Supplemental Materials: Evolutionary Expansion of DNA Hypomethylation in the Mammalian Germline Genome

Contents

Supplemental Figures

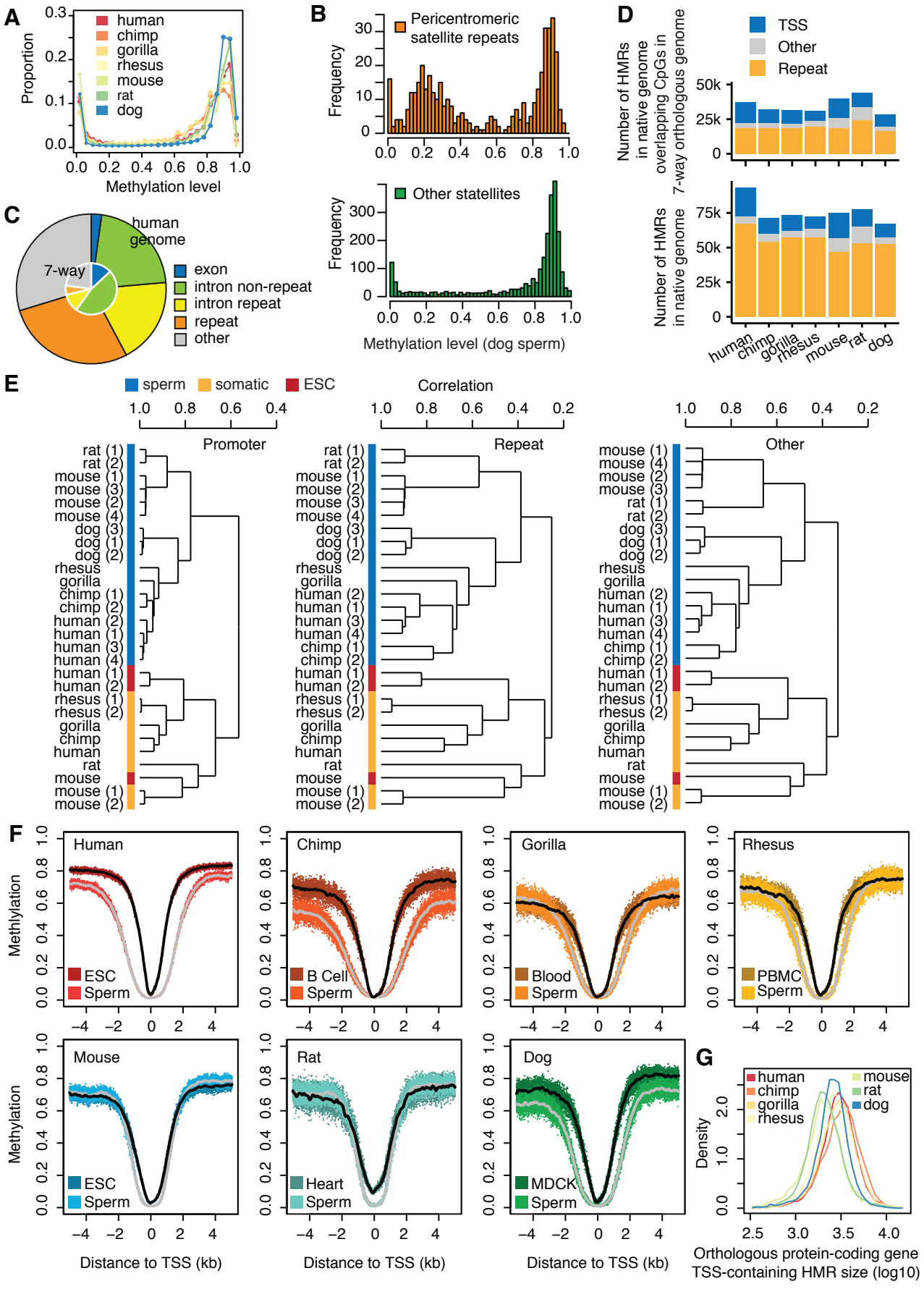

Figure S1: (Caption on next page)

Figure S1: Mammalian sperm methylome characteristics. (A) Distribution of single CpG methylation levels in each species. CpGs with less than 10x coverage were excluded. Y-axis shows the proportions of CpGs in 25 bins for methylation levels. (B) Distribution of methylation levels at pericentromeric satellites and other satellites in dog sperm. (C) Genomic context composition of the 7-way orthologous genome in comparison with the entire human genome. (D) The total number of HMRs in the entire native genomes, and the number of native HMRs that contain CpGs from the 7-way orthologous genome. (E) Hierarchical clustering of sperm, ESC and somatic methylomes in different genomic contexts. (F) Average methylation around TSS that are hypomethylated in sperm and ESC/somatic cells in the native genome of each species; solid lines represent data smoothed by splines. (G) Log-scale size distribution of TSS-containing sperm HMR for orthologous protein-coding genes in the native genome of each species. HMRs containing more than one TSS are excluded.

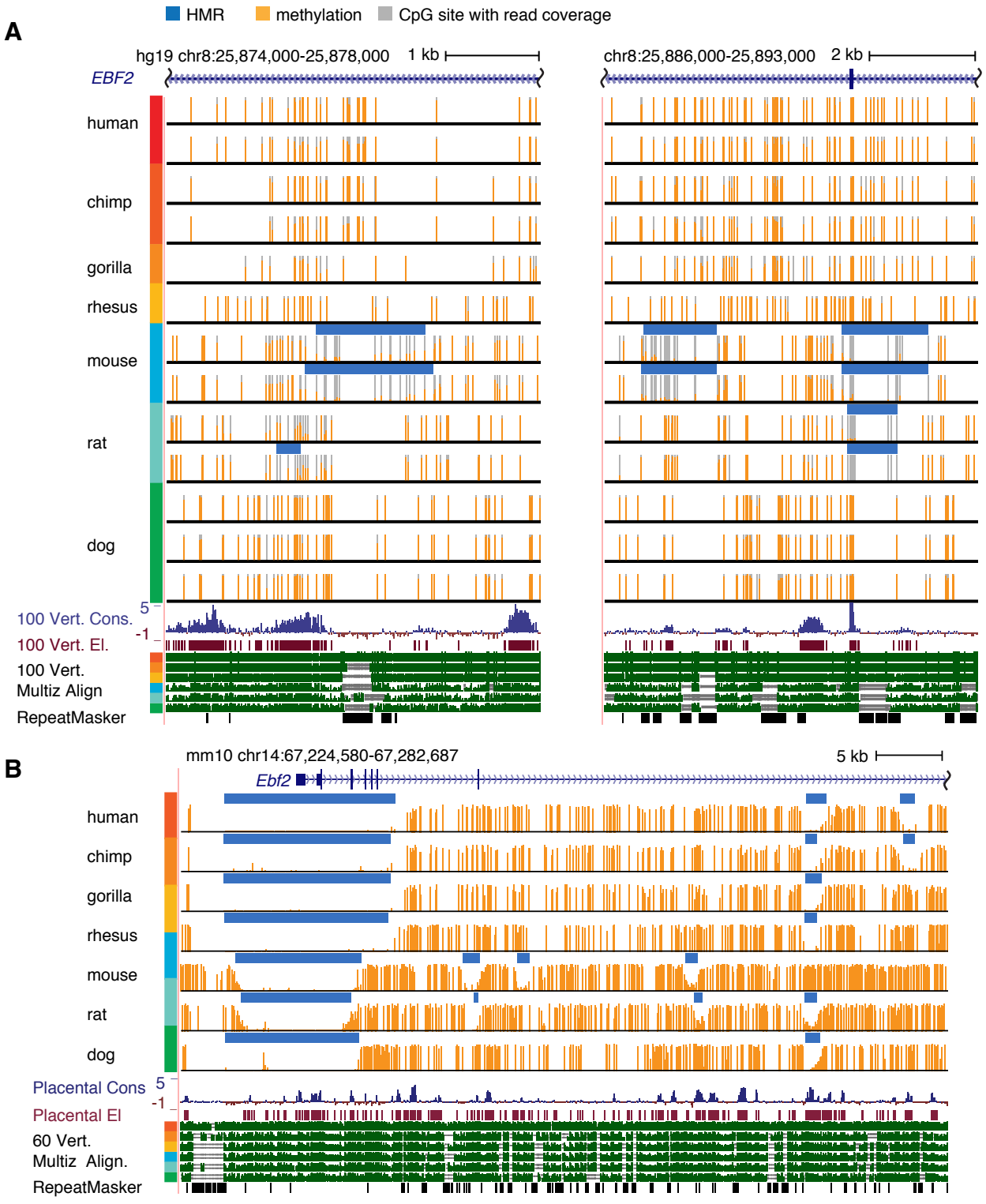

Figure S2: Example region in methylome alignment related to Fig. 1D. (A) Zoomed-in browser track image for dashed-line boxes in Fig. 1D. (B) Mouse-referenced methylome alignment in the orthologous region of Fig. 1D.

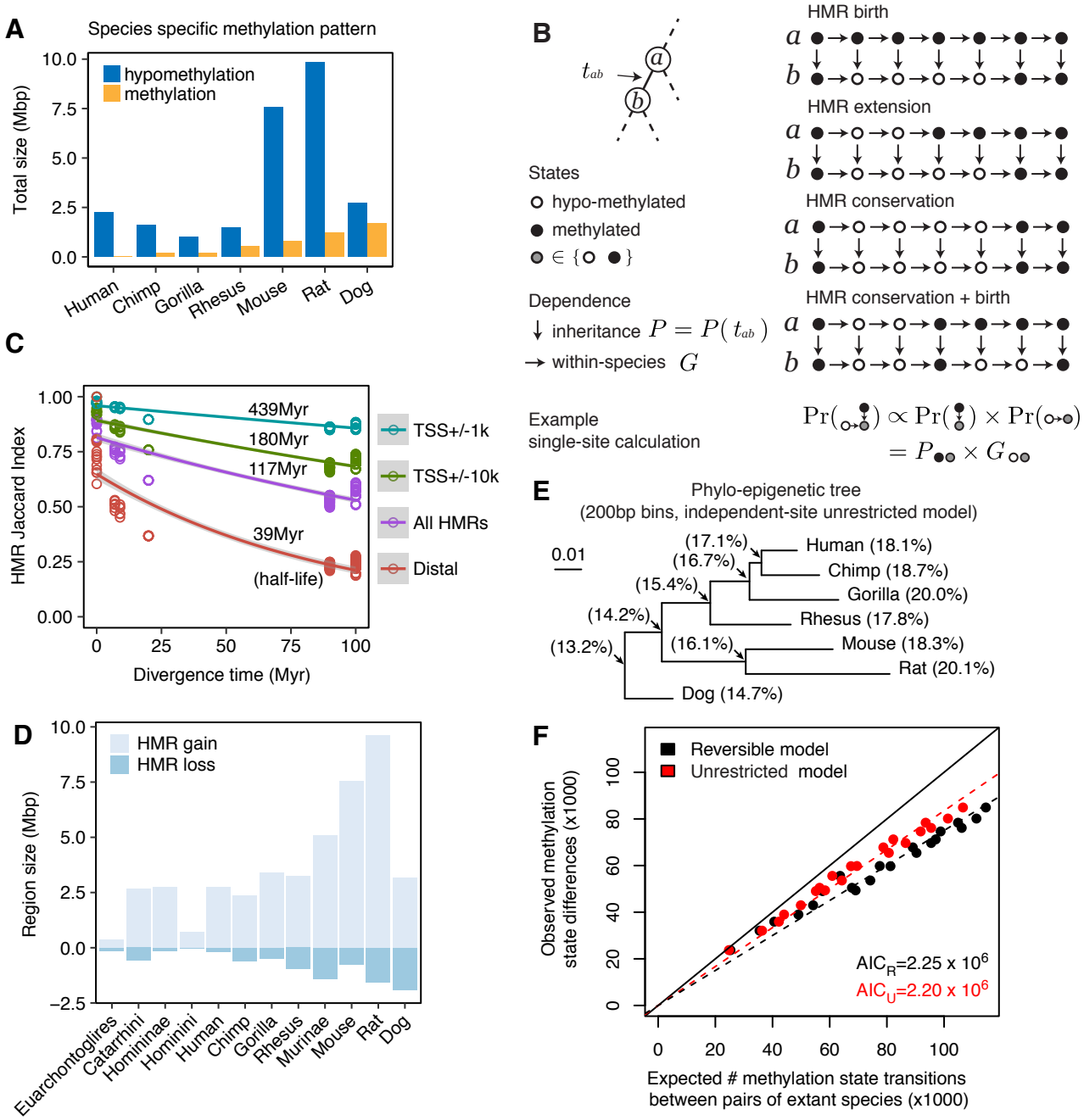

Figure S3: (Caption on next page)

