Role of Germ Cell-Specific Epigenetic Modifications in Gene Expression in the Mouse

September 2013

Rieko IKEDA

Role of Germ Cell-Specific Epigenetic Modifications in Gene Expression in the Mouse

A Dissertation Submitted to
the Graduate School of Life and Environmental Sciences,
the University of Tsukuba
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Science

Rieko IKEDA

Table of Contents

Abstract

To understand the epigenetic regulation required for germ cell-specific gene expression, I analyzed DNA methylation profiles of developing germ cells using a microarray-based assay adapted for a small number of cells. This microarray-based method provides the genome-wide assay of DNA methylation using only a subnanogram quantity of genomic DNA. I obtained DNA methylation profiles for mouse primordial germ cells (PGC) of different developmental stages and for stem cells derived from embryos or germ cells. Cluster analysis of the data revealed that each cell type possesses its own characteristic DNA methylation profile, enabling classification of the cell types. This classification is generally consistent with that based on gene expression profiles except for PGCs, whose genome is globally hypomethylated. Among the differentially methylated sites thus identified, I focused on a group of genomic sequences hypomethylated specifically in germline cells as candidate regions involved in the epigenetic regulation of germline gene expression. These hypomethylated sequences tend to be clustered, forming large (10 kb to ~9 Mb) genomic domains particularly on the X chromosome of male germ cells. Most of these hypomethylated regions designated here as *L*arge Hyp*o*methylated *D*omain (LoD) correspond to segmentally duplicated regions that contain gene families showing germ cell- or testis-specific expression, including cancer testis antigen genes. I found an inverse correlation between DNA methylation level and expression of genes in these domains. Most LoDs appear to be enriched with H3 lysine 9 dimethylation (H3K9me2), usually regarded as a repressive histone modification, although some LoD genes can be expressed in male germ cells. It thus appears that such a

unique epigenomic state associated with the LoDs may constitute a basis for the specific expression of genes contained in these genomic domains.

List of abbreviations

Abbreviation	Meaning
C	cytosine
T	thymine
CGI	CpG island
PGCs	primordial germ cells
E	embryonic day
P	postnatal day
e.g.	(exempli gratia) for example
i.e.	(id eat) that is
LoD	large hypomethylated domain
CTA	cancer testis antigen
R	correlation coefficient
bp	base pair
IR	inverted repeats
LOCKs	Large organized chromatin K9 modifications
ES	Embryonic Stem
EG	Embryonic Germ
GS	Germline Stem

1. Introduction

Embryonic development of multicellular organisms is initiated after fertilization; a totipotent fertilized egg will go through cleavage stage, increase number of cells and will be differentiated into various cell types that constitute functional structures and eventually a whole organism. During this developmental process, various parts of genome in each cell are activated or inactivated to drive the developmental gene expression underlying the morphological changes. Although regulations of gene expression can be achieved by several layers of mechanisms, importance of "epigenetic" regulations is increasingly evident.

Epigenetics is an academic discipline for the study of mechanisms that influence gene expression without changing the DNA sequence of genome [1]. Epigenetic regulations are involved in genomic modifications acquired during development such as DNA methylation or various modifications (acetylation, methylation, phosphorylation, etc.) of histone tail. Genomic modification by DNA methylation is not found in all organisms, but it is known to be extremely important in regulations of gene expression in mammals including human and in some plants [2].

In mammalian genome, addition of methyl groups usually occurs at cytosine of CpG dinucleotide. As methylated cytosine is prone to mutation, Cytosine (C) tends to be converted to thymine (T), resulting in underrepresentation of CpG dinucleotide in

mammalian genome. However, there are regions that contain a high frequency of CpG sequences within mammalian genome and these clusters of CpGs are designated as CpG islands (CGIs). Length of CGIs ranges from 300 bp to 3000 bp with GC percentage greater than ~50% and with an observed/expected CpG ratio of > 0.6. In mammalian genomes, there are approximately 15,000 CGIs, and these CGIs have been found in or near to promoters of mammalian genes [3]. About 40% of mammalian gene promoters contain CGIs. CGIs are typically free of DNA methylation in most cell types. However, hypermethylation of CGIs in promoters of tumor suppressor genes were detected in cancer cells [4]. In female mammals, one of two X-chromosomes is inactivated for gene dosage compensation. CGIs of X-linked gene promoters on inactive X chromosome are also known to be hypermethylated, resulting in silencing of gene expression [5]. These studies therefore suggest that DNA methylation is important for repression of gene expression. In mice, there are at least three distinct DNA methyltransferases exist; *Dnmt1* is required for maintenance of DNA methylation, while *Dnmt3a* and *Dnmt3b* are used for *de novo* DNA methylation [6, 7]. Since mice with disrupted *Dnmt* genes die during development, DNA methylation is thought to be essential for regulation of developmental gene expression and the nuclear organization of chromatin [2, 8-11].

It is thus obvious that DNA methylation-dependent control is responsible for dynamic changes in gene expression during development. However, genomic information about which sequences undergo methylation or demethylation and in what order, or

relationships of these methylation changes with gene expression have been poorly understood. It has been known that genomic DNA methylation changes in developmentally regulated manner, but these studies investigated only a limited number of genes [12, 13], falling short of demonstrating global pictures of developmental changes in DNA methylation. Immunohistochemistry with anti-5-methyl cytosine antibody has been used to observe "global DNA methylation" pattern in developing cells [14]. This method is useful to roughly demonstrate DNA methylation pattern in individual cells. However, the staining pattern likely to reflect global methylation pattern of heterochromatic repetitive sequences and the method cannot detect DNA methylation patterns of individual gene sequences.

In this study, I aim to investigate dynamic changes in epigenetic states of early embryonic cells and developing germ cells as well as stem cell lines derived from these cells. Epigenetic state of cells in embryos is thought to be changed every moment. Moreover, in mammalian development, large scale changes in epigenetic states called "epigenetic reprogramming"[12, 15] take place at least twice; once just after fertilization [16] and once during specification of primordial germ cells (PGCs) [17] (Fig. 1). At both stages, genome-wide changes in epigenetic modifications such as DNA methylation and histone modifications should occur followed by re-establishment of cell-type-specific epigenetic status. After fertilization, demethylation of genomic DNA initiates and continues until the morula stage, by then the global DNA methylation level becomes quite

low[16]. *De novo* methylation is thought to occur at some stage after implantation, although precise timing of such an epigenetic remodeling is still unknown.

Major reprogramming also takes place in PGCs. In mice, PGCs are first identified as a cell population of about 45 cells at the base of allantois at embryonic day (E)7.25 [18-20]. In developing PGCs, epigenetic reprogramming such as reactivation of the inactive X chromosome or erasure of genomic imprints take place [21, 22], because PGCs need to remove such an epigenetic "parental legacy" in the genome before transmitting their genome to the next generation (Fig. 1). As a result of the epigenetic reprogramming, genomes of PGCs will become extensively hypomethylated. This DNA demethylation is known to be initiated around the time when PGCs enter genital ridges, i.e. embryonic day (E) 11.5, and completed between E11.5 and 13.5 [12, 15]. However, some other studies suggested that DNA demethylation may start earlier than E11.5, along with “epigenetic reprogramming” initiated early in PGC development [17, 22, 23].

As just described, early embryonic cells and germ cells possess vital biological functions to reprogram their epigenomic status. Despite the evident significance of reprogramming events in this cell lineage, the precise timing and kinetics of epigenetic modifications are still largely unknown. While unbiased and genome-wide studies of DNA methylation have recently been carried out for cultured cells, attempts to delineate DNA methylation changes during mammalian development have been hampered, at least partly, due to technical reasons. For developmental epigenetic analyses, materials can often be

very limited in quantity, precluding conventional analytical techniques. Amount of genomic DNA in one diploid mouse cell is approximately 7 picogram. Number of cells that constitute early embryo is known to be small; e.g. single blastocyst embryo comprises about 60 cells, from which only 0.4 nanogram of DNA can be isolated. As shown in Fig. 2, there have been several distinct methods for DNA methylation analysis. However, for genome-wide analysis of DNA methylation, conventional techniques require microgram level of genomic DNA. Therefore I devised an experimental method that allow global analysis of the DNA methylation status using only a subnanogram of genomic DNA. In this study, I used a proven method of DNA methylation analysis called the HELP (*Hpa*II tiny fragment enrichment by ligation-mediated PCR) assay [24-26]. Because this method uses linker-mediated PCR, it can be adapted for the small-scale analysis of developing germ cells. Oda et al.[26] developed an improved version of the method, nanoHELP. Here, I have fine-tuned the protocol further. This modified nanoHELP method provides a global analysis of the DNA methylation status of CCGG sites using only a subnanogram (≥0.5 ng) quantity of genomic DNA.

Another point of DNA methylation study would be which regions of the genome should be investigated. Traditionally, CGIs and gene promoter regions have been the main targets in most DNA methylation studies [27]. However, the importance of DNA methylation in genomic regions outside the promoters is becoming increasingly apparent [2, 11]. For example, as shown in Fig. 3, differential DNA methylation of some nongenic

sequences located in intergenic regions appear to be correlated with expression of nearby genes [11]. While CGIs are almost always unmethylated, differential methylation in regions close to CGIs have been reported [28]. These "CGI shores" show differential methylation in tissue- or cancer-specific manners. Promoters of actively transcribed genes are normally DNA hypomethylated, but it is unexpectedly found that bodies of active genes tend to be hypermethylated compared with those of inactive genes [2]. These findings suggest the importance of DNA methylation studies in regions outside the promoters/ CGIs. Although recent research has advanced my understanding of the PGC epigenome [23, 29-31], further studies are still required to gain more detailed information on epigenomic features of germline cells and their involvement in defining germ cell-specific gene expression.

Therefore I used a custom-made genomic microarray that can assay DNA methylation status of intergenic regions as well as the promoters and gene bodies of known genes. This microarray may provide novel information about previously unexplored but potentially informative parts of epigenome. The custom-made genomic microarray used in this study is unusual in that the CCGG sites of the intergenic regions as well as the promoters and gene bodies of the RefSeq genes could be tested. This HELP microarray may provide new information about previously unexplored parts of the germ cell epigenome. I applied this method to analyze DNA methylation in the mouse X chromosome. I reasoned that epigenomic features specific to germ cells could be found

by focusing on the X chromosome, because the X chromosome carries many germ cell-expressed genes [32, 33] and undergoes major epigenetic changes (e.g. X chromosome reactivation) during germ cell development [22].

In this thesis, I describe details of newly established method of DNA methylation analysis and applications of the method for the analysis of DNA methylation in the genome of PGCs and stem cell lines derived from early embryos and germ cells. Through these analyses, I found for the first time a group of sequences that are specifically hypomethylated on the X chromosome of male germline cells. These sequences form relatively large genomic domains that harbor gene families displaying specific expression in germ cells. These regions are termed here as large hypomethylated domains (LoDs). LoDs have not been detected in previous studies, including recent whole-genome bisulfite sequence analyses [31] probably because mapping of bisulfite-converted short sequence reads onto locally duplicated regions such as LoDs is technically challenging. By contrast, the experimental design of the HELP assay, which involved removal of the potentially confounding effect of copy number difference [24] and inclusion of a probe design that selects unique sequences for hybridization, was effective in finding LoDs.

Interestingly, LoDs contain many genes with homologies to human cancer testis antigen (CTA) genes. CTA genes are normally expressed only in the germline, and are also expressed in some tumor cell types [34]. The results presented in this study may shed light

on the epigenetic basis for the germline gene expression program and its relationship with oncogenesis.

2. Results

2.1. DNA methylation analysis with subnanogram amounts of genomic DNA

For epigenomic analyses, materials can be very limited in quantity, precluding conventional analytical techniques. One of my goals is to describe comprehensively the epigenomic changes during the development of early embryos and germ cells in the mouse. Toward this goal, I use the HELP assay [24], a proven, microarray-based method for the analysis of DNA methylation [24]. The original HELP protocol requires 10 μg of genomic DNA as the starting material [24], but Oda et al. [26] established an improved version of the method, nanoHELP that is adapted to accommodate a limited amount of starting DNA. Here I have further fine-tuned the protocol for the analysis of 0.5–2 ng of starting material. The detail of this method is described in the Methods section, and the flow of the data analysis is presented in Fig. 4.

The HELP assay is a microarray-based method that detects the subset of unmethylated *Hpa*II fragments in the genome, with the corresponding, methylation-insensitive *Msp*I representations serving as a control. The M-value, an index of the methylation level, is calculated as log2(*Hpa*II signal/*Msp*I signal) as described in Section 2: unmethylated segment has a value of $\geqq 0$ and methylated segment has a negative value

of <0. As shown in Fig. 5A–D, the modified nanoHELP generated good correlations between the M-value and the data obtained by the original protocol. To validate these results, bisulfite pyrosequencing analysis of six CpG sites was performed as described [24]. The results showed that the modified nanoHELP assay could generate reliable data (Fig. 5E). I have exported the data as a custom track for the UCSC Genome Browser to present the methylation data associated with genomic annotations (Fig. 6). The results appear to be reproducible for the different samples and can detect differentially methylated regions specific to particular samples.

2.2. Custom HELP microarray used in this study

The number of restriction sites for *Hpa*II, 5'-CCGG, in the mouse genome is 1,588,546, covering ~7.5% of the total CpG dinucleotides in the genome (Table 1A). The CCGG sites are almost evenly distributed over the mouse genome and do not show an apparent bias to a particular genomic context. Thus, the use of CCGG sites is suitable for obtaining a chromosome-wide view of CpG methylation profiles.

In this experiment, I designed a custom microarray harboring 382,018 oligoprobes. I first selected *Hpa*II or *Msp*I fragments (designated here as CCGG segments) with a size range of 200 bp to 2000 bp, mostly from mouse chromosome 7 and X, and designed 10 probes of 50 nucleotides long per each CCGG segment. The custom

microarray can detect 22,128 CCGG segments on chromosome 7 and 14,472 on the X chromosome, which correspond to 46% and 39% of the total segments on each chromosome, respectively (Table 1B). Table 1C, D describe the categorization of the CCGG segments based on the genome annotations. About 47% of the segments map to intergenic regions, 5% map to promoter regions, and 46% to bodies of RefSeq genes (Table 1C). About 40% of RefSeq genes on chromosome 7 and 75% of X-linked RefSeq genes are covered by this HELP microarray (Table 1D). About 1% of CGIs annotated in the UCSC mm8 genome assembly can be assayed by this array (Table 1E).

CGIs and gene promoter regions have been the main targets in most DNA methylation studies. However, the importance of DNA methylation in genomic regions outside the promoters is becoming increasingly apparent [2, 11]. It is expected that this microarray method should be appropriate for the analysis of previously unexplored and potentially informative parts of the genome.

2.3. Analysis of DNA methylation profiles of stem cells and germline cells

I performed DNA methylation profiling of the following samples: embryonic stem (ES) cells from male and female blastocysts, male and female embryonic germ (EG) cells established from PGCs of embryonic day 12.5 (E12.5), germline stem (GS) cells derived from spermatogonia, and male and female PGCs purified from embryos in various stages by fluorescence activated cell sorting (FACS). PGCs were isolated from male and

female E10.5, E13.5, and E17.5 embryos. PGCs have not entered the gonads at E10.5, and PGCs are colonized within the gonads in E13.5 embryos. At E17.5, PGCs are subjected to mitotic arrest in male gonads, and female PGCs are arrested in the early phase of meiosis [20]. I isolated germ cells from newborn ovary and testis. Whole adult testis, thymus, and brain were isolated from male mice and used for the analysis. Germ cells in the adult testis were purified by FACS from *Mvh* (mouse Vasa homolog)-*Venus* transgenic mouse) [35]. Gene expression profiling of all samples was conducted using my custom 44K microarray. Fig.7 shows the results of principal component analysis (PCA) and hierarchical cluster analysis of the DNA methylation profiles and the gene expression profiles. In this comparison, pluripotent stem cells (i.e. ES and EG cells) and PGCs from various stages show similar but distinct expression profiles; ES and EG cells are positioned more closely (blue circle) relative to PGCs (red circle) (Fig. 7A and 7C). This result confirms my previous findings that PGCs possess a distinct transcription program from ES cells, although both share the expression of common 'signature genes'[36]. In contrast, analysis of the DNA methylation profiles showed the differences between samples more clearly. PGC samples could be classified into two groups: one comprising female PGCs and early male PGCs (i.e. E10.5 and E13.5 in the red circle) and E17.5 and P0.5 male germ cells that formed a cluster together with GS cells and testis (green circle) (Fig. 7B and 7D). Male PGCs in different stages appeared to be more distantly related to each other than to female PGCs, suggesting that the DNA methylation profiles change more drastically during male

PGC development. These results suggest that cell type can be classified by their DNA methylation profiles and that, in some cases, DNA methylation profiling can display differences in the cellular state more effectively.

2.4. K-means cluster analysis of DNA methylation profiles revealed cell-type specific differentially methylated regions

To visualize differences in the DNA methylation profiles of the samples tested, nonhierarchical k-means analysis (k=12) was performed using the data obtained from the 28,217 informative CCGG segments; the result is shown as a heat map in Fig. 8A. One of the most conspicuous trends was that the genomes of the PGCs examined are mostly hypomethylated except for male E17.5 PGCs. Box plot of the M-value for each sample is shown in Fig. 8B. Global levels of DNA methylation are lower in the E10.5 PGC genome than in ES and EG cells, and the levels are even lower at E13.5. At E17.5, the methylation level of male PGCs is increased, whereas female PGCs maintain a hypomethylated status similar to those at E10.5 or E13.5. The difference in methylation level between male and female germ cells is most prominent in neonates: male spermatogonia possess a highly methylated genome. GS cells derived from spermatogonia also have a globally hypermethylated genome. Genomes of the somatic organs, thymus and brain, are also relatively hypermethylated, although the brain genome is less methylated than the thymus. GS cells possess a similar DNA methylation profile to that of P0.5 spermatogonia (R = 0.80). Epigenetic features of GS cells have not been reported to date, and this result suggests that GS cells should provide a valuable in vitro model for epigenetic studies of spermatogonial cells. Adult testis, which comprises both germ cells and somatic cells, has

a slightly lower DNA methylation level relative to spermatogonia (P0.5 male) and GS cells (Fig. 8B).

Although most of the PGC genomes are hypomethylated, genomic regions classified as cluster 12 remain methylated at a level comparable to the other cell types examined. The rest of the clusters showed some cell or tissue specificities in DNA methylation. For example, cluster 11 regions are hypomethylated in stem cells and germ cells but are hypermethylated in the somatic cell types examined. There are 1,111 CCGG segments classified in the cluster 11,169 of which correspond to promoter regions of known RefSeq genes. Gene Ontology (GO) analysis showed that terms like meiotic sister chromatid cohesion, meiotic chromosome segregation, female meiosis, and oogenesis are enriched in these genes. Genes related to germ cell differentiation or meiosis, such as *Stra8* [37], *Sycp3* [38], and *Figla* [39] are included in this cluster.

