Evolutionary Expansion of DNA Hypomethylation in the
Mammalian Germline Genome

Jianghan Qu¹, Emily Hodges², Antoine Molaro³,⁴, Pascal Gagneux⁵, Matthew D. Dean¹, Gregory
J. Hannon³,⁶, and Andrew D. Smith *¹

¹Molecular and Computational Biology Section, Division of Biological Sciences, University of
Southern California, Los Angeles, CA 90089, USA
²Department of Biochemistry and Vanderbilt Genetics Institute, Vanderbilt University, Nashville,
TN 37232, USA
³Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
11724, USA
⁴Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
⁵Division of Comparative Pathology and Medicine, Department of Pathology, Glycobiology
Research and Training Center, University of California San Diego, La Jolla, CA 92093, USA
⁶Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA; Cancer Research UK
Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK; The New
York Genome Center, 101 Avenue of the Americas, New York, NY 10013, USA.

Running title: Mammalian sperm methylome evolution
Keywords: DNA methylation, sperm, mammalian evolution

*andrewds@usc.edu
Abstract

DNA methylation in the germline is among the most important factors influencing the evolution of mammalian genomes. Yet little is known about its evolutionary rate or the fraction of the methylome that has undergone change. We compared whole-genome, single-CpG DNA methylation profiles in sperm of seven species: human, chimpanzee, gorilla, rhesus macaque, mouse, rat and dog, to investigate epigenomic evolution. We developed a phylo-epigenetic model for DNA methylation that accommodates the correlation of states at neighboring sites and allows for inference of ancestral states. Applying this model to the sperm methylomes we uncovered an overall evolutionary expansion of the hypomethylated fraction of the genome, driven both by the birth of new hypomethylated regions and by extensive widening of hypomethylated intervals in ancestral species. This expansion shows strong lineage-specific aspects, most notably that hypomethylated intervals around transcription start sites have evolved to be considerably wider in primates and dog than in rodents, while rodents show evidence of a greater trend towards birth of new hypomethylated regions. Lineage-specific hypomethylated regions are enriched near sets of genes with common developmental functions and significant overlap across lineages. Rodent-specific and primate-specific hypomethylated regions are enriched for binding sites of similar transcription factors, suggesting that the plasticity accommodated by certain regulatory factors is conserved, despite substantial change in the specific sites of regulation. Overall our results reveal substantial global epigenomic change in mammalian sperm methylomes and point to a divergence in trans-epigenetic mechanisms that govern the organization of epigenetic states at gene promoters.

INTRODUCTION

Mammalian DNA methylation is an epigenetic modification that occurs primarily on cytosines in the context of CpG dinucleotide and has wide-ranging connections to mammalian development (Reik et al. 2001; Smith & Meissner 2013). A typical mammalian genome contains 20-30 million CpG sites with average methylation level between 70-80% in most cell types (Lister et al. 2013; Kundaje et al. 2015). Hypomethylated regions punctuate this globally high methylation profile, and often coincide with gene regulatory elements that comprise gene promoters, enhancers and insulators (Jones 2012). Programmed changes to these methylome features, for example altering the methylation state through a promoter, correlate with gene expression changes during development and differentiation (Meissner et al. 2008; Hodges et al. 2011). These dynamics can exhibit high precision, for example when boundaries of hypomethylated regions shift to uncover CpG
island “shores” and allow additional regulatory elements to receive signals (Irizarry et al. 2009).

DNA methylation has substantially impacted the large-scale evolution of mammalian genomes. Spontaneous deamination elevates the mutation rate at methylated cytosines (Bird 1980). As a consequence, germline DNA methylation profiles have shaped the CpG landscape of mammalian genomes (Cohen et al. 2011), resulting in the CpG island phenomenon (Gardiner-Garden & Frommer 1987). In germ cells, DNA methylation is necessary for suppressing germline retro-element proliferation (Walsh et al. 1998; Bourc’his & Bestor 2004). Mounting evidence supports a contribution of germ cell DNA methylation to the epigenomic state of embryonic cells and the early activation of developmental genes (Brykczynska et al. 2010; Carrell & Hammoud 2010; Smith et al. 2012; Branco et al. 2016; Atsem et al. 2016). Examination of germline DNA methylation is therefore of particular interest in understanding how the requirements of epigenomic regulatory programming are balanced against the increased mutational burden of DNA methylation.

Taking a broader view, patterns of DNA methylation and the way they vary between species afford us a glimpse into the evolution of gene regulation. As highlighted by the seminal work of King & Wilson (1975) and validated in the current genomics area, species with high genetic similarity in their protein coding regions can show substantial phenotypic differences which may be related to differences in gene expression and regulation (Karaman et al. 2003; Odom et al. 2007; Bauernfeind et al. 2015). Epigenetic divergence reflects altered protein-DNA interactions, which may associate with changes in both the sequence of cis-regulatory elements (Heijmans et al. 2007; Lienert et al. 2011) and the binding preference of trans-acting factors (Cheng et al. 2014). Thus, characterizing epigenetic divergence is a first step towards understanding the evolution of regulatory systems in different species.

Existing comparative analyses of mammalian methylomes have increased our understanding of DNA methylation divergence. Methylation states across multiple primate species recapitulate their phylogenetic relationship (Martin et al. 2011), and greater variation has been observed between tissues than between species (Pai et al. 2011; Molaro et al. 2011). Regions with species-specific DNA methylation have been reported to harbor excessive species-specific substitutions (Hernando-Herraez et al. 2015). However, such studies have been limited to species most closely related to human, and have focused on species-specific differentially methylated regions, instead of systematically studying methylation changes in all lineages of
the phylogenetic tree without bias.

It remains unclear how and why DNA methylation patterns have evolved in mammalian species: where in the genome, and when during evolutionary history have changes occurred? Addressing these questions involves modeling an evolutionary process that may be viewed as analogous to sequence evolution. In homogeneous healthy cell populations, and over nearly the entire genome, continuous methylation levels at CpG sites can be functionally understood as having either high or low methylation, and a methylome can be modeled as a sequence of binary states. Phylogenetic modeling for molecular sequence evolution has a rich history (Felsenstein 1981; Yang 1994; Siepel & Haussler 2004), and approaches have been applied to DNA methylation changes in tumor growth and development (Siegmund et al. 2009; Capra & Kostka 2014). However, modeling methylome evolution at individual CpG sites presents a unique challenge: the majority of those sites have mutated between distant species due to their hypermutability. Focusing on CpG sites that are fully conserved in the relevant species dramatically reduces the portion of the orthologous genome that can be analyzed. Solutions that aggregate data, for example in pre-determined windows along the genome, lower resolution while only partially solving the problem. At the same time, it is well known that methylation states of neighboring CpG sites are highly interdependent and a phenotypically relevant epigenomic change tends to involve multiple consecutive CpG sites. Horizontal dependence provides information that may be leveraged to overcome this challenge.