Figure S3: Modeling methylome evolution reveals overall methylation loss. (A) Total size of species-specific HMRs and non-HMRs in 6 species, with dog as the out-group species. (B) Schematic presentation of how the intergeneration inheritance, P, modeled by a continuous time Markov process, and interdependence of neighboring sites within species, G, modeled by a discrete time Markov chain, are combined. In the example evolution scenarios from an ancestor species a to a descendant species b, the number of hypomethylated sites in b is the same in all cases, but the probabilities differ in each case under this model. (C) Empirically determined HMR divergence rates (HMR half-lives) by the fraction of conserved orthologous HMRs between pairs of species as a function of divergence time in million years. (D) Total size of HMR gain and loss on individual branches, estimated by the interdependent-site phylo-epigenetic model at single CpG resolution. (E) Evolutionary tree and hypomethylation fraction at individual species estimated by independent-site unrestricted model using discretized methylation states in 200-bp bins with observations from all 7 species. (F) The unrestricted model (without the reversibility assumption) fits the observed methylation divergence better than assuming reversible state substitution process by Akaike information criterion (AIC).

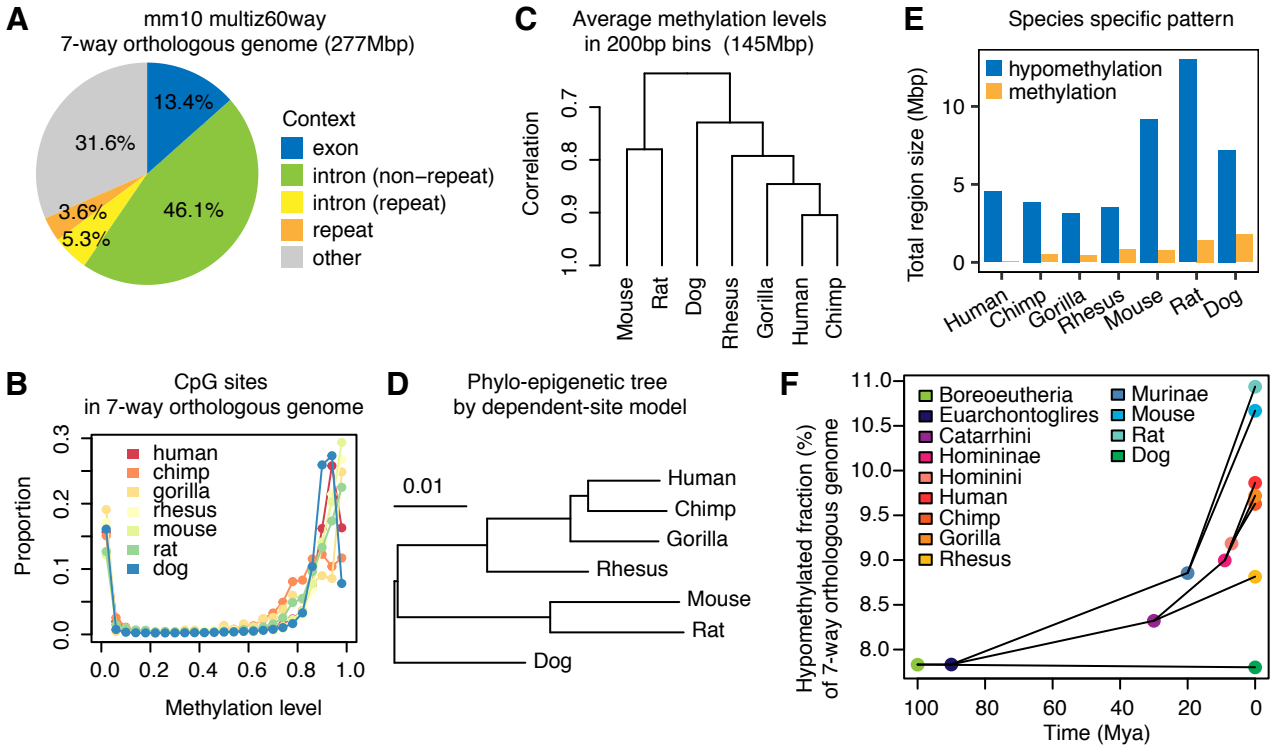

Figure S4: Evolution of sperm DNA methylation in mouse-referenced 7-way orthologous genome. (A) Genomic context composition of mouse-referenced 7-way orthologous genome. (B) Distribution of single-CpG methylation levels in sperm for CpG sites in mouse-referenced 7-way orthologous genome. (C) Hierarchical clustering of species based on correlation of average methylation levels in 200-bp bins. (D) Evolutionary tree estimated under interdependent-site phylogenetic model. (E) Total size of species-specific hypomethylated and methylated regions. (F) The fraction of 7-way orthologous genome inferred to be hypomethylated in individual species in the phylogeny.

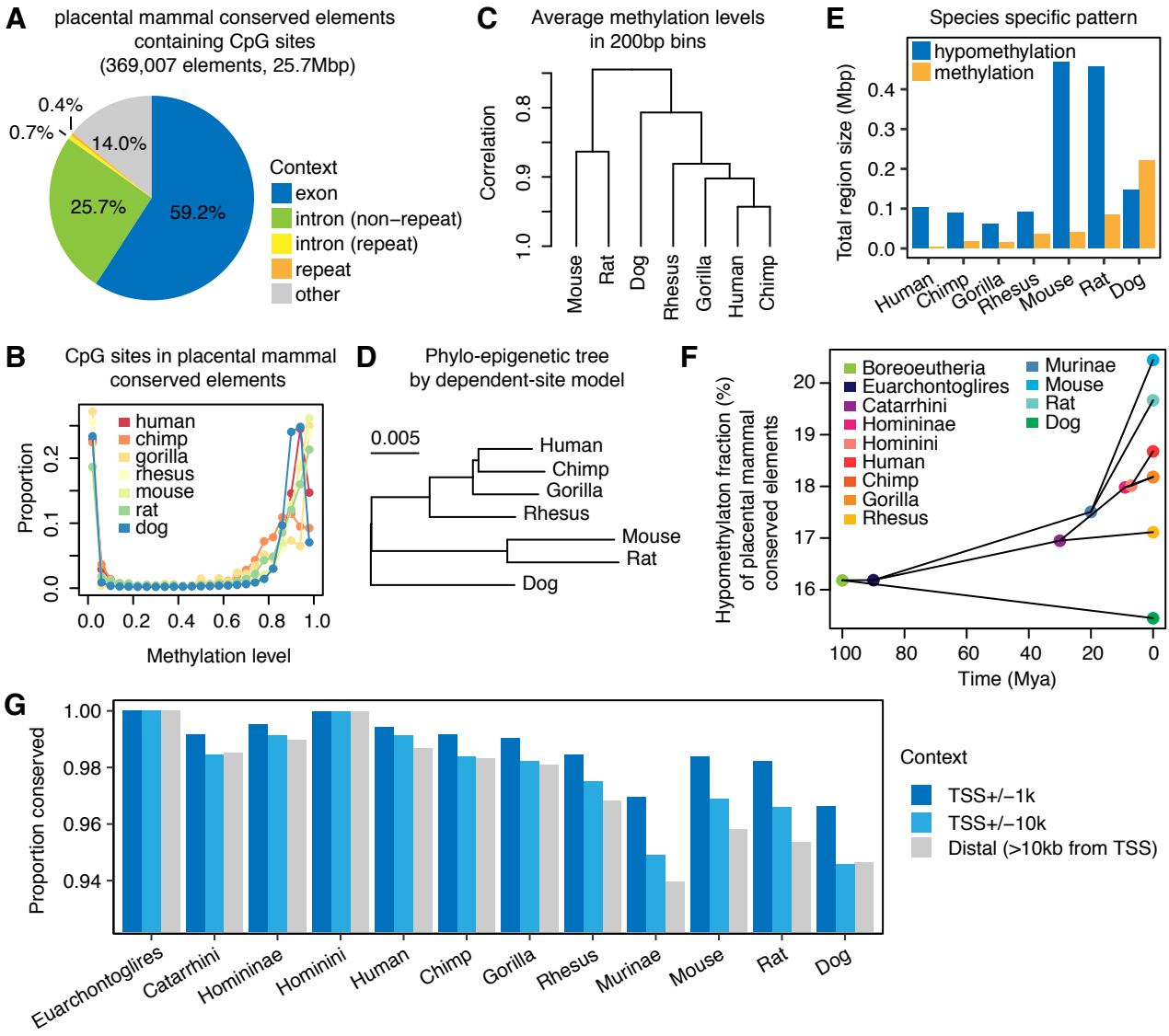

Figure S5: Evolution of sperm DNA methylation in placental mammal conserved elements. (A) Genomic context composition of placental mammal conserved elements. Elements that do not contain CpG sites with observed data in one or more species are excluded. (B) Distribution of single-CpG methylation levels in sperm. (C) Hierarchical clustering of species based on correlation of average sperm methylation levels in 200-bp bins. (D) Evolutionary tree estimated under the interdependent-site phylo-epigenetic model. (E) Total size of species-specific hypomethylated and methylated regions at placental mammal conserved elements. (F) Fraction of the total size of conserved elements inferred as hypomethylated in individual species in the phylogeny. (G) Size proportion of conserved elements inferred to have conserved methylation states along individual branches.

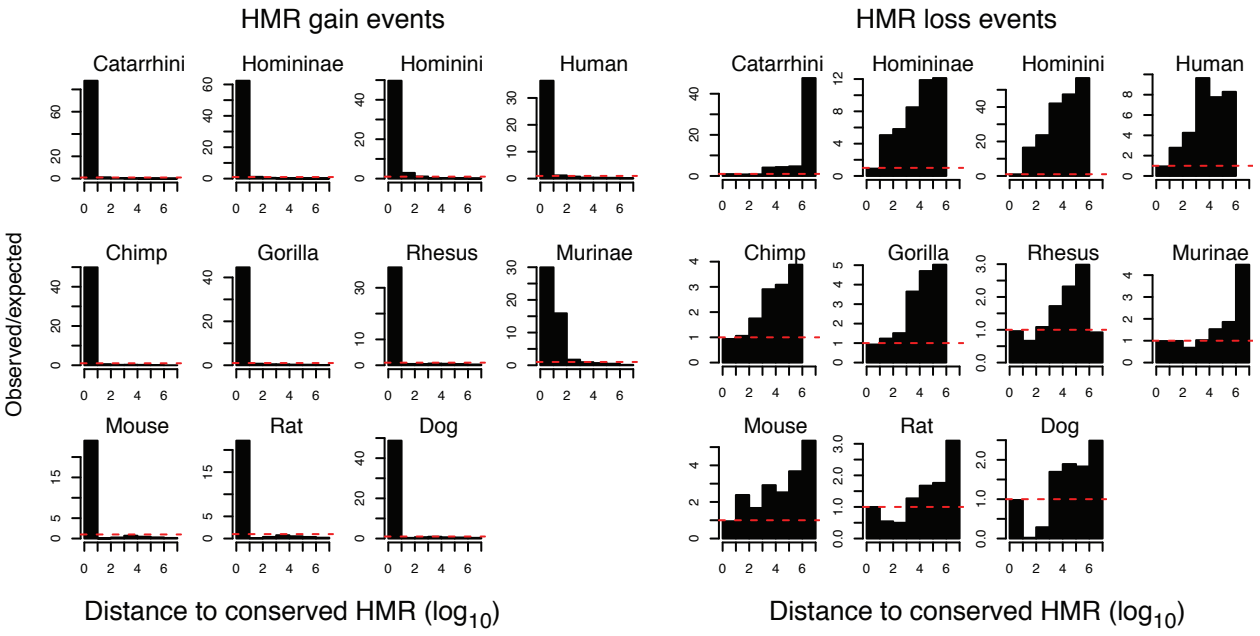

Figure S6: HMR extension and HMR death are the enriched type of HMR gain and loss events. Mutation events were inferred using interdependent-site phylo-epigenetic model on single-CpG methylation probabilities from extant species. These gain and loss events on each branch are shuffled within genomic regions available for the respective type of changes. The relative abundance of events (observed over randomly shuffled) is computed at different distances to inherited HMRs in the descendant species. Close-to-zero distances from inherited HMRs (left-most bin) indicate extension/contraction events. Bigger distances from inherited HMRs indicate birth/death events.

