IDENTIFICATION OF DIFFERENTIALLY METHYLATED REGIONS (DMRs) BETWEEN EMBRYONIC STEM CELLS (ESCs) AND EMBRYONIC GERM CELLS (EGCs) BY DNA METHYLATION STUDIES

KHOO JO LYNN

UNIVERSITI SAINS MALAYSIA

2018

IDENTIFICATION OF DIFFERENTIALLY METHYLATED REGIONS (DMRs) BETWEEN EMBRYONIC STEM CELLS (ESCs) AND EMBRYONIC GERM CELLS (EGCs) BY DNA METHYLATION STUDIES

by

KHOO JO LYNN

Thesis submitted in fulfilment of the requirements

for the degree of

Doctor of Philosophy

May 2018

ACKNOWLEDGEMENT

First and foremost, I would like to express my heartfelt thanks to my advisor, Professor Dr. Shaharum Shamsuddin for his assistance and guidance in getting my graduate career started on the right foot. I greatly appreciated his scientific counsel, support and insightful contributions that have helped me overcome many hurdles. His support for the research work carried out for my thesis is greatly appreciated. Not forgetting also Dr. Edinur Hisham Atan and Dr. Ezzeddin Kamil Mohamed Hashim for their assistance in the completion of this thesis.

I would also like to express my sincere gratitude to Prof. Kuniya Abe for the opportunity to join his laboratory and for providing a collaborative environment of seemingly limitless scientific freedom. His continued support and guidance during my graduate studies, in addition to his earnest outlook and indomitable conviction towards solving problems, are things that I have always looked up to and hopefully learnt from him. Working with him has been a rewarding journey and I cannot be grateful enough for his encouragement, faith and patience during my work on this thesis. His tireless enthusiasm and mentoring have made my time in the laboratory an amazing experience, one that I will always cherish.

I am also grateful to both past and present members of Mammalian Genome Dynamics Department, RIKEN BioResource Centre, Tsukuba, Japan, for their continuous support and encouragement all this while by creating a zestful and fun–filled environment during the difficult periods of my research work. In this regard, I would like to thank Madam Masayo Kondo, Madam Yumiko Koga, Dr. Michihiko Sugimoto, Dr. Hirosuke Shiura, Dr. Hiroki Ura, and Dr. Shinnosuke Suzuki for their assistance and discussions that have helped me during my research work. Not forgetting Madam Miwako Kusayama who has greatly assisted me during my stay in Japan.

Last but not least, I would like to thank my parents and family members for their unending love, support and encouragement, for without them, this would not have been possible.

TABLE OF CONTENTS

CHAPTER THREE – MATERIALS AND METHODS

3.1 Materials

CHAPTER FOUR – RESULTS

CHAPTER FIVE – DISCUSSION

CHAPTER SIX – CONCLUSION AND FUTURE STUDIES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CGIs	CpG islands
DMEM	Dulbecco’s Modified Eagle’s Medium
DMRs	Differentially methylated regions
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
E	Embryonic day
EB	Embryoid bodies
EGCs	Embryonic germ cells
ESCs	Embryonic stem cells
FBS	Fetal bovine serum
FFPE	Formalin–fixed paraffin–embedded
gDMRs	Germ line differentially methylated regions
GMEM	Glasgow Minimum Essential Medium
HATs	Histone acetyltransferases
HDACs	Histone deacetylases
H3K4	Histone H3 at lysine 4
H3K9	Histone H3 at lysine 9
H4K20	Histone H3 at lysine 20
H3K27	Histone H3 at lysine 27
ICM	Inner cell mass
ICRs	Imprinting control regions
LINEs	Long interspersed nuclear elements

LMRs	Low–methylated regions
LTR	Long terminal repeats
MECP2	Methyl–CpG–binding protein 2
MEF	Mouse embryonic fibroblast
MPC	2–methacryloyloxyethyl phosphorylcholine
PcG	Polycomb group
PGCs	Primordial germ cells
RNA	Ribonucleic acid
RRBS	Reduced representation bisulfite sequencing
SETDB1	SET domain bifurcated
SINEs	Short interspersed nuclear elements
SNPs	Single nucleotide polymorphisms
SureSelect–PBAT	SureSelect^{XT} Methyl–Seq Combined with Post–Bisulfite Adaptor Tagging
TrxG	Trithorax group
UHRF1	Ubiquitin Like With PHD And Ring Finger Domains 1
WGBS	Whole–genome bisulfite sequencing
5mC	5–methylcytosine

LIST OF SYMBOLS

α	Alpha
β	Beta
µL	Micro litres
µM	Micro molar
µg	Micro gram
mL	Milli litres
ng	Nano gram
pg	Pico gram
nm	Nanometer
mm	Millimeter
cm	Centimetre
g	Gravity
bp	Base pairs
Kb	Kilo base pairs
° C	Degree Celcius
%	Percentage

LIST OF APPENDICES

IDENTIFIKASI KAWASAN TERMETILASI TERBEZA (DMRs) ANTARA STEM SEL EMBRIONIK (ESCs) DAN STEM SEL GERMA (EGCs) MELALUI KAJIAN DNA METILASI

ABSTRAK

Epigenetik adalah pengajian berkenaan perubahan terwaris yang tidak disebabkan oleh perubahan jujukan DNA. Ia melibatkan modifikasi pos–translasi histon dan modifikasi sitosina yang natijahnya adalah perubahan jangka panjang potensi transkripsi dalam sel yang tidak semestinya di wariskan. Satu daripada kejadian utama epigenetic adalah peneraan genomik, sejenis fenomena di mana sebahagian gen di ekspreskan berdasarkan sifat asal usul ibu bapa. Sehingga kini terdapat lebih dari 100 gen yang telah diketahui dikawalatur oleh peneraan genomik yang mana mempunyai peranan yang signifikan dalam embriogenesis, pembentukan plasenta dan perkembangan otak. Dalam kajian ini, stem sel embrionik (ESCs) dan stem sel germa (EGC) telah digunakan. ESC adalah sel stem pluripoten daripada jasad dalaman sel (ICM) embrio pra–implantasi yang di panggil blastosis, manakala EGC pula adalah sel stem pluripoten yang berasal dari sel germa primodial (PGC). Walaupun genom sel ICM lazimnya terhipo–metilasi, penekanan metilasi DNA asas adalah kekal dalam ICM dan derivasinya, i.e. ESCs. Oleh itu, pengekspresan gen tertekan monoalel akan dilengkapkan dalam ESC semasa pembezaan. Sebagai perbandingan, PGC yang menguasai dalam gonad telah melalui DNA metilasi global dengan mana setiap metilasi DNA tertekan sepatutnya dihapuskan. Oleh yang demikian, ESC dan EGC adalah sel serupa kecuali pada satu aspek yakni status `tanggap–tiru'. Perbandingan ESC dan EGC mungkin mendedahkan modifikasi yang diperlukan untuk pengekspresan gen tertekan. Di akhir kajian, ESC dan EGC dibezakan dengan pembentukan jasad embroid dan dituai setiap hari, diverifikasi secara qPCR dan pengekspresan alel spesifik. Metilasi pada imprint disahkan kekal pada ESC tapi tidak dalam EGC pada kedua–dua tahap non–pembezaan dan pembezaan menggunakan RRBS dan SureSelectXT Methyl–Seq digabungkan dengan kaedah PBAT (SureSelect–PBAT). Beberapa kawasan termetilasi terbeza (DMR) selain dari kawasan yang sedia diketahui tertekan telah dikenalpasti buat kali pertama antara ESC dan EGC. Yang menariknya, DMR yang bersifat hipermetilasi dalam EGC manakala ter-hipometilasi dalam ESC juga berjaya dikenalpasti. Walaupun beberapa DMR ini telah divalidasi secara teknikalnya, terdapat perbezaan tahap metilasi antara kaedah RRBS dan SureSelect–PBAT. Walaubagaimana pun, kesemua DMR ini masih menunjukan keadaan hipermetilasi/hipometilasi yang sama pada ESC/EGC. Kajian pengekspresan gen menggunakan RNA–Seq menunjukan hanya sebahagian kecil gen dengan 2 atau lebih kali ganda perbezaan dalam pengekspresan antara ESG dan EGC boleh dikaitkan dengan tahap metilasi kerana kebanyakan DMR dikenalpasti pada intron dan kawasan intergenik dan bukannya pada kawasan promoter.