In this study, we examined whole genome bisulfite sequencing data from the sperm genomes of 7 mammalian species – human, chimpanzee, gorilla, rhesus macaque, mouse, rat and dog – to investigate the evolution of germline DNA methylation. We designed a phylo-epigenetic model that accounts for loss of CpG sites between species by incorporating the interdependence of methylation states at neighboring CpG sites to infer the history of methylation divergence on each branch of the phylogenetic tree. We were able to detect genome-wide trends and lineage-specific features of methylome evolution, identify evolutionary events at a high resolution, and investigate the relationship with divergence of genomic DNA, the association with other epigenetic modifications and the functional relevance of recently evolved methylome features.
RESULTS

Comparison of sperm methylomes reveals lineage-specific features

We generated whole-genome bisulfite sequencing (WGBS) data for sperm cells from gorilla, dog and rat. Sperm methylomes of human, chimpanzee, rhesus macaque and mouse were produced in previous WGBS studies (Molaro et al. 2011; Lu et al. 2015; Hammoud et al. 2014). For each species, the sequencing had a total depth above 10\times, and covered the entire mappable fraction of the corresponding reference genome (Supplemental Table S1). In parts of our study, we used methylomes of embryonic stem cells (ESC) and somatic cells (B-cell, whole blood, peripheral blood mononuclear cells and left ventricle) from different species for comparison with sperm methylomes (Methods). We used Ensembl gene annotations (Release 75) (Flicek et al. 2014) and defined gene promoter as the ±1 kbp region surrounding the transcription start site (TSS). We identified HMRs from individual methylomes using a hidden Markov model as previously described (Molaro et al. 2011).

All species exhibited global methylation in sperm, with genome-wide average DNA methylation levels ranging between 0.68 and 0.79 (Supplemental Table S2). The methylation levels at single CpG sites formed a bimodal distribution (Supplemental Fig. S1A). Sperm methylomes contained more and wider HMRs than those in the examined ESC and somatic methylomes (Fig. 1A, Supplemental Table S2). Sperm HMRs overlapped retrotransposons and pericentromeric satellites more frequently than ESC and somatic HMRs (Supplemental Table S3, Supplemental Fig. S1B). These observations are consistent with patterns observed in human, chimpanzee and mouse (Molaro et al. 2011, 2014).

To study epigenomic evolution across species, we aligned non-human methylomes and HMRs to the human reference genome at single CpG resolution (Methods). The genomic composition of the 7-way orthologous genome is shown in Supplemental Fig. S1C. The orthologous genome contains, on average, 17% (4.3 million) of the total CpG sites in individual species. The majority of the CpG sites have diverged between species, and only 0.4 million CpG sites are conserved across all species, the majority (59.7%) of which reside in codons. We therefore do not focus on fully conserved CpG sites, as they are limited in number and may be confounded by evolutionary pressures associated with protein coding regions. The number of HMRs located in the orthologous genome ranges between 24k and 40k, and the average size
ranges between 1.1 kbp and 1.5 kbp for individual species (Supplemental Table S4, Fig. S1D).

For a global view of the relationships between sperm methylomes, we clustered the sperm methylomes along with ESC and somatic methylomes based on pair-wise correlation of average methylation levels in 200-bp bins (Fig. 1B, Supplemental Fig. S1E). There was a clear separation of sperm from ESC and somatic methylomes, consistent with tissue-specific characteristics of methylation (Pai et al. 2011; Molaro et al. 2011). The hierarchy in the sperm methylome cluster was largely consistent with the known phylogeny, except that dog and primates were more similar to each other than either was to rodents, despite a more recent common ancestor for rodents and primates (dos Reis et al. 2012). This may not be unexpected. Comparative genomic studies have reported a three-fold higher neutral mutation rate in rodents compared with primates (Gibbs et al. 2004; Lindblad-Toh et al. 2011).
Figure 1

A

Mean HMR size (kbp)

Number of HMRs (x1000)

B

Correlation (7-way orthologous 200bp bins)

Number of HMRs

C

conserved in all species

non-promoter

promoter

D

5 kb

Promoter HMR size (bp)

E

25% 75%

p < 2.2x10^-16

Figure 1: (Caption on next page)
Figure 1: Mammalian sperm methylome characteristics. (A) The number and average size of HMRs in native assemblies. (B) Hierarchical clustering of aligned 7-way orthologous methylomes of multiple species and cell types. (C) The number of promoter HMRs and non-promoter HMRs in 7-way orthologous sperm methylomes. Dashed lines indicate the average number of conserved HMRs across species. (D) Sperm DNA methylation of seven species in an example orthologous region. Methylome alignment is shown along with conservation tracks from Multiz alignment of 100 vertebrates, and human repeat elements by RepeatMasker. Regions in solid boxes show divergent methylation states at well-conserved elements. See zoomed-in browser image for dashed boxes in Supplemental Fig. S2A. (E) The median size of hypomethylated regions up and down-stream of TSS in somatic (orange) and sperm (blue) methylomes of different species. HMR sizes are measured in respective native genome. Whiskers mark 25th and 75th percentiles of HMR sizes up or down stream of TSS. Wilcoxon rank-sum tests for all pairs of species between rodent species and non-rodent species showed significantly narrower HMRs in rodents.

**Methylation features are conserved at promoters and divergent distally**

All species have comparable numbers of promoter HMRs, while rodents have many more non-promoter HMRs than other species (Fig. 1C). The intersection of HMRs across all species contained a total of 15k regions, with average 1.0 kbp in size. If a HMR contains a that region is hypomethylated in all species, we call it a conserved HMR. Conserved HMRs were located predominantly at gene promoters, and constitute on average 74% of the hypomethylated promoters in each species (Fig. 1C). Non-promoter HMRs were far less conserved across species, with conserved HMRs constituting 8-22% of the non-promoter HMRs in each species. The example interval in Fig. 1D (Supplemental Fig. S2B) shows the conserved presence of an HMR at the promoter in each species. The gene within this interval, EBF2 in human, is expressed in a variety of somatic cells, but has low expression in testis (Uhlén et al. 2015). This promoter HMR size varied between species, and there were lineage-specific HMR gain and loss events more distal to TSS. These observations are consistent with slower change at promoters, and more rapid evolution of enhancers, as has been observed in a comparative study of histone modifications focused on mammalian liver (Villar et al. 2015).

