学位論文

STUDY ON THE SEQUENCE-DEPENDENCY OF DNA METHYLATION IN MEDAKA EMBRYOS
（メダカ胚における DNA methylation の配列依存性の研究）

平成29年7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

張 國 鳴
CHEUNG KWOK MING

Abstract

The heavily methylated vertebrate genomes are punctuated by stretches of poorly methylated DNA sequences that usually mark gene regulatory regions. It is known that the methylation state of these regions confers transcriptional control over their associated genes. Given its governance on the transcriptome, cellular functions and identity, genome-wide DNA methylation pattern is tightly regulated and evidently predefined. However, how is the methylation pattern determined *in vivo* remains enigmatic. Based entirely on *in silico* and *in vitro* evidence, recent studies proposed that the regional hypomethylated state is primarily determined by local DNA sequence, e.g., high CpG density and presence of specific transcription factor binding sites. Nonetheless, the dependency of DNA methylation on nucleotide sequence has not been carefully validated *in vivo*.

Herein, with the use of blastula embryos of medaka (*Oryzias latipes*) as a model, the sequence dependency of DNA methylation was rigorously tested *in vivo*. Statistical modelling confirmed the strong statistical association between nucleotide sequence pattern and methylation state in the medaka genome. In particular, consecutive CpG and CGnull-repeats were found highly enriched within the hypomethylated genomic loci of medaka (Chapter 1). However, disruption of these DNA motifs in multiple hypomethylated loci using CRISPR-Cas9 failed to induce any change to the local hypomethylated state in both F0 and F1 embryos. Moreover, by manipulating the methylation state of a substantial number of genomic sequences and reintegrating them into medaka embryos, it was demonstrated that artificially conferred DNA methylation states were predominantly and robustly maintained *in vivo*, regardless of their sequences and endogenous states (Chapter 2).

Hence, despite the observed statistical association, nucleotide sequence was unable to autonomously determine its own methylation state *in vivo*. The results presented herein argue against the general presumption of the governance on DNA methylation by nucleotide sequence, but instead suggest the involvement of other epigenetic factor(s) in defining and maintaining the DNA methylation landscape of vertebrate genomes.

Table of Contents

Abstract ... i

List of figures ... v

List of tables ... vii

List of abbreviations ... viii

General introduction ... 1

Biological importance of nucleotide sequence ... 1

Epigenetics: the bridge between genotype and phenotypes ... 2

The epigenetic element under spotlight: DNA methylation ... 3

Establishment and regulation of DNA methylome ... 5

Experimental outline ... 7

Chapter 1 Statistical association between genomic sequence and DNA methylation state *in vivo* ... 9

1.1 Introduction ... 9

1.2 Results ... 12

1.2.1 HypoMDs and HyperMDs contain distinct sequence patterns ... 12

1.2.2 Consecutive CpG and CGnull-repeat are strongly enriched in HypoMDs ... 13

1.2.3 Only a small subset of the enriched motifs could be linked to transcription factor binding sites ... 13

1.3 Discussion ... 14

1.4 Chapter conclusion ... 17

Chapter 2 Interrogation of the autonomy in DNA methylation determination by genomic sequence ... 25

2.1 Introduction ... 25

2.2 Results ... 27

2.2.1 Disruption of $(CG)_3$ and CGnull-repeat did not lead to ***de novo*** DNA methylation of HypoMDs in F0 blastula embryos ... 27

2.2.2 Methylation state of the $(CG)_3$- or CGnull-repeat-disrupted HypoMDs remained unchanged in F1 blastula embryos ... 29

2.2.3 DNA methylation states are not autonomously determined by nucleotide sequence at ectopic genomic positions ... 29

2.2.4 DNA methylation state was maintained independent of sequence and position context 32

2.3 Discussion 34

2.4 Chapter conclusion 36

General discussion and perspectives 55

Materials and Methods 62

HypoMD and HyperMD calling 62

Supervised classification of HypoMD and HyperMD using support vector machine .. 63

In vivo disruption of $(CG)_3$ and ACGT-repeat in HypoMDs 63

In silico motif analysis 65

High-throughput transplantation of CpG-rich genomic loci 65

Transplantation of Hypo-/Hyper-MDs to specific genomic locus via site-specific recombination 72

In vivo 'methylome editing' via homology-directed repair 74

General procedure: In vitro transcription for the generation of mRNA and sgRNA 76

General procedure: Genomic DNA extraction from medaka blastula embryos 77

General procedure: Genotyping 78

References 79

Appendix 92

Acknowledgements 97

List of figures

Figure 1.1 Genome browser view of a representative locus (approx. 62 kb) in the HdrR medaka genome showing CpG methylation rate, called HypoMDs and HyperMDs, SVM classification results, as well as DNase I hypersensitivity and called DNase I hypersensitive sites (“DHS”). 18

Figure 1.2 Violin plots showing the distribution of (A) length, (B) GC content, and (C) CpG density of HypoMDs and HyperMDs. 19

Figure 1.3 Precision-recall curves of the kmer-SVM models trained for binary classification of HypoMDs and HyperMDs. 20

Figure 1.4 Matthew’s correlation coefficients (MCC) of the kmer-SVM models trained for binary classification of HypoMDs and HyperMDs 21

Figure 1.5 Compact representation of CpG and CGnull-repeats enrichment inside all 2080 possible canonical 6-mers. 22

Figure 1.6 Binding sequence of (A) ZBTB14, (B) E2F2, and (C) E2F3. 23

Figure 2.1 Schematic diagram illustrating the disruption of $(CG)_3$ or CGnull-repeat in a HypoMD via CRISPR-Cas9 and the expected outcomes depending on whether these sequence motifs are responsible for the local hypomethylated state. 37

Figure 2.2 Methylation state of six HypoMDs after disruption of the $(CG)_3$ (Locus 1-3) or CGnull-repeat (Locus 4-6) by the use of CRISPR-Cas9........ 38

Figure 2.3 Schematic diagram illustrating the capturing and processing of genomic fragments for the interrogation of their autonomy in methylation state determination. 39

Figure 2.4 Efficient removal of injected but unintegrated libraries via PEG precipitation and DpnI digestion. 40

Figure 2.5 Violin plots showing the distribution of (A) length, (B) GC content, and (C) CpG density of the unmethylated or pre-methylated fragments that were successfully integrated into genome and subsequently assayed. 41

Figure 2.6 Sampling origins (from HypoMDs, HyperMDs, or elsewhere in the genome) of the assayed CpGs on the integrated genomic fragments..... 42

Figure 2.7 Correlation between the methylation rate of the same CpGs at their endogenous versus at reintegrated/ectopic positions, (A) without- or (B) with- artificial methylation prior to injection and genome integration. . 43

Figure 2.8 Distributions of the methylation rates of CpGs on the integrated genomic fragments, (left) without- or (right) with- artificial methylation prior to injection. 44

Figure 2.9 Differential sensitivity of untreated and M.SssI-treated DNA samples to MspI and HpaII restriction enzymes. 45

Figure 2.10 Correlation of CpG methylation state at ectopic versus native positions with respect to the endogenous local chromatin accessibility. 46

Figure 2.11 Model performance of kmer-SVMs trained for classification of CpGs and their flanking sequences that underwent demethylation. 47

Figure 2.12 Correlation between the overall methylation rate and (A) length or (B) CpG density of the integrated fragments. Integrated fragments derived from the (left) unmethylated and (right) M.SssI-treated libraries were further segregated according to their endogenous methylation state (top vs bottom). 48

Figure 2.13 Interrogation of DNA methylation recapitulation by full-length HyperMDs and HypoMDs at pre-specified, inert genomic location. 49

Figure 2.14 Methylation state of six full-length HyperMDs at their endogenous loci and at ectopic location after being cloned and integrated into genome via PhiC31-mediated site-specific recombination. 50

Figure 2.15 Methylation state of eleven full-length HypoMDs (Locus 1 to 11) at their endogenous loci and at ectopic location after being cloned, artificially methylated, then integrated into genome via PhiC31-mediated site-specific recombination. 51

Figure 2.16 Schematic diagram illustrating the principle of the *in vivo* methylome editing on the targeted HypoMDs via homology-directed repair (HDR) and the use of artificially methylated repair template. 52

Figure 2.17 Methylation state of two HypoMDs after *in vivo* methylome editing mediated by CRISPR-Cas9-induced homology-directed repair (HDR) and pre-methylated repair templates. 53

Figure 2.18 Methylation state of the *in situ*-edited HypoMDs at multiple embryonic stages. 54

List of tables

Table 1 The 30 canonical 6-mers with the highest absolute weights in the SVM model trained for classifying/predicting HypoMDs and HyperMDs. 24

Table 2 List of HypoMDs used for testing the anticipated governing role of $(CG)_3$ and CG-null repeats on local methylation state. 92

Table 3 List of oligo used. 93

Table 4 List of HyperMD loci synthesized and assayed for autonomy in methylation recapitulation. 95

Table 5 Endogenous origins of the integrated fragments from (left) unmethylated and (right) artificially methylated library. 96

List of abbreviations

A-NHEJ — alternative non-homologous end joining pathway

AUC — area under curve

CpG — a dinucleotide context of 5'—cytosine—phosphate—guanine—3'

CRISPR — Clustered Regularly Interspaced Short Palindromic Repeats system

DHS — DNase I hypersensitive site

DNA — Deoxyribonucleic acid

DSB — DNA double-strand break

DNMT1 — DNA methyltransferase 1

dpf — day-post-fertilization

ESC — embryonic stem cells

F0 — the founding generation of transgenic animal

F1 — hybrids obtained by F0×F0 or F0×wide-type cross

F2 — offspring obtained by F0×F0 inter cross

GC content — total percentage of guanine and cytosine

H3 — histone H3

H4 — histone H4

HDR — homology-directed repair

HypoMD — hypomethylated domain

HyperMD — hypermethylated domain

kmer — short stretch of nucleotide sequence with a length of finite number, ***k***

MCC — Matthew's correlation coefficient

MMEJ — microhomology-mediated end joining

NHEJ — non-homologous end joining

PCR — polymerase chain reaction

PRC — precision-recall curve

RRBS — reduced representation bisulfite sequencing

TDG — thymine-DNA glycosylase

TF — transcription factor

TpG — a dinucleotide context of 5'—thymine—phosphate—guanine—3'

SVM — support vector machine

General introduction

Biological importance of nucleotide sequence

Since the discovery by Frederick Griffith in 1928, it is now beyond doubt that DNA is *the* hereditary molecule that convey genetic information in virtually all organisms. The subsequent unravelling of DNA structure by Watson and Crick in 1953, as well as the complete decryption of genetic codes by Nirenberg and Leder in 1964 set on an avalanche of researches on this simple (as a linear polymer composed of only four types of nucleotides as repeating unit), yet highly complicated (as the arrangement, i.e. sequence, of nucleotides serves as cryptogram of genetic information), molecule. As is stated by Frederick Sanger, the Noble Prize laureate who invented the revolutionary “plus and minus” DNA sequencing method (a.k.a. “chain termination” or simply Sanger method), in his bibliography (Sanger 2005), “… *knowledge of sequences could contribute much to our understanding of living matter.*” Indeed, the “central dogma of molecular biology” put forward by Francis Crick in 1958 laid down the core mechanistic framework that links nucleotide sequences to biological phenomena. This classical dogma stated that (1) heritable genetic information is hard-coded into the DNA molecule as specific sequences/arrangement of nucleotides, (2) when the information is retrieved, the DNA nucleotide sequence is transcribed into transportable RNA molecule that carries essentially the same nucleotide sequence (except that uracil is used in place of thymine on the RNA transcript), and (3) the nucleotide sequence in the RNA cassette (which mirrors the DNA template) is then translated into a corresponding chain of amino acids, i.e. peptide/protein, which forms specific cellular structure or catalyses specific biochemical reaction that leads to specific biological traits.

Given the postulated direct governance of phenotype by the nucleotide sequence, most of the genetic researches in the past decades have a primary focus of unravelling which specific stretches of nucleotide sequence (e.g., genes) along the genome determine which biological traits.

Epigenetics: the bridge between genotype and phenotypes

It is, however, incontestable that nucleotide sequence by itself is not sufficient for explaining many fundamental aspects of biology. With the rapid technological advancement since the beginning of the 21st century, we can now easily unravel the nucleotide sequence of the entire genome of an organism within hours. The bloom of genome information has given rise to a whole new, exciting era of biological science, and greatly accelerates genetic and medical researches. However, in spite of all the technological breakthroughs, it remains poorly understood that how exactly a single genome (i.e. from one fertilized egg) can give rise to the diverse cell types that constitute the body of a multicellular organism. While the genome has been praised as "the blueprint of life", it is incontestable that complete decryption of the "blueprint" requires a lot more than obtaining only the genomic sequence.

The recent conception of epigenetics has provided a crucial link between the static genomic sequence and the dynamics in phenotype (Dunham et al. 2012). Under the definition that "*[epigenetics is] the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence*" (Bird 2002), epigenetics accounts for all mechanisms that can lead to heritable change to the accessibility, hence usability and expression state, of different parts of the genome without incurring changes to the underlying nucleotide sequence. Examples of epigenetic mechanisms includes DNA methylation, histone modifications, spatial organization of chromatin, to name but a few. It is now known that different cell states/types acquire different epigenetic landscapes (a.k.a. "epigenomes") along differentiation, and hence can only utilize different subsets of the same genome, resulting in different phenotypes (Waddington 1942, reprinted in 2012; Goldberg et al. 2007). The existence of epigenetic regulatory mechanisms on top of the genome has provided an elegant solution to the aforementioned enigma that the very same genome can result in the high level of heterogeneity and complexity as observed among individual cells within a multicellular organism.

The epigenetic element under spotlight: DNA methylation

Among all currently known epigenetics mechanisms, DNA methylation is by far the most intensively studied and best understood epigenetic element (Beisel and Paro 2011), particularly in vertebrates. DNA methylation is the natural occurring, enzymatically catalysed, covalent attachment of methyl group ($-CH_3$) onto DNA molecules. While organisms along the evolutionary tree display diverse DNA methylation landscapes, higher eukaryotes within a same clade generally possess highly conserved genome-wide methylation pattern (Noyer-Weidner and Trautner 1993; Colot and Rossignol 1999). Vertebrate animals, in particular, demonstrate peculiar DNA methylomes that have drawn most of the research focus for the past two decades.

In a vertebrate genome, cytosines within the CpG (i.e. 5'-CG-3') dinucleotide-context are extensively methylated, while unmethylated CpGs are almost exclusively found clustered at high density inside gene regulatory elements, such as promoters and enhancers. Having an extensively methylated genome is paradoxical since methylated cytosine is prone to mutation into thymine through spontaneous deamination (Holliday and Grigg 1993). It was evidenced that the vertebrate genomes have undergone extensive loss of CpGs over the course of evolution due to the mutagenic nature of cytosine methylation (Bird 1980). Since mutation is often detrimental, it is generally believed that the hypermethyated state of vertebrate genomes must confer critical biological functions that outweigh the elevated risk of mutation (Bird 1980).

Intensive researches for the past two decades suggest that gene expression is tightly linked to the DNA methylation state within gene regulatory regions. Mechanistic studies have proven that DNA methylation governs gene expression by preventing binding of methylation-sensitive transcription activators, attracting transcription repressors, as well as inducing stable compacting of chromatin (i.e. heterochromatin) that precludes the access of transcription machinery (Siegfried and Cedar 1997). DNA methylation was also shown to be crucial in suppressing the activity of transposable elements, which constitute a significant portion of the vertebrate genomes (Slotkin and Martienssen 2007), and, thereby, help maintain genome stability. The importance of DNA methylation in vertebrates is further highlighted by the study of Li et al. (1992) that genome-wide depletion of DNA methylation in mice, via knockout of DNMT1 (i.e. DNA methyltransferase 1; Also known as the "maintenance" enzyme that copies methylation state from the template strand to nascent strand during DNA replication), is lethal in early embryogenesis. Moreover, aberrant changes in methylation pattern are commonly found associated with cellular dysfunction, developmental abnormality, and a spectrum of diseases (Robertson and Wolffe 2000; Egger et al. 2004). It is now beyond doubt that DNA methylation is of vital importance to vertebrates that it directly governs cell state and identity.

Furthermore, DNA methylation is, at the time of this writing, the only known epigenetic element that can be faithfully inherited (in a semiconservative manner as the underlying DNA) during DNA replication. This unique property allows DNA methylation to precisely convey epigenetic information across cell division and along cell lineage. Given its inheritability, it is believed that DNA methylation may serve as the ultimate epigenetic mark that restores or guides the establishment of other epigenetic signatures after cell division (Jin et al. 2011). Consequently, unravelling the logic of DNA methylation patterning is fundamental to thorough understanding of the establishment of- and regulation on the entire epigenome.

Establishment and regulation of DNA methylome

Given its vital importance, the methylation landscape has to be precisely specified and modulated. The DNA methylation pattern is established and maintained through highly dynamic biological processes, in which the methylome undergoes substantial, yet precise, changes. For instance, differentiating cells faithfully acquire specific methylation landscapes that are unique to their committed cell types (Spivakov and Fisher 2007; Mohn et al. 2008; Houseman et al. 2012). Remarkably, in human and mice, the DNA methylomes are extensively erased and fully reconstituted during gametogenesis and early embryonic development (Wu and Zhang 2010; Kohli and Zhang 2013). These facts suggest that the methylation landscape is pre-defined by genetic information. Deciphering how the methylation pattern is encoded is a prerequisite for understanding of differentiation processes and the pathogenesis of various diseases. Yet, by what means the methylation pattern is defined and established *in vivo* (i.e. inside a living vertebrate animal) remains enigmatic.

Given that CpG dyads are predominantly methylated unless they are clustered at high density (Lister et al. 2009), it is generally presumed that hypermethylation is the "default" state of vertebrate genomes and specific regions (i.e. gene regulatory elements) are protected from *de novo* methylation, rendering them hypomethylated (Bird 2002; Edwards et al. 2010; Lienert et al. 2011; Reeve and Black 2013; Schübeler 2015; Takahashi et al. 2017). Intensive researches for the past decade have demonstrated that the protection on the genomic loci is possibly mediated by nucleosome positioning (Jones and Liang 2009; Jones 2012; Baubec et al. 2015) and/or the recruitment of a myriad of proteins (C. Xu et al. 2011; Y. Xu et al. 2011; Wu et al. 2012; Kohli and Zhang 2013; Marchal and Miotto 2015; Castillo-Aguilera et al. 2017) which eventually (1) block off local access of DNA methyltransferases or (2) remove methylation on cytosines in vicinity through oxidation and thymine DNA glycosylase (TDG)-mediated base excision repair. However, little is known about how these factors are specifically predisposed on the preselected loci.

Recent researches proposed that DNA methylation pattern is ultimately governed by local sequence context. In particular, *in silico* analyses asserted that there is strong statistical association between sequence variants and differential DNA methylation states in vertebrates, from fish (Uno et al. 2016) to human (Kaminsky et al. 2009). A number of recent *in vitro* studies, which utilized mammalian cell lines that had been stably maintained in dishes or flasks, further demonstrated that high CpG density and/or the presence of specific transcription factor binding sites is capable of autonomously determining local hypomethylation in the globally methylated genome (Lienert et al. 2011; Stadler et al. 2011; Krebs et al. 2014; Takahashi et al. 2017). These recent *in silico* and *in vitro* reports propel the notion that DNA methylation pattern is primarily and autonomously determined by local sequence context (Lienert et al. 2011). However, the anticipated sequence-dependency of DNA methylation is in contradiction to the pioneer *in vitro* experiments in early days of DNA methylation researches (Pollack et al. 1980; Wigler 1981; Stein et al. 1982), in which the methylation status of exogenous DNAs (either artificially CpG-methylated or completely unmethylated) was found maintained for many cell generations upon stable genome integration. Given the opposing results, whether or not the DNA methylation landscape is autonomously determined by the underlying genomic sequence appears to be less concrete than what has recently been proposed. Moreover, the postulated relationship between DNA methylation landscape and local sequence context has never been rigorously validated *in vivo*.

Experimental outline

Given the above contentious *in vitro* findings and a void of *in vivo* proof for the postulated sequence-dependency in the patterning of DNA methylation landscape, this research project was conceived with a primary objective to rigorously test the governance of genomic sequence on its own DNA methylation state *in vivo*. The small laboratory fish, medaka (*Oryzias latipes*), was chosen as an experimental model for their relatively small genome size (approx. 700 Mbp), ease of *in vivo* genetic manipulation, oviparity, in addition to their capability of producing 10–20 fertilized embryos per pair on daily basis (Kinoshita et al. 2009; Kirchmaier et al. 2015). These traits enabled the use of large number of embryos for *in vivo* experimentation (Chapter 2), which is hardly feasible with conventional mammalian models, such as rodents. The blastula stage was specifically chosen as the endpoint in all experiments throughout this study due to its homogenous cell type composition (i.e. mostly pluripotent cells) and the ample amount of genetic materials (2000–4000 diploid cells, i.e. 4000–8000 copies of the genome, per embryo) available for precise methylation state measurements. The experiments, as documented and structured in this dissertation, aimed to:

1. (Chapter 1) demonstrate that the medaka genome also displays the characteristic, strong statistical association between DNA methylome and genomic sequence as observed in aforementioned mammalian-based *in silico* studies. This observation not only validates the medaka as a suitable vertebrate model for the study on sequence-dependency of DNA methylation, but also highlights that the anticipated linkage between the DNA methylome and the underlying genomic sequence is a conserved feature along the evolution of vertebrates, from fish to mammals. And,
2. (Chapter 2) scrutinize whether genomic sequences can autonomously determine their own DNA methylation state via large-scale genetic engineering, including the use of genome and methylome editing, *in vivo*.

As presented herein, the results suggest that nucleotide sequence, by itself, cannot dictate its own methylation state *in vivo* in medaka, which apparently contradicts the prevailing view of DNA methylation in vertebrates and call for a new wave of pursuit of factors that genuinely determine the vertebrate DNA methylome.

Chapter 1 Statistical association between genomic sequence and DNA methylation state *in vivo*

1.1 Introduction

Since the dawn of bioinformatics, there has been unceasing endeavours and ever-accumulating informatics tools crafted to help predict or identify the functional role of any particular segments of genome *in silico*. Many of these computational tools could predict the locations of genes (Mathe 2002; Stanke and Waack 2003; Mount 2004; Wasserman and Sandelin 2004) and their regulatory regions (Pedersen et al. 1999; Scherf et al. 2000; Bajic et al. 2002; Noble et al. 2005; Fiedler and Rehmsmeier 2006; Su et al. 2010; Lee et al. 2015) with satisfactory accuracy and sensitivity, greatly facilitating further functional characterization of the genomes.

The success behind all these prediction attempts is underpinned by the fact that genomic loci that serve similar functions frequently share similar nucleotide sequence patterns. For instance, many eukaryotic promoters contain binding sequences of the general transcription factors (TFs) that involve in transcription initiation, e.g., TATA element (canonical sequence: 5’-TATAAA-3’) that is recognized by TATA-binding protein; B recognition element (canonical sequence: 5’-SSRCGCC-3’, where S = C or G, R = A or G) that is recognized by Transcription Factor II B; E-box (canonical sequence: 5’-CACGTG-3’) that is recognized by a wide range of TFs of the basic helix-loop-helix (bHLH) family. Similarly, the exon-intron boundaries inside most open reading frames are characterized by the presence specific sequences, e.g., consensus sequences of 5’-MAGGTRAGT-3’ (where M = A or C) and 5’-CAGG-3’ in the 5' and 3’ splice sites, respectively, of all U2 introns (the major class of intron). In other words, conserved patterns of nucleotide sequence are repeatedly and extensively used along the genome to encode conserved, pre-specified genetic information.

The DNA methylation states in multiple mammalian cell types have also been repeatedly shown to strongly correlate with, hence highly predictable by, underlying nucleotide sequences (Bhasin et al. 2005; Rollins 2005; Bock et al. 2006; Das et al. 2006; Fang et al. 2006; Wang et al. 2006; Bock and Lengauer 2008; Fan et al. 2008; Kim et al. 2008; Previti et al. 2009; Xie et al. 2012; Zhou et al. 2012; Zheng et al. 2013; Ma et al. 2014; Zhang et al. 2015; Angermueller et al. 2017). In particular, locally high CpG density is proven to be the molecular signature of hypomethylated regions (Krebs et al. 2014; Takahashi et al. 2017). The widely observed statistical linkage between DNA methylation state and sequence forms the logic basis of the general belief that local methylation state is encoded by nucleotide sequence (Bock et al. 2006; Edwards et al. 2010).

However, such dependency between DNA methylation and sequence was inferred from the analyses of human and mouse genomes only and has never evidenced in any non-mammalian species. From the evolutionary standpoint, the acquisition of genome-wide hypermethylation is the hallmark event that underpinned the advent of vertebrates from their invertebrate ancestors (Bird 1995). Coupling with the fact that the overall genome-wide methylation landscape appears to be highly similar among vertebrate species (globally methylated with specific gene regulatory regions rendered hypomethylated), the molecular logic underlying the patterning of DNA methylome is also likely conserved through vertebrate evolution. Yet, it remains uncertain whether different clades of vertebrates (i.e. fish, amphibians, reptiles, bird, and mammals) share similar degree of sequence-dependency in methylation state determination. Such uncertainty severely undermines the generalisability of the notion that DNA methylome is defined by genomic sequence *in vertebrates*.

In order to evident the expected association between genomic sequence and DNA methylation in non-mammalian vertebrate species, the statistical linkage between the genome and the methylome of medaka (i.e. a teleost) was determined. Specifically, differential sequence composition in hypomethylated (a.k.a. hypomethylated domains, "HypoMDs") and hypermethylated loci (a.k.a. hypermethylated domains, "HyperMDs") in the medaka genome is revealed and modelled by machine learning. Support vector machine (SVM) was chosen for this purpose since it is by far the most frequently used machine learning algorithm for the classification/prediction of DNA methylation state with respect to nucleotide sequence and has shown consistently high accuracy and sensitivity (Bhasin et al. 2005; Bock et al. 2006; Das et al. 2006; Fang et al. 2006; Fan et al. 2008; Previti et al. 2009; Zhou et al. 2012; Zheng et al. 2013; Ma et al. 2014).

The results described in this chapter ascertain that the strong correlation between local DNA methylation state and the underlying nucleotide is not restricted to mammals, but can also be found in a teleost (i.e. the medaka). This finding suggests that the expected coupling of nucleotide sequence and its methylation state is a conserved genomic feature among the clade of vertebrates.

1.2 Results

1.2.1 HypoMDs and HyperMDs contain distinct sequence patterns

Statistical association between medaka genomic sequences and local methylation states was modelled using support vector machine (kmer-SVM) (Fletez-Brant et al. 2013). HypoMDs and HyperMDs at the blastula stage (Stage 11 according to Iwamatsu, 2004) were identified by the same criteria as described by Nakamura et al. (2014) (see also Figure 1.1, as well as Materials and Methods). While HypoMDs and HyperMDs are not readily discernible in terms of length (Figure 1.2A) and GC composition (Figure 1.2B), they bear conspicuous difference in their sequence pattern, allowing robust classification and prediction of the methylation states based solely on nucleotide sequence information (Figure 1.1: "SVM classification" track, c.f. "HyoMD" and "HyperMD" tracks; Figure 1.3A: area under precision-recall curve ≥ 0.83, versus 0.08 from the random classifier; Figure 1.4A: maximum Matthew's correlation coefficient = 0.76).

