学位論文

Epigenetic regulation of developmental genes in vertebrates

（脊椎動物における発生関連遺伝子のエピジェネティック制御）

平成 25 年 12 月博士（理学）申請

東京大学大学院理学系研究科

生物科学専攻

中村　遼平

Contents

Abbrebiations

- ChIP — chromatin immunoprecipitation
- *Da* — *Double anal fin*
- dpf — days post fertilization
- ESC — embryonic stem cell
- GO — gene ontology
- H3K4me1 — histone H3 Lys4 monomethylation
- H3K4me2 — histone H3 Lys4 dimethylation
- H3K4me3 — histone H3 Lys4 trimethylation
- H3K27ac — histone H3 Lys27 acetylation
- H3K27me3 — histone H3 Lys27 trimethylation
- hESC — human embryonic stem cell
- qPCR — quantitative polymerase chain reaction
- SVM — support vector machine
- *wt* — wild type
- *zic1*/*zic4* — *zic1* and *zic4*

Abstract

DNA methylation is a fundamental epigenetic modification in vertebrate genomes and a small fraction of genomic regions is hypomethylated. Previous studies have implicated hypomethylated regions in gene regulation, but their functions in vertebrate development remain elusive. To address this issue, I generated epigenomic profiles that include base-resolution DNA methylomes and histone modification maps from both pluripotent cells and mature organs of medaka fish and compared the profiles with those of human ES cells. As a result, I found a subset of hypomethylated domains harbor H3K27me3 (K27HMDs) and their size positively correlates with the accumulation level of H3K27me3. Large K27HMDs are conserved between medaka and human pluripotent cells, and predominantly contain promoters of developmental transcription factor genes. Those key genes were found to be under strong transcriptional repression, as compared with other developmental genes with smaller K27HMDs. Furthermore, human-specific K27HMDs show an enrichment of neuronal activity-related genes, which suggests a distinct regulation of those genes in medaka and human. In mature organs, some of the large HMDs become shortened by elevated DNA methylation, and associate with sustained gene expressions. Finally, by using support vector algorithm, I demonstrate that large K27HMDs can be predicted from DNA sequence. Unexpectedly, some of the K27HMD-specific sequence motifs are located in developmental gene exons, suggesting the involvement of coding sequence in the regulation of epigenetic states. This study highlights the significance of domain size in epigenetic gene regulation, and I propose that large K27HMDs play a crucial role in pluripotent cells by strictly repressing key developmental genes, while their shortening consolidate long-term gene expression in adult differentiated cells.

General Introduction

Development of multicellular organisms proceeds with cell differentiations that are controlled by strict and complex gene regulations. Key developmental genes (or transcriptional regulator genes) govern cell-type-specific transcriptional profiles, and their expression states are considered as dominant determinants of cell identities. Indeed, forced expressions of some of those genes can change cell identities (Takahashi et al., 2007). Under natural conditions, however, cell differentiations are mostly forward-moving processes characterized by continuous specifications. Thus, cells are thought to possess "epigenetic" information that influences gene regulation and determines differentiation competence or restricts a cellular fate. Chemical modifications to DNA or chromatin could be transmitted to daughter cells and affect gene regulations, and are promising candidates for epigenetic information. However, it remains poorly understood what form of the modifications actually functions as such epigenetic machinery during development.

Cytosine methylation of CpG dinucleotides of the genomic DNA is one of the essential and heritable modifications in vertebrates. The genomes of all vertebrates studied thus far are globally methylated, and a small fraction of genomic regions is hypomethylated (Hendrich and Tweedie, 2003; Tweedie et al., 1997). Recent genome-wide analyses have revealed that the majority of gene promoters are hypomethylated (Lister et al., 2009). For many years, DNA methylation was widely accepted as a repressive modification (Bird, 2002), and hypomethylated promoters were considered to be active or permissive for transcription. The notable feature of DNA methylation is that this modification pattern is precisely copied to daughter cells through cell division, mediated by the ubiquitously expressed maintenance DNA methyltransferase (Dnmt1) which catalyzes the methylation of hemimethylated CpGs produced during DNA replication (Jeltsch, 2006). DNA methylation is therefore suggested to

maintain a repressed state across the cell generations and stably silence promoter activity (Suzuki and Bird., 2008).

In addition to DNA methylation, histone modifications influence promoter activity (Zhou et al., 2011). Histone H3 lysine 4 (H3K4) methylation distributes exclusively on hypomethylated DNA regions (Cedar and Bergman, 2009; Hu et al., 2009; Ooi et al., 2007), and positively regulates promoter activity. However, hypomethylated promoters are not exclusively found in genes with active transcription. In embryonic stem cells (ESCs), promoters of developmentally regulated genes are also hypomethylated, even though their transcriptions begin later during cell differentiation (Suzuki and Bird, 2008; Xie et al., 2013). These promoters are frequently marked by repressive histone H3 lysine 27 (H3K27) methylation and proposed to be "poised" for immediate induction upon stimulation (Bernstein et al., 2006; Zhao et al., 2007). This poised state, but not simply silenced state, of developmental genes is thought to be essential for pluripotent cells to maintain the undifferentiated state with pluripotency. H3K27me3 and hypomethylation at developmental gene promoters were also reported in zebrafish and *Xenopus* early embryos (Akkers et al., 2009; Bogdanovic et al., 2011; Lindeman et al., 2011; Potok et al., 2013; Vastenhouw et al., 2010), suggesting that repression of hypomethylated developmental gene promoters by H3K27me3 is an essential and conserved feature in vertebrate development. The nature of these hypomethylated promoters marked by H3K27me3 is beginning to be elucidated in the context of development, for example, the mechanism that regulates the accumulation of H3K27me3 and their actual role in control of cell identities.

Recent genome-wide analyses using mammalian cells have suggested the antagonistic relationship between DNA methylation and H3K27me3 (Brinkman et al., 2012; Lindroth et al., 2008). In post-natal mouse brains, DNA methylation at regions flanking proximal promoters was shown to facilitate transcription of neuronal genes by antagonizing

H3K27 methylation (Wu et al., 2010). These studies suggest the diverse function of DNA methylation depending on the location it occurs, which in turn points out the potential role of DNA methylation or hypomethylation outside the proximal promoter in developmental gene regulation. Importantly, more recent studies have found large genomic domains (e.g. several tens of kb) with hypomethylation and H3K27me3, the size of which is extremely larger than other promoters (Jeong et al., 2014; Long et al., 2013; Xie et al., 2013). Interestingly, those large domains frequently contain promoters of developmental transcription factors, suggesting their important role in vertebrate development. However, previous studies have not addressed to the functions or biological significance of those domains. Furthermore, the mechanism underlying the hypomethylation or H3K27me3 accumulation at specific developmental promoter regions is still incompletely understood. Although some DNA binding factors or chromatin modifiers have reported to be involved, the genomic sequences that determine the chromatin modification patterns are unknown.

Here, in my doctoral thesis, I analyzed chromatin modifications of the medaka genome and propose a role for hypomethylated domains in control of pluripotency and matured cell identities. In Chapter 1, I characterized the hypomethylated domains in pluripotent cells of medaka, and found the domain-size-dependent transcriptional regulation. I also revealed the conservation and species-specific changes in hypomethylated domains by comparative analysis between medaka and human pluripotent cells. In Chapter 2, I found dynamic reduction in the size of hypomethylated domains in adult medaka tissues, which generally associates with persistent expression of developmental genes. In Chapter 3, by using a support vector machine algorithm, I explored the DNA sequences that can predict both DNA methylation and H3K27me3 pattern, and propose that coding exons of developmental genes contribute significantly to the formation of K27HMDs.

Chapter 1

Role of large hypomethylated domains in vertebrate development

Introduction

Development initiates from cells that possess totipotency or pluripotency. Various observations suggest that chromatin structures have important roles in maintaining those cellular states. Polycomb group proteins repress developmental genes by catalyzing H3K27me3 at those gene promoters and play an essential role in maintaining the undifferentiated state of ES cells (Lee et al., 2006; O'Carroll et al., 2001). At the same time, however, those promoters are hypomethylated and marked by active histone marks (H3K4 methylations) and proposed to be poised (Bernstein et al., 2006). Furthermore, accessible chromatin regions continuously decrease during ES cell differentiations (Stergachis et al., 2013b). These studies suggest that although developmental genes should be repressed in ES cells, they should be kept inducible. Thus, the pluripotent state may not be a simply "undifferentiated" state, but also "prepared" state for future differentiations.

Although hypomethylation at developmental promoters are believed to function for rapid induction upon differentiation, some hypomethylated domains at developmental promoters are extremely large, and the significance of such large size is unknown. Antagonistic relationships have been reported between DNA methylation and H3K27me3, but the function of DNA methylation in pluripotent cells remains elusive. In contrast to H3K27me3, DNA methylation was shown to be dispensable for ES cell self-renewal (Tsumura et al., 2006). Furthermore, despite the fact that global DNA methylation and promoter hypomethylation is conserved among the vertebrate genomes, it is not a general feature in other animal species (Suzuki and Bird, 2008). In invertebrates, DNA methylation is either absent or varies considerably in its pattern from vertebrates (Bird, 2002). As the role and pattern of DNA methylation vary widely among organisms, it is important to ask to what extent their patterns are conserved among vertebrates to understand the function and

evolution of hypomethylated domains in the vertebrate lineage.

In this chapter, I investigate the epigenomic profiles of medaka blastula embryos (pluripotent cells) by using single-base resolution DNA methylomes from the previous study (Qu et al., 2012) and also by generating genome-wide histone modification maps. In blastula embryos, large compartmentalized genomic domains with H3K27me3 and DNA hypomethylation (K27HMDs) were found at developmental transcription factor genes. Strikingly, I found that the K27HMD size positively correlates with the transcriptional repression. I also performed comparative analyses between medaka and human pluripotent cells, and revealed both conservation and species-specific changes in the distribution of K27HMDs.

Results

Identification of hypomethylated DNA domains in the medaka genome

Cells from medaka blastula embryos are known to retain pluripotency (Yi et al., 2009). To obtain the global genomic distribution of epigenetic modifications in medaka blastula embryos, I performed ChIP-seq analyses using antibodies against H3K27me3, H3K4me1, H3K4me2, H3K4me3 and H3K27ac (supplementary material Table S1), and integrated these results with previously established medaka single-base DNA methylomes (Qu et al., 2012). In medaka blastula embryos, contiguous regions of a low level of methylation appeared obvious on a genome browser with sharp boundaries (Fig. 1), as methylation frequency of individual CpG site has a clear bimodal distribution (Fig. 2). By scanning the whole genome, I identified 15,145 regions with more than 9 contiguous low methylated CpG sites (methylation frequency < 0.4), termed hypomethylated domains or HMDs (supplementary material Table S4). Gene promoters and CpG islands accounted for the majority of HMDs (Fig. 3). H3K4 methylations exclusively distributed to HMDs, as previously reported (Hu et al., 2009; Ooi et al., 2007), and some portions of those regions are co-enriched with other histone modifications (Fig. 1).

Large HMDs mark key transcription factors for vertebrate development

It is well known that DNA methylation at proximal promoters silences cell type-specific genes, and that permissive promoters are generally hypomethylated (Smith and Meissner, 2013). In some cases, hypomethylated domains expand beyond promoters and cover much larger regions (Bogdanovic et al., 2011; Jeong et al., 2014; Laurent et al., 2010; Long et al., 2013; Xie et al., 2013). However, the significance of such large domain size in gene expression has not been addressed. Interestingly, in the medaka genome, while most

HMDs were several kb in size and highly enriched with H3K4me2 and other active histone modifications (Fig. 4A), I found that a subset of HMDs was extremely large and those regions were frequently enriched with H3K27me3 (Fig. 4B). Furthermore, I noticed a tendency that large HMDs are more frequently marked by H3K27me3 than small ones (Fig. 5). To examine the relationship between the HMD size and H3K27me3, I first classified the HMDs by the existence of H3K27me3 enrichment. 2,398 HMDs were found to contain H3K27me3 peaks detected by QuEST software (Valouev et al., 2008), and classified as K27HMDs (supplementary material Table S4). The H3K27me3 marked hypomethylated regions were also identified by ChromHMM software that can annotate the regions in a statistically principled manner (Ernst and Kellis, 2012). I confirmed that the K27HMDs largely overlap with the regions identified by ChromHMM (Fig. 6). However, regions called by ChromHMM were frequently divided into smaller fragments, and therefore, I chose K27HMDs that were suitable for the further analysis focusing on domain size. While H3K27me3-free HMDs (nonK27HMDs) had a high enrichment of H3K4me2, the majority of K27HMDs also had a low but significant enrichment of H3K4me2 (Fig. 7).

I next investigated the characteristics of large HMDs. Notably, 12% of K27HMDs were larger than 4 kb in size, whereas 99.8% of 12,747 nonK27HMDs were less than 4 kb (Fig. 8A, B). As the majority of HMDs overlapped with promoter regions, I linked medaka genes to promoter-associated HMDs (supplementary material Table S5). Intriguingly, large K27HMD (> 4 kb) associated genes (317) and small K27HMDs (< 4 kb) associated genes (1,295) showed specific features in their functions. Gene ontology analysis showed that terms related to transcription regulation and developmental processes are highly enriched in large K27HMDs (Fig. 9A). Indeed, 65% of large K27HMD genes encoded DNA binding factors (Fig. 9B; supplementary material Table S6), and had crucial functions for embryonic development. In contrast, small K27HMDs showed enrichment of terms for developmental

processes, signal transduction, and cell communication, with relatively low enrichment for transcription factors (Fig. 9C). Thus, the large K27HMDs mark specific set of developmental genes, and may have important role in vertebrate development. I also confirmed that the HMD size does not reflect gene length, as large K27HMDs tended to have shorter genes than small ones (Fig. 10).

It is known that the teleost underwent whole genome duplication (Jaillon et al., 2004; Kasahara et al., 2007; Taylor et al., 2003), and I found that some duplicated genes encoding transcription factors showed different HMD size. I picked two duplicated gene pairs (*pax6* (ENSORLG00000009913 and ENSORLG00000000847), and *tbx2* (ENSORLG00000014792 and ENSORLG00000010011)) for expression analysis, because for each pair one gene associated with a large K27HMD and the other with a small K27HMD (Fig. 11A, B). *In situ* hybridization of those gene pairs in medaka embryos demonstrated that the genes with large HMDs are expressed in a tissue-specific pattern (Fig. 11C, D), which is nearly identical to the conserved expression pattern in vertebrates including mouse (Harrelson et al., 2004; Puschel et al., 1992). In contrast, the genes with small HMDs showed no or partial expression of the conserved pattern (Fig. 11C, D). These results suggest the conserved function for large K27HMD genes, and strengthen the possibility that the size of HMD reflects the gene function.

The size of K27HMD and the number of CpGs in the domain correlate with H3K27me3 level

I assumed that the size of the HMDs contribute to the transcriptional regulation of developmental transcription factor genes. The fact that most of the large HMDs are enriched with H3K27me3 led me to hypothesize that the size of HMD is one of the determinant of the level of H3K27me3 accumulation. Comparison of the K27HMD size and ChIP signal

enrichment or ChIP peak intensity revealed a significant correlation (Fig. 12A, B and C). By contrast, negative correlation was found between the domain size and active modification levels (H3K4me1, H3K4me2, H3K4me3, and H3K27ac). Among those active modifications, the negative correlation was most significant for H3K4me2 (Fig. 13). These results indicate that a larger K27HMD have a stronger repressive property for transcription. Given that large K27HMD genes mostly encode transcription factors crucial for development, the strong repression of those genes in blastula embryos might prevent improper cell differentiation and maintain stemness.

Previously, Polycomb group (PcG) proteins that mediate H3K27me3 accumulation were reported to preferentially bind to CpG islands (Deaton and Bird, 2011; Ku et al., 2008; Woo et al., 2010), but DNA methylation reduces their bindings (Hagarman et al., 2013; Wu et al., 2010). I therefore counted the number of low methylated CpG sites inside the K27HMDs and examined the relationship with H3K27me3 level. A significant correlation was observed between the low methylated CpG count and the H3K27me3 ChIP peak intensity (Fig. 14), supporting the possibility that a large K27HMD provides a large number of unmethylated CpGs that can potentially recruit PcG proteins.

Comparison of HMDs between medaka and human pluripotent stem cells

To test if the importance of the HMD size holds true for other vertebrates, I applied my analysis pipeline to a whole-genome dataset from human ESCs (hESCs) (Lister et al., 2009; Lister et al., 2011). hESCs were chosen because I assumed that they are equivalent to medaka blastula cells in terms of pluripotency and their chromatin modification patterns were extensively studied. First, I examined the epigenetic status at promoters of orthologous genes and its conservation. All medaka genes annotated to single human orthologous genes in Ensembl database (13,301 genes) were used for comparison between the two species. I first

found that the majority of medaka promoters (59%) have similar DNA and histone modifications in human ES cells (Pearson's Chi-squared test P<2.2E-16; Fig. 15; supplementary material Table S7). Furthermore, the comparison of HMD size between medaka and human further revealed a conserved tendency: while nonK27HMD largely reside in small size fraction in both medaka and human, genes marked by large K27HMDs in medaka embryos are also marked by large K27HMDs in hESCs (Fig. 16A, B). I defined HMDs larger than 8 kb as large HMDs in human, so that the number of genes designated as large K27HMD is equal to large K27HMD (> 4 kb) genes in medaka. I confirmed that the genes associated with conserved large K27HMDs are highly enriched for GO terms related to transcription regulation and developmental processes (Fig. 17A). Furthermore, at the level of protein sequence, genes found in large K27HMDs are more conserved between the two organisms than those in small K27HMDs or nonK27HMDs, and genes with no HMD (i.e. methylated) showed the lowest conservation (Fig. 17B), indicating that large K27HMD genes have conserved functions among vertebrates.

Like in medaka blastula embryos, large K27HMDs in hESCs accumulated higher levels of H3K27me3 (Fig. 18A), and the correlation between the number of low methylated CpG sites and H3K27me3 level was also significant in hESCs (Fig. 18B). Thus, large K27HMDs seem to be more repressive than small K27HMDs. I tested this idea by analyzing the transcriptome data set from the previous study (Lister et al., 2011). First, as expected, nonK27HMD genes exhibit high levels of expression, while majority of no HMD (methylated) genes are silenced (Fig. 18C). Consistent with the poised model (Bernstein et al., 2006; Pan et al., 2007), the expression levels of K27HMD genes tend to be low. Importantly, large K27HMD genes showed significantly lower expression than small ones, almost similar levels to methylated genes (Fig. 18C). This tendency was also observed in blastula embryos, but not so evident (data not shown), probably due to the presence of maternally derived

transcripts and due to unusual transcriptional environment of mid-blastula transition that the blastula embryos had just experienced (Aizawa et al., 2003). Taken together, these results suggest the large size of K27HMD at key transcription factor genes is conserved among vertebrate species, and have strong repressive effect on gene expressions in pluripotent cells (Fig. 21).

Human-specific K27HMDs mark genes related to neuronal activities

Although the epigenetic status of homologous genes is largely conserved between medaka and human pluripotent cells, a subset of genes has been subjected to differential DNA and histone methylations. I looked into these differences focusing on human-specific K27HMD genes, because they could reflect changes in function that occurred during vertebrate evolution. For this purpose, K27HMD genes in hESCs were classified into three categories according to the epigenetic status of their medaka counterparts. Class 0 is a category of K27HMD genes shared by the two species as described above, while genes of the latter two became K27HMD in hESCs human from nonK27HMD (Class I) or methylated (Class II) in medaka, respectively (Fig. 15). I then determined GO terms enriched for these three classes of hESC K27HMDs. I first confirmed that developmental genes are highly enriched in Class 0 K27HMDs (Fig. 19A). In contrast, no such high enrichment for developmental genes was observed for class I and class II genes. Furthermore, the large K27HMDs were mostly included in Class 0 (Fig. 19B), strongly suggesting that this epigenetic machinery is essential for development and thus conserved among vertebrates. Interestingly, however, class II genes are more associated with signal transduction and neuronal activities (Fig. 20A, B; supplementary material Table S8), while Class I genes did not show any enrichment of specific terms. I also failed to find any preference of GO terms for genes in medaka-specific K27HMDs (data not shown).

The human-specific HMDs could be due to a different differentiation state between the two types of cells, medaka blastula cells and hESCs, despite their presumed pluripotency. However, sequence conservation of three classes of human K27HMDs among vertebrates revealed that Class I and II HMDs are more divergent than Class 0 (Fig. 20C), suggesting that genetic variations in cis-elements rather than cellular characters of the cells account for the observed differences in epigenetic modifications between medaka and human.

Discussion

Large K27HMDs serve as strong repressive machinery in pluripotent cells

Previous studies have demonstrated that the H3K27me3 domains often occupy several kb around promoters of developmental genes, but at some gene loci, the domains exceed to cover much larger regions (Zhao et al., 2007). Very recent studies also identified the large genomic domains with DNA hypomethylation at transcription factor gene loci in various vertebrate species (Jeong et al., 2014; Long et al., 2013; Xie et al., 2013). However, the biological significance of the size of these genomic domains has not been addressed. In this chapter, I have focused on the relationship between the HMD size and H3K27me3 level, and found that the large size associates with strong transcriptional repression.

H3K27me3 marked promoters are proposed to keep developmental genes poised for immediate activation in pluripotent cells (Bernstein et al., 2006). Indeed, the transcriptional activity of K27HMD genes in human ES cells was found to be low, but slightly higher than that of methylated genes (Fig. 18C, small K27HMD genes vs. methylated genes). These data suggest that keeping the promoter activity poised may pass leaky transcription under polycomb-mediated repression. Importantly, however, I found that the H3K27me3 level correlates with K27HMD size and the transcription level of genes in large K27HMDs is significantly lower than that in small ones in both differentiated medaka tissues and human ESCs. In pluripotent cells, the large K27HMDs preferentially mark transcription factor genes with crucial function in development, while the smaller ones tend to mark genes related to signal transduction. Transcription factors in large K27HMDs, thus, are strictly shut off, if not required. Previous study in zebrafish sperm also reported that genes with high levels of H3K27me3 at proximal promoters mostly encode transcription factors (Wu et al., 2011b), which are largely overlapped with genes identified by the domain size (large

K27HMD genes) in this study. This further emphasizes the generality of size-dependent H3K27me3 accumulation and strong repression of transcription factor genes in vertebrates. The strict repression is important presumably because the derepression of those transcription factors would result in inappropriate cell differentiation (Boyer et al., 2006; Fujikura et al., 2002) or malignancy such as cancer (Darnell, 2002). Large K27HMDs may thus play critical role in maintaining pluripotency by keeping key developmental genes both inducible and robustly repressed (Fig. 21).

Significance of the size of epigenetically modified domains in transcriptional regulation

The precise mechanism regulating the amount of H3K27me3 awaits future studies. However, previous studies demonstrated that PcG proteins are recruited to CpG islands, but DNA methylation inhibits their binding to chromatin. Hence, our data suggest that large HMDs provide a broad platform with unmethylated CpGs that potentially recruit PcG proteins to enrich H3K27me3, while small HMDs have limited capability of binding of PcG proteins.

Recent study identified that the super-enhancer in which large domains occupied by clusters of enhancers drive strong expression of transcriptional regulator genes and control cell identities (Whyte et al., 2013). Super-enhancers are characterized by their domain size and high enrichment of active histone modifications and transcription factor bindings. This also suggests that the size of epigenetically modified domain actually affect gene regulation. The concept of the large K27HMD is reminiscent of that of the super-enhancer, but the two large genomic domains have opposite functions, robust repression vs. strong activation. These large domains with dense epigenetic modifications may strengthen the repression or activation effect on developmental genes and cooperatively contribute to the control of cell identity.

Evolution of vertebrate epigenomes

Comparative analysis in this study demonstrated that the overall pattern of DNA methylation is conserved between medaka and human. Moreover, the majority of gene promoters shared the same epigenetic states between the two systems. In particular, large K27HMDs were highly conserved, thus, pre-marking of key transcription factors by large K27HMD and its strong repression are shared features in the vertebrate development. Sequence conservation analyses showed that region inside the K27HMDs are more conserved than methylated region (Fig. 20C). This suggests that functional sequences are enriched inside the K27HMDs. These sequences may include elements that induce hypomethylation or H3K27me3 accumulation. The sequence features of K27HMDs will be addressed in Chapter 3.

Importantly, my analysis also identified human-specific K27HMD genes and found that those methylated in medaka pluripotent cells (Class II) are related to neuronal activities. Human-specific K27HMDs had lower sequence conservation than those common to both human and medaka, suggesting that the changes in cis-regulatory elements have led to the differential epigenetic modifications. Indeed, regulatory elements under human constraint, but not conserved in mammals, were found to be associated with neuronal activities (Ward and Kellis, 2012). These neuronal genes are not expressed in both medaka and human pluripotent cells, but the repression is mediated by distinct modifications, DNA methylation and H3K27me3, respectively. Then what is the biological significance of such differential epigenetic marking? One possibility is that K27HMD may enable flexible changes in regulation of those genes; the poised state or sustained expression can be achieved through regulation of histone modification and the HMD size (described in Chapter 2). Interestingly, a significant number of glutamate receptor genes were found at Class II human-specific K27HMDs (Fig. 20B). Glutamate receptors are known to have a key function in synaptic plasticity and important for learning and memory (Lancaster and Dalmau, 2012; Salter and

Kalia, 2004). The flexible regulation of these genes could be required for the sophisticated neuronal network in human. However, I could not exclude the possibility that the differential epigenetic state between medaka blastula embryos and human ES cells comes from different differentiation state. To reveal the differential regulation of neuronal genes in human, the analysis using other cell types or other species is necessary.

Chapter 2

Dynamics of DNA methylation in adult tissues

Introduction

During embryonic development, a cell differentiates into a more specialized cell type and loses differentiation competency. In developing embryos, cells are frequently exposed to transiently provided stimuli that will permanently alter their cell identity. In particular, vertebrate cells maintain their identities during the animal's long lifetime, even though most of the cells constituting tissues are continuously turned over. Chromatin structures have been suggested to play a role in such cellular memory formation (Stergachis et al., 2013b). In mouse macrophage, transient stimulation was reported to alter chromatin accessibility at enhancer regions that will changes the cell behavior against later stimulations at least over the short term (Ostuni et al., 2013). However, it is unknown whether this system can maintain long-term cellular memory.

Persistent expressions of some key developmental genes have been reported to maintain cell identities in vertebrates throughout life. It was previously reported that, in human feet, embryonic *HOX* expression pattern is maintained in adult fibroblasts, and are required for the site-specific cell identities (Rinn et al., 2008). Although the study showed that the expression of *HOX* itself is cell-autonomously and epigenetically maintained, the precise mechanism for maintenance of the long-term gene expression has not been elucidated.

In fish development, specification of the dorsal and ventral trunk and its regulation by *zic1* and *zic4* (*zic1*/*zic4*) genes has been well studied (Kawanishi et al., 2013). *zic1*/*zic4*, organized head-to-head with a small intergenic sequence (Fig. 22), encode zinc finger transcription factors and function in the neuronal development and the specification of dorsal fate in the trunk (Kawanishi et al., 2013; Merzdorf, 2007). They are thought to share cis-regulatory elements and their expression is nearly identical. During somitogenesis, it has

been shown that *zic1/zic4* expression is activated in the dorsal somite by secreted factors, and their expression patterns depends on the surrounding tissues. At later stages, however, their regulation switches to a cell-autonomous manner, and the expression is maintained throughout life in the dorsal parts of somite derivatives, i.e. the myotome, dermis and vertebrae (Kawanishi et al., 2013). Therefore, I speculated that epigenetic machinery is involved in the regulation of the two genes. Interestingly, I noticed that *zic1*/*zic4* genes are covered by a particularly large K27HMD (26 kb) in medaka blastula embryos (Fig. 22). To understand the epigenetic regulation of *zic1*/*zic4*, and to investigate the role of large K27HMDs in later development, I analyzed the chromatin modifications in tissues from adult and larval medaka.

In this chapter, I demonstrated a dramatic reduction in domain size of the large K27HMD at *zic1*/*zic4* loci in adult dorsal myotome. I term this phenomenon 'HMD shortening'. Importantly, HMD shortening is generally associated with sustained gene expression in adult tissues. These results are consistent with the correlation between K27HMD size and transcriptional repression in pluripotent cells described in Chapter 1, suggesting the HMD-size-dependent transcriptional regulation also in matured cells.

