Supplementary Material

Large hypomethylated domain serves as strong repressive machinery for key developmental genes in vertebrates

Ryohei Nakamura[1], Tatsuya Tsukahara[1], Wei Qu[2], Kazuki Ichikawa[2], Takayoshi Otsuka[1], Katsumi Ogoshi[4], Taro L Saito[2], Kouji Matsushima[4], Sumio Sugano[3], Shinichi Hashimoto[4,5], Yutaka Suzuki[2], Shinichi Morishita[2,*], and Hiroyuki Takeda[1,*]

Figure S1 – S15

Table S10

Supplementary methods

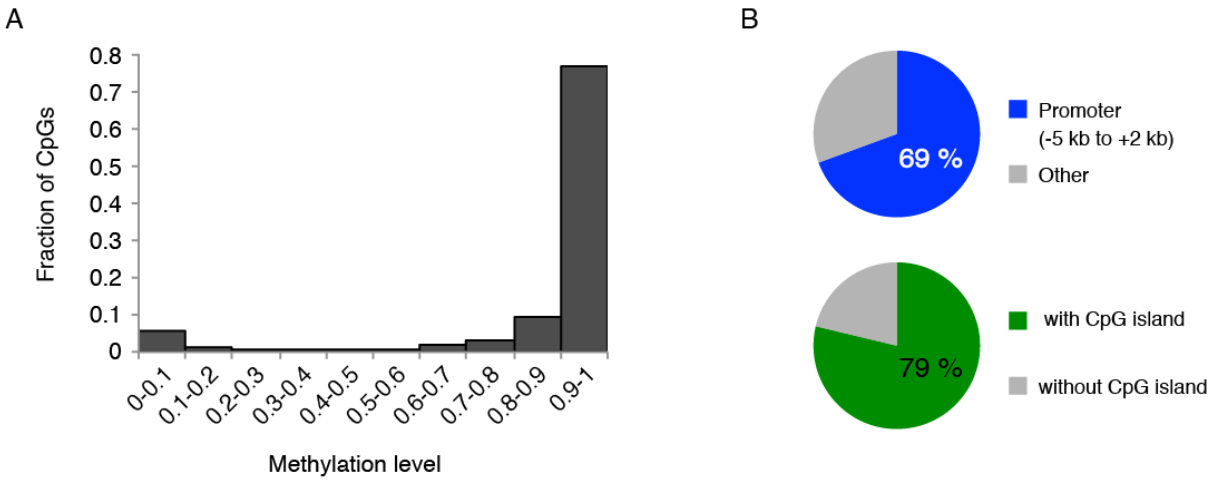

Figure S1. Characteristics of hypomethylation in medaka blastula embryos

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

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

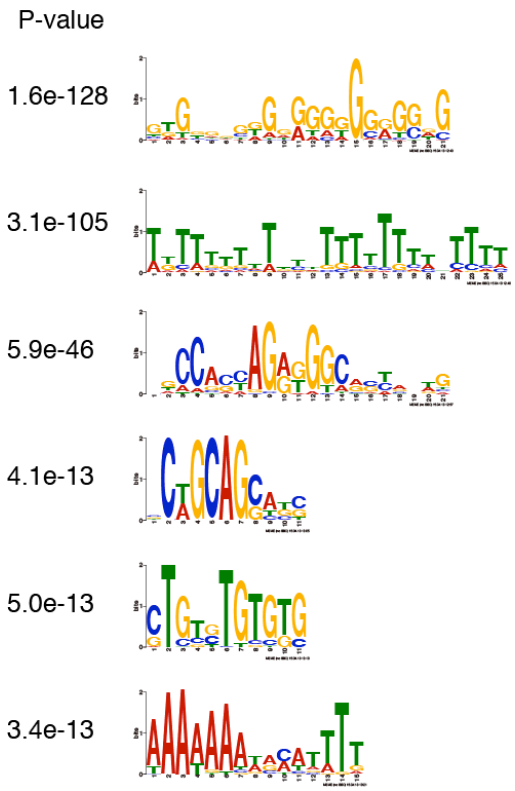