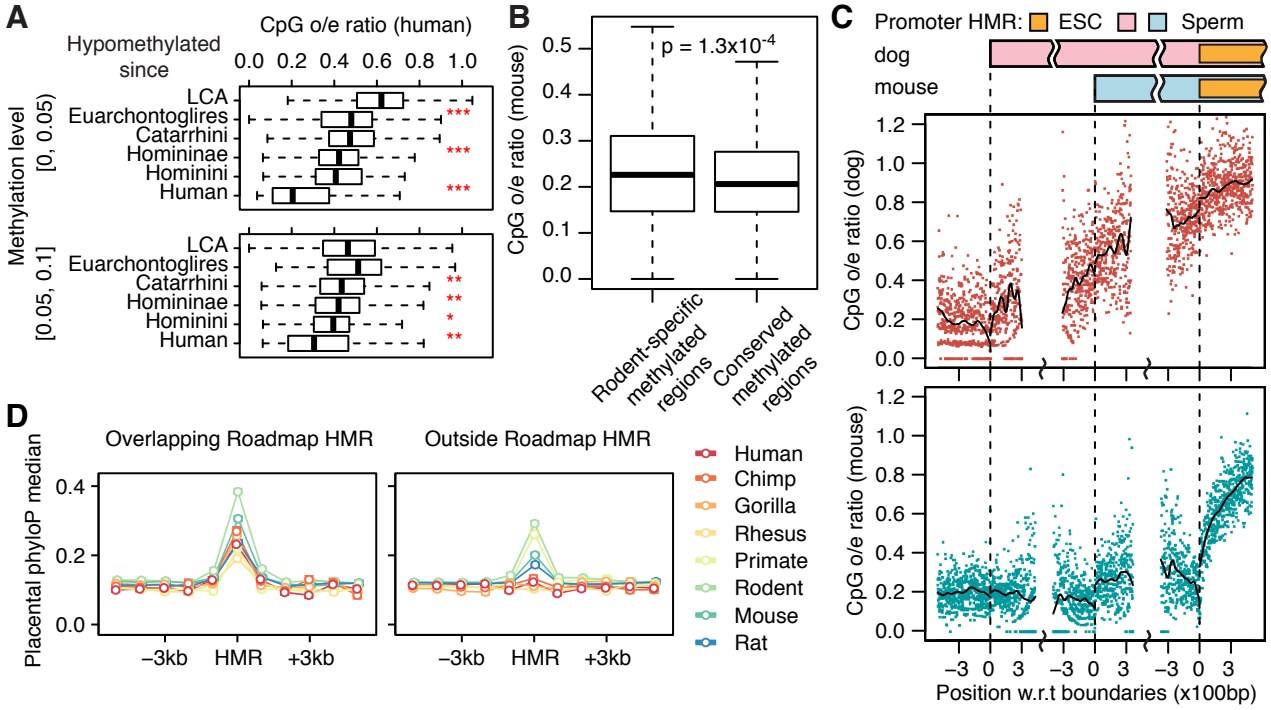

Figure S7: Sequence signatures driven by sperm methylome evolution. (A) CpG enrichment in Human HMRs stratified by HMR age and average methylation level. (B) Mouse CpG enrichment in rodent-specific methylated regions and in regions with conserved methylation across all seven species. (C) CpG enrichment profiles at orthologous gene promoters where dog sperm HMRs are wider than mouse sperm HMRs. (D) Median sequence conservation level (placental mammal phyloP score) in lineage-specific sperm HMRs and neighboring regions (1-kbp bins), separated by whether having overlap with human somatic HMRs (Roadmap).

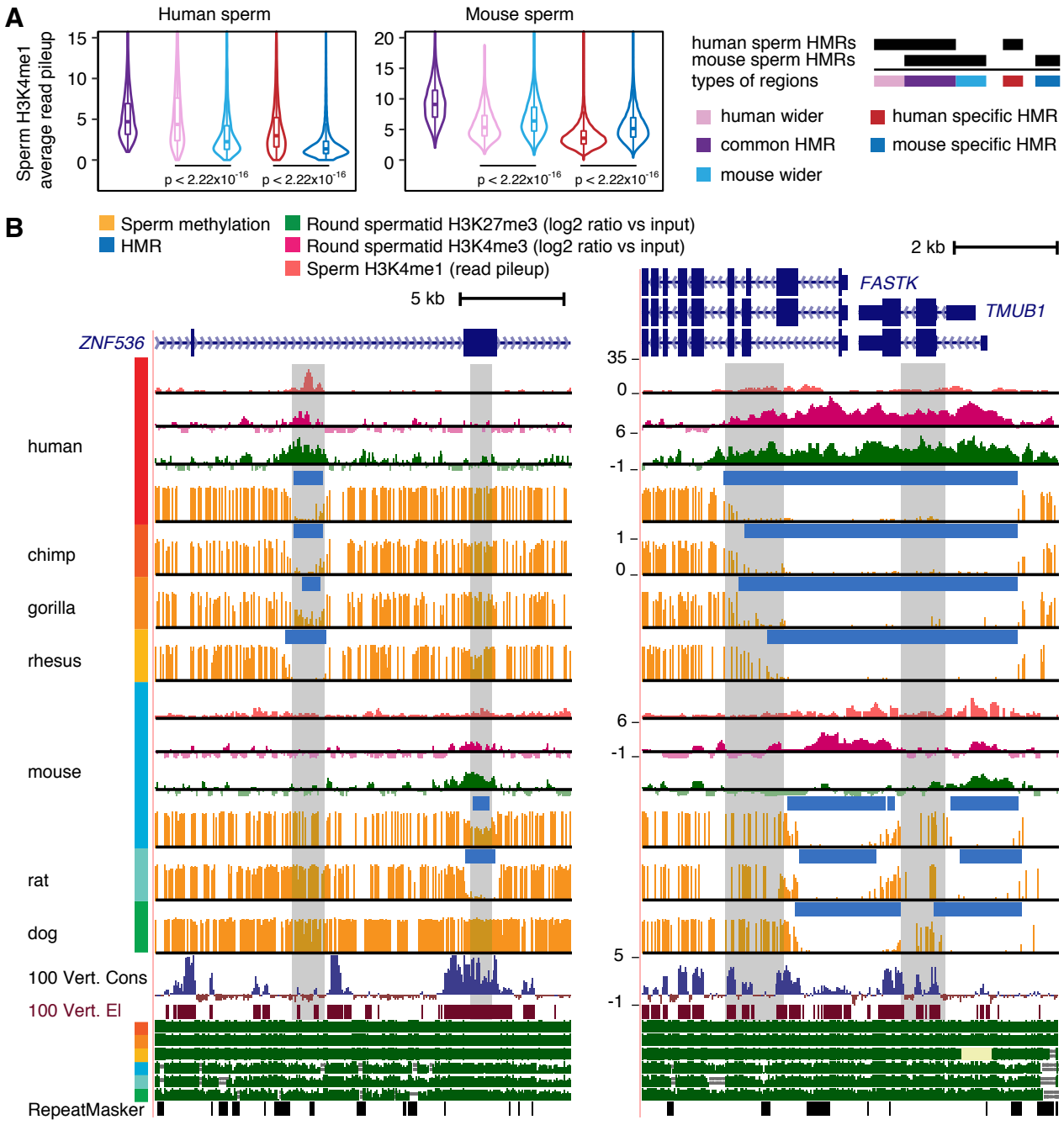

Figure S8: Lineage-specific HMRs are associated with lineage-specific enrichment of histone modifications. (A) Enrichment of H3K4me1 in shared HMRs and lineage-specific HMRs in human and mouse sperm. (B) Example regions showing lineage-specific sperm HMR births and HMR extensions associated with lineage-specific histone modifications in round spermatid and mature sperm.

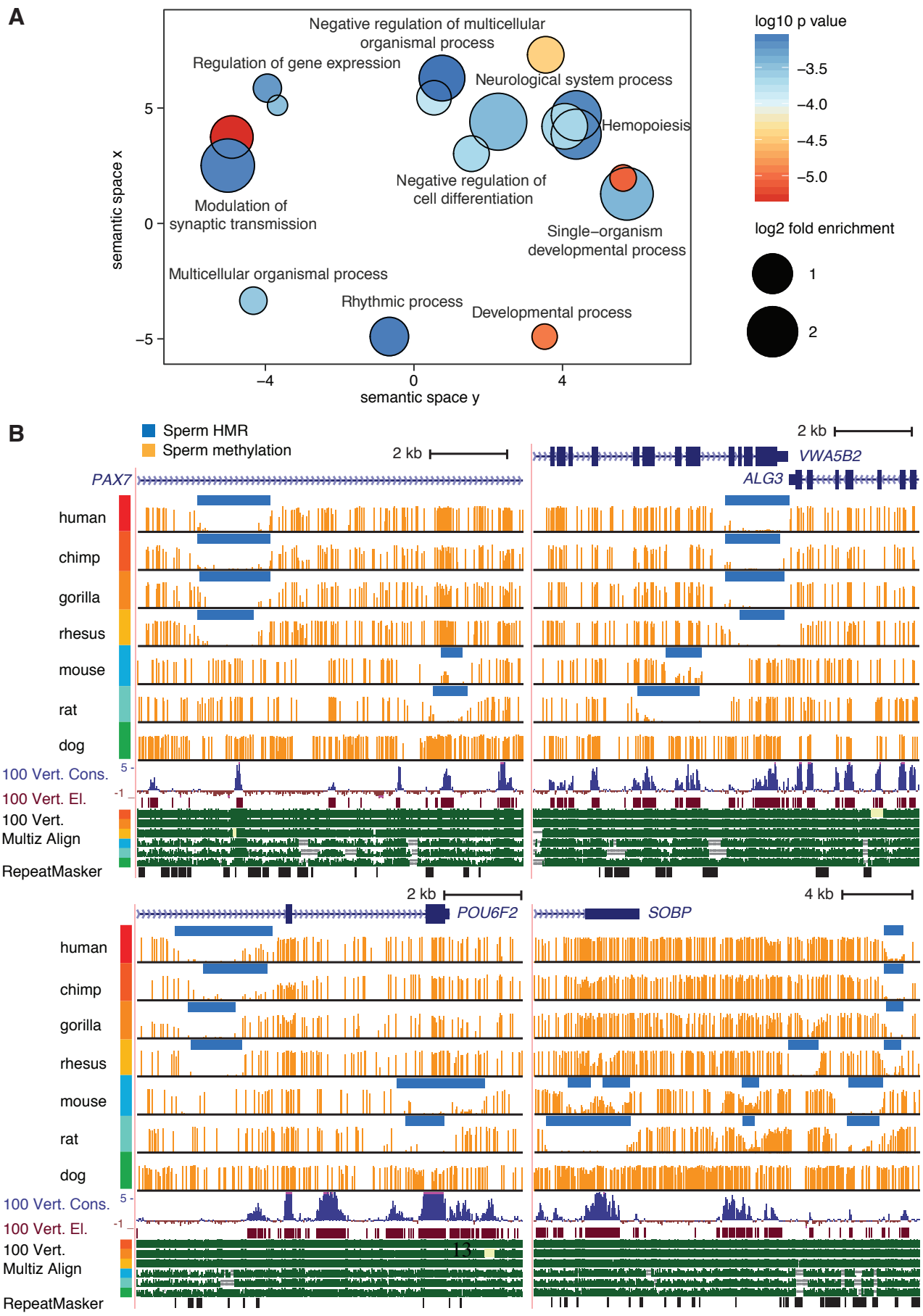

Figure S9: (Caption on next page)

Figure S9: HMR gains on parallel lineages. (A) Enriched biological processes associated with human-lineage promoter HMR extensions. (B) Examples of lineage-specific HMR births on parallel lineages located in proximity to the same genes.