IDENTIFICATION OF DIFFERENTIALLY METHYLATED REGIONS (DMRs) BETWEEN EMBRYONIC STEM CELLS (ESCs) AND EMBRYONIC GERM CELLS (EGCs) BY DNA METHYLATION STUDIES

ABSTRACT

Epigenetic is the study of heritable changes that are not caused by the changes in the DNA sequence. It involves post–translational modifications of histones and cytosine modifications, resulting in long–term alterations in the transcriptional potential of a cell that are not necessarily heritable. One of the major epigenetic events, known as genomic imprinting, is a phenomenon in which some genes are expressed in a parent–of–origin–specific manner. To date, over 100 genes are known to be regulated by genomic imprints, many of which have significant roles for embryogenesis, placental formation and brain development. In this study, embryonic stem cells (ESCs) and embryonic germ cells (EGCs) were used. ESCs are pluripotent stem cells derived from inner cell mass (ICM) of pre–implantation embryos called blastocyst, while EGCs are pluripotent stem cells originated from primordial germ cells (PGCs). Although genomes of ICM cells are globally hypomethylated, primary DNA methylation imprints are retained in ICM and its derivatives, i.e. ESCs. Therefore, monoallelic imprinted gene expression will be established in ESCs upon differentiation. In contrast, PGCs colonised in the gonads have undergone global DNA demethylation, by which DNA methylation imprints are supposed to be erased. Thus, ESCs and EGCs are highly similar cells except for one aspect, that is, the imprinting status. Comparison of ESCs with EGCs may uncover epigenetic modifications required for the establishment of imprinted gene expression. Toward this end, ESCs and EGCs were differentiated by embryoid body formation and harvested every other day, being verified by qPCR and allele specific expression. Methylation of imprints were confirmed to be retained in ESCs but not EGCs at both undifferentiated and differentiated state by Reduced Representative Bisulfite Sequencing (RRBS) and SureSelect^{XT} Methyl–Seq Combined with Post–Bisulfite Adaptor Tagging (SureSelect–PBAT) methods. A number of differentially methylated regions (DMRs) other than known imprinted regions were identified for the first time between ESCs and EGCs. Interestingly, DMRs that are hypermethylated in EGCs while being hypomethylated in ESCs were also able to be identified. Although some of these DMRs were technically validated, methylation levels showed were different from that of RRBS and SureSelect–PBAT method. Nonetheless, these DMRs still exhibited the same hypermethylated/hypomethylated state in ESCs/EGCs. Gene expression studies by RNA–seq demonstrated that only a small proportion of genes with 2–or more–fold difference in expression between ESCs and EGCs were able to be correlated with their methylation levels, as most of the DMRs were identified at intron and intergenic regions, not promoter regions.

CHAPTER ONE

INTRODUCTION

1.1 Introduction

Epigenetics is the study of heritable changes that are not caused by the changes in the DNA sequence (Berger *et al.,* 2009). It resides in post–translational modifications of histones and cytosine modifications, which result in long–term alterations in the transcriptional potential of a cell that are not necessarily heritable. One of the major epigenetic events, known as genomic imprinting, is a phenomenon in which some genes are expressed in a parent–of–origin–specific manner (McGrath and Solter, 1984; Surani *et al.*, 1984). Primary imprints that determine which allele is expressed from two allelic loci are laid onto the genome during gametogenesis. However, monoallelic expression for some of the imprinted genes occur during embryonic development. To date, over 150 genes are known to be regulated by genomic imprints, many of which have significant roles for embryogenesis, placental formation and brain development (Blake *et al.*, 2010). Although genes and factors required for imprinting have been studied, epigenetic mechanisms for the establishment of imprinted gene expression are still poorly understood.

Embryonic stem cells (ESCs) and embryonic germ cells (EGCs) possess the same properties of self–renewal and pluripotency, hence are often considered as essentially equivalent cell types (Sharova *et al.*, 2007). Both ESCs and EGCs are able to form embryoid bodies (EBs) when cells are induced to differentiate into different cell lineages (Rossant, 2008; Ma and Chen, 2005; Takahashi and Yamanaka, 2006). As EBs from both ESCs and EGCs are able to give rise to all three germ layers upon differentiation, they are suitable to be used as a model system that mimics developmental programmes *in vivo* (Murry and Keller, 2008).

It has been reported that ESCs and EGCs differ in imprinting status (Sharova *et al.*, 2007; Kagiwada *et* al., 2013; Seisenberger *et al.*, 2012). Although genomes of inner cell mass (ICM) cells are globally hypomethylated, primary DNA methylation imprints are retained in ICM and its derivatives, such as ESCs. Therefore, monoallelic imprinted gene expression will be established in ESCs upon differentiation. In contrast, primordial germ cells (PGCs), for which EGCs are derived from, colonise the gonads and undergo global DNA demethylation. This step of global DNA demethylation removes cytosine methylation in a time and locus–specific manner, which includes the DNA methylation of imprints. Therefore, comparison of ESCs and EGCs at undifferentiated and differentiated stages will enable us to uncover differentially methylated regions (DMRs) during development.

The ultimate aim of this study is to verify DNA methylation difference between ESCs and EGCs at known germ line differentially methylated regions (gDMRs) as well as to uncover novel DMRs at undifferentiated and differentiation state.