Figure S2. Characteristics of HMD boundary

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

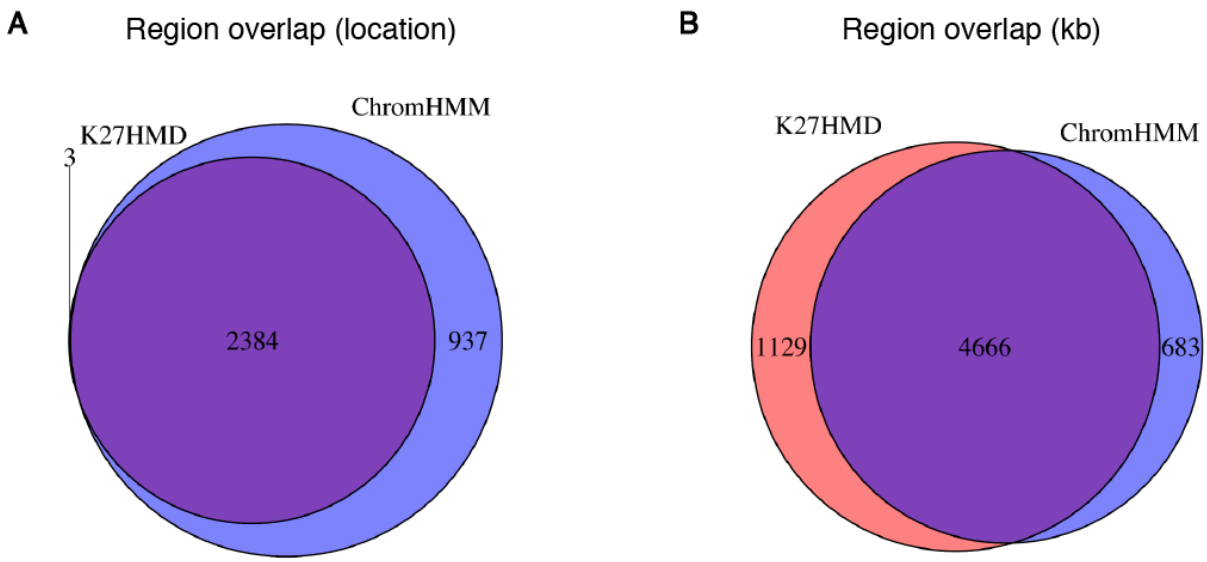

Figure S3. Comparison of K27HMD and region called by ChromHMM

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

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

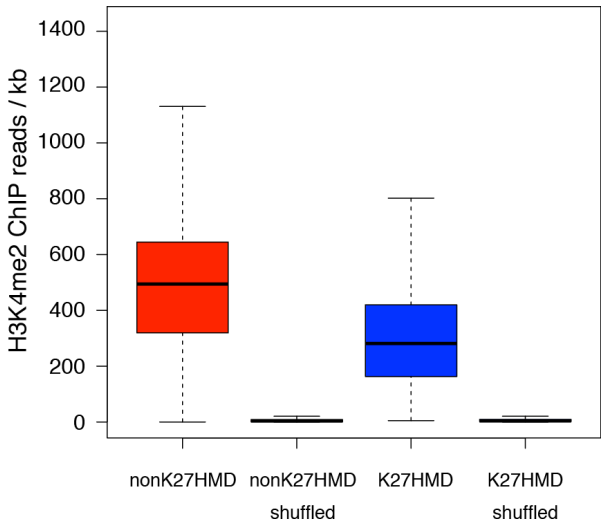

Figure S4. H3K4me2 enrichment at HMDs

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

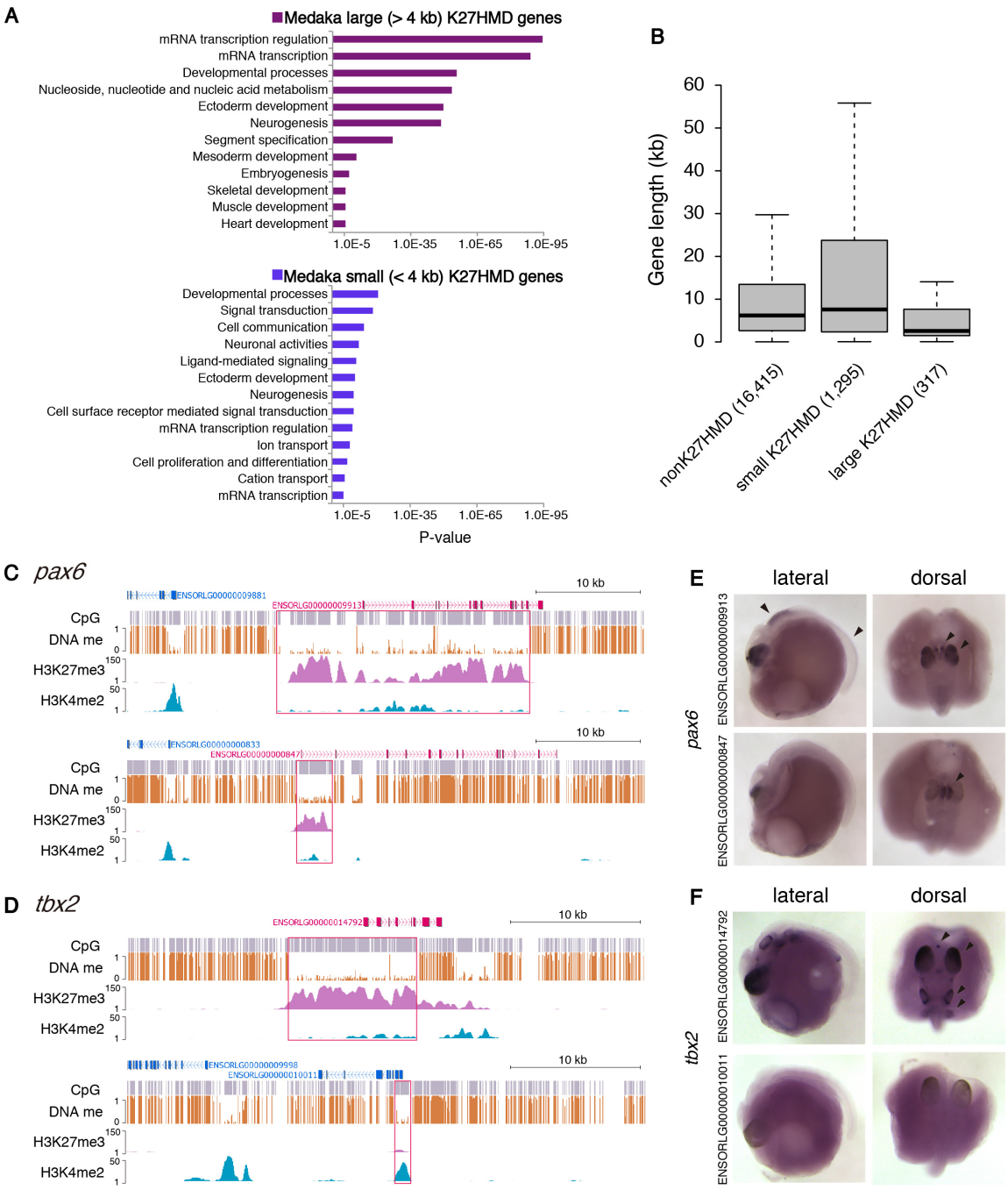

Figure S5. Large K27HMDs mark key transcription factors important for development

(A) Full lists of GO terms significantly enriched for genes marked by large (top) and small (bottom) K27HMDs. PANTHER