DNA methylation can be found together in differentiated cells throughout most of the genome, they remain mutually exclusive in CpG islands (Brinkman et al., 2012; Statham et al., 2012), showing that the repulsive action of DNA methylation requires high CpG density.

The exclusion of Polycomb from methylated and CpG-dense regions appears to be an intrinsic feature of PRC2 complexes. In an attempt to mechanistically understand what triggers the interaction of proteins with chromatin, reconstituted nucleosomes were incubated with whole Hela cell extracts and used as a bait to pull-down interacting proteins. In this assay, PRC1 and PRC2 complexes could be recovered when using unmethylated nucleosomes, but the interaction was lost when nucleosomal DNA was methylated beforehand (Bartke et al., 2010). DNA sequences used as baits in these experiments had very high CpG contents; whether CpG-poor and methylated nucleosomes would also block the interaction with PRC complexes was, unfortunately, not investigated. However, other studies reported that PRC2 could methylate H3K27 in DNA methylated nucleosomes *in vitro* or upon artificial tethering of PRC2 to methylated region in ES cells, even with high CpG density, suggesting that DNA methylation alone is not sufficient to prevent PRC2 activity (Cooper et al., 2014). These results are somehow contradictory, but might suggest that PRC2 binding, but not its activity, is repelled by DNA methylation, and that PRC2 complexes act differentially in differentiated and pluripotent cells. The *in vivo* existence of large GC-poor domains with high H3K27 and DNA methylation in differentiated cells, but not ES cells, shows at least that sparsely distributed methylation is not a major impediment for Polycomb activity.

The observation that H3K27me3 can spread in differentiated cells but is constrained to unmethylated CpG islands in ES cells is intriguing and indicates that Polycomb has different property in pluripotent and differentiated cells. In particular, in *Dnmt*-tKO ES cells, H3K27me3 extends from CpG islands and covers large domains, in a pattern reminiscent of the one observed in differentiated cells; this supports the concept that DNA methylation can constrain Polycomb to unmethylated CpG islands in ES cells, but not in differentiated

contexts (Brinkman et al., 2012). Interestingly, promoters marked by H3K27me3 frequently become DNA methylated during ES cell differentiation or carcinogenesis (Mohn et al., 2008; Ohm et al., 2007), and a controversial study showed that PRC2 could physically interact with DNMT1, DNMT3a and DNMT3b, leading to DNA methylation deposition at target promoters in Hela cells (Viré et al., 2006). One possible explanation for this observation is that silencing of these genes is initiated by Polycomb in pluripotent cells, and that Polycomb further recruit DNA methylation to ensure long term silencing in differentiated cells. Interestingly, artificial tethering of PRC2 to a transgene in ES cells leads to the recruitment of Dnmt3a, but do not result in *de novo* methylation (Rush et al., 2009). This shows that interaction of PRC2 and DNMTs is not restricted to differentiated cells, but that ES cells have additional mechanisms to ensure that H3K27me3-enriched CpG islands remain unmethylated.

Several mechanisms probably cooperate to prevent *de novo* methylation of PRC2 targets in ES cells. For example, PRC2 was shown to recruit TET enzymes and promote active demethylation of target sequences (Neri et al., 2013a). The majority of CpG islands marked by H3K27me3 are also enriched for H3K4me3. Since the ADD domain of DNMT3 enzymes is repelled by this mark, the bivalency of H3K27me3 domains could allow protection against *de novo* methylation. Moreover, DNMT3l has also the ability to bind to PRC2, and its presence outcompetes the interaction with DNMT3a and 3b. Since DNMT3l is only expressed in undifferentiated cells, the exclusion of DNMT3a and 3b from PRC2 complexes by DNTM3l ensures that PRC2 targets do not become *de novo* methylated, which is observed in ES cell upon DNMT3l depletion (Neri et al., 2013b). Of note however, the interaction between EZH2 and DNMT3l could not be reproduced in another study (Vlachogiannis et al., 2015).

Enigmatically, the CxxX-containing protein KDM2b was also shown to be required for the protection of H3K27me3-CpG islands. KDM2b is bound to virtually every unmethylated CpG island promoters, and at a subset of them also associate with RYBP-PRC1.1 complexes (Farcas et al., 2012; Wu et al., 2013). Deletion of KDM2b barely affects the enrichment of PRC1 and PRC2 (Blackledge et al., 2014; He et al., 2013), but surprisingly two-third of PRC-bound CpG islands become hypermethylated (Boulard et al., 2015). This result is puzzling, because it suggests that in absence of KDM2b, PRC complexes and DNA methylation can be found together in GC-dense regions, which stands against much of current knowledge.

The observation that Polycomb and DNA methylation are mutually exclusive appears to be the consequence of a subtle equilibrium between several counterbalancing forces: in one

side, high density of unmethylated CpGs is sufficient for recruiting H3K27me3 and PRC complexes are mechanistically excluded from methylated nucleosomes; on the other side, PRC2 can directly recruit the DNA methylation machinery and H3K27me3-rich CpG islands need to be actively protected against *de novo* deposition.

2.5.3 H3K27 and H3K9 methylations

In pluripotent and differentiated cells, H3K27 and H3K9 methylation are almost never found together and their domains, large or small, do not overlap (Hawkins et al., 2010; Mikkelsen et al., 2007). This is intriguing because the two marks do not seem mutually exclusive *per se*. PRC2 was even shown to be preferentially attracted *in vitro* by nucleosomes marked with H3K9me3 (Bartke et al., 2010). However, this interaction is lost if nucleosomes are additionally DNA methylated. Since H3K9me2/3 is always found together with DNA methylation, and H3K27me3 requires unmethylated sequences (at least in ES cells), it is tempting to speculate that the mutual exclusion of H3K27me3 and H3K9me3 in most cell types is caused by DNA methylation. In line with this hypothesis, the rare known sequences where H3K27 and H3K9 methylation are observed together harbor low DNA methylation.

In *Suv39h*-KO ES cells, major satellite repeats lose both H3K9me3 and DNA methylation and interestingly, some cells show concomitantly a strong H3K27me3 enrichment at chromocenters (Lehnertz et al., 2003; Peters et al., 2003). Similarly, in zygotes lacking SUV39h enzymes, H3K9me3 is replaced in pericentric regions by PRC1 (Puschendorf et al., 2008). These observations suggest that PRC complexes can occupy pericentric regions upon loss of H3K9me3. However, in DNA methylation-free ES cells, H3K9me3 is not affected, but H3K27me3 relocalizes nonetheless to major satellite repeats and forms bright foci in chromocenters that overlap with H3K9me3 (Cooper et al., 2014; Marks et al., 2012; Saksouk et al., 2014). PRC2 and RYBP-PRC1, but not CBX-PRC1, are present in pericentric heterochromatin in *Dnmt*-tKO ES cells. This shows that in the absence of DNA methylation, H3K9 and Polycomb can coexist in pericentric regions. Major satellite repeats are AT-rich and are not transcriptionally activated in *Dnmt*-tKO cells, making it unlikely that Polycomb complexes are recruited by RNAs or by a high GC-content. Recruitment of PRC2 was in fact shown to rely on the DNA-binding protein BEND3, which specifically binds to major satellite repeats in absence of DNA methylation (Dai et al., 2013; Saksouk et al., 2014). BEND3 could potentially recognize a motif in major satellite repeats in a methylation-dependent manner.

A second example of co-occurrence of H3K27me3 and H3K9me3 in absence of DNA methylation can be observed in primordial germ cells. At this stage, the genome is globally unmethylated; in this context, H3K27me3 and H3K9me3 are present together in transposons, including L1s and class I and II ERVs (Liu et al., 2014). Interestingly, both marks decrease upon loss of ESET, indicating that their deposition might be functionally linked.

These limited examples suggest that in some specific contexts, H3K27me3 and H3K9me3 could be present together, either at pericentric regions or in transposable elements. At which extent is the absence of DNA methylation a necessary condition, or whether the two marks can be present in the same nucleosome or the same H3 tail remain unaddressed questions. In differentiated cells, H3K27me3 and H3K9me3 form large domains independently of DNA methylation. However, these domains do not overlap, suggesting that at least in differentiated cells, other mechanisms than DNA methylation may restrict the two marks into their respective territory.

In ES cells lacking DNA methylation, H3K9me3 and H3K27me3 do not overlap, but occupy distinct subdomains of pericentric heterochromatin, as determined by cytological tools (Cooper et al., 2014). Therefore, H3K9me3 and H3K27me3 might be antagonistic, and in line with this hypothesis, depletion of SUV39h in *Dnmt*-tKO cells leads to broader enrichment of H3K27me3 in chromocenters. Since PRC2 activity is not blocked by H3K9 methylation (Schmitges et al., 2011), the observed antagonism is probably indirect. In the zygote, it was shown that HP1 proteins, but not H3K9me3, prevent the binding of CBX-PRC1 and PRC2 in pericentric heterochromatin, suggesting a mechanism for the mutual exclusion of the two marks (Tardat et al., 2015). In contrast, PRC1 recruitment and H2Aub do not appear to be antagonized by H3K9 methylation and both H3K9me3 and H2Aub overlap at chromocenters in in *Dnmt*-tKO cells (Cooper et al., 2014).

Overall, H3K27me3 and H3K9me3 appear to be mutually exclusive, but what mechanistically causes this exclusion is far for being understood.

Further reading:

The following reviews were used as starting material to write this chapter.

van Bemmel, 2012: historical perspective
Luger et al., 2012: chromatin core structure
Maison and Almouzni, 2004: HP1 and SUV39
Shinkai and Tachibana, 2011: G9a
Iyengar and Farnham, 2011: KAP1 and ESET
Margueron and Reinberg, 2011; Simon and Kingston, 2013: Polycomb

Chédin, 2011; Schübeler, 2015: DNA methylation
Du et al., 2015; Rose and Klose, 2014: links between histone and DNA methylation

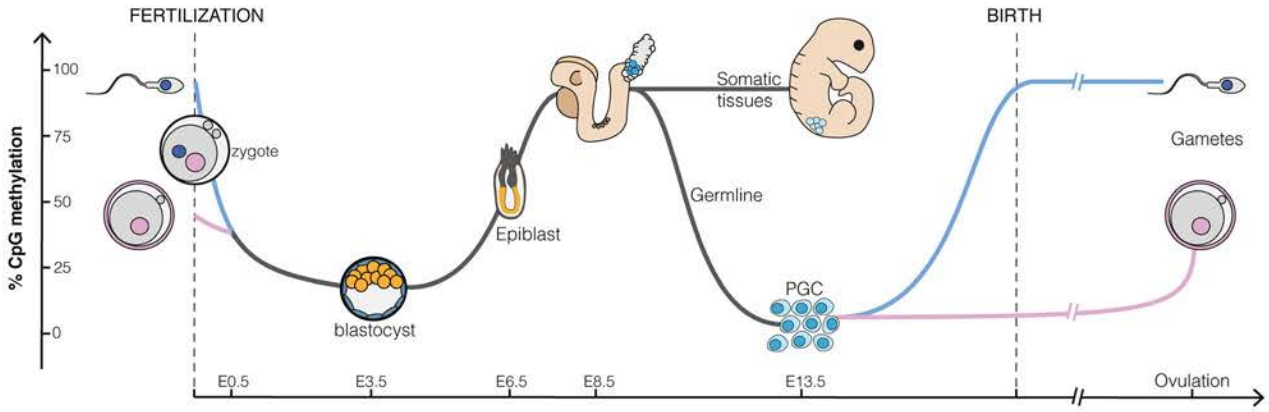
The following studies are cited in figure legends:
Blackledge et al., 2014; Bulut-Karslioglu et al., 2014; Canzio et al., 2011; Castro-Diaz et al., 2014; Ciferri et al., 2012; Du et al., 2015; Elsässer et al., 2015; Gao et al., 2012; Jacobs et al., 2014; Luger et al., 2012; Nishino et al., 2012; Olins and Olins, 2003; Rodriguez-Paredes and Esteller, 2011; Trojer et al., 2011; Whitcomb et al., 2007; Yu et al., 2012.

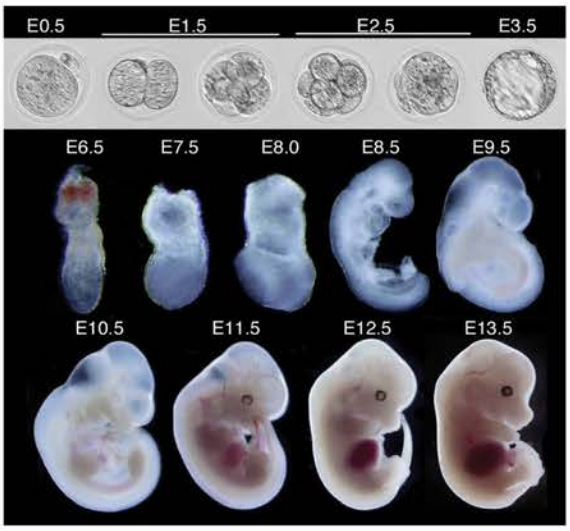
3 REPROGRAMMING

What ultimately defines a cell is not its chromatin landscape, but the ensemble of genes that are expressed at a given moment. Consequently, cell type-specific transcription factors are at the heart of cellular identity and control the gene expression profiles in an intricate balance between antagonistic forces. Cells in culture are usually maintained in a stable state by inhibiting or promoting defined regulatory pathways. Cell fate can also easily be switched *in vitro* from pluripotent or multipotent to more differentiated states by targeting few key signaling pathways, promoting the development towards a given cell fate while repressing alternative possibilities (Graf and Enver, 2009). Conversely, differentiated cells can be reversed artificially to a pluripotent state. John Gurdon showed that the introduction of the nucleus of terminally differentiated intestinal cell into an enucleated *Xenopus* oocyte resulted in the development of normal frogs, proving that the acquired somatic state of the nucleus can return to a totipotent state, under the control of cytoplasmic proteins from the oocyte (Gurdon, 1962; Gurdon et al., 1958). Shinya Yamanaka further showed that reprogramming of somatic cells back to pluripotency only necessitates the expression of four core pluripotency transcription factors (Takahashi and Yamanaka, 2006).

Transcription factors are sufficient by themselves to promote reprogramming, but the process is inefficient and the chromatin landscape of differentiated cells represents a barrier that has to be overcome. Somatic cells have large and static domains of repressive chromatin and comparatively few active regions, whereas pluripotent stem cell are characterized by a very plastic, dynamic and globally permissive chromatin environment (Meshorer et al., 2006; Zhu et al., 2013). During reprogramming, the large repressive domains of differentiated cells need to be released: accordingly, removal of DNA or H3K9 methylation greatly improves the efficiency of somatic cell reprogramming (Chen et al., 2013; Mikkelsen et al., 2008).

During mammalian development, the majority of cell fate transitions are mono-directional, from potent progenitors to more committed cell types. The process is driven by the finely tuned temporal and spatial expression of transcription factors, promoting the progressive compaction of the chromatin landscape. However, during two keys periods of embryonic development, committed cells are reprogrammed and their developmental potency is restored.





Introductory Figure 21. Dynamics of DNA methylation during mammalian development

Two waves of DNA methylation during mammalian development. After fertilization, the paternal genome (blue) is actively demethylated before the first cellular division. After this, both paternal and maternal genome (pink) loses DNA methylation passively until the blastocyst stage (E3.5). Around the time of implantation, the genome becomes *de novo* DNA methylated, reaching its maximum level around E8.5. Primordial germ cells (PGCs) arise around E7.25 and migrate from E8.5 to E13.5 from to allantois to the developing gonads. The second wave of demethylation occurs during this PGC migration in both an active and passive manner. Minimum DNA methylation is attained around E13.5. DNA methylation is then reestablished in a sex-specific manner: Male germ cells immediately regain DNA methylation and pattern are fully established around birth, while non-growing oocyte remain hypomethylated and gain DNA methylation only during the last phases of oocyte maturation.

3.1 DNA methylation reprogramming *in vivo*

In mammalian development, cell fate reprogramming is intimately linked to the loss of DNA methylation. Twice during development, DNA methylation patterns are erased on a global scale before being reestablished (**Introductory figure 21**). The first DNA demethylation event occurs in germ cells, the early fetal progenitors of spermatozoa and oocytes; the second occurs in the zygote just after fertilization and culminates at the blastocyst stage (Kafri et al., 1992; Monk et al., 1987). Resetting DNA methylation is absolutely required for establishing male- and female-specific methylation patterns at imprinting control regions, but it is very unlikely that global reprograming evolved for this function. Indeed, *Xenopus* embryos appear to similarly go through a phase of global hypomethylation in the first hours after fertilization and in their germline, while genomic imprinting is restricted to placental mammals (Stancheva et al., 2002; Venkatarama et al., 2010). DNA methylation reprogramming occurs concomitantly to the restoration of developmental potency, and it is often assumed that hypomethylation is a constitutive feature of pluripotent states. However, hypermethylated ES cells remain pluripotent and more importantly, early zebrafish embryos stay hypermethylated throughout early development. The methylome inherited from the zebrafish father remains even literally untouched, showing that hypomethylation is not a requirement for pluripotency *per se*, at least in this species (Jiang et al., 2013; Potok et al., 2013). Little is known in fact about the biological significance of DNA methylation erasure in mammals, albeit it was proposed to be a way to erase any acquired epimutations, preventing them from being transmitted to the next generation (Heard and Martienssen, 2014; Seisenberger et al., 2013). In direct relevance with the focus of my thesis work, the following paragraphs will review the known mechanisms that govern erasure and reestablishment of DNA methylation in the early mammalian embryo.

3.1.1 DNA methylation erasure

In mouse, primordial germ cells (PGCs) arise from precursor cells in the epiblast at around 7.25 day after fertilization (E7.25) (Ginsburg et al., 1990). First localized at the base of the allantois around E8, PGCs migrate through the hindgut and reach the developing gonads around E11.5 (Anderson et al., 2000; Molyneaux et al., 2001). Epiblast cells have already acquired somatic developmental programs at this stage, and in particular germline genes are methylated and silenced. The erasure of the somatic chromatin landscape appears necessary to turn on the germline developmental program. Interestingly, even though the germline

lineage is a unipotent program that goes from early epiblast progenitors to very specialized gametes, PGCs acquire the characteristics of pluripotent cells, such as the expression of pluripotent transcription factors OCT4 and NANOG. Moreover, i*n vitro*-derived embryonic germ cells are in fact nearly indistinguishable from ES cells and can similarly form potent chimeras, which can give rise to all somatic lineages and functional gametes when injected into embryos (Leitch et al., 2013; Stewart et al., 1994).

Demethylation of PGCs occurs in two steps. From E8 to E10.5, DNA methylation maintenance is impaired by downregulating and excluding DNMT1-UHRF1 from the nucleus, and global levels of CpG methylation drop from around 70% in the E6.5 epiblast to 20-30% at E10.5 (Kobayashi et al., 2013; Seisenberger et al., 2012). While the bulk of the genome is hypomethylated, some specific sequences such as imprinting control regions, germline gene promoters and CpG islands on the X chromosome maintain methylation. This remaining methylation requires a local protection against active demethylation, with both hydroxy-methylation by TET1 and TET2 enzymes and the base excision repair pathway (Hackett et al., 2013; Hajkova et al., 2010). The PGC genome reaches a minimum of 5-7% of CpG methylation at E13.5, and the only sequences that remain significantly methylated at that point are IAP transposable elements and other ERVs (Guibert et al., 2012). Interestingly, H3K9me2 levels are concomitantly globally erased from E8 to E13.5, whereas H3K9me3 remains unaltered and H3K27me3 greatly increases (Hajkova et al., 2008; Seki et al., 2007). Importantly, the global erasure of DNA methylation in PGCs was recently shown to be conserved in humans, indicating that germline reprogramming is a general characteristic of mammalian development, albeit the time scale is very different. The minimum of DNA methylation is reached at around 9-10 weeks of gestation and levels remain low for several weeks (Gkountela et al., 2015; Guo et al., 2015; Tang et al., 2015).

The second wave of DNA demethylation starts in the zygote just after fertilization. The DNA methylation pattern inherited from the sperm and the oocyte are markedly different; moreover, the two parental pronuclei follow distinct paths during the first cellular divisions. The sperm nucleus is highly methylated (~90%) and tightly packed with protamines (Kobayashi et al., 2012; Smith et al., 2012). The sperm genome is actively and rapidly demethylated and loses half of its methylation before the first cellular division (Wang et al., 2014b). The TET3 enzyme is highly enriched on the paternal pronucleus and converts a substantial amount of methylated cytosines to hydroxymethylated cytosines (Gu et al., 2011). How precisely demethylation occurs is still unclear, but active removal of methylated cytosines was shown to occur before and during DNA replication and to involve base excision repair

(Hajkova et al., 2010; Santos et al., 2013). Analysis of *Tet3*-KO zygotes shows that hydroxy-methylation is involved in active demethylation of 30-50% of the genome, but the global increase of methylation in absence of Tet3 is fairly limited (~10%), indicating that passive dilution during DNA replication is probably the main driver of DNA demethylation before the first cellular division (Guo et al., 2014; Peat et al., 2014)

In comparison, the oocyte genome is markedly less methylated (~40%) than the sperm, and the maternal pronucleus is protected against hydroxymethylation by the protein STELLA (Nakamura et al., 2007, 2012). Active demethylation seems to occur nonetheless at some extent in the maternal genome, but global maternal DNA methylation levels remain almost constant during the first cellular division (Guo et al., 2014; Wang et al., 2014b). From the two-cell embryo to the blastocyst stage, the remaining DNA methylation is thought to be lost passively from the two parental genomes, to reach a minimal level of around 20% in the early blastocyst. At that time, most of the genome is hypomethylated, and IAP transposons and imprinting control regions are among the few regions that remain significantly methylated.

3.1.2 *De novo* DNA methylation

With the exception of the female germline, global hypomethylation is a very transient state during mouse development. *De novo* DNA methylation begins in the late blastocyst around E4, reaches ~60% at E6.5 and is complete around E7.5-8.5 (~80%) (Wang et al., 2014b). This rapid gain of methylation correlates with pluripotency loss. For example, injection of ES cells derived from the E2.5 to E4.5 inner cell mass into a blastocyst can form chimaeras and contribute to the offspring, whereas epiblast stem cells (EpiSCs) originating from E5.5 to E7.5 embryo fail to do so (Brons et al., 2007; Tesar et al., 2007).

Precise analysis of *de novo* methylation by RRBS (reduced-representation bisulfite sequencing) between E3.5 and E8.5 shows that 50% of the gain occurs in a single day between E4.5 and E5.5 and then progresses more slowly between E5.5 and E7.5 (Auclair et al., 2014). Around 700 CpG islands are highly methylated at E8.5, around half of which were already methylated in the E4.5 blastocyst. The majority of these CpG islands overlap with exons, preferentially of important developmental genes, and most of the methylated CpG island promoters regulate genes linked to the germline expression program (Auclair et al., 2014; Borgel et al., 2010). The dynamics of *de novo* methylation is probably specific for every gene. For example, methylation of the transcription factor *Elf5* is thought to occur early, as this gene governs the early separation between the epiblast and the trophectoderm lineages (Ng et al., 2008). By contrast, *Oct4* is still expressed in the E6.5 epiblast and its promoter gains

methylation only later during differentiation. *De novo* methylation in the peri-implentation embryo relies mostly on DNMT3b, and absence of DNMT3a results in limited DNA methylation defects (Auclair et al., 2014).

Because of the difficulty to isolate peri-implantation embryos, the methylome of E4.5 and E5.5 stages has not been analyzed with whole-genome base-pair precision level so far. The technique used by Auclair et al., 2014, (RRBS) is biased toward CpG-rich sequences. As a consequence, a precise overview *de novo* methylation dynamics at the level of the genome is still missing, especially at CpG-poor regions such as gene bodies or intergenic sequences. The reason for the different kinetics of DNA methylation is unclear. For example, DNMT3b was shown in cellular systems to methylate preferentially CpG-rich sequences and to recognize H3K36me3 in the body of highly transcribed genes, but whether this is the case *in vivo* as well remains largely unexplored (Baubec et al., 2013; Morselli et al., 2015). Absence of DNMT3l in this period is not critical since *Dnmt3l*-KO embryos develop to term and have no detectable methylation defect; however, some studies have suggested it may have some subtle role both in the embryo and in differentiating ES cells (Arand et al., 2012; Guenatri et al., 2013; Ooi et al., 2010).

By contrast, DNMT3l is absolutely required in the germline and both males and females cannot develop functional gametes in its absence (Bourc'his and Bestor, 2004; Bourc'his et al., 2001). PGCs start to acquire sex-specific features around E12.5, when DNA methylation reaches its minimum level. Male PGCs undergo mitotic arrest shortly after and remain non-diving until they become spermatogonial stem cells a few days after birth, which is around 3-4 post-natal days (P3-4) in the laboratory mouse (Ewen and Koopman, 2010). A day-by-day timeline of *de novo* methylation is missing, but global levels rise from ~7% at E13.5 to ~55% at E16.5, are mostly completed before the beginning of meiosis (~75% at P2.5 or P10), and reach a final level of 90% in the sperm (Kobayashi et al., 2012; Pastor et al., 2014). Conditional knockout of DNMT3b has no consequences, whereas absence of DNMT3a or DNMT3l leads to dramatic TE activation, meiosis arrest and germ cell apoptosis (Bourc'his and Bestor, 2004; Kaneda et al., 2004). Consequently, DNMT3a and DNMT3l are assumed to be the main responsible for *de novo* methylation in the male germline, but a comprehensive analysis of the respective contribution of the three enzymes during that period is missing.

Interestingly, *de novo* DNA methylation of a subset of sequences relies on an original pathway involving 25-30nt-long small RNAs and specialized Argonaute proteins (piRNAs and PIWI proteins, respectively) in the fetal male germline. Most fetal piRNAs are complementary to TE sequences and the piRNA pathway appears to be mostly dedicated to transposon

control, at least during this developmental window. Shortly, piRNAs originating from long single-stranded RNA precursors are loaded into PIWI proteins in the cytoplasm and repress TEs at the post-transcriptional level by degrading TE mRNAs, using especially MILI endonuclease activity (De Fazio et al., 2011). A subset of PIWI-piRNA complexes relocalizes to the nucleus and is thought to induce *de novo* methylation at complementary genomic targets by a yet unknown mechanism (Aravin et al., 2008; Suh and Blelloch, 2011). Strikingly, *Dnmt3l, Mili* or *Miwi2* (two of the three PIWI proteins) knockout mouse have similar phenotypes, with strong upregulation of young LINEs and ERVs and a meiotic catastrophe (Aravin et al., 2007; Bourc'his and Bestor, 2004). In fact, even though piRNA-mediated DNA methylation represents a very small proportion of the global methylation, piRNAs appear to target and silence specifically ~17,000 active transposons (Molaro et al., 2014). In absence of piRNAs or DNA methylation, TEs acquire some active chromatin marks during meiosis, causing the relocalization of meiotic homologous recombination hotspots to TEs and major meiotic defects (Zamudio et al., 2015).

At E12.5, female PGCs take a different developmental trajectory: they enter the prophase of meiosis I at E13.5 and arrest at the end of metaphase I around the time of birth. Oocytes remain blocked at this stage until meiosis resumes at puberty (Ewen and Koopman, 2010). Oocytes remain unmethylated during that whole period; *de novo* DNA methylation occurs only during oocyte maturation, reaching a final level of around 40% (Kobayashi et al., 2012; Smallwood et al., 2011). Interestingly in oocytes, high levels of genomic methylation are highly correlated with transcription, and most of the DNA methylation is observed in the body of active genes. As a consequence and contrary to the sperm where most of the genome is methylated except CpG islands, intergenic regions are globally hypomethylated in oocytes. On the other hand, CpG islands localized in transcriptional units are hypermethylated, which is important for the setting of maternal genomic imprinting. DNA methylation is totally abolished in absence of either DNMT3a or DNMT3l. These methylation-free mutant oocytes develop to term and can be successfully fertilized, but the resulting embryos die of imprinting defects at mid-gestation (Bourc'his et al., 2001). In contrast, mutations in DNMT3b or DNMT1 have no effect (Shirane et al., 2013).

While DNMT3b-dependent deposition of DNA methylation is especially targeted to the body of active genes in oocytes, the situation is different in ES cells where DNMT3b is significantly enriched in gene bodies (Baubec et al., 2015; Smallwood et al., 2011). This would suggest that DNMT3a and DNMT3b act differently depending on the context. Indeed, why embryonic methylation relies on DNMT3b and germline methylation on DNMT3a and

DNMT3l remains a question largely unanswered. It is interesting to note that in both male ad female germlines, *de novo* DNA methylation occurs in non-replicative cells, whereas in the embryo cells are actively dividing. It would be interesting to analyze, in cell culture systems for example, whether DNMT3a and DNMT3b have similar efficiency in dividing and non-dividing cells.

3.2 Reprogramming in ES cells

In vivo, events of DNA methylation reprogramming occur in developmental stages that are difficult to collect and that represent entities with a very limited number of cells. As a consequence, analyzing chromatin modifications and transcriptional changes during these periods is technically challenging. Technical innovations that allow performing ChIP, RNA sequencing (RNA-seq) or whole genome DNA methylation mapping in a reduced number of cells or even in single cells have only started to emerge recently (Brind'Amour et al., 2015; Smallwood et al., 2014). These innovations will be instrumental for characterizing how chromatin and transcriptional landscapes are dynamically regulated during critical developmental periods.

Cultured ES cells have been very invaluable tools for developmental studies. Still, cells in culture are artificially maintained in a steady state and consequently do not represent an appropriate model to reproduce the dynamic events inherent to embryonic development. ES cells can be easily differentiated, but since DNA methylation is already high when ES cells are grown in classical serum-based conditions, even differentiation is not a good system to recapitulate genome-wide *de novo* methylation. In the last years, the development of alternative culture conditions in ES cells has offered the possibility to drastically modify the chromatin and DNA methylation landscapes.

ES cells are derived from the inner cell mass of the blastocyst. They can proliferate indefinitely *in vitro* without losing their identity and keep the potential to differentiate into any embryonic cell lineages (but usually not into extra-embryonic lineages). Mouse ES cells were initially established on feeder cells, usually immortalized embryonic fibroblasts, and cultured in presence of fetal bovine serum (Evans and Kaufman, 1981; Martin, 1981). Feeders can be removed when the cells are grown in presence of leukemia inhibitory factor (LIF), a small cell signaling protein that is otherwise provided by the feeders (Smith et al., 1988; Williams et al., 1988). "Serum" cells naturally express FGF4, a signaling protein that binds to FGF receptors on the cellular membrane and can induce differentiation by activating the ERK/MAP kinase

pathway (Kunath et al., 2007). LIF and BMP4, a signaling protein present in the serum, counteract FGF4-induced differentiation by activating alternative intracellular signaling pathways. In particular, downstream transcription factors localize on the promoters of core pluripotent factors such as NANOG, OCT4 and SOX2 and activate pluripotency autoregulatory networks (Hirai et al., 2011). Of note, LIF blocks endodermal and mesodermal differentiation, whereas BMP4 suppresses neural differentiation (Ying et al., 2003).

Importantly, FGF4 is expressed in serum-grown ES cells and the pluripotent pathways induced by LIF and BMP4 do not prevent the expression of the FGF/ERK/MAP kinase pathway, but act downstream to block differentiation (Ying et al., 2008). As a consequence, serum-grown ES cells express both pluripotent and early differentiation genes. They are thought to represent a metastable state primed for differentiation and express many genes linked to developmental processes, such as mesoderm and ectoderm development (Marks et al., 2012). In particular, DNMT3a, 3b and 3l are markers of early differentiation and are highly expressed in serum conditions, which probably explains the hypermethylation of serum-grown ES cells. These cells are also highly heterogeneous in terms of morphology and gene expression. Key pluripotency genes such as *Nanog*, *Rex1* or *Stella* show a high intercellular heterogeneity in expression levels, which impacts on the developmental potential of individual cells (Hayashi et al., 2008; Kalmar et al., 2009; Toyooka et al., 2008).

Disruption of the ERK/MAP kinase by small-molecule kinase inhibitors enables the derivation of ES cells in minimal, serum-free media. The molecule PD0325901 inhibits the MEP protein and blocks the MAP kinase pathway, preventing spontaneous differentiation (Ying et al., 2008). Addition of CHIRON99021 inhibits the GSK3 protein and stimulates the WNT pathway, enhancing ES cells self-renewal and allowing robust cell propagation. Together, the two inhibitors (or 2i) are sufficient for the maintenance of pluripotency in the absence of serum and LIF. There are around 3,500 genes that are differentially expressed between serum and 2i and the two culture conditions are thought to represent two markedly different states of pluripotency (Marks et al., 2012). ES cells in 2i appear closer to early blastocyst cells, whereas serum cells are thought to represent later stages. In fact, 2i-grown ES cells are postulated to represent the "ground state" of pluripotency, defined by Austin Smith as "a basal proliferative state that is free of epigenetic restriction and has minimal requirements for extrinsic stimuli" (Wray et al., 2010). Cells in the ground state are very homogeneous in terms of cell morphology and gene expression and do not express developmental genes. In particular, cells in 2i express at a high level the transcriptional

repressor PRDM14, which binds to the promoters of DNMT3a, 3b and 3l and prevent their expression (Ficz et al., 2013; Leitch et al., 2013)

As a result and in contrast to serum-grown ES cells that are hypermethylated, genomic methylation in 2i is severely reduced, with a global level of CpG methylation around 25-30% (Ficz et al., 2013; Habibi et al., 2013; Leitch et al., 2013). The only sequences that retain relatively high levels of DNA methylation are young ERVs, such as IAPs, and imprinted control regions. Interestingly, the methylome of serum and 2i ES cells are inter-convertible, and switching between 2i and serum media promotes gain or loss of DNA methylation. The dynamics of demethylation from serum to 2i is rather slow, taking 2-3 weeks to go from a hypermethylated to a hypomethylated genome. By comparison, *de novo* methylation upon transition from 2i to serum occurs much faster, being completed in only a few days. Interestingly, 2i-induced loss of DNA methylation is accompanied by a reconfiguration of the chromatin landscape (Marks et al., 2012). In particular, H3K27me3 relocates from promoters to satellite repeats. Consequently around two thirds of the H3K27me3/H3K4me3 bivalent domains at promoters disappear during 2i conversion.

In addition, it was recently shown that the addition of vitamin C in the culture medium can promote further demethylation (Blaschke et al., 2013). Vitamin C enhances the catalytic activity of TET enzymes and leads to an increase of cytosine hydroxy-methylation, which is followed by rapid and active loss of cytosine methylation. So far, the effect of vitamin C on global DNA methylation levels has only been analyzed in a short time scale, either after 12 or 72 hours of treatment; the long-term effects of vitamin C remain to be explored.

3.3 Transposable elements during reprogramming

When McClintock discovered TEs, she referred to them as "controlling elements" because of their capacity to affect the expression of neighboring protein-coding genes. Newly integrated elements affect expression in a random manner, depending on their integration site. Interestingly, in some species such as plants and insect, TE mobility is greatly increased upon stressful conditions. It is thought that this increased mobility could serve to create genetic variability, in order to adapt to the stressful event (Capy et al., 2000). Moreover, there is increasing evidence that long-term TE residents play an important regulatory role on the genome. In mammals in particular, the turnover of TEs in the genome is low: TEs fixed in a population remain there for dozen of millions of years, providing enough evolutionary time for old TEs to evolve important regulatory function.

In differentiated tissues, TEs are silenced by DNA methylation and do not appear to significantly contribute to genome functions. However, in early phases of embryonic development, TEs become expressed in a highly controlled manner and contribute to genome regulation. From the oocyte to the blastocyst stages, TEs represent between 15 and 20% of total capped RNAs, with both LINEs and ERVs being highly expressed (Fadloun et al., 2013). Virus-like particles have long been detected in mouse embryos, originating either from IAP, MusD or MERVL elements (Kuff and Lueders, 1988; Ribet et al., 2008). In human, virus-like particles originating from young HERVK(HML-2) elements seem even to have acquire a functional role by boosting the antiviral defense of early embryos (Grow et al., 2015).

Interestingly, TEs that contribute the most to embryonic development are class III ERVs (MalR and MERVL), which incidentally are also the oldest ERVs of the mammalian genomes. For example, MalR expression is particularly high in the unfertilized oocyte, where these elements represent 13% of the transcriptome (Peaston et al., 2004). But the most striking and best-studied case of influence of an ERV on embryonic genome regulation is provided by MERVL. MERVL is one of the earliest transcripts to be expressed at the onset of zygotic genomic activation: they can be already detected eight hours after fertilization, they are upregulated 300-fold between the oocyte and the two-cell stage embryo, and decrease rapidly during the following divisions (Kigami, 2002; Macfarlan et al., 2012). Moreover, MERVL activation impacts on the expression of around 300 genes by the formation of chimeric transcripts between MERVL LTR and exons. Most of these genes are specific to the two-cell stage, suggesting that MERVL activation is an important early activator of embryonic development. Importantly, MERVL expression is a hallmark of totipotency. In serum- based culture, a small proportion of ES cells highly express MERVLs and genes specific to the two-cell stage and almost all cells fluctuate in and out of this two-cell-like state (Macfarlan et al., 2012). Akin to *in vivo* two-cell embryos, this subpopulation does not express the pluripotency proteins OCT4, NANOG and SOX2, and are rather totipotent cells that can contribute to both embryonic and extra-embryonic lineages. This can be explained by the fact that MERVL elements drive the expression of the *Gata4* and *Tead4* genes, whose product is critical for the specification of early extra-embryonic lineages.

By contrast, IAP expression peaks at the blastocyst stage and is rapidly silenced afterwards (Peaston et al., 2004), but little is known as to whether IAPs contribute to transcriptional regulation in a manner similar to MERVLs. In ES cells, numerous genes are transcribed from alternative promoters located in TEs, either LINEs or ERVs, and many new chimeric transcripts are formed upon deletion of ESET or other transposon repressors

(Karimi et al., 2011b). In fact, by rewiring gene-regulatory networks, the use of alternative promoters located in TEs may be determinant for pluripotent states. In both mice and humans, between 5 and 20% of NANOG and OCT4 binding sites are located in TEs, especially class I in humans, and class II in mice (Kunarso et al., 2010). During events of cellular reprogramming, these binding-sites probably serve as regulatory units that contribute to define the pluripotent expression program. For example, naïve subpopulation of human ES cells are characterized by elevated levels of a primate-specific ERV that provides functional binding sites for pluripotency transcription factors and drive the expression of specific alternative transcripts (Wang et al., 2014a).

Overall, TEs appear to play an important role during mammalian development. They have been coopted to fulfill specific regulatory functions that probably contribute to define pluripotency in the early embryo and in the germline.

Further reading:

The following reviews were used as starting material to write this chapter.

Seisenberger et al., 2013; Lee et al., 2014: DNA methylation reprogramming *in vivo*

Marks and Stunnenberg, 2014: Serum- and 2i-grown ES cells

Gifford et al., 2013: Transposons in development

RESULTS

TEs provide important genetic regulatory substrates during early mammalian development. Reprogramming events uncover regulatory units in TEs, which are usually not accessible in differentiated tissues and which define the very specific transcriptional and chromatin landscape of pluripotency. However, alleviating TE silencing during these periods (gametogenesis and early embryogenesis) can have double-edged consequences, especially because any *de novo* insertion of TEs would be transmitted to the following generations and affect the fitness of the species. Mammals have evolved various complementary mechanisms, which control TE expression specifically during reprogramming events. The different pathways that collaborate to control TEs in pluripotent cells have received considerable attention in the field and have been reviewed at length in the introduction. However, most conclusions have been drawn from the study of undifferentiated ES cells, which are in a static state for unlimited periods of time. By contrast, embryonic development is very dynamic, with regulatory patterns changing at almost every cell division. A critical area of investigation is henceforth to study TE regulation in the context of dynamically evolving chromatin and transcriptional landscape.

During my PhD, I precisely attempted to mimic in ES cells some of the dynamic changes that normally occur during embryonic development. In particular, our objective was to develop cellular models where DNA methylation could be turned on and off in a controlled and rapid manner. Such a system would allow in depth analysis of how the genome copes in face of a rapid DNA methylation loss and would be extremely valuable for understanding the interplay between DNA methylation, histone modifications and transcription.

During my first year of PhD, my first attempt was to develop an inducible and reversible system of DNA methylation switch in ES cells. Using tetracycline-responsive promoters as a way to control transcriptional activation of various transgenes (Gossen and Bujard, 1992), I relied on two complementary strategies: *1)* an inducible RNA interference approach against the three active DNA methyltransferases (*Dnmt1*, *Dnmt3a*, *Dnmt3b*) in WT ES cells, using shRNAs combined with the Tet-ON system (Wiederschain et al., 2009); *2)* a transgenic gene approach in *Dnmt*-tKO ES cells, through the re-introduction of Tet-ON-controlled cDNA copies of Dnmt1, Dnmt3a and Dnmt3b (**Figure 0.1A**). I designed, validated and cloned arrays of shRNAs for simultaneous knockdown of the three *Dnmts*, using Zinc Finger Nucleases (ZFN) for a targeted integration of the constructs at the *Rosa26* locus (**Figure 0.1B,C** and Perez-Pinera et al., 2012). However unfortunately, even after integration of arrays of up to 24 shRNAs, with sometime six or seven different shRNA sequences, global DNA methylation could never be reduced below 30% (**Figure 0.1D, E** and **F**). Moreover, the

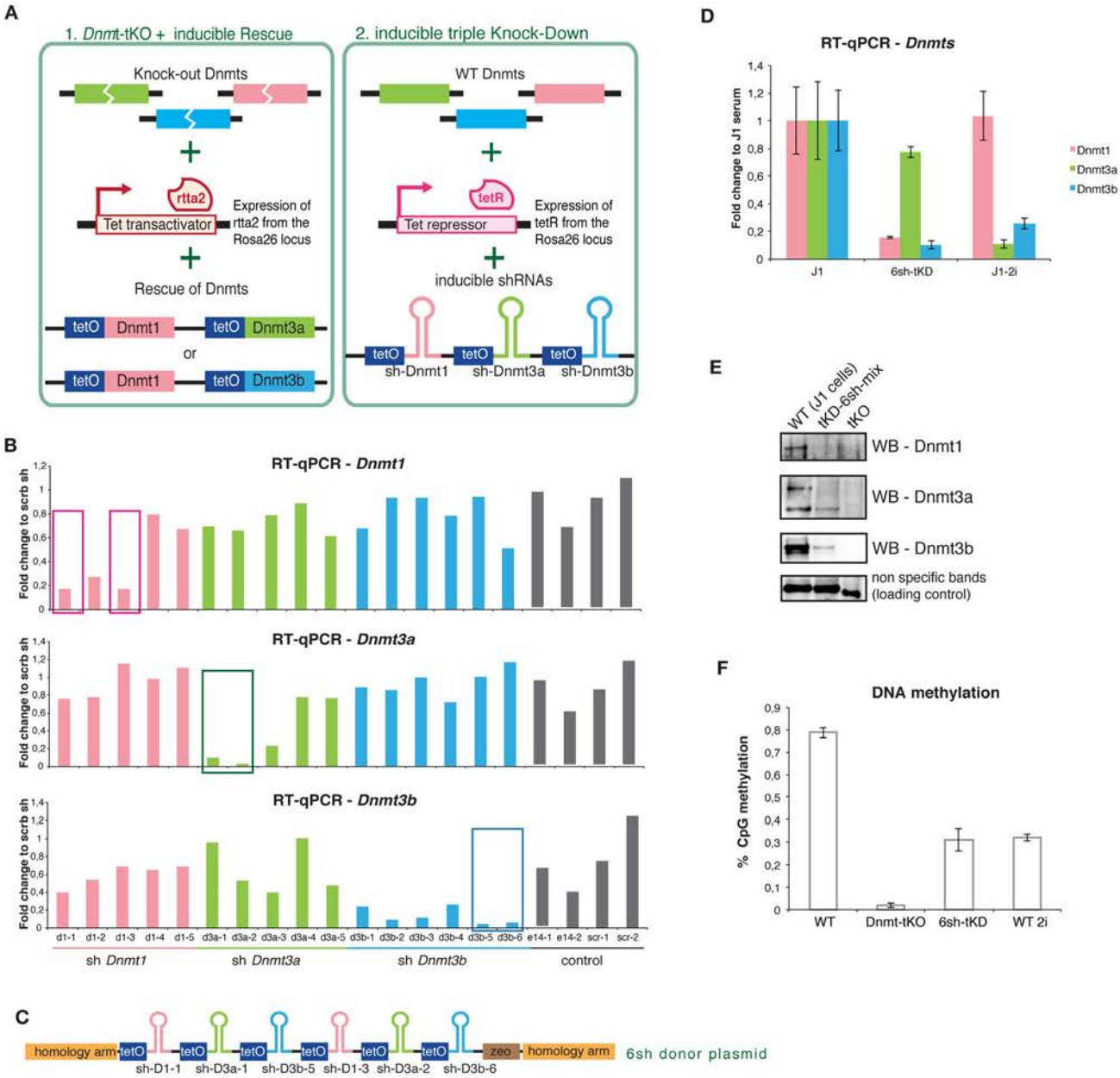


Figure 0.1.

A. Two strategies to develop an inducible and reversible cellular system of DNA methylation switch. 1.Tet-ON-controlled cDNA copies of *Dnmt1*, *Dnmt3a* and *Dnmt3b* are introduced into *Dnmt1*, *Dnmt3a* and *Dnmt3b* Knock-Out ES cell. 2. Tet-ON-controlled shRNAs against *Dnmt1*, *Dnmt3a* and *Dnmt3b* are introduced into WT ES cells. In both case, rtta2 and tetR cDNAs are integrated at the Rosa26 locus using Zinc Finger Nuclease. The tetO sequence allows the control of the expression by rtta2 or tetR, depending of the presence of Doxycycline

B. Test of 16 different shRNAs against either *Dnmt1*, *Dnmt3a* or *Dnmt3b* by RT-qPCR. The pink, green and blue boxes represent the 6 shRNAs that were selected.

C. Donor plasmid containing the 6 selected shRNAs, a zeocin resistance cassette and two homology arms for the *Rosa26* locus.

D. RT-qPCR for the *Dnmts* in WT ES cells grown either in serum or 2i, and in 6sh-tKD grown in serum. Data were normalised using *Rrm2* and *gapdh*, and are presented as the ratio compared to WT ES cells grown in serum. Error bar represent standart error of the mean (SEM) between technical replicates.

E. Western Blot against Dnmt1, Dnmt3a and Dnmt3b in WT, triple knock-down (tKD) and triple knock-out (tKO) ES cells grown in serum.

F. LUMA assay to determine global level of DNA methylation in WT, tKO, tKD ES cells grown in serum, and in WT ES cells grown in 2i conditions. Error bar represent SEM between technical replicates.

design of an inducible *Dnmt* rescue system in *Dnmt*-tKO ES cells proved to be a technical nightmare, principally because *Dnmt*-tKO cells contain in their genome most of the usable antibiotic resistant markers. After a year of fanatical cloning, these two complementary approaches were finally dropped.

Two technological revolutions then allowed my PhD work to go around the wall it had met. The first one is the worldwide revelation that CRISPR/Cas9 editing could be used to manipulate virtually every biological systems, suddenly transforming ES cells into a very powerful genetic model (Cong et al., 2013; Jinek et al., 2012). With CRISPR/Cas9, I was able to knockout seven different genes in less than two years[18]: this allowed me to rapidly and powerfully test hypotheses that would otherwise have remained unanswered. The second revolution has not such a broad impact, and does not even deserve this term outside this manuscript. However, the publication that vitamin C enhanced TET enzyme activity and promoted active demethylation in ES cells was really the starting point of the work presented here (Blaschke et al., 2013).

ES cultured in 2i conditions keep significant DNA methylation levels (~30% of CpG methylation) and the transition between serum- and 2i-based media necessitates several weeks to attain such methylation level (Ficz et al., 2013; Habibi et al., 2013). By contrast, we quickly realized in the lab that conjunction of 2i culture system with vitamin C allowed reaching unprecedented levels of hypomethylation in a short period of time. By switching culture conditions from serum to 2i+vitamin C, I could promote fast and quasi-total demethylation of the ES cell genome, providing clues as to how to answer the main question of my PhD work:

How do TEs react to the rapid loss of DNA methylation, and how the genome adapts to the removal of this essential protective barrier?

By carefully following transposon behavior during global DNA demethylation, I was able to unravel the mechanisms that ensure transposon regulation when DNA methylation is lost. Our findings shed new lights about the way the genome can adapt to DNA methylation reprogramming, with the particularly exciting evidence of the unsuspected role of Polycomb proteins in transposon regulation. The results presented in the following section take the form of a manuscript, which is currently under review at *eLife*, with a few cosmetic differences necessary to fit the general style of this thesis.

[18] So far

In a second time, we realized in the lab that the hypomethylated genome of ES cells grown in 2i+vitamin represented an ideal starting point to study the dynamics of global *de novo* methylation. Since they are already hypermethylated, serum-grown ES cells represent an inadequate system to analyze DNA methylation dynamically. In contrast, by differentiating 2i+vitC-converted ES cells, we could reproduce the global *de novo* methylation events that normally occur in the germline and early embryos. I used this system to characterize genome-wide the contribution of DNMT3l in embryonic *de novo* methylation. Moreover, by comparing the effect of *Dnmt3l* deficiency in differentiated ES cells, *in vivo* post-implantation embryos and germ cells, I attempted to elucidate the developmental requirements of DNMT3l. The results that I present in this thesis are still on-going work and represent a first version of a manuscript that we plan to submit in the following months.

Of important note, massive amounts of genomic data were generated during my PhD, and all of them were analyzed in close collaboration with Aurélie Teissandier, the bioinformatician of our team.

1 An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in ES cells

1.1 Authors and affiliations

Marius Walter[1,2], Aurélie Teissandier[1,2,3], Raquel Pérez Palacios[1], Déborah Bourc'his[1*]

[1] Institut Curie, Dpt of Genetics and Developmental Biology, CNRS UMR3215, INSERM U934, Paris, France
[2] UPMC, University Paris 06, Paris, France
[3] Institut Curie, Inserm U900, Mines Paris Tech, Paris, France

1.2 Abstract

DNA methylation is extensively remodeled during mammalian gametogenesis and embryogenesis. Most transposons become hypomethylated, raising the question of their regulation in the absence of DNA methylation. To reproduce a rapid and extensive demethylation, we subjected mouse ES cells to chemically defined hypomethylating culture conditions. Surprisingly, we observed two phases of transposon regulation. After an initial burst of de-repression, various transposon families were efficiently re-silenced, showing that DNA methylation is involved in transposon repression in ES cells, but that alternative mechanisms can compensate for its loss. This was accompanied by a reconfiguration of the repressive chromatin landscape: while H3K9me3 was stable, H3K9me2 globally disappeared and H3K27me3 accumulated at transposons. Interestingly, we observed that H3K9me3 and H3K27me3 occupy different transposon families or different transposon territories within the same family, defining three functional categories of adaptive chromatin responses to DNA methylation loss. Our work highlights that H3K9me3 and H3K27me3 chromatin pathways can secure the control of a large spectrum of transposons in periods of intense DNA methylation change, ensuring longstanding genome stability.

1.3 Introduction

Millions of transposable elements reside in mammalian genomes, far surpassing in number the approximately 25,000 protein-coding genes (Lander et al., 2001). Most of these

elements are retrotransposons, which utilize an RNA intermediate to duplicate and mobilize. Through their activity or their mere presence, transposons can be both beneficial for the evolution of the host genome and deleterious for its integrity. They can modify gene functions through insertional mutagenesis, influence gene transcriptional outputs by acting as promoters or enhancers or induce chromosomal rearrangements through non-allelic recombination (Goodier and Kazazian, 2008). Accordingly, erratic transposon-related events have been linked to congenital diseases, cancer and infertility (Kaer and Speek, 2013).

Successive waves of transposon expansion and decline have shaped mammalian genomes over evolution, leading to a current occupancy rate of approximately half of the genomic space. Reflecting their various evolutionary origin and multiplication success, resident elements are greatly diverse in structures, numbers and functional properties, which define discrete families of transposons. Long Terminal Repeat (LTR) sequences characterize endogenous retroviruses (ERVs, 12% of the mouse genome), which can be further subdivided into three families (ERV1, ERVK and ERVL), according to the infectious retroviruses they derive from (Stocking and Kozak, 2008). Non-LTR elements comprise Long and Short INterspersed Elements (LINEs and SINEs, 20% and 8% of the genome, respectively), and also consist of specific sub-families (Babushok et al., 2007). The majority of transposons have accumulated nullifying mutations and truncations, but around 1-2% of LINEs and ERVs have intact sequences that embed the protein coding information necessary for their mobilization. Notably, ERVK elements show the greatest level of activity, which causes at least 10% of spontaneous mutations in laboratory mice (Maksakova et al., 2006).

To minimize their impact on genome fitness, multiple layers of control antagonize transposons at different steps of their life cycle (Zamudio and Bourc'his, 2010). Notably, restraining mechanisms can differ between cell types. In somatic cells and in the male differentiating germline, DNA methylation is the main transcriptional suppressor of LTR and non-LTR transposons. In these contexts, transposable elements are densely methylated (Rollins et al., 2006; Smith et al., 2012) and DNA hypomethylation leads to their de-repression (Bourc'his and Bestor, 2004; Walsh et al., 1998). In contrast, the early germline and the early embryo manage to globally control their transposon burden without DNA methylation. These cells naturally undergo genome-wide loss of DNA methylation, likely as part of the acquisition of a pluripotent, flexible state (Seisenberger et al., 2013). Moreover, genetic studies have demonstrated that mouse embryonic stem (ES) cells can use DNA methylation-independent mechanisms to silence transposons: knocking-out the three active

DNA methyltransferases (*Dnmt*-tKO) does not yield significant de-repression of transposons, except Intracisternal A Particle (IAP) elements (Karimi et al., 2011b; Matsui et al., 2010)

In fact, transposon control in ES cells seems to rely primarily on post-translational histone methylation, notably at lysine 9 of histone H3 (H3K9). H3K9 dimethylation (H3K9me2), which is deposited by the G9a and GLP lysine methyltransferases, directly and specifically represses class L ERVs (Maksakova et al., 2013). H3K9 trimethylation (H3K9me3) can be catalyzed by the SETDB1 (also known as ESET) or the SUV39H enzymes. The SUV39H system targets H3K9me3 at evolutionary young LTR and non-LTR transposons, but *Suvar39h* mutant ES cells up-regulate LINE1 elements only (Bulut-Karslioglu et al., 2014). In parallel, SETDB1, together with its associated co-repressor, the Krüppel-associated box domain (KRAB)-Associated Protein 1 (KAP1, also known as TRIM28), mainly control H3K9me3-dependent suppression of ERVK transposons- a family to which IAP elements belong (Karimi et al., 2011b; Matsui et al., 2010; Rowe et al., 2010). KAP1 is recruited to specific genomic sites via direct interactions with KRAB-zinc finger proteins (Friedman et al., 1996), which are a large family of DNA binding factors that co-evolved with ERVs (Emerson and Thomas, 2009). Therefore, different H3K9 methylation-based mechanisms are utilized to silence different transposons families in ES cells. In contrast, the repressive spectrum of polycomb-mediated H3 lysine 27 trimethylation (H3K27me3) is limited: only Murine Leukemia Virus (MuLV) elements are reactivated upon H3K27me3 deficiency (Leeb et al., 2010).

However, the prevailing view that H3K9 methylation acts as the main transposon controller in ES cells may be biased by two confounding factors. First, conclusions are based on analyses of chromatin modifier mutants, which still harbor high DNA methylation levels. Second, proper transposon repression in *Dnmt*-tKO ES cells may reflect a long-term adaptation to a DNA methylation-free state rather than a lack of significant role of DNA methylation *per se*. In fact, how the ES cell genome transitions from a DNA methylation-dependent to -independent mode of transposon control has never been investigated.

To study the dynamics of transposon regulation upon DNA methylation loss, we modulated the ES cell methylome by using interconvertible culture systems, which do not modify pluripotency potential. ES cells grown in standard serum-based conditions have heavily methylated genomes (~75% of CpG methylation)(Stadler et al., 2011), which is linked to the expression of *de novo* DNA methyltransferases. ES cells grown in presence of two small kinase inhibitors (2i) down-regulate these enzymes, and have reduced DNA methylation levels (Leitch et al., 2013; Ying et al., 2008). Upon transfer from serum to 2i medium, demethylation

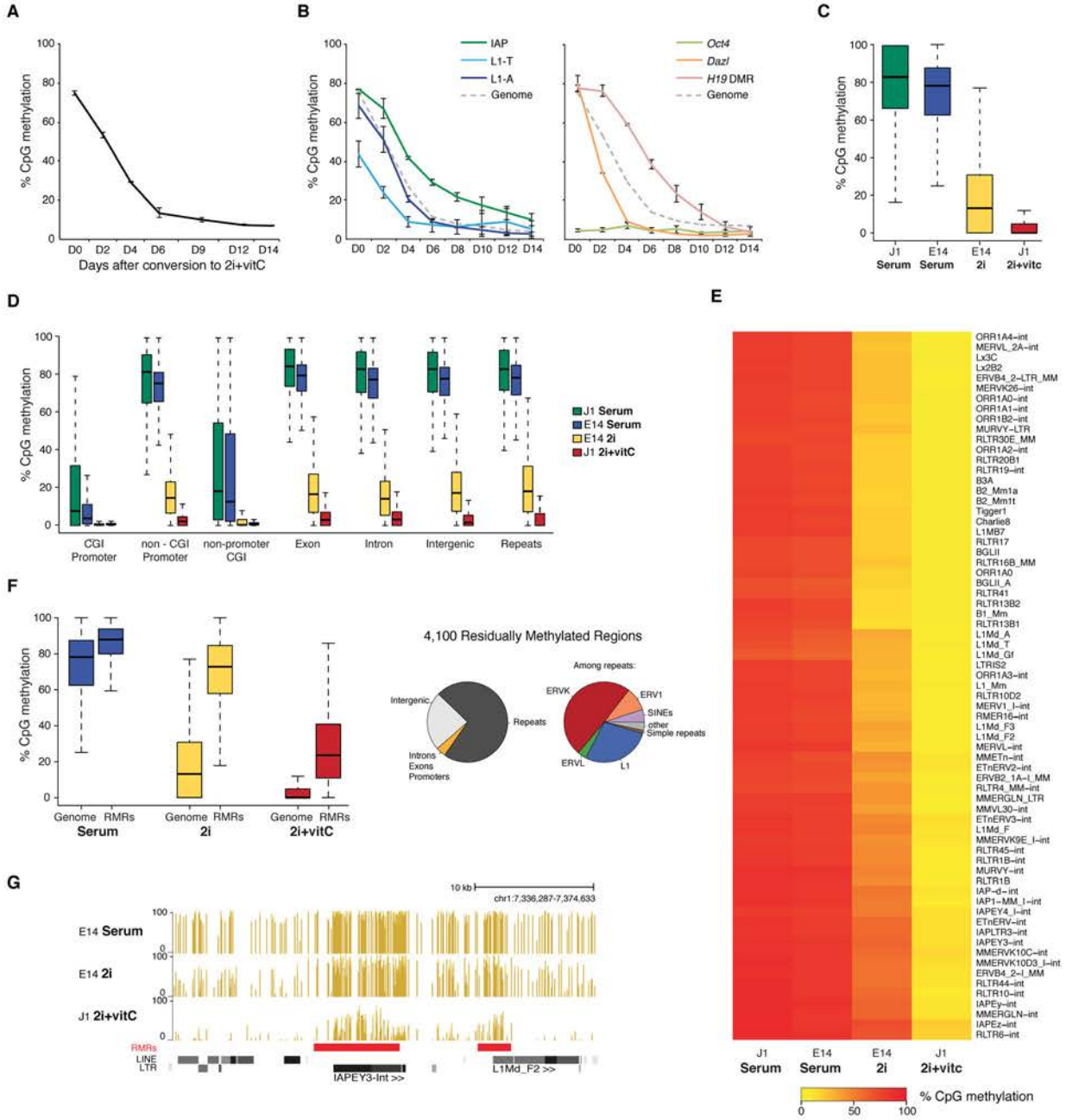


Figure 1.1. Kinetics and extent of DNA methylation loss in ES cells upon serum to 2i+vitC conversion

A. Time course of global CpG methylation loss measured by LUMA over 14 days (D0 to D14) of conversion from serum to 2i+vitC. Data represent mean and Standard Error of the Mean (SEM) between two biological replicates.