To achieve this aim, the following specific objectives were targeted and were done according to the overall work–flow (Figure 1.1) :

1. To determine differentially methylated regions (DMRs) between undifferentiated embryonic stem cells (ESCs) and embryonic germ cells (EGCs).

2. To determine differentially methylated regions (DMRs) between ESCs and EGCs developing *in vitro*.

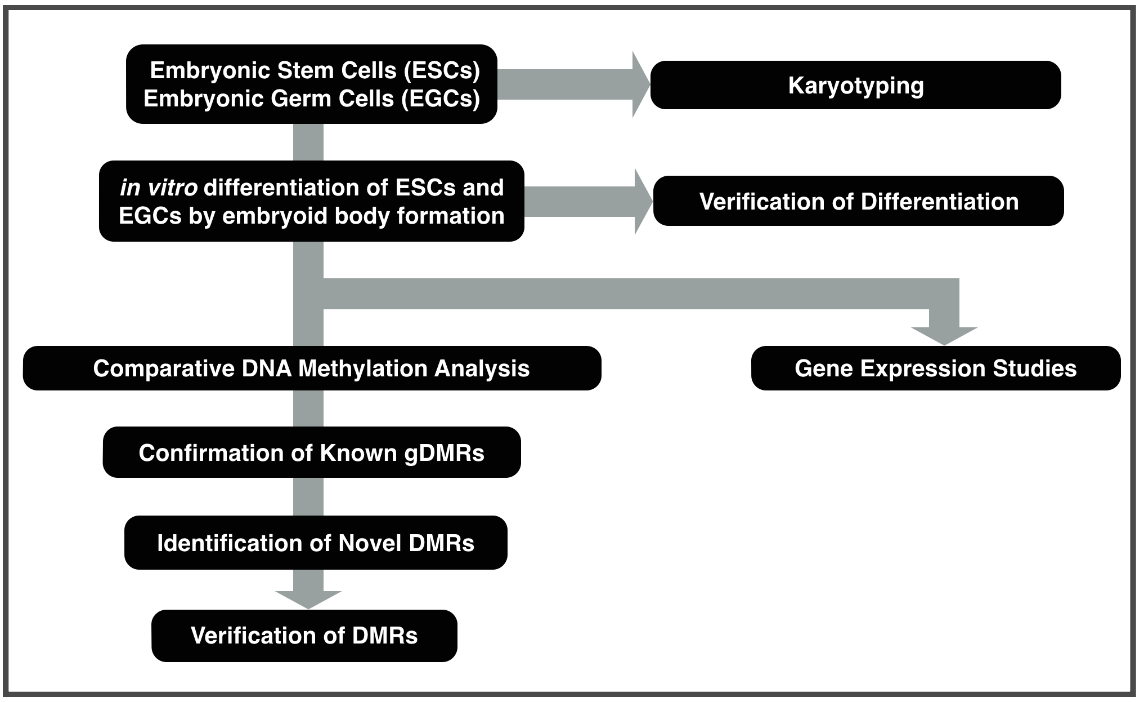
Embryonic Stem Cells (ESCs)
Embryonic Germ Cells (EGCs)
Karyotyping
in vitro differentiation of ESCs and EGCs by embryoid body formation
Verification of Differentiation
Comparative DNA Methylation Analysis
Gene Expression Studies
Confirmation of Known gDMRs
Identification of Novel DMRs
Verification of DMRs

Figure 1.1 : Overall work–flow for this study.

CHAPTER TWO

LITERATURE REVIEW

2.1 Epigenetics

"Epigenetics" was originally introduced by C. H. Waddington in 1942 by combining the words "genetics" and "epigenesis", defining that a phenotype is a result of interactions between genes and their environment (Waddington, 1942). Although at that time the relationship between genes and heredity was still unknown, Waddington suggested that, since the same DNA sequences are available in all cell types, it is only fitting that genes and their interaction with the environment would have an impact on a phenotype (Waddington, 1942). Nevertheless, the definition of epigenetics has evolved through the years. Berger *et al.* (2009) re–defined epigenetics as the process that ensures the propagation of phenotypes through mitosis or meiosis with reversible changes in the genetic sequence. This gave the field of epigenetics a wider scope of study, as it may include the study of alterations in gene expression through post translational modifications, the inheritance of gene expression from mother to daughter cells, and vast environment factors that might have an impact on gene activities.

Epigenetic mechanisms play critical roles and are essential for many cellular processes and normal development as they are able to control gene expression. This can be observed especially in the inactivation of one of the two X–chromosomes in female mammals, in order for females to have the same number of X–chromosome gene products as males (Egger *et al.*, 2004). Various epigenetic modifications have since been identified, which include modifications that have important implication on diseases such as leukaemia, Prader–Willi syndrome and Fragile X syndrome. The most studied epigenetic modifications are DNA methylation and histone modifications, with the former usually associated with gene regulation by transcriptional silencing (Bird, 2002), and the latter, either active or repressive effect on transcription (Shilatifard, 2006). Other modifications include, and not limited to, Polycomb (PcG) and Trithorax (TrxG) group of proteins (Bird, 2007) along with noncoding RNA, for which these different modifications usually work interdependently.

2.1.1 DNA Methylation

2.1.1(a) DNA Methylation and DNA Methyltransferases

Occurring at carbon 5 of cytosines, 5–methylcytosine (5mC) is a result of DNA methylation and is mainly associated with chromatin structure and transcriptional repression. The presence of 5mC was first observed by Hotchkiss (1948) even before it was revealed that DNA was the true genetic information carrier. This was further verified by Wyatt (1951) and Kornberg *et al.* (1959), in which the latter proposed that 5mC might be added to the DNA by post–replicative mechanism. Abundant in vertebrates and at lower levels in invertebrates (Suzuki and Bird, 2008), 5mC usually occurs as a symmetrical mark at CpG dinucleotides (Ziller *et al.*, 2011; Ramsahoye *et al.,* 2000). In mammals, DNA methylation patterns are established during embryonic development and is catalysed by three conserved enzymes, namely, maintenance enzyme DNA methyltransferase 1 (DNMT1), and *de novo* methyltransferases, DNMT3A, DNMT3B and DNMT3L. These four methyltransferases are responsible for the deposition and maintenance of the methyl group and are essential for normal development (Meissner, 2010; Okano *et al.,* 1999).