biological process terms are shown. The x axes values (in logarithmic scale) correspond to the P-values caluclated by DAVID tool (Huang et al., 2008).

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

(C and D) Genome browser representation of DNA methylation, H3K27me3 and H3K4me2 enrichment

are shown for duplicated genes, *pax6* (B) and *tbx2* (C).

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

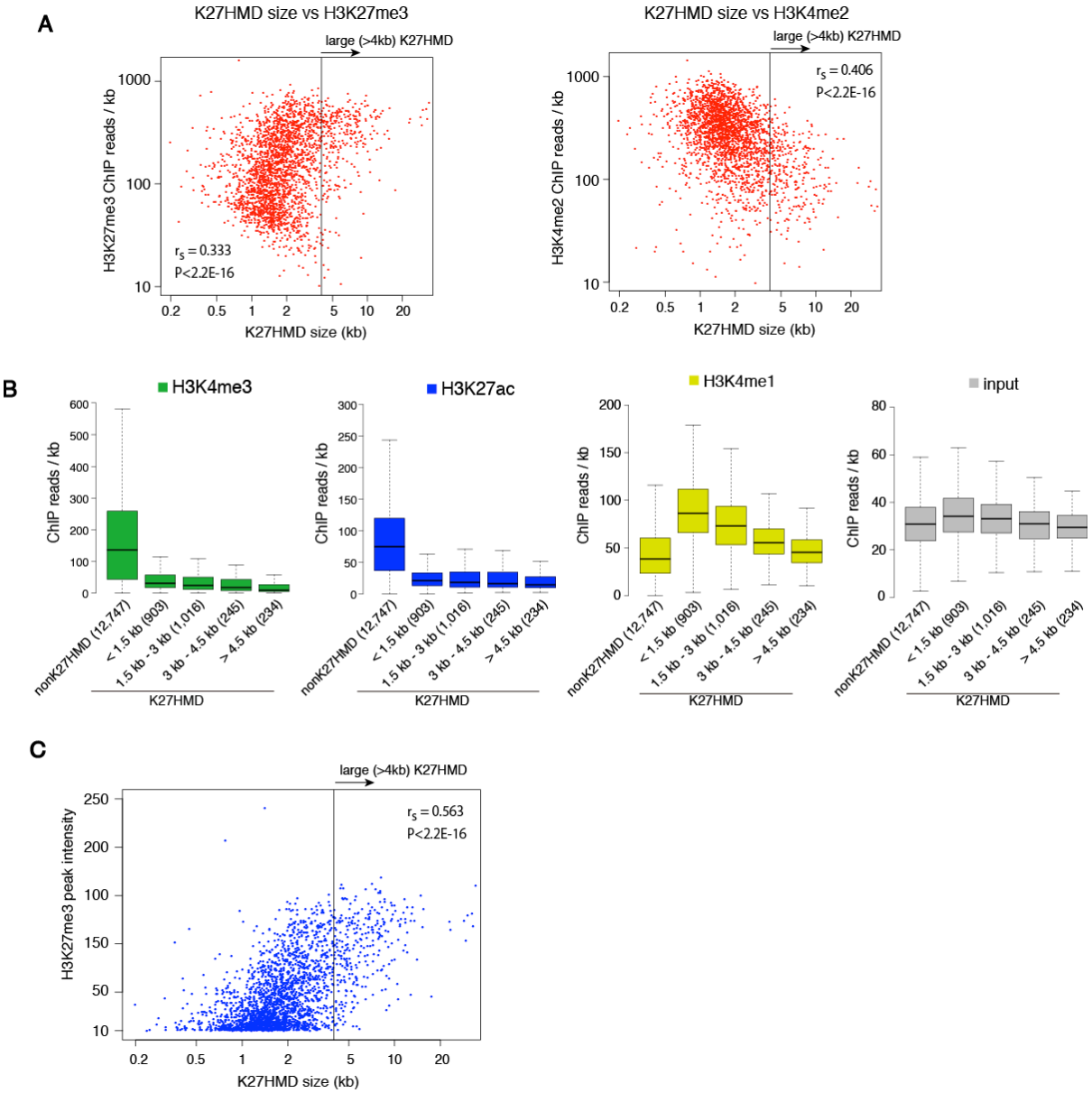

Figure S6. The size of K27HMD correlates with H3K27me3 level

(A) Comparison of the size of K27HMD with mapped ChIP reads per kb for H3K27me3 (left) and H3K4me2 (right). Spearman's rank correlation coefficient (r_s) and p value are shown.