B. Sequence-specific CpG methylation level measured by bisulfite pyrosequencing. Data represent mean ± SEM between two biological replicates.

C. Tukey boxplot representation of genome-wide CpG methylation content as measured by WGBS in different culture conditions. Datasets of J1 Serum, E14 Serum and E14 2i were obtained from previous studies (Habibi et al., 2013; Seisenberger et al., 2012)

D. CpG methylation distribution over different genomic compartments by WGBS.

E. Heatmap and hierarchical clustering of average CpG methylation over 69 transposon families as measured by WGBS.

F. Left panel: Tukey boxplot representation of CpG methylation content in Residually Methylated Regions (RMRs) (n=4,100) compared to the whole genome in various culture conditions. Right panel: pie chart distribution of 2i+vitC RMRs in different genomic compartments (left) and among repeats (right).

G. Example of WGBS profile of a genomic region containing two 2i+vitC RMRs mapping to an IAPEY and a L1 elements. Bars represent the methylation percentage of individual CpG sites, between 0 (unmethylated) and 100% (fully methylated). Location of LINE and LTR transposons (RepeatMasker) are displayed below; the RMRs are highlighted in red.

occurs with a slow kinetics: several weeks are required to reach 20-30% of CpG methylation. Notably, imprinted genes, major satellite repeats and IAP elements maintain persistent DNA methylation after 2i adaptation (Ficz et al., 2013; Habibi et al., 2013). Addition of vitamin C (vitC) can also lower the ES cell methylome. This compound promotes active demethylation by stimulating the TET (Ten Eleven Translocation) enzymes, which oxidize 5-methylcytosines to 5-hydroxymethylcytosines that are potential intermediates towards unmethylated cytosines (Blaschke et al., 2013).

Here, by switching ES cells directly from a serum-based to a 2i+vitC medium, we were able to induce rapid and extensive demethylation genome-wide, mimicking a situation occurring in the early embryo. By combining DNA methylation, chromatin and transcriptional profiling of transposons along with genetic analyses, we found that DNA methylation represses multiple families of transposons in ES cells, but an epigenetic switch towards histone-based control is progressively implemented as DNA methylation disappears. Importantly, we reveal for the first time the specific and overlapping roles of H3K9 and H3K27 trimethylation in controlling distinct transposon families upon DNA demethylation. These findings have important implications for understanding the molecular underpinning of transposon control in the pluripotent cells of the early mammalian embryo.

1.4 Results

DNA methylation is rapidly and extensively lost in ES cells during serum to 2i+vitC media conversion

Dnmt-tKO ES cells are completely devoid of DNA methylation, yet expression levels of most transposable elements remain globally similar to wild-type (WT) ES cells (Karimi et al., 2011b; Tsumura et al., 2006). This may indicate the implementation of alternative mechanisms that compensate for DNA methylation-based repression. To analyze dynamic adaptation, we utilized a culture-based system that results in rapid DNA methylation loss: converting ES cells from serum-based to 2i+vitamin C (2i+vitC) culture conditions. To overcome confounding genetic effects, we used the J1 ES cell line, from which *Dnmt*-tKO mutants were originally derived.

Quantification using the methyl-CpG sensitive restriction enzyme-based LUminometric Methylation Assay (LUMA) (Karimi et al., 2011a) revealed that CpG methylation linearly decreased from 77% to 13% in six days, and reached a minimal level of 6% after 14 days of conversion (**Figure 1.1A**). In comparison, cells grown in serum+vitC or in 2i-only maintained

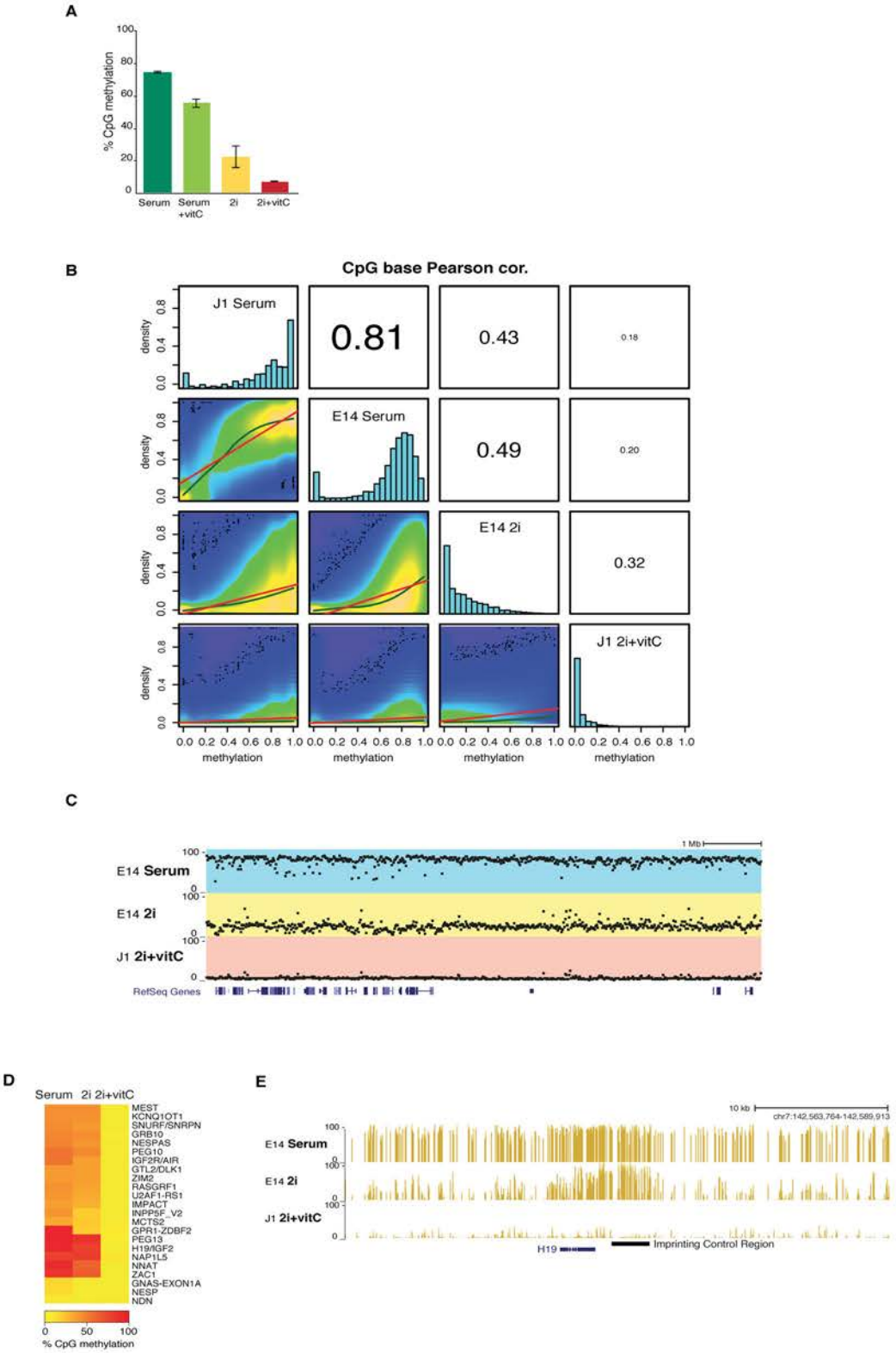


Figure 1.1 - Figure Supplement 1. DNA methylation is almost completely erased in 2i+vitC medium.

A. Global CpG methylation level in J1 ES cells as measured by LUMA. Mean ± SEM between two biological replicates.

B. Smoothed scatter plots of DNA methylation levels at individual CpG between two different conditions as measured by WGBS. Histograms: distribution of CpG methylation levels for all CpG. Pearson correlation between culture conditions.

C. Typical example of methylation profiles in serum, 2i-only and 2i+vitC. Each dot represents the average methylation level over 10-15kb windows.

D. Average methylation level in Imprinted Control Regions (ICRs) as measured by WGBS.

E. Representative genomic region containing the *H19* ICR.

relatively high CpG methylation content after the same treatment duration, with an average of 56% and 22%, respectively (**Figure 1.1 – figure supplement 1A**), in agreement with a previous study (Habibi et al., 2013). This suggests that such a rapid and extensive loss of genomic methylation can only be attained through the synergistic action of 2i-dependent passive demethylation and vitC-dependent active demethylation.

To monitor the demethylation dynamics of specific genomic sequences, we performed quantitative bisulfite-pyrosequencing (**Figure 1.1B**). All analyzed sequences reached very low levels of CpG methylation upon 2+vitC switch, although at various rates. Young LINE1 transposons (L1-A and L1-T) mirrored the dynamics of the genome average, while the CpG-rich promoter of the germline-specific gene *Dazl* was a fast "loser", comparatively. Consistent with their intrinsic ability to maintain high levels of DNA methylation in various contexts of global DNA hypomethylation (Ficz et al., 2013; Seisenberger et al., 2013), the demethylation rate of IAP transposons and the Imprinting Control Region (ICR) of the *H19-Igf2* locus was slower than the rest of genome. Nevertheless, the combination of 2i and vitC eventually overcame chromatin environments that confer protection of these sequences from DNA demethylation.

To determine the extent of DNA demethylation globally in 2i+vitC culture conditions, we carried out whole-genome bisulfite sequencing (WGBS) at the conversion end-point. Quality control indicated high genomic coverage, with approximately 55% of CpGs covered at least five times (**Table S1.1A**). Available methylome maps indicate that 71% and 30% of CpG sites are methylated in serum and in 2i-only conditions, respectively (Habibi et al., 2013; Seisenberger et al., 2012) (**Figure 1.1C**, **Figure 1.1 - figure supplement 1B** and **Table S1.2**); in contrast, ES cells grown in 2i+vitC were almost completely unmethylated, with an average CpG methylation of 4.6%, which is fully consistent with the LUMA quantification (**Figure 1.1C**). Low methylation levels were homogeneously found throughout all genomic compartments, including single-copy genic regions and repeated sequences (**Figure 1.1D** and **Figure 1.1 - figure supplement 1C**). In particular, all transposable element families (ERVs, LINEs and SINEs) were affected by 2i+vitC-induced demethylation (**Figure 1.1E**). In an attempt to identify individual genomic regions with significant DNA methylation traces (Song et al., 2013), we uncovered 4,100 Residually Methylated Regions (RMRs) (**Figure 1.1F** and **G**), which exhibited an average of 26% of CpG methylation after long-term 2i+vitC conversion. These were also regions prone to high DNA methylation retention in 2i-only conditions (65% of CpG methylation). We estimated that nearly 75% of the RMRs overlapped with repeated sequences, among which half belonged to the ERVK class. This

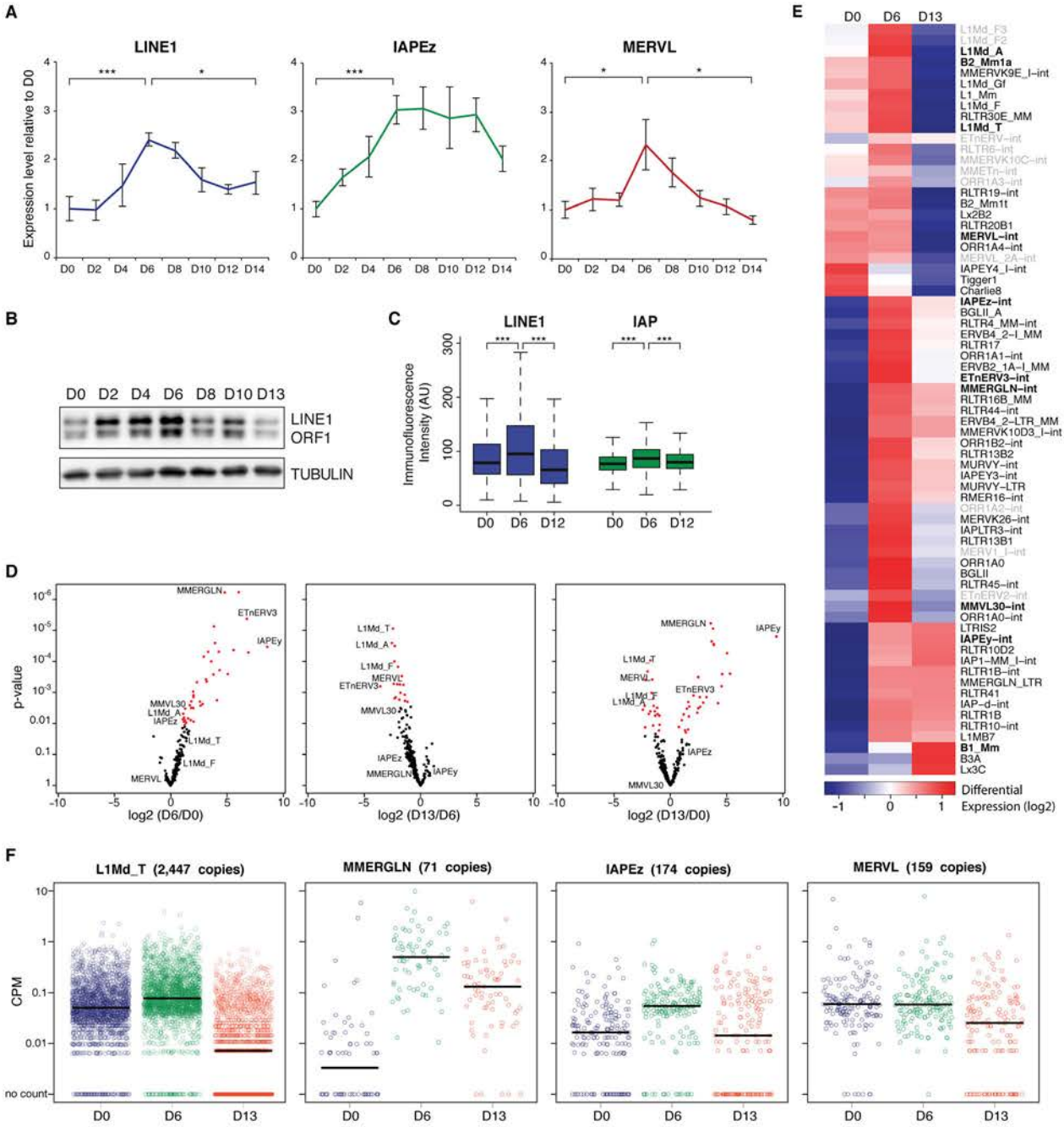


Figure 1.2. Two phases of transposon regulation upon serum to 2i+vitc conversion.

A. Dynamic expression of LINE1, IAPEz and MERVL families upon conversion from serum to 2i+vitC as measured by RT-qPCR. Values were normalized to Gapdh and Rplp0 and are expressed as the fold change to D0. Data represent mean ± SEM from five biological replicates. *p<0.05, **p<0.01 and ***p<0.001 (Student's t-test).

B. Evolution of LINE1-ORF1 protein levels at different time points during medium conversion.

C. Distribution of LINE1-ORF1 and IAP-gag protein levels after ImageJ quantification of immunofluorescence intensity in individual cells. Between 1,000 and 5,000 cells were analyzed per sample. ***p<0.001 (Wilcoxon rank-sum test)

D. Volcano plot representation of up- and down-regulated transposons as measured by RNA-seq between D0 and D6 (left), D6 and D13 (middle), and D0 and D13 (right). Red dots indicate significantly misregulated repeats between two conditions (fold change > 2 and p-value <0.05). RNA-seq mapping allowed multiple hits onto the genome.

E. Heatmap representation and hierarchical clustering of expression changes for 69 transposon families at D0, D6 and D13. Bold names: transposons of specific interest; grey names: transposons that are not significantly up- or down-regulated between any time points. Colors represent on a log2-scale the differential expression between a given time point and the average of the three time points.

F. Expression of individual elements from different transposon families at D0, D6 and D13 in Count per Millions (CPM). Each dot represents a single element. RNA-seq mapping allowed only unique hits in the reference genome; only elements with a minimum of 10 reads in at least one of the sample were conserved. The black bar represents the median of the distribution. Analyzed numbers of distinct transposon copies per family appear into brackets.

confirmed the specific ability of these elements, which includes IAPs, to resist genome-wide erasure of DNA methylation. One quarter of the repeat-associated RMRs overlapped with LINEs, however specifically localized around 5' UTR regions; in contrast, ERVK-associated RMRs encompassed the entire length of these elements (**Figure 1.1G**). Notably, Imprinting Control Regions (ICRs), which are usually protected against DNA methylation erasure in 2i conditions, were devoid of any residual DNA methylation in 2i+vitC (**Figure 1.1 - figure supplements 1D** and **E**).

Our analyses show that only scarce genomic regions retain DNA methylation in 2i+VitC, and even those regions are lowly methylated when compared to other culture systems. Global CpG methylation levels (less than 5%) are unprecedented in male WT cells, both in culture and *in vivo* (Seisenberger et al., 2013). This experimental system provides a valuable means to study the dynamic adaptation of the genome to a loss of DNA methylation

Transposons undergo a biphasic mode of regulation upon serum to 2i+vitC conversion

Using the serum to 2i+vitC medium conversion system, we investigated how transposable elements transcriptionally respond to an acute loss of DNA methylation. Through a time-course RT-qPCR analysis of steady-state levels of three classes of retrotranscripts (LINE1, IAPEz and MERVL), we observed a two-phase pattern: *1)* an initial up-regulation, which culminates at day 6 (D6) of 2i+vitC conversion, when genomic methylation reaches a low plateau, then *2)* re-silencing in the absence of DNA methylation (**Figure 1.2A**). This was confirmed by amplification-free Nanostring nCounter quantification and further extended to VL30 elements (**Figure 1.2 - figure supplement 1A**). To rule out background-specific effects, we exposed serum-cultured E14 ES cells to 2i+vitC (**Figure 1.2 - figure supplement 1B**). Despite some differences in the magnitude of transposon de-repression observed between the J1 and E14 cell lines, the same biphasic pattern of regulation was reproduced. By contrast, the quantity of transposon transcripts remained constant during the conversion from serum to 2i-only or from serum to serum+vitC (**Figure 1.2 - figure supplement 1A**), which further underscores the synergistic effect of 2i and vitC in releasing DNA methylation-based repression of transposons. Importantly, the transposon transcription burst did not occur upon conversion of *Dnmt*-tKO cells (**Figure 1.2 - figure supplement 1C**). Rapid transition from a methylated to an unmethylated state seems to provide a window for transposon reactivation; this is in agreement with the hypothesis that *Dnmt*-tKO ES cells have likely acquired long-term compensatory mechanisms preventing this relaxation.

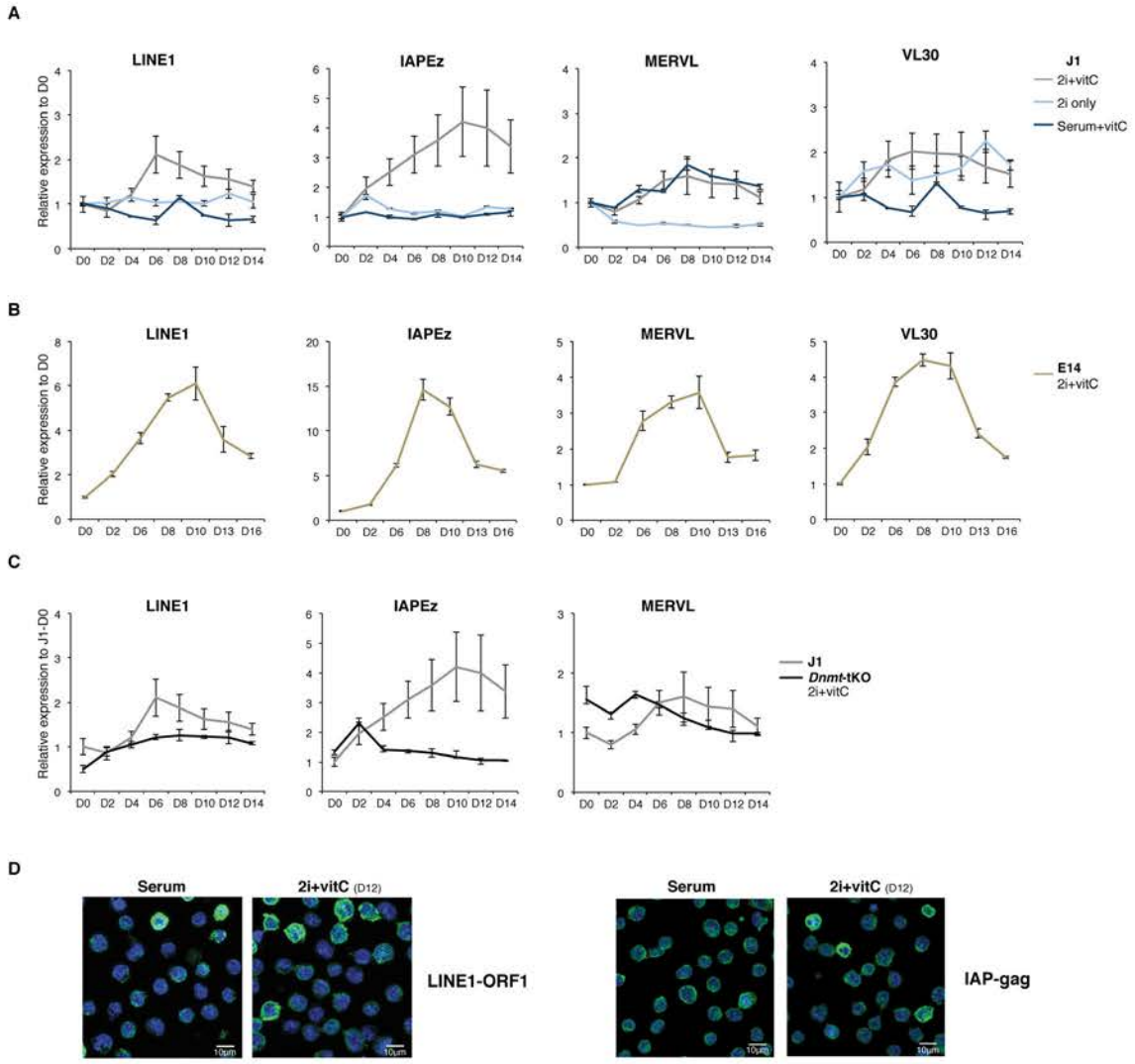


Figure 1.2 -supplement 1. Two phases of transposon regulation upon serum to 2i+vitc conversion.

A. Expression of different transposons in J1 ES cells as measured by Nanostring during the conversion from serum to 2i+vitC (grey), 2i-only (light blue) and serum+vitC (dark blue). Data are expressed as fold change to D0 and represent mean ± SEM between seven (for 2i+vitC) or two (for 2i- and vitC-only) biological replicates.

B. Expression of transposons in E14 ES cells. Data are expressed as fold change to D0 and represent mean ± SEM between two biological replicates.

C. Expression of transposons in Dnmt-tKO and J1 ES cells measured. Data are expressed as fold change to J1 D0 and represent mean ± SEM between three biological replicates (for *Dnmt*-tKO).

D. Immunofluorescence staining for LINE1-ORF1 and IAP-gag proteins in serum- and 2i+vitC-grown ES cells.

We further found that the burst of transposon expression also occurs at the protein level: both LINE1-ORF1 and IAP-gag proteins presented a peak of expression at D6, which we detected by western blotting (**Figure 1.2B**) and by quantification of immunofluorescence signals (**Figure 1.2C** and **Figure 1.2 - figure supplement 1D**). While IAP-gag staining was uniform among cells at a given time point, LINE1 protein intensity showed great inter-cellular variability, ranging from intense to no signal, in both in serum (D0) and 2i+vitC conditions (D14). In an attempt to link the heterogeneity of LINE1 signal with fluctuating levels of pluripotency or DNA damage, we performed co-staining against NANOG and phosphorylated-H2AX, respectively, but we could not detect any correlation (data not shown).

We wanted to rule out that the repression phase we observed was not simply a reflection of positive selection of a subset of cells that maintained transposon repression throughout the medium conversion. We found cell proliferation to remain globally constant over the 14-day period of media conversion, as measured by division rate (**Figure 1.2 - figure supplement 2A**), transcriptional level of different proliferation markers (**Figure 1.2 - figure supplement 2B**) or percentage of histone H3 Serine 10 phosphorylation-positive cells (**Figure 1.2 - figure supplement 2C**). Similarly, we did not observe increased cell death/apoptosis at any days during the conversion (**Figure 1.2 - figure supplement 2C**). Finally, despite the transient release of transposon silencing at D6, we failed to detect transposon multiplication or transposon-induced chromosome rearrangements: genomic copy numbers of LINE1 and IAPEz elements as well as karyotypes were globally similar between cells before (D0) and after the transposon burst (D14) (**Figure 1.2 - figure supplement 2D,E**). In sum, 2i+vitC induces a transient up-regulation of transposon transcription and trans lation, but cellular viability and genome integrity remain largely intact.

Transposon silencing release occurs at the familial and individual level

To gain a qualitative and quantitative view of the transcriptional dynamics of transposons upon acute loss of DNA methylation, we performed paired-end RNA-seq at D0, D6 and D13 of serum to 2i+vitC conversion, in biological replicates. Typically, to map transposons, the choice is either to allow multiple hits at the expense of specificity, or to consider unique reads only and lose substantial information. Here, we combined the two methods, in order to provide in-depth characterization of the dynamics of transposon regulation at the familial level, while bringing insights into intra-familial heterogeneity. We further improved transposon mapping by correcting the RepeatMasker annotation (**Figure**

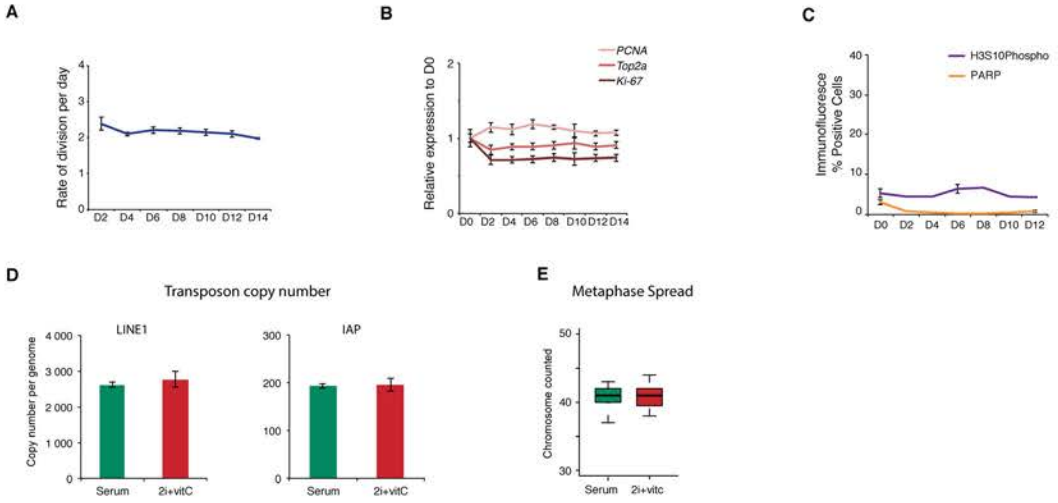


Figure 1.2 -supplement 2.

A. Mitotic division rate of J1 ES cells during the conversion from serum to 2i+vitC. Mean ± SEM between three biological replicates.

B. Quantification of proliferation marker expression by Nanostring. Data are expressed as fold change to D0 and represent mean ± SEM between seven biological replicates.

C. Percentage of cells immunostained against the apoptosis marker PARP and the mitotic marker H3S10 phosphorylation. Between 5,000 and 10,000 cells were counted at each time point. When available, data represent mean ± SEM between two biological replicates.

D. Absolute copy number of L1-ORF2 and IAPΔ1 fragments assayed by qPCR on genomic DNA. Values are expressed as copies per genome, representing the mean ± SEM between two biological replicates.

E. Chromosome numbers counted in 20 metaphase spreads in serum and 2i+vitC medium.

1.2 - figure supplement 3A), which tends to overestimate the number of transposon entities by counting a unique element as several individual fragments. This is systematic for ERVs, which are split into internal and LTR sequences, but can also concern any type of transposons with small internal deletions or insertions. Using bioinformatic resources allowing the assembly of different fragments of an element (Bailly-Bechet et al., 2014), our reconstructed version gave a census of 588,739 LINEs and 497,706 ERVs (**Table S1.3**), while the original annotation roughly doubles these numbers, with 989,411 LINEs and 969,096 ERVs. Finally, we assigned an integrity score to each element (1 being the maximum), taking into account deletions, insertions and the divergence rate from the consensus sequence. Using a cutoff of 0.8, we predicted a number of 37,194 relatively intact LINEs (6.3% of total LINE elements) and 15,604 ERVs (3.1%) in the mouse reference genome.

Quality control of our RNA-seq datasets indicated high genomic coverage (**Table S1.1B**) and consistency between replicates (**Figure 1.2 - figure supplement 3B**). Notably, by excluding transposon-derived reads mapping to RefSeq exons, only autonomously transcribed transposons were considered for this analysis. By allowing multiple hits and by weighting each read by its hit number, a total of 58 transposon families were found differentially expressed between at least two of the time points of medium conversion (**Figure 1.2D,E**). Volcano plots show that almost all families underwent significant up-regulation from D0 to D6 (**Figure 1.2D**, left panel), ranging from modest (LINEs) to robust (MMERGLN) fold changes. Silencing restoration also occurred globally between D6 and D13, except for IAPEy or B1 elements, which remained at constant levels (**Figure 1.2D**, middle panel). Comparison of transposon expression levels between the two end-points (D0 and D13) indicated skewing in both directions (**Figure 1.2D**, right panel). Some families, such as MERVL, SINEs B2 or any LINE1 types, were more strongly repressed after the 2i+vitC conversion at D13 than initially at D0 in serum. Others, like MMERGLN, ETnERV3 and IAPEz, underwent repression from D6 to D13, but not to the full extent when compared to D0. As a general rule, these data show that non-LTR (LINEs and SINEs) and LTR elements belonging to the K, L and 1 classes—albeit very different in terms of evolutionary origins and genomic structures—adopt common fates upon acute loss of DNA methylation.

To examine whether the burst of transcription observed from bulk RNA profiling emanated from a few discrete elements or reflected a general trend within each family, we measured the transcriptional output of individual transposon copies by allowing unique read mapping only. We found that 7,163 uniquely identifiable LINEs and 2,372 ERVs showed activity throughout the conversion, which represented 1.2% and 3.8% of the total number of

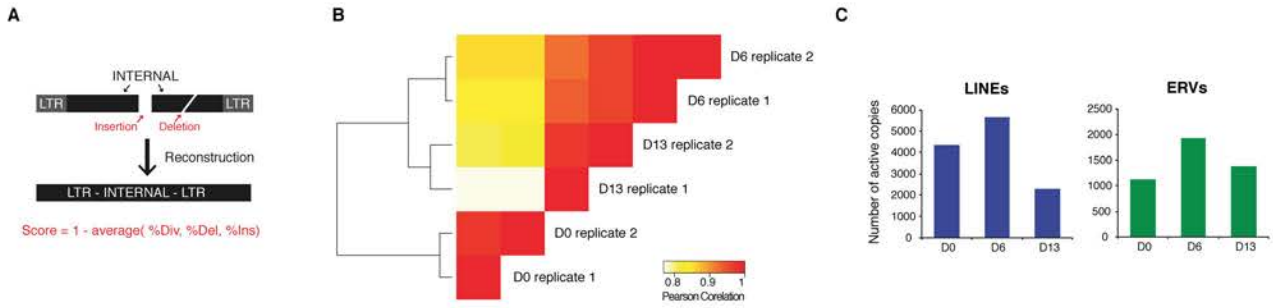


Figure 1.2 - supplement 3. Genome-wide characterization of transposon relaxation.
A. Reconstruction principles of RepeatMasker annotation.
B. Pearson correlation and hierarchical clustering between RNA-seq experiments.
C. Number of individual LINEs and ERVs with detectable expression at D0, D6 and D13 of 2i+vitC conversion. Elements with a minimum of 10 uniquely mappable reads were considered.

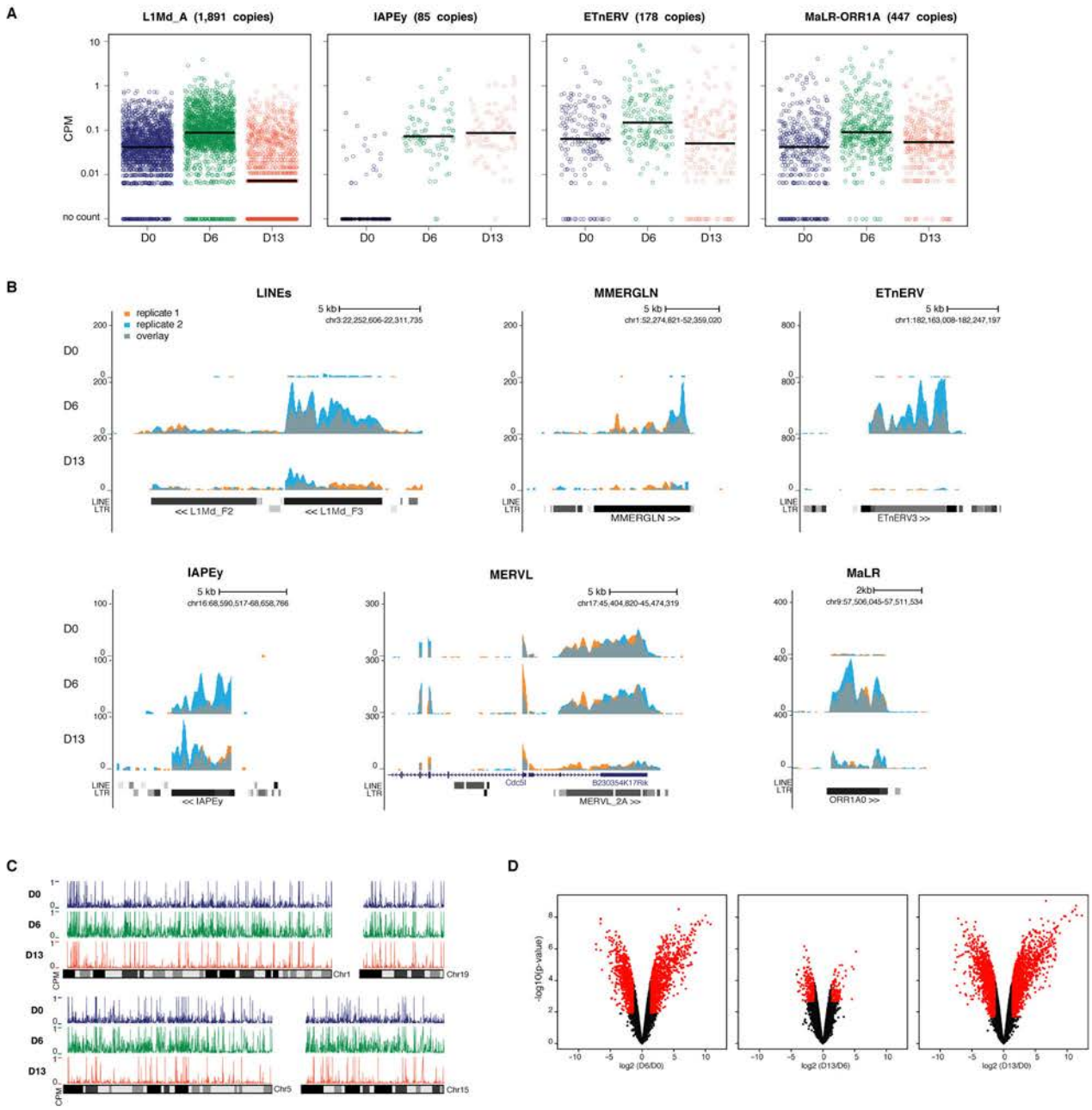


Figure 1.2 - supplement 4. Genome-wide characterization of transposon relaxation.

A. Expression of individual elements from different transposon families at D0, D6 and D13 in Count per Millions (CPM). Each dot represents a single element. RNA-seq mapping allowed only unique hits in the reference genome and only elements that had a minimum of 10 reads in at least one of the sample were conserved. The black bar represents the median of the distribution. Analyzed numbers of distinct transposon copies per family appear into brackets.

B. Representative genomic regions comprising LINE1 and ERV repeats. RNA-seq coverage for the two biological replicates is represented in blue and orange, and the overlap in grey. Data represent normalized read density.

C. Expression of individual transposable elements and their localization on chromosomes 1, 5, 15 and 19 at D0, D6 and D13 of conversion. Each bar represents the expression of a single element in Count per Millions (CPM). Elements with a minimum of 10 uniquely mappable reads in at least one sample were considered.

D. Volcano plot representation of up- and down-regulated genes as measured by RNA-seq between D0 and D6, D6 and D13, and D0 and D13. Red dots indicate significantly misregulated genes (fold change > 4 and p-value <0.01).

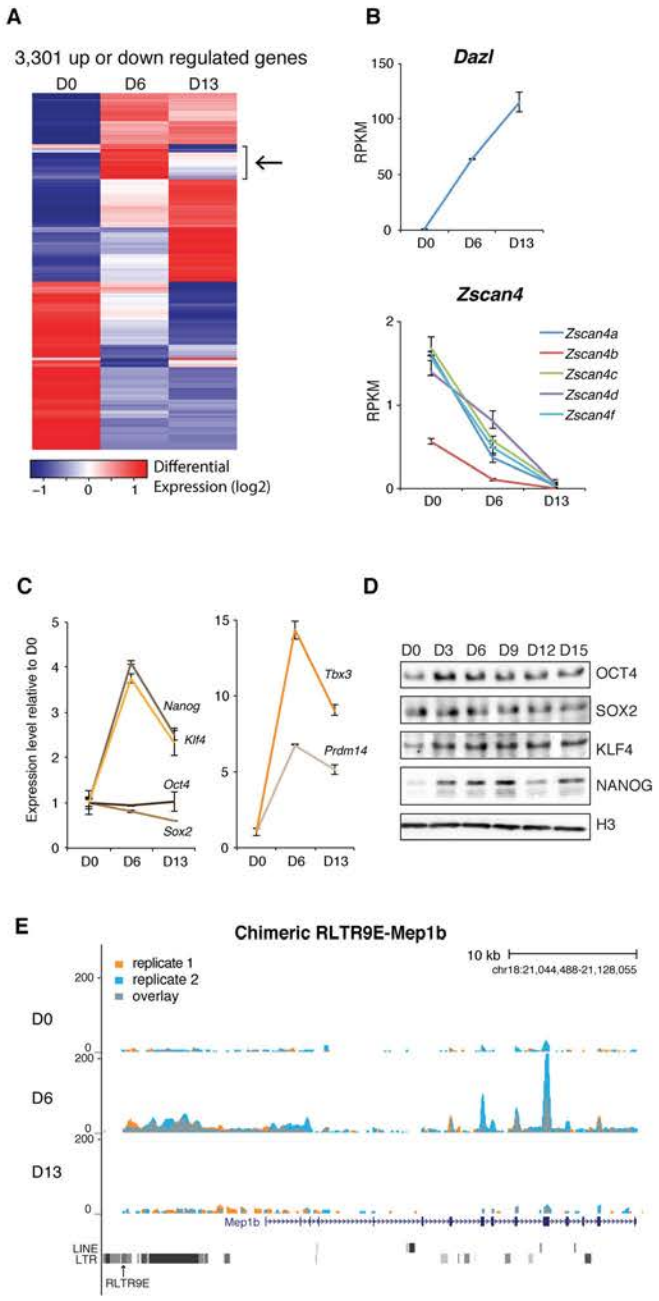


Figure 1.3. Gene expression analysis upon serum to 2i+vitC conversion.

A. Heatmap representation of genes (n=3,301) that are significantly misregulated between at least two time points of D0, D6 and D13 during medium conversion. Color codes as in Figure 2E. The arrow highlights a subset of genes whose expression transiently peaks at D6.

B. Monotonic expression patterns of *Dazl* and *Zscan4* family genes as measured by RNA-seq at D0, D6 and D13, expressed in RPKM (read per kb per millions). Mean ± SEM between two biological replicates.

C. Dynamic expression of pluripotency transcription factor genes as measured by RNA-seq. Data is expressed as fold change to D0 and represent mean ± SEM between two biological replicates.

D. Expression of core pluripotency transcription factors by western blot.

E. RNA-seq track showing a chimeric transcript between a RLTR9E transposon element and the Mep1b gene specifically expressed at D6. Data represent normalized read density

LINE1s and ERVs, respectively, or 19.3% and 15.2% of the intact elements of these families (integrity score >0.8). Importantly, these numbers are likely underestimated because active but identical copies cannot be discriminated, and are discarded from the analysis. Within all families, the number of significantly expressed elements was higher at D6 than at D0 or D13 of conversion (**Table S1.3** and **Figure 1.2 - figure supplement 3C**). Generally, not only were more elements active, but individual copies also gained expression at D6 (**Figure 1.2F** and **Figure 1.2 - figure supplement 4A,B**). Finally, active transposons were evenly distributed along chromosomes, with no particular genomic hotspot (**Figure 1.2 - figure supplement 4C**).

As a whole, the unique mapping analysis confirmed the class-specific features previously inferred from the familial analysis, regarding the degree of activation at D6 (from modest for LINEs to intense for MMERGLN) and silencing restoration at D13 (strong for LINEs, intermediate for MMERGLN and nonexistent for IAPEy). Most importantly, it uncovered unprecedented details into the diverse regulation of individual transposons. Expression levels were the most homogeneous among elements of the same family during the D6 de-repression phase. Comparatively, at the D13 silencing restoration time-point, we observed heterogenic regulation at the inter- and intra-familial levels (**Figure 1.2F** and **Figure 1.2 - figure supplement 4A**). Some families, such as LINEs and MMERGLN, displayed collective behaviors, with the vast majority of elements simultaneously undergoing repression. In contrast, IAPEz, MERVL or ETn elements showed the widest distribution in individual expression. In particular, IAPEz elements were split into two categories at D13, one that maintained high expression, and the other that underwent complete silencing.

Globally, our analysis reveals that transposons undergo a transient relaxation of silencing upon DNA methylation loss followed by an expression reduction phase. However, family- and element-specific behaviors provide nuance to this general trend. It should be stressed here that a certain degree of heterogeneity is frequently inaccessible for young and highly conserved families of transposons, such as IAPEz and MMERVK10C, for which mapping reads to precise genomic locations is ambiguous, if not impossible.

Silencing release is specific to transposons

Compared to transposons, protein-coding genes followed different dynamics during 2i+vitC conversion (**Figure 1.3A** and **Figure 1.2 - figure supplement 4D**). The vast majority exhibited stable expression, while 3,301 genes were either up- or down-regulated; these numbers are similar to previous reports of a serum to 2i transcriptional switch (Marks et

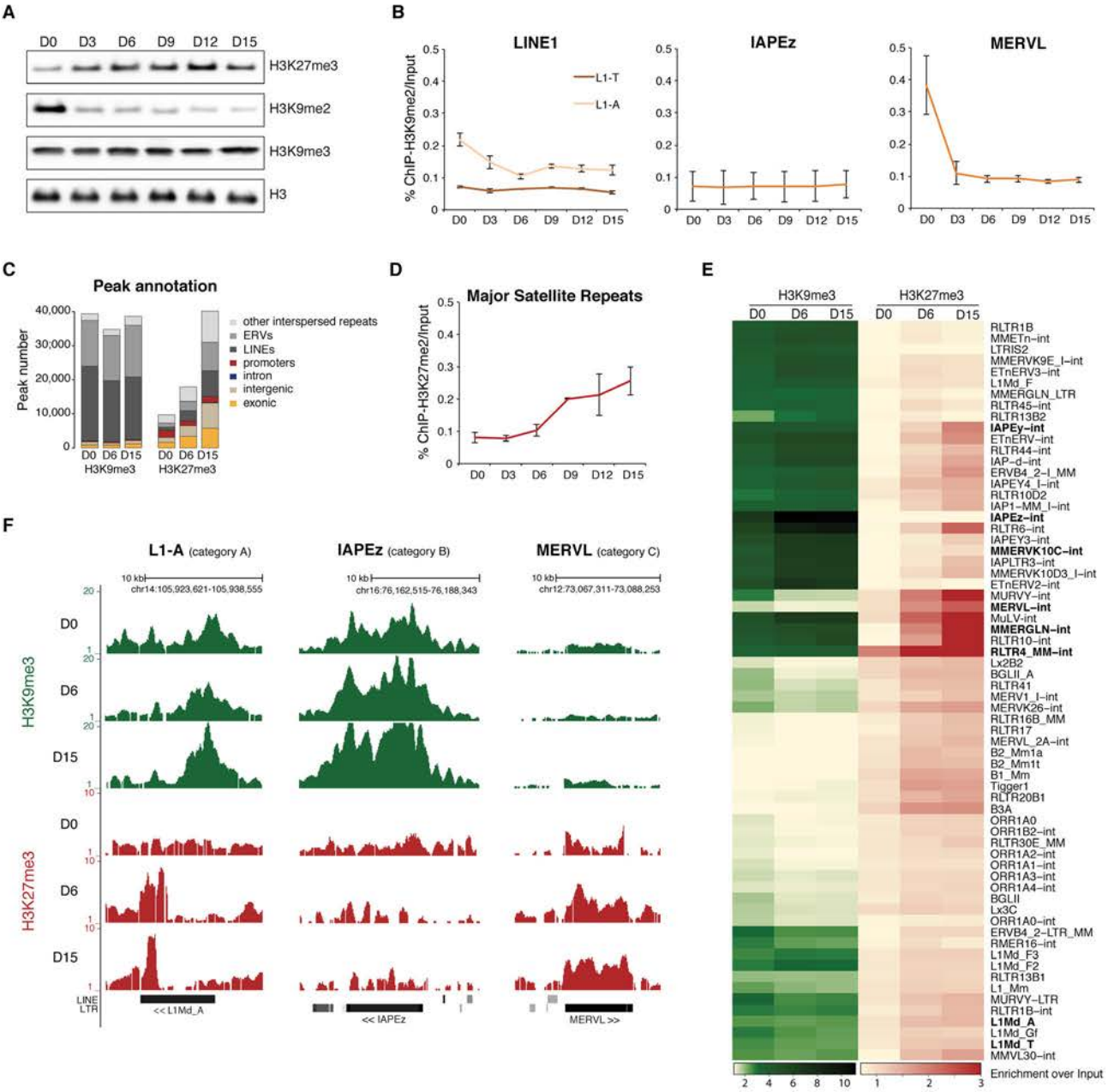


Figure 1.4. Repressive chromatin reorganization upon loss of DNA methylation.

A. Western blot analysis of global levels of repressive histone modifications during the course of serum to 2i+vitC conversion.

B. H3K9me2 enrichment levels at three transposon families as measured by ChIP-qPCR. Quantitative data are expressed as the percentage of ChIP over Input. Data represents mean ± S.E.M. of two biological replicates.

C. Genomic annotation of ChIP-seq peaks. Data represent the number of annotated peaks. H3K9me3 data are representative of two biological replicates, while H3K27me3 data represent only one, as peak calling could not be performed successfully on the second replicate.

D. H3K27me3 enrichment at major satellite repeats as measured by ChIP-qPCR. Data are represented as in 4B.

E. Heatmap and hierarchical clustering of average H3K9me3 and H3K27me3 levels in 69 transposons families at D0, D6 and D15. Colors represent the average read count for an element in a given family, relative to input (average between two biological replicates). Only intact (score>0.8) elements were considered.

F. Representative genomic region depicting evolution of H3K9me3 (green) and H3K27me3 (red) at L1-A (category A), IAPEz (category B) and MERVL (category C) transposons. Data represent normalized read density

al., 2012). While the general expression trend for transposons was biphasic, most differentially expressed genes displayed a monotonic pattern. A relevant example is the *Dazl* gene, which was continuously up-regulated from D0 to D13 (**Figure 1.3B**), reflecting its sensitivity to vitC (Blaschke et al., 2013). Conversely, expression of genes encoding transcription factors of the ZSCAN4 family progressively decreased during the conversion, with undetectable transcripts by D13 (**Figure 1.3B**). As expression of these factors reflects a subpopulation of ES cells exhibiting a transcriptional profile akin to 2-cell stage embryos (Macfarlan et al., 2012), our results imply that 2-cell-like cells exist in serum-based conditions but disappear in 2i+vitC medium.

The burst of expression at D6 appears specific to transposons, and is not a general trend of the genome. Nevertheless, 156 genes adopted a transposon-type pattern, with a peak at D6 followed by subsequent down-regulation at D13 (**Figure 1.3A** and **Table S1.4**). These genes were linked to ontology categories such as organismal development and were significantly enriched for transcription factors, most notably those related to pluripotency (**Table S1.4**). Further examination indicated that transcription of *Nanog*, *Klf4*, *Tbx3* and *Prmd14* peaked at D6 (**Figure 1.3C**); enhanced production of pluripotency-related proteins was also observed by western blot within the first few days of 2i+vitC conversion (**Figure 1.3D**). Therefore, the peak of transposon transcription at D6 coincides with a maximum availability of pluripotency regulators.

It was previously shown that LTR sequences of ERVs can direct transcription of nearby genes in ES cells and early embryos, forming chimeric transcripts (Karimi et al., 2011b; Macfarlan et al., 2012). We detected several dozens of genes that used a promoter located in a transposable element (ERV or LINE1), independently of the medium composition (**Table S1.5**). A particular case was *Mep1b*, which clusters with the 156 "transposon-like" genes. This gene was induced ten-fold at D6, concomitant with the activation of the RLTR9E element driving its expression, before returning to its initial level at D13 (**Figure 1.3E**). This example indicates that the burst of transposon expression at D6 can coordinate the transient activation of adjacent genes. However, apart from a few examples, it can be concluded that the genome-wide relaxation of transposons had generally a minimal effect on protein-coding gene expression.

Reconfiguration of the repressive chromatin landscape upon 2i+vitC conversion

To gain insight into the basis for transposon regulatory dynamics, we examined the chromatin state of cells undergoing serum to 2i+vitC conversion. By western blot and

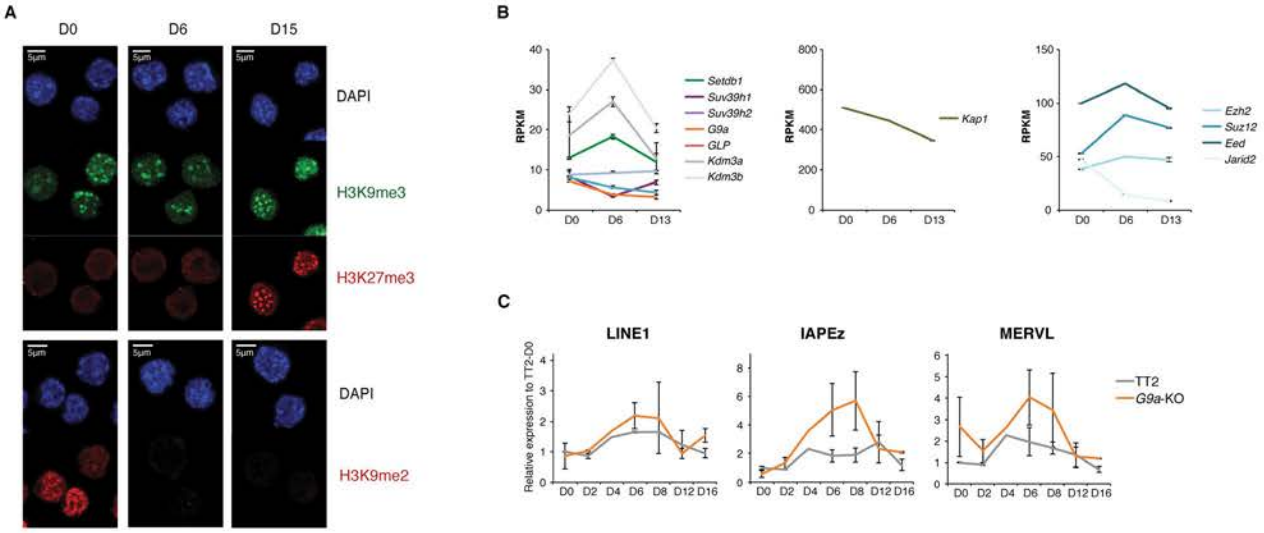


Figure 1.4 - figure supplement 1. Repressive chromatin reorganization during conversion from serum to 2i+vitC.
A. Immunofluorescence experiment against H3K9me2, H3K9me3 and H3K27me3 in J1 ES cells during serum to 2i+vitC conversion.
B. RPKM expression values of H3K9 HKMTs, Kap1 and PRC2 members as measured by RNA-seq. Mean ± SEM between two biological replicates.
C. Transposon expression in *G9a*-KO and TT2 ES cells (WT) as measured by Nanostring. Data are expressed as fold changes to WT D0 and represent mean ± SEM between two biological replicates.

immunostaining, we observed large-scale reorganization of histone modifications linked to transcriptional repression. While H3K9me3 marks remained globally constant, H3K9me2 levels were strongly reduced and, inversely, H3K27me3 levels increased from the first days of conversion (**Figure 1.4A** and **Figure 1.4 - figure supplement 1A**). The global dynamics of histone marks were not correlated with the changes in the availability—or lack thereof—of H3K9me2 modifiers and components of the polycomb machinery (**Figure 1.4 - figure supplement 1B**).

ChIP-qPCR measurement confirmed either the constitutive absence or the rapid removal of H3K9me2 at several transposon types upon 2i+vitC conversion (**Figure 1.4B**), making it unlikely that this mark could participate to long-term transposon silencing in the absence of DNA methylation. To functionally exclude this possibility, we examined ES cells lacking the G9a H3K9 dimethyltransferase (**Figure 1.4 - figure supplement 1C** and Tachibana et al., 2002). As previously described, when cultured in serum, *G9a*-KO ES cells did not exhibit significant up-regulation of transposons as measured by Nanostring, with the exception of MERVL elements (Macfarlan et al., 2012; Maksakova et al., 2013). Upon 2i+vitC conversion, while LINE1 elements behaved as in WT cells, IAPEz and MERVL expression was enhanced around D6 in *G9a*-KO cells. As H3K9me2 cannot be detected after D3 at these sequences (**Figure 1.4B**), this relative up-regulation likely occurs through indirect effects. Most importantly, *G9a* mutants exhibited transposon re-silencing after D6, which excludes a role for H3K9me2 in compensating the loss of DNA methylation-dependent repression.