Supplemental Methods

Sperm sample collection and library preparation

The gorilla sample was obtained from 43 year old western lowland Gorilla at Zoo Atlanta after review and approval by the Zoos scientific review committee. The sample was collected opportunistically (cage floor after animal masturbated) and shipped to San Diego overnight in BWW medium with penicillin/streptomycin. Two bisulfite-converted libraries were constructed as previously described (Molaro et al. 2011). Paired-end sequencing was performed on Illumina HiSeq2000 platform (paired-end 76 bp read lengths).

Dog sperm samples were from three individuals, each from a different breed: Doberman, Labrador retriever and Portuguese water dog. Frozen dog semen samples were collected by Dr. Beckie Williams at Yorba Regional Animal Hospital, Anaheim, CA. Genomic DNA was extracted from each dog semen sample using a user-developed protocol (QIAamp DNA Mini kit and QIAamp DNA Blood Mini kit 11/2007, Isolation of genomic DNA from sperm using the QIAamp DNA Mini kit; protocol 2). Extracted DNA samples were sent to BGI for bisulfite-conversion, library construction and sequencing. Bisulfite sequencing libraries were prepared using standard Illumina protocol. Paired-end sequencing was performed on Illumina HiSeq2000 platform (paired-end 90 bp read lengths).

Genomic DNA from rat sperm (two biological replicates) was extracted according to previously described methods (Molaro et al. 2011, 2014). Whole genome bisulfite libraries were constructed using tagmentation-based methods as previously described (Wang et al. 2013b) with the following modifications: Transposomes were assembled with Tn5 enzyme purified in house using a plasmid construct kindly provided by the Sandberg lab and according to protocols previously described (Picelli et al. 2014). Tagmentation of 50ng genomic DNA was performed in Tris-DMF buffer (10 mM Tris-HCl pH 7.5, 5 mM MgCl2, 10% DMF) at 55°C for 8 minutes. Reactions were stopped with 0.04% SDS (final concentration) and incubated for 7 minutes at 55°C. Samples were incubated with 2ul 10uM replacement oligo Tn5mC-Repl01 at 45°C for 10 minutes. Gap repair was performed in T4 DNA Ligase Buffer with 10mM ATP (NEB), 0.1mM each dNTPs (final concentration), 1ul T4 DNA ligase (400,000 units/mL, NEB) and 1ul T4 DNA polymerase (3000 units/mL, NEB) and incubated at 37°C for 15 minutes followed by 25°C for 10 minutes. Reactions were heat inactivated at 75°C for 20 minutes. Bisulfite conversion of DNA was performed using the EZ DNA Methylation Lightning kit (Zymo cat # D5030) according to the manufacturers recommendations. After desulphonation and purification of sodium bisulfite treated libraries, samples were amplified for 14-18 cycles with the Kapa HiFi HotStart Uracil+ Ready Mix according to the manufacturers instructions. Oligonucleotide sequences are detailed in (Wang et al. 2013b). Each replicate library was barcoded, pooled and sequenced on 3 lanes of an Illumina HiSeq 2500 (paired-end 100 bp read lengths).

Hierarchical clustering

In the hierarchical clustering analyses of methylomes, we used 1 - Pearson correlation of average DNA methylation levels in 200-bp genomic bins as pair-wise distance, with complete linkage. The somatic methylomes are from human, chimpanzee and mouse B-cells, gorilla peripheral whole blood, rhesus macaque PBMC and rat left ventricle.

Orthologous promoter HMR sizes

For each non-human species, annotation of protein-coding genes orthologous to human genes were extracted using BioMart from Ensembl release 75. HMRs containing a single transcription start site of orthologous

protein-coding genes were selected from each somatic and sperm methylome. The somatic methylomes used for this analysis were human B-cell, chimpanzee B-cell, gorilla whole blood, rhesus macaque PBMC, mouse B-cell, rat left ventricle and dog MDCK cell line. The methylome of dog MDCK cell line contains partially methylated domains (PMDs). We only used dog MDCK HMRs located outside of PMDs to measure promoter HMR size.

Ultra-conserved HMRs

We defined core HMRs as the intersection of sperm HMRs from all seven species. Sperm HMRs from all species that overlapped with a same core HMR were merged into a single region, which we call conserved HMR. The ratio between the core HMR size and the corresponding conserved HMR size is a measurement of the local conservation of DNA methylation pattern. We chose a lower cutoff at 0.7 for definition of ultra-conserved HMR, which resulted in 250 regions. A more stringent cutoff 0.8 resulted in 42 regions. We used `GOrilla` (Eden et al. 2009) to identify enriched biological processes, molecular functions and cellular components in the target set of genes whose TSS are located within the ultra-conserved HMRs, in contrast with a background gene set, which are genes with TSS hypomethylated in the sperm of all seven species. Enriched GO terms with p-value $< 1 \times 10^{-3}$ are reported in Supplemental Table S6.

Species-specific hypomethylation and methylation

We used HMRs from individual methylomes (aligned to reference genome) to determine species-specific hypomethylated regions and species-specific methylated regions (Supplemental Fig. S3A, S4D, S5D). As an alternative, we also used methylation level cutoffs to determine methylation states and species-specific methylation patterns. Average methylation levels in each species were measured in 200-bp bins along the hg19 genome. Bins with methylation level less than a threshold were considered hypomethylated, and methylated otherwise (Supplemental Table S7).

Phylogenetic tree from multiple genome alignment

We estimated the phylogenetic tree branch lengths under the unrestricted single-nucleotide model (Yang 1994), using the R package `RPHAST` (Hubisz et al. 2011). Multiple genome alignment of the seven species in this study was extracted from the multiple alignments of 99 vertebrate genomes with human genome (hg19/GRCh37, Feb. 2009) (Blanchette et al. 2004) downloaded from UCSC Genome Browser. Phylogenetic tree was estimated independently from alignments in different autosomes. The results were very stable across chromosomes. The average branch lengths were used as the final branch lengths for the phylogenetic tree for the entire orthologous genome across the seven species.

State space and units of measurement for DNA methylation

DNA methylation is often discussed in terms of levels at individual CpG sites, and the level reflects the fraction of cells that have the discrete methyl mark at that site (more accurately molecules with the mark, in the case of non-haploid cells). In multiple studies since 2009, when considering cells that are relatively pure in terms of phenotype, the methylation levels have been observed to fall into two categories: high and low levels. This is seen in the global bimodal distribution of methylation levels, and when one observes profiles of DNA methylation in a genome browser. There are special cases of intermediate methylation levels, for example cancers have large domains of partial methylation. In addition, imprinted loci have intermediate

methylation (one allele methylated through an imprinting control region in most somatic cells). However, for the vast majority of the sites, major phenotypic differences among healthy cells usually involve methylation changing from low to high, or from high to low. For this reason, all our modeling is in terms of low and high methylation states, and we use the corresponding state space $\{0, 1\}$. This allows for a distribution of the observed levels associated with the "low" methylation state, and another distribution for the observed levels at sites occupied by the "high" state.

Although this restriction to a discrete state space is justified, our modeling approach still allows for an interpretation that is consistent with the fact that methylation is usually measured as levels between 0 and 1. In our modeling we ultimately make use of probabilities over the state space, which behave very much like a continuous level. Given a probability for a site occupying a "high" methylation state, the expected methylation level can be obtained by using that probability as the weight in a convex combination of the expected methylation levels associated with the high and low states. At the same time, the discrete state space is highly convenient, and the preponderance of evidence indicates that in most cases the state carries almost the same information as the level (and may be more robust to artifacts, noise and sampling error).

Phylo-epigenetic model with independent sites

Our goal is to model the evolution of DNA methylation states in multiple extant species with a common ancestor. We first focus on a single-site to derive the likelihood function, and then extend the model to multiple sites but for which epigenomic evolution is assumed to be independent.

We assume that the methylation state at a single CpG site evolves according to a two-state continuous-time Markov process. Let $\pi = (\pi_0, \pi_1)$ be the initial distribution of the methylation state at the root node, and let the transition rate matrix be

$$Q = \begin{bmatrix} -\lambda & \lambda \\ \eta & -\eta \end{bmatrix}.$$

The transition probability matrix between two time points separated by time interval of length ℓ is $P(\ell) = \exp(Q\ell)$, where

$$\exp(Q\ell) = \sum_{k=0}^{\infty} \frac{1}{k!} \ell^k Q^k$$

is the matrix exponential. $P(\ell)$ is determined by two terms $(\ell(\lambda + \eta), \lambda/\eta)$. Let $\lambda + \eta = 1$, so that the mutation rate and branch length parameters are identifiable.

Let $\tau = \{\mathcal{V}, \mathcal{E}\}$ be the phylogenetic tree with known topology and unknown branch lengths, where $\mathcal{V}$ is the set of known vertices and $\mathcal{E}$ is the set of branches with unknown lengths. The model parameter space is thus

$$\Theta = \{\tau, \pi, Q\}.$$

Model inference with complete leaf observation With complete observations at leaf nodes, we use maximum likelihood method to estimate model parameters. Before writing out the observed data likelihood, we require some notation for representing nodes and their relationships in the tree. We use r to denote the root node of the phylogenetic tree, and use u, v and c to denote 3 consecutive nodes on a lineage, such that $(u, v) \in \mathcal{E}$ and $(v, c) \in \mathcal{E}$. Let ℓ_v be the length of edge $(u, v) \in \mathcal{E}$. Each node $u \in \mathcal{V}$ is associated with a random variable for methylation state. For simplicity of notation, we also use u to denote this random variable. We use j and k to denote methylation states, $j, k \in \{0, 1\}$. In general, for a parent-child pair, we associate j with the parent, and k with the child. We use $\mathcal{L}$ to denote the set of all leaf nodes, $\mathcal{I}$ for all internal nodes, and therefore $\mathcal{V} = \mathcal{I} \cup \mathcal{L}$. Let random variable Y be the methylation states at all nodes in the

tree, Z be the states at all internal nodes, and X be the states at all leaf nodes. We use $X(u)$ to denote the set of methylation states at leaf nodes that are descendants of u, and it follows that $X(r) = X$.

For node v, given that its parent u has methylation state j, the conditional probability of observing states $X(v)$ at terminal descendants of node v is

$$p_j(v) = \Pr(X(v)|u = j, \Theta).$$

For notational convenience we define

$$q_k(v) = \begin{cases} \Pr(v = k) & \text{if } v \text{ is a leaf node,} \\ \prod_{c \in \text{child}(v)} p_k(c) & \text{otherwise,} \end{cases}$$

where $\Pr(v = k) \in \{0, 1\}$ when a binary methylation state is observed, and $\Pr(v = k) \in (0, 1)$ when the observed data are continuous levels representing a probability distribution over the state space.

We can then write the probability $p_j(v)$ as the recurrence

$$p_j(v) = \sum_k \Big(P(\ell_v)_{jk} \times q_k(v) \Big). \tag{S1}$$

The likelihood of the observed data for a single site is then

$$L(\Theta|X) = \Pr(X(r)|\Theta) = \sum_{j \in \{0,1\}} \pi_j p_j(r).$$

The recurrence in (S1) is the basis of Felsenstein's pruning algorithm for efficiently computing the likelihood of a tree topology, branch lengths and transition rate, given data at leaf nodes (Felsenstein 1981).