Since HypoMDs have a higher average CpG density than HyperMDs (Figure 1.2C), CpG density might act as a confounding factor that outweighs and conceals non-CpG-containing sequence features. The impact of CpG density was, hence, controlled for by masking all CpG dinucleotides (i.e. from 'CG' to 'NN') and SVM models were retrained. Conspicuously, CpG-masking could still result in models with modest classification performance (Figure 1.3B: area under precision-recall curve ≥ 0.53, versus 0.08 from the random classifier; Figure 1.4B: maximum Matthew's correlation coefficient = 0.51), suggesting that CpG-free DNA motifs are also differentially enriched in HypoMDs and HyperMDs.

1.2.2 Consecutive CpG and CGnull-repeat are strongly enriched in HypoMDs

The weight of individual *k*-mers (i.e. 6-mer in this case) in kmer-SVM model directly reflects the relative importance of the corresponding sequence in the classification/prediction of the methylation states. Since the weight of each of the *k*-mers is mathematically assigned as a function of relative enrichment in HypoMDs versus that in HyperMDs, these values were used to identify consensus patterns that are overrepresented in HypoMDs. Of all 2080 possible canonical 6-mers, those with *consecutive* CGs were highly overrepresented in HypoMDs (odd ratio = 9.54; Figure 1.5: the 3rd – 4th lanes versus the 2nd lane; see also Table 1). On the other hand, as revealed by SVM modelling after CpG masking, specific 6-mers that contain no CpG are also found highly enriched in HypoMDs (Table 1: right columns). Interestingly, the top most enriched, CpG-free 6-mers are all derived from the same, but shifted, repeat pattern of $(AGCT)_n$, where n ≥ 1.5 (i.e. ≥ 6 bp): AGCTAG, GCTAGC (palindromic), CTAGCT (a reverse complement of AGCTAG), or TAGCTA (palindromic) (odd ratio = 2.50; Figure 1.5: the 6th lane; see also Table 1). For the sake of simplicity, these 6 bp-long, CpG-free repeating sequences are collectively referred to as “CGnull-repeats” from here onwards.

1.2.3 Only a small subset of the enriched motifs could be linked to transcription factor binding sites

To further identify whether consecutive CpGs and CGnull-repeats could be the recognition/binding sequences of transcription factors, the 6-mers that are highly enriched in HypoMDs and contain consecutive CpG or CGnull-repeat were matched against databases of vertebrates’ transcription factor binding sites using TomTom (part of the MEME Suite) (Gupta et al. 2007). However, among all 56 valid 6-mers (52 with CGCG, 1 CGCGCG, and 3 CGnull-repeats), only 7 motifs (GCGCGC, CGCGCG, CGCGGA, CGCGCA, CCGCGG, CCCGCG, and CGCGCC) could be confidently mapped to 3 known transcription factor binding sites: ZBTB14, E2F2, and E2F3 (Figure 1.6). CGnull-repeats could not be matched to any known transcription factor binding site regardless of statistical stringency.

1.3 Discussion

This study is the first to demonstrate that there is the strong statistical association between nucleotide sequences and local methylation states in a non-mammalian genome. By means of statistical modelling, it is clearly shown that the HypoMDs and HyperMDs in medaka genome are characterized by distinctly different sequence patterns, allowing highly precise and sensitive *in silico* classification/prediction of local methylation state based solely on nucleotide sequence information. This indicates that the anticipated linkage between DNA methylome and the underlying genomic sequence is not exclusive to mammals, but also exists in teleost, hence possibly a conserved genomic feature among vertebrate animals.

While the findings as presented herein fully agree with all previous *in silico* studies on mammalian genomes that high CpG density is a common feature of HypoMDs, the above SVM results further highlight that the "high CpG density" is frequently manifested as *consecutive* CpGs in the medaka genome. Such feature is not unique to medaka. A previous study on mice oocytes' methylome (Saadeh and Schulz 2014) also showed that genomic regions resistant to *de novo* methylation are enriched with the 5'-CGCGC-3' motif. The authors further demonstrated that this motif was indeed bound by E2F1/2 transcription factors, which concurs with the above *in silico* motif analyses. While direct and mechanistic relationship between E2F family proteins and DNA methylation is yet to be identified, E2F1–3 are known to be capable of inducing H3 and H4 acetylation, as well as indirectly interacting with SWI/SNF chromatin remodelling complex, which, in both cases, can lead to protection of local DNA from *de novo* DNA methyltransferases (Saadeh and Schulz 2014), resulting in a local hypomethylated state.

Intriguingly, despite all possible DNA sequence patterns that contain consecutive CpGs are highly enriched in HypoMDs, only a small subset of those motifs is known/predicted to be associated with TF(s). While this could be due to the fact that the binding sequences have not been exhaustively characterised for all TFs, this might also indicate that the chained CpGs may play a role other than as a recognition target of TFs. Indeed, it has long been recognized that a succession of CpG has critical impacts on the local conformation of DNA molecule by destabilizing the B-form (the most common double helical structure of DNA) and favouring the formation of Z-form DNA (Drew and Dickerson 1981; Tran-Dinh et al. 1983; Shakked and Rabinovich 1986; Peticolas et al. 1988). Z-form DNA has been found associated with transcriptional initiation and is stably bound by a myriad of DNA binding proteins with high affinity (Rich and Zhang 2003). Since the biological functions of Z-DNA remain poorly studied, it is largely unknown if Z-DNA can directly regulate local DNA methylation. However, it is possible that the binding between Z-form DNA and its recognising proteins can physically hinder the local attachment of DNA methyltransferases, thereby protecting the local region from methylation.

The above SVM results also highlight that certain CpG-free DNA motifs, especially the CGnull-repeats, might be associated with the hypomethylated state in HypoMDs. As introduced in *General Introduction*, methylated cytosines are prone to spontaneous mutation into thymine. It has been reasoned that the natural conversion of methylated CpG into TpG (or CpA on the reverse strand) over an evolution time scale eventually led to a depletion of CpGs in HyperMDs, while unmethylated CpG within HypoMDs were not affected (Bird 1980). Consequently, the widely observed "enrichment" of CpGs within HypoMDs could possibly be a statistical illusion caused by the historical loss of CpGs elsewhere in the genome (i.e. methylation states caused the relative "enrichment" of CpGs in HypoMDs, but not the other way around), opposing the recently postulated role of high CpG density as a determinant of local hypomethylation. However, this argument cannot explain the observed enrichment of non-CpG-containing DNA motifs in HypoMDs as cytosines outside the context of CpG dinucleotide are rarely methylated, and, thus, not susceptible to deamination and conversion into thymine. Hence, the overrepresentation of CpG-free sequence patterns inside HypoMDs reinforces the notion that local methylation state is closely linked to the underlying sequence pattern. Nonetheless, the biological function(s) of these non-CpG-containing patterns remain obscured. In fact, this study is the first report on the statistical enrichment of CpG-free sequence patterns, in particular the CGnull-repeats, in HypoMDs of a vertebrate genome. The biological roles of these peculiar DNA motifs, as well as their association with HypoMDs, deserve future investigation.

1.4 Chapter conclusion

By the use of statistical modelling, the DNA methylation landscape of medaka genome was unveiled to be strongly associated with the underlying genomic sequence. HypoMDs and HyperMDs carry distinctly different sequence patterns that cannot be completely accounted for by their differential CpG density. In particular, consecutive CpGs and CGnull-repeats were found highly enriched in HypoMDs, suggesting their relationship with the local hypomethylated state. The above results concur with the general belief that DNA methylation landscape is primarily determined by genomic sequence. However, it should be noted that statistical association does not necessary imply causal relationship. Therefore, to scrutinize the postulated dependency of DNA methylome on genomic sequence *in vivo*, a series of experiments was specifically and carefully designed and conducted as documented in the next chapter.

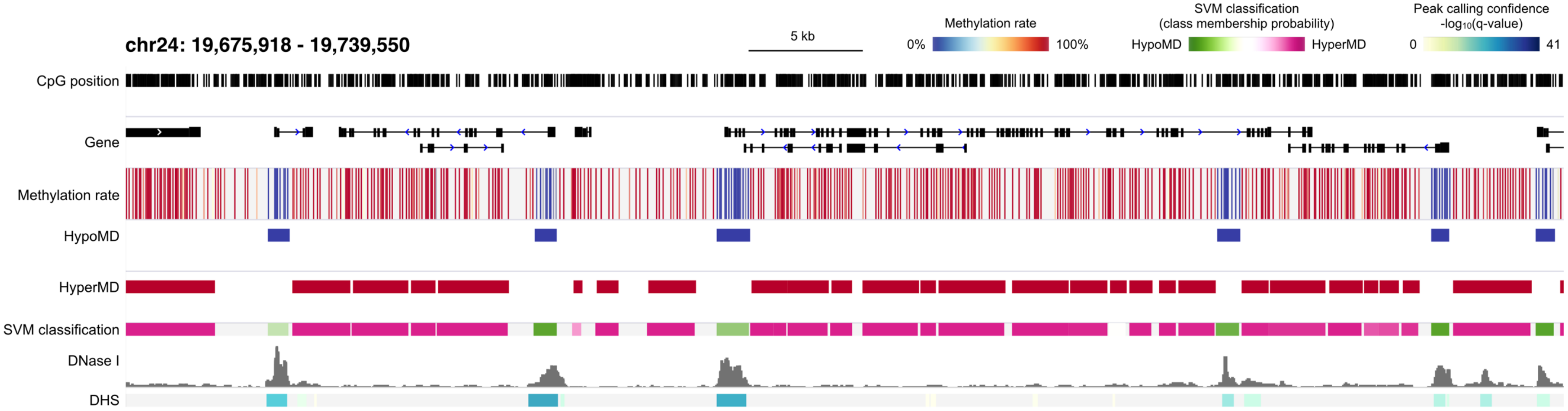

Figure 1.1 Genome browser view of a representative locus (approx. 62 kb) in the HdrR medaka genome showing CpG methylation rate, called HypoMDs and HyperMDs, SVM classification results, as well as DNase I hypersensitivity and called DNase I hypersensitive sites (“DHS”).

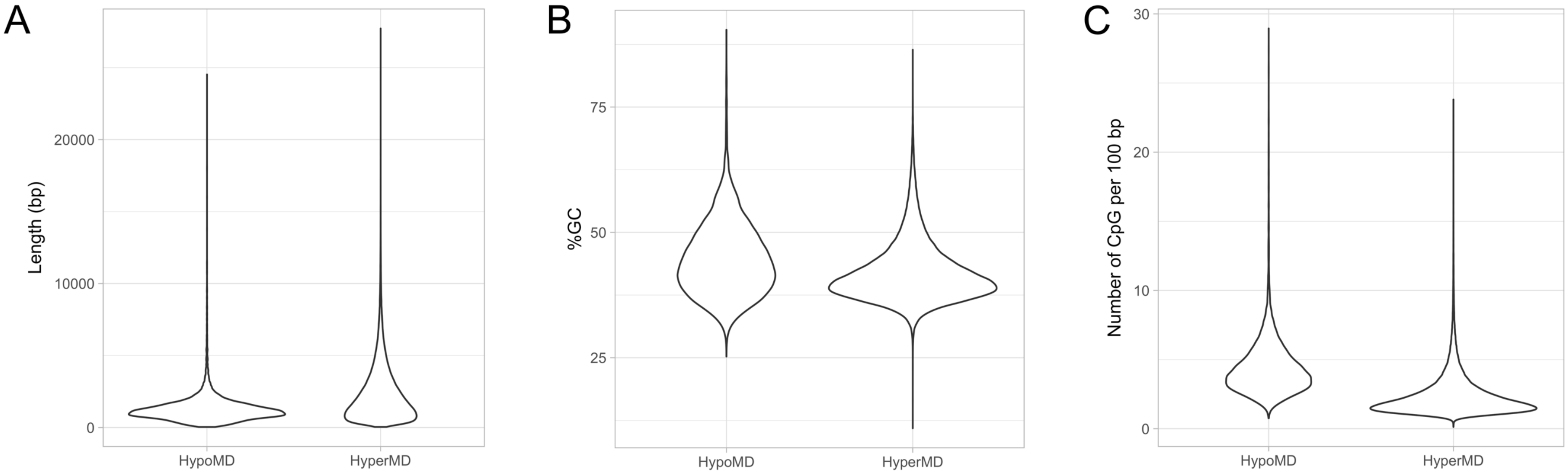

Figure 1.2 Violin plots showing the distribution of (A) length, (B) GC content, and (C) CpG density of HypoMDs and HyperMDs.

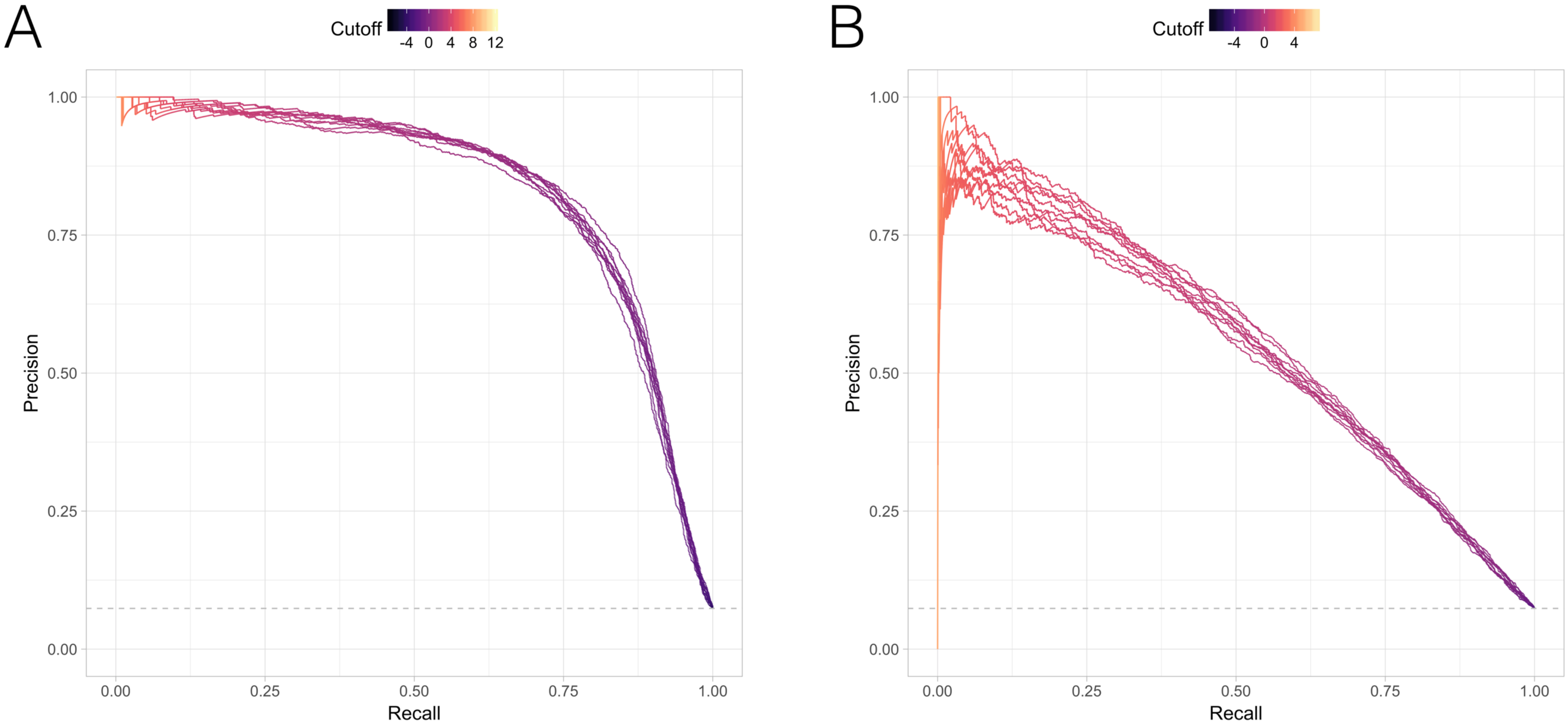

Figure 1.3 Precision-recall curves of the kmer-SVM models trained for binary classification of HypoMDs and HyperMDs.

(A) Without- or (B) with- CpG-masking. HypoMD and HyperMD sequences were assigned to positive and negative classes, respectively. Solid, colored lines are individual precision-recall curves derived from 10-fold cross-validation. The colors represent the cut-off values for binary classification. Area-under-curve (AUC): (A) minimum = 0.83, maximum = 0.84; (B) minimum = 0.53, maximum = 0.56. Random classifier is represented by horizontal dashes at the bottom of both panels and has an AUC of 0.08.

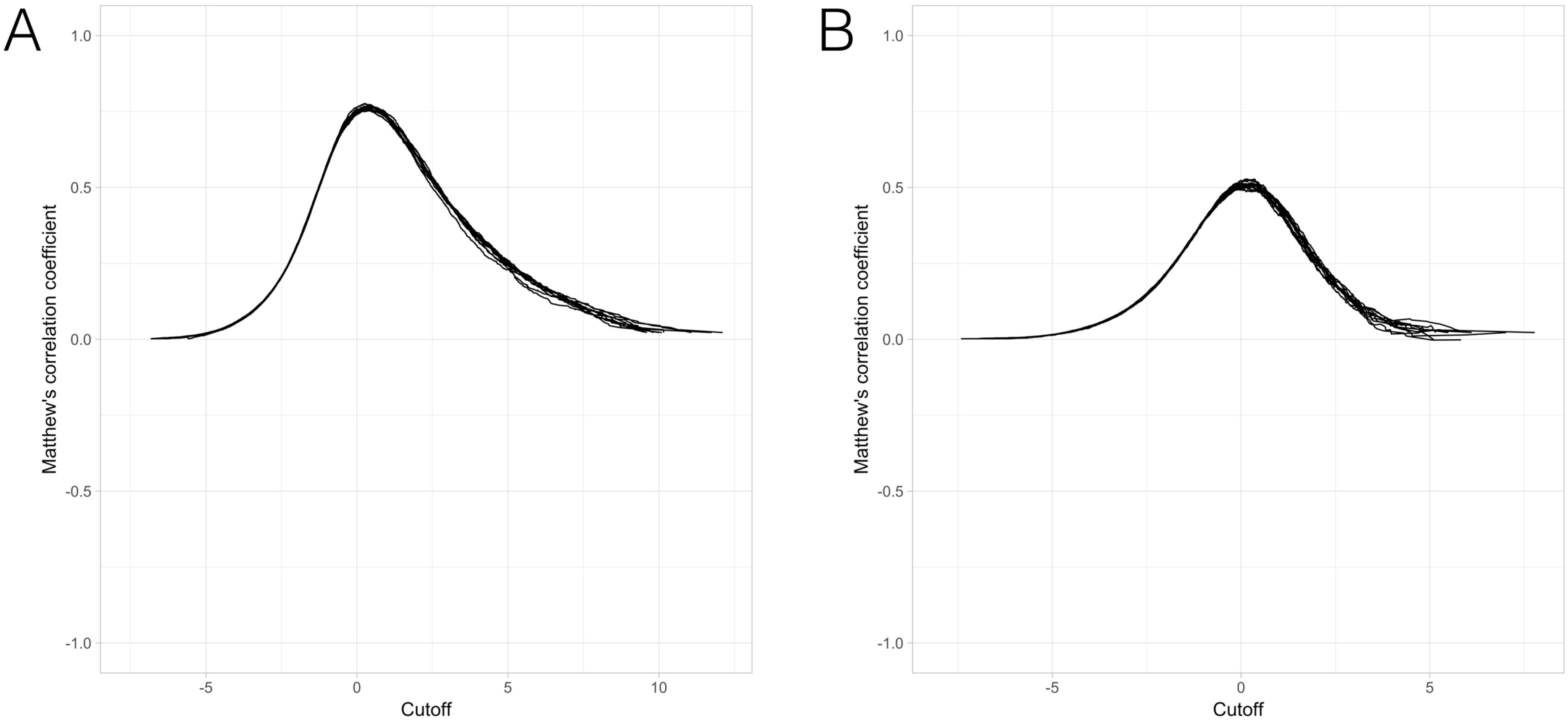

Figure 1.4 Matthew’s correlation coefficients (MCC) of the kmer-SVM models trained for binary classification of HypoMDs and HyperMDs

(A) without- and (B) with- CpG-masking. HypoMD and HyperMD sequences were assigned to positive and negative classes, respectively. Lines represent the MCC derived from individual round of 10-fold cross-validation across various cutoff values. Maximum MCC attained: (A) 0.76, (B) 0.51.

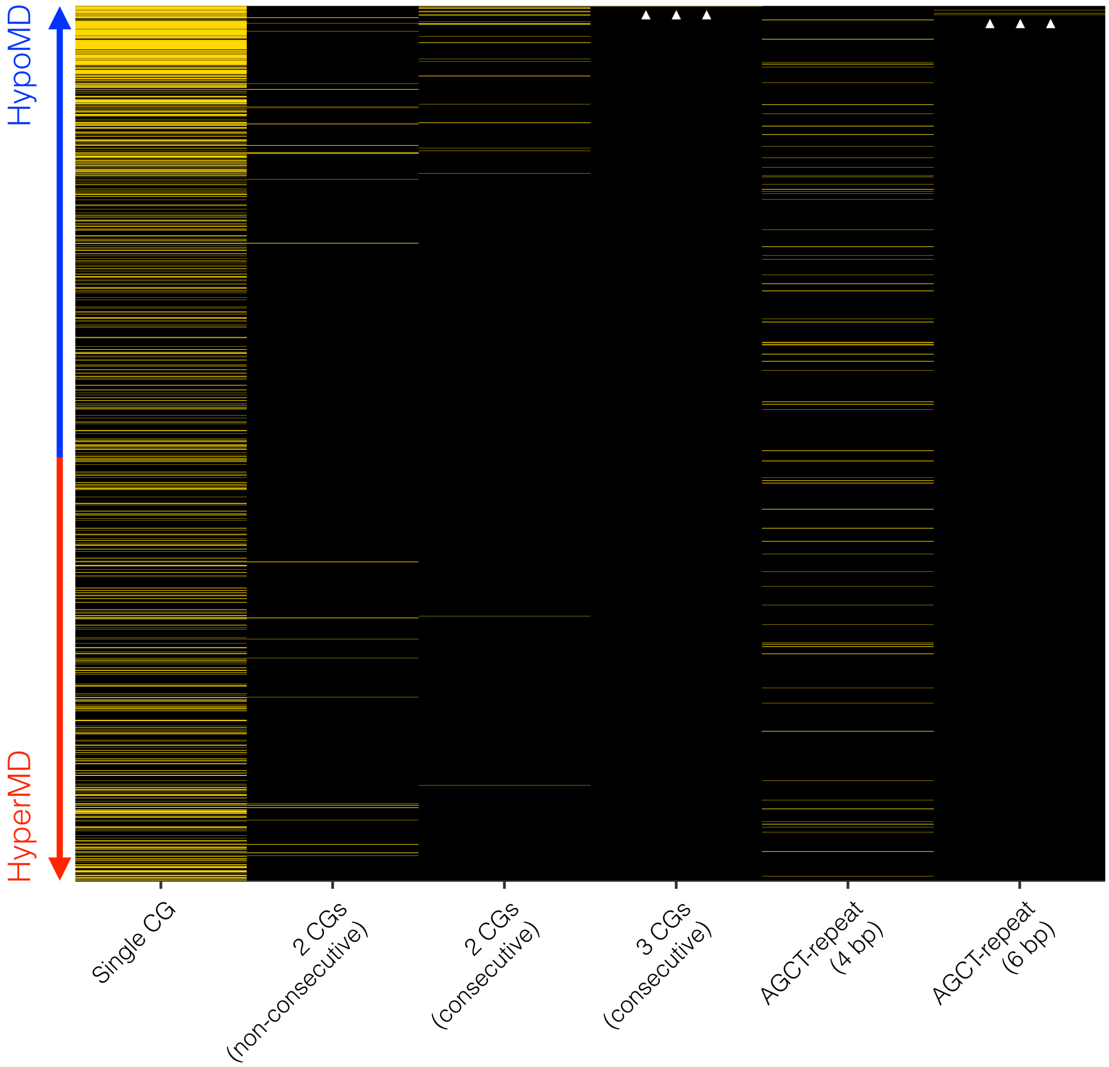

Figure 1.5 Compact representation of CpG and CGnull-repeats enrichment inside all 2080 possible canonical 6-mers.

The 6-mers (i.e. individual rows) are sorted vertically according to their importance (i.e. SVM weights) in classifying/predicting HypoMDs and HyperMDs. The 6-mers that are more important (i.e. higher absolute weight) in the classification/prediction are ranked towards the top and bottom, respectively. The ranked 6-mers (rows) are annotated according to their CpG and CGnull-repeat content with golden lines in the applicable columns. For instance, the 6-mer "CGAACG" (having 2 non-consecutive CpGs) would be highlighted gold in the 1st and 2nd column, whereas "AACGCG" (having 2 consecutive CpGs) would be highlighted gold in the 1st and 3rd column. Note that 6-mers with consecutive CpGs (the 3rd and 4th column) are highly enriched in HypoMDs. White arrowheads in the 4th and 6th lane denote the golden lines that represent $(CG)_3$ and CGnull-repeats, respectively.

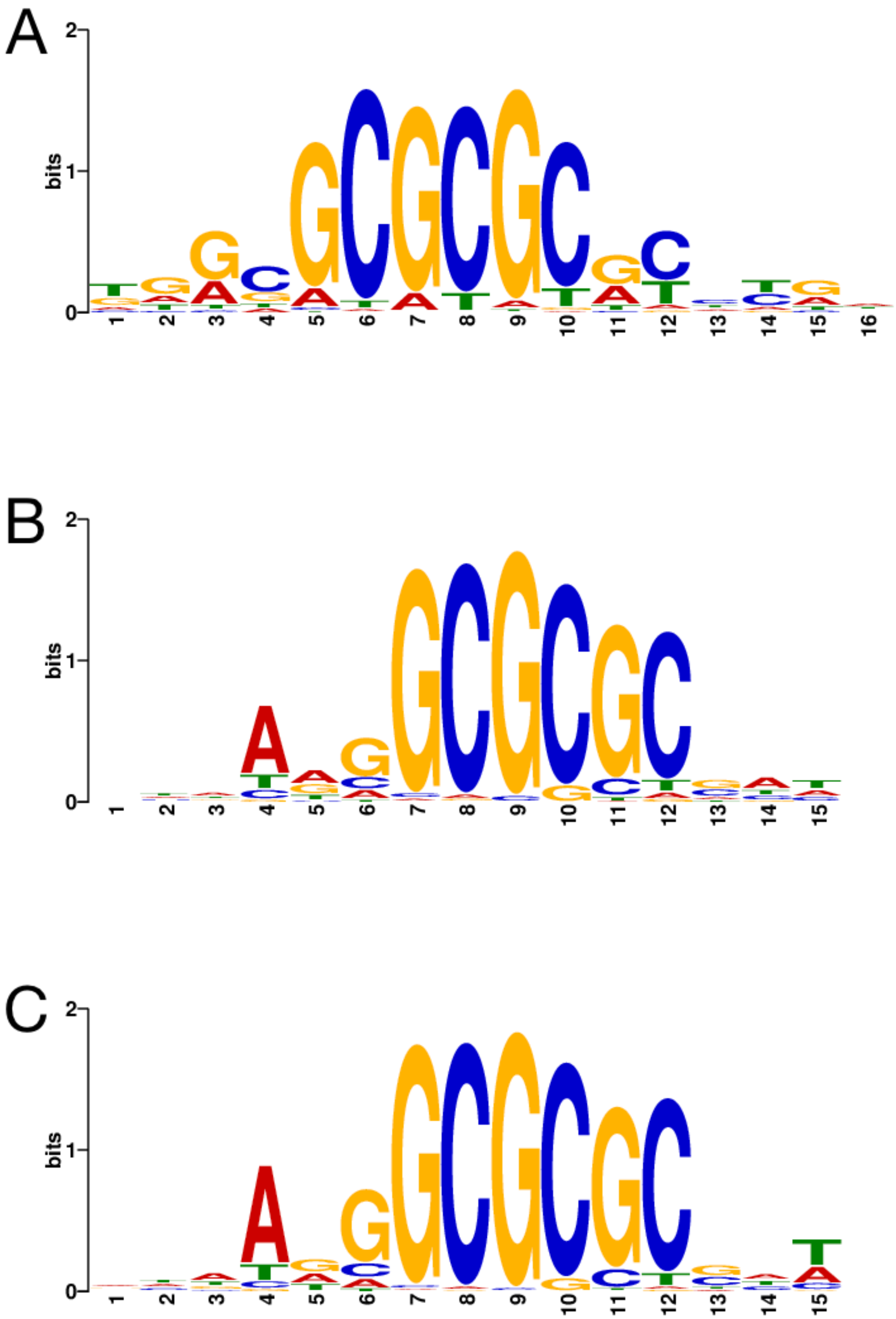

Figure 1.6 Binding sequence of (A) ZBTB14, (B) E2F2, and (C) E2F3.