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

(C) Comparison of the K27HMD size with H3K27me3 ChIP peak intensity. Spearman's rank correlation coefficient (r_s) and p value are shown.

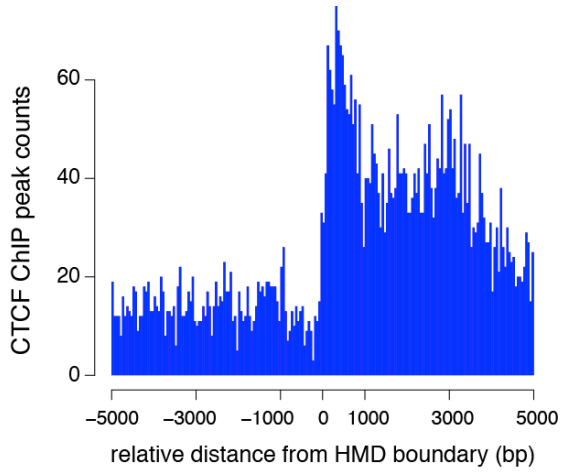

Figure S7. CTCF peak distribution at HMD in hESCs

Distribution of the CTCF ChIP-seq peaks around boundaries of HMDs larger than 3 kb in hESCs. The number of CTCF ChIP peak centers in 50 bp window were counted.

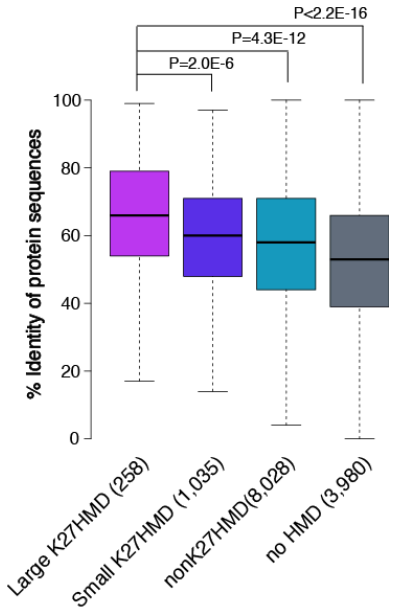

Figure S8. Large K27HMD genes are highly conserved in protein sequence

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

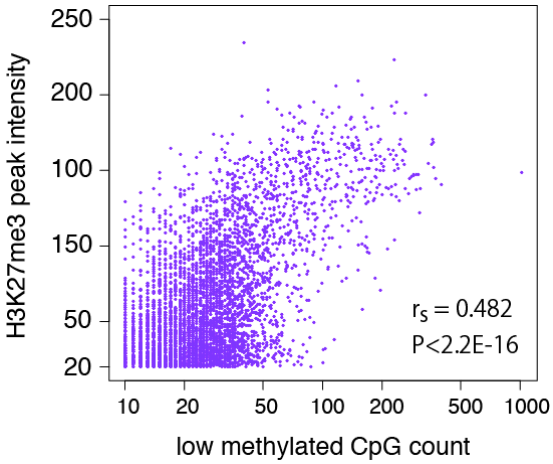

Figure S9. Correlation between low methylated CpG count and H3K27me3 level in hESCs

