学位論文

Comparative analysis of genome and epigenome between two polymorphic medaka populations

（2 種の多型メダカ集団におけるゲノム・エピゲノム比較解析）

平成 27 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

宇野　絢子

Contents

Abbreviations

- HMD hypomethylated domain
- TSS transcription start site
- RPKM reads per kilobase of exon per million mapped sequence reads
- SNP single nucleotide polymorphism
- indel insertion and deletion
- MI mutation index
- DMR differentially methylated region
- TF transcription factor
- DHS DNase I hypersensitive site
- Tg transgenic

Abstract

The genomes of vertebrates are globally methylated, but a small portion of genomic regions is known to be hypomethylated. Although hypomethylated domains (HMDs) have been implicated in transcriptional regulation in various ways, how a HMD is determined in a particular genomic region remains elusive.

In Chapter 1, to search for DNA motifs essential for the patterning of HMDs, I performed the genome-wide comparative analysis of genome and DNA methylation patterns of the two medaka inbred lines, Hd-rRII1 (referred to as Hd-rR) and HNI-II (referred to as HNI), which are established from two closely related species in Japan, *Oryzias latipes* and *Oryzias sakaizumii*, respectively, and exhibit high levels of genetic variations between them (SNP, ~ 3%). I successfully mapped > 70% of HMDs in both genomes and found that the majority of those mapped HMDs are conserved between the two lines (common HMDs). While a large part of the common HMDs resided in gene promoters, more than half of species-specific HMDs were located in gene bodies or outside genes. Unexpectedly, the average genetic variation rates were similar between the common HMDs and other genome regions. However, I identified well-conserved short motifs (6-mers) that are specifically enriched in HMDs, suggesting that they could function in the patterning of HMDs in the medaka genome.

In Chapter 2, I selected 40 motifs (20 with CpGs and 20 without CpGs) from the above identified motifs which are highly conserved in the common HMDs, and further characterized them by relating their positions to accessible chromatin across the genome. First, I examined DNase-seq signals around selected motifs and found that DNase-seq signal exhibits a periodic pattern around some of those motifs, specifically within HMDs. Combining these data with nucleosome core positions determined with

MNase-seq, I revealed that these 6-mers reside specifically in linker regions within HMDs. Furthermore, I indicated that the preferential localization of most of those motifs in linker regions does not reflect simple base compositions, suggesting that they function in HMD formation as motifs that regulate nucleosome positioning within HMDs.

In Chapter 3, to examine if intrinsic local DNA sequences are responsible for differential DNA methylation pattern between the two medaka species, I made transgenic medaka carrying constructs including the Hd-rR or HNI-type sequences of those HMDs (or its methylated counterparts). I then examined the methylation pattern of F1 or F2 blastula-stage embryos of these transgenic lines by bisulfite analysis. Unexpectedly, I found that DNA methylation did not occur or occurred only partially, if any, in all transgenes irrespective of their original methylation status. These results indicated that, unlike in mammals, *de novo* methylation fails to target exogenous DNA fragments in medaka.

In summary, my comparative analyses of genomes and epigenomes between Hd-rR and HNI and subsequent transgenic analyses provide unique insights into the mechanisms underlying HMD formation in the vertebrate genomes.

General introduction

Nowadays, the term ‘epigenetics’ is considered to refer to heritable changes in gene expression that does not involve changes in underlying DNA sequences. This term, which was coined by Waddington in 1942, was derived from the Greek word “epigenesis”, which originally described the influence of genetic processes on development (http://www.whatisepigenetics.com/fundamentals/). In his report, Waddington described that “between genotype and phenotype lies a whole complex of development processes”, for which he proposed the name ‘epigenotype’. Furthermore, he insisted on the need to discover the processes involved in the mechanism by which the genes of the genotype bring about phenotypic effects, and pointed out that the important part of such task is to discover the causal mechanisms at work, and to relate them as far as possible to what experimental embryology has revealed of the mechanics of development. He named such studies ‘epigenetics’ (Waddington, 1942, reprinted in 2012).

Since he emphasized the importance of epigenetics, from 1942 until now, 2016, the world of epigenetics has continued to expand. During over last 70 years, we have obtained so much information in epigenetics from various aspects, which helped us to understand the complicated processes linking genotype to phenotype. Now we have some fundamental knowledge in this field, which was unknown about 70 years ago. For example, DNA is wrapped around histone octamers, thereby consisting of a structure called ‘nucleosome.’ Nucleosome composes the higher-order structure, called chromatin (for review, see Szerlong and Hansen, 2011). The questions of how DNA is wrapped around histone octamers and how nucleosomes are packed have been addressed to understand gene regulation, because the resulting chromatin structure greatly affects

gene expression (for review, see Wallrath et al., 1994; Li et al., 2007), which could lead to the phenotypic changes. One of the so-called ‘epigenetic modifications,’ chemical modification to histone proteins such as acetylation and methylation, can alter the chromatin structure directly or indirectly (for review, see Li et al., 2007). In addition, as another epigenetic modification, DNA methylation is closely related to nucleosome packing, and in particular DNA methylation at gene promoters are known to function in stable repression of gene expression (for review, see Bird, 2002). The dynamic changes in such epigenetic modifications are considered to be essential for development, growth and differentiation of eukaryotes.

From a larger point of view beyond developmental processes, environment, aging, and even our lifestyles can affect the epigenetic status in the genome, which is sometimes related to diseases. Cancer is the first disease which was reported to be associated with epigenetic changes (Feinberg and Vogelstein, 1983). Currently, abnormal hypermethylation of tumor-repressing genes and/or hypomethylation in oncogenes are known to be strongly associated with cancers (Akhavan-Niaki and Samadani, 2013).

However, despite the accumulated knowledge as described above, we are far from complete understanding of epigenetics. One example is the patterning of DNA methylation. Needless to say, DNA methylation is one of the most fundamental and well-studied epigenetic modifications. Indeed, in addition to gene silencing at promoter regions, a wide variety of functions of DNA methylation at gene bodies and intergenic regions have been reported (for review, see Jones, 2012). Intriguingly, while the pattern of DNA methylation affects cell differentiation, the large part of methylation patterns, especially most of the hypomethylated domains, established by the blastula stage, are

largely maintained during development and growth (Laurent et al., 2010; Stadler et al., 2011; Potok et al., 2013; Lee et al., 2015). Thus, it is essential to ask how the methylation patterns in these pluripotent cells are determined in the specific regions (where to be highly-methylated and where to be hypomethylated) but it still remains largely unknown.

Previous studies suggest that a local sequence rule determines nearby methylation status (Lienert et al., 2011; Schilling et al., 2009; Stadler et al., 2011), although its entity remains elusive. In this context, I thought that the Japanese killifish, medaka, is a very attractive model organism from several reasons (for review, see Takeda and Shimada, 2010). The big advantage is that inbred lines are established, and high quality genomes are available for two lines, Hd-rRII1 and HNI-II (Kasahara et al., 2007). It was reported that there is a substantial genetic variation between them (Kasahara et al., 2007) but they can mate and produce healthy offspring under laboratory conditions. I thought that they have the genome which can be aligned reliably to the other one but show high incidence of genetic variations, which should be useful to identify conserved sequences between Hd-rR and HNI within HMDs.

Furthermore, medaka has a compact genome size (~ 800 Mb), which is only one-third of the size of human genome (~ 3 Gb). This makes calculation time relatively shorter in genome-wide computational analysis. In addition, the data of epigenetic modifications including DNA methylation by the previous studies of my laboratory and collaborators are available in medaka (Nakamura et al., 2014; Nakatani et al., 2015; Qu et al., 2012; Sasaki et al., 2009). Furthermore, most epigenetic studies has focused on mammals (human and mouse) and used cultured cells such as ES cells, and thus the study using medaka, which is evolutionary distant from mammals, should give us novel

insights into distinct and conserved mechanisms of epigenetics in the vertebrate lineage. Indeed, recent advances in experimental techniques such high-throughput sequencing allowed us to investigate various organisms, leading to the notion that the mechanisms discovered in some model organisms are sometime not applicable to other organisms. One example is the existence or absence of global DNA demethylation during early development. While mammals show global demethylation and re-establishment in early embryogenesis (for review, see Wu and Zhang, 2010), some other vertebrates are suggested to lack such global clearance of methylation patterns (Macleod et al., 1999; Veenstra and Wolffe, 2001; Walter et al., 2002).

In my doctoral thesis, to understand the patterning mechanisms of DNA methylation, I use the medaka system, in particular the two inbred lines Hd-rRII1 and HNI-II, which were established from two closely related species in Japan, *Oryzias latipes* and *Oryzias sakaizumii*, respectively, focusing that there exist a high incidence of genetic variations between them. My doctoral thesis consists of three chapters. In Chapter 1, I compared DNA hypomethylated domains (HMDs) at blastula cells genome-wide between the two medaka species, and identified short DNA sequences which are conserved and enriched in the HMDs shared by the two species. In Chapter 2, I examined the relationship between identified short sequences and chromatin open structure using DNase-seq and MNase-seq data in medaka. In Chapter 3, to examine whether sequence differences between Hd-rR and HNI account for the difference in methylation status seen in species-specific HMDs, I made transgenic medaka carrying the sequences of HMD and performed bisulfite analysis for those fish.

Chapter 1:

Identification of conserved sequence preferences for DNA hypomethylated domains

Introduction

Methylation of cytosine at CpG dinucleotides is one of the most fundamental epigenetic modifications of vertebrate genomes. DNA methylation is often described as 'silencing' epigenetic mark, as DNA methylation at gene promoters is associated with stable repression of gene expression (for review, see Bird, 2002). In vertebrates, a small portion of genomic regions are known to be hypomethylated, and such hypomethylated domains (HMDs) are often seen in gene promoters (Hendrich and Tweedie, 2003). Most of those HMDs serve as a site for binding of transcription factors and accumulate histone modification, mostly active and sometimes repressive-type (Andersen et al., 2012; Jeong et al., 2014; Nakamura et al., 2014) and thereby contribute to transcriptional regulation of nearby genes. In addition to these promoter-associated HMDs, some of the HMDs are seen in the regions distant from promoters. Recent studies have reported a wide variety of functions of DNA methylation at gene bodies and intergenic regions such as regulation of transcriptional elongation, splicing, alternative promoters, enhancers, and insulators (for review, see Jones, 2012). Hence, the establishment of HMDs, in particular, how a HMD is determined in a particular genomic region, has been a subject of intense studies in genome science.

Cis-regulatory sequences are thought to initially determine the epigenetic code, a combination of DNA methylation and histone modifications. Indeed, the analysis using hybrid mice of two inbred lines demonstrated that DNA methylation patterns are regulated by cis-sequences (Schilling et al., 2009). Consistent with this, a transgenic approach has revealed that the methylation patterns of inserted DNA sequences maintained their original status (Lienert et al., 2011). The strong association between genotype and DNA methylation in human family also supports the importance of

cis-elements (Gertz et al., 2011). However, consensus DNA sequences that regulate the pattern of DNA methylation remain elusive. A simple approach to look for such essential cis-elements is to find out evolutionary conserved genomic sequences among closely related species and relate them to the epigenetic code. Recent advances in DNA sequencing technology have facilitated this approach (Heinz et al., 2013; Kasowski et al., 2013; McVicker et al., 2013). However, we still have difficulties to identify conserved motifs even in human and mouse which have rich genome and epigenome resources, because of their low frequency of genetic variations within populations (~ 0.1%).

In this context, the medaka is a particularly useful model system with the high quality draft genome (Kasahara et al., 2007) and base-resolution methylome (Qu et al., 2012). Importantly, the medaka has polymorphic inbred lines from two geographically separated populations living in the northern and southern part of Japan. The two populations were separated by an appropriate evolutionary distance (4 - 18 million years) that is close enough to reliably align noncoding sequences but also entails sufficient sequence variations (SNP, ~ 3%) (Kasahara et al., 2007; Setiamarga et al., 2009; Takeda and Shimada, 2010; Takehana et al., 2003). The two populations were originally considered as one species, *Oryzias latipes*, but recently the northern one was described as a new species, *Oryzias sakaizumii* (Asai et al., 2011). However, the two species are biologically similar to each other; they can mate and produce healthy offspring under laboratory conditions, even showing hybrid vigor. Thus, the transcriptional and epigenetic profiles of the two species might be largely conserved under such large genetic variations. Thus, the comparison of the two genomes and methylomes thus would provide insights into mechanisms of HMD formation mediated

by cis-elements.

In Chapter 1, I performed the genome-wide comparison of genome and DNA methylation patterns of the two medaka inbred lines, Hd-rRII1 and HNI-II, from southern and northern species, respectively. I focused on the genome of blastula in which all cells retain pluripotency, and the epigenome of this stage is so called ‘ground-state’. In the aligned genome regions of the two species, the majority of HMDs were found to be conserved between the two species (common HMDs). Unexpectedly, common HMDs still accumulate genetic variations at a comparable level to that of the methylated regions (~ 2.8%). However, I identified short well-conserved motifs that are enriched in HMDs.

Results

A majority of HMDs commonly exist in the closely related medaka species

I first calculated the proportion of HMDs shared by the two inbred lines, Hd-rRII1 (referred to as Hd-rR) and HNI-II (referred to as HNI). Dr. R. Nakamura in my laboratory previously reported 15,145 HMDs containing at least 10 continuous low-methylated (methylation rate < 0.4) CpGs in Hd-rR blastula embryos (Nakamura et al., 2014). Based on the same criteria, I identified 16,361 HMDs in the HNI blastula embryos using the previously obtained bisulfite-sequencing data (Qu et al., 2012) and a newly assembled genome of HNI (available from http://mlab.cb.k.u-tokyo.ac.jp/~yoshimura/Medaka/#!Assembly.md). I mapped HMD sequences in one species' to the other species' genome and checked if the HMDs are shared by the two species (**Fig. 1-1A**). Due to repetitive sequences and deletions (or insertions), about 13% and 23% of Hd-rR and HNI HMDs failed to be mapped to the other genome, respectively. Of the uniquely mapped HMDs (13,165 in Hd-rR, 12,660 in HNI), approximately 95% (that is, ~ 83% of total Hd-rR HMDs, ~ 74% of the total HNI HMDs) was commonly found in the two genomes (referred to as 'common HMDs') (**Fig. 1-1B, 1-2**). Only small populations (618 or 598 HMDs in Hd-rR or HNI, respectively) had no corresponding HMDs in the other species' genome (referred to as 'species-specific HMDs'), even though the sequences were uniquely mapped in both genomes. The size of these species-specific HMDs was relatively small compared to that of the common HMDs (**Fig. 1-3**).

As the HMD generally overlaps with the gene promoter (Nakamura et al., 2014), I examined if such tendency is also the case for each set of HMDs. I defined the position of transcription start sites (TSSs) according to the Ensembl genome database

(http://www.ensembl.org) and classified genomic regions into three regions as follows, 1. promoter regions (the regions from 5 kb upstream to 2 kb downstream of TSSs), 2. gene bodies (the regions from 2 kb downstream of TSSs to the end of the genes) and 3. the regions outside gene. I found that a large part of the common HMDs (76.1%) are located at promoter regions (**Fig. 1-4**, left). On the other hand, less than one-third of species-specific HMDs were at gene promoters (29.4% for Hd-rR specific and 27.9% for HNI) (**Fig. 1-4,** middle and right). Instead, about half of the species-specific HMDs and about one-fifth of them were found in the region outside genes and gene bodies (both exon and intron), respectively.

Species-specific HMDs affect gene transcription

Next, I examined how each type of HMDs (common or species-specific) is reflected in gene transcription of the two species, by conducting a comparison of RNA-seq data. I newly obtained about 62.5 million reads from HNI blastula cells, and for Hd-rR, I utilized the previous RNA-seq data from d-rR (Nakatani et al., 2015), a closed colony line from which the Hd-rR inbred line had been established. After mapping them to the Hd-rR genome, genes were isolated and classified according to those having common HMDs or species-specific HMDs in their promoter regions or in gene bodies. As for the HMDs located in intergenic regions, I searched for their nearest genes. In order to compare the relative expression level, I calculated the ratio of RPKM (reads per kilobase of exon per million mapped sequence reads), the gene expression level normalized by the total number of the mapped reads and the length of exon, of d-rR to that of HNI (d-rR / HNI) for each gene. In the genes with common HMDs in their promoters, the median of the RPKM ratio was 0.86 (**Fig. 1-5**, left, green), which deviate a little from the ideal figure, 1.0, probably due to slightly different conditions

(e.g. sampling timing and experimental procedures) in the two independent RNA-seq experiments. In spite of this, I found a significant tendency; the expression ratio of the genes with Hd-rR-specific HMDs and HNI-specific HMDs in their promoters was significantly higher (0.97) and lower (0.60) than those with common HMDs in their promoters, respectively (**Fig. 1-5**, left, pink and blue). This suggests that the genes of which promoters are marked by HMDs tend to express at higher levels than their unmarked counterparts. The expression level of each gene in two species which has a species-specific HMD in the promoter is provided in **Table 1** and **Table 2** (Hd-rR specific in **Table 1**, HNI specific in **Table 2**). On the other hand, in the genes which have a HMD in gene bodies or are nearest to each HMD existing in intergenic regions, the ratio of RPKM did not significantly change between each gene category (**Fig. 1-5**, middle and right).

Genetic variations between Hd-rR and HNI in HMDs

High conservation of HMDs in the two divergent genomes could be explained if genetic mutation occurs less frequently in those HMD regions. To test this idea, I investigated the rate of sequence variations within the common HMDs, species-specific HMDs and methylated regions. Unexpectedly, however, the average frequency of single nucleotide polymorphisms (SNPs) did not show a big difference among those regions; the median was 2.77, 2.75, 2.83 and 2.96% for methylated, common, Hd-rR specific and HNI specific, respectively (**Fig. 1-6**, left). Thus, the incidence of genetic variations in common HMDs is comparable to that in the methylated regions.

In contrast with SNP, the indel (insertion / deletion) rate was higher in the common HMDs (**Fig. 1-6**, right). This might suggest that HMDs marking the promoter are open in chromatin structure and more susceptible to insertion/deletion events than compact

methylated regions. Indeed, the indel rate was reported to show peaks in the regions with the low nucleosome-occupancy downstream of TSSs (Sasaki et al., 2009). The indel event, however, is far less frequent as compared with SNP (0.64% (indel) vs 2.75% (SNP) in common HMDs) and does not affect much on the overall mutation rate.

Taken together, blastula-stage HMDs are well-conserved between the two medaka species in spite of high incidence of genetic variations.

Specific DNA motifs are conserved and enriched in common HMDs

The above fact that common HMDs exhibit comparable levels of SNPs led me to speculate the presence of short crucial DNA sequences that are specifically conserved during speciation. To search for such sequences, I examined the occurrence of short oligomers and their conservation between the two species in HMDs or in the methylated regions. For each of the 2,080 sequences (reverse compliment is excluded) of 6 bp long DNA oligomers (6-mers), I calculated their occurrence and mutation index (the proportion of mutated to all found 6-mers, **Fig. 1-7A**) in each region.

Given that HMDs are predominantly found at gene promoters, certain DNA motifs could be enriched simply because they are required for gene transcription, but irrelevant to DNA methylation state. To efficiently extract the candidate 6-mers essential for HMD patterning, I looked at their mutation index in species-specific HMDs where 6-mers relevant to HMD patterning were expected to be normally mutated. For this, I utilized the ratio of the mutation index of common HMDs to that of species-specific HMDs for assessment. The low value of this ratio indicates that 6-mer is preferentially conserved in common HMDs, but not in species-specific HMDs. Furthermore, since CpGs tend to be more conserved within HMDs, as they are easily lost when methylated (Bird, 1980; Coulondre et al., 1978; Shen et al., 1994), I classified

all 2,080 6-mers into two categories by the presence of CpG, and compared their ratio separately. As expected, the histograms of oligomers of each category (**Fig. 1-7B**) demonstrate that most of the 6-mers with CpGs and about a half of the 6-mers without CpGs are more conserved in common HMDs (the ratio of mutation index is < 1.0). This result further confirmed the higher conservation level of non-methylated CpGs in HMDs. Then, top 20 most conserved 6-mers in common HMDs (the ratio of mutation index is < 0.455 for CpG and < 0.664 for non-CpG) were selected as (**Fig. 1-7C**, see **Table 3** and **Table 4** to see the ratio of mutation index) and subjected to further analyses. They are specifically conserved in common HMDs, and could play a role in HMD patterning.

Then, to examine the enrichment levels of each DNA motif in common HMDs, I calculated the ratio of the frequency within common HMDs to that within the methylated regions for each 6-mer. The 6-mers with low ratio of mutation index tended to be highly enriched in common HMDs. In particular, top 20 selected 6-mers of both type (CpG and non-CpG) exhibited significantly higher enrichment levels in common HMDs compared to the methylated regions (**Fig. 1-8A**) or species-specific HMDs (**Fig. 1-8B**). These 6-mers are thus specifically enriched HMDs and at the same time, well protected against genetic mutations. Finally, I examined the distribution pattern of the conserved 6-mers in common HMDs and found that top 20 6-mers of CpG and non-CpG are highly accumulated in the HMD region (**Fig. 1-9**).

Discussion

The initial pattern of blastula-stage HMDs examined in this study has a profound effect on gene expression throughout life. Although some methylated genes are later activated by demethylation at their promoters in a cell-type specific manner, the majority of HMDs in blastula cells are largely maintained during development and growth (Laurent et al., 2010; Lee et al., 2015; Potok et al., 2013; Stadler et al., 2011).

The medaka system has provided a unique tool to gain insights into genome evolution and speciation (for review, see Takeda and Shimada, 2010). In this study, I performed the comparative analyses of genome, expression profile and DNA methylome of the two closely related medaka species, and successfully identified the candidate DNA motifs that may participate in the patterning of HMDs. The estimated divergence time of the two regional species varies depending on a method of estimation, 4 - 5 million years ago by a molecular clock hypothesis (Takehana et al., 2003) and 18 million years ago by a Bayssian model (Setiamarga et al., 2009). In spite of high accumulation of genetic variations during this long separation time, the two populations had long been considered as a single species, *Oryzias latipes*. In 2011, however, the northern population was described as a new species, *Oryzias sakaizumii* (Asai et al., 2011), which is still controversial in the medaka community. In any case, their divergent genetic backgrounds with nearly identical biological features allowed me to survey functional cis-elements throughout the genome. Furthermore, the high quality draft genome of HNI (http://mlab.cb.k.u-tokyo.ac.jp/~yoshimura/Medaka/#!Assembly.md), recently produced in addition to Hd-rR, greatly facilitated aligning homologous sequences in the two genomes.