Note that all three motifs contain a series of interleaving cytosine and guanine at the center.

Table 1 The 30 canonical 6-mers with the highest absolute weights in the SVM model trained for classifying/predicting HypoMDs and HyperMDs.

The lists are sorted according to the absolute weights in descending order. Solitary and consecutive CpGs are coloured orange and red, respectively. The 6-bp CGnull-repeats, i.e. AGCTAG, GCTAGC, CTAGCT (reverse complement of AGCTAG), and TAGCTA, are coloured purple.

Without CpG-masking				CpG-masked			
Weighed towards HypoMDs		Weighed towards HyperMDs		Weighed towards HypoMDs		Weighed towards HyperMDs	
kmer	Weight	kmer	Weight	kmer	Weight	kmer	Weight
GCGCGC	9.05	ACGCCA	-4.11	GCTAGC	10.54	AAATTT	-7.06
AAAACG	8.59	AAACGG	-4.06	AGCTAG	9.26	AAATTG	-5.27
ACGCGC	8.21	AAACGT	-3.82	TAGCTA	6.75	AAAAAT	-4.31
CGCGCG	8.11	CTCGCC	-3.61	GCTAAC	6.25	AAATTC	-4.01
ATGCGC	6.86	AAACGC	-3.24	CAAAAA	5.02	AAGCTT	-3.78
CGCGAG	6.59	CACTAC	-3.09	ACTTAC	4.75	ATATTG	-3.56
CGTTTA	6.46	ATCGCA	-2.74	AGCTAA	4.54	CCCTAG	-3.51
CGCGGA	5.88	GGCCCA	-2.69	GAAAAA	4.28	GCTAGA	-3.41
GCGCGA	5.80	ATACGC	-2.68	ACTCAC	4.27	AAAAAC	-3.32
CGTTTC	5.64	CAACCG	-2.64	CTTACC	4.24	GCCCTA	-3.25
TGCGCA	5.36	AGCGTA	-2.59	AAGAAG	3.66	AAAAAG	-3.19
CAAACG	5.27	CCCCCA	-2.54	TTTAAA	3.45	AAGGCC	-3.17
GCTAGC	5.24	CCCTAG	-2.43	CTAGCA	3.45	AAGCAC	-3.15
GCGCAC	5.08	AGAGCG	-2.39	AATTTA	3.26	CATGTA	-3.07
CGCGCA	5.00	AACCGT	-2.36	AGGTAA	3.24	GAGCTA	-3.06
CGGAAG	4.98	ACCGTA	-2.32	TTAAAA	3.23	GACATA	-3.05
ACGCGG	4.92	AAGGCC	-2.28	ATCATG	3.19	ACTATG	-3.03
TACGCA	4.87	GCTCTA	-2.24	TAAAAA	3.15	GCTCTA	-2.98
GCGTAA	4.66	CGCCTA	-2.24	CTAAAA	3.12	CACTAC	-2.95
AGCTAG	4.65	GGGTCA	-2.23	AGGGGG	3.09	AGGTCA	-2.94
CCGGAA	4.54	AAATTT	-2.19	AATGAA	3.05	AGGCCC	-2.91
CACGTG	4.48	ATGGGG	-2.19	AGGCTA	3.01	GAGGCC	-2.85
CCGCGG	4.44	CTCGTA	-2.19	GAATAA	2.99	ACATTG	-2.81
TAGCTA	4.35	ACGGAT	-2.17	AGCAAA	2.98	AAGCCT	-2.81
CCGCGC	4.33	CACGTC	-2.16	CTTAGC	2.92	CATTTC	-2.76
CTGCGC	4.27	CGCTCA	-2.16	AGGCTG	2.90	CACTGG	-2.72
CTTACC	4.24	AACCCT	-2.14	GGAAAA	2.89	AACCTT	-2.69
GCGCAA	4.22	AAGCCG	-2.12	AGTGAA	2.77	ACATGC	-2.68
AATTCG	4.17	GTGTGA	-2.11	TCAAAA	2.73	CCCATA	-2.67
GCGGAA	4.11	GGCACC	-2.11	CTCACC	2.68	CTAGGA	-2.67

Chapter 2 Interrogation of the autonomy in DNA methylation determination by genomic sequence

2.1 Introduction

As detailed in the previous chapter, DNA methylation state along vertebrate genomes can be robustly predicted using local nucleotide sequence information alone. *In silico* analyses by previous studies on mammalian genomes, as well as by this study on the medaka genome, clearly demonstrated that hypomethylated and hypermethylated genomic loci are indeed characterized by distinct nucleotide sequence composition, which underpins the high precision and sensitivity achieved by the *in silico* predictions of local methylation state. The strikingly strong statistical association between local methylation state and nucleotide sequence suggests that nucleotide sequence is associated with, or even directly determine, the patterning of the DNA methylation landscape.

As briefly reviewed in *General Introduction*, a number of recent *in vitro* studies have provided critical experimental evidence for the anticipated causal relationship between DNA methylation and the underlying nucleotide sequence (Lienert et al. 2011; Stadler et al. 2011; Krebs et al. 2014; Takahashi et al. 2017). Most importantly, these *in vitro* reports further highlighted that local hypomethylation is autonomously determined by the locally high CpG density (Krebs et al. 2014; Takahashi et al. 2017) and/or the presence of specific transcription factor bind sites (Lienert et al. 2011; Stadler et al. 2011; Krebs et al. 2014). Yet, the above *in vitro* observations were all made on cultured embryonic stem cells (ESCs). It is irrefutable that ESCs are very versatile *in* *vitro* experimental models and have repeatedly enabled important discoveries in relation to the mechanistic actions of multiple epigenetic machinery (Dawlaty et al. 2014; Baubec et al. 2015). Yet, their use in the characterization of the molecular logic behind the patterning of DNA methylation landscape is dubious as cultured ESCs are known to have a high degree of epigenetic instability (Rebuzzini et al. 2016). More alarmingly, the use of mouse ESCs (Lienert et al. 2011; Stadler et al. 2011; Krebs et al. 2014; Takahashi et al. 2017), in particular, is unjustifiable. This is because mouse ESCs can thrive normally with a demethylated genome (Tsumura et al. 2006), which is in stark contrast to human ESCs (Liao et al. 2015) and early mouse embryos (Li et al. 1992) that major loss of CpG methylation is guaranteed to be lethal. Therefore, the unconventional epigenetic characteristics of ESCs call into question the generalizability of the aforementioned *in vitro* findings.

Moreover, the anticipated dependency of DNA methylome on genomic sequence has never been carefully validated in a proper genetic context *in vivo*. Long et al. (2016) were among the first to attempt to gain insight into the sequence dependency of DNA methylation *in vivo*. These authors examined the DNA methylation state of the 42-Mbp fragment of human chromosome 21 in the Tc1 trans-chromosomic mice and showed that the hypomethylated regions that are natively found in human remained poorly methylated in Tc1 mice. In addition, they demonstrated that natively hypomethylated mouse genomic sequence also remained unmethylated after being transposed into the zebrafish genome. Based on these results, the authors inferred that *in vivo* hypomethylated state is autonomously conferred by evolutionarily conserved sequence signatures. Notwithstanding, their observations were made on non-native sequences (i.e. examining human genomic sequence in mouse, or mouse genomic sequence in zebrafish), and their method-of-choice, i.e. Bio-CAP, had a strong bias towards sequences with high CpG density, intrinsically underrepresenting loci that were less dense in CpG (Blackledge et al. 2012). Likewise, Li et al. (2015) examined the methylation status of one, and only one, transgene across three generations in rat and found the stable acquisition and inheritance of DNA methylation pattern, but the transgene examined was composed of a mouse promoter and human gene. Thus, it is difficult to draw a general conclusion with these studies on the causal relationship between DNA sequence and methylation in native context *in vivo*.

Therefore, a cascade of experiments, as documented in this chapter, was specially designed and conducted to rigorously scrutinize the long-anticipated sequence dependency of DNA methylation *in vivo*. Specifically, these experiments aimed to (1) validate if the HypoMDs in the medaka genome is genuinely conferred by the highly enriched DNA motifs as identified by the statistical modelling in Chapter 1, namely the consecutive CpGs and CGnull-repeats that are highly enriched in HypoMDs, and (2) examine if native (i.e. medaka) genomic sequences can autonomously determine their own methylation state *in vivo*. Unexpectedly, in spite of the strong statistical association, the results suggest that nucleotide sequence, by itself, cannot dictate its own methylation state in medaka *in vivo*.

2.2 Results

2.2.1 Disruption of $(CG)_3$ and CGnull-repeat did not lead to *de novo* DNA methylation of HypoMDs in F0 blastula embryos

To verify if there is any causal relationship between the local hypomethylated state and the presence of consecutive CpGs or CGnull-repeats that were identified in Chapter 1, the $(CG)_3$ or CGnull-repeats in 6 HypoMDs were disrupted by inducing sequence insertions or deletions (indel mutations) *in vivo* using CRISPR-Cas9. Medaka embryos were injected with Cas9 mRNA and target-specific sgRNAs at 1-cell stage and allowed to develop to the blastula stage (2000–4000 cells; at approx. 8 hour-post-fertilization; see also Figure 2.1 for a schematic illustration of the experiment). The methylation state of the targeted HypoMDs was determined by bisulfite PCR, subcloning and Sanger sequencing. Alleles (i.e. individual sequencing reads) were filtered for the presence of indels in the targeted $(CG)_3$ and CGnull-repeats, i.e. successful disruption. In spite of the strong statistical enrichment of $(CG)_3$ and CGnull-repeat in HypoMDs as shown in Chapter 1, disruption of these motifs did not result in any alteration to the hypomethylated state of the targeted HypoMDs. In fact, all of these HypoMDs remained completely unmethylated (i.e. 0% methylation across all CpGs) after the CRISPR-Cas9-mediated disruption (Figure 2.2). These results suggest that either (1) DNA sequences were not perused until genome-wide reprogramming is required (e.g., in primordial germ cells and during gametogenesis), hence the effect of motif disruption could not be manifested, (2) the local hypomethyation state is redundantly conferred by unknown sequence pattern(s) in addition to $(CG)_3$ or CGnull-repeats, or (3) the targeted sequence patterns are simply not related to the hypomethylated state, in spite of their strong statistical association. These possibilities were thoroughly tested in subsequent experiments as follows.

2.2.2 Methylation state of the $(CG)_3$- or CGnull-repeat-disrupted HypoMDs remained unchanged in F1 blastula embryos

To delineate whether the disruption of $(CG)_3$ and CGnull-repeat could lead to alteration in local methylation state after genome-wide reprogramming, the CRISPR-Cas9 injection experiment was repeated. The injected embryos were reared to adults (approx. 3 month-post-hatch) and genotyped to identify founder mutants, which were than inter-crossed to produce F1 generation that inherited the indel mutations through germline transmission. The methylation state of the edited loci was examined in F1 embryos at the blastula stage. In spite of substantial amount of time was passed since the disruption of the $(CG)_3$ and CGnull-repeat (> 3 months) and of the expected genome-wide reprogramming during primordial germ cells formation and gametogenesis in the F0 founders, as well as immediately after the fertilization of their gametes, the targeted HypoMDs remained completely unmethylated (methylation rate = 0% across all examined CpGs, as in Figure 2.2) in the F1 offspring.

2.2.3 DNA methylation states are not autonomously determined by nucleotide sequence at ectopic genomic positions

To address the possible redundancy of sequence determinants in HypoMDs and to gain genome-wide insight into the autonomy of methylation state determination by genomic sequence, medaka genomic DNA was fragmented, captured, then integrated back into the genome with or without prior artificial methylation. In brief (see Materials and Methods for details), medaka genomic DNA was digested and enriched for CpG-containing fragments (40–220 bp; extended to 180–360 bp with adaptors) using a library preparation method identical to that was designed for reduced representation bisulfite sequencing (RRBS) (Gu et al. 2011). The PCR-amplified (hence, unmethylated) fragments were labelled (methylation at the N6 position of the adenine in the Dam sites, 5'-GATC-3', inside the adapters), which is followed by (or without) artificial CpG methylation *in vitro*, then introduced into medaka zygotes and allowed for random genome integration via highly efficient I-SceI-mediated random genome integration (see Figure 2.3 for graphical procedures). Integration was expected to occur at 1-cell stage, immediately after injection (Thermes et al. 2002). The unintegrated fragments were then removed by size-selection and DpnI-digestion (see Materials and Methods for details and Figure 2.4 for removal efficiency). The methylation state of the integrated fragments was subsequently gauged at the blastula stage (2000–4000 cells) using bisulfite PCR and high-throughput sequencing. The assayed integrated fragments encompassed nearly the entire range of GC content and CpG density of HypoMDs and HyperMDs (Figure 2.5 vs Figure 1.2). Approximately equal number of CpGs from HypoMDs and HyperMDs were assayed (Figure 2.6).

In spite of the strong statistical association between nucleotide sequence and methylation states as described in Chapter 1, the integrated genomic fragments could not recapitulate their endogenous methylation state at ectopic locations. The methylation rate at endogenous loci and that at ectopically integrated locations showed essentially zero statistical correlation: Spearman's $\rho \leq 0.08$, Kendall's $\tau \leq 0.07$ (see Figure 2.7 for biplots). Without prior artificial methylation, CpGs on the integrated fragments were almost entirely unmethylated regardless of their endogenous state (Figure 2.8: upper-left vs lower-left). The lack of sequence dependency was further illustrated by a drastically different ectopic methylation pattern when the genomic fragments were artificially methylated prior to injection and genome integration

(Figure 2.7: panel A vs B; Figure 2.8: left vs right). The sharp contrast in the ectopic methylation patterns suggested that nucleotide sequence alone does not carry adequate information for the determination of its own methylation state.

Notwithstanding, the artificially methylated, integrated fragments contained a substantial number of unmethylated CpGs when examined at the blastula stage (Figure 2.8: the peaks at 0% in both upper- and lower-right panels). These unmethylated CpGs were unlikely due to incomplete artificial methylation prior to injection since: (1) the methylase (CpGs-specific DNA methyltransferase SssI, a.k.a. "M.SssI") used is known to completely methylate CpGs in all sequence context (Fatemi 2005), and (2) evidently complete methylation was routinely achievable by the optimized reaction regimen (see Figure 2.9 for examples using bacterial genomic DNA and vector library that have higher CpG frequencies per unit weight of DNA than the medaka genome). The observed unmethylated CpGs could be caused by demethylation in the injected embryos. However, such demethylation could not be directly inferred as recapitulation of the endogenous methylation state, since there was an absence of correlation between the endogenous and ectopic states (Figure 2.7: panel B; Also see Figure 2.8: upper-right vs lower-right panels). In addition, the observed loss of premethylated state was unrelated to the endogenous chromatin accessibility (hence, potential binding of- or recognition by- transcription factors), as CpGs originated from heterochromatin and euchromatin were equally susceptible to the loss of methylation (right panels of Figure 2.10; note the peaks at 0% methylation rate in the histograms along Y-axes). To clarify if the observed demethylation could be due to sequence features intrinsic to the integrated fragments, the nucleotide sequences (10 bp from both up- and down-stream) encompassing CpGs that were demethylated were compared with those that were maintained as hypermethylated using kmer-SVM with the same parameters as Chapter 1. However, the resultant SVMs were highly imprecise and insensitive (Figure 2.11: area under precision-recall curve ≤ 0.47, versus 0.43 from random classifier; Matthew's correlation coefficient ≤ 0.07). Also, the overall ectopic methylation states, as well as the demethylation, of the integrated fragments do not correlate with their size or CpG density (Figure 2.12). Together, the observed demethylated state could not be related to intrinsic sequence features of the genomic fragments.

Given that the injected genomic fragments were (1) only partial fragments of HypoMDs or HyperMDs and may lack the presumed sequence features that are required for autonomous determination of methylation state, and (2) integrated into random genomic positions where they might be influenced by local chromatin state, it is logical to speculated that the observed demethylation might be due, at least in part, to the local epigenetic state of the integrated loci (i.e. position effect; E.g., integrated into pre-existing HypoMDs or somewhere under the influence of *trans*-acting hypomethylation determining elements, hence rendered hypomethylated). Further experiments were thus conducted at pre-specified genomic loci to control for the possible position effect.

2.2.4 DNA methylation state was maintained independent of sequence and position context

In order to examine whether *full-length* HypoMDs and HyperMDs can autonomously determine their own methylation state at an inert genomic location, six unmethylated HyperMDs and eleven pre-methylated HypoMDs with length of 300–400 bp were randomly selected, cloned (with or without artificial CpG methylation), and injected into one-cell stage medaka embryos and integrated into the gene desert region (Kirchmaier et al. 2013) presumably devoid of any possible influence of active regulatory elements (see also Figure 2.13A). The integration was achieved by the highly efficient, PhiC31 integrase-mediated site-specific integration in medaka and was expected to occur at the one-cell stage (Kirchmaier et al. 2013). Methylation state of the integrated sequences were examined at the blastula stage. Autonomy in methylation state determination by the full-length, integrated sequences would manifest as remethylation of the unmethylated HyperMDs, as well as active or passive loss of the methyl groups on the premethylated HypoMDs, after genome integration (see also Figure 2.13B for illustration of the logic of the experiment).

All of the unmethylated, integrated HyperMDs failed to acquire methylation (Figure 2.14). Likewise, the artificially methylated, integrated HypoMDs remained hypermethylated (Figure 2.15), with very limited number of CpG dinucleotides (i.e. only 4 out of the 202 CpGs inspected) showing no methylation (Figure 2.15: blue dots on the integrated, ectopic copies of HypoMDs/Loci 1, 4, 6, and 9). Since the determination of methylation state of these distinct CpGs in the premethylated plasmid library is infeasible (as plasmid DNA converts very poorly in bisulfite reaction), it is possible that these CpGs were not fully methylated prior to injection. However, as aforementioned, the M.SssI methyltransferase used in the pretreatment has no known sequence specificity. The observed absence of methylation probably reflects highly localized loss of methyl groups on these specific CpGs. Collectively, the above results indicate that the overall, ectopically introduced nucleotide sequences were not perused and the artificially conferred methylation states (i.e. hypomethylation in the HyperMDs, and hypermethylation in the HypoMDs) were robustly maintained *in vivo*.

Finally, to ascertain that the above unexpected observations were not artifacts incurred by ectopic genome locations, the methylation state of two HypoMDs were edited *in situ* via CRISPR-Cas9-triggered homology-directed repair (HDR) and artificially methylated repair template (see Figure 2.16 for illustration of concept behind the experiment). Consistent with the aforementioned observations, in spite of their original hypomethylated state, the loci were rendered largely hypermethylated after editing (Figure 2.17). Since the observed lack of restoration of native methylation state could be due to the seemingly limited time allowed for recapitulation (from injection to sampling, i.e. from 1-cell stage to blastula: approx. 8 hr, encompassing 11-12 rounds of cell divisions), the editing experiment was repeated and the endpoint was extended to later developmental stages: Stage 31 and 39, at 3 and 7 day-post-fertilization (i.e. day-post-injection), respectively. Yet, the edited alleles remained hypermethylated in the mid-/late-stage embryos (Figure 2.18). Significant loss of methyl groups could only be observed on two distinct, adjacent CpGs in one of the two edited loci (Figure 2.18, panel B: the 1st and 2nd CpG). Taken together, these observations indicate that genomic sequence and its methylation state were not coupled even at the endogenous position.

2.3 Discussion

The above experimental results definitely showed that local methylation state is not autonomously determined by nucleotide sequence in medaka *in vivo*, in spite of the strong statistical association between DNA methylation state and the underlying nucleotide sequence as reported in Chapter 1. The disruption of highly enriched DNA motifs (i.e. consecutive CGs and CGnull-repeat) in HypoMDs failed to obliterate the hypomethylated state in both F0 and F1 medaka embryos. Also, by artificially manipulating the methylation state of native genomic sequences and reintegrating them into the medaka genome, the artificially conferred methylation states were shown predominantly maintained *in vivo*, regardless of the sequences and their endogenous states. These findings strongly argue against the general belief that the DNA methylation landscape in vertebrates is sequence-dependent.

These unexpected results are in clear contradiction to the recent *in vitro* studies as reviewed in this chapter's *Introduction*. As aforementioned, those *in vitro* evidence is derived almost exclusively from mouse ESCs (Lienert et al. 2011; Stadler et al. 2011; Krebs et al. 2014), which are known to be epigenetically unstable and can thrive without DNA methylation unlike most other vertebrate cell types both *in vitro* and *in vivo*. Due to the atypical epigenetic characteristics of cultured mouse ESCs, whether and to what extent can those *in vitro* findings be extrapolated to an *in vivo* context remains unknown.

On the contrary, the findings as presented in this chapter are coherent to the pioneer experiments (Pollack et al. 1980; Wigler 1981; Stein et al. 1982) on the inheritance of DNA methylation in cultured mouse L cells. In those reports, the authors demonstrated that the methylation states of exogenous DNA (regardless of being artificially CpG-methylated or completely unmethylated) could be maintained with appreciable fidelity for many cell generations upon stable integration. Taken together, the recently proclaimed sequence-dependency of DNA methylation is far from concrete.

However, this study does not completely rule out the existence of highly confined, local sequence-dependent DNA methylation. As proposed by Richards (2006), the sequence-dependency of epigenetic signatures may vary with actual sequence-context, i.e. some nucleotide sequences may favour or even mandate a certain methylation state, while others may be completely independent of DNA methylation. Although the artificially established hypermethylated state of the sequences examined in this study was mostly maintained after genome integration, there was complete loss of methyl groups on some CpGs in the eleven pre-methylated HypoMDs, as well as within one of the *in situ* edited loci. This suggests the presence of local sequence elements that facilitate demethylation on specific CpGs, although their effect was spatially confined. As previously demonstrated *in vitro*, some DNA motifs, in particular several transcription factor binding sites (Blattler and Farnham 2013), are indeed instructive to DNA methylation and may account for the change of methylation state in specific loci upon differentiation (Meissner et al. 2008; Hodges et al. 2011). Importantly, their effect was also demonstrated to be limited to no more than a few tens of base pairs up- and down-stream (Stadler et al. 2011; Krebs et al. 2014). It is thus likely that the restricted governing range (< 100 bp) of these DNA sequences is insufficient to account for the span of HypoMDs (median length > 1 kb).

2.4 Chapter conclusion

In spite of the observed statistical association (Chapter 1), nucleotide sequence was not capable of autonomously determining its own methylation state in medaka *in vivo*. The above experimental results clearly demonstrated that DNA methylation state was robustly maintained *in vivo*, independent of the underlying sequence and the endogenous state. This unexpected finding apparently contradicts the notion of the governance on DNA methylation by genomic sequence in vertebrates, but instead suggest the involvement of other epigenetic factors in defining and maintaining the DNA methylation landscape.

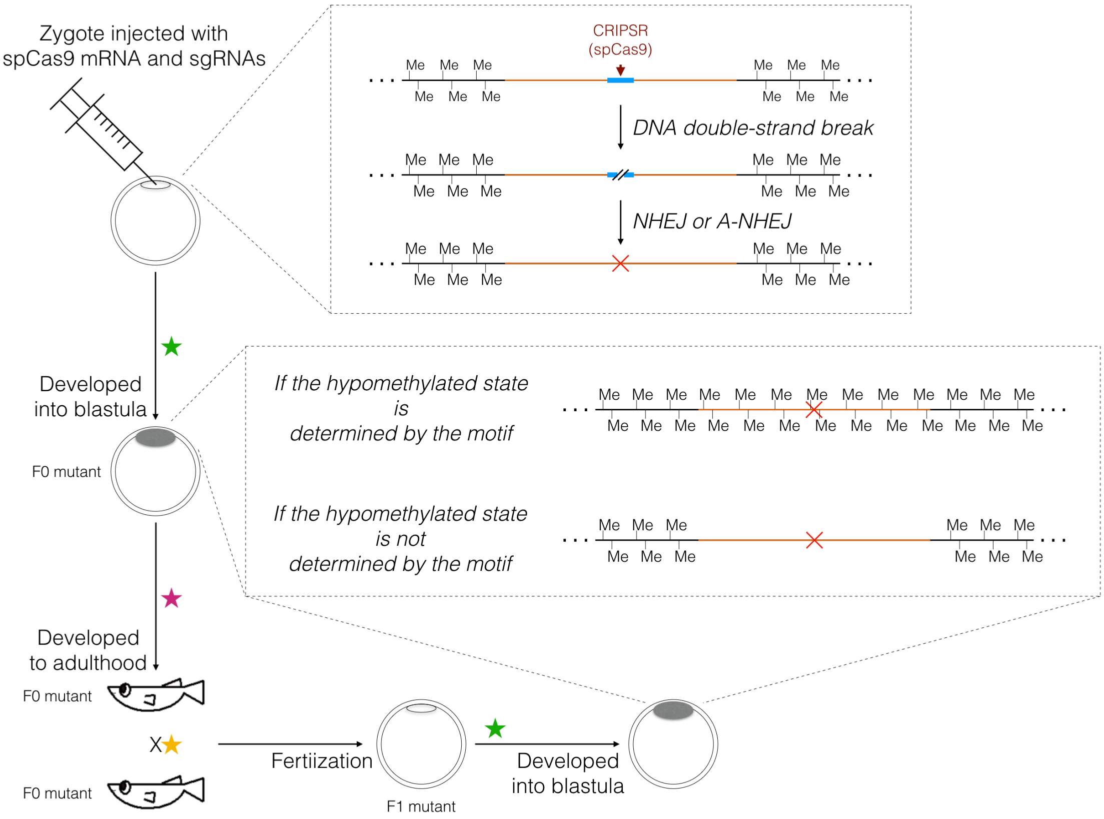

Figure 2.1 Schematic diagram illustrating the disruption of $(CG)_3$ or CGnull-repeat in a HypoMD via CRISPR-Cas9 and the expected outcomes depending on whether these sequence motifs are responsible for the local hypomethylated state.

The $(CG)_3$ or CGnull-repeat is represented by the blue box inside the targeted HypoMD (orange segment). The motif is disrupted by CRISPR-induced double-strand break and the subsequent indel (red cross) formed via non-homologus end-joining (NHEJ) or alternative non-homologous end joining (A-NHEJ, a.k.a. microhomology-mediated end joining / MMEJ). Major genome-wide *de novo* methylation (denoted by green stars) for the acquisition of zygotic DNA methylome is known to occur at some point between 64-cell and blastula stage (Walter et al. 2002). Epigenetic reprogramming is expected to occur during primordial germ cell development (denoted by purple star) and gametogenesis (denoted by yellow star). The artistic drawings of adult medaka is a courtesy of Dr. Ayako Uno.