Results

Dynamics of the epigenetic state of *zic1*/*zic4* genes during development

I isolated dorsal (*zic*-positive) and ventral (*zic*-negative) myotome separately from adult fish and generated methylomes and histone modification maps for each half (supplementary material Table S1 and S3). Although the overall methylation pattern was similar between dorsal and ventral myotome, I noticed a significant difference in the chromatin state at the *zic1*/*zic4* locus (Fig. 22). In the dorsal myotome, the *zic1*/*zic4* locus showed low H3K27me3 levels and high H3K4me2 levels, consistent with the state of active transcription. Surprisingly, however, I found large blocks of DNA hypermethylation at regions outside the H3K4me2-enriched promoter regions, which led to the shortened HMD around the promoter region (Fig. 22, blue box). I termed this phenomenon "HMD shortening". By contrast, the ventral myotome maintained the blastula-type epigenetic state; high H3K27me3 and low DNA methylation. Given that the HMD size correlates with transcriptional repression level, this HMD shortening might promote *zic1*/*zic4* expression. To examine when this HMD shortening occurs during development, I additionally investigated the DNA methylation pattern in dorsal and ventral myotome at the hatching stage, a stage when the dorsal specific expression of *zic1*/*zic4* is already induced (Kawanishi et al, 2013). As expected, ChIP-qPCR revealed that the pattern of active and repressive histone modifications was already established in the dorsal and ventral myotome by the hatching stage (Fig. 23). By contrast, the dorsal specific DNA hypermethylation was not detected at this stage (Fig. 22). Bisulfite sequencing at the larval stages revealed a gradual increase in DNA methylation as larval development progressed (Fig. 24). Thus, HMD shortening occurs after the establishment of active histone modifications and gene expression in dorsal myotome. These results suggest that DNA hypermethylation in the *zic1/zic4* HMD is not necessary for the initial induction

but rather for the maintenance of gene expression.

I hypothesized that HMD shortening at later stages depends on the active chromatin state. To explore this possibility, I determined the DNA methylation pattern in the *zic1*/*zic4* mutant, *Double anal fin* (*Da*). The *Da* mutant has a large transposon insertion that impairs the mesoderm enhancer of *zic1*/*zic4*, leading to a dramatic reduction in myotome expression and ventralized dorsal structures in the trunk (Fig. 25; (Moriyama et al., 2012). ChIP-qPCR analysis confirmed the high enrichment of H3K27me3 remained in the dorsal myotome of *Da* (Fig. 23). Strikingly, bisulfite-seq showed a complete absence of the dorsal specific DNA hypermethylation in the myotome of *Da* (Fig. 22), suggesting the requirement of *zic1*/*zic4* activation for HMD shortening. To further elaborate this possibility, I made *Da* heterozygous fish and investigated DNA methylation of each allele. *wt* and *Da* alleles were distinguished by single base nucleotide variations. In the dorsal myotome of heterozygous fish, only *zic1/zic4* locus of *wt* allele is activated and that of *Da* allele remained repressive because of the transposon insertion. Bisulfite sequencing revealed that the *wt* allele was highly methylated, while *Da* allele remained hypomethylated (Fig. 26). Thus, I conclude that the activation of the *zic1*/*zic4* locus, but not Zic1/Zic4 proteins or their downstream factors, induces HMD shortening. Collectively, these results suggest that once *zic1*/*zic4* are induced during embryogenesis, they are autonomously subject to HMD shortening during growth.

HMD shortening associates with sustained gene expressions in matured organs

The epigenetic profile of the *zic1*/*zic4* locus suggested that large K27HMDs have a repressive function, while the shortening of the K27HMD associates with their sustained expression. I also noticed that the large K27HMD at *hox* gene clusters underwent the remarkable DNA hypermethylation that led to shortened HMDs at active promoter in adult myotome (Fig. 27). To address whether HMD shortening at active gene loci is a general

feature in adult tissues, I compared the epigenomes of the adult dorsal myotome and liver to that of blastula embryos. For liver epigenome data, I generated histone modification maps (supplementary material Table S1), and integrated with methylome data from previous study (Qu et al., 2012). I found that the majority of K27HMDs have unchanged methylation levels in adult tissues, but a significant proportion of HMDs were subjected to DNA hypermethylation in which more than 5% of their CpG sites became highly methylated (> 0.4) (Fig. 28A; supplementary material Table S4 and Table S5). Indeed, 8.3% of K27HMDs identified at the blastula stage acquired elevated DNA methylation in both adult myotome and liver, and 9.6% and 5.3% showed elevated DNA methylation specifically in adult myotome and adult liver, respectively (Fig. 28B). Large K27HMDs (> 4kb) tend to be more frequently methylated in adult tissues, 15.7% for both, and 20.9% and 5.2% only for myotome and liver, respectively (Fig. 28B). Strikingly, the large K27HMDs with elevated DNA methylation still retained low methylation levels around TSSs in adult tissues, resulting in the HMD shortening toward promoter region (Fig. 28A and 29A, B). Similar to *zic1/zic4*, this change in DNA methylation occurred mainly after hatching stage in myotome (Fig. 29A).

Consistent with the fact that larger K27HMD have higher H3K27me3 enrichment in pluripotent cells, shortened K27HMD in adult tissues showed significantly lower levels of H3K27me3 than unchanged K27HMDs (Fig. 30A, B). Lower levels of H3K27me3 associated with elevated DNA methylation because the levels of H3K27me3 accumulation in blastula embryos did not show any difference between the two groups (elevated methylation and unchanged methylation in adult) of K27HMDs (Fig. 30C). To examine whether HMD shortening is correlated with gene expression, I performed RNA-seq for adult myotome and liver (supplementary material Table S2). Consistent with the reduced H3K27me3 level, gene expression levels were significantly higher in the K27HMDs with elevated methylation than

in ones with unchanged methylation (Fig. 30D, E). Thus, K27HMD shortening is associated with reduced H3K27me3 levels and active gene expression. By contrast, DNA methylation occurred in some nonK27HMDs, and in this case, the negative correlation between DNA methylation and gene expression was observed (Fig. 31). These results suggest the opposite functions of DNA methylation between nonK27HMDs and K27HMDs in gene expression.

Importantly, a correlation between the HMD size and H3K27me3 levels was also observed for unchanged K27HMDs in adult myotome and liver (Fig. 30A, B; unchanged small vs. unchanged large). Furthermore, the expression level was significantly lower for unchanged large K27HMDs than small ones (Fig. 30D, E; unchanged small vs. unchanged large). Taken together, these data suggest that the large K27HMDs have strong repressive effect on gene expression in pluripotent and adult matured cells, but the shortening of the large K27HMDs may facilitate persistent expression in adult tissues (Fig. 33).

HMD shortening associates with active gene expression also in zebrafish

To further investigate whether HMD shortening occur in other vertebrate species, I analyzed methylome and transcriptome data from zebrafish sphere stage embryos (an equivalent stage for medaka blastula) and adult myotome (Potok et al., 2013). I identified 237 HMDs larger than 8 kb in zebrafish sphere embryos, and 139 of those showed elevated methylation in adult myotome (supplementary material Table S9). Expression levels of genes associate to elevated methylation were significantly higher than that of unchanged HMD genes (Fig. 32). These results suggest that HMD shortening also associates with active gene expression in zebrafish. Therefore, the regulation of HMD size could be an important epigenetic mechanism for the strict and stable regulation of key developmental genes in vertebrates.

Discussion

The role of HMD shortening in key developmental gene regulation

During growth and differentiation, large K27HMDs undergo substantial changes in histone modification and DNA hypermethylation when their genes are activated. In Chapter 1, I found that K27HMD size correlates with transcriptional repression levels. Furthermore, DNA methylation was reported to antagonize H3K27me3 (Wu et al., 2010). These facts suggest that HMD shortening functions to promote transcription. The analysis of the large K27HMD harboring the *zic1/zic4* genes revealed that DNA hypermethylation is locus activity-dependent and occur after hatching, suggesting that HMD shortening is irrelevant to gene induction, but may consolidate the continuous expression of a gene that has been activated. Consistently, two distinct regulations of *zic1/zic4* transcription were previously demonstrated; the expression in somites is induced by secreted signals but is later maintained throughout life in a cell-autonomous manner (Kawanishi et al., 2013). Similar to *zic1/zic4*, it was reported that the embryonic *HOX* expression pattern is epigenetically maintained in fibroblasts of the human adult foot and is required to maintain its site-specific identity (Rinn et al., 2008). Notably, in our dataset, very large K27HMDs covering *hox* clusters observed at the blastula stage undergo HMD shortening at active loci in adult myotome. Given that DNA methylation is a stable epigenetic modification (Smith and Meissner, 2013), this K27HMD shortening could serve as cellular memory by marking developmental genes once activated (Fig. 33). The general correlation between elevated DNA methylation and active transcription at K27HMDs also supports this idea. I propose the two-step regulation of key developmental genes: histone modification-dependent induction and HMD shortening-dependent long-term maintenance. However, to test the model, further analyses that inhibit HMD shortening will be needed.

Molecular mechanisms underlying HMD shortening

Interestingly, Polycomb-target genes in ESCs often acquire hypermethylation after differentiation (Mohn et al., 2008; Rush et al., 2009), but the underlying mechanism is unknown (Deaton and Bird, 2011). The analysis of the large K27HMD harboring the *zic1/zic4* genes revealed that DNA hypermethylation is locus activity-dependent and gradually occurs after the histone modification pattern has changed. A likely scenario is that the reduction of H3K27me3 allows the accumulation of DNA methylation outside the H3K4me2-marked proximal promoter. In this model, the H3K27me3 and DNA methylation are likely to be mutually antagonistic. Indeed, a very recent study showed that, in ES cells, H3K27me3 regions recruit Dnmt3L (catalytically inactive DNA methyltransferase), which inhibits DNA methylation via competition with Dnmt3a and Dnmt3b (catalytically active DNA methyltransferase) (Neri et al., 2013). Although Dnmt3L is not found in non-mammalian vertebrates (Yokomine et al., 2006), a similar mechanism may exist in those species.

The above model, however, cannot directly explain HMD shortening events, as the elevation of DNA methylation gradually occurs after hatching, even though most of the genes have already been activated during embryogenesis. A positive regulator of DNA methylation that triggers HMD shortening after hatching may exist, and further molecular analyses will be required.

HMD shortening in other vertebrate species

The analyses with zebrafish data revealed that HMD shortening also occur in species other than medaka. Importantly, the elevated DNA methylation in large HMDs in zebrafish also associated with active transcription. Furthermore, a very recent study reported that, in human

medulloblastoma, hypermethylation occurs at some of the large hypomethylated domains and positively correlated with gene expression (Hovestadt et al., 2014). DNA methylation at the promoter regions of these genes were still at low levels, suggesting that this is an equivalent phenomenon to the HMD shortening in medaka and zebrafish. Thus, HMD shortening is a general feature among vertebrate species, and may also be involved in tumorigenesis.

Chapter 3

Sequence characteristics of hypomethylated domains

Introduction

How totipotency or pluripotency is established after fertilization is a central issue in developmental biology. Specific chromatin structures in pluripotent cells are believed to be crucial for maintaining the undifferentiation state and differentiation competence. The data in Chapter 1 and 2 suggest that the establishment of large K27HMDs at the loci of key developmental genes is an indispensable event. However, the underlying molecular mechanisms are not completely understood; for example, how are the future HMD regions determined in the genome and marked by specific chromatin modifications? To establish large K27HMD, the large continuous region in the genome should be first specified, and then marked by DNA hypomethylation and H3K27me3. Many studies have shown the involvement of various molecules and interplay between them.

First, it has been suggested that each chromatin modifications relies on each other. For instance, H3K4 methylation has been shown to maintain hypomethylation by protecting genome from DNA methyltransferase (Hu et al., 2009; Ooi et al., 2007). Indeed, HMDs are found to be exclusively marked by H3K4me2, and K27HMDs also have significant (but low) levels of H3K4me2 enrichment in medaka blastula embryos. However, DNA methylation was also reported to exclude H3K4 methylation (Okitsu and Hsieh, 2007), indicating the mutual antagonism between them. As described in Chapter 2, H3K27me3 and DNA methylation are also suggested to be mutually antagonistic. Thus, these studies suggest that chromatin states of K27HMDs are maintained by crosstalk of these modifications.

Second, in spite of their crosstalk, other studies have shown that each modification has the potential to takes place on target regions independently. H3K4 and H3K27 methylation are deposited at CpG islands by CXXC zinc-finger-domain-containing proteins Cfp1 and Kdm2b, respectively (Blackledge et al., 2014; Farcas et al., 2012; He et al., 2013;

Thomson et al., 2010). Tet proteins, considered as mediators of DNA demethylation, are also known to possess CXXC zinc-finger-domain, and suggested to target CpG islands (Wu et al., 2011a). However, the facts that not all CpG islands become HMD and only a portion of HMDs acquire H3K27me3 indicate that there are additional factors involved in specification of HMD regions. Indeed, several DNA binding factors and non-coding RNAs were reported to mediate H3K4 and H3K27 methylation, and DNA hypomethylation by interaction with chromatin modifiers (Arnold et al., 2013; Stadler et al., 2011; Wang et al., 2011). Together, above studies suggest that multiple factors act in concert to achieve the establishment of HMDs, and understanding how certain genomic regions are specified to large K27HMDs is a challenging problem.

An alternative approach to understand the logic of the domain establishment is to look for the features of DNA sequences inside the K27HMDs. Recent study showed that nonK27HMDs and K27HMDs can be predicted from DNA sequences with high accuracy using a supervised machine learning algorithm, Support Vector Machine (SVM) (van Heeringen et al., 2014). By using K27HMD sequences and nonK27HMD sequences as a training data set, this study showed that nonK27HMDs are enriched with housekeeping TF binding motifs, whereas K27HMDs are enriched with motifs for developmental regulators, and suggests that chromatin modification patterns are determined by underlying genomic sequences. However, we still do not understand how these short TF binding sequences could specify large contiguous domains.

In this chapter, I focus on the features of DNA sequences which determine HMDs. Medaka blastula embryos have strictly definable HMD boundaries, and thus, it is a suitable model for searching sequence characteristics for the domains. Indeed, I found unique characteristics of DNA sequences at HMD boundaries. Next, by using SVM, I found that large K27HMDs have specific short sequences densely along a entire domain. Unexpectedly,

some of the specific sequences are distributed on the coding exons of developmental genes., suggesting that coding regions may contribute to determine the epigenetic state of the genes.

Results

Specific sequence features of HMD boundary

We noticed that HMDs generally had sharp boundaries (Fig. 34A), suggesting that the boundaries may harbor specific sequences. Because HMDs largely overlap with CpG islands, we analyzed the CpG density around the HMD boundary. We found that the CpG density was significantly higher inside of HMD than outside. Interestingly, the CpG density dropped off around the HMD boundary and a CpG poor region spanned just outside of the HMD (Fig. 34B). CpG poor regions at HMD boundaries were also reported in human (Molaro et al., 2011), and may have a specific function in the establishment of the boundaries or may contribute to the sharpness. Furthermore, we identified motifs that are significantly enriched around the boundary sequence (Fig. 35). Interestingly, we noticed that one motif showed a high similarity to the CTCF binding sequence (Fig. 36A). Mapping of CTCF binding motifs on the medaka genome confirmed that they are indeed highly enriched around HMD boundaries (Fig. 36B). Furthermore, to confirm that the HMD boundaries are actually bound by CTCF, I analyzed CTCF ChIP-seq data in hESCs. Mapping of CTCF ChIP-seq peaks from ENCODE Project (2011) around HMDs showed the highest enrichment at human HMD boundary (Fig. 36C). Thus, HMD boundaries actually have binding sites of CTCF, and it is a conserved feature of HMD boundary between medaka and human. CTCF is known to mediate chromatin looping and functions as a barrier to separate distinct chromatin regions with different modifications (Handoko et al., 2011). Together, these results suggest that HMD boundaries are strictly confined and have specific features in genomic sequence.

Previous studies reported that nucleosome positioning is affected by CTCF binding as well as DNA methylation and histone modifications (Fu et al., 2008; Segal and Widom, 2009; Valouev et al., 2011). To determine the nucleosomal distribution around HMD

boundaries bearing a CTCF motif, we re-analyzed the previous data of nucleosome core distribution in medaka blastula embryos (Sasaki et al., 2009). We found clear peaks of nucleosome core signals, and a transition of the periodic pattern of the nucleosomal core position around the CTCF site; the average score of nucleosome core position exhibits a clear 170 bp periodic pattern outside HMDs but the peak becomes low and less defined inside HMDs (Fig. 37). The binding motif site showed no significant peak, indicating actual CTCF binding at the nucleosome-free region. These results suggest that the nucleosome structure also changes at the HMD boundary from "packed" to "loose".

Prediction of K27HMD and nonK27HMD from DNA sequence

Although specific sequence features exist at HMD boundaries, it is still difficult to predict from DNA sequence, which genomic regions become HMD and how their histone modification patterns are determined. In particular, how chromatin acquires DNA hypomethylation and histone modification along continuous large domains remains largely elusive. To address this issue, I utilized the SVM algorithm, a supervised machine learning approach that distinguishes two sets of data (Lee et al., 2011; van Heeringen et al., 2014). In brief, SVM finds a decision boundary that maximally distinguishes two classes of genomic sequences (positive and negative sequences) using frequencies of all sequences of length k (*k*-mers). In this study, I used 6-mers, and the basic approach is outlined in Figure 38. First, SVM was trained using sequences from all nonK27HMDs, except for those on chromosome 8, and similar number of methylated sequences from randomly selected regions. I refer to this trained SVM as SVM_{DNA}, as it is to classify DNA hypomethylated domains and methylated regions. To test the performance of SVM_{DNA}, the remaining nonK27HMD sequences and methylated sequences on chromosome 8 were classified by the trained machine. As there is trade-off between true positive rate and false positive rate, I calculated the area under the

ROC curve (ROCauc). High ROCauc (~ 1) corresponds to high accuracy and sensitivity. Consistent with the previous study (van Heeringen et al., 2014), SVM_{DNA} could distinguish nonK27HMD sequences from methylated sequences with high accuracy (Fig. 39A, C). Furthermore, even though I used nonK27HMD sequences for positive training data, it could also distinguish K27HMD sequences from methylated sequences (Fig. 39B, C). This result suggests that K27HMD and nonK27HMD share the same sequence characteristics for their hypomethylated state. Next, the machine was trained to distinguish K27HMD sequences from nonK27HMD sequences. I refer to this machine as SVM_{K27}, as it is to classify K27HMDs and nonK27HMDs. K27HMDs from chromosome 8 was excluded from the training data, and was used to test the accuracy of prediction. SVM_{K27} could accurately distinguish K27HMD from nonK27HMD in chromosome 8 (Fig. 40A, B). Thus, by using SVM_{DNA} and SVM_{K27}, K27HMDs, nonK27HMDs and methylated regions could be predicted from DNA sequences (Fig. 41).

To examine whether the HMD have specific sequence features along the entire domain, I calculated the SVM scores of 2 kb sequences along the all HMDs (both K27HMDs and nonK27HMDs) and comparable number of methylated sequences using SVM_{DNA} and SVM_{K27}. Most of the sequences were appropriately classified (Fig. 42), suggesting that epigenetic modification state of most of the regions are determined by their neighboring DNA sequences. For example, at the large K27HMD covering *hoxA* cluster, prediction scores for both SVM_{DNA} and SVM_{K27} were mostly positive throughout the entire domain (Fig. 43). These results suggest that large K27HMDs are clusters of K27HMD-specific sequences.

The SVM outputs a set of weights, which determine the contribution of each of all *k*-mers to a prediction (Table 1, 2). *k*-mers with large positive weights are sequence features specific to positive sequences (HMD sequences, in case of SVM_{DNA}), and *k*-mers with large negative weights are absent from positive sequences but present in negative sequences

(methylated sequences, in case of SVM_{DNA}). I examined the distribution patterns of *k*-mers with large positive weights at K27HMDs, and confirmed that K27HMD regions are specifically enriched with *k*-mers with large positive weights, and the frequencies greatly drop outside the boundary (Fig. 44). Thus, this suggests that continuous domains require specific sequences at a certain density, and regions that are poor with those specific sequences become boundary of the domains.

Exons of developmental genes have high enrichment of predictive sequence and H3K27me3

Interestingly, I found that some of the exons enrich large positive weight *k*-mers (Fig. 45), suggesting that the sequences of those exons contribute to K27HMD predictions by the SVM. Consistently, exons tend to have similar or higher H3K27me3 enrichment than introns, and in some cases, exons appear to locate exactly at the K27HMD boundaries (Fig. 46A, B). These results suggest that DNA sequences of exons contribute to the formation of K27HMDs. Furthermore, I noticed that some of those exons encodes important protein domains (e.g. homeobox domain, zinc-finger lim domain) of developmental genes. Indeed, comparison of the 6-mer frequencies in exon sequences revealed that homeobox-coding-exons have significantly higher enrichment of K27HMDs-specific 6-mers than other exons (Fig. 47). Thus, these results suggest that exons encoding functional protein domains serve as determinants for hypomethylation and H3K27me3 accumulation and contribute to the enlargement of K27HMDs.

Discussion

The logic for large K27HMD establishment

In this chapter, I demonstrated that the SVM can predict K27HMDs and K4HMDs accurately from DNA sequences in the medaka blastula genome. These results suggest that the information for the domain establishment in pluripotent cells is encoded in the sequence of genomic DNA, and do not need any inheritable information. The sequences that highly contribute to the classification of K27HMD, nonK27HMD and methylated sequences may have important function in regulating the chromatin modifications in cells. Indeed, HMD-specific sequences tend to contain CpG dinucleotide, consistent with the fact that CXXC zinc-finger-domain-containing proteins mediate H3K4 and H3K27 methylations at CpG rich regions. However, it is well known that DNA methylation affects the mutation rate, and as a result, hypomethylated regions tend to become CpGs rich over a long evolutionary period. Thus, whether the HMD-specific sequences feature found by the SVM actually a primary determinant for hypomethylation or just a consequence of hypomethylation is unknown. To my knowledge, however, the different mutation tendencies between H3K4me and H3K27me regions have not been reported. Therefore, it is more likely that at least *k*-mers with large weight for K27HMD and nonK27HMD classification have actual functions in accumulating each histone modification. It will be interesting to ask which factors recognize these sequences. In the recent study using Xenopus, Heeringen et al. reported that nonK27HMDs are enriched with housekeeping TF binding motifs, whereas K27HMDs are enriched with motifs for developmental regulators (van Heeringen et al., 2014). Linking high weighted *k*-mers with genome-wide DNase I footprinting data will provides cues for actual binding of the transcription factors at the specific *k*-mers.

Furthermore, the SVM could also predict continuous domains. In large K27HMDs,

sequences that show high SVM scores distribute contiguously along the entire domain. This suggests that large K27HMDs are formed by cluster of K27HMD-specific sequences, and this way of formation could assure the robustness of epigenetic functions of the domain. I showed in Chapter 1 that large K27HMD sequences are highly conserved between medaka and human, and have also been reported to be free from transposable elements and repeats (Jeong et al., 2014). Transposon insertions may interfere with large domain formation and thus those sequences might be excluded from large K27HMD regions during genome evolution.

Epigenetic state of developmental genes determined by their own coding sequences

It is widely accepted that the information for transcriptional regulation is encoded in non-coding regions of the genome. Indeed, non-coding regions are known to contain most of the regulatory elements, such as promoters, enhancers, silencers, and insulators. However, it has been reported that some coding exons can potentially function as enhancers or silencers in vertebrates (Birnbaum et al., 2012; Khan et al., 2012; Ritter et al., 2012). More recently, genome-wide DNase I footprinting in human cells revealed that transcription factors frequently bind to coding exons (Stergachis et al., 2013a). This study suggested that ~15% of human codons are dual-use codons ("duons") that simultaneously encode both amino acids and transcription factor binding sites. Thus, these studies have suggested the involvement of coding sequences in transcriptional regulation.

In this chapter, I examined if some of the coding exons of developmental genes posses the K27HMD-specific sequence features. Indeed, K27HMDs tend to contain two or more exons, while most nonK27HMDs only contain single exon. These exons actually have 6-mers with large positive SVM weights, suggesting that those regions are important for K27HMD formation. Although further analyses by reporter assays and genome editing

techniques will be needed before I conclude the actual functions of these exons, strong signals of H3K27me3 at the exons strengthen their role in H3K27me3 accumulation. Furthermore, I found that exons coding homeobox domain tends to be enriched with K27HMD-specific sequences higher than other exons. This suggests that specific protein domains have propensity toward hypomethylation and H3K27me3, and may partly explain the specific marking of developmental genes by poised chromatin modifications. The identification of factors that recognize the exons and their functions await further experiments.

Conclusions

Using medaka as a model organism, I revealed several novel characteristics of unique but highly conserved genomic domains, the large K27HMD. This particular domain has been reported independently in several previous studies, because of its unusual size and distinct localization at genes crucial for development. However, the biological significance has not been addressed so far. On the basis of the findings in my study, I propose the large K27HMD as key epigenetic machinery for both pluripotency and long-term cell fate maintenance as described below (Fig. 48).

In pluripotent cells, differentiation competence is exerted by the repressed yet inducible (poised) state of developmental genes. Hypomethylated promoters with both active and repressive histone modifications are considered to adopt the poised state, but by nature they cannot avoid leaky transcriptions. The large K27HMD reconciles this dilemma by assuring strong accumulation of repressive histone modifications, especially for genes encoding key transcriptional regulators that have profound impacts on cell fate determination.

After embryogenesis, vertebrate cells have to maintain their fates for long periods of time, even though most of the stimuli that have induced the fates no longer exist. During animal growth, cells deposit DNA methylation at gene loci that are actively transcribed, leading to shortening of large HMDs. This change makes the promoters of key developmental genes active-type such as those of housekeeping genes (small HMD with active histone modifications), which are protected from repressive histone modification. Consolidations of key developmental gene expression then would play a major role in the maintenance of the cell identity.

Although large K27HMDs are conserved among vertebrates, given that DNA methylation is not observed in some invertebrate species, this modification would not

generally be essential for animal development. Meanwhile, DNA methylation is utilized in a wide range of organisms (bacteria, fungi, plants, and animals) and involved in various biological processes, such as restriction modification system in bacteria, transposon silencing in plants, and gene imprinting in mammals. Intriguingly, vertebrates have also achieved global DNA methylation, which is unique in animal species. Although the role of global DNA methylation is still largely unknown, it is possible that it was adopted to regulate the size of K27HMDs in vertebrates, which is needed to sophisticate epigenetic gene regulation for their long lifetime.

Materials and Methods

Fish strains

I used medaka d-rR strains as wild type. *Da* mutant used in this study was the same strain as previously described (Moriyama et al., 2012). I crossed the *Da* homozygous mutant and d-rR to generate *Da* heterozygous mutants. Medaka fishes were maintained and raised under standard condition. Identification of developmental stages was performed as previously described (Iwamatsu, 2004).

Whole mount in situ hybridization

Whole mount in situ hybridization was performed as previously described (Takashima et al., 2007). For *tbx2a* and *2b*, primers for DIG-labeled RNA probes were listed on Supplementary Table 3. For *pax6a* and *6b*, 3'UTR was cloned by SMARTer RACE cDNA Amplification kit (Clontech) and used for RNA probe synthesis. Primers are listed on supplementary material Table 3.

RT-qPCR

RNA was isolated from adult wild type and *Da* mutant muscle (dorsal and ventral part were separated) using ISOGEN (NIPPON GENE) according to the manufacture's protocol. SuperSucript III (Invitrogen) was used for cDNA synthesis. qPCR was performed with the Stratagene Mx3000P system (agilent technologies) using the THUNDERBIRD SYBR qPCR mix (TOYOBO). All primers were listed on supplementary material Table 3.

Bisulfite sequence

Genomic DNA was isolated from adult and larva muscle (dorsal and ventral part were separated) and performed bisulfite treatment using methyl easy DNA Bisulphite Modification Kit (Human Genetic Signatures). Bisulfite converted DNA was subjected to PCR using Ex Taq (TaKaRa) and TOPO-TA cloning (life technologies). Amplified fragments were sequenced and analyzed and visualized by the QUMA software (Kumaki et al., 2008). All primers were listed on Table 3.