Taking advantage of the precision afforded by WGBS, we compared the sizes of promoter-associated HMRs across species. HMRs at orthologous gene promoters had similar sizes in ESC and somatic cells.
across species. In sperm, by contrast, HMRs showed substantial variation (Fig. 1E, Supplemental Fig. S1F, G). Rodent sperm promoter HMRs are significantly shorter compared with primates and dog (Wilcoxon rank-sum test $p < 2.2 \times 10^{-16}$). Within primates, rhesus macaque had significantly shorter HMRs than other species examined (Wilcoxon rank-sum test $p < 2.2 \times 10^{-16}$). Similar levels of variation were evident at replicate level between species as well (Supplemental Table S2). These sperm promoter HMR size differences indicate a global divergence in the organization of the sperm epigenome.

The cross-species HMR size differences in orthologous regions indicate that an ancestral HMR has evolved to have different boundary locations in different species. For a conserved HMR in human, on average, the core region of hypomethylation shared in all species only constituted 53% of the HMR. This proportion was 64% for rodents. Boundary variation was also common for HMRs specific to a lineage. Altogether, a considerable amount of cross-species epigenomic variation took the form of adjustments in boundaries of these putative regulatory intervals – in addition to complete gain or loss. Below we implement a formal model to quantify such boundary movements.

Despite prevalent HMR boundary variation, a subset of orthologous HMRs had relatively conserved boundary locations across species. We identified 250 such “ultra-conserved” HMRs (Supplemental Methods), all of which overlapped phastCons conserved elements for placental mammals (Siepel et al. 2005), and 92% overlapped gene transcription start sites (Supplemental Table S5). Relative to genes with promoter HMRs in all seven species, those with ultra-conserved promoter HMRs were enriched in multiple biological processes including developmental process, signaling and regulation of multicellular organismal process (Supplemental Table S6).

Methylation loss exceeds methylation gain across species

We compared sperm HMRs across the seven species and identified species-specific HMRs and species-specific methylated regions. For each species, there was more species-specific hypomethylation than methylation (Supplemental Fig. S3A, Table S7). This hints at a non-stationary process, and prompted us to ask if global trends and rates of epigenomic change can be extracted from these methylomes. We designed a model that combines two processes: (1) the “horizontal” process that induces autocorrelation of methylation states along the genome within a species, and (2) the “vertical” inheritance of
methylation states from the ancestral methylome. The vertical process is modeled with a continuous-time Markov process in the usual way (Felsenstein 1981). To model the horizontal dependency between the methylation states of two neighboring sites in the same methylome, we introduce a transition probability to model the conditional distribution of a site’s methylation state given the state at the site’s neighbors (Supplemental Fig. S3B). Combining these two processes, our model is a Bayesian network that describes the probability of a site having a particular methylation state as a function of its ancestral state and the contemporaneous state of its previous neighbor (Methods). The inclusion of the horizontal process allows us to treat non-conserved CpG sites as sites with latent data to be modeled using a computationally tractable set of dependencies (Supplemental Fig. S3B).

We estimated the sperm methylome evolutionary tree with the proposed model (Fig. 2A, Supplemental Table S8). Branches within rodents were significantly longer than the primate lineage (Wilcoxon signed-rank test $p = 6.6 \times 10^{-9}$, Supplemental Methods), indicating faster epigenomic evolution in sperm of rodents than primates. The rat sperm methylome exhibited faster evolution than mouse (Wilcoxon signed-rank test $p = 1.6 \times 10^{-8}$), which coincided with a higher nucleotide substitution rate in rat (Gibbs et al. 2004). However, the methylome evolution rates were much higher at the terminal branches than internal branches, suggesting relatively fast turnover of sperm HMRs in mammals. With an exponential decay model, we estimated the half-life of HMRs to be 117Myr overall, 439Myr at gene promoters and 39Myr for distal HMRs (Supplemental Fig. S3C).

The inferred methylation states in ancestral species indicate an increasing proportion of hypomethylation in the orthologous genome along each branch of the phylogenetic tree (Fig. 2C). We estimated that HMRs made up 7.0% of the orthologous genome in the last common ancestor of the seven species; 7.6% in the primate common ancestor and 8.1% in the rodent common ancestor; 8.3-9.2% in extant primates and 9.9-10.2 in extant rodents (Fig. 2C). The overall increase in the hypomethylated fraction of the orthologous genome was a result of DNA methylation loss exceeding gain along all branches of the tree, with the dog coming closest to balancing the two (Supplemental Fig. S3D). To examine whether the methylome evolutionary process was stationary, we applied independent-site phylogenetic models on discretized methylation states in 200-bp bins (Methods). We compared a reversible model assuming stationary evolutionary process with an unrestricted model (i.e. lacking the reversibility assumption). The unrestricted model fit the observed
data better, and estimated an evolutionary process with increasing hypomethylation fraction (Supplemental Fig. S3E, F).
Figure 2

A. Reference phylogeny

B. Ancestral primate and rodent segmentation and branch assignment

C. Phylogenetic tree (genome)

D. Phylogenetic tree (methylome)

E. Phylogenetic tree (methylome) with evolution events

Figure 2: (Caption on next page)
Figure 2: Methylation loss exceeds methylation gain during evolution. (A) Phylogenetic tree representing consensus divergence time (Hedges et al. 2015), genome evolution and sperm methylome evolution. Unit branch length represents 1 million year, 1 substitution/site and 1 methylation state change/CpG site, respectively. (B) Schematic segmentation of the orthologous genome according to the history of methylation evolution and annotation of methylation mutation events. (C) Fraction of orthologous genome hypomethylated in extant and ancestral species, estimated by interdependent-site phylo-epigenetic model. (D) Total size of different types of methylation evolution events on individual branches. (E) The distribution of distances from methylome evolution events to closest TSS in the 7-way orthologous genome. Same color legend as (D).

To examine whether our observations above are robust to the technical aspect of using human-referenced alignment, we performed the same analyses using mouse-referenced alignment (Supplemental Methods). In addition, to investigate whether sperm methylomes have similar evolutionary features at well-conserved sequence elements as in the entire orthologous genome, we restricted analyses to CpG sites located within placental mammal conserved elements (Siepel et al. (2005), Supplemental Methods). The genomic context composition and distribution of CpG methylation levels vary slightly for mouse-referenced 7-way orthologous genome (Supplemental Fig. S4A, B). However, they vary substantially for well-conserved elements that are mostly protein coding sequences and important regulatory elements under strict evolutionary constraints (Supplemental Fig. S5A, B). Nevertheless, the results in both analyses are consistent with results presented above. Rodents display faster evolution rates (Supplemental Fig. S4C, E, S5C, E). Sperm methylomes have experienced expansion of hypomethylation in the overall orthologous genome (Supplemental Fig. S4D, F) and in well-conserved elements (Supplemental Fig. S5D, F). Methylation patterns are more conserved at promoter-proximal elements than distal elements (Supplemental Fig. S5G).