As expected from their similar biological features, the pattern of HMDs was found to be highly conserved between the two species. However, I identified a small population of the HMDs (~ 5% of the mapped HMDs of each species) that were only found in one species. I found that the genes of which promoters are marked by species-specific HMDs tend to express at higher levels than their unmarked counterparts (**Fig. 1-5**, left). This result demonstrated that species-specific HMDs in promoter regions could contribute to species-specific gene transcription. This result is consistent with the previous report of human family that allele-specific DNA methylation accounts for differences in gene expression levels between alleles (Gertz et al., 2011). However, it should be noted that interpretation of the blastula RNA-seq data is complicated by the presence of maternal transcripts, although the maternal expression profile is expected to reflect the initial HMD pattern as the blastula HMDs tend to be largely maintained during development and growth (Laurent et al., 2010; Lee et al., 2015; Potok et al., 2013; Stadler et al., 2011).

Interestingly, the majority of the species-specific HMDs marks the gene bodies or intergenic regions. This is a sharp contrast to the common HMDs which mostly reside at gene promoters. Consistent with my finding, Hernando-Herraez et al. (2015) reported that most of human-specific DMRs (differentially methylated regions) identified by comparison with non-human primates are located in regions distal to TSSs, although they examined differentiated cells, blood cells, with different methods for identification of the targeted regions. DNA methylation in gene bodies or promoter-distal regions is thought to have diverse functions depending on context, such as transcriptional elongation, alternative splicing, control of alternative promoter usage, and alteration of activity of enhancer or insulators (for review, see Jones, 2012), and

thereby affects gene expression either positively or negatively. Indeed, in my study, species-specific HMDs located outside gene promoters did not show any correlation with the average relative transcription levels. Notably, the comparison between chick inbred lines demonstrated that DMRs responsible for differences in immune response reside in gene bodies as well as promoters (Li et al., 2015). Taken together, although about 13 or 23% of the HMDs are unmapped in each species, the species-specific HMDs most likely confer species-specific morphological and physiological characters in medaka species identified in previous studies (Ishikawa et al., 1999; Kimura et al., 2007; Tsuboko et al., 2014) and thus will be interesting targets for the future study of speciation.

Species-specific HMDs greatly helped in identifying the conserved short sequences in HMDs. These sequences are specifically enriched in the common HMDs. Furthermore, they have been protected against genetic mutations for 4 - 18 million years. Importantly, this specific protection is not observed, when they are located outside the HMD. These facts suggest that the identified short sequences play an important role in initial patterning of HMDs in the blastula genome (**Fig. 1-10**). Thus far, many attempts have been made to identify essential sequences for DNA hypomethylation (Brandeis et al., 1994; Dickson et al., 2010; Lienert et al., 2011; Macleod et al., 1994). Recently, computational analyses addressed how DNA motifs determine the epigenetic status (Luu et al., 2013; Whitaker et al., 2015). My identified sequences only partially overlapped with those reported motifs, raising the possibility that essential motifs vary among species or unknown logic works behind these various motifs. Furthermore, while some of my identified sequences partially overlapped with known binding motifs, more than half of them exhibited no similarity with known motifs (**Table 3** and **Table 4**). In

any case, I believe that further functional studies of the identified motifs will provide insight into molecular mechanisms underlining the establishment of HMDs, an essential process of genome function.

Chapter 2:

Analysis of the relationship between conserved motifs and chromatin open structure

Introduction

In Chapter 1, through comparative analysis between the two genomes of the polymorphic medaka species, Hd-rR and HNI, I identified the short sequences (6-mers) that are well conserved specifically within the common HMDs even under high incidence of genetic variations. These sequences are indeed significantly enriched in the common HMDs, suggesting that they are good candidates of the DNA motifs essential for the formation of HMDs in the medaka genome. In this chapter, I further characterized those sequences by relating them to accessible chromatin across the genome.

Regulatory DNA regions in the genome have often been analyzed by the DNase I sequencing technique (DNase-seq) that identified accessible chromatin regions as DNase I hypersensitive sites (DHSs). This technique, combined with high-throughput sequencing, can globally identify accessible chromatin regions (Neph et al., 2012). Indeed, mapping DHSs has historically been a valuable tool for identifying all different types of regulatory elements because accessible chromatin harbors promoters, enhancers, silencers, insulators and locus control regions (Boyle et al., 2008; Crawford et al., 2006). When the chromatin state around genes is closed, the genes are no longer accessible to most transcription factors. Furthermore, DNase I is known to selectively digest nucleosome linkers when the chromatin state is open, while DNA regions tightly wrapped in nucleosome is intact, i.e. closed state of the chromatin (**Fig. 2-1**). This allows for mapping the nucleosome position in open chromatin genome-wide (Zhong et al., 2016). Interestingly, it was reported that DHSs tend to exhibit low DNA methylation levels (Thurman et al., 2012), suggesting a connection between chromatin accessibility and epigenetic modification.

Nucleosome positioning along the DNA is known to play a crucial role in chromatin accessibility (Bassett et al., 2009). The nucleosome is a basic packaging unit of chromatin consisting of 147 base pairs (bp) DNA wrapped around a histone octamer. Nucleosomes are connected by linker DNA with a variable length in the range about 20 - 90 bp, forming nucleosomal arrays (one-dimensional 'beads on a string'), which is the fundamental building block of chromatin structures (for review, see Szerlong and Hansen, 2011). Positioning of nucleosomes affects accessibility of DNA binding proteins to DNA and thereby influences gene transcription (Li et al., 2007; Wallrath et al., 1994). In many eukaryotes, nucleosome arrays have been reported to be highly phased downstream of TSSs, in other words, nucleosomes exhibit binding to a specific region rather than more normal random binding (Chen et al., 2013; Lantermann et al., 2010; Mavrich et al., 2008; Ponts et al., 2010; Wu et al., 2014; Yuan et al., 2005), which could facilitate gene transcription. Intriguingly, the nucleosome structure is known to change according to the epigenetic status. Nakamura et al. (2014) reported in the medaka genome that, across the boundary of some HMDs, the chromatin status shifts from 'packed' (methylated) to 'loose' (hypomethylated); the average nucleosome core signals exhibits a clear 170 bp periodic pattern outside HMDs but the peak becomes low and less defined inside HMDs (Nakamura et al., 2014). Taken together, epigenetic modifications, nucleosome positioning and chromatin accessibility could collectively regulate gene transcription in the genome. This notion led me to speculate that the short sequences I identified in Chapter 1 could participate in any of these processes.

In Chapter 2, I characterized the identified short sequences by DNase-seq and found that in HMDs, some of them preferentially localize in the linker region of the nucleosome array. I will discuss the significance of this finding in terms of DNA-guided

nucleosome positioning in the vertebrate genome, which is still controversial in vertebrate genomes.

Results

I selected top 20 of conserved short sequences, from those with CpGs and without CpGs (**Fig. 2-2**, top), which are enriched in HMDs, and focused on these sequences in the following of my experiments. They will be sometimes referred to as selected top 20 with CpGs and without CpGs, respectively.

DNase-seq signals often show periodic distribution around the selected motifs

To relate the location of the selected top 20 to the chromatin accessible region, I utilized the DNase-seq data of d-rR blastula cells done by Dr. R. Nakamura in my laboratory (unpublished). Dr. Nakamura observed that DHSs were highly enriched in HMDs (**Supplementary figure S1**); 84.8% of HMDs contained at least one DHS and 40.7% of DHSs are found in the HMD which constitutes only 3% of the blastula genome. Notably, DNase-seq signal in HMDs showed the periodic pattern of peaks of approximately 200 bp intervals (**Supplementary figure S1**), suggesting that the DNase I cleavage pattern in the medaka blastula genome represents arrays of accessible nucleosome linkers in HMDs (Nakamura et al., unpublished).

I first examined the profiles of DNase-seq signal around each identified 6-mer. **Fig. 2-2** shows the DNase-seq profile centered by selected top 20 sequences with CpGs (left) and without CpG (right), together with the average of selected top 20 and all other 6-mers in each category (CpG or non-CpG). At first glance, the pattern varies from sequence to sequence, and a majority of them show essentially a pattern similar to that of other non-selected 6-mers in HMDs. However, I noticed that some of the selected top 20 with CpGs and without CpGs exhibit a strong periodicity within HMD, while the others do not. In the methylated region, no such periodic pattern was observed.

Although my classification was rather arbitrary, I categorized the top 20 into four categories in terms of periodicity, strong, intermediate, weak and no periodicity. Out of 40 selected 6-mers, five showed strong periodicity, three for intermediate and ten for weak periodicity in their vicinity. No such periodicity was found in the remaining twenty-two 6-mers. The peaks are highest at the center and the intervals of them were approximately 200 bp, both of which tendency were shared by all 6-mers showing the periodical pattern. These results suggest that some selected 6-mers tend to be located in nucleosome linker regions in HMDs, as the peak of DNase-seq signals in open chromatin is known to correspond to the nucleosome linker region (Zhong et al., 2016).

The selected motifs are distributed in linker regions within HMD

The above findings led me to examine the relationship between the position of the selected 6-mers and nucleosomes in HMDs. I focused on the selected 6-mers of which the DNase-seq profiles show strong and intermediate periodicity (altogether eight, three for CpG containing and five for non-CpG), and related their positions to that of nucleosome linkers and cores.

Nucleosomes are known to be highly phased downstream of TSSs in many organisms (Chen et al., 2013; Lantermann et al., 2010; Mavrich et al., 2008; Ponts et al., 2010; Wu et al., 2014; Yuan et al., 2005). Given that most of the HMDs overlap promoter regions (Nakamura et al., 2014), phased nucleosome patterns could be reflected in the observed periodic pattern of DNase-seq signals around the selected short sequences within HMD. To test this possibility, I examined the spatial relationship between selected 6-mers and nucleosome cores within HMDs. The positioning score of nucleosome cores was calculated from the previously generated data of micrococcal

nuclease-digested chromatin (MNase-seq) from Hd-rR blastula-stage embryos (Sasaki et al., 2009). MNase-seq has widely been used to determine nucleosome occupancy genome-wide and nucleosome core positioning score can be used as an indicator of the probability that nucleosome core are located to each region. I designated the position of each 6-mer within or outside HMD as position 0 and examined the profiles of nucleosome core positioning score around each selected 6-mer.

Around each 6-mer showing strong periodicity (**Fig. 2-3**, upper), nucleosome core positioning score determined by MNase-seq (blue line) showed strong periodic patterns and the valleys of the nucleosome core positioning score were mostly in phase with DNase-seq signal peaks (red line). This indicates that nucleosomes are highly phased around these 6-mers within HMDs and that the DNase-seq peak corresponds to the linker DNA region. Notably, the nucleosome core positioning score showed the lowest value at position 0, indicating that these 6-mers reside preferentially in linker DNA regions within HMDs. As for the three 6-mers with intermediate periodicity (**Fig. 2-3**, lower), they also exhibit nucleosome phasing in their neighboring regions and their preferential localization in linker regions, but such tendency is weak as compared with 6-mers showing strong periodicity.

Preferential localization of the selected motifs in linker regions does not mostly reflect the simple base composition

In general, the nucleosome core is known to favor GC-rich sequences and disfavor AT-rich sequences such as poly (dA:dT) (Nelson et al., 1987; Tillo and Hughes, 2009) (see Discussion). This fact raised the possibility that the five selected 6-mers with strong periodicity of DNase-seq signature distributed in linker regions simply because

their base composition is unfavorable for nucleosome core formation. To test this possibility, I examined the distribution pattern of the 6-mers in which the base composition is the same as the selected 6-mers but the order was reversed (**Fig. 2-4**). I examined the five 6-mers with strong periodicity and obtained the similar results for these 6-mers, except for CGCTAG. For example, while one of the selected 6-mer, GCTAGC, which showed low ratio of the mutation index (common / Hd-rR specific), exhibited the strong periodic distribution of nucleosome core positioning score within HMDs (blue line), its base-reversed version, CGATCG, exhibited no periodicity (green line). This result indicates that the distribution of specific motifs in linker regions does not simply reflect their base composition. As for CGCTAG, the reverse sequence also exhibits periodicity (**Fig. 2-4**), suggesting that the base composition is important for periodic distribution in this case.

Discussion

The selected 40 6-mers (selected top 20 with CpG and without CpGs) I examined here were selected based on the conservation between Hd-rR and HNI and specific enrichment in common HMDs. I speculated that those 6-mers tend to reside in accessible chromatin in HMDs, because they may need to interact with nuclear proteins and epigenetic machinery to exert their effects. However, the DNase-seq analysis demonstrated that nearly half of the 6-mers exhibited no such preferential localization to the accessible chromatin region. A part of them show similarity with the binding sites of known transcription factors (TFs) (**Table 3** and **Table 4**), suggesting that they could recruit such TFs and direct transcriptional activation. At the moment, I do not know the reason why they are not localized in accessible chromatin; they might work at later stages.

By contrast, a few, but not many, 6-mers shows preferentially localization to the accessible regions. The analysis with the MNase-seq data further demonstrated that they tend to be located in the nucleosome linker. Importantly, this pattern was specifically observed within HMDs, suggesting their HMD-specific roles in nucleosome positioning. Interestingly, the 6-mers with strong periodicity do not show any homology with sequences of known TFs (**Table 3** and **Table 4**). A simple idea is that the structure of these sequences may be intrinsically unfavorable for nucleosome core formation, although I cannot rule out the possibility that unknown proteins bind to those 6-mers and influence nucleosome positioning.

In principle, nucleosome organization can be guided both by intrinsic sequence preference and by the action of trans-acting factors (Beh et al., 2015; Hughes et al., 2012; Kaplan et al., 2009; Struhl and Segal, 2013; Zhang et al., 2009). Intrinsic

determinants or DNA preferences have been proposed; for example, AT-rich sequences for linker DNAs and GC-rich for nucleosome cores (Nelson et al., 1987; Tillo and Hughes, 2009). Genome-wide nucleosome mapping in the yeast (Kaplan et al., 2009) and Tetrahymena (Beh et al., 2015) genomes also demonstrated the strong dependency of nucleosome positioning on local DNA sequences, i.e. DNA-guided nucleosome positioning. However, these intrinsic sequence-based rules have failed to work in the genomes of more complex organisms such as human (Valouev et al., 2011), suggesting much greater roles of trans-acting factors in these organisms. In this context, my finding of periodic 6-mers is very important in that it suggests the presence of a novel DNA-guided nucleosome positioning in the vertebrate genome. Notably, the sequence feature of those 6-mers apparently contradicts the previously reported global sequence preference of nucleosomes; in spite of their preferential localization in the linker region, they are not AT-rich. Furthermore, most of them may work as motifs but not as simple sequence-composite preferences.

Taken together, although the molecular mechanisms remain unknown, my study focusing on the conserved motifs in the common HMDs provides insights into sequence-based mechanisms for HMD formation and nucleosome positioning.

Chapter 3:

Analysis of DNA methylation patterns of transgenic medaka carrying HMD sequence

Introduction

During DNA methylation processes at CpG sites, two distinct mechanisms are known to work; one is *de novo* methylation, and the other is maintenance methylation. In early development of mammals, global DNA demethylation occurs genome-wide after fertilization, and a new methylation pattern is established subsequently (for review, see Wu and Zhang, 2010). This establishment process is governed by *de novo* methyltransferases, DNMT3a and DNMT3b (Okano et al., 1999). These methyl marks are inherited to daughter cells during development through maintenance methyltransferase, DNMT1, which has a preference for hemi-methylated DNA (Bestor et al., 1988; Bestor and Ingram, 1983; Hermann et al., 2004). While the global demethylation is a hallmark of early embryogenesis in mammals, several studies demonstrated the absence of global demethylation in other animals such as zebrafish and *Xenopus* (Macleod et al., 1999; Veenstra and Wolffe, 2001). Also in medaka, there is a report suggesting the lack of global demethylation during early embryogenesis (Walter et al., 2002), although it only investigated DNA methylation at limited sites (CCGG).

In Chapter 1, I demonstrated that the majority of the mapped HMDs are shared by the two medaka species, Hd-rR and HNI, while a small portion of those HMDs (~ 5% of the mapped HMDs of each species) are species-specific HMDs which have the methylated counterparts in the other species' genome. I identified the 6-mers which are specifically conserved in the common HMDs. These results suggest that these motifs could act for the patterning of HMDs. Furthermore, the enrichment level of these conserved 6-mers was significantly low in species-specific HMDs than in common HMDs (**Fig. 1-8B**). These observations led me to speculate that differences in the

genomic sequences itself could account for differentially methylated patterns between the two species. In other words, differentially methylated patterns are intrinsically created by DNA sequence motifs.

In mouse stem cells, introduced DNA fragment recapitulated the methylation patterns of their endogenous sites (Lienert et al., 2011), which was the basis of my above speculation. I further speculated that the methylation pattern of a transgene could recapitulate that of its endogenous site also in medaka. However, given that the establishment processes of DNA methylation in early embryos may vary between medaka and mammals, I wanted to investigate whether introduced sequences of HMD and their methylated counterparts could recapitulate the DNA methylation status of their original sites. To address this, I made transgenic medaka fish which carry the sequence of HMDs (or its methylated counterparts) and its flanking regions, and examined DNA methylation patterns in those regions of F1 or F2 blastula-stage embryos by bisulfite analysis.

Results

Transgenes are partially methylated in HMD-containing transgenic medaka

In order to examine whether the difference in the methylation patterns in identified species-specific HMDs are caused by intrinsic sequence differences between the two medaka genomes, I made a series of transgenic medaka which contain HMD sequences. The overview of the experiment is shown in **Fig. 3-1**. For making transgenic lines, I selected the HMDs which cover promoter regions, and made constructs which contain the sequence (0.6 – 1.5 kb) of a species-specific HMD or its methylated counterpart. The constructs contained the β-actin promoter that drives the GFP expression in order to detect the presence of transgenes. These sequences were flanked by I-SceI sites (Rembold et al., 2006) to facilitate integration. The constructs I made included those containing the HNI methylated domain for the analysis of Hd-rR specific HMDs (**Fig. 3-2**), Hd-rR methylated domains for HNI specific HMDs (**Fig. 3-3**), and hypomethylated domains for common HMDs (**Fig. 3-4**). For a technical reason (Hd-rR fish spawn less eggs than d-rR), the injected host was always the d-rR line, a closed colony line from which the Hd-rR inbred line had been established. The methylation status of injected DNA fragments and their counterpart host regions was analyzed in genome DNAs extracted from F2 blastula embryos, unless otherwise noted (**Fig. 3-4**, HMD-1).

In total, seventeen transgenic lines of seven HMDs were established in d-rR hosts. As for the Hd-rR specific HMD, I obtained one line which has the methylated counterpart of the HNI genome (**Fig. 3-2**). As for HNI specific HMD, I obtained total seven lines for three HMDs (**Fig. 3-3**); five of them have Hd-rR methylated domains (left, one for HMD-1, two for HMD-2 and two for HMD-3) and two have HNI

counterparts (right). As for the common HMD (**Fig. 3-4**), I obtained total nine lines of three HMDs; four of them have Hd-rR sequence of the common HMD (left, three for HMD-1 and one for HMD-3) and five have the HNI counterparts (right, one for HMD-1, two for HMD-2 and two for HMD-3). The sequences of transgenes derived from the HNI genome was distinguished by SNPs within the regions between Hd-rR and HNI in these experiments as shown in **Fig. 3-2**. However, I was unable to distinguish Hd-rR-derived transgenes from d-rR host genome sequences because of high similarity between Hd-rR and d-rR and thus the results were presented as a mixture of endogenous and introduced sequences. In this case, the methylation status of endogenous sequences was deduced if data are available in a transgenic line having its HNI counterpart (for example, see HMD-1 in **Fig. 3-3**).

I originally thought that the methylation status in transgenes follows their original one, i.e. if a methylated fragment in a donor species is introduced into the d-rR host genome, it would regain the methylated status in descendant embryos. Unexpectedly, however, this was not the case; all injected genomic fragments were found to remain hypomethylated irrespective of their original methylation status. In some cases, methylation was detected in introduced fragments but it was very limited (for example, see **Fig. 3-3**, HMD-2, HNI type (introduced) in HNI-type introduced fish). Regarding the HNI specific HMD (**Fig. 3-3**), the sequences from two HNI specific HMDs were almost hypomethylated in two lines in which HNI specific HMDs were introduced (**Fig. 3-3**, right, HMD-1 and HMD-2, HNI type (introduced) in HNI type-introduced fish), and their counterpart endogenous sequences in the host recapitulated the same pattern as Hd-rR, i.e. methylated (**Fig. 3-3**, right, HMD-1 and HMD-2, Hd-rR type (endogenous) in HNI type-introduced fish). On the other hand, in

the fish which have the methylated counterparts of HNI specific HMDs (their corresponding sequences of Hd-rR which are highly methylated *in vivo*), both mostly hypomethylated reads and mostly methylated reads were obtained (**Fig. 3-3**, left, HMD-1, HMD-3, Hd-rR type-introduced fish). In these fish, although I was unable to distinguish the introduced Hd-rR sequences and endogenous d-rR sequences, I reasoned that the substantial hypomethylated reads were derived from the introduced sequences.

I then examined the methylation pattern of introduced sequences of the common HMDs. I found that almost all introduced sequences exhibited the hypomethylated status (**Fig. 3-4**), although some CpGs were partially methylated in the fish to which the Hd-rR or HNI type sequence of HMD-3 was introduced (**Fig. 3-4**, HMD-3, Hd-rR-type introduced or HNI-type introduced fish).

Since I failed to obtain any positive results of clearly DNA methylation in transgenes, I suspected that DNA methylation failed to occur even in originally methylated regions in both species. For this, I reexamined the methylated status of highly-methylated regions that reside within the introduced sequences and flank HMD in the transgenic lines of HNI-specific HMD-1 and HMD-2 (**Fig. 3-3**) and common HMD-1 and HMD-2 (**Fig. 3-4**). As a result, the methylation was none or only partial at all the introduced HNI sequences, while its endogenous sites were highly methylated (**Fig. 3-5**).

The partial methylation of the transgenes is observed in other transgenic medaka and differentiated cells

The failure of DNA methylation in introduced sequences could be due to the short period that passed after integration. Indeed, I examined only F1 or F2 embryos. To

test this idea, I examined whether the observed tendency was applicable to transgenic fish that passed many generations after establishment. I chose one transgenic medaka fish which were established previously and had been maintained in my laboratory. This transgenic line carries the BAC construct including *zic1/4* genes (referred to as zicTg) (Kawanishi et al., 2013). I performed the bisulfite analysis targeting blastula–stage embryos of this transgenic fish and successfully amplified three regions that are known to be highly methylated in the original genome of d-rR background. First, I confirmed that all reads were mostly methylated in two of these regions derived from the host d-rR genome at the blastula stage (**Fig. 3-6**, lower). For the transgenes, although the results were again presented as a mixture of endogenous and transgenic fragments due to their nearly identical sequences (**Fig. 3-6**, upper), while deduced endogenous sequences were highly methylated, a substantial number of the fragments remained hypomethylated.