Whole genome bisulfite sequence (WGBS)

For adult and larva muscle (dorsal and ventral part were separated), sample preparation, library construction, sequencing and mapping for WGBS were performed as previously described (Qu et al., 2012). Briefly, Genomic DNA from medaka was isolated and sonicated to a desired size range (100–400 bp). The DNA fragments were treated with DNA polymerase to generate blunt ends and were ligated with double-stranded DNA adaptors containing methylated cytosines, which were designed to amplify only those DNA fragments carrying bisulfite-converted adaptor sequences at both ends. Followed by 7–10 cycles of PCR, 250 –450-bp size- fractionated DNA was sequenced using an Illumina GAIIx genome analyzer. All cytosines in reads and in both the Watson and Crick strands of the reference genome were converted to thymines for primary mapping and used Smith-Waterman alignments between the original sequences of primary best hits. The level of methylation of a particular cytosine was estimated by dividing the number of mapped reads reporting a cytosine (C) by the total number of reads reporting a C or T (thymine).

ChIP

ChIP was performed using the following antibodies: H3K27me3 (Millipore, 07-449), H3K4me1 (Millipore, 07-436), H3K4me2 (Millipore, 07-030), H3K4me3 (Millipore, 07-473), H3K27Ac (abcam, ab4729). ChIP was peformed as previously described with modifications (Lindeman et al., 2009). Cells were dissociated using a 21G needle and fixed with 1% formaldehyde for 10 min at RT then quenched by adding Glycine (200 mM final). After washing with PBS containing 20 mM Na-butylate, complete protease inhibitor (Roche) and 1 mM PMSF, cell pellets were suspended by Lysis buffer (50 mM Tris-HCl pH8.0, 10 mM EDTA, 1 % SDS, 20 mM Na-butylate, complete protease inhibitors, 1 mM PMSF), sonicated for 10 times at vol. 5 of sonifier (Branson), and centrifuged for collecting chromatin lysates. The chromatin lysates were diluted with ChIP RIPA buffer (10 mM Tris-HCl pH8.0, 140 mM NaCl,1 mM EDTA, 0.5 mM EGTA, 1 % Triton-X100, 0.1 % SDS, 0.2 % Na-Deoxycholate, 20 mM Na-butylate, complete protease inhibitors, 1 mM PMSF) and rotated with antibody/protein A Dynabeads complex for over night at 4 ^{o}C. Immunoprecipitated materials were washed three times with ChIP RIPA buffer and once with TE buffer, followed by elusion with Lysis buffer at 65 ^{o}C over night. Eluted samples were treated with RNase A for 2 h at 37 ^{o}C and proteinase K for 2 h at 55 ^{o}C, and DNA was purified by Phenol/Chloroform extraction and ethanol precipitation. Input DNA was simultaneously treated from the elution step. Input and immunoprecipitated DNAs were analyzed by quantitative PCR with the Stratagene Mx3000P system (agilent technologies) using the THUNDERBIRD SYBR qPCR mix (TOYOBO). All primers were listed on supplementary material Table 3.

ChIP-seq

For ChIP-seq analysis, I used ~10^6 cells for one modification. Using the input and immunoprecipitated samples of ChIP, ChIP-seq templates were prepared using the TruSeq DNA sample prep kit (version 2; illumina) and sequenced using Genome analyzer IIx (illumina). Sequence was read by 36-base-single end sequencing.

RNA-seq

RNA was isolated from adult muscle (dorsal and ventral part were separated) and adult liver using RNeasy mini kit (QIAGEN) or ISOGEN (NIPPON GENE) according to the manufacture's protocol. Purified RNAs were treated with Ribominus eukaryote kit for RNA-seq (life technologies). RNA-seq analysis was conducted basically following the instructions from the manufacturer. Briefly, the RNA-seq template was prepared using TruSeq RNA-seq sample prep kit (version 2; Illumina) omitting the polyA selection procedure. The double stranded PCR products were purified and size fractionated by bead-mediated method using AMPure (Ambion). Sequencing was conducted on Genome analyzer IIx platform (Illumina) using TruSeq Cluster generation kit. At least 20 million sequences of 36-base-single read were generated per sample.

Data access

All sequence data are deposited at the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) (accession number SRP029233)

Alignment for ChIP-seq and RNA-seq reads

After removing low-quality reads (those containing five or more undetectable bases in 36mer-length reads), remaining reads were mapped using Burrows-Wheeler Aligner mapping

software (http://bio-bwa.sourceforge.net); no more than three mismatches and no gap were allowed. Only uniquely mapped reads were used for the further analysis. The medaka genome and predicted gene sequences were downloaded from Ensembl database (http://www.ensembl.org).

ChIP-seq data processing and analysis

After the mapping, redundant reads were excluded to avoid potential PCR bias. QuEST software (version 2.4 (Valouev et al., 2008)) was used to detect H3K27me3 peak in blastula embryos. WIG data generated by QuEST were also used for ChromHMM and for the visualization of other modifications. The following QuEST 2.4 settings were used for blastula H3K27me3: KDE (kernel density estimation) bandwidth = 100, ChIP seeding fold enrichment = 10, ChIP extension fold enrichment = 3, ChIP-to-background fold enrichment = 3. For other ChIP data visualization, WIG files were generated using following KDE bandwidth settings: H3K4me1: KDE bandwidth = 100; H3K4me2: KDE bandwidth = 60; H3K4me3: KDE bandwidth = 60; H3K27me3: KDE bandwidth = 100; H3K27Ac: KDE bandwidth = 60.

RNA-seq data processing and analysis

The BAM files of mapped reads were used for calculating the normalized RPKM with SAMMATE software (version 2.6.1 (Xu et al., 2011)).

Identification of HMDs

In this study, medaka genomic regions on scaffolds and ultracontigs were excluded from the analysis. I searched for contiguous regions of low methylated (methylation ratio lower than 0.4) CpG sites. ‘CpG site’ represents a single CpG in the genome. The two close low

methylated regions were connected and considered as one domain when they were divided by less than four high methylated CpG sites (methylation ratio higher than 0.4). Regions with more than 9 low methylated CpG sites were defined as hyomethylated domains (HMDs). The boundaries of HMD were defined at the first low methylated CpG site inside the HMD regions. The position of TSS was defined according to the ensemble database (http://www.ensembl.org). Promoter region in this study was defined as regions from 5 kb upstream to 2 kb downstream of the TSS. The HMD overlapped with the promoter region was assigned to the gene, and used to classify genes. When more than two HMD overlapped with promoter region, the closest one to the TSS was selected. CpG island was defined as GC >= 50%, O/E >= 0.6 and length = 200 bp, according to the previous study (Kasahara et al., 2007).

Identification of K27HMDs

K27HMDs were identified as the presence of H3K27me3 peaks inside the HMD, whereas nonK27HMDs were identified as the absence of H3K27me3 peaks according to QuEST analysis. For figures 2G, 3E and 6D, the number of mapped reads inside HMD of each category was counted and normalized by the length of HMD. For figures 2H, the highest peak signal intensity within the K27HMD calculated by QuEST was compared with low methylated CpG counts (the number of low methylated CpG sites).

Identification of H3K27me3 enriched hypomethylated regions by ChromHMM

We calculated K27HMD regions from methylation levels at individual CpG sites and H3K4me2/ H3K27me3 ChIP-seq data. First, for each running window of 200bp in length, we computed the average of 10(1 – x), where x was the methylation level at each CpG site, and the averages of H3K4me2/H3K27me3 ChIP-seq signal levels calculated by QuEST in the

window. Subsequently, we input these epigenetic values into ChromHMM, which used the Poission distribution to model the input data, and set a real value to 0 if its p-value < 0.0001, and to 1, otherwise. After this binarization, ChromHMM decomposed the input DNA sequence into regions of six chromatin states using the Hidden Markov Model. Finally, we selected such regions that the methylation level was low while H3K27me3 ChIP-seq signal was high, and examined the overlap with K27HMD regions.

Identification of HMD with elevated DNA methylation in adult tissues

The HMDs in which methylation increased were identified as >5 % of low methylated CpGs inside the HMD at blastula became high methylated (> 0.4) in adult myotome (dorsal) or liver. For figure 4C, average CpG methylation ratio in 25 bp window was calculated.

Heatmap generation

For DNA methylation, average methylation in 250 bp window was calculated. The level of windows without CpG site was set to 1. For histone modifications, the numbers of mapped reads in 250 bp window were counted. Heatmap was visualized in log scale using Java TreeView software.

Motif analysis at HMD boundaries

For the identification of motifs around boundaries of large K27HMDs (> 4 kb), I analyzed the sequence of 200 bp (100 bp each for upstream and downstream) using MEME software. Discovered motifs were subjected to TOMTOM software and the CTCF motif was identified. Next I analyzed the genome-wide distribution of CTCF motifs using FIMO software (I set the threshold of p value 1.0E-5) and counted based on the position from the center of HMD. All softwares used here were on MEME suite (http://meme.nbcr.net/meme/).

Analysis of nucleosome positioning around CTCF motifs

For analyzing the nucleosome positioning aroung CTCF motifs, I utilized the genome-wide nucleosome positioning data sets from the previous report (Sasaki et al., 2009). The nucleosome density around CTCF motifs located near HMD boundary (100 bp upstream to 400 bp downstream from the HMD boundary) was averaged.

Gene Ontology (GO) analysis

For the GO analysis, I selected the genes that have human orthologous genes because of the absence of a GO platform in medaka. Among 18,028 medaka ensemble genes (genes on scaffolds and ultracontigs were excluded), 13,301 genes have their single human ortholog. GO analysis was performed by DAVID program (Huang et al., 2009). All human orthologous genes (13,301) were used for background. PANTHER biological processes category was used.

Sequence conservation analysis

For the analysis of sequence conservation among vertebrates, the average PhastCons scores profiles around the HMD boundaries were generated with the Conservation Plot tool, part of the Cistrome Analysis pipeline (http://cistrome.dfci.harvard.edu/ap/). The identity of amino acid sequence between medaka and human proteins were downloaded from Ensembl database and compared among groups of DNA methylation state.

Analysis of human ESCs data

For the comparative analysis between medaka and human, ChIP-seq, and RNA-seq data sets were downloaded from previous reports (Lister et al., 2009; Lister et al., 2011) (GEO

accession no.: GSE16256). Mapping to hg19 reference genome was performed using bowtie program (Langmead et al., 2009). For WGBS, I downloaded processed wig data from Lister et al., 2009, and converted to hg19 using UCSC LiftOver. Low methylated CpG was defined as < 0.6 in human. Other analyses were performed in the same way as medaka data sets. The following QuEST settings were used for H3K27me3 peak detection and visualization; H3K27me3: KDE bandwidth = 100, ChIP seeding fold enrichment = 20, ChIP extension fold enrichment = 3, ChIP-to-background fold enrichment = 3; H3K4me2: KDE bandwidth = 60.

CTCF ChIP-seq peak distribution around HMD boundaries in human

CTCF ChIP-seq peak position file was downloaded from the UCSC genome browser (The ENCODE Project Consortium 2011; GSM733672). The relative position to the HMD boundary was counted.

Analysis of zebrafish data

Methylome and RNA-seq data were downloaded from previous reports ((Potok et al., 2013) Accession number: SRP020008). For Bisulfite data, only one end of the paired-end reads was used in this study. Analyses were performed in the same way as medaka data sets.

Support vector machine analyses

For SVM analysis, I used the method previously described (Lee et al., 2011). For SVM_{DNA}, I used all nonK27HMD sequences except for those on chromosome 8 as positive training data. The sequences larger than 2 kb were divided into two. All nonK27HMD regions were also randomized on methylated regions and used as negative training data. For SVM_{K27}, I used all nonK27HMD and K27HMD sequences except for those on chromosome 8 as positive and negative training data. The sequences were divided into fragments shorter than 1.5 kb. For

both SVM_{DNA} and SVM_{K27}, the sequences on chromosome 8 were used for testing performance. For Fig. 38C and 39B, 2 kb sequences were extracted from the example regions (chr8:12400000-12500000) by 500 bp steps, and classified by each SVM. For Figure 42, 2 kb sequences were extracted along the HMDs by 500 bp steps. HMDs were also randomized on methylated regions and sequences were extracted similarly. SVM scores were calculated, and plotted. The sequences next to each other are connected with a line.

Statistical analysis and the data visualization

The statistical analysis and graph visualization were performed using R software (version 2.14.2). For the visualization of genome-wide data, I integrated the data into UTGB genome browser (Saito et al., 2009).

Figures

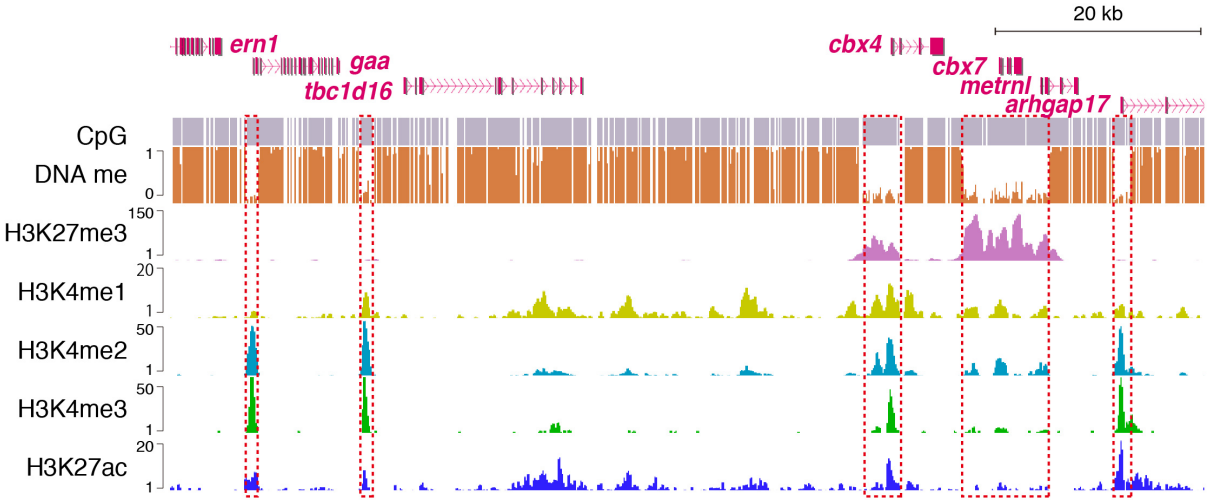

Figure 1. DNA and histone modification patterns in medaka blastula embryos

Genome browser representations of DNA methylation, H3K27me3, H3K4me1, H3K4me2, H3K4me3, and H3K27ac in medaka blastula embryos are shown. Red dashed boxes indicate HMDs.

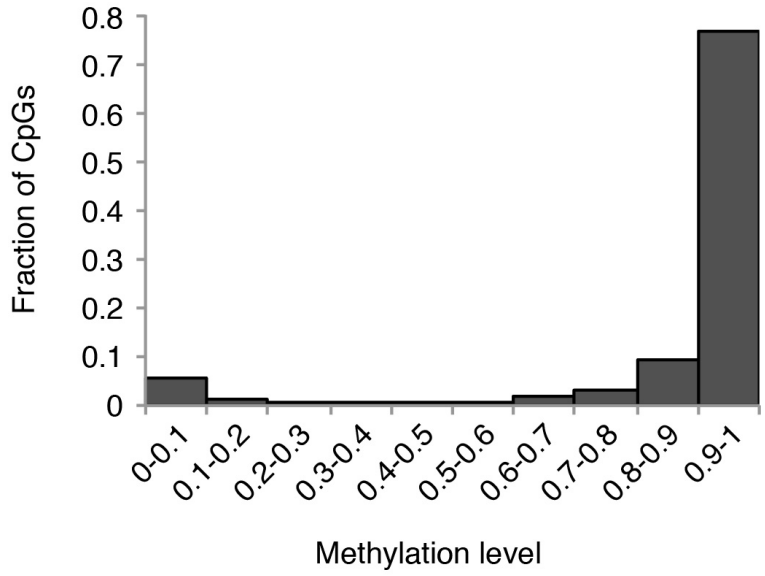

Figure 2. Bimodal distribution of methylation frequency of individual CpG site

Histogram of methylation level (methylation frequency) at each CpG site for blastula embryos.

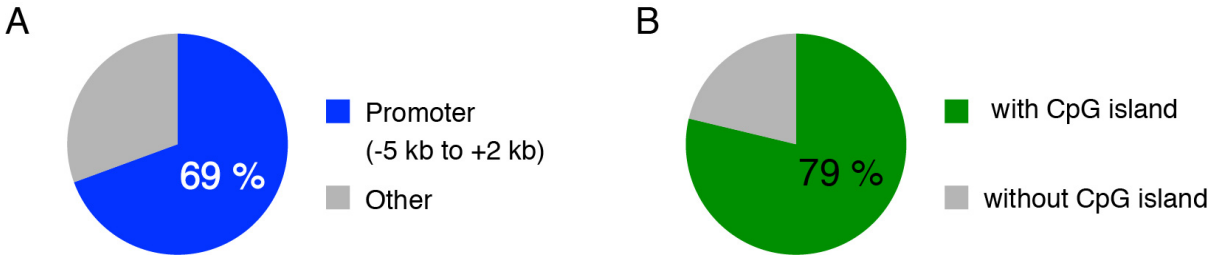

Figure 3. DNA and histone modification patterns in medaka blastula embryos

Fractions of HMDs overlapped with promoters (defined as regions from 5 kb upstream to 2 kb downstream to the TSSs) (A) and CpG islands (B).

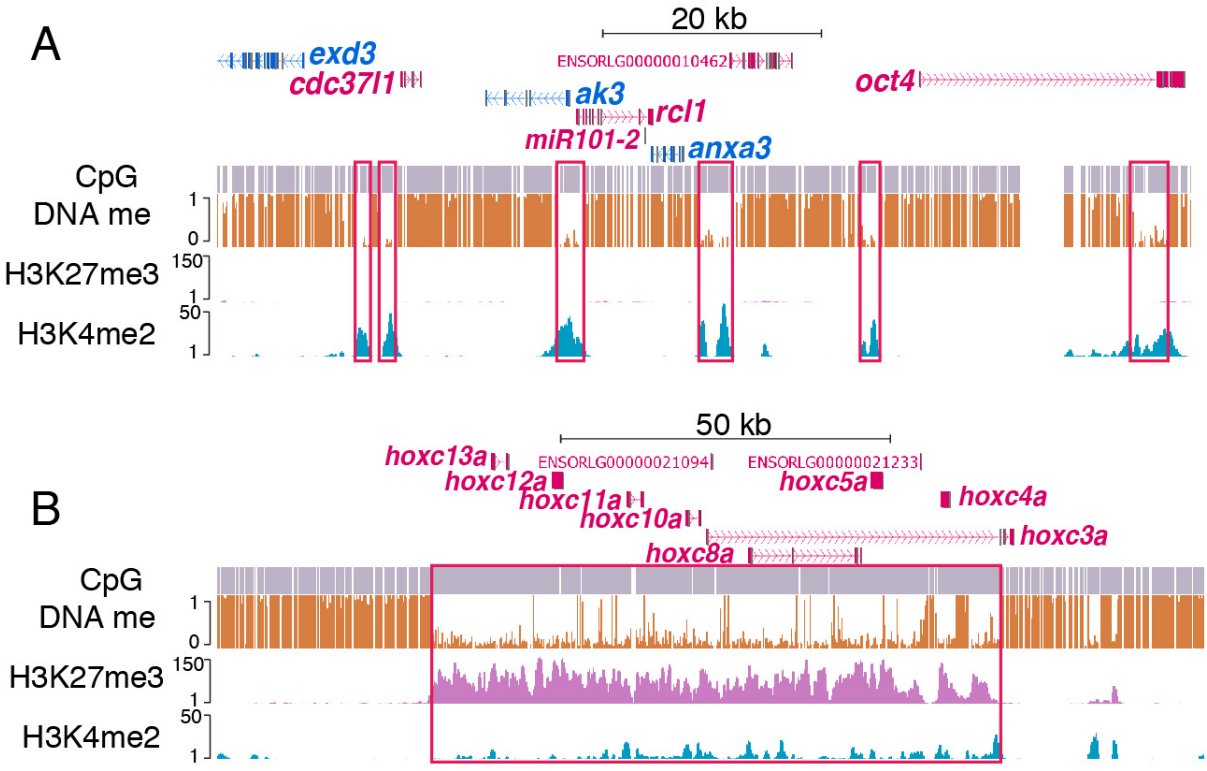

Figure 4. Examples of typical and large HMDs in medaka blastula embryos

Genome browser representations of DNA methylation, H3K27me3 and H3K4me2 in medaka blastula embryos are shown for representative typical HMD (A) and large HMD (B). Red boxes indicate HMDs.

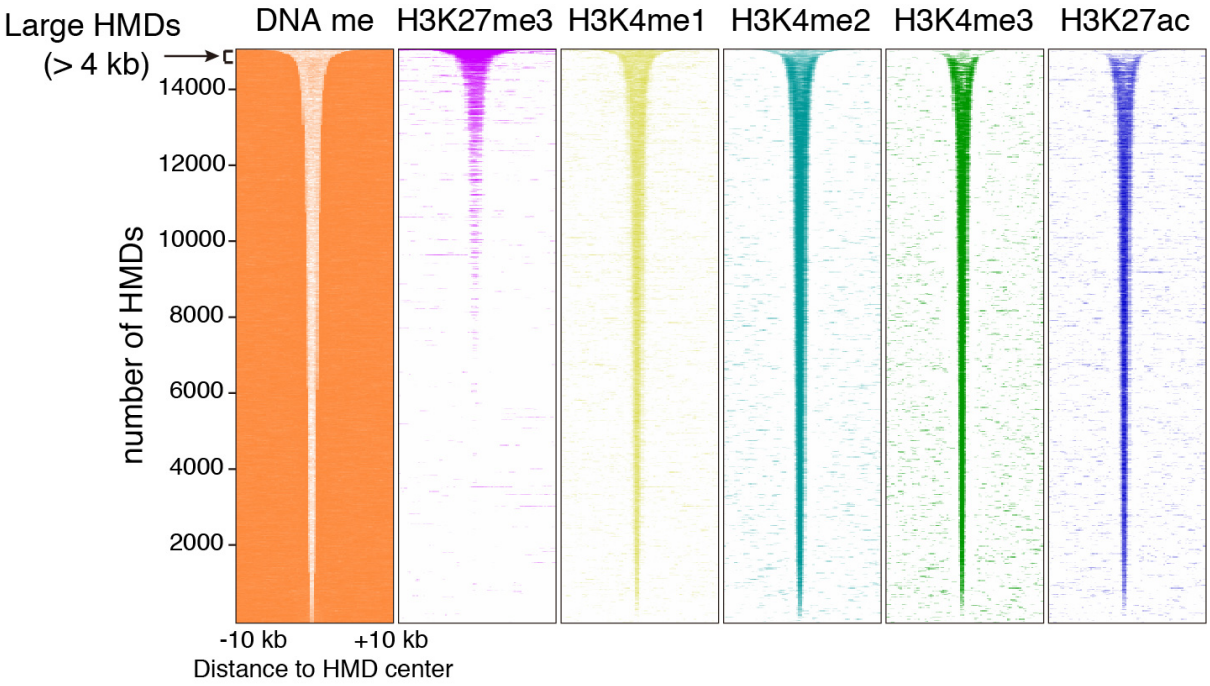

Figure 5. Histone modification patterns at all HMDs

Heat map representations of DNA methylation, H3K27me3, H3K4me1, H3K4me2, H3K4me3 and H3K27ac are shown for all HMDs in medaka blastula embryos. HMDs were ordered by size. Dark colors represent high signal, and white represents low signal.

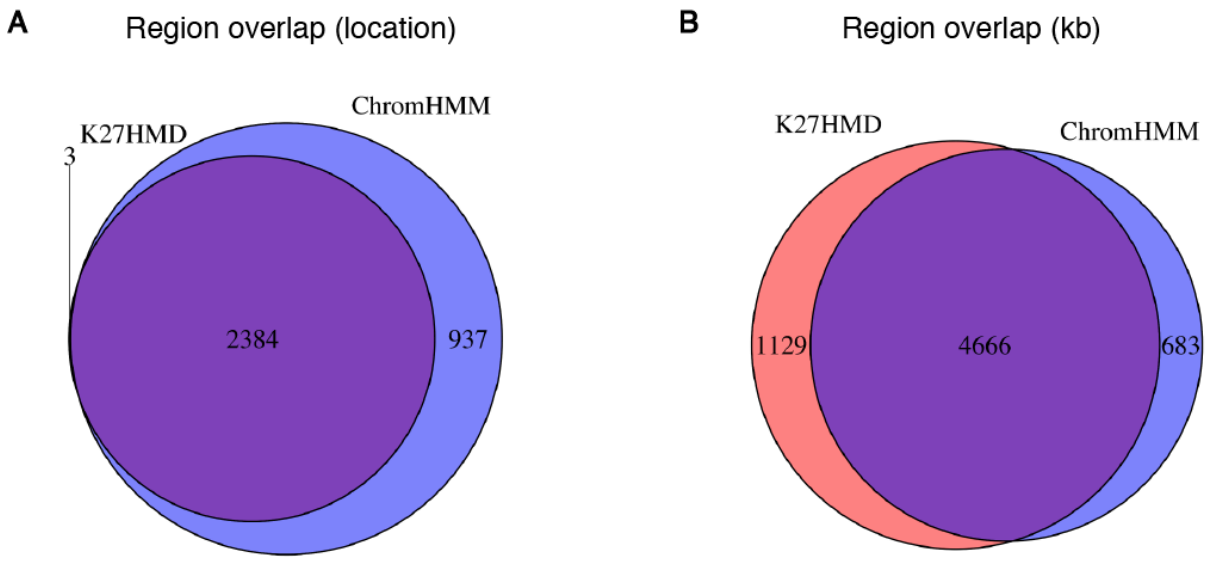

Figure 6. Comparison of K27HMD and region called by ChromHMM

(A) Venn diagram showing the number of K27HMD only, ChromHMM called only, and shared locations. Most of the K27HMDs overlapped with regions called by ChromHMM.

(B) Venn diagram showing the overlapped length of genomic regions annotated as K27HMD or ChromHMM.

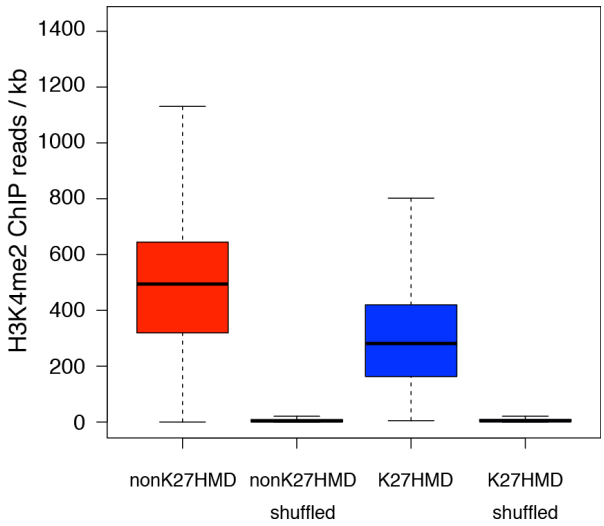

Figure 7. H3K4me2 enrichment at HMDs

Boxplots showing the enrichment of H3K4me2 ChIP reads for nonK27HMD, K27HMD and randomized regions. In the box plots, the bottom and top of the boxes correspond to the 25th and 75th percentiles and the internal band is the 50th percentile (median). The plot whiskers extending outside the boxes correspond to the lowest and highest datum within 1.5 interquartile range of the lower and upper quartiles, respectively.

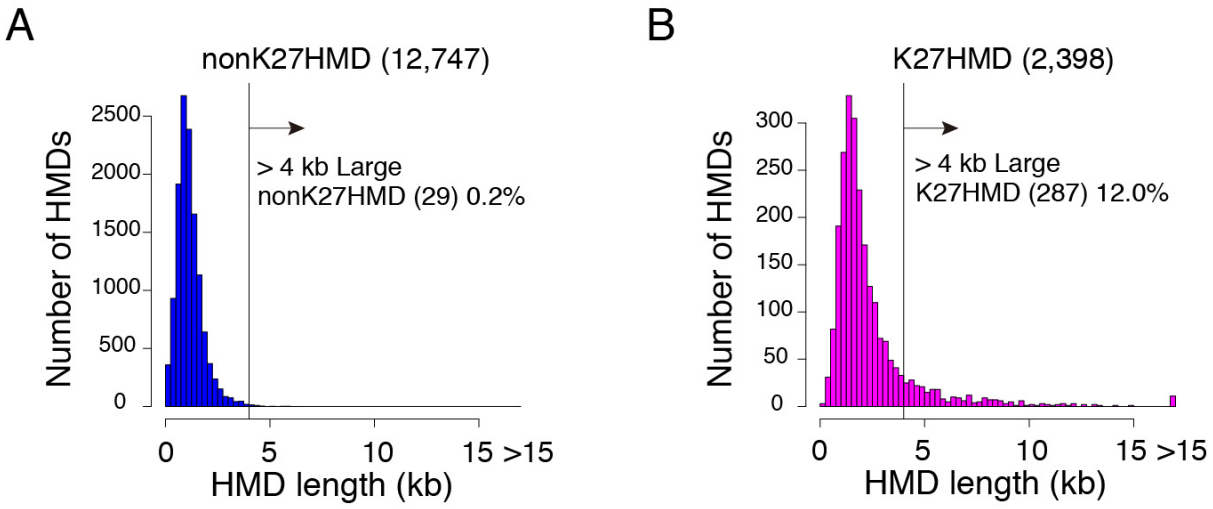

Figure 8. Size distribution of HMDs

Size distributions of nonK27HMDs and K27HMDs in blastula embryos are shown for nonK27HMDs (A) and K27HMDs (B).