**Widening of hypomethylated regions has contributed significantly to methylation loss**

Loss of DNA methylation during evolution reflects an increase in either the size or the number of HMRs. We classified methylation loss (HMR gain) events into HMR birth events or HMR extension events, and similarly classified HMR loss events into HMR deaths or HMR contractions (Fig. 2B, D, Methods). The HMR birth and death events showed different genomic distribution from the extensions and contractions
Distal (more than 10 kbp away from TSS) events constitute 56% of HMR births and 40% of deaths, in contrast to 30% of extensions and 16% of contractions.

Only a small proportion (less than 10%) of the orthologous genome was hypomethylated. However, HMR extensions comprise 21-74% of total HMR gains (Fig. 2D, Supplemental Table S9), which correspond to over 20-fold more than expected if HMR gains were to randomly occur in the genome (Supplemental Fig. S6, Methods). Among HMR loss events, deaths occurred more frequently than expected (Supplemental Fig. S6). These observations suggest that existing HMRs may facilitate the evolutionary formation of a new HMR in the adjacent region, meanwhile loss of HMR is more likely to occur as methylation through an entire ancestral HMR rather than only a fraction of it.

We found two forms of evidence in line with our hypothesis that once established, a typical HMR can gradually evolve to span a larger genomic region. First, we compared the HMR sizes between older and younger groups of, and found positive correlation between the age and the size of HMRs (Fig. 3A), with faster expansion at promoter HMRs than non-promoter HMRs. Second, we found that HMRs in the orthologous genome were consistently larger in size within each genomic context group than those in species-specific genomic regions (Fig. 3B). An alternative hypothesis states that HMRs may randomly arise in the genome with substantial variation in size, but wider HMRs are more likely to survive through evolution. Although this alternative hypothesis may also explain the size-age correlation of HMRs, it offers no explanation to the substantial cross-species size variation of conserved HMRs. Therefore, we find progressive extension a more likely hypothesis. Together, these observations suggest that recently appearing epigenomic features encompass smaller genomic intervals than more ancient features. Of note, such a global trend would have progressed independently and may have different rates along parallel evolutionary lineages.
Figure 3: Older and more conserved sperm HMRs are wider. (A) Mean sizes of promoter and non-promoter HMRs in the 7-way orthologous genome of human and mouse sperm, grouped by estimated HMR emergence time in mammalian evolution. Error bars mark standard errors of the mean. Significant Wilcoxon rank-sum test $p$-values are shown. The “Boreoeutheria” group contains HMRs conserved in all seven species, representing HMRs formed before and during the speciation of Boreoeutheria. Other groups represent HMRs formed during the speciation of respective lineages. (B) Distribution of HMR sizes in native genome assemblies, grouped by genomic context and conservation level.
Sequence signatures driven by sperm methylome divergence

Because methylated CpGs experience elevated mutation rates driven by deamination, we predicted that observed/expected (o/e) ratio of CpG frequency in a region is related to the length of time a region has been methylated. To test this prediction, we measured CpG frequency in human HMRs separated into 6 different age groups based on estimated ancestral methylation states. As predicted, older HMRs had higher CpG o/e ratios compared to younger HMRs (Fig. 4A). In addition, older HMRs displayed lower levels of DNA methylation than younger HMRs, suggesting that conserved elements are subject to more stringent epigenetic regulation and maintenance while younger elements may be involved in more dynamic interactions. These findings remained unchanged after we controlled for DNA methylation level (Supplemental Fig. S7A).

The relationship between germline methylation history and CpG decay through a genomic interval suggests we might use CpG o/e ratios to polarize the promoter HMR size divergence observed in Fig. 1D. The substantial number of promoter HMRs that are much wider in human than in mouse may have arisen from a general widening along the human lineage or a narrowing along the mouse lineage. Fig. 4B illustrates a method for distinguishing these possibilities. We selected gene promoters that were hypomethylated in sperm and ESC of both human and mouse, with a wider sperm HMR than ESC HMR in both species, and with a wider sperm HMR in human than in mouse. The mouse CpG o/e ratios on opposite sides of positions orthologous to human sperm HMR boundaries (sections i vs ii) showed no significant difference (Fig. 4C), suggesting that both sides in the mouse genome shared a common methylation history during evolution. Meanwhile, regions corresponding to rodent-lineage-specific sperm HMR loss showed higher CpG o/e ratios than the methylated genomic background level (Wilcoxon rank-sum test $p = 1.3 \times 10^{-4}$, Supplemental Fig. S7B), which suggests that ancestral CpGs have not had enough evolutionary time to decay following recent increases in methylation. In contrast, there was a significant drop of human CpG o/e ratios (Wilcoxon rank-sum test $p = 1.3 \times 10^{-31}$) crossing the mouse sperm HMR boundaries from section iii to ii. Although regions on both sides were hypomethylated in present-day human sperm, the outer side (section ii) must have become hypomethylated later than the inner side (section iii). Dog sperm methylome also showed wider promoter HMRs than mouse (Fig. 1D), and we observed similar profiles of CpG o/e ratios in the dog-mouse comparison (Supplemental Fig. S7C). These observations suggest that although sperm promoter
HMR sizes may have increased on parallel lineages, the changes likely occurred at different rates, and that ancestral promoter HMR size in sperm was closer to that in rodents, with both dog and primates evolving wider intervals of hypomethylation around promoter in sperm.

Figure 4: Sequence signatures driven by sperm methylome evolution. (A) CpG enrichment (observed/expected ratio of CpG occurrences) and regional average DNA methylation level in human HMRs in 7-way orthologous genome grouped by estimated hypomethylation age. (B) Two opposite scenarios explaining the evolution path to different promoter HMR sizes in two extant species, and expected CpG enrichment schematic profile in the extant species as a result of methylation-induced CpG-decay. (C) Human and mouse CpG enrichment profile at hypomethylated orthologous gene promoters where human sperm HMRs are wider than mouse. Profiles show CpG enrichment in 2-bp windows by distance to indicated HMR boundaries. Smoothed lines were generated with local polynomial regression fitting (LOESS).
Methylation divergence is associated with histone modification and sequence divergence

To examine whether sperm DNA methylation divergence is accompanied by histone modification divergence, we analyzed H3K4me3 and H3K27me3 ChIP-seq data from round spermatids of human and mouse (Lesch et al. 2016), and H3K4me1 ChIP-seq data from human and mouse sperm (Hammoud et al. 2014; Jung et al. 2017). Lineage-specific sperm HMR births and extensions in human and mouse were associated with lineage-specific enrichment of histone modifications (Fig. 5A, Supplemental Fig. S8A, B). In both human and mouse, promoter HMRs showing species-specific widening in sperm, especially those showing the same widening in ESC, were enriched with bivalent histone marks in round spermatid (Supplemental Table S10). The association between divergent DNA methylation and divergent histone modification in sperm indicates that sperm DNA methylation is an integral part of the germline epigenome evolution.