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

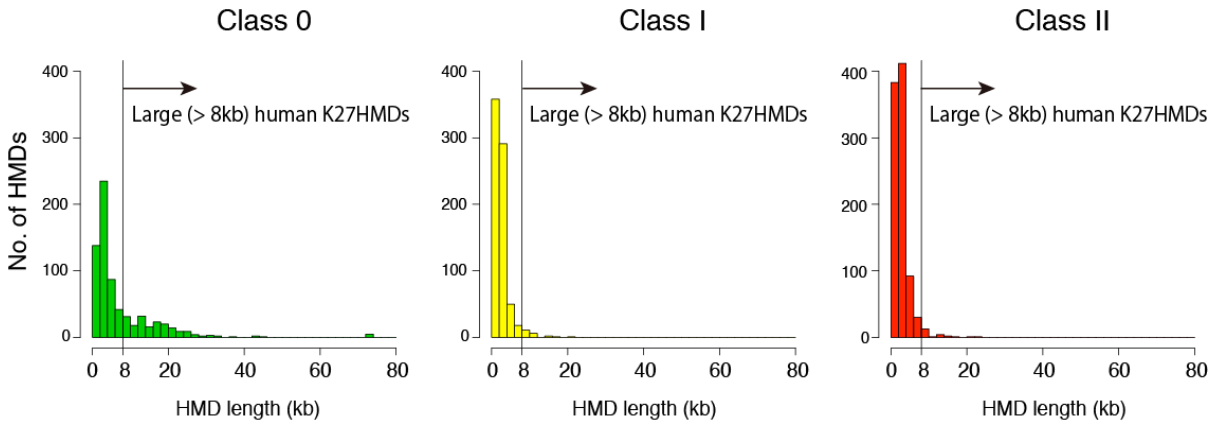

Figure S10. Large K27HMD is conserved between human and medaka

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

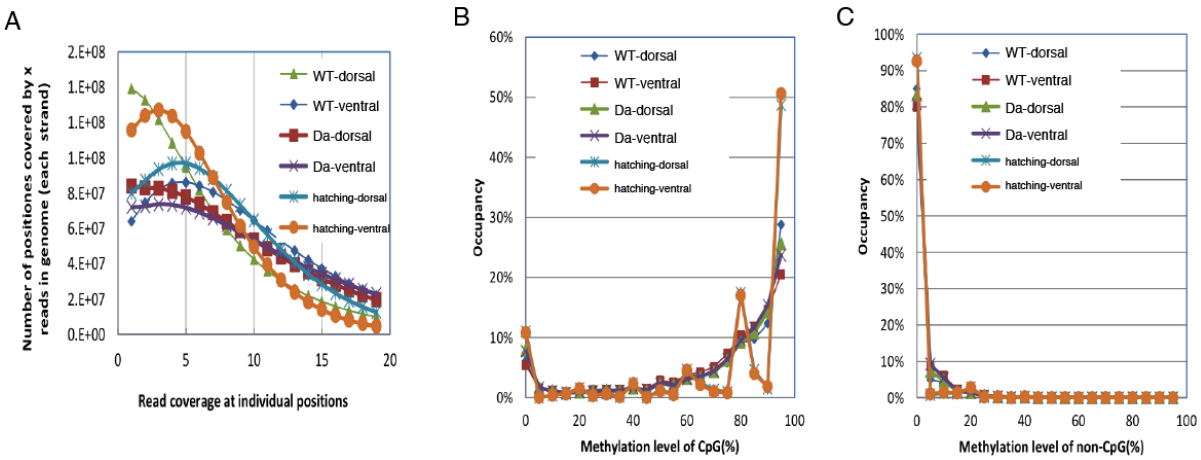

Figure S11. Summary of whole genome bisulfite sequencing data

(A) Read coverage at individual positions of bisulfite sequencing.

(B and C) Distributions of methylation rate for CpG cytosine sites (A) and Non-CpG cytosine sites (B).

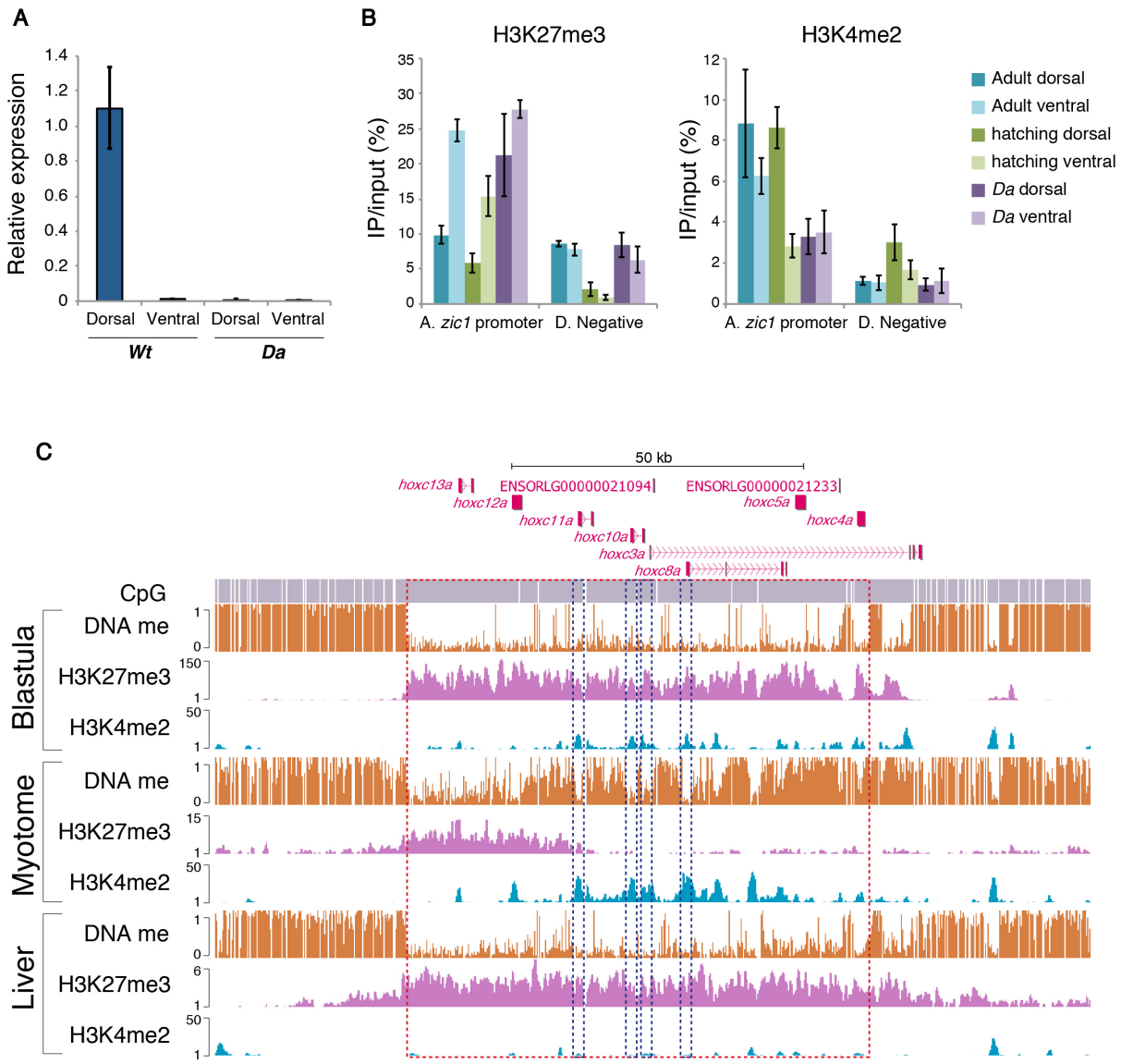

Figure S12. Active chromatin dependent K27HMD shortening in adult myotome

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

(B) Biological replicate for ChIP-qPCR in Figure 5B. Error bars represent s.d. from three technical replicates.