2.5. Characterization of germline-specific hypomethylated CCGG segments on the X chromosome

The clusters were then characterized by examining their tissue specificities in DNA methylation patterns. I noticed that Cluster 4 comprises the segments hypomethylated only in PGCs, GS cells and the testis, whereas these segments are hypermethylated in ES and EG cells, and somatic organs. These putative germline-specific hypomethylated segments were characterized further. Although no particular GO terms are enriched, Cluster 4 contains genes expressed in the germ cells of testis such as *Xmr* (*Xlr*-related, meiosis regulated)[40, 41] or CTA genes; e.g. the Mage (melanoma antigen) gene family[42]. The number of Cluster 4 CCGG segments mapped onto the X chromosome is disproportionately high. There are 1,004 segments in Cluster 4, and 715 (71.2%) are on the X chromosome, which has 14,472 CCGG segments in total. In contrast, 219 of Cluster 4 segments (21.8%) are on Chromosome 7, which carries 22,128 segments. Because genes expressed in germ cells or testis are known to be enriched on the X chromosome[32, 33], I focused on the Cluster 4 segments of the X chromosome as candidates involved in the epigenetic regulation of germ cell-specific gene expression. As shown in Fig. 9A-D, I plotted the M-values of all the CCGG segments along the X chromosome using the data obtained from each sample (grey dots). To translate the M-value measurements into regions of equal M-value, I used a circular binary

segmentation program, which is used normally for comparative genomic hybridization analysis [43]. Using this program, I drew lines (black horizontal lines) to show regions of equal M-value. By tracing the line, I could identify the genomic regions in which the M-value changes significantly from the flanking regions. The M-values of the segments belonging to Cluster 4 are overlaid as red dots. The distributions of the M-values in the DNA of somatic cells (i.e. brain and thymus) along the entire X chromosome are similar to each other: the average M-value is less than –1, with some local exceptions. The Cluster 4 dots are mapped even below the average line, indicating that, as expected, Cluster 4 segments are hypermethylated in both brain and thymus (Fig. 9A). In ES and EG cells (Fig. 3 and Fig. 9B), the average M-values of the Cluster 4 segments do not change significantly along the X chromosome and are positioned below –1, suggesting that Cluster 4 segments are largely hypermethylated in the genomes of ES or EG cells. In sharp contrast, in E17.5 male PGC DNA, it appears that the average M-value line is often discontinuous, and that hypomethylated CCGG segments exist over relatively large, contiguous genomic regions (Fig. 9A-D). For example, the average M-value of the segments within the 9 MB genomic region harboring the *Xmr* gene cluster (double-headed arrows) is close to 0, and Cluster 4 segments are enriched in this region. It is obvious that the distribution of the Cluster 4 segments is not uniform, and that these Cluster4 segments form 'hypomethylated domains' compared with their flanking regions. These trends persist in P0.5 spermatogonia and GS cells derived from P0.5 spermatogonia (Fig. 9A), with a few cell type-specific differences.

In testis DNA, the overall methylation pattern of the Cluster 4 segments is essentially similar to that found in the male germline cells, described above (Fig. 9A and C). I also examined earlier stages of male PGCs (Fig. 9C). In E10.5 male PGCs, formation of hypomethylated domains, e.g. *Xmr* region, is not as obvious as seen in E17.5 male PGCs. In E13.5 male PGC DNA, the distribution of the Cluster 4 segments is similar to that found in E17.5 male PGCs. It thus appears that clustering of hypomethylated DNA segments become increasingly evident on the X chromosome during the development of male germ cells.

2.6. Discovery of large genomic regions hypomethylated specifically in male germline cells

Although it is clear that segment DNAs possess generally lower methylation levels in female PGCs than in cells of somatic organs, the formation of hypomethylated DNA regions as seen in male PGCs is not evident in female PGCs (Fig. 9C and D). I therefore decided to focus on the male germline-specific hypomethylated DNA regions that comprise Cluster4 segments. To visualize the hypomethylated DNA regions in the male germline from a different viewpoint, I plotted fold differences in the methylation level between somatic and male germ cell DNA along the X chromosome (Fig. 10A). Because methylation patterns of Cluster 4 segments are essentially similar in testis, E17.5 and P0.5 male PGCs, the testis was chosen for this analysis. Brain was also used as somatic tissue for this analysis. As the data were plotted with a log2 scale, a negative value indicates the lower level of DNA methylation in the testis than in the brain. The plot revealed broad domains with lower methylation levels in the testis and therefore, in male germ cells of late stages (colored light blue in Fig. 10A). These broad domains of hypomethylated DNA described above are distinct from CGIs, which are generally located within or near a promoter and have a typical length of 300– 3000 bp [44]. The broad and hypomethylated domains identified here are often much larger than CGIs and do not show preferential localization at promoter regions. Thus, these broad domains do not correspond to the known hypomethylated regions and may represent a hitherto unknown epigenomic

entity. For convenience of discussion, I here designate such a broad and hypomethylated domain as a LoD. By definition, a LoD has a size of >10 kb and shows more than a 2-fold difference in M-value between germline and somatic cells (testis and brain in this case). Each LoD should also have at least one Cluster 4 segment.

2.7. Overlap of LoDs with segmentally duplicated regions

Using the definition described above, I list the LoDs of the X chromosome in Table 3. There are 16 LoDs on the X chromosome (Table 3), and their sizes are generally large: 11 of the16 LoDs are >100 kb (mean:1,219,252 bp), and six of the large LoDs are ~1 Mb. The mammalian genome is replete with segmentally duplicated regions [45]. Although segmental duplications can be found on every chromosome, they are particularly abundant on the sex chromosomes. Because LoDs are generally large and contain gene families such as *Xmr,* I asked whether LoDs overlap with segmentally duplicated regions. As shown in Fig. 10A and B and Table 3, all LoDs on the X chromosome are found to contain segmentally duplicated regions. The use of the *Msp*I control represents an unusual strength of the HELP assay to remove the potentially confounding effect of copy number variation [24]. Combined with a probe design that selects only unique sequences for hybridization, these aspects ensure that the DNA methylation readout from regions of constitutive segmental duplication accurately reflects the underlying DNA methylation and is not influenced by DNA copy number. Hypomethylation of two such domains, LoD 10 and 12, was confirmed by Southern blot analysis (Fig. 10C and D). Since LoD 10 and 12 contain homologous, locally repeated sequences, a hybridization probe can be used to assess the methylation status of both regions. The genomic DNAs of the thymus, brain and testis were digested by either methylation-sensitive *Hpa*II or the methylation-insensitive *Msp*I. In the *Hpa*II digests of thymus and brain DNA, no bands were detected except for a

hybridization signal in the unresolved part of the lanes, indicating that the genomic region is hypermethylated in somatic organs. In the *Hpa*II digest of testis DNA, many bands were detected, and the band pattern was essentially the same as that found in the *Msp*I digest, clearly indicating that this region is largely unmethylated in the testis. Given that the testis comprises both germ and somatic cells, I asked whether LoDs are hypomethylated in germ cells. I used an *Mvh-Venus* reporter transgenic mouse line [46], in which germ cells are marked by Venus fluorescence protein. I also used FACS to purify the *Venus*-positive germ cells from the adult testis and performed Southern blot analysis. The results indicated that the genomic regions in the purified germ cells are indeed hypomethylated (Fig. 10C and D).

2.8. Predominance of genes expressed in male germ cells or in the testis in LoDs

We noticed that most LoDs contain genes that are expressed in the testis. For example, *Gmcl1l* (germ cell-less protein-like 1-like), *Ssx9*, *Fthl17*, *Xmr*, *Mageb*, *Ott*, *Samt4* and *Magea* are expressed in the testis and are included in LoDs 1, 2, 3, 4, 11, 14 and 15, respectively (Table 3). Expression of these genes is also detected in germ cells purified from adult testes. If I omit LoDs 10, 12 and 16, which do not carry known genes, only LoDs 6 and 13 do not contain genes predominantly expressed in germ cells (Table 3 and Table 4). The mean expression levels of genes contained in LoDs are shown in Fig. 11A. Genes within LoDs show significantly higher expression in the testis than in the brain. Figure 11B shows the mean levels of DNA methylation within and outside LoDs on the mouse X chromosome. Fig.11 indicates that there is an inverse correlation between the level of DNA methylation and the expression of genes in LoDs.

2.9. Genomic structures of LoDs: *Xmr*/*Slx* and *Mageb* regions

In addition, I demonstrate the detailed structures of two LoD regions (Fig. 12). LoD 4 is ~ 9.1 Mb in size (chrX: 22,991,291–32,117,922) and contains three distinct genes/gene families, all of which are expressed specifically in the testis. Because *Xmr* is a synonymous gene with *Slx* [40, 41], I call this gene/gene family either *Xmr* or *Xmr*/*Slx* in this study. *Xmr*/*Slx* is known to be expressed in spermatids, where it encodes a protein, *SLX*/*XMR*, normally localized in cytoplasm [40, 41]. *Xmr*/*Slx* represents a locally duplicated multi gene family, whose copy number is at least 28 in LoD 4. *Gmcl1l* and *LOC236749* are included in the same LoD, and both are expressed in the testis and in purified male germ cells (Fig. 12A). The LoD 4 region represents one of the largest segmentally duplicated regions on the mouse X chromosome (Katsura and Satta, personal communication) and can be divided into four subregions (Fig. 13). Subregion I spans~3 Mb and harbours tandemly repeated *Xmr* genes. Subregion II spans ~3.8 Mb and comprises both tandem and inverted repeats (IRs) of *Xmr* genes. Subregion III contains tandem and IRs of *Gmcl1l* genes, which are duplicated on two distant sites on the X chromosome; the other site is also classified as LoD 1 (Table 3). Subregion IV is, 1 Mb and contains tandem repeats of Xmr genes. One hundred and forty-one CCGG segments are mapped within LoD 4, and the fold difference in methylation level (brain versus testis) of these segments were calculated as described in Fig. 10A. The mean value is –1.4745,

suggesting that the CCGG segments in this region are generally hypomethylated in the testis genome (Table 3 and Figs 9 and 10). Because of the repetitive nature of the LoD region, HELP probes cannot be assigned for most of the subregions II and IV. To examine DNA methylation in these regions, I performed Southern analysis of the testis and brain DNA digested with either *Hpa*II or *Msp*I, and hybridized with an *Xmr* cDNA probe. As shown in Fig. 12A and Fig. 13. the *Xmr* cDNA probe should be able to assess the methylation status of 161 restriction fragments. These fragments are distributed evenly within subregions I, II and IV, and fill the gaps of information provided by the nanoHELP assay, which tests only unique sequences. The results of the Southern blot analysis shown in Fig. 12B demonstrate that the *Xmr* region is highly methylated in the brain and liver, whereas a considerable proportion of the restriction fragments appear unmethylated in the testis. It has been suggested that transcriptionally active genes are hypomethylated in their promoter region, while their gene bodies tend to be hypermethylated [2]. However, a magnified view of the LoD 4 region (Fig. 14) indicates that all CCGG segments in this region are hypomethylated in the testis and male PGCs regardless of their positions with respect to the *Xmr* genes. Both the probes positioned near the transcription start sites and the probes positioned at introns or even at intergenic regions are unmethylated in the testis, GS and male PGCs. The Southern blot analysis data suggest that the CCGG segments containing exons of the *Xmr* genes appear to be relatively hypomethylated in the testis (Fig. 12B). These results imply that methylation of the whole LoD 4 is subjected to a

region-wide regulation. This feature is shared by other LoDs not described here. *Mageb* belongs to the *Mage* (melanoma antigen) gene family, which is expressed in spermatogenic cells and in some cancer cells [40]. Figure 12C shows a genomic region spanning ~1 Mb that contains *Mageb1* and *Mageb2* genes. This region represents a large IR with arms of ~400 kb in length. At the ends of both arms, LoDs 10 and 12 are located 4–2 kb upstream of the transcription start sites of *Mageb1* and *Mageb2*, respectively. These two LoDs do not contain the *Mageb* locus itself (Fig. 15). Both LoDs are highly homologous and ~14 kb long, and comprise repeat sequences with a unit size of ~3 kb. These sequences are both tandem and IRs (Fig. 12C magnified part), are found only in these LoD regions and are clearly hypomethylated only in germ cells (Fig. 15). Hypomethylation of LoDs 10 and 12 was confirmed by Southern blot analysis as described (Fig. 10C and D).

2.10. Developmental changes in the methylation levels of LoDs

LoDs are hypomethylated in the testis, GS cells and male PGCs. The methylation heat maps of LoDs 10 and 12 shown in Fig. 12D illustrate how the DNA methylation of LoDs changes during germ cell development. During development, the PGC genome undergoes global DNA demethylation, which is known to be completed between E11.5 and E13.5 [15]. In E10.5 PGCs, the LoDs tested here are not unmethylated completely, whereas demethylation of LoD DNA progresses in PGCs by E13.5. At E17.5, the LoD regions are largely unmethylated in both male and female PGCs. This trend persists in later stages of male germ cells, where as the methylation levels of the LoDs appear to increase in newborn oocytes. The results together with the results shown in Fig. 9 suggest that, in general, LoDs begin to form between E10.5 and E13.5, and distinct hypomethylated domains are established around E13.5 in the male germline. Despite the global increase in DNA methylation at later stages of male germline development (Fig. 8B), hypomethylation of LoD DNAs is maintained in male germ cells. Although LoDs 10 and 12 are shared by male and female PGCs, the overall DNA methylation patterns are not identical, suggesting that a distinct epigenomic status is generated in male and female germlines (Fig. 9C and 9D).

2.11. Coincidence of most LoDs with broad domains of the repressive histone mark, H3K9 dimethylation

I have shown that most LoDs are broad genomic domains with low DNA methylation levels that form boundaries between the LoDs and other methylated parts of the genome. The mammalian genome can be divided into broad domains of distinct histone modifications[47, 48]. For example, LOCKs (large organized chromatin K9 modifications) are genomic domains with histone H3 lysine 9 dimethylation (H3K9me2) modification thought to be involved in region-wide gene repression [48]. To investigate the relationship between LoDs and the repressive histone mark, ChIP-on-chip analysis was performed to detect H3K9me2 enrichment in GS cells as a representative of germ cells in this test and in cumulus (somatic cells in the ovary) cells as a somatic cell control [49]. Figure 16 shows the H3K9me2 modification patterns on the X chromosome in both GS and cumulus cells. The overall pattern of H3K9me2 modifications along the X chromosome in GS cells is essentially similar to that in cumulus cells (Fig. 16A and C). Enrichment of the modifications along the LoD regions (colored light blue) is seen in both GS and cumulus cells (Fig. 16A and C). In contrast, as expected, DNA methylation levels in the LoD regions are high in cumulus cells and low in GS cells (Fig. 16B and D). Figure 16E shows a magnified view of LoD 12, indicating that the hypomethylated region has the H3K9me2 mark. A significant enrichment of H3K9me2 is found in most (11 of 16) LoDs in GS cells (Fig. 17). Expression of six genes contained in the LoDs was examined in cumulus, GS,

testis and two other somatic cell types by quantitative RT-PCR analysis (Fig. 16F). *Fthl17*, *Ott*, *Mageb* and *Magea* are included in the LoDs and are expressed in the testis; the expression level of these genes is much higher in GS cells, but only negligible expression is detected in somatic cells. This result indicates that genes in hypomethylated LoDs can be expressed even though the same region has continuous H3K9me2 modifications (Fig. 16A , C and Fig. 17), demonstrating peculiar epigenomic features of LoD regions. It is reasonable to expect that *Ssx* and *Xmr* are barely detectable in GS cells, because these genes become active in postmeiotic stages [32], whereas GS cells are derived from pre-meiotic spermatogonia. It is probable that, in GS cells, other factors required for the expression of postmeiotic genes (e.g. transcription factors) are lacking. These results suggest that DNA hypomethylation in LoDs may not be sufficient by itself, but is a prerequisite for the expression of LoD genes.

3. Discussion

In the present study, I have analyzed the DNA methylation profiles of developing germ cells using the modified nanoHELP method, which requires only a limited amount of DNA. Recent studies by Guibert et al. [29] using the MeDIP analysis of a promoter array and Seisenberger et al. [31] using a whole-genome bisulfite sequencing suggest that DNA demethylation of the PGC genome is initiated earlier than previously thought [12, 15]. The finding that the PGC genome is substantially hypomethylated already at E10.5 is consistent with the result of Seisenberger et al. [31], confirming the technical reliability of my method. My data from developing germ cells revealed for the first time the presence of large, hypomethylated DNA domains on the X chromosome of male germline cells in mice.

3.1. Discovery of large hypomethylated domains of epigenomic organization

Traditionally, epigenetic studies have focused on modifications of genes or elements adjacent to genes. However, with the development of genome-wide assays, recent studies have revealed marked clustering of particular histone modifications over relatively large genomic regions; e.g. LOCKs and BLOCs (broad local enrichments) enriched with the histone marks H3K9me2 and H3K27me3, respectively [47, 48]. These large epigenetic marks, LOCKs in particular, are thought to be involved in gene silencing. DNA methylation is found throughout the mammalian genome except for short unmethylated regions, CGIs, which typically occur around the transcription start sites of genes [2]. The LoDs described in this work are also hypomethylated genomic regions, but are distinct from CGIs in terms of their size, tissue specificity and genomic structure. To my knowledge, large differentially methylated DNA regions showing germ cell specificities, such as LoDs, have not been previously reported. This may be because previous studies have focused only on methylation of gene promoters and not broader genomic contexts in germ cell samples. In contrast, my custom HELP chip method could assess the DNA methylation status of both genic and intergenic regions using the meager amounts of DNA that could be sampled from germ cell genomes in this study. Seisenberger et al. [31] recently reported the results of whole-genome bisulfite sequencing analysis of the mouse PGC genome. I analyzed their data on E16.5 male PGCs to determine whether LoDs could be found at the single-nucleotide level and found

that the number of sequence reads mapped to LoDs were significantly lower than that mapped to the flanking regions. Given that mapping of bisulfite converted short sequence reads onto locally duplicated regions is technically challenging, the probability of finding LoDs using the currently available bisulfite sequencing data seems low. In contrast, the HELP assay uses an *Msp*I control to remove the potentially confounding effect of copy number variation [24] along with a probe design that selects unique sequences for hybridization. These ensure that the DNA methylation readout from regions of segmental duplication is genuinely reflective of the underlying DNA methylation. Oda et al. [50] reported that CGI methylation of an X-linked homeobox gene cluster spanning ~1Mb is under long-range regulation in a tissue-specific manner. Therefore, widespread changes in DNA methylation could occur depending on the cellular phenotype or differentiation status.

3.2. Peculiar epigenomic features of LoDs

LoDs have been detected based on arbitrary criteria, but most share common features. Most LoDs represent segmental duplications that harbor germline expressed genes and overlap with large H3K9me2-enriched domains. Wen et al. [48] described large H3K9me2-enriched chromatin blocks, LOCKs, in the human and mouse. The occurrence of LOCKs is differentiation specific: there are more LOCKs in differentiated cells, and genes contained in the LOCKs tend to be repressed during differentiation. Because LOCKs substantially overlap with lamin B-associated domains, a gene-silencing mechanism based on three-dimensional subnuclear organization has been proposed [48] (Fig. 18).

I found that most LoDs are enriched with H3K9me2 modifications, and that at least four LoDs—1, 2, 3 and 4—correspond to the LOCKs described by Wen et al. [48]. Overlaps of other LoDs with LOCKs cannot be checked because LOCKs data are not available for the rest of the mouse X chromosome. Overlap of LoDs with LOCKs is counterintuitive because LOCKs are supposed to repress gene expression, whereas genes can be highly expressed within LoDs. This may be reconciled if gene silencing in LoDs is complete when both DNA methylation and H3K9me2 marks are established, but is derepressed in the absence of DNA methylation. Consistent with this idea, somatic cells such as cumulus cells, which have both marks, do not express the LoD genes, although germ cell genes can be active in DNA-hypomethylated but H3K9-dimethylated LoDs.

The H3K9me2 histone methyltransferases, G9a and GLP, are required for DNA methylation in ES cells, but not in cancer cells [51]. It is thus likely that DNA methylation and the H3K9me2 modification are not always interdependent, and that they can be regulated independently in the LoD regions of male germ cells and cancer cells.