To examine the observed partial methylation in the introduced genes were only seen in blastula cells, I examined the methylation pattern in the differentiated cells. As differentiated cells, I chose liver cells since liver has a substantial size and easy to extract from body of adult fish. I extracted genomic DNA from liver of my transgenic fish (two lines) and zicTg respectively, and performed bisulfite analysis for these three lines at the same five regions with **Fig. 3-5** (HNI specific HMD-1, HNI specific HMD-2 left and right) and **Fig. 3-6** (Region A and Region C). As shown in **Fig. 3-7**, among the three regions which I investigated, one exhibited relatively hypomethylated status in the endogenous site in adult liver (HMD-2 left, Hd-rR type), the other two regions remained highly-methylated in the endogenous sites in adult liver (HMD-1 and HMD-2 right, Hd-rR type). However, the methylated status was still incomplete in these introduced sequences (HMD-1 and HMD-2 right, HNI type). This incomplete methylation in the

introduced sequences was also confirmed in liver cells of adult fish of zicTg (**Fig. 3-8**). These results suggest that methylation occur only partial even in differentiated cells.

Discussion

In this chapter, in order to examine whether the differences in DNA methylation pattern seen in the identified species-specific HMDs are caused by intrinsic genomic sequence differences, I made transgenic medaka carrying HMD or methylated sequences, then performed the bisulfite analysis with F1 or F2 embryos of these lines. However, unexpectedly, I found that at the blastula stage, DNA methylation did not occur or partially, if any, in all the introduced sequences, irrespective of their original methylation status (**Fig. 3-2, 3-3, 3-4**). Furthermore, even in the HMD-flanking sequences of which the original sites are highly methylated *in vivo* in both species (**Fig. 3-5**), DNA methylation was limited. This lack of DNA methylation could be a general phenomenon for exogenously introduced DNA sequences, because the same result was obtained with the blastula cells of transgenic line (zicTg) which was established long time ago in my laboratory (**Fig. 3-6**) and with other cells (liver cells) of my transgenic fish and zicTg (**Fig. 3-7, 3-8**). This is a sharp contrast with previous results with mouse stem cells (Lienert et al., 2011). Given that DNA fragments to be injected were methylation-free during preparation of DNA constructs, my present results imply that *de novo* methylation fails to target exogenous DNA fragments in medaka.

Why do introduced DNA fragments maintain the hypomethylated status *in vivo*? One possibility is that the exogenous sequence included in the constructs (*β-actin* promoter and GFP) may affect DNA methylation status of nearby regions. Another possibility would be that exogenous DNAs, once introduced, are marked by some unknown tags, which specifically protect them from *de novo* methylation. Integration sites could also affect the efficiency of *de novo* methylation.

The lack of or limited global demethylation in fish may need to be considered

in interpreting my present results. As described in Introduction, during early embryogenesis, mammals are known to experience global demethylation and the subsequent *de novo* methylation (for review, see Wu and Zhang, 2010). In contrast, in some organisms, the absence of global demethylation process has been suggested (Macleod et al., 1999; Veenstra and Wolffe, 2001). This is also the case for medaka (Walter et al., 2002). Although dynamics of DNA methylation still remains largely elusive in medaka, recent studies in zebrafish showed that the methylation pattern of sperm is inherited to embryonic cells, while the oocyte methylome is reprogrammed to a pattern similar to that of sperm after fertilization (Jiang et al., 2013; Potok et al., 2013). Like zebrafish, the methylation status of medaka blastula cells seems highly similar to that of sperm (at least as for HMDs, > 95% of each stage's HMDs was commonly seen between blastula cells and sperm in my analysis), suggesting that medaka adopts the zebrafish-type methylation process, rather than mammalian-type. There is a possibility that such fundamental difference may be related to the difference in DNA methylation to exogenous sequences between medaka and mammals in part, but at the moment I do not have evidence which discriminates these possibilities and it still remains to be addressed.

It is, however, worth noting that DNA methylation seemed to occur at some sites, in a part of my transgenic fish (For example, see HNI–type reads (right) in **Fig. 3-2**). Thus, in spite of relatively loose methylation situation in medaka, the wave of *de novo* methylation seems to exist. Under such situation, the conservation and enrichment of the identified 6-mer could contribute to HMD formation and/or maintenance. Anyway, further studies will be required to elucidate the factors that cause the difference in methylation pattern seen in transgenes between mammals and medaka.

General Discussion

In my doctoral thesis, I compared DNA hypomethylated domains (HMDs) in the two medaka inbred lines, Hd-rR and HNI, which are established from the two closely related species, *Oryzias latipes* and *Oryzias sakaizumii*, respectively. I demonstrated that the majority of HMDs in blastula cells are shared by Hd-rR and HNI, but that a small portion of HMDs only exist in one species (species-specific HMDs). Genes in or nearby species-specific HMDs tend to show species-specific expression levels and thus are expected to contribute to the species-specific characters in these medaka species (Ishikawa et al., 1999; Kimura et al., 2007; Tsuboko et al., 2014). The studies identifying the differentially methylated regions in inbred lines or closely related species have been very limited, probably due to the lack of high quality genome and genome-wide base-resolution methylomes. In this context, these species-specific HMDs identified in medaka in my study may be interesting targets for the future research of DNA methylation-based phenotypic differences.

Hd-rR and HNI are known to show high incidence of genetic variations (Kasahara et al., 2007). This was confirmed in my study, and furthermore, I revealed that even HMDs shared by the two species accumulate genetic variations at similar rates to other methylated regions. At first glance, this finding was curious but my subsequent analysis identified some short sequences which are highly conserved in HMDs under such high genetic variations. Furthermore, I found that some of the highly-conserved 6-mers (showing low mutation index rate (common / Hd-rR specific)) reside in the nucleosome linker region within HMDs. The downstream region of TSSs is known to have highly-phased nucleosome arrays, but the mechanism of nucleosome positioning is

still controversial. The dependence of positioning on the intrinsic sequences varies among organisms. In this context, my results in Chapter 2 should support the existence of a DNA-guided mechanism in medaka, in that these identified motifs are suggested to function in nucleosome positioning. Notably, previous studies demonstrated that the nucleosome core is known to disfavor AT-rich sequences such as poly (dA:dT), but in my study, such bias was not observed in these motifs' base composition. Furthermore, although these sequences showed no similarity to binding motifs of known TFs, most of them are suggested to function as motif. Therefore, they may compose an intrinsically disfavorable structure for nucleosome core positioning, or serve as unknown binding sites of TFs, thereby positioning nucleosome core stably in specific regions. Future studies focusing on these motifs will give us further novel insights into the mechanism of nucleosome positioning.

In Chapter 3, I made transgenic medaka carrying HMD sequences and found that DNA methylation failed to target introduced DNA fragments in medaka irrespective of the methylation status of its endogenous site. This revealed a clear difference from a previous report using mouse stem cells. However, my data are still limited at present and further studies are required to discuss *de novo* methylation in medaka.

Through my doctoral thesis, the medaka system was further recognized as a very attractive model for epigenetic research. Medaka has a big advantage such as the established inbred lines and high-quality draft genomes. As described above, the comparison of DNA methylation patterns and genomic sequences between the two medaka inbred lines and the subsequent analyses in my study provided novel insights and interesting targets for future study. More than 10 inbred lines of medaka are currently maintained in Japan, including one from the Korean medaka (HSOK), and

efforts are being made to create additional strains from different regional populations, including close relatives (Takeda and Shimada, 2010). Although the resource in genomes and epigenetic modifications is not sufficient for other inbred lines for now, their comparative analysis will give us further insights into how genomic sequences are interpreted as the epigenetic code, and how such changes lead to changes in phenotypes.

Material and Methods

Fish strains

I used medaka Hd-rRII1 (referred to as Hd-rR), d-rR, and HNI-II (referred to as HNI). Medaka fishes were maintained and raised under standard condition. All experimental procedures and animal care were carried out according to the animal ethics committee of the University of Tokyo.

Identification of common HMDs and species-specific HMDs

First, I mapped the bisulfite-treated reads collected from of HNI blastula-stage embryos (Qu et al., 2012) to the HNI genome (version 2) which became available recently (http://mlab.cb.k.u-tokyo.ac.jp/~yoshimura/Medaka/#!Assembly.md), according to the mapping condition previously described (Qu et al., 2012). Then, based on the same criteria as the previous report from my laboratory (Nakamura et al., 2014), I identified the region containing at least 10 continuous low-methylated (methylation rate < 0.4) CpGs as HMDs in HNI blastula embryos.

Next, I mapped the identified HMD sequences of each species to the genome of the other species using BLAT (tileSize=18, oneOff=1) (Kent, 2002), as the mapping with such parameters are compatible with both high sensitivity (> 99.9% are expected, data not shown) and short calculation time. Among the outputs, due to partial similarities, queries were sometimes mapped to much longer genomic regions. A majority of such cases seemed mapping errors, because insertion or deletion events of > 2 kb regions were rare in the regions which were reliably aligned between the two species. Thus, in order to obtain reliable comparison, I did not include the outputs for further analysis in which the mapped region's length is > 2 kb longer than that of query

HMD. After removing these outputs, I further isolated query sequences (hypomethylated sequences in Hd-rR or HNI) which were uniquely mapped or multiply mapped to other genomic regions. I set a criterion that 80% of query's sequences were aligned in the other species' genome. This criterion excluded 1% (Hd-rR mapped to HNI) or 4% (HNI to Hd-rR) of uniquely mapped pairs and 92% (both cases) of multiple mapped pairs. To further isolate reliable pairs from the remaining multiply mapped outputs, I extracted pair as reliable ones of which the best matching rate of such pair (the ratio of the number of the base matches to the whole query size) was > 50% higher than that of any other pairing.

Subsequently, from the remaining results, I selected those in which the mapped genomic region of the query HMD was unique and was not covered by any other query HMDs. Last, I extracted the mapping results in which the query HMD was anchored to the same chromosome.

Next, with the remaining results, I checked if each HMD of the target genome overlapped with the mapped region of the query HMD. If the test was negative, I regarded that such an HMD had no corresponding HMD in the other species and identified it as a 'species-specific HMD'; otherwise, I treated it as a 'common HMD' that is shared in common in both species. Since > 94% of the common HMDs which were identified from the mapping of Hd-rR HMDs to HNI genome overlapped with the common HMDs which were identified from the mapping of HNI HMDs to Hd-rR genome, I used the former set as 'common HMDs' in all analyses.

RNA-seq

For d-rR blastula cells, the previously obtained data was used (Nakatani et al.,

2015). For HNI blastula cells, RNA was isolated using ISOGEN (Nippon Gene) and RNeasy mini kit (QIAGEN) and treated with Ribominus eukaryote kit for RNA-seq (Life Technologies). RNA-seq library was prepared using TruSeq RNA-seq sample prep kit (Illumina). The PCR products were purified and size fractionated using a bead-mediated method (AMPure, Ambion). Sequencing was conducted on HiSeq 2500 platform (Illumina). Sequences were mapped using BWA (Burrows-Wheeler Alignment tool) (Li and Durbin, 2009) and RPKM (reads per kilobase of exon per million mapped reads) was calculated using SAMMATE software (Xu et al., 2011).

Calculation of the incidence of genetic variations between Hd-rR and HNI

I categorized the HMD sequences into three HMD groups, 'common HMDs', 'Hd-rR specific HMDs' and 'HNI specific HMDs' and similarly classified the corresponding regions on the other species' genome, and performed the alignment of reciprocally best matching pairs of sequences with LASTZ (Harris, 2007) (--format=axt) for each group. As the LASTZ sometimes produced multiple outputs with different size for the same region or outputs that partially overlapped with each other, I removed the relatively short outputs such that the whole aligned region of the query was covered by the longest or second-longest alignments for the same query, then extracted the alignments that were independent and did not overlap with each other. The sequences of gene exons were also excluded from the further analyses. Then, from the remaining output alignments, I counted the single nucleotide polymorphisms (SNPs), insertions and deletions. As a small portion of the mapped regions of common HMDs was methylated in HNI genome (~ 10% of all mapped regions), I excluded such methylated regions from further analysis of common HMDs. When the alignment of

one HMD was separated into more than one block, the mutations of the separated alignments were summed. Then, the mutation rate for each HMD was calculated by dividing the total number of mutations by the length (bp) of the investigated region. For the negative control data set, the original Hd-rR HMD genome-coordinate set was randomly distributed on methylated regions using bedtools (ver. 2.17.0) (Quinlan and Hall, 2010). Then, the obtained sequences of the methylated regions were treated as well as HMDs and used for the calculation of the incidence of genetic variations.

Calculation of 6-mer's mutation index

Using the output of LASTZ alignment, I examined whether a 6-mer is mutated or not by searching the query and the aligned regions for the 6-mer. To take into account the case that short indels occur within a 6 bp aligned region, but 6-mer is still conserved between two medaka genomes in spite of such indels, I extracted the aligned regions flanked by the 8 bp with no-mismatch, and examined whether the 6-mer is conserved within the extracted regions. If the 6-mer in the query was not found in the aligned region, the 6-mer was regarded as 'mutated'. Then, the mutation index was calculated by dividing the number of 'mutated' 6-mers by the total number of the 6-mers in the query. The calculation results of the motif and its reverse complement were combined for each 6-mer.

Motif analyses

TOMTOM (Gupta et al., 2007) was used to search motifs similar to top 20 selected 6-mers. JASPAR Vertebrates and UniPROBE Mouse databases were used as target motifs. I set the significance threshold (q value < 0.1) in the selection of outputs.

Making transgenic medaka

For each three selected HMDs from each set of HMDs (common HMDs, Hd-rR specific HMD and HNI specific HMDs), I cloned the sequence of each HMD and its flanking region (~ 2 kb length from both HMD boundaries) from the genomic DNA of Hd-rR adult liver or HNI adult liver with Phusion (NEW ENGRAND BioLabs). Then, I made the constructed in which each cloned sequence is preceded by *β-actin* promoter and followed by GFP coding sequence and flanked by I-SceI sites as shown in **Fig. 3-1**, with InFusion kit (Clontech). All the sequences of HMD and its flanking regions of the constructs were confirmed by sequencing. All primers for making constructs and sequence confirmation were listed on **Table 5** and **Table 6**.

I injected these constructs to d-rR embryos at 1-cell stage with I-SceI, and selected and raised the injected embryos with GFP-positive cells. I crossed each fish with d-rR adult fish and isolated and GFP-positive offspring, then raised them as F1 fish.

Bisulfite analysis of transgenic medaka

For most lines, I crossed F1 adult male fish and F1 adult female fish and extracted genomic DNA from about 30 - 100 offspring at blastula stage. For 1 line (**Fig. 3-4**, HMD-1), I extracted genomic DNA from F1 blastula embryos instead of F2 embryos for further procedures. I performed bisulfite treatment of the extracted genomic DNA using MethylEasy Xceed 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 for PCR with

bisulfite-converted DNA were listed on **Table 7**.

For analysis of liver cells, I extracted genomic DNA from liver of 3 - 4 F2 adult fish and performed bisulfite analysis and PCR as described above.

Methylation patterns of liver cells of Hd-rR and HNI adult fish

As well as blastula embryos, I mapped the bisulfite-treated reads collected from of liver cells of Hd-rR and HNI (Qu et al., 2012) to Hd-rR genome (http://www.ensembl.org) and the HNI genome (version 2) which became available recently (http://mlab.cb.k.u-tokyo.ac.jp/~yoshimura/Medaka/#!Assembly.md), respectively, according to the mapping condition previously described (Qu et al., 2012).

Statistical analysis and the data visualization

The statistical analysis and graph visualization were performed using R software (version 3.2.0). For the visualization of genome-wide data, we integrated the data into UTGB genome browser (Saito et al., 2009).

Data access

All sequence data are deposited at the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra) (accession number SRP070096).

Figures and Tables

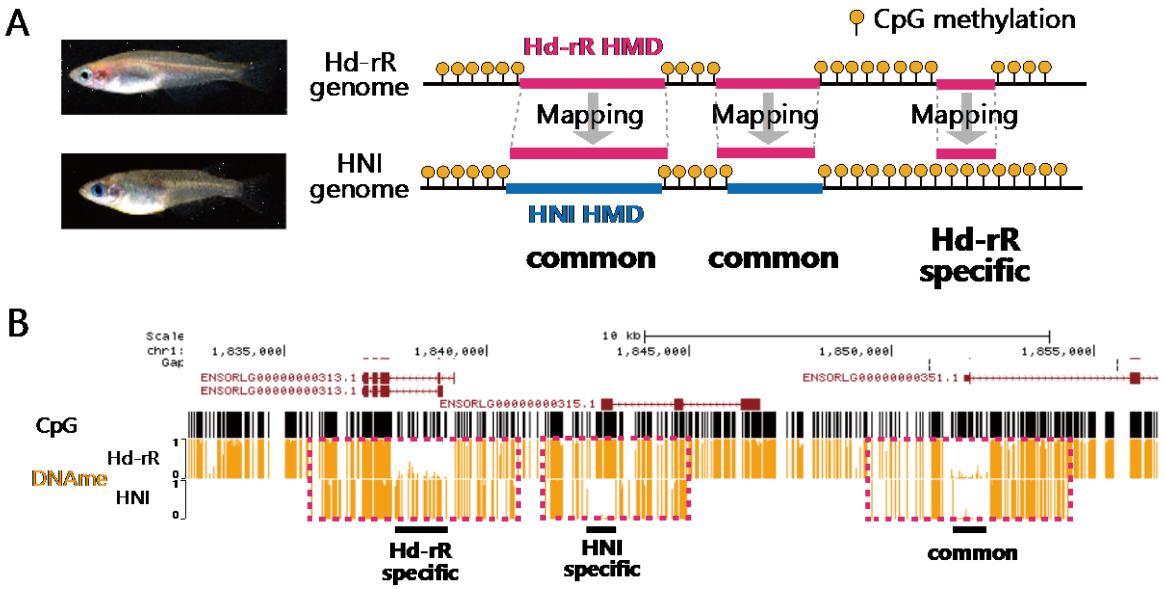


Figure 1-1. Identification of common HMD and species-specific HMD

A. Schematic representation of HMDs in aligned genomic regions. Hd-rR blastula HMDs were mapped to the genome of HNI to identify common HMDs and Hd-rR specific HMDs. I also performed the mapping of HNI HMDs to Hd-rR genome for identifying HNI specific HMDs (not shown). Each picture shows male Hd-rR (upper) or HNI (lower) adult fish.

B. Genome browser view showing the example of common HMDs, Hd-rR specific HMDs and HNI specific HMDs. The distribution of CpG is shown in black vertical lines and the methylation level is shown in orange ones. Black horizontal bars indicate the position of each HMD. The sequences of each HMD and its 2 kb flanking regions were mapped to the other species' genome, and the methylation status of the two species was compared in aligned regions (red-dotted boxes).

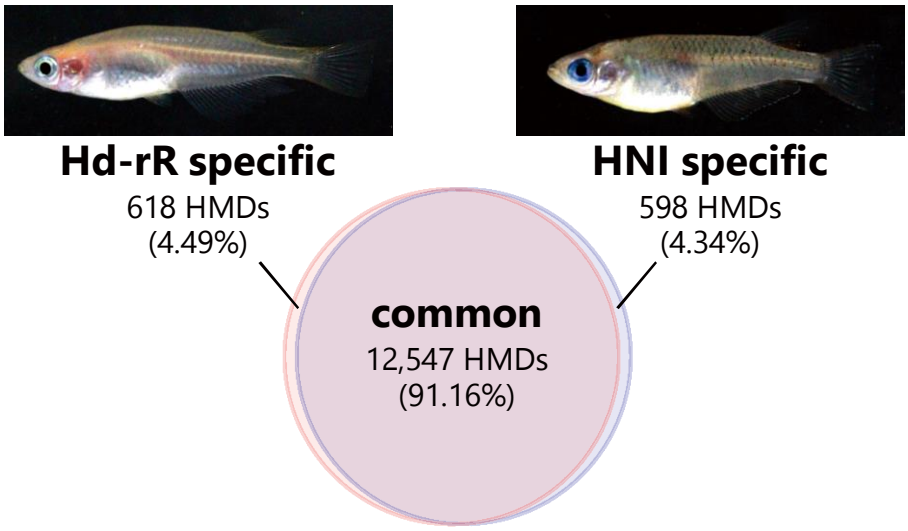


Figure 1-2. Proportion of common HMDs and species-specific HMDs

Venn diagram showing the overlap of HMDs between Hd-rR and HNI. Each picture above the diagram shows male Hd-rR or HNI adult fish.

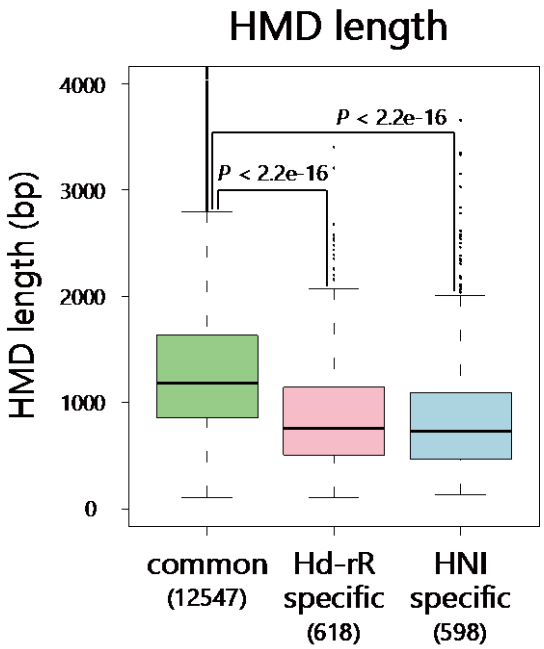


Figure 1-3. The size of each HMD

Boxplots showing the length of each HMD in common HMDs, Hd-rR specific HMDs and HNI specific HMDs. P-values were calculated using Wilcoxon rank sum test. In the box plots, the bottom and top of the boxes correspond to the 25th and 75th percentiles; 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 ranges of the lower and upper quartiles, respectively.

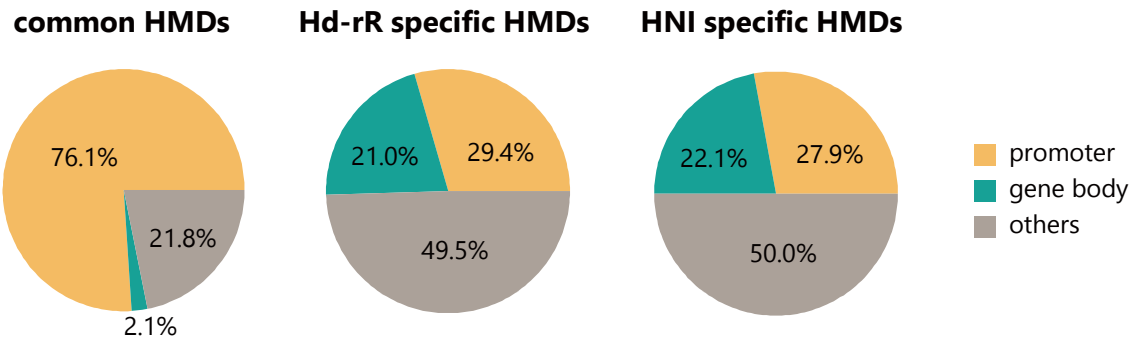


Figure 1-4. The positions of each HMD sets

Pie charts showing the proportion of HMD type (promoter (orange), gene body (green) and others (gray)) in common HMDs (left), Hd-rR specific HMDs (middle) and HNI specific HMDs (right). For categories, see text.