We then focused our analysis on the distribution of H3K9me3 and H3K27me3 marks by chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) in biological replicates at D0, D6 and D15 of conversion, allowing multiple mapping with random allocation (**Table S1.1B** and **Figure 1.4 - figure supplement 2A**). Neither the total number of H3K9me3 peaks (39,424 at D0 and 38,554 at D15), nor their preferential occurrence on transposons was significantly altered during the conversion (**Figure 1.4C**). In contrast, the number of H3K27me3-enriched regions raised four fold from D0 to D15 (9,663 to 40,098). The vast majority of newly gained H3K27me3 peaks were located in ERV and LINE1 repeats, at the expense of gene promoters (**Figure 1.4C**). We also observed a gradual H3K27me3 re-localization to pericentric heterochromatin during the conversion, by ChIP-qPCR at major satellite repeats (**Figure 1.4D**), by immunostaining (**Figure 1.4 - figure supplement 1A**) and by mapping ChIP-seq reads to the major satellite consensus sequence (**Figure 1.4 - figure supplement 2B**). Redistribution of H3K27me3 from gene promoters

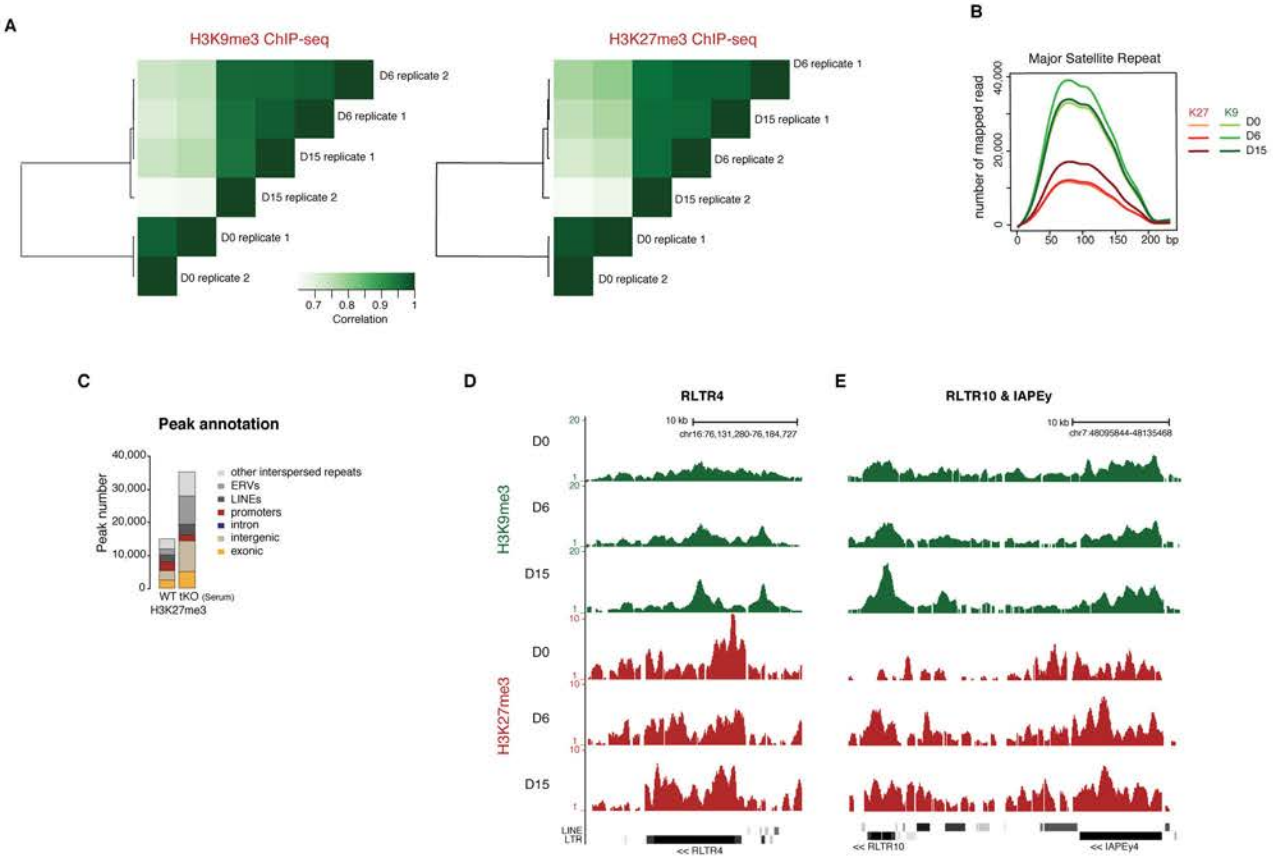


Figure 1.4 - figure supplement 2. Repressive chromatin reorganization during conversion from serum to 2i+vitC.
A. Pearson correlation and hierarchical clustering between samples after peak calling in H3K9me3 and H3K27me3 ChIP-seq experiments.
B. Number of H3K9me3 and H3K27me3 ChIP-seq read mapping to major satellite repeat consensus sequence
C. Genomic annotation of ChIP-seq peaks in ES grown in serum but lacking DNA methylation (*Dnmt3*-dKO, *Dnmt1*-KD). On the contrary to Figure 4C, peak calling was performed without taking input data into account. ChIP-seq datasets were taken from. Brinkman et al., 2012.
D. Representative genomic region depicting evolution of H3K9me3 and H3K27me3 at RLTR4 repeats. Data represent normalized read density.
E. Representative genomic region depicting evolution of H3K9me3 and H3K27me3 at RLTR10 and IAPEy repeats.

towards satellite repeats was previously reported in 2i-only conditions (Marks et al., 2012). However, increased H3K27me3 levels and subsequent accumulation at different transposon repeats seems specific to the globally hypomethylated genome of 2i+vitC-cultured cells. Accordingly, hypomethylated *Dnmt*-tKO ES cells grown in serum displayed similar H3K27me3 redistribution towards transposons when we analyzed available ChIP-seq data (**Figure 1.4 - figure supplement 2C**).

Relative H3K9me3 and H3K27me3 enrichments define three categories of transposons

We next measured relative H3K9me3 and H3K27me3 levels over different transposon families, focusing our analysis on elements that were scored as intact. At D0 in serum, most transposon families were occupied by H3K9me3 to various extents, but lacked H3K27me3 (**Figure 1.4E**). One noticeable exception was RLTR4, which exhibited a strong H3K27me3 signal (**Figure 1.4 - figure supplement 2D**). Interestingly, this element is 90% identical to MuLV, which is one of the few transposons up-regulated in polycomb-deficient ES cells (Leeb et al., 2010). Upon 2i+vitC conversion, H3K9me3 levels remained largely constant, although patterns observed in serum tended to be exacerbated: families with the initial highest enrichment (IAPEz, RLTR6 and MMERVK10C) were further enriched for this mark, while families with modest enrichment (MERVL, MURVY or the MalR-class L ORR1A and ORR1B) tended to lose it. Meanwhile, H3K27me3 progressively accumulated at most transposons (**Figure 1.4E**), and this gain was variable among families: from inexistent for IAPEz to moderate for LINEs and VL30, and to strong for various ERVs. Remarkably, different kinetics were observed for H3K9me3- and H3K27me3-related changes: H3K9me3 levels were rapidly modified between D0 and D6, while H3K27m3 gain lagged behind, reaching its full extent between D6 and D15. Although the whole picture is quite complex, it can be concluded that medium-induced DNA methylation profoundly remodels the repressive chromatin landscape of transposons. From a universal H3K9me3 occupancy in serum, transposon families exhibited three general trends in 2i+vitC: *A)* co-occupancy of H3K9me3 and H3K27me3 (LINEs, MMERGLN, RLTR6, RLTR10, IAPEy, VL30), *B)* exclusive H3K9me3 occupation (IAPEz, MMERVK10C), and *C)* complete switch from H3K9me3 to H3K27me3-regulated chromatin (MERVL and MURVY) (**Figure 1.4E,F** and **Figure 1.4 - figure supplement 2E**). Our analysis therefore provides a classification of the different transposon families into three main categories (A, B, and C), according to the chromatin signature they adopt upon DNA methylation loss.

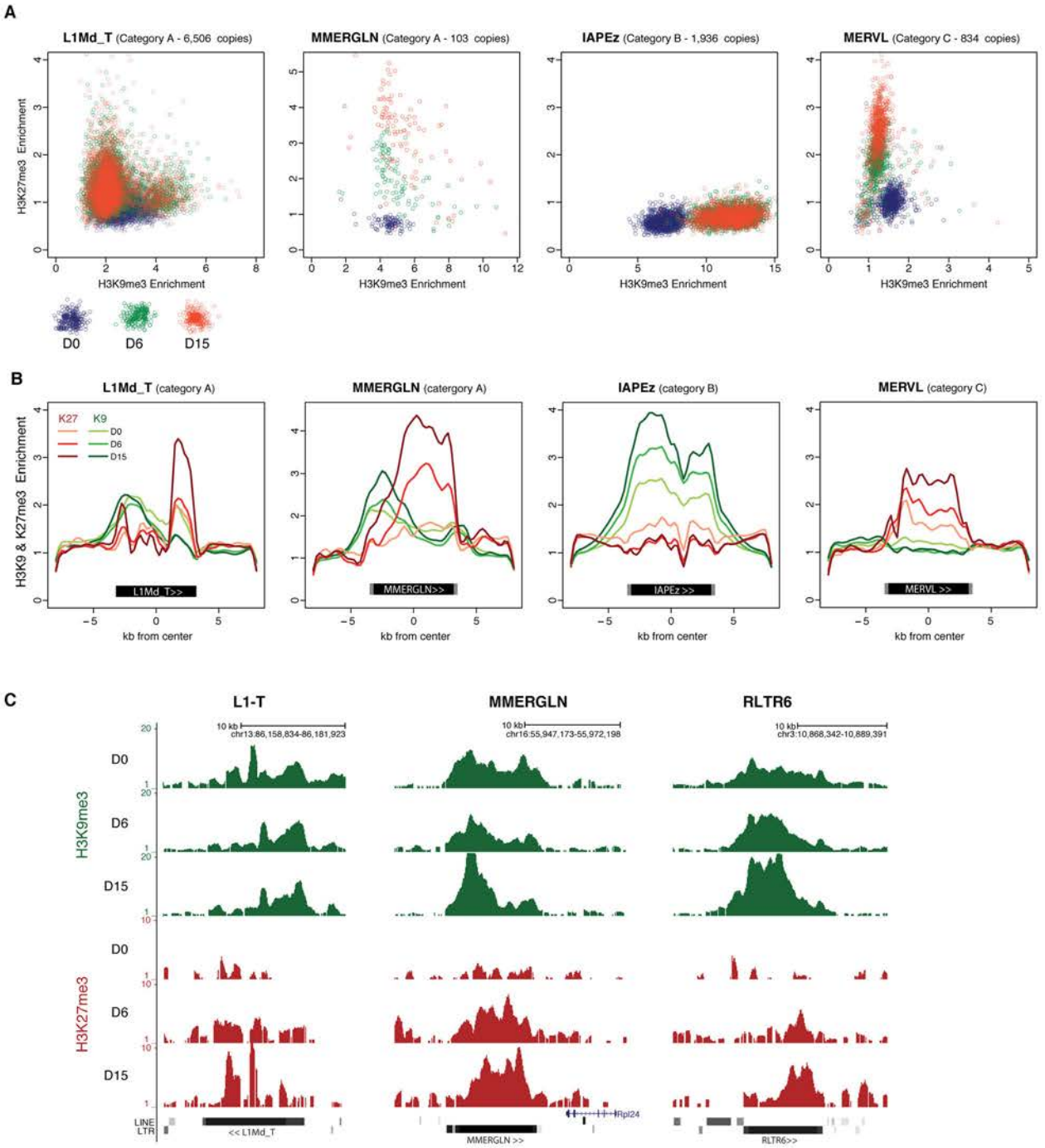


Figure 1.5. H3K9me3 and H3K27me3 mark the same transposons but do not spatially overlap.

A. Normalized H3K9me3 and H3K27me3 enrichment over input at individual elements from different transposon families. Each dot represents a single element at D0 (blue), D6 (green) and D15 (red). Only intact (integrity score>0.8) elements were considered. Data represent the average between two biological replicates. Analyzed numbers of distinct transposon copies per family appear into brackets.

B. Composite profile showing H3K9me3 (green) and H3K27me3 (red) coverage along different transposon sequences at D0, D6 and D15 of medium conversion.

D. Representative genomic regions comprising LINE1 and ERV repeats that gain H3K27me3 (red) in their 3' end during the conversion, while maintaining H3K9me3 in the 5' end. Data represent normalized read density.

To assess the behavior of individual elements among these three generic patterns, we attempted to analyze unique mappers, but the coverage on individual transposons was too low to extract reliable information. Nevertheless, to gain insight into the question of intra-familial heterogeneity, we plotted H3K9me3 and H3K27me3 enrichment for every intact transposons (score>0.8) per family during the conversion. We found that elements of the B and C categories tended to be very homogeneous. IAPEz elements (category B) collectively gained H3K9me3 from D0 to D6; the MERVL and the Y-specific MURVY families (category C) also showed compact patterns, with individual elements transitioning together from H3K9m3 enrichment at D0 to H3K27me3 at D15 (**Figure 1.5A** and **Figure 1.5 - figure supplement 1B,C**). The A category, which is enriched for both H3K9me3 and H3K27me3, was more diverse, with some families displaying homogeneous patterns, while others showed intra-familial dispersion in chromatin fates upon conversion. Within the MMERGLN, RLTR6 or RLTR10 families, all elements gained H3K27me3 while maintaining or gaining high levels of H3K9me3. Within the L1-T and IAPEy families, the majority of elements gained H3K27me3, but a subset maintained H3K9me3 without acquiring H3K27me3 (**Figure 1.5A** and **Figure 1.5 - figure supplement 1A**). Another case of intra-familial heterogeneity is provided by RLTR4, which specifically carries H3K27me3 marks at D0 in serum: we demonstrate here that this enrichment was restricted to a small proportion of elements, as was previously suspected (**Figure 1.5 - figure supplement 1D** and Reichmann et al., 2012). By extracting single-element information from RNA-seq and ChIP-seq data, it is clear that both transcriptional and chromatin heterogeneity exists among some transposon families. Our analysis reveals that caution should be taken when interpreting average familial behaviors, as they may be representative of only a few individual elements inside a given family.

H3K27me3 occupies DNA methylation- and H3K9me3-free territories of transposon sequences

H3K27me3 and H3K9me3 marks usually do not occur concomitantly (Mikkelsen et al., 2007). We were therefore intrigued to observe that H3K27me3 and H3K9me3 were simultaneously enriched at transposon families of the A category in 2i+vitC medium (**Figure 1.4E** and **Figure 1.4 - figure supplement 1G,H**). To map the relative position of H3K9me3 and H3K27me3, we determined their average profile over full-length individual elements of all transposon families, including their immediate genomic vicinity (+/- 5kb from the center of each element). Notably, H3K9me3 domains often spread out on adjacent genomic regions, whereas H3K27me3 was confined to transposon sequences (**Figure 1.5B,C**

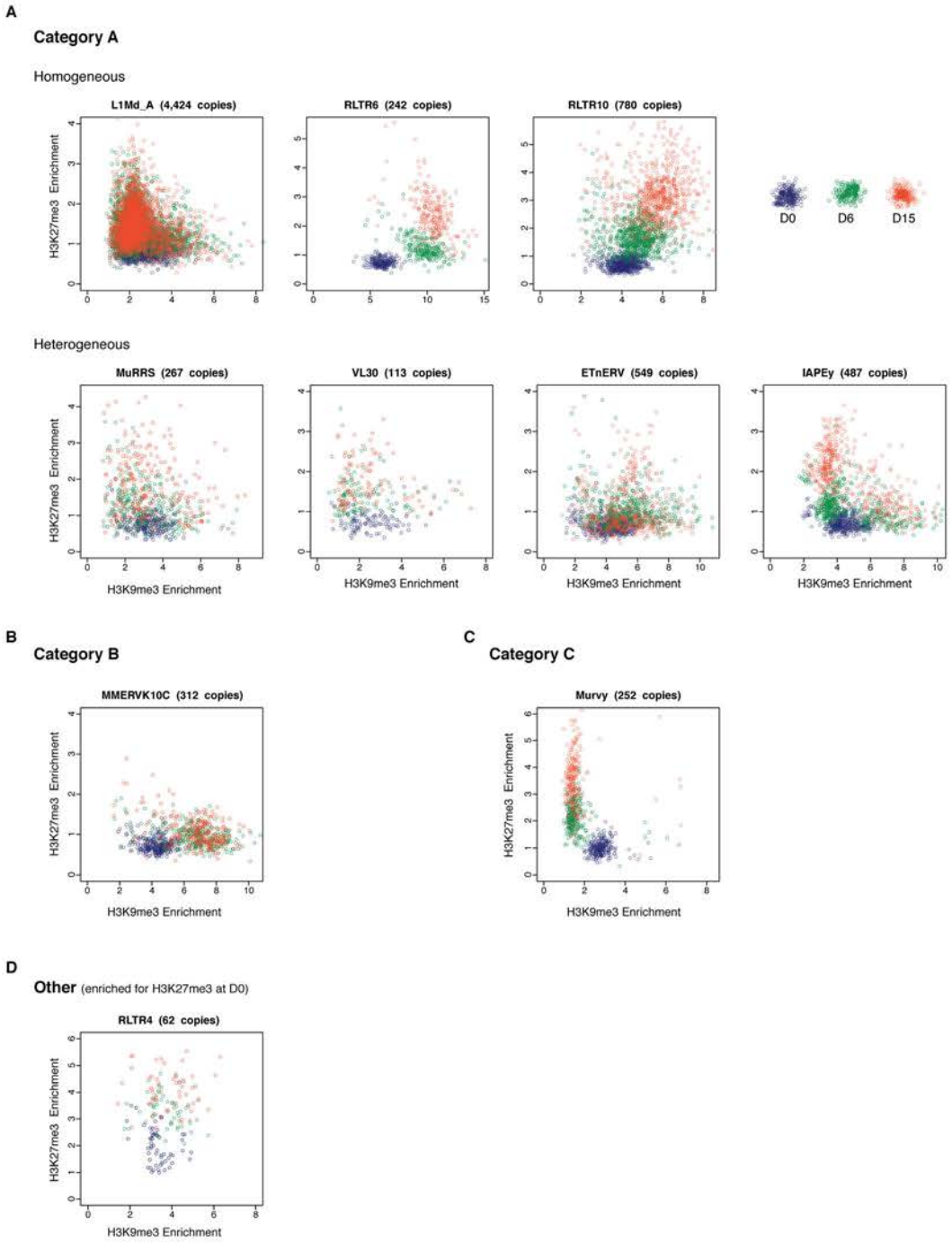


Figure 1.5 - supplement 1. Intra-familial heterogeneity of H3K9me3 and H3K27me3 in transposons.
Normalized H3K9me3 and H3K27me3 enrichment over input at individual elements from different transposon families. Each dot represents a single element at D0 (blue), D6 (green) and D15 (red). Only "intact" (integrity score>0.8) elements were considered. Data represent the average between two biological replicates. The number of distinct transposon copies considered is indicated next to the name of the family.

A. Member of the A category of transposons that are marked by both H3K9me3 and H3K27me3 in 2i+vitC

B. B category, with exclusive H3K9me3

C. C category, with exclusive H3K27me3

D. RLTR4 is the only transposon marked by H3K27me3 in serum

and **Figure 1.5 - figure supplement 2A**). It was previously described that H3K9me3 enrichment is restricted to the 5' UTR of LINEs, while being evenly distributed along the entire length of ERVs (Bulut-Karslioglu et al., 2014; Pezic et al., 2014). In fact, we found this to be valid for specific ERVK elements only, namely IAPEz, IAPEy and MMERKV10C. Our most striking finding was the observation of a spatial separation between the two marks in category A transposons: H3K9me3 tended to occupy the 5' end, while H3K27me3 preferentially targeted the 3' end. This was observed for a significant proportion of LINEs and for several ERVs of the 1 or K classes (MMERGLN, RLTR6, MuRRS, RLTR10) (**Figure 1.5C** for visual examples). However, some category A families showed H3K9me3 and H3K27me3 co-localization in their 5' region (VL30, IAPEy, ETnERV). Notably, these transposon families harbor the greatest individual chromatin heterogeneity upon conversion (**Figure 1.5 - figure supplement 1A**): we presume that the metaplot figures likely represent an average among different individual elements and/or different cell populations.

Having demonstrated that culture-induced DNA demethylation leads to increased and family-specific distribution of H3K27me3 on transposon sequences, we reasoned that similar features might occur upon genetically induced DNA demethylation. Fulfilling this prediction, analysis of available ChIP-seq datasets (Brinkman et al., 2012) showed concordant H3K27me3 patterns between serum-grown *Dnmt*-tKO ES cells and 2i+vitC-grown ES cells: entire coverage of MERVL sequences, 3' localization in MMERGLN and RLTR6, and 5' localization in VL30 and ETnERV (**Figure 1.5 - figure supplement 2B**). These results support the notion that the pattern of H3K27me3 distribution on transposon sequences corresponds to an adaptation to the lack of DNA methylation.

In summary, upon 2i+vitC-induced DNA demethylation, H3K27me3 and H3K9me3 can converge on category A transposon sequences, but they occupy different territories. IAPEz (category B) and MERVL (C) represent extreme cases of exclusivity, with the former being entirely covered by H3K9me3, and the latter by H3K27me3. Our study provides unprecedented evidence that H3K27me3 deposition at transposons is a default response to the absence of both DNA methylation and H3K9me3.

Chromatin silencing pathways play diverse roles at transposons

Our analysis reveals that H3K9me3 and H3K27me3 jointly or separately decorate transposon sequences upon DNA methylation loss. Through genetic analyses, we aimed to discern the functional relevance of these marks in controlling the three categories of transposons that we defined. Regarding H3K9me3-dependent pathways, we used

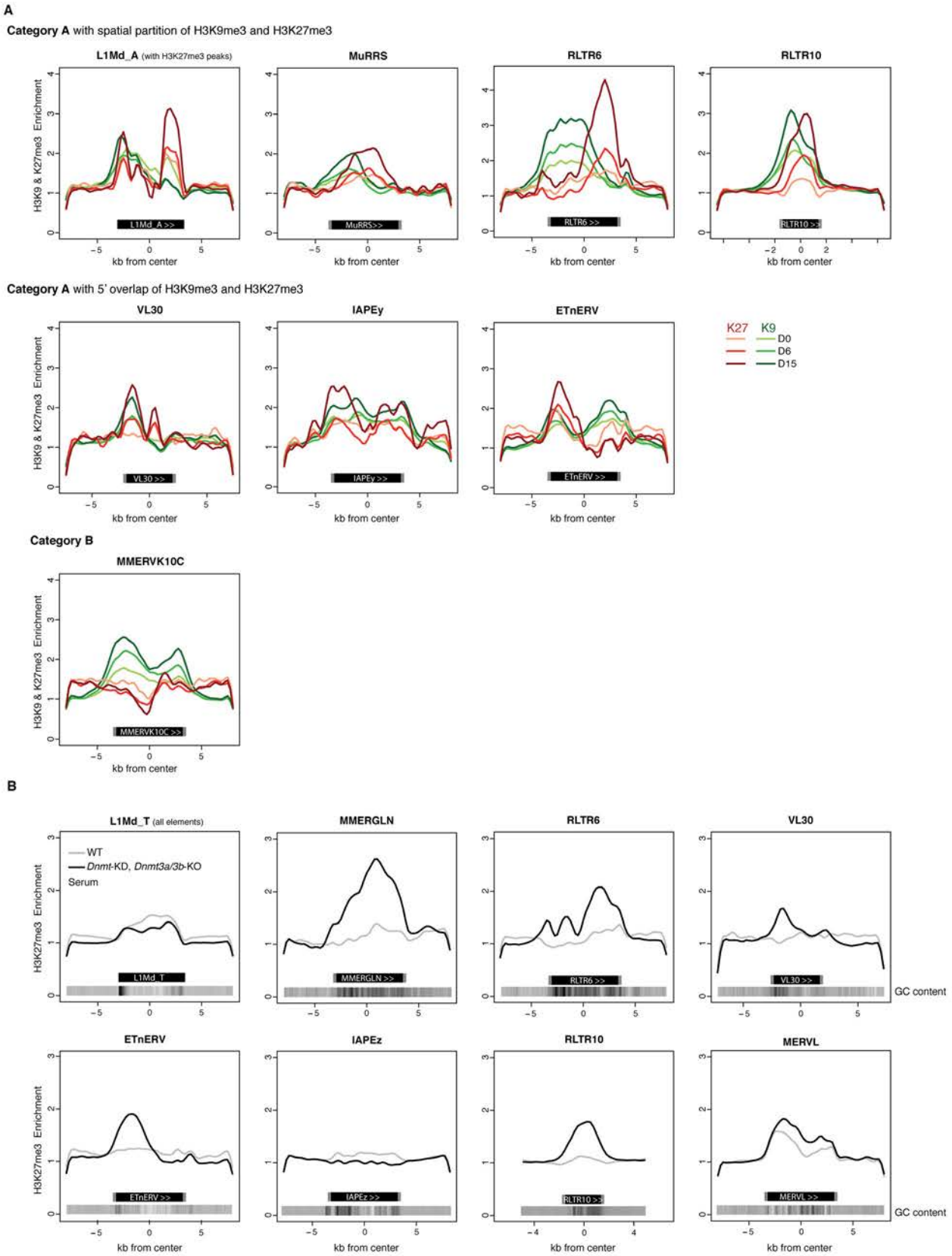


Figure 1.5 - supplement 2. H3K9me3 and H3K27me3 mark transposon sequences but do not spatially overlap.
A. Composite profile of H3K9me3 and H3K27me3 coverage on and +/- 5kb around different transposon sequences at D0, D6 and D15 of 2i+vitC conversion. Only intact (score >0.8) and full-length elements were analysed. For LINEs specifically, coverage was only represented on elements where an H3K27me3 peak was detected upon peak calling.
B. Composite profile of H3K27me3 coverage on and +/- 5kb around different transposon sequences in WT and Dnmt-tKO (*Dnmt3A/3B*-dKO; *Dnmt1* KD) DNA methylation-deficient ES cells. ChIP-seq datasets were taken from Brinkman et al., 2012.

CRISPR/Cas9 editing to generate a double-knockout ES cell line for the H3K9 trimethyltransferases, SUV39H1 and SUV39H2 (*Suvar39*-dKO). Additionally we created an haploinsufficient mutant for the KAP1 co-repressor (*Kap1+/-)*, which interacts with the H3K9 trimethyltransferase SETDB1 (**Figure 1.6 - figure supplement 1A,B**); complete KAP1 removal is not compatible with ES cell survival (Rowe et al., 2010). The role of H3K27me3 was studied in mutant ES cells for the EED protein (*Eed*-KO, Schoeftner et al., 2006), which is required for H3K27me3 catalysis by the Polycomb Repressive Complex 2 (PRC2) (Margueron and Reinberg, 2011). Nanostring quantification of transposon transcripts was performed upon serum to 2i+vitC transfer of these three cell lines, which, importantly, share the same J1 cell background.

As representatives of transposon A category, young LINE1 elements maintain H3K9me3 while gaining H3K27me3 during medium conversion. Previous studies concluded that L1 repression in serum relies on SUV39H-dependent H3K9me3 (Bulut-Karslioglu et al., 2014): through the analysis of our genetic mutants, we found this was the case for L1-A, and very modestly for L1-T elements (**Figure 1.6A**). While L1-A elements appear under exclusive SUV39H-dependent H3K9me3 control, in contrast, L1-T remained elevated in *Eed*-KO cells specifically, indicating that the gain of H3K27me3 plays a role in silencing these elements as DNA methylation is lost. The category B of transposons is exemplified by IAPEz elements, which harbor exclusive H3K9me3 enrichment in all culture conditions. Although this profile would predict a continuous and exclusive dependence towards H3K9me3 upon medium adaptation, we observed complex behaviors in the different mutants (**Figure 1.6B**). During conversion, *Kap+/-* cells showed enhanced IAPEz up-regulation and repression failure after D6, in line with a major role of SETDB1-related H3K9me3 for controlling these elements in ES cells (Matsui et al., 2010). However, SUV39H depletion led to an unexpected IAPEz suppression upon conversion. One possible explanation is that *Suvar39h*-dKO cells have acquired long-term compensatory mechanisms that prevent transient IAPEz activation upon DNA methylation loss. Moreover, IAPEz elements were more strongly expressed in *Eed*-KO compared to WT cells during conversion, which is at odds with their apparent lack of H3K27me3 enrichment in ChIP-seq data (**Figures 1.4E** and **F**). This could be due to indirect effects of the *Eed* deficiency and/or from a few H3K27me3-enriched elements, which we were unable to detect by ChIP-seq.

Finally, the H3K9me3- to H3K27me3-chromatin transition undergone by category C elements was very clearly illustrated in chromatin modifier mutants (**Figure 1.6C**). At D0, H3K9me3-driven silencing of MERVL dominantly relied on SUV39H in serum. While

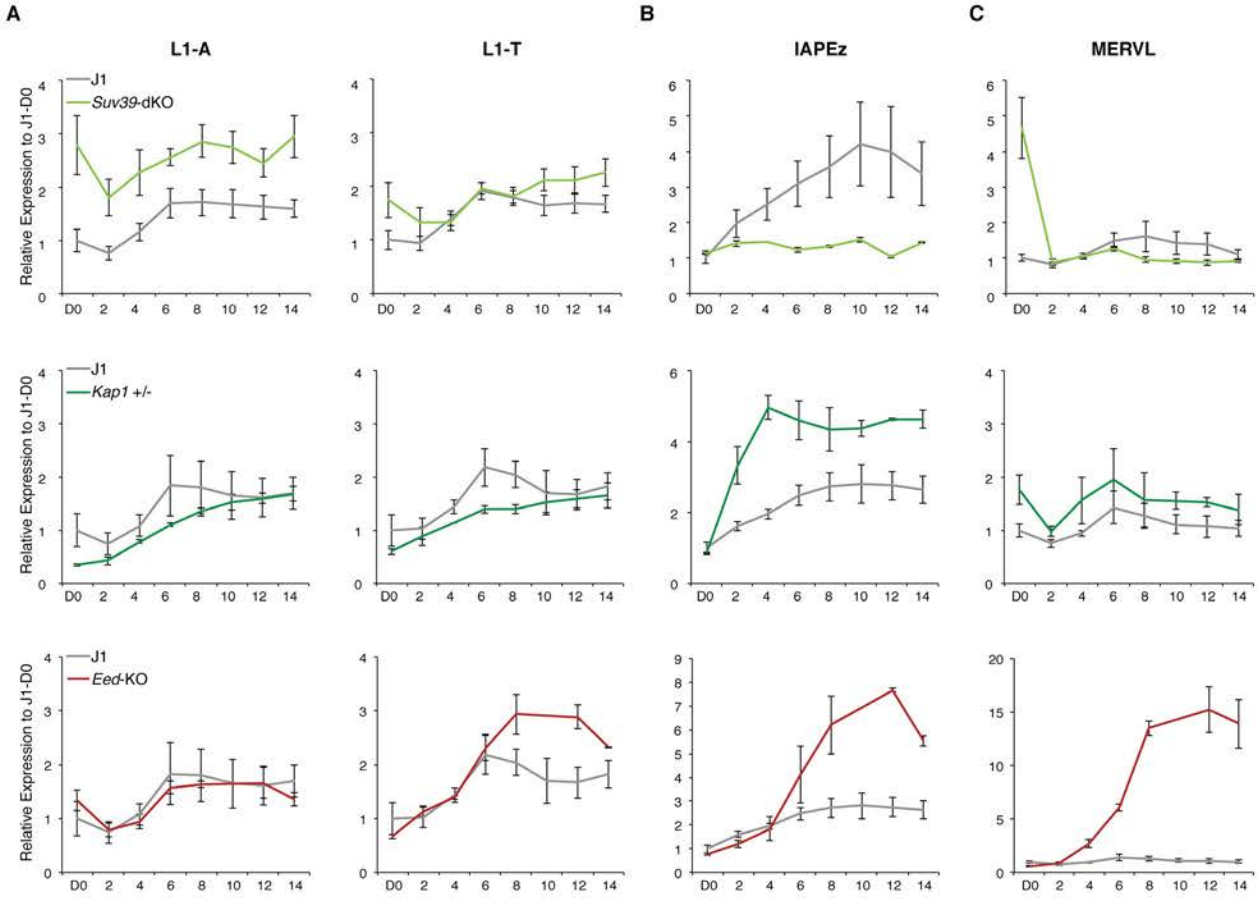


Figure 1.6. Complex regulation of transposons by SUV39H, KAP1 and EED upon loss of DNA methylation.
Expression levels in *Suv39*-dKO, *Kap1+/-*, *Eed*-KO and WT J1 ES cells for:
A. L1-A and L1-T (category A)
B. IAPEz (category B)
C. MERVL (category C)
Expression levels were measured by Nanostring nCounter. Data are expressed as fold changes to WT D0 and represent mean ± SEM between two (*Kap1* and *Eed*) or three (*Suvar39*) biological replicates.

H3K9me3 became dispensable upon conversion, the switch towards H3K27me3 control was perfectly correlated with a 15-fold expression increase we observed in *Eed*-KO cells. MERVL represent a striking model of epigenetic switch, which occurs subsequently to DNA methylation loss.

1.5 Discussion

Our study provides unprecedented insight into the dynamic adaptation of the pluripotent genome to a loss of DNA methylation-based control of transposons. This was achieved through detailed kinetic assessment of transcription and chromatin states during conversion of WT ES cells from serum to 2i+vitC media, as a way to reproduce the DNA methylation erasure that occurs during embryogenesis. Despite their heterogeneous origins and structures, we found that various transposon families residing in the mouse genome adopted a common regulatory fate: after an initial transcriptional burst, repression was re-established in a DNA methylation-independent manner. Distinct combinations of H3K9me3 and H3K27me3 were observed among transposon families, defining three functional categories of chromatin-based responses to DNA methylation loss: joint H3K9me3 and H3K27m3 (A), H3K9me3-exclusive (B), and H3K27me3-exclusive (C) (**Figure 1.7**). Importantly, *Dnmt*-tKO cells, which have endured long-term adaptation to a DNA methylation-free state, displayed similar transposon-specific chromatin patterns when grown in serum, which excludes a medium-related effect. In conclusion, our work revises the previous assumption that DNA methylation is dispensable for transposon silencing in ES cells; rather, we reveal here that various histone-based repression strategies are implemented upon DNA methylation loss, thereby safeguarding pluripotent cells against a multitude of heterogeneous transposon entities.

Upon 2i+vitC-mediated DNA methylation loss, the repertoire of repressive histone marks of transposons is profoundly remodeled (**Figure 1.7A**): H3K9me2 enrichment decreases, while H3K27me3 is enhanced. In contrast, H3K9me3 levels are globally constant: this highlights that DNA methylation does not exert significant control over H3K9me3-targeting of transposons in ES cells. Interestingly, we consistently observed that regions of persistent DNA methylation (RMRs) coincide with high H3K9me3 enrichment on transposon sequences in fully 2i+vitC-converted cells, *e.g.* on the 5' end of LINE1 elements and throughout the length of ERVK elements. This supports previous evidence that H3K9me3 can confer protection against DNA demethylation (Leung et al., 2014). Inversely, the rapid

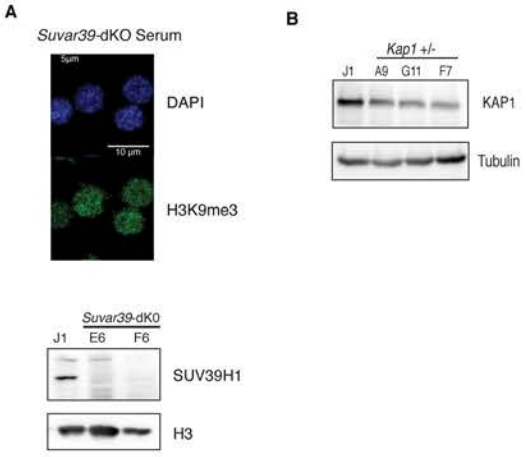


Figure 1.6 – Figure Supplement 1. Caracterization of *Suvar39*-dKO and *Kap1+/-* cells
A. Characterization of *Suvar39*-dKO ES cells: loss of H3K9me3 at pericentric heterochromatin assessed by immunofluorescence (upper panel, compare with S4A at D0 for WT) and lack of SUV39H1 protein assessed western blot (lower panel).
B. Characterization of CRISPR/Cas9- generated *Kap1+/-* ES cells: western blot showing the level of reduction of KAP1 protein in three independent clones. Clone F7 was used for subsequent experiments.

disappearance of H3K9me2 upon serum to 2i+vitC conversion could reflect a direct role of DNA methylation in the maintenance of these marks. Accordingly, H3K9me2 reduction was also observed in DNA methylation-free *Dnmt*-tKO ES cells grown in serum (data not shown). Coupled losses of DNA methylation and H3K9me2 have also been previously reported *in vivo*, during normal primordial germ cell development (Hajkova et al., 2008; Seki et al., 2005) and in DNA methylation-deficient spermatocytes (Zamudio et al., 2015). A mechanism of H3K9me2 methyltransferase recruitment via DNA methylation has been resolved in plants (Du et al., 2015); the evolution of an analogous mechanism in mammals should be explored.

Of particular importance to this study is our observation of an epigenetic switch occurring between DNA methylation- and H3K27me3-based control. H3K27me3 is barely detectable at transposons in DNA hypermethylated WT ES cells grown in serum; in contrast, transposons accumulate H3K27me3 in both 2i+vitC-converted cells and in serum-grown *Dnmt*-tKO cells. This is in line with the prevailing notion that DNA methylation and H3K27me3 are mutually exclusive genome-wide and that DNA methylation antagonizes H3K27me3 deposition (Brinkman et al., 2012; Jermann et al., 2014; Tanay et al., 2007). Saliently, this raises the question as to how transposons acquire H3K27me3 upon DNA methylation loss. In mammalian genomes, polycomb is typically targeted to unmethylated GC-rich gene promoters (Jermann et al., 2014; Mendenhall et al., 2010). Notably, transposon sequences have a GC content superior to the genome average (**Figure 1.5 - figure supplement 2B**): this signature could be sufficient to attract polycomb-mediated H3K27me3 deposition in the absence of DNA methylation. Intermediate methyl-sensitive DNA binding proteins may be involved: the BEND3 protein was recently identified as a sensor of DNA methylation states at pericentromeric repeats, recruiting polycomb-dependent H3K27me3 marks in *Dnmt*-tKO ES cells (Saksouk et al., 2014). Interestingly, we also observed H3K27me3 relocalization towards pericentromeric repeats in hypomethylated 2i+vitC ES cells. Comparable mechanisms might be at play for the recruitment of H3K27me3 at hypomethylated transposons, involving BEND3 and/or other methyl-sensitive DNA binding proteins.

Thus, based on previous observations, we posit that H3K27me3 invades the transposon space left unmarked by DNA methylation upon 2i+vitC conversion. Moreover, we provide evidence that the three possible chromatin configurations that the different transposon families adopt are further determined by H3K9me3 occupancy (**Figure 1.7B**). Mutual exclusion between H3K9me3 and H3K27me3 marks has been previously documented at gene promoters and pericentromeric repeats (Mikkelsen et al., 2007; Peters et

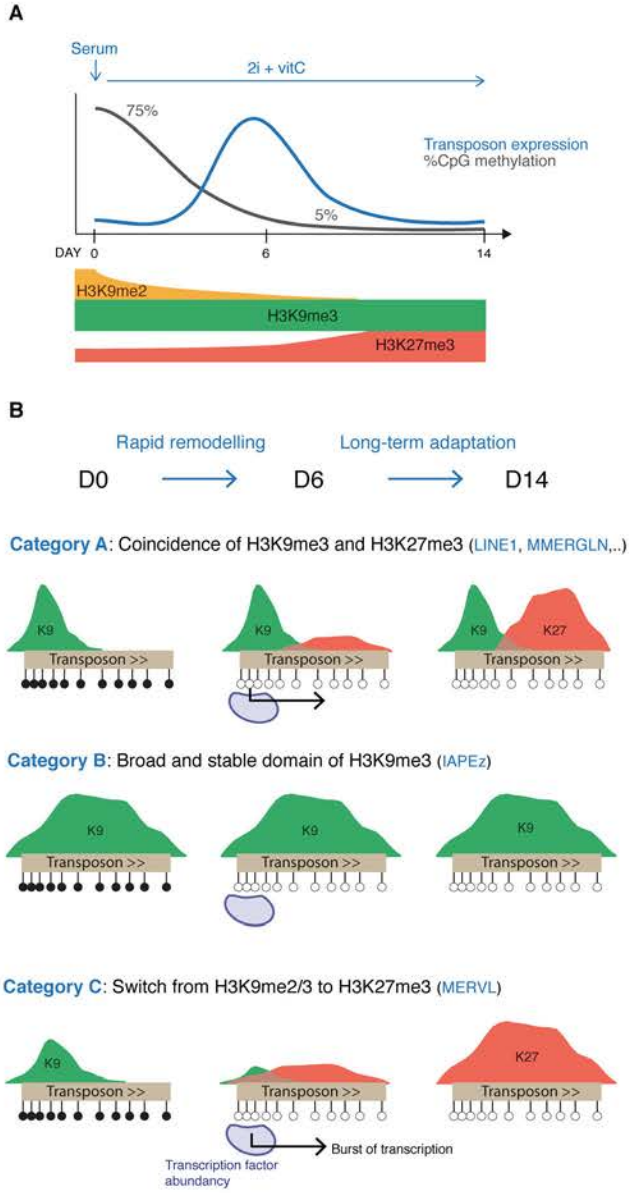


Figure 1.7. Model for the acquisition of H3K27me3 at transposons during genome-wide demethylation.
A. Summary of chromatin and transcriptional changes during conversion from serum to 2i+vitC. DNA methylation and H3K9me2 are rapidly erased, H3K9me3 remains stable and H3K27me3 increases. Transposon expression peaks at D6.
B. Model for the acquisition of H3K27me3 at transposons: upon loss of DNA methylation, H3K27me3 appears at GC-rich, H3K9me3 free-regions. Relative enrichments in H3K9me3 and H3K27me3 define three main types of repressive chromatin organization. Category A transposons are marked by H3K9me3 on their 5' end and gain H3K27me3 on their 3' region. Category B transposons are fully covered by H3K9me3 and do not gain H3K37me3. Category C transposons lose H3K9me2 and H3K9me3 and acquire H3K27me3 decoration on their full length. At D6 of 2i+vitC conversion, the abundance of pluripotent transcription factors and the loose chromatin environment likely contribute to the burst of transposon expression.

al., 2003), but not at transposons. We found that category B transposons, which constantly maintain H3K9me3 marks throughout their entire length, do not acquire H3K27me3-based chromatin even though they lose DNA methylation. In contrast, category C transposons, exemplified by MERVL elements, become strongly enriched for H3K27me3 as H3K9me3 depletes during medium conversion. Finally, category A elements provide a striking illustration of the physical segregation of H3K9me3 and H3K27me3: as H3K9me3 constitutively marks the 5' end of this transposon category, only their 3' end is accessible to H3K27me3 deposition upon DNA methylation loss. The presence of H3K27me3 at the 3' end of transcription units has not been described before, and its functional relevance remains to be investigated.

The main message conveyed by our work is that compensatory histone-based mechanisms ensure transposon silencing when DNA methylation-based control is alleviated in ES cells. We cannot rule out that other mechanisms-such as small RNA-based post-transcriptional repression-could also participate to transposon control; but importantly, genetic analyses globally confirmed the functionality of the chromatin patterns that we identified. In particular, H3K27me3-dependency of transposon categories A (LINE1 T) and C (MERVL) was very well illustrated by the failure to repress these elements in *Eed*-KO ES cells undergoing medium-based DNA methylation loss. However, the transposon category A (IAPEz), which remains enriched for H3K9me3 throughout media conversion, gave complex, disparate phenotypes in the mutants of the different H3K9me3 pathways. While these elements failed to be repressed in *Kap1*-deficient ES cells, the complete suppression of IAPEz reactivation in *Suvar39*-dko cells was unexpected. We suspect that alternative repressive processes likely obscure IAPEz transcriptional responses to DNA methylation loss in this mutant. This is akin to *Dnmt*-tKO cells, which also exhibit global transposon repression. Thus, our analyses highlight the possible unexplained phenotypes of mutant cells that have adapted to long-term chromatin-based deficiencies.

Finally, one important point to raise is that the epigenetic switch from a DNA methylation-dependent to -independent mode of transposon silencing is not perfectly synchronized: ES cells experience an acute burst of transposon expression at D6 of medium conversion. At this time point, we showed that DNA methylation has been mostly erased but H3K27me3 patterns have not been established yet. Interestingly, the stability of H3K9me3 marks at category A and B transposons is not sufficient to ensure their continuous silencing upon conversion. This may imply that H3K9me3 readers are transiently deficient in this system. The lag between DNA methylation loss and subsequent implementation of histone-

based repression could create an opportunistic window for transposon reactivation, provided that adequate transcription factors are available. Several studies have previously pointed out that transposons are enriched in pluripotency transcription factor binding motifs (Kunarso et al., 2010; Wang et al., 2014a), in particular for NANOG and OCT4, and that upregulation of these transcription factors was sufficient to promote transposon expression (Grow et al., 2015).

We propose that the simultaneous disappearance of DNA methylation marks and increased availability of pluripotency activators create favorable conditions to transposon expression at D6 of serum to 2i+vitC conversion (Figure 7B). After a brief silencing release, functional repressive chromatin is recovered, in an H3K9me3 and/or H3K27me3-dependent manner. Notably, we repeatedly observed a peak of massive cell death of H3K27me3-deficient *Eed*-KO ES cells between D6 and D10 of medium conversion, when DNA methylation has mostly disappeared (data not shown). This observation supports the critical role for H3K27me3 in supplementing DNA methylation-based control in ES cells.

1.6 Experimental procedures

ES cell lines

J1 and *Dnmt*-tKO ES cells were a gift from M. Okano (Tsumura et al., 2006). E14 ES cells were kindly provided by E. Heard. WT TT2 and *G9a*-KO ES cells (Tachibana et al., 2002), and *Eed*-KO (Schoeftner et al., 2006) (on a J1 background) were gifts from Y. Shinkai and A. Wutz, respectively. *Kap1-/+* and *Suvar39*-dKO were generated from J1 ES cells using CRISPR/Cas9 editing. Briefly, guide-RNAs specific to the target sequences were designed using the online CRISPR Design Tool (Hsu et al., 2013)(**Table S1.6**) and incorporated into the X330 backbone (Cong et al., 2013). Five millions J1 ES cells grown in serum were transfected with 1-3µg of plasmid using Amaxa 4d nucleofector (Lonza) and plated at a low density. Individual clones were picked and screened by PCR; mutated alleles were confirmed by Sanger sequencing. *Suvar39*-dKO cells were obtained by creating a frame-shift in *Suvar39h1* exon 4 and by deleting *Suvar39h2* exon 4; *Kap1+/-* cells were generated by deleting exon 3.

ES cell culture

ES cells were grown in two different media, serum and 2i, defined as follow. Serum: Glasgow medium (Sigma), 15% FBS (Gibco), 2mM L-Glutamine, 0.1mM MEM non essential amino acids (Gibco), 1mM sodium pyruvate (Gibco), 0.1mM β-mercaptoethanol, 1000U/mL leukemia inhibitory factor (LIF, Miltenyi); 2i: 50% neurobasal medium (Gibco), 50%

DMEM/F12 (Gibco), 2mM L-glutamine (Gibco), 0.1mM β-mercaptoethanol, Ndiff Neuro2 supplement (Milipore), B27 serum-free supplement (Gibco), 1000U/mL LIF, 3μM Gsk3 inhibitor CT-99021, 1μM MEK inhibitor PD0325901. Vitamin C (Sigma) was added at a concentration of 100ug/mL (Blaschke et al., 2013).

When in serum, J1, *Dnmt*-tKO, E14, *Kap-/+* and *Suvar39*-dKO ES cells were grown in feeder-free conditions on gelatin-coated plates. TT2, *G9a*-KO and *Eed*-KO were cultured on a monolayer of mitomycin C-treated mouse embryonic fibroblasts. ES cells were passaged with TrypLE Express Enzyme (Life Technologies). All 2i ES cells were grown in gelatin-coated plates and passaged every two or three days with Accutase (Life Technologies).

DNA methylation analyses

Genomic DNA was isolated using the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma) with RNase treatment. Global CpG methylation levels were assessed using LUminometric Methylation Assay (LUMA) as described previously (Karimi et al., 2011a; Richard Pilsner et al., 2010). Briefly, 500ng of genomic DNA was digested with *MspI/EcoRI* and *HpaII/EcoRI* (NEB) in parallel reactions. *HpaII* is a methylation-sensitive restriction enzyme and *MspI* is its methylation insensitive isoschizomer. *EcoRI* is included as an internal reference. The overhangs created by the enzymatic digestion were quantified by Pyrosequencing (PyroMark Q24, Qiagen) with the dispensation order: GTGTGTCACACAGTGTGT. Global CpG methylation levels were calculated from the peak heights at the position 7,8,13,14 as follows: 1-sqrt($[p8*p14/p7*p13]_{HpaII}$ /$[p8*p14/p7*p13]_{MspI}$)

CpG methylation at specific loci was assessed by bisulfite-pyrosequencing using the Imprint DNA modification Kit (Sigma) for conversion. PCR and sequencing primers (**Table S1.6**) were designed with the PyroMark Assay Design Software and quantification of DNA methylation was performed according to the recommended protocol.

Whole-Genome Bisulfite Sequencing libraries were prepared from 50ng of bisulfite-converted genomic DNA using the EpiGnome/Truseq DNA Methylation Kit (Illumina) following the manufacturer instructions. Sequencing was performed in 100pb paired-end reads at a 30X coverage using the Illumina HiSeq2000 platform (**Table S1.1C**).

RNA expression analyses

Total RNA was extracted using Trizol (Life Technologies). cDNAs were prepared after DNase treatment (Turbo DNase, Ambion) using random priming with Superscript III (Life Technologies). Real-time quantitative PCR was performed using the SYBR Green Master

Mix on the Viia7 thermal cycling system (Applied Biosystem). Relative expression levels were normalized to the arithmetic mean of the housekeeping genes *Gapdh* and *Arp0* and to WT-D0 with the ΔΔCt method. Primers are given in Table S4.

Nanostring nCounter quantification was performed using 100ng of total RNA per sample on a custom expression Codeset (sequences available upon request). *Actin, Ppia, Gapdh* and *Arp0* were used for normalization. Data are presented as the fold change compared to WT-D0. Expression data for the different mutants are presented next to WT data that were processed on the same Nanostring run. The same WT data can be used in several figures. When necessary and in order to calculate mean and standard error of the mean between replicates every two days, we extrapolated linearly the expression value of a given day using data of immediately adjacent time points (for both RT-qPCR and Nanostring).

RNA-seq libraries were prepared from 500ng of DNase-treated RNA with the TruSeq Stranded mRNA kit (Illumina). Sequencing was performed in 100pb paired-end reads using the Illumina HiSeq2000 platform (**Table S1.1B**).

Chromatin Immunoprecipitation (ChIP)

Cells were cross-linked directly in culture plates with 1% formaldehyde (culture medium supplemented with 1% formaldehyde, 0.015M NaCl, 0.15mM EDTA, 0.075mM EGTA, 15mM Hepes pH 8). After quenching with 0.125M glycine, cells were washed in PBS and pelleted. Cells were then incubated at 4°C for 10 min in buffer 1 (Hepes-KOH pH 7.5 50mM, NaCl 140mM, EDTA pH 8.0 1mM, glycerol 10% NP-40 0.5%, Triton X-100 0.25% and the protease inhibitors: PMSF 1mM, Aprotinin 10µg/ml, leupeptin 1µg/ml and pepstatin1µg/ml), then at room temperature for 10 min in buffer 2 (NaCl 200mM, EDTA pH 8.0 1mM, EGTA pH 8.0 0.5mM and 10mM Tris pH 8 and the same protease inhibitors as buffer 1) and finally resuspended in buffer 3 (EDTA pH 8.0 1mM, EGTA pH 8.0 0.5mM, Tris pH8 10mM, N-lauroyl-sarcosine 0.5%; protease inhibitors as buffer 1). Chromatin was sonicated with a Bioruptor (Diagenode) to reach a fragment size around 200bp. Chromatin corresponding to 10µg of DNA was incubated overnight at 4°C with 3-5µg of antibody in incubation buffer (buffer 3 supplemented with 0.5 volume of 3% Triton, 0.3% sodium deoxycholate, 15mM EDTA; protease inhibitors). A fraction of chromatin extracts (5%) were taken aside for inputs. Antibody-bound chromatin was recovered using Protein G Agarose Columns (Active Motif). Briefly, the antibody-chromatin mix was incubated in the column for 4 hours, washed eight times with modified RIPA buffer (Hepes pH7.6 50mM, EDTA pH 8.0 10mM, sodium deoxycholate 0.7%, NP-40 1%, LiCl 500mM, PMSF 1 mM, 1µg/ml

leupeptin and 1µg/ml pepstatin), and washed one last time with TE-NaCl (50mM Tris pH 8.0, 10mM EDTA, 50mM NaCl). Chromatin was eluted with pre-warmed TE-SDS (50mM Tris pH 8.0, 10mM EDTA, 1% SDS). ChIP-enriched sample and inputs were then reverse cross-linked at 65°C overnight and treated with RNase A and proteinase K. DNA was extracted with phenol/chloroform/isoamyl alcohol, precipitated with glycogen in sodium acetate and ethanol and finally resuspended in TE. Enrichment compared to input was analyzed by qPCR. A quantity of 20ng of ChIP- or input-DNA were used for ChIP-seq. Remaining large DNA fragments were first eliminated using SPRIselect beads (Beckman Coulter) and libraries were prepared using the TruSeq ChIP Sample Prep kit (Illumina). Sequencing was performed in 50pb paired-end reads using the Illumina HiSeq2000 platform (**Table S1.1C**). Antibodies are listed in **Table S1.7**

Western blotting

To prepare total protein extracts, cells were resuspended in BC250 lysis buffer (25mM Tris pH 7.9, 0.2mM EDTA, 20% Glycerol, 0.25M KCl and protease inhibitor coktail from Roche), sonicated and centrifuged to pellet debris. To prepare nuclear protein extracts, cells were incubated for 10 min on ice in buffer A (Hepes pH 7.9 10mM, MgCl2 5mM, Sucrose 0.25M, NP40 0.1%, DTT 1mM and protease inhibitors) and centrifuged. The pellet was resuspended in buffer B (Hepes pH 7.9 25mM, glycerol 20%, MgCl2 1.5mM, EDTA 0.1 mM, NaCl 700mM, DTT 1mM and protease inhibitors), sonicated and centrifuged to pellet debris. Total and nuclear proteins were quantified by Bradford assay. Proteins (10-20µg per gel lane) were separated by electrophoresis in 8-15% poly-acrylamide gels and transferred onto nitrocellulose membranes using the Trans-Blot turbo transfer system (Biorad). After incubation with primary antibodies and HRP-conjugated secondary antibodies, signal was detected using ECL prime kit (Amersham) and ImageQuant Las-4000 mini biomolecular Imager. Antibodies are listed in **Table S1.7.**

Immunofluorescence

Cells were harvested with Trypsin or Accutase, resuspended in PBS and plated for 10min on Poly-L-Lysine-coated glass cover slips. Cells were first fixed with 3% paraformaldehyde for 10 min at room temperature, then rinsed three times with PBS and permeabilized for 4 min with 0.5X Triton on ice. After blocking in 1% BSA for 15 min, samples were incubated at room temperature for 40 min with primary antibodies, 45 min with secondary antibodies and 3 min in 0.3µg/mL DAPI. Slides were mounted with Prolong

Gold mounting media (Invitrogen). Images were obtained with an Upright Widefield microscope (Leica) or a Zeiss LSM700 inverted confocal microscope. Quantification of immunofluorescence intensity in individual cells was performed using custom ImageJ and R scripts. Between 1,000 and 5,000 cells were analyzed per sample.

Metaphase spreading

Cells were cultured for two hours with 0.04µg/mL colchicine and harvested by trypsinization. Cell pellets were incubated in hypotonic buffer (15% FBS in water) for 7 min at 37°C and fixed with 66% acetic acid/33% ethanol. After centrifugation, cells were resuspended in 1.5mL fixative and dropped from ~1m height onto glass slides. Slides were dried and DNA was stained with DAPI. Chromosomes were counted with an Upright Widefield microscope (Leica). Around 20 cells were analyzed per cell line.

Quantification of transposon genomic copy number

Absolute copy numbers of IAP and LINE1 were calculated by qPCR by establishing standard curves plotting absolute Ct values of genomic DNA against serial dilutions of PCR targets cloned into the pCR2.1-TOPO vector (Life Technologies), as described in Zamudio et al., 2015.

Reconstruction of RepeatMasker

As described in Bailly-Bechet et al., 2014, a dictionary was constructed for LTR retrotransposons that associated elements corresponding to the internal sequence and those corresponding to LTR sequences. With the latter and the RepeatMasker database, fragments of transposable elements corresponding to the same copy were merged. Divergence, deletion and insertion percentages were recalculated from RepeatMasker and an integrity score for each transposon were calculated as follow: score = 1-average(%divergence, %deletions, %insertions)

WGBS data analysis

Whole-genome bisulfite sequencing reads generated in this study or recovered from available datasets were treated as follow. The first eight base pairs of the reads were trimmed using FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Adapter sequences were removed with Cutadapt v1.3 (https://code.google.com/p/cutadapt/) and reads shorter than 16bp were discarded. Cleaned sequences were aligned onto the Mouse reference genome (mm10) using Bismark v0.12.5 (Krueger and Andrews, 2011) with Bowtie2-2.1.0 (Langmead and Salzberg, 2012) and default parameters. Only reads mapping uniquely on the

genome were conserved. Methylation calls were extracted after duplicate removal. Only CG dinucleotides covered by a minimum of 10 reads were conserved for the rest of the analysis.

The R-package Methylkit v0.9.2 (Akalin et al., 2012) was used to provide Pearson's correlation scores between samples. To analyze the distribution of CpG methylation in different genomic compartments, the mouse genome was divided into different partitions. The RefSeq gene annotation and the RepeatMasker database were downloaded from UCSC table browser and used for transcript and repeat annotations, respectively. Promoters were defined as the -1kb to +100pb region around transcription start sites. CpG islands (CGIs) were defined as in Illingworth et al., 2010. Intergenic partitions were defined as genomic regions that did not overlap with promoters, CGI, exons, introns or repeats. Whole-genome mapping of CpG methylation was then intersected with the different genomic compartments using Bedtools (Quinlan and Hall, 2010).

Average CpG methylation on individual transposons was extracted from RepeatMasker with Bedtools, average CpG methylation in the different transposon families was calculated and plotted using R. Heatmap for average CpG methylation in Imprinted control regions (ICRs) was generated similarly after retrieving ICR genomic coordinates from the WAMIDEX database (Schulz et al., 2008). Residually methylated regions (RMRs) in 2i+vitC samples were identified using the MethPipe pipeline (Song et al., 2013) with default parameters. RMRs located less than 1kb from each others were concatenated.

RNA-seq data analysis

In order to quantify gene expression, Paired-end 2x100bp reads were mapped onto mm10 using Tophat v2.0.6 and RefSeq gene annotation (Kim et al., 2013) allowing five mismatches. Gene-scaled quantification was performed with HTSeq v0.6.1 (Anders et al., 2014).

In order to quantify transposon expression, reads mapping to ribosomal RNA (rRNA) sequences (GenBank identifiers: 18S NR_003278.3, 28S NR_003279.1, 5S D14832.1, 5.8S K01367.1) were first removed with Bowtie v1.0.0 allowing three mismatches. The rRNA-depleted libraries were then mapped onto mm10 using Bowtie v1.0.0 allowing zero mismatch and 10,000 best alignments per read. Exonic reads were removed. In order to count reads mapping to transposable elements, reads were weighted by the number of mapping sites and each library was intersected with the reconstructed RepeatMasker annotation, conserving only reads overlapping at least at 80% with a given transposon.

For each library, read counts for genes and transposons were combined into a single table. TMM normalization from the edgeR package v3.6.2 (Robinson and Oshlack, 2010) was first applied. As described in the guideline of limma R-package v3.20.4, normalized counts were processed by the voom method (Law et al., 2014) to convert them into log2 counts per million with associated precision weights. The differential expression was estimated with the limma package. Genes and transposons were called differentially expressed when two criteria were met: 1) the fold-change between two conditions was higher than four and two, respectively, and 2) the adjusted p-value using the Benjamini Hochberg procedure was below 0.05.

For the analysis of RNA-seq libraries with uniquely mapped reads, the mapping was performed as previously with Bowtie v1.0.0, except that only uniquely mapping reads were conserved. Read counts on individual reconstructed element were quantified using HTSeq v0.6.1. Only elements with at least 10 reads in at least one sample were conserved for further analysis and read counts were subsequently normalized by the library size. Normalized read counts for individual elements belonging to different families were then plotted using custom R script. Tracks were created using HOMER software v4.7 (Heinz et al., 2010).

In order to identify and characterize chimeric transcripts, reads were mapped onto mm10 using Tophat v2.0.6, without providing a gene annotation. Cufflinks v2.2.1 (Trapnell et al., 2010) was used to reconstruct the transcriptome and quantify the different isoforms. Transcripts were considered chimeric when the first exon overlapped with a transposon annotated in Repeatmasker and one of the other exon was annotated in RefSeq.