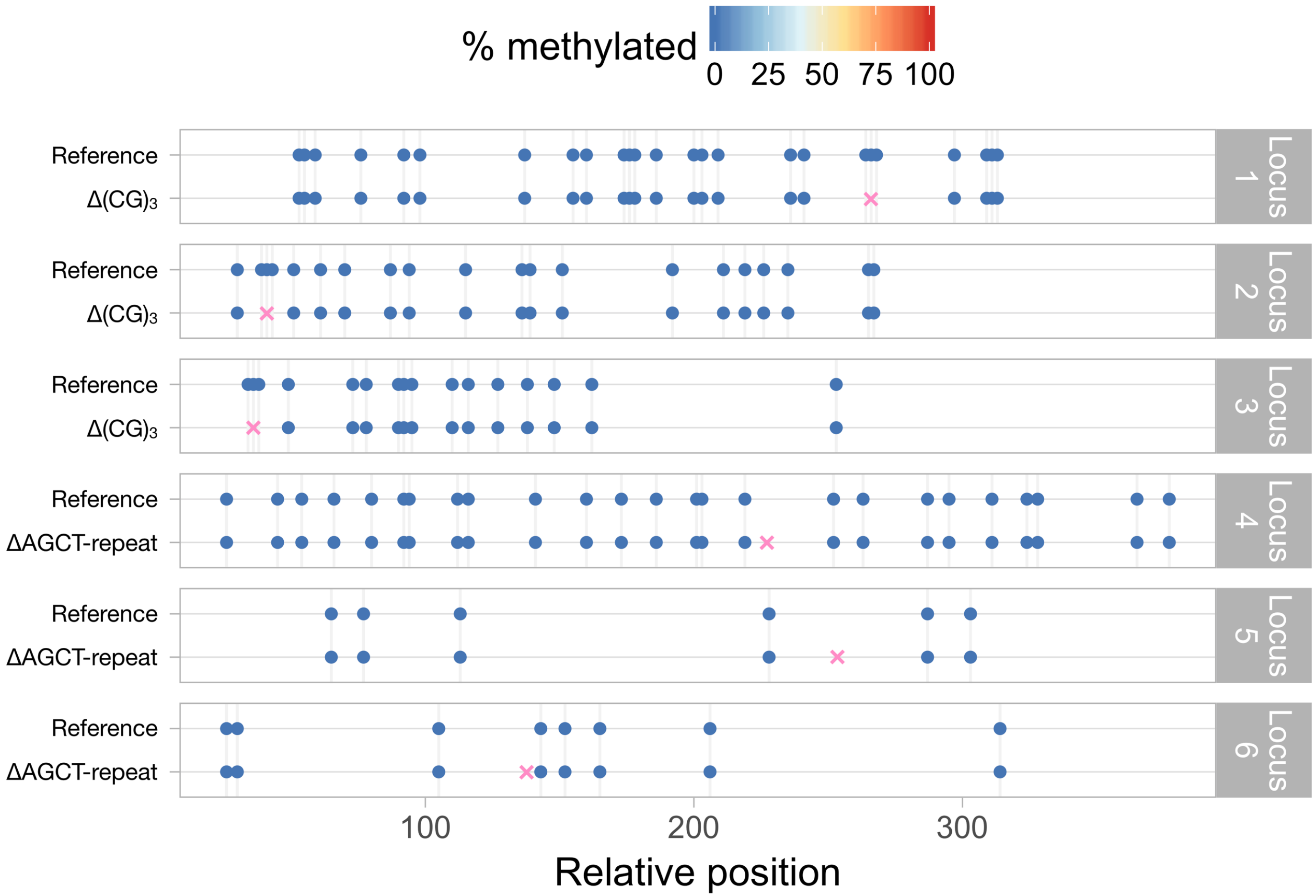

Figure 2.2 Methylation state of six HypoMDs after disruption of the $(CG)_3$ (Locus 1-3) or CGnull-repeat (Locus 4-6) by the use of CRISPR-Cas9.

Native methylation state was overlaid as reference. Pink crosses indicate the CRISPR-Cas9-induced indel locations. Note that the methylation rate is zero across all CpGs (i.e. completely unmethylated) with or without editing. Mean coverage = 5×.

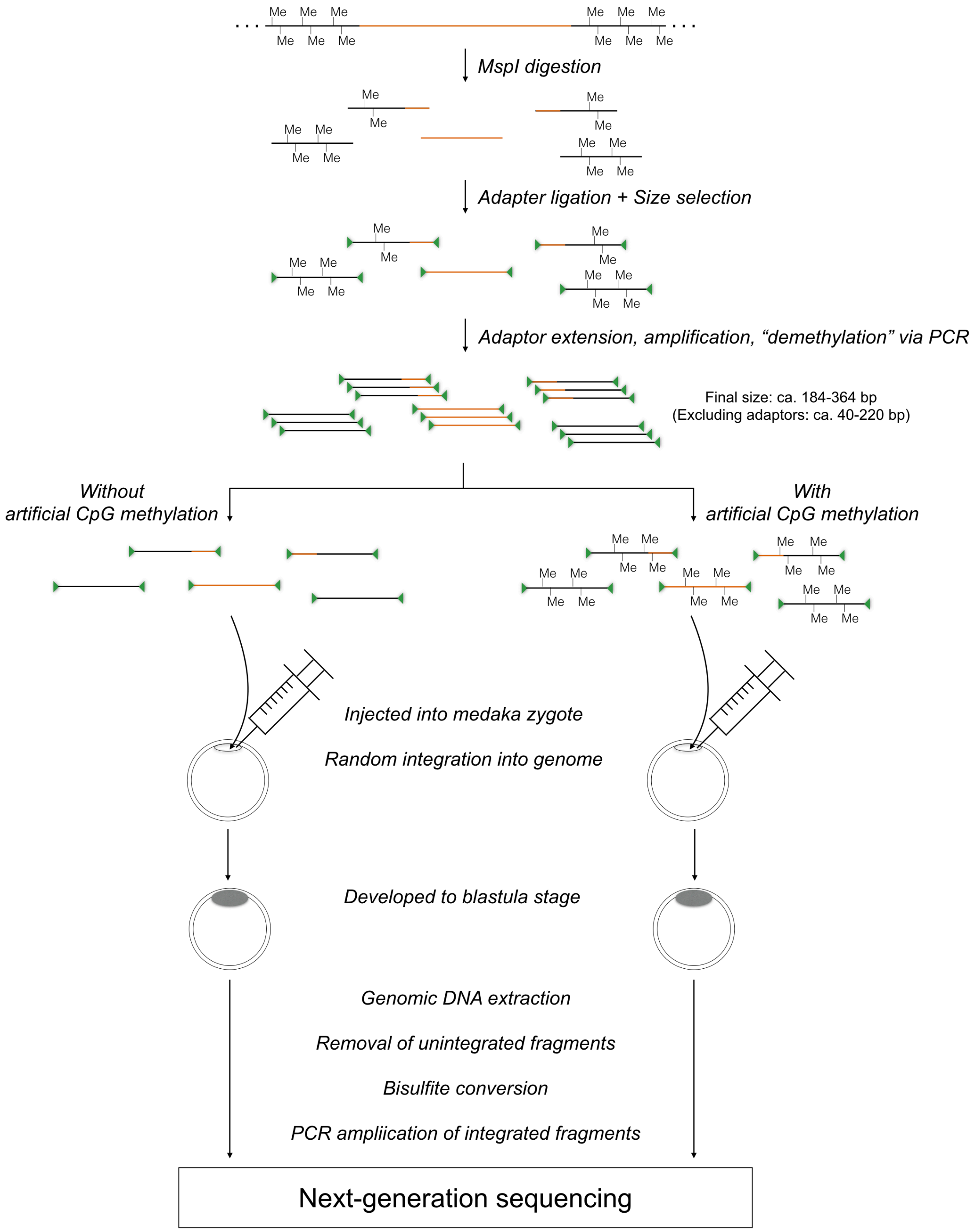
Me Me Me
Me Me Me
Me Me Me
Me Me Me
MspI digestion
Adapter ligation + Size selection
Adaptor extension, amplification, "demethylation" via PCR
Final size: ca. 184-364 bp
(Excluding adaptors: ca. 40-220 bp)
Without
artificial CpG methylation
With
artificial CpG methylation
Injected into medaka zygote
Random integration into genome
Developed to blastula stage
Genomic DNA extraction
Removal of unintegrated fragments
Bisulfite conversion
PCR ampliication of integrated fragments
Next-generation sequencing

Figure 2.3 Schematic diagram illustrating the capturing and processing of genomic fragments for the interrogation of their autonomy in methylation state determination.

The orange segment represents genomic region that is endogenously hypomethylated.

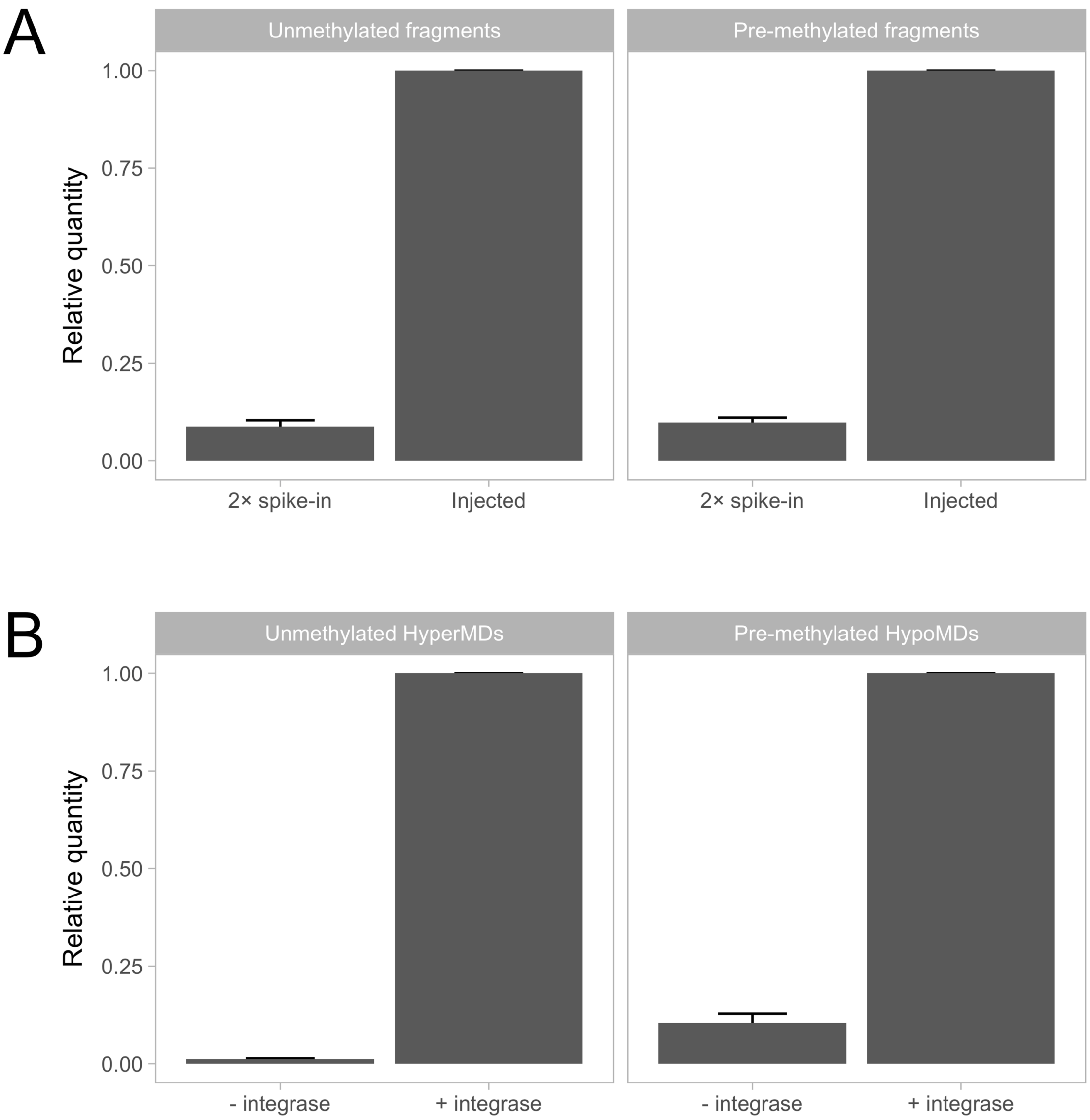

Figure 2.4 Efficient removal of injected but unintegrated libraries via PEG precipitation and DpnI digestion.

(A) MspI-captured fragments with or without pre-methylation via CpG methyltransferase M.SssI. Since it is technically infeasible to prevent the spontaneous integration of linear DNA, an integration-free surrogate control ("2× spike-in") was generated by spiking-in the injection mixtures directly into fresh lysate of uninjected blastula embryos. Approximately twice the amount of the injection mix consumed by genuinely injected embryos ("Injected"), i.e. ca. 34 pL per embryo, was spiked-in to provide conservative estimation of the removal efficiency of unintegrated fragments. (B) Plasmid libraries containing unmethylated HyperMDs or M.SssI-methylated HypoMDs. Since spontaneous integration of circular DNA (i.e. plasmids) into the genome is generally very rare in the absence of integrase, integration-free surrogate control ("- integrase") was generated by injecting PhiC31 medaka embryos without PhiC31 integrase mRNA (in contrast to 100 ng/μL of the mRNA for the integration experiment, i.e. "+ integrase"). Error bars represent 95% confidence intervals.

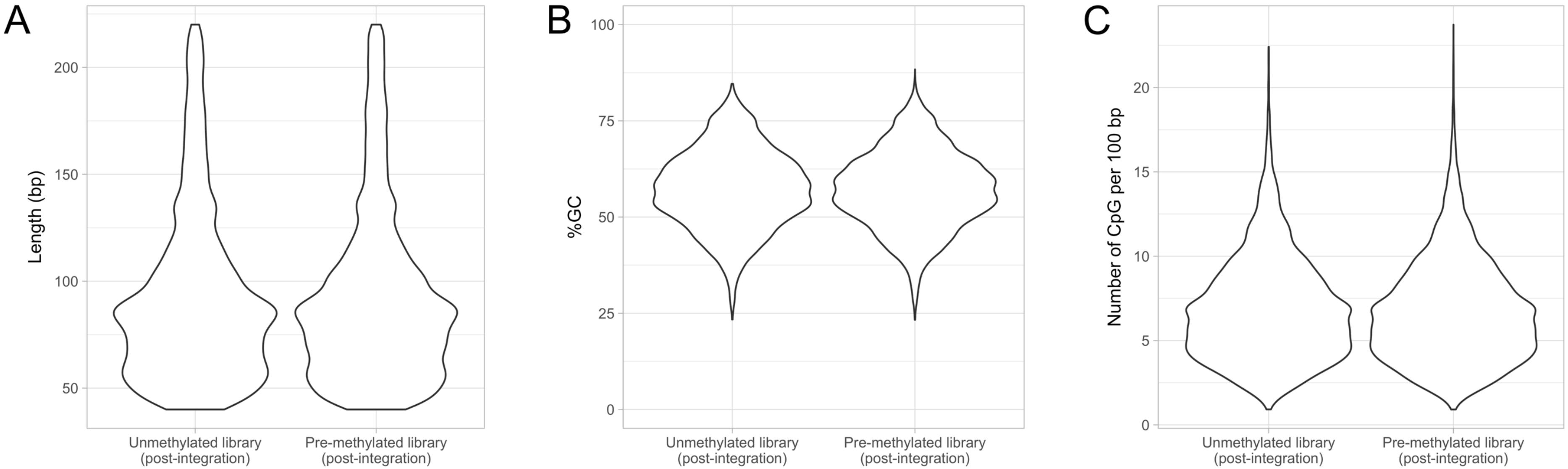

Figure 2.5 Violin plots showing the distribution of (A) length, (B) GC content, and (C) CpG density of the unmethylated or pre-methylated fragments that were successfully integrated into genome and subsequently assayed.

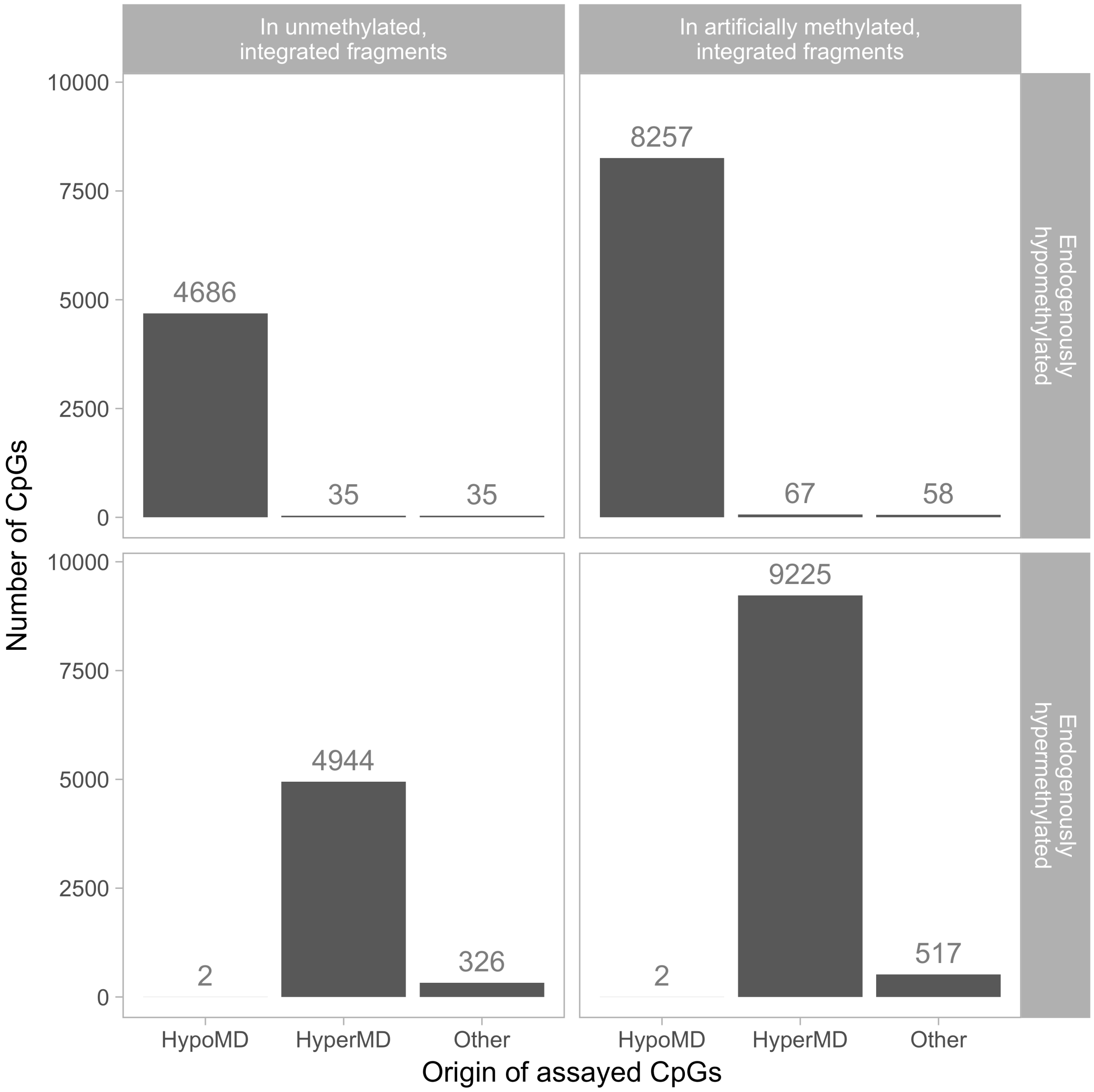

Figure 2.6 Sampling origins (from HypoMDs, HyperMDs, or elsewhere in the genome) of the assayed CpGs on the integrated genomic fragments.

Note that CpGs from HypoMDs and HyperMDs were nearly equally represented (upper panels vs lower panels).

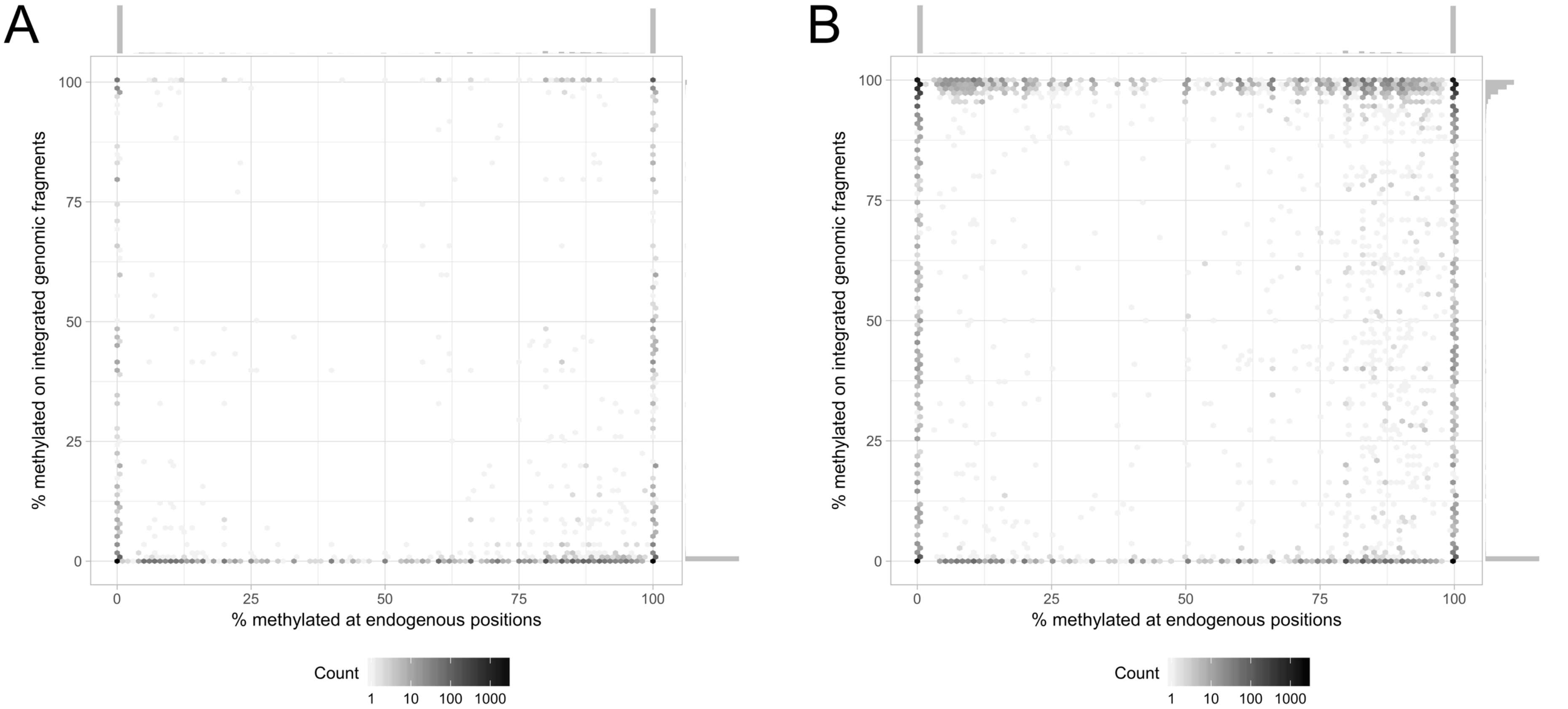

Figure 2.7 Correlation between the methylation rate of the same CpGs at their endogenous versus at reintegrated/ectopic positions, (A) without- or (B) with- artificial methylation prior to injection and genome integration.

The methylation rates were bimodal and strongly skewed towards either 0% or 100%. To circumvent over-plotting, individual CpGs, *N* = (A) 10251 and (B) 18537, were consolidated into hexes (bin width = 1%), with the shade of the hex representing the number of CpGs included (in logarithmic scale). Bars on the top and right-hand side of each of the scatterplots are the histograms that show the density of CpGs along the corresponding axes (bin width = 1%). Correlation coefficients: Spearman's ρ = (A) 0.08, (B) 0.02; Kendall's τ = (A) 0.07, (B) 0.01.

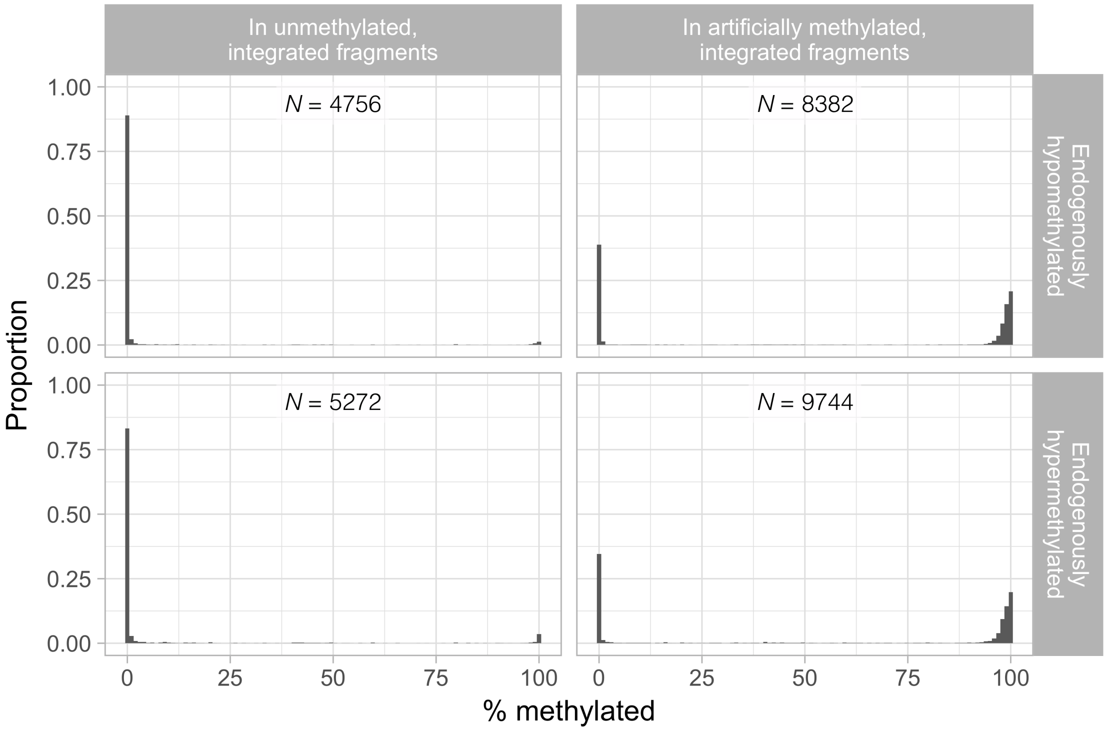

Figure 2.8 Distributions of the methylation rates of CpGs on the integrated genomic fragments, (left) without- or (right) with- artificial methylation prior to injection.

The distributions were displayed separately for CpGs that are endogenously (upper) hypomethylated and (lower) hypermethylated. Bin width = 1%. Note that the histograms in the upper panels (i.e. CpGs that are endogenously hypomethylated) strongly resemble those in the lower panels (i.e. CpGs that are endogenously hypermethylated). "*N*" denotes the number of CpGs included in the corresponding histogram.