Moving from single site to multiple sites, let N be the total number of sites in the methylome. Let X_n, for $1 \leq n \leq N$, be the set of methylation states associated with all leaf nodes at site n. The variables $X = X_1, \ldots, X_N$ denote the observed methylation states at all leaf nodes. Under the assumption that methylation states at distinct sites evolve independently, the likelihood for observed data at multiple sites is

$$L(\Theta|X) = \textstyle\prod_{n=1}^{N} L(\Theta|X_n).$$

The likelihood and partial derivatives are recursively computed in the same spirit of the pruning algorithm (Felsenstein 1981). In our implementation, we optimize parameters using gradient ascent, and can accurately estimate model parameters in simulated datasets. With the maximum likelihood estimates (MLE) of model parameters, we then compute the joint and marginal posterior probabilities of the states at internal nodes at each site. This can be achieved through a dynamic programming algorithm with time complexity linear to the number of nodes in the tree (Pupko et al. 2000).

Estimating transition rates and branch lengths under can be done with existing software, such as the `RPHAST` R package (Hubisz et al. 2011). However, to estimate methylation probabilities at ancestral species in the phylogenetic tree, we implemented this independent-site model in our `Epiphyte` package.

Phylo-epigenetic model with interdependent sites

In this section, we first introduce our model for epigenome evolution with dependence between neighboring sites. Then, we discuss its relationship with alternative models for epigenome evolution, including the context-dependent model introduced for DNA/RNA sequence evolution (Siepel & Haussler 2004), and an ideal graphical model for epigenome evolution.

Phylo-eipgenetic model with interdependent sites We develop a model to allow for two processes that jointly describe the observed mammalian methylome: one is the process of methylation state inheritance from ancestral species, and the other is the process governing the correlation observed between methylation states at neighboring sites within a species. As in the independent-site model, we use $\tau = \{\mathcal{V}, \mathcal{E}\}$ to denote the phylogenetic tree relating extant species, with known topology and unknown branch lengths. The inheritance process is defined by Q, as previously introduced. At the root species, the dependence between neighboring CpG sites is described with a discrete-time Markov chain over the state space $\{0, 1\}$, where the time points correspond to CpG sites in the genome. Let the initial distribution be π in the root species. Let the transition probability matrix between neighboring CpG sites of the root species be

$$F = \begin{bmatrix} f_0 & 1 - f_0 \\ 1 - f_1 & f_1 \end{bmatrix}.$$

In non-root species, we also use a transition probability matrix to describe the autocorrelation relationship of CpGs within individual methylomes. The transition probability matrix

$$G = \begin{bmatrix} g_0 & 1 - g_0 \\ 1 - g_1 & g_1 \end{bmatrix}$$

is assumed to be homogeneous in all non-root species. We use different horizontal transition probability matrices for the root species and non-root species because the horizontal process at non-root species interacts with the vertical inheritance process to determine the genomic distribution of methylation states, while the the horizontal process at root species alone determines the methylation states in the genome. The model parameter space is now

$$\Theta = \{\tau, \pi, Q, F, G\}.$$

We assume the CpG sites are ordered from 1 to N by their position within the reference human genome, ignoring chromosome boundaries for simplicity. Consider two neighboring CpG sites $n-1$ and n ($1 < n \leq N$) and a branch $(u, v) \in \mathcal{E}$ in the phylogenetic tree, with u being the parent of v. Let random variable v_n denote the methylation state of site n in species v. We assume the conditional probability of methylation state k is proportional to $G_{ik}P(\ell_v)_{jk}$, and thus define the conditional distribution of v_n, given the previous site's state v_{n-1} and its ancestral state u_n, as:

$$\begin{aligned} p_v(i, j, k) &= \Pr(v_n = k | v_{n-1} = i, u_n = j) \\ &= \frac{G_{ik}P(\ell_v)_{jk}}{\sum_{k'=0,1} G_{ik'}P(\ell_v)_{jk'}}, \text{ where } i, j, k \in \{0, 1\} \text{ and } 1 < n \leq N. \quad \text{(S2)} \end{aligned}$$

This conditional probability distribution models the majority of sites, but there are important special cases: (i) the states of all nodes at position $n = 1$, and (ii) the states of all sites at the root node r. For sites at position $n = 1$, the methylation state evolution is described by the initial distribution π, the continuous-time Markov process with transition rate matrix Q, and the branch lengths in the phylogenetic tree τ. The methylome of the root species is modeled only with the discrete-time Markov chain F and the initial distribution π.

The parameter interpretations of Q and branch lengths under the independent-site model can not be applied to the interdependent-site model. To scale branch lengths to have unit length represent approximately 1 expected methylation state change, we compute a scaling factor in the following way. Given the model parameters, the methylation probability averaged along the whole methylome of the LCA is $p_1 = (1 -$

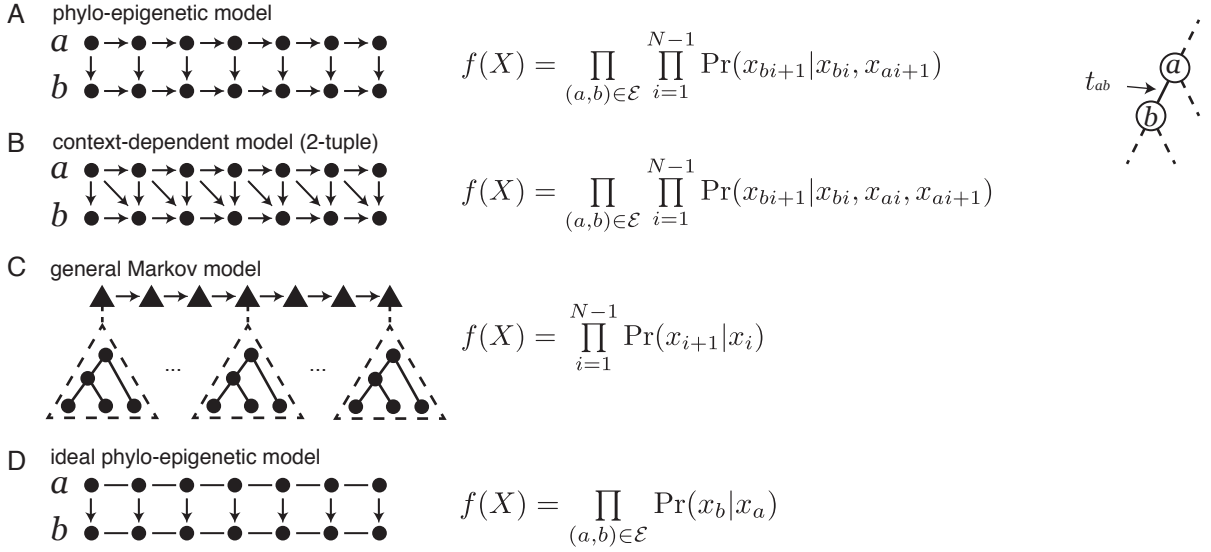

Figure S10: Graphical models for epigenome evolution and main factors in probability mass function factorization. Species a is the parent of species b in the phylogenetic tree. Circles represent methylation states at individual sites of individual species. Triangles represent tree-wise methylation states at individual sites. Vertex set x_i contains all vertices at site i. In the chain graph model, x_a, x_b are the vertex sets corresponding to the chain components for species a and b, containing all sites in each species.

$f_0)/(2 - f_0 - f_1)$. The expected frequency of two neighboring sites having both hypomethylated states is $p_{00} = (1 - p_1)f_0$, both methylated states $p_{11} = p_1 \cdot f_1$, and different methylation states $p_{01} = p_{10} = (1 - p_1) \cdot (1 - f_0)$. We construct a scaling factor:

$$\beta = p_{00} \cdot \frac{1 - g_0}{g_0} \cdot \lambda + p_{10} \cdot \frac{g_1}{1 - g_1} \cdot \lambda + p_{01} \cdot \frac{g_0}{1 - g_0} \cdot (1 - \lambda) + p_{11} \cdot \frac{1 - g_0}{g_1} \cdot (1 - \lambda).$$

Multiplying β with raw branch lengths approximately scales the unit length to represent 1 expected methylation state change per site.

Alternative models for epigenome evolution To model epigenome evolution, the independent-site assumption needs to be relaxed because of the well characterized autocorrelation of epigenetic marks, especially when modeling at high-resolution, focusing on individual sites. Introducing dependence between neighboring sites as we described above is a major improvement over independent-site models. It not only captures the biological autocorrelation, also permits effective means of solving the missing data problem due to the substantial divergence of CpG sites between species.

In the context of genome evolution, sequence context-dependent mutation rates have been incorporated to extend the standard phylogenetic models. The standard phylogenetic model and many of its extensions can be considered as graphical models. Siepel & Haussler (2004) introduced a context-dependent phylogenetic model to allow mutation rates to vary depending on the identity of neighboring nucleotides. This model is equivalent to a Bayesian network where factors in the density function are conditional probabilities of a descendant site given its parent site, previous parent site and its previous site (Figure S10B). The conditional distributions are derived from a continuous-time Markov model for dinucleotide evolution. Computing the

data likelihood of observations at leaf species of the phylogenetic tree is intractable. Several approximation methods have been applied (Jensen & Pedersen 2000; Lunter & Hein 2004; Siepel & Haussler 2004; Jojic et al. 2004). For example, Siepel & Haussler (2004) originally ignored the dependencies between ancestral states of neighboring sites and simplified likelihood computation to that of an (k-1)st order Markov chain for sequences in extant species when the substitution process is jointly modeled for k-tuples; Jojic et al. (2004) used structured variational methods to approximate the data-likelihood, and showed that preserving the phylogenetic tree structure at individual sites provides a better approximation.

Our model for methylome evolution also defines a Bayesian network, where vertical edges exist between a site in one species and the homologous site in the parent species, and describe the causal relationship between ancestor and descendant methylation states. Inter-site dependence is modeled with directed edges between neighboring sites within the same species. The probability mass function can be factorized into conditional probabilities of a descendant site given its previous site and its ancestral site (Figure S10A). Our formulation of the conditional probability distribution offers flexibility at combining the horizontal dependence and vertical inheritance relationships. For example, when $g_0 = g_1 = 0.5$, the methylation state evolution processes at different sites are independent, while the states at neighboring sites in a non-root species are still correlated as a result of the correlation of the homologous sites in the root species. When the rows of F are the initial distribution π, in addition to $g_0 = g_1 = 0.5$, this special case is equivalent to the independent-site model. Potential extensions can be made to the modeling of horizontal processes to capture more characteristics of the methylome and its evolution dynamics, which we did not pursue in this study. For example, the horizontal process can be modeled with a continuous time Markov chain to capture the inverse-relationship between pair-wise methylation state correlation and inter-site distance. In addition, different horizontal processes can be assumed for individual non-root species to capture lineage-specific property of local correlation.

Both of these two Bayesian network models are special cases of a general Markov model for transitions between neighboring sites over tree-wise epigenomic states (Figure S10C). The transition probability between tree states can be expressed by the product of local conditional probabilities at individual nodes of the phylogenetic tree. Compared with a general Markov model, these two Bayesian network models are expressive of the phylogenetic relationship between species and dramatically reduce the number of parameters involved in the tree-state transition probabilities from $O(2^{2|\mathcal{V}|})$ to $O(|\mathcal{V}|)$.

Ideally, the epigenome evolution can be modeled with a chain graph (Lauritzen 1996). Each node in the graph corresponds to the methylation state of a site in a species. Undirected edges exist between neighboring sites within the same species to describe the autocorrelation of methylation states. Directed edges link each site in each internal species to its descendants in the child species, and describe the causal relationship between parent and descendant sites. A generative model for such a chain-graph structure would involve dynamic processes that generate samples iteratively until convergence to some kind of equilibrium. In the context of modeling epigenome evolution, for example, this equilibrium may be measured by the distribution of the number and sizes of HMRs. This iterative dynamic process allows a descendant site to see past the immediate neighboring sites and its direct parent site in the chain graph and incorporate information from much wider range of sites in both the parent species and the descendant species. Compared with the generative process for Bayesian network models based on conditional probabilities, a generated chain-graph sample is a more realistic picture of epigenome evolution. However, factorization of a probability function on this graph involves potentials over all complete subsets of the complete graph induced by all nodes in a chain component (Lauritzen 1996). In contrast with Bayesian network models, even computing complete data likelihood for this chain-graph is intractable because each chain component may inflict state space of size exponential to the number of sites in the genome.