ChIP-seq data analysis

Paired-end 2x50bp reads were mapped onto mm10 using Bowtie v1.0.0 allowing 3 mismatches. Reads mapping to multiple locations were randomly allocated. Duplicate reads were removed using Picard v1.65 (http://broadinstitute.github.io/picard/). Tracks were created using HOMER software v4.7 (Heinz et al., 2010) and Peak calling was performed with MACS2 v2.0.10 (Zhang et al., 2008b) using the broad option and a 5% FDR threshold. Detected peaks were annotated using RefSeq and RepeatMasker databases. In order to construct the heatmap and the scatter plots, the total number of read counts for every annotated transposable element was computed using Bedtools and the reconstructed RepeatMasker annotation. Enrichment was normalized by the size of the element and Input data. Metaplots for average enrichment and GC content on and around different transposons

were obtained using HOMER V4.7. Only full-length (>6kb) and intact (integrity score >0.8) elements were used for the metaplots.

1.7 Annexes

Accession numbers

All next generation sequencing data have been deposited to the NCBI Gene Expression Omnibus (GEO) database under ######.

Author contribution

M.W. and D.B conceived the study, analyzed the data and wrote the manuscript. M.W. performed all the experiments, at the exception of the immunofluorescence studies, which were carried out by R.P.-P. Bioinformatics analyses were conducted by A.T. and M.W.

Acknowledgments

We thank the members of D.B's laboratory, especially M. Greenberg, J. Barau and T. Chelmicki for critical input and experimental help. We thank M. Okano, A. Wutz, Y. Shinkai and E. Heard for the gift of mouse ES cells lines; E. Heard, G. Almouzni, R. Margueron, B. Cullen and A. Bortvin for antibodies. We acknowledge the PICTIBiSA@BDD for microscopy; the Institut Curie NGS platform supported by the ANR-10-EQPX-03 and ANR10-INBS-09-08 grants and the Canceropôle Ile-de-France for high-throughput sequencing; the Genomic Platform for Nanostring ncounter analysis. D.B.'s laboratory is part of the Laboratoire d'Excellence (LABEX) entitled DEEP (11-LBX0044). This research was supported by grants from the European Research Council (ERC) and ANR ("ABS4NGS"-ANR-11-BINF-0001). M.W. is recipient of a PhD fellowship from the Ecole Polytechnique.

1.8 Supplemental tables

#Sample identifier	Biological identifier	Number of total sequenced tags	Number of cleaned sequenced tags	Number of paired-end alignments with a unique best hit	Mapping efficiency	Total count of deduplicated leftover sequences
ERR192357	J1 Serum	242714563	241941728	108882056	45,00342165	86908726
GSM1027571	E14 Serum	467149751	466901697	316209649	67,72510167	273132665
GSM1027572	E14 2i	466450045	465957183	301890930	64,78941435	255104546
A274-B113T1	J1 2i+vitC	447494671	447404318	270945884	60,55951476	190029868

#Sample identifier	meC in CpG context	meC in CHG context	meC in CHH context	%CpGs covered by at least 1x	% covered by at least 5x	% covered by at least 10x	GEO number	Publication
ERR192357	77,1	2,2	2,3	72,2	31,0	6,9	ERP001953	Seisengerber 2012
GSM1027571	71,3	0,7	0,6	81,7	63,4	47,2	GSE41923	Habibi 2013
GSM1027572	30,3	0,1	0,2	80,9	61,7	44,8	GSE41923	Habibi 2013
A274-B113T1	4,6	0,7	0,8	81,1	55,3	33,4		this study

Table S1.1A – WGBS sequencing statistics.

#Sample identifier	Biological identifier	Number of total reads	Number of mapped reads	% of mapped reads
B99T1	J1 D0	213459951	151096208	70,78%
B99T2	J1 D6	210340434	142149513	67,58%
B99T3	J1 D13	204453602	139059777	68,02%
B99T4	J1 D0	230009813	159157855	69,20%
B99T5	J1 D6	229534739	156779528	68,30%
B99T6	J1 D13	216338499	138319471	63,94%

Table S1.1B – RNA sequencing statistics.

#Sample identifier	Biological identifier	Number of total reads	Number of mapped reads	% of mapped reads	% of duplicates
A291-A292C1	Input_K9-1	75468914	69370052	91,92%	7,87%
A291-A292C2	ChIP_K9-1_D0	79172270	69473090	87,75%	20,51%
A291-A292C3	ChIP_K9-1_D6	84177608	73522216	87,34%	26,32%
A291-A292C4	ChIP_K9-1_D15	75428406	66968060	88,78%	17,82%
A291-A292C5	Input_K9-2	83496606	75945018	90,96%	10,37%
A291-A292C6	ChIP_K9-2_D0	75060252	65940744	87,85%	20,19%
A291-A292C7	ChIP_K9-2_D6	87029208	76798684	88,24%	22,30%
A291-A292C8	ChIP_K9-2_D15	85637102	75413994	88,06%	23,67%
A291-A292C9	Input_K27-1	86256562	78822332	91,38%	9,87%
A291-A292C10	ChIP_K27-1_D0	82339150	75835212	92,10%	7,19%
A291-A292C11	ChIP_K27-1_D6	89508604	82104576	91,73%	7,61%
A291-A292C12	ChIP_K27-1_D15	87891782	79856614	90,86%	10,95%
A291-A292C13	Input_K27-2	79127682	72111116	91,13%	9,94%
A291-A292C14	ChIP_K27-2_D0	69144980	63468582	91,79%	7,30%
A291-A292C15	ChIP_K27-2_D6	67961062	62111844	91,39%	6,61%
A291-A292C16	ChIP_K27-2_D15	70720000	63983318	90,47%	8,20%

Table S1.1C – ChIP sequencing statistics.

	number (and percentage) of unique elements expressed at				total number of elements
	D0	**D6**	**D13**	**throughout conversion**	
All LINE1s	4,334 (0.7%)	5,659 (1%)	2,303 (0.4%)	7,163 (1.2%)	588,739
All ERVs	1,229 (2.0%)	1,925 (3.1%)	1,381 (2.2%)	2,372 (3.8%)	62,098
L1Md_A	1,068 (8.1%)	1,620 (12.3%)	425 (3.2%)	1,891 (14.4%)	13,172
L1Md_T	1,631 (9.2%)	2,129 (12.0%)	438 (2.5%)	2,447 (13.8%)	17,740
IAPEz	45 (1.5%)	127 (4.3%)	63 (2.1%)	174 (5.9%)	2,958
IAPEy	9 (1.8%)	66 (12.9%)	64 (12.5%)	85 (16.6%)	512
ETnERV	119 (6%)	159 (8.1%)	94 (4.8%)	178 (9.0%)	1,969
MERVL	118 (3.4%)	110 (3.2%)	56 (1.6%)	159 (4.6%)	3,444
VL30	4 (1.2%)	12 (3.5%)	3 (0.9%)	13 (3.8%)	343
MMERGLN	7 (2.3%)	68 (22,6%)	53 (17,6%)	71 (23,6%)	300
malR-ORR1A	210 (3.9%)	349 (6.6%)	265 (5.0%)	447 (8.5%)	5,256

Table S1.2 - Number and percentage of active transposable elements at D0, D6 and D13 during conversion from serum to 2i+vitC. Elements were considered as “active” when they were covered by at least 10 uniquely mapped reads at one of the time point. Percentages represent the proportion of active copies relative to the total number of elements in a given family, as estimated from the reconstructed version of RepeatMasker.

Biological Process	P-value
multicellular organismal development (GO:0007275)	2.890e-05
single-multicellular organism process (GO:0044707)	1.004e-04
multicellular organismal process (GO:0032501)	1.875e-04
anatomical structure development (GO:0048856)	2.711e-04
single-organism developmental process (GO:0044767)	2.929e-04
developmental process (GO:0032502)	3.279e-04
anatomical structure morphogenesis (GO:0009653)	3.383e-04
system development (GO:0048731)	5.315e-04
organ development (GO:0048513)	1.275e-03
single-organism process (GO:0044699)	3.169e-03

Molecular Function	P-value
nucleic acid binding transcription factor activity (GO:0001071)	9.731e-03
sequence-specific DNA binding transcription factor activity (GO:0003700)	9.731e-03
regulatory region nucleic acid binding (GO:0001067)	5.069e-02
regulatory region DNA binding (GO:0000975)	5.069e-02
transcription regulatory region DNA binding (GO:0044212)	5.069e-02
sequence-specific DNA binding RNA polymerase II transcription factor activity (GO:0000981)	8.025e-02
molecular_function (GO:0003674)	1.229e-01
RNA polymerase II core promoter proximal region sequence-specific DNA binding transcription factor activity (GO:0000982)	2.048e-01
RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription (GO:0001227)	2.774e-01
DNA binding (GO:0003677)	6.186e-01

List of 156 Refseq genes significantly upregulated between D0 and D6 and without significant differences between D0 and D13 :

Cttnbp2	Cep57l1	Fut10	Mex3b	Ppargc1b	Trp53cor1
Gpr160	Clca4	Gbgt1	Mgat5	Ptgfrn	Tspan1
Lrrn2	Clcnka	Gbp9	Mpeg1	Rab39	Ttll6
Mep1b	Cntn1	Gm3604	Mpp1	Rcsd1	Ttn
Prss35	Cobl	Grsf1	Mppe1	Relb	Tubb2b
Rnf135	Coro6	Grtp1	Mpv17l	Rnf125	Ube2dnl1
Shf	Crisp1	H2-M5	Msc	Rnf144b	Ugdh
Smyd1	Cyp2c55	Hap1	Myo1f	Rpgrip1	Ulk1
Tmem181b-ps	Cyp4a30b	Herc3	Myof	Rtn4rl1	Xirp2
Trim43a	Cyp7b1	Hes1	Nanog	Sema5b	Zeb1
Zfhx2	Cysltr2	Ifi35	Nanos3	Sema6c	Zfp78
Zfp951	D130040H23Rik	Ifitm7	Nfix	Sesn1	Zp3

Ablim1	D1Pas1	Igsf23	Nim1k	Sh3tc1	1700030C10Rik
Agpat2	D230025D16Rik	Irak4	Nkx6-2	Siah2	1700084C01Rik
AI427809	Dact2	Jam2	Nlrp4g	Slc15a1	4930452B06Rik
Ak7	Dnajc6	Kank2	Nod1	Slc16a14	4930519F16Rik
Anapc10	Dsg2	Kcnh3	Noxo1	Slfn9	
Ankrd45	Duox1	Kcns3	Npr3	Smad3	
Axin2	Ehhadh	Klf4	Omp	Sobp	
Baz2b	Eif2s2	Klf8	Pabpc6	Spint1	
Bcl2l14	Elf3	Klhl22	Pcsk1	Src	
Bcl3	Eng	Klhl42	Pex3	Sync	
C030039L03Rik	Epha4	Lap3	Phf11a	Tbx15	
Capn6	Fam102a	Lax1	Pip5k1b	Tfeb	
Capsl	Fdft1	Ldlr	Pla2g4e	Tmem181c-ps	
Casp8	Fgf10	Lmo7	Plscr1	Tnrc18	
Catip	Flrt2	Mblac2	Podn	Tob1	
Cbr3	Fry	Mcf2	Popdc3	Triml1	

Table S1.4 – Gene Ontology enrichment analysis for genes specifically up-regulated between D0 and D6. Gene ontology analysis for biological processes and molecular functions was performed for the 156 genes that were significantly up-regulated between D0 and D6 but with no significant differences between D0 and D13. The ten most significant terms are shown.

		number of reads at the junction transposon-2nd exon			RPKM		
Gene	**Transposon**	**D0**	**D6**	**D13**	**D0**	**D6**	**D13**
Prrc1	RLTR16B_MM	542	4202	4583	5.1	49.7	70.7
Bglap3	IAP-d-int	29	2420	3061	0.1	5.7	4.1
Gab1	RLTR15	2823	1714	1261	142.5	113.1	100.0
Mylpf	ID_B1	537	1691	1093	28.1	143.3	116.3
Pecam1	RLTR11B	869	1493	1530	20.2	36.5	33.8
Usp7	ORR1A4	994	1360	1077	40.7	55.2	48.0
Plcb4	RLTR17	209	1093	716	5.7	4.2	2.8
Cyp2b23	ERVB7_2B-LTR_MM	16	860	356	0.2	4.2	3.3
Nfu1	Lx7	785	637	468	5.8	4.6	3.7
Mep1b	RLTR9E	66	625	159	0.3	3.1	0.6
Dopey2	Lx7	74	591	258	2.2	16.1	11.0
1500012F01Rik	B2_Mm2	571	534	198	20.7	20.5	8.1
Cdyl2	RLTR15	230	496	482	5.1	13.6	12.9
Serpina3m	MMERVK9C_I-int	2	488	507	0.0	6.9	6.6
Wdr95	L1MB2	24	443	433	0.2	5.1	5.2
4930548H24Rik	BGLII	88	419	385	0.6	4.1	4.2
Kdm2b	B4A	14	401	158	1.1	26.5	16.6
Ppm1a	MT2B	87	393	265	0.8	2.6	1.9
Atg4b	RLTR12D	60	358	311	2.1	9.6	8.0
Pla2g1b	RLTR11A	874	357	155	21.2	6.5	3.9
Fbrsl1	RLTR11B	324	325	168	12.7	13.7	10.8
Hcrtr2	ORR1B2	47	307	220	0.3	1.2	0.4
Xpot	RMER6B	250	297	132	5.9	7.0	4.0
Sec24d	ORR1D1	424	276	127	2.0	2.9	0.5
Oas1f	MTD-int	4	257	544	0.1	9.7	21.7
Atg13	L1MB3	155	243	57	3.8	7.2	2.1
Cul5	RLTR20C1	98	236	125	4.5	7.3	8.0
Mybl2	B1_Mur3	224	223	121	9.2	9.4	6.3
Aoah	BLII_Mus	101	212	89	5.1	10.6	2.8
Ccbl2	RLTR44-int	2	201	148	0.0	1.7	1.6

Table S1.5 – List of chimeric transcripts identified during conversion from serum to 2i+vitC. The 30 genes with the highest number of chimeric reads at D6 are ranked here. Numbers represent the absolute read count at the junction between the transposon (first exon) and the second exon of the gene, and the normalized read count of the whole transcript in RPKM.

RT-qPCR and ChIP-qPCR	
RplP0 F	TCCAGAGGCACCATTGAAATT
RplP0 R	TCGCTGGCTCCCACCTT
Gapdh F	TCCATGACAACTTTGGCATTG
Gapdh R	CAGTCTTCTGGGTGGCAGTGA
LINE1 F (ORF2)	GGAGGGACATTTCATTCTCATCA
LINE1 R (ORF2)	GCTGCTCTTGTATTTGGAGCATAA
IAPEz F (IAPdelta1 subfamily)	AACGCTGCTGCTTTAACTCC
IAPEz R (IAPdelta1 subfamily)	TGCACATAAAGCTGGCACA
MERVL F (LTR)	CAATGGGAAGGTCCAGAAGA
MERVL R (LTR)	CCTTGTTACCTCGGAATCCA
L1-T F (T monomer)	CAGCGGTCGCCATCTTG
L1-T R (T monomer)	CACCCTCTCACCTGTTCAGACTAA
L1-A F (A monomer)	GGATTCCACACGTGATCCTAA
L1-A R (A monomer)	TCCTCTATGAGCAGACCTGGA
Major Satellite F	GACGACTTGAAAAATGACGAAATC
Major Satellite R	CATATTCCAGGTCCTTCAGTGTGC

CRISPR guide sequence (without PAM)	
Kap1 exon 3 5'	GCAAGTAAATACAGGTCTGC
Kap1 exon 3 3'	GCAGACTTTGGAGGTTTAGG
Suvar39h1 exon 4	GGCCAGATCTACGACCGCCA
Suvar39h2 exon 4 5'	TCTTCACTTGTGATCACCTA
Suvar39h2 exon 4 3'	ACAGTGGATGCAGCTCGATA

Bisulfite pyrosequencing	
Dazl F	TTTAGGATTTATTTTATAGGGGT
Dazl R [Btn]	CAAAAAAAACCAAAAAACCCA
Dazl Seq	GGGGGGTTAGGGAGTG
H19 F	GGGTTTTTTTGGTTATTGAATTTTAA
H19 R [Btn]	AATACACACATCTTACCACCCCTATA
H19 Seq	TGTTATGTGTAATAAGGGAA
Oct4 F	AGGGGTGAGAGGATTTTGAA
Oct4 R [Btn]	ACCTCTCCCTCCCCAATC
Oct4 Seq	GGTTGAAAATGAAGGTTT
IAP F	GAGGGTGGTTTTTTATTTTATGTGT
IAP R [Btn]	ATCACTCCCTAATTAACTACAACC
IAP Seq	TTTTTATTTTATGTGTTTTGTTTTT
L1-T F	GGTTGGGGAGGAGGTTTAAGTTATA
L1-T R [Btn]	CTACCTATTCCAAAAACTATCAAATTCTT
L1-T Seq	GGGAGGAGGTTTAAGTTATAGTA
L1-A F	AGATTGAGGTATATAGGGAAGTAGGTT
L1-A R [Btn]	ATCCACTCACCAAAAATCTTAAAAT
L1-A Seq	GGTATATAGGGAAGTAGGTTA

Table S1.6 – Primer list

Antigen target	Supplier/Reference	Usage (dilution)
LINE1-ORF1	gift from A. Bortvin	Western (1/1000) - IF (1/1000)
IAP-GAG	gift from B. Cullen	IF (1/200)
KAP1	Abcam ab10483	Western (1/1000)
PARP	Cell Signaling	IF (1/200)
H3S10phospho	Cell signaling 9706	IF (1/500)
H3K9me2	Abcam ab1220	Western (1/1000) - IF (1/200) - ChIP (5µg)
H3K9me3	Abcam ab8898	Western (1/1000) - IF (1/200) - ChIP (3µg)
H3K27me3	Cell Signaling C36B11	Western (1/1000) - IF (1/200) - ChIP (5µL)
OCT4	Abcam ab19857	Western (1/1000)
NANOG	Abcam ab70482	Western (1/1000)
KLF4	Santa Cruz sc20691	Western (1/1000)
SOX2	Millipore AB5603	Western (1/1000)
H3	Abcam ab1791	Western (1/5000)
TUBULIN	Millipore CP06	Western (1/5000)
SUV39H1	Cell signaling 8729S	Western (1/1000)

Table S1.7 – Antibody list

2 Various requirements for DNMT3L during events of genome-wide *de novo* methylation

2.1 Authors and affiliations

Marius Walter[1,2], Aurélie Teissandier[1,2,3], Julian Iranzo[1], Sebastien A Smallwood[4], William A Pastor[5], Steven E Jacobsen[5], Gavin Kelsey[4], **Déborah Bourc'his**[1]*

[1] Institut Curie, Dpt of Genetics and Developmental Biology, CNRS UMR3215, INSERM U934, Paris, France

[2] UPMC, University Paris 06, Paris, France

[3] Institut Curie, Inserm U900, Mines Paris Tech, Paris, France

[4] Babraham Institute, Cambridge, UK

[5] Department of Molecular, Cell and Developmental Biology, University of California, Los Angeles, Los Angeles, United States

2.2 Introduction[19]

In eukaryotes, chromatin plays a fundamental role in the establishment and maintenance of cell fate, allowing tight control of transcriptional programs in a tissue-specific and developmentally regulated manner. Nucleosome organization, histone post-translational modifications, histone variants and many other players contribute to the specification of multiple functional chromatin states. Methylation of cytosines at the fifth position of the six-atoms aromatic ring, or DNA methylation, is a widespread DNA modification found in the majority of eukaryote kingdoms (Zemach et al., 2010). In mammals, DNA methylation is essentially found in CpG dinucleotides, and the symmetry of this sequence context allows for the faithful transmission of DNA methylation patterns during cell divisions. In mammals, DNA methylation is critical for many important biological processes such as genomic imprinting, X-chromosome inactivation in females, transposon repression and regulation of gene expression (Smith and Meissner, 2013). In particular, aberrant DNA methylation patterns is a hallmark of cancer and many human diseases are linked to imprinting defects (Peters, 2014; You and Jones, 2012).

[19] Some paragraphs of this introduction are almost directly taken from the main introduction of this thesis manuscript.

In mammalian somatic cells, 70-80% of CpG dinucleotides are methylated, which corresponds to 5-6% of total cytosines (Lister et al., 2009; Stadler et al., 2011). The only exception for this global hypermethylation resides in a specific category of GC- and CpG-rich sequences termed CpG islands (CGI). Aside from repeated sequences, there are around 23,000 CGIs in the mouse genome (Waterston et al., 2002). They are major regulatory units and around 50% of CGIs are located in gene promoter regions, while another 25% lie in gene bodies, sometimes serving as alternative promoters. Reciprocally, around 60-70% of genes have a CGI in their promoter region (Illingworth et al., 2010; Saxonov et al., 2006). The majority of CGIs are constitutively unmethylated and enriched for permissive chromatin modifications such as H3K4 methylation. In somatic tissues, only 10% of CGIs are methylated, the majority of them being located in intergenic and intragenic regions.

Twice during embryonic development, DNA methylation patterns are reset on a global scale. The first DNA demethylation event occurs in the zygote just after fertilization and culminates at the blastocyst stage; the second one occurs in primordial germ cells (PGCs), the early progenitors of spermatozoa and oocytes (Seisenberger et al., 2013). Whereas maintenance of DNA methylation during cell divisions relies on the DNA methyltransferase DNMT1, *de novo* methylation is catalyzed by DNMT3a and DNMT3a; a third member of the de novo DNA machinery, DNMT3l is catalytically inactive but stimulates DNMT3a/b (Chédin, 2011). DNMT3 expression peaks during periods of *de novo* methylation in the peri-implantation embryo and in germ cells. They are also highly expressed in ES cells, but downregulated after differentiation (Okano et al., 1998, 1999). *Dnmt3a*-knockout mice survive until birth but die in the next four weeks, while *Dnmt3b*-knockout is lethal around mid-gestation. By contrast, mice depleted for *Dnmt3a* in germ cells specifically are sterile, whereas *Dnmt3b* knockout has not effect (Kaneda et al., 2004). *Dnmt3l* knockout mice develop to term and have no obvious phenotype, but males and females are infertile due to major germline methylation defect (Bourc'his et al., 2001).

DNMT3a and DNMT3b have non-redundant functions but share the same protein structure. They contain a PWWP motif, a PHD-related domain termed ADD (ATRX-DNMT3-DNMT3l) and a catalytic domain that catalyzes the transfer of a methyl group using S-Adenosyl methionine as a methyl donor. By contrast, DNMT3l lacks the PWWP domain and its catalytic sites are mutated, but it possesses a functional ADD domain (Chédin, 2011). DNMT3l does not bind DNA but its pseudo-catalytic site physically interacts with the catalytic domain of DNMT3a and DNMT3b to stimulate their activity (Chen et al., 2005; Suetake et al., 2004). Crystallographic studies showed that DNMT3l and DNMT3a form

tetramers composed of DNMT3a and DNMT3l dimers, whereas in absence of DNMT3l, DNMTa forms long oligomers with reduced activity and progressivity (Holz-Schietinger and Reich, 2010; Jia et al., 2007). PWWP and ADD domains regulate the interaction of DNMT3 proteins with chromatin. The PWWP domain binds to H3K36me3, a mark associated with transcription elongation, and DNMT3b was shown to be indeed recruited to actively transcribed regions in ES cells (Baubec et al., 2015; Dhayalan et al., 2010; Morselli et al., 2015). The ADD domain binds to histone H3 tails, and this interaction is blocked by H3K4 methylation, a mark linked to transcriptional activity (Ooi et al., 2007; Otani et al., 2009). As a consequence, the *de novo* machinery is excluded from H3K4me3-enriched regions, such as promoters and enhancers.

There has been extensive characterization of the DNMT3 proteins in *in vitro* biochemical assays, but a precise analysis of how DNA methylation patterns are shaped *in vivo* is still incomplete, in particular regarding the contribution of DNMT3l. While DNMT3l is absolutely required for germline methylation, it is mainly dispensable in the embryo. *Dnmt3l*-KO animals develop normally and do not have methylation defects. Nonetheless, we previously showed that some genomic sequences, unique or repeated, acquired DNA methylation at a slower rate in absence of DNMT3l in early post-implantation embryo (Guenatri et al., 2013). Moreover, DNMT3l was shown to be involved in methylating the genome of ES cells (Arand et al., 2012; Ooi et al., 2010): however, these studies were performed on serum-grown ES cells that have a fully methylated genome (Leitch et al., 2013) and could therefore only infer DNMT3l role in the maintenance of DNA methylation. In general, little is known about the mechanisms by which DNMT3l contributes to establish correct DNA methylation patterns, its sequence specificity and interplay with underlying chromatin contexts. To address this question, we generated *Dnmt3l*-knockout ES cells (*Dnmt3l*-KO). By differentiating ES cells originally grown in cultured-induced hypomethylated conditions, we could recapitulate genome-wide events of *de novo* methylation, in presence and in absence of DNMT3L. To get the most comprehensive view of the biological contribution of DNMT3L in different *de novo* DNA methylation events, we further compared this cellular system with *in vivo* post-implantation embryos and male germ cells.

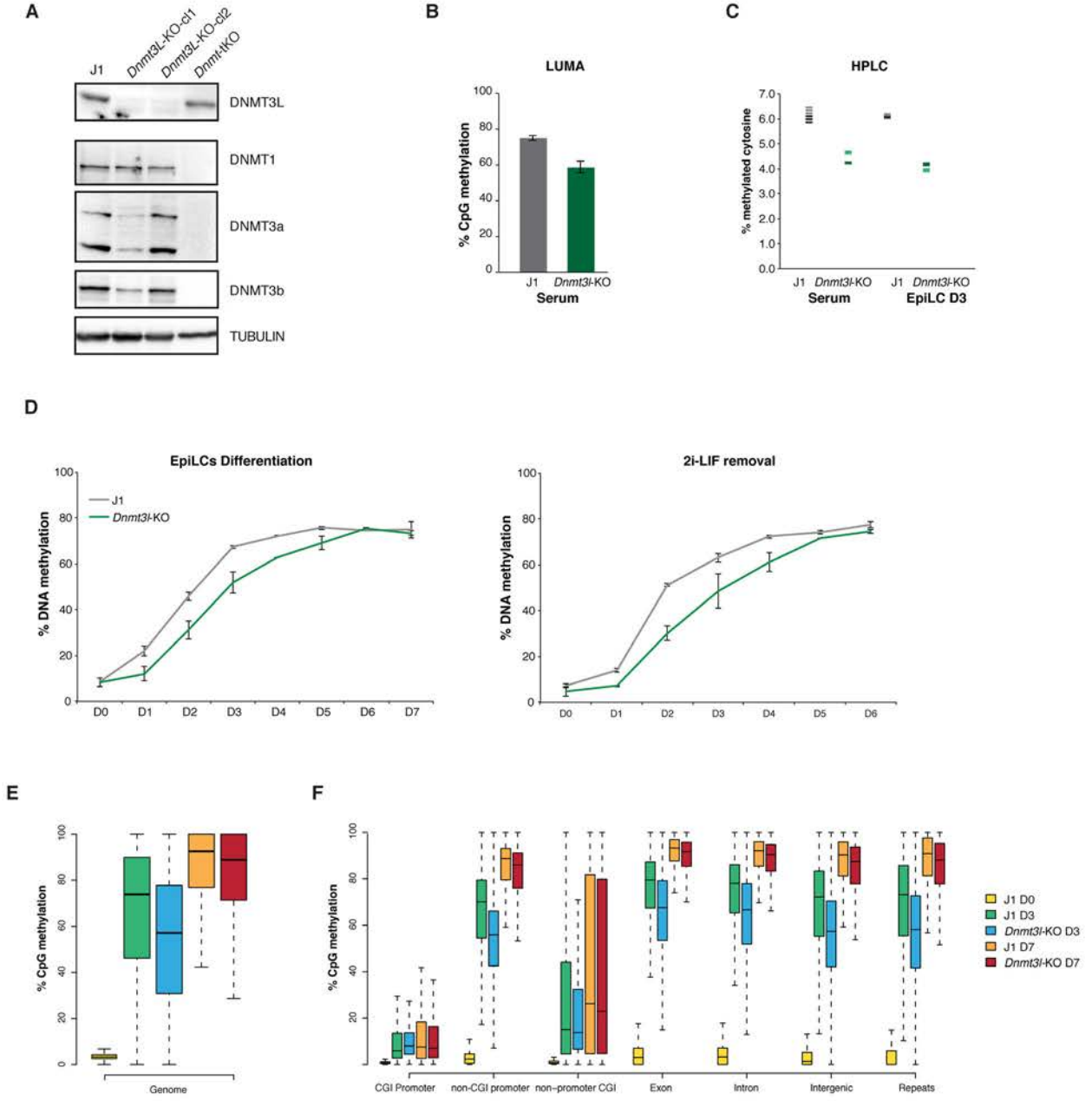


Figure 2.1. DNA methylation is globally delayed upon differentiation of *Dnmt3l*-KO ES cells

A. Dnmt protein levels in *Dnmt3l*-KO and *Dnmt*-tKO ES cells grown in serum

B. Global CpG methylation measured by LUMA in *Dnmt3l*-KO ES cells grown in serum. Data represent mean and Standard Error of the Mean (SEM) between two biological replicates.

C. Global cytosine methylation levels measured by liquid chromatography followed by mass spectrometry in serum-grown WT and *Dnmt3l*-KO ES cells and at D3 of EpiLC differentiation. Data represent two biological replicates (light and dark colors), with three technical replicates for each measurement.

D. Time course of global CpG methylation loss measured by LUMA over 6 or 7 days (D0 to D7) of ES cell differentiation, either in EpiLC medium or by 2i-LIF removal. Data represent mean ± SEM between two biological replicates.

E. Tukey boxplot representation of genome-wide CpG methylation content as measured by WGBS during differentiation of WT and *Dnmt3l*-KO ES cells.

F. CpG methylation distribution over different genomic compartments by WGBS.

2.3 Results

Acquisition of DNA methylation is globally delayed in absence of DNMT3l during ES cell differentiation

In order to generate *Dnmt3l*-KO ES cells, we used CRISPR/Cas9 editing to delete *Dnmt3l* second exon in J1 WT ES cells, introducing a stop codon that efficiently knocked out DNMT3l protein (**Figure 2.1A**). CRISPR/Cas9-mediated deletion produced two independent knockout cell lines that we used as biological replicates during this study. Importantly, protein levels of DNMT1, DNMT3a and DNMT3b were globally unaltered by DNMT3l absence, even though DNMT3a and DNMT3b appeared slightly reduced in one of the replicate (**Figure 2.1A**). We measured global CpG methylation in serum-grown ES cells using LUminometric Methylation Assay (LUMA) and observed a reduction from ~75% to ~60% of CpG methylation in absence of DNMT3l (**Figure 2.1B**). Similarly, measurement of the total proportion of methylated cytosines by liquid chromatography followed by mass spectrometry showed a reduction from ~6% to ~4% in *Dnmt3l*-KO cells (**Figure 2.1C**). Importantly, hydroxymethylation levels remained identical in absence of DNMT3l, suggesting that active demethylation catalyzed the TET enzymes is unlikely to explain the hypomethylation of *Dnmt3l*-KO ES cells (data not shown). The decreased level of DNA methylation in *Dnmt3l*-KO cells is in line with the observation that serum-grown ES cells lacking DNMT3a or DNMT3b have also reduced levels of DNA methylation (Arand et al., 2012; Okano et al., 1999). This indicates that maintenance of hypermethylated state in serum-grown ES cell necessitates a continuous involvement of the *de novo* DNA methylation machinery.

In order to study DNMT3l contribution to *de novo* DNA methylation, we first reduced global DNA methylation levels in *Dnmt3L*-KO and J1 WT ES cells by culturing them for three weeks in presence of two small kinase inhibitors (2i) and vitamin C (Blaschke et al., 2013; Ying et al., 2008), two conditions that promote almost complete erasure of DNA methylation in ES cells, as we showed previously (Walter et al., submitted). Cells were then differentiated, either into epiblast-like cells (EpiLC) by switching to EpiLC-specific medium (Guo et al., 2009), or by 2i-LIF removal. In both differentiating conditions, global DNA methylation increased rapidly in WT J1 ES cells and flattened out at roughly 75% after four days (**Figure 2.1D**). In *Dnmt3l*-KO ES cells, DNA methylation acquisition was globally delayed, with 10-20% of decreased methylation observed from day 1 (D1) to D5 of differentiation. However, *Dnmt3l*-

A

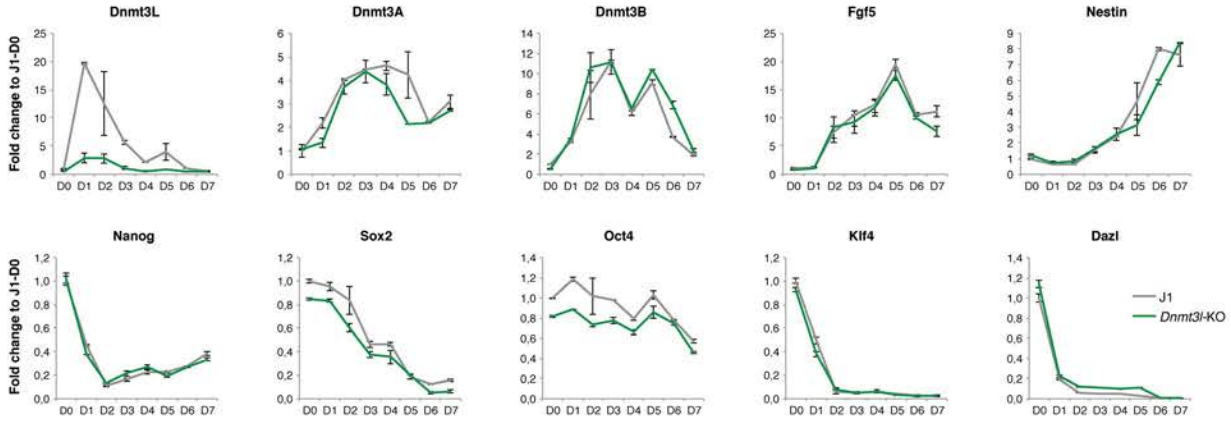


Figure 2.2. The dynamic of differentiation is not affected by Dnmt3l absence

A. Relative expression levels of *Dnmt* enzymes, pluripotent and differentiation markers during EpiLC differentiation. Expression levels were measured by Nanostring nCounter. Data are expressed as fold changes to WT D0 and represent mean ± SEM between two biological replicates

KO ES cells eventually reached a level of DNA methylation similar to their WT counterparts around D6, in both differentiating conditions.

To determine the extent of DNA methylation defects between WT and *Dnmt3l*-KO, we carried out whole-genome bisulfite sequencing (WGBS) at D3 and D7 of EpiLC differentiation for both WT and *Dnmt3l*-KO cells; the D0 WT time point was based on previously generated data (Walter et al., submitted). Quality control indicated high genomic coverage (around 30x), with approximately 50-60% of CpGs covered at least five times (**Table S2.1A**). CpG methylation reached an average of 65.2% at D3 in WT, but was reduced at 52.2% in *Dnmt3l*-KO cells. By contrast, DNA methylation was similar at D7 of differentiation, with average levels of 79.6% and 76.3% in WT and *Dnmt3l*-KO contexts, respectively (**Figure 2.1E**). Importantly, this 15% methylation delay at D3 affected every genomic compartments, including gene bodies, repeated elements and intergenic regions (**Figure 2.1F**). Overall, these observations indicate that DNMT3l acts essentially by accelerating the deposition of DNA methylation genome-wide, but becomes dispensable in the long term.

Transcriptional patterns are barely affected by the absence of DNMT3l during ES cell differentiation

In order to analyze if the DNA methylation delay caused by DNMT3l deficiency could impact proper cellular differentiation, we followed the expression of pluripotent and differentiation markers during ES to EpiLC transition using Nanostring nCounter quantification (**Figure 2.2A**). As expected, in WT cells, *Dnmt3a* and *Dnmt3b* expression peaked at the beginning of differentiation, while the ectoderm marker *Fgf5* and *Nestin* increased progressively. Conversely, expression of pluripotent transcription factors such as *Nanog, Sox2* and *Klf4* decreased rapidly, while *Oct4* levels remained high. Importantly, no significant difference could be noted between WT and *Dnmt3l*-KO cells, suggesting that DNMT3l deficiency does not affect the dynamics of ES cell differentiation and their cellular fate. Incidentally, this excludes that the delay in methylation acquisition reflects a delay in the kinetics of differentiation, and rather supports an intrinsic default of the *de novo* DNA methylation process in *Dnmt3l*-KO ES cells.

To determine if the DNA methylation delay impacted on transcriptional programs, we performed paired-end RNA-seq at D0, D3 and D7 of EpiLC differentiation, in two biological replicates of WT and *Dnmt3l*-KO ES cells. Quality controls indicated high genomic coverage (**Table S2.1**) and consistency between replicates (**Figure 2.3A**). A total of 4,513 genes were significantly up- or downregulated between the beginning and the end of the differentiation,

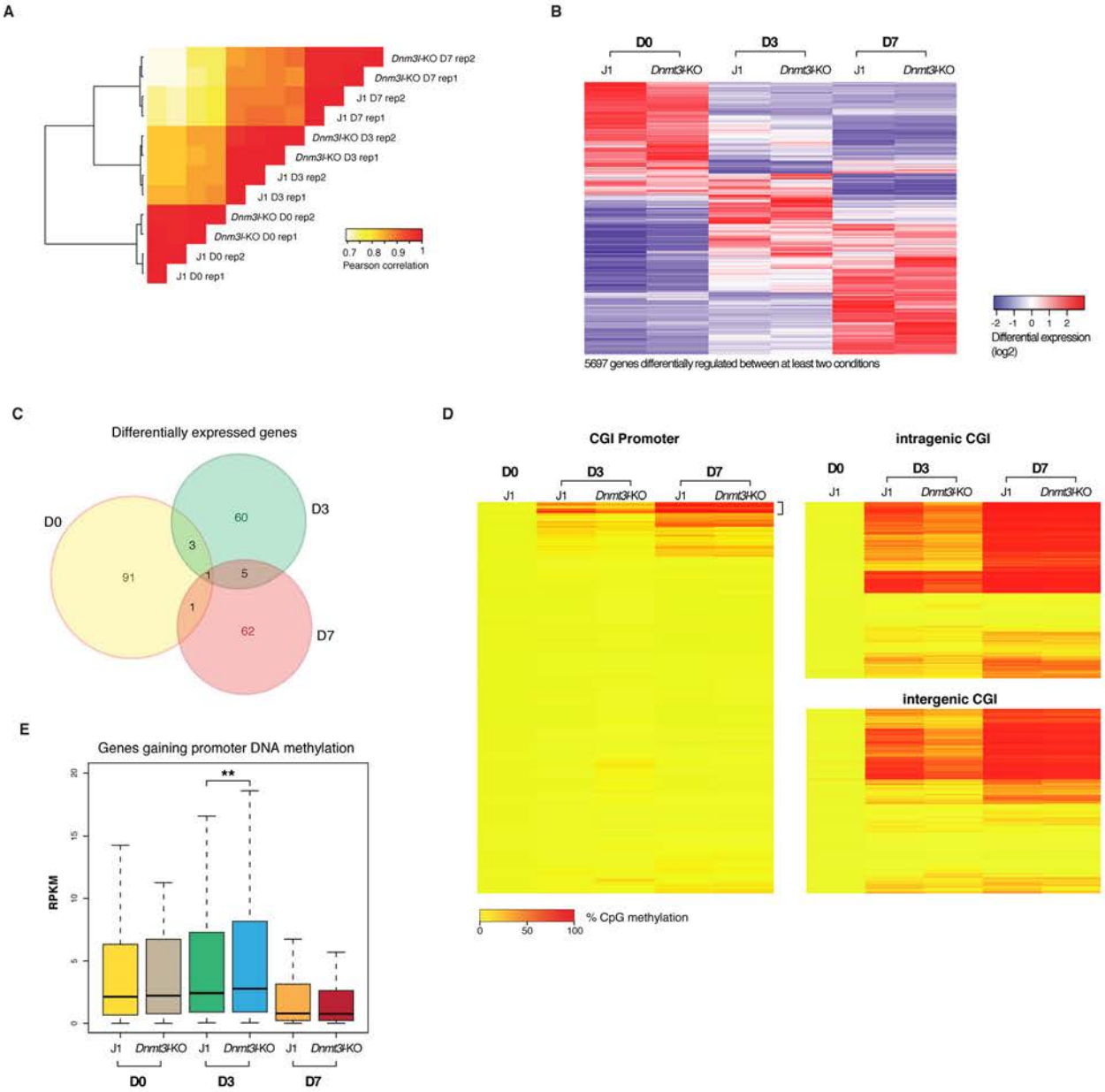


Figure 2.3. Mild transcriptional perturbation during Dnmt3l-KO ES cells differentiation

A. Pearson correlation and hierarchical clustering between RNA-seq experiments.

B. Heatmap representation of genes (n=3,301) that are significantly misregulated between at least two time points of D0, D3 and D7 of EpiLC differentiation. Colors represent on a log2-scale the differential expression between a given condition and the average of the six conditions.

C. Venn diagram of differentially expressed genes at D0, D3 and D7 between *Dnmt3l*-KO and WT genotypes. *Dnmt3l* is the only gene differentially expressed in every conditions.

D. Heatmap and hierarchical clustering of average DNA methylation at CpG islands as measured by WGBS

E. Box plot of average expression level of 420 genes that contained CGI promoters gaining DNA methylation during EpiLC differentiation. Data represent read per kb per millions (RPKM). **p<0.01 (Wilcoxon rank-sum test).

and WT and *Dnmt3l*-KO followed globally the same trend (**Figure 2.3C**). Closer analysis revealed that 96, 69 and 69 genes were differentially expressed between WT and *Dnmt3l*-KO at D0, D3 and D7 of differentiation, respectively (**Table S2.2** and **Figure 2.2D**).

Around 75% of genes misrelegulated at D3 were upregulated in *Dnmt3l*-KO cells, and this upregulation primarily affected genes that gained promoter DNA methylation during differentiation. For example, the germline genes *Dazl, Asz1* or *Tuba3a* were significantly upregulated at D3, but not at D7, which correlates with a delayed acquisition of DNA methylation in their CGI promoters (**Figure 2.3 – supplemental figure 1A**). Of note, many of these genes were strongly expressed at D0 and experienced dramatic repression upon differentiation, being almost completely silenced at D7. The delayed acquisition of DNA methylation at their promoter correlates therefore with a delayed transcriptional repression. By contrast, at D7, the majority (~80%) of the 69 differentially expressed genes were downregulated in *Dnmt3L*-KO cells, suggesting indirect mechanisms. However, a few genes, like for example the testis-specific *Ldhc* gene, were upregulated at D7 between *Dnmt3l*-KO WT and WT cells, which correlated with reduced methylation of their promoters (**Figure 2.3 – supplemental figure 1B**).

We reasoned that delayed acquisition of DNA methylation in the *Dnmt3l*-KO context would principally affect transcription of genes that gain promoter DNA methylation during differentiation. During WT EpiLC differentiation, around 10% of CGIs located in promoter regions and roughly 40% of non-promoter CGIs, acquire methylation (**Figure 2.3D**). More specifically, we distinguished 420 genes that gained promoter DNA methylation during differentiation, a number in good agreement with the 411 CpG-rich promoters that were found highly methylated in 9.5 days (E9.5) embryos (Borgel et al., 2010). As expected, gene ontology revealed that the majority of these genes were involved in germline development (**Table S2.3**). Among these 420 genes, 180 were significantly transcribed at some point during the differentiation. As anticipated, in average, their expression decreased importantly over the course of the differentiation. In *Dnmt3l*-KO EpiLC differentiation, as for the rest of the genome, CGI methylation was reduced after three days of differentiation, but attained WT levels after seven days. The 180 germline genes that are expressed showed a slight but significant upregulation could at D3 of *Dnmt3l*-KO differentiation (**Figure 2.3E**).

Another class of genomic elements that gained methylation upon differentiation was transposable elements. We followed 69 representative families of transposons over the course of the EpiLC differentiation and observed that virtually every families became heavily methylated, with a delay in *Dnmt3l*-KO cells similar to the rest of the genome (**Figure 2.3 –**

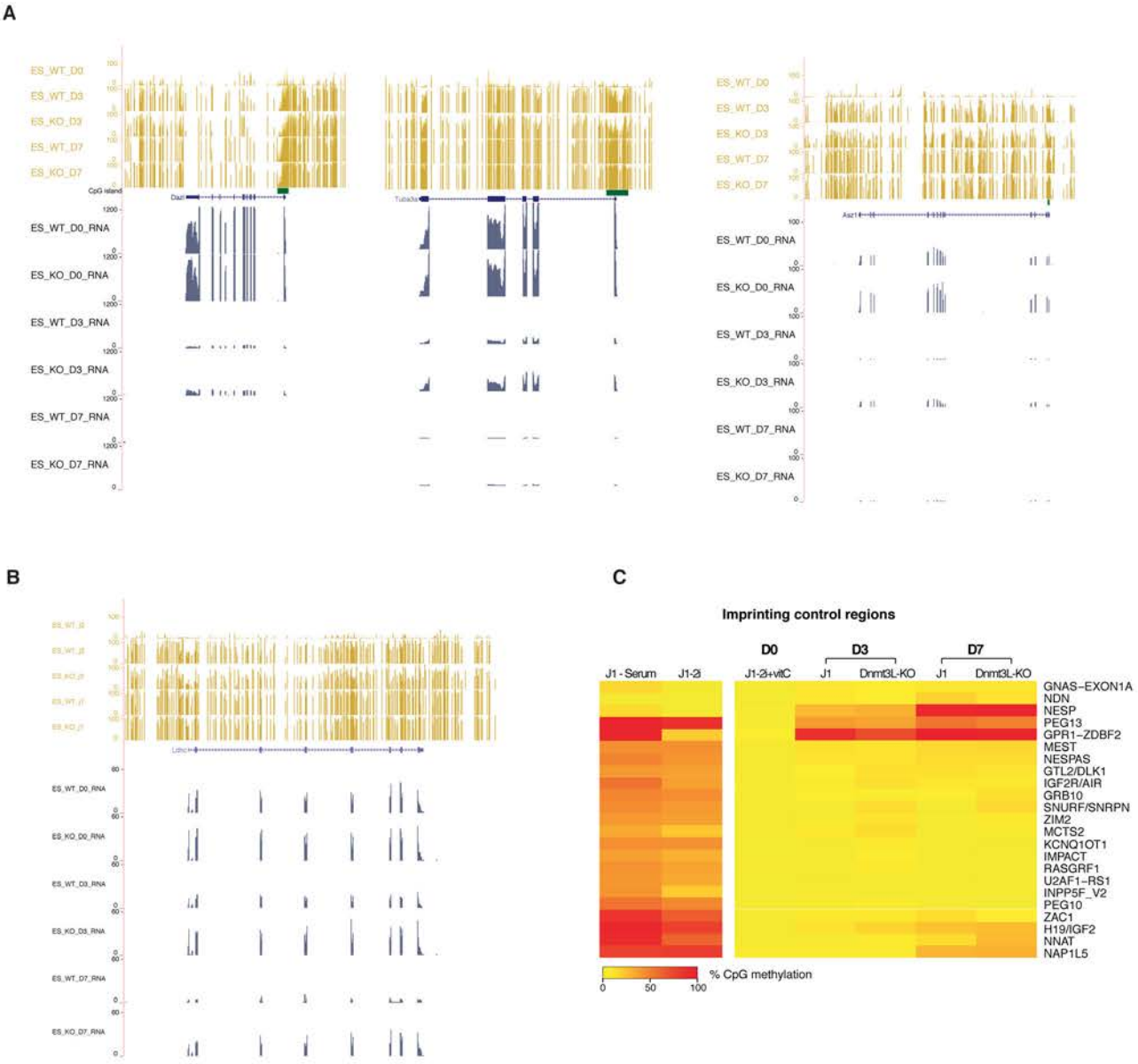


Figure 2.3 - supplement figure 1

A. Examples of WGBS profile of genes with delayed DNA methylation acquisition at CGI promoters and corresponding transcriptional profile. Bars represent the methylation percentage of individual CpG sites, between 0 (unmethylated) and 100% (fully methylated). RNA-seq data represent normalized read count.

B. Same as (A)

C. Average methylation level in Imprinted Control Regions (ICRs) as measured by WGBS.

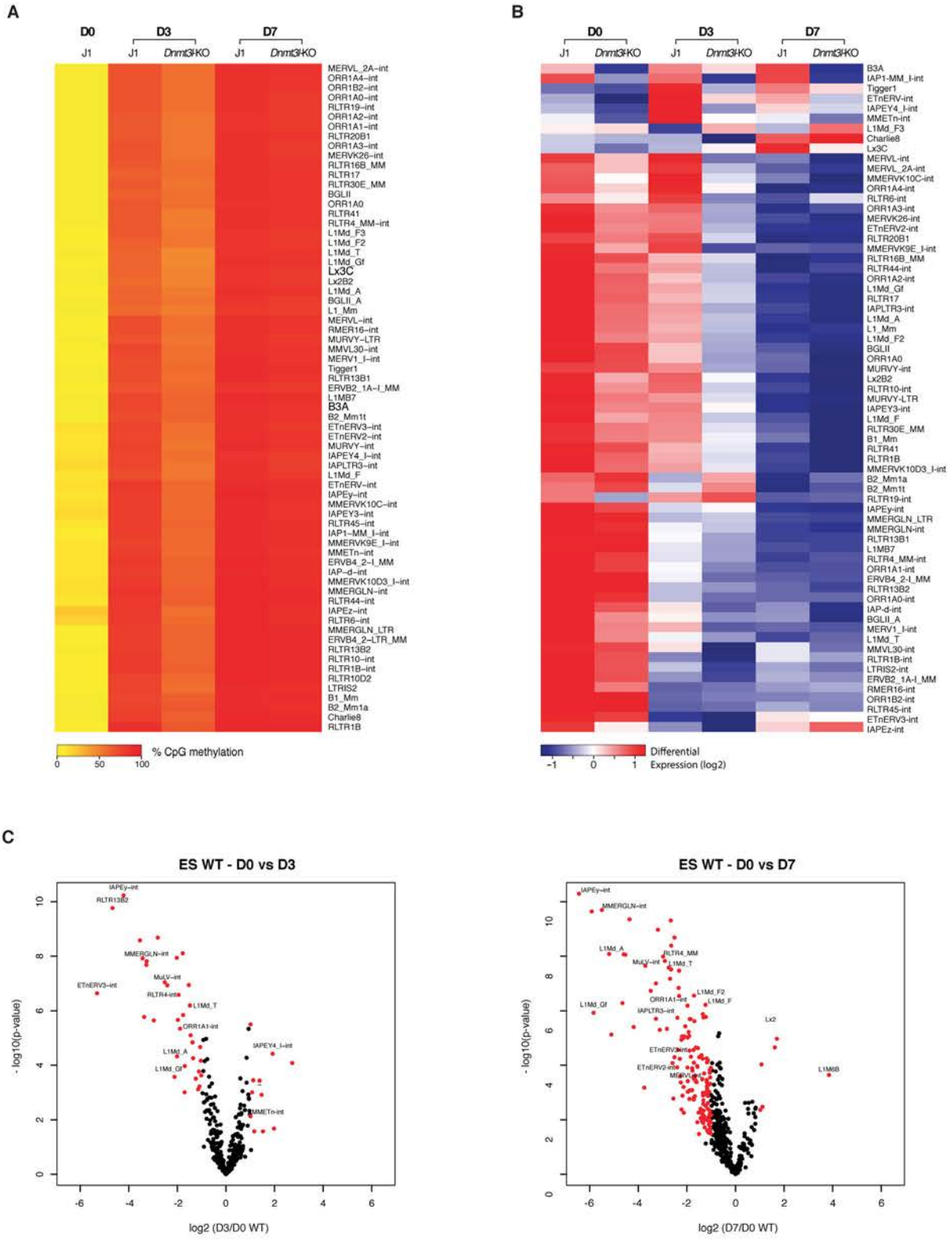


Figure 2.3 - supplement figure 2.

A. Heatmap and hierarchical clustering of average CpG methylation over 69 transposon families as measured by WGBS.
B. Heatmap representation and hierarchical clustering of expression changes for 69 transposon families. Colors represent on a log2-scale the differential expression between a given condition and the average of the six conditions.
C. Volcano plot representation of up- and down-regulated transposons as measured by RNA-seq between D0 and D3 (left), D0 and D7 (middle) of ES cell differentiation. Red dots indicate significantly misregulated repeats between two conditions (fold change > 2 and p-value <0.05). RNA-seq mapping allowed multiple hits onto the genome.

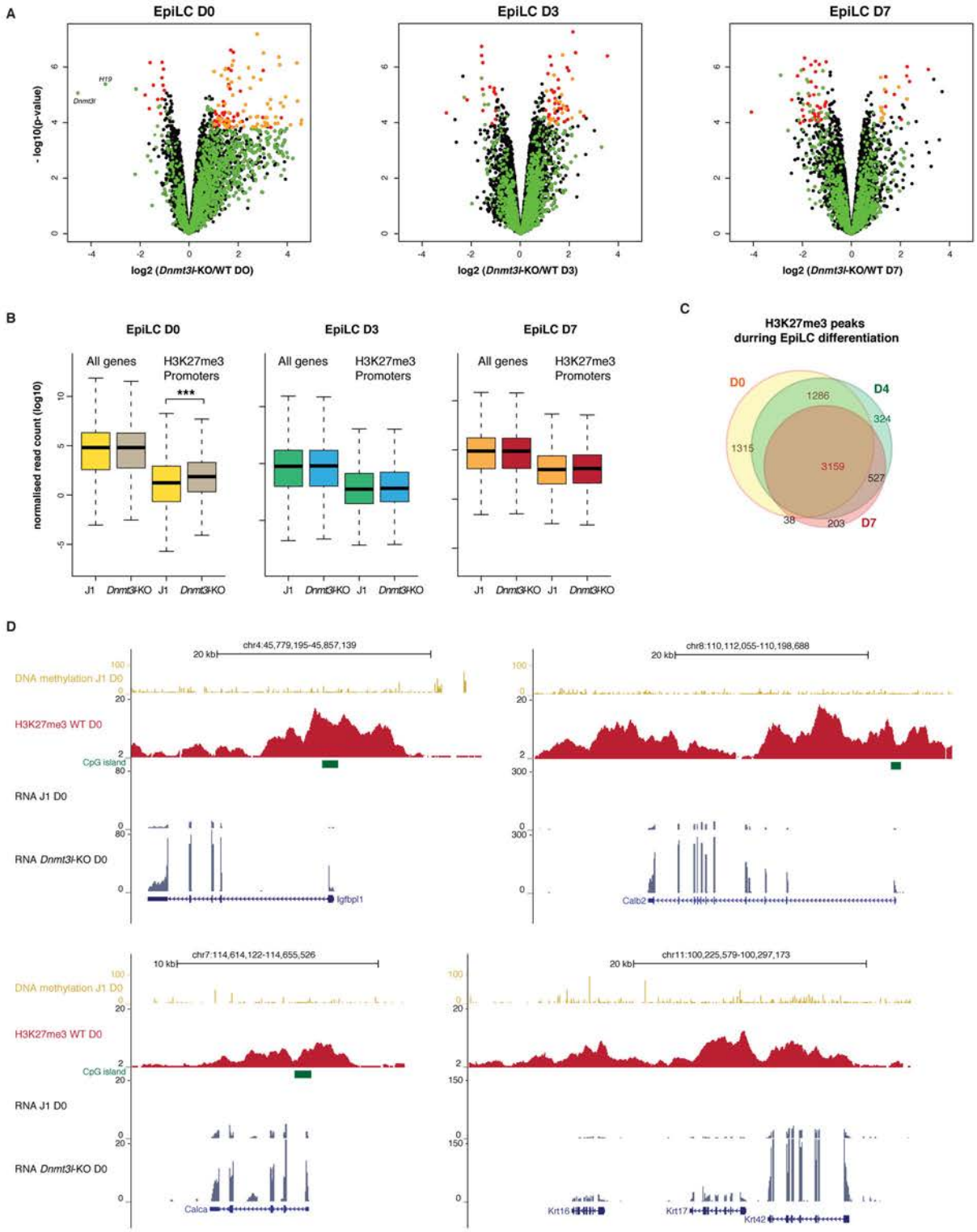


Figure 2.4. Polycomb target genes are upregulated in undifferentiated 2i+vitc-grown *Dnmt3l*-KO

A. Volcano plot representation of up- and down-regulated genes as measured by RNA-seq between WT and *Dnmt3l*-KO cells at D0, D3 and D7 of EpiLC differentiation. Red dots indicate significantly misregulated genes between two conditions (fold change > 2 and False discovery rate (FDR) <0.1). Green dots represent genes with H3K27me3 marking their promoter. Orange dots represent genes that are both significantly upregulated and with H3K27me3 marking their promoter.

B. Gene expression for all genes or only Polycomb target genes in WT and *Dnmt3l*-KO at D0, D3 and D7 of differentiation. Only genes with detectable expression in at least one condition are represented (n= 12236). For Polycomb target genes, only genes with detectable expression in at least one condition and with an H3K27me3 peak at the given day were considered, which represents 1730 genes at D0, 1881 at D3 and 1569 at D7.

C. Venn Diagram representing the overlap between all annotated H3K27me3 peaks during differentiation.

D. Example of genes marked by H3K27me3 and upregulated in *Dnmt3l*-KO ES cells RNA-seq and ChIP-seq data represent normalized read count.

supplemental figure 2A). DNA methylation is described to repress transposable elements in differentiated cells (Hutnick et al., 2010; Walsh et al., 1998), and indeed the gain of DNA methylation correlated with the strong repression of the majority of transposon family during WT differentiation (**Figure 2.3 – supplemental figure 2B,C**). Surprisingly, at both D0 and D3, transposons appeared more expressed in WT than in *Dnmt3l*-KO cells. Considering the naturally weak expression of *Dnmt3l* and the global absence of DNA methylation at D0, this result was puzzling and difficult to interpret.

Notably, we followed the dynamics of *de novo* methylation in imprinting control regions (**Figure 2.3 – supplemental figure 1C**). In globally hypermethylated ES cells grown in serum, the majority of these regions showed, as expected, a methylation level of around 50%, while some had lost their imprint, by gain or loss of methylation, according to the genetic drift undergone by these cells (Greenberg and Bourc'his, 2015). As shown previously, DNA methylation was completely lost in 2i+vitC (Walter et al., submitted). Interestingly, during differentiation, the majority of the imprinting control regions did not undergo *de novo* methylation, while a small subset became by contrast fully hypermethylated (*Nesp*, *Peg13* and *Gpr1-Zdbf2*). This showed that 2i+vitC treatment durably erases imprinting, and that the two alleles of each imprinted regions are epigenetically similar in this context.