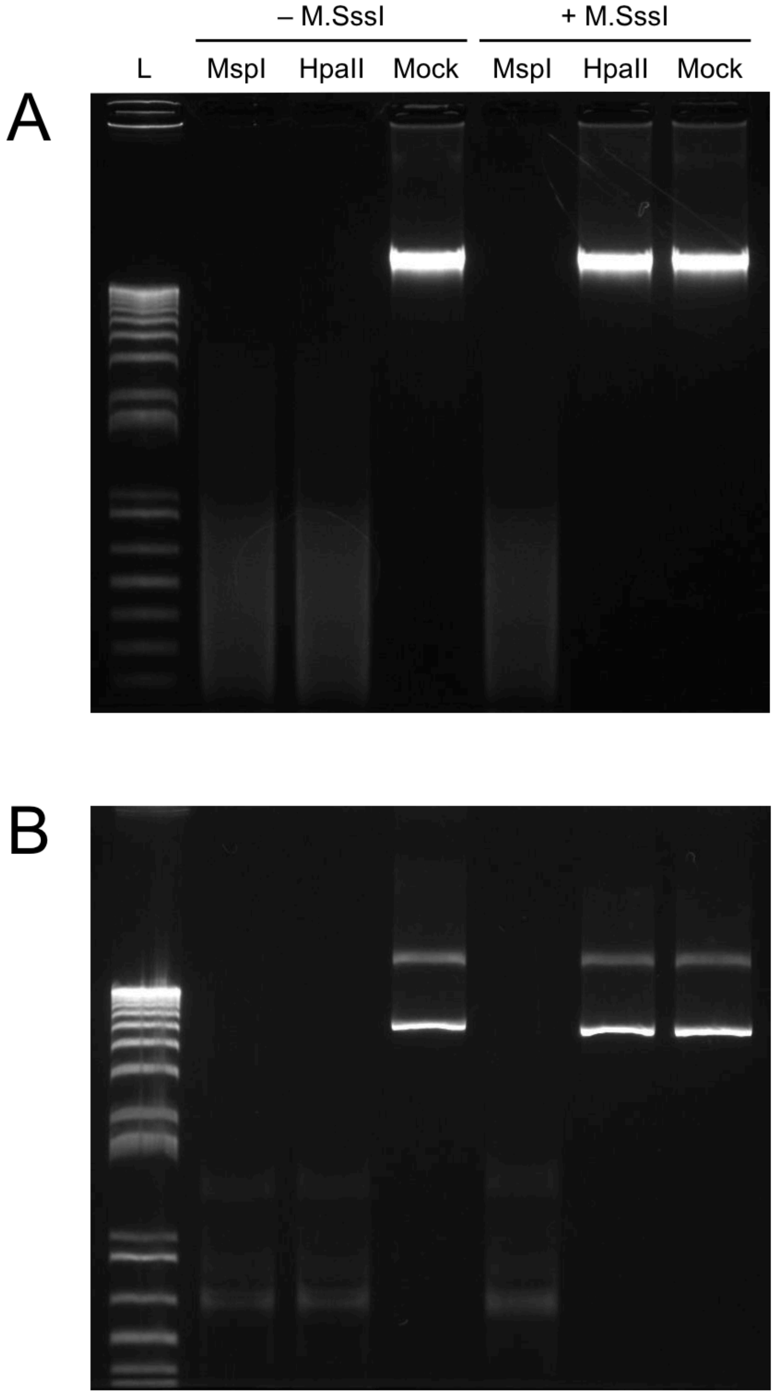

Figure 2.9 Differential sensitivity of untreated and M.SssI-treated DNA samples to MspI and HpaII restriction enzymes.

(A) *E. coli* genomic DNA. (B) Plasmid library that was used to generate results shown in Figure 2.15. Approximate CpG and MspI/HpaII restriction site densities (counts per kilobase pair): (A) 76 and 5, (B) 62 and 4, which are much higher than those in medaka genome, i.e. 23 and 1. Note that there is no observable cleavage by HpaII after pretreatment with the methyltransferase, suggesting complete methylation was achieved using the reaction regimen as described in Materials and Methods. “L”: Thermo Fisher Scientific 1Kb Plus DNA ladder. “Mock”: control reaction without restriction enzyme.

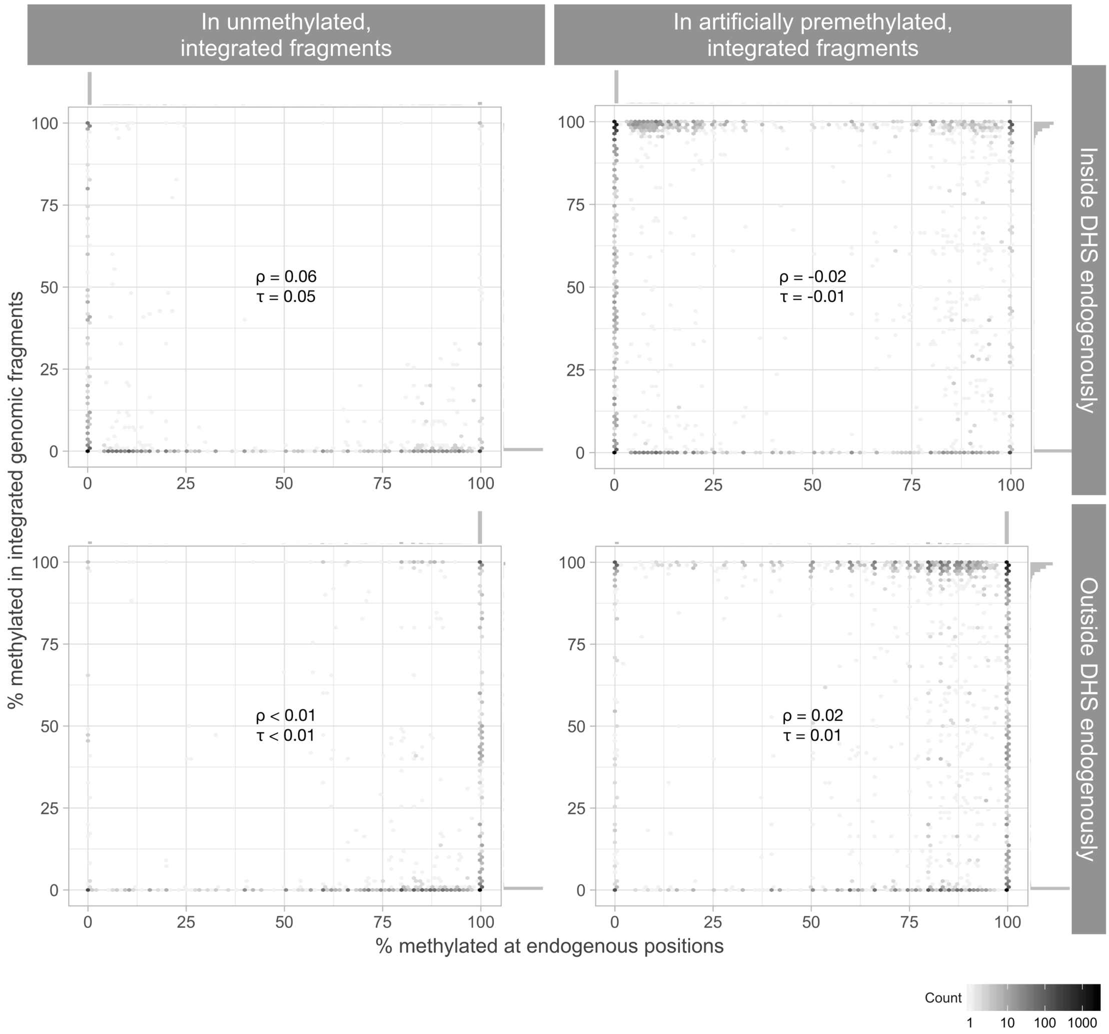

Figure 2.10 Correlation of CpG methylation state at ectopic versus native positions with respect to the endogenous local chromatin accessibility.

CpGs from the inside and outside of DNaseI hypersensitive sites (DHS) were graphed separately (i.e. upper versus lower panels). To circumvent over-plotting, overlapping CpGs were consolidated into hexes (bin width = 1%), with the shade of the hex representing the number of fragments included (in logarithmic scale). Numerical figures denoted at the center of each of the biplots are the correlation coefficients. ρ = Spearman's rho; τ = Kendall's tau.

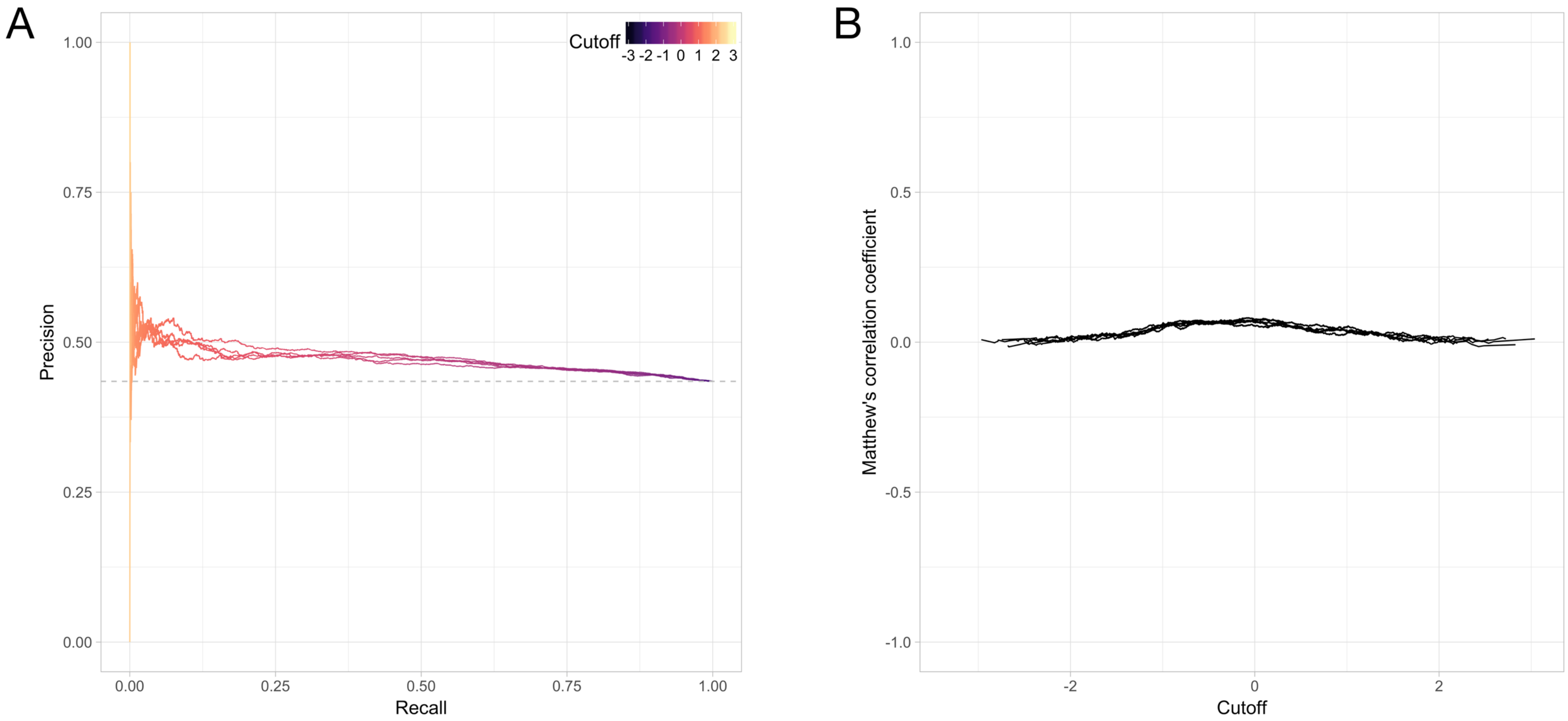

Figure 2.11 Model performance of kmer-SVMs trained for classification of CpGs and their flanking sequences that underwent demethylation.

Demethylated (N = 23655) and hypermethylated (N = 30760) sequences (including 10 bp from both up- and down-stream of the CpG) were assigned to positive and negative classes, respectively. (A) Precision-recall curves. Solid, colored lines are individual precision-recall curves derived from 10-fold cross-validation. The colors represent the cut-off values for binary classification/prediction of the testing pool in each rounds of cross-validation. Area-under-curve (AUC): minimum = 0.46, maximum = 0.47. Random classifier is represented by horizontal dashes at the center and has an AUC of 0.43. (B) Matthew's correlation coefficients (MCC). Individual lines are the MCCs, with respect to all possible cutoff values, derived from 10-fold cross-validation. Maximum MCC achieved = 0.07.

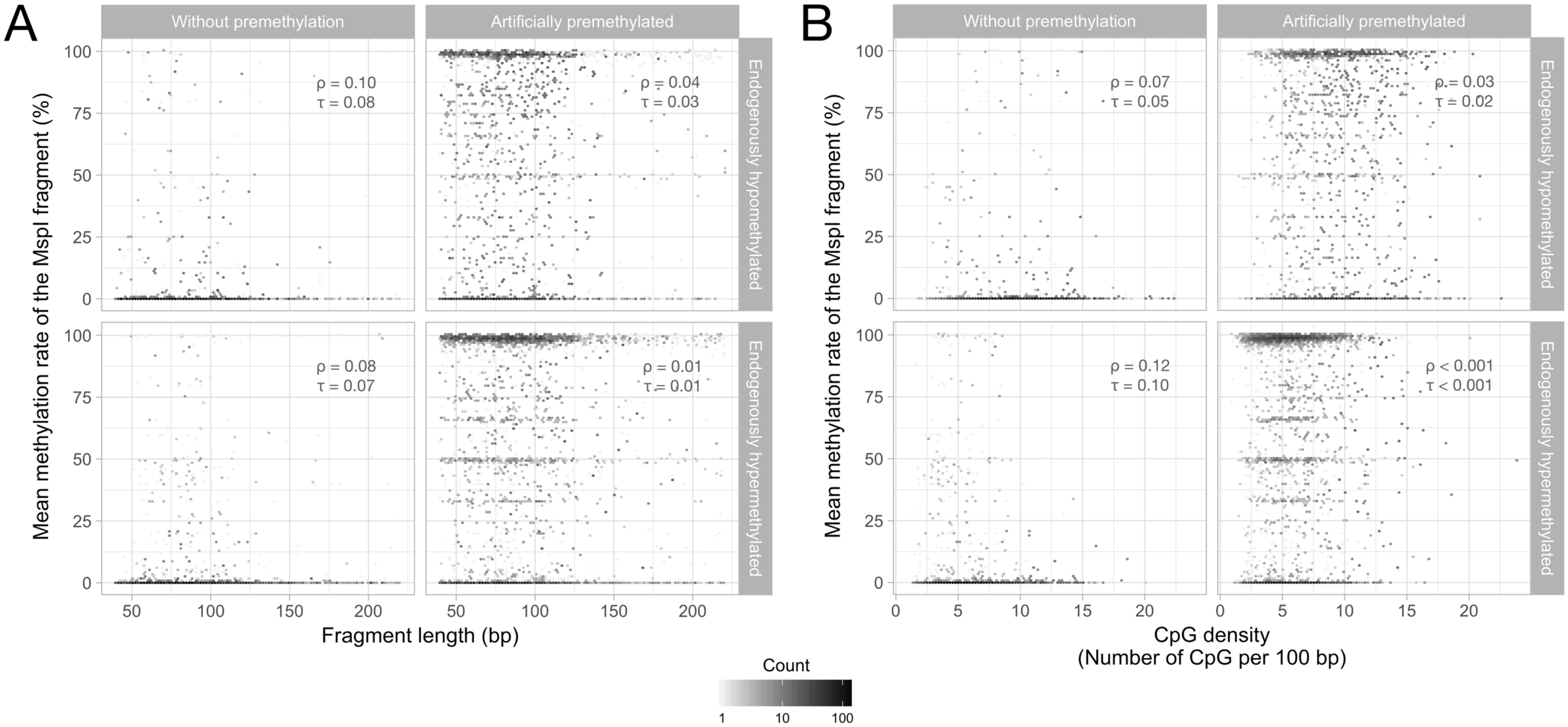

Figure 2.12 Correlation between the overall methylation rate and (A) length or (B) CpG density of the integrated fragments. Integrated fragments derived from the (left) unmethylated and (right) M.SssI-treated libraries were further segregated according to their endogenous methylation state (top vs bottom).

Fragments are defined as endogenously (top) hypomethylated or (bottom) hypermethylated if they have a mean CpG-methylation rate of < 40% or > 60%, respectively. To circumvent over-plotting, fragments with similar methylation rate and (A) length or (B) CpG density were consolidated into hexes (number of bins = 100, both horizontally and vertically), with the shade of the hex representing the number of fragments included (in logarithmic scale). "ρ" and "τ" indicate the Spearman's rho and Kendall's tau correlation coefficients of the corresponding scatterplots.

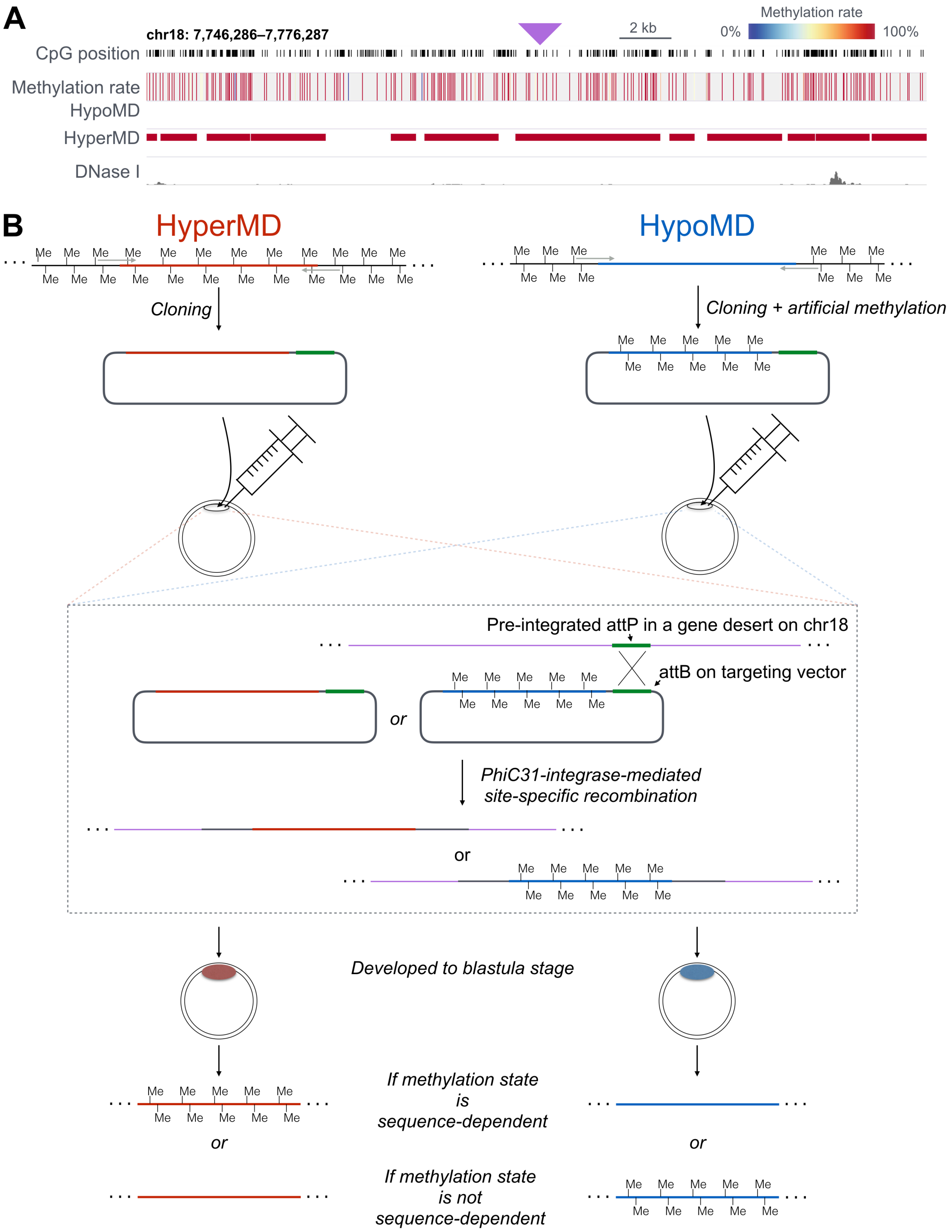
A
chr18: 7,746,286–7,776,287
2 kb
Methylation rate
0%
100%
CpG position
Methylation rate
HypoMD
HyperMD
DNase I
B
HyperMD
HypoMD
Me
Cloning
Cloning + artificial methylation
Pre-integrated attP in a gene desert on chr18
attB on targeting vector
or
PhiC31-integrase-mediated site-specific recombination
or
Developed to blastula stage
If methylation state is sequence-dependent
or
If methylation state is not sequence-dependent
or

Figure 2.13 Interrogation of DNA methylation recapitulation by full-length HyperMDs and HypoMDs at pre-specified, inert genomic location.

(A) Genome browser view of methylation state (as in HdrR strain) of the genomic locus that contains a landing site for PhiC31-mediated site-specific recombination (approximate location is denoted by the purple triangle). (B) Schematic diagram illustrating the irreversible, site-specific integration of the subcloned, unmethylated HyperMDs and pre-methylated HypoMDs via PhiC31-integrase-mediated site-specific recombination and the expected outcomes depending on whether the integrated sequences can autonomously determine their own methylation state.

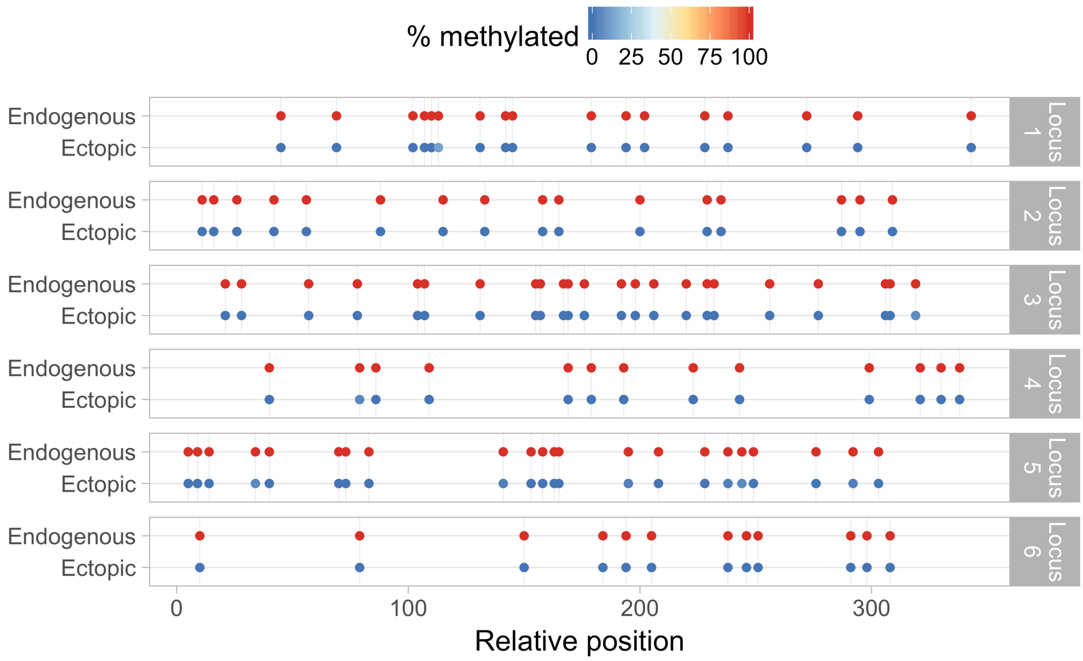

Figure 2.14 Methylation state of six full-length HyperMDs at their endogenous loci and at ectopic location after being cloned and integrated into genome via PhiC31-mediated site-specific recombination.

Note that all of the integrated sequences failed to recapitulate their endogenous hypermethylated state. Mean coverage = 20×.

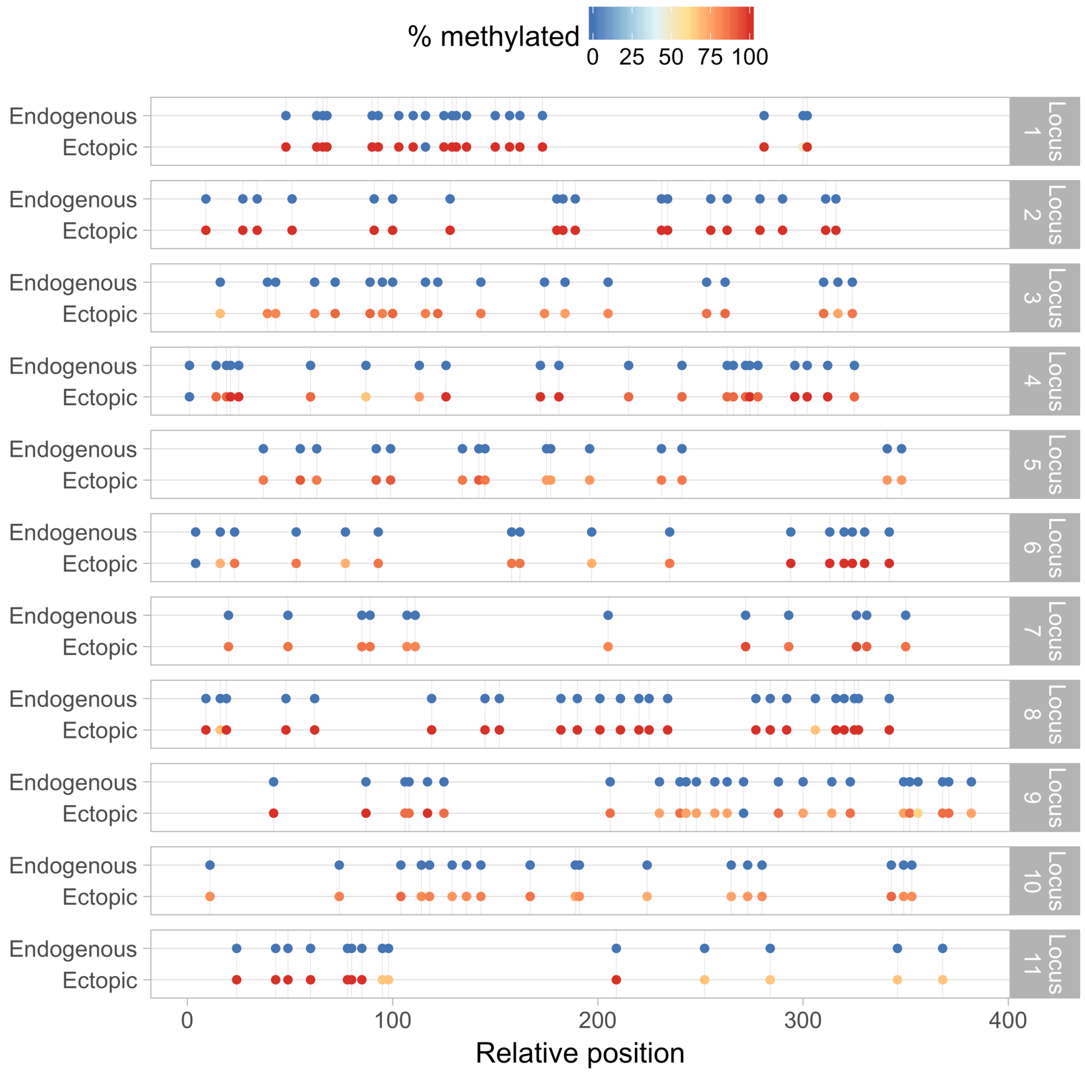

Figure 2.15 Methylation state of eleven full-length HypoMDs (Locus 1 to 11) at their endogenous loci and at ectopic location after being cloned, artificially methylated, then integrated into genome via PhiC31-mediated site-specific recombination.