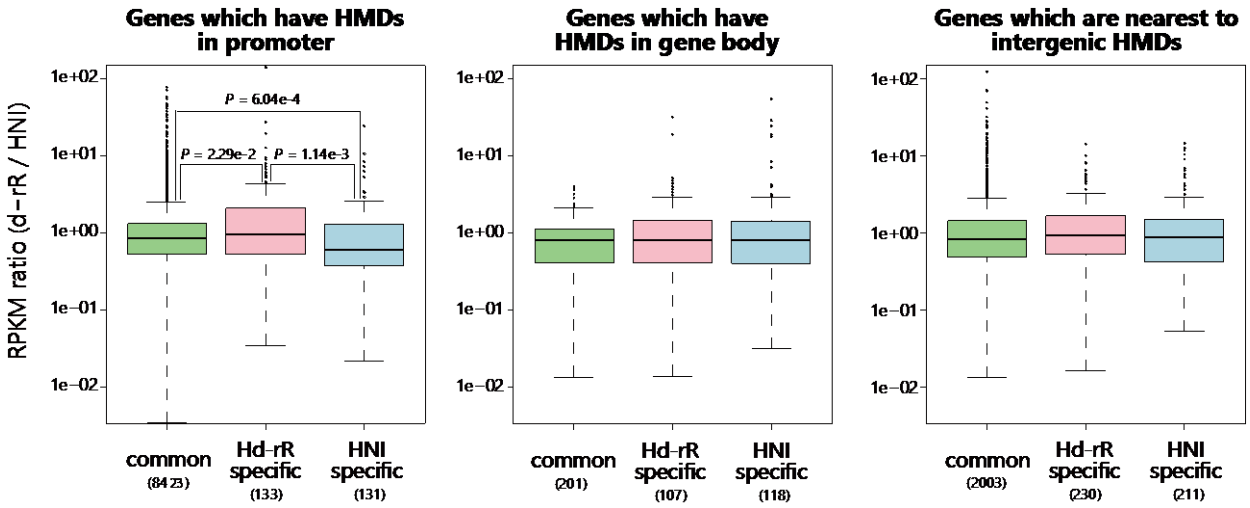


Figure 1-5. Relative expression level of the genes marked by HMDs

Boxplots showing the ratio of RPKM of d-rR to HNI (d-rR / HNI) of the genes marked by common HMDs (green), Hd-rR specific HMDs (pink) and HNI-specific HMDs (blue). Genes were classified according to the position marked by HMDs, promoters (left), gene bodies (middle) and intergenic regions (right). In the calculation of the ratio of RPKM, the genes in which RPKM of the either species is 0 are excluded. P-values were calculated using Wilcoxon rank sum test. In the box plots, the bottom and top of the boxes correspond to the 25th and 75th percentiles; 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 ranges of the lower and upper quartiles, respectively.

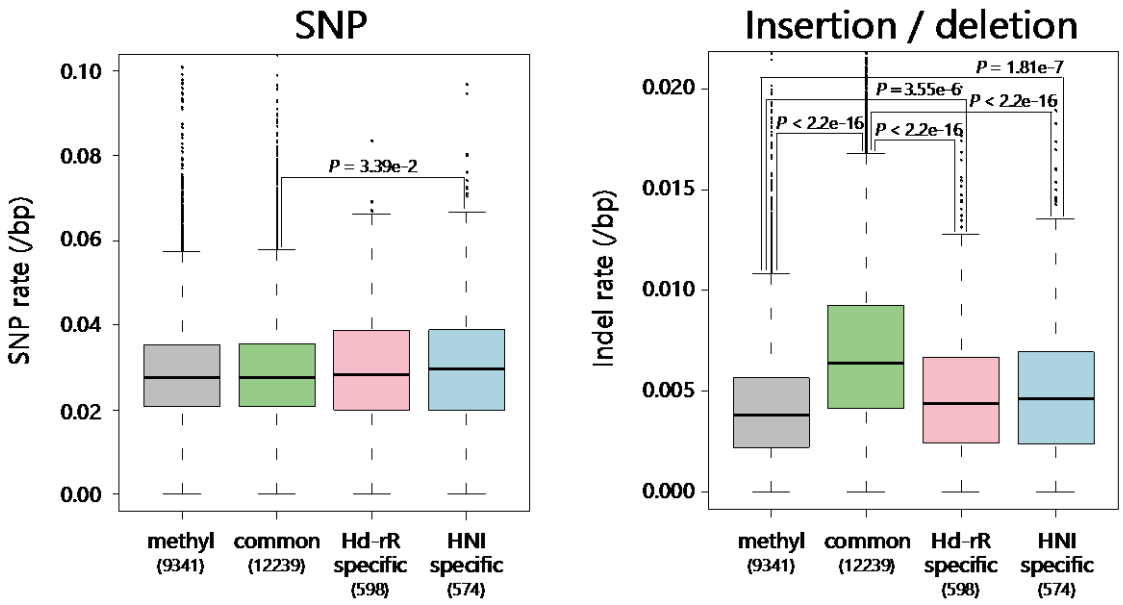


Figure 1-6. Genetic variations between Hd-rR and HNI in HMDs

Boxplots showing the incidence of genetic variations between Hd-rR and HNI. The left figure shows the rate of single nucleotide polymorphisms (SNPs) per base pair, and the right one shows the rate of insertions and deletions per base pair in the methylated regions (gray), common HMDs (green), Hd-rR specific HMDs (pink) and HNI specific HMDs (blue). Note that exons in the Hd-rR genome and their aligned regions in HNI genome were excluded in this analysis, because the proportions of those regions could vary among the investigated HMD set. P-values were calculated using Wilcoxon rank sum test. In the box plots, the bottom and top of the boxes correspond to the 25th and 75th percentiles; 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 ranges of the lower and upper quartiles, respectively.

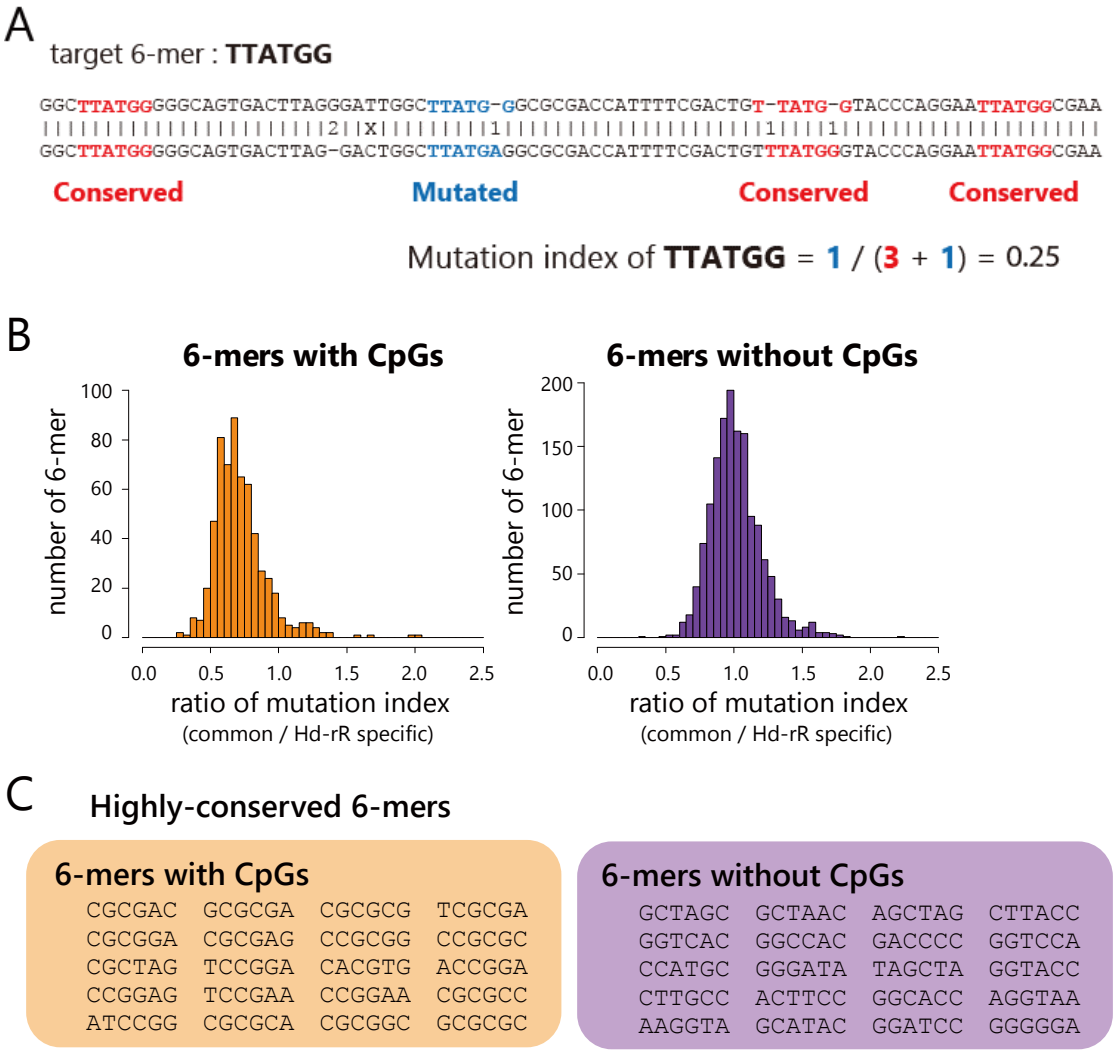


Figure 1-7. Identification of the conserved sequences within common HMDs

A. An example of calculation of mutation index. The target 6-mer TTATGG is found at four regions in the upper sequences of aligned sequences and is mutated at one of them (blue), so mutation index of TTATGG is 0.25 in this region.

B. Histograms showing the distributions of the ratio of mutation index (common HMDs / Hd-rR specific HMDs) for 6-mers with CpGs (left) and without CpGs (right).

C. Lists of top 20 most conserved 6-mers with CpGs (left) and without CpGs (right), which have the lowest values in the ratio of mutation index (common HMDs / Hd-rR specific HMDs).

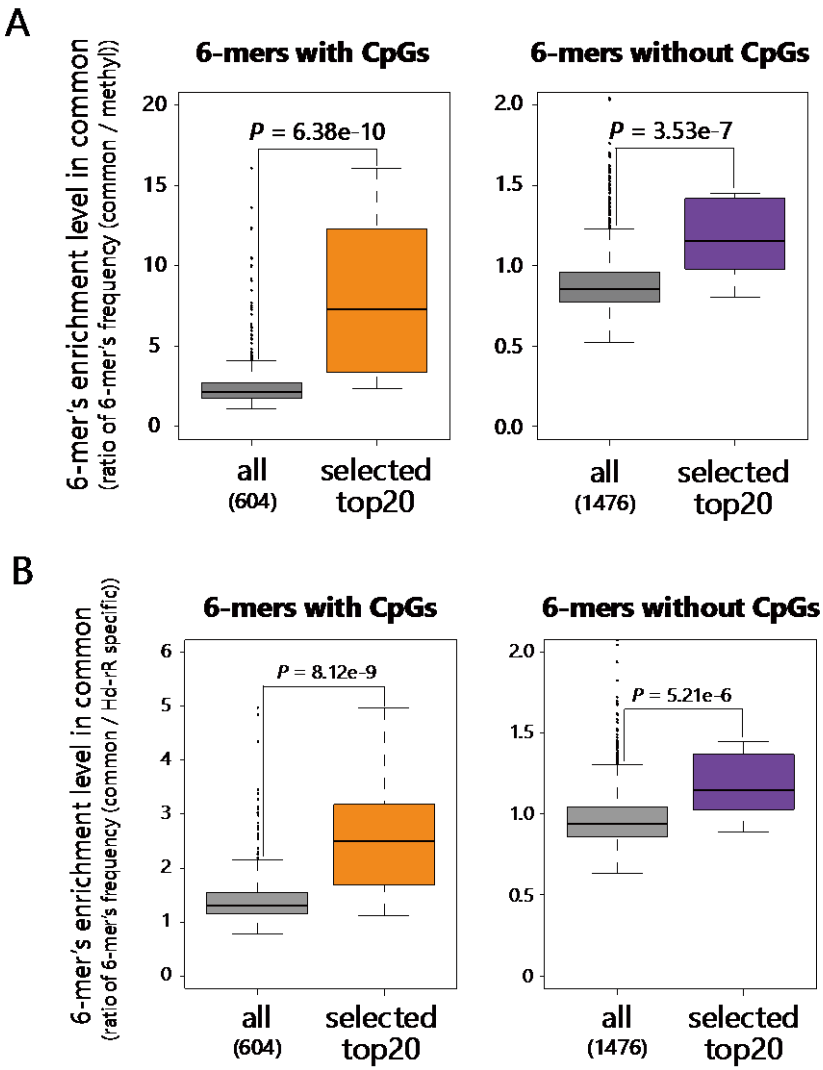


Figure 1-8. Enrichment levels of the conserved sequences within common HMDs

A. Boxplots showing the 6-mer's enrichment levels in common HMDs (the ratio of each 6-mer's frequency (common HMDs / methylated regions)).

B. Boxplots showing the 6-mer's enrichment levels in common HMDs (the ratio of each 6-mer's frequency (common HMDs / species-specific HMDs)).

Gray boxes represent the all 6-mers, while orange and purple boxes represent the top 20 most conserved 6-mers with and without CpGs, respectively. P-values were calculated using Wilcoxon rank sum test. In the box plots, the bottom and top of the boxes correspond to the 25th and 75th percentiles; 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 ranges of the lower and upper quartiles, respectively.

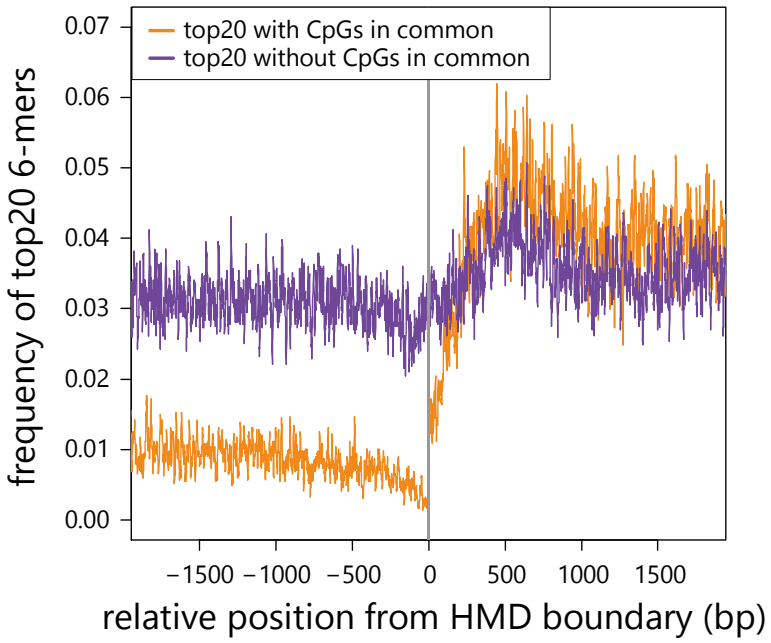


Figure 1-9. Distribution of the conserved sequences around HMD boundary
Distribution pattern of the top 20 most conserved 6-mers with CpGs (orange) or without CpGs (purple) in the 2 kb region around the boundary of the HMDs of which the size is > 2 kb. The boundaries of HMD were defined at the first low-methylated CpG site inside the HMD. X axis shows the length of each position from the HMD boundary. Downstream regions are hypomethylated.

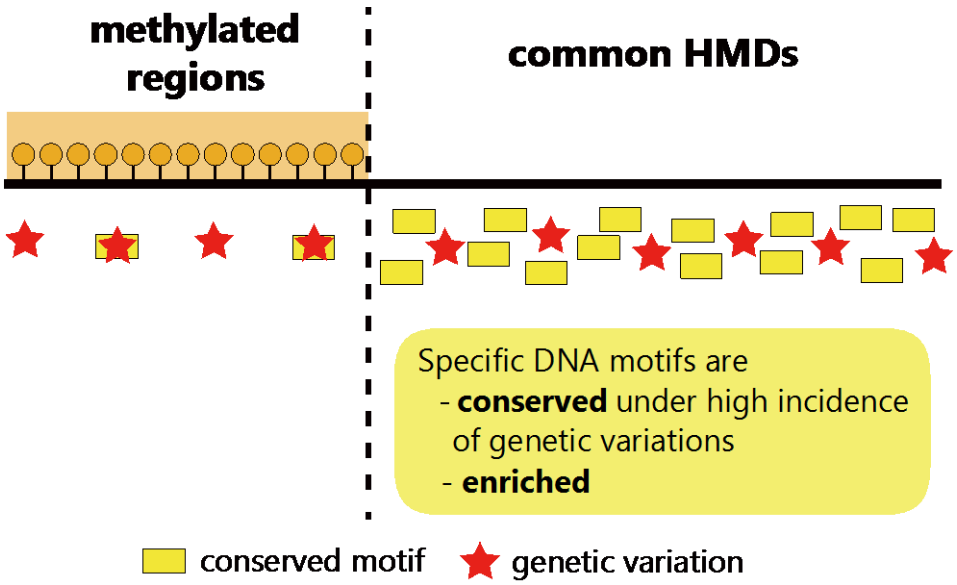


Figure 1-10. Schematic representation of conserved DNA motifs and genetic variations in common HMDs and the methylated regions

In common HMDs, specific DNA motifs (yellow rectangles) are conserved under high incidence of genetic variations (red stars). These motifs are enriched in common HMDs compared to methylated regions.

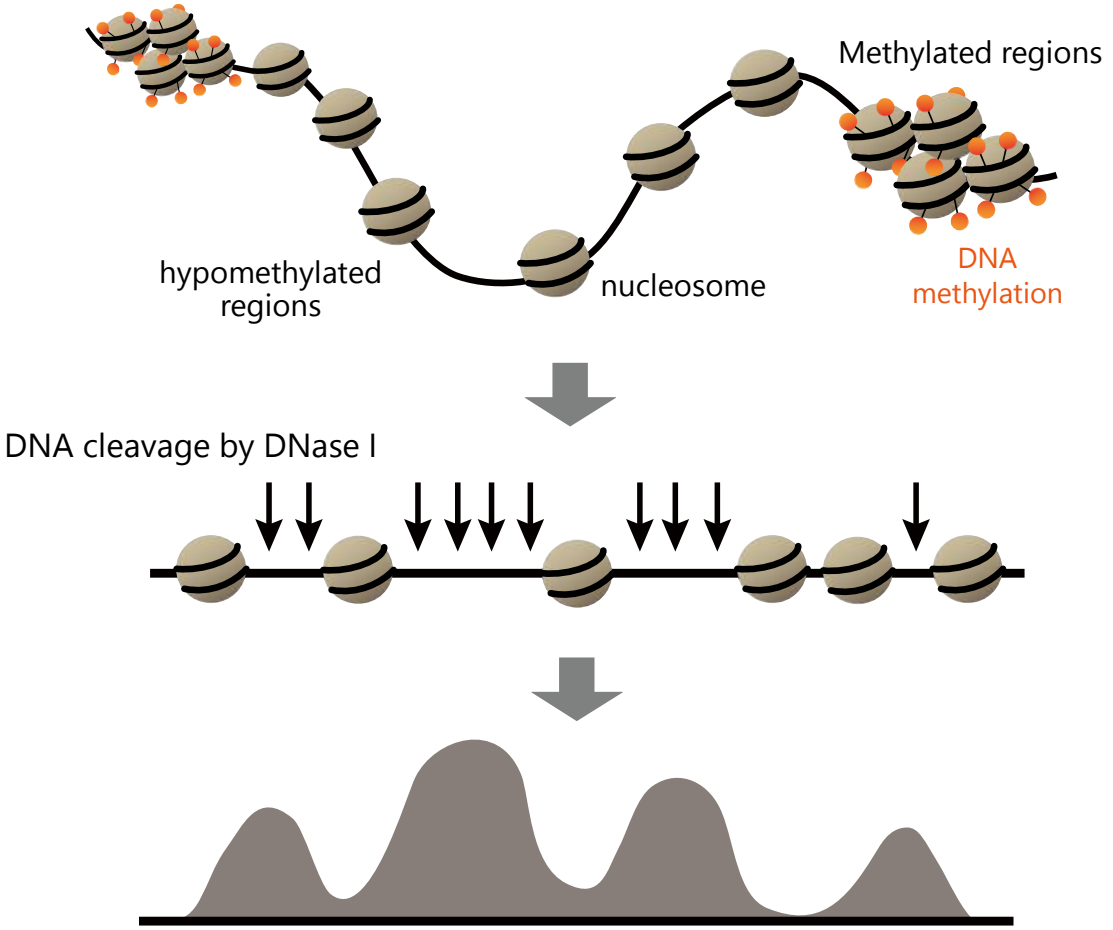


Figure 2-1. Basic principle of DNase-seq technique to reveal accessible chromatin

DNase I can digest accessible DNA which is depleted from nucleosome, thereby releasing DNA fragments. The high-throughput sequencing of them and the subsequent mapping of the reads to the genome can reveal accessible chromatin regions genome-wide.