Overall this analysis showed that transcriptional patterns are barely affected in general by the methylation delay in differentiating *Dnmt3l*-KO cells. Nonetheless, a subset of genes that usually gain promoter methylation showed delayed repression during EpiLC differentiation, with apparently no long-term consequences.

The majority of genes up-regulated in undifferentiated *Dnmt3l*-KO cells have H3K27me3-enriched promoters

The observation that many genes were differentially expressed at D0 in *Dnmt3l*-KO cells, the majority of them (~85%) being upregulated, is puzzling. Indeed, *Dnmt3l* is globally repressed in 2i+vitC-grown ES cells (Leitch et al., 2013), and the genome is also completely unmethylated in this context, making it difficult to propose an explanation for this upregulation. Closer analysis revealed in fact that 722 genes were upregulated more than two-fold between *Dnmt3l*-KO and WT cells. However, most of these genes were lowly expressed, preventing them to meet our significance threshold.

To investigate what could explain this unexpected upregulation, we used H3K27me3 ChIP-seq data generated in the lab during EpiLC differentiation of E14 ES cells, another male ES cells from a similar 129Sv background than J1 cells. ChIP-seq was carried at D0, D4 and D7 of differentiation. Peak calling and annotation indicated that 6,084 genes had a

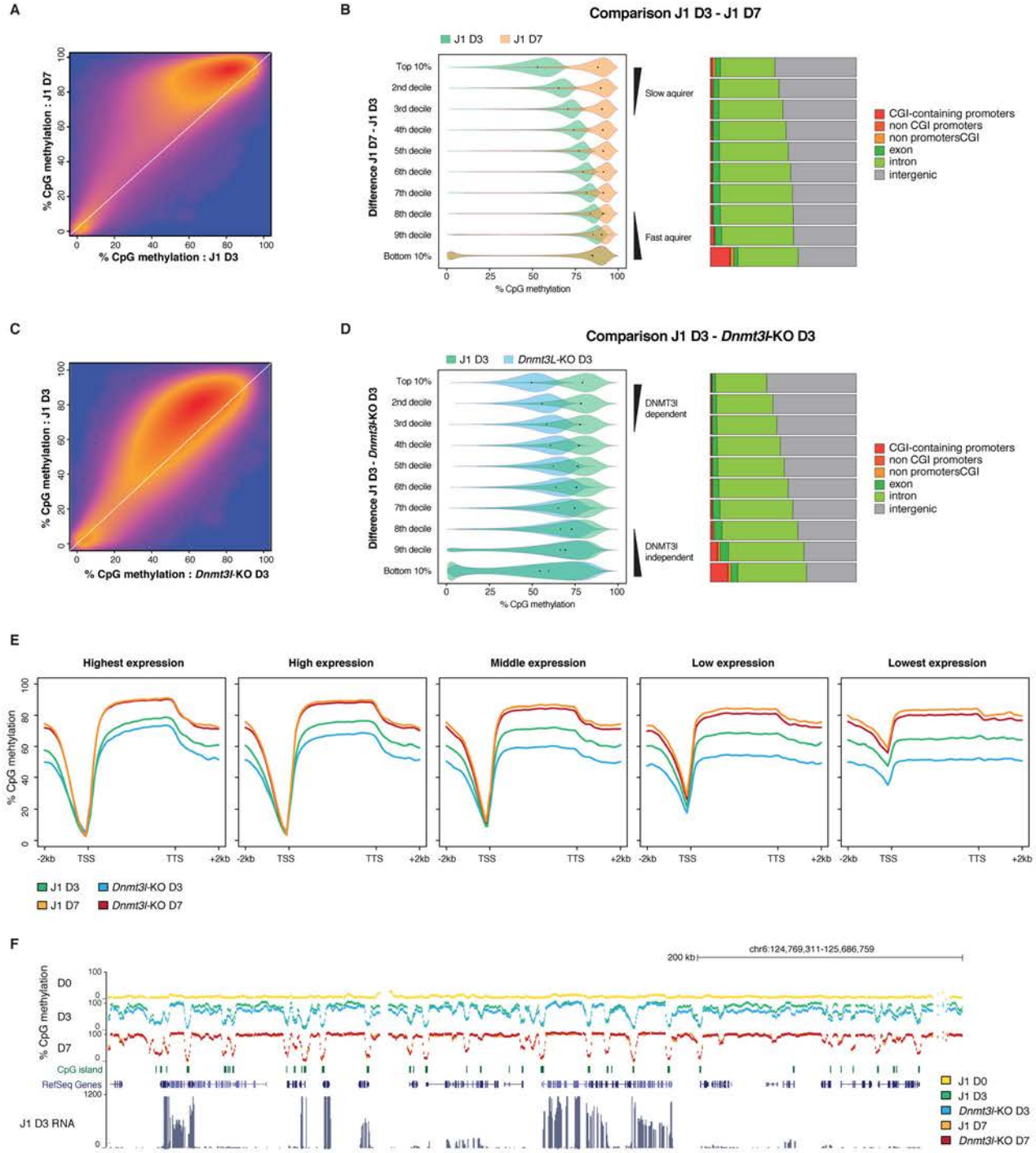


Figure 2.5. DNA methylation is deposited faster in highly transcribed regions and in a DNMT3l-independent manner

A. Density map comparing average CpG methylation at 1kb sliding windows between WT ES cells at D3 and D7 of ES cells differentiation.

B. Left panel: Violin plots representing average CpG methylation at 1kb sliding windows. The genome was separated into 10 groups depending on the DNA methylation differences between D3 and D7 of WT ES cells differentiation. Deciles separate regions that gain DNA methylation slowly or rapidly during differentiation. Right panel: Annotation of each 1kb window into different genomic compartments.

C. Density map comparing average CpG methylation at 1kb sliding windows between WT and *Dnmt3l*-KO ES cells at D3 of differentiation.

D. Left panel: Violin plots representing average CpG methylation at 1kb sliding windows. The genome was separated into 10 groups depending on the DNA methylation differences betwee *Dnmt3l*-KO and WT ES cells at D3 of differentiation. Deciles separate regions that gain DNA methylation in a DNMT3l-dependant or independent manner. Right panel: Annotation of each 1kb window into different genomic compartments.

E. Average CpG methylation over metagenes, separated in five classes depending on transcriptional level at D3 of differentiation (in WT).

F. Representative genomic region showing DNA methylation level in transcriptionally silent or active regions. RNA-seq data represent normalized read count

promoter marked by H3K27me3 in D0 undifferentiated 2i+vitC ES cells. Among them, 1,730 genes had detectable expression in at least one of the day of differentiation. Surprisingly, most of the genes upregulated at D0 in *Dnmt3l*-KO cells have H3K27me3-enriched promoters (**Figure 2.4A**, left panel). More specifically, out of the 720 genes upregulated in *Dnmt3l*-KO cells, 462 (65 %) are marked by H3K27me3: therefore, 25% of genes both marked by H3K27me3 and with detectable expression get upregulated in *Dnmt3l*-KO cells at D0 (hypergeometric test: p-value < 10^{-229})(**Figure 2.4B**, left panel).

These results indicate that H3K27me3 marks the majority of genes upregulated in *Dnmt3l*-KO cells at D0 of differentiation, and that reciprocally Polycomb-marked genes tended to be up-regulated in absence of DNMT3l (**Figure 2.4A, B**, left panel, **Figure 2.4D** for visual examples). Importantly the link between H3K27me3 and DNMT3l could only be observed at D0 of differentiation, and disappeared during differentiation (**Figure 2.4A** and **B**, mid and right panel, **Figure 2.4C**). This suggests that this putative link between DNMT3land the Polycomb machinery is specific to the hypomethylated context of 2i+vitC-grown cells and therefore independent of DNA methylation.

Overall, we show that genes upregulated in *Dnmt3l*-ko ES cells with an hypomethylated states had promoters marked by H3K27me3, suggesting that Polycomb-mediated repression could be alleviated in absence of DNMT3l. This result is vey preliminary but points towards an unsuspected mechanistic link between DNMT3l and the Polycomb machinery in absence of DNA methylation.

DNMT3l is dispensable for DNA methylation acquisition in highly transcribed regions in ES cells

To gain more insight into the dynamics of DNA methylation, we partitioned the genome into 1kb sliding windows, with a 500pb overlap between adjacent windows. We then calculated the average methylation of every windows in the different samples. Comparison between WT D3 and D7 showed, as expected, that most of the genome had low methylation levels at D3 (**Figure 2.5A**). However, differences between D3 and D7 were not homogeneous. Some regions had comparable methylation levels between the two time points, while others were markedly different, indicating that the rate of DNA methylation acquisition differ depending on the genomic region. To understand what distinguished fast and slow methylation acquirers, we then calculated for every 1kb window the difference of methylation between D7 and D3, and used the deciles of this difference to separate the genome into 10 groups (**Figure 2.5B**, left panel). The first decile contains regions with an important methylation difference between D3 and D7: they acquire DNA methylation at a slow rate. By

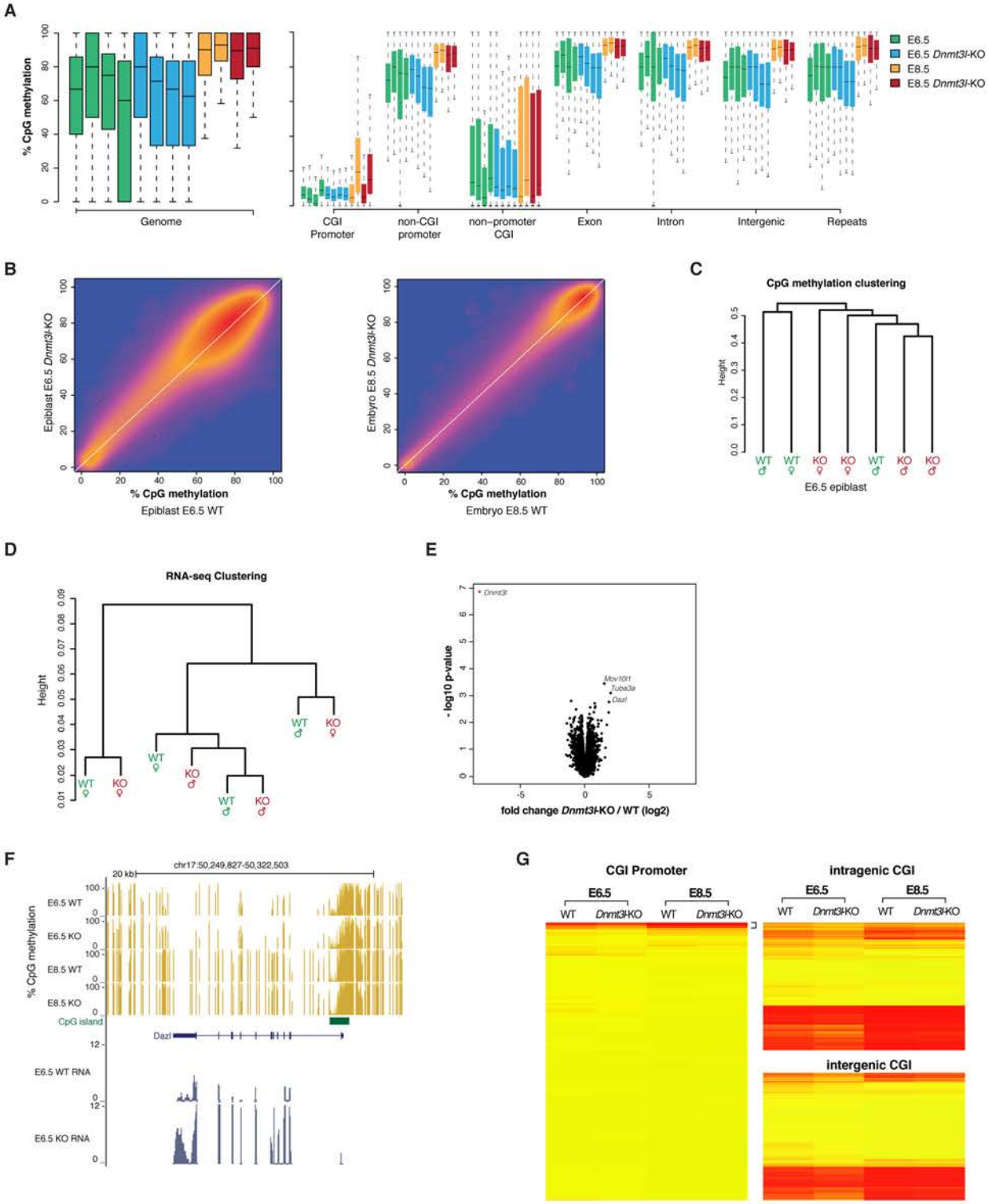


Figure 2.6. DNMT3l is dispensable for embryonic de novo methylation

A. Left panel: Genome-wide CpG methylation content as measured by WGBS in E6.5 epiblasts and E8.5 embryos, in WT and *Dnmt3l*-KO genotypes, in four biological replicates for E6.5 and two for E8.5. Right panel: CpG methylation distribution over different genomic compartments by WGBS.

B. Density map comparing average CpG methylation at 1kb sliding windows between E6.5 WT and *Dnmt3l*-KO epiblasts (left), or WT E6.5 epiblasts and E8.5 embryos (right).

C. Hierarchical clustering of E6.5 epiblasts based on WGBS data (ward's method)

D. Hierarchical clustering of E6.5 embryos based on RNA-seq data (ward's method)

E. Volcano plot representation of up- and downregulated genes as measured by RNA-seq between E6.5 *Dnmt3l*-KO and WT epiblasts. Red dots indicate significantly misregulated genes between two conditions (fold change > 2 and FDR <0.1).

F. Screenshot of *Dazl* gene showing slightly reduced promoter methylation and increased expression in E6.5 epiblast.

G. Heatmap and hierarchical clustering of average DNA methylation at CpG islands as measured by WGBS

contrast, the last decile represents fast acquiring regions that have already attained their final level at D3. Repartition of the windows of the 10 groups into different genomic compartments showed that fast acquiring regions were more often found into gene bodies that slow acquiring ones (**Figure 2.5B**, right panel). Of note, most CGIs cluster within the lower decile, which was to be expected since they remain hypomethylated. However, a noticeable proportion of CGIs appeared in the upper decile, and probably represent islands that gained DNA methylation during the differentiation, although at a slow rate.

We repeated a similar analysis at D3 of differentiation between *Dnmt3l*-KO and WT samples. Most of the genome had reduced methylation in *Dnmt3l*-KO cells (**Figure 2.5C**), allowing again partitioning the genome into 10 groups, based this time upon the differences between *Dnmt3l*-KO and WT cells (**Figure 2.5D**). Top deciles represented regions that require DNMT3l for enhanced *de novo* methylation. These showed above 25% of methylation decrease in *Dnmt3l*-KO compared to WT cells. By contrast, lower deciles represented DNMT3l-independent regions that showed no methylation difference between *Dnmt3l*-KO and WT cells. Interestingly, DNMT3l-independent regions tended to locate more often in intragenic sequences than DNMT3l-dependent ones (**Figure 2.5D**, right panel). To sum up, this shows that gene bodies have a faster rate of methylation than the rest of the genome and do not require DNMT3l.

It was recently shown that DNMT3b is recruited to the body of actively transcribed genes *via* binding of its PWWP domain to H3K36 methylation (Baubec et al., 2015; Morselli et al., 2015). To analyze whether fast-acquiring and DNMT3l-independent regions that localized in gene bodies were affected by the transcriptional status of the underlying gene, we used the RNA-seq data of WT D3 sample to separate genes into five classes depending on their transcriptional status: lowest expression, low expression, mid expression, high expression, highest expression. We then plotted average DNA methylation on and around meta-genes representing these five classes (**Figure 2.5E**). As expected, gene bodies were hypermethylated and promoter regions hypomethylated. By comparing the five transcriptional classes, we could observe that gene body methylation was systematically higher for highly transcribed genes than lowly expressed ones, for both WT and *Dnmt3l*-KO cells, at both D3 and D7. This showed that highly transcribed regions tend to methylate faster. Moreover, differences between *Dnmt3l*-KO and WT cells were minimal for highly transcribed genes (**Figure 2.5E**, left panels), whereas lowly expressed genes behave like the rest of the genome, with a delay of around 20% methylation in absence of DNMT3l (**Figure 2.5E**, right panels). Screenshots of representative genomic regions showed indeed that DNA methylation levels were similar

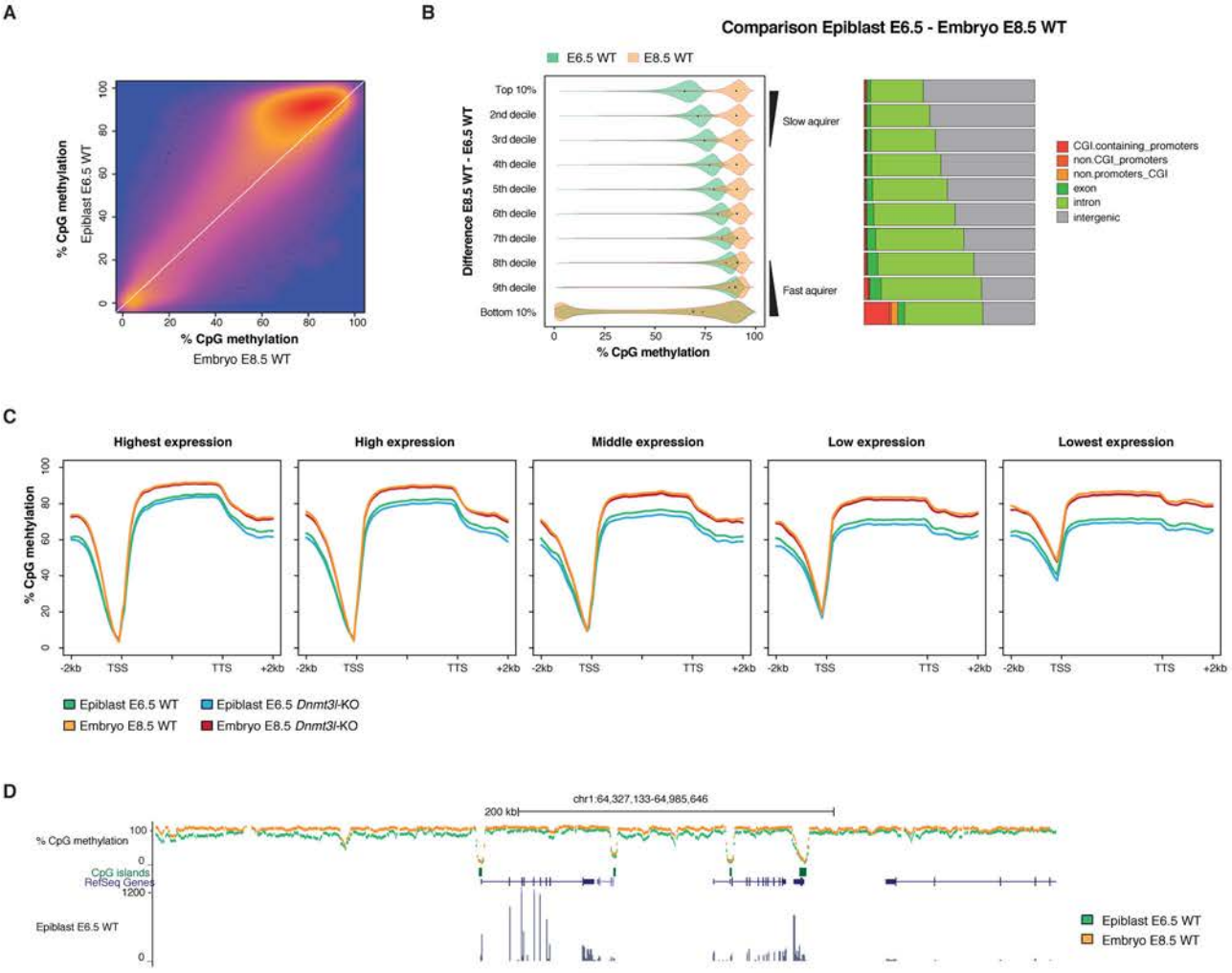


Figure 2.7. Highly transcribed regions acquire methylation faster in post-implantation embryos

A. Density map comparing average CpG methylation at 1kb sliding windows between E6.5 and E8.5 WT embryos.

B. Left panel: Violin plots representing average CpG methylation at 1kb sliding windows. The genome was separated into 10 groups depending on DNA methylation differences between E6.5 and E8.5 WT embryos. Deciles separate regions that gain DNA methylation slowly or rapidly during differentiation. Right panel: Annotation of each 1kb window into different genomic compartments.

C. Average CpG methylation over metagenes, separated in five classes depending on transcriptional level in E6.5 epiblasts.

D. Representative genomic region showing DNA methylation level in transcriptionally silent or active regions. RNA-seq data represent normalized read count.

between *Dnmt3l*-KO and WT cells in the body of highly transcribed genes, whereas DNA methylation acquisition was retarded in the rest of the genome (**Figure 2.5F**)

Overall, these results suggested that in ES cells, highly transcribed regions tend to methylate faster than the rest of the genome, and this occurs in a DNMT3l-independent manner.

DNMT3l is mostly dispensable *in vivo* for embryonic *de novo* methylation

The differentiation protocol that we used to follow the dynamic of *de novo* methylation – from ES cells to EpiLC – is supposed to recapitulate the transition from ground state blastocyst cells to committed epiblast cells. In the hope of observing *in vivo* a similar delayed methylation in absence of DNMT3l, we performed WGBS in 6.5 days (E6.5) epiblasts and E8.5 embryos, in WT and *Dnmt3l*-KO littermates. Two biological replicates were used for E8.5 embryos, and four for E6.5 epiblasts, separating males and females: this was motivated by the fact that female embryos have a delayed kinetics of development (Schulz et al., 2014). Sequencing was carried out at a reduced depth compared to ES cells, with around 10x coverage for each sample (**Table S2.1**). In good agreement with other studies (Wang et al., 2014b), average CpG methylation ranged between 60.0 and 68.9% in the eight E6.5 epiblast samples, and between 78.0 and 84.5% in the four E8.5 embryos. Importantly, no global difference could be observed between *Dnmt3l*-KO and WT samples, either at the level of the genome or in any genomic compartments (**Figure 2.6A** and **B**). Moreover, clustering of E6.5 samples failed to group WT and *Dnmt3l*-KO datasets together, indicating that the impact of DNMT3L in embryonic methylation is minimal (**Figure 2.6C**).

We also carried RNA-seq of E6.5 epiblasts, in four biological replicates for WT and *Dnmt3l*-KO contexts, with again in total four males and four females (**Table S2.1**). As for the methylation analysis, clustering of E6.5 epiblast RNA-seq datasets failed to separate WT and *Dnmt3l*-KO samples (**Figure 2.6D**). More detailed inspection revealed in fact that *Dnmt3l* was the one and only gene that was significantly misregulated between WT and *Dnmt3l*-KO embryos (**Figure 2.6E**). The germline genes *Dazl*, *Mov10l1* and *Tuba3a* were nonetheless upregulated around 2-3 fold in average between *Dnmt3l*-KO and WT epiblasts, but could not reach our significativity threshold, probably because these genes were very lowly expressed in both genotypes. The upregulation of these genes correlated with a modest reduction of promoter CGI methylation in E6.5 epiblasts, as exemplified by *Dazl* in **Figure 2.6F**.

In general, among the subset of CGIs that gained DNA methylation, we could observe a very slight delay in E6.5 *Dnmt3l*-KO epiblasts, for CGIs localizing either in promoters or in

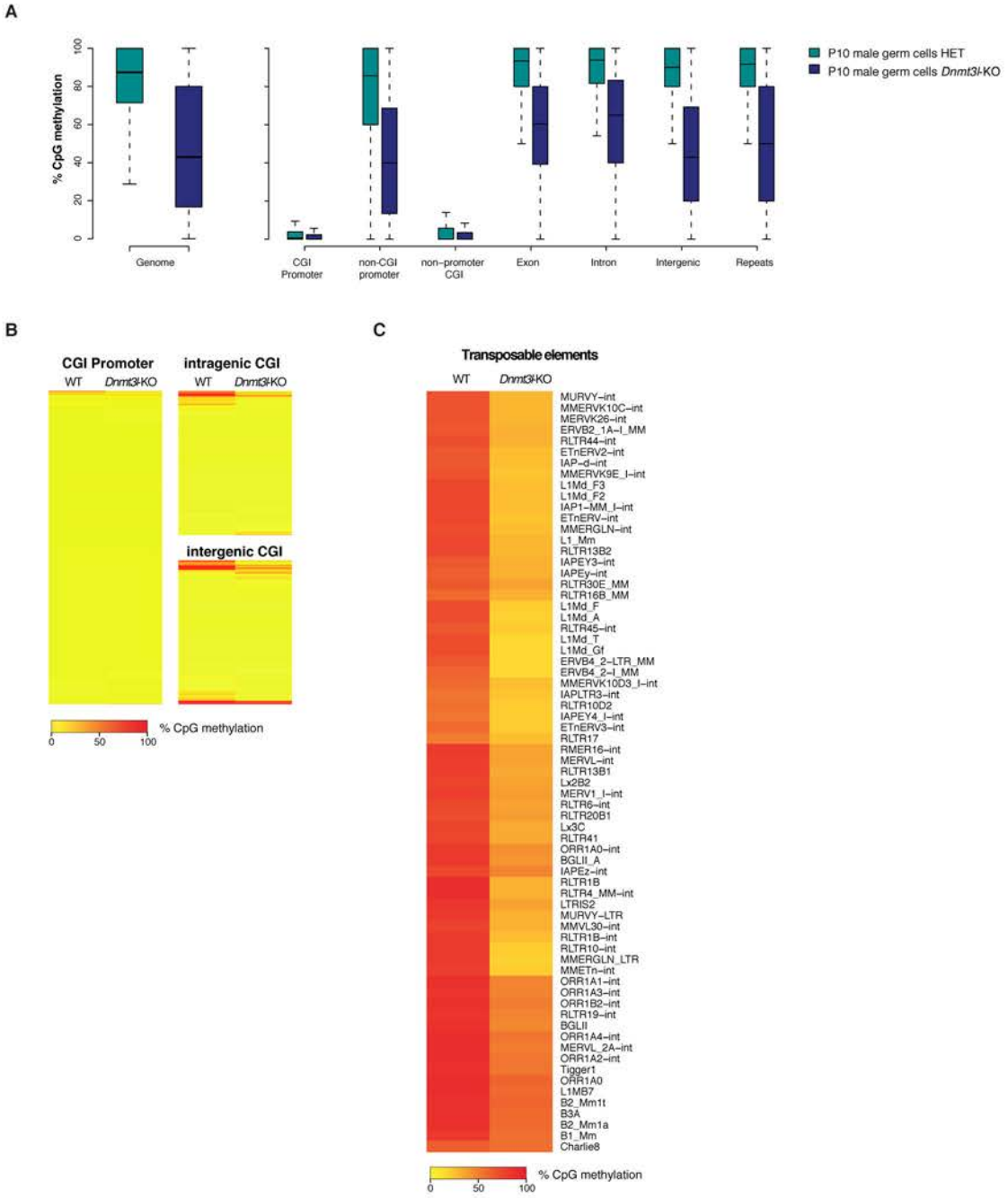


Figure 2.8. Male germ cells are severely hypomethylated in absence of DNMT3l

A. Left panel: Genome-wide CpG methylation content as measured by WGBS in P10 male germ cells, in WT and *Dnmt3l*-KO. Three (for WT) and two (for *Dnmt3l*-KO) biological replicates were combined. Right panel: CpG methylation distribution over different genomic compartments by WGBS.

B. Heatmap and hierarchical clustering of average DNA methylation at CGIs as measured by WGBS.

C. Heatmap and of average CpG methylation over 69 transposon families as measured by WGBS.

inter- and intragenic regions (**Figure 2.6G**). No difference could be observed for the methylation of imprinted control regions (around 50% as anticipated) or transposable elements (**Figure 2.6 – supplemental figure 1A, B** and **C**). We found that 225 CGI promoters had gained significant DNA methylation in E8.5 embryos, a number markedly smaller than the 431 CGI promoters that we found *de novo* methylated during ES cell differentiation. Interestingly in general, the majority of CGIs that gained DNA methylation *in vivo* also gained DNA methylation in ES cells, but conversely many methylated CGIs in ES remained unmethylated *in vivo,* especially those associated with promoter regions (**Figure 2.6 – supplemental figure 1D** and **E**). This indicates that the EpiLC differentiation protocol that we used did not faithfully reproduce *de novo* methylation *in vivo*.

DNA methylation is acquired faster at highly transcribed regions during embryonic *de novo* DNA methylation

In E6.5 epiblasts, deposition of DNA methylation is not terminated yet, with an average level of ~65% of CpG methylation compared to 85% in E8.5 embryos. Levels of methylation are, however, not homogeneous genome-wide at E6.5, with some regions acquiring DNA methylation faster than others (**Figure 2.7A**). To understand what distinguished fast and slow acquirers of DNA methylation *in vivo,* we partitioned the genome into 1kb windows and separated the genome into 10 groups, depending on the DNA methylation difference between E6.5 and E8.5 embryos (**Figure 2.7B**). As during ES cell differentiation and in an even more pronounced manner, we observed that fast acquiring regions tended very often to be in intragenic position compared to slow acquiring regions. To analyze if the preferential localization of fast acquiring regions in gene bodies correlated with transcriptional levels, we used E6.5 RNA-seq data to separate genes into five groups depending on their transcriptional strength (**Figure 2.7C** and **D**). As during ES cell differentiation, highly transcribed genes tended to gain DNA methylation faster than transcriptionally silent regions. However, in that case DNMT3l absence had barely any consequences.

This observation confirmed that during embryonic *de novo* methylation, highly transcribed regions methylate at a faster rate than the rest of the genome.

Male germ cells are severely hypomethylated in absence of DNMT3l

Male and female *Dnmt3l*-KO mice are infertile. *Dnmt3l*-KO oocytes are completely unmethylated and their progeny dies of imprinting defects around mid-gestation (Bourc'his et al., 2001; Smallwood et al., 2011). Male germ cells on the other hand cannot complete meiosis and *Dnmt3l*-KO males do not produce mature sperm. In particular transposable elements are

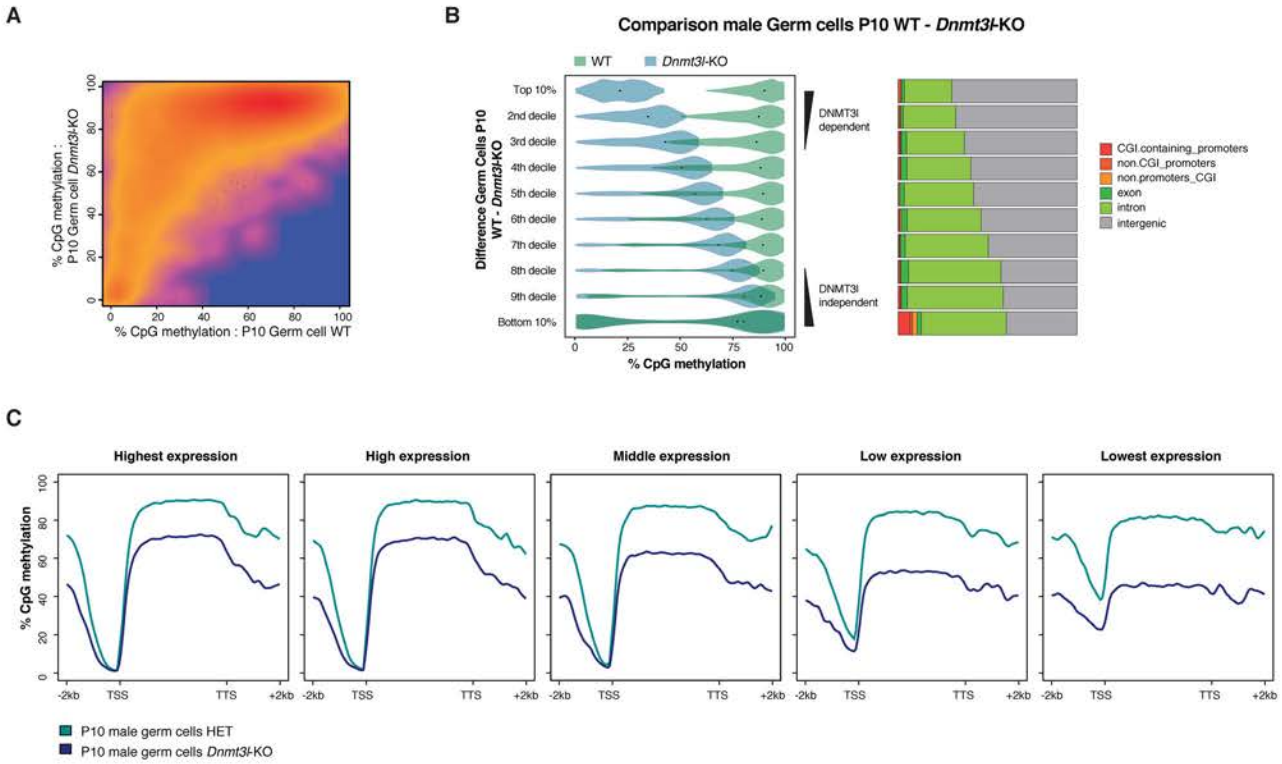


Figure 2.9. Hypomethylation of *Dnmt3l*-KO male germ cells is attenuated in highly transcribed regions

A. Density map comparing average CpG methylation at 1kb sliding windows between WT and Dnmt3l-KO P10 male germ cells

B. Left panel: Violin plots representing average CpG methylation at 1kb sliding windows. The genome was separated into 10 groups depending on the DNA methylation differences between *Dnmt3*l-KO and WT P10 male germ cells. Deciles separate regions that gain DNA methylation in a Dnmt3l-dependent or independent manner. Right panel: Annotation of each 1kb window into different genomic compartments.

C. Average CpG methylation over metagenes, separated in five classes depending on transcriptional level in P10 male germ cells.

severely hypomethylated and reactivated, which modifies the regulatory landscape of meiotic recombination and leads to spermatogenesis interruption (Bourc'his and Bestor, 2004; Zamudio et al., 2015). However, DNA methylation defects caused by DNMT3l deficiency were never analyzed genome-wide in male germ cells. We therefore performed WGBS in male germ cells of ten days old (P10) *Dnmt3l*-KO pups and their heterozygote littermates. At P10, germ cells just enter meiosis. In particular, *Dnmt3l*-KO testes have not experienced yet massive germ cell loss and the transcriptional differences between WT and *Dnmt3l*-KO germ cells are minimal (Zamudio et al., 2015).

WGBS was performed at a low coverage, and showed severe hypomethylation of *Dnmt3l*-KO germ cells. CpG methylation levels were reduced from 71% to 43.5%, and this hypomethylation affected every genomic compartments (**Figure 2.8A**). Compared to ES cells or post-implantation embryos, few CGIs were significant methylated, especially for CGI located in gene promoters (**Figure 2.8B**). This is in agreement with previous published studies of DNA methylation profiling performed at later stages of male germ cell development, in spermatozoa (Smallwood et al., 2011). As the rest of the genome, the few CGIs that gained DNA methylation in P10 germ cells were severely hypomethylated in absence of DNMT3l. Similarly, most transposable element families had severely reduced DNA methylation (**Figure 2.8C**), probably contributing to the massive up-regulation of transposons during meiosis.

DNA methylation defects caused by DNMT3l deficiency in male germ cells are attenuated in highly transcribed regions

DNA methylation defects in *Dnmt3l*-KO male germ cells are not homogeneous genome-wide. Some regions have highly reduced DNA methylation, while others seem almost unaffected (**Figure 2.9A**). We partitioned the genome into 1kb windows and separated the genome into 10 groups, depending on the DNA methylation differences between *Dnmt3l*-KO and WT germ cells (**Figure 2.9B**). Top deciles represented regions that are highly dependent upon DNMT3l for *de novo* methylation, as attested by the important differences between and *Dnmt3l*-KO and WT methylation levels (around 60%). By contrast, lower deciles represented DNMT3l-independent regions that showed no striking differences in *Dnmt3l*-KO and WT genotypes. As during ES cell differentiation, DNMT3l-independent regions tended to be more often found in intragenic sequences than DNMT3l-dependent ones (**Figure 2.9B**). It appears therefore that regions that get methylated in a DNMT3l-independent manner were often found in gene bodies.

To analyze if the preferential localization of DNMT3l-independent regions into gene bodies correlated with transcriptional levels, we used RNA-seq data generated in a previous

study, to separate genes into five transcriptional groups (**Figure 2.9C** and Zamudio et al., 2015). As ES cells and in early embryos, highly transcribed genes were more densely methylated than transcriptionally silent regions in WT post-natal germ cells. More importantly, the hypomethylation created by DNMT3l deficiency was less pronounced in the body of highly transcribed regions. This confirms our original observation made in ES cells that highly transcribed regions can be methylated efficiently, even in absence of DNMT3l.

2.4 Discussion

This work analyzes for the first time the contribution of DNMT3l during different developmental contexts of *de novo* methylation, highlighting common and specific themes in the rules governing the establishment of DNA methylation patterns. Specifically, we analyzed *de novo* methylation in three different contexts: during ES cell differentiation, in post-implantation embryos and in the pre-pubere male germline. In these three cases, the global dynamics of DNA methylation followed similar rules, but differed in term of DNMT3l dependency.

DNMT3a and DNMT3b physically recognize H3K36 methylation via their PWWP domain, a chromatin mark associated with transcriptional elongation and present mostly in the body of highly transcribed regions. As a consequence, in ES cells or upon artificial expression in yeast, DNMT3b physically localizes to highly transcribed regions and H3K36me3 mediates this attraction (Baubec et al., 2015; Dhayalan et al., 2010; Morselli et al., 2015). Here we demonstrate that recruitment of *de novo* DNA methylation machinery to H3K36me3 is a potent feature of *de novo* methylation.

Our second finding is that DNMT3l deficiency had barely any effect on the methylation rate of highly transcribed regions, but significantly impacted on DNA methylation deposition in the rest of the genome. Consistent with the absence of developmental defects in *Dnmt3l*-KO embryos, analysis of DNA methylation and transcriptional patterns of *Dnmt3l*-KO E6.5 epiblasts failed to detect any difference with WT, showing that DNMT3l plays virtually no role in this context. By contrast, *Dnmt3l*-KO male germ cells were severely hypomethylated before meiosis onset. In particular, work in the lab showed with my contribution that the absence of methylation in TEs lead to their reactivation and the persistence of activating chromatin marks. This causes a relocalization of developmentally programmed meiotic double strand breaks to TE sequences and cause major meiotic catastrophe (Zamudio et al., 2015).

During ES cell differentiation, DNA methylation was globally delayed in absence of DNMT3l, except in the body of highly transcribed genes where the difference was minimal. Even in severely hypomethylated *Dnmt3l*-KO male germ cells, DNA methylation differences were milder in highly transcribed regions. DNMT3l lacks a PWWP domain and has therefore no particular affinity for H3K36me3. Our results showed that DNMT3a/b enzymes do not require DNMT3l for efficient targeting and DNA methylation deposition at H3K36me3-enriched regions. DNMT3a/b enzymes are probably efficiently targeted through their PWWP domain. By contrast, DNMT3l is necessary to accelerate DNA methylation deposition in regions where DNMT3a and b are not naturally attracted.

ES cells are grown in 2i+vitamin C are almost completely unmethylated and *Dnmt3a, 3b* and *3l* are expressed at low level (Leitch et al., 2013). I was surprising to notice that during ES cells differentiation, the biggest effect on gene expression of DNMT3l depletion occurred at D0, before differentiation even started. The majority of differentially expressed genes were moreover upregulated in *Dnmt3l*-KO ES cells, indicating that DNMT3l could play a repressive role independently of DNA methylation deposition. Unexpectedly, the majority of upregulated genes had promoters marked by H3K27me3, suggesting that Polycomb-mediated repression could be alleviated in absence of DNMT3l.

It has been reported in hypermethylated serum-grown ES cells that Dnmt3 enzymes physically interact with PRC2 (Neri et al., 2013b). In the model proposed in this study, PRC2 naturally recruits DNMT3a/3b, while DNMT3l presence prevents this interaction and, excludes DNA methylation deposition at Polycomb targets. As a consequence, in serum-based conditions and in absence of DNMT3l, some Polycomb targets such as *Rhox* genes get *de novo* methylated and repressed. However, we did not observe similar effects during ES cell differentiation. In particular, we could not detect any Polycomb-target genes gaining DNA methylation in absence of DNMT3l, therefore revising the model proposed in Neri et al., 2013b.

What form could take a mechanistic link between PRC2 and DNMT3l in hypomethylated cells is a novel question. If validated, these preliminary results could lead to interesting area of investigation

2.5 Experimental procedures

ES cell lines

J1 mouse ES cells, from a male 129S4/SvJae background were used as WT in this study. *Dnmt3l*-KO ES cells were generated from J1 ES cells using CRISPR/Cas9 editing. Briefly, guide-RNAs specific to the target sequences were designed using the online CRISPR Design Tool (Hsu et al., 2013)(Table S2.?) and incorporated into the X330 backbone (Cong et al., 2013). Five millions J1 ES cells grown in serum were transfected with 1µg of plasmid using Amaxa 4d nucleofector (Lonza) and plated at a low density. Individual clones were picked and screened by PCR; mutated alleles were confirmed by Sanger sequencing. *Dnmt3l*-KO cells were obtained by deleting exon 2.

ES cell culture

Undifferentiated ES cells were grown in two different media, serum and 2i, defined as follow. Serum: Glasgow medium (Sigma), 15% FBS (Gibco), 2mM L-Glutamine, 0.1mM MEM non essential amino acids (Gibco), 1mM sodium pyruvate (Gibco), 0.1mM β-mercaptoethanol, 1000U/mL leukemia inhibitory factor (LIF, Miltenyi); 2i: N2B27 medium (50% neurobasal medium (Gibco), 50% DMEM/F12 (Gibco), 2mM L-glutamine (Gibco), 0.1mM β-mercaptoethanol, Ndiff Neuro2 supplement (Milipore), B27 serum-free supplement (Gibco)) supplemented with 1000U/mL LIF, 3µM Gsk3 inhibitor CT-99021, 1µM MEK inhibitor PD0325901. Vitamin C (Sigma) was added at a concentration of 100ug/mL (Blaschke et al., 2013).

Cells were grown in feeder-free conditions on gelatin-coated plates. ES cells in serum were passaged with TrypLE Express Enzyme (Life Technologies), whereas. 2i ES cells were harvested with Accutase (Life Technologies).

To induce differentiation towards EpiLCs, cells were plated at a density of 10^4 cells/cm^2 on Fibronectin (10 µg/ml) coated plates in ground state medium. After one day, cells were gently washed with 1xPBS, and cultured in N2B27 medium supplemented with 12 ng/ml Fgf2 (R&D) and 20ng/ml Activin A (R&D). EpiLCs were passaged with Accutase at day 4 of differentiation.

Mice

Dnmt3l-KO and WT littermates mice were derived and bred in the pathogen-free Animal Care Facility of the Institut Curie (agreement number: C 75-05-18). All experimentation was approved by the Institut Curie Animal Care and Use Committee and

adhered to European and National Regulation for the Protection of Vertebrate Animals used for Experimental and other Scientific Purposes (Directive 86/609 and 2010/63). For embryo collection, females were euthanized by cervical dislocation. *Dnmt3l*-KO mice were originally described in (Bourc'his and Bestor, 2004; Bourc'his et al., 2001). E6.5 epiblasts were manually dissected from extra embryonic tissues.

Dnmt3l-KO and WT P10 germ cells were isolated and purified in the lab of Steve Jacobsen, as described in Pastor et al., 2014.

DNA methylation analyses

Genomic DNA was isolated using the GenElute Mammalian Genomic DNA Miniprep Kit (Sigma) with RNase treatment. Global CpG methylation levels were assessed using LUminometric Methylation Assay (LUMA) as described previously (Karimi et al., 2011a; Richard Pilsner et al., 2010). Briefly, 500ng of genomic DNA was digested with *MspI/EcoRI* and *HpaII/EcoRI* (NEB) in parallel reactions. *HpaII* is a methylation-sensitive restriction enzyme and *MspI* is its methylation insensitive isoschizomer. *EcoRI* is included as an internal reference. The overhangs created by the enzymatic digestion were quantified by Pyrosequencing (PyroMark Q24, Qiagen) with the dispensation order: GTGTGTCACACAGTGTGT. Global CpG methylation levels were calculated from the peak heights at the position 7,8,13,14 as follows: 1-sqrt($[p8*p14/p7*p13]_{HpaII}$ /$[p8*p14/p7*p13]_{MspI}$)

Quantification of global cytosine methylation by liquid chromatography and mass spectrometry was performed by Zymo Research.

Whole-Genome Bisulfite Sequencing libraries from cells and E8.5 days whole embryos were prepared from 50ng of bisulfite-converted genomic DNA using the EpiGnome/Truseq DNA Methylation Kit (Illumina) following the manufacturer instructions. Sequencing was performed in 100pb paired-end reads at a 30X coverage using the Illumina HiSeq2000 platform (Table S1.1C).

Whole-Genome Bisulfite sequencing libraries from E6.5 epiblast were prepared essentially as described in Smallwood et al., 2014. Whole-Genome Bisulfite sequencing libraries from P10 germ cells were prepared as described in Pastor et al., 2014.

RNA expression analyses

For cells, total RNA was extracted using Trizol (Life Technologies). Nanostring nCounter quantification was performed using 100ng of total RNA per sample on a custom expression Codeset (sequences available upon request). *Actin, Ppia, Gapdh* and *Arp0* were used

for normalization. Data are presented as the fold change compared to WT-D0. RNA-seq libraries were prepared from 500ng of DNase-treated RNA with the TruSeq Stranded mRNA kit (Illumina). Sequencing was performed in 100pb paired-end reads using the Illumina HiSeq2000 platform.

For E6.5 epiblast, total RNA was extracted using PicoPure RNA isolation kit (Applied Biosystem), including DNAse treatment (Qiagen). RNA-seq libraries were prepared from 10-20ng of DNase-treated RNA with the Ovation RNA-seq system (Nugen). Sequencing was performed in 100pb paired-end reads using the Illumina HiSeq2000 platform.

Western blotting

To prepare total protein extracts, cells were resuspended in BC250 lysis buffer (25mM Tris pH 7.9, 0.2mM EDTA, 20% Glycerol, 0.25M KCl and protease inhibitor coktail from Roche), sonicated and centrifuged to pellet debris. Proteins were quantified by Bradford assay. Proteins (10-20µg per gel lane) were separated by electrophoresis in 10% poly-acrylamide gels and transferred onto nitrocellulose membranes using the Trans-Blot turbo transfer system (Biorad). After incubation with primary antibodies and HRP-conjugated secondary antibodies, signal was detected using ECL prime kit (Amersham) and ImageQuant Las-4000 mini biomolecular Imager. Antibodies are listed in Table S1.7.

WGBS data analysis

Whole-genome bisulfite sequencing reads generated in this study were treated as follow. The first eight base pairs of the reads were trimmed using FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Adapter sequences were removed with Cutadapt v1.3 (https://code.google.com/p/cutadapt/) and reads shorter than 16bp were discarded. Cleaned sequences were aligned onto the Mouse reference genome (mm10) using Bismark v0.12.5 (Krueger and Andrews, 2011) with Bowtie2-2.1.0 (Langmead and Salzberg, 2012) and default parameters. Only reads mapping uniquely on the genome were conserved. Methylation calls were extracted after duplicate removal. Only CG dinucleotides covered by a minimum of 5 reads were conserved for the rest of the analysis.

The R-package Methylkit v0.9.2 (Akalin et al., 2012) was used to provide Pearson's correlation scores between samples. To analyze the distribution of CpG methylation in different genomic compartments, the mouse genome was divided into different partitions. The RefSeq gene annotation and the RepeatMasker database were downloaded from UCSC table browser and used for transcript and repeat annotations, respectively. Promoters were defined as the -1kb to +100pb region around transcription start sites. CGI (CGIs) were defined as in

Illingworth et al., 2010. Intergenic partitions were defined as genomic regions that did not overlap with promoters, CGI, exons, introns or repeats. Whole-genome mapping of CpG methylation was then intersected with the different genomic compartments using Bedtools (Quinlan and Hall, 2010).

Average CpG methylation on individual transposons was extracted from RepeatMasker with Bedtools, average CpG methylation in the different transposon families was calculated and plotted using R. Heatmap for average CpG methylation in Imprinted control regions (ICRs) and CGI was generated similarly after retrieving ICR genomic coordinates from the WAMIDEX database (Schulz et al., 2008).

In order to compare DNA methylation level between conditions, the mouse genome was partitioned into 1-kb sized windows with an overlap of 500 bp. Windows overlapping with satellite annotated in the Repeat Masker database were removed. Windows with at least 5 CpGs were kept. Windows with less than 50% of CpGs covered at least 5 times were removed.

Metaplots showing DNA methylation across gene bodies were created using deepTools v1.5.11 (Ramírez et al., 2014).

RNA-seq data analysis

In order to quantify gene expression, Paired-end 2x100bp reads were mapped onto mm10 using Tophat v2.0.6 and RefSeq gene annotation (Kim et al., 2013) allowing five mismatches. Gene-scaled quantification was performed with HTSeq v0.6.1 (Anders et al., 2014).

In order to quantify transposon expression, reads mapping to ribosomal RNA (rRNA) sequences (GenBank identifiers: 18S NR_003278.3, 28S NR_003279.1, 5S D14832.1, 5.8S K01367.1) were first removed with Bowtie v1.0.0 allowing three mismatches. The rRNA-depleted libraries were then mapped onto mm10 using Bowtie v1.0.0 allowing zero mismatch and 10,000 best alignments per read. Exonic reads were removed. In order to count reads mapping to transposable elements, reads were weighted by the number of mapping sites and each library was intersected with the reconstructed RepeatMasker annotation, conserving only reads overlapping at least at 80% with a given transposon.

For each library, read counts for genes and transposons were combined into a single table. TMM normalization from the edgeR package v3.6.2 (Robinson and Oshlack, 2010) was first applied. As described in the guideline of limma R-package v3.20.4, normalized counts were processed by the voom method (Law et al., 2014) to convert them into log2

counts per million with associated precision weights. The differential expression was estimated with the limma package. Genes and transposons were called differentially expressed when two criteria were met: 1) the fold-change between two conditions was higher than two and 2) the adjusted p-value using the Benjamini Hochberg procedure was below 0.05.

2.6 Annexes

Supplemental information

Author contribution

M.W. and D.B. conceived the study, analyzed the data and wrote the manuscript. M.W. performed most of the experiments. Bioinformatics analyses were conducted by A.T. and M.W. E6.5 and E8.5 embryo dissection were carried out by M. W and J. I. WGBS libraries for E6.5 embryo were performed in the lab of G.K. by S.S. Germ cells isolation and subsequent WGBS were performed in the lab of S.J. by W.P.

Acknowledgments

We thank a lot of people

2.7 Supplemental tables

#Sample identifier	Biological identifier	Number of total sequenced tags	Number of cleaned sequenced tags	Number of paired-end alignments with a unique best hit	Mapping efficiency	Total count of deduplicated leftover sequences	C methylated in CpG context	C methylated in CHG context	C methylated in CHH context	%CpGs covered by at least 1x	% covered by at least 5x
A274-B113T1	J1 2i+vitC	447494671	447404318	270945884	60,55951476	190029868	4,6	0,7	0,8	81,14354349	55,37614374
A274-B113T2	EpiLC J1 D3	550122845	549935878	317207185	57,68075837	220289229	65,2	3,4	2,8	81,52976072	57,33153603
A274-B113T3	EpiLC KO D3	538454334	538125998	299973985	55,74419116	212560285	79,6	2	1,3	80,39517103	55,74751023
A274-B113T4	EpiLC J1 D7	566803502	566680527	348404646	61,48166902	247015079	52,2	10,8	11,8	86,08823084	64,08559144
A274-B113T5	EpiLC KO D7	527720401	527587798	323564790	61,32908896	224245029	76,3	2,9	2,1	82,76150438	59,35955473
B100B1	E8.5 WT	610329483	609315652	319648760	52,46029032	183871786	78,5	1,1	0,8	83,16348068	54,11754003
B100B2	E8.5 KO	533246070	532262528	269953708	50,71815012	154703039	78	1,1	0,9	79,90874387	47,32324819
BC1	E8.5 WT	348589379	348589379	266905669	76,56735548	228119988	84,5	3,6	3,5	88,79648483	60,41644453
BC2	E8.5 KO	342063592	342063592	253532693	74,11858465	233939963	82,7	5,9	6,3	91,87460813	66,72885093
Epi3L_WT_male1	Epi3L_WT_male1	225605727	224227120	145322057	64,81020538	127733853	62,7	4,8	4	70,30083246	20,58354644
Epi3L_WT_female1	Epi3L_WT_female1	122437726	122136436	70738474	57,91758489	44074208	64,2	4,4	4,5	46,53280663	1,904735017
Epi3L_KO_female1	Epi3L_KO_female1	210040937	208732518	135374116	64,85530731	119657367	60	5	4,1	69,91864345	18,20863403
Epi3L_KO_male1	Epi3L_KO_male1	220260012	218742717	138739177	63,42573545	122655087	67,6	4,2	3,3	70,77727468	20,71301051
Epi3L_KO_female2	Epi3L_KO_female2	175015536	174604690	118730239	67,9994558	97675677	60	4,6	3,8	66,20796834	12,19263522
Epi3L_KO_male2	Epi3L_KO_male2	171227941	170793658	112665161	65,9656584	92624051	64,2	4,5	3,6	65,29097488	11,35107765
Epi3L_WT_female2	Epi3L_WT_female2	209379216	208464692	136405545	65,43340443	117492835	66,7	7,3	6,2	70,25303031	16,47620131
Epi3L_WT_male2	Epi3L_WT_male2	211493323	210824841	140798915	66,78478415	101361122	68,9	2,4	1,9	66,86457527	14,58466009
Het1_Lane1	male PGC het	27455451	27455073	19640096	71,53539894	18923786	71,6	0,5	0,5	26,34501731	0,018833296
Het1_Lane2	male PGC het	29782808	29782479	21524255	72,27153589	20723098	71,6	0,5	0,5	28,28962131	0,021631366
Het2_Lane1	male PGC het	25472258	25471992	18541610	72,79214755	17868882	71,8	0,5	0,5	24,93416106	0,017160392
Het2_Lane2	male PGC het	27461362	27461072	20213282	73,60703908	19466264	71,7	0,5	0,5	26,70729808	0,019381041
Het3_Lane1	male PGC het	26083368	26083094	18505854	70,94961204	17757826	69,6	0,5	0,5	24,7208806	0,017628257
Het3_Lane2	male PGC het	28193625	28193380	20262037	71,86806619	19427922	69,5	0,5	0,5	26,58278653	0,019737075
KO1_Lane1	male PGC ko	36162242	36161645	25899315	71,62095364	24885773	43,8	0,5	0,4	33,15463916	0,032691242
KO1_Lane2	male PGC ko	39123890	39123297	28381821	72,54455319	27245125	43,8	0,5	0,4	35,43423711	0,041359306
KO2_Lane1	male PGC ko	42267424	42267028	30278284	71,63570621	28990945	43,1	0,5	0,4	36,80892622	0,049226292
KO2_Lane2	male PGC ko	45569627	45569221	33092354	72,61996864	31649121	43,1	0,5	0,4	39,17474149	0,064186575

<u>Table S2.1A</u> – WGBS sequencing statistics.

#Sample identifier	Biological identifier	Number of total reads	Number of mapped reads	% of mapped reads
A250T1	J1 D0 WT	202041982	137720883	68.16%
A250T2	J1 D0 WT	236100808	157823606	66.85%
A250T3	J1 D0 KO	210173508	145849106	69.39%
A250T4	J1 D0 KO	238756957	163872599	68.64%
A250T5	J1 D3 WT	193691939	131803348	68.05%
A250T6	J1 D3 WT	217936286	149504559	68.60%
A250T7	J1 D3 KO	181017215	126727859	70.01%
A250T8	J1 D3 KO	176780767	124912879	70.66%
A250T9	J1 D7 WT	140166270	97857990	69.82%
A250T10	J1 D7 WT	135965099	93437301	68.72%
A250T11	J1 D7 KO	165732540	116499471	70.29%
A250T12	J1 D7 KO	156744423	110423954	70.45%
A250T13	E6.5 WT male	110443427	42217084	38.23%
A250T14	E6.5 WT male	93742425	42633013	45.48%
A250T15	E6.5 WT female	94637491	40965659	43.29%
A250T16	E6.5 WT female	89240951	40922311	45.86%
A250T17	E6.5 KO male	100724151	43152272	42.84%
A250T18	E6.5 KO male	115414837	48718190	42.21%
A250T19	E6.5 KO female	84415189	37058100	43.90%
A250T20	E6.5 KO female	110165726	48943458	44.43%

Table S2.1B – RNA sequencing statistics.

#Sample identifier	Biological identifier	Number of total reads	Number of mapped reads	% of mapped reads	% of duplicates
A312C1	InputWT	41896775	40959296	97.76%	18%
A312C2	WTD0H3K27m3	43140990	42347220	98.16%	17%
A312C3	WTD4H3K27m3	38225230	37384240	97.80%	13%
A312C4	WTD7H3K27m3	42228291	41313444	97.83%	12%
A312C5	Input_Liz	31467692	30883037	98.14%	17%
A312C6	LizD0H3K27m3	46172687	45336055	98.19%	8%
A312C7	LizD4H3K27m3	44463904	43487132	97.80%	10%
A312C8	LizD7H3K27m3	38301482	37404120	97.66%	11%

Table S2.1C – ChIP sequencing statistics.