Model learning and inference for phylo-epigenetic model with interdependent sites

Complete data likelihood and sufficient statistics Assume the complete-data Y – methylation states at every site $n \in \{1, \ldots, N\}$ and in every species $v \in \mathcal{V}$ – are observed, including ancestral species. For convenience below we will use the symbol v_n, which we previously defined as a random variable, to denote the observed state of that variable. With model parameters Θ, the complete data likelihood is

$$L(\Theta|Y) = \Pr(Y|\Theta) = \Pr(Y_1|\Theta) \prod_{n=2}^{N} F_{r_{n-1}r_n} \prod_{u \in \mathcal{I}} \prod_{v \in \text{child}(u)} p_v(v_{n-1}, u_n, v_n).$$

With u denoting the parent of v and r denoting the root, we define the following:

$$\begin{aligned} w_j &= \mathbb{1}\{r_1 = j\} && \text{(corresponds to site 1 in root species)} \\ w_{jk}(v) &= \mathbb{1}\{u_n = j, v_n = k\} && \text{(site 1 in non-root species)} \\ w_{ik} &= \textstyle\sum_{n=2}^{N} \mathbb{1}\{r_{n-1} = i, r_n = k\} && \text{(remaining root sites)} \\ w_{ijk}(v) &= \textstyle\sum_{n=2}^{N} \mathbb{1}\{v_{n-1} = i, u_n = j, v_n = k\} && \text{(all other sites)} \end{aligned} \tag{S3}$$

The log-likelihood can then be written as

$$\log L(\Theta|Y) =$$

$$\sum_{\substack{v=r \\ j \in \{0,1\}}} w_j \log \pi_j \; + \sum_{\substack{v \neq r \\ j,k \in \{0,1\}}} w_{jk}(v) \log P(\ell_v)_{jk} \; + \sum_{i,k \in \{0,1\}} w_{ik} \log F_{ik} \; + \sum_{\substack{v \neq r \\ i,j,k \in \{0,1\}}} w_{ijk}(v) \log p_v(i,j,k). \tag{S4}$$

Therefore, the following are sufficient statistics for our model parameters:

$$W = \big\{w_j, w_{ik}, w_{jk}(v), w_{ijk}(v) : i, j, k \in \{0,1\}, v \in \mathcal{V} \setminus \{r\}\big\}.$$

Among these statistics, $\{w_{ijk}\}$ represent the vast majority of information from the data, while $\{w_j\}$ and $\{w_{jk}\}$ carry negligible information. Under complete data, we can use these sufficient statistics to efficiently compute the MLE for model parameters by numerical methods. As with the independent-site model, our implementation also uses gradient ascent to arrive at the MLE of parameters in the interdependent-site model.

Learning model with EM algorithm from incomplete data In reality, observations about DNA methylation states are only available at a subset of sites in extant species. Given the model parameter θ and observed states X, we can obtain expected values of the sufficient statistics $\mathrm{E}_{Z|X,\theta} W$. Given the sufficient statistics in (S3), we can derive the MLE of model parameters by maximizing the complete-data likelihood in (S4). These two processes are exactly the two steps in an expectation-maximization (EM) algorithm. In the expectation (E) step, we compute $Q(\theta|\theta^{(t)})$, which is defined as the expected value of the complete-data log-likelihood $\log L(\theta|X, Z)$ with respect to the unknown data Z given the observed data X and parameter

estimates $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{Z|X,\theta^{(t)}} \log L(\theta|X,Z) =$$

$$\begin{aligned} &\sum_{\substack{v=r \\ j\in\{0,1\}}} \mathrm{E}(w_j|X,\theta^{(t)}) \log \pi_j \quad + \sum_{\substack{v\neq r \\ j,k\in\{0,1\}}} \mathrm{E}(w_{jk}(v)|X,\theta^{(t)}) \log P(\ell_v)_{jk} \; + \\ &\sum_{i,k\in\{0,1\}} \mathrm{E}(w_{ik}|X,\theta^{(t)}) \log F_{ik} \; + \sum_{\substack{v\neq r \\ i,j,k\in\{0,1\}}} \mathrm{E}(w_{ijk}(v)|X,\theta^{(t)}) \log p_v(i,j,k). \end{aligned} \tag{S5}$$

The E-step is equivalent to computing the expected value of $W|X,\theta^{(t)}$ as explained below. The maximization (M) step can be expressed as:

$$\theta^{(t+1)} = \arg\min_{\theta} Q(\theta|\theta^{(t)}).$$

E-step and Markov chain Monte Carlo approximation In the E-step, we use Markov chain Monte Carlo (MCMC), specifically Gibbs sampling, to estimate the expectation of the sufficient statistics given the model parameters and observations at leaf nodes. Recall that we used Z to denote (unobserved) methylation states at internal nodes in previous sections, and assumed that observation at leaf nodes are complete. Here we describe the algorithm for a more general situation, where observations at a number of leaf nodes may be missing at a number of sites. Hereafter, we use Z to denote the methylation states that are not observed, including those at all internal species and any leaf species with missing data at some sites. We use X to denote all the observed methylation states, which must come from leaf nodes. We apply MCMC to approximate the conditional distribution of $\Pr(Z|X,\theta)$. Let $Z^{(t)}$ be the t^{th} sample in the chain. Let $\boldsymbol{MB}(v)$ be the Markov blanket of node v in the Bayesian network. Given the states of all variables in $\boldsymbol{MB}(v)$ as b, the conditional distribution $\Pr(v = i|\boldsymbol{MB}(v) = b)$ can be computed easily (see Table S14).

Let $\boldsymbol{MB}(v,t)$ denote the states of nodes in $\boldsymbol{MB}(v)$ in $Z^{(t)}$. The sampling procedure for $Z^{(t)}$ is as follows:

1. Draw a starting sample $Z^{(t=0)}$, which is a specific initiation of the states for all nodes with unobserved methylation states.

2. For $t = 1, 2, \ldots$:
   Iterate over all sites from site 1 to site N. At each site, iterate over the nodes of the phylogenetic tree according to a post-order traversal. If the node has an unobserved methylation state $z \in Z$:

   (a) Make a proposal to change the state of z to: $z^{\text{prop}} = 1 - z^{(t-1)}$.

   (b) Accept the proposal $z^{(t)} = z^{\text{prop}}$ with probability

   $$\alpha = \Pr(z^{\text{prop}}|\boldsymbol{MB}(z,t-1)).$$

   (c) If the proposal is rejected, let $z^{(t)} = z^{(t-1)}$.

The chain $\{Z^{(t)}\}$ is positive recurrent and aperiodic on a finite state space. It follows that the chain is uniformly ergodic (Roberts & Polson 1994). The same are true for the chain of sample statistics $\{W^{(t)}\}$ derived from $\{Z^{(t)}\}$. Theories about MCMC (Theorem 3.1 of Gilks et al. (1995)) guarantees that

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T} W^{(t)} \to \mathrm{E}_{Z|X,\theta} W\right) = 1.$$

Our goal in the E-step is to approximate $\mathrm{E}_{Z|X,\theta}W$ using MCMC samples. For any scalar sample statistic $w \in W$, the output from MCMC is summarized in terms of ergodic averages of the form

$$\bar{w}_t = \frac{\sum_{i=1}^{t} w^{(i)}}{t}.$$

The Markov chain central limit theorem states that

$$\sqrt{t}(\bar{w}_t - \mathrm{E}_{Z|X,\theta}[w]) \xrightarrow{d} N(0, \sigma_w^2), \text{ as } t \to \infty,$$

where σ_w is a constant given X and θ. Given an estimate of σ_w, we can form a confidence interval for $\mathrm{E}_{Z|X,\theta}[w]$, and continue sampling until the confidence interval is sufficiently small. We adopt a fixed-width stopping rule for the Markov chain based on a consistent batch means method, i.e. the CBM method referred to by Jones et al. (2006). In each E-step, let t be the current chain length. Let the batch size b_t, and batch number a_t be functions of t:

$$b_t = \left\lfloor t^{\frac{1}{2}} \right\rfloor, a_t = \lfloor t/b_t \rfloor .$$

Let $\bar{w}_j$ be the mean of the $j\mathrm{th}$ batch:

$$\bar{w}_j = \frac{1}{b_t} \sum_{i=(j-1)b_t+1}^{jb_t} w^{(i)}, \quad \text{for } j = 1, \ldots, a_t.$$

The CBM estimate of σ_w^2 is

$$\hat{\sigma}^2_{w,\mathrm{CBM}} = \frac{b_t}{a_t - 1} \sum_{j=1}^{a_t} (\bar{w}_j - \bar{w}_t)^2.$$

We stop sampling the first time that

$$t_\star \frac{\hat{\sigma}_{w,\mathrm{CBM}}}{\sqrt{t}} + p(t) < \epsilon, \text{ for all } w \in W,$$

where $t_\star$ is the appropriate quantile of Student's t-distribution with $a_t - 1$ degree of freedom, and $p(t) = \epsilon \times \mathbb{1}\{t \leq t_{min}\}$, where t_{min} is a minimum sampling effort (for example 100). In our implementation, we chose $\epsilon = N/10^4$ as the half 95%-confidence interval width cutoff. Then, $\mathrm{E}_{Z|X,\theta^{(t)}}W$ is approximated with the sample average of the second half of the chain, which we denote with $\tilde{W}$.

M-step Let $\tilde{Q}(\theta|\theta^{(t)})$ be an approximation to $Q(\theta|\theta^{(t)})$ with $\mathrm{E}_{Z|X,\theta^{(t)}}W$ substituted by the MCMC approximation $\tilde{W}$. We use gradient ascent to maximize

$$\begin{aligned}\tilde{Q}(\theta|\theta^{(t)}) = &\sum_{\substack{v=r \\ j\in\{0,1\}}} \tilde{w}_j \log \pi_j \quad + \sum_{\substack{v\neq r \\ j,k\in\{0,1\}}} \tilde{w}_{jk}(v) \log P(\ell_v)_{jk} \; + \\ &\sum_{i,k\in\{0,1\}} \tilde{w}_{ik} \log F_{ik} \; + \sum_{\substack{v\neq r \\ i,j,k\in\{0,1\}}} \tilde{w}_{ijk}(v) \log p_v(i,j,k).\end{aligned}$$

Partial derivatives required for parameter optimization are shown in Table S15.

Summary of model inference procedure Together, the procedure for model parameter estimation and ancestral state reconstruction is summarized as below:

1. Choose start point for model parameters $\theta^{(t)}$.
2. Iterate the following EM procedure
   - E-step: Use Gibbs sampling to approximate $\mathrm{E}_{Z|X,\theta^{(t)}} W$.
   - M-step: update model parameters to $\theta^{(t+1)}$.

   until convergence: $\|\theta^{(t+1)} - \theta^{(t)}\| < \epsilon$.
3. Given the final model parameter estimates, generate MCMC samples as in the E-step. After discarding the first half of samples, construct marginal posterior distribution with the second half of samples at individual sites. Use the MAP estimates as the methylation states at individual sites.

Application on methylome data The input data for a single-CpG resolution phylo-epigenetic model are the posterior methylation probabilities at individual CpG sites, which were estimated in their native genome with the `hmr` program in `MethPipe` package (Song et al. 2013).

We applied the interdependent-site phylo-epigenetic model on sperm methylomes of the seven species at single CpG resolution. Sites in species without observed data were treated as missing data. We separated the orthologous methylome into 30 equal-sized regions, and estimated parameters from individual genomic fragment with the `epiphy-est` program inside our `Epiphyte` package. The estimated parameters are stable across different genomic fragments. The median of parameter estimates are reported in Supplemental Table S8, and the scaled evolutionary tree for methylomes is shown in Fig. 2A. We estimated that $g_0 = 0.99 > g_1 = 0.89$, which indicates a stronger force to maintain consecutive hypomethylated sites than consecutive methylated sites.