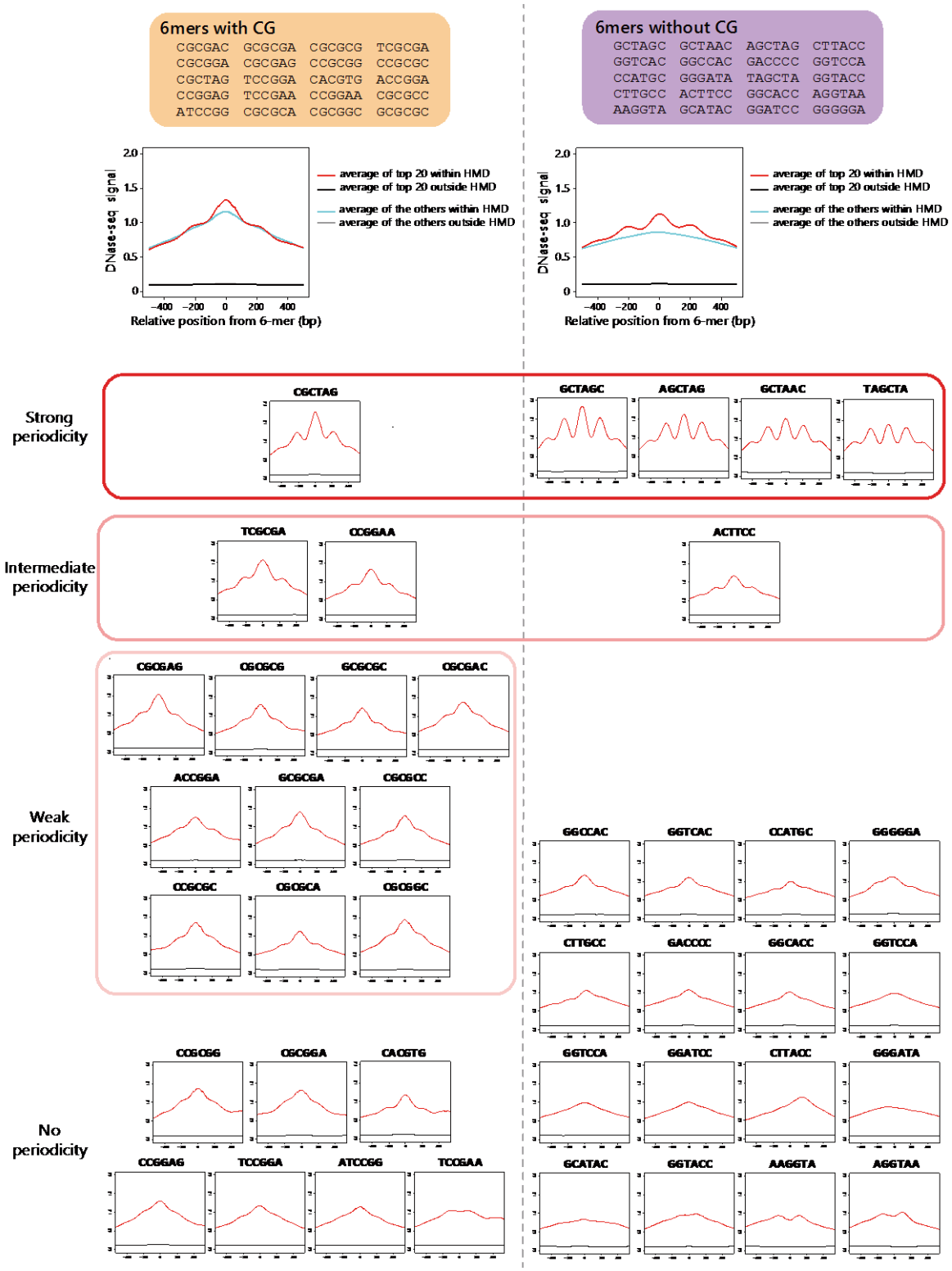


Figure 2-2. The distribution profile of DNase-seq signals around each selected top 20 6-mers with CpGs or without CpGs

The Upper two graphs show the average distribution profile of DNase-seq signal

around selected top 20 with CpGs (left) and without CpGs (right). The other graphs show distribution profile of DNase-seq signal around each 6-mer. Red and black line of each graph shows the signal within HMDs and outside HMDs, respectively. For categories, see text.

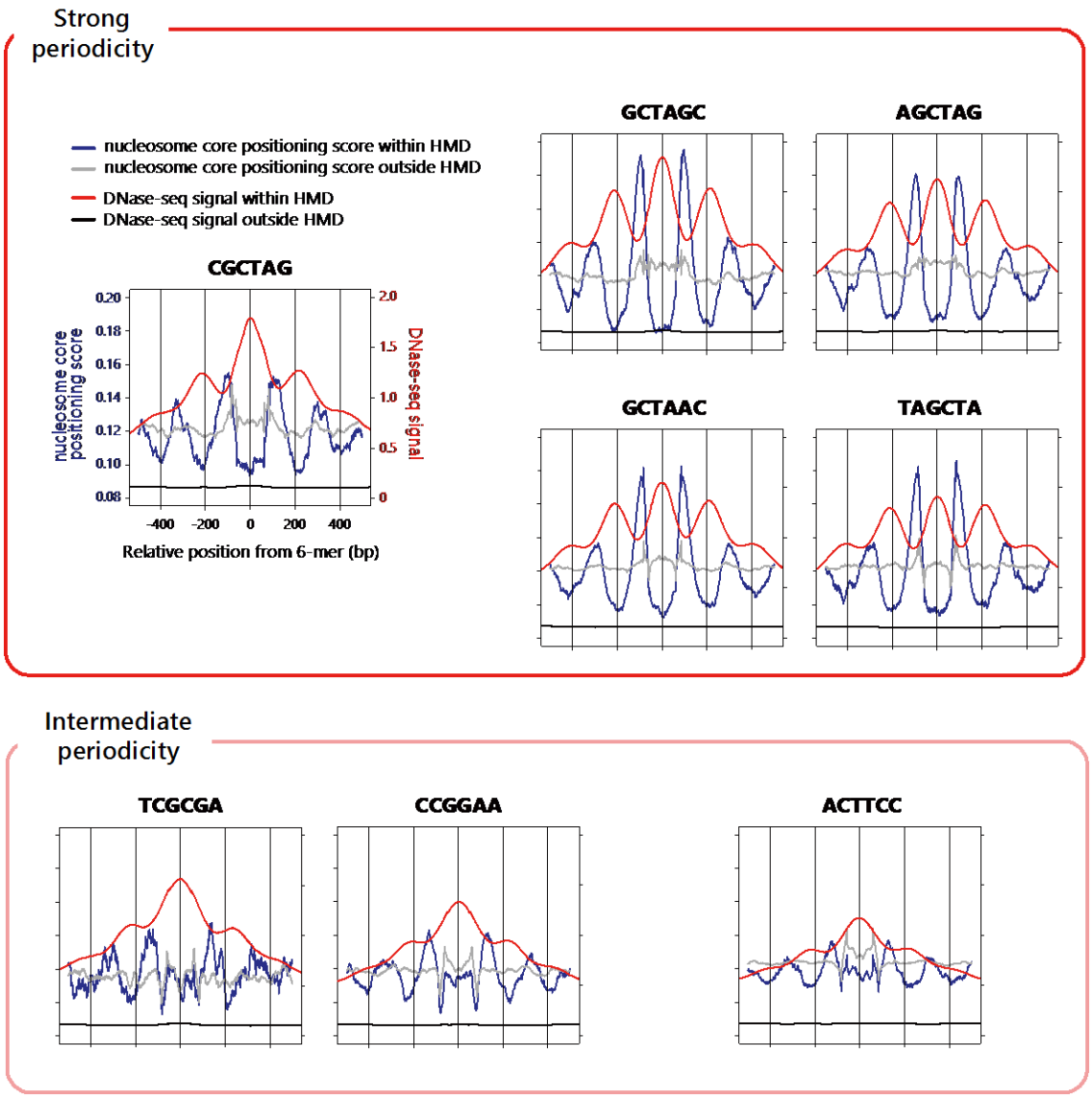


Figure 2-3. Nucleosome core positioning score and DNase-seq signal around the selected 6-mers with strong or intermediate periodicity of DNase-seq signal

The distributions of nucleosome core positioning score within HMDs (blue) or without HMDs (gray) and DNase-seq signal within HMDs (red) or without HMDs (black) around the selected 6-mers with strong or intermediate periodicity. For details, see text.

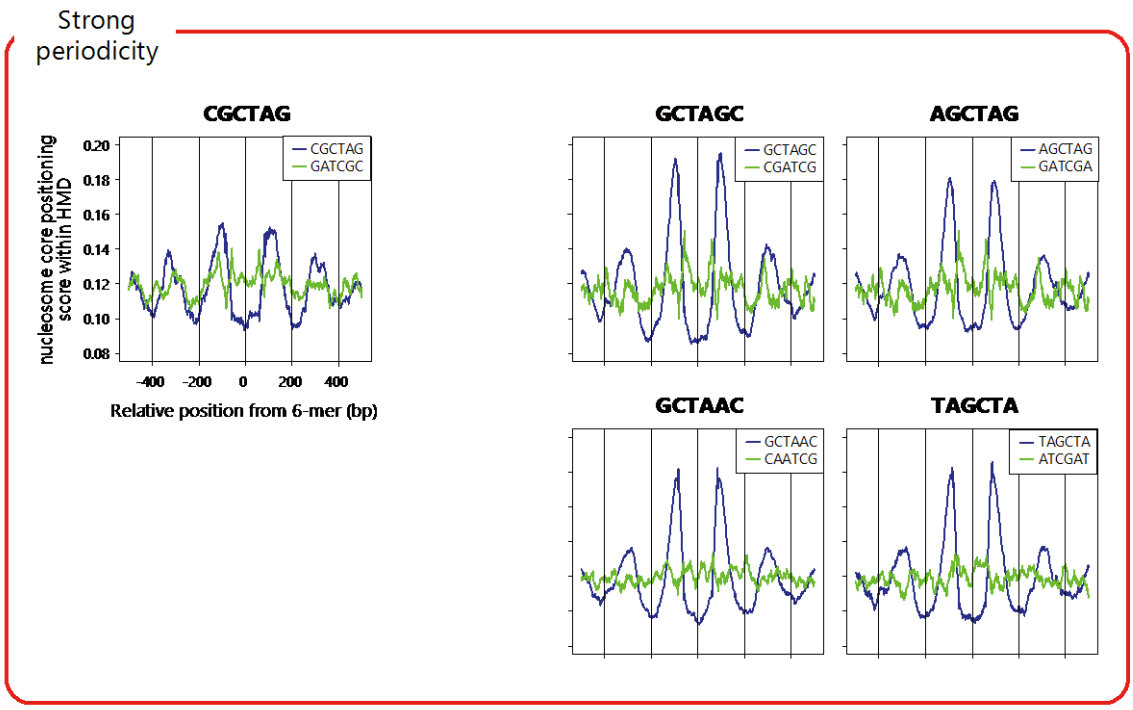


Figure 2-4. Nucleosome core positioning score around the selected 6-mer with strong periodicity of DNase-seq signal or its reverse 6-mer within HMDs

Blue line shows the 6-mer with periodicity and green line shows its reverse 6-mer. Reverse 6-mers, except for the case of CGCTAG, do not have periodic nucleosome core positioning score in their neighboring regions. For details, see text.

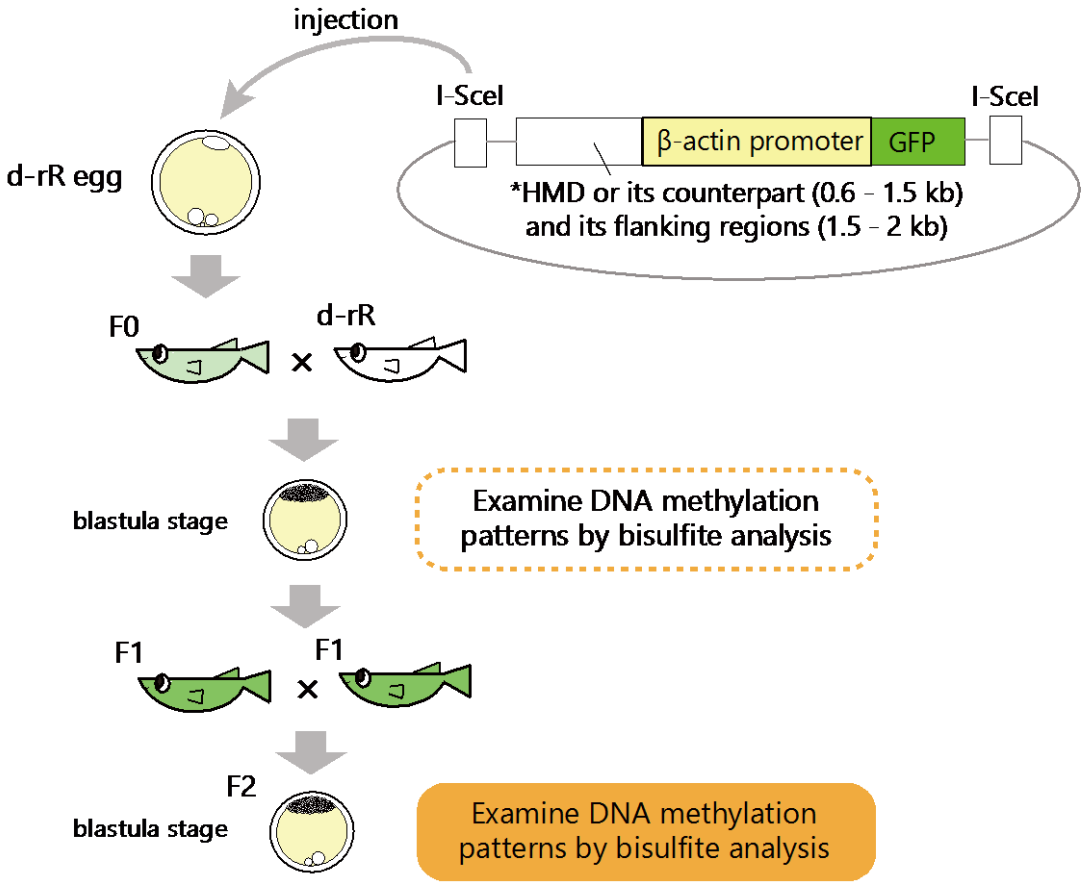


Figure 3-1. The overview of the experiment of bisulfite analysis with transgenic medaka

The constructs carrying HMD-sequence and its flanking regions were injected to d-rR 1 cell-stage embryos, and then GFP positive embryos were raised as F0. Adult F0 was crossed with d-rR, then, among the obtained embryos GFP positive embryos were raised as F1. The genomic DNA was extracted from F1 embryos or the offspring of F1 parents (F2 embryos) at blastula-stage and DNA methylation patterns of them were analyzed by bisulfite conversion and PCR.

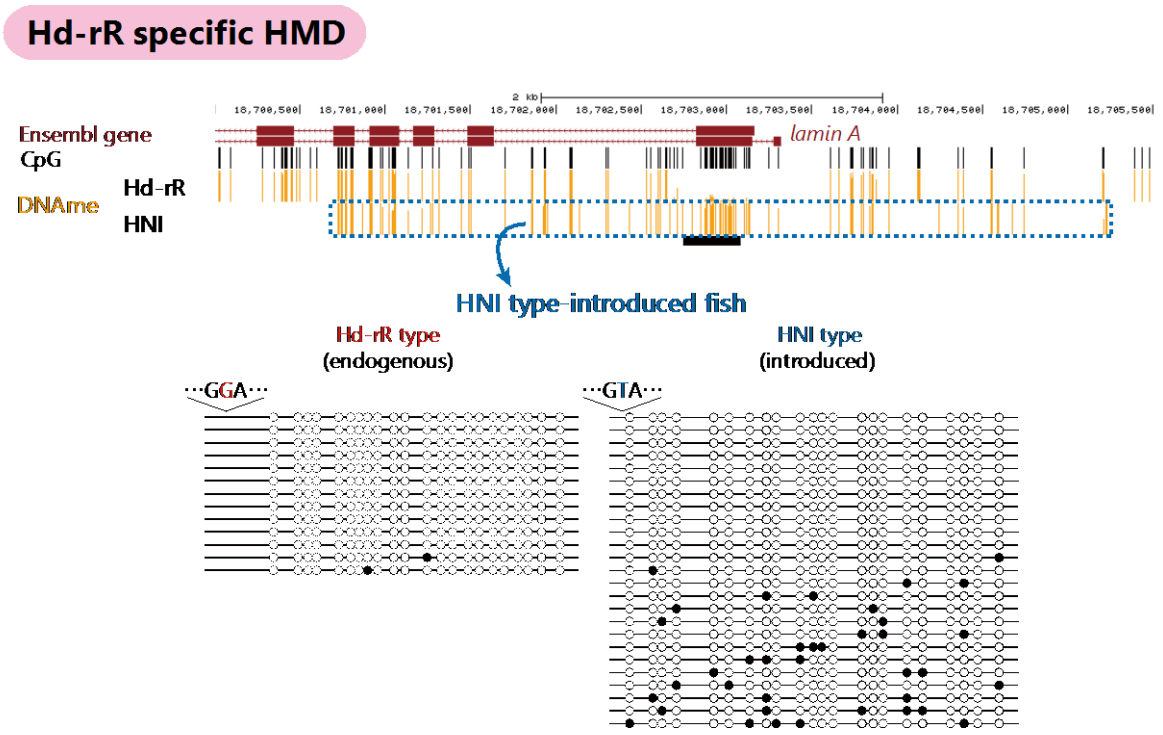


Figure 3-2. Bisulfite sequencing in F2 blastula embryos of transgenic medaka to which HNI-type sequence (methylated) of Hd-rR specific HMD is introduced

The upper figure is genome browser image showing the methylation pattern of blastula embryos of Hd-rR and HNI around the HMD. The sequence of the HMD and its 2 kb flanking regions was mapped to the other species' genome, and the methylation status of the two species was compared in aligned regions. The distribution of CpG is shown in black vertical lines and the methylation levels are shown in orange ones. A blue-dotted box shows the introduced region to the transgenic fish and a black horizontal bar shows the position of the amplified region from bisulfite-converted genomic DNA. In lower two figures, the positions of circle indicate the positions of CpG in each read. Unmethylated CpGs are shown as white circles and methylated CpGs are shown as black circles. The short sequences above each methylation patterns show an example of SNP between Hd-rR and HNI seen in the region.

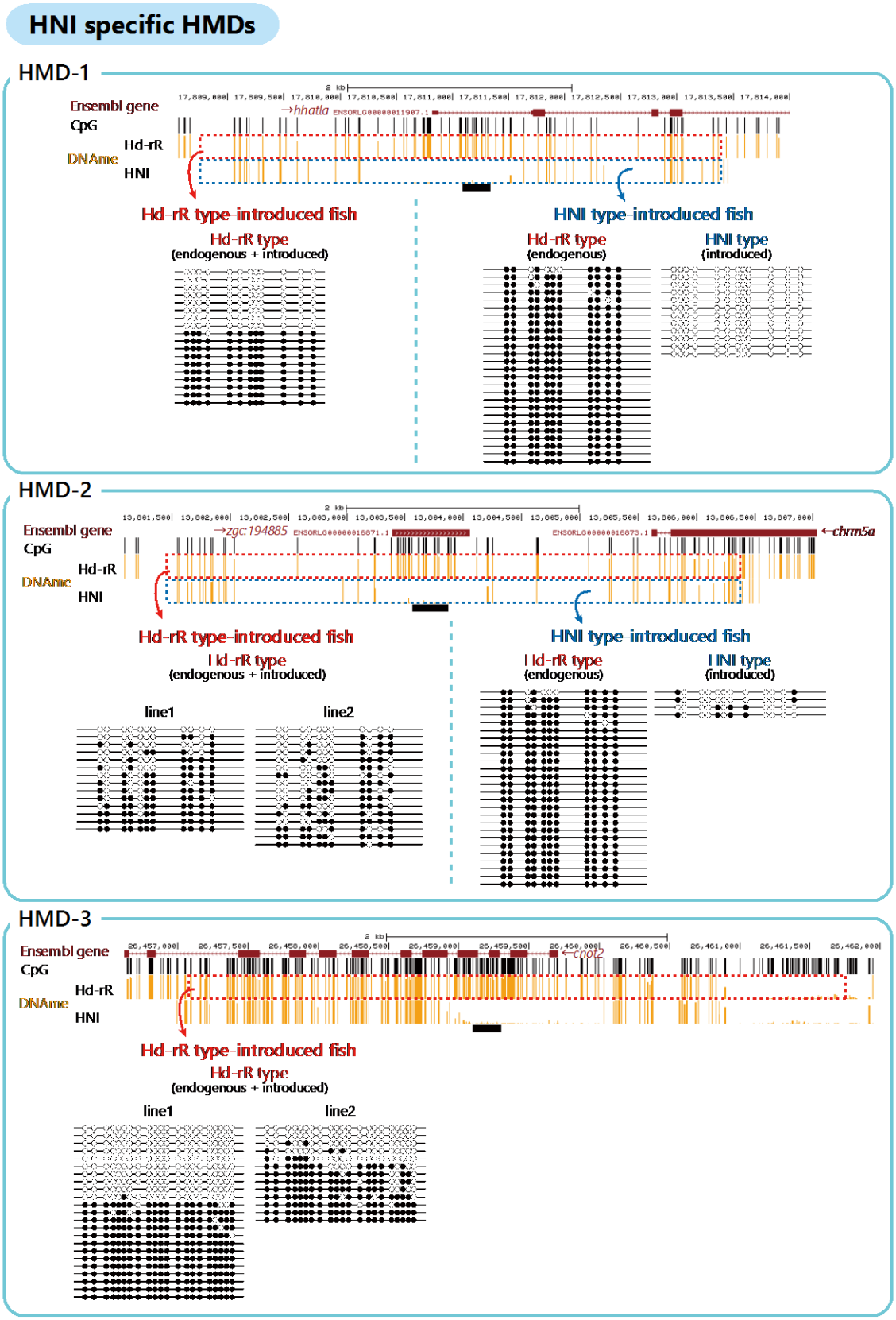


Figure 3-3. Bisulfite sequencing in F2 blastula embryos of transgenic medaka to which Hd-rR-type sequence (methylated) or HNI-type sequence (hypomethylated) of HNI specific HMD is introduced

Genome browser images show the methylation pattern of blastula embryos of Hd-rR and HNI around each HMD. The sequences of each HMD and its 2 kb flanking regions were mapped to the other species' genome, and the methylation status of the two species was compared in aligned regions. The distribution of CpG is shown in black vertical lines and the methylation levels are shown in orange ones. Red-dotted or blue-dotted boxes show the introduced region to the transgenic fish and black horizontal bars show the positions of the amplified regions from bisulfite-converted genomic DNA. The figures below each genome browser image show methylation status of the amplified regions. The positions of circle indicate the positions of CpG in each read. Unmethylated CpGs are shown as white circles and methylated CpGs are shown as black circles.