Diff D0	Diff D3	Diff D7
A2m	Ace	Abcg1
Acap3	Anxa8	Ap3b2
Arhgef25	Aoah	Capn5
Atp6v0e2	Arl4d	Cpe
B4galt2	Asz1	Cpt1a
BC068157	Bcas1	Crlf2
Bcam	Cnpy1	D2hgdh
Bend7	Creld1	Dmrt3
Calb2	Ctcfl	Dnmt3l
Calca	Cyp2b23	Dppa4
Camk2b	D1Pas1	Efhc2
Cap2	D2hgdh	Epb4.1l1
Cbln1	Dazl	Esrrb
Ccng2	Dnmt3l	EU599041
Ccnjl	Dusp27	Fam178b
Crispld2	Epha4	Fgf15
Cyp2b23	Gabarapl1	Fgf4
D430019H16Rik	Glrx	Fzd10
D8Ertd82e	Gm10416	Ggt7
Dlg4	Gm10451	Gjb3
Dnmt3l	Gm364	Gm10324
Dpysl5	Gm6880	Gm13242
Dusp8	Hck	Gm13247
Ece1	Hsh2d	Gm53
Emb	Ifi35	Gm7325
Emp2	Irf2	Kifc3
Epb4.1l3	Lgals3bp	Kirrel2
Etv4	M1ap	Lama1
Evc2	Magea5	Laptm5
Evc	Magohb	Ldhc
Fah	Map7d2	Lefty2
Fam19a4	Mcf2	Ly6g6e
Fam64a	Mfap5	Magohb
Fgf17	Mgl2	Map7d2
Gap43	Nap1l5	Mpc1
Gbp9	Nlrp14	Nab2
Gm11627	Nlrp4c	Ndrg3
Gng3	Nlrp4f	Nphs1
Gpm6a	Nmi	Nxf3
H2-Eb1	Nup62cl	Pcsk1
Ido2	Nxf2	Pla2g5
Igfbpl1	Nxt2	Plcd1
Kif7	Oasl2	Plekhg5
Krt17	Parp9	Pramef12
Krt19	Plet1	Prdm1
Krt42	Plxnd1	Prdm14
Lppr2	Prtg	Pycard
Lppr3	Prune2	Rap1gap2
Map3k8	Reep1	Rex2
Mark1	Retsat	Rfx2
Mef2b	Scml2	Rnf17
Mfap4	Sdc3	Serpina3m
Mfsd7c	Sema4f	Slco3a1
Mmp14	Serpina3m	Snca
Nacad	Slc25a31	Sox2
Neil2	Smarca5-ps	Tfap2c
Nkd1	Sp110	Triml1
Nnat	Sp140	Triml2
Nos1	Taf9b	Trpv2
Nuak2	Tnfrsf19	Tubb3
Pcsk9	Tuba3a	Uba1y
Pianp	Tuba3b	Xlr5a
Pkhd1	Xaf1	Xlr5b
Plcd1	Xlr	Zcchc12
Podxl2	Zfp936	Zfp42
Prnp	Zfp951	Zfp600
Pros1	Znf41-ps	Zfp936
Ptrf	1700030C10Rik	1700048O20Rik
Rftn1	4930506M07Rik	2410141K09Rik
Rnf103		
Rnf144b		
Robo3		
Scara5		
Sdc2		
Sema5b		
Sft2d2		
Slain1		
Slc11a1		
Slc16a13		
Slc16a3		
Slc22a23		
Smad9		
Smarca5-ps		
Smarcd3		
Stac2		
Tdrkh		
Tgm1		
Tia1		
Trib2		
Tsc22d3		
Tuba1a		
Tubb4a		
Ube2ql1		
Zhx1		
1700030C10Rik		
5730559C18Rik		

Table S2.2. List of significantly differentiated genes during ES cells differentiations

GO biological process experimental only	upload_1 (fold Enrichment)	upload_1 (P-value)
meiotic DNA repair synthesis	> 5	3.49E-03
piRNA metabolic process	> 5	8.64E-11
DNA synthesis involved in DNA repair	> 5	8.42E-03
DNA methylation involved in gamete generation	> 5	3.21E-08
synaptonemal complex assembly	> 5	6.21E-05
reciprocal meiotic recombination	> 5	1.69E-03
reciprocal DNA recombination	> 5	1.69E-03
synaptonemal complex organization	> 5	1.98E-04
gene silencing by RNA	> 5	2.40E-03
synapsis	> 5	1.87E-07
DNA methylation	> 5	2.09E-06
DNA alkylation	> 5	2.09E-06
male meiosis	> 5	2.76E-06
DNA modification	> 5	9.36E-07
homologous chromosome segregation	> 5	1.52E-06
DNA methylation or demethylation	> 5	1.23E-05
meiosis I	> 5	2.22E-10
chromosome organization involved in meiosis	> 5	3.73E-06
meiotic nuclear division	> 5	6.15E-14
meiotic cell cycle process	> 5	9.08E-14

Table S2.3 – Gene Ontology enrichment analysis for genes gaining DNA methylation during ES cell differentiation. Gene ontology analysis for biological processes was performed for the 420 genes that gained CGI promoter DNA methylation.

DISCUSSION

3 Discussion

3.1.1 A cellular system to reproduce *in vivo* DNA methylation reprogramming

Living systems, either single cells or whole organisms, have the amazing and intrinsic capacity to continuously adjust their biology to respond to internal or external perturbations. Cells adapt to their environment and maintain an apparent stable state, or homeostasis, that reflects in fact a balance between countless dynamically regulated processes (Cannon, 1926). Upon perturbation and given enough time, cells reach a novel dynamically stable state that reflects the new conditions. Cells in culture are a particularly good example and maintain a stable state for unlimited periods of time, which can be different depending on the culture condition or the presence of genetic mutations. By contrast, cells in a developing mammalian embryo are continuously unbalanced, and cell fate and internal processes are constantly reacting to changing conditions. As a consequence, cells grown in a steady state in culture represent a poor model to reproduce the fundamentally dynamic evolution of early embryogenesis. ES cell differentiation is an easy and efficient way to introduce regulatory dynamism in cell culture and has been widely used to mimic the evolution of post-implantation embryos. Reproducing the dynamic changes that occurs in pre-implantation embryos or developing germ cells is on the other hand more challenging. During these reprogramming events, cells are dedifferentiating and no good cellular system has yet emerged to reproduce that event. Generation of iPS cells is, for example, a notoriously long and inefficient process that fails to mimic the rapid and highly coordinated reprogramming event of early embryos.

The principal interest of this thesis is to provide a model that reproduces in ES cells the dynamic evolution of one key aspect of embryonic reprogramming: DNA methylation erasure. By switching cells from serum to 2i+vitC media, CpG methylation reached 30% in four days – roughly the level of an E3.5 blastocyst – and ~10% in six days, which is approximately the time that migrating PGCs take to demethylate from E7.5 to E13.5. Moreover, upon loss of DNA methylation, we observed that H3K9me2 rapidly disappeared, that H3K9me3 remained globally stable, whereas H3K27me3 greatly increased, which is precisely what has been observed in E13.5 PGCs (Hajkova et al., 2008; Seki et al., 2007). Demethylation upon 2i+vitC conversion occurs at a pace comparable to *in vivo* events and appears to reproduce important feature of *in vivo* chromatin reorganization. This comforted us

that our cellular system might represent a valuable tool to study how the genome adapts to DNA methylation loss.

Our analysis focused both on short and long-term consequences of DNA methylation loss. In the early embryo and in the mouse male germline, periods of hypomethylation are very transient and are immediately followed by genome-wide events of *de novo* methylation. In humans however, male germ cells remain hypomethylated for several weeks (Gkountela et al., 2015; Guo et al., 2015; Tang et al., 2015). In addition, female germ cells remain completely hypomethylated for very long periods of time, several weeks in the mouse and up to 50 years in humans. Long-term hypomethylation does not exist only in cell culture, but represens a constitutive feature of germ cell development. Considering how critical is germ cell genetic integrity, it is likely that germ cells have developed alternative mechanisms that compensate for DNA methylation loss, especially related to TE regulation. By unraveling how TEs are regulated in hypomethylated cells, this work therefore provides insight into the mechanisms that could potentially be at play in germ cells.

Upon loss of DNA methylation in ES cells, we could identify two important phases. Correlating with DNA methylation disappearance, most of TEs in the genome were first activated, before being silenced again on the long term. Importantly, H3K27 methylation and Polycomb proteins become important regulators of TE expression in hypomethylated contexts. Moreover, some transposon families such as MERVL and probably others operate a complete switch from H3K9- to H3K27 methylation-based repressive pathways upon loss of DNA methylation. This observation was unexpected and represents one of the most interesting finding of this work. It would be of great interest to investigate in male and female germlines whether what we observed in our cellular system could be also relevant *in vivo,* for example by checking the effect of Polycomb disruption specifically in PGCs. A recent study performed ChIP-seq for H3K9me3 and H3K27me3 in E13.5 PGCs, and observed that H3K27me3 was precisely localizing in transposons (Liu et al., 2014). However, their data are intriguing because it appears than all the regions marked by H3K9 methylation are also enriched for H3K27me3. Reciprocally, when re-analyzing their data we observed that most regions that are typically exclusively marked by H3K27me3, such as *Hox* gene clusters, showed also important enrichment for H3K9me3. This raises questions about the specificity of antibodies used in this study, and prevents us to draw strong conclusion from this data or to use it to confirm our results (**Figure 3.1A**).

Overall, whether the work presented in this thesis has *in vivo* implications, and especially whether Polycomb could regulate TEs in fetal germ cells needs to be investigated.

3.1.2 DNA methylation has a role in TE repression in ES cells

In differentiated cells, loss of DNA methylation leads to important TE reactivation, especially of IAP elements (Hutnick et al., 2010; Walsh et al., 1998). By contrast *Dnmt*-tKO ES cells show only very modest reactivation of TEs, and this observation had led to the conclusion that TE repression occurs in a DNA methylation-independent manner in ES cells (Hutnick et al., 2010; Karimi et al., 2011b). Our work showed that this interpretation is erroneous. Indeed, upon conversion from serum to 2i+vitamin C, TEs were strongly up-regulated genome-wide, and this activation involved the majority of TE families. Moreover, this burst of expression could not be observed upon conversion of *Dnmt*-tKO ES that are constitutively unmethylated, showing that medium conversion alone does not cause TE activation. Importantly also, while the majority of changes in protein-coding gene expression occurred at the beginning of the conversion, TEs became activated only when DNA methylation reached a very low level. These observations strongly suggest that the loss of DNA methylation caused TE reactivation and indicates that DNA methylation is a potent suppressor of TE expression in ES cells. In our system, the silencing release is, however, very brief and alternative mechanisms take over rapidly to ensure the stability of TE expression.

The existence of compensatory mechanisms explains why DNA methylation was considered dispensable for TE repression. ES cells lines used to draw this conclusion, like *Dnmt*-tKO ES cells, have been in cultured for extended periods of time, giving enough time to implement DNA methylation-independent compensation.

One of the most serious limitation of our system is that 2i- and serum-grown ES cells are transcriptionally very distinct. In consequence, transient activation of TEs could potentially be essentially a consequence of medium conversion. However, while this work was being conducted, we became aware that other groups were observing a similar phenomenon, but using completely different systems. These studies are still unpublished but rely on conditional depletion of Dnmt1 in serum ES as a way to promote rapid demethylation. In a very reassuring manner, these groups observed a similar burst of TEs expression followed by a second phase of silencing. Moreover, and somewhat surprisingly, the timing of TE peak of expression coincided exactly with our own results, occurring exactly six days after Dnmt1 depletion. These independent results confirm our own experiments and strongly suggest that DNA methylation rapid loss caused this transient TE reactivation (Rebecca Berrens and Wolf Reik, 2015 EMBO Mobile Genome Meeting).

3.1.3 Pluripotency transcription factors may play a role in transient TE activation

Intriguingly, we observed that the burst of transposon expression coincided with an increased availability of core pluripotency transcription factors such as OCT4 or NANOG. In mice and humans, an important proportion of pluripotency binding motifs are localized in TEs (Kunarso et al., 2010) and at least in human ES cells, pluripotency factors were shown to drive TE expression (Grow et al., 2015; Wang et al., 2014a). The specific presence of pluripotent transcription factors therefore probably contributes to the global reactivation of TEs, but is difficult to explain. Indeed, pluripotent factors are among the only protein-coding genes that show specifically an increased expression at the time of transposon burst. In theory, this burst of transcription factor expression could be the sole responsible of the observed TE transient activation, except that it would then fail to explain why this reactivation is not observed in *Dnmt*-tKO ES cells as well.

Testing directly whether the presence of pluripotent factors contribute to TE expression is difficult to achieve. In serum ES cells, levels of some pluripotency factors such as NANOG are highly heterogeneous from cells to cells, but we could not detect by immunofluorescence any correlation between NANOG and IAP or LINE1 protein expression. Moreover, functional genetic studies of the role of pluripotency factors on TE expression are not really amenable, as depletion of any of the core pluripotent factors in ES cells would lead to loss of pluripotency and cell differentiation, changing cell fate too radically to hope drawing any conclusion. Since loss of function experiments would not be informative, I attempted to artificially over-express pluripotent transcription factors in ES by using an OSKM construct originally designed to generate iPS cells[20]. My hope was that pluripotent factors over-expression, either in serum or 2i+vitamin C conditions, would lead to TE transient reactivation. Unfortunately and probably because they are already very highly expressed in ES cells, I could only induce minimal over-expression of pluripotency factors and did not observe any effect on transposon expression.

Serum and 2i ES cells maintain pluripotency by targeting different regulatory pathways (Marks and Stunnenberg, 2014). Serum ES cells represent a metastable and primed pluripotent state, while 2i ES cells are considered to be in a "ground state" of pluripotency, which expresses pluripotent transcription factors at an higher and more homogeneous level (Marks et al., 2012). The transition from serum to 2i+vitamin C is an important perturbation of the equilibrium of regulatory pathways, and the first days of conversion probably represents

[20] In other words, I attempted to reprogram pluripotent cells.

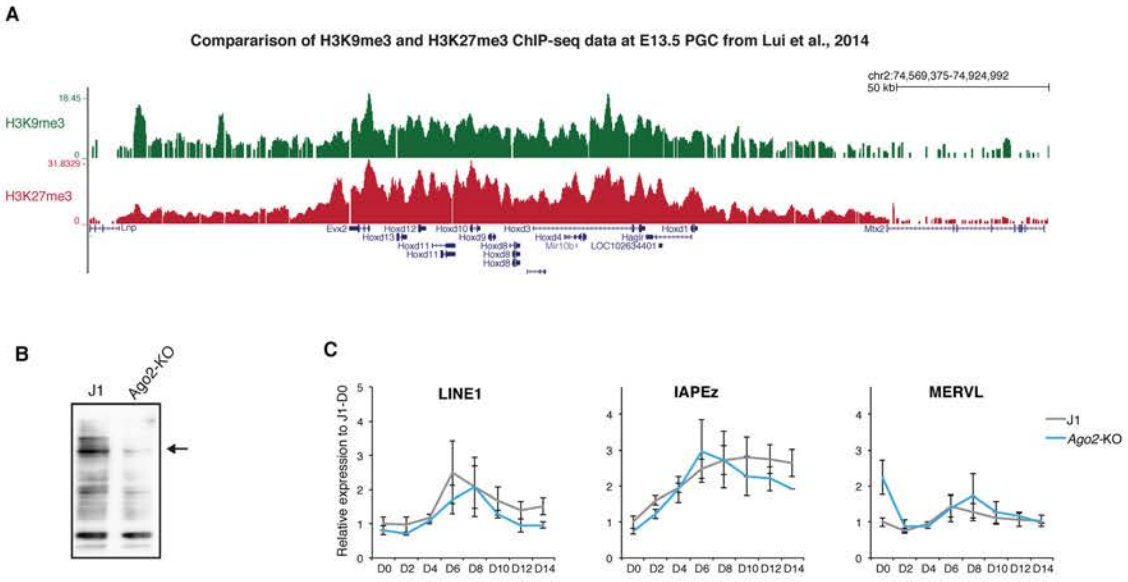


Figure 3.1 Small-RNA pathway is not involved in transposon repression upon loss of DNA methylation

A. ChIP-seq data from Liu et al., 2014 showing intriguing cololocalization of H3K27me3 and H3K9me3 in HoxD cluster in male E13.5 germ cells.

B. Western Blot showing loss of Ago2 protein in *Ago2*-KO ES cells.

C. Expression levels were measured by Nanostring nCounter. Data are expressed as fold changes to J1D0 and represent mean ± SEM between two biological replicates.

an unbalanced cellular state that oscilllates between two distinct equilibria. Speculatively, the transient abundance of pluripotency factors and the decrease that follows could be seen as an oscillation of regulatory networks, that would go "too far" before adjusting back to equilibrium condition. Whether this oscillation is mechanistically linked to DNA methylation loss, or whether the two events are merely correlative would be difficult to address.

3.1.4 What about the role of small RNA pathways?

Transposon repression upon 2i+vitC-induced loss of DNA methylation correlates with a reorganization of repressive chromatin marks, a finding that we explored at length. It is known that small RNA-based mechanisms can provide another layer of transposon repression, acting at the post-transcriptional level. In the male germline, the piRNA pathway is critical for TE silencing, but piRNAs are not found in any other cellular contexts including ES cells. In ES cells and elsewhere, small RNA production relies on the RNAse III ribonuclease Dicer that cleaves specific RNA precursors into ~21pb small RNAs, either miRNAs or small interfering RNAs (siRNAs). MiRNAs and siRNAs are loaded into Argonaute proteins and the complex can recognize complementary RNA targets, causing either translational inhibition or RNA degradation (Siomi and Siomi, 2009). AGO2 is the only Argonaute protein that can mediate cleavage of RNA targets and the three other members of the family lack this specific slicing activity (Liu et al., 2004). Importantly, Ago2 is absolutely required in the female germline (Kaneda et al., 2009) and associated siRNAs have been shown to repress TEs in the oocyte (Watanabe et al., 2008, 2011). Whether siRNA-mediated repression is active in other cellular contexts than the oocyte is under passionate debate. Small RNA silencing has been shown to silence *bona fide* viruses in ES cells and is often proposed to play a role in TE control as well (Maillard et al., 2013).

In order to test whether small RNA-based silencing could be implicated in TE repression in our system of induced loss of DNA methylation, I generated *Ago2*-KO ES cells with CRISPR/Cas9. Because AGO2 slicing activity is not required for the miRNA pathway, *Ago2* deletion would supposedly suppress siRNA silencing without affecting miRNAs. The knockout was generated by deleting *Ago2* second exon in J1 WT ES and was confirmed by Western Blot. Upon serum to 2i+vitC, I could not detect any significant difference in transposon expression between *Ago2*-KO and WT cells, except for MERVLs that were slightly upregulated in serum. Ago2-KO cells followed the same pattern of activation-repression, indicating that small RNA pathways (at least siRNAs) are unlikely to play a role in transposon control in our system (**Figure 3.1B** and **C**).

3.1.5 H3K9 methylation pathways play a complex role upon loss of DNA methylation

This work improves our comprehension of how H3K9-related pathway maintains TE silencing. In particular, analysis of H3K9-KMT mutants had until now been conducted almost exclusively in highly methylated serum ES cells and this work shed important light on how H3K9me3-silencing adapt to hypomethylated contexts.

Of interest, H3K9me2 is reduced to almost undetectable levels upon conversion from serum to 2i+vitamin C, and a similar reduction was observed in *Dnmt*-tKO ES cells, albeit not at the same extent. It is established from numerous reports that G9a can recruit Dnmt3a and b and accordingly, DNA methylation is reduced at TEs, major satellite repeats and promoters in *G9a*-KO ES cells (Dong et al., 2008; Epsztejn-Litman et al., 2008). Our results indicate conversely that DNA methylation presence might be necessary for G9a/GLP activity and H3K9me2 deposition. Interestingly, we observed that upon conversion from serum to 2i+vitamin C, the burst of transposon expression at day 6 was amplified in *G9a*-KO, even though H3K9me2 had already disappeared at this time point in WT cells. DNA methylation recruitment by G9a was shown to be independent of its catalytic activity and our results suggest similarly that during 2i+vitC conversion, G9a might play a repressive role independent of H3K9me2 deposition (Dong et al., 2008). How G9a, H3K9me2 and DNA methylation mechanistically interact would need to be properly investigated.

Upon loss of DNA methylation, H3K9me3 pattern remain globally untouched. Surprisingly, many TE families with a high enrichment for this repressive mark are nonetheless transiently reactivated, showing that H3K9me3 itself is not sufficient to maintain silencing. This suggests that H3K9me3 readers might be destabilized by DNA methylation disappearance or by the 2i+vitC medium composition . The MBD protein MeCP2 associates with Suv39h enzymes, binds to methylated DNA and repress transcription of TEs in certain contexts (Lunyak et al., 2002; Muotri et al., 2010; Nan et al., 1998). Upon loss of DNA methylation, MeCP2-mediated repression at H3K9me3 sites could be alleviated, potentially contributing to repression release. It would be also of interest to analyze whether downstream actors of the H3K9 methylation pathway, such as HP1 or KAP1, remain bound to TEs when they are activated but still enriched in H3K9me3.

Even thought H3K9 methylation is not sufficient to prevent TE burst of activation, our results show H3K9-mediated repression is critical for long term re-silencing. KAP1 is critical for ES cell survival (Rowe et al., 2010); therefore, I could only study Kap1 heterozygous mutants. Interestingly, half the dose of KAP1 protein was sufficient to maintain IAP

repression in serum, likely because these cells are hypermethylated, but not in 2i+vitamin C, when DNA methylation disappears. Moreover, H3K9me3 levels increased at IAPs in 2i+vitamin C, suggesting that upon loss of DNA methylation, long-term maintenance of silencing of IAP elements may require a reinforcement of Kap1/H3K9me3-related repression. To further test that hypothesis, I recently deleted one allele of ESET in ES cells (complete *Eset*-KO is also cell lethal, Matsui et al., 2010), which will allow me to compare IAP repression in serum and 2i+vitamin C and determine if ESET-mediated silencing efficiency is affected by DNA methylation loss. If validated, this would suggest that in absence of DNA methylation, IAP silencing is indeed achieved by a reinforcement of H3K9me3-related silencing. This strategy might be specific to the IAP superfamily. Indeed, IAP and related MMERVK10C elements are the most strongly reactivated TEs in *Eset* knockout (Karimi et al., 2011b). They are also among the only transposons that are covered on their full-length by H3K9me3 and that never gain any observable H3K27me3, indicating that their dependency on ESET/KAP1 is particularly strong.

Except for MERVL elements that mostly rely on G9a, the majority of TEs in serum-grown ES cells are repressed in a KAP1/ESET-dependent manner (Karimi et al., 2011b; Matsui et al., 2010; Rowe et al., 2010). It was recently proposed that SUV39h enzymes were adding another layer of repression on TEs to form large H3K9me3 domains spreading around from KAP1 binding sites (Bulut-Karslioglu et al., 2014). However, we found that knockout of SUV39h has little effect on TE repression, at the exception of MERVL elements and of one specific LINE family (L1-A). Upon loss of DNA methylation in *Suv39h*-dKO, we were puzzled to observe that IAP expression remained at a constant level and did not go trough the typical burst of transcription. Of note, the flat curve of IAP expression during conversion is identical in *Suv39h*-dKO and *Dnmt*-tKO cells. In *Dnmt*-tKO cells, IAP elements have already adapted to the absence of DNA methylation, which explains that they are not reactivated. Considering the numerous mechanistic links between SUV39h enzymes and DNA methylation, it is tempting to speculate that in *Suv39h*-dKO, IAP have also already adapted to the lack of SUV39h, and that this adaptation suppresses the dependency on DNA methylation. In this scenario, lack of SUV39h or DNA methylation would have the same effect: the implementation of DNA methylation-independent repression.

What form would take these SUV39h- and DNA methylation-independent mechanisms is an interesting question. One particular challenge of transposon biology is to take into account the variability of elements inside a single family. It is probably wrong to assume that every elements are regulated the same way, especially because repressive strategies likely

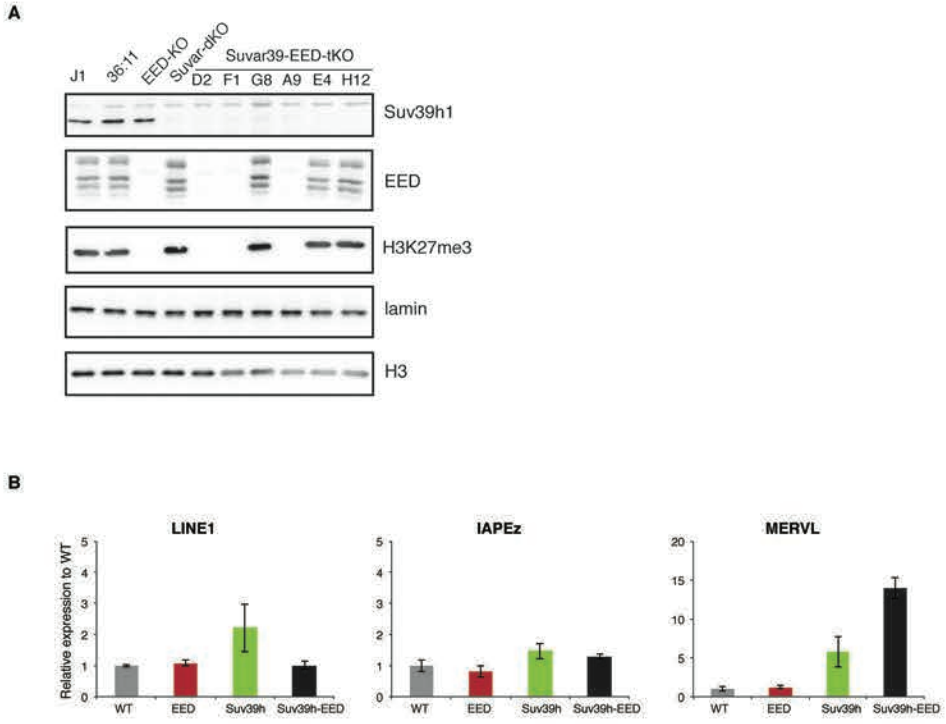


Figure 3.2. Polycomb represses MERVL transposon in absence of Suv39h

A. Western Blot showing loss of Suv39h1, EED and H3K27me3 in *Suv39h-EED*-tKO ES cells. Clone D2, F1 and A9 were used as biological replicate in **B**.

C. Expression levels were measured RT-qPCR in ES cells grown in serum. Data are expressed as fold changes to WT (J1 ES cells) and represent mean ± SEM between two (for WT and *EED*-KO) or three (*Suv39h*-dKO and *EED-Suv39h*-tKO) biological replicates

depend upon the genomic context. For example, even though H3K27me3 cannot be detected at IAP sequences by ChIP-seq, IAP elements are strongly upregulated in 2i+vitC in *Eed*-KO ES cells, showing that IAPs are at least genetically dependent upon PRC2. IAPs are young TEs that are almost impossible to distinguish from one to another. In consequence, only a few elements could rely on Polycomb, which would be impossible to observe by ChIP but would nonetheless result in strong transcriptional up-regulation. Upon loss of DNA methylation, MERVL elements operate a complete switch from G9a- and SUV39h-based repression to Polycomb-mediated silencing. I reasoned that this hidden minority of IAPs could behave like MERVL family, and that the adaptation to SUV39h loss could involve Polycomb repression.

To test that hypothesis, I very recently generated *Eed-Suv39h*-tKO ES cells, which survived in serum-based conditions, albeit growing at a reduced pace. In serum, I found that MERVL elements were unaltered in *Eed*-KO, upregulated by five-fold in *Suv39h*-dKO and further upregulated by 15-fold in *Eed-Suv39h*-tKO (**Figure 3.2A** and **B**). This shows without much doubt that Polycomb compensates for SUV39h absence in MERVL elements, even in globally hypermethylated ES cells, and supports our claim that MERVLs operate a switch from H3K9 to H3K27-based repression. However, IAP expression was unaffected by single or joint depletion of *Eed* and *Suv39h1/2* (**Figure 3.2 B**). This indicates that, in serum at least and contrary to MERVL, IAP elements do not replace Suv39h by Polycomb.

Upon conversion to 2i+vitamin C, I observed that *Eed-Suv39h*-tKO ES cells died after six-eight days: ES cell viability is therefore impaired when DNA methylation, Suv39h and PRC2 are depleted together, but not when one of these marks is still present. In absence of DNA methylation or SUV39h, H3K27me3 was shown to relocalize to pericentric heterochromatic (Cooper et al., 2014; Peters et al., 2003; Saksouk et al., 2014). It would be very informative to analyze whether *Eed-Suv39h*-tKO death correlates with massive pericentric defects, and also to follow transposon expression during the first day of conversion.

Overall, H3K9me3-related pathways are not affected by 2i+vitC-induced loss of DNA methylation. However, closer analysis reveals a more complex picture: ESET/KAP1-repression is reinforced, while SUV39h and Polycomb engage in an intricate and complex crosstalk.

3.1.6 A new function of H3K27me3 in repressing TEs

One of the main findings of my work is the observation that H3K27me3 relocalizes to TEs upon loss of DNA methylation. In the case of MERVLs at least, it is clear that Polycomb acts as a potent transcriptional repressor. In MERVL elements, H3K27me3 is localized in

promoter regions, which correlate well with its silencing effect. Other families, such as ETn and VL30, gain H3K27me3 in their 5'-region as well and we can expect that these families would similarly be reactivated. By contrast, a majority of TE families gain H3K27me3 in their 3' regions, while their 5' remained marked by H3K9me3. To investigate whether the presence of H3K27me3 in TE body could be functionally relevant and be involved in silencing, we are currently performing RNA-seq for *Eed*-KO ES cells grown in serum and 2i+vitamin C conditions.

The presence of Polycomb in TE bodies could affect transcriptional elongation. H3K27me3 usually recruit PRC1 through the binding of CBX proteins to H3K27me3 residues, and PRC1 subsequently catalyzes local H2Aub deposition. H2A ubiquitination and PRC1 are thought to block transcriptional initiation or elongation (Brookes et al., 2012; Stock et al., 2007). It is therefore tempting to speculate that PRC1 could play a similar role in TEs. Whether PRC1 and H2A ubiquitination are recruited to TEs upon loss of DNA methylation needs to be investigated.

3.1.7 H3K27me3 is attracted to unmethylated TEs

One of the main finding of my work is the observation that H3K27me3 relocalize to TEs upon loss of DNA methylation, raising the interesting question of how Polycomb proteins are targeted there. The question of the determinants of Polycomb recruitment remains largely unanswered in mammals. In ES cells grown is serum, the majority of H3K27me3-marked regions nucleate around CGIs and do not usually spread into DNA methylated territories (Brinkman et al., 2012; Ku et al., 2008; Statham et al., 2012). By contrast, upon loss of DNA methylation, H3K27me3 leaves CGIs and starts colonizing new genomic compartments, in particular TEs and pericentric repeats.

Recruitment of PRC2 to pericentric heterochromatin in absence of DNA methylation was shown to rely on the (putative) DNA binding protein BEND3, probably by recognizing specific motifs in major satellite repeats (Saksouk et al., 2014). It is therefore unlikely that BEND3 could be implicated in the targeting of PRC2 to TEs. Polycomb recruitment can also be mediated by long-non-coding RNAs (Pandey et al., 2008; Plath et al., 2003; Rinn et al., 2007), and it would be interesting to investigate whether PRC2 physically interacts with TE RNAs. Intriguingly, TE burst of transcription and H3K27me3 deposition are not synchronized: H3K27me3 reaches its full potential only after transcription has stopped. This raises interesting question about the mechanistic links between TE transcription and H3K27me3 deposition in TE bodies. Several hypotheses are possible, which would need to be

tested: *1)* H3K27me3 deposition blocks transcription; 2) H3K27me3 can only be laid down after unrelated mechanisms have suppressed transcription; *3)* Transcription recruits H3K27me3.

The observation that H3K27me3 invades TEs only upon loss of DNA methylation confirms the general assumption that these two modifications are mutually exclusive in ES cells. PRC1 complexes associated with RYBP were recently shown to be involved in PRC2 recruitment (Blackledge et al., 2014; Cooper et al., 2014). In particular, the CxxX-protein KDM2b associates with RYBP-PRC1 and is thought to mediate PRC1 and PRC2 recruitment at unmethylated CpG-rich regions, such as CGIs. Considering the role of CxxX-domains in recognizing unmethylated CpGs, KDM2b represents therefore a very seducing model that could explain how H3K27me3 is recruited at unmethylated TEs. Indeed, most TEs and especially young and active ones tend to be CpG-rich, and KDM2b could mediate recruitment of PRC1, which in turn would recruit PRC2 and H3K27me3.

In fact, our system represents a very interesting model to study the mechanisms of Polycomb recruitment. Indeed, thousand of new genomic locations become enriched for H3K27me3 and could help unraveling how Polycomb machinery is recruited *de novo*. Unfortunately, the necessity to write this thesis prevented me so far to generate *Kdm2b* and *Rybp* knockout ES cells and I could not directly test whether depleting KDM2b and RYBP-PRC1 would affect PRC2 recruitment at TE upon loss of DNA methylation.

From the literature, it appears that depletion of KDM2b in serum-grown ES cells barely affect H3K27me3 enrichment at CGIs, which indicates that KDM2b is not the unique player involved in Polycomb recruitment at CGIs (Blackledge et al., 2014). Interestingly, a study reported that in serum-grown ES cells, numerous germline genes such as *Dazl*, *Mov10l1*, *Mael* or *Ddx4* were bound by RYBP and RING1b and analysis of ChIP-seq data from another study confirms this observation (**Figure 3.3A** and Hisada et al., 2012; Morey et al., 2013). Interestingly, the majority of these germline genes (at the notable exception of *Dazl*) were shown to be bound by E2F6, a transcription factor member of one specific RYBP-PRC1 complex (Gao et al., 2012; Velasco et al., 2010), which could explain how PRC1 is recruited to these regions. These germline genes have methylated CGI promoters, suggesting that RYBP-PRC1 binding might not be affected by DNA methylation. Interestingly, RYBP-PRC1 presence at germline genes was not associated with PRC2 and CBX-PRC1 (**Figure 3.3A**), which is expected considering their high level of DNA methylation. However, we observed that upon loss of DNA methylation, some of these germline genes such as *Dazl* or *Sycp3* are gaining H3K27me3, raising the interesting possibility that when DNA methylation

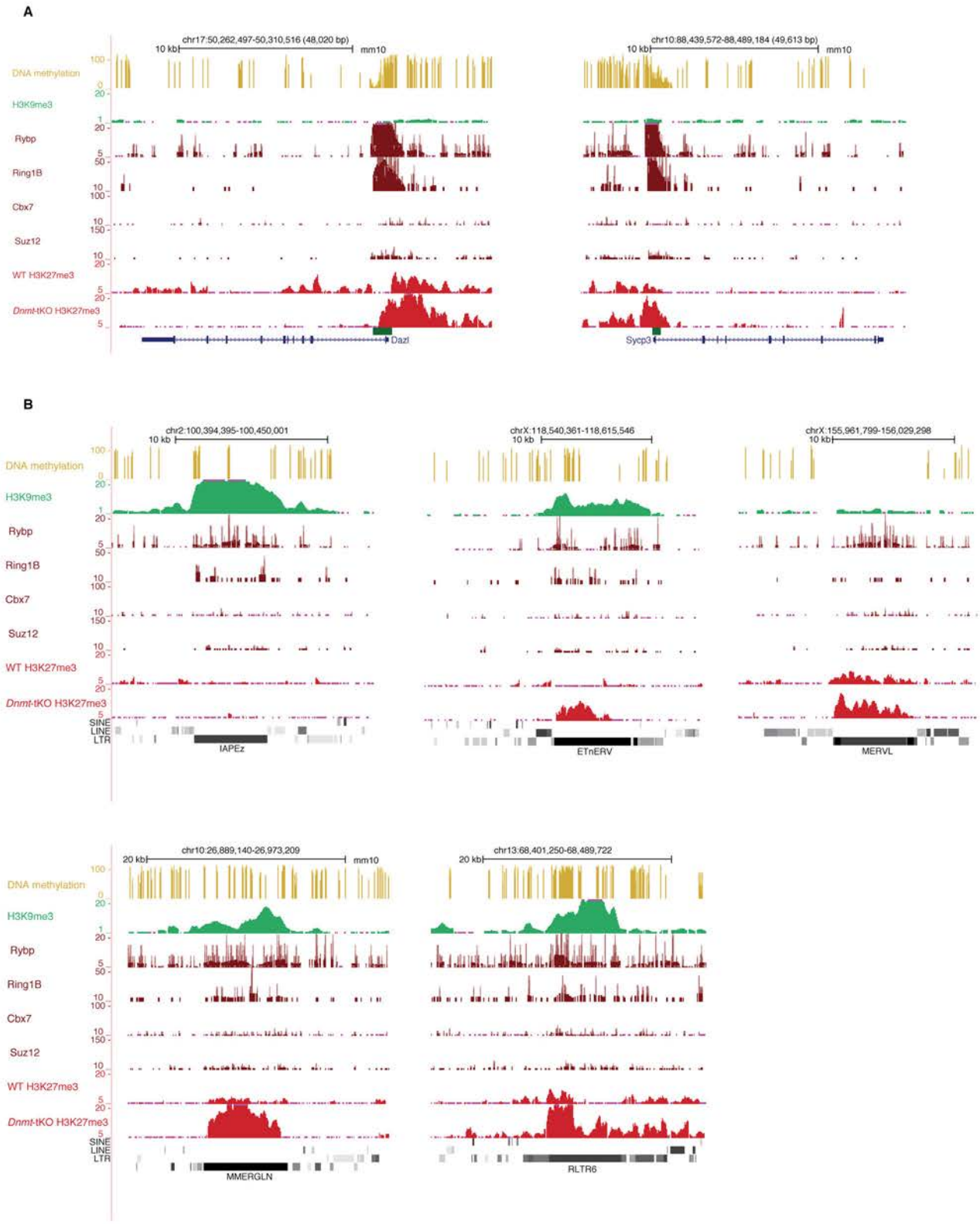


Figure 3.3. PRC1 presence at transposon in serum grown ES cells could promote the recruitment of H3K27me3 upon loss of DNA methylation.

ChIP-seq (and DNA methylation) data for ES cells grown in serum and hypermethylated (except for the *Dnmt*-tKO ES cells). DNA methylation and H3K9me3 are from this study, Rybp, Ring1b, Cbx7 and Suz12 are from Morey et al., 2013. WT and Dnmt-tKO H3K27me3 data are from Cooper et al., 2014.

A. Rybp-PRC1, but not Cbx-PRC1 or PRC2, is present in the DNA hypermethylated promoter of the germline gene *Dazl* and *Sycp3*. Both these promoter gain H3K27me3 upon loss of DNA methylation, potentially recruited by Rybp-PRC1.

B. Various ERVs of class I (MMERGLN, RLTR6), II (IAP, ETn) and III (MERVL) are modestly enriched for Rybp-PRC1 in hypermethylated ES cells and gain H3K27me3 upon loss of DNA methylation, except in regions enriched for H3K9me3.

disappears, RYBP-PRC1 could trigger the recruitment of PRC2 at these unmethylated targets (**Figure 3.3A**).

Remarkably, RYBP was also shown to be present at the LTR of TEs such as IAP, MERVL or MusD, albeit at comparatively lower levels for the latter,. Moreover, MERVL elements are activated in *Rybp*-KO cells (Hisada et al., 2012). The presence of RYBP in hypermethylated TEs could serve to prepare for the future recruitment of PRC2 and CBX-PRC1 upon loss of DNA methylation. Inspection of ChIP-seq tracks from available datasets show that several TEs appear indeed bound by RYBP in serum cells, which we found to,gain H3K27me3 upon loss of DNA methylation, precisely around RYBP binding site (**Figure 3.3B**). This observation is very preliminary and would necessitate thorough investigation. Nonetheless, it can be speculated that **Rypb**, probably in partner with PRC1 complexes and DNA binding protein, could act as the seed already in hypermethylated contexts that promote the recruitment of PRC2 upon loss of DNA methylation.

3.1.8 H3K27 and H3K9 methylation occupy different TE territories

This work is focused on TE regulation, but our detailed analysis of repressive chromatin reorganization has a broader impact and shed particular light on how H3K9 and H327 repressive pathway interact. In particular both repressive marks were apparently enriched in the same TE sequences, either in LINEs or class I and II ERVs, but were nonetheless occupying non-overlapping territories, a pattern that was never observed before. H3K27 and H3K9 methylation are mutually exclusive in almost all cell types (Hawkins et al., 2010; Mikkelsen et al., 2007). However, since H3K27me3 antagonizes DNA methylation whereas H3K9me3 positively associates with, the mutual exclusion of the two marks could be simply explained by the presence DNA methylation in most contexts.

Since the two marks do not overlap even in 2i+vitamin C medium, our results suggest that mutual exclusion of H3K9me3 and H3K27me3 is independent of DNA methylation. In numerous transposon sequences, H3K27 and H3K9 methylation are in very close physical proximity, probably only a few nucleosomes away from each others. Despite the natural spreading propensity of these two repressive marks, H3K9me3 and H3K27me3 remain physically separated, which might suggest that the two repressive compartments are actively repelling each others. Similarly, in chromocenters that also contain both marks upon loss of DNA methylation, H3K9me3 and H3K27me3 occupy in fact distinct territories (Cooper et al., 2014). The two marks appear antagonistic but what mechanistically cause this exclusion is far for being understood. Binding or catalytic activity of PRC2 and H3K9-KMT is not

affected by any of the two chromatin modifications (Bartke et al., 2010; Schmitges et al., 2011). Antagonism between the marks is therefore probably indirect and it was proposed that HP1 could prevent the recruitment of PRC1 and PRC2 (Tardat et al., 2015).

Our analysis offers resolution at H3K9me3 and H3K27me3 separation, showing that they can be immediately adjacent. However, TEs arre repeated sequences and this is therefore not possible to completely exclude the possibility that this observation is a mapping artifact. One way to circumvent this issue would be to create artificially an H3K9me3-H3K27me3 bivalent domain, for example by inserting KAP1-binding sequences inside H3K27me3-marked CGI or into an *Hox* cluster. Such a system could be very instrumental to understand mechanistically what triggers the mutual exclusivity of the two marks.

Overall, the fine analysis of repressive chromatin reorganization at TEs gives very fundamental insight into the intricate relationship that exists between chromatin modifications. From a regulatory point of view, the mutual exclusivity of H3K27me3 and H3K9me3 can probably be explained by the fundamentally different biological roles for these two repressive pathways. H3K9 methylation is mostly associated with permanent silencing of the repeated fraction of the genome, whereas Polycomb is highly plastic and involved in genic regulation. Structurally separating the two repressive domains may ensure a tight control of genome regulation, for example ensuring that TEs bound by Kap1 and H3K9me3 do not perturb the regulatory capacity of neighboring Polycomb domains.

REFERENCES

Aagaard, L., Laible, G., Selenko, P., Schmid, M., Dorn, R., Schotta, G., Kuhfittig, S., Wolf, A., Lebersorger, A., Singh, P.B., et al. (1999). Functional mammalian homologues of the Drosophila PEV-modifier Su(var)3-9 encode centromere-associated proteins which complex with the heterochromatin component M31. EMBO J. *18*, 1923–1938.

Aapola, U., Kawasaki, K., Scott, H.S., Ollila, J., Vihinen, M., Heino, M., Shintani, A., Minoshima, S., Krohn, K., Antonarakis, S.E., et al. (2000). Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. Genomics *65*, 293–298.

Abad, P., Vaury, C., Pélisson, A., Chaboissier, M.C., Busseau, I., and Bucheton, A. (1989). A long interspersed repetitive element--the I factor of Drosophila teissieri--is able to transpose in different Drosophila species. Proc. Natl. Acad. Sci. U. S. A. *86*, 8887–8891.

Adey, N.B., Schichman, S.A., Graham, D.K., Peterson, S.N., Edgell, M.H., and Hutchison, C.A. (1994). Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. Mol. Biol. Evol. *11*, 778–789.

Akalin, A., Kormaksson, M., Li, S., Garrett-Bakelman, F.E., Figueroa, M.E., Melnick, A., and Mason, C.E. (2012). methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol. *13*, R87.

Alisch, R.S., Garcia-Perez, J.L., Muotri, A.R., Gage, F.H., and Moran, J. V (2006). Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev. *20*, 210–224.

Alkema, M.J., Bronk, M., Verhoeven, E., Otte, A., van 't Veer, L.J., Berns, A., and van Lohuizen, M. (1997). Identification of Bmi1-interacting proteins as constituents of a multimeric mammalian polycomb complex. Genes Dev. *11*, 226–240.

Allis, C.D., Berger, S.L., Cote, J., Dent, S., Jenuwien, T., Kouzarides, T., Pillus, L., Reinberg, D., Shi, Y., Shiekhattar, R., et al. (2007). New nomenclature for chromatin-modifying enzymes. Cell *131*, 633–636.

Aloia, L., Di Stefano, B., and Di Croce, L. (2013). Polycomb complexes in stem cells and embryonic development. Development *140*, 2525–2534.

Anders, S., Pyl, P.T., and Huber, W. (2014). HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Anderson, R., Copeland, T.K., Schöler, H., Heasman, J., and Wylie, C. (2000). The onset of germ cell migration in the mouse embryo. Mech. Dev. *91*, 61–68.

Anxolabehere, D., Kidwell, M., and Periquet, G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. Mol. Biol. Evol. *5*, 252–269.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature *408*, 796–815.

Arand, J., Spieler, D., Karius, T., Branco, M.R., Meilinger, D., Meissner, A., Jenuwein, T., Xu, G., Leonhardt, H., Wolf, V., et al. (2012). In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases. PLoS Genet. *8*, e1002750.

Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744–747.

Aravin, A.A., Bourc'his, D., and Bourc'his, D. (2008). Small RNA guides for de novo DNA methylation in mammalian germ cells. Genes Dev. *22*, 970–975.

Arita, K., Isogai, S., Oda, T., Unoki, M., Sugita, K., Sekiyama, N., Kuwata, K., Hamamoto, R., Tochio, H., Sato, M., et al. (2012). Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. Proc. Natl. Acad. Sci. U. S. A. *109*, 12950–12955.

Arkhipova, I., and Meselson, M. (2000). Transposable elements in sexual and ancient asexual taxa. Proc. Natl. Acad. Sci. U. S. A. *97*, 14473–14477.

Arrigoni, R., Alam, S.L., Wamstad, J.A., Bardwell, V.J., Sundquist, W.I., and Schreiber-Agus, N. (2006). The Polycomb-associated protein Rybp is a ubiquitin binding protein. FEBS Lett. *580*, 6233–6241.

Athanikar, J.N., Badge, R.M., and Moran, J. V (2004). A YY1-binding site is required for accurate human LINE-1 transcription initiation. Nucleic Acids Res. *32*, 3846–3855.

Auclair, G., Guibert, S., Bender, A., and Weber, M. (2014). Ontogeny of CpG island methylation and specificity of DNMT3 methyltransferases during embryonic development in the mouse. Genome Biol. *15*, 545.

Aucott, R., Bullwinkel, J., Yu, Y., Shi, W., Billur, M., Brown, J.P., Menzel, U., Kioussis, D., Wang, G., Reisert, I., et al. (2008). HP1-beta is required for development of the cerebral neocortex and neuromuscular junctions. J. Cell Biol. *183*, 597–606.

Avery, O.T., Macleod, C.M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. J. Exp. Med. *79*, 137–158.

Babushok, D. V, Ostertag, E.M., Courtney, C.E., Choi, J.M., and Kazazian, H.H. (2006). L1 integration in a transgenic mouse model. Genome Res. *16*, 240–250.

Babushok, D. V, Kazazian, H.H., and Jr, Ã. (2007). Progress in understanding the biology of the human mutagen LINE-1. Hum. Mutat. *28*, 527–539.

Bailey, J.A., Liu, G., and Eichler, E.E. (2003). An Alu transposition model for the origin and expansion of human segmental duplications. Am. J. Hum. Genet. *73*, 823–834.

Baillie, G.J., van de Lagemaat, L.N., Baust, C., and Mager, D.L. (2004). Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. J. Virol. *78*, 5784–5798.

Bailly-Bechet, M., Haudry, A., and Lerat, E. (2014). "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mob. DNA *5*, 13.

Bannister, A.J., Zegerman, P., Partridge, J.F., Miska, E.A., Thomas, J.O., Allshire, R.C., and Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. Nature *410*, 120–124.

Barau, J. (2015). SINEs tether the Synaptonemal Complex to gene-rich regions to prevent harmful recombination. In Proceedings of Late and Unproven Labnight Hypothesis in the Bourc'his Lab, pp. 10–15.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell *143*, 470–484.

Bartolome, C., Maside, X., and Charlesworth, B. (2002). On the Abundance and Distribution of Transposable Elements in the Genome of Drosophila melanogaster. Mol. Biol. Evol. *19*, 926–937.

Baubec, T., Ivánek, R., Lienert, F., and Schübeler, D. (2013). Methylation-dependent and -independent genomic targeting principles of the MBD protein family. Cell *153*, 480–492.

Baubec, T., Colombo, D.F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A.R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature *520*, 243–247.

Baust, C., Gagnier, L., Baillie, G.J., Harris, M.J., Juriloff, D.M., and Mager, D.L. (2003). Structure and Expression of Mobile ETnII Retroelements and Their Coding-Competent MusD Relatives in the Mouse. J. Virol. *77*, 11448–11458.

Beall, E.L., and Rio, D.C. (1996). Drosophila IRBP/Ku p70 corresponds to the mutagen-sensitive mus309 gene and is involved in P-element excision in vivo. Genes Dev. *10*, 921–933.

Beall, E.L., Admon, A., and Rio, D.C. (1994). A Drosophila protein homologous to the human p70 Ku autoimmune antigen interacts with the P transposable element inverted repeats. Proc. Natl. Acad. Sci. U. S. A. *91*, 12681–12685.

Beard, C., Li, E., and Jaenisch, R. (1995). Loss of methylation activates Xist in somatic but not in embryonic cells. Genes Dev. *9*, 2325–2334.

Becker, K.G., Swergold, G.D., Ozato, K., and Thayer, R.E. (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. Hum. Mol. Genet. *2*, 1697–1702.

Bellefroid, E.J., Poncelet, D.A., Lecocq, P.J., Revelant, O., and Martial, J.A. (1991). The evolutionarily conserved Krüppel-associated box domain defines a subfamily of eukaryotic multifingered proteins. Proc. Natl. Acad. Sci. U. S. A. *88*, 3608–3612.

Van Bemmel, J. (2012). Chromatin Flavors: Chromatin composition and domain organization in Drosophila melanogaster. In Thesis Dissertation, (Erasmus University Rotterdam), pp. 11–23.

Bénit, L., De Parseval, N., Casella, J.F., Callebaut, I., Cordonnier, A., and Heidmann, T. (1997). Cloning of a new murine endogenous retrovirus, MuERV-L, with strong similarity to the human HERV-L element and with a gag coding sequence closely related to the Fv1 restriction gene. J. Virol. *71*, 5652–5657.

Bénit, L., Lallemand, J.-B.B., Casella, J.-F.F., Philippe, H., Heidmann, T., Benit, L., Lallemand, J.-B.B., Casella, J.-F.F., Philippe, H., and Heidmann, T. (1999). ERV-L Elements: a Family of Endogenous Retrovirus-Like Elements Active throughout the Evolution of Mammals. J. Virol. *73*, 3301–3308.

Bennet, M., and Leicht, I. (2012). Plant DNA C-values database (release 6.0, Dec. 2012).

Bennett, M.D. (1971). The Duration of Meiosis. Proc. R. Soc. London. Ser. B, Biol. Sci. *178*, 277–299.

Bennett, M.D. (1972). Nuclear DNA Content and Minimum Generation Time in Herbaceous Plants. Proc. R. Soc. B Biol. Sci. *181*, 109–135.

Bensasson, D., Petrov, D.A., Zhang, D.-X.X., Hartl, D.L., and Hewitt, G.M. (2001). Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol. Biol. Evol. *18*, 246–253.

Berndsen, C.E., and Wolberger, C. (2014). New insights into ubiquitin E3 ligase mechanism. Nat. Struct. Mol. Biol. *21*, 301–307.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006a). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., Snyder, M., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bernstein, E., Duncan, E.M., Masui, O., Gil, J., Heard, E., and Allis, C.D. (2006b). Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. Mol. Cell. Biol. *26*, 2560–2569.

Bestor, T.H., and Ingram, V.M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. Proc. Natl. Acad. Sci. U. S. A. *80*, 5559–5563.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. J. Mol. Biol. *203*, 971–983.

Bian, C., Chen, Q., and Yu, X. (2015). The zinc finger proteins ZNF644 and WIZ regulate the G9a/GLP complex for gene repression | eLife Lens.

Bickmore, W.A., and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Biemont, C. (2010). A Brief History of the Status of Transposable Elements: From Junk DNA to Major Players in Evolution. Genetics *186*, 1085–1093.

Biémont, C. (1994). Dynamic equilibrium between insertion and excision of P elements in highly inbred lines from an M' strain of Drosophila melanogaster. J. Mol. Evol. *39*, 466–472.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8*, 1499–1504.

Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature *321*, 209–213.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91–99.

Biterge, B., and Schneider, R. (2014). Histone variants: key players of chromatin. Cell Tissue Res. *356*, 457–466.

Blackledge, N.P., Zhou, J.C., Tolstorukov, M.Y., Farcas, A.M., Park, P.J., and Klose, R.J. (2010). CpG islands recruit a histone H3 lysine 36 demethylase. Mol. Cell *38*, 179–190.

Blackledge, N.P., Farcas, A.M., Kondo, T., King, H.W., McGouran, J.F., Hanssen, L.L.P., Ito, S., Cooper, S., Kondo, K., Koseki, Y., et al. (2014). Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. Cell *157*, 1445–1459.

Blaschke, K., Ebata, K.T., Karimi, M.M., Zepeda-Martínez, J. a, Goyal, P., Mahapatra, S., Tam, A., Laird, D.J., Hirst, M., Rao, A., et al. (2013). Vitamin C induces Tet-dependent DNA demethylation and a blastocyst-like state in ES cells. Nature *500*, 222–226.

Boeke, J.D., and Stoye, J.P. (1997). Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements.

Boissinot, S., and Furano, A. V. (2001). Adaptive Evolution in LINE-1 Retrotransposons. Mol. Biol. Evol. *18*, 2186–2194.

Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T., and Weber, M. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. Nat. Genet. *42*, 1093–1100.

Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science *317*, 1760–1764.

Boulard, M., Edwards, J.R., and Bestor, T.H. (2015). FBXL10 protects Polycomb-bound genes from hypermethylation. Nat. Genet. *47*, 479–485.

Bourc'his, D., and Bestor, T.H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature *431*, 96–99.

Bourc'his, D., Xu, G.L., Lin, C.S., Bollman, B., and Bestor, T.H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536–2539.

Bowen, N.J., and McDonald, J.F. (2001). Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res. *11*, 1527–1540.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349–353.

Brady, T., Lee, Y.N., Ronen, K., Malani, N., Berry, C.C., Bieniasz, P.D., and Bushman, F.D. (2009). Integration target site selection by a resurrected human endogenous retrovirus. Genes Dev. *23*, 633–642.

Brind'Amour, J., Liu, S., Hudson, M., Chen, C., Karimi, M.M., and Lorincz, M.C. (2015). An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nat. Commun. *6*, 6033.

Brinkman, A.B., Gu, H., Bartels, S.J.J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., et al. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res. *22*, 1128–1138.

Bromham, L., Clark, F., and McKee, J.J. (2001). Discovery of a novel murine type C retrovirus by data mining. J. Virol. *75*, 3053–3057.

Brons, I.G.M., Smithers, L.E., Trotter, M.W.B., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S.M., Howlett, S.K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R.A., et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. Nature *448*, 191–195.

Brookes, E., de Santiago, I., Hebenstreit, D., Morris, K.J., Carroll, T., Xie, S.Q., Stock, J.K., Heidemann, M., Eick, D., Nozaki, N., et al. (2012). Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs. Cell Stem Cell *10*, 157–170.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J. V, and Kazazian, H.H. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. U. S. A. *100*, 5280–5285.

Brown, S.W. (1966). Heterochromatin. Science *151*, 417–425.

Brown, J.P., Bullwinkel, J., Baron-Lühr, B., Billur, M., Schneider, P., Winking, H., and Singh, P.B. (2010). HP1gamma function is required for male germ cell survival and spermatogenesis. Epigenetics Chromatin *3*, 9.

Brulet, P., Kaghad, M., Xu, Y.S., Croissant, O., and Jacob, F. (1983). Early differential tissue expression of transposon-like repetitive DNA sequences of the mouse. Proc. Natl. Acad. Sci. *80*, 5641–5645.

Buchenau, P. (1998). The Distribution of Polycomb-Group Proteins During Cell Division and Development in Drosophila Embryos: Impact on Models for Silencing. J. Cell Biol. *141*, 469–481.

Bulut-Karslioglu, A., DeLaRosa-Velázquez, I. a., Ramirez, F., Barenboim, M., Onishi-Seebacher, M., Arand, J., Galán, C., Winter, G.E.E., Engist, B., Gerle, B., et al. (2014). Suv39h-Dependent H3K9me3 Marks Intact Retrotransposons and Silences LINE Elements in Mouse Embryonic Stem Cells. Mol. Cell *55*, 277–290.

Burke, W.D., Malik, H.S., Lathe, W.C., and Eickbush, T.H. (1998). Are retrotransposons long-term hitchhikers? Nature *392*, 141–142.

Van den Bussche, R.A., Longmire, J.L., and Baker, R.J. (1995). How bats achieve a small C-value: frequency of repetitive DNA in Macrotus. Mamm. Genome *6*, 521–525.

Calisher, C.H., Childs, J.E., Field, H.E., Holmes, K. V, and Schountz, T. (2006). Bats: important reservoir hosts of emerging viruses. Clin. Microbiol. Rev. *19*, 531–545.

Callinan, P.A., and Batzer, M.A. (2006). Retrotransposable elements and human disease. Genome Dyn. *1*, 104–115.

Cammas, F., Mark, M., Dolle, P., Dierich, A., Chambon, P., and Losson, R. (2000). Mice lacking the transcriptional corepressor TIF1beta are defective in early postimplantation development. Development *127*, 2955–2963.

Cannon, W.B. (1926). Physiological regulation of normal states: some tentative postulates concerning biological homeostatics. Ses Amis, Ses Coll. Ses Elev.

Canzio, D., Chang, E.Y., Shankar, S., Kuchenbecker, K.M., Simon, M.D., Madhani, H.D., Narlikar, G.J., and Al-Sady, B. (2011). Chromodomain-mediated oligomerization of HP1 suggests a nucleosome-bridging mechanism for heterochromatin assembly. Mol. Cell *41*, 67–81.

Cao, R., and Zhang, Y. (2004). SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. Mol. Cell *15*, 57–67.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in Polycomb-group silencing. Science *298*, 1039–1043.

Capy, P., Vitalis, R., Langin, T., Higuet, D., and Bazin, C. (1996). Relationships between transposable elements based upon the integrase-transposase domains: Is there a common ancestor? J. Mol. Evol. *42*, 359–368.

Capy, P., Langin, T., Higuet, D., Maurer, P., and Bazin, C. (1997). Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica *100*, 63–72.

Capy, P., Gasperi, G., Biémont, C., and Bazin, C. (2000). Stress and transposable elements: co-evolution or useful parasites? Heredity (Edinb). *85*, 101–106.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.-J., Anderson, S., Yates, J., Washburn, M.P., et al. (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581–592.

Castro-Diaz, N., Ecco, G., Coluccio, A., Kapopoulou, A., Yazdanpanah, B., Friedli, M., Duc, J., Jang, S.M., Turelli, P., and Trono, D. (2014). Evolutionally dynamic L1 regulation in embryonic stem cells. Genes Dev. *28*, 1397–1409.