DNMT1 has been reported to be predominantly expressed and is responsible for the faithful propagation of DNA methylation during cell division. This is achieved by depositing a methyl group on the newly synthesised DNA strand at hemimethylated CpG sites created during replication (Jones and Liang, 2009). The main function of DNMT1 as a maintenance methyltransferase was verified by Arand *et al.* (2012), with the findings that DNMT1 localises at DNA replication sites during S phase. Furthermore, the study also demonstrated that DNMT1 is preferentially active on hemimethylated CpG sites brought on by replication. Studies have shown that in the absence of DNMT1, mice generally undergo early embryonic death, global genome hypomethylation with an increase of hemimethylated CpG sites, as well as an imprints failing to maintain their methylation (Arand *et al.*, 2012; Hirasawa *et al.*, 2008). The study by Bostick *et al.* (2007) furthermore showed that DNMT1 interacts with UHRF1 at hemimethylated sites, where ubiquitine–like containing PHD and RING finger domains 1 (UHRF1) is responsible for the recruitment of DNMT1 at hemimethylated sites. Although DNMT1 has been reported to contain a CXXC–type zinc–finger domain that mediate the binding of unmethylated CpGs, it has been shown that the CXXC domain is able to create an autoinhibitory conformational change which prevents interaction between DNA and the catalytic domain when DNMT1 is bound to unmethylated CpGs. This formation, however, was not found when DNMT1 is bound to hemimethylated CpGs, explaining the preference of DNMT1 for hemimethylated CpG sites over unmethylated CpG sites (Frauer *et al.,* 2011; Song *et al.,* 2011; Song *et al.*, 2012).

In 1989, usage of the prokaryotic cytosine DNA methyltransferase for a homology search led to the discovery of three genes with a shared set of conserved protein motifs. Posfai *et al.* (1989) hypothesised that these protein motifs could encode for novel DNA methyltransferases (Figure 2.1). One of the protein, known as DNMT2, was shown to have the least DNA methyltransferase activity *in vitro* and had detectable effect on neither maintenance nor *de novo* methylation of DNA. However, the other two genes, DNMT3A and DNMT3B, had catalytic activity that had no inclination towards hemimethylated DNA (Okano *et al.*, 1998; Okano *et al.*, 1998). DNMT3A is abundant in differentiated cells whereas DNMT3B is prevalently found in early embryos and are responsible for establishing DNA methylation after implantation (Borgel *et al.*, 2010). Both DNMT3A and DNMT3B are responsible for the methylation of proviral genomes and repetitive elements in embryos and embryonic stem cells (Okano *et al.,* 1999), indicating that DNMT3A and DNMT3B were accountable for the synergistic functions during early development. DNMT3A knockout mice survive birth but die around four weeks of age whereas DNMT3B knockout mice die during gestation. Furthermore, DNMT3A is also pivotal for the establishment of DNA methylation imprints in the germ line (Kato *et al.*, 2007), and along with DNMT3B, work hand in hand with DNMT1 to efficiently maintain DNA methylation during replication (Arand *et al.*, 2012; Jones and Liang, 2009). Despite the absence of a functional catalytic domain, DNMT3L functions as a cofactor for DNMT3A and DNMT3B by stimulating their activity (Ooi *et al.,* 2007; Jia *et al.*,

2007) and is highly expressed in germ cells. DNMT3L is critical in establishing DNA methylation particularly in the male and female germ line, for without DNMT3L, sterility ensues (Barlow and Bartolomei, 2014; Kobayashi *et al.*, 2012; Smallwood *et al.,* 2011).

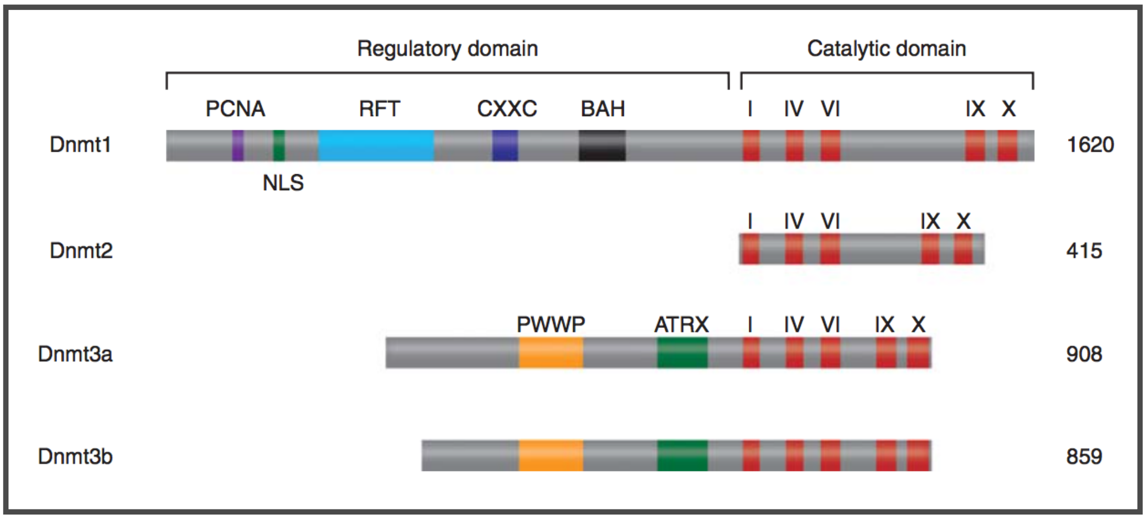

Figure 2.1 : Mammalian DNA methyltransferases. Although the amino–terminal regulatory domains of Dnmt1, Dnmt2, and the Dnmt3 have minimum similarity, their catalytic domains are conserved. PCNA, PCNA–interacting domain; NLS, nuclear localisation signal; RFT, replication foci–targeting domain; CXXC, a cysteine–rich domain implicated in binding DNA sequences containing CpG dinucleotides; BAH, broom–adjacent homology domain implicated in protein–protein interactions; PWWP, a domain containing a highly conserved "proline–tryptophan–tryptophan–proline" motif involved in heterochromatin association; ATRX, and ATRX–related cysteine–rich region containing a C2–C2 zinc finger and an atypical PHD domain implicated in protein–protein interactions (Adapted from Li and Zhang, 2014).

2.1.1(b) Targets and Functions of DNA Methylation

In order to better understand DNA methylation, we first need to look into the targets as wells as functions of DNA methylation in different genomic contexts, as different genomic contexts leads to different transcriptional levels.

It is known that mammalian genomes have low CpG frequency, and is estimated that of the 28 million CpGs in the human genome, only 70 % of the CpGs are methylated (Ehrlich *et al.,* 1982). Some DNA regions have CpGs that occur with a higher CpG density, termed "CpG islands" (CGIs). Prevalent at transcription start sites of housekeeping and tissue–specific genes, CGIs are typically 0.5–2.0 kb long and are mainly unaffected by DNA methylation (Deaton and Bird, 2011). According to Cohen *et al.* (2011), these CGIs are not a result of positive selection for CpGs, but the fact that these CpGs are unmethylated even in the germ line demonstrate that these regions are not likely to undergo CpG loss during evolution. CGIs, especially those that are associated with promoters, are highly conserved between human and mice, implying that these regions are of significant importance (Illingworth *et al.*, 2010).