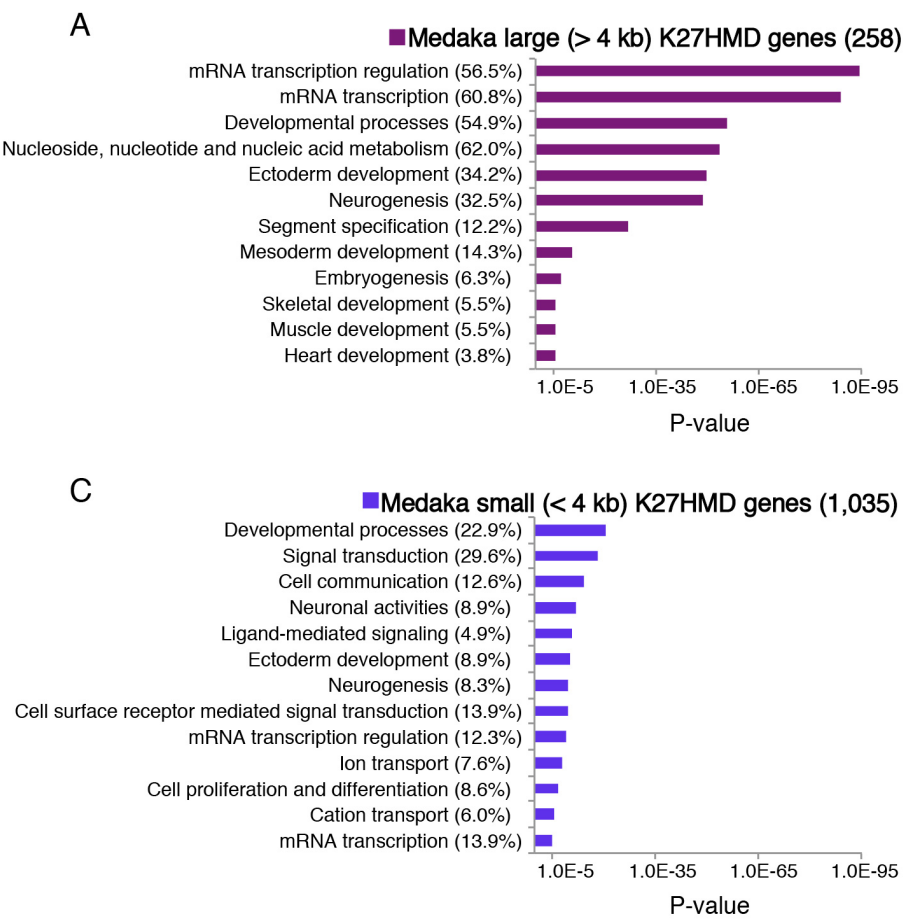

Figure 9. Gene ontology analyses of small and large K27HMDs

(A) Functional annotations of genes marked by large K27HMDs. The over-represented Gene Ontology (GO) PANTHER biological process terms and fraction of genes associated with each GO term are shown. The values of x axes (in logarithmic scale) correspond to the P-values calculated by the DAVID tool (Huang da et al., 2009).

(B) Examples of transcription factors included in large K27HMD genes.

(C) Functional annotations of genes marked by small K27HMDs. The over-represented Gene Ontology (GO) PANTHER biological process terms and fraction of genes associated with each GO term are shown.

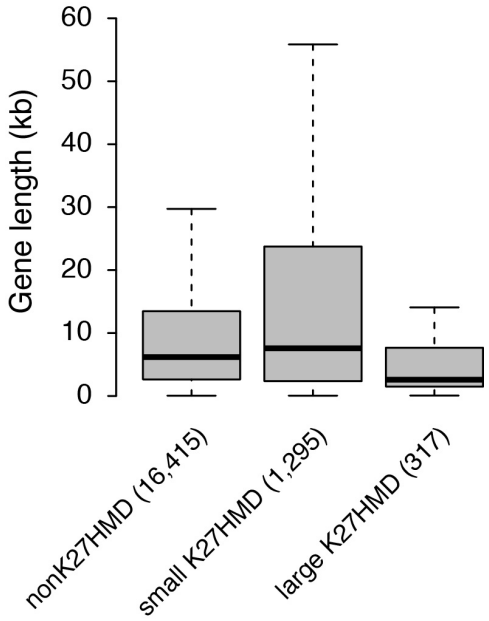

Figure 10. Comparison of gene length between HMD categories

Boxplots show the gene length of each HMD category. In the boxplots, the bottom and top of the boxes correspond to the 25th and 75th percentiles and the internal band is the 50th percentile (median). The plot whiskers extending outside the boxes correspond to the lowest and highest datum within 1.5 interquartile range of the lower and upper quartiles, respectively.

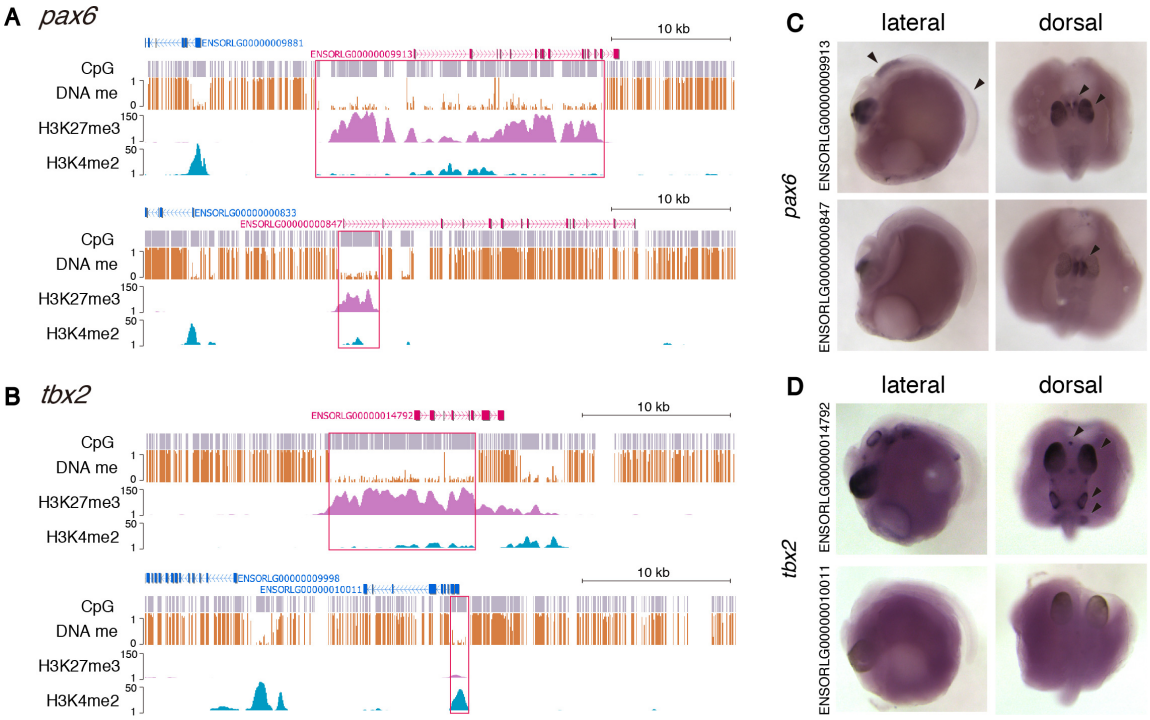

Figure 11. Comparison of gene length between HMD categories

(A and B) Genome browser representation of DNA methylation, H3K27me3 and H3K4me2 enrichment are shown for duplicated genes, *pax6* (A) and *tbx2* (B).

(C and D) in situ hybridization for *pax6* (C), and *tbx2* (D) in somite stage medaka embryos (st. 27). Lateral view (left) and dorsal view (right) are shown. Arrowheads point to specific expressions.

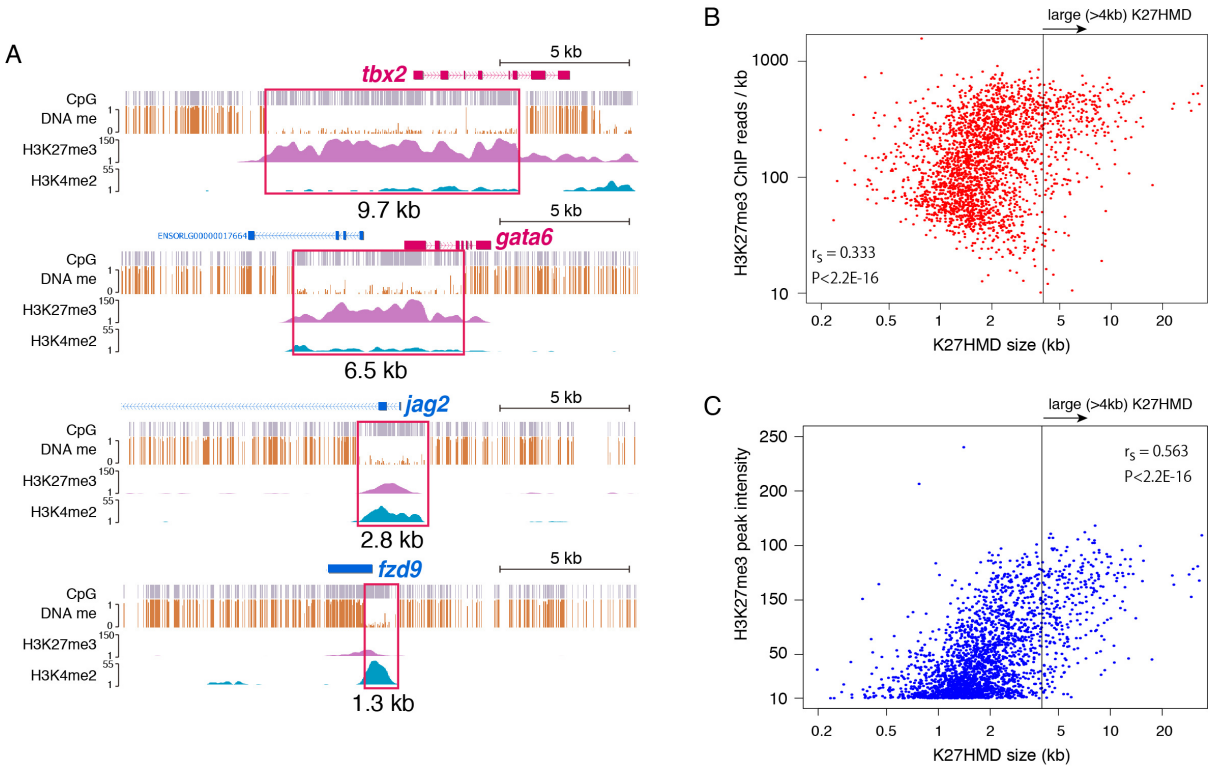

Figure 12. Size of K27HMD correlates with H3K27me3 level

(A) Genome browser representations of DNA methylation, H3K27me3 and H3K4me2 in medaka blastula embryo are shown for large K27HMDs (*tbx2* and *gata6*; top two) and small K27HMDs (*jag2* and *fzd9*; bottom two)).

(B and C) Comparison of the size of K27HMD with mapped ChIP reads per kb (B) and ChIP peak intensity calculated by QuEST (C) for H3K27me3. Spearman's rank correlation coefficient (r_s) and p value are shown.

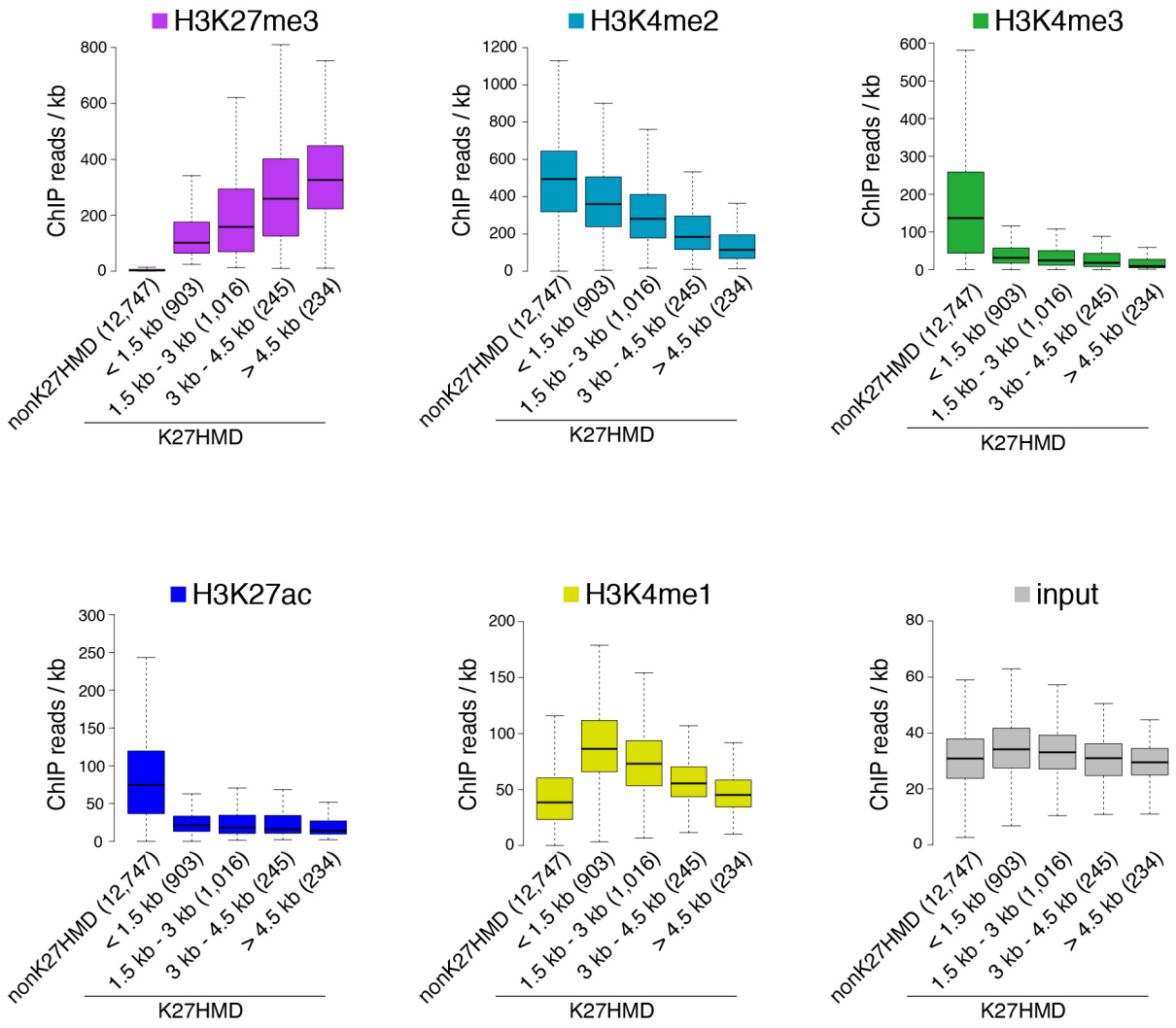

Figure 13. Relationships between K27HMD size and histone modification enrichment

Boxplots show the correlation between the HMD length and mapped ChIP reads per kb for indicated histone modifications and input DNA. In the boxplots, the bottom and top of the boxes correspond to the 25th and 75th percentiles and the internal band is the 50th percentile (median). The plot whiskers extending outside the boxes correspond to the lowest and highest datum within 1.5 interquartile range of the lower and upper quartiles, respectively.

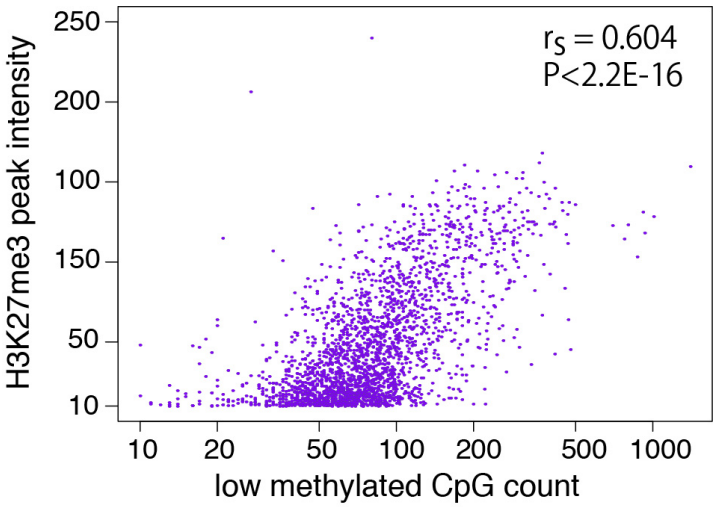

Figure 14. Number of low methylated CpG sites correlates with H3K27me3 level

Comparison of the number of low methylated CpG sites and the highest H3K27me3 ChIP peak intensity inside K27HMD. Spearman's rank correlation coefficient (r_s) and p value are shown.

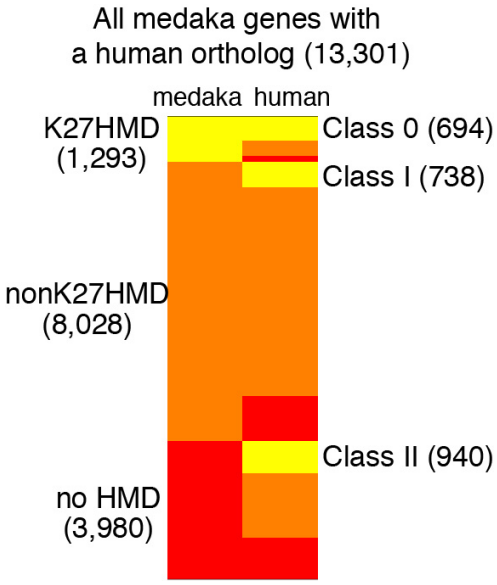

Figure 15. Comparison of epigenetic state of gene promoters between medaka blastula embryos and human ES cells

HMD status in medaka blastula embryos and human ES cells is shown for promoters of 13,301 genes that had an orthologous human gene annotated in the Ensembl database. Yellow indicates the presence of a K27HMD; orange indicates nonK27HMD; red indicates no HMD at promoter regions.

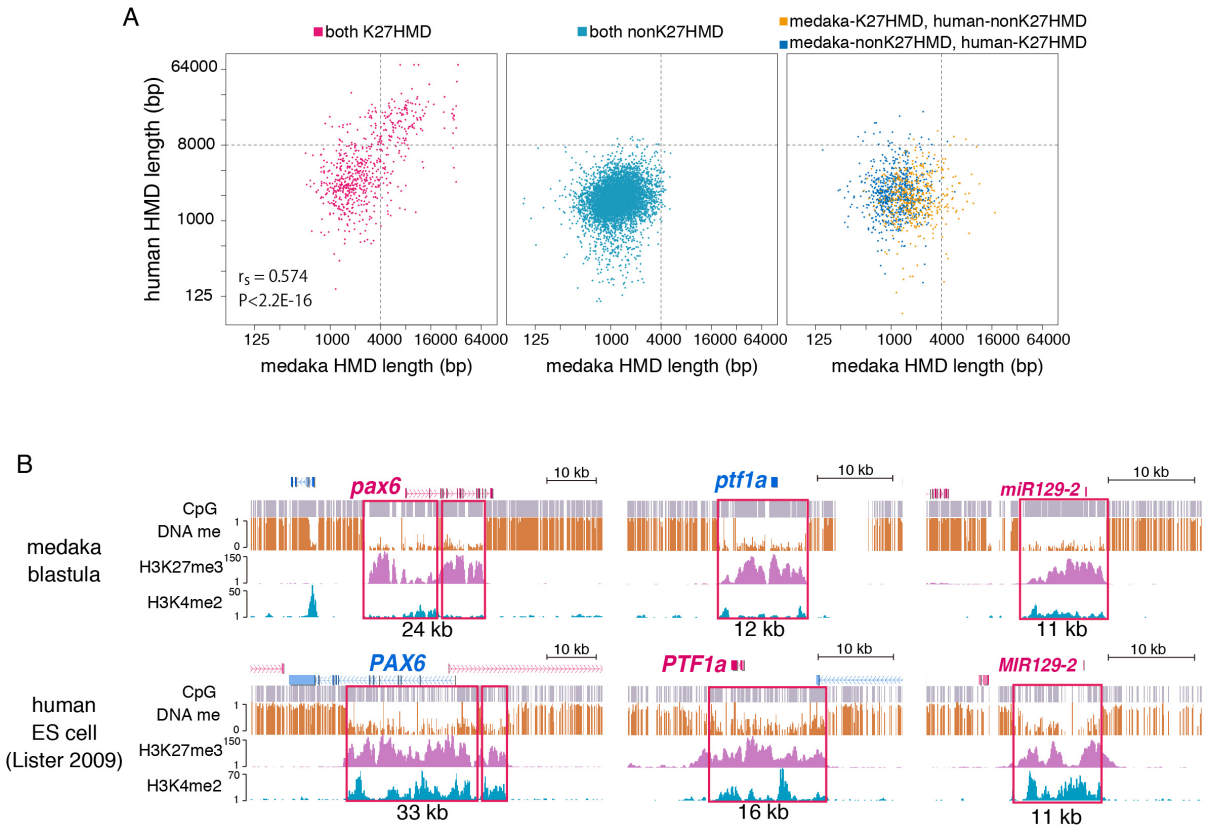

Figure 16. Comparison of HMD size between medaka blastula embryos and human ES cells

(A) Comparison of the size of HMDs associating with gene promoters for medaka gene and its human ortholog. The HMD size correlation is shown in each panel for genes marked by the same type of HMDs in medaka and human (left: K27HMD, middle: nonK27HMD) and different types of HMDs (right). Spearman's rank correlation coefficient (r_s) and p value are shown for genes with conserved K27HMDs (left panel).

(B) Genome browser representations of DNA methylation, H3K27me3 and H3K4me2 in medaka (top) and human (bottom) are shown for conserved large K27HMDs (*pax6*, *ptf1a* and *miR129-2*).

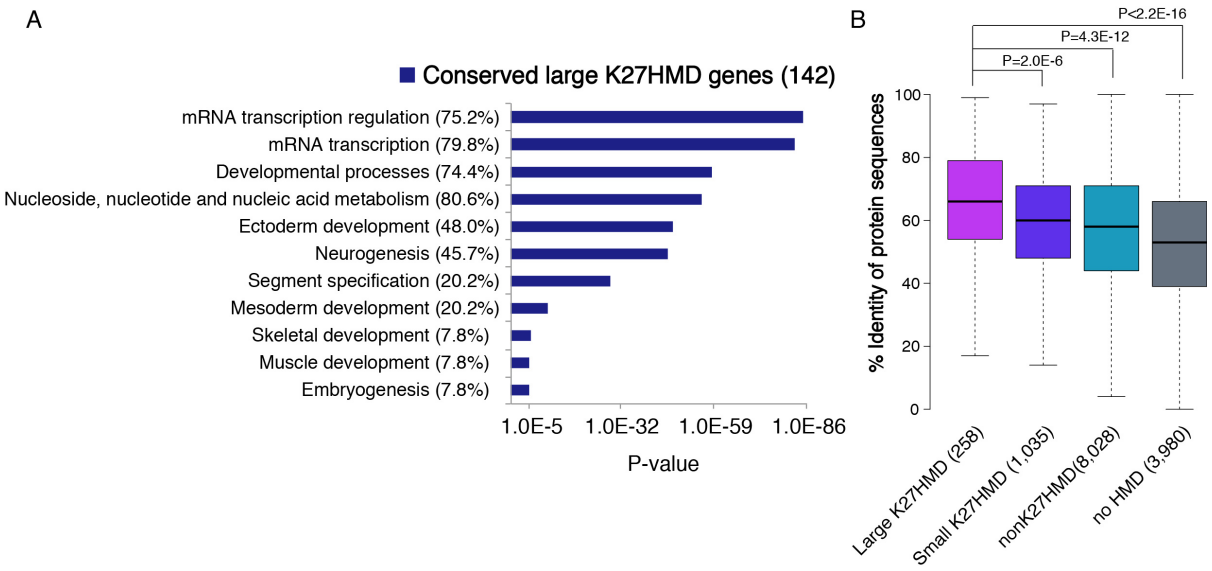

Figure 17. Large K27HMD genes are conserved between medaka and human

(A) Functional annotation of genes marked by conserved large (> 4 kb for medaka and > 8 kb for human) K27HMDs. The over-represented GO PANTHER biological process terms and fraction of genes associated with each GO term are shown. The values of x axes (in logarithmic scale) correspond to the P-values calculated by the DAVID tool.

(B) Boxplots show percent identity of medaka protein sequences to human orthologues. P-values were calculated using non-paired Wilcoxon tests.

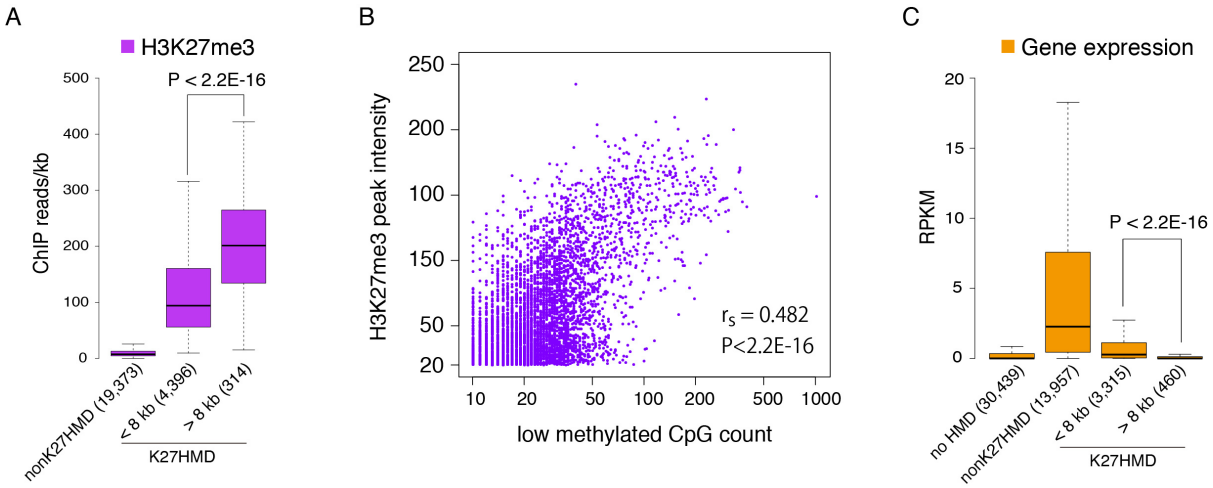

Figure 18. Large K27HMD genes are strongly repressed in human ES cells

(A) Boxplots show H3K27me3 ChIP reads per kb in human ES cells for each HMD category. The number of HMDs are shown under the boxes. P-values were calculated using non-paired Wilcoxon tests.

(B) Comparison of the number of low methylated CpG sites and the highest H3K27me3 ChIP peak intensity inside K27HMD. Spearman's rank correlation coefficient (r_s) and p value are shown.

(C) Boxplots show gene expression levels in human ES cells for each HMD category. The number of genes linked to each HMD category are shown under the boxes. P-values were calculated using non-paired Wilcoxon tests.