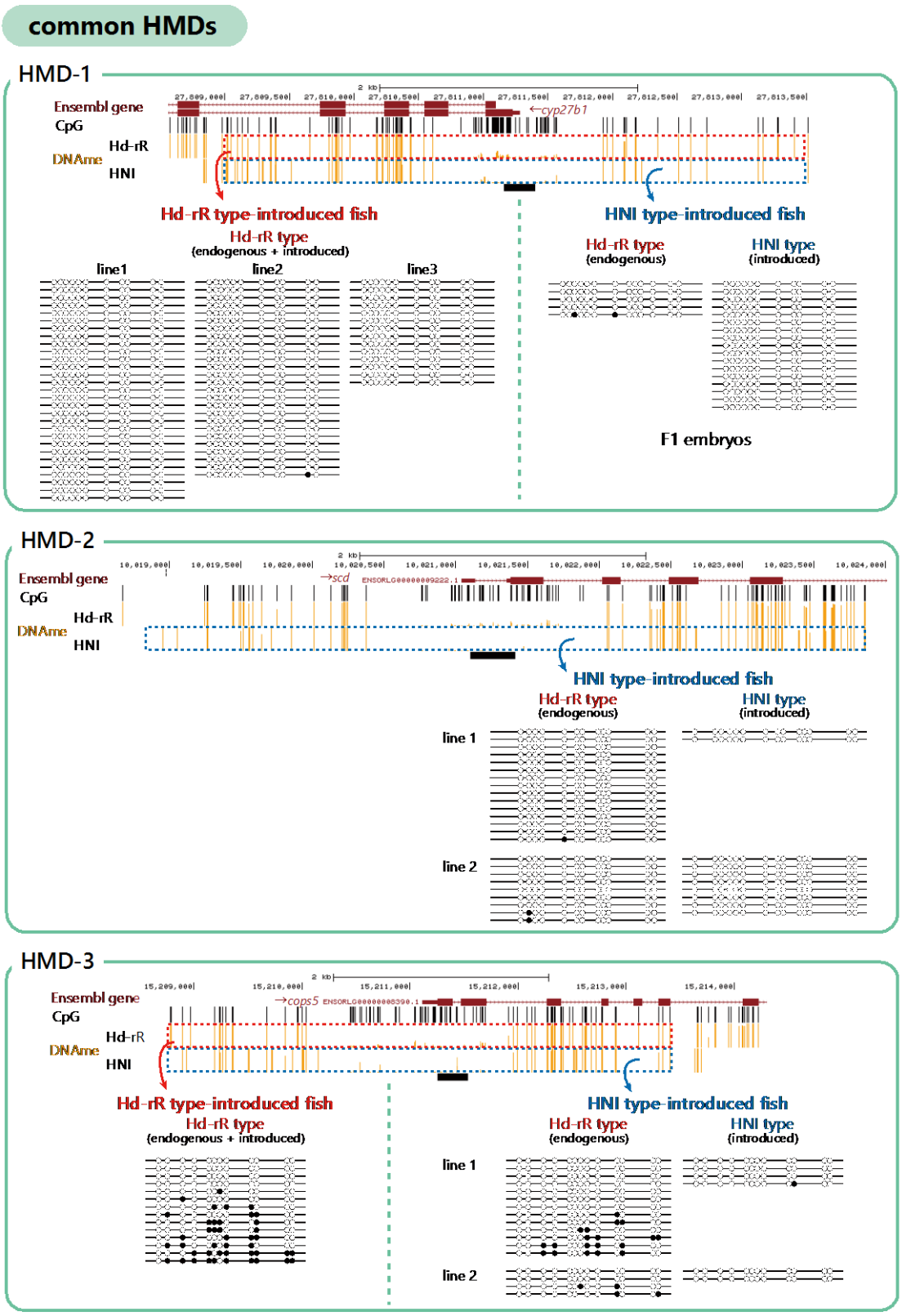


Figure 3-4. Bisulfite sequencing in F1 or F2 blastula embryos of transgenic medaka to which Hd-rR-type sequence (hypomethylated) or HNI-type sequence (hypomethylated) of common HMD is introduced

Genome browser images show the methylation pattern of blastula embryos of Hd-rR and HNI around each HMD. The sequences of each HMD and its 2 kb flanking regions were mapped to the other species' genome, and the methylation status of the two species was compared in aligned regions. The distribution of CpG is shown in black vertical lines and the methylation levels are shown in orange ones. Red-dotted or blue-dotted boxes show the introduced region to the transgenic fish and black horizontal bars show the positions of the amplified regions from bisulfite-converted genomic DNA. The figures below each genome browser image show methylation status of the amplified regions. The positions of circle indicate the positions of CpG in each read. Unmethylated CpGs are shown as white circles and methylated CpGs are shown as black circles.

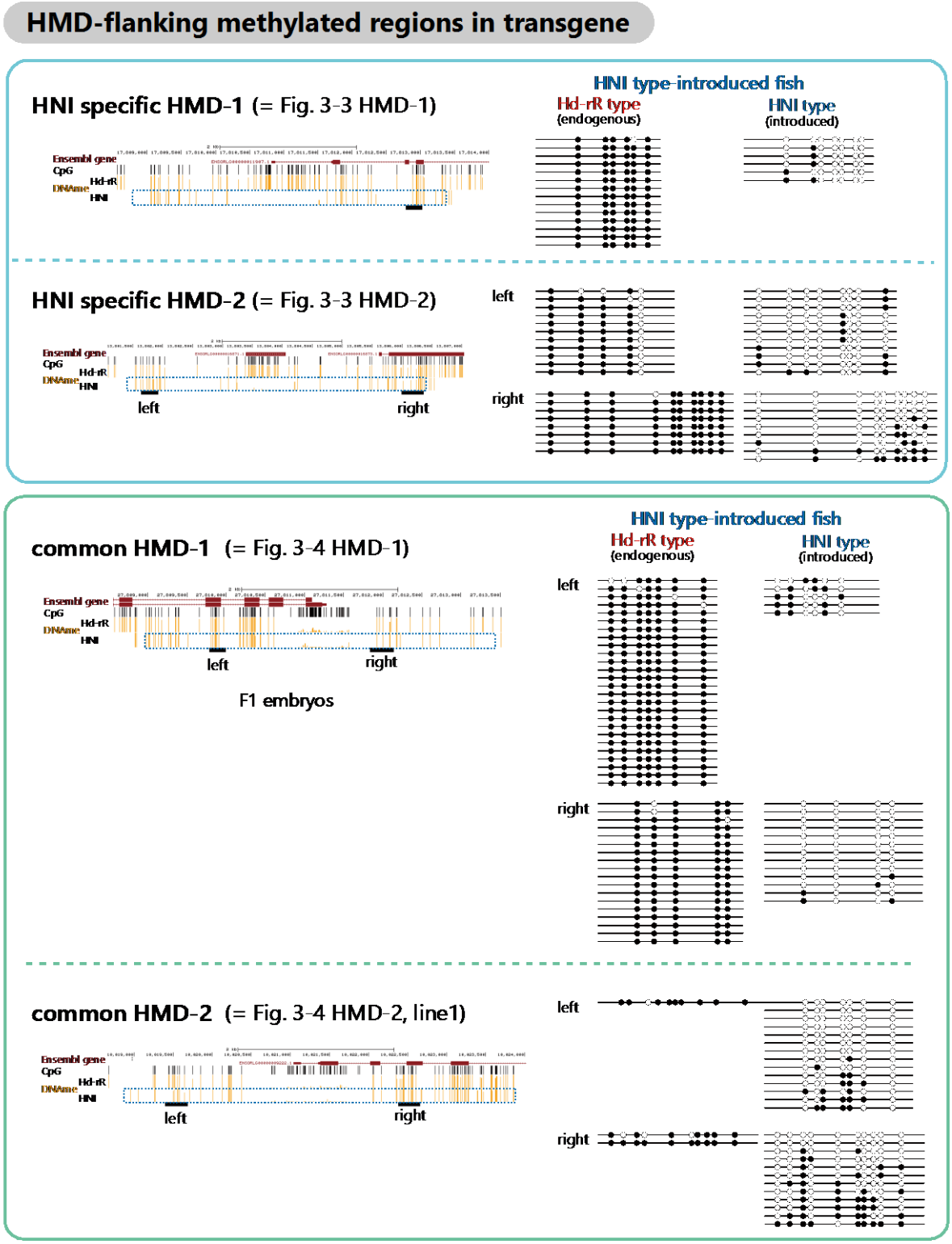


Figure 3-5. Bisulfite sequencing at HMD-flanking regions in F1 or F2 blastula embryos of transgenic medaka to which HNI-type sequence (hypomethylated) of HNI specific HMD or common HMD is introduced

Genome browser images show the methylation pattern of blastula embryos of

Hd-rR and HNI around each HMD. The sequences of each HMD and its 2 kb flanking regions were mapped to the other species' genome, and the methylation status of the two species was compared in aligned regions. The distribution of CpG is shown in black vertical lines and the methylation levels are shown in orange ones. Blue-dotted boxes show the introduced region to the transgenic fish and black horizontal bars show the positions of the amplified regions from bisulfite-converted genomic DNA. The figures below each genome browser image show methylation status of the amplified regions. The positions of circle indicate the positions of CpG in each read. Unmethylated CpGs are shown as white circles and methylated CpGs are shown as black circles.

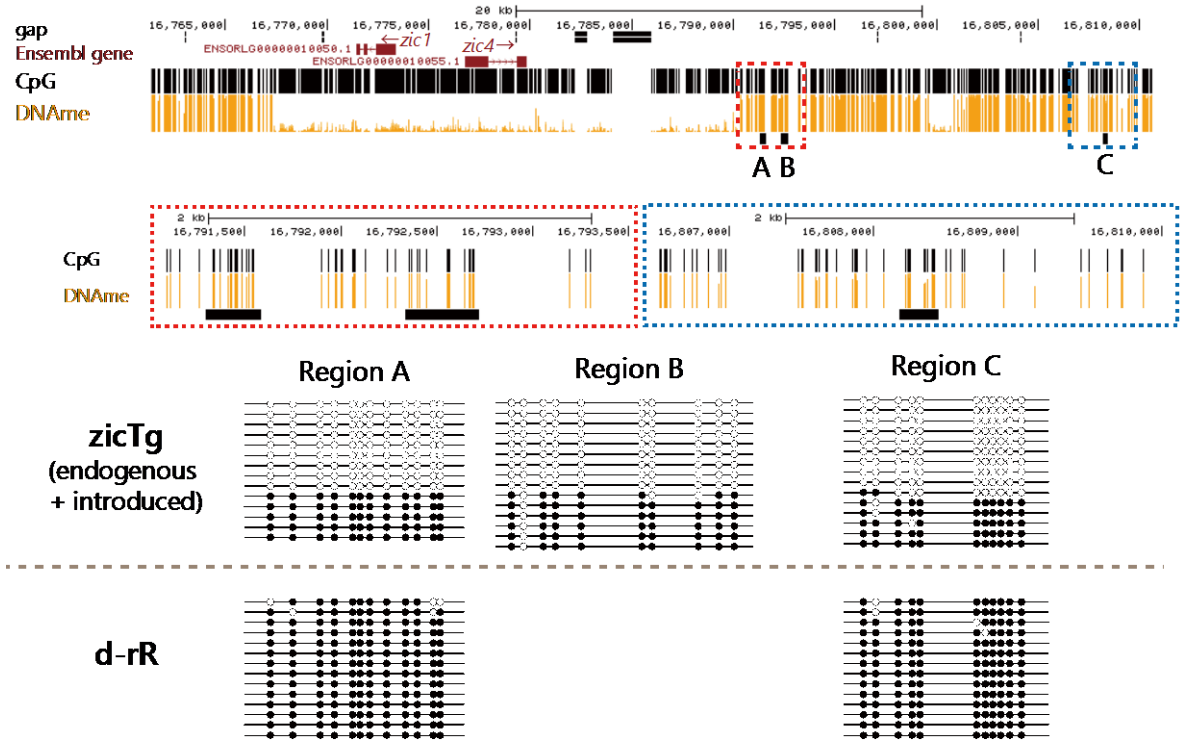


Figure 3-6. Bisulfite sequencing at methylated regions in blastula embryos of zicTg

Genome browser images show the methylation pattern of blastula embryos of Hd-rR around *zic1/4* genes. The distribution of CpG is shown in black vertical lines and the methylation levels are shown in orange ones. Black horizontal bars show the positions of the amplified regions from bisulfite-converted genomic DNA. The two magnified genome browser images show the same regions with those within red or blue-dotted boxes in the top image. The figures below the genome browser image show methylation status of the amplified regions. The positions of circle indicate the positions of CpG in each read. Unmethylated CpGs are shown as white circles and methylated CpGs are shown as black circles.

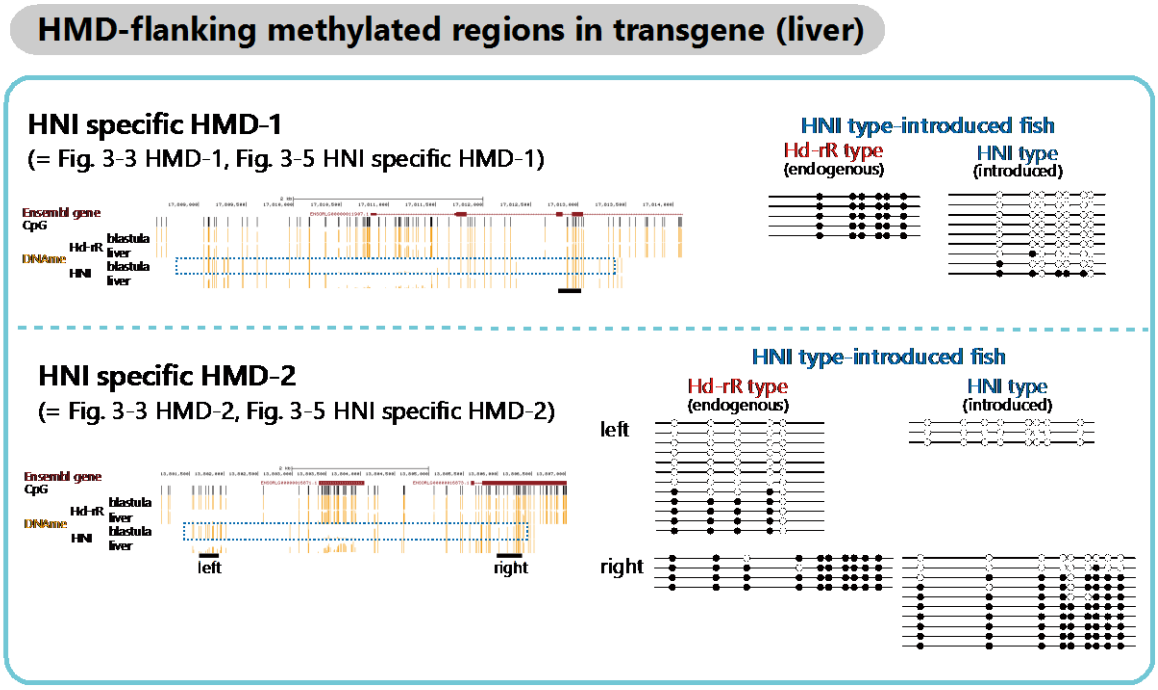


Figure 3-7. Bisulfite sequencing at HMD-flanking regions in liver cells of F2 transgenic medaka to which HNI-type sequence (hypomethylated) of HNI specific HMD is introduced

Genome browser images show the methylation pattern of blastula embryos and liver cells in Hd-rR and HNI around each HMD. The sequences of each HMD and its 2 kb flanking regions were mapped to the other species' genome, and the methylation status of the two species was compared in aligned regions. The distribution of CpG is shown in black vertical lines and the methylation levels are shown in orange ones. Blue-dotted boxes show the introduced region to the transgenic fish and black horizontal bars show the positions of the amplified regions from bisulfite-converted genomic DNA. The figures below each genome browser image show methylation status of the amplified regions. The positions of circle indicate the positions of CpG in each read. Unmethylated CpGs are shown as white circles and methylated CpGs are shown as black circles.

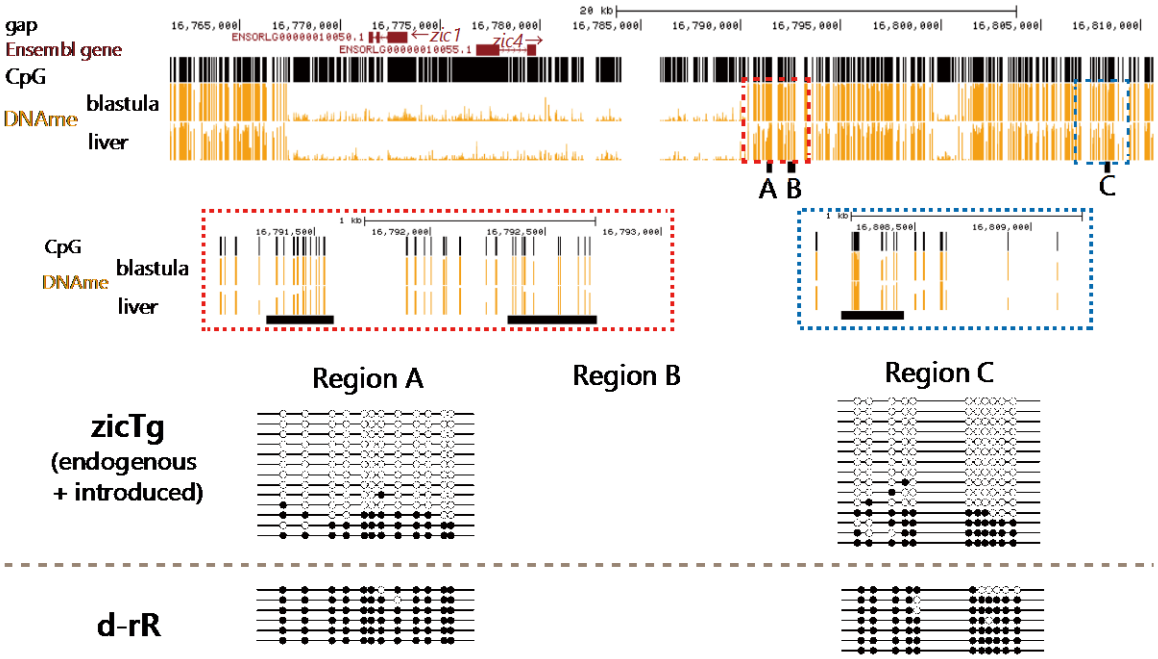
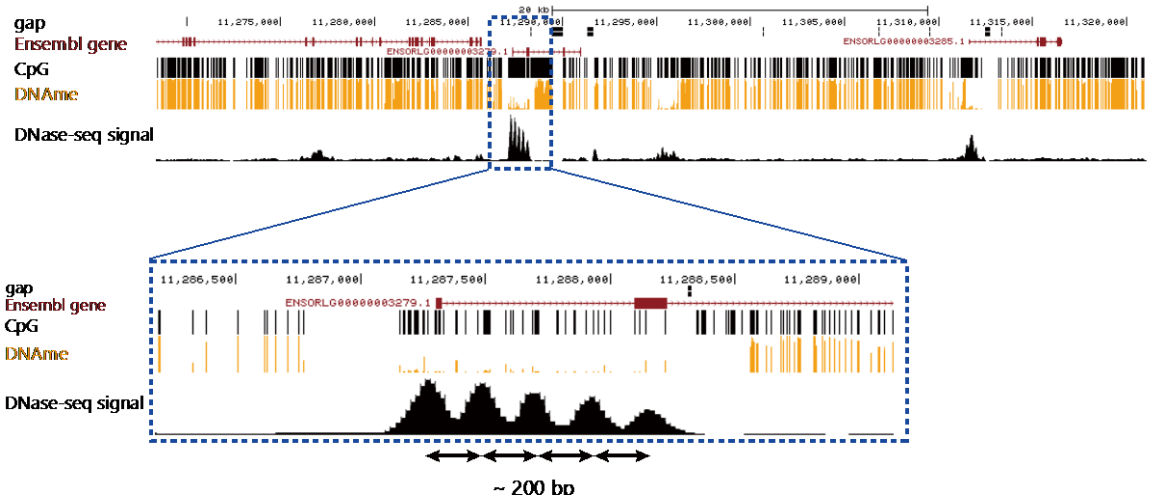


Figure 3-8. Bisulfite sequencing at methylated regions in liver cells of zicTg
Genome browser images show the methylation pattern of blastula embryos and liver cells in Hd-rR around *zic1/4* genes. The distribution of CpG is shown in black vertical lines and the methylation levels are shown in orange ones. Black horizontal bars show the positions of the amplified regions from bisulfite-converted genomic DNA. The two magnified genome browser images show the same regions with those within red or blue-dotted boxes in the top image. The figures below the genome browser image show methylation status of the amplified regions. The positions of circle indicate the positions of CpG in each read. Unmethylated CpGs are shown as white circles and methylated CpGs are shown as black circles.



Supplementary Figure S1. Genome browser view of the DNA methylation and DNase-seq signals.

The distribution of CpG is shown in black vertical lines, the methylation level is shown in orange ones and DNase-seq signal is shown in black. DNase-seq signal within the HMD shows the periodic pattern of peaks of approximately 200 bp intervals.

Gene ID	Gene name	RPKM in d-rR	RPKM in HNI
ENSORLG00000000045	NoName	8.0407	4.6834
ENSORLG00000000081	ptp4a1	359.0151	257.4057
ENSORLG00000000213	ARPC2(1of2)	0.5345	0.0000
ENSORLG00000000293	si:ch211-255i20.3	0.2021	0.7691
ENSORLG00000000313	lygl1	1.3207	0.0000
ENSORLG00000000335	slit3	1.0544	0.6687
ENSORLG00000000509	ptgs1(1of2)	0.8027	0.6108
ENSORLG00000000640	ints4	12.5449	14.8229
ENSORLG00000000741	NoName	1.7088	0.2167
ENSORLG00000000754	RHBDF2	36.4830	27.3777
ENSORLG00000000833	dus2	31.7360	50.7089
ENSORLG00000001169	NoName	4873.4900	728.2306
ENSORLG00000001183	si:ch73-56p18.4	20.3897	20.5034
ENSORLG00000001307	NoName	0.0000	1.2257
ENSORLG00000001585	mfsd10	51.5743	43.0059
ENSORLG00000001627	pcsk9	0.3819	1.0897
ENSORLG00000001697	KIF2A(1of2)	1.3342	5.0763
ENSORLG00000001769	slc30a8	33.6519	6.0972
ENSORLG00000001776	NoName	0.0000	0.0000
ENSORLG00000001861	pard6b	9.3640	20.8220
ENSORLG00000001880	dpm1	20.4857	30.2149
ENSORLG00000002073	slc22a15	11.9408	14.3375
ENSORLG00000002252	myt1b	1.8694	5.1026
ENSORLG00000002339	slc1a8a	8.8715	6.0684
ENSORLG00000002460	CLINT1(1of2)	13.4532	11.8540
ENSORLG00000002526	fermt3b	1.5570	0.8078
ENSORLG00000002622	ldb3a	0.1239	0.0000
ENSORLG00000002741	lrrc34	1.1900	0.6174
ENSORLG00000002748	TXK	0.3370	0.4274
ENSORLG00000002764	si:dkeyp-86f7.4	123.0693	132.9525
ENSORLG00000002766	sept4a	0.8262	0.0000
ENSORLG00000002806	ttc9c	4.1224	5.3778
ENSORLG00000003248	si:dkey-7e14.3	0.9927	2.8329
ENSORLG00000003303	SEPT9(1of2)	23.4424	53.9376
ENSORLG00000003649	NoName	0.1338	0.0000
ENSORLG00000003726	stard13a	0.4812	0.0000
ENSORLG00000003819	atic	59.2824	75.6765
ENSORLG00000003841	cabp2b	0.0000	0.0000
ENSORLG00000003911	taf2	13.4743	20.1411
ENSORLG00000003959	HRH2(1of2)	0.6422	1.2217
ENSORLG00000004413	gckr	0.3343	0.6360
ENSORLG00000004424	asph	20.7803	21.4418
ENSORLG00000004792	adgrl3.1	0.5426	0.1214
ENSORLG00000004819	dhx15	63.3069	68.5875
ENSORLG00000004845	PAXBP1	56.0877	50.0914
ENSORLG00000005181	slc27a2b	7.8703	26.0275
ENSORLG00000005308	lim2.3	0.5365	0.0000
ENSORLG00000005381	lrp2b	0.3798	0.0619
ENSORLG00000005557	NoName	0.7960	1.0096
ENSORLG00000005581	NoName	0.4811	1.8305
ENSORLG00000005592	rps15	409.4277	160.3491
ENSORLG00000005630	mcf2la	6.2056	22.2997
ENSORLG00000005961	NoName	155.9052	1.1161
ENSORLG00000005964	cldnd1a	9.2129	9.4646
ENSORLG00000005993	oacyl	0.0000	0.0000
ENSORLG00000006157	slc17a6b	0.0816	0.0000
ENSORLG00000006195	si:ch73-67c22.3(14of37)	6.1828	4.3244
ENSORLG00000006223	si:dkey-185e18.6	3.4746	1.6525
ENSORLG00000006283	smu1b	1.1513	0.7301
ENSORLG00000006359	wu:fd14a01(4of6)	0.0989	0.1882

Table 1. RPKM of the genes which have Hd-rR specific HMDs in their promoters.