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

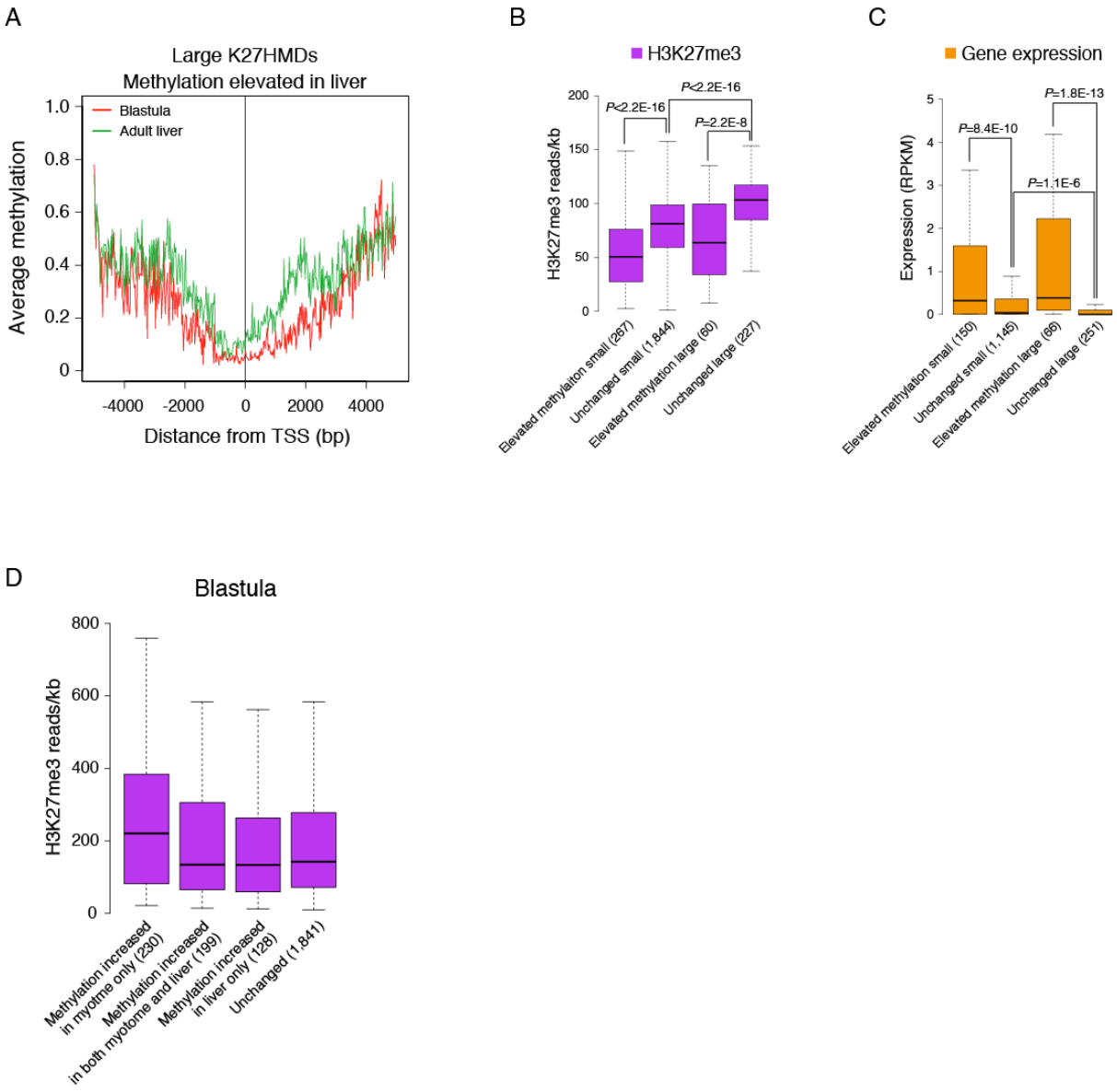

Figure S13. HMD shortening associates with sustained gene expression in adult tissues

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

(B and C) Boxplots show H3K27me3 enrichment (B) and gene expression (C) at hypermethylated and unchanged K27HMDs in adult liver. P-values were calculated using non-paired Wilcoxon tests.

(D) Boxplots show H3K27me3 enrichment for each K27HMD category in blastula embryos.