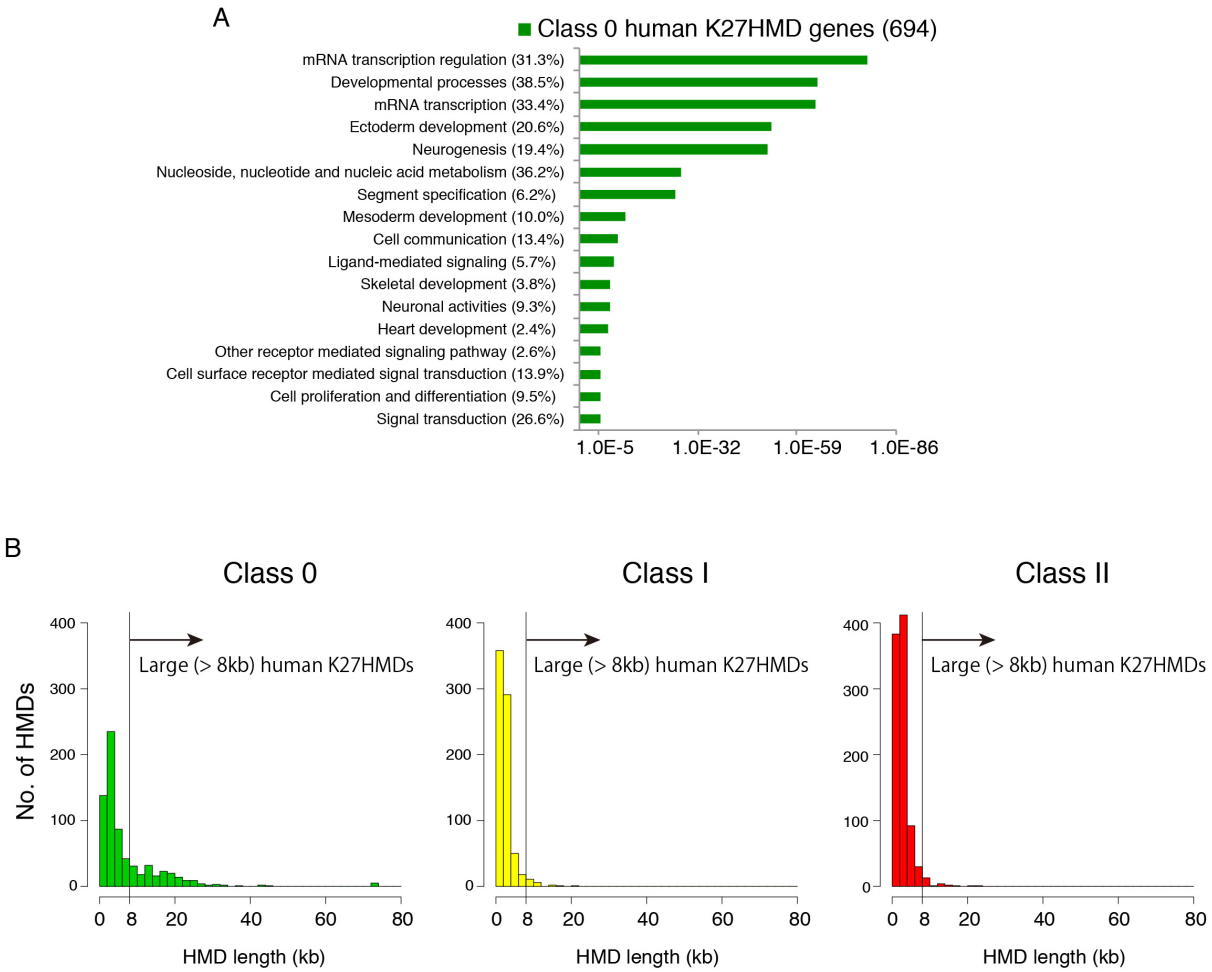

Figure 19. Large K27HMD genes are strongly repressed in human ES cells

(A) Functional annotation of genes marked by Class 0 (K27HMD also in medaka) human K27HMDs. The top over-represented GO PANTHER biological process terms and fraction of genes associated with each GO term are shown. The values of x axes (in logarithmic scale) correspond to the P-values calculated by the DAVID tool.

(B) Size distributions of Class 0 (K27HMD in medaka; green), Class I (nonK27HMD in medaka; orange) and Class II (methylated in medaka; red) human K27HMDs in human ESCs.

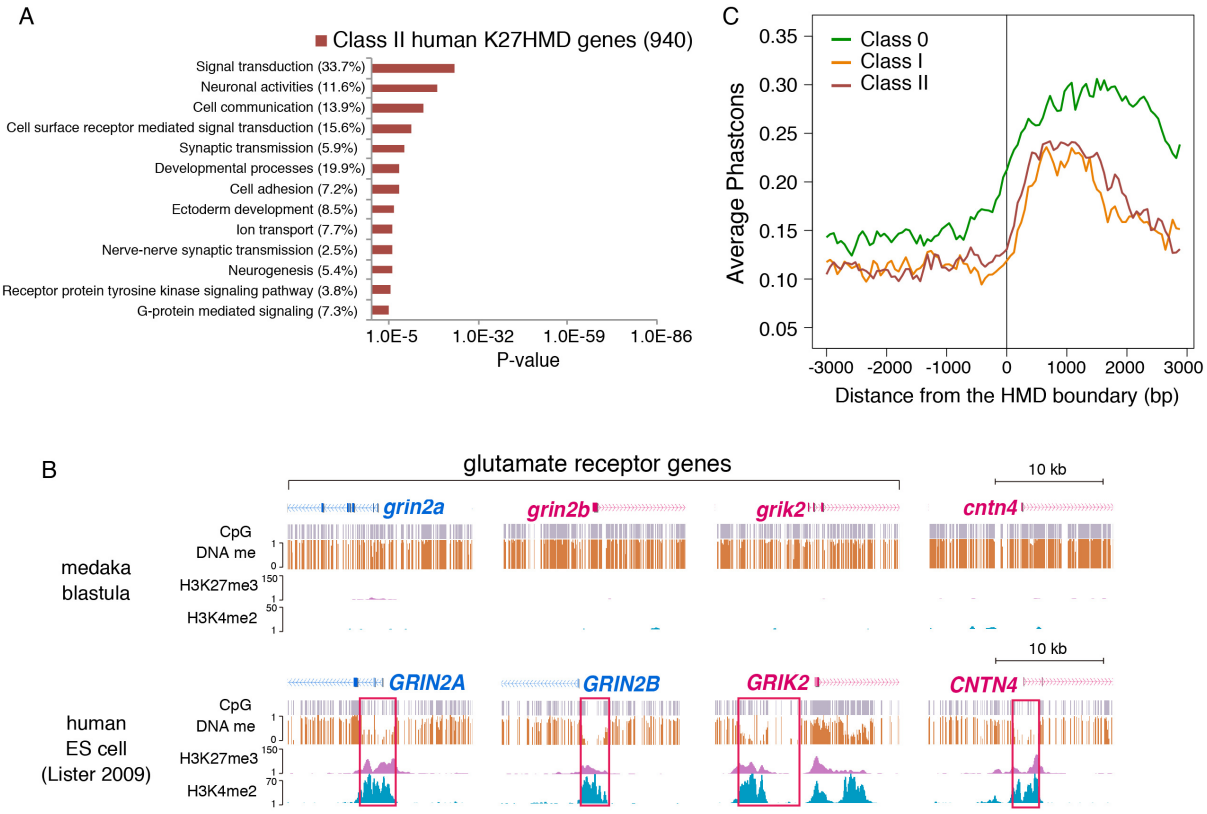

Figure 20. Large K27HMD genes are strongly repressed in human ES cells

(A) Functional annotation of genes marked by Class II (methylated in medaka) human K27HMDs. The top over-represented GO PANTHER biological process terms and fraction of genes associated with each GO term are shown. The values of x axes (in logarithmic scale) correspond to the P-values calculated by the DAVID tool

(B) Genome browser representations of DNA methylation, H3K27me3 and H3K4me2 in medaka (top) and human (bottom) are shown for Class II genes (*grin2a*, *grin2b*, *grik2*, and *cntn4*).

(C) Average vertebrate PhastCons profiles around the boundaries of Class 0 (K27HMD in medaka; green), Class I (nonK27HMD in medaka; orange) and Class II (methylated in medaka; red) human K27HMDs.

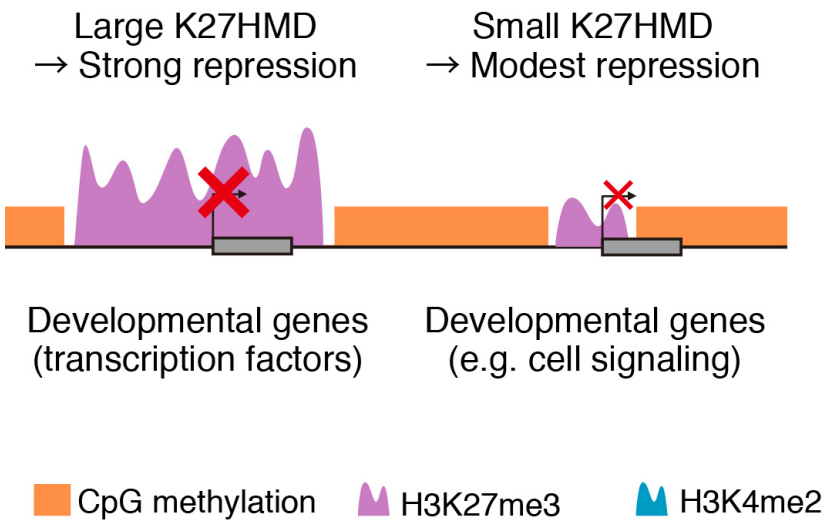

Figure 21. Large K27HMD provides strong repression to key developmental genes

A model for the HMD size dependent repression of developmental genes in pluripotent cells.

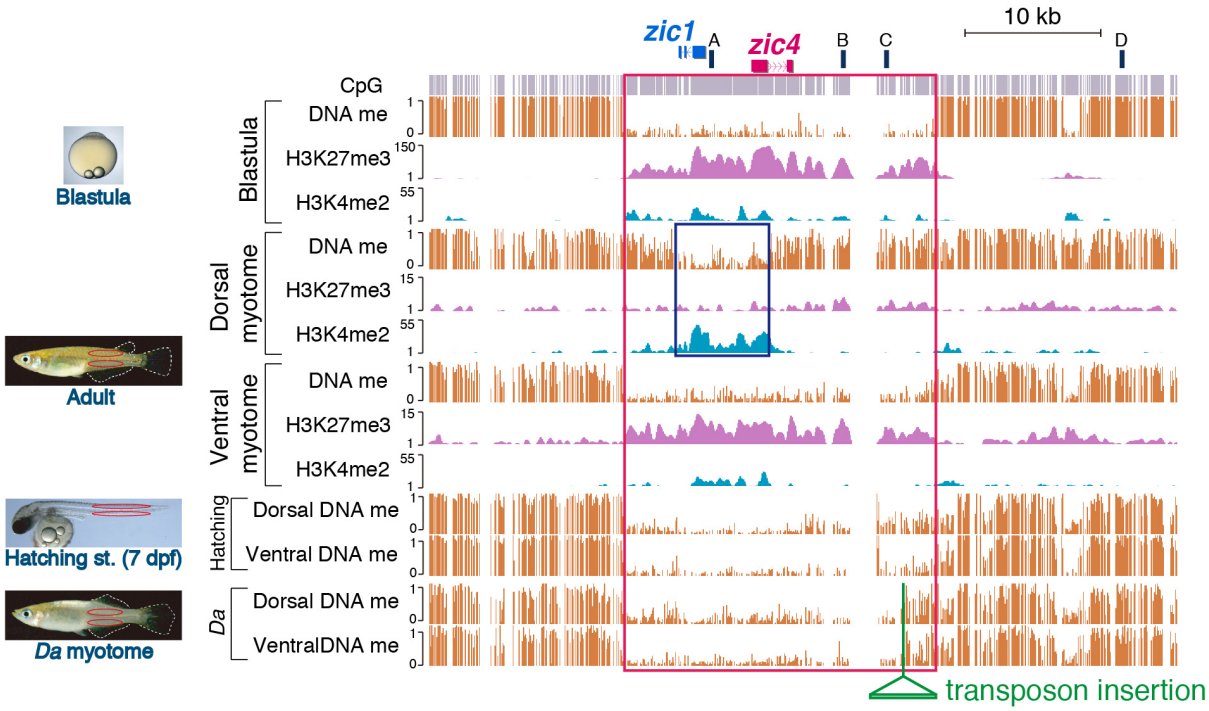

Figure 22. Large K27HMD at *zic1*/*zic4* locus undergoes shortening in adult dorsal myotome

(A) Genome browser representation at the *zic1*/*zic4* loci. DNA methylation, H3K27me3 and H3K4me2 are shown for blastula embryos, adult dorsal and ventral myotome, and DNA methylation for dorsal and ventral myotome of hatching stage larvae and adult *Da* mutant. Red box indicates large K27HMD identified in the blastula embryo and blue box indicates shortend HMD in adult dorsal myotome. Navy bars indicate regions examined for ChIP-qPCR and bisulfite sequencing in Figure 23 and 24.

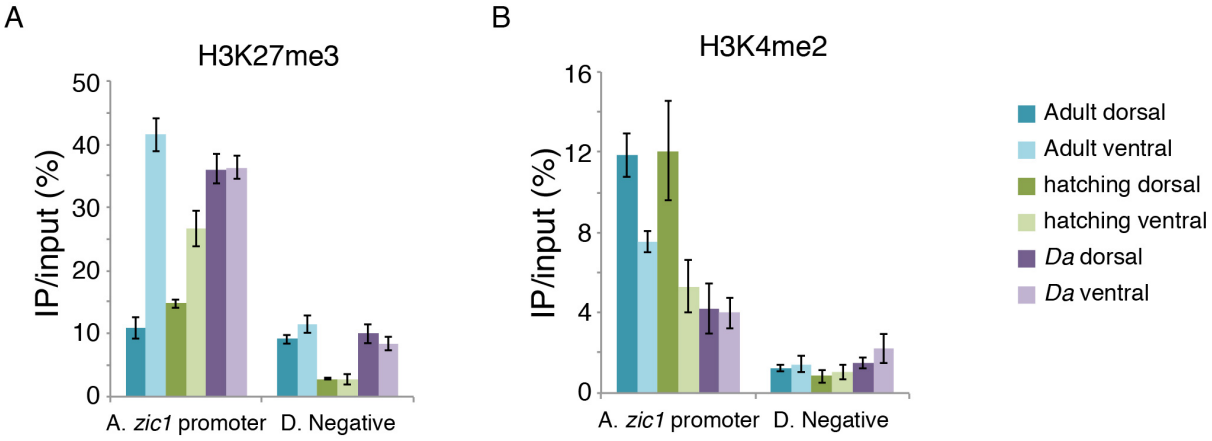

Figure 23. ChIP-qPCR analysis of dorsal and ventral myotome

ChIP-qPCR analysis of dorsal and ventral myotome from *wt* adult, hatching stage larvae, and *Da* adult for H3K27me3 (A) and H3K4me2 (B) at regions indicated in Figure 22. Error bars represent s.d. from three technical replicates.

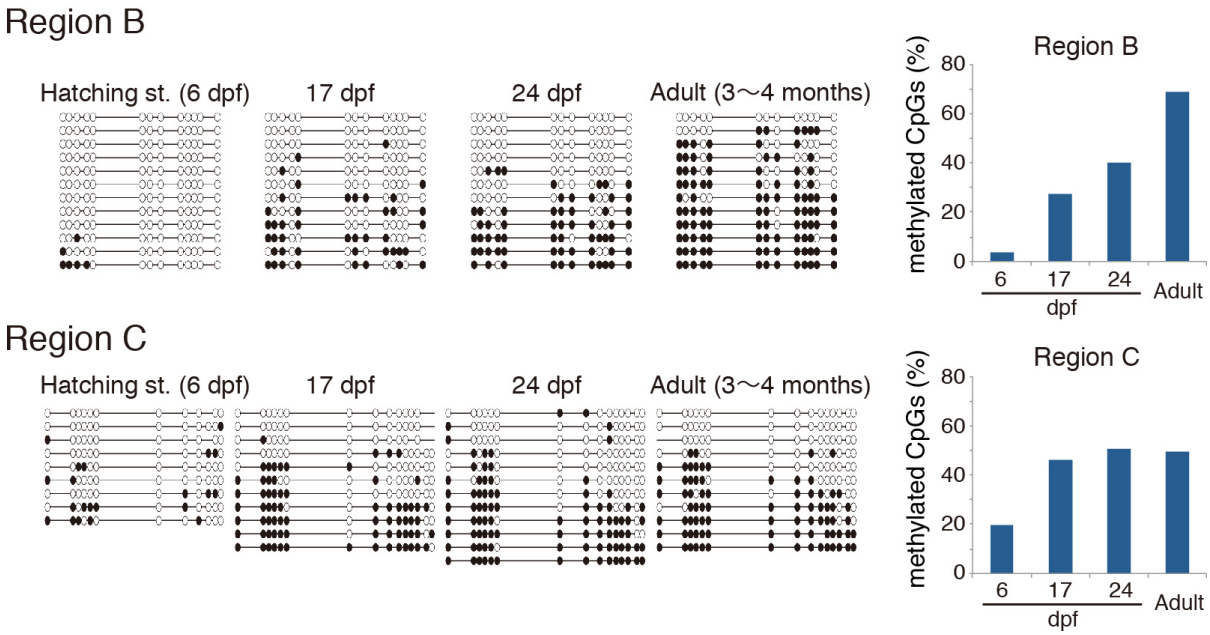

Figure 24. HMD shortening gradually occur after hatching

Bisulfite sequencing in dorsal myotome at multiple stages after hatching. Graphs indicate the ratio of methylated CpGs at each stage. The examined regions are indicated in Figure 22.

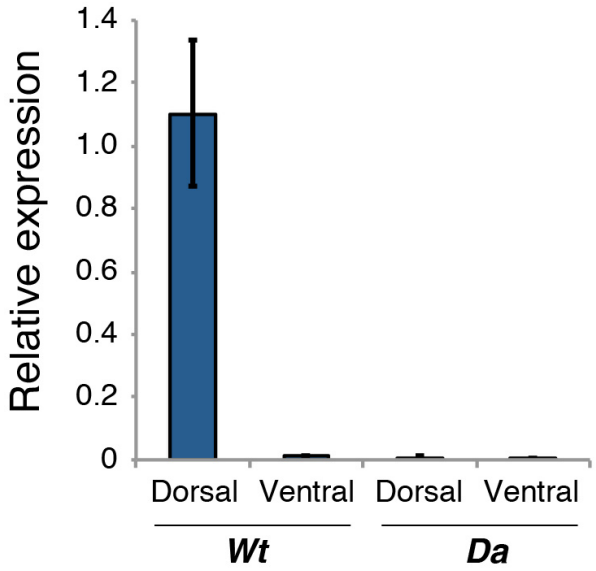

Figure 25. *zic1*/*zic4* expression is reduced in *Da*

Relative expression of *zic1* measured by RT-qPCR in dorsal and ventral myotome from *Wt* and *Da*. Expression is normalized to that of the housekeeping gene *ef1a*. Error bars represent s.d. from three biological replicates.

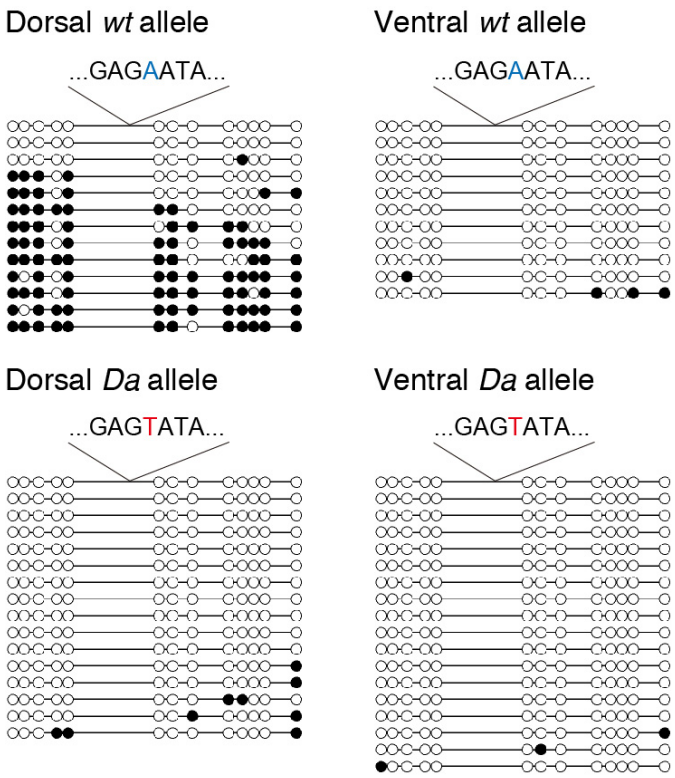

Figure 26. Only *wt* allele acquire hypomethylation in *Da* heterozygous fish

Bisulfite sequencing in adult *Da* heterozygous fish at region C in Figure 22. Even in the same nucleus, only *wt* allele shows DNA hypermethylation.

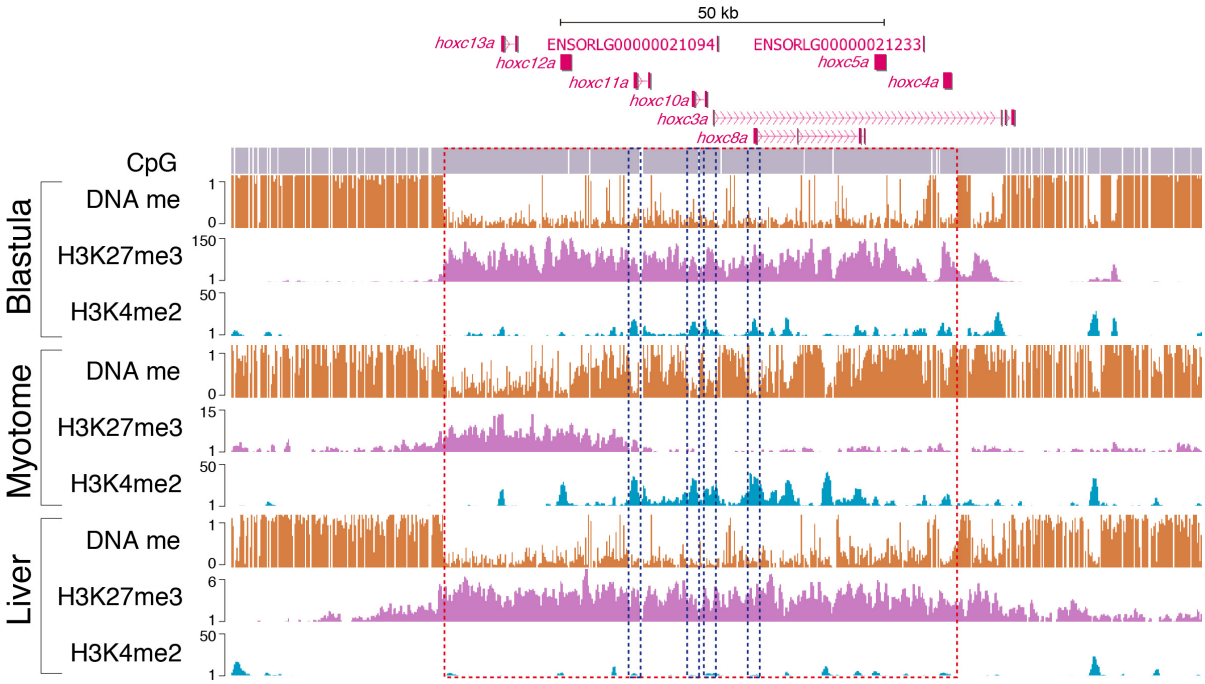

Figure 27. Large K27HMD at *hoxc* cluster undergoes shortening in adult myotome

Genome browser representation of DNA methylation, H3K27me3 and H3K4me2 enrichment in the blastula embryo, adult myotome and liver are shown for large K27HMD covering *HoxC* cluster. Red dashed box indicates large K27HMD regions. Note that the methylation levels of promoter regions of *hoxc11a*, *10a*, *3a*, and *8a* remain low in adult myotome (blue dashed boxes).

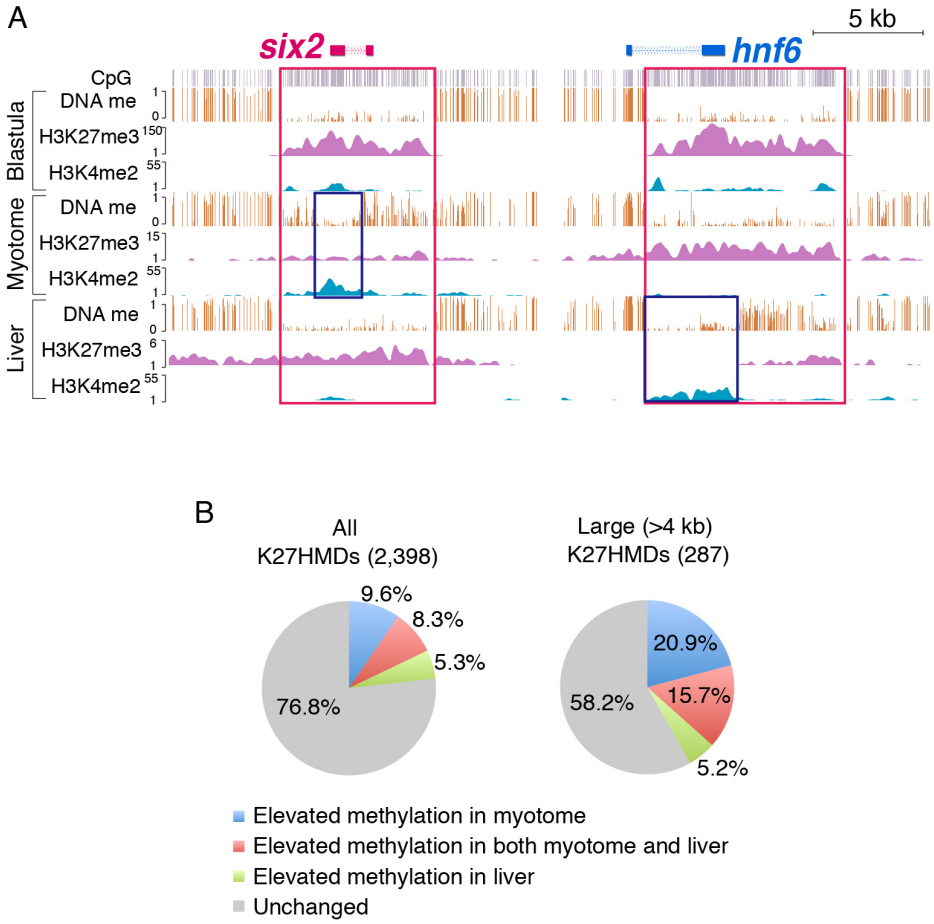

Figure 28. HMD shortening in adult myotome and liver

(A) Genome browser representation of DNA methylation, H3K27me3 and H3K4me2 enrichment are shown for K27HMD with myotome-specific hypermethylation (*six2*; left) and liver-specific hypermethylation (*hnf6*; right).

(B) Fraction of K27HMDs with DNA hypermethylation in adult myotome and liver for all K27HMD (left) and large K27HMD (right).

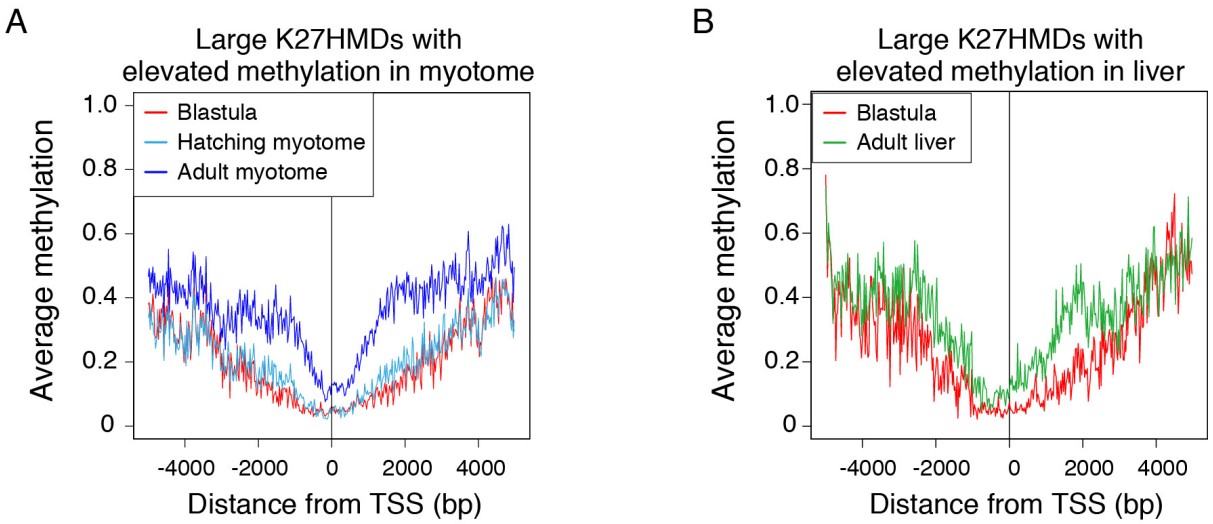

Figure 29. HMD shortening in adult myotome and liver

(A) Average DNA methylation around TSSs marked by large K27HMD with elevated DNA methylation in myotome at multiple stages.