Note that all of the pre-methylated, integrated sequences failed to recapitulate their endogenous hypomethylated state. Complete loss of methylation was only observed in very small number of CpGs in three of the examined sequences: the 9th CpG of Locus 1, the 1st CpGs of Locus 4, the 1st CpG of Locus 6, and the 14th CpG of Locus 9. Mean coverage = 15×.

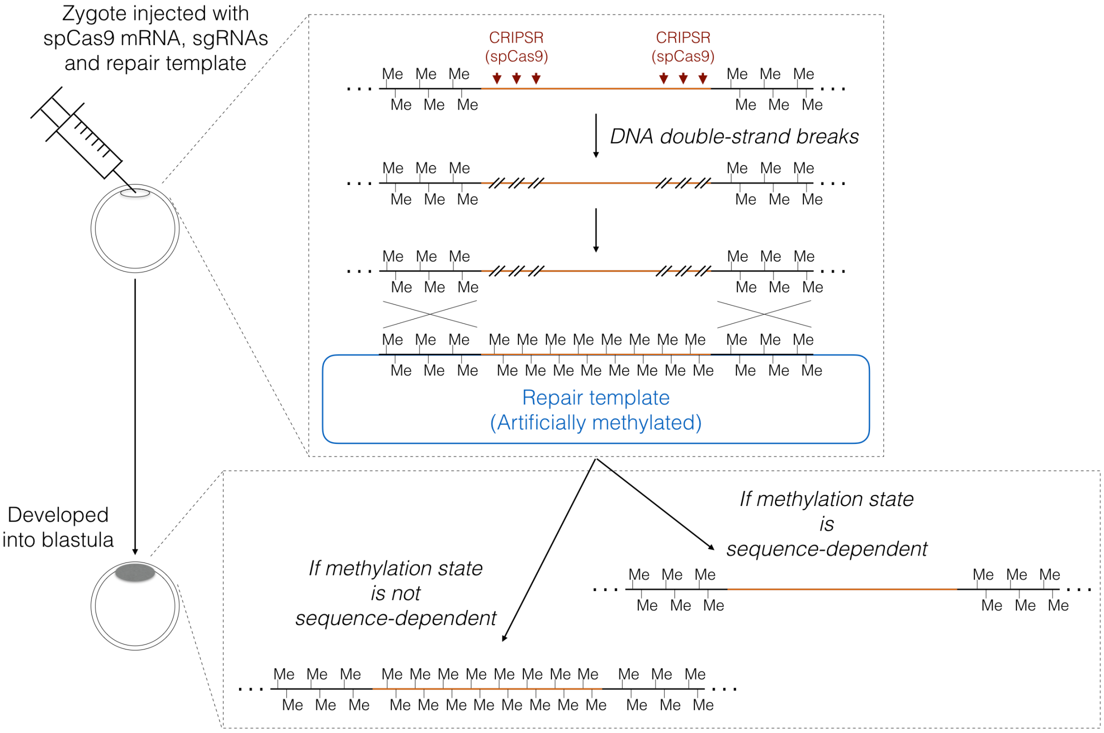
Zygote injected with spCas9 mRNA, sgRNAs and repair template
CRIPSR (spCas9)
CRIPSR (spCas9)
Me Me Me
DNA double-strand breaks
Repair template (Artificially methylated)
Developed into blastula
If methylation state is sequence-dependent
If methylation state is not sequence-dependent

Figure 2.16 Schematic diagram illustrating the principle of the *in vivo* methylome editing on the targeted HypoMDs via homology-directed repair (HDR) and the use of artificially methylated repair template.

HDR was triggered by CRISPR-Cas9 induced DNA double-strand breaks (DSB) at the targeted loci. The repair template contains the subcloned HypoMD (with substitutions in the spCas9's PAM sites, from 5'-NGG-3' to 5'-NGC-3') along with approximately 800 bp flanking regions that served as homology arms. Note that multiple DSBs was made using a cocktail of sgRNAs that guided spCas9 to six different positions along the targeted HypoMD to enhance DSB, hence HDR, rate.

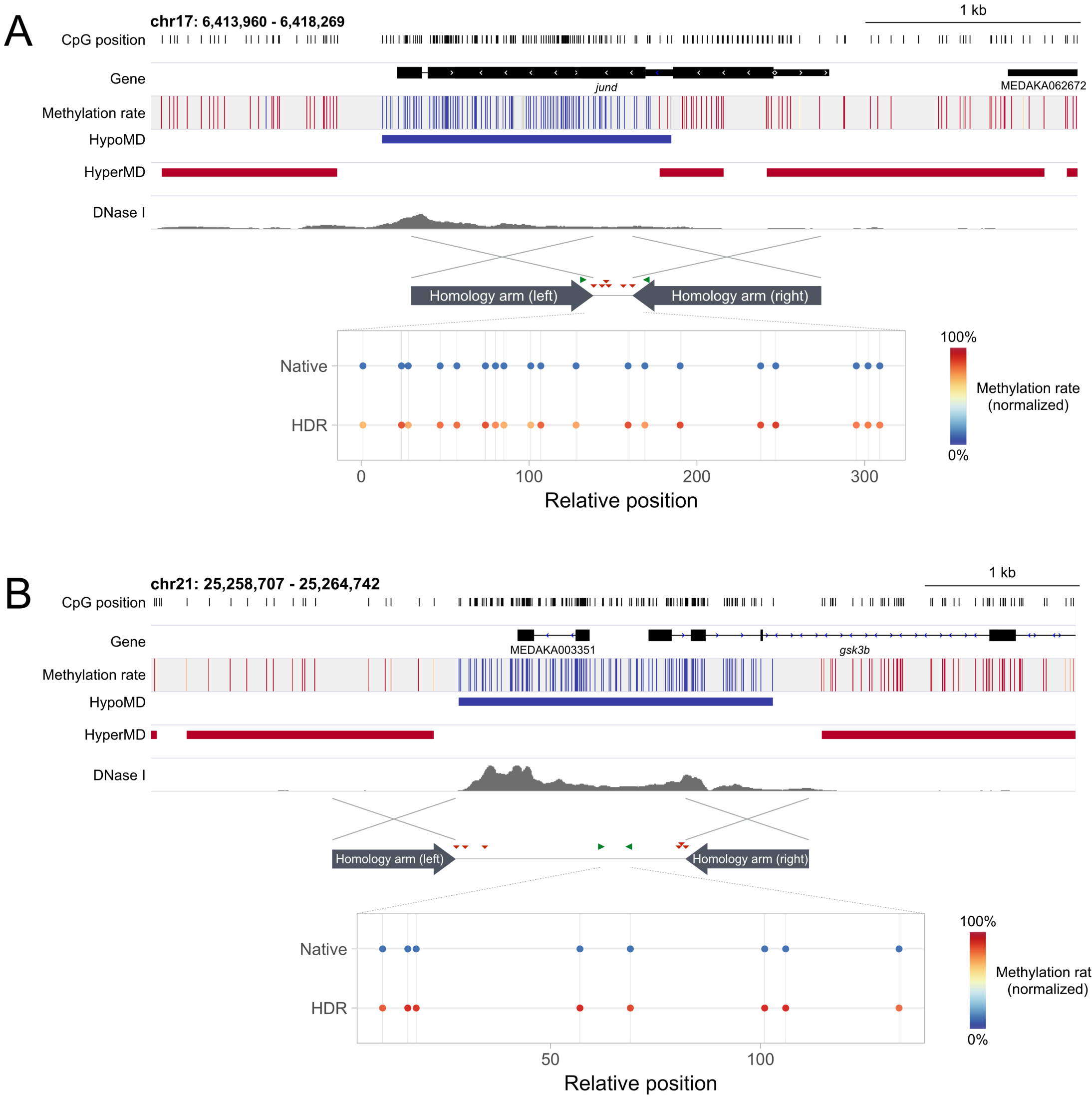

Figure 2.17 Methylation state of two HypoMDs after *in vivo* methylome editing mediated by CRISPR-Cas9-induced homology-directed repair (HDR) and pre-methylated repair templates.

Green arrowheads: bisulfite PCR and sequencing primers; red triangles: binding positions of the sgRNAs.

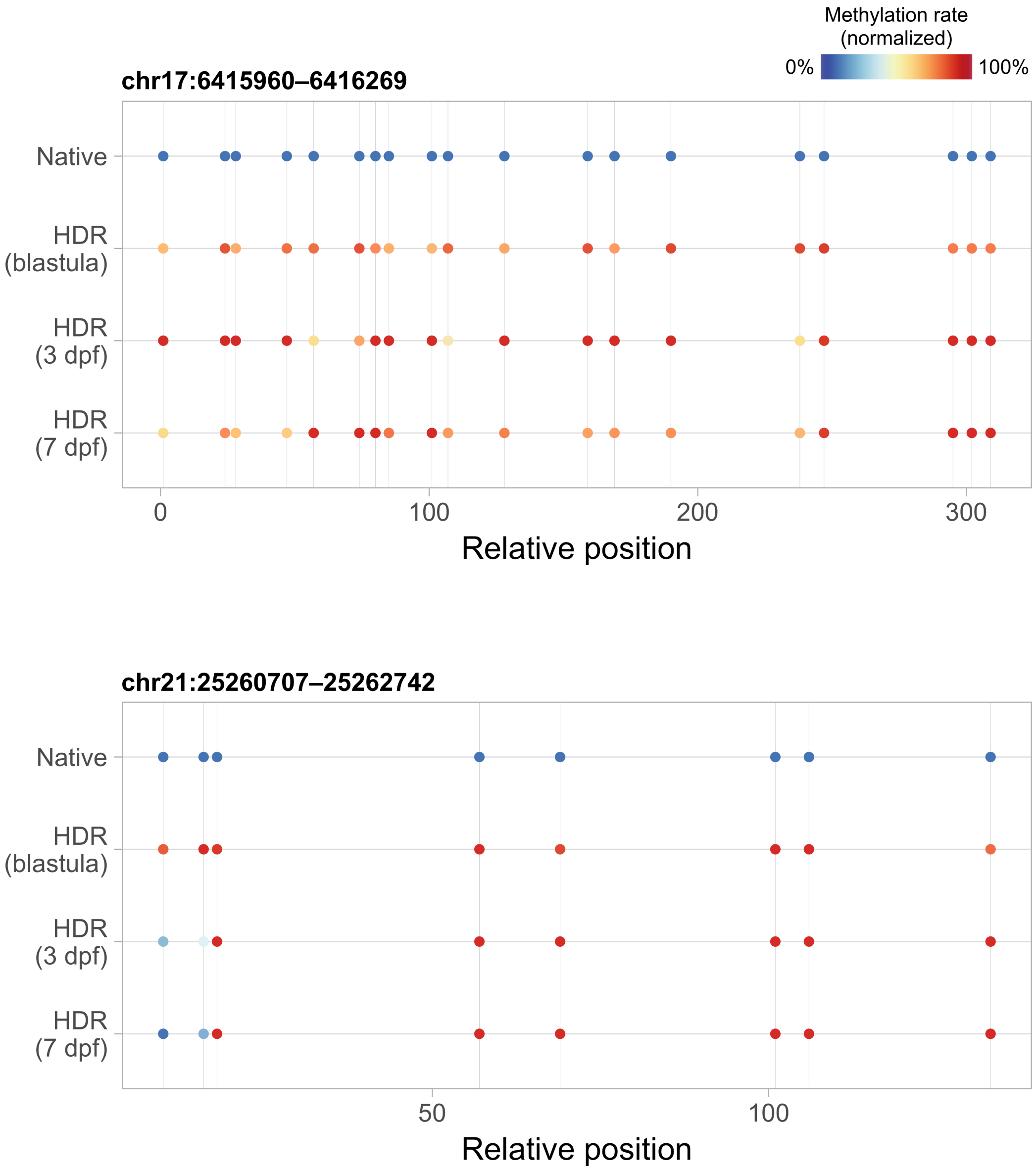

Figure 2.18 Methylation state of the *in situ*-edited HypoMDs at multiple embryonic stages.

Edited embryos were sampled at early (blastula, 0 day-post-fertilization; dpf), mid (3 dpf), and late (7 dpf; hatching) embryonic stages. To enable comparison across sampling time-points with variable editing efficiency, the estimated methylation rates were normalized against the estimated editing rate.

General discussion and perspectives

Based almost exclusively on *in silico* and *in vitro* evidence, it is generally believed that the vertebrate genomes are hypermethylated "by default", and that specific nucleotide sequence features (specifically, high CpG density and the presence of certain transcription factor binding sites) encode and autonomously determine the local hypomethylated state at specific loci. Since (1) nucleotide sequences do confer adequate complexity for specific regional recognition along the genome and (2) the genome is the only hereditary material that can be faithfully replicated and passed on to progenies, the surmised dependency between DNA methylation state and the underlying sequence features does provide the most parsimonious and logically appealing explanation to the precise patterning and inheritance of the DNA methylome in vertebrates.

The results presented herein, however, definitively showed that, despite the strong statistical association (Chapter 1), there is no immediate connection between DNA methylation state and the underlying nucleotide sequence *in vivo* (Chapter 2). Importantly, disruption of the top enriched DNA motifs by the use of CRISPR-Cas9 failed to induce any change to the methylation state of the targeted hypomethyated regions. Moreover, by manipulating and controlling the methylation state of the genomic fragments prior to reintegration into the genome, it is clearly demonstrated that the artificially granted methylation states were predominantly maintained *in vivo*, independent of the nucleotide sequences and their endogenous methylation states. More importantly, the artificially conferred methylation states appear to be stably maintained in the medaka embryos, independent of the native states. Such robustness is corroborated and further exemplified by a previous medaka transgenic study (Uno

2015) that the endogenously hypermethylated genomic sequences were found hypomethylated after subcloning (hence unmethylated) and stable genome reintegration as transgenes, and such hypomethylated state were faithfully cascaded to the F2 generation, i.e. from F0 to F2: across > 6 months and 3 animal generations. The faithful maintenance of artificially conferred methylation states is in contradiction to the aforementioned prevailing belief that the vertebrate genome is hypermethylated "by default" except for the loci that contain the supposed hypomethylation determinants. Thus, the experimental results presented herein not only disproved the inferred determining role of DNA sequence on the methylation landscape, but also defied the long-standing belief that there is a default state (i.e. hypermethylation) for vertebrate genomes.

In fact, the postulated strict sequence-dependency is paradoxical to the concept of epigenetics itself. There are accumulating reports for the last two decades that DNA methylation could be perturbed by transient physiological stress or chemical exposure. More importantly, the perturbed states could be persistent and sometimes inheritable, while the underlying genomic sequence remains unchanged (Richards 2006; Bird 2007; Feil and Fraga 2012; Heard and Martienssen 2014). These observations highlighted that DNA methylation pattern is not directly coupled with the underlying genomic sequence *in vivo*, in spite of what has been repeatedly shown *in silico* and recently demonstrated *in vitro*.

The proven lack of immediate sequence-dependency in DNA methylome *in vivo* might appear to contradict the widely observed statistical association between genomic sequence and its DNA methylation pattern. While it is true that statistical correlation/association does not necessarily imply a causal relationship, strong statistical correspondence does often reflect the existence of direct or indirect relationship. Indeed, the correlation between methylation and nucleotide sequence pattern is well explainable. First, as is shown by the statistical modelling in Chapter 1, the precision and sensitivity of methylation state prediction and classification by nucleotide sequence (hence statistical association) rely substantially on the fact that hypomethylated regions tend to have higher CpG density than the hypermethylated regions. In fact, CpG density is well acknowledged to be one of the key *in silico* predictors of methylation state (Illingworth and Bird 2009). Yet, as explained in Chapter 1's *Discussion*, the relative enrichment of CpGs in hypomethylated regions can be accounted for by the spontaneous deamination, hence mutation, of methylated cytosine into thymine in the rest of the genome during vertebrate evolution. Therefore, the observed statistical association is, in part, indicative of the mutagenic consequence of the global methylation of vertebrate genomes; Second, as described in *General Introduction*, hypomethylated regions usually encompass gene regulatory elements, where binding sites of transcriptional regulators reside. The observed statistical association may hence reflect general enrichment of these binding sites in functional gene regulatory regions. For instance, calmodulin-binding transcription activators (CAMTAs), a family of general transcription activators that regulates a broad range of genes, also have a core binding motif of CGCG (Finkler et al. 2007), which is highly enriched in medaka's HypoMDs (see Chapter 1). Taken together, DNA methylation state and the underlying genomic sequence are conceivably related, albeit not necessarily in a causal relationship.

While this study does argue against the notion that that the genome-wide DNA methylation landscape is *autonomously* and *primarily* determined by the genomic sequence in vertebrates, the experimental results presented herein does not rule out the possibility that nucleotide sequence might serve as a secondary/companion element in patterning of DNA methylome. This hypothesis is corroborated by a preliminary study conducted by Fukushima et al. (personal communication) in Takeda's laboratory, the University of Tokyo. In their experiment, they surveyed the methylation state of a transgene construct that comprises ~140 kb-long medaka genomic region (except with two open reading frames replaced by coding sequences of fluorescent proteins). The transgenic line of medaka that hosts this construct (Moriyama et al. 2012) had been curated for at least 4 years and 16 animal generations. The authors identified that the periphery of the transgene construct in the long-curated transgenic line did largely, albeit imperfectly, recapitulate the methylation pattern of its endogenous counterpart, whereas the centre of the construct remained poorly methylated, unlike the endogenous locus. The distinct difference in methylation state between the peripheral and the centre of the construct suggests that the recapitulation of DNA methylation pattern on ectopic sequence could be a progressive process that spans a timescale of years and/or require many rounds of epigenetic reprogramming (which is triggered in-between animal generations, i.e. during gametogenesis and fertilization). The authors' observation also strongly indicates that local nucleotide sequence might serve as a guide that defines the ultimate DNA methylation pattern. Indeed, the possible secondary role of nucleotide sequence in DNA methylation patterning also implies possible mechanistic relationship between DNA sequence and methylation, which further help explain the strong statistical association between nucleotide sequence and DNA methylation state.

Given that the nucleotide sequence itself is not sufficient to determine its own methylation state, the patterning of DNA methylation landscape must involve other epigenetic factors as previously suggested by Kaminsky et al. (2009). Indeed, DNA sequences examined in Chapter 2, as well as in the aforementioned transgenic studies by Uno (2015) and Fukushima et al. (unpublished data), were all purified prior to being introduced into the genome, hence initially lacked any associated factor(s) that could participate in the determination/modulation of local DNA methylation. While the identity of these presumed determinants remains elusive, they are unlikely to be transcription factors as they could not be efficiently recruited onto the integrated DNA in a sequence-dependent manner within 6 months or 3 animal generations (Uno 2015). Nevertheless, it is beyond doubt that the identification for the genuine methylation determinant(s) is of utmost importance and urgently needed as this is imperative to the deciphering of the mechanism and logic behind DNA methylome establishment *in vivo*.

The strong link between DNA methylation, nucleosome position, and histone modifications (Fuks 2005; Cedar and Bergman 2009; Chodavarapu et al. 2010) suggests that local chromatin state could be one of the ultimate governors of DNA methylation state. However, proving this postulation experimentally is currently technically infeasible as molecular tools for manipulation of nucleosomes and histone modifications *in vivo* are not yet available. It is currently known that (1) there are 15 possible modifications (or 28 if the absence of functional group attachment is also considered as one form of 'modification') on 13 different amino acid residuals of the 4 histones (i.e. H2A, H2B, H3, and H4) that constitute the nucleosome core, (2) many of these modifications can co-exist or vary within the same nucleosome, and (3) H2A can be replaced by H2A.X (which can also be modified on 2 different amino acid residuals) or H2A.Z in some genomic loci and/or under certain cellular conditions. The overwhelming number of combinations of different modifications on different residuals of different histones is a formidable hurdle for identifying and proving which particular sets of modifications and variants are responsible for DNA methylation determination. Moreover, while there were a few recent *in vitro* endeavours in achieving site-specific alteration of histone states by the use of dCas9-tethered histone modifiers (Kearns et al. 2015; Zentner and Henikoff 2015; O'Geen et al. 2017), their usability in an *in vivo* context remains unproven. In fact, similar tools for the manipulation of most of the histone modifications and variants are not yet available nor currently conceivable both *in vitro* and *in vivo*, since many histone modifiers are associated with large multiprotein complexes (Bannister and Kouzarides 2011) that complicates construction of fusion proteins for tethering. Above all, unlike DNA methylation, how nucleosome and histone modifications are inherited across cell divisions remains poorly known. Further investigation into the mechanism of their deposition onto specific locations of the genome, and of their inheritance across cell division and animal generations is very much needed before they can be mechanistically related to the patterning of DNA methylome.

To conclude, this study strongly argues against the recent proposition that genome-wide DNA methylation pattern is primarily and autonomously determined by the underlying genomic sequence in vertebrates *in vivo*, but instead provides insights into potential involvement of other epigenetic factor(s) in defining the DNA methylation landscape. The data presented herein demonstrate that the DNA methylation landscape and genomic sequence are not directly coupled, which underpin the widely-observed plasticity of DNA methylation along differentiation, as well as the transgenerational inheritance of perturbed DNA methylation *in vivo*. However, it is worth noting that vertebrates could have variable DNA methylation dynamics, particularly during early embryonic development. This is true even within the same clade of vertebrate species (e.g., in mammals, the genome-wide methylation erasure immediately after fertilization is highly extensive in mice and human but very subtle or virtually absent from sheep; Young and Beaujean 2004). Further investigation in other vertebrate models will definitely be needed before generalization of the current observations made in medaka.

Materials and Methods

HypoMD and HyperMD calling

Published whole genome bisulfite sequencing reads of medaka blastula embryos (Qu et al. 2012) were fetched from the Data Bank of Japan (Accession number: SRX149583). Individual reads were trimmed to remove primers, adapters sequences, and low quality basecalls (Phred score ≤ 3) using BBDuk from the BBTools (Bushnell) ver. 35.85. Trimmed reads were mapped to the latest (as of the time of this writing) medaka genome assembly ver. 2.2.4 (all genome coordinates reported herein refer to this assembly version) using bwa-meth ver. 0.2.0. Methylation rates of the mapped CpG dyads were then extracted using MethylDackel ver. 0.2.1 with the default quality filters of MAPQ score ≥ 10 and Phred score ≥ 5. Only those CpG dinucleotides with at least 5× coverage were considered as valid calls (Qu et al. 2012) and the final mean coverage after filtering was 8×. The same filtering criteria were also applied to all experiments throughout this study, wherever they are applicable. The endogenous methylation states of sequences assayed in this study were directly extracted from this mapped, filtered dataset.

HypoMDs calling followed the same definition as previously published (Nakamura et al. 2014; Uno et al. 2016). Specifically, any stretch of ten or more hypomethylated (methylation rate < 40%) CpGs with no more than 4 interleaving non-hypomethylated (methylation rate ≥ 40%) or undetermined (unsampled, unmappable or low coverage) dyads were called as HypoMD. HyperMDs were analogously defined as any stretch of at least ten hypermethylated (methylation rate > 60%) CpG dyads containing no more than 4 interleaving non-hypermethylated (methylation rate ≤ 60%)

or undetermined CpGs. Length, GC content and CpG density of the called HypoMDs and HyperMDs were summarized in Figure 1.2.

Supervised classification of HypoMD and HyperMD using support vector machine

To elucidate whether HypoMDs and HyperMDs contain distinct sequence features, genomic sequences of all called HypoMDs (N = 18435) and HyperMDs (N = 231516) were subjected to supervised classification using kmer-SVM (support vector machine with string-, i.e. nucleotide sequences-, based spectrum kernel) (Fletez-Brant et al. 2013). The default, recommended parameters and k = 6 (i.e. 6-mer) were used. Proportionally higher weights were assigned to HypoMDs (weight = 231516 ÷ 18435 = 12.56) than HyperMDs (weight = 1) to offset the imbalanced sample sizes. Classification performance was gauged by 10-fold cross-validation and the area under precision-recall curves, as well as Matthew's correlation coefficient. In parallel, the possible confounding effect of differential CpG density between HypoMDs and HyperMDs was controlled for by masking all CpG dinucleotides (i.e. from 'CG' to 'NN') and the SVM model was retraining using the same parameters as listed above.

In vivo disruption of $(CG)_3$ and ACGT-repeat in HypoMDs

A total of 6 HypoMDs that contain a single $(CG)_3$ or CGnull-repeat immediately followed by spCas9's PAM (i.e. "NGG") were randomly selected for targeted disruption. spCas9 mRNA was produced from pMLM3613 (a gift from Keith Joung; Addgene plasmid #42251) via *in vitro* transcription. sgRNAs were manually designed to target the 18-nt immediately upstream of the PAM site. spCas9 mRNA and sgRNA were *in* *vitro* transcribed and purified as described in the section "*General procedure: In vitro transcription for the generation of mRNA and sgRNA*" below. spCas9 mRNA and sgRNA were co-injected (final concentration: 100 and 10 ng/μL, respectively, along with 0.05% phenol red) into drR medaka embryos at 1-cell stage. Approximately 200 embryos were injected with each sgRNA. The injected embryos were reared at 28°C to blastula stage (approx. 8 hours). Normally developed embryos (approx. 90%) were homogenized for genomic DNA extraction (see "*General procedure: Genomic DNA extraction from medaka blastula embryos*" below). The purified genomic DNA was bisulfite converted using MethylEasy Xceed Rapid DNA Bisulphite Modification Kit (Human Genetic Signatures, Australia) following manufacturer's instructions, except that the DNA was denatured at 42°C for 20 mins. The targeted loci were PCR-amplified using target-specific bisulfite PCR (BSP) primers designed in MethPrimer (Li and Dahiya 2002), TA-cloned using TOPO TA Cloning Kit, Dual Promoter (Thermo Fisher Scientific, USA) and transformed into *E. coli*, which were spread on lysogeny broth (LB) agar plate with ampicillin (50 μg/μL) to obtain single colonies. Individual colonies were randomly picked and expanded in LB with ampicillin (50 μg/μL). Plasmid from each of the expanded clones was extracted using QIAprep Miniprep Kit (Qiagen, USA) with the silica columns substituted by EconoSpin mini spin column (Epoch Life Science Inc., USA), then Sanger sequenced in ABI PRISM 3100 Genetic Analyzer using the BigDye 3.1 chemistry (ThermoFisher Scientific). Sequences containing indels at the sgRNA-target sites were selected and aligned to reference genomic sequence using MAFFT version 7. Methylation rate of CpGs within the target loci was manually enumerated (unmethylated C becomes T, while methylated C remains as C, after bisulfite PCR).

To obtain F1 generation that inherits disrupted $(CG)_3$ or CGnull-repeat in the HypoMDs, the injection experiment was repeated and the injected embryo were reared to adulthood (i.e. for approx. 3 months). The adult fish were genotyped (see "*General procedure: Genotyping*" below). Five pairs of male and female fish with indel mutation at the targeted locus were intercrossed to produce F1 offsprings. The F1 embryos were allowed to develop to blastula stage, then homogenized and extracted for genomic DNA, which was bisulfite converted, PCR amplified to obtain methylation state of the target loci as aforementioned.