The bulk of CGIs in somatic cells are unmethylated and are characterised by nucleosome–depleted regions at the transcription start site of active genes (Choi, 2010). These nucleosome–depleted regions are tightly associated with histone proteins and are usually marked with trimethylation of histone H3 at lysine 4 (H3K4me3), which is a mark for active transcription (Mikkelsen *et al.*, 2007). According to Carninci *et al.* (2006), despite the fact that around 50 % of CGIs are associated with known transcription start sites that enhance the binding of transcription factors, most of the CGIs are often depleted of common promoter elements such as TATA–boxes. Furthermore, the other 50 % of CGIs that are remote from known CGI–promoters, often known as ‘orphan’ CGIs, also show similar epigenetic features, such as transcriptional initiation and dynamic expression during development. These orphan CGIs are frequently methylated during development and might be associated with nuanced or different functional roles (Illingworth *et al.*, 2010).

DNA methyltransferases need to be actively and constantly excluded at CGIs to maintain a hypomethylated state. Studies have shown that these CGIs remain unmethylated with the help of transcription factor bindings, such as SP1 elements, for with the depletion of transcription factor bindings, methylation occurs (Lienert *et al.,* 2011; Stadler *et al.*, 2011). Additionally, CGIs are usually bound by CXXC finger protein 1 that recruits H3K4 methyltransferases to maintain transcription in an unmethylated state (Thomson *et al.*, 2010). It has also been reported that the binding of MLL family H3K4 methyltransferases, which is a positive global regulator of gene transcription, protects promoters of development genes from methylation through their CXXC domain (Erfurth *et al.*, 2008). The *de novo* methyltransferase 3 (DNMT3) family enzyme also plays a role in maintaining an unmethylated state. DNMT3 family enzyme has an ATRX–DNMT3–DNMT3L (ADD) domain that is responsible for the recognition of unmodified H3 and is inhibited by H3K4 methylation (Otani *et al.*, 2009). Moreover, the histone variation H2A.Z has been reported by Conerly *et al.* (2010) to be enriched at unmethylated, active promoters. H2A.Z–containing nucleosomes is able to spread to neighbouring regions when azacytidine, which is a DNA methylation inhibitor, was used. This finding indicates that histone variants and epigenetic modifications that are associated with transcription can be affected by the surrounding DNA methylation (Yang *et al.*, 2012).

Although CGI–promoters are usually associated with expressed genes, some repressed genes can have methylated promoter CGIs as well. These methylated promoter CGIs can be a result of various mechanisms, such as mediation by Polycomb proteins, which usually result in stable silencing of gene expression (Taberlay *et al.,* 2011; Mohn *et al.*, 2008). Regulation of gene expression by methylation at CGIs is especially important for the establishment of imprints (Choi *et al.*, 2005), genes located on the inactive X chromosome, and genes expressed during development and differentiation, especially those that are exclusively expressed in germ cells (Meissner *et al.*, 2008; Fouse *et al.*, 2008). The stable suppression of CGIs by DNA methylation can last over a 100–year lifespan and has no impact on the existence of other CGIs as deamination events would not be passed on to subsequent generations.

There are also genes with non–CGI transcription start site. These genes, such as those that are expressed in primordial germ cells, are unmethylated at the transcription start site, whereas genes that are expressed in tissue–specific genes often show methylation in sperm but not in oocytes (Farthing *et al.,* 2008). For instance, OCT4 and NANOG are expressed in stem cells but not in differentiated or somatic cells as both of these marks are essential for the maintenance of stem cell state.

The role of CGIs in regulating gene expression is still being studied. For now, methylation of CGIs is understood to impair transcription factor binding, recruit repressive factors, such as methyl–binding proteins, as well as the stable silencing of gene expression. Nonetheless, CGIs associated with gene promoters are hardly ever methylated. Studies are still being conducted to determine DNA methylation impact on CGIs during the regulation on gene expression.

The majority of CpG sites in the mammalian genome are methylated, hence these methylated CpG sites should be distributed along genes as well. Gene body has been defined as the region of the gene right after the first exon (Brenet *et al.*, 2011). Most of the gene bodies are CpG–poor and are highly methylated, which is usually associated with higher levels of gene expression in dividing cells (Aran *et al.*, 2011). However, gene body methylation is not associated with increased gene expression if it occurs in non dividing or slowly dividing cells (Aran *et al.* 2011; Xie *et al.*, 2012). Extensive studies have shown positive correlations between gene body methylation and active transcription. These studies include gene body methylation on the X chromosome (Hellman and Chess, 2007), as well as gene body methylation studies in plant and animal genomes as determined by shotgun bisulphite sequencing (Lister *et al.*, 2009; Feng *et al.*, 2010).

Although gene body CGIs are highly methylated, this state does not impair transcription elongation. Gene body CGIs that are methylated are usually marked by H3K9me3 and are bound by methyl–CpG–binding protein 2 (MECP2). When these marks are associated with transcription start sites, transcription repression occurs, whereas the opposite is observed when these marks are associated with gene bodies.

This situation in itself leads to a contradiction, as promoter methylation is negatively correlated with gene expression, whereas gene body methylation is directly correlated with expression (Jones, 1999).

Almost half of the mammalian genome is made up of endogenous transposable elements. The three major classes of endogenous transposable elements are long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs). Some of the LINE and LTR elements have strong promoters which permits transcription and affect the integrity of the genome if not repressed. Therefore, they are usually constitutively hypermethylated in order to repress their expression. Although DNMT1 is responsible for the maintenance of methylation, it has been reported that DNMT1 alone is not sufficient to stably repress these sites and assistance from DNMT3 is required (Liang *et al.*, 2002). Furthermore, studies have shown that H3K9 methyltransferase SET domain bifurcated 1 (SETDB1), which targets specific sequences through zinc finger protein, fosters persistent DNA methylation at LTR retrotransposons by acting upstream of DNMT recruitment (Leung *et al.*, 2014; Rowe *et al.*, 2010; Matsui *et al.*, 2010). This means that LTR retrotransposons are primarily repressed by histone methyltransferase activity of SETDB1, whereas DNA methylation only functions as a secondary repressor (Karimi *et al.*, 2011).

It was initially thought that gene body methylation functions solely to silence repetitive DNA elements, such as LINE1 elements, retroviruses, and *Alu* elements. However, recent whole–genome studies have shown that there might be other functions for gene body methylation. Laurent *et al.* (2010) demonstrated that exons are more methylated than introns, and the transition in the methylation levels usually occur at exon–intron boundaries, suggesting a role for gene body methylation in regulating splicing. In fact, Schwartz *et al.* (2009) also reported that exons showed increased nucleosome occupancy levels as compared to introns, for which DNA methylation preferentially occurs at nucleosomes (Chodavarapu *et al.*, 2010).