(B) Average DNA methylation around TSSs marked by large K27HMD with elevated DNA methylation in liver.

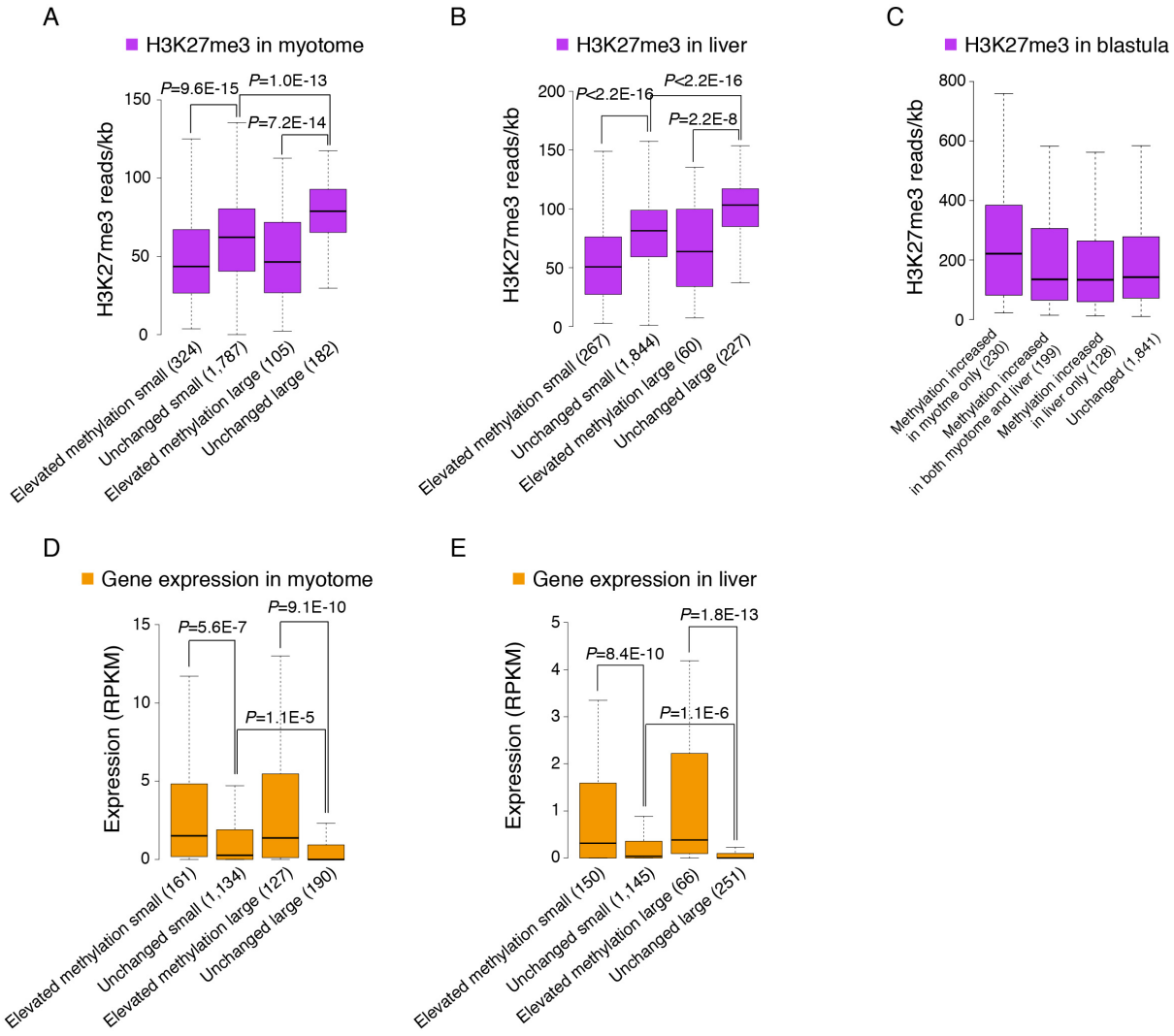

Figure 30. HMD shortening in adult myotome and liver

(A - C) Boxplots show H3K27me3 enrichment at hypermethylated (elevated methylation) and unchanged K27HMDs in adult myotome (A), liver (B) and blastula (C). The number of HMDs are shown under the boxes. P-values were calculated using non-paired Wilcoxon tests. (D and E) Boxplots show gene expression at hypermethylated (elevated methylation) and unchanged K27HMDs in adult myotome (C) and liver (D). The number of genes linked to each HMD category are shown under the boxes. P-values were calculated using non-paired Wilcoxon tests.

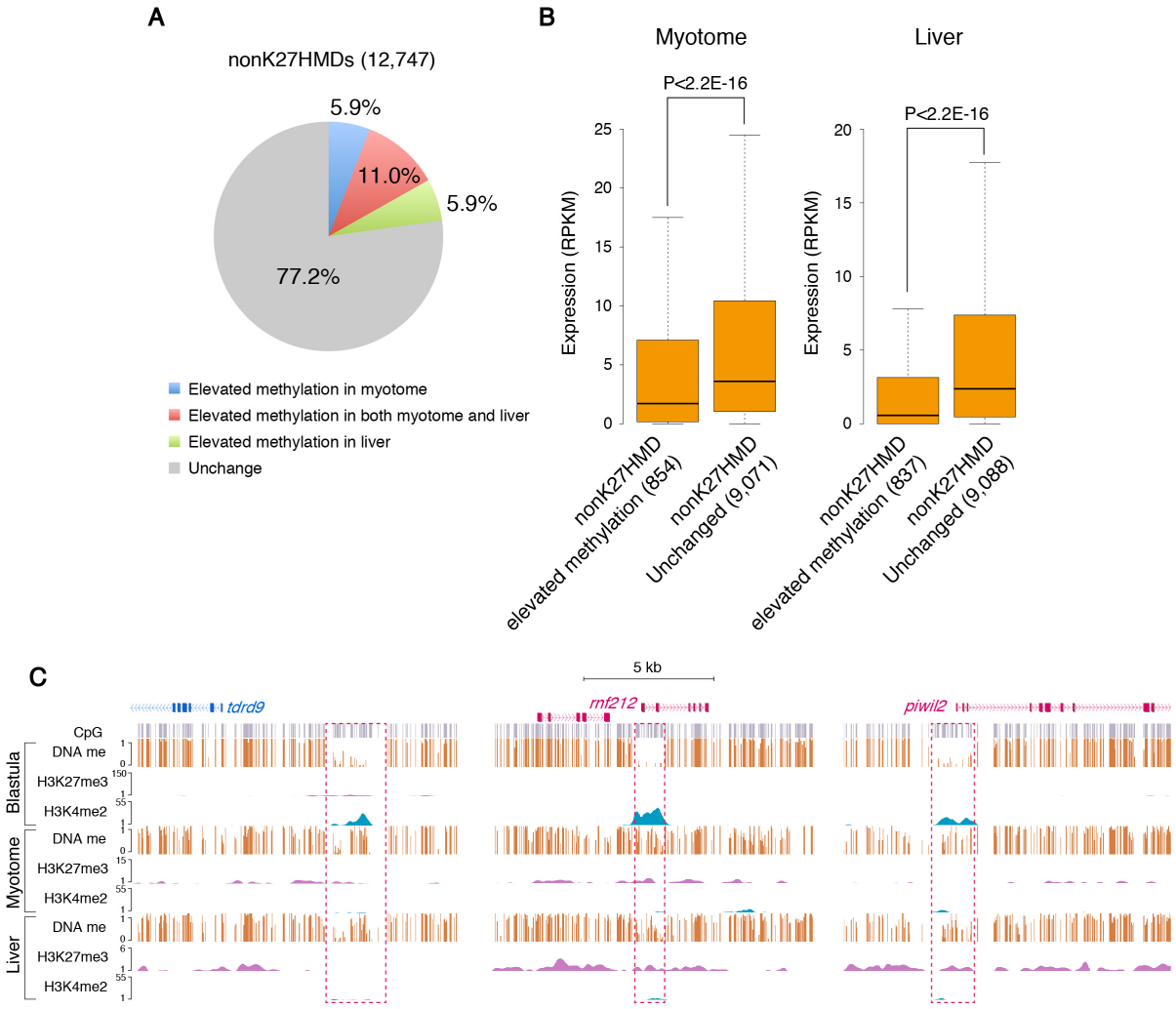

Figure 31. DNA hypermethylation of nonK27HMD associates with gene silencing

(A) Fraction of nonK27HMDs with DNA hypermethylation in adult myotome and liver

(B) Boxplots show gene expression levels (RPKM) at hypermethylated and unchanged nonK27HMDs for adult myotome and liver. P-values were calculated using non-paired Wilcoxon tests.

(C) Genome browser representation of DNA methylation, H3K27me3 and H3K4me2 enrichment in the blastula embryo, adult myotome and liver are shown for nonK27HMD with hypermethylation (for example, *tdrd9*; left, *rnf212*; middle, *piwil2*; right).

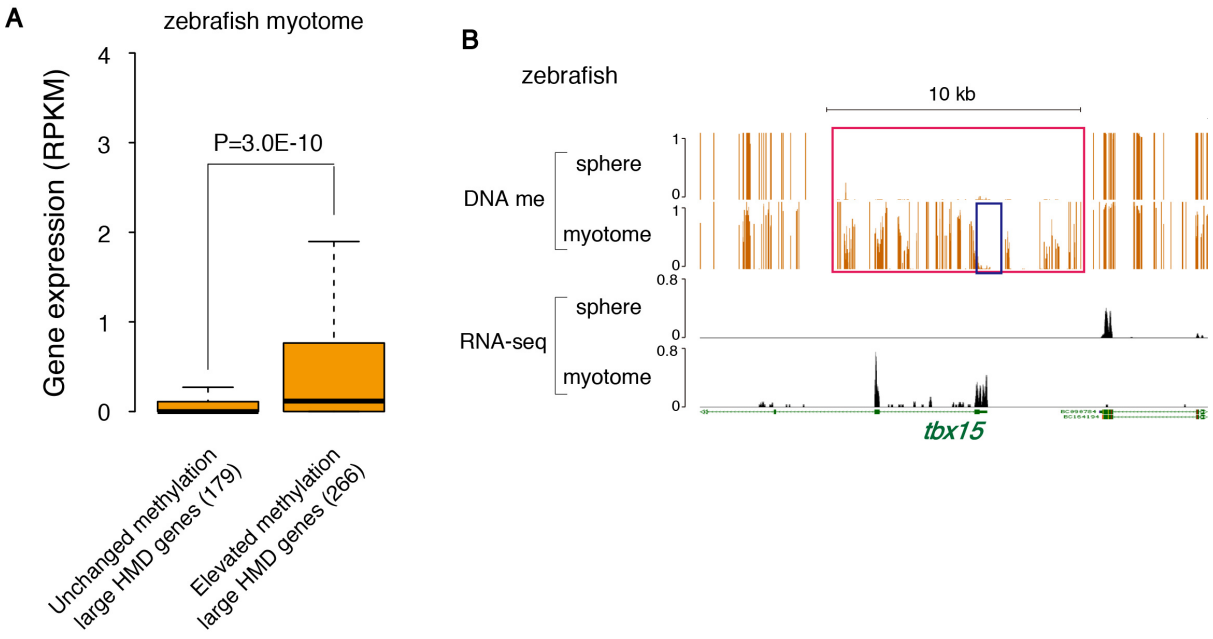

Figure 32. HMD shortening associates with active gene expression also in zebrafish

(A) Boxplots show gene expression levels (RPKM) of genes associate to large (>8 kb) HMDs with unchanged methylation and elevated methylation in adult myotome. P-values were calculated using non-paired Wilcoxon tests.

(B) Genome browser representation of DNA methylation and gene expression in the zebrafish sphere stage embryo and adult myotome are shown for large HMD which undergo shortening (for example, *tbx15*). Relative read coverage normalized by total mapped reads are shown for RNA-seq tracks. Red and blue boxes represent a large HMD in sphere and a shortened HMD in myotome, respectively.

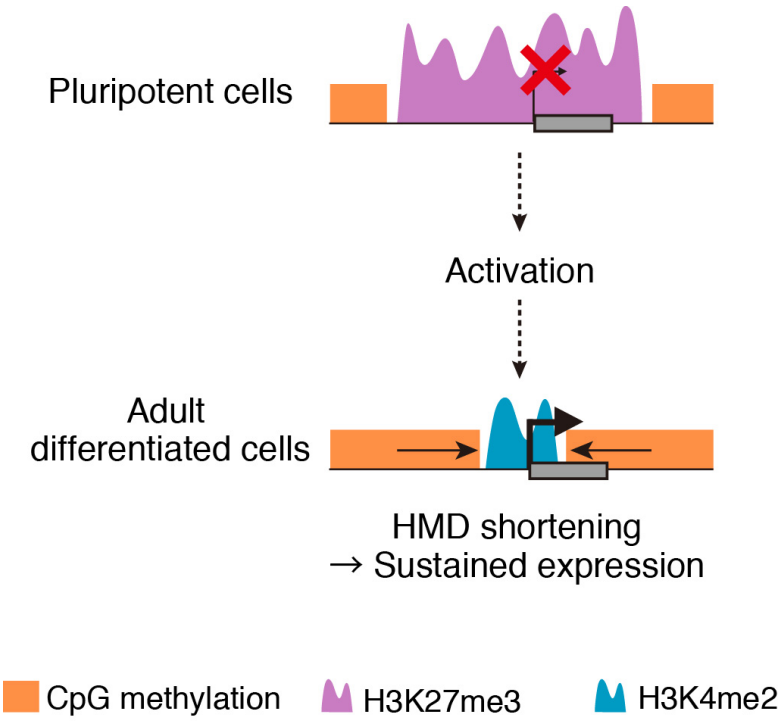

Figure 33. HMD shortening consolidates developmental gene expressions

A model for the HMD shortening consolidating long-term gene expressions in adult differentiated cells.

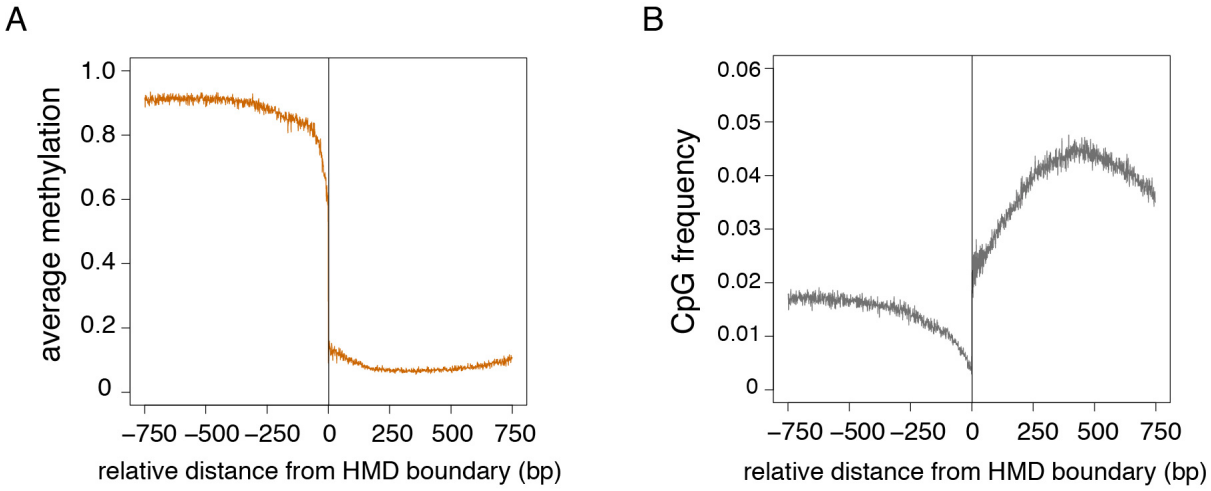

Figure 34. HMDs have sharp boundaries and CpG density greatly changes at the boundaries

(A) Average methylation ratio around all HMD boundaries. Position -1 and 1 were excluded as there is no CpG from the definition of the boundary (Position 0 is always cytosine of CpG).

(B) Frequency of CpG around all HMD boundaries. Position -1, 0, and 1 were excluded.

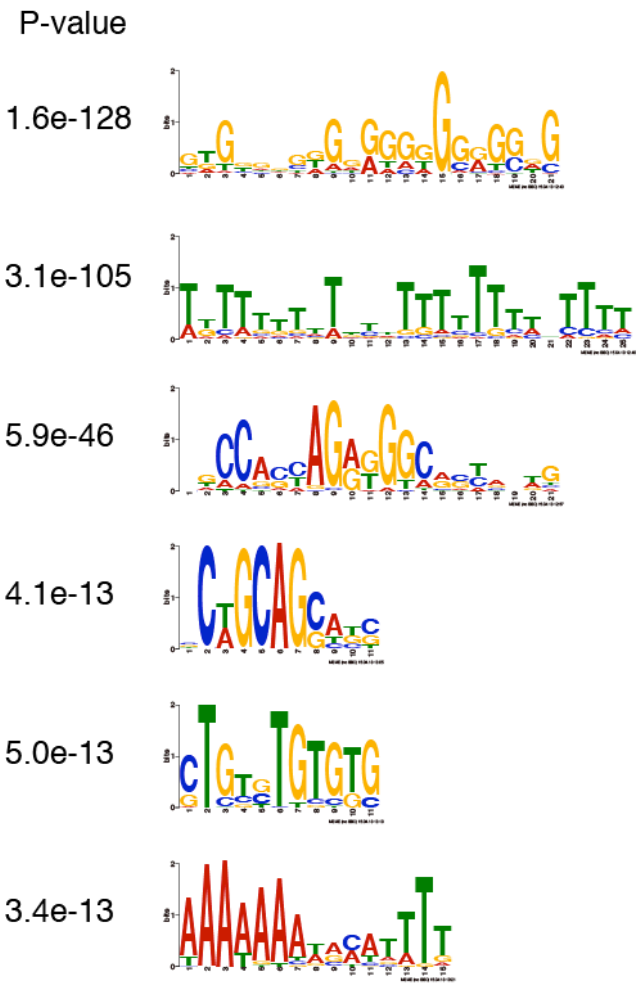

Figure 35. Specific sequence feature at HMD boundary

The DNA motifs significantly enriched at HMD boundaries. P-value was determined by the MEME program.

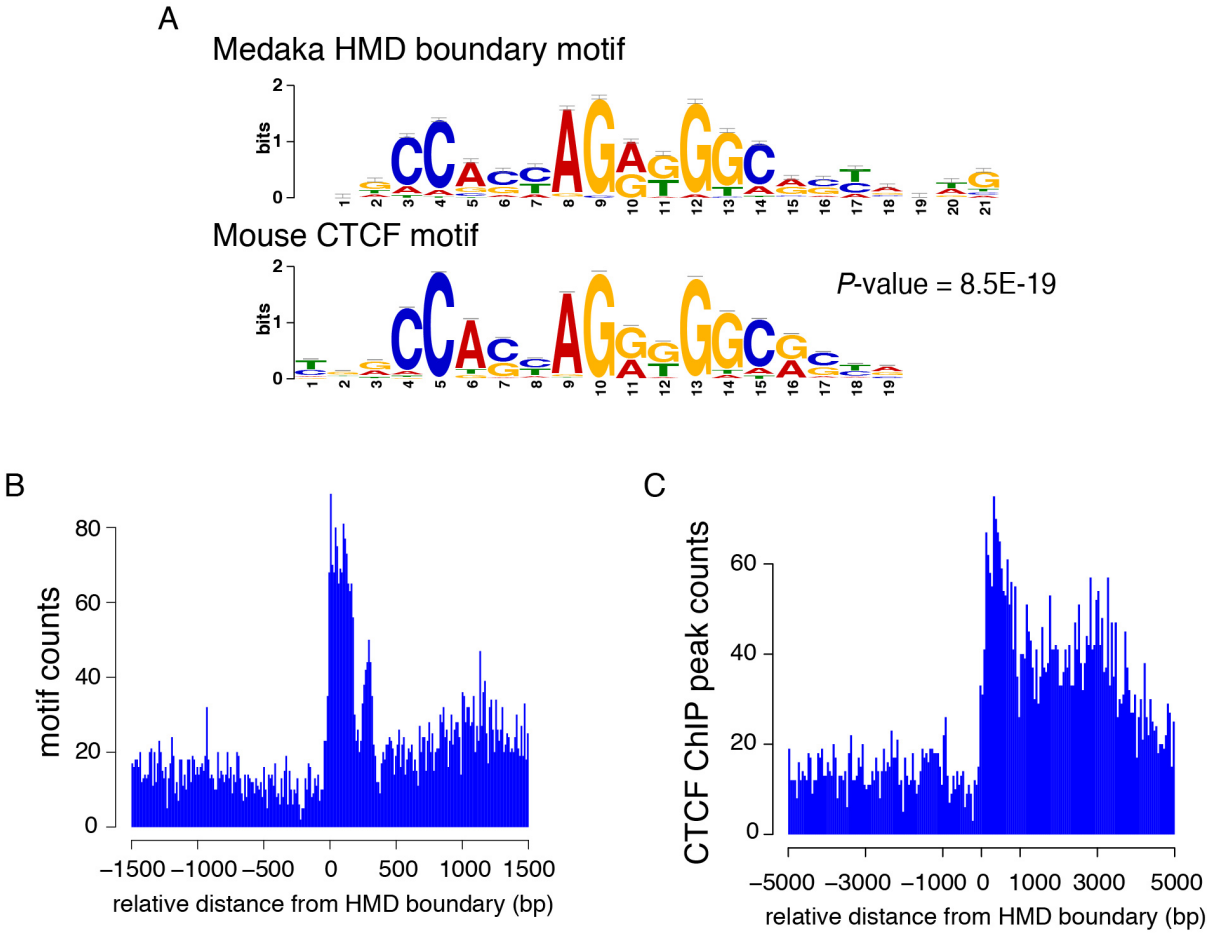

Figure 36. CTCF binding sites are enriched at HMD boundaries

(A) The DNA motif enriched at HMD boundaries (top) and Mouse CTCF motif (bottom). P-value was determined by the TOMTOM program(Gupta et al., 2007).

(B) Distribution of the CTCF motif around boundaries of HMDs larger than 1 kb in medaka. Number of CTCF motif centers in 10 bp window were counted.

(C) Distribution of the CTCF ChIP-seq peaks around boundaries of HMDs larger than 3 kb in human ES cells. The number of CTCF ChIP peak centers in 50 bp window were counted.

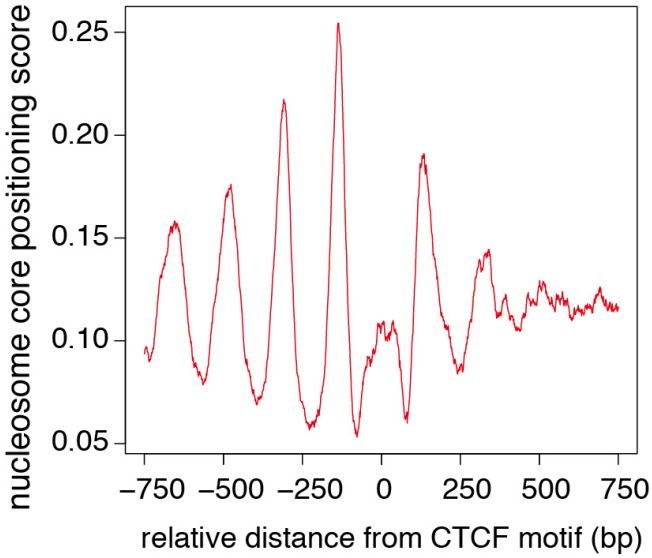

Figure 37. Nucleosome positioning around CTCF motifs at HMD boundaries

The average local dyad positioning score around boundary-associated CTCF motifs. Downstream regions are hypomethylated.

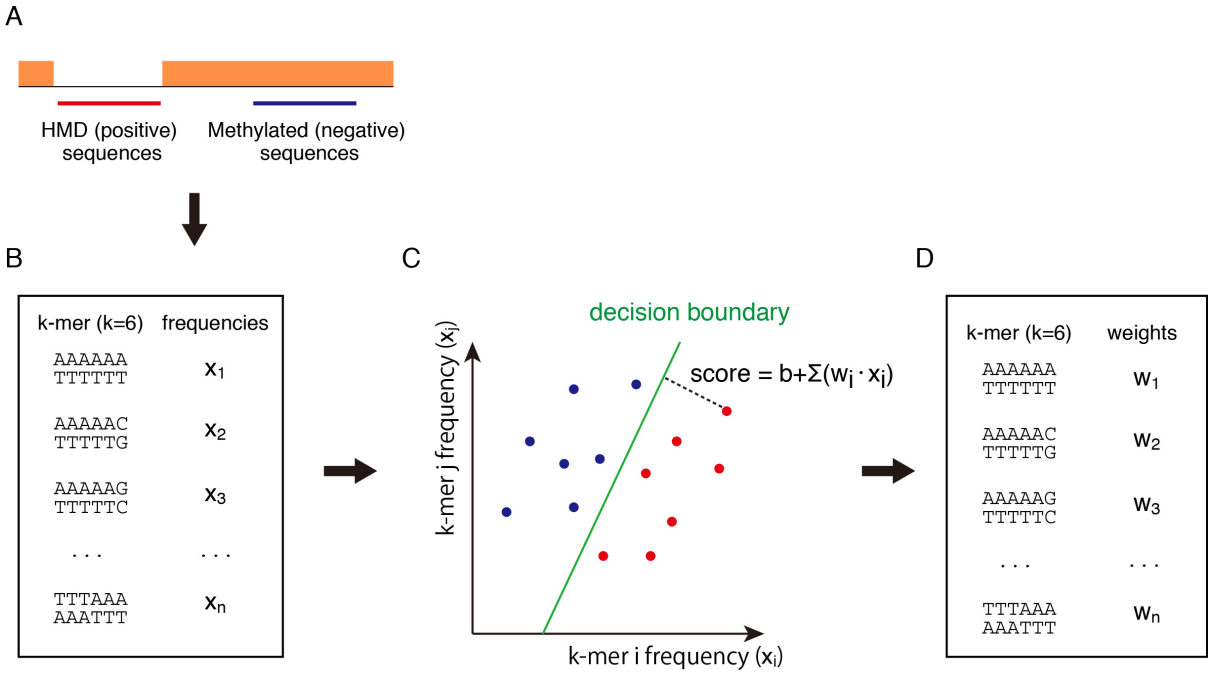

Figure 38. Schematic of SVM algorithm

Overview of the method is shown.

(A) For SVM_{DNA}, HMD sequences were used as positive data, and sequences from randomly selected methylated regions were used as negative data.

(B) SVM calculates the frequencies of all *k*-mers for each of the positive and negative sequences.

(C) Feature vectors (x_1, …, x_n) calculated in (B) are used to find a decision boundary, which most accurately separate the positive (red) and negative (blue) training data sets.

(D) Weights, w_i, are obtained from the decision boundary, and are used to predict the epigenetic modification states of novel sequences.