3.3. Segmental duplication, hypomethylation and gene expression in germ cells and cancer cells

Through the analysis of germ cell-specific hypomethylated regions, I found that LoDs overlap with large segmentally duplicated regions, within which germ cell-expressed genes are commonly found. Some of these genes, such as *Xmr*, are found only in rodents. In contrast, the *Mage* gene family genes, *Ssx* and *Fthl17*, are conserved in the human genome and are known as CTA genes, which are expressed specifically in germ cells and in some tumor cell types. More than 260 CTA genes have been detected in the human (http://www.cta.lncc.br/), and half of them are on the X chromosome. Most of the X-linked CTA genes are organized as multicopy gene families [34]. Warburton et al. [52] searched the IR structures in the human genome and found that the X chromosome is replete with large IRs harboring testis-expressed genes, most of which encode CTA genes. More than 40% of large IRs found in the mouse genome are on the X chromosome, and *Ssx*, *Fthl17* and the *Xmr* loci are contained in such regions. Thus, three kinds of studies with different starting points reached the same conclusion: the X chromosome is abundant with duplicated regions containing germ cell-expressed genes, including CTA genes. To this, I add the new observation that these regions also have unique epigenomic features, i.e. widespread DNA hypomethylation and H3K9me2 enrichment. The epigenomic features of the LoDs could account for the finding that CTA genes can be activated by inhibition of DNA methylation but not by a reduction in H3K9 dimethylation [51, 53],

and suggest that DNA methylation is the key epigenetic mechanism involved with the regulation of LoD–CTA genes. It is not fully understood how DNA methylation regulates the coordinated expression of CTA genes in a cell type-specific manner. It is also necessary to clarify whether CTA gene expression contributes directly to oncogenesis or whether it simply reflects global chromatin changes that occur during tumor formation. Simpson et al. [54] postulated an intriguing hypothesis that the aberrant expression of germline genes in cancer reflects the activation of the gametogenic program, which is normally silenced in somatic cells. The gametogenic program is normally repressed because germline specific products would be harmful for normal somatic cells, whereas they would be advantageous for cancer cells. To test this hypothesis, it will be essential to elucidate the activation mechanism for the germline gene expression program, as well as the epigenetic and chromatin status required for the operation of this program. As shown in this study, widespread DNA hypomethylation may be a prerequisite for the activation of LoD genes, including CTA genes. In addition to DNA methylation, the nuclear chromatin environments within germ cells and/or tumor cells may also be important for long-range transcriptional control over large genomic regions, because LOCKs [48], LoDs and the partially methylated domains found in colorectal cancer [10] are correlated with nuclear lamin associated domains. Therefore, further studies of epigenomic features and the nuclear architecture of LoDs may shed light on the germline gene expression program and its relationship to oncogenesis [54].

4. Conclusions and future directions

In this study, I first established a method that allow DNA methylation analysis of subnanogram quantity of genomic DNA by improving the method known as HELP assay [24]. With this modified nanoHELP method, I analyzed genomes of PGCs and showed for the first time that the PGC genome is substantially hypomethylated already at E10.5, suggesting that DNA demethylation during PGC development may initiates earlier than previously thought. Currently my colleagues and I try to determine when DNA demethylation of PGC genomes starts using other genome-wide technique based on a massively parallel sequencing.

Through the analysis of PGCs of various developmental stages as well as stem cells derived from embryos and germ cells, I discovered large hypomethylated DNA domains (LoDs). Although large-scale enrichments of a particular histone modifications have been reported, large differentially methylated DNA regions showing germ cell-specificities such as LoDs have not previously been reported to my knowledge. The germ cell-specific LoDs contain gene families showing germ cell-specific expression, and I demonstrate that DNA methylation is the key epigenetic mechanism involved in the regulation of LoD genes. Among these, there are mouse genes with homology to human cancer testis antigen genes, suggesting epigenetic regulations common to both germ cells and tumor cells. Remarkably, most LoDs coincide with broad domains of the repressive histone mark, H3K9 dimethylation, and indeed overlap with LOCKs reported by Wen et al.

[48]. Therefore, expression regulation of genes within the LoDs may be dependent on changes in nuclear architecture. Analyses of DNA methylation profiles and chromosomal positioning of LOCKs/LoDs in cancer cells and germ cells of both mouse and human should be performed in future.

5. Materials and Methods

5.1. Sample preparations and purifications of DNA and RNA

TMA5 cells are male ES cells derived from the 129/Sv mouse [55]. The female ES#5 line was from F1 hybrid mice between TgN(deGFP)20^{Imeg} (RBRC No. 00822) and MSM/Ms (RBRC No. 00209) [36]. The EG cell lines used in this study were TMA55G (male) and TMA58G (female) [55]. These ES and EG cell lines were cultured on mitomycin C-treated primary mouse embryonic fibroblasts in Dulbecco's modified Eagle's medium (Sigma-Aldrich, St. Louis, MO) supplemented with 14% Knockout Serum Replacement (Life Technologies, Carlsbad, CA), 1% fetal bovine serum (Life Technologies), 1000 U/ml Leukemia Inhibitory Factor, 2-mercaptoethanol, 1x nonessential amino acids, and penicillin–streptomycin. The *Oct3/4*–GFP transgenic mouse line, TgN(deGFP)18^{Imeg} (RBRC No. 00821) [56], was used to collect PGCs from developing mouse embryos, as described previously [36]. GS cells were obtained from the RIKEN BRC Cell Bank (RCB1968) and were cultured on a feeder layer as described by Kanatsu-Shinohara et al. [57]. Germ cells expressing the *Venus* reporter were purified by FACS from the *Mvh[35]-Venus* BAC transgenic mouse line (N. Mise and K. Abe, in preparation). All animal experiments were approved by the Institutional Animal Experiment Committee of the RIKEN BioResource Center. DNA and RNA were extracted simultaneously from the same samples using an AllPrep DNA/RNA micro kit (Qiagen, Hilden, Germany). The amount of DNA was measured using a Qubit dsDNA High

Sensitivity Kit (Life Technologies). RNA was quantified by NanoDrop (Thermo Scientific, Wilmington, DE), and the quality of RNA samples was checked using a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA).

5.2. Gene expression profiling

A 44K custom microarray was used for gene expression profiling throughout this study [58]. This custom array covers all the known protein-coding genes as well as ESTs derived from PGC cDNA libraries, and was manufactured by Agilent Technologies. Total RNA was labeled with Cy3-CTP with a Quick Amp labeling kit (Agilent Technologies). Hybridization was performed according to the protocol suggested by the supplier. Hybridized slides were scanned using a microarray scanner (Agilent Technologies), and the signals were processed with the Feature Extraction software ver. 10.5.1.1 (Agilent Technologies). The processed signal data were normalized and analyzed by Gene Spring GX11.5 software (Agilent Technologies). The microarray experiments were conducted using biologically duplicated samples.

5.3. Modified nanoHELP: linker-mediated amplification and hybridization

The nanoHELP assay, a microarray-based DNA methylation analysis, was performed according to my previous reports [24, 26] with modifications. Briefly, genomic DNA (0.5–2 ng) was digested by *Hpa*II or *Msp*I in 100 μl of reaction mixture at 37°C overnight. This was followed by DNA purification with the MinElute Reaction Cleanup Kit (Qiagen), and the digested DNA was ligated to linker adapters, NHpaII12/NhpaII24 and JHpaII12/JhpaII24 [26] overnight at 16°C. After removing the linker adapters, the ligated DNA was added to a total of 50 μl PCR reaction mixture containing 1.5 μl each of 20 μM primer (NHpaII24, 5'-GCAACTGTGCTATCCGAGGGAAGC-3'; JHpaII24, 5'-CGACGTCGACTATCCATGAAC AGC-3'), 10 μl of 5 M betaine (Sigma-Aldrich), 200 μM of dNTPs, and 2.5 units of ExTaq DNA polymerase (TaKaRa Bio Inc., Otsu, Japan) in a buffer supplied by the manufacturer. The mixture was heated at 72°C for 10 minutes and subjected to PCR amplification with the following parameters: 15 cycles at 95°C for 30 seconds and 72°C for 3 minutes, with a final extension at 72°C for 10 minutes. After the first round of amplification, one-tenth of the volume of the reaction was added into a fresh PCR reaction mix containing the same primers and amplified for an additional 10–15 cycles [59] with the same PCR parameters as described above. The PCR products were purified using the MinElute Kit (Qiagen). An additional column-washing step with 750 μl of 35% guanidine hydrochloride (Nacalai Tesque Inc., Kyoto, Japan) solution was performed to remove the residual primer–adapters. The amplified DNA originally digested

with *Hpa*II was labeled with Cy5-labeled Random 9-mers (TriLink Biotechnologies, San Diego, CA), while *Msp*I-digested DNA was labeled with Cy3-Random 9-mers (TriLink Biotechnologies). The labeled DNAs were mixed and hybridized with a custom microarray (Roche NimbleGen, Madison, WI) using a NimbleGen Array Hybridization Kit (http://www.nimblegen.com/products/lit/lit.html). After washing with the NimbleGen Array wash kit, the microarrays were scanned on an Agilent Technologies Scanner G2505C with a setting of 5 μm resolution. The HELP array experiments were carried out using biological replicates. The raw data were processed using NimbleScan 2.4 data extraction software (NimbleGen) to obtain the processed $\log_2$ (Cy5/Cy3) ratio data.

ChIP-on-chip experiments using cumulus cells and GS cells were performed as described previously [49].

5.4. Microarray design

The microarrays were designed to represent restriction fragments with 5'-CCGG restriction sites in a size range of 200 to 2000 bp (=CCGG segments) on mouse chromosome 7 and the X chromosome. Ten 50-mer oligonucleotide probes were designed from each CCGG segment sequence, avoiding repeat-masked regions and sequence ambiguities. Probe sequences were selected using a score-based selection algorithm, as described[24]. Detailed information for the coverage of genomic regions on each chromosome and annotations of the CCGG segments are described in Table 1. Information about the positions of probes, M-values obtained from different samples, and k-means cluster number are described gff files of the HELP array data are available at the web site (http://www.brc.riken.go.jp/lab/mcd/mcd2/protocol/nanoHELP.html).

5.5. Data analysis

The steps in the HELP data analysis are shown schematically in Fig. 4. Briefly, hybridization signal noise is first removed from the processed data by cutting off the values in the range of random sequence probes. On the microarray, 10 oligonucleotide probes are normally assigned to each CCGG segment. The median signal intensity of the 10 probes is calculated and used to define the segment's signal intensity. Using the median signal values, the *Hpa*II/*Msp*I ratio is then calculated for each CCGG segment and converted to a $\log_2$ value to obtain the M-value. After normalization of the microarray ratio data, hypomethylated and hypermethylated segments are distinguished using an R script (http://www.r-project.org/) that determines the threshold values based on a binarization method [60]. The marginal width of the threshold is calculated using the Mahalanobis distance [61]. The $\log_2$ value at the threshold is set as 0 so that hypomethylated segments have a positive value (>0) and hypermethylated segments have a negative value (<0). For interarray normalization of the *Hpa*II/*Msp*I ratio, the threshold value of each array data set is scaled to 0.

5.6. Accession number

Gene expression microarray data and the HELP array data are available at Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) (Accession Number; GSE39895).

Tables and Figures

Table 1A. Number of CG sites and CCGG sites in the mouse genome (mm8).

Chromosome	Number of CCGG sites on each chromosome	Number of CG sites on each chromosome	% of CG sites contained in CCGG sites on each chromosome	% of CCGG segments of 200-2000 base in total CCGG segments
Chr1	106,823	1,471,329	7.26%	43.62%
Chr2	113,291	1,520,279	7.45%	45.75%
Chr3	84,044	1,190,876	7.06%	42.36%
Chr4	104,061	1,323,303	7.86%	46.50%
Chr5	103,127	1,346,153	7.66%	47.22%
Chr6	84,219	1,165,757	7.22%	41.05%
Chr7	95,605	1,215,853	7.86%	46.21%
Chr8	84,776	1,130,161	7.50%	47.62%
Chr9	79,735	1,066,932	7.47%	47.08%
Chr10	79,703	1,103,558	7.22%	46.11%
Chr11	93,717	1,163,245	8.06%	48.64%
Chr12	70,512	954,683	7.39%	45.20%
Chr13	69,182	976,005	7.09%	45.40%
Chr14	67,206	932,261	7.21%	43.79%
Chr15	65,477	869,622	7.53%	46.12%
Chr16	53,535	754,258	7.10%	44.49%
Chr17	65,436	849,625	7.70%	47.78%
Chr18	51,155	722,011	7.09%	45.17%
Chr19	41,982	557,394	7.53%	47.22%
ChrX	73,738	948,046	7.78%	36.90%
ChrY	1,222	15,675	7.80%	38.35%
total	1,588,546	21,277,026	7.47%	44.89%

Reference genome assembly used is UCSC Genome Browser, mm8 (Feb. 2006).

Table 1B. Coverage of CCGG segments on each chromosome by HELP array.

Chromosome	Number of CCGG sites on each chromosome	Number of segments on each chromosome covered by HELP array	% coverage of CCGG sites on each chromosome by HELP array	Proportion of the probes mapped to each chromosome per total probes on HELP array
chr7	95605	22128	46%	54%
chrX	73738	14472	39%	35%
chr11	93717	1200	3%	3%
chr6	84219	1187	3%	3%
chr12	70512	655	2%	2%

Reference genome assembly used is UCSC Genome Browser, mm8 (Feb. 2006).

Table 1C. A breakdown of CCGG segments annotated to different genomic categories

Genomic category	Segments number on HELP array	% of segments annotated to each category
intergenic	19306	47%
gene body	18892	46%
promoter	2008	5%
* promoter & gene body	372	1%
**could not transfer mm6 to mm8	342	1%
Total	40920	100%

"Promoter" is defined as a region between -2kbp and +0.5kbp of transcription start site.

* In this category, single CCGG segment overlaps with promoter and gene body.

** The HELP array was originally designed based on mm6 genome assembly, and re-annotated with mm8 genomic informations.

Reference genome assembly used is UCSC Genome Browser, mm8 (Feb. 2006).

Table 1D. Coverage of RefSeq genes by HELP array

Chromosome	Number of RefSeq genes covered by HELP array	Number of RefSeq genes on each chromosome (mm8)	% coverage of RefSeq genes by HELP array
chr7	837	2088	40%
chrX	758	1017	75%
chr11	414	1743	24%
chr6	47	1236	4%
chr12	32	792	4%

RefSeq Genes are processed non redundant symbol of refFlat file from California Santa Cruz Genome Browser.

Table 1E. Number of CCGG segments annotated as repeat sequences, RefSeq genes and CpG islands

Category	Number of segments covered by HELP array /Total number In mm8
LTR	4285/782603
LINE	7219/926605
SINE	8401/1413261
RefSeq	2164/23255
CpG island	165/15963

Reference genome assembly used is UCSC Genome Browser, mm8 (Feb. 2006).
RefSeq Genes are processed non redundant symbol of refFlat file from California Santa Cruz Genome Browser.

Table 2. Primer sets used for Bisulfite Pyrosequencing

Seq ID	Gene Symbol	Position (mm8)	Name	1st Forward primer (5'-3')	1st Reverse primer (5'-3')
MM6MSPIS00447193	Msi2	chr11:88474756-88475961	Msi2_A	GGTTGATAAATGAGGTTAGGGT	AATATCTCTATACCACACCCCAC
			Msi2_B	AGTGGGGGATATTTATTAAAAG	CTACAATTTTAAAACAACCAAAA
MM6MSPIS00442963	Nos2	chr11:78747440-78747664	Nos2_A	TGAGTTTGAGATTAGTTTGAGTT	CTAAATAAAACCCCTATTCCC
			Nos2_B	TGAGTTTGAGATTAGTTTGAGTT	CTAAATAAAACCCCTATTCCC
MM6MSPIS00290873	Clns1a	chr7:97574487-97575533	Clns1a-A	GGAATTATATTTGGGGGAGGG	TCTCAAAATCTATCCACACAACAA
			Clns1a-B	GATGGAGATGGGAATGGAGAT	TCACATTAAAACTCACAACCCCT
MM6MSPIS00286335	Ntrk3	chr7:78381272-78381662	Ntrk3-A	TGTTGTAAGGTTAAGATAGGGT	CACAAACAACTACCATTCTCTA
			Ntrk3-B	TGTTGTAAGGTTAAGATAGGGT	CACAAACAACTACCATTCTCTA
MM6MSPIS00290518	Odz4	chr7:96441242-96441877	Odz4-A	GGTTGTTTATGGTTTTATTTGG	TCTCCCATCTTTACAAAAAACT
			Odz4-B	ATTTTAAGTGGGAAATAAATGG	AAATCCCTTCCCTTATATAACC
MM6MSPIS00302223	Fam53b	chr7:132631377-132632958	Fam53b-A	TTGGGTTAGAGAGAGTGTTTTG	CACCTAACCTTCTCCAAAAAAT
			Fam53b-B	TTTTGAGTTTAGGTTATTTGGG	TCTAATACATCTTAATCCTCCCC

Seq ID	Gene Symbol	Position (mm8)	Name	2nd Forward primer (5'-3')	2nd Reverse primer (5'-3')	SequencePrimer (5'-3')
MM6MSPIS00447193	Msi2	chr11:88474756-88475961	Msi2_A	GAGGTTAGGGTTAGAGTTTGGTGATT	[Bio]AAAAATCCICCTACCTCTACTTCCC	TTTTAAAAAAAGAATATGAT
			Msi2_B	GAGAGATTGGTTGGAAGAAATTTATAGTTA	[Bio]ACAACACTCAATAATCCTAAAAATCATACT	GTGTATTATTTAGTTATG
MM6MSPIS00442963	Nos2	chr11:78747440-78747664	Nos2_A	GTTTTTTTGGGTTGATATTGGAGTT	[Bio]TTTCCCTAAAAAATCTACTTTCCTTTCTA	TTGTTTTTTGAGGAGTTAGT
			Nos2_B	GTTTTAGGAAAGGAGAGGGGAGTTA	[Bio]CTATTCCCCACTTCATCCCA	TGGGGTTATTGTGAGTATAT
MM6MSPIS00290873	Clns1a	chr7:97574487-97575533	Clns1a-A	TTTGGGGTTTGTTTGATTTGTA	[Bio]TTTAAAATATTTCCCCACCTCCCA	TATTGTATTTTGGTTATAGG
			Clns1a-B	[Bio]AGGGATGTTTTATGTTGAGATGG	AAACTCACAACCCCTAAACTTCT	ACCAAAATAATAAAATTACC
MM6MSPIS00286335	Ntrk3	chr7:78381272-78381662	Ntrk3-A	[Bio]AAATAAGAGATTGTGTGAGGTAGA	CCCCACRAATACACTAATTTAAAA	AAACTCAAAAAAACAACTAC
			Ntrk3-B	GTTTAAGGGGATGTGGTTGTATAT	[Bio]CATTCTCTAAATCTTCCCACRTAA	TGATTAATAGTTTTTGAGGT
MM6MSPIS00290518	Odz4	chr7:96441242-96441877	Odz4-A	TTGGTTTTTATTTTAGGGATGAT	[Bio]TAACAATTTCTCACRAAACAACC	TTAGGGATGATTTTTAGGTA
			Odz4-B	TGGGTTTTTGTATTTTGGTTTAGA	[Bio]CATACTTAAAAACCGCCCTAACA	TTTAAAGAAYGGGAGTG
MM6MSPIS00302223	Fam53b	chr7:132631377-132632958	Fam53b-A	GYGGGGGATTTAATGGTTATAT	[Bio]CCAACCCTTTACCTCTCTCCTA	TGTTAYGGGATGAGTG
			Fam53b-B	[Bio]TYGTTTTGAGAAATTAAGGTTTTA	ACCGTCAAAATATACRTTTAACCC	TCCCCTCACTTTCTTTA