Enhancers are mostly CpG–poor and tend to have variable methylation. Studies have found that the enhancers in mouse genome are neither 100 % methylated nor unmethylated, hence defining them as 'low–methylated regions (LMRs)' (Stadler *et al.*, 2011). This might suggest that the LMRs are in a dynamic state and that the methylation levels might change with time, in which case, methylation of enhancers could result in reduced enhancer activity. Insulators, elements that block the interaction between an enhancer and promoter, are usually bound by CCCTC–binding factor (CTCF) protein. A well studied CTCF containing insulator is the imprinted IGF2–H19 locus, where CTCF plays an important role in controlling the enhancer–promoter interactions (Lee *et al.*, 2010). Although it was thought that CTCF might play an important role in the methylation of insulators, recent global studies in mouse embryonic stem cells as well as differentiated cells suggest that CTCF binding within CpG–poor region is generally not affected by the methylation status of the binding sites, but rather, the binding of CTCF itself initiates local demethylation (Stadler *et al.*, 2011; Singh *et al.*, 2012). Since lesser studies on the methylation at enhancers and insulators are done, therefore the methylation mechanism at these regions are less understood. Further studies are needed to determine how DNA methylation of each of these sites regulate gene expression.

2.1.2 Histone Modifications

Chromatin of eukaryotic cells consists of DNA that is packaged into highly conserved basic proteins called histone proteins. Four different histone proteins are arranged into an octameric structure, each with highly similar structural motif. These four histone proteins encompass a histone fold domain with a common structure and an N–terminal histone tail which is variable in length (Arents and Moudrianakis, 1995). Histone fold domain is globular whereas the histone N–terminal tails, and to a lesser extent, histone C–terminal tails, are usually unstructured, with most of them protruding out from the nucleosome core particle. These histone proteins can be modified at many sites, to which more than 60 different modifications on histone tails have been detected according to Kouzarides (2007). These modifications, termed post translational modifications, regulate chromatin structure as well as recruit remodelling enzymes to reposition nucleosomes (Bannister and Kouzarides, 2011). Histones are modified by various post translational modifiers, such as phosphorylation, acetylation, methylation, sumoylation and ubiquitination.

Occurring at lysine residues in the amino–terminal tails of the histones, acetylation is able to neutralise positive charges of the histones, thereby decreasing their affinity towards DNA (Bhaumik *et al.*, 2007). Histone acetylation may have an impact in the changes of nucleosomal conformation, leading to chromatin architectural change which cause chromosomal domains to be more accessible. This dynamic mechanism is regulated by histone acetyltransferases (HATs) and histone deacetylases (HDACs). These two opposing enzymes are responsible for the acetylation and deacetylation of histones, which leads to transcription activation and transcription repression, respectively. A permissive chromatin structure arises when SWI2/SNF2 family, a nucleosome remodelling complex, is recruited and that leads to weakening of the histone–DNA interaction (Awad and Hassan, 2008; Fry and Peterson, 2001). On the other hand, a repressive chromatin structure arises when deacetylation of histones contribute to the condensation of the nucleosomal fibres, and along with other processes such as histone methylation and DNA methylation, lead to transcription repression (Kimura *et al.,* 2005). Not only is histone acetylation important for transcription regulation, it also has an impact on other processes such as chromatin replication, DNA repair and site–specific recombination (Ge *et al.*, 2013; Bird *et al.*, 2002; Roth *et al.*, 2001).

The dynamic phosphorylation of histones often occurs at serine, threonine, and tyrosine residues during cell division (Xhemalce *et al.*, 2011). Phosphorylation of histones modifies the overall charge of the protein, leading to changes in the overall structure as well as the function of the local chromatin environment. Phosphorylation has been shown to provide a binding platform for various factors, such as proteins that are essential for chromatin remodelling that leads to transcription regulation (Sawicka and Seiser, 2012; Rossetto *et al.*, 2012). It is regulated by two enzymes, namely kinases and phosphatases. These two enzymes add and remove phosphate groups, and in doing so, alter charges to the histone that changes the chromatin structure. Although histone phosphorylation generally occurs at the N–terminal tails, some of the modifications do occur within the core region, such as the H3Y41 phosphorylation (Dawson *et al.,* 2009). It has been reported that kinases and phosphatases often work together to modulate a number of cellular processes. This include DNA repair, mitosis and apoptosis (Sharma *et al.,* 2012; Medema and Lindqvist, 2011; Cook *et al.*, 2009). For example, Hsu *et al.* (2000) reported that mitotic kinase IPl1 activity is neutralised by the phosphatase activity of Glc7, while WSTF, a novel tyrosine kinase, work together with EYA1/3 phosphatases in order to repair DNA double–stranded breaks during DNA damage response (Cook *et al.*, 2009; Xiao *et al.*, 2009).

Histone methylations, a more stable histone modifications as compared to other modifications, are maintained through cell division and generally occurs at lysine and arginine residues. Unlike acetylation and phosphorylation which alter the charges of the histones, histone methylation only affects the basicity and hydrophobicity of the histones (Migliori *et al.,*2010). Histone methylation acts both as a repressor and an activator, depending on the site of methylation. For instance, methylation of H3K4, H3K36 and H3K79 (Pekowska *et al.*, 2011; Xu *et al.*, 2008; Steger *et al.*, 2008) is often correlated with transcriptional activation, whereas methylation of H3K9, H3K27 and H4K20 is responsible for transcriptional repression (Barski *et al.,* 2007; Stewart *et al.*, 2005; Boros *et al.*, 2014; Kouzarides, 2007; Brinkman *et al.*, 2006). Histone lysine methylation has been reported to occur in combination. For instance, H3K27me3 and H3K9me3 work together to maintain heterochromatin protein at chromatin, while H3K9me2, in combination with H3K27me3 and H4K20me1, was seen to be enriched during X–chromosome inactivation (Boros *et al.,* 2014; Escamilla–Del–Arenal *et al.*, 2011; Sims *et al.*, 2006). Furthermore, opposing marks can also co–exist, such as H3K4me3 and H3K27me3. This coordinated positioning of active and inactive marks assists in transcriptional competence by maintaining appropriate gene expression during development (Voigt *et al.*, 2013; Vastenhouw and Schier, 2012; Sachs *et al.*, 2013). This predicament enables H3K4me3 to activate lineage–regulatory genes during differentiation while H3K27me3 represses these genes during pluripotency.