Gene ID	Gene name	RPKM in d-rR	RPKM in HNI
ENSORLG00000006375	EGR4	0.9197	0.0000
ENSORLG00000006490	brd4	23.0851	32.0824
ENSORLG00000006833	cers5	6.3449	0.7100
ENSORLG00000006850	wnt8b	0.1333	0.0000
ENSORLG00000007124	bmp8a	0.8013	0.0000
ENSORLG00000007296	si:ch1073-280h16.1	28.4137	65.8713
ENSORLG00000007558	gcnt3	8.6037	0.3148
ENSORLG00000007631	rnf26	29.4956	28.7356
ENSORLG00000007789	march1	0.3581	0.0000
ENSORLG00000007861	rarga	14.6317	13.1855
ENSORLG00000007932	rcn2	49.5532	61.1183
ENSORLG00000008028	zgc:92360	0.2497	0.7125
ENSORLG00000008045	slc25a26	4.4875	3.4148
ENSORLG00000008114	SLCO5A1(3of3)	20.3409	21.0646
ENSORLG00000008204	tusc5a	0.0000	0.0000
ENSORLG00000008287	asb12b	0.1330	3.7962
ENSORLG00000008568	akr1a1a	4.1885	4.1653
ENSORLG00000008655	acsl1b	0.6878	0.0000
ENSORLG00000008718	NoName	0.0000	0.0000
ENSORLG00000008851	txlnbb	0.3233	0.4101
ENSORLG00000008978	TMEM233	1.9991	0.0000
ENSORLG00000008984	LMNA(1of2)	0.8929	0.3640
ENSORLG00000009025	kcnip3b	0.9319	0.0000
ENSORLG00000009162	kcnk12l	0.3508	0.0000
ENSORLG00000009204	acer1	6.4976	0.3341
ENSORLG00000009220	gorasp2	32.3853	18.3576
ENSORLG00000009491	NoName	0.2984	0.0000
ENSORLG00000009564	NoName	0.0000	0.0000
ENSORLG00000009585	GAB3	0.4103	0.1301
ENSORLG00000009853	NoName	60.8808	107.5492
ENSORLG00000010101	zgc:101785	10.8423	2.5087
ENSORLG00000010130	jmjd7	6.3679	9.8431
ENSORLG00000010131	chst2b	9.7510	0.0000
ENSORLG00000010188	NoName	6.5718	0.0000
ENSORLG00000010534	CYP46A1(2of2)	53.7104	0.0000
ENSORLG00000010738	FCHSD1	0.3510	0.1335
ENSORLG00000010811	trappc11	13.7378	26.4642
ENSORLG00000010872	vrk1	94.6376	60.1372
ENSORLG00000011003	EPHB1(1of2)	1.5064	1.1642
ENSORLG00000011018	NoName	24.5735	18.8639
ENSORLG00000011521	NoName	0.1645	0.0000
ENSORLG00000011646	RASA2	5.7228	15.4650
ENSORLG00000011676	NoName	1.0537	4.5102
ENSORLG00000011698	ddias	15.2902	7.2722
ENSORLG00000011703	NoName	5.5518	11.0648
ENSORLG00000011885	hdhd2	6.8434	12.1396
ENSORLG00000011950	NoName	15.4618	66.4322
ENSORLG00000012156	ano10b	6.0518	1.3116
ENSORLG00000012194	srd5a2b	0.3224	0.8178
ENSORLG00000012440	auts2a	0.3767	0.0896
ENSORLG00000012675	NoName	0.9816	1.2450
ENSORLG00000012838	bace2	13.6916	105.2484
ENSORLG00000012875	poll	24.2972	30.1093
ENSORLG00000013041	PTCHD3	0.1059	0.0000
ENSORLG00000013244	vmhcl	0.0000	0.2830
ENSORLG00000013368	lrrc53	0.2320	0.0000
ENSORLG00000013616	slc16a7	0.5559	0.1763
ENSORLG00000013691	parp1	130.2761	64.7385
ENSORLG00000013731	pkd2l1	1.4113	0.2685
ENSORLG00000013769	zgc:92107	68.5084	67.5954

Table 1 (continued)

Gene ID	Gene name	RPKM in d-rR	RPKM in HNI
ENSORLG00000013831	APOH	0.0000	0.5752
ENSORLG00000013910	tnfrsf9a	12.5085	12.0707
ENSORLG00000013920	FADS6	0.4054	0.0000
ENSORLG00000013983	NoName	34.8152	10.7085
ENSORLG00000013993	keap1a	3.5495	20.5650
ENSORLG00000014020	NoName	34.8443	5.9070
ENSORLG00000014089	NOX5	0.3230	1.9664
ENSORLG00000014119	map3k1	0.3483	0.6626
ENSORLG00000014180	si:ch1073-416j23.1	41.0016	99.9458
ENSORLG00000014235	PPAP2C(1of2)	55.4012	88.6299
ENSORLG00000014287	kif3a(2of2)	0.0000	0.0000
ENSORLG00000014312	serpinh1b	2.0347	0.2150
ENSORLG00000014430	NoName	0.4261	0.0000
ENSORLG00000014608	adoa	36.4081	0.0000
ENSORLG00000014644	sv2bb	0.2876	0.2189
ENSORLG00000014670	slc17a9a	3.2270	6.8213
ENSORLG00000014673	LRRC52(2of2)	0.3378	0.6427
ENSORLG00000014798	vps8	12.1707	26.6421
ENSORLG00000014811	cyp27a7	0.1774	0.1688
ENSORLG00000014932	nt5e(1of2)	1.4675	2.4816
ENSORLG00000014942	atad1a	10.3360	0.0000
ENSORLG00000014988	abcc12	0.2196	0.0000
ENSORLG00000015149	adamts18	18.3555	9.5434
ENSORLG00000015165	zgc:162161	7.9848	12.7920
ENSORLG00000015284	rab11bb(1of2)	41.4360	43.6665
ENSORLG00000015350	clip2	8.6013	10.2766
ENSORLG00000015360	tat	4.0353	0.5583
ENSORLG00000015514	NoName	0.6748	0.0000
ENSORLG00000015540	racgap1	157.3527	57.8054
ENSORLG00000015707	si:dkeyp-110c7.4(1of2)	3.3471	1.2735
ENSORLG00000015733	CTSS(2of2)	433.8264	0.6663
ENSORLG00000015853	C3orf38	7.9447	7.9549
ENSORLG00000015962	NoName	2.6085	0.2757
ENSORLG00000016315	nr2f5	0.5732	0.2181
ENSORLG00000016388	hdac7b	0.1032	0.2944
ENSORLG00000016512	tfa	0.3048	0.1160
ENSORLG00000016536	nrip2	1.5134	0.0000
ENSORLG00000016606	rad23b(2of2)	147.4874	163.2942
ENSORLG00000016659	NoName	60.1106	59.2109
ENSORLG00000016707	si:dkey-88e18.8	0.3075	0.0000
ENSORLG00000016718	NoName	0.1409	0.0000
ENSORLG00000016741	msxe	13.3216	5.2662
ENSORLG00000016848	MEGF9	0.1603	1.3721
ENSORLG00000016853	NoName	0.0747	0.1421
ENSORLG00000016916	NoName	0.1908	1.0887
ENSORLG00000016942	gchfr	7.5563	3.5939
ENSORLG00000017060	nfe2l1a	2.2402	1.3394
ENSORLG00000017104	usp6nl	9.3099	12.1766
ENSORLG00000017108	agbl4	0.2044	0.3888
ENSORLG00000017231	ripk1l	23.2415	23.6167
ENSORLG00000017248	myo3b	0.0796	0.9088
ENSORLG00000017362	fdft1	43.9636	46.7932
ENSORLG00000017883	clul1	1.5316	0.6475
ENSORLG00000018176	cntnap5a	6.2632	0.4906
ENSORLG00000018215	prkag3b	0.6916	1.3158
ENSORLG00000020923	NoName	0.0000	0.0000
ENSORLG00000021171	NoName	0.0000	0.0000
ENSORLG00000021310	Y_RNA	9.6219	0.0000
ENSORLG00000021579	SNORA62	0.0000	0.0000

Table 1 (continued)

Gene ID	Gene name	RPKM in d-rR	RPKM in HNI
ENSORLG00000000055	slit1b	0.0316	0.0000
ENSORLG00000000313	lygl1	1.3207	0.0000
ENSORLG00000000403	si:ch211-240g9.1	0.0000	0.0000
ENSORLG00000000463	pygo2	25.1205	32.2708
ENSORLG00000000499	pitpnb	70.7332	64.9043
ENSORLG00000000540	INPP5A	3.6483	2.1690
ENSORLG00000000542	emilin1b	1.3526	4.0982
ENSORLG00000000548	VDAC3(1of2)	6.4160	4.7867
ENSORLG00000000758	hspb11(1of2)	0.0000	0.4449
ENSORLG00000000793	nde1	22.3312	18.7719
ENSORLG00000000801	nr2c2ap	53.5434	89.8793
ENSORLG00000000804	znf277	22.8568	25.3819
ENSORLG00000000905	rbfox3l	1.1105	4.0491
ENSORLG00000000950	NoName	0.4816	0.0000
ENSORLG00000001052	xylt1(1of2)	1.2229	3.3235
ENSORLG00000001152	NoName	7.0499	4.3589
ENSORLG00000001307	NoName	0.0000	1.2257
ENSORLG00000001446	ppiab	6.3199	3.6434
ENSORLG00000001510	NoName	4.3239	8.2260
ENSORLG00000001586	rel	12.7531	11.4378
ENSORLG00000001598	arhgap4b	1.9218	5.0714
ENSORLG00000001685	usp19	41.1792	22.2688
ENSORLG00000002023	AP3B2	4.0463	0.1673
ENSORLG00000002076	NoName	135.6203	258.8627
ENSORLG00000002212	nrm	1.4307	5.4437
ENSORLG00000002236	CLEC3B(1of2)	0.9729	6.4780
ENSORLG00000002298	NoName	0.0985	0.5620
ENSORLG00000002317	kcnh6a	0.1627	0.0774
ENSORLG00000002333	SLC39A3	42.5310	119.7607
ENSORLG00000002545	C2orf42	3.1925	7.3522
ENSORLG00000002571	rims1b	2.1086	7.8558
ENSORLG00000002997	fbp1b	0.3682	7.7054
ENSORLG00000003086	NoName	0.6554	1.6624
ENSORLG00000003099	hyal3	1.7975	0.2137
ENSORLG00000003156	lingo4a	0.0000	0.0000
ENSORLG00000003190	rasgrf2b	0.1695	0.0645
ENSORLG00000003346	C18orf8	16.3846	19.2566
ENSORLG00000003363	rhbdl3	0.3892	0.2468
ENSORLG00000003367	kcng3	0.2206	0.4197
ENSORLG00000003424	FlGN(1of2)	1.9252	1.0988
ENSORLG00000003657	NoName	0.7530	1.1460
ENSORLG00000003673	chd1l	5.7648	6.6333
ENSORLG00000003687	CSGALNACT1(1of2)	0.3617	0.1720
ENSORLG00000003722	PRSS23	0.2451	0.4662
ENSORLG00000003894	CELF2(1of2)	130.3021	185.4956
ENSORLG00000004195	ppih	119.2936	119.6126
ENSORLG00000004207	fam78bb	0.0000	0.0000
ENSORLG00000004237	lypd6	1.7582	4.4599
ENSORLG00000004268	rbfox2(1of2)	3.4511	8.9103
ENSORLG00000004398	XKR6(1of2)	0.5157	0.4905
ENSORLG00000004412	IFFO2	10.5506	1.2677
ENSORLG00000004415	BCAP29(1of2)	0.3162	0.6016
ENSORLG00000004671	grtp1a	36.2792	30.5028
ENSORLG00000004723	tyrp1b	0.0842	0.0000
ENSORLG00000004944	COLQ(1of2)	0.0000	0.4239
ENSORLG00000005044	amh	0.8501	5.3908
ENSORLG00000005065	inpp5kb	0.2674	0.5088
ENSORLG00000005450	HOMER3(1of2)	0.5466	3.3794
ENSORLG00000005497	nxnl2	0.3098	4.1261
ENSORLG00000005630	mcf2la	6.2056	22.2997

Table 2. RPKM of the genes which have HNI specific HMDs in their promoters.

Gene ID	Gene name	RPKM in d-rR	RPKM in HNI
ENSORLG00000005778	ZBTB7C	1.0575	2.3776
ENSORLG00000005873	RFESD(1of2)	12.9117	14.8525
ENSORLG00000005927	NoName	688.5746	205.5557
ENSORLG00000005990	zcchc8	65.3991	49.3559
ENSORLG00000006014	TCTN3	17.0466	13.7904
ENSORLG00000006079	slc22a6l	0.0845	3.8582
ENSORLG00000006354	ggact.2	12.0794	32.2967
ENSORLG00000006450	shbg	0.3007	1.1441
ENSORLG00000006454	cyp4f3	2.9168	14.1409
ENSORLG00000006547	neurl1b	3.0870	10.5000
ENSORLG00000007057	naa10	141.9277	168.1562
ENSORLG00000007325	ppp1r16a	32.7395	19.3132
ENSORLG00000007367	rnf13	20.0241	49.2444
ENSORLG00000007539	padi2(2of2)	5.1596	8.6220
ENSORLG00000007547	SVEP1	2.1389	4.3507
ENSORLG00000007762	fbxw7	23.2529	39.3220
ENSORLG00000007768	lgals3bpa	11.0685	1.0454
ENSORLG00000008091	MYO1E	24.6136	47.5657
ENSORLG00000008516	gpr31(1of2)	0.2948	0.2804
ENSORLG00000008863	gng2	3.3781	0.0000
ENSORLG00000009012	mapk12b	4.1546	6.6690
ENSORLG00000009111	TGFB3(1of2)	0.0954	0.1815
ENSORLG00000009155	si:dkey-266m15.5	10.9221	22.7989
ENSORLG00000009179	fuom	14.1224	16.7173
ENSORLG00000009218	TIMP3	4.5250	27.1171
ENSORLG00000009234	si:ch211-161h7.8	182.6892	195.1739
ENSORLG00000009473	ZC3H12A(1of2)	4.7369	11.9577
ENSORLG00000009487	cx39.9	0.4088	0.2592
ENSORLG00000009655	NoName	35.2372	41.0798
ENSORLG00000009687	tnk2b	0.3980	7.2351
ENSORLG00000009931	tbcela	8.4950	7.8882
ENSORLG00000010093	si:dkey-19e4.5	28.8426	17.8059
ENSORLG00000010300	P2RY2	0.1293	0.2459
ENSORLG00000010320	NoName	0.0000	0.0000
ENSORLG00000010446	NoName	0.1838	0.0000
ENSORLG00000010709	si:ch73-127m5.1	0.1923	0.0000
ENSORLG00000010726	mmp17b	0.9576	0.1822
ENSORLG00000010745	trdn	0.0000	0.0000
ENSORLG00000010844	fam83fb	0.3276	1.2464
ENSORLG00000010994	kif19	7.8797	23.4854
ENSORLG00000011034	gpr186	0.4409	0.2796
ENSORLG00000011206	dachd	1.3898	1.1941
ENSORLG00000011305	NoName	22.4814	83.2881
ENSORLG00000011393	cln5	6.7269	18.8045
ENSORLG00000011512	cacng6b	0.3771	3.5869
ENSORLG00000011532	syt14b	0.4390	0.1670
ENSORLG00000011726	camkk1a	8.2541	16.3310
ENSORLG00000011875	SSC4D	0.8501	1.9766
ENSORLG00000011907	hhatla	1.7950	0.8537
ENSORLG00000012181	NoName	0.1561	0.0000
ENSORLG00000012186	NoName	0.0000	0.0000
ENSORLG00000012428	NoName	48.7206	61.6463
ENSORLG00000012482	FHL2(2of2)	0.1647	0.0000
ENSORLG00000012690	NoName	17.6965	21.4685
ENSORLG00000012714	grin2aa	0.4975	2.9743
ENSORLG00000012758	bcl2l10	44.5684	79.9116
ENSORLG00000012858	MAT1A(2of2)	1.2141	2.3097
ENSORLG00000012895	slc2a6	0.2716	0.1722
ENSORLG00000013093	C1orf116	1.2096	2.9587
ENSORLG00000013122	pltp	1.0832	15.5486

Table 2 (continued)

Gene ID	Gene name	RPKM in d-rR	RPKM in HNI
ENSORLG00000013190	C9orf172(1of2)	0.4171	0.2976
ENSORLG00000013258	TMEM229A	0.5826	1.6624
ENSORLG00000013293	NoName	0.2688	0.0000
ENSORLG00000013536	myh11a	1.9920	3.3218
ENSORLG00000013703	C17orf85	27.8851	37.2822
ENSORLG00000013751	fundc1	86.6177	55.5254
ENSORLG00000014110	lgi3	0.4989	0.0000
ENSORLG00000014204	ldlrap1a	2.4347	11.0009
ENSORLG00000014434	lrrc3	0.0000	0.0000
ENSORLG00000014485	HEPACAM(2of2)	0.5649	0.0000
ENSORLG00000014564	PDPR	16.4770	31.5463
ENSORLG00000014584	ggt5a	0.5138	1.4663
ENSORLG00000014639	CCDC134	6.0029	12.6015
ENSORLG00000014694	NoName	0.0000	1.3014
ENSORLG00000014855	has3	3.3051	9.5122
ENSORLG00000015086	kcnj1b	1.5045	0.2385
ENSORLG00000015155	rev3l	20.2256	34.4790
ENSORLG00000015474	tldc1	52.2080	41.9648
ENSORLG00000015577	anapc13	47.6781	14.0317
ENSORLG00000015735	cbln11	0.0000	0.0000
ENSORLG00000015785	FAM177B	0.0000	1.2525
ENSORLG00000015895	SPTBN1(2of3)	0.3644	0.0000
ENSORLG00000015981	myl1	0.0000	3.4391
ENSORLG00000016178	kcng1	0.3793	0.9020
ENSORLG00000016228	pusl1	5.5149	2.1521
ENSORLG00000016234	zdhhc22	0.3572	0.3398
ENSORLG00000016244	paqr6	0.5519	1.4700
ENSORLG00000016454	cnot2	32.5479	45.5099
ENSORLG00000016609	AMDHD1	1.4339	2.0985
ENSORLG00000016638	gpr55a	0.2870	0.0000
ENSORLG00000016750	NoName	9.7005	1.3471
ENSORLG00000016871	zgc:194887	0.2191	3.3349
ENSORLG00000016897	C2CD4C(2of2)	0.1363	0.0000
ENSORLG00000017095	C15orf52	0.6949	1.3220
ENSORLG00000017219	NoName	1.3933	1.6372
ENSORLG00000017301	tspan4b	0.1554	0.0000
ENSORLG00000017367	KCNMB4	0.4611	8.3332
ENSORLG00000017368	NoName	49.2462	35.4681
ENSORLG00000017415	matk	0.2221	0.0000
ENSORLG00000017918	ackr4a	1.4345	0.4962
ENSORLG00000018069	chga	1.4961	0.4379
ENSORLG00000018199	ahr1b	0.0797	0.1517
ENSORLG00000021061	NoName	0.0000	0.0000
ENSORLG00000021195	NoName	0.0000	0.0000
ENSORLG00000021244	NoName	0.0000	0.0000
ENSORLG00000021563	SNORA62	0.0000	0.0000

Table 2 (continued)

6mer	ratio of MI	enrichment	periodicity	matched known motifs (q value < 0.1)
CGCGAC	0.295	6.021	weak	
GCGCGA	0.296	9.754	weak	
CGCGCG	0.315	48.092	weak	Zfp161,E2F2,E2F3
TCGCGA	0.359	5.200	intermediate	ZBTB33
CGCGGA	0.360	11.797	No	
CGCGAG	0.364	9.981	weak	ZBTB33
CCGCGG	0.364	8.035	No	
CCGCGC	0.386	13.042	weak	Zfp161
CGCTAG	0.388	3.302	strong	
TCCGGA	0.391	3.228	No	Spdef
CACGTG	0.400	2.606	No	Mycn,Arnt,MYC::MAX,Max,Bhlhb2, Bhlhe40,USF1,Myc,HIF1A::ARNT,USF2
ACCGGA	0.412	3.175	weak	Gabpa
CCGGAG	0.419	3.363	No	
TCCGAA	0.420	2.367	No	
CCGGAA	0.427	3.726	intermediate	Gabpa,ELK4,ELK1,Ehf
CGCGCC	0.435	11.455	weak	E2F3,E2F2,Zfp161
ATCCGG	0.446	2.328	No	Spdef
CGCGCA	0.450	15.178	weak	Zfp161
CGCGGC	0.450	9.141	weak	
GCGCGC	0.455	24.526	weak	E2F2,E2F3,Zfp161

Table 3. The list of the possible transcription factors which could bind to selected top 20 6-mers with CpGs.

For each 6-mer, the ratio of mutation index (common HMDs / Hd-rR specific HMDs), enrichment level within the common HMDs (frequency in common HMDs / frequency in methylated regions) and the intensity of periodicity of DNase-seq signal around itself are also shown.

6mer	ratio of MI	enrichment	periodicity	matched known motifs (q value < 0.1)
GCTAGC	0.335	4.486	strong	
GCTAAC	0.498	2.650	strong	
AGCTAG	0.502	2.525	strong	
CTTACC	0.527	1.407	No	
GGTCAC	0.588	1.059	No	ESRRA,PPARG,Rara,NR4A2,ESR1,USF1, Nr2f2,USF2,ESR2
GGCCAC	0.597	1.002	No	
GACCCC	0.608	1.126	No	Glis2,Hnf4a,Esrra,Rxra,Zfp281,Rara,
GGTCCA	0.619	1.040	No	
CCATGC	0.620	0.885	No	
GGGATA	0.620	0.928	No	
TAGCTA	0.628	2.359	strong	
GGTACC	0.630	1.168	No	Plagl1
CTTGCC	0.633	0.810	No	
ACTTCC	0.640	1.184	intermediate	ELF1,Gabpa,Spi1,Erg,FLI1,Sfpi1,Ehf,ELK4, Ets1,Ehf,ELF5,FEV,Elf3,
GGCACC	0.641	1.234	No	
AGGTAA	0.641	1.136	No	
AAGGTA	0.643	0.890	No	
GCATAC	0.644	0.788	No	
GGATCC	0.652	1.373	No	
GGGGGA	0.664	1.705	No	Obox2,Pitx3,Obox3,Zfp740,MZF1,Zfp281

Table 4. The list of the possible transcription factors which could bind to selected top 20 6-mers without CpGs.