In silico motif analysis

The canonical 6-mers with positive weights (i.e. relatively enriched HypoMDs) and screened for those containing consecutive CpG (i.e. CGCG), along those that constitute CGnull-repeats (i.e. AGCTAG, GCTAGC, TAGCTA), were matched against the 'Vertebrates (*in vivo* and *in silico*)' databases in TomTom ver. 4.11.4 (Gupta et al. 2007). Only matches with false discovery rate < 0.1 were retained and reported.

High-throughput transplantation of CpG-rich genomic loci

To ascertain whether nucleotide sequences can autonomously determine their own methylation state *in vivo* at genome-wide scale, CpG-rich genomic fragments were captured and injected into medaka zygotes for random reintegration into the genome, then fished out to check for their methylation state. The capturing method was akin to those described for reduced representation bisulfite sequencing (RRBS). In fact, procedures up to the size selection of adaptor-ligated genomic fragments closely followed those optimized for RRBS (Gu et al. 2011).

In details, to obtain bulk genomic DNA for library preparation, adult medaka (drR strain; ca. 14-month-old) was anesthetized in ice-cold water and decapitated. Muscles were scraped from the tail using a scalpel. The muscles were ground to slurry in approx. 500 µL of lysis buffer (100 mM Tris-HCl, 50 mM EDTA, and 1% SDS). The slurry was incubated with 100 µg/mL proteinase K (Sigma-Alrich, USA) at 56°C for two hours with occasional, gentle mixing, cooled to room temperature and extracted with equal volume of Tris-saturated phenol-chloroform (pH 8) followed by chloroform extraction. Nucleic acids were precipitated by the addition of 0.6× volume of isopropanol and pelleted at 17900×g for 15 mins, desalted by washing with 70% ethanol, air-dried, then re-dissolved in 50 µL TE buffer. RNAs were degraded with 10 µg/mL RNase A (Wako Pure Chemical Industries, Japan) at 37°C for 1 hour. Proteinase K digestion, organic extraction, and isopropanol precipitation were repeated to remove the RNase and most of the degraded RNA. The resultant nucleic acid pellet was redissolved in 50 µL of 10 mM Tris-HCl, pH 8.0 and quantitated using Nanodrop 2000 (Thermo Fisher Scientific). Approximately 20 ng of the extracted DNA was pre-stained with 1:600 GelRed (Biotium, USA) and eletrophoresized in 1% agarose gel, 0.5X TAE buffer for 1 hour to check for integrity.

For fragmentation and capturing, 1 µg of the genomic DNA was digested with 20 units of MspI (New England BioLabs, USA; a.k.a. NEB) in 50 µL of 1× NEB Buffer 2 at 37°C overnight. Two µL of the digestion product were electrophoresized in 1% agarose gel (0.5× TAE) to ensure complete digestion. The digestion product was cleaned up via phenol-chloroform extraction, ethanol precipitated and re-dissolved in 10 mM Tris-HCl, pH 8. End-filling and dA-tailing of the digestion product were carried out simultaneously: 300 ng of MspI-digested genomic fragments were incubated with

5 units of Klenow fragments (3'→5' exo-) and dNTP mix (10 mM dATP, 1 mM dCTP, and 1 mM dGTP) in 20 μL of 1× NEB Buffer 2 at 30°C for 20 mins, then at 37°C for another 20 mins. The product was purified via phenol-chloroform extraction and ethanol precipitation and were re-dissolved in 15 μL of 10 mM Tris-HCl, pH 8. The purified product was ligated to 0.75 μM of custom-made adapters using 2000 units of T4 DNA ligase (NEB) in 20 μL of 1× NEB T4 ligation buffer at 16°C overnight. The adaptor was prepared immediately before ligation by annealing 1 μL each of F3-02top and F3-02bottom (Appendix: Table 3) (synthesized by Thermo Fisher Scientific) in 10 μL of annealing buffer (10 mM Tris-HCl, pH8, 50 mM NaCl, and 1 mM EDTA) via denaturation at 95°C for 2 mins and ramping down to 25°C at a rate of -0.1°C/second in a PCR machine. Ligation product was electrophoresized in parallel with the 20 bp DNA ladder (Takara Bio, Japan) in 3% agarose gel, 0.5× TBE buffer at 100V for 2 hours. The gel was post-stained in 3× GelGreen (Biotium) in 0.5× TBE buffer. The gel lane containing the fragments was excised for 160-340 bp dsDNA equivalent (corresponds to 40-220 bp adaptor-ligation fragments due to the Y-shaped adaptor that retards the molecules' mobility in gel). Fragments were extracted from the excised gel and purified using Zymoclean Gel DNA Recovery Kit (Zymo Research, USA) according manufacturer's instructions, except that 500 ng of sheared salmon sperm DNA were spiked into the dissolved gel prior to column loading to minimize sample loss during washing. Purified DNA was eluted in 20 μL 10 mM Tris-HCl, pH 8 into low-binding microfuge tube (Eppendorf, Germany).

Adaptor-ligated fragments were enriched and amplified by PCR: 0.5 μL of eluted product, 200 μM dNTP, 300 nM each of F3-03F and F3-03R primers (containing, from 5' to 3' in this order, non-template I-SceI restriction sites, bisulfite PCR primer

F3-01 sites, and Dam site), and 10 units of PfuTurbo Cx polymerase in 200 μL (split into 8 tubes of 25 μL) of 1× PfuTurbo Cx reaction buffer (Agilent, USA) at 95°C 2 mins, 18 cycles of 95°C for 30 s, 65°C for 30s, and 72°C for 45 s, then followed by 72°C for 5 mins, finally held at 10°C until further processing. PCR product were pooled and purified with 1.8× volume of homemade SPRI magnetic beads (1:50-diluted carboxylated Sera-Mag Magnetic SpeedBeads in the SPRI buffer mix described below) (GE Healthcare, USA) and eluted in 42.5 μL of 10 mM Tris-HCl, pH 8. Two and a half microliters of the purified product were run in 3% agarose gel, 0.5× TBE buffer to check for properly selected sizes. Negative control was processed in parallel using identical procedures, except MspI was replaced with Milli-Q water in the beginning, and resulted in no amplification product.

The amplified fragments were then Dam-methylated by incubating with Dam methylase to facilitate downstream counter-selection of unintegrated fragments. In details, the captured, amplified genomic fragments have two Dam sites (5'-GATC-3') on the ligated adapters (one on each end; downstream of the BSP primer binding sites). The enriched CpG-rich fragments was tagged on the Dam sites via Dam methylation using 8 units of *dam* methyltransferase in 50 μL of 1× *dam* Methyltransferase Reaction Buffer (New England Biolabs) at 37°C overnight. Reaction product was ethanol precipitated, re-dissolved and re-incubated with fresh *dam* methyltransferase reaction mix overnight. Dam-methylated products were purified via phenol-chloroform extraction and ethanol precipitation, then re-dissolved in 20 μL of 10 mM Tris-HCl, pH 8. One microliter of the purified product was used for photometric quantification using Nanodrop 2000 (Thermo Fisher Scientific). Nine microliters were aliquoted for injection and the rest was subjected to artificial CpG methylation as follows.

CpG methylation of the DNA fragments *in vitro* was mediated by CpG methyltransferase M.SssI: 10 µL of Dam-methylated fragments were incubated with 4 units of M.SssI methyltransferase, 640 nM of fresh S-adenosylmethionine (a.k.a. SAM) in 50 µL of 1× NEB Buffer 2 at 37°C overnight. Reaction product was ethanol precipitated, re-dissolved and re-incubated overnight in the same volume of fresh M.SssI methyltransferase reaction mix, then purified via phenol-chloroform extraction and ethanol precipitation, finally re-dissolved in 10 µL of 10 mM Tris-HCl, pH 8. The purified product was quantitated using Nanodrop as above. All ethanol precipitation described above was carried out with spike-in of 1 µL of Ethachinmate (long, linear acrylamidic polymer) (Nippon Gene, Japan) as carrier to maximize DNA recovery.

Immediately prior to injection, the methylated fragments (final concentration: 10 ng/µL) were pre-treated with 5 units of I-SceI meganuclease in 20 µL of 1× I-SceI digestion buffer (New England Biolabs) at room temperature for 1 hour. Medaka zygotes were injected with Dam-methylated or Dam+CpG-methylated fragments at 1-cell stage following standard procedures (Kinoshita et al. 2009). Around 500 embryos were injected with each pools of fragments and were allowed to develop to the blastula stage, i.e. Stage 11 by Iwamatsu (2004), at 28°C. The embryos were visually inspected under dissecting microscope with dead or malformed embryos discarded. Ultimately, 496 (86%) and 433 (92%) embryos injected with Dam-methylated and Dam&CpG-methylated fragments, respectively, developed normally to the blastula stage, and from which genomic DNA with fragments integrated was extracted (see "*General procedure: Genomic DNA extraction from medaka blastula embryos*" below). While most of the unintegrated fragments were presumably removed using our optimized DNA extraction method that includes size selection by PEG precipitation, carryover was further minimized by incubating the extracted DNA with 2 μL of FastDigest DpnI (Thermo Fisher Scientific) in a 20 μL reaction volume for a total of 72 hours at 37°C in an incubator. This was followed by routine phenol-chloroform extraction and isopropanol precipitation. The precipitated DNA was finally re-dissolved in 20 μL of freshly dispensed Milli-Q water (Merck Millipore, USA).

Efficient removal of unintegrated fragments was indicated by the parallel use of uninjected, spike-in control. Approximately twice the amount of the injection cocktail was spiked into the lysate of uninjected blastula embryos, which was then processed as described above. Relative quantity of library with or without integration was gauged by real-time PCR (THUNDERBIRD SYBR qPCR Master Mix, TOYOBO, Japan; in Agilent Stratagene Mx3000P, USA) using the library-specific primers F3-01F and F3-01R. In parallel, input DNA was also quantitated using primers F3-04F and F3-04R. Amplification plots were imported into qpcR v1.4.0 (Ritz and Spiess 2008), where the relative quantities were enumerated after sigmoidal modelling (all adjusted $R^2 = 1.00$).

The purified genomic DNA was then bisulfite-converted as described above. Integrated fragments were enriched via PCR using BSP primers F3-01F and F3-01R. The BSP products were dA-tailed and ligated to Illumina TruSeq adapters, pooled, and sequenced using Illumina MiSeq system. In detail, the integrated fragments were enriched from the bisulfite converted genomic DNA by PCR amplification using two high fidelity polymerases, ExTaq (Takara Bio) and KAPA HiFi Uracil+ (Kapa Biosystems, USA) separately. The BSP primers were originally designed and optimized for the use in ExTaq reaction. However, KAPA HiFi Uracil+ was also included for its low amplification bias and serve as a cross-reference to the ExTaq library. For ExTaq reaction: 2 μL of bisulfite converted genomic DNA was used as template in 50 μL of

ExTaq reaction with the primers F3-01F and F3-01R (500 nM each) at 95°C for 2 mins, 35 cycles of 95°C for 30 s, 60°C for 30 s, 72°C for 30 s, then 72°C for 10 mins, and finally hold at 10°C until further processing; For Kapa HiFi Uracil+ reaction: 2 μL of bisulfite converted genomic DNA was used as template in 50 μL of KAPA HiFi HotStart Uracil+ ReadyMix (Kapa Biosystems, USA) reaction with the same primers (300 nM each) at 95°C for 3 mins, 35 cycles of 98°C for 20 s, 60°C for 15 s, 72°C for 15 s, then 72°C for 30 s, and finally hold at 10°C until further processing. PCR products were purified using 1.8× volume of homemade SPRI beads and eluted in 10 μL of 10 mM Tris-HCl, pH 8. Yield was measured via fluorometric quantification using Qubit dsDNA HS Assay Kit and Qubit 2.0 Fluorometer (Thermo Fisher Scientific). The remaining eluates (405 ng and 423 ng from embryos injected with Dam-methylated or Dam&CpG-methylated fragments, respectively) were ligated with FastGene Adapter Kit for Illumina (Nippon Genetics, Japan) using KAPA Hyper Prep Kit (Kapa Biosystems) according to manufacturer's instructions. No post-ligation library amplification was performed. Routine quality check was carried out using High Sensitivity DNA Kit on the 2100 Bioanalyzer (Agilent, USA). Quantitation of the adaptor-ligated library was accomplished using GenNext NGS Library Quantification Kit (TOYOBO, Japan) in Mx3005P qPCR System (Agilent). Libraries were pooled in equal molar amount, denatured and diluted to a final concentration of ca. 13 pM with 20% PhiX control spiked-in for sequencing (MiSeq Reagent Kit v2; Illumina, USA). The automated sequencing run was conducted in paired-end mode (150 bp from each ends). A cluster density of ca. 1200/mm^2 was achieved with over 85% of Phred score > 30. Sequencing outputs were minimally trimmed, mapped to genome, and called for methylation rate as aforementioned, except bwa-mem's "-U" switch was set to its default.

In order to relate the methylation state of the integrated fragments to possible binding or recognition by DNA-binding proteins (e.g., transcription factors), DNase I hypersensitive sites (DHS) were identified by remapping the publicly available DNase-seq dataset of drR medaka blastula embryos (accession number: SRX1032807) (Nakamura et al. 2017) to the medaka genome assembly v2.2.4. Adaptor trimming and alignment was accomplished using BBmap v37.36 (Bushnell) with default parameters. Aligned reads were filtered for a minimum MAPQ of 20. MACS v2.1.1.20160309 (Zhang et al. 2008) was subsequently used to called 112987 peaks (DHS) with the following switches: "-g 6.3e+8 --nomodel --shift -50 --extsize 100 -q 0.01". Vast majority (> 96%) of the assayed fragments were originated completely from either inside or outside, but not spanning across the boundaries, of DHS (Appendix: Table 5).

Transplantation of Hypo-/Hyper-MDs to specific genomic locus via site-specific recombination

An engineered transgenic line that carries an attP site inside a gene desert on chromosome 18 for PhiC31 integrase-mediated recombination was used for site-specific integration of the full-length, unmethylated HyperMDs (i.e. PCR-amplified, cloned, and without pretreating with M.SssI) and pre-methylated HypoMDs (i.e. PCR-amplified, cloned, and pretreated with M.SssI) with lengths of 300–400 bp.

PhiC31 integrase coding sequence was amplified from pPGK-PhiC31o-bpA (a gift from Philippe Soriano; Addgene plasmid #13795) and attached to SV40 nuclear localization sequence (NLS) using primer pair F4-01 and Phusion polymerase (Thermo Fisher Scientific), then blunt-end-cloned using Zero Blunt PCR Cloning Kit (Thermo Fisher Scientific). Cloning direction and proper coding sequence were checked via

Sanger sequencing (by FASMAC Co). PhiC31 integrase mRNA was generated from the constructed template via *in vitro* transcription (see "*General procedure: In vitro transcription for the generation of mRNA and sgRNA*" below).

Six HyperMDs (see Appendix: Table 4) with flanking BSP primer binding sites (for F3-01F and F3-01R) and Dam-sites (downstream of the BSP primer sites) were directly synthesized by Thermo Fisher Scientific and Integrated DNA Technologies (USA) as double-stranded DNA and cloned into the targeting vector pEx_MCS-attBtagRFPt (a gift from Joachim Wittbrodt; Addgene plasmid #48876). Eleven HypoMDs were amplified from drR genomic DNA and extended to include BSP primer binding sites and Dam-sites on both ends using primer sets primer sets F4-02 through F4-12, then cloned into the targeting vector. HyperMD-containing targeting vectors were propagated in *dam*$^+$ *E. coli* (DH5α) (Thermo Fisher Scientific), pooled in approximately equimolar amount. HypoMD-containing targeting vectors were similarly processed, except that the pooled library was further artificially methylated with CpG methyltransferase M.SssI and purified as aforementioned. Individual plasmid libraries (final concentration: 10 ng/μL) was injected with PhiC31 integrase mRNA (100 ng/μL) into >200 embryos of PhiC31 transgenic strain (Kirchmaier et al. 2013) at 1-cell stage. Injected embryos were reared at 28°C to blastula stage, screened for normal development (> 85%), homogenized, and extracted for genomic DNA (see "*General procedure: Genomic DNA extraction from medaka blastula embryos*" below). The extracted DNA was digested with DpnI to degrade unintegrated vectors, re-purified, bisulfite-converted, subjected to PCR via ExTaq polymerase, TA-cloned, Sanger sequenced, and quantified for methylation rate as aforementioned.

To ensure the injected but unintegrated vectors were efficiently removed, the above injection was also carried out without PhiC31 integrase mRNA. These injected embryos were processed in parallel with those injected with integrase mRNA up to DpnI digestion. The relative abundance of undigested libraries (both unintegrated and integrated) was quantified and normalized to amount of input genomic DNA using real-time PCR as described above (Figure 2.4, panel B).

In vivo 'methylome editing' via homology-directed repair

Homology-directed repair was triggered by CRISPR-Cas9-induced double-strand breaks. spCas9 mRNA was produced from pMLM3613 (a gift from Keith Joung; Addgene plasmid #42251) via *in vitro* transcription. The HypoMDs, chr17:6415960–6416269 (Locus 1) and chr21:25260707–25262742 (Locus 2), were randomly chosen as targets for editing. sgRNAs targeting these regions were designed using CCTop (Stemmer et al. 2015). The 6 top-ranked guide sequence designs (sets F5-01 and F5-02, for Locus 1 and 2, respectively) were synthesized (Thermo Fisher Scientific) and *in vitro* transcribed (see "*General procedure: In vitro transcription for the generation of mRNA and sgRNA*" below). To construct the repair template, these genomic regions (with 6 mutations to the targeted spCas9 PAMs, i.e. from 'NGG' to 'NGC', in order to protect the template from being cleaved by spCas9) along with their up- and down-stream sequences (800 bp on both sides) as homology arms were synthesized by Integrated DNA Technologies (USA), assembled, cloned into pCR-BluntII vector (Thermo Fisher Scientific) using NEBuilder assembly mix (New England Biolabs) and propagated by *dam*$^+$ E. coli. The repair templates were artificially methylated *in vitro* using CpG methyltransferase M.SssI and purified as described above. For each of the target regions, sgRNA cocktail, spCas9 mRNA, and artificially Dam&CpG-methylated repair template were co-injected into approx. 200 medaka (drR strain) embryos at 1-cell stage at high concentrations (25 ng/μL each, 600 ng/μL and 10 ng/μL, respectively, i.e. 750 ng/μL of RNA and 10 ng/μL DNA in total) to maximize editing rate. Injected embryos were reared at 28°C for approx. 8 hours to blastula stage, screened for normal development (> 75%) and extracted for genomic DNA, which was DpnI-treated to degrade the repair template, re-purified, and bisulfite-converted as aforementioned. The BSP primer pairs (F5-03 and F5-04 for Locus 1 and 2, respectively) were designed using MethPrimer 2.0 and screened for the presence of Dam-site (5'-GATC-3') within the target range. The amplification products were Sanger sequenced from both ends. The methylation rate of each CpG was estimated from the sequencing chromatograms as: C ÷ (C + T) × 100%, where C and T are the called peak height in the 'cytosine' (i.e. methylated cytosines, after bisulfite PCR) and 'thymine' (i.e. unmethylated cytosines, which were converted to uracil by bisulfite treatment, then to thymine by PCR) channels, respectively. The signal intensities were extracted in R v3.3.3 (https://www.r-project.org/) using the sangerseqR package v1.12.0 (Hill et al. 2014).

To estimate the editing rate, regions containing the sgRNA target sites was PCR-amplified from unconverted DNA using primer sets F5-05 through F5-09. Editing rate was gauged by the relative frequency of mutated sgRNA PAMs (5'-NGC-3'; on the edited alleles) versus the native PAMs (5'-NGG-3'; i.e. unedited alleles) from the Sanger sequencing trace using the same approach as described above. Editing efficiency was estimated to be 92.04% and 85.10% for Locus 1 and 2, respectively.

To collect edited embryos at later developmental stages (3 and 7 day-post-fertilization; dpf), the above cocktail was diluted 10-fold (in Milli-Q water; Merck

Millipore) immediately prior to injection to reduce the toxicity (manifested after gastrulation) of ultra-high nuclei acid concentration at the expense of efficient editing. DNA extraction and subsequent processing were carried as above. Estimated editing efficiency for Locus 1 = 9.56% (at 3 dpf) and 7.81% (at 7 dpf); Locus 2 = 27.16% (at 3 dpf) and 18.69% (at 7 dpf). In order to enable comparison across sampling time-points with variable editing rates, the estimated methylation rates were normalized to the editing efficiency (i.e. "normalized methylation rate" = "methylation rate" ÷ "editing rate").

General procedure: *In vitro* transcription for the generation of mRNA and sgRNA

To produce mRNA for injection, the *in vitro* transcription template was amplified from the vector (using available standard sequencing primers other than T7 or SP6 promoter) and gel purified using Zymoclean Gel DNA Recovery Kit (Zymo Research). *In vitro* transcription and polyA-tailing were carried out using HiScribe ARCA mRNA Kit (NEB) according to manufacturer's instructions, except that the T7 enzyme mix was replaced by SP6 polymerase mix (Thermo Fisher Scientific) if the template was to be transcribed from the SP6 promoter. Capped, polyA-tailed synthetic mRNA was purified using RNeasy Mini Kit (Qiagen, USA); sgRNAs for CRISPR-Cas9 were produced using EnGen sgRNA Synthesis Kit, S. pyogenes (NEB) following the supplied protocol. *In vitro* transcribed RNAs were purified via RNA Clean & Concentrator-5 (Zymo Research). For *in vivo* methylome editing, 6 sgRNAs designed for the same locus were pooled in equal molar amount, ethanol precipitated, and re-dissolved in freshly dispensed Milli-Q water (Merck Millipore). Purified RNAs were all quantitated using

Nanodrop 2000 (Thermo Fisher Scientific). Approx. 100 ng of the RNA was electrophoresized in 1% agarose gel, 0.5× TAE (i.e. under native condition) to check for integrity.

General procedure: Genomic DNA extraction from medaka blastula embryos

Up to 20 blastula embryos were dispensed into individual 1.5 mL microfuge tubes and were homogenized in 200 μL of 2× CTAB extraction buffer containing 2% CTAB (cetrimonium bromide) (Wako Pure Chemical Industries, Japan), 100 mM Tris-HCl (pH 8), 20 mM EDTA, and 1.4 M NaCl, supplemented with 100 μg/mL of proteinase K, and incubated at 65°C overnight. The product was extracted with an equal volume of chloroform. Up to four aqueous layers were pooled to minimize pipetting. CTAB was replenished by the addition of 0.1× volume of 10% CTAB solution (pre-heated to 65°C to re-dissolve any CTAB precipitate). The mixture was mixed gently by inversion and re-extracted with roughly equal volume of chloroform. Nucleic acids were precipitated by mixing with 0.5× vol. of SPRI buffer (20% PEG6000, 2.5 M NaCl, 10 mM Tris-HCl, 1 mM EDTA, 0.05% Tween-20, pH 8) and pelleted at 17900×g, room temperature for 30 mins, desalt by washing with 70% ethanol after complete removal of the supernatant, re-spun, air-dried, and re-dissolved in 10 μL freshly dispensed Milli-Q water. Dissolution was carried out at 65°C for ca. 15 mins in an incubator, then cooled on ice. RNA degradation was carried out by the addition of RNase A (Wako Pure Chemical Industries) to a final concentration of 1 μg/μL and incubation at 37°C for 60 mins. To check the DNA integrity and quantity, 0.5 μL of the extracted genomic DNA was electrophoresized in 1% agarose gel, 0.5× TAE. The purified DNA was immediately subjected to downstream enzymatic treatment or bisuflite conversion without being stored frozen. Under these conditions, all of the purified sample DNA achieved bisulfite conversion rate of about 99%.

General procedure: Genotyping

Founder F0 medaka with indel mutation at the targeted locus were identified by routine genotyping procedure. Genomic DNA was crudely extracted from fin clip using the "HotSHOT" method: fin clip was completely submerged and lysed in lysis solution (25mM NaOH, 0.2mM EDTA) at 95°C for 15 mins. The crude lysate was cooled down to 4°C, then neutralized with equal volume of 40 mM Tris-HCl, pH 8, mixed and centrifuged briefly. The supernatant was directly pipetted from the liquid surface and used as PCR template. The target locus was amplified by routine PCR with amplicon size < 500 bp. The PCR product was then denatured at 95°C for 2mins, then renatured by ramped down to 25°C at a rate of -0.1°C/second. One μL of the renatured product was digested with 0.5 units of T7 Endonuclease I (T7E1) in 5 μL 1× NEB Buffer 2 at 37°C for 30 mins. To generate wild-type control, the same procedures were repeated using wild-type fish (i.e. not injected with CRISPR). The digestion products were electrophoresed in 2% agarose gel. Presence of indel would result in extra bands (relative to the control) at lower molecular weights, due the formation of heteroduplex, which was cleaved by T7E1, after renaturation.

References

Angermueller C, Lee HJ, Reik W, Stegle O. 2017. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 18:90. doi:10.1186/s13059-017-1189-z.

Bajic VB, Seah SH, Chong A, Zhang G, Koh JLY, Brusic V. 2002. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics 18:198–199.

Bannister AJ, Kouzarides T. 2011. Regulation of chromatin by histone modifications. Cell Res. 21:381–395. doi:10.1038/cr.2011.22.

Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, Akalin A, Schübeler D. 2015. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature 520:243–247. doi:10.1038/nature14176.

Beisel C, Paro R. 2011. Silencing chromatin: comparing modes and mechanisms. Nat. Rev. Genet. 12:123–135. doi:10.1038/nrg2932.

Bhasin M, Zhang H, Reinherz EL, Reche PA. 2005. Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett. 579:4302–4308. doi:10.1016/j.febslet.2005.07.002.

Bird A. 2002. DNA methylation patterns and epigenetic memory. Genes Dev. 16:6–21. doi:10.1101/gad.947102.

Bird A. 2007. Perceptions of epigenetics. Nature 447:396–398. doi:10.1038/nature05913.

Bird AP. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. 8:1499–1504. doi:10.1093/nar/8.7.1499.

Bird AP. 1995. Gene number, noise reduction and biological complexity. Trends Genet. 11:94–100. doi:10.1016/S0168-9525(00)89009-5.

Blackledge NP, Long HK, Zhou JC, Kriaucionis S, Patient R, Klose RJ. 2012. Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. Nucleic Acids Res. 40:e32–e32. doi:10.1093/nar/gkr1207.

Blattler A, Farnham PJ. 2013. Cross-talk between site-specific transcription factors and DNA methylation states. J. Biol. Chem. 288:34287–34294. doi:10.1074/jbc.R113.512517.

Bock C, Lengauer T. 2008. Computational epigenetics. Bioinformatics 24:1–10. doi:10.1093/bioinformatics/btm546.

Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, Walter J. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet. 2:e26. doi:10.1371/journal.pgen.0020026.

Bushnell B. BBMap. [accessed 2017 May 28]. https://sourceforge.net/projects/bbmap/.

Castillo-Aguilera O, Depreux P, Halby L, Arimondo P, Goossens L. 2017. DNA methylation targeting: the DNMT/HMT crosstalk challenge. Biomolecules 7:3. doi:10.3390/biom7010003.

Cedar H, Bergman Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. Nat. Rev. Genet. 10:295–304. doi:10.1038/nrg2540.

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, Hetzel JA, Kuo F, Kim J, Cokus SJ, et al. 2010. Relationship between nucleosome positioning and DNA methylation. Nature 466:388–392. doi:10.1038/nature09147.