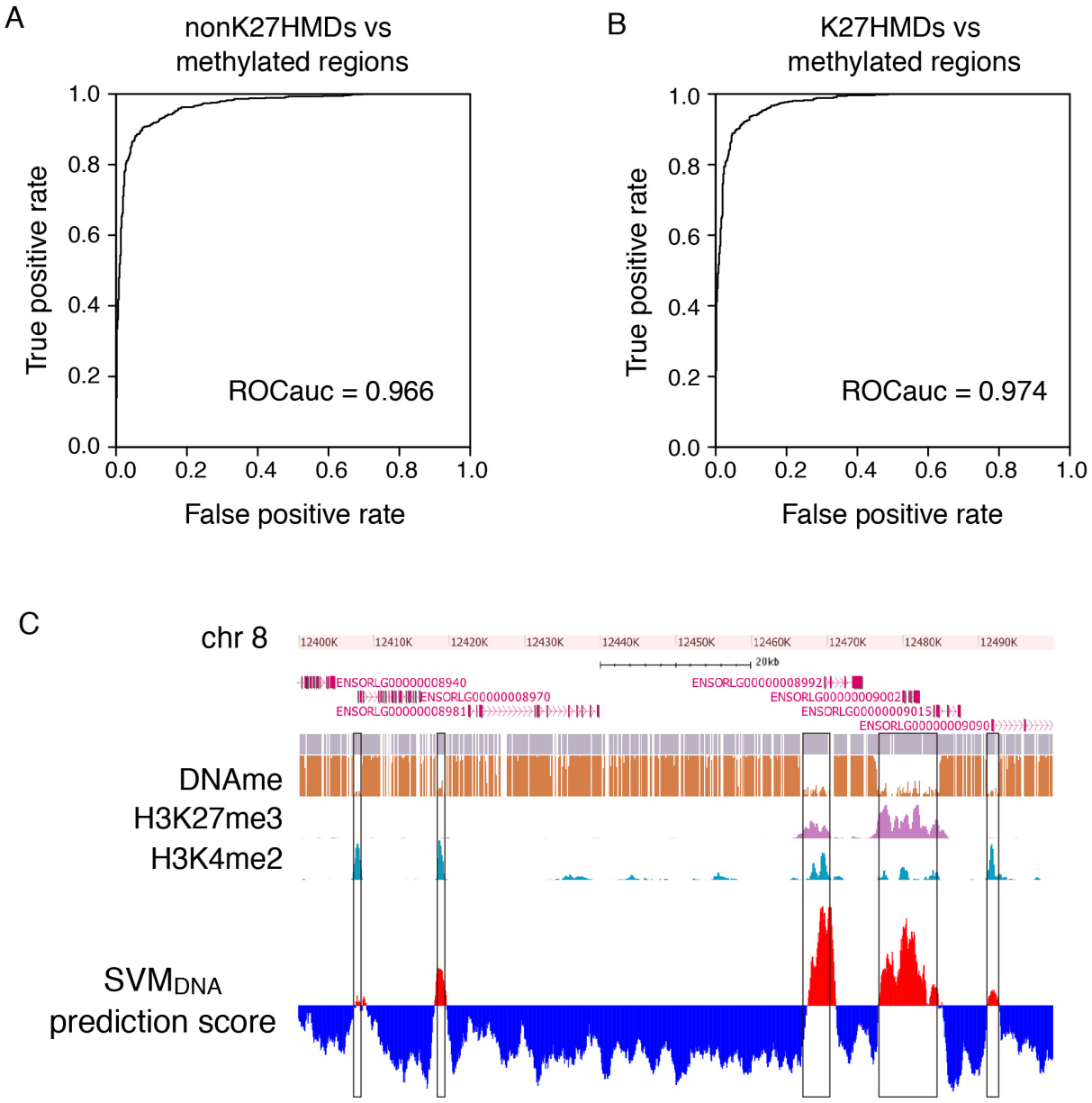

Figure 39. SVM can distinguish hypomethylated sequences and methylated sequences

(A) Classification of nonK27HMD vs. methylated sequences on chromosome8 by SVM_{DNA}.

(B) Classification of all K27HMD vs. same number of methylated sequences by SVM_{DNA}.

(C) Example of HMD prediction on chromosome 8 by SVM_{DNA}. The prediction scores are visualized in the bottom track (red is a positive score, and blue is negative score). DNA methylation, H3K27me3, and H3K4me2 in medaka blastula embryo is visualized in the top tracks. Black boxes represent HMDs.

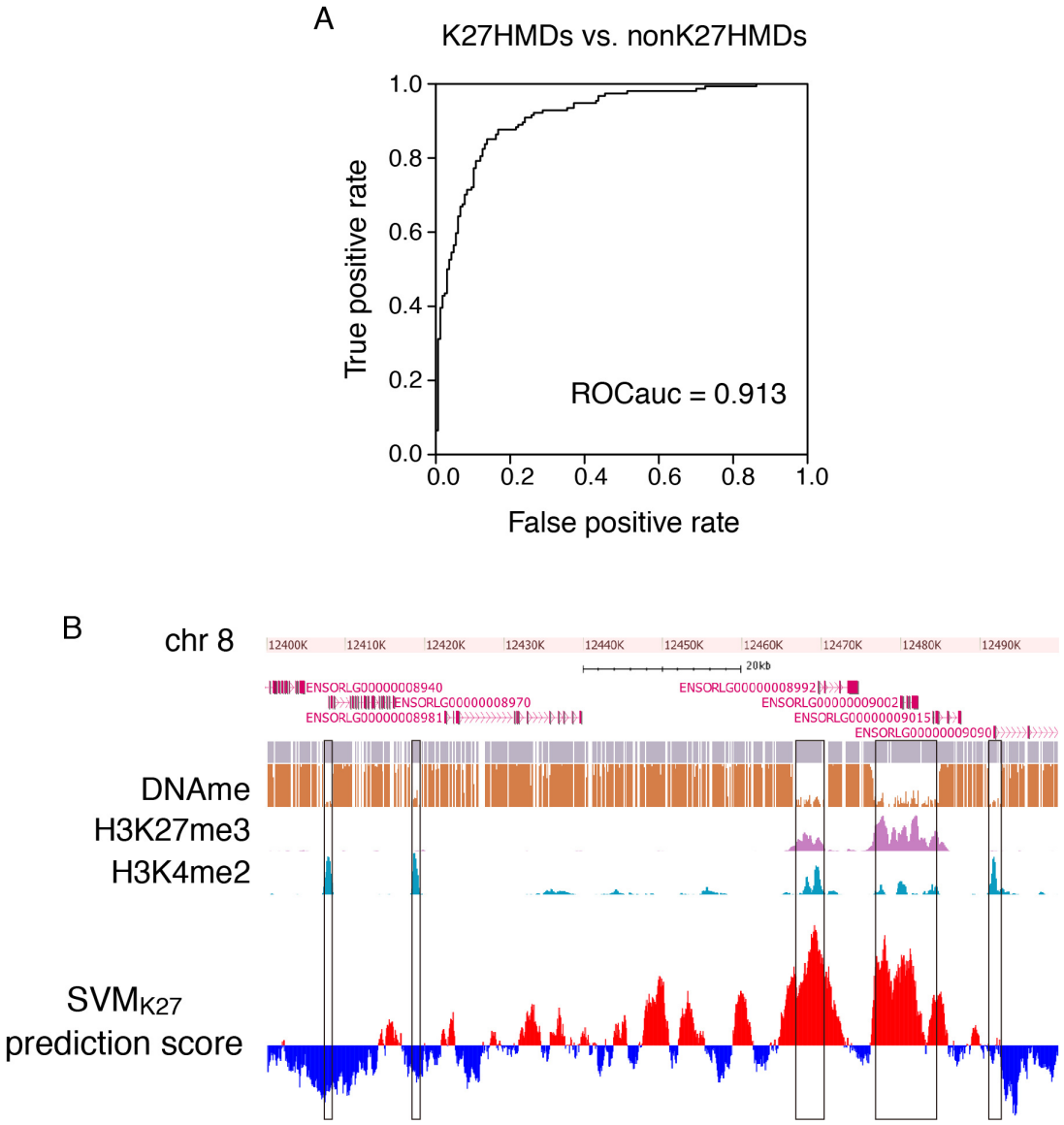

Figure 40. SVM can distinguish K27HMD sequences and nonK27HMD sequences

(A) Classification performance of K27HMD vs. nonK27HMD sequences on chromosome8 by SVM_{K27}.

(B) Example of K27HMD prediction on chromosome 8 by SVM_{K27}. The prediction scores are visualized in the bottom track (red is a positive score, and blue is negative score). DNA methylation, H3K27me3, and H3K4me2 in medaka blastula embryo is visualized in the top tracks. Note that nonK27HMD regions tend to have negative scores. Black boxes represent HMDs

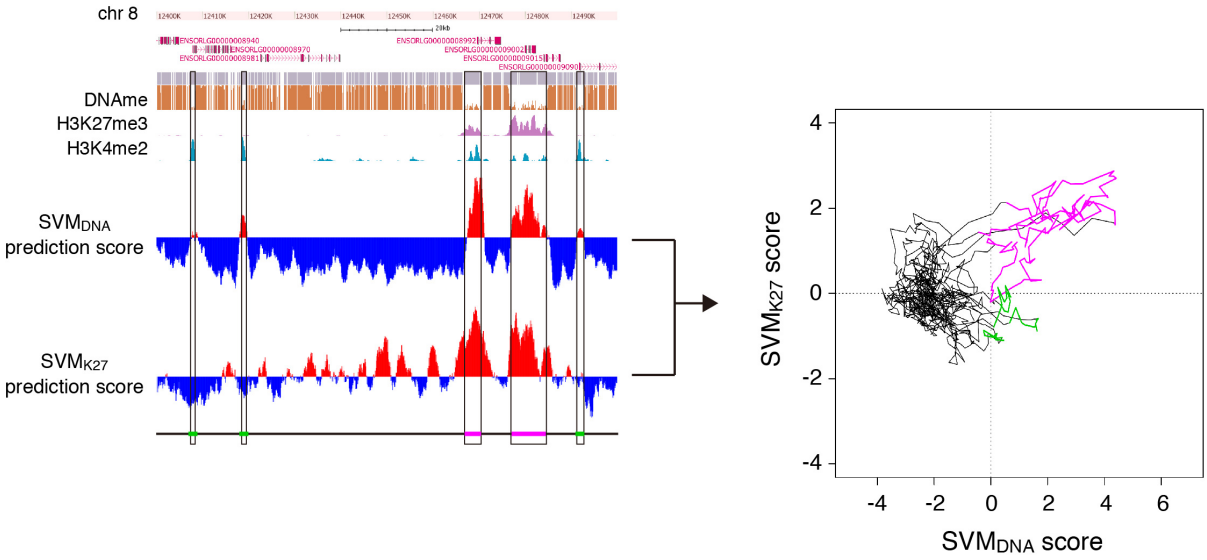

Figure 41. K27HMD, nonK27HMD, and methylated sequences can be predicted by SVM

An example of genomic region from chromosome 8 was plotted by the prediction score calculated by SVM_{DNA} (x-axis) and SVM_{K27} (y-axis). Green regions represent nonK27HMD regions, and magenta regions represent K27HMD regions. By using both SVM_{DNA} and SVM_{K27}, K27HMD and nonK27HMD can be predicted in medaka genome.

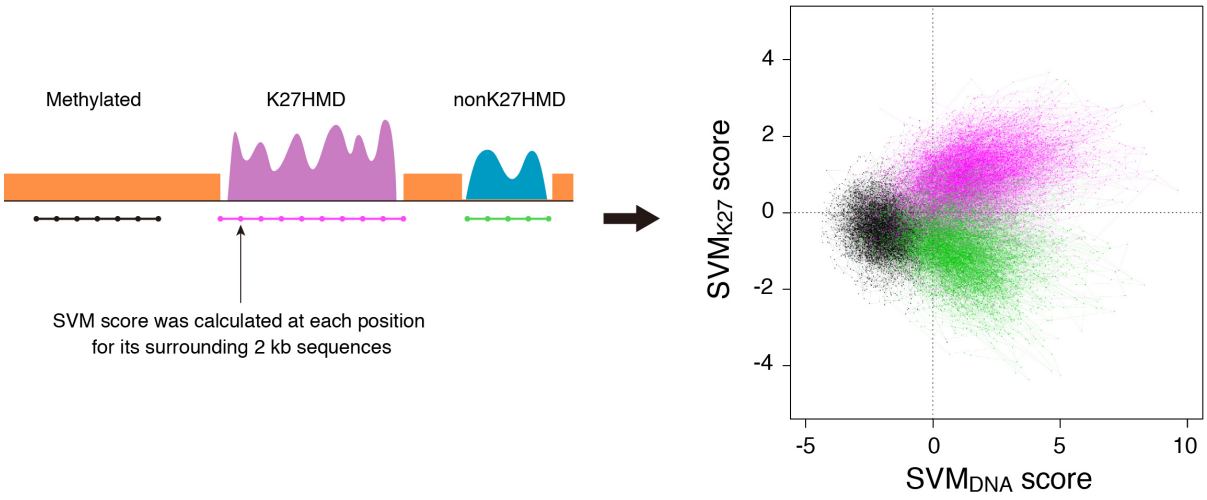

Figure 42. Epigenetic modification state of most of the regions are determined by their neighboring DNA sequences

SVM score for all HMD regions were calculated and plotted on the graph (right). The sequences next to each other on the genome are connected with a line.

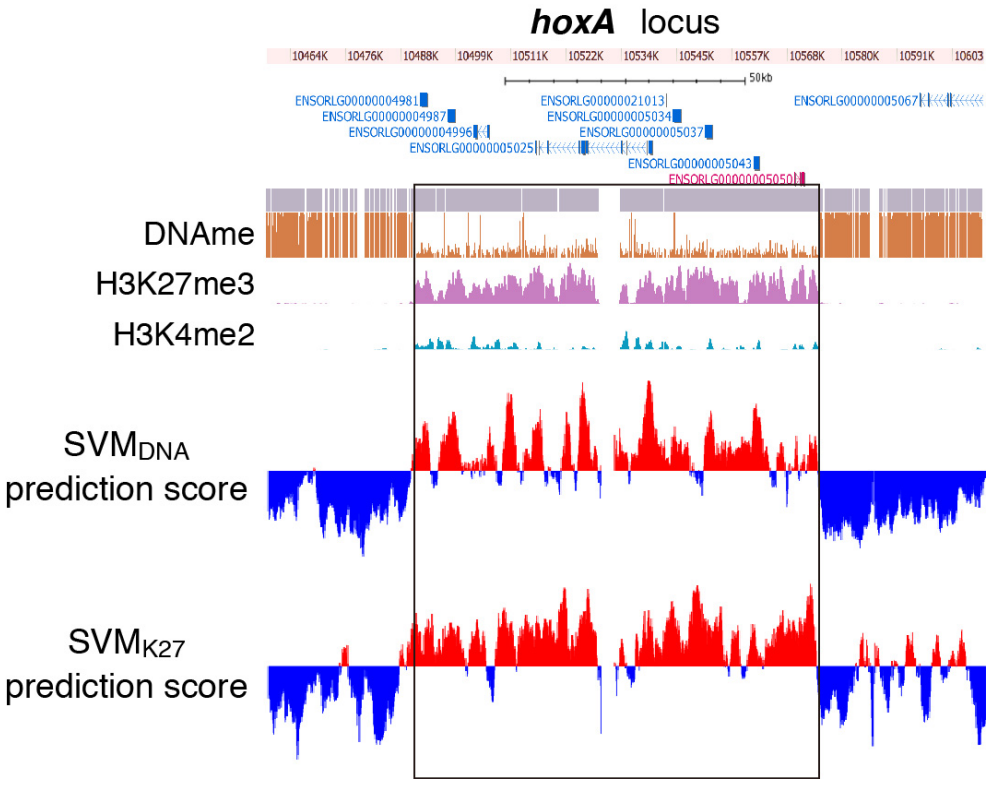

Figure 43. Large K27HMD have specific sequence features along entire domain

An example of SVM prediction of large K27HMD. Most of the genome fragments inside the large K27HMD at *hoxA* locus show high positive scores.

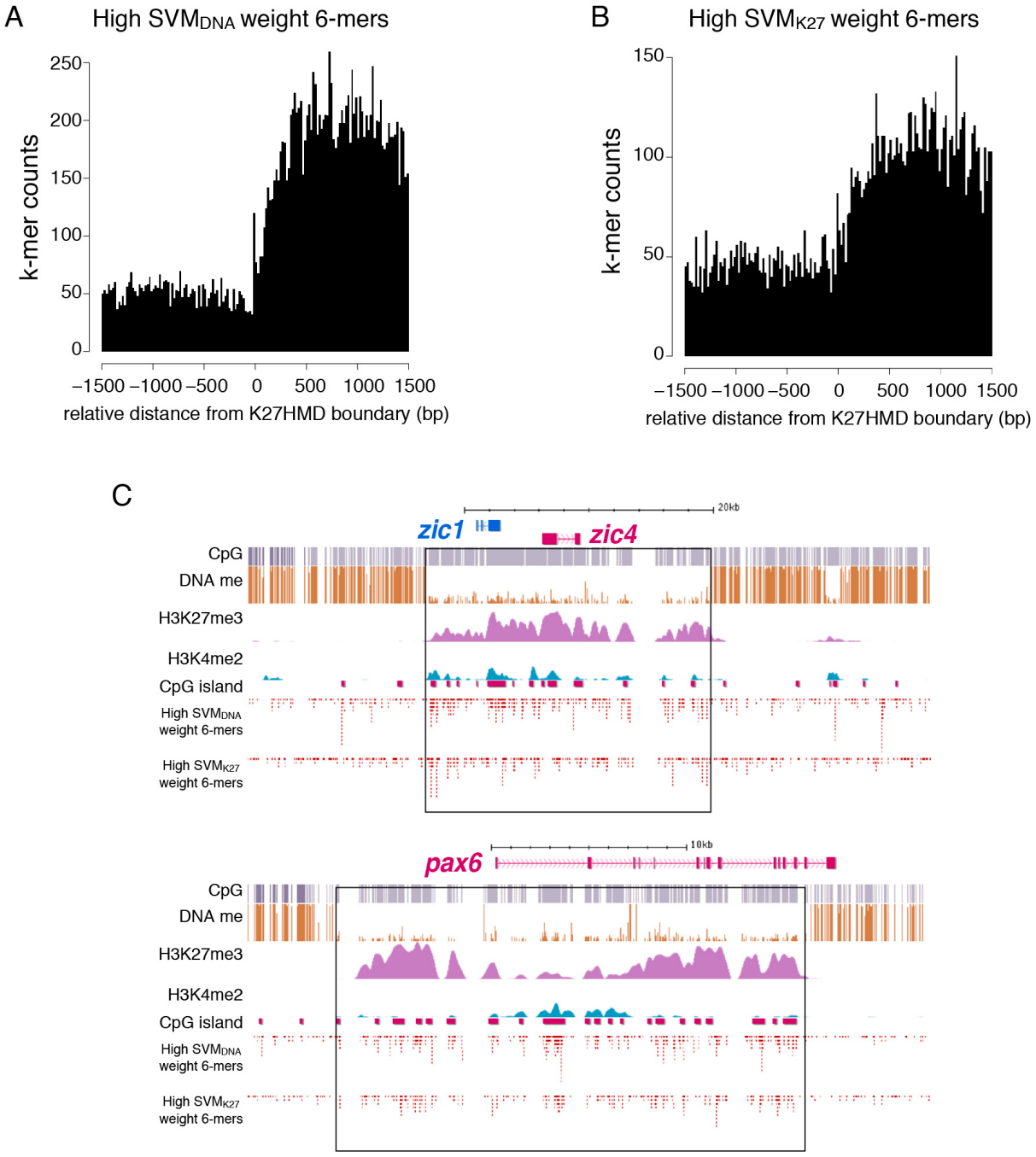

Figure 44. Distribution pattern of 6-mers with the largest positive SVM weights

(A and B) Distribution of 6-mers with large positive weights from SVM_{DNA} (A) and SVM_{K27} (B) relative to K27HMD boundary are shown. Top 50 6-mers and top 20 6-mers are counted for SVM_{DNA} (A) and SVM_{K27} (B), respectively.

(C) Genome browser representation of large K27HMDs (black boxes) and distribution of CpG islands and 6-mers with large positive SVM weight.

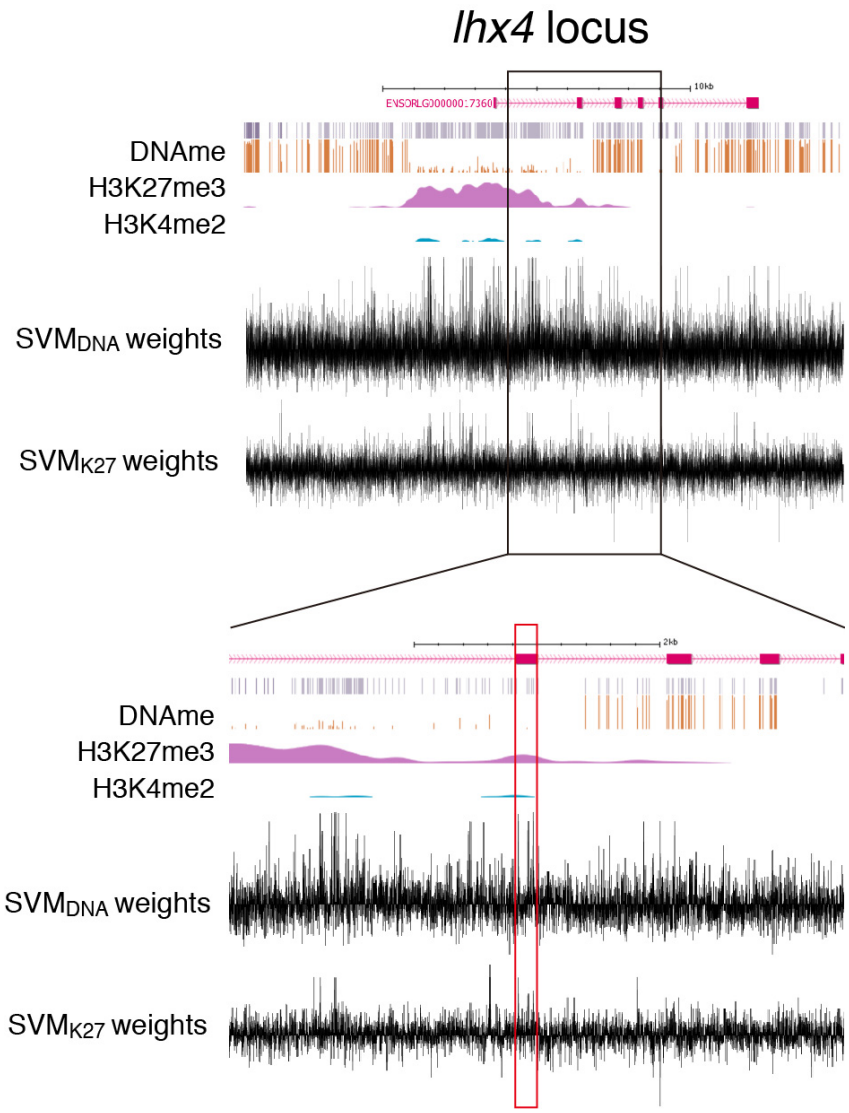

Figure 45. Some exons of developmental genes posses K27HMD-specific sequences

An example of developmental gene exon that have several 6-mers with large positive SVM weights. The SVM weights for every 6-mers along the *lhx4* locus are visualized in the bottom two tracks.

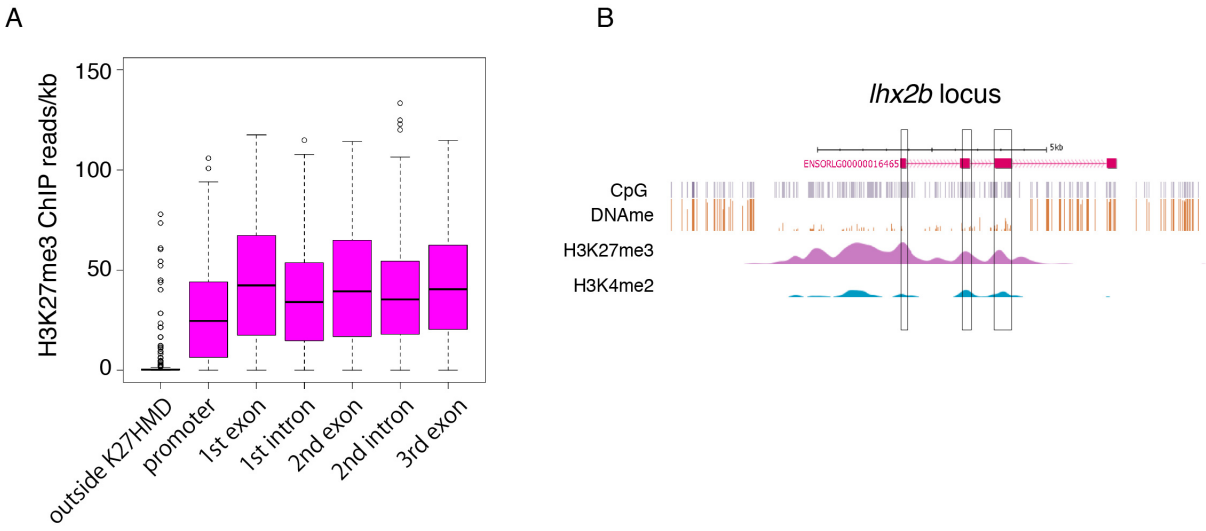

Figure 46. Exons tend to have high H3K27me3 enrichment

(A) Boxplots showing the enrichment of H3K27me3 ChIP reads at exons, introns, promoters, and outside K27HMD regions. 1 kb upstream region from TSS was used as promoter region. In the box plots, the bottom and top of the boxes correspond to the 25th and 75th percentiles and the internal band is the 50th percentile (median). The plot whiskers extending outside the boxes correspond to the lowest and highest datum within 1.5 interquartile range of the lower and upper quartiles, respectively.

(A) An example of H3K27me3 accumulation at exons. The 3rd exon locates at the K27HMD boundary. Black boxes represent exon regions inside the K27HMDs.

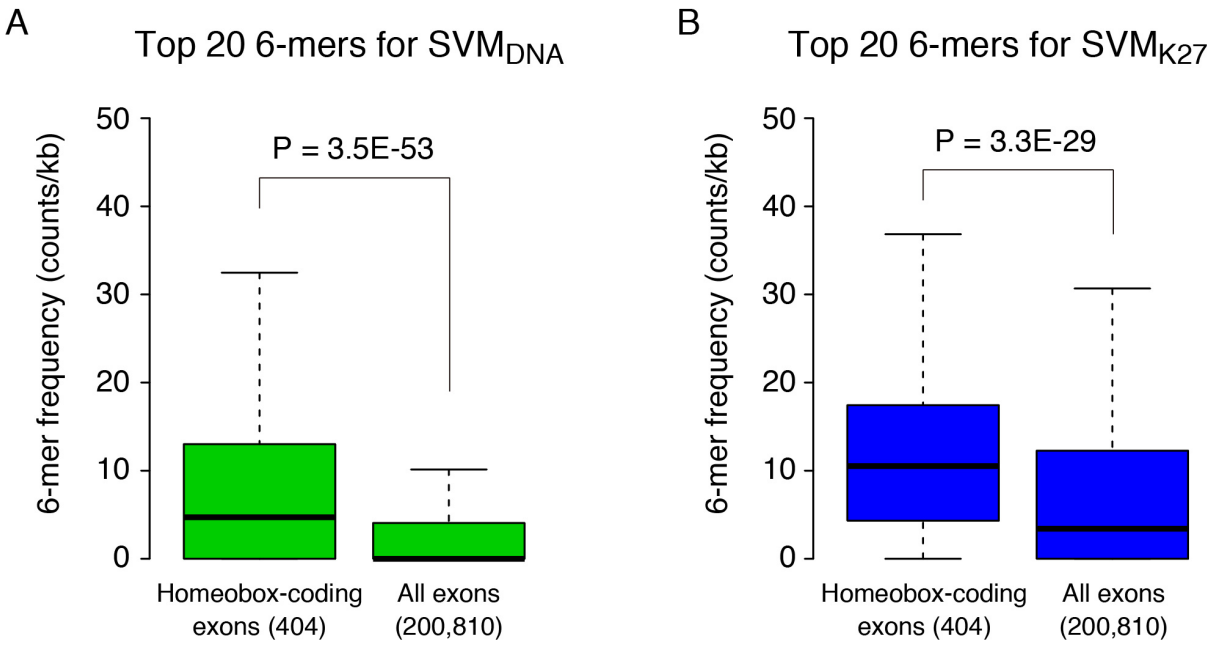

Figure 47. Homeobox-coding exons tend to have high K27HMD sequence features

Boxplots showing the frequencies of 6-mers with the largest positive weights from SVM_{DNA} (A) and SVM_{K27} (B) for homeobox-coding and all exons. In the box plots, the bottom and top of the boxes correspond to the 25th and 75th percentiles and the internal band is the 50th percentile (median). The plot whiskers extending outside the boxes correspond to the lowest and highest datum within 1.5 interquartile range of the lower and upper quartiles, respectively. P-values were calculated using non-paired Wilcoxon tests.