We also applied the independent-site model at single CpG resolution after imputing methylation probabilities at most sites with missing data, to provide extra validation for the expansion of hypomethylation fraction. Sites in species without observed data were interpolated with the average methylation probabilities at the two closest (within 1 kbp distance up and downstream) observed sites if the two sites belong to the same HMR or nonHMR in native genome, and treated as missing data otherwise. Although the data interpolation step violates the independent site assumption, it is a reasonable step because of the empirically observed spatial correlation of methylation states in the genome. The model is estimated with `epiphy-est` program with option `-d 0`, which sets CpG desert size cut off at 0 and effectively treats all sites as evolving independently. Estimated model parameters are shown in Supplemental Table S8. The fraction of sites hypomethylated was estimated to be 13.6% for the LCA, while it is between 15.6-19.6% for extant species.

The inference step estimates the posterior methylation probabilities at unobserved sites, and can be separated from model learning. We can estimate ancestral methylation probabilities using the MCMC sampling method as described above, with any given set of parameters for the interdependent-site phylo-epigenetic model. With the posterior probabilities, we identified hypomethylated regions in ancestral methylomes by collapsing neighboring sites with less than 0.5 posterior methylation probability. We observed that methylation states inferred using parameter estimates from the interdependent-site model are not parsimonious in some cases. Although parsimony reconstruction of ancestral states may not be the most likely reconstruction under a likelihood-based method, we tried to reach a balance between these two. We keep the root horizontal parameters F estimated by the interdependent-site model, and fixed the phylogenetic tree branch lengths and methylation mutation rate parameters to the values estimated by the independent-site model.

The descendant horizontal process parameters estimated by the interdependent-site model are $g_0 = 0.99$ and $g_1 = 0.89$. We experimented with different values within this range (0.89, 0.99), and chose 0.95 as the value for both g_0 and g_1, so that the ancestral methylation state estimates are parsimonious given the input extant species methylation states showing clade-specific methylation patterns. Hypomethylated regions in ancestral species, as well as all types of methylation evolution events were identified based on the posterior methylation probabilities estimated in this parameter setting with programs `epiphy-post` and `epiphy-seg`.

If the current node $v = r$ is the root:

$n = 1$	$\boldsymbol{MB}(v_n) = \{c_n : c \in \text{child}(v)\} \cup \{v_{n+1}\}$ $\Pr(v_n = i\|b) \propto \prod_{c\in\text{child}(v)} \Pr(c_n\|v_n = i)$
$n = N$	$\boldsymbol{MB}(v_n) = \{c_{n-1}, c_n : c \in \text{child}(v)\} \cup \{v_{n-1}\}$ $\Pr(v_n = i\|b) \propto \Pr(v_n = i\|v_{n-1}) \prod_{c\in\text{child}(v)} p_c(c_{n-1}, i, c_n)$
$1 < n < N$	$\boldsymbol{MB}(v_n) = \{c_{n-1}, c_n : c \in \text{child}(v)\} \cup \{v_{n-1}, v_{n+1}\}$ $\Pr(v_n = i\|b) \propto \Pr(v_n = i\|v_{n-1}) \Pr(v_{n+1}\|v_n = i) \prod_{c\in\text{child}(v)} p_c(c_{n-1}, i, c_n)$

If the current node $v \in \mathcal{L}$ is a leaf:

$n = 1$	$\boldsymbol{MB}(v_n) = \{v_{n+1}\} \cup \{u_n, u_{n+1}\}$ $\Pr(v_n = i\|b) \propto \Pr(v_n = i\|u_n) p_v(i, u_{n+1}, v_{n+1})$
$n = N$	$\boldsymbol{MB}(v_n) = \{v_{n-1}\} \cup \{u_n\}$ $\Pr(v_n = i\|b) \propto p_v(v_{n-1}, u_n, i)$
$1 < n < N$	$\boldsymbol{MB}(v_n) = \{v_{n-1}, v_{n+1}\} \cup \{u_n, u_{n+1}\}$ $\Pr(v_n = i\|b) \propto p_v(v_{n-1}, u_n, i) p_v(i, u_{n+1}, v_{n+1})$

If the current node $v \notin \mathcal{L} \cup \{r\}$ is internal:

$n = 1$	$\boldsymbol{MB}(v_n) = \{c_n : c \in \text{child}(v)\} \cup \{v_{n+1}\} \cup \{u_n, u_{n+1}\}$ $\Pr(v_n = i\|b) \propto \Pr(v_n = i\|u_n) p_v(i, u_{n+1}, v_{n+1}) \prod_{c\in\text{child}(v)} \Pr(c_n = i\|v_n = u)$
$n = N$	$\boldsymbol{MB}(v_n) = \{c_n, c_{n-1} : c \in \text{child}(v)\} \cup \{v_{n-1}\} \cup \{u_n\}$ $p_b(v_n) = p_{v_{n-1}} p_{u_n} \prod_{c\in\text{child}(v)} p_{c_{n-1}} p_{c_n}$ $\Pr(v_n = i\|b) \propto p_v(v_{n-1}, u_n, i) \prod_{c\in\text{child}(v)} p_c(c_{n-1}, i, c_n)$
$1 < n < N$	$\boldsymbol{MB}(v_n) = \{c_n, c_{n-1} : c \in \text{child}(v)\} \cup \{v_{n-1}, v_{n+1}\} \cup \{u_n, u_{n+1}\}$ $\Pr(v_n = i\|b) \propto p_v(v_{n-1}, u_n, i) p_v(i, u_{n+1}, v_{n+1}) \prod_{c\in\text{child}(v)} p_c(c_{n-1}, i, c_n)$

Table S14: The Markov blanket and probabilities involved in MCMC sampling procedure. These are broken down for each of the 9 separate cases involving combinations of {root, leaf, internal} nodes and {first, internal, last} sites. In each expression where it appears, $i \in \{0, 1\}$ is the methylation state of v_n, $b \in B(v_n)$ is the joint state set for $\boldsymbol{MB}(v_n)$, the indicated Markov blanket. In our notation, the node u is the parent of node v.

$$
\begin{aligned}
\frac{\partial \tilde{Q}}{\partial \pi_0} &= \frac{\tilde{w}_0}{\pi_0} - \frac{\tilde{w}_1}{1-\pi_0} \\
\frac{\partial \tilde{Q}}{\partial f_i} &= \frac{\tilde{w}_{ii}}{f_i} - \frac{\tilde{w}_{i(1-i)}}{1-f_i}, \quad i = 0, 1 \\
\frac{\partial \tilde{Q}}{\partial g_i} &= \sum_{\substack{v \neq r \\ j,k \in \{0,1\}}} \tilde{w}_{ijk}(v) \left\{ \frac{(-1)^k}{G_{ik}} - \frac{\sum_{k' \in \{0,1\}} (-1)^{k'} P(\ell_v)_{jk'}}{\sum_{k' \in \{0,1\}} G_{ik'} P(\ell_v)_{jk'}} \right\}, \quad i = 0, 1 \\
\frac{\partial \tilde{Q}}{\partial \lambda} &= \sum_{\substack{v \neq r \\ j,k \in \{0,1\}}} \tilde{w}_{jk}(v) \frac{(-1)^{1-k} T_v}{P(\ell_v)_{jk}} \quad + \\
&\quad \sum_{\substack{v \neq r \\ i,j,k \in \{0,1\}}} \tilde{w}_{ijk}(v) T_v \left\{ \frac{(-1)^{1-k}}{P(\ell_v)_{jk}} - \frac{\sum_{k' \in \{0,1\}} G_{ik'} (-1)^{1-k'}}{\sum_{k' \in \{0,1\}} G_{ik'} P(\ell_v)_{jk'}} \right\} \\
\frac{\partial \tilde{Q}}{\partial T_v} &= \sum_{\substack{v \neq r \\ j,k \in \{0,1\}}} \tilde{w}_{jk}(v) \frac{\lambda^{1-j} (\lambda - 1)^j (-1)^{1-k}}{P(l_v)_{jk}} \quad + \\
&\quad \sum_{\substack{v \neq r \\ i,j,k \in \{0,1\}}} \tilde{w}_{ijk}(v) \lambda^{1-j} (\lambda - 1)^j \left\{ \frac{(-1)^{1-k}}{P(l_v)_{jk}} - \frac{\sum\limits_{k' \in \{0,1\}} G_{ik'} (-1)^{1-k'}}{\sum\limits_{k' \in \{0,1\}} G_{ik'} P(\ell_v)_{jk'}} \right\}, \\
&\quad \text{where } T_v = 1 - \exp(-l_v), \;\; v \in \mathcal{V} \setminus \{r\}
\end{aligned}
$$

Table S15: Partial derivatives with respect to model parameters required in the maximization step of EM algorithm (Section).

Sperm methylome evolution at well-conserved elements

The placental mammal conserved elements (Siepel et al. 2005) are highly fragmented, with median region size 16 bp, and a substantial fraction (70%) of these elements do not contain any CpG sites from the 7-way orthologous genome. We restricted analysis to the 369,007 placental mammal conserved elements with total size 25.7 Mbp that have observed methylation data from all seven species, and applied our interdependent-site phylo-epigenetic model (Supplemental Table S8).

Sperm methylome evolution with mouse as reference species

We examined the impact of choice of reference genome on phylo-epigenetic analyses by using mouse as the reference species. We used multiple genome alignment of 59 vertebrate genomes with mouse (Blanchette et al. 2004) (available from UCSC Genome Browser) to map CpG locations in the other 6 sperm methylomes onto the mouse mm10 reference genome, following the same procedure described for human-referenced alignment. The resulting 7-way orthologous regions have a total size of 277 Mbp in the mouse mm10 reference genome.

Hypomethylation expansion

The trends we observe are genome-wide averages, and although we attempted to analyze specific sequence changes associated with widening, we have so far not seen any obvious connections with specific types of sequence changes (other than what we have already described). If we focus on individual examples, we can find some where a progressive expansion seems to be taking place along an individual lineage. Our modeling suggests the existence of HMRs that have widened on parallel lineages. But when we examine any individual HMR in most cases we do not have enough information (without many additional species) to distinguish parallel widening vs. ancestral widening or contraction along another lineage. We can only speak to the averages indicated by our modeling and our supporting analyses (e.g. Figure 3A). We favor the view that any individual HMR almost certainly has a "stable" size for most of its evolution, with widening likely the result of discrete evolutionary events. Any events that would have made an HMR more narrow seem to be comparatively rare. This leads to widening of HMRs on average (a) over the genome, (b) across lineages, and (c) over time. Although we can detect and measure these averages, isolating the individual events is more difficult and will require more data.

The sizes of newly arising HMRs can give insight on the birth size and expansion rate. The difference in discernible age matters when we consider the size distribution of HMRs. For example, the HMR births inferred on the human branch formed after the human-chimp divergence, and the mouse HMR births formed after the mouse-rat divergence. These two categories of HMR birth events represent different average ages of hypomethylation. The size of an observed HMR birth contains the initial size at birth and subsequent expansion in that species. We compared the size distribution of HMR birth events on the human branch and on the mouse branch. The size of HMR births ranges from hundreds of bases to thousands of bases for both human (standard deviation 372 bp) and mouse (standard deviation 415 bp). Although the average or median mouse HMR birth size is only larger than that of human by 10-30 bp, the difference between two distributions are significant (single-sided Wilcoxon ranksum test $p = 1.11 \times 10^{-8}$). We examined HMR births in chimp and rat as well. The HMR birth sizes in mouse or rat are significantly larger than those in human and chimp. Increasing the species sampling density in the phylogenetic tree to represent more species divergence events along a single lineage will help calibrate HMR expansion rates along individual lineages.