The sequence conservation levels (Pollard et al. 2010) were substantially stronger in lineage-specific HMRs than their flanking regions, suggesting that regulatory elements may be conserved at the sequence level, but diverged at the epigenetic level. Moreover, all groups of lineage-specific sperm HMRs showed a sizable (over 30%) overlap with human Roadmap somatic HMRs, which are potentially active regulatory elements in human somatic tissues (Kundaje et al. 2015) (Supplemental Table S11). Overlapping with human somatic HMRs considerably increased the sequence conservation level in lineage-specific sperm HMRs (Supplemental Fig. S7D). Rodent lineage-specific HMRs not overlapping human somatic HMRs remained at higher sequence conservation level than neighboring regions (Supplemental Fig. S7D), suggesting conserved regulatory regions between human and mouse yet to be observed. These observations are consistent with exaptation of existing elements as a major source for species to evolve new regulatory interactions (Villar et al. 2015).

To examine whether lineage-specific HMRs are associated with lineage-specific substitutions, we computed the relative sequence substitution rate (RSSR) in orthologous regions between two parallel lineages, and compared the RSSR in lineage-specific HMRs against randomly sampled regions from the background orthologous genome (Methods). We studied four pairs of sister or parallel lineages, and observed that lineage-specific HMR births were associated with significantly elevated sequence substitution rate on the same lineage (Fig. 5B). After excluding CpG sites, the differences in relative substitution rates remained significant (Supplemental Table S12). These observations, further extending previous results, reveal a local
link between sequence substitutions and evolution of new epigenetic features.

Figure 5

A  

Human round spermatid

Mouse round spermatid

H3K27me3 signal
(H3K27me3 ratio vs input)

H3K4me3 signal
(H3K4me3 ratio vs input)

B  

HMR births specific to
denominator lineage
numerator lineage

RSSR distribution
by random sampling
90% CI of mean RSSR

HMR in
lineage-specific HMRs

Density
Relative sequence substitution rate
#human subs.  #chimp subs.

Density
Relative sequence substitution rate
#gorilla subs.  #rhesus subs.

Density
Relative sequence substitution rate
#murinae subs.  #catarrhini subs.

p = 9.32x10^-5
p = 8.70x10^-6
p = 2.31x10^-8

p = 7.33x10^-10
p < 2.22x10^-16
p < 2.22x10^-16

Figure 5: (Caption on next page)
Figure 5: Sperm DNA methylation divergence is accompanied by sperm histone modification and DNA sequence divergence. (A) Mouse and human-specific sperm HMRs and HMR extensions show species specific H3K4me3 and H3K27me3 enrichment. (B) Lineage-specific HMRs are associated with relatively more sequence substitutions. For each pair of parallel lineages, expected distribution of RSSR is shown as a gray histogram, and fit with normal distribution (black curve). Yellow area marks 90% confidence interval for mean RSSR. Upper (lower) tail of the distribution indicates significantly more substitutions in the numerator (denominator) lineage. Each lollipop marks the observed RSSR in HMRs specific to the indicated lineage, with one-sided $p$-value shown on top.

**Lineage-specific changes involve common transcription factors and developmental genes**

DNA hypomethylation reflects *cis*-regulatory interactions (Bird 2002; Hodges et al. 2011; dos Santos et al. 2015). We tested whether certain transcription factors were enriched in lineage-specific promoter HMR extensions and lineage-specific non-promoter HMR gains in human and mouse (Fig. 6A, Methods). At lineage-specific non-promoter HMRs, a number of transcriptional activators were enriched on both lineages, such as EP300, activator protein 1 (AP-1) subunits FOS, FOSL1 and JUN, and AP-1 interacting factor MAFK. In addition, chromatin looping factors including CTCF and cohesin components RAD21 and SMC3 were overrepresented on both lineages. Of the CTCF binding sites found in recently evolved lineage-specific non-promoter sperm HMRs, 33% and 43% of the sites were also lineage-specific hypomethylated in the ESCs of human and mouse. These observations suggest a degree of evolutionary plasticity for these factors in their roles associated with 3D organization of the genome and long-range promoter-enhancer interaction; in light of results presented above, this plasticity seems to be increasing the number of regions involved in these functions.

Lineage-specific promoter HMR extensions on both lineages showed enriched binding sites of EP300 and RNA polymerase II, while extensions on the human lineage showed enrichment of many transcriptional activators that are also enriched in human lineage-specific non-promoter HMR gains. Genes with primate-lineage-specific promoter HMR extension showed enrichment in developmental process and nervous system process (Supplemental Fig. S9A, Table S13). About 12% of primate-lineage-specific HMR gains in human sperm were also hypomethylated in human ESC, and as much as 82.5% overlap with ENCODE transcription
factor binding sites, DNase I hypersensitive sites (Thurman et al. 2012) or Roadmap somatic HMRs. This suggests that recently evolved sperm HMRs likely can take on roles of regulatory elements in developmental stages beyond germ cell development.

Despite being lineage-specific, HMR births on these two parallel lineages sometimes occurred in each other’s neighborhood in the orthologous genome (Supplemental Fig. S9B), and they shared a sizeable list of common associated genes (Jaccard index = 0.41, Fisher exact test $p < 2.2 \times 10^{-16}$, Methods). These common genes were enriched for functions in embryonic developmental processes, including mesoderm and ectoderm development and nervous system development (Fig. 6B). Meanwhile, analysis of associated genes that were specific to one species did not yield any enrichment of biological processes. Together, these results suggest that independently evolving features of the sperm epigenome are involved in regulating common genes that are important for embryonic development. This may be viewed as a form of convergent evolution: regulatory features are emerging independently for the same genes.
Figure 6: HMRs gained on parallel lineages are associated with similar transcription factors and developmental genes. (A) Enrichment of transcription factor binding sites in HMR birth and extension regions on the human lineage and mouse lineage. ("*" indicates factors enriched in both lineages.) (B) Lineage-specific HMRs are associated with gene sets with significant overlap, and that are enriched in developmental processes.
DISCUSSION

We used high-resolution whole genome DNA methylation profiles and quantitatively measured the divergence among the sperm methylomes of seven mammals with a newly developed phylo-epigenetic model. Several key findings emerged from our analysis. There is a global trend towards expansion of the hypomethylated fraction of the genome on all phylogenetic lineages. Hypomethylated intervals around transcription start sites have evolved to be substantially wider in primates and dog than in rodents. Genomic intervals with lineage-specific hypomethylation are associated with common developmental processes, and are often close to sets of genes with significant overlap. Analysis of transcription factor binding data indicates that regions of rodent-specific and primate-specific hypomethylation are enriched for binding sites of orthologous transcription factors.