Table 3. List of LoDs on mouse X chromosome

LoD No	position	length (bp)	Segment number on HELP array	Methytaion: Brain M value /Tesits M value	CGI	Seg Dups	Gene families	Description	Number of duplications: Total	within LoD	outside LoD on chrX	human CTA genes
1	chrX:3035387-4561626	1,526,239	12	-2.52		1344	Gmcl1l	germ cell-less protein-like 1	11	8	3	
2	chrX:7857031-7924410	67,379	5	-1.924		88	Ssx9	synovial sarcoma, X breakpoint 9	11	4	7	CTA
3	chrX:8116622-8236784	120,162	19	-1.82		45	Fthl17	ferritin, heavy polypeptide-like 17	7	6	1	CTA
4	chrX:22991291-32117922	9,126,631	100	-1.4745		10697	Xmr Gmcl1l LOC236749	XMR protein (Xlr-related, meiosis regulated). Germ cell-less homolog 1 (Drosophila)-like hypothetical protein LOC236749	32 11 1	28 2 1	4 9 0	
5	chrX:50380769-52184422	1,803,653	48	-1.75	1	1971	similar to Xmr protein	Adult male testis cDNA, RIKEN full-length enriched library, clone:4930527E24 product:weakly similar to XMR PROTEIN.	30	16	14	
6	chrX:57970833-58090999	120,166	5	-2.14		23	Ldoc1	leucine zipper, down-regulated in cancer 1	1	1	0	
7	chrX:58775958-58815172	39,214	13	-3.45		10	*1700019B21Rik	Mus musculus adult male testis cDNA,Mus musculus RIKEN cDNA 1700019B21 gene (1700019B21Rik), transcript variant 1, non-coding RNA.	1	1	0	
8	chrX:72448732-72544743	96,011	4	-1.59		67	LOC238829	hypothetical protein LOC238829 (AK133378 - Mus musculus adult male testis cDNA, RIKEN full-length enriched library, clone:4933402E19 product:hypothetical protein, full insert sequence)	4	1	3	
9	chrX:84663998-86056286	1,392,288	23	-1.2207		334	Pet2 4932429P05Rik	plasmacytoma expressed transcript 2 Mus musculus RIKEN cDNA 4932429P05 gene (4932429P05Rik), mRNA.(Mus musculus adult male testis cDNA)SMEK homolog 3, putative	1 1	1 1	0 0	
10	chrX:87564706-87579103	14,397	20	-3.70		1		(up site of Mageb1,Mageb2)				
11	chrX:87598220-88271067	672,847	15	-1.0483		43	Mageb5 Mageb1	melanoma antigen, family B, 5 melanoma antigen, family B, 1	2 2	2 1	0 1	CTA CTA
12	chrX:88271564-88285701	14,137	21	-3.76		1		(up site of Mageb1,Mageb2)				
13	chrX:103040719-103447968	407,249	10	-1.186		158	A630033H20Rik Gpr23 P2ry10 Zcchc5	hypothetical protein LOC213438(product:similar to PUTATIVE PURINERGIC RECEPTOR P2Y10) G protein-coupled receptor 23 purinergic receptor P2Y, G-protein coupled 10 zinc finger, CCHC domain containing 5	1 1 1 1	1 1 1 1	0 0 0 0	
14	chrX:142924085-145405975	2,481,890	112	-1.0089	2	1666	Ott	ovary testis transcribed	17	16	1	CTA
15	chrX:149333243-150403389	1,070,146	41	-1.4149		211	Magea Samt4	melanoma antigen family A 2,3,5,6,8 Mus musculus spermatogenesis associated multipass transmembrane protein 4 (Samt4), mRNA.(hypothetical protein LOC75185)	7 1	5 1	2 0	CTA
16	chrX:160673416-161229042	555,626	13	-1.13		274						

*noncording RNA

Table 4. Gene expression in LoDs on the X chromosome

LoD No	position	Gene families	Testis Expression			Brain Expression		
			core	extended	full	core	extended	full
1	chrX:3035387-4561626	Gmcl1l	4.19	0.46	0.46	-0.13	0.08	0.02
2	chrX:7857031-7924410	Ssx9	0.92	0.08	0.15	0.12	-0.01	0.07
3	chrX:8116622-8236784	Fthl17	0.46	0.18	-0.14	0.14	0.13	-0.01
4	chrX:22991291-32117922	Xmr Gmcl1l LOC236749	3.30	1.09	0.28	-0.06	0.02	0.04
5	chrX:50380769-52184422	similar to Xmr protein		1.37	0.25		-0.09	0.00
6	chrX:57970833-58090999	Ldoc1	0.30		-0.07	0.48		0.05
7	chrX:58775958-58815172	*1700019B21Rik		0.37	0.02		0.12	0.03
8	chrX:72448732-72544743	LOC238829	1.16	0.44	0.31	-0.10	0.16	0.08
9	chrX:84663998-86056286	Pet2 4932429P05Rik	0.96	0.66	0.08	-0.01	0.17	0.02
10	chrX:87564706-87579103				0.24			-0.02
11	chrX:87598220-88271067	Mageb5 Mageb1	0.46	0.49	0.24	0.01	0.17	0.00
12	chrX:88271564-88285701				0.33			0.05
13	chrX:103040719-103447968	A630033H20Rik Gpr23 P2ry10 Zcchc5	-0.43	-0.09	-0.03	-0.37	-0.02	-0.04
14	chrX:142924085-145405975	Ott	1.32	0.68	0.37	0.07	0.04	0.07
15	chrX:149333243-150403389	Magea Samt4	0.70	0.96	0.14	0.06	-0.01	0.02
16	chrX:160673416-161229042			0.19	0.08		0.28	0.08

Gene expression data were obtained from the Affy Exon tissue track[62]. Exon probe set intensities are represented as log ratios relative to median values across the dataset. Expression values of exon probe sets contained in each LoD is averaged and color-coded (high in red=3; low in green=-3). *Testis and Brain Expression have three categories. Exon probe set expression data in each category, i.e. core, extended, full, are retrieved from the following gene datasets described below.

Core: RefSeq transcripts, full-length GenBank mRNAs. Extended: dbEst alignments, Ensembl annotations, syntenic mRNA from human, rat and mouse, microRNA annotations, MITOMAP annotations, Vega genes, Vega pseudogenes. Full: Geneid genes, Genscan genes, Genscan Subopt, Exoniphy, RNA genes, SGP genes, Twinscan genes.

Table 5. Primer sets used for quantitative RT-PCR

Gene Symbol	Source	Forward or Reverse	primer	product sequence	PCR product size (bp)	Universal probe library set probe No of Mouse
Ssx9	NM_199063.2	F	tgtcaccatgaaccaaccag	tgtcaccatgaaccaaccagttttcatgcgtggcaaggagc aggccaagcaatccctgttcgaaggcattgaagttcatgac agtgatgttgaatgctttgaagg	105	68
		R	ccttcaaagcattcaacatcac			
Fthl17	NM_031261.2	F	gggactgtgttctttgctcct	gggactgtgttctttgctcctgccctgcccctgtgagcgggtg gggaatcgctgagcctgtggttgtacagc	72	100
		R	gctgtacaaccacaggctca			
Gm4836(Xmr)	NM_009529.3	F	atgagaatatgccgcctcac	atgagaatatgccgcctcacgtagaagcagatgaagatat aagagatgaacaagacagtatgttggataaatctggagaa aacgtaagtttctcagaggaatggcagcgttttgcacgttctg tagaga	129	6
		R	tctctacagaacgtgcaaaacg			
Mageb1	NM_010759.1	F	ttcctgtctgccagctcttt	ttcctgtctgccagctcttttacttcagccctgagcacagtcaa catgcctaggggtcaaaa	62	99
		R	ttttgacccctaggcatgtt			
Magea8	NM_020020.3	F	tgccttgccatttgaggt	tgccttgccatttgaggtccctggtatcaaggagctgaaatcc tggtaaagcagagtttaactaga	66	17
		R	tctagttaaactctgctttaccagga			
Ott	NM_011022.1	F	gacacacctcagcaagtgga	gacacacctcagcaagtggatctctgtgctttgaggtggctg ggactcccacaggaggtggcaaagtgatggcgaaccatg aagacgaaattagacagctga	102	10
		R	tcagctgtctaatttcgtcttca			
Xlr4a	NM_001081642.	F	caaaatcaagggcagacctc	caaaatcaagggcagacctcccaagcaaccaaaggtga ctcctgctctgccatctgacgattcacagcagcttc	74	53
		R	gaagctgctgtgaatcgtca			
Rhox2a	NM_029203.2	F	agagcttcaatgtgctgcaa	agagcttcaatgtgctgcaactacaagagctggagagcat cttccagtgcaatcactacatcagcactaaggaggcaaatc gcct	85	78
		R	aggcgatttgcctccttagt			

Universal probe library set, Mouse (catalogNo. 04 683 641 001) from Roche applied Science.

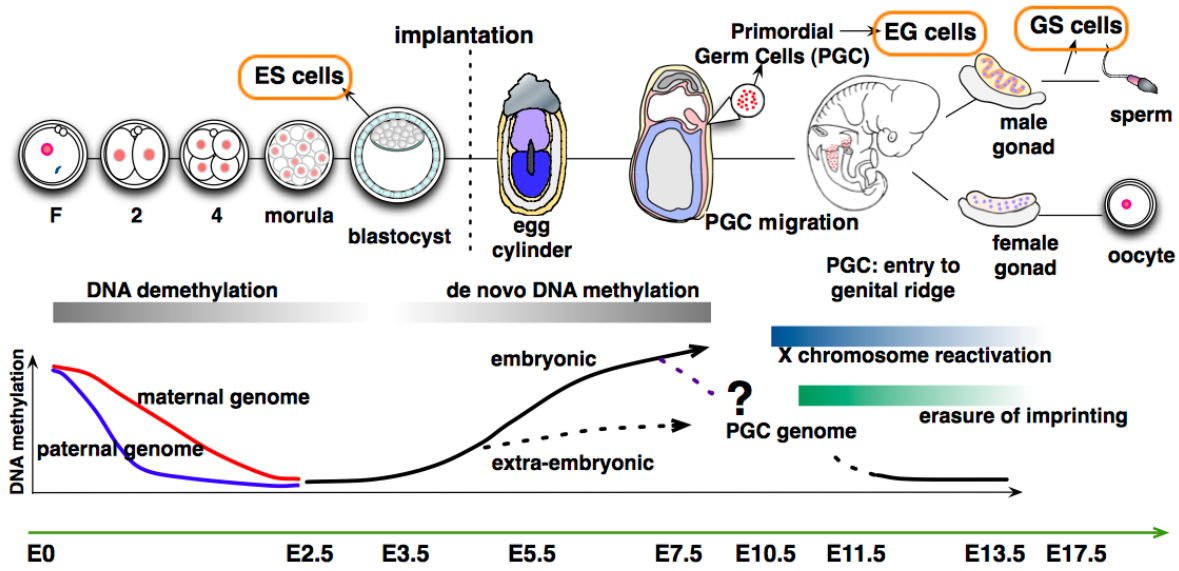


Figure 1. Early development and germ cell differentiation in mice.

Developmental processes of mouse early embryos and germ cells are schematically shown. Fertilized eggs (F) undergo cleavage stage to give rise 2, 4, 8, 16 cell stage embryos and morula comprised of about 32 cells. Blastocyst consist of trophectoderm and inner cell mass is formed at embryonic (E) day 3.5. From inner cell mass, pluripotent embryonic stem (ES) cells can be derived. After implantation to uterus, blastocyst becomes egg cylinder stage embryos, and gastrulation begins at around E7.5. Primordial germ cells (PGCs) emerge at the base of allantois as a population of about 45 cells at E7.25. PGCs migrate through mesentery to reach genital ridges (=gonads) at E10.5 and entry into gonads is completed at E11.5. Embryonic Germ (EG) cells are pluripotent stem cells with characteristic highly similar to ES cells. EG cells can be derived from E8.5-E12.5 PGCs. PGCs undergo gametogenesis process within gonads to produce oocytes and sperm eventually. Germline Stem (GS) cells are self renewing stem cells with distinct characterstics from ES or EG cells. GS cells can be derived from spermatogonia of new born testis. When transplanted to testis, GS cells can be differentiated into functional sperm. During developmental processes described above, DNA methylation levels of cells in this cell lineage are dynamically changing. Upon fertilization, DNA demethylation initiates, although kinetics of demethylation appears to be different between paternal (blue line) and maternal (red line) genome. Global level of DNA methylation becomes very low at E2.5-E3.5. At late blastocyst stage, de novo DNA methylation is supposed to occur: DNA methylation levels of epiblast (blue portion of egg cylinder) and embryonic ectoderm (light blue part of E7.5 embryo) show higher methylation level compared with blastocyst or extra-embryonic part of these embryos. PGCs are known to be derived from embryonic ectoderm, and undergo global DNA demethylation, but precise kinetics of DNA demethylation has not been determined. During PGC development, epigenetic reprogramming revealed by X chromosome reactivation (blue gradient) or erasure of genomic imprinting (green) is known to take place.

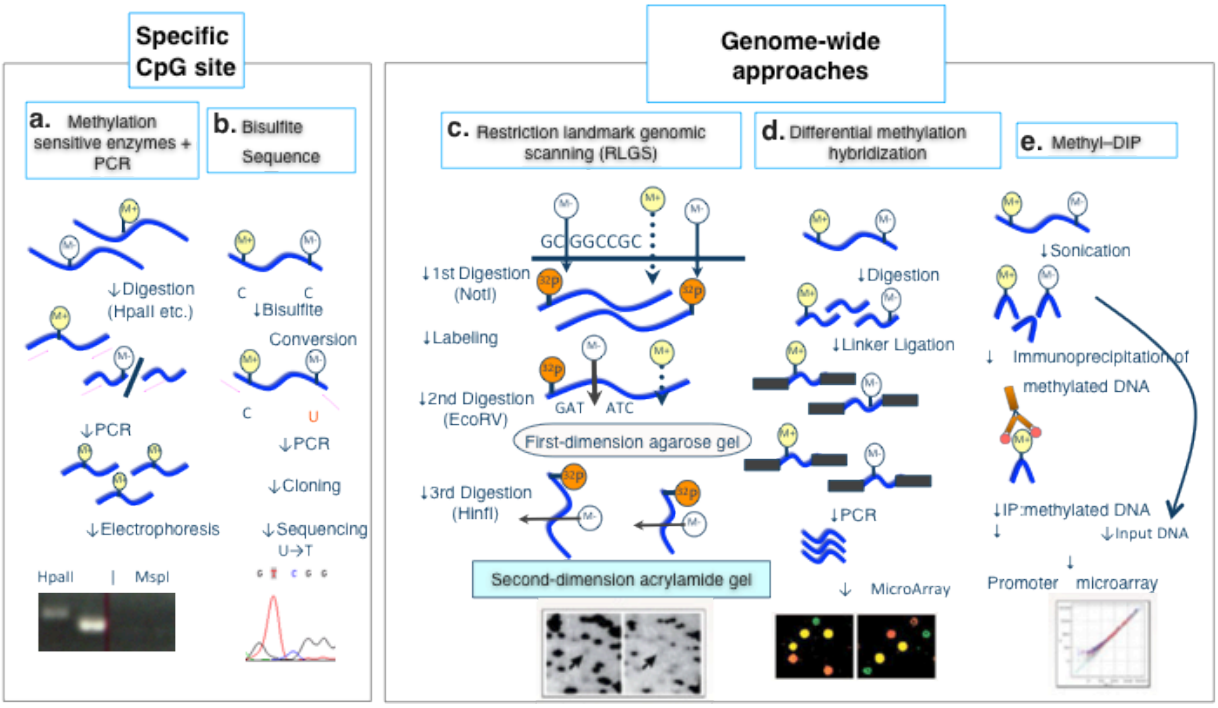


Figure 2. Techniques for DNA methylation analysis

Analytical technique for specific CpG site: (a) Digestion with methylation-sensitive enzyme followed by PCR. Genomic DNA is digested by either methylation-sensitive *Hpa*II or insensitive isoschizomer *Msp*I. After digestion, DNA is PCR amplified using primers that flank the recognition site for the enzymes. If the CpG is methylated, PCR product can be obtained from *Hpa*II-digested sample, whereas no amplification will be obtained from the *Msp*I sample.

(b) Bisulfite sequence analysis. By treating genomic DNA with bisulfite reagent, unmethylated cytosine will be converted to U, whreas methylated C remains as C. DNA is PCR amplified using primers that flank the test CpG sequence, cloned into vector, and sequences of the amplified products are determined.

Genome-wide approaches: (c) Restriction landmark genomic scanning (RLGS). Genomic DNA is first digested by *Not*I, a rare cutter with methylation sensitivity, and the ends of the fragments are radiolabeled with ^{32}P. The DNA is digested with 6 base cutter such as *Eco*RV and separated on an agarose gel. The agarose strip of 1st-dimension gel is subjected to restriction digestion with another enzyme such as *Hinf*I, and then to second-dimension agarose gel electrophoresis. After autoradiography of the 2D gel, presence or absence of spots are examined to seek for differentially methylated spots. (d) Differential methylation hybridization. Genomic DNA is digested first by methylation insensitive enzyme such as *Mse*I and the fragments are tagged by linker adapter. Then the linker-ligated DNAs are digested by methylation-sensitive enzyme, and amplified by linker sequence as a primer so that only hypermethylated sequences will be amplified. The amplified products will be hybridized with CpG island microarray to detect methylated CGIs.

(e) Methyl-DIP analysis. Genomic DNA is fragmented by sonication, and the DNA fragments are subjected to immunoprecipitation with anti-5' methyl-cytosine antibody. The precipitated, methylated DNA is hybridized with promoter microarray to detect methylated promoters.

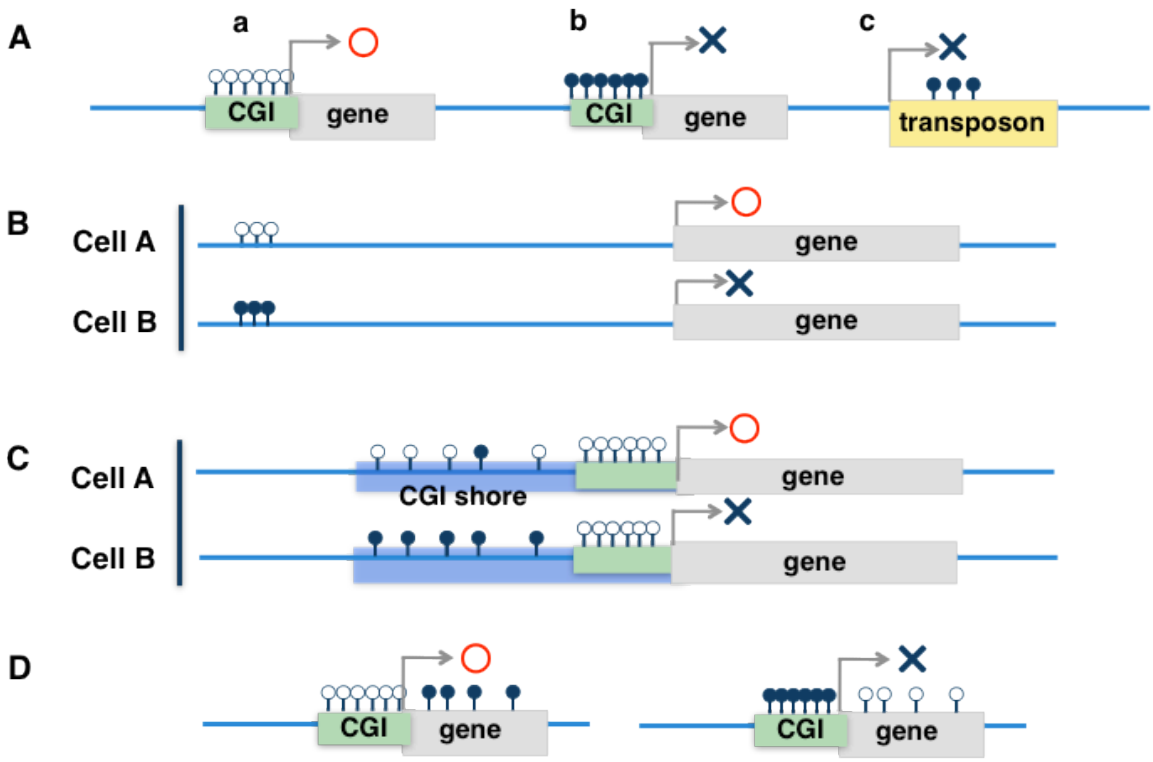


Figure 3. DNA methylation and gene expression regulation.

A. It is thought that most parts of the mammalian genome are hypermethylated. However, there are regions that contain a high frequency of CpG sequences within mammalian genome and these clusters of CpGs are designated as CpG islands (CGIs) shown as light green boxes. (a) CGIs are mostly methyl-free (open circle), and gene expression can be achieved. (b) A small fraction of CGIs may be methylated (solid circle), and, normally, genes with methylated CGIs are not transcribed (X). (c) Mammalian genome is replete with transposon-derived repetitive sequences, and these sequences are normally hypermethylated and not transcribed.