There are also other histone modifications that are important in the control of gene expression. This includes ubiquitination, sumoylation, ADP–ribosylation, proline isomerization and histone tail clipping (Bannister *et al.*, 2011). Ubiquitin is a 76–amino acid protein that is found in almost all tissues and is highly conserved in eukaryotic organisms. Ubiquitination occurs with the help of three enzymes, namely, E1–activating, E2–conjugating and E3–ligating enzymes. A ubiquitin molecule is formed through the activity of these three enzymes which determine the target and type of ubiquitination. Ubiquitination has been associated with various processes including protein degradation, DNA repair, cell–cycle control, protein interactions and transcription (Mukhopadhyay and Riezman, 2007; Jentsch, 2011; Brown and Jackson, 2015; Jason *et al.*, 2002; Geng *et al.*, 2012; Ndoja *et al.*, 2014). Although ubiquitination is responsible for a wide range of modification, it is a very dynamic modification as it is reversible through de–ubiquitin enzymes, thus leading to both gene activity and silencing (Reyes–Turcu *et al.*, 2009). According to Weake and Workman (2008), the incorporation of a ubiquitin molecule to a protein involves the three ubiquitin enzymes whereas deubiquitination is achieved by ubiquitin–specific proteases. In higher eukaryotes, the most ubiquitinated conjugates are H2A and H2B, for which H2A is usually mono–ubiquitinated at lysine 119 and H2B at lysine 120 (Osley, 2006). H2Aub1 generally plays a role in gene silencing whereas H2Bub1 is responsible for the regulation of transcriptional initiation and elongation (Aranda *et al.*, 2015, Kim and Sung, 2014; Fuchs *et al.*, 2014; Fleming *et al.*, 2008). A member of ubiquitin–like protein family, the small ubiquitin–like modifier (SUMO) is another modification available at lysine residues. Sumoylation occurs via the action of E1–activating, E2–conjugating, and E3–ligating enzymes, similar to ubiquitination (Gareau and Lima, 2010). Histone sumoylation has been reported to play a role in transcription repression, although more work is still needed to elucidate the role of sumoylation on chromatin (Lyst and Stancheva, 2007; Yang and Sharrocks, 2004; Perdomo *et al.*, 2012).

2.2 DNA Methylation Changes During Mammalian Development

2.2.1 Preimplantation Development

DNA methylation undergoes drastic changes during early embryonic development in mammals. This change in methylation levels is pivotal for the establishment of pluripotency, through which, global demethylation and a lineage–specific methylome is established. Initial studies in mice using immunofluorescence and restriction enzymes demonstrated that global DNA demethylation occurs after fertilisation up to blastocyst stage, for which DNA methylation is re–established after implantation (Kobayashi *et* al., 2012; Monk *et al.*, 1987; Santos *et al.*, 2002; Figure 2.2). Reprogramming of DNA methylation occurs in the inner cell mass of the pre–implantation embryo (E3.5), which coinsides with the establishment of pluripotent cells, forming embryonic stem (ES) cells when cultured *in vitro* (Smith *et al.*, 2012).

Although DNA methylation knockout ES cells are viable and able to self–renew (Tsumura *et al.*, 2006), these knockout ES cells go through apoptosis during *in vitro* differentiation and Dnmt1–null embryos die around E8.5, indicating that DNA methylation is essential for the establishment and maintenance of cell differentiation in embryonic lineages (Schmidt *et al.*, 2012; Takebayashi *et al.*, 2007; Okano *et al.,* 1999; Li *et al.*, 1992). Hence, demethylation in the preimplantation embryo leads to an epigenetic state that is instrumental in embryonic lineage–specification, and along with *de novo* methylation, establish specific cellular identity.

Reprogramming of DNA methylation during preimplantation occurs through multiple mechanisms, resulting in the rapid loss of 5–methylcytosine (5mC) in the zygote of paternal DNA, and a much slower, replication dependent demethylation in the maternal DNA (Smith *et al.*, 2012; Inoue and Zhang*,* 2011; Wossidlo *et al.*, 2010; Santos *et al.*, 2002). Many studies have been done on active DNA demethylation and it has been shown that 5mC can be oxidised to 5–hydroxymethylcytosine (5hmC), 5–formylcytosine (5fC), and 5–carboxylcytosine (5caC) by ten–eleven translocation methyl cytosine dioxyenases (Tet dioxygenase) (Tahiliani *et al.*, 2009; He *et al.*, 2011; Ito *et al.*, 2011). This family of protein is comprise of three members, namely, TET1 which is highly expressed in ES cells, TET2 which is highly expressed in ES cells and hematopoietic cells, and TET3 which is mostly present in oocytes. Knockouts for the *Tet* genes have been shown to lead to developmental delay and failures, subfertility, causes diseases and severe hematopoietic defects, as well as impair differentiation of ES cells (Yamaguchi *et al.*, 2013a; Dawlaty *et al.*, 2014; Dawlaty *et al.*, 2013; Ko *et al.*, 2011; Gu *et al.*, 2011b; Abdel–Wahab, 2009). As for the maternal DNA, whereby replication dependent demethylation occurs, demethylation occurs most probably as a result of lack of maintenance activity (Saitou *et al.*, 2012).

Overcoming the difficulty in working with small numbers of cells through the years, several groups have applied genome–wide bisulfite sequencing to quantify the dynamics of DNA methylation of early cleavage–stage mice embryos (Guo *et al.*, 2013; Guo *et al.*, 2014; Smith *et al.*, 2014; Kobayashi *et al.*, 2012; Smith *et al.*, 2012). Through these studies, it was found that oocytes have relatively hypomethylated genome compared to sperm. Furthermore, methylation levels is also reduced to a minimal level in preimplantation blastocyst upon fertilisation. Although bisulfite sequencing detects both 5mC and 5hmC, these studies prove that a major epigenetic reprogramming event occurs after fertilisation.

After implantation of the embryo, *de novo* remethylation, which is directed by DNMT3A and DNMT3B, is initiated. A number of mechanisms have been proposed as to how DNA methylation is established during development. Some of the suggested mechanisms include the establishment of DNA methylation through the interactions of various modifications on the chromatin as well as transcription read–through, protection by the DNA–binding factors, and guidance by small RNAs (Baubec *et al.*, 2015; Feldmann *et al.*, 2013; Stadler *et al.*, 2011; Lienert *et al.*, 2011). These mechanisms most likely work interdependently, in a locus and context specific manner. DNA methylation is actively recruited by CpG–poor regions as well as to a small subset of CpG islands. Nevertheless, there are some sequences that are able to escape demethylation after fertilisation through the protective roles of *Zfp57* and

PGC7/Stella (Nakamura *et al.*, 2012; Nakamura *et al.*, 2007; Li *et al.*, 2008). These sequences include imprinted loci and repeats such as intracisternal A particles, L1Md_A elements, and LTR ERV V1 elements (Hackett *et al.*, 2013; Smith *et al.*, 2012; Chan *et al.*, 2012; Guibert *et al.*, 2012). Furthermore, studies have also shown that there are CpG islands that manage to evade DNA demethylation in preimplantation embryos (Kobayashi *et al.*, 2012; Smith *et al.*, 2012). These oocyte–derived DNA methylated CpG islands typically do not maintain their maternal–specific DNA methylation after implantation as they are demethylated after implantation, or the paternal alleles undergo methylation at the time of implantation, resulting in similar levels of methylation in both paternal and maternal DNA (Smith *et al.*, 2012; Proudhon *et al.*, 2012). This phenomenon demonstrates that gametic DNA methylation almost never translates into imprinted methylation.