Cavalier-Smith, T. (1982). Skeletal DNA and the evolution of genome size. Annu. Rev. Biophys. Bioeng. *11*, 273–302.

Cerveau, N., Leclercq, S., Bouchon, D., and Cordaux, R. (2011). Evolutionary Dynamics and Genomic Impact of Prokaryote Transposable Elements. In Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution SE - 17, P. Pontarotti, ed. (Springer Berlin Heidelberg), pp. 291–312.

Charlesworth, B., and Charlesworth, D. (1983). The population dynamics of transposable elements. Genet. Res. (Camb). *42*, 1–27.
Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature *371*, 215–220.
Chédin, F. (2011). The DNMT3 family of mammalian de novo DNA methyltransferases (Elsevier Inc.).
Chen, C., Ara, T., and Gautheret, D. (2009). Using Alu elements as polyadenylation sites: A case of retroposon exaptation. Mol. Biol. Evol. *26*, 327–334.
Chen, J., Liu, H., Liu, J., Qi, J., Wei, B., Yang, J., Liang, H., Chen, Y., Chen, J., Wu, Y., et al. (2013). H3K9 methylation is a barrier during somatic cell reprogramming into iPSCs. Nat. Genet. *45*, 34–42.
Chen, T., Tsujimoto, N., and Li, E. (2004). The PWWP domain of Dnmt3a and Dnmt3b is required for directing DNA methylation to the major satellite repeats at pericentric heterochromatin. Mol. Cell. Biol. *24*, 9048–9058.
Chen, Z.-X., Mann, J.R., Hsieh, C.-L., Riggs, A.D., and Chédin, F. (2005). Physical and functional interactions between the human DNMT3L protein and members of the de novo methyltransferase family. J. Cell. Biochem. *95*, 902–917.
Chénais, B. (2013). Transposable elements and human cancer: a causal relationship? Biochim. Biophys. Acta *1835*, 28–35.
Cheutin, T., McNairn, A.J., Jenuwein, T., Gilbert, D.M., Singh, P.B., and Misteli, T. (2003). Maintenance of stable heterochromatin domains by dynamic HP1 binding. Science *299*, 721–725.
Chin, H.G., Estève, P.-O., Pradhan, M., Benner, J., Patnaik, D., Carey, M.F., and Pradhan, S. (2007). Automethylation of G9a and its implication in wider substrate specificity and HP1 binding. Nucleic Acids Res. *35*, 7313–7323.
Chuang, L.S. (1997). Human DNA-(Cytosine-5) Methyltransferase-PCNA Complex as a Target for p21WAF1. Science (80-.). *277*, 1996–2000.
Ciferri, C., Lander, G.C., Maiolica, A., Herzog, F., Aebersold, R., and Nogales, E. (2012). Molecular architecture of human polycomb repressive complex 2. Elife *1*, e00005.
Cohen, C.J., Lock, W.M., and Mager, D.L. (2009). Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene *448*, 105–114.
Collins, R.E., Northrop, J.P., Horton, J.R., Lee, D.Y., Zhang, X., Stallcup, M.R., and Cheng, X. (2008). The ankyrin repeats of G9a and GLP histone methyltransferases are mono- and dimethyllysine binding modules. Nat. Struct. Mol. Biol. *15*, 245–250.
Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.
Cooper, S., Dienstbier, M., Hassan, R., Schermelleh, L., Sharif, J., Blackledge, N.P., De Marco, V., Elderkin, S., Koseki, H., Klose, R., et al. (2014). Targeting polycomb to pericentric heterochromatin in embryonic stem cells reveals a role for H2AK119u1 in PRC2 recruitment. Cell Rep. *7*, 1456–1470.
Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat Rev Genet *10*, 691–703.
Cordaux, R., Hedges, D.J., Herke, S.W., and Batzer, M.A. (2006). Estimating the retrotransposition rate of human Alu elements. Gene *373*, 134–137.
Cordonnier, A., Casella, J.F., and Heidmann, T. (1995). Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. J. Virol. *69*, 5890–5897.
Corradi, N., Pombert, J.-F., Farinelli, L., Didier, E.S., and Keeling, P.J. (2010). The complete sequence of the smallest known nuclear genome from the microsporidian Encephalitozoon intestinalis. Nat. Commun. *1*, 77.
Cortijo, S., Wardenaar, R., Colomé-Tatché, M., Gilly, A., Etcheverry, M., Labadie, K., Caillieux, E., Hospital, F., Aury, J.-M., Wincker, P., et al. (2014). Mapping the epigenetic basis of complex traits. Science *343*, 1145–1148.
Cost, G.J., Feng, Q., Jacquier, A., and Boeke, J.D. (2002). Human L1 element target-primed reverse transcription in vitro. EMBO J. *21*, 5899–5910.
Craig, N.L., Craigie, R., Gellert, M., and Lambowltz, A. (2002). Mobile DNA II (ASM Pres).
Crick, F. (1956). Ideas on Protein Synthesis.
Cui, K., Zang, C., Roh, T.-Y., Schones, D.E., Childs, R.W., Peng, W., and Zhao, K. (2009). Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell *4*, 80–93.
Cui, P., Liu, W., Zhao, Y., Lin, Q., Zhang, D., Ding, F., Xin, C., Zhang, Z., Song, S., Sun, F., et al. (2012). Comparative analyses of H3K4 and H3K27 trimethylations between the mouse cerebrum and testis. Genomics. Proteomics Bioinformatics *10*, 82–93.
Dai, Q., Ren, A., Westholm, J.O., Serganov, A.A., Patel, D.J., and Lai, E.C. (2013). The BEN domain is a novel sequence-specific DNA-binding domain conserved in neural transcriptional repressors. Genes Dev. *27*, 602–614.

Dalton, A.J., Potter, M., and Merwin, R.M. (1961). Some ultrastructural characteristics of a series of primary and transplanted plasma-cell tumors of the mouse. J. Natl. Cancer Inst. *26*, 1221–1267.

Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G., and Chovnick, A. (1990). Evidence for horizontal transmission of the P transposable element between Drosophila species. Genetics *124*, 339–355.

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (J. Murray).

DeBerardinis, R.J., and Kazazian, H.H. (1999). Analysis of the promoter from an expanding mouse retrotransposon subfamily. Genomics *56*, 317–323.

Deininger, P.L., and Batzer, M.A. (1999). Alu Repeats and Human Disease. Mol. Genet. Metab. *67*, 183–193.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

Devos, K.M., Brown, J.K.M., and Bennetzen, J.L. (2002). Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. *12*, 1075–1079.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. *35*, 41–48.

Dhayalan, A., Rajavelu, A., Rathert, P., Tamas, R., Jurkowska, R.Z., Ragozin, S., and Jeltsch, A. (2010). The Dnmt3a PWWP domain reads histone 3 lysine 36 trimethylation and guides DNA methylation. J. Biol. Chem. *285*, 26114–26120.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dodge, J.E., Kang, Y.-K., Beppu, H., Lei, H., and Li, E. (2004). Histone H3-K9 Methyltransferase ESET Is Essential for Early Development. Mol. Cell. Biol. *24*, 2478–2486.

Dong, K.B., Maksakova, I.A., Mohn, F., Leung, D., Appanah, R., Lee, S., Yang, H.W., Lam, L.L., Mager, D.L., Sch|[uuml]|beler, D., et al. (2008). DNA methylation in ES cells requires the lysine methyltransferase G9a but not its catalytic activity. EMBO J. *27*, 2691–2701.

Donnelly, S.R., Hawkins, T.E., and Moss, S.E. (1999). A conserved nuclear element with a role in mammalian gene regulation. Hum. Mol. Genet. *8*, 1723–1728.

Doolittle, W.F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. Nature *284*, 601–603.

Dorigo, B., Schalch, T., Bystricky, K., and Richmond, T.J. (2003). Chromatin fiber folding: requirement for the histone H4 N-terminal tail. J. Mol. Biol. *327*, 85–96.

Drake, J.W., Charlesworth, B., Charlesworth, D., and Crow, J.F. (1998). Rates of Spontaneous Mutation. Genetics *148*, 1667–1686.

Druker, R., and Whitelaw, E. (2004). Retrotransposon-derived elements in the mammalian genome: a potential source of disease. J. Inherit. Metab. Dis. *27*, 319–330.

Du, J., Johnson, L.M., Jacobsen, S.E., and Patel, D.J. (2015). DNA methylation pathways and their crosstalk with histone methylation. Nat. Rev. Mol. Cell Biol. *16*, 519–532.

Dupressoir, A., Marceau, G., Vernochet, C., Bénit, L., Kanellopoulos, C., Sapin, V., and Heidmann, T. (2005). Syncytin-A and syncytin-B, two fusogenic placenta-specific murine envelope genes of retroviral origin conserved in Muridae. Proc. Natl. Acad. Sci. U. S. A. *102*, 725–730.

Eickbush, T.H. (1997). Telomerase and retrotransposons: which came first? Science *277*, 911–912.

Elsässer, S.J., Noh, K.-M., Diaz, N., Allis, C.D., and Banaszynski, L.A. (2015). Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells. Nature *522*, 240–244.

Emerson, R.O., and Thomas, J.H. (2009). Adaptive evolution in zinc finger transcription factors. PLoS Genet. *5*, e1000325.

Endoh, M., Endo, T.A., Endoh, T., Fujimura, Y., Ohara, O., Toyoda, T., Otte, A.P., Okano, M., Brockdorff, N., Vidal, M., et al. (2008). Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. Development *135*, 1513–1524.

Endoh, M., Endo, T.A., Endoh, T., Isono, K., Sharif, J., Ohara, O., Toyoda, T., Ito, T., Eskeland, R., Bickmore, W.A., et al. (2012). Histone H2A mono-ubiquitination is a crucial step to mediate PRC1-dependent repression of developmental genes to maintain ES cell identity. PLoS Genet. *8*, e1002774.

Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., Deplus, R., Fuks, F.F., Shinkai, Y., Cedar, H., et al. (2008). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nat. Struct. Mol. Biol. *15*, 1176–1183.

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. *24*, 363–367.

Estève, P.-O., Chin, H.G., Smallwood, A., Feehery, G.R., Gangisetty, O., Karpf, A.R., Carey, M.F., and Pradhan, S. (2006). Direct interaction between DNMT1 and G9a coordinates DNA and histone methylation during replication. Genes Dev. *20*, 3089–3103.

Eustermann, S., Yang, J.-C., Law, M.J., Amos, R., Chapman, L.M., Jelinska, C., Garrick, D., Clynes, D., Gibbons, R.J., Rhodes, D., et al. (2011). Combinatorial readout of histone H3 modifications specifies localization of ATRX to heterochromatin. Nat. Struct. Mol. Biol. *18*, 777–782.

Evans, M.J., and Kaufman, M.H. (1981). Establishment in culture of pluripotential cells from mouse embryos. Nature *292*, 154–156.

Evgen'ev, M.B., and Arkhipova, I.R. (2005). Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet. Genome Res. *110*, 510–521.

Ewen, K.A., and Koopman, P. (2010). Mouse germ cell development: from specification to sex determination. Mol. Cell. Endocrinol. *323*, 76–93.

Fadloun, A., Le Gras, S., Jost, B., Ziegler-Birling, C., Takahashi, H., Gorab, E., Carninci, P., and Torres-Padilla, M.-E. (2013). Chromatin signatures and retrotransposon profiling in mouse embryos reveal regulation of LINE-1 by RNA. Nat. Struct. Mol. Biol. *20*, 332–338.

Falandry, C., Fourel, G., Galy, V., Ristriani, T., Horard, B., Bensimon, E., Salles, G., Gilson, E., and Magdinier, F. (2010). CLLD8/KMT1F is a lysine methyltransferase that is important for chromosome segregation. J. Biol. Chem. *285*, 20234–20241.

Fang, J., Feng, Q., Ketel, C.S., Wang, H., Cao, R., Xia, L., Erdjument-Bromage, H., Tempst, P., Simon, J.A., and Zhang, Y. (2002). Purification and Functional Characterization of SET8, a Nucleosomal Histone H4-Lysine 20-Specific Methyltransferase. Curr. Biol. *12*, 1086–1099.

Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R., Lee, S., Sims, D., Cerase, A., Sheahan, T.W., et al. (2012). KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. Elife *1*, e00205.

De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., Funaya, C., Antony, C., Moreira, P.N., Enright, A.J., and O'Carroll, D. (2011). The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature *480*, 259–263.

Fedoroff, N. V. (2012). Transposable Elements, Epigenetics, and Genome Evolution. Science (80-.). *338*, 758–767.

Feng, Q., Wang, H., Ng, H.H., Erdjument-Bromage, H., Tempst, P., Struhl, K., and Zhang, Y. (2002). Methylation of H3-Lysine 79 Is Mediated by a New Family of HMTases without a SET Domain. Curr. Biol. *12*, 1052–1058.

Feng, S., Cokus, S.J., Zhang, X., Chen, P.-Y., Bostick, M., Goll, M.G., Hetzel, J., Jain, J., Strauss, S.H., Halpern, M.E., et al. (2010). Conservation and divergence of methylation patterning in plants and animals. Proc. Natl. Acad. Sci. U. S. A. *107*, 8689–8694.

Feschotte, C. (2004). Merlin, a new superfamily of DNA transposons identified in diverse animal genomes and related to bacterial IS1016 insertion sequences. Mol. Biol. Evol. *21*, 1769–1780.

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. *41*, 331–368.

Feschotte, C., Jiang, N., and Wessler, S.R. (2002). Plant transposable elements: where genetics meets genomics. Nat. Rev. Genet. *3*, 329–341.

Ficz, G., Hore, T. a., Santos, F., Lee, H.J., Dean, W., Arand, J., Krueger, F., Oxley, D., Paul, Y.L., Walter, J., et al. (2013). FGF signaling inhibition in ESCs drives rapid genome-wide demethylation to the epigenetic ground state of pluripotency. Cell Stem Cell *13*, 351–359.

Filion, G.J., van Bemmel, J.G., Braunschweig, U., Talhout, W., Kind, J., Ward, L.D., Brugman, W., de Castro, I.J., Kerkhoven, R.M., Bussemaker, H.J., et al. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell *143*, 212–224.

Finch, J.T., and Klug, A. (1976). Solenoidal model for superstructure in chromatin. Proc. Natl. Acad. Sci. *73*, 1897–1901.

Finnegan, D. (1992). Transposable elements. In The Genome of Drosophila Melanogaster, D. Lindsley, and G. Zimm, eds. (Academic Press, San Diego), pp. 1096–1015.

Finnegan, D.J. (1989). Eukaryotic transposable elements and genome evolution. Trends Genet. *5*, 103–107.

Finnegan, J.E., and Dennis, E.S. (1993). Isolation and identification by sequence homology of a putative cytosine methyltransferase from Arabidopsis thaliana. Nucleic Acids Res. *21*, 2383–2388.

Firestein, R., Cui, X., Huie, P., and Cleary, M.L. (2000). Set Domain-Dependent Regulation of Transcriptional Silencing and Growth Control by SUV39H1, a Mammalian Ortholog of Drosophila Su(var)3-9. Mol. Cell. Biol. *20*, 4900–4909.

Fischle, W., Wang, Y., Jacobs, S.A., Kim, Y., Allis, C.D., and Khorasanizadeh, S. (2003). Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains. Genes Dev. *17*, 1870–1881.

Fleischmann, A., Michael, T.P., Rivadavia, F., Sousa, A., Wang, W., Temsch, E.M., Greilhuber, J., Müller, K.F., and Heubl, G. (2014). Evolution of genome size and chromosome number in the carnivorous plant genus Genlisea (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. Ann. Bot. *114*, 1651–1663.

Flemming, W. (1882). Zellsubstanz, Kern und Zelltheilung (Leiptiz: F.C.W Vogel).
Francis, N.J., Kingston, R.E., and Woodcock, C.L. (2004). Chromatin compaction by a polycomb group protein complex. Science *306*, 1574–1577.
Frauer, C., Hoffmann, T., Bultmann, S., Casa, V., Cardoso, M.C., Antes, I., and Leonhardt, H. (2011). Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain. PLoS One *6*, e21306.
French, N.S., and Norton, J.D. (1997). Structure and functional properties of mouse VL30 retrotransposons. Biochim. Biophys. Acta - Gene Struct. Expr. *1352*, 33–47.
Friedman, J.R., Fredericks, W.J., Jensen, D.E., Speicher, D.W., Huang, X.P., Neilson, E.G., and Rauscher, F.J. (1996). KAP-1, a novel corepressor for the highly conserved KRAB repression domain. Genes Dev. *10*, 2067–2078.
Friend, C. (1957). CELL-FREE TRANSMISSION IN ADULT SWISS MICE OF A DISEASE HAVING THE CHARACTER OF A LEUKEMIA. J. Exp. Med. *105*, 307–318.
Friesen, P.D., and Nissen, M.S. (1990). Gene organization and transcription of TED, a lepidopteran retrotransposon integrated within the baculovirus genome. Mol. Cell. Biol. *10*, 3067–3077.
Fritsch, L., Robin, P., Mathieu, J.R.R., Souidi, M., Hinaux, H., Rougeulle, C., Harel-Bellan, A., Ameyar-Zazoua, M., and Ait-Si-Ali, S. (2010). A Subset of the Histone H3 Lysine 9 Methyltransferases Suv39h1, G9a, GLP, and SETDB1 Participate in a Multimeric Complex. Mol. Cell *37*, 46–56.
Fuks, F. (2003). The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. Nucleic Acids Res. *31*, 2305–2312.
Gall, J. (1963). Chromosome fibers from an interphase nucleus. Science *139*, 120–121.
Gao, X., Havecker, E.R., Baranov, P. V, Atkins, J.F., and Voytas, D.F. (2003). Translational recoding signals between gag and pol in diverse LTR retrotransposons. RNA *9*, 1422–1430.
Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012). PCGF homologs, CBX proteins, and RYBP define functionally distinct PRC1 family complexes. Mol. Cell *45*, 344–356.
Gardiner-Garden, M., and Frommer, M. (1987). CpG Islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.
Georgel, P.T., Horowitz-Scherer, R.A., Adkins, N., Woodcock, C.L., Wade, P.A., and Hansen, J.C. (2003). Chromatin compaction by human MeCP2. Assembly of novel secondary chromatin structures in the absence of DNA methylation. J. Biol. Chem. *278*, 32181–32188.
Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature *428*, 493–521.
Gifford, R., Kabat, P., Martin, J., Lynch, C., and Tristem, M. (2005). Evolution and distribution of class II-related endogenous retroviruses. J. Virol. *79*, 6478–6486.
Gifford, W.D., Pfaff, S.L., and Macfarlan, T.S. (2013). Transposable elements as genetic regulatory substrates in early development. Trends Cell Biol. *23*, 218–226.
Gilbert, C., Schaack, S., Pace, J.K., Brindley, P.J., and Feschotte, C. (2010). A role for host-parasite interactions in the horizontal transfer of transposons across phyla. Nature *464*, 1347–1350.
Gilbert, N., Lutz-Prigge, S., and Moran, J. V. (2002). Genomic Deletions Created upon LINE-1 Retrotransposition. Cell *110*, 315–325.
Ginsburg, M., Snow, M.H., and McLaren, A. (1990). Primordial germ cells in the mouse embryo during gastrulation. Development *110*, 521–528.
Gkountela, S., Zhang, K.X., Shafiq, T.A., Liao, W.-W., Hargan-Calvopiña, J., Chen, P.-Y., and Clark, A.T. (2015). DNA Demethylation Dynamics in the Human Prenatal Germline. Cell *161*, 1425–1436.
Goll, M.G., and Bestor, T.H. (2005). Eukaryotic cytosine methyltransferases. Annu. Rev. Biochem. *74*, 481–514.
Goodier, J.L., and Kazazian, H.H. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell *135*, 23–35.
Goodier, J.L., Ostertag, E.M., and Kazazian, H.H. (2000). Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. *9*, 653–657.
Goodier, J.L., Ostertag, E.M., Du, K., and Kazazian, H.H. (2001). A novel active L1 retrotransposon subfamily in the mouse. Genome Res. *11*, 1677–1685.
Gopalakrishnan, S., Sullivan, B.A., Trazzi, S., Della Valle, G., and Robertson, K.D. (2009). DNMT3B interacts with constitutive centromere protein CENP-C to modulate DNA methylation and the histone code at centromeric regions. Hum. Mol. Genet. *18*, 3178–3193.
Gossen, M., and Bujard, H. (1992). Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc. Natl. Acad. Sci. U. S. A. *89*, 5547–5551.
Graf, T., and Enver, T. (2009). Forcing cells to change lineages. Nature *462*, 587–594.
Grau, D.J., Chapman, B.A., Garlick, J.D., Borowsky, M., Francis, N.J., and Kingston, R.E. (2011). Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. Genes Dev. *25*, 2210–2221.

Graur, D., Shuali, Y., and Li, W.H. (1989). Deletions in processed pseudogenes accumulate faster in rodents than in humans. J. Mol. Evol. *28*, 279–285.
Greenberg, M.V., and Bourc'his, D. (2015). Cultural relativism: maintenance of genomic imprints in pluripotent stem cell culture systems. Curr. Opin. Genet. Dev. *31*, 42–49.
Gregory, T.R. (2001). The bigger the C-value, the larger the cell: genome size and red blood cell size in vertebrates. Blood Cells. Mol. Dis. *27*, 830–843.
Gregory, T.R. (2015). Animal Genome Size Database http://www.genomesize.com.
Gregory, T.R., Andrews, C.B., McGuire, J.A., and Witt, C.C. (2009). The smallest avian genomes are found in hummingbirds. Proc. Biol. Sci. *276*, 3753–3757.
Griffiths, A.J.F., Wessler, S.R., Lewontin, R.C., Fixsen, W.D., and Carroll, S.B. (2008). Introduction to Genetic Analysis, 9th Ed + Solutions Manual (Macmillan Higher Education).
Grotkopp, E., Rejmánek, M., Sanderson, M.J., and Rost, T.L. (2004). Evolution of genome size in pines (Pinus) and its life-history correlates: supertree analyses. Evolution *58*, 1705–1729.
Grow, E.J., Flynn, R. a., Chavez, S.L., Bayless, N.L., Wossidlo, M., Wesche, D.J., Martin, L., Ware, C.B., Blish, C. a., Chang, H.Y., et al. (2015). Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. Nature.
Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. Nature *389*, 349–352.
Gu, T.-P., Guo, F., Yang, H., Wu, H.-P., Xu, G.-F., Liu, W., Xie, Z.-G., Shi, L., He, X., Jin, S., et al. (2011). The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes. Nature *477*, 606–610.
Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., et al. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.
Guenatri, M., Bailly, D., Maison, C., and Almouzni, G. (2004). Mouse centric and pericentric satellite repeats form distinct functional heterochromatin. J. Cell Biol. *166*, 493–505.
Guenatri, M., Duffié, R., Iranzo, J., Fauque, P., and Bourc'his, D. (2013). Plasticity in Dnmt3L-dependent and -independent modes of de novo methylation in the developing mouse embryo. Development *140*, 562–572.
Guibert, S., Forné, T., Weber, M., and Forne, T. (2012). Global profiling of DNA methylation erasure in mouse primordial germ cells. Genome Res. *22*, 633–641.
Guo, F., Li, X., Liang, D., Li, T., Zhu, P., Guo, H., Wu, X., Wen, L., Gu, T.-P., Hu, B., et al. (2014). Active and passive demethylation of male and female pronuclear DNA in the mammalian zygote. Cell Stem Cell *15*, 447–458.
Guo, F., Yan, L., Guo, H., Li, L., Hu, B., Zhao, Y., Yong, J., Hu, Y., Wang, X., Wei, Y., et al. (2015). The Transcriptome and DNA Methylome Landscapes of Human Primordial Germ Cells. Cell *161*, 1437–1452.
Guo, G., Yang, J., Nichols, J., Hall, J.S., Eyres, I., Mansfield, W., and Smith, A. (2009). Klf4 reverts developmentally programmed restriction of ground state pluripotency. Development *136*, 1063–1069.
Gurdon, J.B. (1962). The Developmental Capacity of Nuclei taken from Intestinal Epithelium Cells of Feeding Tadpoles. J Embryol Exp Morphol *10*, 622–640.
Gurdon, J.B., Elsdale, T.R., and Fischberg, M. (1958). Sexually mature individuals of Xenopus laevis from the transplantation of single somatic nuclei. Nature *182*, 64–65.
Habibi, E., Brinkman, A.B., Arand, J., Kroeze, L.I., Kerstens, H.H.D., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N.C., et al. (2013). Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. Cell Stem Cell *13*, 360–369.
Hackett, J.A., Sengupta, R., Zylicz, J.J., Murakami, K., Lee, C., Down, T.A., and Surani, M.A. (2013). Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine. Science *339*, 448–452.
Hajkova, P., Ancelin, K., Waldmann, T., Lacoste, N., Lange, U.C., Cesari, F., Lee, C., Almouzni, G., Schneider, R., and Surani, M.A. (2008). Chromatin dynamics during epigenetic reprogramming in the mouse germ line. Nature *452*, 877–881.
Hajkova, P., Jeffries, S.J., Lee, C., Miller, N., Jackson, S.P., and Surani, M.A. (2010). Genome-wide reprogramming in the mouse germ line entails the base excision repair pathway. Science *329*, 78–82.
Hammoud, S.S., Nix, D.A., Zhang, H., Purwar, J., Carrell, D.T., and Cairns, B.R. (2009). Distinctive chromatin in human sperm packages genes for embryo development. Nature *460*, 473–478.
Han, K., Lee, J., Meyer, T.J., Remedios, P., Goodwin, L., and Batzer, M.A. (2008). L1 recombination-associated deletions generate human genomic variation. Proc. Natl. Acad. Sci. U. S. A. *105*, 19366–19371.
Hansen, J.C. (2002). Conformational Dynamics of the Chromatin Fiber in Solution: Determinants, Mechanisms, and Functions. Annu. Rev. Biophys. Biomol. Struct. *31*, 361–392.
De Harven, E., and Friend, C. (1958). Electron microscope study of a cell-free induced leukemia of the mouse: a preliminary report. J. Biophys. Biochem. Cytol. *4*, 151–156.

Hashimoto, H., Liu, Y., Upadhyay, A.K., Chang, Y., Howerton, S.B., Vertino, P.M., Zhang, X., and Cheng, X. (2012). Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Res. *40*, 4841–4849.

Hathaway, N.A., Bell, O., Hodges, C., Miller, E.L., Neel, D.S., and Crabtree, G.R. (2012). Dynamics and Memory of Heterochromatin in Living Cells. Cell.

Havecker, E.R., Gao, X., and Voytas, D.F. (2004). The diversity of LTR retrotransposons. Genome Biol. *5*, 225.

Hawkins, R.D., Hon, G.C., Lee, L.K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L.E., Kuan, S., Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell Stem Cell *6*, 479–491.

Hayashi, K., Lopes, S.M.C. de S., Tang, F., and Surani, M.A. (2008). Dynamic equilibrium and heterogeneity of mouse pluripotent stem cells with distinct functional and epigenetic states. Cell Stem Cell *3*, 391–401.

Hayward, B.E., Zavanelli, M., and Furano, A. V (1997). Recombination creates novel L1 (LINE-1) elements in Rattus norvegicus. Genetics *146*, 641–654.

He, J., Shen, L., Wan, M., Taranova, O., Wu, H., and Zhang, Y. (2013). Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. Nat. Cell Biol. *15*, 373–384.

He, Q., Kim, H., Huang, R., Lu, W., Tang, M., Shi, F., Yang, D., Zhang, X., Huang, J., Liu, D., et al. (2015). The Daxx/Atrx Complex Protects Tandem Repetitive Elements during DNA Hypomethylation by Promoting H3K9 Trimethylation. Cell Stem Cell *17*, 273–286.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science *333*, 1303–1307.

Heard, E., and Martienssen, R.A. (2014). Transgenerational epigenetic inheritance: myths and mechanisms. Cell *157*, 95–109.

Hedges, S.B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. Mol. Biol. Evol. *32*, 835–845.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Heitz, E. (1928). Das heterochromatin der Moose. Jahrbuch. Wiss. Bot. 762–818.

Heitz, E. (1929). Heterochromatin, chromocentren, chromomeren (Komm. Fischer).

Heitz, E. (1932). Geschlechtschromosomen bei einem Laubmoos: vorläufige Mitteilung.

Hendrich, B., and Bird, A. (1998). Identification and characterization of a family of mammalian methyl-CpG binding proteins. Mol. Cell. Biol. *18*, 6538–6547.

Hirai, H., Karian, P., and Kikyo, N. (2011). Regulation of embryonic stem cell self-renewal and pluripotency by leukaemia inhibitory factor. Biochem. J. *438*, 11–23.

Hisada, K., Sánchez, C., Endo, T.A., Endoh, M., Román-Trufero, M., Sharif, J., Koseki, H., and Vidal, M. (2012). RYBP represses endogenous retroviruses and preimplantation- and germ line-specific genes in mouse embryonic stem cells. Mol. Cell. Biol. *32*, 1139–1149.

Holliday, R., and Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science *187*, 226–232.

Holz-Schietinger, C., and Reich, N.O. (2010). The inherent processivity of the human de novo methyltransferase 3A (DNMT3A) is enhanced by DNMT3L. J. Biol. Chem. *285*, 29091–29100.

Hong, L., Schroth, G., Matthews, H., Yau, P., and Bradbury, E. (1993). Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 "tail" to DNA. J. Biol. Chem. *268*, 305–314.

Houck, M., Clark, J., Peterson, K., and Kidwell, M. (1991). Possible horizontal transfer of Drosophila genes by the mite Proctolaelaps regalis. Science (80-.). *253*, 1125–1128.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. *31*, 827–832.

Hu, G., Kim, J., Xu, Q., Leng, Y., Orkin, S.H., and Elledge, S.J. (2009). A genome-wide RNAi screen identifies a new transcriptional module required for self-renewal. Genes Dev. *23*, 837–848.

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.-F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., et al. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat. Genet. *43*, 476–481.

Huff, J.T., and Zilberman, D. (2014). Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes. Cell *156*, 1286–1297.

Hughes, A.L., and Hughes, M.K. (1995). Small genomes for better flyers. Nature *377*, 391.

Hughes, J.F., and Coffin, J.M. (2001). Evidence for genomic rearrangements mediated by human endogenous retroviruses during primate evolution. Nat. Genet. *29*, 487–489.

Huntley, S., Baggott, D.M., Hamilton, A.T., Tran-Gyamfi, M., Yang, S., Kim, J., Gordon, L., Branscomb, E., and Stubbs, L. (2006). A comprehensive catalog of human KRAB-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res. *16*, 669–677.

Hurles, M. (2005). How homologous recombination generates a mutable genome. Hum. Genomics *2*, 179–186.

Hutnick, L.K., Huang, X., Loo, T.-C.T.-C., Ma, Z., and Fan, G. (2010). Repression of Retrotransposal Elements in Mouse Embryonic Stem Cells Is Primarily Mediated by a DNA Methylation-independent Mechanism. J. Biol. Chem. *285*, 21082–21091.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet. *6*, e1001134.

International Chicken Genome Sequencing Consortium (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature *432*, 695–716.

Isono, K., Endo, T.A., Ku, M., Yamada, D., Suzuki, R., Sharif, J., Ishikura, T., Toyoda, T., Bernstein, B.E., and Koseki, H. (2013). SAM domain polymerization links subnuclear clustering of PRC1 to gene silencing. Dev. Cell *26*, 565–577.

Iwase, S., Xiang, B., Ghosh, S., Ren, T., Lewis, P.W., Cochrane, J.C., Allis, C.D., Picketts, D.J., Patel, D.J., Li, H., et al. (2011). ATRX ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome. Nat. Struct. Mol. Biol. *18*, 769–776.

Iyengar, S., and Farnham, P.J. (2011). KAP1 protein: an enigmatic master regulator of the genome. J. Biol. Chem. *286*, 26267–26276.

Jacob, F. (1977). Evolution and tinkering. Science *196*, 1161–1166.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Jacobs, F.M.J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A.D., Katzman, S., Paten, B., Salama, S.R., and Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature *516*, 242–245.

James, T.C., and Elgin, S.C. (1986). Identification of a nonhistone chromosomal protein associated with heterochromatin in Drosophila melanogaster and its gene. Mol. Cell. Biol. *6*, 3862–3872.

Jamieson, K., Rountree, M.R., Lewis, Z.A., Stajich, J.E., and Selker, E.U. (2013). Regional control of histone H3 lysine 27 methylation in Neurospora. Proc. Natl. Acad. Sci. U. S. A. *110*, 6027–6032.

Jenuwein, T., and Allis, C.D. (2001). Translating the histone code. Science *293*, 1074–1080.

Jenuwein, T., Laible, G., Dorn, R., and Reuter, G. (1998). SET domain proteins modulate chromatin domains in eu- and heterochromatin. Cell. Mol. Life Sci. *54*, 80–93.

Jermann, P., Hoerner, L., Burger, L., Schubeler, D., and Schübeler, D. (2014). Short sequences can efficiently recruit histone H3 lysine 27 trimethylation in the absence of enhancer activity and DNA methylation. Proc. Natl. Acad. Sci. U. S. A. *111*, E3415–E3421.

Jern, P., Sperber, G.O., and Blomberg, J. (2005). Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology *2*, 50.

Jia, D., Jurkowska, R.Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature *449*, 248–251.

Jiang, L., Zhang, J., Wang, J.-J., Wang, L., Zhang, L., Li, G., Yang, X., Ma, X., Sun, X., Cai, J., et al. (2013). Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. Cell *153*, 773–784.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Jirtle, R.L., and Skinner, M.K. (2007). Environmental epigenomics and disease susceptibility. Nat. Rev. Genet. *8*, 253–262.

Jobse, C., Buntjer, J.B., Haagsma, N., Breukelman, H.J., Beintema, J.J., and Lenstral, J.A. (1995). Evolution and recombination of bovine DNA repeats. J. Mol. Evol. *41*, 277–283.

Jockusch, E.L. (1997). An Evolutionary Correlate of Genome Size Change in Plethodontid Salamanders. Proc. Biol. Sci. *264*, 597–604.

Johannes, F., Porcher, E., Teixeira, F.K., Saliba-Colombani, V., Simon, M., Agier, N., Bulski, A., Albuisson, J., Heredia, F., Audigier, P., et al. (2009). Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet. *5*, e1000530.

Johnson, T.B., and Coghill, R.D. (1925). RESEARCHES ON PYRIMIDINES. C111. THE DISCOVERY OF 5-METHYL-CYTOSINE IN TUBERCULINIC ACID, THE NUCLEIC ACID OF THE TUBERCLE BACILLUS 1. J. Am. Chem. Soc. *47*, 2838–2844.

Johnson, M.E., Rowsey, R.A., Shirley, S., Vandevoort, C., Bailey, J., and Hassold, T. (2013). A specific family of interspersed repeats (SINEs) facilitates meiotic synapsis in mammals. Mol. Cytogenet. *6*, 1.

Jolicoeur, P., Rosenberg, N., Cotellessa, A., and Baltimore, D. (1978). Leukemogenicity of clonal isolates of murine leukemia viruses. J. Natl. Cancer Inst. *60*, 1473–1476.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet.

Jørgensen, S., Schotta, G., and Sørensen, C.S. (2013). Histone H4 lysine 20 methylation: key player in epigenetic regulation of genomic integrity. Nucleic Acids Res. *41*, 2797–2806.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. U. S. A. *94*, 1872–1877.

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. *110*, 462–467.

Kaer, K., and Speek, M. (2013). Retroelements in human disease. Gene *518*, 231–241.

Kafri, T., Ariel, M., Brandeis, M., Shemer, R., Urven, L., McCarrey, J., Cedar, H., and Razin, A. (1992). Developmental pattern of gene-specific DNA methylation in the mouse embryo and germ line. Genes Dev. *6*, 705–714.

Kajikawa, M., and Okada, N. (2002). LINEs Mobilize SINEs in the Eel through a Shared 3′Sequence. Cell *111*, 433–444.

Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., and Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. PLoS Biol. *7*, e1000149.

Kaneda, M., Okano, M., Hata, K., Sado, T., Tsujimoto, N., Li, E., and Sasaki, H. (2004). Essential role for de novo DNA methyltransferase Dnmt3a in paternal and maternal imprinting. Nature *429*, 900–903.

Kaneda, M., Tang, F., O'Carroll, D., Lao, K., and Surani, M.A. (2009). Essential role for Argonaute2 protein in mouse oogenesis. Epigenetics Chromatin *2*, 9.

Kapitonov, V. V, and Jurka, J. (1999). Molecular paleontology of transposable elements from Arabidopsis thaliana. Genetica *107*, 27–37.

Kapitonov, V. V, and Jurka, J. (2003). Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proc. Natl. Acad. Sci. U. S. A. *100*, 6569–6574.

Kapitonov, V. V, and Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. *23*, 521–529.

Kapitonov, V. V., and Jurka, J. (2005). RAG1 Core and V(D)J Recombination Signal Sequences Were Derived from Transib Transposons. PLoS Biol. *3*, e181.

Kapitonov, V. V, Tempel, S., and Jurka, J. (2009). Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene *448*, 207–213.

Karimi, M., Luttropp, K., Ekström, T.J., Mikeska, T., Felsberg, J., Hewitt, C. a, and Dobrovic, A. (2011a). Epigenetics Protocols. *791*.

Karimi, M.M., Goyal, P., Maksakova, I.A., Bilenky, M., Leung, D., Tang, J.X., Shinkai, Y., Mager, D.L., Jones, S., Hirst, M., et al. (2011b). DNA Methylation and SETDB1/H3K9me3 Regulate Predominantly Distinct Sets of Genes, Retroelements, and Chimeric Transcripts in mESCs. Cell Stem Cell *8*, 676–687.

Kayne, P.S., Kim, U.J., Han, M., Mullen, J.R., Yoshizaki, F., and Grunstein, M. (1988). Extremely conserved histone H4 N terminus is dispensable for growth but essential for repressing the silent mating loci in yeast. Cell *55*, 27–39.

Kazazian, H.H. (1998). Mobile elements and disease. Curr. Opin. Genet. Dev. *8*, 343–350.

Kazazian, H.H. (1999). An estimated frequency of endogenous insertional mutations in humans. Nat. Genet. *22*, 130.

Keeney, S., Lange, J., and Mohibullah, N. (2014). Self-organization of meiotic recombination initiation: general principles and molecular pathways. Annu. Rev. Genet. *48*, 187–214.

Kelley, J.L., Peyton, J.T., Fiston-Lavier, A.-S., Teets, N.M., Yee, M.-C., Johnston, J.S., Bustamante, C.D., Lee, R.E., and Denlinger, D.L. (2014). Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. Nat. Commun. *5*, 4611.

Khan, H., Smit, A., and Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. *16*, 78–87.

Kidwell, M.G., Kidwell, J.F., and Sved, J.A. (1977). Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. Genetics *86*, 813–833.

Kigami, D. (2002). MuERV-L Is One of the Earliest Transcribed Genes in Mouse One-Cell Embryos. Biol. Reprod. *68*, 651–654.

Kim, A., Terzian, C., Santamaria, P., Pélisson, A., Purd'homme, N., and Bucheton, A. (1994). Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *91*, 1285–1289.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Kim, F.J., Battini, J.-L., Manel, N., and Sitbon, M. (2004). Emergence of vertebrate retroviruses and envelope capture. Virology *318*, 183–191.

Kim, I.S., Lee, M., Park, K.C., Jeon, Y., Park, J.H., Hwang, E.J., Jeon, T.I., Ko, S., Lee, H., Baek, S.H., et al. (2012). Roles of Mis18α in epigenetic regulation of centromeric chromatin and CENP-A loading. Mol. Cell *46*, 260–273.

Kim, J., Kollhoff, A., Bergmann, A., and Stubbs, L. (2003). Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. Hum. Mol. Genet. *12*, 233–245.

Kirik, A., Salomon, S., and Puchta, H. (2000). Species-specific double-strand break repair and genome evolution in plants. EMBO J. *19*, 5562–5566.

Knight, C.A., Clancy, R.B., Götzenberger, L., Dann, L., and Beaulieu, J.M. (2010). On the relationship between pollen size and genome size. J. Bot. *2010*.

Kobayashi, H., Sakurai, T., Imai, M., Takahashi, N., Fukuda, A., Yayoi, O., Sato, S., Nakabayashi, K., Hata, K., Sotomaru, Y., et al. (2012). Contribution of Intragenic DNA Methylation in Mouse Gametic DNA Methylomes to Establish Oocyte-Specific Heritable Marks. PLoS Genet *8*, e1002440.

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

Kohli, R.M., and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. Nature *502*, 472–479.

Kokura, K., Sun, L., Bedford, M.T., and Fang, J. (2010). Methyl-H3K9-binding protein MPP8 mediates E-cadherin gene silencing and promotes tumour cell motility and invasion. EMBO J. *29*, 3673–3687.

De Koning, A.P.J., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet. *7*, e1002384.

Kordis, D., and Gubensek, F. (1997). Bov-B long interspersed repeated DNA (LINE) sequences are present in Vipera ammodytes phospholipase A2 genes and in genomes of Viperidae snakes. Eur. J. Biochem. *246*, 772–779.

Kordis, D., and Gubensek, F. (1998). Unusual horizontal transfer of a long interspersed nuclear element between distant vertebrate classes. Proc. Natl. Acad. Sci. U. S. A. *95*, 10704–10709.

Kornberg, R.D. (1974). Chromatin Structure: A Repeating Unit of Histones and DNA. Science (80-.). *184*, 868–871.

Kossel, A. (1884). Ueber einen peptonartigen Bestandtheil des Zellkerns. Zeitschrift Für Physiol. Chemie *8*, 511–515.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693–705.

Kramerov, D.A., and Vassetzky, N.S. (2001). Structure and origin of a novel dimeric retroposon B1-diD. J. Mol. Evol. *52*, 137–143.

Kramerov, D.A., and Vassetzky, N.S. (2005). Short retroposons in eukaryotic genomes. Int. Rev. Cytol. *247*, 165–221.

Kramerov, D.A., and Vassetzky, N.S. (2011). Origin and evolution of SINEs in eukaryotic genomes. Heredity (Edinb). *107*, 487–495.

Kreahling, J., and Graveley, B.R. (2004). The origins and implications of Aluternative splicing. Trends Genet. *20*, 1–4.

Kress, C., Thomassin, H., and Grange, T. (2006). Active cytosine demethylation triggered by a nuclear receptor involves DNA strand breaks. Proc. Natl. Acad. Sci. U. S. A. *103*, 11112–11117.

Krouwels, I.M., Wiesmeijer, K., Abraham, T.E., Molenaar, C., Verwoerd, N.P., Tanke, H.J., and Dirks, R.W. (2005). A glue for heterochromatin maintenance: stable SUV39H1 binding to heterochromatin is reinforced by the SET domain. J. Cell Biol. *170*, 537–549.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics *27*, 1571–1572.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., et al. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS Genet. *4*, e1000242.

Kuff, E.L., and Lueders, K.K. (1988). The intracisternal A-particle gene family: structure and functional aspects. Adv. Cancer Res. *51*, 183–276.

Kulpa, D.A., and Moran, J. V (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat. Struct. Mol. Biol. *13*, 655–660.

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

Kunath, T., Saba-El-Leil, M.K., Almousailleakh, M., Wray, J., Meloche, S., and Smith, A. (2007). FGF stimulation of the Erk1/2 signalling cascade triggers transition of pluripotent embryonic stem cells from self-renewal to lineage commitment. Development *134*, 2895–2902.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature *410*, 116–120.

Lambowitz, A.M., and Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb. Perspect. Biol. *3*, a003616.

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

Lane, N., Dean, W., Erhardt, S., Hajkova, P., Surani, A., Walter, J., and Reik, W. (2003). Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. Genes. (New York, N.Y. 2000) *35*, 88–93.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Laurent, A.M., Puechberty, J., Prades, C., Gimenez, S., and Roizès, G. (1997). Site-specific retrotransposition of L1 elements within human alphoid satellite sequences. Genomics *46*, 127–132.

Law, J.A., and Jacobsen, S.E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet *11*, 204–220.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Lee, J.Y., and Lee, T.-H. (2012). Effects of DNA methylation on the structure of nucleosomes. J. Am. Chem. Soc. *134*, 173–175.

Lee, D.Y., Hayes, J.J., Pruss, D., and Wolffe, A.P. (1993). A positive role for histone acetylation in transcription factor access to nucleosomal DNA. Cell *72*, 73–84.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K., et al. (2012). Landscape of somatic retrotransposition in human cancers. Science *337*, 967–971.

Lee, H.J., Hore, T.A., and Reik, W. (2014). Reprogramming the methylome: erasing memory and creating diversity. Cell Stem Cell *14*, 710–719.

Lee, J., Han, K., Meyer, T.J., Kim, H.-S., and Batzer, M.A. (2008). Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. PLoS One *3*, e4047.

Leeb, M., Pasini, D., Novatchkova, M., Jaritz, M., Helin, K., and Wutz, A. (2010). Polycomb complexes act redundantly to repress genomic repeats and genes. Genes Dev. *24*, 265–276.

Lehnertz, B., Ueda, Y., Derijck, A.A.H.A.H.A., Braunschweig, U., Perez-Burgos, L., Kubicek, S., Chen, T., Li, E., Jenuwein, T., and Peters, A.H.F.M.F.M. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr. Biol. *13*, 1192–1200.

Lei, H., Oh, S.P., Okano, M., Juttermann, R., Goss, K.A., Jaenisch, R., Li, E., Jüttermann, R., Goss, K.A., Jaenisch, R., et al. (1996). De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development *122*, 3195–3205.

Leitch, H.G., McEwen, K.R., Turp, A., Encheva, V., Carroll, T., Grabole, N., Mansfield, W., Nashun, B., Knezovich, J.G., Smith, A., et al. (2013). Naive pluripotency is associated with global DNA hypomethylation. Nat. Struct. Mol. Biol. *20*, 311–316.

Lerat, E., Brunet, F., Bazin, C., and Capy, P. (1999). Is the evolution of transposable elements modular? Genetica *107*, 15–25.

Leung, D., Du, T., Wagner, U., Xie, W., Lee, A.Y., Goyal, P., Li, Y., Szulwach, K.E., Jin, P., Lorincz, M.C., et al. (2014). Regulation of DNA methylation turnover at LTR retrotransposons and imprinted loci by the histone methyltransferase Setdb1. Proc. Natl. Acad. Sci. U. S. A. *111*, 6690–6695.

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

Levine, S.S., Weiss, A., Erdjument-Bromage, H., Shao, Z., Tempst, P., and Kingston, R.E. (2002). The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. Mol. Cell. Biol. *22*, 6070–6078.

Levine, S.S., King, I.F.G., and Kingston, R.E. (2004). Division of labor in polycomb group repression. Trends Biochem. Sci. *29*, 478–485.

Levinson, G., and Gutman, G.A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. *4*, 203–221.

Levis, R.W., Ganesan, R., Houtchens, K., Tolar, L.A., and Sheen, F. (1993). Transposons in place of telomeric repeats at a Drosophila telomere. Cell *75*, 1083–1093.

Lewis, E.B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565–570.

Lewis, P.W., Elsaesser, S.J., Noh, K.-M., Stadler, S.C., and Allis, C.D. (2010). Daxx is an H3.3-specific histone chaperone and cooperates with ATRX in replication-independent chromatin assembly at telomeres. Proc. Natl. Acad. Sci. U. S. A. *107*, 14075–14080.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.
Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. Nature *366*, 362–365.
Li, G., Margueron, R., Ku, M., Chambon, P., Bernstein, B.E., and Reinberg, D. (2010). Jarid2 and PRC2, partners in regulating gene expression. Genes Dev. *24*, 368–380.
Li, H., Rauch, T., Chen, Z.-X.X., Szabó, P.E., Riggs, A.D., and Pfeifer, G.P. (2006a). The histone methyltransferase SETDB1 and the DNA methyltransferase DNMT3A interact directly and localize to promoters silenced in cancer cells. J. Biol. Chem. *281*, 19489–19500.
Li, J., Kannan, M., Trivett, A.L., Liao, H., Wu, X., Akagi, K., and Symer, D.E. (2014). An antisense promoter in mouse L1 retrotransposon open reading frame-1 initiates expression of diverse fusion transcripts and limits retrotransposition. Nucleic Acids Res. *42*, 4546–4562.
Li, P.W.-L., Li, J., Timmerman, S.L., Krushel, L.A., and Martin, S.L. (2006b). The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: implications for retrotransposition. Nucleic Acids Res. *34*, 853–864.
Licht, L.E., and Lowcock, L.A. (1991). Genome size and metabolic rate in salamanders. Comp. Biochem. Physiol. Part B Comp. Biochem. *100*, 83–92.
Lin, I.G., Han, L., Taghva, A., O'Brien, L.E., and Hsieh, C.-L. (2002). Murine de novo methyltransferase Dnmt3a demonstrates strand asymmetry and site preference in the methylation of DNA in vitro. Mol. Cell. Biol. *22*, 704–723.
Lin, X.L., Lin, Y.Z., Koelsch, G., Gustchina, A., Wlodawer, A., and Tang, J. (1992). Enzymic activities of two-chain pepsinogen, two-chain pepsin, and the amino-terminal lobe of pepsinogen. J. Biol. Chem. *267*, 17257–17263.
Lisch, D., and Slotkin, R.K. (2011). Strategies for silencing and escape: the ancient struggle between transposable elements and their hosts. Int. Rev. Cell Mol. Biol. *292*, 119–152.
Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.
Lister, R., Mukamel, E.A., Nery, J.R., Urich, M., Puddifoot, C.A., Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global epigenomic reconfiguration during mammalian brain development. Science *341*, 1237905.
Liu, J., Carmell, M.A., Rivas, F. V, Marsden, C.G., Thomson, J.M., Song, J.-J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437–1441.
Liu, S., Amour, J.B., Karimi, M.M., Shirane, K., Bogutz, A., Lefebvre, L., Sasaki, H., Shinkai, Y., and Lorincz, M.C. (2014). Setdb1 is required for germline development and silencing of H3K9me3-marked endogenous retroviruses in primordial germ cells. 2041–2055.
Liu, X., Gao, Q., Li, P., Zhao, Q., Zhang, J., Li, J., Koseki, H., and Wong, J. (2013). UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. Nat. Commun. *4*, 1563.
Long, H.K., Blackledge, N.P., and Klose, R.J. (2013). ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. Biochem. Soc. Trans. *41*, 727–740.
Loyola, A., Tagami, H., Bonaldi, T., Roche, D., Quivy, J.P., Imhof, A., Nakatani, Y., Dent, S.Y.R., and Almouzni, G. (2009). The HP1alpha-CAF1-SetDB1-containing complex provides H3K9me1 for Suv39-mediated K9me3 in pericentric heterochromatin. EMBO Rep. *10*, 769–775.
Lu, X., Simon, M.D., Chodaparambil, J. V, Hansen, J.C., Shokat, K.M., and Luger, K. (2008). The effect of H3K79 dimethylation and H4K20 trimethylation on nucleosome and chromatin structure. Nat. Struct. Mol. Biol. *15*, 1122–1124.
Luan, D.D., Korman, M.H., Jakubczak, J.L., and Eickbush, T.H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell *72*, 595–605.
Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251–260.
Luger, K., Dechassa, M.L., and Tremethick, D.J. (2012). New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? Nat. Rev. Mol. Cell Biol. *13*, 436–447.
Lunyak, V. V, Burgess, R., Prefontaine, G.G., Nelson, C., Sze, S.-H., Chenoweth, J., Schwartz, P., Pevzner, P.A., Glass, C., Mandel, G., et al. (2002). Corepressor-dependent silencing of chromosomal regions encoding neuronal genes. Science *298*, 1747–1752.
Lynch, M. (2010). Evolution of the mutation rate. Trends Genet. *26*, 345–352.

Lynch, M.D., Smith, A.J.H., De Gobbi, M., Flenley, M., Hughes, J.R., Vernimmen, D., Ayyub, H., Sharpe, J.A., Sloane-Stanley, J.A., Sutherland, L., et al. (2011). An interspecies analysis reveals a key role for unmethylated CpG dinucleotides in vertebrate Polycomb complex recruitment. EMBO J. *31*, 317–329.

Maatouk, D.M., Kellam, L.D., Mann, M.R.W., Lei, H., Li, E., Bartolomei, M.S., and Resnick, J.L. (2006). DNA methylation is a primary mechanism for silencing postmigratory primordial germ cell genes in both germ cell and somatic cell lineages. Development *133*, 3411–3418.

Macfarlan, T.S., Gifford, W.D., Agarwal, S., Driscoll, S., Lettieri, K., Wang, J., Andrews, S.E., Franco, L., Rosenfeld, M.G., Ren, B., et al. (2011). Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. Genes Dev. *25*, 594–607.

Macfarlan, T.S., Gifford, W.D., Driscoll, S., Lettieri, K., Rowe, H.M., Bonanomi, D., Firth, A., Singer, O., Trono, D., and Pfaff, S.L. (2012). Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature *487*, 57–63.

Maddison, D.R., Schulz, K.-S., and Maddison, W.P. (2007). The tree of life web project. Zootaxa *1668*, 19–40.

Maeshima, K., Imai, R., Tamura, S., and Nozaki, T. (2014). Chromatin as dynamic 10-nm fibers. Chromosoma *123*, 225–237.

Mager, D.L., and Freeman, J.D. (2000). Novel mouse type D endogenous proviruses and ETn elements share long terminal repeat and internal sequences. J. Virol. *74*, 7221–7229.

Mager, D.L., and Goodchild, N.L. (1989). Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. Am. J. Hum. Genet. *45*, 848–854.

Maillard, P. V, Ciaudo, C., Marchais, A., Li, Y., Jay, F., Ding, S.W., and Voinnet, O. (2013). Antiviral RNA interference in mammalian cells. Science *342*, 235–238.

Maison, C., and Almouzni, G. (2004). HP1 and the dynamics of heterochromatin maintenance. Nat. Rev. Mol. Cell Biol. *5*, 296–304.

Maksakova, I.A., Romanish, M.T., Gagnier, L., Dunn, C.A., van de Lagemaat, L.N., and Mager, D.L. (2006). Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet. *2*, e2.

Maksakova, I.A., Goyal, P., Bullwinkel, J., Brown, J.P., Bilenky, M., Mager, D.L., Singh, P.B., and Lorincz, M.C. (2011). H3K9me3-binding proteins are dispensable for SETDB1/H3K9me3-dependent retroviral silencing. Epigenetics Chromatin *4*, 12.

Maksakova, I.A., Thompson, P.J., Goyal, P., Jones, S.J.M., Singh, P.B., Karimi, M.M., and Lorincz, M.C. (2013). Distinct roles of KAP1, HP1 and G9a/GLP in silencing of the two-cell-specific retrotransposon MERVL in mouse ES cells. Epigenetics Chromatin *6*, 15.

Malik, H.S. (2000). Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses. Genome Res. *10*, 1307–1318.

Malik, H.S. (2005). Ribonuclease H evolution in retrotransposable elements. Cytogenet. Genome Res. *110*, 392–401.

Malik, H.S., and Eickbush, T.H. (2001). Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res. *11*, 1187–1197.

Malik, H.S., Burke, W.D., and Eickbush, T.H. (1999). The age and evolution of non-LTR retrotransposable elements. Mol. Biol. Evol. *16*, 793–805.

Del Mar Lorente, M., Marcos-Gutiérrez, C., Pérez, C., Schoorlemmer, J., Ramírez, A., Magin, T., and Vidal, M. (2000). Loss- and gain-of-function mutations show a polycomb group function for Ring1A in mice. Development *127*, 5093–5100.

Margueron, R., and Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. Nat. Rev. Genet. *11*, 285–296.

Margueron, R.R., and Reinberg, D. (2011). The Polycomb complex PRC2 and its mark in life. Nature *469*, 343–349.

Margueron, R., Li, G., Sarma, K., Blais, A., Zavadil, J., Woodcock, C.L., Dynlacht, B.D., and Reinberg, D. (2008). Ezh1 and Ezh2 maintain repressive chromatin through different mechanisms. Mol. Cell *32*, 503–518.

Margueron, R., Justin, N., Ohno, K., Sharpe, M.L., Son, J., Drury 3rd, W.J., Voigt, P., Martin, S.R., Taylor, W.R., De Marco, V., et al. (2009). Role of the polycomb protein EED in the propagation of repressive histone marks. Nature *461*, 762–767.

Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V., and Voinnet, O. (2013). Reconstructing de novo silencing of an active plant retrotransposon. Nat. Genet. *45*, 1029–1039.

Marks, H., and Stunnenberg, H.G. (2014). Transcription regulation and chromatin structure in the pluripotent ground state. Biochim. Biophys. Acta *1839*, 129–137.

Marks, H., Kalkan, T., Menafra, R., Denissov, S., Jones, K., Hofemeister, H., Nichols, J., Kranz, A., Stewart, a F., Smith, A., et al. (2012). The transcriptional and epigenomic foundations of ground state pluripotency. Cell *149*, 590–604.

Marmorstein, R., and Roth, S.Y. (2001). Histone acetyltransferases: function, structure, and catalysis. Curr. Opin. Genet. Dev. *11*, 155–161.

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. *3*, e357.

Martin, G.R. (1981). Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. Proc. Natl. Acad. Sci. U. S. A. *78*, 7634–7638.

Martin, S.L. (2006). Mini-Review Article The ORF1 Protein Encoded by LINE-1 : Structure and Function During L1 Retrotransposition. *2006*, 1–6.

Martin, S.L., and Bushman, F.D. (2001). Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol. Cell. Biol. *21*, 467–475.

Martin, S.L., Branciforte, D., Keller, D., and Bain, D.L. (2003). Trimeric structure for an essential protein in L1 retrotransposition. Proc. Natl. Acad. Sci. U. S. A. *100*, 13815–13820.

Martin, S.L., Cruceanu, M., Branciforte, D., Wai-Lun Li, P., Kwok, S.C., Hodges, R.S., and Williams, M.C. (2005). LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. J. Mol. Biol. *348*, 549–561.

Matsui, T., Leung, D., Miyashita, H., Maksakova, I.A., Miyachi, H., Kimura, H., Tachibana, M., Lorincz, M.C., and Shinkai, Y. (2010). Proviral silencing in embryonic stem cells requires the histone methyltransferase ESET. Nature *464*, 927–931.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., et al. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253–257.

McBryant, S.J., Adams, V.H., and Hansen, J.C. (2006). Chromatin architectural proteins. Chromosome Res. *14*, 39–51.

McCarthy, E.M., McDonald, J.F., and others (2004). Long terminal repeat retrotransposons of Mus musculus. Genome Biol. *5*, R14.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proc. Natl. Acad. Sci. U. S. A. *36*, 344–355.

McClintock, B. (1983). Nobel Lecture: The Significance of Responses of the Genome to Challenge.