B. Differential DNA methylation of nongenic sequences can be detected in intergenic regions. This differential methylation may be cell differentiation-dependent, and correlates with expression of genes distantly located.

C. While CGIs are almost always unmethylated, differential methylation in regions close to CGIs have been reported. These “CGI shores” show differential methylation in tissue- or cancer-specific manners [28].

D. Promoters of actively transcribed genes are normally DNA hypomethylated (left). However, it is unexpectedly found that bodies of active genes tend to be hypermethylated compared with those of inactive genes (right) [2].

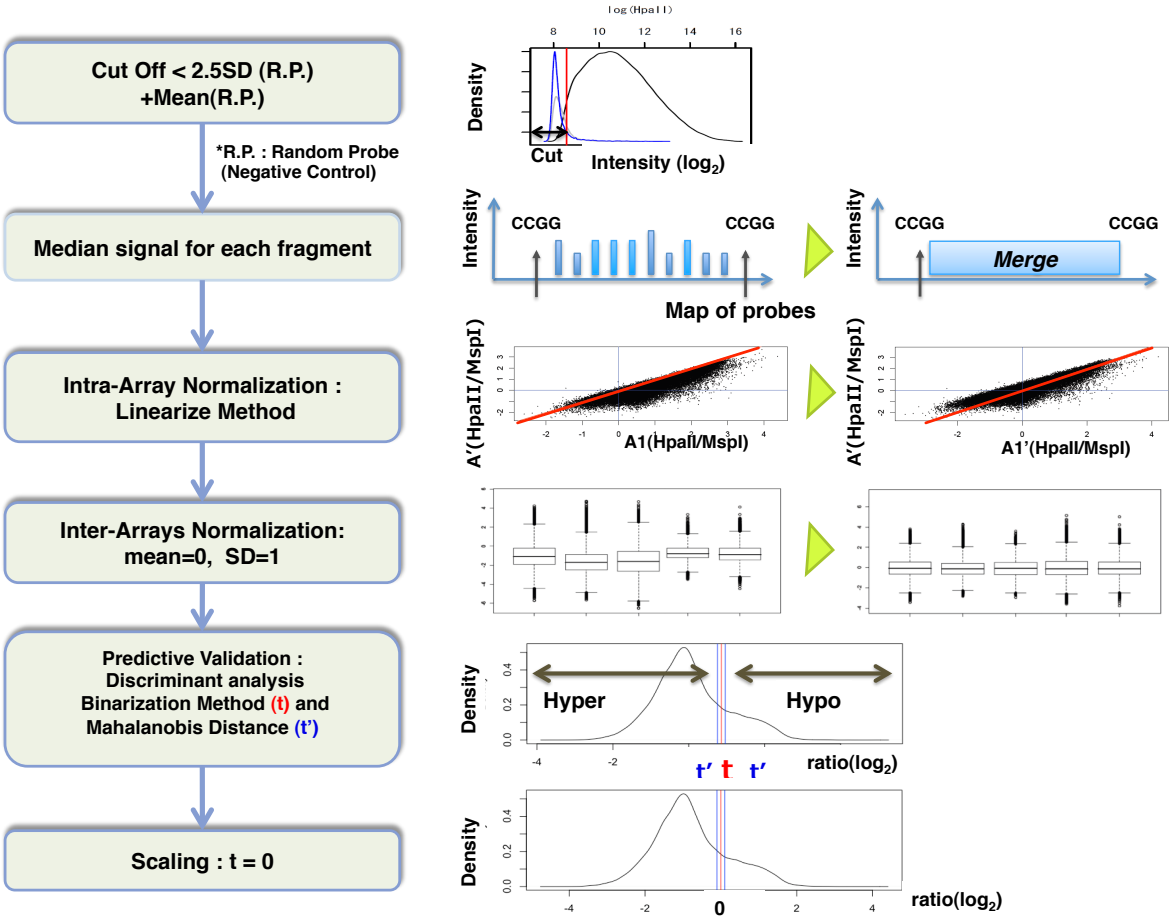


Figure 4. Flow of data analysis in the modified nanoHELP method

The steps of the data analysis are shown schematically. After removing the background signal noises, the median signal intensity of the 10 probes assigned for each CCGG segment is calculated and used to define the segment's signal intensity. After normalizations of the microarray data, hypomethylated and hypermethylated segments are distinguished using an R script that determines the threshold values based on a binarization method [60]. The marginal width of the threshold is calculated using the Mahalanobis distance [61]. The $\log_2$ value at the threshold is set as 0 so that hypomethylated segments have a positive value (>0) and hypermethylated segments have a negative value (<0).

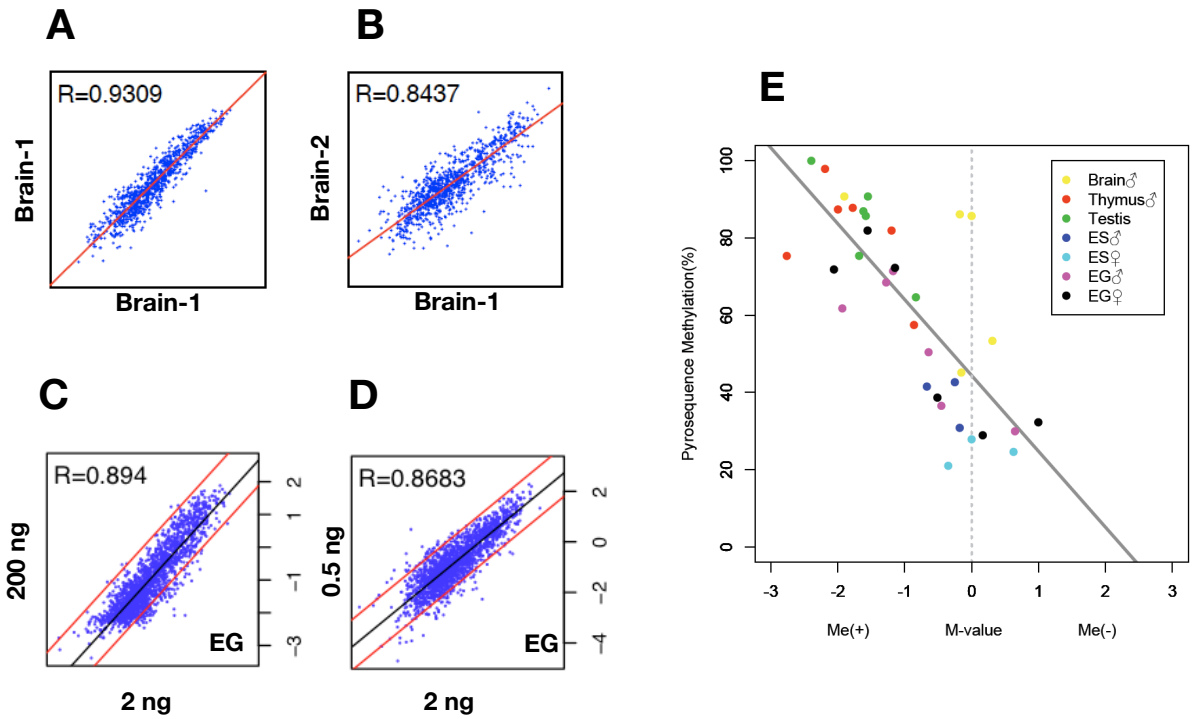


Figure 5. Validations of the modified nanoHELP method

A, B. Reproducibility of the nanoHELP experiments. Technical replicates (A). Brain DNA sample (2 ng) from the same mouse was processed separately and used for the two independent microarray experiments. A scattergram of the $\log_2$ (*Hpa*II/*Msp*I) is shown. Biological replicates (B). Two-nanogram samples of brain DNA from two different C57BL/6 mice were used separately for the two independent assays.

C, D. The modified nanoHELP method with a limited amount of sample. One microgram of EG cell genomic DNA was restriction digested and ligated to the adapters, and 10 nanogram of the EG DNA was processed similarly. Either a 200 ng or 2 ng equivalent of the adapter-tagged DNAs was used for the nanoHELP experiments. In addition, 0.5 ng of the EG DNA was restriction digested, ligated to the adapter, and amplified for the nanoHELP experiment. Scatter plots of the $\log_2$ (*Hpa*II/*Msp*I) are depicted in Fig. 5C (200 ng vs 2 ng) and Fig. 5D (0.5 ng vs 2 ng). The results show the profiles obtained from three different DNA amounts to be highly concordant.

E. Bisulfite pyrosequencing analysis. Six CpG sites were selected and the methylation status of these sites in seven different cells and organs were determined by bisulfite pyrosequencing. The y-axis represents the methylation percentage obtained by pyrosequencing, and the x-axis represents the M-values. Primers used for the bisulfite pyrosequencing analysis are listed in Table 2.

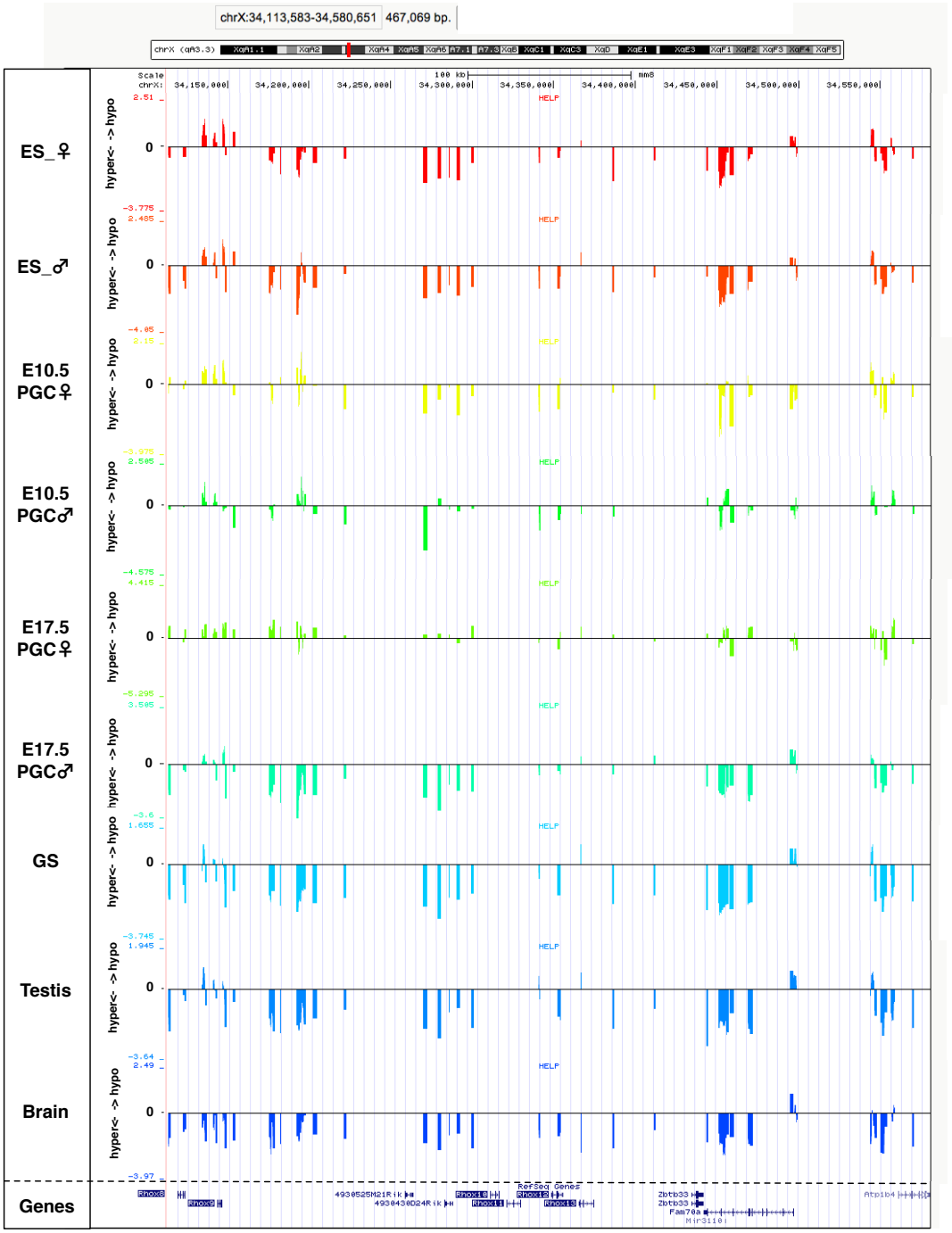


Figure 6. Data presentation using the UCSC genome browser

The modified nanoHELP data were exported to the UCSC genome browser to allow presentation of methylation pattern of the 467 kb region (chrX: 34,113,583–34,580,651) corresponding to *Rhox* gene clusters. (mm8)

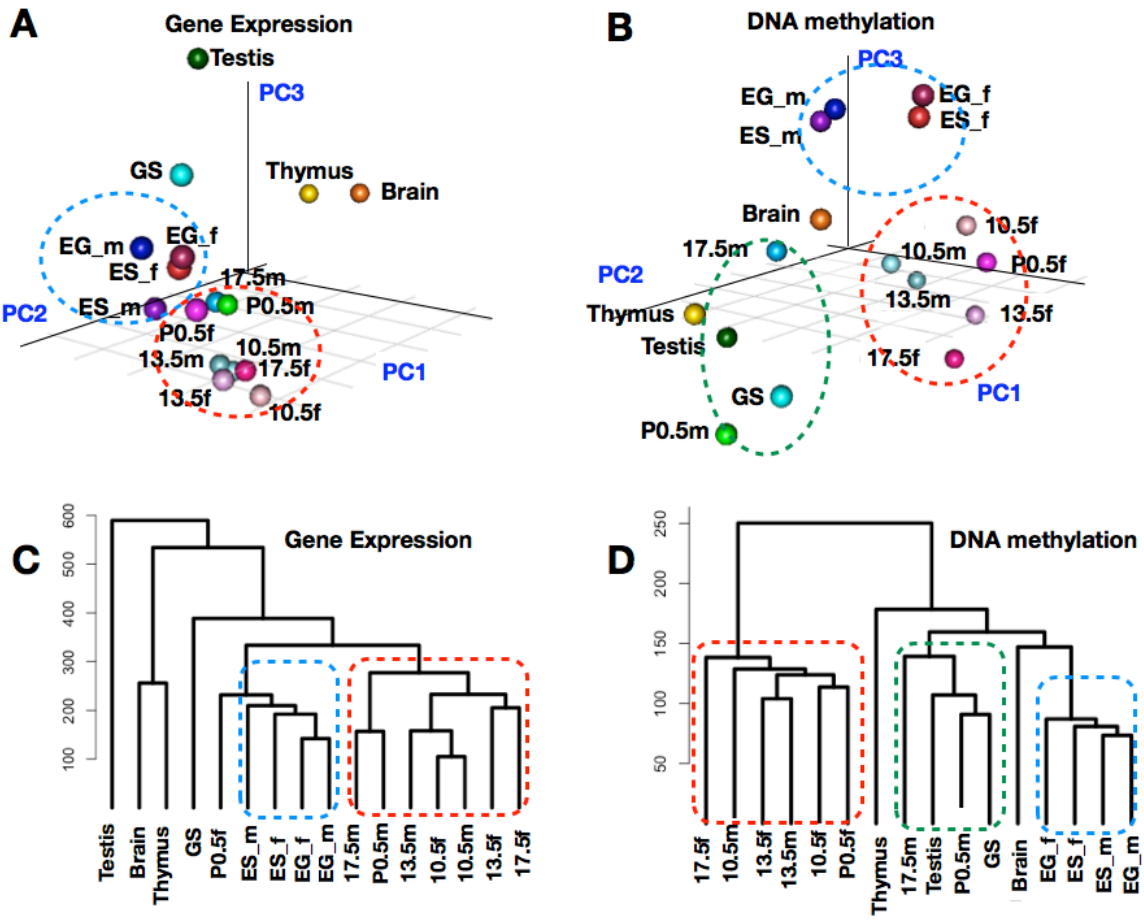


Figure 7. Profiling of gene expression and DNA methylation in germ cells and stem cells.

(A) PCA of the expression profiles of germ cells, stem cells and adult organs. ES_m (male ES cells), ES_f (female ES cells), EG_m (male EG cells), EG_f (female EG cells), 10.5m (PGCs from male E10.5 embryos), 10.5f (female E10.5 PGCs), 13.5m (male E13.5 PGCs), 13.5f (female E13.5 PGCs), 17.5m (male E17.5 PGCs), 17.5f(female E17.5 PGCs), P0.5m (spermatogonia from P0.5 neonates), P0.5f (oocytes from P0.5 neonates) and GS cells. Testis, thymus and brain were isolated from male adult mice.

(B) PCA analysis of DNA methylation profiles of germ cells, stem cells and adult organs.

(C) Hierarchical clustering of gene expression profiles.

(D) Hierarchical clustering of DNA methylation profiles.

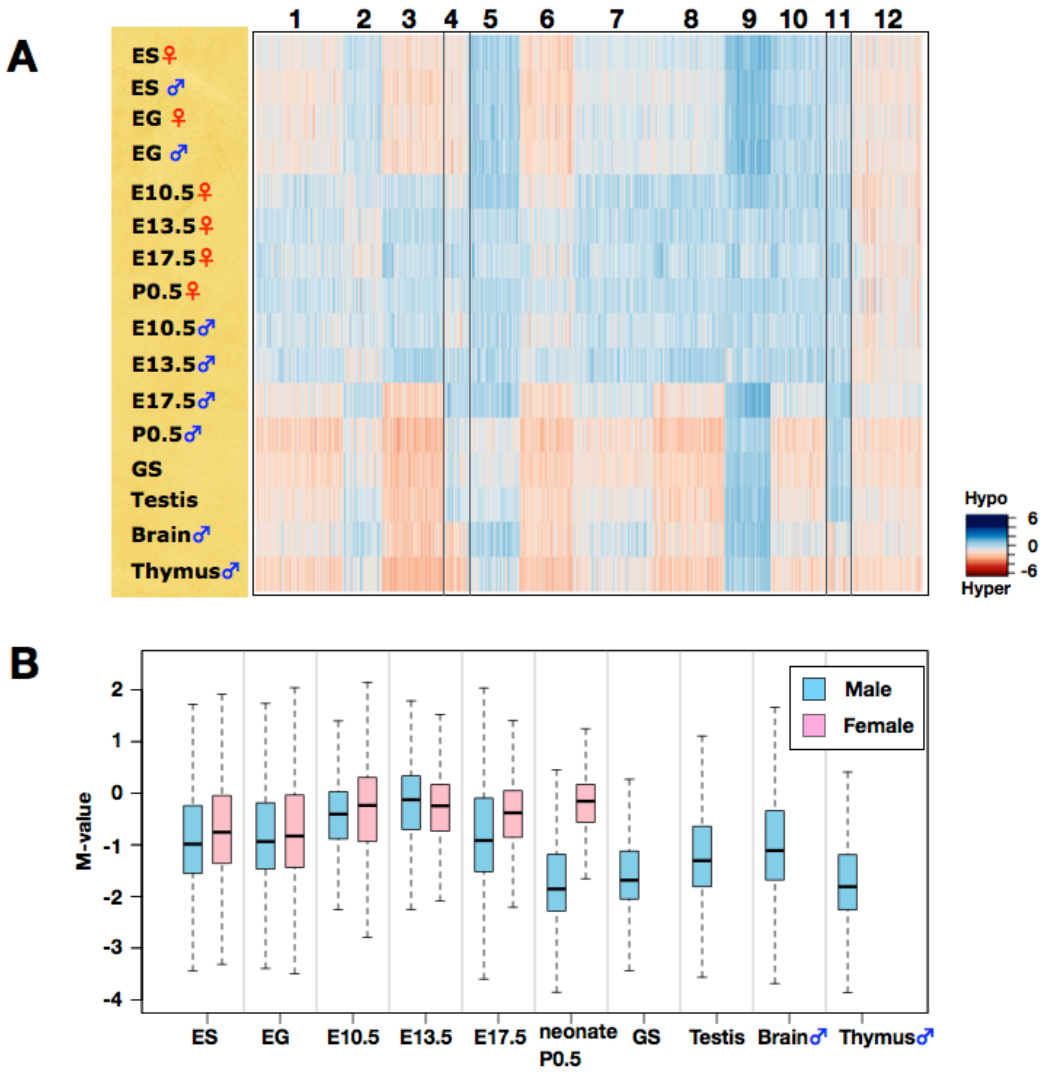


Figure 8. DNA methylation dynamics during germ cell development.