2.2.2 Gametes

In general, embryos at blastocyst stage have low levels of DNA methylation. This hypomethylated state changes to a hypermethylated state upon implantation through *de novo* methylation. Mice primordial germ cells (PGCs) are specified from the posterior proximal epiblast cells at embryonic day (E) 6.25, and during this point of time, they express specific markers such as *Prdm14, Prdm1/Blimp1,* as well as *Dppa3/Stella*. A cluster of around 40 cells is formed by PGCs at around E7.5 will subsequently migrate and colonise the genital ridges by E10.5, where they continue to proliferate until E13.5. Derived from somatic cells, these PGCs possess a somatic epigenetic profile that must be reprogrammed in order for the germ line–specific genes to be primed. The reprogramming process require modifications of histones,

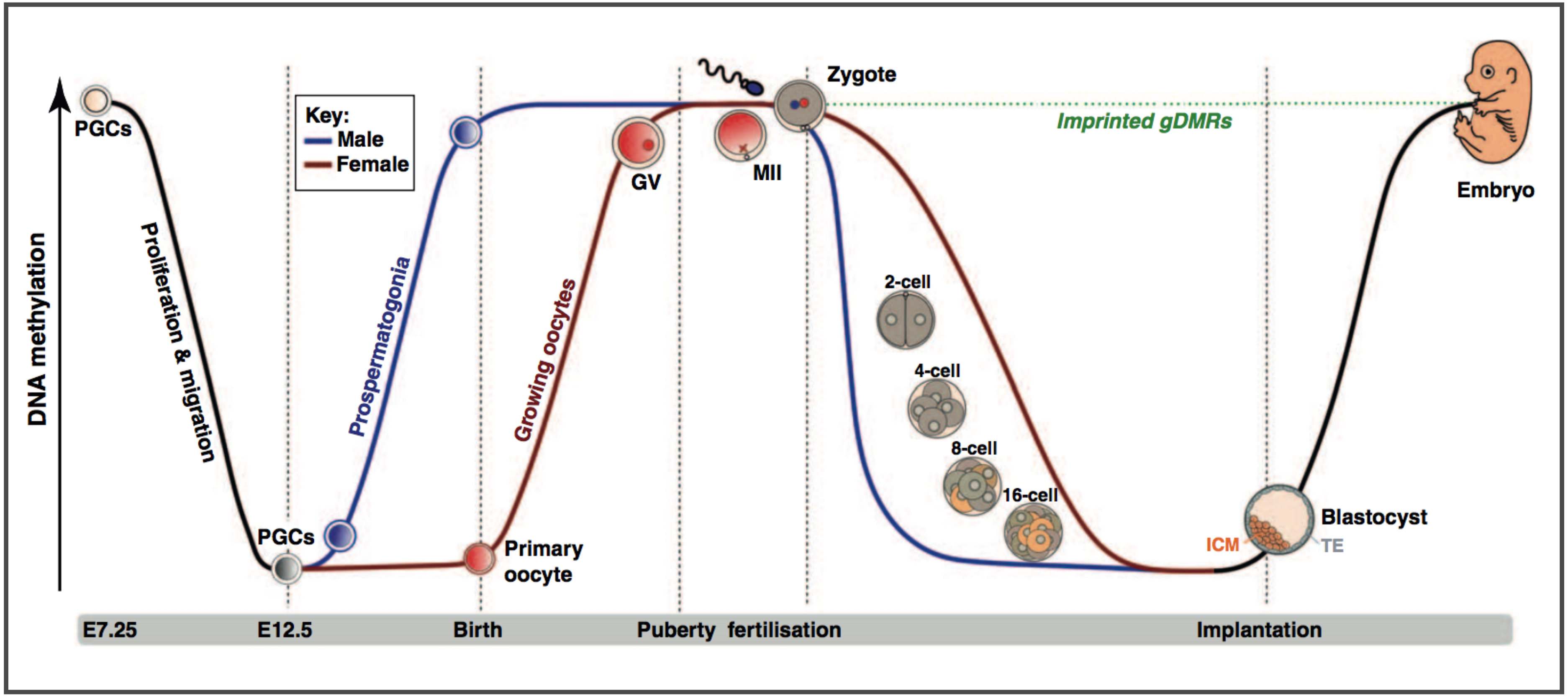

Figure 2.2 : DNA methylation changes during developmental epigenetic reprogramming. PGCs lose global DNA methylation from E7.25 to E12.5 and DNA methylation is restored in a sex–dependent manner. After fertilisation, paternal genome undergoes rapid 5mC erasure by active mechanism whereas maternal genome undergoes DNA replication dependent slow erasure to reach a low point in the blastocyst. These post–fertilisation demethylation events do not include imprinted genes (green dotted line), which results in parental–allele specific methylation, and consequently parental–allele–specific expression of associated imprinted genes. After implantation, DNA methylation levels increase to establish new methylation landscapes which is associated with cellular differentiation. Adapted from Smallwood and Kelsey (2012).

repression of somatic genes, as well as global DNA demethylation (Hackett *et al.*, 2012b; Saitou *et al.*, 2012). The genome–wide demethylation in PGCs is demonstrated by a global loss of 5mC immunostaining signal starting from E8.0 (Yamaguchi *et al.*, 2013; Seki *et al.*, 2005). Between E8.5 and E9.5, DNA demethylation mainly occurs on promoters, CpG islands, exons, introns and intergenic regions (Deaton and Bird, 2011; Jones, 2012; Seisenberger *et al.*, 2012). Only after PGCs enter the genital ridges at around E10.5 does the imprinted genes start to be erased (Guibert *et al.*, 2012; Hackett *et al.*, 2012a). Studies by bisulfite sequencing and immunoprecipitation had demonstrated that PGCs at E13.5 have a demethylated genome compared to somatic cells (Guibert *et al.*, 2012; Seisenberger *et al.*, 2012). This genome–wide demethylation includes the erasure of methylation at imprinted loci, gene bodies, intergenic regions, transposable elements as well as all methylated CGI promoters accessible in early embryos, demonstrating that demethylation of PGCs is more comprehensive than in preimplantation embryos. These sequential demethylation of 5mCs may point to the existence of multiple demethylation mechanisms in PGCs to gain full demethylation (Hackett *et al.*, 2012b), and have been reported to be important to restore a germ–line epigenetic state as well as to erase any erroneous epimutations that might have occurred (Guibert *et al.*, 2012; Seisenberger *et al.*, 2012).

Mechanisms of DNA demethylation in PGCs have been greatly debated. Recent time–course analysis showed that DNA methylation in both PGCs and somatic genes gradually decreased beginning from E8.5, with rapid demethylation ensuing between E10.5 and E12.5 (Guibert *et al.*, 2012). Hence, it has been proposed