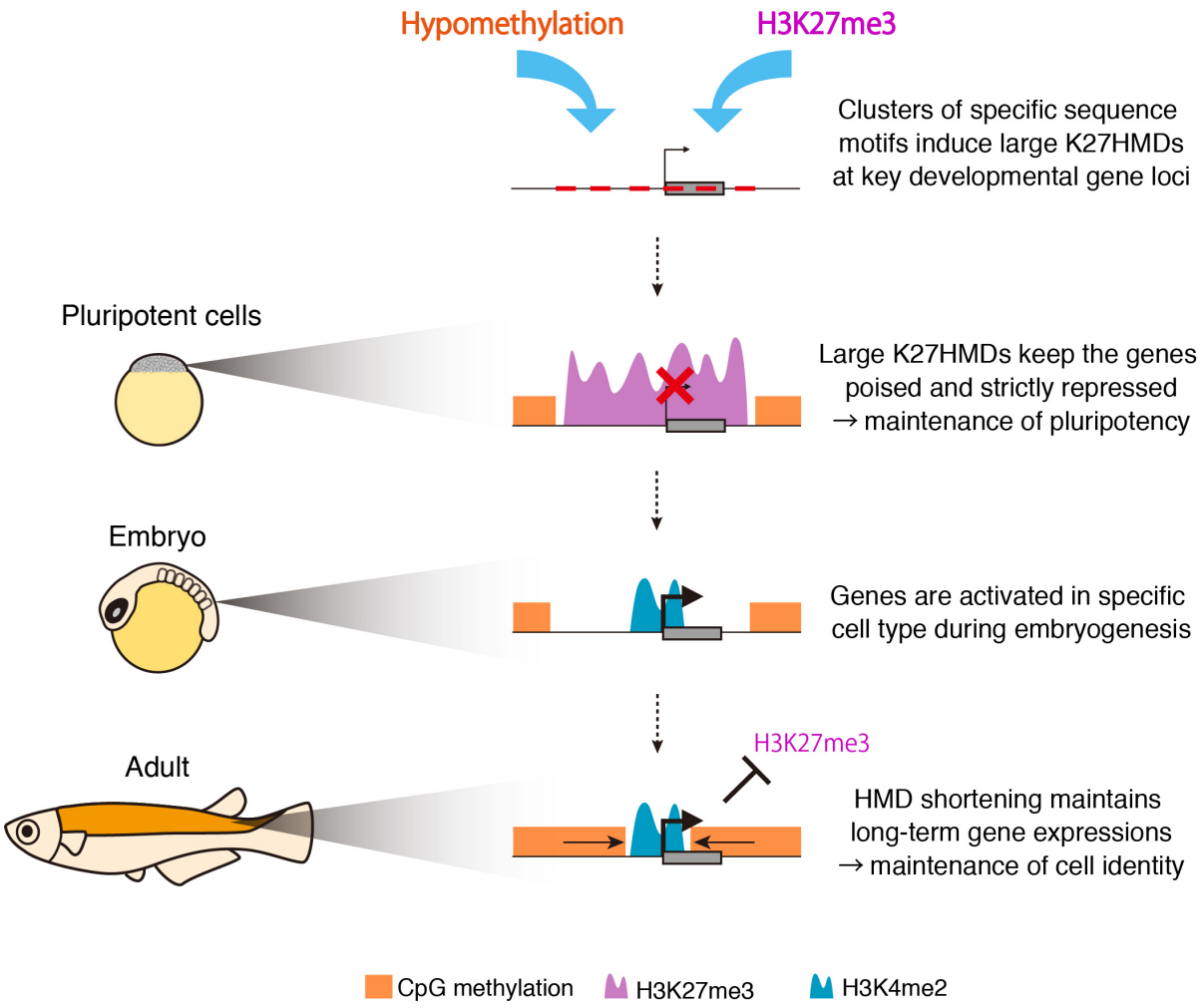

Figure 48. Model for epigenetic regulation of key developmental genes in vertebrates

Schematic illustration of transcriptional regulation of key developmental genes by large K27HMDs. First, the future large K27HMD regions are specified by cluster of specific DNA motifs. In pluripotent cells, large K27HMDs maintain strict repression of poised promoter by strong accumulation of H3K27me3, which ensure the pluripotency. During development, H3K27me3 is depleted and transcription activates in specific cell type. After embryogenesis, accumulation of DNA methylation shorten the large HMDs at activated gene loci and protects from H3K27me3 to consolidate the gene expressions, and the cell identity is 'memorized'.

Tables

Table 1. HMD-specific and methylated-region-specific 6-mers

Top 20 6-mers with the largest positive SVM_{DNA} weights			Top 20 6-mers with the largest negative SVM_{DNA} weights		
6-mer	Reverse complement	weight	6-mer	Reverse complement	weight
CGCGCA	TGCGCG	4.305434531	ACGCCA	TGGCGT	−2.355760447
ACGCGC	GCGCGT	4.252922865	GGGTCA	TGACCC	−1.984257473
GCGCAC	GTGCGC	4.228706635	ACTCCT	AGGAGT	−1.940315639
GCTAAC	GTTAGC	3.896890691	CTTCCA	TGGAAG	−1.882896423
CGCGCG	CGCGCG	3.89490491	AGACGG	CCGTCT	−1.863347171
CTGCGC	GCGCAG	3.892881458	CAGGGA	TCCCTG	−1.855712115
AGCTAG	CTAGCT	3.874144547	CCCCAG	CTGGGG	−1.790161205
CGGAAG	CTTCCG	3.868014368	CTTACA	TGTAAG	−1.769817506
GCGCGC	GCGCGC	3.84644681	ACCCAG	CTGGGT	−1.75642334
CCGGAA	TTCCGG	3.816108759	CCCCCA	TGGGGG	−1.721958675
GCTAGC	GCTAGC	3.475428531	AGTGTG	CACACT	−1.70674528
ATGCGC	GCGCAT	3.438250235	CCCAAC	GTTGGG	−1.662635454
ACGTGA	TCACGT	3.408373025	ACTCAT	ATGAGT	−1.646326319
CGCGGA	TCCGCG	3.27767622	ACATTG	CAATGT	−1.636411463
CACGTG	CACGTG	3.245874108	TGCCCA	TGGGCA	−1.611770933
GCGGAA	TTCCGC	3.205584156	CCCATA	TATGGG	−1.58180312
CACGCG	CGCGTG	3.178938699	CCCAGG	CCTGGG	−1.57819384
CGCGAG	CTCGCG	3.036252997	CAGAAG	CTTCTG	−1.572077365
CGCACG	CGTGCG	2.959328025	CCCTGA	TCAGGG	−1.566889826
TGCGCA	TGCGCA	2.925913607	AAGATT	AATCTT	−1.561635785

Table 2. K27HMD-specific and nonK27HMD-specific 6-mers

Top 20 6-mers with the largest positive SVM_{K27} weights			Top 20 6-mers with the largest negative SVM_{K27} weights		
6-mer	Reverse complement	weight	6-mer	Reverse complement	weight
GCGTAA	TTACGC	3.398194506	GCTAAC	GTTAGC	−3.432375513
TACGCA	TGCGTA	2.981769745	AGCTAG	CTAGCT	−2.636486892
ACGCAC	GTGCGT	2.942618816	CTAGCA	TGCTAG	−2.30494733
GCGCAA	TTGCGC	2.920872811	GCTAGC	GCTAGC	−2.27823116
GCGCAC	GTGCGC	2.681206916	AGCTAA	TTAGCT	−2.165999614
GGCGCA	TGCGCC	2.244584896	GGCTAA	TTAGCC	−1.902085779
CGCACA	TGTGCG	2.171333158	AACTAC	GTAGTT	−1.850377838
CTGCGC	GCGCAG	2.1368818	ACGTCA	TGACGT	−1.812292803
CGCGCA	TGCGCG	2.097184089	AAGCTA	TAGCTT	−1.634492419
AGCGCA	TGCGCT	1.979078653	AAACTA	TAGTTT	−1.620951824
ATGAAG	CTTCAT	1.908650275	AAGTAA	TTACTT	−1.618814886
ATGCGC	GCGCAT	1.761047908	CCGCCC	GGGCGG	−1.606978751
GAGGCC	GGCCTC	1.69103026	ACACAT	ATGTGT	−1.577567907
AATTAT	ATAATT	1.67353149	CCGGAA	TTCCGG	−1.555293362
CACCTG	CAGGTG	1.606315671	ACTAGC	GCTAGT	−1.515947403
TGCGCA	TGCGCA	1.605519282	CGTTAG	CTAACG	−1.504264332
CGCAAA	TTTGCG	1.568937934	AAGTTA	TAACTT	−1.49653345
AGGCCT	AGGCCT	1.554196458	AGCAGG	CCTGCT	−1.478922927
GATGAA	TTCATC	1.541396967	ACAACA	TGTTGT	−1.47081956
GCAAAA	TTTTGC	1.537971655	TAAACA	TGTTTA	−1.441560869

Table 3. Primers used in this study

Primer Name	Sequence
zic1_promoter_F	CATCAGATGAGCGTTGTAGG
zic1_promoter_R	CTGAGACGACTGAGAGCAG
zic1_negative_F	ACGCTGCATGCATCAAACAAGGC
zic1_negative_R	TGTCACACAACCCGGGCACAC
bisulfite_B_F	TGGGAAGTTGTATTAATAAGTTTTTT
bisulfite_B_R	AAATATACCACATACTTCACACCTAC
bisulfite_C_F	GAGTTTTTTTGGAGTAGTAGGGATG
bisulfite_C_R	AACTTAACCTTTACCTTTATATTTCCCC
tbx2-1_F	AACGTGCACTGACAGTGAAC
tbx2-1_R	TGGGTGAAACAACAGTGGTG
tbx2-2_F	GTCTTTTTCCCCCACAGATG
tbx2-2_R	CCCAATGACATCTGTCCTGG
pax6-1_3race	TGTCCAAGTCCCAGGGAGCGAGCCT
pax6-2_3race	GGTCCAAGTTCCAGGAAGTGAAGCA
zic1_RTqPCR_F	AGCCCTTTCCGTGTCCGTTCC
zic1_RTqPCR_R	CCGACGTGTGGACGTGCATGT
ef1a_RTqPCR_F	AAGGCTGAGCGTGAGCGTGG
ef1a_RTqPCR_R	CTCACCAACGCCAGCAGCGA

Supplementary Tables

The supplementary tables are available in the online version of the published paper at http://dev.biologists.org. (doi:10.1242/dev.108548)

References

(2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, e1001046.

Aizawa, K., Shimada, A., Naruse, K., Mitani, H. and Shima, A. (2003). The medaka midblastula transition as revealed by the expression of the paternal genome. *Gene Expr Patterns* **3**, 43-47.

Akkers, R. C., van Heeringen, S. J., Jacobi, U. G., Janssen-Megens, E. M., Francoijs, K. J., Stunnenberg, H. G. and Veenstra, G. J. (2009). A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Dev Cell* **17**, 425-434.

Arnold, P., Scholer, A., Pachkov, M., Balwierz, P. J., Jorgensen, H., Stadler, M. B., van Nimwegen, E. and Schubeler, D. (2013). Modeling of epigenome dynamics identifies transcription factors that mediate Polycomb targeting. *Genome Res* **23**, 60-73.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev* **16**, 6-21.

Birnbaum, R. Y., Clowney, E. J., Agamy, O., Kim, M. J., Zhao, J., Yamanaka, T., Pappalardo, Z., Clarke, S. L., Wenger, A. M., Nguyen, L., et al. (2012). Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res* **22**, 1059-1068.

Blackledge, N. P., Farcas, A. M., Kondo, T., King, H. W., McGouran, J. F., Hanssen, L. L., Ito, S., Cooper, S., Kondo, K., Koseki, Y., et al. (2014). Variant PRC1 Complex-Dependent H2A Ubiquitylation Drives PRC2 Recruitment and Polycomb Domain Formation. *Cell* **157**, 1445-1459.

Bogdanovic, O., Long, S. W., van Heeringen, S. J., Brinkman, A. B., Gomez-Skarmeta, J. L., Stunnenberg, H. G., Jones, P. L. and Veenstra, G. J. (2011). Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis. *Genome Res* **21**, 1313-1327.

Boyer, L. A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L. A., Lee, T. I., Levine, S. S., Wernig, M., Tajonar, A., Ray, M. K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349-353.

Brinkman, A. B., Gu, H., Bartels, S. J., Zhang, Y., Matarese, F., Simmer, F., Marks, H., Bock, C., Gnirke, A., Meissner, A., et al. (2012). Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. *Genome Res* **22**, 1128-1138.

Cedar, H. and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. *Nat Rev Genet* **10**, 295-304.

Darnell, J. E., Jr. (2002). Transcription factors as targets for cancer therapy.

Nat Rev Cancer **2**, 740-749.

Deaton, A. M. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev* **25**, 1010-1022.

Ernst, J. and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods* **9**, 215-216.

Farcas, A. M., Blackledge, N. P., Sudbery, I., Long, H. K., McGouran, J. F., Rose, N. R., Lee, S., Sims, D., Cerase, A., Sheahan, T. W., et al. (2012). KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *Elife* **1**, e00205.

Fu, Y., Sinha, M., Peterson, C. L. and Weng, Z. (2008). The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4**, e1000138.

Fujikura, J., Yamato, E., Yonemura, S., Hosoda, K., Masui, S., Nakao, K., Miyazaki Ji, J. and Niwa, H. (2002). Differentiation of embryonic stem cells is induced by GATA factors. *Genes Dev* **16**, 784-789.

Hagarman, J. A., Motley, M. P., Kristjansdottir, K. and Soloway, P. D. (2013). Coordinate regulation of DNA methylation and H3K27me3 in mouse embryonic stem cells. *PLoS One* **8**, e53880.

Handoko, L., Xu, H., Li, G., Ngan, C. Y., Chew, E., Schnapp, M., Lee, C. W., Ye, C., Ping, J. L., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat Genet* **43**, 630-638.

Harrelson, Z., Kelly, R. G., Goldin, S. N., Gibson-Brown, J. J., Bollag, R. J., Silver, L. M. and Papaioannou, V. E. (2004). Tbx2 is essential for patterning the atrioventricular canal and for morphogenesis of the outflow tract during heart development. *Development* **131**, 5041-5052.

He, J., Shen, L., Wan, M., Taranova, O., Wu, H. and Zhang, Y. (2013). Kdm2b maintains murine embryonic stem cell status by recruiting PRC1 complex to CpG islands of developmental genes. *Nat Cell Biol* **15**, 373-384.

Hendrich, B. and Tweedie, S. (2003). The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet* **19**, 269-277.

Hovestadt, V., Jones, D. T., Picelli, S., Wang, W., Kool, M., Northcott, P. A., Sultan, M., Stachurski, K., Ryzhova, M., Warnatz, H. J., et al. (2014). Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. *Nature*.

Hu, J. L., Zhou, B. O., Zhang, R. R., Zhang, K. L., Zhou, J. Q. and Xu, G. L. (2009). The N-terminus of histone H3 is required for de novo DNA methylation in chromatin. *Proc Natl Acad Sci U S A* **106**, 22187-22192.

Huang, d. W., Sherman, B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57.

Iwamatsu, T. (2004). Stages of normal development in the medaka Oryzias latipes. *Mech Dev* **121**, 605-618.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., et al. (2004). Genome duplication in the teleost fish Tetraodon

nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-957.

Jeltsch, A. (2006). On the enzymatic properties of Dnmt1: specificity, processivity, mechanism of linear diffusion and allosteric regulation of the enzyme. *Epigenetics* **1**, 63-66.

Jeong, M., Sun, D., Luo, M., Huang, Y., Challen, G. A., Rodriguez, B., Zhang, X., Chavez, L., Wang, H., Hannah, R., et al. (2014). Large conserved domains of low DNA methylation maintained by Dnmt3a. *Nat Genet* **46**, 17-23.

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y., et al. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714-719.

Kawanishi, T., Kaneko, T., Moriyama, Y., Kinoshita, M., Yokoi, H., Suzuki, T., Shimada, A. and Takeda, H. (2013). Modular development of the teleost trunk along the dorsoventral axis and zic1/zic4 as selector genes in the dorsal module. *Development* **140**, 1486-1496.

Khan, A. H., Lin, A. and Smith, D. J. (2012). Discovery and characterization of human exonic transcriptional regulatory elements. *PLoS One* **7**, e46098.

Ku, M., Koche, R. P., Rheinbay, E., Mendenhall, E. M., Endoh, M., Mikkelsen, T. S., Presser, A., Nusbaum, C., Xie, X., Chi, A. S., et al. (2008). Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. *PLoS Genet* **4**, e1000242.

Kumaki, Y., Oda, M. and Okano, M. (2008). QUMA: quantification tool for methylation analysis. *Nucleic Acids Res* **36**, W170-175.

Lancaster, E. and Dalmau, J. (2012). Neuronal autoantigens--pathogenesis, associated disorders and antibody testing. *Nat Rev Neurol* **8**, 380-390.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C. T., Low, H. M., Kin Sung, K. W., Rigoutsos, I., Loring, J., et al. (2010). Dynamic changes in the human methylome during differentiation. *Genome Res* **20**, 320-331.

Lee, D., Karchin, R. and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* **21**, 2167-2180.

Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K., et al. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* **125**, 301-313.

Lindeman, L. C., Andersen, I. S., Reiner, A. H., Li, N., Aanes, H., Ostrup, O., Winata, C., Mathavan, S., Muller, F., Alestrom, P., et al. (2011). Prepatterning of developmental gene expression by modified histones before zygotic genome activation. *Dev Cell* **21**, 993-1004.

Lindeman, L. C., Vogt-Kielland, L. T., Aleström, P. and Collas, P. (2009). Fish'n ChIPs: chromatin immunoprecipitation in the zebrafish embryo. *Methods Mol Biol* **567**, 75-86.

Lindroth, A. M., Park, Y. J., McLean, C. M., Dokshin, G. A., Persson, J. M., Herman, H., Pasini, D., Miro, X., Donohoe, M. E., Lee, J. T., et al. (2008). Antagonism between DNA and H3K27 methylation at the imprinted Rasgrf1 locus. *PLoS Genet* **4**, e1000145.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315-322.

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., et al. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68-73.

Long, H. K., Sims, D., Heger, A., Blackledge, N. P., Kutter, C., Wright, M. L., Grutzner, F., Odom, D. T., Patient, R., Ponting, C. P., et al. (2013). Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife* **2**, e00348.

Merzdorf, C. S. (2007). Emerging roles for zic genes in early development. *Dev Dyn* **236**, 922-940.

Mohn, F., Weber, M., Rebhan, M., Roloff, T. C., Richter, J., Stadler, M. B., Bibel, M. and Schubeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Mol Cell* **30**, 755-766.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W. R., Hannon, G. J. and Smith, A. D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**, 1029-1041.

Moriyama, Y., Kawanishi, T., Nakamura, R., Tsukahara, T., Sumiyama, K., Suster, M. L., Kawakami, K., Toyoda, A., Fujiyama, A., Yasuoka, Y., et al. (2012). The medaka zic1/zic4 mutant provides molecular insights into teleost caudal fin evolution. *Curr Biol* **22**, 601-607.

Neri, F., Krepelova, A., Incarnato, D., Maldotti, M., Parlato, C., Galvagni, F., Matarese, F., Stunnenberg, H. G. and Oliviero, S. (2013). Dnmt3L antagonizes DNA methylation at bivalent promoters and favors DNA methylation at gene bodies in ESCs. *Cell* **155**, 121-134.

O'Carroll, D., Erhardt, S., Pagani, M., Barton, S. C., Surani, M. A. and Jenuwein, T. (2001). The polycomb-group gene Ezh2 is required for early mouse development. *Mol Cell Biol* **21**, 4330-4336.

Okitsu, C. Y. and Hsieh, C. L. (2007). DNA methylation dictates histone H3K4 methylation. *Mol Cell Biol* **27**, 2746-2757.

Ooi, S. K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S. P., Allis, C. D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714-717.

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S. and Natoli,

G. (2013). Latent enhancers activated by stimulation in differentiated cells. *Cell* **152**, 157-171.

Pan, G., Tian, S., Nie, J., Yang, C., Ruotti, V., Wei, H., Jonsdottir, G. A., Stewart, R. and Thomson, J. A. (2007). Whole-genome analysis of histone H3 lysine 4 and lysine 27 methylation in human embryonic stem cells. *Cell Stem Cell* **1**, 299-312.

Potok, M. E., Nix, D. A., Parnell, T. J. and Cairns, B. R. (2013). Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. *Cell* **153**, 759-772.

Puschel, A. W., Gruss, P. and Westerfield, M. (1992). Sequence and expression pattern of pax-6 are highly conserved between zebrafish and mice. *Development* **114**, 643-651.

Qu, W., Hashimoto, S., Shimada, A., Nakatani, Y., Ichikawa, K., Saito, T. L., Ogoshi, K., Matsushima, K., Suzuki, Y., Sugano, S., et al. (2012). Genome-wide genetic variations are highly correlated with proximal DNA methylation patterns. *Genome Res* **22**, 1419-1425.

Rinn, J. L., Wang, J. K., Allen, N., Brugmann, S. A., Mikels, A. J., Liu, H., Ridky, T. W., Stadler, H. S., Nusse, R., Helms, J. A., et al. (2008). A dermal HOX transcriptional program regulates site-specific epidermal fate. *Genes Dev* **22**, 303-307.

Ritter, D. I., Dong, Z., Guo, S. and Chuang, J. H. (2012). Transcriptional enhancers in protein-coding exons of vertebrate developmental genes. *PLoS One* **7**, e35202.

Rush, M., Appanah, R., Lee, S., Lam, L. L., Goyal, P. and Lorincz, M. C. (2009). Targeting of EZH2 to a defined genomic site is sufficient for recruitment of Dnmt3a but not de novo DNA methylation. *Epigenetics* **4**, 404-414.

Saito, T. L., Yoshimura, J., Sasaki, S., Ahsan, B., Sasaki, A., Kuroshu, R. and Morishita, S. (2009). UTGB toolkit for personalized genome browsers. *Bioinformatics* **25**, 1856-1861.

Salter, M. W. and Kalia, L. V. (2004). Src kinases: a hub for NMDA receptor regulation. *Nat Rev Neurosci* **5**, 317-328.

Sasaki, S., Mello, C. C., Shimada, A., Nakatani, Y., Hashimoto, S., Ogawa, M., Matsushima, K., Gu, S. G., Kasahara, M., Ahsan, B., et al. (2009). Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites. *Science* **323**, 401-404.

Segal, E. and Widom, J. (2009). What controls nucleosome positions? *Trends Genet* **25**, 335-343.

Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nat Rev Genet* **14**, 204-220.

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E. J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480**, 490-495.

Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., Raubitschek, A., Ziegler, S., LeProust, E. M., Akey, J. M., et al. (2013a). Exonic transcription factor binding directs codon choice and

affects protein evolution. *Science* **342**, 1367-1372.

Stergachis, A. B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S. L., Vernot, B., Cheng, J. B., Thurman, R. E., Sandstrom, R., et al. (2013b). Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. *Cell* **154**, 888-903.

Suzuki, M. M. and Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**, 465-476.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872.

Takashima, S., Shimada, A., Kobayashi, D., Yokoi, H., Narita, T., Jindo, T., Kage, T., Kitagawa, T., Kimura, T., Sekimizu, K., et al. (2007). Phenotypic analysis of a novel chordin mutant in medaka. *Dev Dyn* **236**, 2298-2310.

Taylor, J. S., Braasch, I., Frickey, T., Meyer, A. and Van de Peer, Y. (2003). Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res* **13**, 382-390.

Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R., Deaton, A., Andrews, R., James, K. D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**, 1082-1086.

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H. R., et al. (2006). Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. *Genes Cells* **11**, 805-814.

Tweedie, S., Charlton, J., Clark, V. and Bird, A. (1997). Methylation of genomes and genes at the invertebrate-vertebrate boundary. *Mol Cell Biol* **17**, 1469-1475.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M. and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Methods* **5**, 829-834.

Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z. and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520.

van Heeringen, S. J., Akkers, R. C., van Kruijsbergen, I., Arif, M. A., Hanssen, L. L., Sharifi, N. and Veenstra, G. J. (2014). Principles of nucleation of H3K27 methylation during embryonic development. *Genome Res* **24**, 401-410.

Vastenhouw, N. L., Zhang, Y., Woods, I. G., Imam, F., Regev, A., Liu, X. S., Rinn, J. and Schier, A. F. (2010). Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464**, 922-926.

Wang, K. C., Yang, Y. W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B. R., Protacio, A., Flynn, R. A., Gupta, R. A., et al. (2011). A long noncoding RNA maintains active chromatin to

coordinate homeotic gene expression. *Nature* **472**, 120-124.

Ward, L. D. and Kellis, M. (2012). Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**, 1675-1678.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I. and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319.

Woo, C. J., Kharchenko, P. V., Daheron, L., Park, P. J. and Kingston, R. E. (2010). A region of the human HOXD cluster that confers polycomb-group responsiveness. *Cell* **140**, 99-110.

Wu, H., Coskun, V., Tao, J., Xie, W., Ge, W., Yoshikawa, K., Li, E., Zhang, Y. and Sun, Y. E. (2010). Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science* **329**, 444-448.

Wu, H., D'Alessio, A. C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Sun, Y. E. and Zhang, Y. (2011a). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature* **473**, 389-393.

Wu, S. F., Zhang, H. and Cairns, B. R. (2011b). Genes for embryo development are packaged in blocks of multivalent chromatin in zebrafish sperm. *Genome Res* **21**, 578-589.

Xie, W., Schultz, M. D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J. W., Tian, S., Hawkins, R. D., Leung, D., et al. (2013). Epigenomic analysis of multilineage differentiation of human embryonic stem cells. *Cell* **153**, 1134-1148.

Xu, G., Deng, N., Zhao, Z., Judeh, T., Flemington, E. and Zhu, D. (2011). SAMMate: a GUI tool for processing short read alignments in SAM/BAM format. *Source Code Biol Med* **6**, 2.

Yi, M., Hong, N. and Hong, Y. (2009). Generation of medaka fish haploid embryonic stem cells. *Science* **326**, 430-433.

Yokomine, T., Hata, K., Tsudzuki, M. and Sasaki, H. (2006). Evolution of the vertebrate DNMT3 gene family: a possible link between existence of DNMT3L and genomic imprinting. *Cytogenet Genome Res* **113**, 75-80.

Zhao, X. D., Han, X., Chew, J. L., Liu, J., Chiu, K. P., Choo, A., Orlov, Y. L., Sung, W. K., Shahab, A., Kuznetsov, V. A., et al. (2007). Whole-genome mapping of histone H3 Lys4 and 27 trimethylations reveals distinct genomic compartments in human embryonic stem cells. *Cell Stem Cell* **1**, 286-298.

Zhou, V. W., Goren, A. and Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* **12**, 7-18.

Acknowledgements

I would like to express my deepest and sincere gratitude to my supervisor, Dr. Hiroyuki Takeda (The University of Tokyo) for providing me with the opportunity to study in a splendid environment.

I would like to express my sincere appreciation to Dr. Tatsuya Tsukahara (The University of Tokyo) for his support and discussion of my experiment. Without his help, this thesis would not have been materialized.

I would also like to express my gratitude to Dr. Morishita Shinichi (The University of Tokyo) and Dr. Wei Qu (The University of Tokyo) for the collaboration and the generous help about the computational analyses.

I would like to thank Mr. Kazuki Ichikawa (The University of Tokyo) for ChromHMM analysis, Mr. Otsuka Takayoshi (The University of Tokyo) for ChIP using medaka liver, Dr. Taro L Saito (The University of Tokyo) for utilization of UTGB, Dr. Shinichi Hashimoto (Kanazawa University) and Dr. Katsumi Ogoshi (The University of Tokyo) for whole genome bisulfite sequencing, and Dr. Yutaka Suzuki (The University of Tokyo) for the sequencing by next generation sequencer.

I am truly grateful to the member of Takeda Laboratory (the Laboratory of Embryology, Department of Biological Sciences, Graduate School of Science, The University of Tokyo) for all they have done for my life in the laboratory.

Finally, I am greatly indebted to my parents for their heartfelt support and generous affection, without which I would not have accomplished this study. I dedicate this doctoral thesis to them.