(A) k-means clustering of DNA methylation profiles of germ cells, stem cells and adult organs. DNA methylation levels of the CCGG segments are represented as a heat map(unmethylated segments in dark blue, M-value = 6.00; highly methylated segments in dark red, M-value = –6.00). (B) Changes in global DNA methylation levels during PGC development. The M-value was calculated for each sample and is shown as a box plot. The bottom and top of the boxes are the 25th and 75th percentile, respectively.

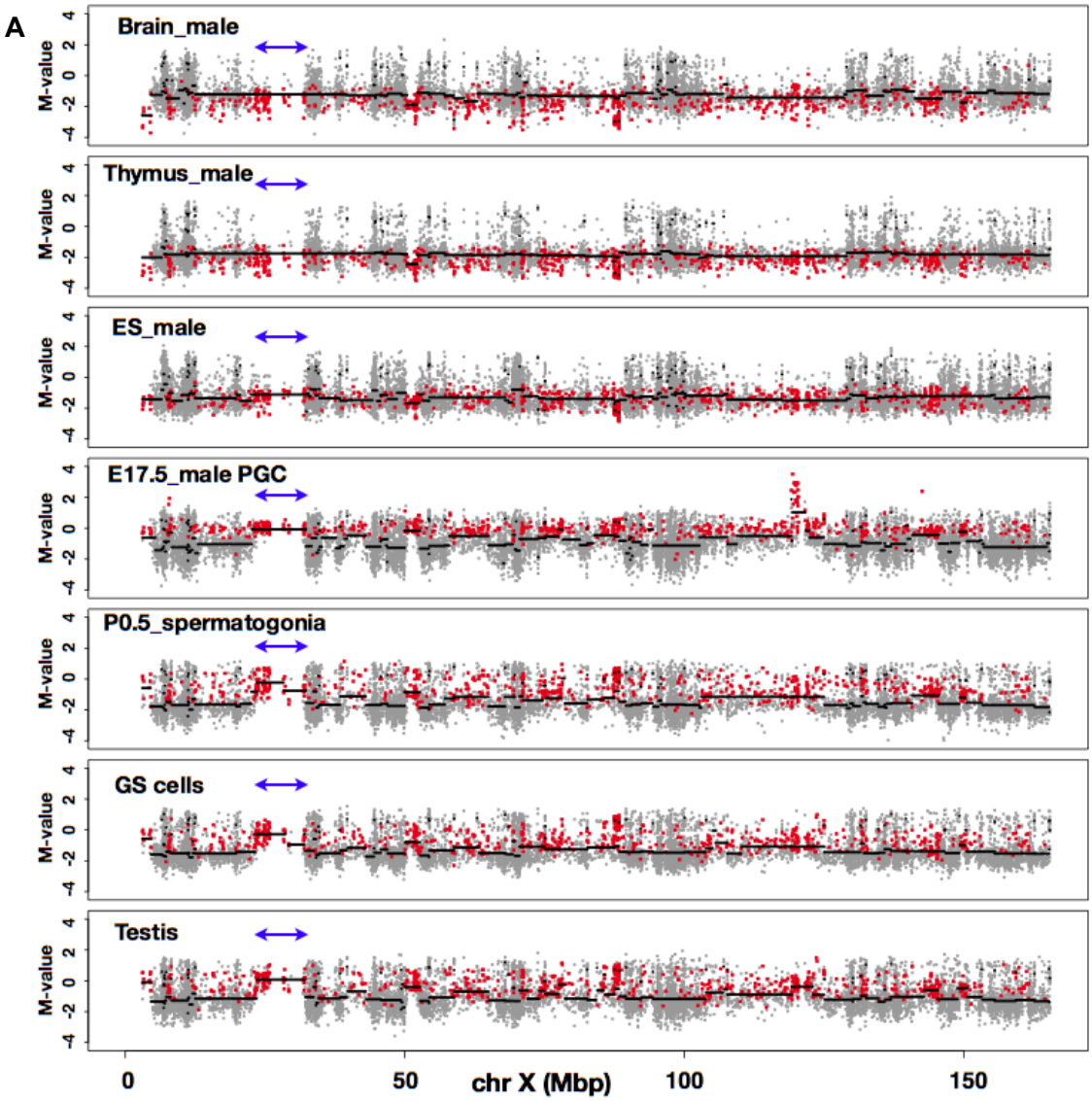


Figure 9. Discovery of large, contiguous genomic regions with low DNA methylation modifications in male germ cells.

The methylation profiles of genomic DNA along the mouse X chromosome. Samples used for the analysis are indicated in each figure. The M-value of each CCGG segment obtained from the analysis of each sample DNA is plotted on the mouse X chromosome (grey dots). The y-axis represents the M-value; log2(HpaII/MspI). The black line was drawn using DNAcopy, a circular binary segmentation program obtained from http://www.bioconductor.org/packages/ 2.3/ bioc/html/DNAcopy.html.35 Red dots represent CCGG segments belonging to Cluster 4. Double-headed arrows indicate the position of the Xmr gene cluster.

(A) Mainly male tissue, ES and male germ cells. (Brain, Thymus, ES, E17.5 PGC, P0.5 spermatogonia, GS cells, Testis)

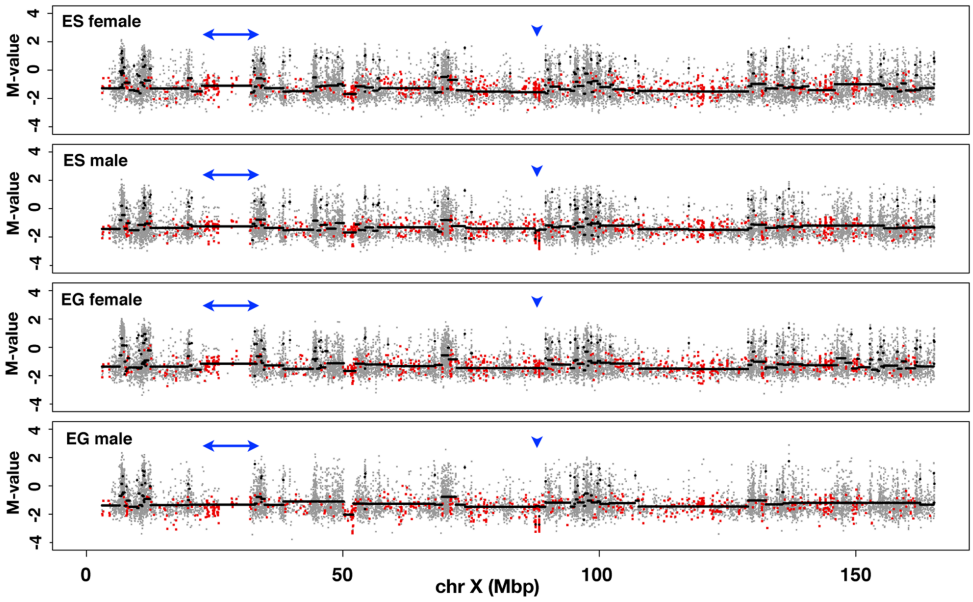
(B) ES cells and EG cells of female and male.

(C) Male PGC of E10.5, E13.5, E17.5 and P0.5 spermatogonia.

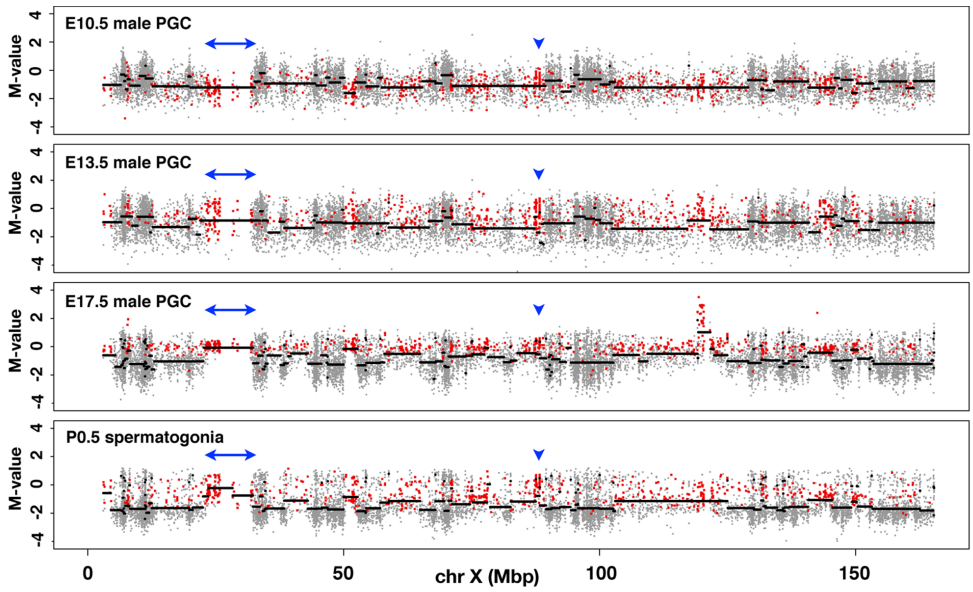
(D) Female PGC of E10.5, E13.5, E17.5 and P0.5oocyte.

Figure 9.

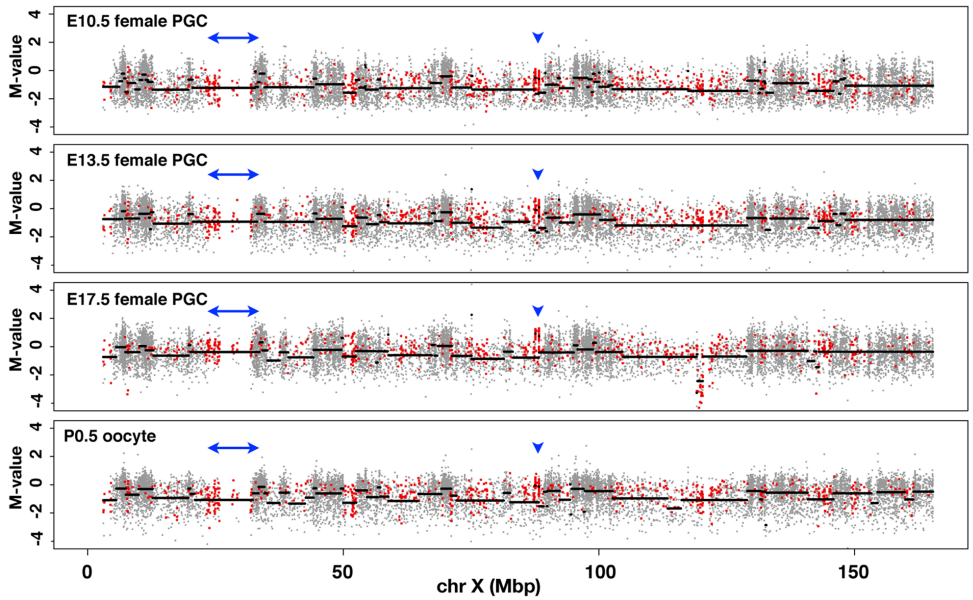
B


ES female
ES male
EG female
EG male
M-value
chr X (Mbp)

C


E10.5 male PGC
E13.5 male PGC
E17.5 male PGC
P0.5 spermatogonia
M-value
chr X (Mbp)

D


E10.5 female PGC
E13.5 female PGC
E17.5 female PGC
P0.5 oocyte
M-value
chr X (Mbp)

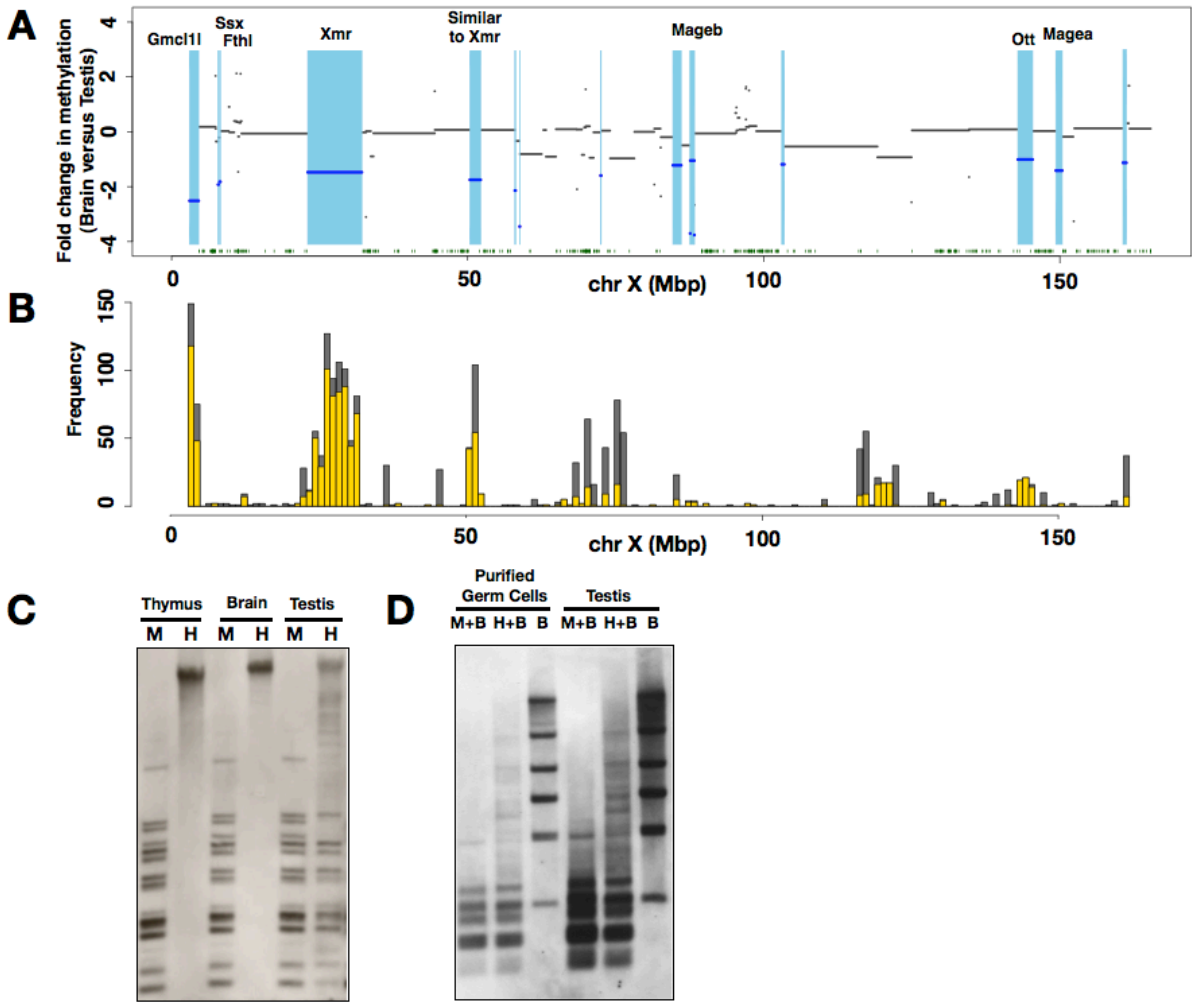


Figure 10. Demonstration of LoDs on the mouse X chromosome.

(A) Differences in methylation levels between brain and testis genomic DNA along the mouse X chromosome. Fold changes in the methylation level, i.e. brain M-value versus testis M-value, were calculated for each CCGG segment, and plotted using the log2 scale along the X chromosome. The blue lines indicate genomic regions showing more than a 2-fold difference between the brain M-value and testis M-value. Light blue represents hypomethylated regions in the testis relative to the brain. Green dots at the bottom represent the positions of CGIs. (B) The positions of segmentally duplicated regions along the mouse X chromosome. Segmentally duplicated regions >1000 bases with >98% similarity are counted, and the frequencies of duplications (y-axis) are shown. (The data were from UCSC Genome Browser.) Grey bars represent duplications occurring on the X and other chromosomes, and yellow bars represent the frequencies of duplications mapped only on the X chromosome. (C) Methylation analysis of LoDs 10 and 12 by Southern blot hybridization. Genomic DNAs of the male thymus, male brain and testis were digested by either methylation-sensitive *Hpa*II (H) or methylation-insensitive isoschizomer, *Msp*I (M). The Southern blot was hybridized with a probe targeted to LoDs 10 and 12. A primer pair (FW: 50-GCTGGGTCCAGCTTCCCTGG-30 , RV: 50-TGGCACCCCTCCTGCCTGAT-30) was used to amplify a 807-bp sequence using testis cDNA for generation of the probe. The 807-bp probe contains locally repeated sequences and corresponds to both LoDs 10 and 12 located upstream of

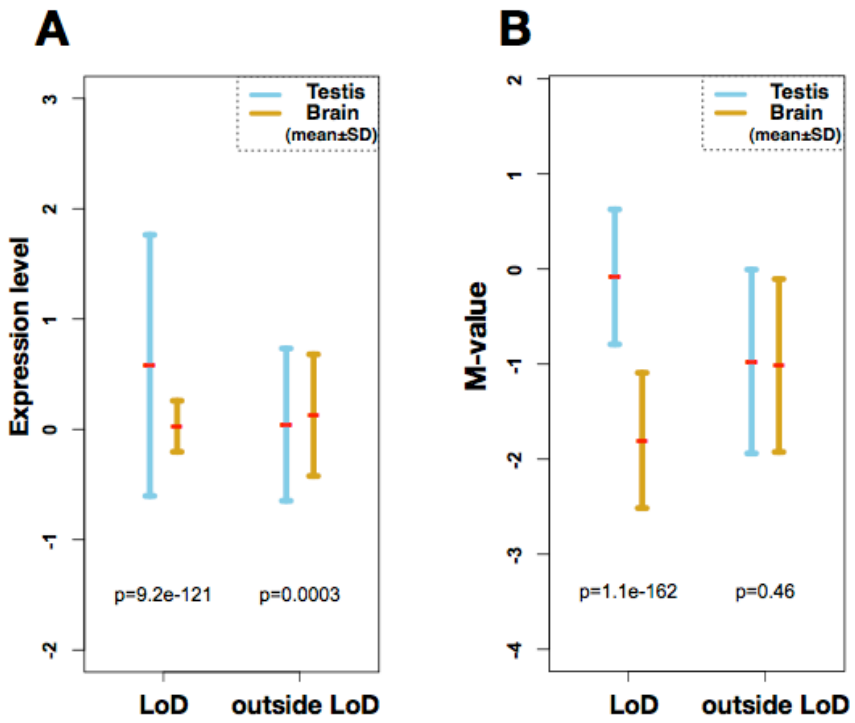


Figure 11. Inverse relationship between DNA methylation and expression of genes within the LoDs.

(A) A plot of the expression levels of genes contained in LoDs, and regions outside LoDs on the mouse X chromosome. Gene expression data were obtained from the Affymetrix Exon array data set [62]. The expression values of exons contained in LoDs were averaged and plotted, and the data points from regions outside LoDs were similarly averaged. For statistical analysis of the data from regions outside LoDs, the same number of data points as those used to analyze within LoDs were randomly selected and used. (B)Mean methylation levels of genomic DNA within LoDs. The mean M-values from the CCGG segments contained in LoDs or in regions outside LoDs are shown. Statistical significance was tested by Wilcoxon t-test, and P-values are shown within the figures.