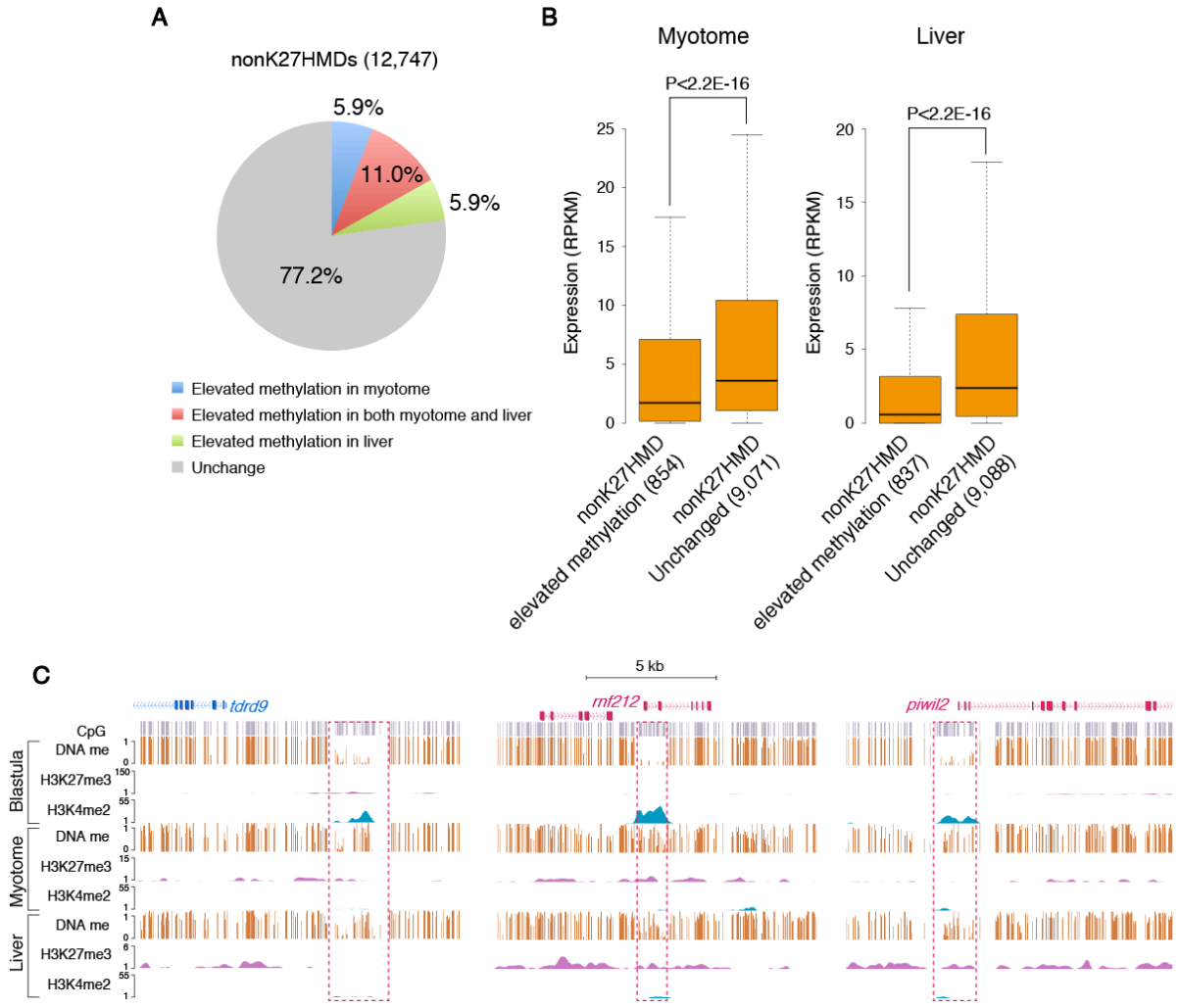

Figure S14. DNA hypermethylation of nonK27HMD associates with gene silencing

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

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

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

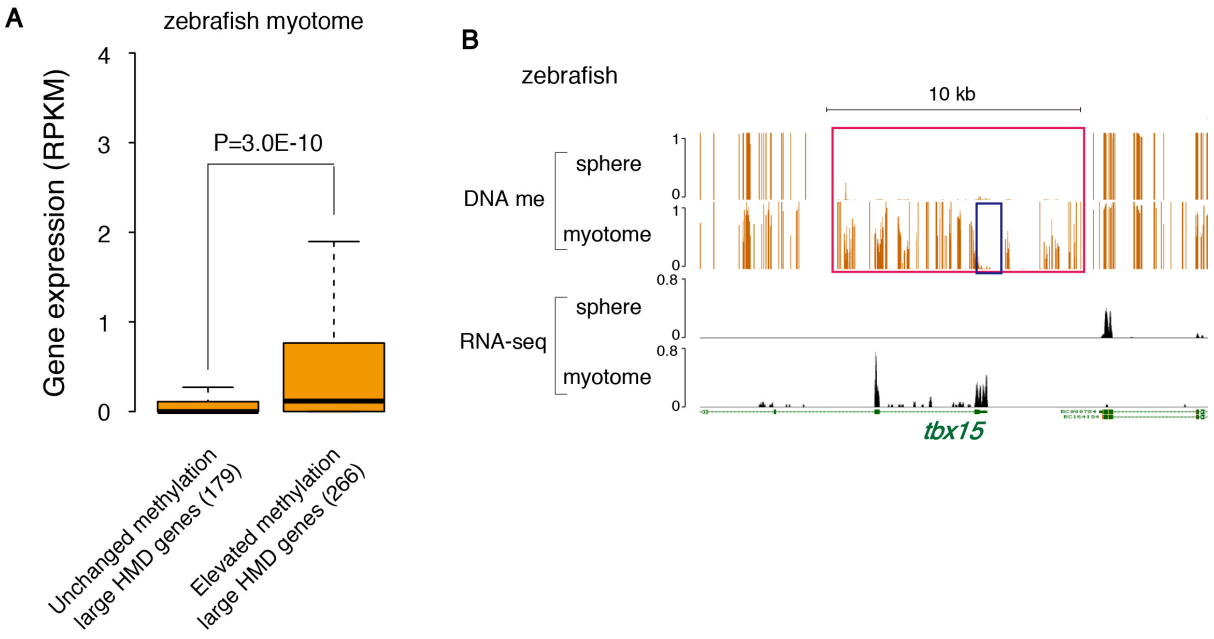

Figure S15. HMD shortening associates with active gene expression also in zebrafish

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

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

Download Table_S1_S2_S3

Download Table S4

Download Table S5

Download Table S6

Download Table S7

Download Table S8

Download Table S9

Table S10. Primers used in this study

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

Methods for data analyses

Alignment for ChIP-seq and RNA-seq reads

After removing low-quality reads (those containing five or more undetectable bases in 36mer-length reads), we mapped the remaining reads using Burrows-Wheeler Aligner mapping software (http://bio-bwa.sourceforge.net); no more than three mismatches and no gap were allowed. Only uniquely mapped reads were used for the further analysis. The medaka genome and predicted gene sequences were downloaded from Ensembl database (http://www.ensembl.org).

ChIP-seq data processing and analysis

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

RNA-seq data processing and analysis

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

Identification of HMDs

In this study, medaka genomic regions on scaffolds and ultracontigs were excluded from the analysis. We searched for contiguous regions of low methylated (methylation ratio lower than 0.4) CpG sites. ‘CpG site’ represents a single CpG in the genome. The two close low methylated regions were connected and considered as one domain when they were divided by less than four high methylated CpG sites (methylation ratio higher than 0.4). Regions with more than 9 low methylated CpG sites were defined as hyomethylated domains (HMDs). The boundaries of HMD were defined at the first low methylated CpG site inside the HMD regions. The position of TSS was defined according to the ensemble database (http://www.ensembl.org). Promoter region in this study was defined as regions from 5 kb upstream to 2 kb downstream of the TSS. The HMD overlapped with the promoter region was assigned to the gene, and used to classify genes. When more than two HMD overlapped with promoter region, the closest one to the TSS was selected. CpG island was defined as GC >= 50%, O/E >= 0.6 and length = 200 bp, according to

the previous study (Kasahara et al., 2007).

Identification of K27HMDs

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

Identification of H3K27me3 enriched hypomethylated regions by ChromHMM