Colot V, Rossignol JL. 1999. Eukaryotic DNA methylation as an evolutionary device. BioEssays 21:402–411. doi:10.1002/(SICI)1521-1878(199905)21:5<402::AID-BIES7>3.0.CO;2-B.

Das R, Dimitrova N, Xuan Z, Rollins RA, Haghighi F, Edwards JR, Ju J, Bestor TH, Zhang MQ. 2006. Computational prediction of methylation status in human genomic sequences. Proc. Natl. Acad. Sci. 103:10713–10716. doi:10.1073/pnas.0602949103.

Dawlaty MM, Breiling A, Le T, Barrasa MI, Raddatz G, Gao Q, Powell BE, Cheng AW, Faull KF, Lyko F, et al. 2014. Loss of Tet Enzymes Compromises Proper Differentiation of Embryonic Stem Cells. Dev. Cell 29:102–111. doi:10.1016/j.devcel.2014.03.003.

Drew HR, Dickerson RE. 1981. Conformation and dynamics in a Z-DNA tetramer. J. Mol. Biol. 152:723–736. doi:10.1016/0022-2836(81)90124-8.

Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R, et al. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. doi:10.1038/nature11247.

Edwards JR, O'Donnell AH, Rollins RA, Peckham HE, Lee C, Milekic MH, Chanrion B, Fu Y, Su T, Hibshoosh H, et al. 2010. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Res. 20:972–980. doi:10.1101/gr.101535.109.

Egger G, Liang G, Aparicio A, Jones PA. 2004. Epigenetics in human disease and prospects for epigenetic therapy. Nature 429:457–463. doi:10.1038/nature02625.

Fan S, Zhang MQ, Zhang X. 2008. Histone methylation marks play important roles in predicting the methylation status of CpG islands. Biochem. Biophys. Res. Commun. 374:559–564. doi:10.1016/j.bbrc.2008.07.077.

Fang F, Fan S, Zhang X, Zhang MQ. 2006. Predicting methylation status of CpG islands in the human brain. Bioinformatics 22:2204–2209. doi:10.1093/bioinformatics/btl377.

Fatemi M. 2005. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic Acids Res. 33:e176–e176. doi:10.1093/nar/gni180.

Feil R, Fraga MF. 2012. Epigenetics and the environment: emerging patterns and implications. Nat. Rev. Genet. 13:97–109. doi:10.1038/nrg3142.

Fiedler T, Rehmsmeier M. 2006. jPREdictor: a versatile tool for the prediction of cis-regulatory elements. Nucleic Acids Res. 34:W546–W550. doi:10.1093/nar/gkl250.

Finkler A, Ashery-Padan R, Fromm H. 2007. CAMTAs: calmodulin-binding transcription activators from plants to human. FEBS Lett. 581:3893–3898. doi:10.1016/j.febslet.2007.07.051.

Fletez-Brant C, Lee D, McCallion AS, Beer MA. 2013. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. Nucleic Acids Res. 41:W544–W556. doi:10.1093/nar/gkt519.

Fuks F. 2005. DNA methylation and histone modifications: teaming up to silence genes. Curr. Opin. Genet. Dev. 15:490–495. doi:10.1016/j.gde.2005.08.002.

Goldberg AD, Allis CD, Bernstein E. 2007. Epigenetics: a landscape takes shape. Cell 128:635–638. doi:10.1016/j.cell.2007.02.006.

Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat. Protoc. 6:468–481. doi:10.1038/nprot.2010.190.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble W. 2007. Quantifying similarity between motifs. Genome Biol. 8:R24. doi:10.1186/gb-2007-8-2-r24.

Heard E, Martienssen RA. 2014. Transgenerational epigenetic inheritance: myths and mechanisms. Cell 157:95–109. doi:10.1016/j.cell.2014.02.045.

Hill JT, Demarest BL, Bisgrove BW, Su YC, Smith M, Yost HJ. 2014. Poly peak parser: Method and software for identification of unknown indels using sanger sequencing of polymerase chain reaction products. Dev. Dyn. 243:1632–1636. doi:10.1002/dvdy.24183.

Hodges E, Molaro A, Dos Santos CO, Thekkat P, Song Q, Uren PJ, Park J, Butler J, Rafii S, McCombie WR, et al. 2011. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. Mol. Cell 44:17–28. doi:10.1016/j.molcel.2011.08.026.

Holliday R, Grigg GW. 1993. DNA methylation and mutation. Mutat. Res. 285:61–67. doi:10.1016/0027-5107(93)90052-H.

Houseman E, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13:86. doi:10.1186/1471-2105-13-86.

Illingworth RS, Bird AP. 2009. CpG islands - A rough guide? FEBS Lett. 583:1713–1720. doi:10.1016/j.febslet.2009.04.012.

Iwamatsu T. 2004. Stages of normal development in the medaka *Oryzias latipes*. Mech. Dev. 121:605–618. doi:10.1016/j.mod.2004.03.012.

Jin B, Li Y, Robertson KD. 2011. DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy? Genes Cancer 2:607–617. doi:10.1177/1947601910393957.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13:484–492. doi:10.1038/nrg3230.

Jones PA, Liang G. 2009. Rethinking how DNA methylation patterns are maintained. Nat. Rev. Genet. 10:805–811. doi:10.1038/nrg2651.

Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GHT, Wong AHC, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, et al. 2009. DNA methylation profiles in monozygotic and dizygotic twins. Nat. Genet. 41:240–245. doi:10.1038/ng.286.

Kearns NA, Pham H, Tabak B, Genga RM, Silverstein NJ, Garber M, Maehr R. 2015. Functional annotation of native enhancers with a Cas9–histone demethylase fusion. Nat. Methods 12:401–403. doi:10.1038/nmeth.3325.

Kim S, Li M, Paik H, Nephew K, Shi H, Kramer R, Xu D, Huang TH. 2008. Predicting DNA methylation susceptibility using CpG flanking sequences. Pac. Symp. Biocomput. Pac. Symp. Biocomput.:315–326.

Kinoshita M, Murata K, Naruse K, Tanaka M, editors. 2009. Medaka: biology, management, and experimental protocols. Ames, Iowa: Wiley-Blackwell.

Kirchmaier S, Hockendorf B, Moller EK, Bornhorst D, Spitz F, Wittbrodt J. 2013. Efficient site-specific transgenesis and enhancer activity tests in medaka using PhiC31 integrase. Development 140:4287–4295. doi:10.1242/dev.096081.

Kirchmaier S, Naruse K, Wittbrodt J, Loosli F. 2015. The genomic and genetic toolbox of the teleost medaka (*Oryzias latipes*). Genetics 199:905–918. doi:10.1534/genetics.114.173849.

Kohli RM, Zhang Y. 2013. TET enzymes, TDG and the dynamics of DNA demethylation. Nature 502:472–479. doi:10.1038/nature12750.

Krebs AR, Dessus-Babus S, Burger L, Schübeler D. 2014. High-throughput engineering of a mammalian genome reveals building principles of methylation states at CG rich regions. eLife 3:e04094. doi: 10.7554/eLife.04094

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. Nat. Genet. 47:955–961. doi:10.1038/ng.3331.

Li E, Bestor TH, Jaenisch R. 1992. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell 69:915–926. doi:10.1016/0092-8674(92)90611-F.

Li LC, Dahiya R. 2002. MethPrimer: designing primers for methylation PCRs. Bioinformatics 18:1427–1431.

Li Q, Xu W, Cui Y, Ma L, Richards J, Li W, Ma Y, Fu G, Bythwood T, Wang Y, et al. 2015. A preliminary exploration on DNA methylation of transgene across generations in transgenic rats. Sci. Rep. 5:8292. doi:10.1038/srep08292.

Liao J, Karnik R, Gu H, Ziller MJ, Clement K, Tsankov AM, Akopian V, Gifford CA, Donaghey J, Galonska C, et al. 2015. Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. Nat. Genet. 47:469–478. doi:10.1038/ng.3258.

Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schübeler D. 2011. Identification of genetic elements that autonomously determine DNA methylation states. Nat. Genet. 43:1091–1097. doi:10.1038/ng.946.

Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322. doi:10.1038/nature08514.

Long HK, King HW, Patient RK, Odom DT, Klose RJ. 2016. Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved. Nucleic Acids Res. 44:6693–6706. doi:10.1093/nar/gkw258.

Ma B, Wilker EH, Willis-Owen SAG, Byun HM, Wong KCC, Motta V, Baccarelli AA, Schwartz J, Cookson WOCM, Khabbaz K, et al. 2014. Predicting DNA methylation level across human tissues. Nucleic Acids Res. 42:3515–3528. doi:10.1093/nar/gkt1380.

Marchal C, Miotto B. 2015. Emerging concept in DNA methylation: role of transcription factors in shaping DNA methylation patterns. J. Cell. Physiol. 230:743–751. doi:10.1002/jcp.24836.

Mathe C. 2002. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30:4103–4117. doi:10.1093/nar/gkf543.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, et al. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454:766–771. doi:10.1038/nature07107.

Mohn F, Weber M, Rebhan M, Roloff TC, Richter J, Stadler MB, Bibel M, Schübeler D. 2008. Lineage-specific polycomb targets and *de novo* DNA methylation define restriction and potential of neuronal progenitors. Mol. Cell 30:755–766. doi:10.1016/j.molcel.2008.05.007.

Moriyama Y, Kawanishi T, Nakamura R, Tsukahara T, Sumiyama K, Suster ML, Kawakami K, Toyoda A, Fujiyama A, Yasuoka Y, et al. 2012. The medaka zic1/zic4 mutant provides molecular insights into teleost caudal fin evolution. Curr. Biol. CB 22:601–607. doi:10.1016/j.cub.2012.01.063.

Mount DW. 2004. Bioinformatics: sequence and genome analysis. 2nd ed. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press.

Nakamura R, Tsukahara T, Qu W, Ichikawa K, Otsuka T, Ogoshi K, Saito TL, 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. doi:10.1242/dev.108548.

Nakamura R, Uno A, Kumagai M, Morishita S, Takeda H. 2017. Hypomethylated domain-enriched DNA motifs prepattern the accessible nucleosome organization in teleosts. Epigenetics Chromatin 10:44. doi:10.1186/s13072-017-0152-2.

Noble WS, Kuehn S, Thurman R, Yu M, Stamatoyannopoulos J. 2005. Predicting the in vivo signature of human gene regulatory sequences. Bioinformatics 21:i338–i343. doi:10.1093/bioinformatics/bti1047.

Noyer-Weidner M, Trautner TA. 1993. Methylation of DNA in Prokaryotes. In: Jost JP, Saluz HP, editors. DNA Methylation. Vol. 64. Basel: Birkhäuser Basel. (EXS). p. 39–108.

O'Geen H, Ren C, Nicolet CM, Perez AA, Halmai J, Le VM, Mackay JP, Farnham PJ, Segal DJ. 2017. dCas9-based epigenome editing suggests acquisition of histone methylation is not sufficient for target gene repression. Nucleic Acids Res. 45:9901–9916. doi:10.1093/nar/gkx578.

Pedersen AG, Baldi P, Chauvin Y, Brunak S. 1999. The biology of eukaryotic promoter prediction—a review. Comput. Chem. 23:191–207. doi:10.1016/S0097-8485(99)00015-7.

Peticolas W, Wang Y, Thomas G. 1988. Some rules for predicting the base-sequence dependence of DNA conformation. Proc. Natl. Acad. Sci. 85:2579–2583.

Pollack Y, Stein R, Razin A, Cedar H. 1980. Methylation of foreign DNA sequences in eukaryotic cells. Proc. Natl. Acad. Sci. 77:6463–6467.

Previti C, Harari O, Zwir I, del Val C. 2009. Profile analysis and prediction of tissue-specific CpG island methylation classes. BMC Bioinformatics 10:116. doi:10.1186/1471-2105-10-116.

Qu W, Hashimoto S, Shimada A, Nakatani Y, Ichikawa K, Saito TL, 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. doi:10.1101/gr.140236.112.

Rebuzzini P, Zuccotti M, Redi CA, Garagna S. 2016. Achilles' heel of pluripotent stem cells: genetic, genomic and epigenetic variations during prolonged culture. Cell. Mol. Life Sci. 73:2453–2466. doi:10.1007/s00018-016-2171-8.

Reeve ECR, Black I, editors. 2013. Genetics of cells organelles, structure and functions. In: Encyclopedia of genetics. 3rd ed. New York: Routledge. p. 752.

Rich A, Zhang S. 2003. Z-DNA: the long road to biological function. Nat. Rev. Genet. 4:566–572. doi:10.1038/nrg1115.

Richards EJ. 2006. Inherited epigenetic variation — revisiting soft inheritance. Nat. Rev. Genet. 7:395–401. doi:10.1038/nrg1834.

Ritz C, Spiess A-N. 2008. qpcR: an R package for sigmoidal model selection in quantitative real-time polymerase chain reaction analysis. Bioinforma. Oxf. Engl. 24:1549–1551. doi:10.1093/bioinformatics/btn227.

Robertson KD, Wolffe AP. 2000. DNA methylation in health and disease. Nat. Rev. Genet. 1:11–19. doi:10.1038/35049533.

Rollins RA. 2005. Large-scale structure of genomic methylation patterns. Genome Res. 16:157–163. doi:10.1101/gr.4362006.

Saadeh H, Schulz R. 2014. Protection of CpG islands against *de novo* DNA methylation during oogenesis is associated with the recognition site of E2f1 and E2f2. Epigenetics Chromatin 7:26. doi:10.1186/1756-8935-7-26.

Sanger F. 2005. Frederick Sanger - Biographical. Nobelprize.org. [accessed 2017 Jun 13]. https://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980/sanger-bio.html.

Scherf M, Klingenhoff A, Werner T. 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J. Mol. Biol. 297:599–606. doi:10.1006/jmbi.2000.3589.

Schübeler D. 2015. Function and information content of DNA methylation. Nature 517:321–326. doi:10.1038/nature14192.

Shakked Z, Rabinovich D. 1986. The effect of the base sequence on the fine structure of the DNA double helix. Prog. Biophys. Mol. Biol. 47:159–195. doi:10.1016/0079-6107(86)90013-1.

Siegfried Z, Cedar H. 1997. DNA methylation: a molecular lock. Curr. Biol. 7:R305–R307. doi:10.1016/S0960-9822(06)00144-8.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. 8:272–285. doi:10.1038/nrg2072.

Spivakov M, Fisher AG. 2007. Epigenetic signatures of stem-cell identity. Nat. Rev. Genet. 8:263–271. doi:10.1038/nrg2046.

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, Wirbelauer C, Oakeley EJ, Gaidatzis D, Tiwari VK, et al. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature 480:490–495. doi:10.1038/nature10716.

Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19:ii215–ii225. doi:10.1093/bioinformatics/btg1080.

Stein R, Gruenbaum Y, Pollack Y, Razin A, Cedar H. 1982. Clonal inheritance of the pattern of DNA methylation in mouse cells. Proc. Natl. Acad. Sci. U. S. A. 79:61–65.

Stemmer M, Thumberger T, Del Sol Keyer M, Wittbrodt J, Mateo JL. 2015. CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. PLOS ONE 10:e0124633. doi:10.1371/journal.pone.0124633.

Su J, Zhang Y, Lv J, Liu H, Tang X, Wang F, Qi Y, Feng Y, Li X. 2010. CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. Nucleic Acids Res. 38:e6. doi:10.1093/nar/gkp882.

Takahashi Y, Wu J, Suzuki K, Martinez-Redondo P, Li M, Liao HK, Wu MZ, Hernández-Benítez R, Hishida T, Shokhirev MN, et al. 2017. Integration of CpG-free DNA induces *de novo* methylation of CpG islands in pluripotent stem cells. Science 356:503–508. doi:10.1126/science.aag3260.

Thermes V, Grabher C, Ristoratore F, Bourrat F, Choulika A, Wittbrodt J, Joly JS. 2002. I-SceI meganuclease mediates highly efficient transgenesis in fish. Mech. Dev. 118:91–98. doi:10.1016/S0925-4773(02)00218-6.

Tran-Dinh S, Taboury J, Neumann JM, Huynh-Dinh T, Genissel B, Gouyette C, Igolen J. 1983. B and Z double helical conformations of d-(m5C-G-C-G-m5C-G) in aqueous solution. FEBS Lett. 154:407–410. doi:10.1016/0014-5793(83)80192-6.

Tsumura A, Hayakawa T, Kumaki Y, Takebayashi S, Sakaue M, Matsuoka C, Shimotohno K, Ishikawa F, Li E, Ueda HR, et al. 2006. Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes Cells 11:805–814. doi:10.1111/j.1365-2443.2006.00984.x.

Uno A. 2015. Comparative analysis of genome and epigenome between two polymorphic medaka populations [Doctoral dissertation]. [Japan]: The University of Tokyo.

Uno A, Nakamura R, Tsukahara T, Qu W, Sugano S, Suzuki Y, Morishita S, Takeda H. 2016. Comparative analysis of genome and epigenome in closely related medaka species identifies conserved sequence preferences for DNA hypomethylated domains. Zoolog. Sci. 33:358–365. doi:10.2108/zs160030.

Waddington CH. 1942. The epigenotype. Endeavour:18–20.

Waddington CH. 2012. The epigenotype (reprint). Int. J. Epidemiol. 41:10–13. doi:10.1093/ije/dyr184.

Walter RB, Li HY, Intano GW, Kazianis S, Walter CA. 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. doi:10.1016/S1096-4959(02)00144-6.

Wang Z, Willard HF, Mukherjee S, Furey TS. 2006. Evidence of influence of genomic DNA sequence on human X chromosome inactivation. PLoS Comput. Biol. 2:e113. doi:10.1371/journal.pcbi.0020113.

Wasserman WW, Sandelin A. 2004. Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. 5:276–287. doi:10.1038/nrg1315.

Wigler M. 1981. The somatic replication of DNA methylation. Cell 24:33–40. doi:10.1016/0092-8674(81)90498-0.

Wu H, Tao J, Sun YE. 2012. Regulation and function of mammalian DNA methylation patterns: a genomic perspective. Brief. Funct. Genomics 11:240–250. doi:10.1093/bfgp/els011.

Wu SC, Zhang Y. 2010. Active DNA demethylation: many roads lead to Rome. Nat. Rev. Mol. Cell Biol. 11:607–620. doi:10.1038/nrm2950.

Xie W, Barr C, Kim A, Yue F, Lee A, Eubanks J, Dempster E, Ren B. 2012. Base-resolution analyses of sequence and parent-of-origin dependent DNA methylation in the mouse genome. Cell 148:816–831. doi:10.1016/j.cell.2011.12.035.

Xu C, Bian C, Lam R, Dong A, Min J. 2011. The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. Nat. Commun. 2:227. doi:10.1038/ncomms1237.

Xu Y, Wu F, Tan L, Kong L, Xiong L, Deng J, Barbera AJ, Zheng L, Zhang H, Huang S, et al. 2011. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. Mol. Cell 42:451–464. doi:10.1016/j.molcel.2011.04.005.

Young L., Beaujean N. 2004. DNA methylation in the preimplantation embryo: the differing stories of the mouse and sheep. Anim. Reprod. Sci. 82–83:61–78. doi:10.1016/j.anireprosci.2004.05.020.

Zentner GE, Henikoff S. 2015. Epigenome editing made easy. Nat. Biotechnol. 33:606–607. doi:10.1038/nbt.3248.

Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. 2015. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. 16:14. doi:10.1186/s13059-015-0581-9.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9:R137. doi:10.1186/gb-2008-9-9-r137.

Zheng H, Wu H, Li J, Jiang SW. 2013. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. BMC Med. Genomics 6:S13. doi:10.1186/1755-8794-6-S1-S13.

Zhou X, Li Z, Dai Z, Zou X. 2012. Prediction of methylation CpGs and their methylation degrees in human DNA sequences. Comput. Biol. Med. 42:408–413. doi:10.1016/j.compbiomed.2011.12.008.

Appendix

Table 2 List of HypoMDs used for testing the anticipated governing role of $(CG)_3$ and CG-null repeats on local methylation state.

Locus 1–3 contain single $(CG)_3$ motif, whereas Locus 4–6 contain single CGnull-repeat motif. The motifs are in close proximity to spCas9's PAM (5'-NGG-3'), allowing efficient knockout using CRISPR.

Locus	Genome coordinate	sgRNA guide sequence (5' to PAM)
1	chr2:3940468-3940729	GTGACGCAAATCCGCGCG
2	chr 5:2430624-2430994	GGGCGCTAGGACCGCGCG
3	chr 6:18690493-18691397	CCGCACTTTCTCCGCGCG
4	chr 17:5828017-5828757	TGACAGCCAGCTAGCTCG
5	chr 10:25324929-25325607	GTGCTAATGCAAGCTAGC
6	chr 13:14404288-14405391	TTGTAGTTGTTGCTAGCT

Table 3 List of oligo used.

Name	Sequence (5′ to 3′)
F3-01F	GGAGTGAAGGAGGTTAGGGGTAAGT
F3-01R	AAAAACCATAAAACCCTATACCTAATCTATC
F3-02top	ACACTCTTTCCCTACACGACGCTCTTCCGATC-phosphorothioate-T
F3-02bottom	5′phosphate-GATCGGAAGAGCTCGTATGCCGTCTTCTGCTTG
F3-03F	NNNNNNTAGGGATAACAGGGTAATGGAGTGAAGGAGGTTAGGG- -GTAAGTACACTCTTTCCCTACACGACGCTCTTCCGATCT
F3-03R	NNNNNNTAGGGATAACAGGGTAATAAAAACCATAAAACCCTATA- -CCTAATCTATCCAAGCAGAAGACGGCATACGAGCTCTTCCGATCT
F3-04F	AAAAGTCTCAACACTGCCTCC
F3-04R	AGAGCCTTCCATGTTTGACC
F4-01F	GCCGCCACCATGGATACCTA
F4-01R	CTATCACACTTTCCGCTTTTTCT
F4-02F	GGTTAGGGGTAAGTAGATCTGAATTAAAGAAAGTTCTACAAAC
F4-02R	ATACCTAATCTATCAGATCTACATCATCCTCTTCATCATTGA
F4-03F	GGTTAGGGGTAAGTAGATCTGGGCAGAAACGCATTTTGGG
F4-03R	ATACCTAATCTATCAGATCTTTTGAAATTGAAAATCAAGCA
F4-04F	GGTTAGGGGTAAGTAGATCTCGGACATGCAGAGCTTCG
F4-04R	ATACCTAATCTATCAGATCTAGAAGTCTGATCCTTGTCAGA
F4-05F	GGTTAGGGGTAAGTAGATCTGCATTCTGTGCACGAGACG
F4-05R	ATACCTAATCTATCAGATCTGGGAAGAGTGGCGAGTAGTT
F4-06F	GGTTAGGGGTAAGTAGATCTGCACTCAAAGGCAGCAGG
F4-06R	ATACCTAATCTATCAGATCTAGAACATAAACATCAACCACGGT
F4-07F	GGTTAGGGGTAAGTAGATCTGGAAATAACAGCGGACTACGG
F4-07R	ATACCTAATCTATCAGATCTATCTTTTCCTCTCACGTGGC
F4-08F	GGTTAGGGGTAAGTAGATCTGATTCAAAATGCTGCCATGACG
F4-08R	ATACCTAATCTATCAGATCTATGCTAAGTGCATTAGCCGA
F4-09F	GGTTAGGGGTAAGTAGATCTGGCTAACCTCGCATAGCG
F4-09R	ATACCTAATCTATCAGATCTGTCCGGGTTGCTGCATAC
F4-10F	GGTTAGGGGTAAGTAGATCTGGGGCTCAAAGCTGCACATT
F4-10R	ATACCTAATCTATCAGATCTGTTCAGAGCGAGTCACTGC
F4-11F	GGTTAGGGGTAAGTAGATCTGACATGGCTTCCGGGTTAAA
F4-11R	ATACCTAATCTATCAGATCTACTTCCACCTGTGCGCAC
F4-12F	GGTTAGGGGTAAGTAGATCTGTCTGTGTTAACATTGAGCCCT
F4-12R	ATACCTAATCTATCAGATCTATACAAAAATTGCCGCCAAACC

Table 3 List of oligo used (con't).

Name	Sequence (5' to 3')
F5-01a	GGCATTTGGAGGCAGCGATC
F5-01b	GGAAACTGGAAAGGATCTCC
F5-01c	GGACAACCAGGAGCGCATCA
F5-01f	GGACCTGATTGGGCAGAAGC
F5-01e	GGAAAGGATCTCCCGGCTGG
F5-01f	GGTGGCTCAGCTCAAGCAGA
F5-02a	GGTGAGCTCGAGGGCTGAGT
F5-02b	GGTGCATGACCAGCGGTGCT
F5-02c	GGTGACGCGGTGGGGACAGT
F5-02d	GGCAGGGCTAGCATGGCTAG
F5-02e	GGATGCTAACCTCCTCAACT
F5-02f	GGAAAATGACCGAGTTGAGG
F5-03F	AGAYGGTATYGGATATGTAGAGTTT
F5-03R	AAAATCTAATCCTTATCAAATACAAATA
F5-04F	AGGGAGAGGTTTAGATTTATGAT
F5-04R	AAAAAACTCTTTTAATCCCAAAATTC

Table 4 List of HyperMD loci synthesized and assayed for autonomy in methylation recapitulation.

Locus	Genome coordinate
1	chr1: 3398602–3398902
2	chr12: 8743429–8743729
3	chr13: 153400–153700
4	chr15: 2122812–2123112
5	chr18: 174764–175064
6	chr13: 4930579–4930879

Table 5 Endogenous origins of the integrated fragments from (left) unmethylated and (right) artificially methylated library.

Note that vast majority (> 96%) of the assayed fragments were completely derived from either inside or outside DNase I hypersensitive sites (DHS).

Origin	Unmethylated	Pre-methylated
Outside DHSs	7239 (78.7%)	12791 (79.3%)
Inside DHSs	1622 (17.6%)	2775 (17.2%)
Edges of DHSs	332 (3.6%)	568 (3.5%)
Total	9193	16134

Acknowledgements

This dissertation would not have been possible without the support of multiple individuals:

- Prof. Hiroyuki Takeda, who supported and guided me throughout the study. Without the freedom that he granted me, as well as his timely and useful advice, I would not be able to accomplish such coherently and solidly designed research project. Disputing a widespread belief, especially one among scientists, can never be easy. But Prof. Takeda pointed me the path to achieve so.
- Dr. Ryohei Nakamura, who brought me to the field of epigenetics since the very first day that I arrived Takeda lab. I have lost count of how many inspiring ideas I have got from our discussions.
- Mr. (soon Dr.) Kota Abe, my interpreter, who literally saved me from all the troubles caused by language barrier in Japan.
- Special thanks to all my labmates in Takeda Lab for opening my eyes to the beautiful cities and sceneries throughout Japan.
- Brothers and sisters of the Covenant of Grace Church, who gave me the spiritual support in the last, yet most frustrating, year of my PhD work.
- Tsuji International Scholarship Foundation and its founder, Mr. Shintaro Tsuji, for both the financial support and the culture insights that I gained from the monthly gathering and interactions with other international awardees. Mr. Tsuji’s motto — “let's make friends and help each other” — will forever be engraved in my heart.
- My parents, for ultimately supporting me to go for a doctoral degree abroad.

- My beloved Grace, who waited for my return for over 3 years. I thank her so much for all the encouragements, as well as enduring my (sometimes) awful temper. I wish this dissertation can prove that the past 3 years were not in vain.
- Last but not the least, my Lord Jesus, for answering my prayers and granting me the strength and intelligence to finish this challenging project.