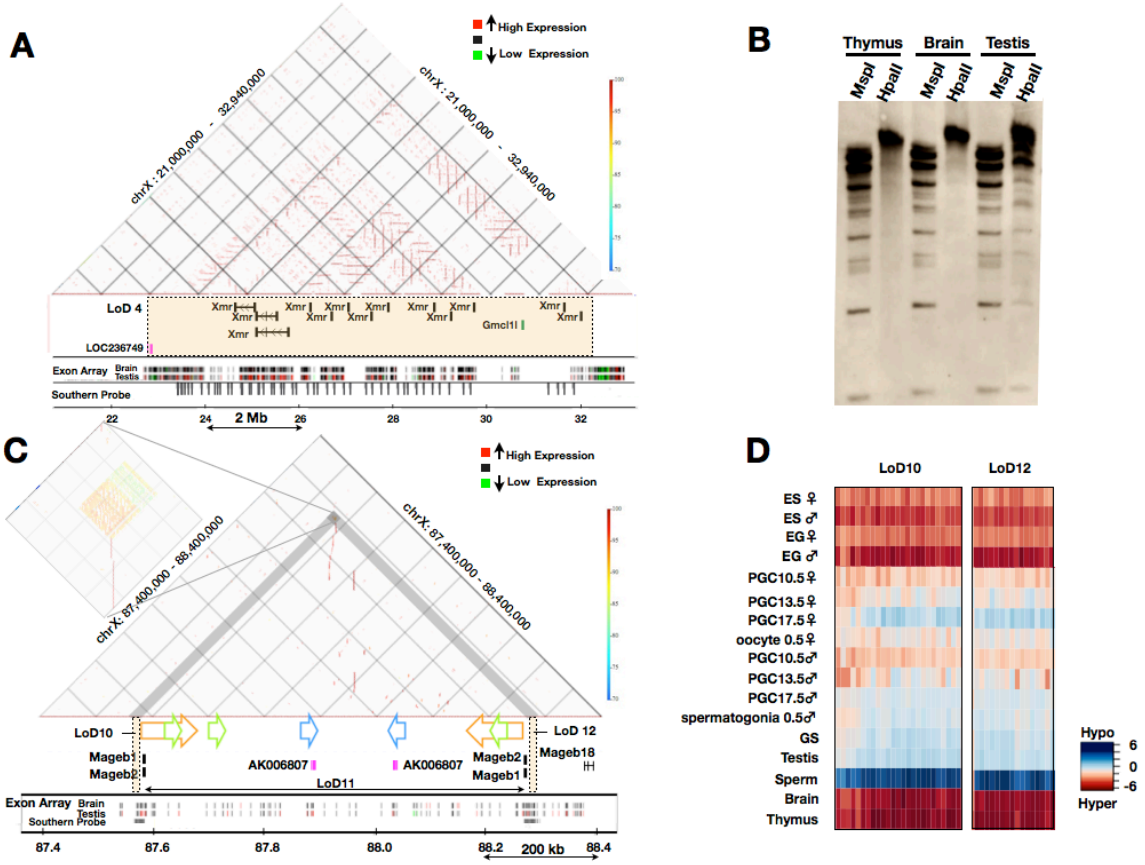


Figure 12. Genomic structures of LoDs: *Xmr* and *Mageb*.

(A) Genomic structure of LoD4 (*Xmr*). The top section represents a similarity dot plot [63]of LoD 4 and flanking regions (chrX: 21,000,000–32,940,000; UCSC mm8). Similarities of the sequences are color-coded as shown by the color bar on the right (high similarity in dark red, 100% similarity; low similarity in blue, 70% similarity). The horizontal lines represent direct repeats, and the vertical lines indicate IRs. The middle section depicts the locations of genes contained in LoD 4 (colored rectangle). Exon array expression data for the brain and testis [62] are shown at the bottom (high gene expression in red and low expression in green).The positions of sequences with homology to the *Xmr* cDNA probe are also shown by grey vertical bars (Southern probe). (B) Southern blot probed with the *Xmr* cDNA. Genomic DNAs from three organs were digested with either *Msp*I or *Hpa*II. *Xmr* cDNA probe (708 bp) was amplified from testis cDNA using primers, Xmr FW: 5'-AAGGGTGCAGTTGTGAAGGT-3' Xmr Rv: 5'-TGTTGGTCTCCATGTTCATCA-3'. The hybridization signals in the unresolved part of the testis DNA blot are likely to reflect non-specific cross-hybridization. Southern blot analysis data using DNA doubly digested by *BamH*I plus either *Hpa*II or *Msp*I confirmed this notion (data not shown). (C) Genomic structure of LoDs 10, 11 and 12 (*Mageb1/b2*). Top: similarity dot plot. The vertical lines and colored arrows indicate positions of IRs. The coloring of the similarities is the same as found in Fig. 12A. Homologous repeats are represented by the same color. A magnified view of the IRs contained in LoDs 10 and 12 is presented on the left. Gene expression data for the brain and testis are shown at the bottom. The positions of sequences with homology to a probe used for Southern blot analysis (Fig. 10C and 10D) are shown by grey bars (Southern probe). (D) A DNA methylation heatmap of LoDs 10 and 12. DNA methylation levels of the CCGG segments are represented as a heatmap (unmethylated segments in dark blue, M-value= 6.00; highly methylated segments in dark red, M-value = –6.00).

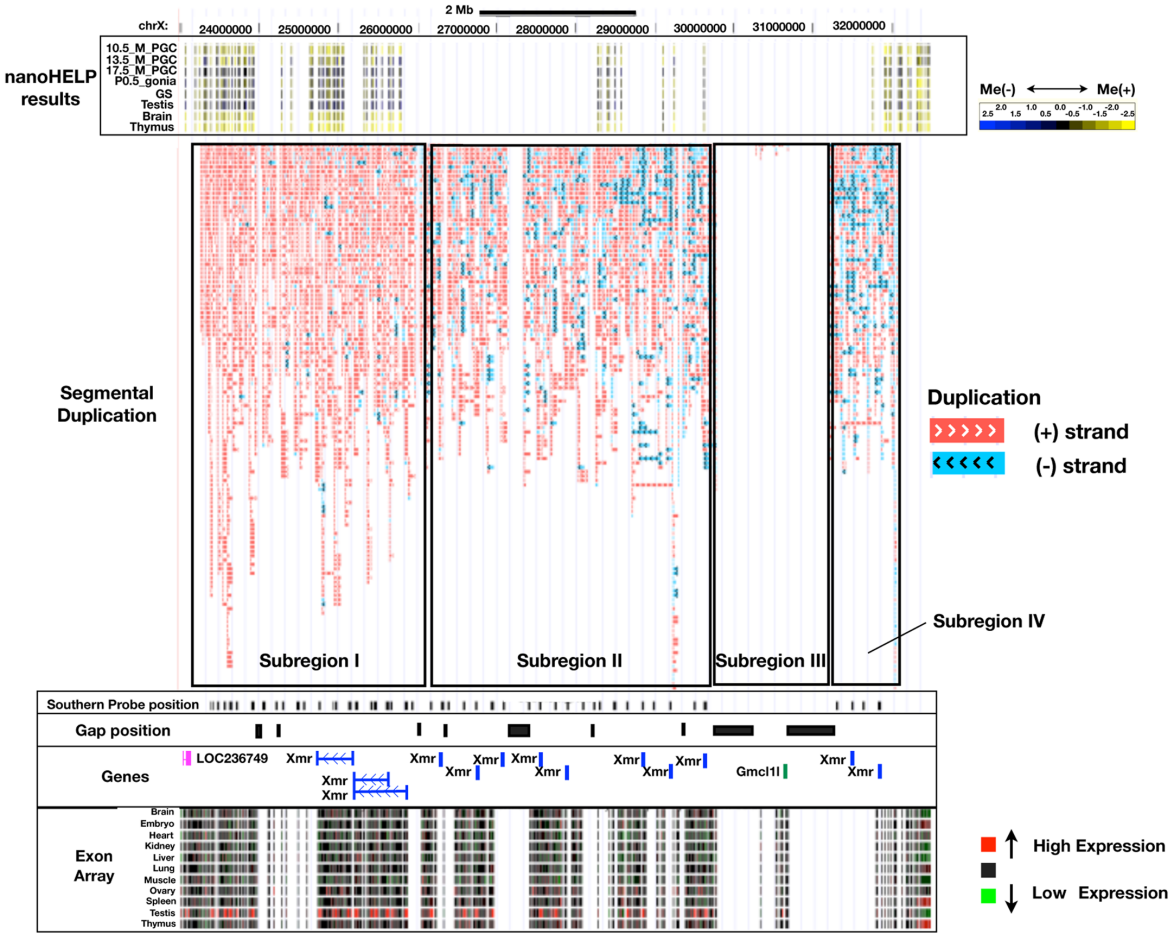


Figure 13. Subregions of LoD 4 (*Xmr*)

The LoD 4 region is divided into four subregions based on the genomic sequence context and patterns of duplications. The *Xmr* gene family is distributed within subregions I, II, and IV. Subregion I contains tandemly duplicated *Xmr* genes, and subregions II and IV have both tandem repeats and inverted repeats. The directions of the repeated units are color-coded (red, + strand; blue, – strand). Subregion III harbors the *Gmcl1l* gene family and sequences of the subregion are distinct from those of subregions I, II and IV. Subregion III is highly homologous to LoD 1, which also contains *Gmcl1l* genes. The positions of sequences with homology to *Xmr* cDNA are shown by gray vertical bars. Positions of sequence gaps are indicated by black rectangles. Exon array expression data for various tissues [62] are shown at the bottom (high gene expression in red and low expression in green).

Figure 14. A Magnified view of the *Xmr* loci

A part of the *Xmr* region (chrX: 24,630,478–25,900,440 (mm8)) is captured and shown with genomic annotations. All the CCGG segments in this region are hypomethylated in germ cells, regardless of their positions within the *Xmr* genes. Methylation level (M-value) is represented as color (high in yellow, M-value= -2.5; low in blue, M-value= 2.5). Exon array expression data for various tissues are shown at the bottom (high gene expression in red and low expression in green).

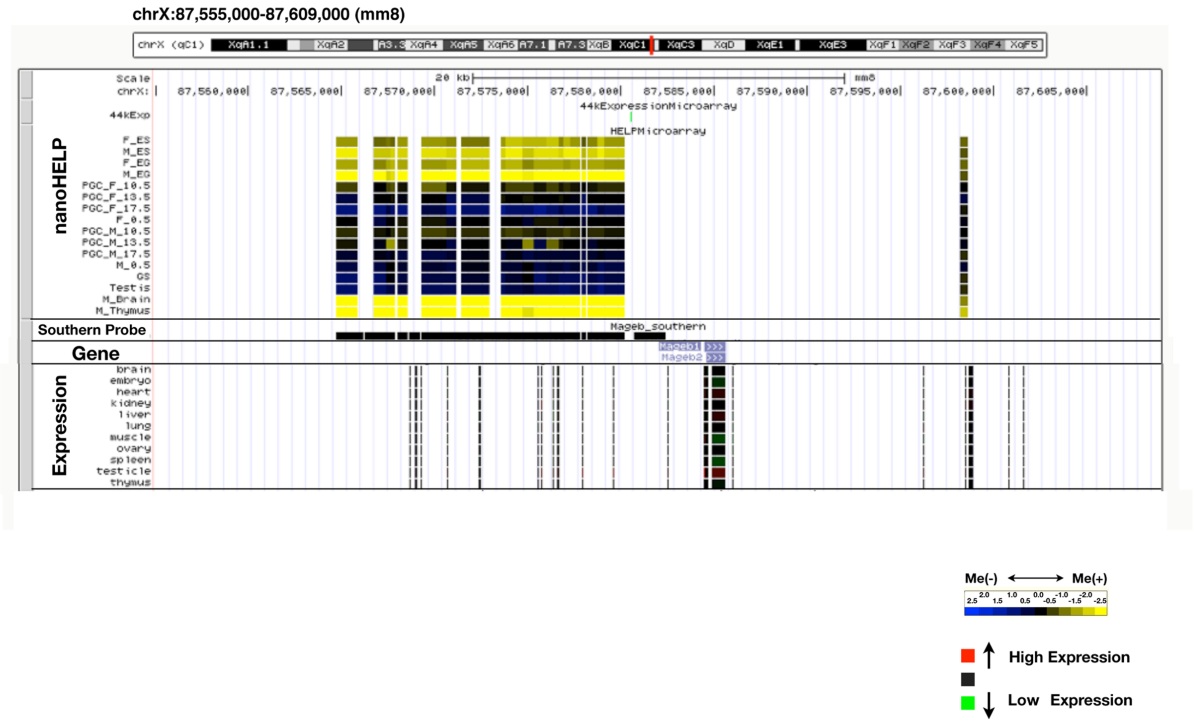


Figure 15. A Magnified view of *Mageb* locus

A part of the *Mageb* region (chrX: 87,555,000–87,609,000 (mm8)) is captured and shown with genomic annotations. Methylation level (M-value) is represented as color (high in yellow, M value=-2.5; low in blue, M-value= 2.5). Exon array expression data for various tissues [62] are shown at the bottom (high gene expression in red and low expression in green).

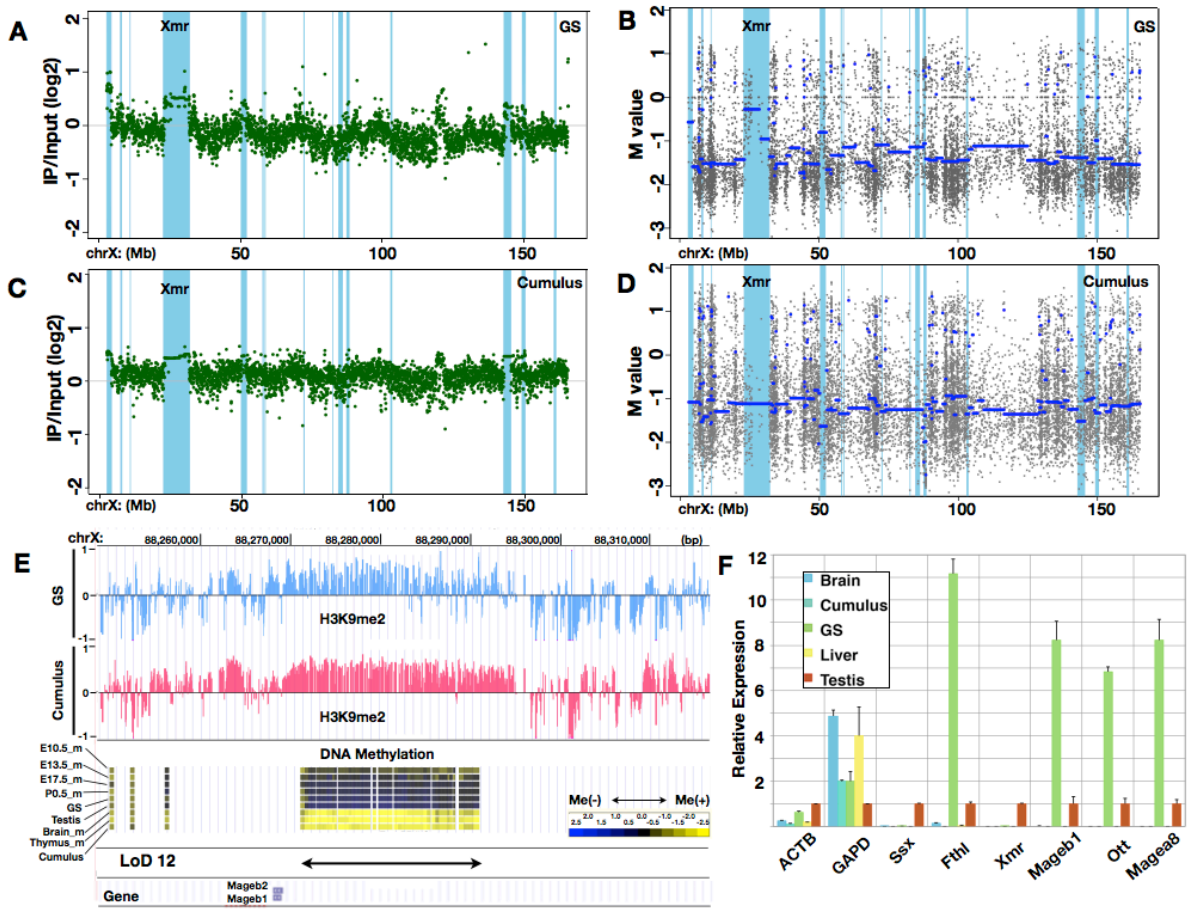


Figure 16. Large stretches of H3K9dimethyl modifications overlapping LoDs.
(A) ChIP-on-chip data for theH3K9me2modification along the X chromosome in GS cells. The green dots represent individual probe data. The positions of LoDs are shownin light blue. (B)DNA methylation profile of GS cells within the X chromosome. The blue line denotes the average M-values. The positions of LoDs are shown in light blue. (C) ChIP-on-chip data for H3K9me2modification along the X chromosome in cumulus cells. (D) The DNA methylation profile of cumulus cells within the X chromosome.(E) A magnified view of the LoD12region. Top, theH3K9me2modification in GS cells (blue) and in cumulus cells (pink). Middle, DNA methylation data for E10.5 male PGCs, E13.5 male PGCs, E17.5 male PGCs, P0.5 spermatogonia, GS cells, testis, brain, thymus and cumulus cells. Bottom, positions of LoD and *Mageb1/b2* genes. (F) Quantitative RT-PCR expression analysis of genes in LoD regions. The *Ssx9* (Ssx), *Fthl17* (Fthl), *Xmr*, *Mageb1*, *Ott* and *Magea8* genes are located within LoDs on the X chromosome. Beta-actin (ACTB) and *Gapdh* (GAPD) genes were also examined. The expression level in the testis is set at 1.0, and the expression levels in other cells and tissues (brain, cumulus cells, GS cells and liver) relative to that in the testis are shown. Primers used for this RT-PCR analysis are listed in Table 5.

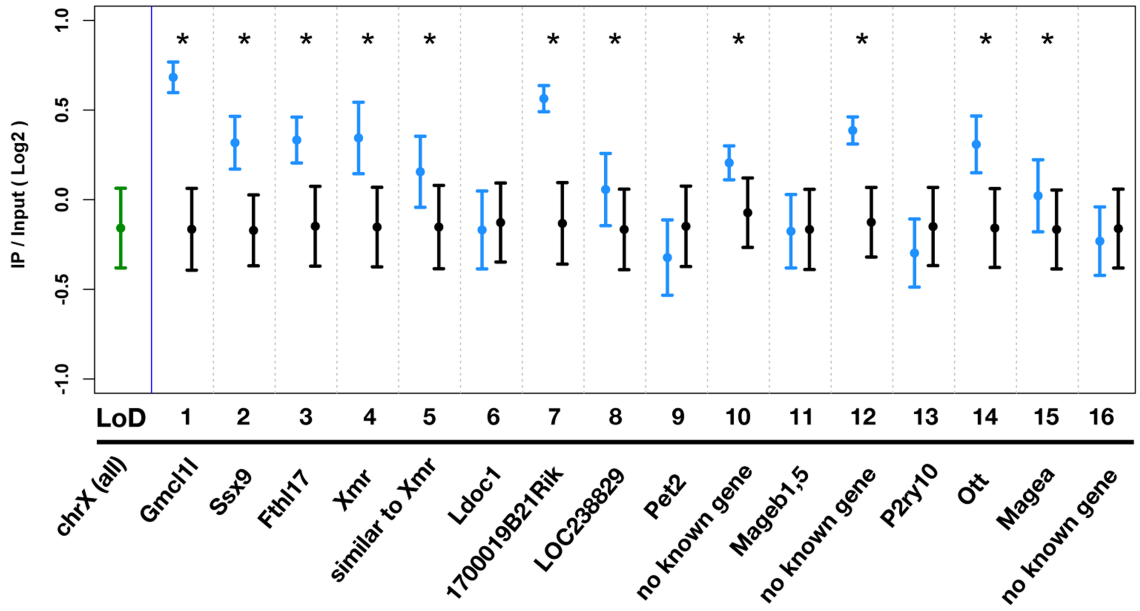


Figure 17. Significant enrichment of H3K9me2 in LoDs

The mean IP/input (log2) value was calculated using signals of the probes mapped in each LoD and is shown with the standard deviation (light blue). In each case, the same number of probes was chosen randomly from the entire X chromosome, and the mean value was obtained from these probes (black). Statistical significance was tested by Wilcoxon t-test. Asterisk indicates significant enrichment of H3K9me2 ($p<0.01$).