DNA hypomethylation reveals cell-type and species-specific distribution of active or poised regulatory elements (Lister et al. 2009; Stadler et al. 2011; ENCODE Project Consortium 2012). Evolution and divergence of DNA methylation patterns reflect the evolution of regulatory elements. We found that the fraction of DNA hypomethylation in the orthologous genome has been expanding, with frequent HMR births and HMR extensions on all lineages of the phylogenetic tree, and provided evidence for progressive widening of HMRs at both promoters and potential distal enhancers. These results suggest a mode by which new regulatory elements may evolve to increase specificity of transcriptional regulation, and are in line with the findings from a recent theoretical study of transcription factor binding site evolution (Tuğrul et al. 2015). Specifically, in promoters and enhancers, which harbor clusters of binding sites, the additive effect of transcription factor binding on gene expression and cooperative binding of transcription factors can accelerate emergence of new transcription factor binding sites. In contrast, evolution of an isolated binding site would typically be much slower. In light of models of regulatory elements, such as the “enhanceosome” and “billboard” models for transcriptional enhancers (Arnosti & Kulkarni 2005), our phylo-epigenetic model provides a phylogenetic tool for elements that lack the degree of sequence conservation often required to infer function.

Our analyses of DNA sequence and methylation divergence revealed principal features of the relationship between genomic and epigenomic evolution. We observed methylome evolution rate differences between lineages, similar to those observed in genomic evolution, reflecting a genetic component underly-
ing epigenetic divergence. Sequence divergence has been associated with gain and loss of some regulatory elements (Schmidt et al. 2010). Linking DNA sequence divergence with DNA methylation changes, Hernando-Herraez et al. (2015) have shown within multiple primate species that species-specific methylation is associated with elevated species-specific nucleotide substitutions. We add to this picture by showing that lineage-specific hypomethylation is associated with increased sequence substitutions compared with the parallel sister lineage in the phylogenetic tree. Interestingly, in the 7-way orthologous genome, sequences in lineage-specific sperm HMRs are usually more conserved than flanking regions, perhaps reflecting a functional exaptation of ancestral sequence and repurposing of ancestral regulatory elements in species and cell type-specific ways.

Germline DNA methylation patterns have shaped the mammalian genomic distribution of CpG dinucleotides. In human germline cells across many generations, the promoter HMRs alternate between a wider size in sperm cells and a narrower size in ESCs, which has left distinctive signatures of deamination-induced CpG depletion in the human genome (Molaro et al. 2011). In this study, we add further support to this theory by delineating the occurrence time during evolution of human sperm HMRs, and showing that CpG enrichment in HMRs is strongly correlated with HMR age. In addition, CpG enrichment around promoter HMR boundaries in primates, rodents and dog offers additional support to our inference of parallel promoter HMR extension in primates and dog by phylo-epigenetic modeling. Moorjani et al. (2016) reported a more clock-like behavior of CpG transitions compared with other substitutions. Understanding germline methylation patterns as well as their evolutionary history in different species would enable better calibration of divergence time by such a CpG-based molecular clock.

Several factors in the mature sperm of human and mouse, such as histone retention and modification, CTCF binding and microRNAs, have been proposed to have regulatory importance in embryonic development (Carone et al. 2014; Hammoud et al. 2014; Rodgers et al. 2015; Teperek et al. 2016). For example, Lesch et al. (2016) reported that epigenetically poising developmental regulator genes with bivalent histone modifications (H3K4me3 and H3K27me3) at gene promoters was a conserved phenomenon in mammalian sperm, and differences in gene poising in sperm were correlated with differential developmental expression between species. However, of all epigenetic modifications, DNA methylation is the only epigenetic mark with known mechanisms of stable transmission across cell divisions (Henikoff & Greally 2016).
though most DNA methylation marks are removed during two waves of epigenetic reprogramming, regulatory elements that can escape demethylation have been observed in both zygotic and primordial germ cell reprogramming (Borgel et al. 2010; Hackett et al. 2013; Wang et al. 2014). In mouse, oocyte-derived methylation is involved in regulating trophoblast development (Branco et al. 2016), and paternal age-related sperm DNA methylation changes are associated with transcriptional dysregulation of developmental genes in the offspring (Milekic et al. 2015). We have observed lineage-specific HMRs associated with conserved developmental processes, and even proximal to the same genes. Meanwhile, these lineage-specific regions are enriched with binding sites of chromatin-looping factors and enhancer binding factors. These observations suggest that the plasticity accommodated by certain regulatory factors is conserved, despite substantial turnover in the specific sites of regulation. Such plasticity may reflect or permit greater divergence between species in how the associated genes are regulated even in the context of developmental programs that are deeply conserved within mammals.

In parallel with emergence of species-specific regulatory elements, divergence in the organization of gene promoters may be of equal significance in mammalian evolution. The striking differences in sperm promoter HMR sizes between species indicate a global divergence in epigenetic regulation and chromatin organization of orthologous genes including key developmental genes. We observed binding sites of repressive factors EZH2 (a subunit of Polycomb repressive complex 2, PRC2) and SIN3A to be enriched respectively in human and mouse lineage promoter HMR extension regions. Although the data from somatic cells does not necessarily imply binding at the same sites in male germ cells, those two repressor complexes have been found indispensable for proper repression of soma-specific genes through chromatin remodeling during germ cell development (Gallagher et al. 2013; Mu et al. 2014). In mouse ESC, it has been observed that both complexes interact with TET1 but at distinct gene promoters (Williams et al. 2011; Neri et al. 2013). The binding profile of TET1 spreads wider at bivalent gene promoters (Neri et al. 2013), possibly due to association with PRC2, which has a proposed mechanism for spreading at bivalent promoters (Margueron et al. 2009). In contrast, TET1 binding is much narrower when it co-binds with SIN3A, which has a narrower binding profile as well (Neri et al. 2013). We also observed lineage-specific enrichment of histone modifications in lineage-specific HMR extensions, indicating that the breadth of sperm histone marks differs between species. Broad H3K4me3 domains at gene promoter have been connected with in-
creased RNA polymerase II pausing, transcriptional consistency and cell identity (Benayoun et al. 2014). Although mature sperm are usually assumed to be transcriptionally inactive, presence of RNA polymerase II has been detected in the nuclei of mature mouse sperm (Lin et al. 2013). Considering these results, it is possible that divergent recruitment of *trans*-acting factors during germ cell development contributes to different sizes of promoter HMRs between species, and has further influence in zygotic gene activation and embryo development. It remains to be examined whether the occupation of these chromatin-modifying and transcription complexes differ between species in the context of germ cell development, and how they are associated with promoter methylation patterns in different species.