We calculated K27HMD regions from methylation levels at individual CpG sites and H3K4me2/ H3K27me3 ChIP-seq data. First, for each running window of 200bp in length, we computed the average of 10(1 – x), where x was the methylation level at each CpG site, and the averages of H3K4me2/H3K27me3 ChIP-seq signal levels calculated by QuEST in the window. Subsequently, we input these epigenetic values into ChromHMM, which used the Poission distribution to model the input data, and set a real value to 0 if its p-value < 0.0001, and to 1, otherwise. After this binarization, ChromHMM decomposed the input DNA sequence into regions of six chromatin states using the Hidden Markov Model. Finally, we selected such regions that the methylation level was low while H3K27me3 ChIP-seq signal was high, and examined the overlap with K27HMD regions.

Identification of HMD with elevated DNA methylation in adult tissues

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

Heatmap generation

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

Motif analysis at HMD boundaries

For the identification of motifs around boundaries of large K27HMDs (> 4 kb), we analyzed the sequence of 200 bp (100 bp each for upstream and downstream) using MEME software. Discovered motifs were subjected to TOMTOM software and the CTCF motif was identified. Next we analyzed the genome-wide distribution of CTCF motifs using FIMO software (we set the threshold of p value 1.0E-5) and counted

based on the position from the center of HMD. All softwares used here were on MEME suite (http://meme.nbcr.net/meme/).

Analysis of nucleosome positioning around CTCF motifs

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

Gene Ontology (GO) analysis

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

Sequence conservation analysis

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

Analysis of human ESCs data

For the comparative analysis between medaka and human, ChIP-seq, and RNA-seq data sets were downloaded from previous reports (Lister et al., 2009; Lister et al., 2011) (GEO accession no.: GSE16256). Mapping to hg19 reference genome was performed using bowtie program (Langmead et al., 2009). For WGBS, we downloaded processed wig data from Lister et al., 2009, and converted to hg19 using UCSC LiftOver. Low methylated CpG was defined as < 0.6 in human. Other analyses were performed in the same way as medaka data sets. The following QuEST settings were used for H3K27me3 peak detection and visualization; H3K27me3: KDE bandwidth = 100, ChIP seeding fold enrichment = 20, ChIP extension fold enrichment = 3, ChIP-to-background fold enrichment = 3; H3K4me2: KDE bandwidth = 60.

CTCF ChIP-seq peak distribution around HMD boundaries in human

CTCF ChIP-seq peak position file was downloaded from the UCSC genome browser (The ENCODE

Project Consortium 2011; GSM733672). The relative position to the HMD boundary was counted.

Analysis of zebrafish data

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

Statistical analysis and the data visualization

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

Supplemental References

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

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

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