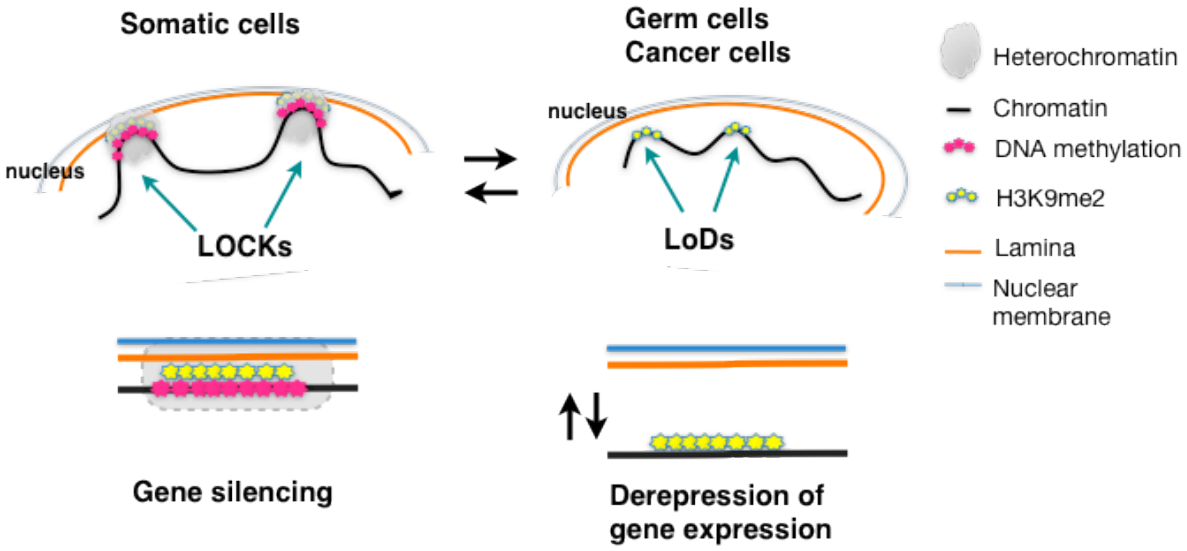


Figure 18. A model for gene silencing and derepression in LOCKs/ LoDs

Genes contained in the LOCKs tend to be repressed in differentiated somatic cells. Because LOCKs coincide with lamin B-associated domains, a gene-silencing mechanism based on three-dimensional subnuclear organization has been proposed [48]. In this model, large portion of genome in the somatic cells possess H3K9me2 LOCKs marks, and via LOCKs chromosome can be closely positioned to nuclear periphery, where gene silencing takes place. I found most LoDs are enriched with H3K9me2 and in fact overlap with some LOCKs, and hypermethylated in somatic cells like liver, thymus and cumulus cells. Therefore, LOCKs in the somatic cells likely to possess both H3K9me2 marks and hypermethylated DNA, and such unique epigenetic features are required for positioning of LOCKs chromatin at nuclear periphery to achieve gene silencing. In contrast, in germ cells or in some cancer cells, LoD/LOCKs regions are hypomethylated and derepression of gene expression is observed. If chromosomal positioning at nuclear periphery is necessary for gene silencing, LoD/LOCKs regions should be distant from the nuclear lamina, leading to expression of genes contained within the corresponding regions.

Acknowledgments

I would especially like to express my sincere gratitude to my thesis supervisor, Professor Kuniya Abe for his expert guidance and encouragement throughout the course of the present study with invaluable suggestions and helpful advice. I would also sincerely like to thank Drs. Hirosuke Shiura, Koji Numata, Michihiko Sugimoto, Ms. Masayo Kondo, Drs. Nathan Mise, Masako Suzuki and John M. Greally for helpful advice.

I wish to thank Drs. T. Suzuki for help with bisulfite pyrosequencing, S. Matoba and A. Ogura for cumulus samples, and Y. Katsura and Y. Satta for information and discussion on segmentally duplicated regions of the X chromosome. I also thank Ms. Y. Koga for help with the experiments. And I wish to thank Ms. M.Kusayama for everything.

This work was supported, in part, by a grant to KA by the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1. Wu Ct, M. and JR, *Genes, genetics, and epigenetics: a correspondence*. Science, 2001. **293**: p. 1103-5.
2. Suzuki, M. and A. Bird, *DNA methylation landscapes: provocative insights from epigenomics*. Nat Rev Genet, 2008. **9**: p. 465-76.
3. Fazzari, M. and J. Greally, *Epigenomics: beyond CpG islands*. Nat Rev Genet, 2004. **5**: p. 446-55.
4. Esteller, M., *Cancer epigenomics: DNA methylomes and histone-modification maps*. Nat Rev Genet, 2007. **8**: p. 286-98.
5. Avner, P. and E. Heard, *X-chromosome inactivation: counting, choice and initiation*. Nat Rev Genet, 2001. **2**: p. 59-67.
6. Bestor, T., A. Laudano, R. Mattaliano, and V. Ingram, *Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases*. J Mol Biol, 1988. **203**: p. 971-83.
7. Okano, M., S. Xie, and E. Li, *Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases*. Nat Genet, 1998. **19**: p. 219-20.
8. Li, E., T. Bestor, and R. Jaenisch, *Targeted mutation of the DNA methyltransferase gene results in embryonic lethality*. Cell, 1992. **69**: p. 915-26.

9. Okano, M., D. Bell, D. Haber, and E. Li, *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development*. Cell, 1999. **99**: p. 247-57.

10. Berman, B., D. Weisenberger, J. Aman, T. Hinoue, Z. Ramjan, Y. Liu, H. Noushmehr, C. Lange, C. van Dijk, R. Tollenaar, D. Van Den Berg, and P. Laird, *Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains*. Nat Genet, 2011. **44**: p. 40-6.

11. Meissner, A., T. Mikkelsen, H. Gu, M. Wernig, J. Hanna, A. Sivachenko, X. Zhang, B. Bernstein, C. Nusbaum, D. Jaffe, A. Gnirke, R. Jaenisch, and E. Lander, *Genome-scale DNA methylation maps of pluripotent and differentiated cells*. Nature, 2008. **454**: p. 766-70.

12. Hajkova, P., S. Erhardt, N. Lane, T. Haaf, O. El-Maarri, W. Reik, J. Walter, and M. Surani, *Epigenetic reprogramming in mouse primordial germ cells*. Mech Dev, 2002. **117**: p. 15-23.

13. Lee, J., K. Inoue, R. Ono, N. Ogonuki, T. Kohda, T. Kaneko-Ishino, A. Ogura, and F. Ishino, *Erasing genomic imprinting memory in mouse clone embryos produced from day 11.5 primordial germ cells*. Development, 2002. **129**: p. 1807-17.

14. Mayer, W., A. Niveleau, J. Walter, R. Fundele, and T. Haaf, *Demethylation of the zygotic paternal genome*. Nature, 2000. **403**: p. 501-2.

15. Hajkova, P., K. Ancelin, T. Waldmann, N. Lacoste, U. Lange, F. Cesari, C. Lee, G. Almouzni, R. Schneider, and M. Surani, *Chromatin dynamics during epigenetic reprogramming in the mouse germ line*. Nature, 2008. **452**: p. 877-81.

16. Santos, F., B. Hendrich, W. Reik, and W. Dean, *Dynamic reprogramming of DNA methylation in the early mouse embryo*. Dev Biol, 2002. **241**: p. 172-82.

17. Seki, Y., K. Hayashi, K. Itoh, M. Mizugaki, M. Saitou, and Y. Matsui, *Extensive and orderly reprogramming of genome-wide chromatin modifications associated with specification and early development of germ cells in mice*. Dev Biol, 2005. **278**: p. 440-58.

18. Ginsburg, M., M.H. Snow, and A. McLaren, *Primordial germ cells in the mouse embryo during gastrulation*. Development, 1990. **110**(2): p. 521-8.

19. McLaren, A., *Signaling for germ cells*. Genes Dev, 1999. **13**(4): p. 373-6.

20. McLaren, A., *Primordial germ cells in the mouse*. Dev Biol, 2003. **262**: p. 1-15.

21. Chuva de Sousa Lopes, S.M., K. Hayashi, T.C. Shovlin, W. Mifsud, M.A. Surani, and A. McLaren, *X chromosome activity in mouse XX primordial germ cells*. PLoS Genet, 2008. **4**(2): p. e30.

22. Sugimoto, M. and K. Abe, *X chromosome reactivation initiates in nascent primordial germ cells in mice*. PLoS Genet, 2007. **3**: p. e116.

23. Popp, C., W. Dean, S. Feng, S. Cokus, S. Andrews, M. Pellegrini, S. Jacobsen, and W. Reik, *Genome-wide erasure of DNA methylation in mouse primordial germ*

cells is affected by AID deficiency. Nature, 2010. **463**: p. 1101-5.

24. Khulan, B., R. Thompson, K. Ye, M. Fazzari, M. Suzuki, E. Stasiek, M. Figueroa, J. Glass, Q. Chen, C. Montagna, E. Hatchwell, R. Selzer, T. Richmond, R. Green, A. Melnick, and J. Greally, *Comparative isoschizomer profiling of cytosine methylation: the HELP assay*. Genome Res, 2006. **16**: p. 1046-55.

25. Irizarry, R., C. Ladd-Acosta, B. Carvalho, H. Wu, S. Brandenburg, J. Jeddeloh, B. Wen, and A. Feinberg, *Comprehensive high-throughput arrays for relative methylation (CHARM)*. Genome Res, 2008. **18**: p. 780-90.

26. Oda, M., J. Glass, R. Thompson, Y. Mo, E. Olivier, M. Figueroa, R. Selzer, T. Richmond, X. Zhang, L. Dannenberg, R. Green, A. Melnick, E. Hatchwell, E. Bouhassira, A. Verma, M. Suzuki, and J. Greally, *High-resolution genome-wide cytosine methylation profiling with simultaneous copy number analysis and optimization for limited cell numbers*. Nucleic Acids Res, 2009. **37**: p. 3829-39.

27. Borgel, J., S. Guibert, Y. Li, H. Chiba, D. Schübeler, H. Sasaki, T. Forné, and M. Weber, *Targets and dynamics of promoter DNA methylation during early mouse development*. Nat Genet, 2010. **42**: p. 1093-100.

28. Doi, A., I. Park, B. Wen, P. Murakami, M. Aryee, R. Irizarry, B. Herb, C. Ladd-Acosta, J. Rho, S. Loewer, J. Miller, T. Schlaeger, G. Daley, and A. Feinberg, *Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and*

fibroblasts. Nat Genet, 2009. **41**: p. 1350-3.

29. Guibert, S., T. Forné, and M. Weber, *Global profiling of DNA methylation erasure in mouse primordial germ cells*. Genome Res, 2012. **(null)**: p. (null).

30. Hackett, J., R. Sengupta, J. Zylicz, K. Murakami, C. Lee, T. Down, and M. Surani, *Germline DNA Demethylation Dynamics and Imprint Erasure Through 5-Hydroxymethylcytosine*. Science, 2012. **(null)**: p. (null).

31. Seisenberger, S., S. Andrews, F. Krueger, J. Arand, J. Walter, F. Santos, C. Popp, B. Thienpont, W. Dean, and W. Reik, *The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells*. Mol Cell, 2012. **48**(6): p. 849-62.

32. Mueller, J., S. Mahadevaiah, P. Park, P. Warburton, D. Page, and J. Turner, *The mouse X chromosome is enriched for multicopy testis genes showing postmeiotic expression*. Nat Genet, 2008. **40**: p. 794-9.

33. Wang, P., J. McCarrey, F. Yang, and D. Page, *An abundance of X-linked genes expressed in spermatogonia*. Nat Genet, 2001. **27**: p. 422-6.

34. Caballero, O. and Y. Chen, *Cancer/testis (CT) antigens: potential targets for immunotherapy*. Cancer Sci, 2009. **100**: p. 2014-21.

35. Ogonuki, N., K. Inoue, M. Hirose, I. Miura, K. Mochida, T. Sato, N. Mise, K. Mekada, A. Yoshiki, K. Abe, H. Kurihara, S. Wakana, and A. Ogura, *A high-speed congenic strategy using first-wave male germ cells*. PLoS One, 2009. **4**(3): p. e4943.

36. Mise, N., T. Fuchikami, M. Sugimoto, S. Kobayakawa, F. Ike, T. Ogawa, T. Tada, S. Kanaya, T. Noce, and K. Abe, *Differences and similarities in the developmental status of embryo-derived stem cells and primordial germ cells revealed by global expression profiling*. Genes Cells, 2008. **13**: p. 863-77.

37. Anderson, E., A. Baltus, H. Roepers-Gajadien, T. Hassold, D. de Rooij, A. van Pelt, and D. Page, *Stra8 and its inducer, retinoic acid, regulate meiotic initiation in both spermatogenesis and oogenesis in mice*. Proc Natl Acad Sci U S A, 2008. **105**: p. 14976-80.

38. Yuan, L., J. Liu, J. Zhao, E. Brundell, B. Daneholt, and C. Höög, *The murine SCP3 gene is required for synaptonemal complex assembly, chromosome synapsis, and male fertility*. Mol Cell, 2000. **5**: p. 73-83.

39. Hu, W., L. Gauthier, B. Baibakov, M. Jimenez-Movilla, and J. Dean, *FIGLA, a basic helix-loop-helix transcription factor, balances sexually dimorphic gene expression in postnatal oocytes*. Mol Cell Biol, 2010. **30**: p. 3661-71.

40. Cocquet, J., P.J. Ellis, S.K. Mahadevaiah, N.A. Affara, D. Vaiman, and P.S. Burgoyne, *A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse*. PLoS Genet, 2012. **8**(9): p. e1002900.

41. Reynard, L., J. Turner, J. Cocquet, S. Mahadevaiah, A. Touré, C. Höög, and P. Burgoyne, *Expression analysis of the mouse multi-copy X-linked gene Xlr-related, meiosis-regulated (Xmr), reveals that Xmr encodes a spermatid-expressed*

cytoplasmic protein, SLX/XMR. Biol Reprod, 2007. **77**: p. 329-35.

42. Clotman, F., O. De Backer, E. De Plaen, T. Boon, and J. Picard, *Cell- and stage-specific expression of mage genes during mouse spermatogenesis*. Mamm Genome, 2000. **11**: p. 696-9.

43. Olshen, A., E. Venkatraman, R. Lucito, and M. Wigler, *Circular binary segmentation for the analysis of array-based DNA copy number data*. Biostatistics, 2004. **5**: p. 557-72.

44. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes*. J Mol Biol, 1987. **196**: p. 261-82.

45. Bailey, J., A. Yavor, H. Massa, B. Trask, and E. Eichler, *Segmental duplications: organization and impact within the current human genome project assembly*. Genome Res, 2001. **11**: p. 1005-17.

46. Fujiwara, Y., T. Komiya, H. Kawabata, M. Sato, H. Fujimoto, M. Furusawa, and T. Noce, *Isolation of a DEAD-family protein gene that encodes a murine homolog of Drosophila vasa and its specific expression in germ cell lineage*. Proc Natl Acad Sci U S A, 1994. **91**: p. 12258-62.

47. Pauler, F., M. Sloane, R. Huang, K. Regha, M. Koerner, I. Tamir, A. Sommer, A. Aszodi, T. Jenuwein, and D. Barlow, *H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome*. Genome Res, 2009. **19**: p. 221-33.

48. Wen, B., H. Wu, Y. Shinkai, R. Irizarry, and A. Feinberg, *Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells*. Nat Genet, 2009. **41**: p. 246-50.

49. Inoue, K., T. Kohda, M. Sugimoto, T. Sado, N. Ogonuki, S. Matoba, H. Shiura, R. Ikeda, K. Mochida, T. Fujii, K. Sawai, A. Otte, X. Tian, X. Yang, F. Ishino, K. Abe, and A. Ogura, *Impeding Xist expression from the active X chromosome improves mouse somatic cell nuclear transfer*. Science, 2010. **330**: p. 496-9.

50. Oda, M., A. Yamagiwa, S. Yamamoto, T. Nakayama, A. Tsumura, H. Sasaki, K. Nakao, E. Li, and M. Okano, *DNA methylation regulates long-range gene silencing of an X-linked homeobox gene cluster in a lineage-specific manner*. Genes Dev, 2006. **20**: p. 3382-94.

51. Link, P., O. Gangisetty, S. James, A. Woloszynska-Read, M. Tachibana, Y. Shinkai, and A. Karpf, *Distinct roles for histone methyltransferases G9a and GLP in cancer germ-line antigen gene regulation in human cancer cells and murine embryonic stem cells*. Mol Cancer Res, 2009. **7**: p. 851-62.

52. Warburton, P., J. Giordano, F. Cheung, Y. Gelfand, and G. Benson, *Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes*. Genome Res, 2004. **14**: p. 1861-9.

53. Karimi, M., P. Goyal, I. Maksakova, M. Bilenky, D. Leung, J. Tang, Y. Shinkai, D.

Mager, S. Jones, M. Hirst, and M. Lorincz, *DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs*. Cell Stem Cell, 2011. **8**: p. 676-87.

54. Simpson, A., O. Caballero, A. Jungbluth, Y. Chen, and L. Old, *Cancer/testis antigens, gametogenesis and cancer*. Nat Rev Cancer, 2005. **5**: p. 615-25.

55. Tada, T., M. Tada, K. Hilton, S. Barton, T. Sado, N. Takagi, and M. Surani, *Epigenotype switching of imprintable loci in embryonic germ cells*. Dev Genes Evol, 1998. **207**: p. 551-61.

56. Ohbo, K., S. Yoshida, M. Ohmura, O. Ohneda, T. Ogawa, H. Tsuchiya, T. Kuwana, J. Kehler, K. Abe, H. Sch√∂ler, and T. Suda, *Identification and characterization of stem cells in prepubertal spermatogenesis in mice small star, filled*. Dev Biol, 2003. **258**: p. 209-25.

57. Kanatsu-Shinohara, M., N. Ogonuki, K. Inoue, H. Miki, A. Ogura, S. Toyokuni, and T. Shinohara, *Long-term proliferation in culture and germline transmission of mouse male germline stem cells*. Biol Reprod, 2003. **69**: p. 612-6.

58. Kobayashi, S., Y. Fujihara, N. Mise, K. Kaseda, K. Abe, F. Ishino, and M. Okabe, *The X-linked imprinted gene family Fthl17 shows predominantly female expression following the two-cell stage in mouse embryos*. Nucleic Acids Res, 2010. **38**: p. 3672-81.

59. Ko, M., S. Ko, N. Takahashi, K. Nishiguchi, and K. Abe, *Unbiased amplification*

of a highly complex mixture of DNA fragments by 'lone linker'-tagged PCR. Nucleic Acids Res, 1990. **18**: p. 4293-4.

60. Otsu, N., *Threshold Selection Method from Gray-Level Histograms*. Ieee Transactions on Systems Man and Cybernetics, 1979. **9**: p. 62-66.

61. Mahalanobis, P.C., *On the generalized distance in statistics*. Proc. Natl. Inst. Sci. India, 1936. **2**: p. 49.

62. Pohl, A., C. Sugnet, T. Clark, K. Smith, P. Fujita, and M. Cline, *Affy exon tissues: exon levels in normal tissues in human, mouse and rat*. Bioinformatics, 2009. **25**: p. 2442-3.

63. Ohtsubo, Y., W. Ikeda-Ohtsubo, Y. Nagata, and M. Tsuda, *GenomeMatcher: a graphical user interface for DNA sequence comparison*. BMC Bioinformatics, 2008. **9**: p. 376.