In conclusion, we characterized evolutionary changes in mammalian sperm methylomes at high spatial and temporal resolution. Our analyses reveal the global trend and lineage-specific features of expanding DNA hypomethylation during evolution, and provide insights to how evolution of the epigenome may contribute to species-specific fine-tuning of conserved mammalian developmental programs.
METHODS

Whole genome bisulfite sequencing data sets

We collected sperm samples from one gorilla individual, two rat individuals and three dog individuals, and generated whole genome bisulfite-sequencing (WGBS) data sets. See Supplemental Methods for details of sample collection, library preparation and sequencing.

Public WGBS data used in this study include human and chimpanzee sperm (GEO accession GSE49624; Hammoud et al. 2014 and GSE30340; Molaro et al. 2011), human H1 ESC (GSE16256; Lister et al. 2009), human H9 ESC (GSE19418; Laurent et al. 2010), human and chimpanzee B-cell (GSE31971; Hodges et al. 2011), gorilla whole blood (NCBI SRA accession SRP059313; Hernando-Herraez et al. 2015), rhesus macaque sperm (GSE60166; Lu et al. 2015), rhesus macaque PBMC (GSE34129; Tung et al. 2012), rat left ventricle (European Nucleotide Archive accession number ERP002215; Johnson et al. 2014), mouse sperm (GSE49624; Hammoud et al. 2014), mouse B-cell (SRP029721; Kieffer-Kwon et al. 2013), mouse ESC (GSE30206; Stadler et al. 2011), and the dog kidney epithelial cell line MDCK (GSE48527; Carmona et al. 2014).

Mapping and annotation

Reads from whole genome bisulfite sequencing were mapped with the RMAP software (Smith et al. 2009) to respective reference genome of each species: hg19, panTro4, gorGor3, rheMac3, mm10, rn5, and canFam3. Subsequent analyses on duplicate removal, bisulfite conversion rates, methylation levels and HMRs were carried out with tools from the MethPipe package (Song et al. 2013). Mapping statistics are provided in Supplemental Table S1. We did not observe significant non-CpG methylation in any species, and focused our analyses on CpG methylation. We used Ensembl gene annotations (Release 75) (Flicek et al. 2014) for species involved in this study. Coordinates of rhesus macaque gene annotation were converted to assembly rheMac3 from rheMac2 with liftOver from the Genome Browser tool suite (Kent et al. 2002).
**Methylome alignment**

We produced multiple methylome alignment by mapping non-human CpG positions and methylation levels to the human genome assembly hg19, using the 100 vertebrate multiz alignment (hg19) downloaded from UCSC Genome Browser (Rosenbloom et al. 2015). We used UCSC Genome Browser binary utility `mafSpeciesSubset` to filter out sequence records from species other than the seven species of interest. Next, we fused together adjacent alignments blocks that were separated by species not included in this study, using Galaxy script `maf_thread_for_species.py` (Giardine et al. 2005; Blankenberg et al. 2010; Goecks et al. 2010). Then we parsed the locations of CpG sites in each species and the corresponding coordinates in the human reference genome from alignment blocks that contain sequences from all seven species. We assessed whether using the latest human genome assembly GRCh38 would significantly impact the methylome alignment. Although GRCh38 has improved coverage at centromere sequences and increased diversity with alternate loci, sequences in genomic regions conserved with other mammalian species are highly consistent with the older assembly GRCh37/hg19. Only 4596 (0.067%) CpGs in the 7-species alignment blocks under the new assembly (alternate loci and unplaced sequences excluded) were absent from the old assembly, 1308 (0.019%) were non-CpG sites in the old assembly, and 78 (0.001%) had insertion/deletion or non-unique mapping, while the rest of CpG sites had one-to-one mapping between the two versions of human reference genome assemblies.

To define seven-way orthologous genome, we excluded any regions that are more than 1 kbp in size and contain no CpGs or have 0 read coverage in at least one species, and consider the resulting set of genomic regions, covering 370 million bases, as the regions conserved across all seven species. We chose 1 kbp as the minimum CpG desert size because the correlation of methylation levels at neighboring CpG sites drops to background level at distance above 1 kbp.

To map HMRs from native genome assembly to the human reference genome, we kept track of HMR membership of individual CpGs, and then merged consecutive CpGs that belonged to the same HMR in the original native genome and were not separated by CpG deserts into a HMR. We filtered out HMRs with less than 5 CpG sites.
Phylo-epigenetic model considering CpG-site interdependence

We constructed a phylo-epigenetic evolution model $\Theta = \{\tau, \pi, Q, F, G\}$, where $\tau$ is a phylogenetic tree with known phylogeny and unknown branch lengths; $\pi$ is the distribution of methylation states at the first site in the last common ancestor (LCA, root of the phylogenetic tree); $F$ is the transition probability matrix for methylation states between neighboring CpG sites in the LCA; $Q$ is a continuous time Markov chain transition rate matrix for the evolution of methylation state at a single site; $G$ is a transition probability matrix for neighboring CpG sites in descendant species. The methylome of the LCA is modeled by a two-state discrete-time Markov chain with model parameters ($\pi$, $F$). Evolution of the first site in the genome is modeled with continuous-time Markov process $Q$ along the phylogenetic tree $\tau$, as in the standard independent-site phylogenetic model. The main component of our phylo-epigenetic model is the combination of the horizontal inter-CpG dependence process described by $G$ and the vertical inheritance process described by $Q$ and divergence time. Consider a parent-child pair of species $(u, v)$ with branch length $\ell_{uv}$, and let the methylation state at the site $i$ of $v$ be $v_i$. Suppose the methylation states of $v_i$'s parent site and previous site are given. We define the conditional distribution of $v_i$ given $v_{i-1}$ and $u_i$ as

$$
Pr(v_i | u_i, v_{i-1}) = \frac{G_{v_{i-1}v_i} \exp(\ell_{uv})_{u_i v_i}}{\sum_{x \in \{0, 1\}} G_{v_{i-1}x} \exp(\ell_{uv})_{u_i x}}
$$