For each 6-mer, the ratio of mutation index (common HMDs / Hd-rR specific HMDs), enrichment level within the common HMDs (frequency in common HMDs / frequency in methylated regions) and the intensity of periodicity of DNase-seq signal around itself are also shown.

Name	Sequence
GFP-F	ATGGTGAGCAAGGGCGAGGAG
GFP-R	GGTGGCGACCGGTGGATCCA
bactPro-InF-F	CCACCGGTCGCCACCGCAGGAATTCAATTACAGTG
bactPro-InF-R	GCCCTTGCTCACCATGGCTAAACTGGAAAAGAACA
HdrRsp1-InF-F	TAGTGGATCCACCGGGGTCCTGCTCACCTGTTTCT
HdrRsp1-InF-R	TGCGGTGGCGACCGGTTGACTTCTGTTGTGAAGTTAGATG
HNIsp1-InF-F	TAGTGGATCCACCGGTCAACCAAATATTAGTAATGACCCTTT
HNIsp1-InF-R	TGCGGTGGCGACCGGTGCACCACTAAGGTTAAATTGG
HNIsp2-InF-F	TAGTGGATCCACCGGTCTGATGAACAAGGAAAAACCA
HNIsp2-InF-R	TGCGGTGGCGACCGGTTCCAGACCTCCCTCAGAAATG
HNIsp3-InF-F	TAGTGGATCCACCGGGAAAACAAACGGACCCTCAG
HNIsp3-InF-R	TGCGGTGGCGACCGGAGGTCAAAGGCTAAAGGTTACT
common1-InF-F	TAGTGGATCCACCGGACATGTTTGATGTCTCAAGCTAC
common1-InF-R	TGCGGTGGCGACCGGCCACTGAAAGGTCCAGATTCA
common2-InF-F	TAGTGGATCCACCGGTGCAATAAAGCAAATAACTTAAAGGAC
common2-InF-R	TGCGGTGGCGACCGGGAATCCCGATTGTTTTAGAATG
common3-InF-F	TAGTGGATCCACCGGTCTTCACATTGCTGGAACTGAC
common3-InF-R	TGCGGTGGCGACCGGACAAAGCCCGTCACCTACTG

Table 5. Primers used for making transgenic medaka (for cloning of HMD sequences and amplification of b-actin and GFP sequences)

Name	Sequence
bactPro-seq-F1	TTAGAAGGTAACATCATCTG
bactPro-seq-F2	AAGCCACGAATGAATTTAAG
bactPro-seq-F3	TGAGGTGGCATTCTGCTTTC
bactPro-seq-F4	TAGCAGAATTTTGTGGCCAC
bactPro-seq-F4-2	AATTGGAGGTGACCATTAGC
bactPro-seq-F5	GTGTAACAATGGGAGGGAAC
GFP-seq-F1	GTTCGAGGGCGACACCCTGG
GFP-polyA-seq-F1	GGTGGTGCAGATGAACTTCA
M13R_bef_seq	TCCGGCTCGTATGTTGTGTG
HdrRsp1_seq_F1	CTCAGCATCTCATCCTGGAG
HdrRsp1_seq_F2	ATGGAAAATAATGGGAGCAC
HdrRsp1_seq_F3	GAGAAATGAAGACGTACATG
HdrRsp1_seq_F4	CCTTTTGTTCTGGAAACATG
HdrRsp1_seq_F5	GGATCACTGAACACTGACAG
HdrRsp1_seq_F6	CTTCGTCAATTGAATAATAATATG
HNIsp1_seq_F1	TTAAGTGAATTTCTAGAAC
HNIsp1_seq_F2	AGGGGATCAGAAATATAAAC
HNIsp1_seq_F3	CGCAACATCTCGGCTGGCTG
HNIsp1_seq_F4	GCCATCCACAAGACAAAAC
HNIsp1_seq_F5	ACTTTCCCCGCTGGGATTTC
HNIsp1_seq_F6	TGTCCTTCCTTCTGTACAG
HNIsp2_seq_R1	AATATGGTGCTTAACCTTGG
HNIsp2_seq_R2	CCAAATCTGCCTATAAACTC
HNIsp2_seq_R3	ATTGTGGCCTACTGCGCCATG
HNIsp2_seq_R4	CCTCATTTTATTATGAAAGG
HNIsp2_seq_R5	TAGGTAAACTATAAAAGTTG
HNIsp2_seq_R6	AAAGTTTGGTGTTATGTTGC
HNIsp2_seq_R7	AGGTACTTCTGCGAGGCGTC
HNIsp3_seq_R1	TGCACATGTGCAGACGGGAC
HNIsp3_seq_R2	TTCTCCCCGTCTGCATGGAG

Table 6. Primers used for making transgenic medaka (for confirmation of the sequences of the constructs)

Name	Sequence
HNIsp3_seq_R3	CTCGGTAGCTGCGTGCCTTG
HNIsp3_seq_R3-2	CGGTGGTTGGCGTGATATG
HNIsp3_seq_R4	ATACTAACGTCCACTCAAAG
HNIsp3_seq_R5	TCTTCCTCTTCATCAGGGAG
HNIsp3_seq_R6	AAGCCGCACAGCTCTGCATC
common1_seq_R1	TCTGTTGGCTCAGTTGTTGG
common1_seq_R2	CGCTTGCAATGTCGGTGATG
common1_seq_R3	AATCCACATTTACGCGTAGC
common1_seq_R4	TCTGTTGGCTCAGTTGTTGG
common1_seq_R5	AAGGTTACACAAACTAACTC
common1_seq_R5-2	GAATGCTACAATCACAGAGG
common1_seq_R6	CAAAAGTGTCAGAAAACGTC
common2_seq_F1	TAGTTCCCTGTTTGGAGCTC
common2_seq_F2	GTCTTTTAATAAGGATAATG
common2_seq_F3	TAAAATCAAGTTTGGCTGTC
common2_seq_F4	CATTCGCCGGGCTAGACCAC
common2_seq_F5	CACAAGTTATGTAAAAAGAC
common2_seq_F6	TTCACCACAAATACTCAGAG
common2_seq_F7	CCCACATGTGGGGAAACAAG
common3_seq_R1	TCAGTCAGGGTGGCAGCGTC
common3_seq_R2	GCCTCAGTTAAAACCTAGAG
common3_seq_R3	AACTGTTATCTCCCATTAGG
common3_seq_R4	CGTAAACTAATTGTGTTTTC
common3_seq_R4-2	ACTTCAAGTACTGCAAAATC
common3_seq_R4-3	TCTATTGAAGTGTTCTAATC
common3_seq_R5	TAGAAACTAAGCAAGCCACG
common3_seq_R6	TACCCAAAGGTACAGCAAAG
common3_seq_R7	TATTGCTGCTTTTTAGCTGG

Table 6 (continued)

Name	Sequence
HdrRsp1_bs_F	TTTTGTGTATTTTTTATTATTAGAAAAATG
HdrRsp1_bs_R	AAAAACCTCTCCAACCTCAATAAC
HNIsp1_bs_F	TGGATTTGATATATATTTTAATTGT
HNIsp1_bs_R	TAAAACTACAAAACTCTAACACCTC
HNIsp2_bs_F	GGATGTTATAGGTGATTATTGGTTTG
HNIsp2_bs_R	CCTTAAAACTCCAACTTAACACAATTT
HNIsp3_bs_F	GTGTTGTTGTTTATTTTTTTGAT
HNIsp3_bs_R	TTTTCCTACAAATACTATCTTCCCC
common1_bs_F	GTTTATTTTTTTATTTATTTGATTAG
common1_bs_R	CAAATTTTACCCCCATAATTAACTC
common2_bs_F	GAGGAGTTAGAATTTTTTTAAAATTT
common2_bs_R	ATACTACTTTAACTCCAATACATCC
common3_bs_F	TGGTTGGAAGTAGTATAGTTTAGAAAA
common3_bs_R	CATACATCACCATCTTCAACAAAAC
HNIsp1R_bs_F	TTTTAAAGAAAGTGTGAAATTAGGATG
HNIsp1R_bs_R	AAATCTTAACAAAAATCACATAACC
HNIsp2L_bs_F	GTTGAAGGTTTGTGAATTTGAATTT
HNIsp2L_bs_R	TCCAAACAACAATATAACCACTACC
HNIsp2R_bs_F	TTTATGTGAGGATGAAGGTTAGTAGG
HNIsp2R_bs_R	AACCTCCCTCAAAAATACAAAATAC
common1L_bs_F	TGGGGAATAGTTGGTGTAGTTAGTT
common1L_bs_R	ATAAAATCTTTAATCCACTTTCTTACCC
common1R_bs_F	TTTTGATTTGATTTGAATTGGAATT
common1R_bs_R	TAAATAATCTTCCACCAACTATAAA
common2L_bs_F	ATTTGGAGTAGGTGAAAAATGTTGT
common2L_bs_R	AAATTACAAACCCAATTCAATCATC
common2R_bs_F	TTGTAGTTTTTTTTGTTTGAAATAG
common2R_bs_R	CAAATCTCTAAACTCCAACTTCCTAC
zicA_bs_F	TTGTGTGGGTAGTATAGTTATTTTGAG
zicA_bs_R	CCTAATAACAAAACATAAAATCTTTTT
zicB_bs_F	GGATTTTGTTTTAGGTTTTTTAGT
zicB_bs_R	CCATTAATCTCTACATATATACATTTTT
zicC_bs_F	GTTGGTAGTTGTAATTTTTATGGGG
zicC_bs_R	CCCAATTAATAACCCTTCAATTAACT

Table 7. Primers used for bisulfite analysis

References

Akhavan-Niaki, H., and Samadani, A.A. (2013). DNA methylation and cancer development: molecular mechanism. Cell Biochem Biophys *67*, 501-513.

Andersen, I.S., Reiner, A.H., Aanes, H., Alestrom, P., and Collas, P. (2012). Developmental features of DNA methylation during activation of the embryonic zebrafish genome. Genome Biol *13*, R65.

Asai, T., Senou, H., Hosoya, K. (2011). Oryzias sakaizumii, a new ricefish from northern Japan (Teleostei: Adrianichthyidae). Ichthyol Explor Freshwaters *22*, 289-299.

Bassett, A., Cooper, S., Wu, C., and Travers, A. (2009). The folding and unfolding of eukaryotic chromatin. Curr Opin Genet Dev *19*, 159-165.

Beh, L.Y., Muller, M.M., Muir, T.W., Kaplan, N., and Landweber, L.F. (2015). DNA-guided establishment of nucleosome patterns within coding regions of a eukaryotic genome. Genome Res *25*, 1727-1738.

Bestor, T., Laudano, A., Mattaliano, R., and Ingram, V. (1988). Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. J Mol Biol *203*, 971-983.

Bestor, T.H., and Ingram, V.M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. Proc Natl Acad Sci U S A *80*, 5559-5563.

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

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res *8*, 1499-1504.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311-322.

Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. Nature *371*, 435-438.

Chen, R.A., Down, T.A., Stempor, P., Chen, Q.B., Egelhofer, T.A., Hillier, L.W., Jeffers, T.E., and Ahringer, J. (2013). The landscape of RNA polymerase II transcription initiation in C. elegans reveals promoter and enhancer architectures. Genome Res *23*, 1339-1347.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. Nature *274*, 775-780.

Crawford, G.E., Davis, S., Scacheri, P.C., Renaud, G., Halawi, M.J., Erdos, M.R., Green, R., Meltzer, P.S., Wolfsberg, T.G., and Collins, F.S. (2006). DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nat Methods *3*, 503-509.

Dasmahapatra, A.K., and Khan, I.A. (2015). DNA methyltransferase expressions in Japanese rice fish (Oryzias latipes) embryogenesis is developmentally regulated and modulated by ethanol and 5-azacytidine. Comp Biochem Physiol C Toxicol Pharmacol *176-177*, 1-9.

Dasmahapatra, A.K., and Khan, I.A. (2016). Modulation of DNA methylation machineries in Japanese rice fish (Oryzias latipes) embryogenesis by ethanol and 5-azacytidine. Comp Biochem Physiol C Toxicol Pharmacol *179*, 174-183.

Dickson, J., Gowher, H., Strogantsev, R., Gaszner, M., Hair, A., Felsenfeld, G., and West, A.G. (2010). VEZF1 elements mediate protection from DNA methylation. PLoS Genet *6*, e1000804.

Feinberg, A.P., and Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature *301*, 89-92.

Gertz, J., Varley, K.E., Reddy, T.E., Bowling, K.M., Pauli, F., Parker, S.L., Kucera, K.S., Willard, H.F., and Myers, R.M. (2011). Analysis of DNA methylation in a three-generation family reveals widespread genetic influence on epigenetic regulation. PLoS Genet *7*, e1002228.

Harris, RS. (2007). Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

Heinz, S., Romanoski, C.E., Benner, C., Allison, K.A., Kaikkonen, M.U., Orozco, L.D., and Glass, C.K. (2013). Effect of natural genetic variation on enhancer selection and function. Nature *503*, 487-492.

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.

Hermann, A., Goyal, R., and Jeltsch, A. (2004). The Dnmt1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. J Biol Chem *279*, 48350-48359.

Hernando-Herraez, I., Heyn, H., Fernandez-Callejo, M., Vidal, E., Fernandez-Bellon, H., Prado-Martinez, J., Sharp, A.J., Esteller, M., and Marques-Bonet, T. (2015). The interplay between DNA methylation and sequence divergence in recent human evolution. Nucleic Acids Res *43*, 8204-8214.

Hughes, A.L., Jin, Y., Rando, O.J., and Struhl, K. (2012). A functional evolutionary approach to identify determinants of nucleosome positioning: a unifying model for establishing the genome-wide pattern. Mol Cell *48*, 5-15.

Ishikawa, Y., Yoshimoto, M., Yamamoto, N., and Ito, H. (1999). Different brain morphologies from different genotypes in a single teleost species, the medaka (Oryzias latipes). Brain Behav Evol *53*, 2-9.

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

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.

Jiang, L., Zhang, J., Wang, J.J., Wang, L., Zhang, L., Li, G., Yang, X., Ma, X., Sun, X., Cai, J., *et al.* (2013). Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. Cell *153*, 773-784.

Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat

Rev Genet *13*, 484-492.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J., *et al.* (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362-366.

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.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J.B., Kundaje, A., Liu, Y., Boyle, A.P., Zhang, Q.C., Zakharia, F., Spacek, D.V., *et al.* (2013). Extensive variation in chromatin states across humans. Science *342*, 750-752.

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.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.

Kimura, T., Shimada, A., Sakai, N., Mitani, H., Naruse, K., Takeda, H., Inoko, H., Tamiya, G., and Shinya, M. (2007). Genetic analysis of craniofacial traits in the medaka. Genetics *177*, 2379-2388.

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

Lantermann, A.B., Straub, T., Stralfors, A., Yuan, G.C., Ekwall, K., and Korber, P. (2010). Schizosaccharomyces pombe genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of Saccharomyces cerevisiae. Nat Struct Mol Biol *17*, 251-257.

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, H.J., Lowdon, R.F., Maricque, B., Zhang, B., Stevens, M., Li, D., Johnson, S.L., and Wang, T. (2015). Developmental enhancers revealed by extensive DNA methylome maps of zebrafish early

embryos. Nat Commun *6*, 6315.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Li, J., Li, R., Wang, Y., Hu, X., Zhao, Y., Li, L., Feng, C., Gu, X., Liang, F., Lamont, S.J., *et al.* (2015). Genome-wide DNA methylome variation in two genetically distinct chicken lines using MethylC-seq. BMC Genomics *16*, 851.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schubeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. Nat Genet *43*, 1091-1097.

Luu, P.L., Scholer, H.R., and Arauzo-Bravo, M.J. (2013). Disclosing the crosstalk among DNA methylation, transcription factors, and histone marks in human pluripotent cells through discovery of DNA methylation motifs. Genome Res *23*, 2013-2029.

Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. Genes Dev *8*, 2282-2292.

Macleod, D., Clark, V.H., and Bird, A. (1999). Absence of genome-wide changes in DNA methylation during development of the zebrafish. Nat Genet *23*, 139-140.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., *et al.* (2008). Nucleosome organization in the Drosophila genome. Nature *453*, 358-362.

McVicker, G., van de Geijn, B., Degner, J.F., Cain, C.E., Banovich, N.E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J.K. (2013). Identification of genetic variants that affect histone modifications in human cells. Science *342*, 747-749.

Nakamura, R., Tsukahara, T., Qu, W., Ichikawa, K., Otsuka, T., Ogoshi, K., Saito, T.L., Matsushima, K., Sugano, S., Hashimoto, S., *et al.* (2014). Large hypomethylated domains serve as strong repressive machinery for key developmental genes in vertebrates. Development *141*, 2568-2580.

Nakatani, Y., Mello, C.C., Hashimoto, S., Shimada, A., Nakamura, R., Tsukahara, T., Qu, W., Yoshimura, J., Suzuki, Y., Sugano, S., *et al.* (2015). Associations between nucleosome phasing,

sequence asymmetry, and tissue-specific expression in a set of inbred Medaka species. BMC Genomics *16*, 978.

Nelson, H.C., Finch, J.T., Luisi, B.F., and Klug, A. (1987). The structure of an oligo(dA).oligo(dT) tract and its biological implications. Nature *330*, 221-226.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., *et al.* (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83-90.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247-257.

Ponts, N., Harris, E.Y., Prudhomme, J., Wick, I., Eckhardt-Ludka, C., Hicks, G.R., Hardiman, G., Lonardi, S., and Le Roch, K.G. (2010). Nucleosome landscape and control of transcription in the human malaria parasite. Genome Res *20*, 228-238.

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.

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.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Rembold, M., Lahiri, K., Foulkes, N.S., and Wittbrodt, J. (2006). Transgenesis in fish: efficient selection of transgenic fish by co-injection with a fluorescent reporter construct. Nat Protoc *1*, 1133-1139.

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

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.

Schilling, E., El Chartouni, C., and Rehli, M. (2009). Allele-specific DNA methylation in mouse strains is mainly determined by cis-acting sequences. Genome Res *19*, 2028-2035.

Setiamarga, D.H., Miya, M., Yamanoue, Y., Azuma, Y., Inoue, J.G., Ishiguro, N.B., Mabuchi, K., and Nishida, M. (2009). Divergence time of the two regional medaka populations in Japan as a new time scale for comparative genomics of vertebrates. Biol Lett *5*, 812-816.

Shen, J.C., Rideout, W.M., 3rd, and Jones, P.A. (1994). The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res *22*, 972-976.

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.

Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. Nat Struct Mol Biol *20*, 267-273.

Szerlong, H.J., and Hansen, J.C. (2011). Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. Biochem Cell Biol *89*, 24-34.

Takeda, H., and Shimada, A. (2010). The art of medaka genetics and genomics: what makes them so unique? Annu Rev Genet *44*, 217-241.

Takehana, Y., Nagai, N., Matsuda, M., Tsuchiya, K., and Sakaizumi, M. (2003). Geographic variation and diversity of the cytochrome b gene in Japanese wild populations of medaka, Oryzias latipes. Zoolog Sci *20*, 1279-1291.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., *et al.* (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Tillo, D., and Hughes, T.R. (2009). G+C content dominates intrinsic nucleosome occupancy. BMC Bioinformatics *10*, 442.

Tsuboko, S., Kimura, T., Shinya, M., Suehiro, Y., Okuyama, T., Shimada, A., Takeda, H., Naruse, K., Kubo, T., and Takeuchi, H. (2014). Genetic control of startle behavior in medaka fish. PLoS One *9*, e112527.

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.

Veenstra, G.J., and Wolffe, A.P. (2001). Constitutive genomic methylation during embryonic development of Xenopus. Biochim Biophys Acta *1521*, 39-44.

Waddington, C.H. (2012). The epigenotype. 1942. Int J Epidemiol *41*, 10-13.

Wallrath, L.L., Lu, Q., Granok, H., and Elgin, S.C. (1994). Architectural variations of inducible eukaryotic promoters: preset and remodeling chromatin structures. Bioessays *16*, 165-170.

Walter, R.B., Li, H.Y., Intano, G.W., Kazianis, S., and Walter, C.A. (2002). Absence of global genomic cytosine methylation pattern erasure during medaka (Oryzias latipes) early embryo development. Comp Biochem Physiol B Biochem Mol Biol *133*, 597-607.

Whitaker, J.W., Chen, Z., and Wang, W. (2015). Predicting the human epigenome from DNA motifs. Nat Methods *12*, 265-272, 267 p following 272.

Wu, S.C., and Zhang, Y. (2010). Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol *11*, 607-620.

Wu, Y., Zhang, W., and Jiang, J. (2014). Genome-wide nucleosome positioning is orchestrated by genomic regions associated with DNase I hypersensitivity in rice. PLoS Genet *10*, e1004378.

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.

Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J., and Rando, O.J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. Science *309*, 626-630.

Zhang, Y., Moqtaderi, Z., Rattner, B.P., Euskirchen, G., Snyder, M., Kadonaga, J.T., Liu, X.S., and Struhl, K. (2009). Intrinsic histone-DNA interactions are not the major determinant of nucleosome

positions in vivo. Nat Struct Mol Biol *16*, 847-852.

Zhong, J., Luo, K., Winter, P.S., Crawford, G.E., Iversen, E.S., and Hartemink, A.J. (2016). Mapping nucleosome positions using DNase-seq. Genome Res *26*, 351-364.

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 gratitude to Dr. Tatsuya Tsukahara (Harvard Medical School), Dr. Morishita Shinichi (The University of Tokyo) and Dr. Wei Qu (The University of Tokyo) for their supports, advices and discussions about my experiments and computational analyses.

I would like to thank Mr. Yuta Suzuki (The University of Tokyo), Mr. Hayato Sakata (The University of Tokyo) and Dr. Jun Yoshimura (The University of Tokyo) for setting and supporting my computing environment for data analysis, Mr. Kazuki Ichikawa (The University of Tokyo) for assembly of HNI genome, Dr. Yutaka Suzuki (The University of Tokyo) and Dr. Sumio Sugano (The University of Tokyo) for the sequencing by next generation sequencer, and Dr. Kiyoshi Naruse (National Institute for Basic Biology) for providing healthy HNI fish.

I am truly grateful to all the members 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.

I also would like to express my gratitude to GPLLI program (Graduate Program for Leaders in Life Innovation) for giving me a great chance to communicate with various students in other disciplines and financial support, and to the GPLLI teachers and the student members of the program for their advices and warm encouragements.

Finally, I would like to express my endless thankfulness to my dearest parents and sister for their heartfelt support and generous affection. They have always supported me and encouraged me. Without them, I would not have accomplished this study. I dedicate this doctoral thesis to them.