Medstrand, P., van de Lagemaat, L.N., and Mager, D.L. (2002). Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res. *12*, 1483–1495.

Mendenhall, E.M., Koche, R.P., Truong, T., Zhou, V.W., Issac, B., Chi, A.S., Ku, M., and Bernstein, B.E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. PLoS Genet. *6*, e1001244.

Meshorer, E., Yellajoshula, D., George, E., Scambler, P.J., Brown, D.T., and Misteli, T. (2006). Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. Dev. Cell *10*, 105–116.

Messerschmidt, D.M., de Vries, W., Ito, M., Solter, D., Ferguson-Smith, A., and Knowles, B.B. (2012). Trim28 is required for epigenetic stability during mouse oocyte to embryo transition. Science *335*, 1499–1502.

Mi, S., Lee, X., Li, X., Veldman, G.M., Finnerty, H., Racie, L., LaVallie, E., Tang, X.Y., Edouard, P., Howes, S., et al. (2000). Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature *403*, 785–789.

Miescher, F. (1871). Über die chemische Zusammensetzung der Eiterzellen. Hoppe-Seyler, Med. Chem. Unters *4*, 441–460.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Mikkelsen, T.S., Hanna, J., Zhang, X., Ku, M., Wernig, M., Schorderet, P., Bernstein, B.E., Jaenisch, R., Lander, E.S., and Meissner, A. (2008). Dissecting direct reprogramming through integrative genomic analysis. Nature *454*, 49–55.

Minc, E., Courvalin, J.C., and Buendia, B. (2000). HP1gamma associates with euchromatin and heterochromatin in mammalian nuclei and chromosomes. Cytogenet. Cell Genet. *90*, 279–284.

Mitchell, R.S., Beitzel, B.F., Schroder, A.R.W., Shinn, P., Chen, H., Berry, C.C., Ecker, J.R., and Bushman, F.D. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. *2*, E234.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schübeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. Mol. Cell *30*, 755–766.

Molaro, A., Falciatori, I., Hodges, E., Aravin, A.A., Marran, K., Rafii, S., McCombie, W.R., Smith, A.D., and Hannon, G.J. (2014). Two waves of de novo methylation during mouse germ cell development. Genes Dev. *28*, 1544–1549.

Molyneaux, K.A., Stallock, J., Schaible, K., and Wylie, C. (2001). Time-lapse analysis of living mouse germ cell migration. Dev. Biol. *240*, 488–498.

Monk, M., Boubelik, M., and Lehnert, S. (1987). Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development *99*, 371–382.

Moran, J. V., DeBerardinis, R.J., and Kazazian, H.H. (1999). Exon Shuffling by L1 Retrotransposition. Science (80-.). *283*, 1530–1534.

Morey, L., Pascual, G., Cozzuto, L., Roma, G., Wutz, A., Benitah, S.A., and Di Croce, L. (2012). Nonoverlapping functions of the Polycomb group Cbx family of proteins in embryonic stem cells. Cell Stem Cell *10*, 47–62.

Morey, L., Aloia, L., Cozzuto, L., Benitah, S.A., and Di Croce, L. (2013). RYBP and Cbx7 define specific biological functions of polycomb complexes in mouse embryonic stem cells. Cell Rep. *3*, 60–69.

Morgan, T.H. (1915). The mechanism of Mendelian heredity,.

Morrish, T.A., Gilbert, N., Myers, J.S., Vincent, B.J., Stamato, T.D., Taccioli, G.E., Batzer, M.A., and Moran, J. V (2002). DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat. Genet. *31*, 159–165.

Morselli, M., Pastor, W. a, Montanini, B., Nee, K., Ferrari, R., Fu, K., Bonora, G., Rubbi, L., Clark, A.T., Ottonello, S., et al. (2015). In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. Elife *4*, 1–21.

Muñoz-López, M., and García-Pérez, J.L. (2010). DNA transposons: nature and applications in genomics. Curr. Genomics *11*, 115–128.

Muotri, A.R., Marchetto, M.C.N., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. Nature *468*, 443–446.

Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. (2005). A fine-scale map of recombination rates and hotspots across the human genome. Science *310*, 321–324.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., et al. (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat. Biotechnol. *33*, 555–562.

Nakamura, T., Arai, Y., Umehara, H., Masuhara, M., Kimura, T., Taniguchi, H., Sekimoto, T., Ikawa, M., Yoneda, Y., Okabe, M., et al. (2007). PGC7/Stella protects against DNA demethylation in early embryogenesis. Nat. Cell Biol. *9*, 64–71.

Nakamura, T., Liu, Y.-J., Nakashima, H., Umehara, H., Inoue, K., Matoba, S., Tachibana, M., Ogura, A., Shinkai, Y., and Nakano, T. (2012). PGC7 binds histone H3K9me2 to protect against conversion of 5mC to 5hmC in early embryos. Nature *486*, 415–419.

Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386–389.

Neira, J.L., Román-Trufero, M., Contreras, L.M., Prieto, J., Singh, G., Barrera, F.N., Renart, M.L., and Vidal, M. (2009). The transcriptional repressor RYBP is a natively unfolded protein which folds upon binding to DNA. Biochemistry *48*, 1348–1360.

Nekrasov, M., Wild, B., and Müller, J. (2005). Nucleosome binding and histone methyltransferase activity of Drosophila PRC2. EMBO Rep. *6*, 348–353.

Nekrutenko, A., and Li, W.H. (2001). Transposable elements are found in a large number of human protein-coding genes. Trends Genet. *17*, 619–621.

Neri, F., Incarnato, D., Krepelova, A., Rapelli, S., Pagnani, A., Zecchina, R., Parlato, C., and Oliviero, S. (2013a). Genome-wide analysis identifies a functional association of Tet1 and Polycomb repressive complex 2 in mouse embryonic stem cells. Genome Biol. *14*, R91.

Neri, F., Krepelova, A., Incarnato, D., Maldotti, M., Parlato, C., Galvagni, F., Matarese, F., Stunnenberg, H.G., and Oliviero, S. (2013b). Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. Cell *155*, 121–134.

Neumann, P., Pozárková, D., and Macas, J. (2003). Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced. Plant Mol. Biol. *53*, 399–410.

Ng, H.H., Zhang, Y., Hendrich, B., Johnson, C.A., Turner, B.M., Erdjument-Bromage, H., Tempst, P., Reinberg, D., and Bird, A. (1999). MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. Nat. Genet. *23*, 58–61.

Ng, R.K., Dean, W., Dawson, C., Lucifero, D., Madeja, Z., Reik, W., and Hemberger, M. (2008). Epigenetic restriction of embryonic cell lineage fate by methylation of Elf5. Nat. Cell Biol. *10*, 1280–1290.

Nielsen, A.L., Oulad-Abdelghani, M., Ortiz, J.A., Remboutsika, E., Chambon, P., and Losson, R. (2001). Heterochromatin Formation in Mammalian Cells. Mol. Cell *7*, 729–739.

Nielsen, A.L., Sanchez, C., Ichinose, H., Cerviño, M., Lerouge, T., Chambon, P., and Losson, R. (2002a). Selective interaction between the chromatin-remodeling factor BRG1 and the heterochromatin-associated protein HP1alpha. EMBO J. *21*, 5797–5806.

Nielsen, P.R., Nietlispach, D., Mott, H.R., Callaghan, J., Bannister, A., Kouzarides, T., Murzin, A.G., Murzina, N. V., and Laue, E.D. (2002b). Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. Nature *416*, 103–107.

Nishino, Y., Eltsov, M., Joti, Y., Ito, K., Takata, H., Takahashi, Y., Hihara, S., Frangakis, A.S., Imamoto, N., Ishikawa, T., et al. (2012). Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. EMBO J. *31*, 1644–1653.

Nishio, H., and Walsh, M.J. (2004). CCAAT displacement protein/cut homolog recruits G9a histone lysine methyltransferase to repress transcription. Proc. Natl. Acad. Sci. U. S. A. *101*, 11257–11262.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Norton, V.G., Imai, B.S., Yau, P., and Bradbury, E.M. (1989). Histone acetylation reduces nucleosome core particle linking number change. Cell *57*, 449–457.

Nusse, R., and Varmus, H.E. (1982). Many tumors induced by the mouse mammary tumor virus contain a provirus integrated in the same region of the host genome. Cell *31*, 99–109.

O'brochta, D.A., and Handler, A.M. (1988). Mobility of P elements in drosophilids and nondrosophilids. Proc. Natl. Acad. Sci. U. S. A. *85*, 6052–6056.

O'Carroll, D., Scherthan, H., Peters, A.H.F.M., Opravil, S., Haynes, A.R., Laible, G., Rea, S., Schmid, M., Lebersorger, A., Jerratsch, M., et al. (2000). Isolation and Characterization of Suv39h2, a Second Histone H3 Methyltransferase Gene That Displays Testis-Specific Expression. Mol. Cell. Biol. *20*, 9423–9433.

Ogawa, H., Ishiguro, K.-I., Gaubatz, S., Livingston, D.M., and Nakatani, Y. (2002). A complex with chromatin modifiers that occupies E2F- and Myc-responsive genes in G0 cells. Science *296*, 1132–1136.

Ohm, J.E., McGarvey, K.M., Yu, X., Cheng, L., Schuebel, K.E., Cope, L., Mohammad, H.P., Chen, W., Daniel, V.C., Yu, W., et al. (2007). A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat. Genet. *39*, 237–242.

Ohno, S. (1972). So much "junk" DNA in our genome. Evol. Genet. Syst. Brookhaven Sympoisa Biol. *23*, 366–370.

Okada, N., Hamada, M., Ogiwara, I., and Ohshima, K. (1997). SINEs and LINEs share common 3′sequences: a review. Gene *205*, 229–243.

Okano, M., Xie, S., and Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. Nat. Genet. *19*, 219–220.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell *99*, 247–257.

Olins, A.L., and Olins, D.E. (1974). Spheroid chromatin units (v bodies). Science *183*, 330–332.

Olins, D.E., and Olins, A.L. (2003). Chromatin history: our view from the bridge. Nat. Rev. Mol. Cell Biol. *4*, 809–814.

Ono, M., Yasunaga, T., Miyata, T., and Ushikubo, H. (1986). Nucleotide sequence of human endogenous retrovirus genome related to the mouse mammary tumor virus genome. J. Virol. *60*, 589–598.

Ooi, S.K., Wolf, D., Hartung, O., Agarwal, S., Daley, G.Q., Goff, S.P., and Bestor, T.H. (2010). Dynamic instability of genomic methylation patterns in pluripotent stem cells. Epigenetics Chromatin *3*, 17.

Ooi, S.K.T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.-P., Allis, C.D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714–717.

Organ, C.L., Shedlock, A.M., Meade, A., Pagel, M., and Edwards, S. V (2007). Origin of avian genome size and structure in non-avian dinosaurs. Nature *446*, 180–184.

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

Ostertag, E.M., Goodier, J.L., Zhang, Y., and Kazazian, H.H. (2003). SVA elements are nonautonomous retrotransposons that cause disease in humans. Am. J. Hum. Genet. *73*, 1444–1451.

Otani, J., Nankumo, T., Arita, K., Inamoto, S., Ariyoshi, M., and Shirakawa, M. (2009). Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain. EMBO Rep. *10*, 1235–1241.

Ovchinnikov, I., Troxel, A.B., and Swergold, G.D. (2001). Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. Genome Res. *11*, 2050–2058.

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

Pace, J.K., Gilbert, C., Clark, M.S., and Feschotte, C. (2008). Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc. Natl. Acad. Sci. U. S. A. *105*, 17023–17028.

Pandey, R.R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-Dinardo, D., and Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. Mol. Cell *32*, 232–246.

Pascale, E., Valle, E., and Furano, A. V. (1990). Amplification of an ancestral mammalian L1 family of long interspersed repeated DNA occurred just before the murine radiation. Proc. Natl. Acad. Sci. *87*, 9481–9485.

Pasini, D., Bracken, A.P., Hansen, J.B., Capillo, M., and Helin, K. (2007). The polycomb group protein Suz12 is required for embryonic stem cell differentiation. Mol. Cell. Biol. *27*, 3769–3779.

Pasini, D., Malatesta, M., Jung, H.R., Walfridsson, J., Willer, A., Olsson, L., Skotte, J., Wutz, A., Porse, B., Jensen, O.N., et al. (2010). Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. Nucleic Acids Res. *38*, 4958–4969.

Pastor, W.A., Stroud, H., Nee, K., Liu, W., Pezic, D., Manakov, S., Lee, S.A., Moissiard, G., Zamudio, N., Bourc'his, D., et al. (2014). MORC1 represses transposable elements in the mouse male germline. Nat. Commun. *5*, 5795.

Pauler, F.M., Sloane, M.A., Huang, R., Regha, K., Koerner, M. V, Tamir, I., Sommer, A., Aszodi, A., Jenuwein, T., and Barlow, D.P. (2009). H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. Genome Res. *19*, 221–233.

Pearson, M.N., and Rohrmann, G.F. (2002). Transfer, Incorporation, and Substitution of Envelope Fusion Proteins among Members of the Baculoviridae, Orthomyxoviridae, and Metaviridae (Insect Retrovirus) Families. J. Virol. *76*, 5301–5304.

Peaston, A.E., Evsikov, A. V, Graber, J.H., de Vries, W.N., Holbrook, A.E., Solter, D., and Knowles, B.B. (2004). Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev. Cell *7*, 597–606.

Peat, J.R., Dean, W., Clark, S.J., Krueger, F., Smallwood, S.A., Ficz, G., Kim, J.K., Marioni, J.C., Hore, T.A., and Reik, W. (2014). Genome-wide bisulfite sequencing in zygotes identifies demethylation targets and maps the contribution of TET3 oxidation. Cell Rep. *9*, 1990–2000.

Pellicer, J., Fay, M., and Leitch, I. (2010). The largest eukaryotic genome of them all? Bot. J. Linn. Soc. *164*, 10–15.

Peng, J.C., Valouev, A., Swigut, T., Zhang, J., Zhao, Y., Sidow, A., and Wysocka, J. (2009). Jarid2/Jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. Cell *139*, 1290–1302.

Perepelitsa-Belancio, V., and Deininger, P. (2003). RNA truncation by premature polyadenylation attenuates human mobile element activity. Nat. Genet. *35*, 363–366.

Perez-Pinera, P., Ousterout, D.G., Brown, M.T., and Gersbach, C.A. (2012). Gene targeting to the ROSA26 locus directed by engineered zinc finger nucleases. Nucleic Acids Res. *40*, 3741–3752.

Peters, J. (2014). The role of genomic imprinting in biology and disease: an expanding view. Nat. Rev. Genet. *15*, 517–530.

Peters, G., and Glover, C. (1980). tRNA's and priming of RNA-directed DNA synthesis in mouse mammary tumor virus. J. Virol. *35*, 31–40.

Peters, A.H.F.M.A.H.F.M., Kubicek, S., Mechtler, K., O'Sullivan, R.J., Derijck, A.A.H.A.A.A.H.A., Perez-Burgos, L., Kohlmaier, A., Opravil, S., Tachibana, M., Shinkai, Y., et al. (2003). Partitioning and plasticity of repressive histone methylation states in mammalian chromatin. Mol. Cell *12*, 1577–1589.

Peters, A.H.F.M.F.M., Carroll, O., Scherthan, H., Mechtler, K., Sauer, S., Scho, C., Weipoltshammer, K., Pagani, M., Lachner, M., Kohlmaier, A., et al. (2001). Loss of the Suv39h Histone Methyltransferases Impairs Mammalian Heterochromatin and Genome Stability. Cell *107*, 323–337.

Petrov, D.A. (2001). Evolution of genome size: new approaches to an old problem. Trends Genet. *17*, 23–28.

Petrov, D.A. (2002). DNA loss and evolution of genome size in Drosophila. Genetica *115*, 81–91.

Petrov, D.A., and Hartl, D.L. (1998). High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol. Biol. Evol. *15*, 293–302.

Petrov, D.A., and Hartl, D.L. (1999). Patterns of nucleotide substitution in Drosophila and mammalian genomes. Proc. Natl. Acad. Sci. U. S. A. *96*, 1475–1479.

Petrov, D.A., and Hartl, D.L. (2000). Pseudogene evolution and natural selection for a compact genome. J. Hered. *91*, 221–227.

Petrov, D.A., Lozovskaya, E.R., and Hartl, D.L. (1996). High intrinsic rate of DNA loss in Drosophila. Nature *384*, 346–349.

Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., and Shaw, K.L. (2000). Evidence for DNA loss as a determinant of genome size. Science *287*, 1060–1062.

Pezic, D., Manakov, S.A., Sachidanandam, R., and Aravin, A.A. (2014). piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. Genes Dev. *28*, 1410–1428.

Pickeral, O.K., Makałowski, W., Boguski, M.S., and Boeke, J.D. (2000). Frequent Human Genomic DNA Transduction Driven by LINE-1 Retrotransposition. Genome Res. *10*, 411–415.

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M., and van Steensel, B. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat. Genet. *38*, 1005–1014.

Pinheiro, I., Margueron, R., Shukeir, N., Eisold, M., Fritzsch, C., Richter, F.M.M., Mittler, G., Genoud, C., Goyama, S., Kurokawa, M., et al. (2012). Prdm3 and Prdm16 are H3K9me1 Methyltransferases Required for Mammalian Heterochromatin Integrity. Cell *150*, 948–960.

Piskurek, O., and Okada, N. (2007). Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. Proc. Natl. Acad. Sci. U. S. A. *104*, 12046–12051.

Plasterk, R.H.., Izsvák, Z., and Ivics, Z. (1999). Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet. *15*, 326–332.

Plath, K., Fang, J., Mlynarczyk-Evans, S.K., Cao, R., Worringer, K.A., Wang, H., de la Cruz, C.C., Otte, A.P., Panning, B., and Zhang, Y. (2003). Role of histone H3 lysine 27 methylation in X inactivation. Science *300*, 131–135.

Ponger, L., and Li, W.-H. (2005). Evolutionary diversification of DNA methyltransferases in eukaryotic genomes. Mol. Biol. Evol. *22*, 1119–1128.

Potok, M.E., Nix, D.A., Parnell, T.J., and Cairns, B.R. (2013). Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. Cell *153*, 759–772.

Poulter, R.T.M., and Goodwin, T.J.D. (2005). DIRS-1 and the other tyrosine recombinase retrotransposons. Cytogenet. Genome Res. *110*, 575–588.

Pritham, E.J., Putliwala, T., and Feschotte, C. (2007). Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene *390*, 3–17.

Promislow, D.E., Jordan, I.K., and McDonald, J.F. (1999). Genomic demography: a life-history analysis of transposable element evolution. Proc. Biol. Sci. *266*, 1555–1560.

Puschendorf, M., Terranova, R., Boutsma, E., Mao, X., Isono, K., Brykczynska, U., Kolb, C., Otte, A.P., Koseki, H., Orkin, S.H., et al. (2008). PRC1 and Suv39h specify parental asymmetry at constitutive heterochromatin in early mouse embryos. Nat. Genet. *40*, 411–420.

Quenneville, S., Verde, G., Corsinotti, A., Kapopoulou, A., Jakobsson, J., Offner, S., Baglivo, I., Pedone, P.V. V, Grimaldi, G., Riccio, A., et al. (2011). In Embryonic Stem Cells, ZFP57/KAP1 Recognize a Methylated Hexanucleotide to Affect Chromatin and DNA Methylation of Imprinting Control Regions. Mol. Cell *44*, 361–372.

Quenneville, S., Turelli, P., Bojkowska, K., Raclot, C., Offner, S., Kapopoulou, A., and Trono, D. (2012). The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. Cell Rep. *2*, 766–773.

Quentin, Y. (1994). A master sequence related to a free left Alu monomer (FLAM) at the origin of the B1 family in rodent genomes. Nucleic Acids Res. *22*, 2222–2227.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., and Manke, T. (2014). deepTools: a flexible platform for exploring deep-sequencing data. Nucleic Acids Res. *42*, W187–W191.

Rando, O.J. (2012). Combinatorial complexity in chromatin structure and function: revisiting the histone code. Curr. Opin. Genet. Dev. *22*, 148–155.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell 1–16.

Ray, D.A., Pagan, H.J.T., Thompson, M.L., and Stevens, R.D. (2007). Bats with hATs: evidence for recent DNA transposon activity in genus Myotis. Mol. Biol. Evol. *24*, 632–639.

Ray, D.A., Feschotte, C., Pagan, H.J.T., Smith, J.D., Pritham, E.J., Arensburger, P., Atkinson, P.W., and Craig, N.L. (2008). Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res. *18*, 717–728.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D., et al. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature *406*, 593–599.

Reichmann, J., Crichton, J.H., Madej, M.J., Taggart, M., Gautier, P., Garcia-Perez, J.L., Meehan, R.R., and Adams, I.R. (2012). Microarray analysis of LTR retrotransposon silencing identifies Hdac1 as a regulator of retrotransposon expression in mouse embryonic stem cells. PLoS Comput. Biol. *8*.

Ribet, D., Louvet-Vallée, S., Harper, F., de Parseval, N., Dewannieux, M., Heidmann, O., Pierron, G., Maro, B., and Heidmann, T. (2008). Murine endogenous retrovirus MuERV-L is the progenitor of the "orphan" epsilon viruslike particles of the early mouse embryo. J. Virol. *82*, 1622–1625.

Rice, J.C., Briggs, S.D., Ueberheide, B., Barber, C.M., Shabanowitz, J., Hunt, D.F., Shinkai, Y., and Allis, C.D. (2003). Histone Methyltransferases Direct Different Degrees of Methylation to Define Distinct Chromatin Domains. Mol. Cell *12*, 1591–1598.

Richard Pilsner, J., Lazarus, A.L., Nam, D.H., Letcher, R.J., Sonne, C., Dietz, R., and Basu, N. (2010). Mercury-associated DNA hypomethylation in polar bear brains via the LUminometric Methylation Assay: A sensitive method to study epigenetics in wildlife. Mol. Ecol. *19*, 307–314.

Riggs, A.D. (1975). X inactivation, differentiation, and DNA methylation. Cytogenet. Genome Res. *14*, 9–25.
Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell *129*, 1311–1323.
Robertson, H.M. (1993). The mariner transposable element is widespread in insects. Nature *362*, 241–245.
Robertson, H.M. (2000). The Large srh Family of Chemoreceptor Genes in Caenorhabditis Nematodes Reveals Processes of Genome Evolution Involving Large Duplications and Deletions and Intron Gains and Losses. Genome Res. *10*, 192–203.
Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.
Robinson, P.J.J., and Rhodes, D. (2006). Structure of the "30 nm" chromatin fibre: a key role for the linker histone. Curr. Opin. Struct. Biol. *16*, 336–343.
Robinson, P.J.J., Fairall, L., Huynh, V.A.T., and Rhodes, D. (2006). EM measurements define the dimensions of the "30-nm" chromatin fiber: evidence for a compact, interdigitated structure. Proc. Natl. Acad. Sci. U. S. A. *103*, 6506–6511.
Da Rocha, S.T., Boeva, V., Escamilla-Del-Arenal, M., Ancelin, K., Granier, C., Matias, N.R., Sanulli, S., Chow, J., Schulz, E., Picard, C., et al. (2014). Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. Mol. Cell *53*, 301–316.
Rodriguez-Paredes, M., and Esteller, M. (2011). Cancer epigenetics reaches mainstream oncology. Nat Med 330–339.
Rollins, R. a, Haghighi, F., Edwards, J.R., Das, R., Zhang, M.Q., Ju, J., and Bestor, T.H. (2006). Large-scale structure of genomic methylation patterns. Genome Res. *16*, 157–163.
Rose, N.R., and Klose, R.J. (2014). Understanding the relationship between DNA methylation and histone lysine methylation. Biochim. Biophys. Acta - Gene Regul. Mech. *1839*, 1362–1372.
Ross, J.P., Rand, K.N., and Molloy, P.L. (2010). Hypomethylation of repeated DNA sequences in cancer. Epigenomics *2*, 245–269.
Rothbart, S.B., Dickson, B.M., Ong, M.S., Krajewski, K., Houliston, S., Kireev, D.B., Arrowsmith, C.H., and Strahl, B.D. (2013). Multivalent histone engagement by the linked tandem Tudor and PHD domains of UHRF1 is required for the epigenetic inheritance of DNA methylation. Genes Dev. *27*, 1288–1298.
Roudier, F., Ahmed, I., Bérard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., et al. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J. *30*, 1928–1938.
Le Rouzic, A., and Capy, P. (2005). The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics *169*, 1033–1043.
Le Rouzic, A., and Capy, P. (2009). Theoretical approaches to the dynamics of transposable elements in Genomes, Populations, and species. In Transposons and the Dynamic Genome, D. Lankenau, and J.-N. Volff, eds. (Springer Science & Business Media), pp. 1–14.
Le Rouzic, A., Boutin, T.S., and Capy, P. (2007a). Long-term evolution of transposable elements. Proc. Natl. Acad. Sci. U. S. A. *104*, 19375–19380.
Le Rouzic, A., Dupas, S., and Capy, P. (2007b). Genome ecosystem and transposable elements species. Gene *390*, 214–220.
Rowe, H.M., and Trono, D. (2011). Dynamic control of endogenous retroviruses during development. Virology *411*, 273–287.
Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P. V, Layard-Liesching, H., Verp, S., Marquis, J., et al. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. Nature *463*, 237–240.
Rowe, H.M., Friedli, M., Offner, S., Verp, S., Mesnard, D., Marquis, J., Aktas, T., and Trono, D. (2013). De novo DNA methylation of endogenous retroviruses is shaped by KRAB-ZFPs/KAP1 and ESET. Development *140*, 519–529.
Rubin, G.M., Kidwell, M.G., and Bingham, P.M. (1982). The molecular basis of P-M hybrid dysgenesis: The nature of induced mutations. Cell *29*, 987–994.
Rush, M., Appanah, R., Lee, S., Lam, L.L., Goyal, P., and Lorincz, M.C. (2009). Targeting of EZH2 to a defined genomic site is sufficient for recruitment of Dnmt3a but not de novo DNA methylation. Epigenetics *4*, 404–414.
Ryan, R.F., Schultz, D.C., Ayyanathan, K., Singh, P.B., Friedman, J.R., Fredericks, W.J., and Rauscher, F.J. (1999). KAP-1 corepressor protein interacts and colocalizes with heterochromatic and euchromatic HP1 proteins: a potential role for Krüppel-associated box-zinc finger proteins in heterochromatin-mediated gene silencing. Mol. Cell. Biol. *19*, 4366–4378.
Sachs, M., Onodera, C., Blaschke, K., Ebata, K.T., Song, J.S., and Ramalho-Santos, M. (2013). Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. Cell Rep. *3*, 1777–1784.

Sadic, D., Schmidt, K., Groh, S., Kondofersky, I., Ellwart, J., Fuchs, C., Theis, F.J., and Schotta, G. (2015). Atrx promotes heterochromatin formation at retrotransposons. EMBO Rep. *16*, 836–850.

Saksouk, N., Barth, T.K.K., Ziegler-Birling, C., Olova, N., Nowak, A., Rey, E., Mateos-Langerak, J., Urbach, S., Reik, W., Torres-Padilla, M.-E., et al. (2014). Redundant Mechanisms to Form Silent Chromatin at Pericentromeric Regions Rely on BEND3 and DNA Methylation. Mol. Cell *2*, 580–594.

Sandmeyer, S.B., and Clemens, K.A. (2010). Function of a retrotransposon nucleocapsid protein. RNA Biol. *7*, 642–654.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. Nat. Genet. *20*, 43–45.

Santos, F., Peat, J., Burgess, H., Rada, C., Reik, W., and Dean, W. (2013). Active demethylation in mouse zygotes involves cytosine deamination and base excision repair. Epigenetics Chromatin *6*, 39.

Sarma, K., Margueron, R., Ivanov, A., Pirrotta, V., and Reinberg, D. (2008). Ezh2 requires PHF1 to efficiently catalyze H3 lysine 27 trimethylation in vivo. Mol. Cell. Biol. *28*, 2718–2731.

Sarraf, S.A., and Stancheva, I. (2004). Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. Mol. Cell *15*, 595–605.

Sasaki, M., Lange, J., and Keeney, S. (2010). Genome destabilization by homologous recombination in the germ line. Nat. Rev. Mol. Cell Biol. *11*, 182–195.

Sauvageau, M., and Sauvageau, G. (2010). Polycomb group proteins: multi-faceted regulators of somatic stem cells and cancer. Cell Stem Cell *7*, 299–313.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. *103*, 1412–1417.

Saxton, J.A., and Martin, S.L. (1998). Recombination between subtypes creates a mosaic lineage of LINE-1 that is expressed and actively retrotransposing in the mouse genome. J. Mol. Biol. *280*, 611–622.

Schalch, T., Duda, S., Sargent, D.F., and Richmond, T.J. (2005). X-ray structure of a tetranucleosome and its implications for the chromatin fibre. Nature *436*, 138–141.

Schliehe, C., Flynn, E.K., Vilagos, B., Richson, U., Swaminathan, S., Bosnjak, B., Bauer, L., Kandasamy, R.K., Griesshammer, I.M., Kosack, L., et al. (2015). The methyltransferase Setdb2 mediates virus-induced susceptibility to bacterial superinfection. Nat. Immunol. *16*, 67–74.

Schmitges, F.W., Prusty, A.B., Faty, M., Stützer, A., Lingaraju, G.M., Aiwazian, J., Sack, R., Hess, D., Li, L., Zhou, S., et al. (2011). Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks. Mol. Cell *42*, 330–341.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science *326*, 1112–1115.

Schoeftner, S., Sengupta, A.K., Kubicek, S., Mechtler, K., Spahn, L., Koseki, H., Jenuwein, T., and Wutz, A. (2006). Recruitment of PRC1 function at the initiation of X inactivation independent of PRC2 and silencing. EMBO J *25*, 3110–3122.

Schotta, G., Lachner, M., Sarma, K., Ebert, A., Sengupta, R., Reuter, G., Reinberg, D., and Jenuwein, T. (2004). A silencing pathway to induce H3-K9 and H4-K20 trimethylation at constitutive heterochromatin. Genes Dev. *18*, 1251–1262.

Schotta, G., Sengupta, R., Kubicek, S., Malin, S., Kauer, M., Callén, E., Celeste, A., Pagani, M., Opravil, S., De La Rosa-Velazquez, I.A., et al. (2008). A chromatin-wide transition to H4K20 monomethylation impairs genome integrity and programmed DNA rearrangements in the mouse. Genes Dev. *22*, 2048–2061.

Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326.

Schueler, M.G., Higgins, A.W., Rudd, M.K., Gustashaw, K., and Willard, H.F. (2001). Genomic and genetic definition of a functional human centromere. Science *294*, 109–115.

Schuettengruber, B., and Cavalli, G. (2009). Recruitment of polycomb group complexes and their role in the dynamic regulation of cell fate choice. Development *136*, 3531–3542.

Schultz, J. (1936). Variegation in Drosophila and the Inert Chromosome Regions. Proc. Natl. Acad. Sci. U. S. A. *22*, 27–33.

Schultz, J. (1941). The evidence of the nucleoprotein nature of the gene. Cold Spring Harb. Symp. Quant. Biol. *9*, 55–65.

Schultz, D.C., Friedman, J.R., and Rauscher, F.J. (2001). Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2alpha subunit of NuRD. Genes Dev. *15*, 428–443.

Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G., and Rauscher, F.J. (2002). SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes Dev. *16*, 919–932.

Schulz, E.G., Meisig, J., Nakamura, T., Okamoto, I., Sieber, A., Picard, C., Borensztein, M., Saitou, M., Blüthgen, N., and Heard, E. (2014). The two active X chromosomes in female ESCs block exit from the pluripotent state by modulating the ESC signaling network. Cell Stem Cell *14*, 203–216.

Schulz, R., Woodfine, K., Menheniott, T.R., Bourc'his, D., Bestor, T., and Oakey, R.J. (2008). WAMIDEX: a web atlas of murine genomic imprinting and differential expression. Epigenetics *3*, 89–96.

Schulz, W.A., Steinhoff, C., and Florl, A.R. (2006). Methylation of endogenous human retroelements in health and disease. Curr. Top. Microbiol. Immunol. *310*, 211–250.

Seisenberger, S., Andrews, S., Krueger, F., Arand, J., Walter, J., Santos, F., Popp, C., Thienpont, B., Dean, W., and Reik, W. (2012). The Dynamics of Genome-wide DNA Methylation Reprogramming in Mouse Primordial Germ Cells. Mol. Cell *48*.

Seisenberger, S., Peat, J.R., Hore, T.A., Dean, W., and Reik, W. (2013). Reprogramming DNA methylation in the mammalian life cycle : building and breaking epigenetic barriers. *95*, 1–11.

Seki, Y., Hayashi, K., Itoh, K., Mizugaki, M., Saitou, M., and Matsui, Y. (2005). Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice. Dev. Biol. *278*, 440–458.

Seki, Y., Yamaji, M., Yabuta, Y., Sano, M., Shigeta, M., Matsui, Y., Saga, Y., Tachibana, M., Shinkai, Y., and Saitou, M. (2007). Cellular dynamics associated with the genome-wide epigenetic reprogramming in migrating primordial germ cells in mice. Development *134*, 2627–2638.

Sen, S.K., Han, K., Wang, J., Lee, J., Wang, H., Callinan, P.A., Dyer, M., Cordaux, R., Liang, P., and Batzer, M.A. (2006). Human genomic deletions mediated by recombination between Alu elements. Am. J. Hum. Genet. *79*, 41–53.

Sen, S.K., Huang, C.T., Han, K., and Batzer, M.A. (2007). Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. Nucleic Acids Res. *35*, 3741–3751.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shao, Z., Raible, F., Mollaaghababa, R., Guyon, J.R., Wu, C.T., Bender, W., and Kingston, R.E. (1999). Stabilization of chromatin structure by PRC1, a Polycomb complex. Cell *98*, 37–46.

Shapiro, J.A. (1969). Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli. J. Mol. Biol. *40*, 93–105.

Sharif, J., Muto, M., Takebayashi, S., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., et al. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature *450*, 908–912.

Shaver, S., Casas-Mollano, J.A., Cerny, R.L., and Cerutti, H. (2014). Origin of the polycomb repressive complex 2 and gene silencing by an E(z) homolog in the unicellular alga Chlamydomonas. Epigenetics *5*, 301–312.

Shinkai, Y., and Tachibana, M. (2011). H3K9 methyltransferase G9a and the related molecule GLP. Genes Dev. *25*, 781–788.

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

Shogren-Knaak, M., Ishii, H., Sun, J.-M., Pazin, M.J., Davie, J.R., and Peterson, C.L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. Science *311*, 844–847.

Silva, J.C., Loreto, E.L., and Clark, J.B. (2004). Factors that affect the horizontal transfer of transposable elements. Curr. Issues Mol. Biol. *6*, 57–71.

Simon, J.A., and Kingston, R.E. (2013). Occupying chromatin: Polycomb mechanisms for getting to genomic targets, stopping transcriptional traffic, and staying put. Mol. Cell *49*, 808–824.

Siomi, H., and Siomi, M.C. (2009). On the road to reading the RNA-interference code. Nature *457*, 396–404.

Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285.

Smallwood, A., Black, J.C., Tanese, N., Pradhan, S., and Carey, M. (2008). HP1-mediated silencing targets Pol II coactivator complexes. Nat. Struct. Mol. Biol. *15*, 318–320.

Smallwood, A., Hon, G.C., Jin, F., Henry, R.E., Espinosa, J.M., and Ren, B. (2012). CBX3 regulates efficient RNA processing genome-wide. Genome Res. *22*, 1426–1436.

Smallwood, S. a, Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W., and Kelsey, G. (2014). Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods *11*, 817–820.

Smallwood, S. a S.A., Tomizawa, S., Krueger, F., Ruf, N., Carli, N., Segonds-Pichon, A., Sato, S., Hata, K., Andrews, S.R., and Kelsey, G. (2011). Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. Nat Genet *advance on*, 811–814.

Smit, A.F. (1993). Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res. *21*, 1863–1872.
Smit, A.F. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. *9*, 657–663.
Smit, A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0.
Smit, A.F., Tóth, G., Riggs, A.D., and Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J. Mol. Biol. *246*, 401–417.
Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nat. Rev. Genet. *14*, 204–220.
Smith, A.G., Heath, J.K., Donaldson, D.D., Wong, G.G., Moreau, J., Stahl, M., and Rogers, D. (1988). Inhibition of pluripotential embryonic stem cell differentiation by purified polypeptides. Nature *336*, 688–690.
Smith, Z.D., Chan, M.M., Mikkelsen, T.S., Gu, H., Gnirke, A., Regev, A., and Meissner, A. (2012). A unique regulatory phase of DNA methylation in the early mammalian embryo. Nature.
Solomon, M.J., Larsen, P.L., and Varshavsky, A. (1988). Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell *53*, 937–947.
Song, J., Rechkoblit, O., Bestor, T.H., and Patel, D.J. (2011). Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science *331*, 1036–1040.
Song, Q., Decato, B., Hong, E.E., Zhou, M., Fang, F., Qu, J., Garvin, T., Kessler, M., Zhou, J., and Smith, A.D. (2013). A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. PLoS One *8*, e81148.
Sookdeo, A., Hepp, C.M., McClure, M. a, and Boissinot, S. (2013). Revisiting the evolution of mouse LINE-1 in the genomic era. Mob. DNA *4*, 3.
Soriano, P., Meunier-Rotival, M., and Bernardi, G. (1983). The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. Proc. Natl. Acad. Sci. U. S. A. *80*, 1816–1820.
Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol. Cell. Biol. *21*, 1973–1985.
Srikanta, D., Sen, S.K., Huang, C.T., Conlin, E.M., Rhodes, R.M., and Batzer, M.A. (2009). An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. Genomics *93*, 205–212.
Sripathy, S.P., Stevens, J., and Schultz, D.C. (2006). The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. Mol. Cell. Biol. *26*, 8623–8638.
Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.
Stancheva, I., El-Maarri, O., Walter, J., Niveleau, A., and Meehan, R.R. (2002). DNA methylation at promoter regions regulates the timing of gene activation in Xenopus laevis embryos. Dev. Biol. *243*, 155–165.
Statham, A.L., Robinson, M.D., Song, J.Z., Coolen, M.W., Stirzaker, C., and Clark, S.J. (2012). Bisulphite-sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA. Genome Res. *22*, 1120–1127.
Van Steensel, B. (2011). Chromatin: constructing the big picture. EMBO J. *30*, 1885–1895.
Stein, R., Razin, A., and Cedar, H. (1982). In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. Proc. Natl. Acad. Sci. U. S. A. *79*, 3418–3422.
Stewart, C.L., Gadi, I., and Bhatt, H. (1994). Stem cells from primordial germ cells can reenter the germ line. Dev. Biol. *161*, 626–628.
Stewart, M.D., Li, J., and Wong, J. (2005). Relationship between histone H3 lysine 9 methylation, transcription repression, and heterochromatin protein 1 recruitment. Mol. Cell. Biol. *25*, 2525–2538.
Stock, J.K., Giadrossi, S., Casanova, M., Brookes, E., Vidal, M., Koseki, H., Brockdorff, N., Fisher, A.G., and Pombo, A. (2007). Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. Nat. Cell Biol. *9*, 1428–1435.
Stocking, C., and Kozak, C. a (2008). Murine endogenous retroviruses. Cell. Mol. Life Sci. *65*, 3383–3398.
Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41–45.
Subramanian, R.P., Wildschutte, J.H., Russo, C., and Coffin, J.M. (2011). Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. Retrovirology *8*, 90.
Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004). DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. J. Biol. Chem. *279*, 27816–27823.

Suh, N., and Blelloch, R. (2011). Small RNAs in early mammalian development: from gametes to gastrulation. Development *138*, 1653–1661.
Sun, C., Shepard, D.B., Chong, R.A., López Arriaza, J., Hall, K., Castoe, T.A., Feschotte, C., Pollock, D.D., and Mueller, R.L. (2012a). LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol. Evol. *4*, 168–183.
Sun, C., López Arriaza, J.R., and Mueller, R.L. (2012b). Slow DNA loss in the gigantic genomes of salamanders. Genome Biol. Evol. *4*, 1340–1348.
Sun, Y.-B., Xiong, Z.-J., Xiang, X.-Y., Liu, S.-P., Zhou, W.-W., Tu, X.-L., Zhong, L., Wang, L., Wu, D.-D., Zhang, B.-L., et al. (2015). Whole-genome sequence of the Tibetan frog Nanorana parkeri and the comparative evolution of tetrapod genomes. Proc. Natl. Acad. Sci. U. S. A. *112*, E1257–E1262.
Suzuki, M.M., and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. *9*, 465–476.
Suzuki, M.M., Kerr, A.R.W., De Sousa, D., and Bird, A. (2007). CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res. *17*, 625–631.
Sverdlov, E.D. (1998). Perpetually mobile footprints of ancient infections in human genome. FEBS Lett. *428*, 1–6.
Swartz, M.N., Trautner, T.A., and Kornberg, A. (1962). Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J. Biol. Chem. *237*, 1961–1967.
Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human L1 Retrotransposition Is Associated with Genetic Instability In Vivo. Cell *110*, 327–338.
Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol. *3*, research0052.1–research0052.18.
Szarski, H. (1970). Changes in the Amount of DNA in Cell Nuclei during Vertebrate Evolution. Nature *226*, 651–652.
Szarski, H. (1983). Cell size and the concept of wasteful and frugal evolutionary strategies. J. Theor. Biol. *105*, 201–209.
Tachibana, M., Sugimoto, K., Fukushima, T., and Shinkai, Y. (2001). Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. J. Biol. Chem. *276*, 25309–25317.
Tachibana, M., Sugimoto, K., Nozaki, M., Ueda, J., Ohta, T., Ohki, M., Fukuda, M., Takeda, N., Niida, H., Kato, H., et al. (2002). G9a histone methyltransferase plays a dominant role in euchromatic histone H3 lysine 9 methylation and is essential for early embryogenesis. Genes Dev. *16*, 1779–1791.
Tachibana, M., Ueda, J., Fukuda, M., Takeda, N., Ohta, T., Iwanari, H., Sakihama, T., Kodama, T., Hamakubo, T., and Shinkai, Y. (2005). Histone methyltransferases G9a and GLP form heteromeric complexes and are both crucial for methylation of euchromatin at H3-K9. Genes Dev. *19*, 815–826.
Tachibana, M., Nozaki, M., Takeda, N., and Shinkai, Y. (2007). Functional dynamics of H3K9 methylation during meiotic prophase progression. EMBO J. *26*, 3346–3359.
Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930–935.
Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.
Takeshima, H., Suetake, I., and Tajima, S. (2008). Mouse Dnmt3a preferentially methylates linker DNA and is inhibited by histone H1. J. Mol. Biol. *383*, 810–821.
Takeshita, K., Suetake, I., Yamashita, E., Suga, M., Narita, H., Nakagawa, A., and Tajima, S. (2011). Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1). Proc. Natl. Acad. Sci. U. S. A. *108*, 9055–9059.
Tamaru, H., and Selker, E.U. (2001). A histone H3 methyltransferase controls DNA methylation in Neurospora crassa. Nature *414*, 277–283.
Tanay, A., O'Donnell, A.H., Damelin, M., and Bestor, T.H. (2007). Hyperconserved CpG domains underlie Polycomb-binding sites. Proc. Natl. Acad. Sci. U. S. A. *104*, 5521–5526.
Tang, W.W.C., Dietmann, S., Irie, N., Leitch, H.G., Floros, V.I., Bradshaw, C.R., Hackett, J.A., Chinnery, P.F., and Surani, M.A. (2015). A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development. Cell *161*, 1453–1467.
Tardat, M., Albert, M., Kunzmann, R., Liu, Z., Kaustov, L., Thierry, R., Duan, S., Brykczynska, U., Arrowsmith, C.H., and Peters, A.H.F.M. (2015). Cbx2 targets PRC1 to constitutive heterochromatin in mouse zygotes in a parent-of-origin-dependent manner. Mol. Cell *58*, 157–171.
Tarlinton, R.E., Meers, J., and Young, P.R. (2006). Retroviral invasion of the koala genome. Nature *442*, 79–81.
Tate, P.H., and Bird, A.P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. Curr. Opin. Genet. Dev. *3*, 226–231.

Tavares, L., Dimitrova, E., Oxley, D., Webster, J., Poot, R., Demmers, J., Bezstarosti, K., Taylor, S., Ura, H., Koide, H., et al. (2012). RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. Cell *148*, 664–678.

Tchenio, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Res. *28*, 411–415.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D.G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196–199.

Thiagalingam, S., Cheng, K.-H., Lee, H.J., Mineva, N., Thiagalingam, A., and Ponte, J.F. (2003). Histone deacetylases: unique players in shaping the epigenetic histone code. Ann. N. Y. Acad. Sci. *983*, 84–100.

Thoma, F., Koller, T., and Klug, A. (1979). Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. J. Cell Biol. *83*, 403–427.

Thomas, C.A. (1971). The genetic organization of chromosomes. Annu. Rev. Genet. *5*, 237–256.

Thomas, J.H., and Schneider, S. (2011). Coevolution of retroelements and tandem zinc finger genes. Genome Res. *21*, 1800–1812.

Thompson, C.B. (1995). New insights into V(D)J recombination and its role in the evolution of the immune system. Immunity *3*, 531–539.

Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

Toyooka, Y., Shimosato, D., Murakami, K., Takahashi, K., and Niwa, H. (2008). Identification and characterization of subpopulations in undifferentiated ES cell culture. Development *135*, 909–918.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Tremblay, K.D., Duran, K.L., and Bartolomei, M.S. (1997). A 5' 2-kilobase-pair region of the imprinted mouse H19 gene exhibits exclusive paternal methylation throughout development. Mol. Cell. Biol. *17*, 4322–4329.

Tremethick, D.J. (2007). Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber. Cell *128*, 651–654.

Trojer, P., Cao, A.R., Gao, Z., Li, Y., Zhang, J., Xu, X., Li, G., Losson, R., Erdjument-Bromage, H., Tempst, P., et al. (2011). L3MBTL2 protein acts in concert with PcG protein-mediated monoubiquitination of H2A to establish a repressive chromatin structure. Mol. Cell *42*, 438–450.

Tse, C., and Hansen, J.C. (1997). Hybrid trypsinized nucleosomal arrays: identification of multiple functional roles of the H2A/H2B and H3/H4 N-termini in chromatin fiber compaction. Biochemistry *36*, 11381–11388.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. Nature *461*, 423–426.

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H.R., et al. (2006). Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes to Cells *11*, 805–814.

Ueda, J., Tachibana, M., Ikura, T., and Shinkai, Y. (2006). Zinc finger protein Wiz links G9a/GLP histone methyltransferases to the co-repressor molecule CtBP. J. Biol. Chem. *281*, 20120–20128.

Vakoc, C.R., Mandat, S.A., Olenchock, B.A., and Blobel, G.A. (2005). Histone H3 lysine 9 methylation and HP1gamma are associated with transcription elongation through mammalian chromatin. Mol. Cell *19*, 381–391.

Vardimon, L., Kressmann, A., Cedar, H., Maechler, M., and Doerfler, W. (1982). Expression of a cloned adenovirus gene is inhibited by in vitro methylation. Proc. Natl. Acad. Sci. U. S. A. *79*, 1073–1077.

Vassetzky, N.S., and Kramerov, D.A. (2013). SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. *41*, D83–D89.

Velasco, G., Hubé, F., Rollin, J., Neuillet, D., Philippe, C., Bouzinba-Segard, H., Galvani, A., Viegas-Péquignot, E., and Francastel, C. (2010). Dnmt3b recruitment through E2F6 transcriptional repressor mediates germ-line gene silencing in murine somatic tissues. Proc. Natl. Acad. Sci. U. S. A. *107*, 9281–9286.

Venkatarama, T., Lai, F., Luo, X., Zhou, Y., Newman, K., and King, M. Lou (2010). Repression of zygotic gene expression in the Xenopus germline. Development *137*, 651–660.

Vinckenbosch, N., Dupanloup, I., and Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. Proc. Natl. Acad. Sci. U. S. A. *103*, 3220–3225.

Viré, E., Brenner, C., Deplus, R., Blanchon, L., Fraga, M., Didelot, C., Morey, L., Van Eynde, A., Bernard, D., Vanderwinden, J.-M., et al. (2006). The Polycomb group protein EZH2 directly controls DNA methylation. Nature *439*, 871–874.

Vlachogiannis, G., Niederhuth, C.E., Tuna, S., Stathopoulou, A., Viiri, K., de Rooij, D.G., Jenner, R.G., Schmitz, R.J., and Ooi, S.K.T. (2015). The Dnmt3L ADD Domain Controls Cytosine Methylation Establishment during Spermatogenesis. Cell Rep. *10*, 944–956.

Voigt, P., LeRoy, G., Drury, W.J., Zee, B.M., Son, J., Beck, D.B., Young, N.L., Garcia, B.A., and Reinberg, D. (2012). Asymmetrically modified nucleosomes. Cell *151*, 181–193.

Volff, J.-N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays *28*, 913–922.

Voncken, J.W., Roelen, B.A.J., Roefs, M., de Vries, S., Verhoeven, E., Marino, S., Deschamps, J., and van Lohuizen, M. (2003). Rnf2 (Ring1b) deficiency causes gastrulation arrest and cell cycle inhibition. Proc. Natl. Acad. Sci. U. S. A. *100*, 2468–2473.

Vos, J.C., De Baere, I., and Plasterk, R.H. (1996). Transposase is the only nematode protein required for in vitro transposition of Tc1. Genes Dev. *10*, 755–761.

Wallau, G.L., Ortiz, M.F., and Loreto, E.L.S. (2012). Horizontal transposon transfer in eukarya: detection, bias, and perspectives. Genome Biol. Evol. *4*, 689–699.

Walsh, C.P., and Xu, G.L. (2006). Cytosine methylation and DNA repair. Curr. Top. Microbiol. Immunol. *301*, 283–315.

Walsh, C.P., Chaillet, J.R., Bestor, T.H., Dev, M.M.A., Exp, G.M.A., Med, B., Natl, R.J.P., Res, R.J.M., Walsh, C.P., Chaillet, J.R., et al. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. Nat. Genet. *20*, 116–117.

Wang, H., An, W., Cao, R., Xia, L., Erdjument-Bromage, H., Chatton, B., Tempst, P., Roeder, R.G., and Zhang, Y. (2003). mAM Facilitates Conversion by ESET of Dimethyl to Trimethyl Lysine 9 of Histone H3 to Cause Transcriptional Repression. Mol. Cell *12*, 475–487.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R.S., and Zhang, Y. (2004). Role of histone H2A ubiquitination in Polycomb silencing. Nature *431*, 873–878.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., and Batzer, M.A. (2005). SVA Elements: A Hominid-specific Retroposon Family. J. Mol. Biol. *354*, 994–1007.

Wang, J., Xie, G., Singh, M., Ghanbarian, A.T., Raskó, T., Szvetnik, A., Cai, H., Besser, D., Prigione, A., Fuchs, N. V., et al. (2014a). Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature.

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

Wang, R., Taylor, A.B., Leal, B.Z., Chadwell, L. V, Ilangovan, U., Robinson, A.K., Schirf, V., Hart, P.J., Lafer, E.M., Demeler, B., et al. (2010). Polycomb group targeting through different binding partners of RING1B C-terminal domain. Structure *18*, 966–975.

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., Chiba, H., Kohara, Y., Kono, T., Nakano, T., et al. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature *453*, 539–543.

Watanabe, T., Tomizawa, S. -i. S., Mitsuya, K., Totoki, Y., Yamamoto, Y., Kuramochi-Miyagawa, S., Iida, N., Hoki, Y., Murphy, P.J., Toyoda, A., et al. (2011). Role for piRNAs and Noncoding RNA in de Novo DNA Methylation of the Imprinted Mouse Rasgrf1 Locus. Science (80-.). *332*, 848–852.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. *39*, 457–466.

Weichenrieder, O., Wild, K., Strub, K., and Cusack, S. (2000). Structure and assembly of the Alu domain of the mammalian signal recognition particle. Nature *408*, 167–173.

Wen, B., Wu, H., Shinkai, Y., Irizarry, R.A., and Feinberg, A.P. (2009). Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. Nat. Genet. *41*, 246–250.

Whitcomb, S.J., Basu, A., Allis, C.D., and Bernstein, E. (2007). Polycomb Group proteins: an evolutionary perspective. Trends Genet. *23*, 494–502.

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

Wiederschain, D., Susan, W., Chen, L., Loo, A., Yang, G., Huang, A., Chen, Y., Caponigro, G., Yao, Y., and Lengauer, C. (2009). Single-vector inducible lentiviral RNAi system for oncology target validation. Cell Cycle *8*, 498–504.

Wienholz, B.L., Kareta, M.S., Moarefi, A.H., Gordon, C.A., Ginno, P.A., and Chédin, F. (2010). DNMT3L modulates significant and distinct flanking sequence preference for DNA methylation by DNMT3A and DNMT3B in vivo. PLoS Genet. *6*, e1001106.

Williams, R.L., Hilton, D.J., Pease, S., Willson, T.A., Stewart, C.L., Gearing, D.P., Wagner, E.F., Metcalf, D., Nicola, N.A., and Gough, N.M. (1988). Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. Nature *336*, 684–687.

Wolf, D., and Goff, S.P. (2007). TRIM28 mediates primer binding site-targeted silencing of murine leukemia virus in embryonic cells. Cell *131*, 46–57.

Wolf, D., and Goff, S.P. (2009). Embryonic stem cells use ZFP809 to silence retroviral DNAs. Nature *458*, 1201–1204.

Wolf, G., Yang, P., Füchtbauer, A.C., Füchtbauer, E.-M., Silva, A.M., Park, C., Wu, W., Nielsen, A.L., Pedersen, F.S., and Macfarlan, T.S. (2015). The KRAB zinc finger protein ZFP809 is required to initiate epigenetic silencing of endogenous retroviruses. Genes Dev. *29*, 538–554.

Woodcock, C.L., Skoultchi, A.I., and Fan, Y. (2006). Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. Chromosome Res. *14*, 17–25.

Wray, J., Kalkan, T., and Smith, A.G. (2010). The ground state of pluripotency. Biochem. Soc. Trans. *38*, 1027–1032.

Wright, S.I., Agrawal, N., and Bureau, T.E. (2003). Effects of recombination rate and gene density on transposable element distributions in Arabidopsis thaliana. Genome Res. *13*, 1897–1903.

Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Sun, Y.E., and Zhang, Y. (2011a). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature *473*, 389–393.

Wu, H., Chen, X., Xiong, J., Li, Y., Li, H., Ding, X., Liu, S., Chen, S., Gao, S., and Zhu, B. (2011b). Histone methyltransferase G9a contributes to H3K27 methylation in vivo. Cell Res. *21*, 365–367.

Wu, X., Johansen, J.V., and Helin, K. (2013). Fbxl10/Kdm2b recruits polycomb repressive complex 1 to CpG islands and regulates H2A ubiquitylation. Mol. Cell *49*, 1134–1146.

Wyatt, G. (1950). Occurrence of 5-Methyl-Cytosine in Nucleic Acids. Nature *166*, 237–238.

Xing, J., Wang, H., Belancio, V.P., Cordaux, R., Deininger, P.L., and Batzer, M.A. (2006). Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc. Natl. Acad. Sci. U. S. A. *103*, 17608–17613.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A., et al. (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res. *19*, 1516–1526.

Xiong, Y., and Eickbush, T.H. (1988). Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol. Biol. Evol. *5*, 675–690.

Xiong, Y., and Eickbush, T.H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. *9*, 3353–3362.

Xu, P.-F., Zhu, K.-Y., Jin, Y., Chen, Y., Sun, X.-J., Deng, M., Chen, S.-J., Chen, Z., and Liu, T.X. (2010). Setdb2 restricts dorsal organizer territory and regulates left-right asymmetry through suppressing fgf8 activity. Proc. Natl. Acad. Sci. U. S. A. *107*, 2521–2526.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Ying, Q.L., Nichols, J., Chambers, I., and Smith, A. (2003). BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. Cell *115*, 281–292.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. *13*, 335–340.

Yokochi, T., Poduch, K., Ryba, T., Lu, J., Hiratani, I., Tachibana, M., Shinkai, Y., and Gilbert, D.M. (2009). G9a selectively represses a class of late-replicating genes at the nuclear periphery. Proc. Natl. Acad. Sci. U. S. A. *106*, 19363–19368.

Yoshiyama, M., Tu, Z., Kainoh, Y., Honda, H., Shono, T., and Kimura, K. (2001). Possible horizontal transfer of a transposable element from host to parasitoid. Mol. Biol. Evol. *18*, 1952–1958.

You, J.S., and Jones, P.A. (2012). Cancer genetics and epigenetics: two sides of the same coin? Cancer Cell *22*, 9–20.

Youngson, N.A., Kocialkowski, S., Peel, N., and Ferguson-Smith, A.C. (2005). A small family of sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting. J. Mol. Evol. *61*, 481–490.

Yu, M., Mazor, T., Huang, H., Huang, H.-T., Kathrein, K.L., Woo, A.J., Chouinard, C.R., Labadorf, A., Akie, T.E., Moran, T.B., et al. (2012). Direct recruitment of polycomb repressive complex 1 to chromatin by core binding transcription factors. Mol. Cell *45*, 330–343.

Zamudio, N., and Bourc'his, D. (2010). Transposable elements in the mammalian germline: a comfortable niche or a deadly trap&quest. Heredity (Edinb). *105*, 92–104.

Zamudio, N., Barau, J., Teissandier, A., Walter, M., Borsos, M., Servant, N., and Bourc'his, D. (2015). DNA methylation restrains transposons from adopting a chromatin signature permissive for meiotic recombination. Genes Dev. *29*, 1256–1270.

Zemach, A., McDaniel, I.E., Silva, P., and Zilberman, D. (2010). Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science *328*, 916–919.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. *18*, 292–298.

Zhang, K., Mosch, K., Fischle, W., and Grewal, S.I.S. (2008a). Roles of the Clr4 methyltransferase complex in nucleation, spreading and maintenance of heterochromatin. Nat. Struct. Mol. Biol. *15*, 381–388.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008b). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, Z., Harrison, P.M., Liu, Y., and Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. Genome Res. *13*, 2541–2558.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953.

Zhao, X.D., Han, X., Chew, J.L., Liu, J., Chiu, K.P., Choo, A., Orlov, Y.L., Sung, W.-K., Shahab, A., Kuznetsov, V.A., et al. (2007). Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. Cell Stem Cell *1*, 286–298.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P.L., et al. (2013). Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell *152*, 642–654.

Zuo, X., Sheng, J., Lau, H.-T., McDonald, C.M., Andrade, M., Cullen, D.E., Bell, F.T., Iacovino, M., Kyba, M., Xu, G., et al. (2012). Zinc finger protein ZFP57 requires its co-factor to recruit DNA methyltransferases and maintains DNA methylation imprint in embryonic stem cells via its transcriptional repression domain. J. Biol. Chem. *287*, 2107–2118.