This modeling strategy achieves a balance between computational tractability and incorporation of dependence relations between neighboring CpG within and across species (see discussion of alternative models in Supplemental Methods). The input data can be binary methylation states, continuous methylation probabilities at individual CpG sites in extant species and missing data at sites in an extant species when its orthologous sites have observed data in other species. See Supplemental Methods for details about the model and parameter estimation methods. In brief, we used a Monte Carlo Expectation Maximization (MCEM) algorithm (Wei & Tanner 1990), where in the E-step posterior distribution of methylation state at unobserved CpG sites are approximated by Gibbs sampling. In-house software can be downloaded from https://github.com/smithlabcode/Epiphyte.
Model selection between reversible and unrestricted evolution processes

We used Akaike information criterion (AIC) to compare two independent-site phylogenetic models, with and without the assumption of reversible evolutionary process. We discretized average methylation probabilities (cutoff 0.5) to determine the methylation states in 200-bp bins in the 7-way orthologous genome. Only bins with non-zero CpG coverage in all species are included to form complete observation at extant species. The reversible model and the unrestricted model were estimated with R package RPHAST (Hubisz et al. (2011), Table S8).

Types of methylome evolution events

We inferred HMRs in ancestral species using the proposed interdependent-site phylo-epigenetic model. Methylome evolution events – HMR gain in the form of birth and extension and HMR loss in the form of death and contraction – along each branch of the phylogenetic tree were determined by comparison of HMRs in the corresponding parent and child species. We filtered out events of size less than 50 bp or located at the boundaries of 7-way orthologous genome fragments.

To examine the enrichment of de novo (birth or death) and secondary (extension or contraction) methylome evolution events on each phylogenetic tree branch, we computed the distances between HMR gains to the closest ancestral HMRs, and between HMR losses to the closest descendant HMRs. We formed the expected distance distribution by randomly shuffling HMR gain events within all methylated regions in the ancestral methylome and HMR loss events within ancestral HMRs, using BEDTools (Quinlan & Hall 2010). The enrichment of different types of HMR gain and loss events was measured with the ratio between the observed and expected distance distribution (Supplemental Fig. S6).

CpG enrichment at promoter HMRs

In the human-referenced 7-way orthologous methylome alignment, we filtered for human and mouse sperm HMRs such that both their intersection and union contain exactly 1 gene transcription start site (TSS). We obtained human promoter HMR extension regions relative to the mouse promoter HMRs with minimum length of 200-bp. The two boundaries of such an extension region correspond to a human HMR boundary and a mouse HMR boundary. We masked out regions that are not alignable between human and mouse ac-
cording to their pair-wise sequence alignment. We used a sliding 2bp-windows with 1bp step size around the human HMR boundary of the extended HMR, the mouse HMR boundary contained within a wider human HMR, and around TSS along the human genome to measure GC content and CpG occurrences, and produced a meta plot for CpG enrichment relative to the boundaries and TSS. To make these measurements in the mouse genome, we mapped human methylomes to mouse mm10 genome assembly through the mouse-referenced multiz60way alignment, and followed the same analysis procedures. We performed the same analysis for dog-mouse comparison using mouse-referenced and dog-referenced methylome alignments.

**Relative sequence substitution rate**

We used 15-way eutherian mammals Enredo-Pecan-Ortheus (EPO) multiple alignments corresponding to the Release 75 of Ensembl (Herrero et al. 2016) and we extracted the alignment of the 7 species in this study and the ancestral species in the corresponding phylogeny. Sequence substitutions were annotated to individual lineages by comparing the ancestral sequence and the descendant sequence. We only considered alignment columns that have no insertion or deletion in all 7 species. Relative sequence substitution rate (RSSR) in orthologous regions between two sister lineages is defined as the ratio of the total number of substitutions on one lineage to the total number of substitutions on the other lineage. To form the background distribution of RSSR, we randomly sampled regions from the 7-way orthologous genome. The sampled region size was chosen to match the smaller size of the total aligned bases in lineage-specific HMRs between the two lineages. We independently sampled 5000 times, and created the sampling distribution of RSSR for each pairs of sister lineages. The significance (one-sided \( p \)-value) of observed RSSRs in lineage specific HMRs was calculated using a normal distribution fit to the empirical sampling results.

**Enrichment of transcription factor binding sites**

We used ENCODE transcription factor binding sites for human and mouse (ENCODE Project Consortium 2012; Gerstein et al. 2012; Wang et al. 2013), and mouse EZH2 binding sites from studies by Ku et al. (2008) and Peng et al. (2009). To evaluate enrichment of the binding sites of each TF in, for example, human lineage-specific HMRs, we used fisher exact test on the contingency table of binding site counts: \{factor, other factors\} × \{lineage-specific HMRs, other human HMR\}. See Supplemental Methods for
more details.

**Gene ontology analyses**

For genes with promoter HMR extension in primate lineage, we used GOrilla (Eden et al. 2009) to identify enriched biological processes at false discovery rate 0.05, and removed ontology term redundancy with REViGO (Supek et al. 2011). For genes closest to lineage-specific HMR gains that are common between parallel lineages, we used PANTHER overrepresentation test (Mi et al. 2016) to find overrepresented biological processes with $< 0.05$ Bonferroni corrected $p$-value. More details can be found in Supplemental Methods.

**DATA ACCESS**

The whole-genome bisulfite sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO) under accession GSE79566. Methylome alignment, inference of ancestral methylation states and HMRs are provided in UCSC Genome Browser Track Data Hubs (Raney et al. 2014) as part of the public hub DNA Methylation (Song et al. 2013). Custom analysis scripts and in-house software are available as Supplemental Scripts.

**ACKNOWLEDGMENTS**

We thank Dr. Tara Stoinski, Zoo Atlanta for the gorilla sample; Dr. Beckie Williams, Yorba Regional Animal Hospital for dog samples. We thank Michael Kessler for sample collection; Brent Young for dog sperm DNA extraction. We thank Dr. Joseph G Hacia, Dr. Peter Calabrese and Giorgia Battistoni for helpful discussions and comments. We thank Vanderbilt Technologies for Advanced Genomics (VANTAGE) and Beijing Genomics Institute (BGI) for sequencing services. We thank the University of Southern California High Performance Computing Centers for computational support. This research was partially supported by National Institutes of Health grant R01 HG005238 to A.D.S.
DISCLOSURE DECLARATION

The authors declare no conflict of interest.

References