Bisulfite sequencing does not distinguish between methylation (5mC) and hydroxymethylation (5hmC) (Jin et al. 2010). Divergent 5hmC levels could lead to observed methylation (5hmC + 5mC) divergence between species. The modification 5-hydroxymethylcytosine (5hmC) is at low abundance when it is detected and is probably short-lived – it is a transient state for which detection at low levels indicates actively maintained hypomethylation. Such active maintenance will only be observed if there is active *de novo* methylation (otherwise the substrate for hydroxylation would quickly disappear). These dynamics likely play a role at some point in precursors of sperm, but we do not expect to observe it in mature sperm. Knowing the 5hmC distribution in the right precursor cells, for multiple species, would be highly informative.

The ideal type of data for addressing this question would be TAB-seq data from human and mouse sperm, or matched cell types from earlier stages of spermatogenesis. However, to our best knowledge, there are no publicly available TAB-seq data for mammalian sperm. TAB-seq data from human and mouse ESC showed that absolute 5hmC level is almost always less than 0.1 at individual CpG sites (Yu et al. 2012). In human ESC, for example, only 4.0% of CpG sites have greater than 0.1 5hmC level; only 200 CpG sites have greater than 0.5 5hmC estimated level (Yu et al. 2012). If mammalian sperm has similar distribution of 5hmC levels, it is unlikely that the difference in 5hmC has a major contribution to the HMR expansion phenomenon that we observe between species, as the methylation (5hmC + 5mC) levels in these regions in species where the regions are methylated are usually quite high, e.g. above 0.6.

ChIP-seq experiments targeting 5hmC can provide qualitative evidence for the presence of this modification. Two studies used ChIP-seq to profile 5hmC distribution in human and mouse mature sperm, and mouse spermatid and spermatocyte (Wang et al. 2015; Hammoud et al. 2014). We analyzed these data and observed that the read count enrichment in 5hmC peaks is largely independent of regional DNA methylation level (5hmC+5mC), especially in mature sperm. It is likely that 5hmC levels in these regions are uniformly low. In addition, about 40% (28.2%) of promoter HMR extensions in human sperm relative to mouse sperm overlap with mouse sperm 5hmC broad (narrow) peaks. Of these peaks, only 0.1% (14.1%) have above 5 fold read enrichment. Judging from the available information, we think that the phenomenon of promoter HMR size divergence between human and mouse is not due to divergence in 5hmC levels.

Relative sequence substitution rate

Sequence substitution rate varies in different genomic contexts and regions of the genome. This knowledge is the reason we used relative substitution rates in orthologous regions between two parallel lineages. The premise is that matching orthologous pairs of intervals, and comparing their relative substitutions, likely provides some degree of control for the sequence context (gene vs. intergenic), composition (e.g. CpG island or not), and to a lesser degree the expression of genes (depending on the evolutionary closeness of the species, and the function of those genes in spermatogenesis).

To examine whether HMR births located in different genomic contexts share this feature of sequence substitution, we compared HMR births located in intergenic regions, which constitute about 1/3 of all lineage-specific HMR births, to those located in the bodies of genes highly (RPKM$>$1) and lowly (RPKM$\leq$1) expressed in human testis tissue (Melé et al. 2015). Of the total 24 (4 pairs of lineages, 3 genomic contexts) tests, all but one showed significantly increased RSSR in lineage-specific HMR births (permutation test $p < 0.05$).

Enrichment of histone modifications

Human and mouse round spermatid histone modification H3K4me3 and H3K27me3 ChIP-seq data were from the study by Lesch et al. (2016) (GSE68507). Human and mouse sperm H3K4me1 ChIP-seq data

were from studies by Hammoud et al. (2014) (GSE49624) and Jung et al. (2017) (GSE79227). Reads were mapped to respective reference genome (hg19 and mm10) with `Bowtie2` (Langmead et al. 2009). Duplicated reads were removed using `SAMtools` within each sequencing library (Li et al. 2009). Fragment lengths were estimated using the `csaw` Bioconductor package (Lun & Smyth 2016). For H3K4me3 and H3K27me3, enrichment scores were calculated using `deepTools` (Ramírez et al. 2014) as the log2 ratio of number of fragments per 10-bp bin between treatment and input after scaling by sequencing depth. For H3K4me1, we used regional average read coverage to compare enrichment between different regions. The mouse histone mark enrichment scores in 10-bp bins were aligned to the human hg19 coordinates using the UCSC Genome browser utility `liftOver` with option `-minMatch=0.5` (Hinrichs et al. 2006). Regional average enrichment scores were computed with UCSC Genome browser utility `bigWigAverageOverBed` (Kent et al. 2010).

Enrichment of transcription factor binding sites

Human transcription factor binding sites (TFBS) used in our analysis were TFBS clusters (V3) from ENCODE data uniformly processed by the ENCODE Analysis Working Group (Gerstein et al. 2012; Wang et al. 2013a), downloaded from UCSC Genome Browser ENCODE Analysis Hub (hg19) (Raney et al. 2014). We kept all TFBS that overlapped with 7-way orthologous regions and with human sperm HMRs.

For mouse transcription factor binding sites, we downloaded all narrowPeak files for *Mus musculus* transcription factor ChIP-seq experiments from ENCODE covering 46 different transcription factors (mm9) (ENCODE Project Consortium 2012). We also included mouse ESC EZH2 ChIP-seq data from studies by Ku et al. (2008) (GSE13084) and Peng et al. (2009) (GSE18776). Reads were mapped to mouse reference genome (mm10) with bowtie2 (Langmead et al. 2009). Duplicated reads were removed using `SAMtools` within each sequencing library (Li et al. 2009). We used `MACS2` to call broad peaks from each data set with default q-value cutoff (Zhang et al. 2008). The two data set generated similar number of broad peaks (10258 vs 9861). We mapped the binding sites from mm9/mm10 to hg19 using `liftOver` tool (-minMatch=0.5), and only kept binding sites that overlapped with 7-way orthologous regions and with mouse sperm HMRs. Binding sites of the same transcription factor profiled in different cell types were pooled and collapsed.

Gene ontology analyses

Protein-coding genes with TSS located in HMRs in the sperm methylomes of all seven species comprise the background gene list. The subset showing primate-lineage specific promoter HMR extension in human sperm comprise the target gene list. We used `GOrilla` (Eden et al. 2009) to identify enriched biological processes at false discovery rate 0.05. We further removed ontology term redundancy, and visualized the remaining terms in semantic similarity-based scatter plots using `REVIGO` (Supek et al. 2011).

Primate-lineage-specific HMRs are human HMRs in 7-way orthologous genome that contain HMR birth events annotated to the human lineage since the mouse-human common ancestor. Rodent-lineage-specific HMRs are mouse HMRs in 7-way orthologous genome that contain HMR birth events annotated to the mouse lineage since the mouse-human common ancestor. Overlapping HMRs between the two lineages were removed. We established gene-HMR association by annotating an HMR to the closest gene transcription start site. The gene-HMR association for mouse HMRs are established according to the mouse reference genome assembly coordinates after converting HMRs from hg19 to mm10 with `liftOver` tool. The candidate genes for gene-HMR association are orthologous protein-coding genes between human and mouse (Ensembl75) that have gene transcription start sites located within the 7-way orthologous genome. The gene orthologs that are associated with both a human sperm HMR and a mouse sperm HMR located

in the 7-way orthologous genome comprise the background gene list (7293 genes). From this gene list, we further identified the subset that are associated with primate-lineage-specific HMRs (2427 genes), and the subset that are associated with mouse-lineage-specific HMRs (2839 genes). These two subsets have a significant overlap (1518 genes, Fisher exact test $p = 6.49 \times 10^{-202}$). We used PANTHER Classification System (Mi et al. 2016) to find overrepresented PANTHER GO-Slim biological processes associated with the common genes.

References

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14:708–715.

Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10:48.

ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17:368–376.

Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, Mu XJ, Khurana E, Rozowsky J, Alexander R et al. (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature* 489:91–100.

Gilks WR, Richardson S, Spiegelhalter D (1995) *Markov chain Monte Carlo in practice* CRC press.

Hammoud SS, Low DH, Yi C, Carrell DT, Guccione E, Cairns BR (2014) Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell* 15:239–253.

Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F et al. (2006) The ucsc genome browser database: update 2006. *Nucleic Acids Res* 34:D590–D598.

Hubisz MJ, Pollard KS, Siepel A (2011) PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief in Bioinform* 12:41–51.

Jensen JL, Pedersen AMK (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv Appl Probab* 32:499–517.

Jin SG, Kadam S, Pfeifer GP (2010) Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res* 38:e125–e125.

Jojic V, Jojic N, Meek C, Geiger D, Siepel A, Haussler D, Heckerman D (2004) Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics* 20:i161–i168.

Jones GL, Haran M, Caffo BS, Neath R (2006) Fixed-width output analysis for Markov chain Monte Carlo. *J Am Stat Assoc* 101:1537–1547.

Jung YH, Sauria ME, Lyu X, Cheema MS, Ausio J, Taylor J, Corces VG (2017) Chromatin states in mouse sperm correlate with embryonic and adult regulatory landscapes. *Cell reports* 18:1366–1382.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D (2010) Bigwig and bigbed: enabling browsing of large distributed datasets. *Bioinformatics* 26:2204–2207.

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

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

Lauritzen SL (1996) *Graphical models*, Vol. 17 Clarendon Press.

Lesch BJ, Silber SJ, McCarrey JR, Page DC (2016) Parallel evolution of male germline epigenetic poising and somatic development in animals. *Nat Genet* 48:888–894.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R et al. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

Lun AT, Smyth GK (2016) csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res* 44:e45.

Lunter G, Hein J (2004) A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* 20:i216–i223.

Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann JM, Pervouchine DD, Sullivan TJ et al. (2015) The human transcriptome across tissues and individuals. *Science* 348:660–665.

Mi H, Poudel S, Muruganujan A, Casagrande JT, Thomas PD (2016) PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res* 44:D336–D342.

Molaro A, Falciatori I, Hodges E, Aravin AA, Marran K, Rafii S, McCombie WR, Smith AD, Hannon GJ (2014) Two waves of de novo methylation during mouse germ cell development. *Gene Dev* 28:1544–1549.

Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, Smith AD (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* 146:1029–1041.

Peng JC, Valouev A, Swigut T, Zhang J, Zhao Y, Sidow A, Wysocka J (2009) Jarid2/jumonji coordinates control of PRC2 enzymatic activity and target gene occupancy in pluripotent cells. *Cell* 139:1290–1302.

Picelli S, Björklund ÅK, Reinius B, Sagasser S, Winberg G, Sandberg R (2014) Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res* 24:2033–2040.

Pupko T, Pe I, Shamir R, Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol Biol Evol* 17:890–896.

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* 42:W187–W191.

Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D et al. (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* 30:1003–1005.

Roberts GO, Polson NG (1994) On the geometric convergence of the gibbs sampler. *J Roy Stat Soc B* pp. 377–384.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050.

Siepel A, Haussler D (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol* 21:468–488.

Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD (2013) A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PloS ONE* 8:e81148.

Supek F, Bošnjak M, Škunca N, Šmuc T (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS ONE* 6:e21800.

Wang J, Zhuang J, Iyer S, Lin XY, Greven MC, Kim BH, Moore J, Pierce BG, Dong X, Virgil D et al. (2013a) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* 41:D171–D176.

Wang Q, Gu L, Adey A, Radlwimmer B, Wang W, Hovestadt V, Bähr M, Wolf S, Shendure J, Eils R et al. (2013b) Tagmentation-based whole-genome bisulfite sequencing. *Nat Protoc* 8:2022–2032.

Wang XX, Sun BF, Jiao J, Chong ZC, Chen YS, Wang XL, Zhao Y, Zhou YM, Li D (2015) Genome-wide 5-hydroxymethylcytosine modification pattern is a novel epigenetic feature of globozoospermia. *Oncotarget* 6:6535.

Yang Z (1994) Estimating the pattern of nucleotide substitution. *J Mol Evol* 39:105–111.

Yu M, Hon GC, Szulwach KE, Song CX, Jin P, Ren B, He C (2012) Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat Protoc* 7:2159.

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.