# An Analysis of Methylome Evolution in Primates

Arne Sahm,*[1] Philipp Koch,[2] Steve Horvath,[3] and Steve Hoffmann*[1]

[1]Computational Biology Group, Leibniz Institute on Aging—Fritz Lipmann Institute, Jena, Germany
[2]Core Facility Life Science Computing, Leibniz Institute on Aging—Fritz Lipmann Institute, Jena, Germany
[3]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

***Corresponding authors**: E-mails: arne.sahm@leibniz-fli.de; steve.hoffmann@leibniz-fli.de.

**Associate editor:** Li Liu

## Abstract

**Although the investigation of the epigenome becomes increasingly important, still little is known about the long-term evolution of epigenetic marks and systematic investigation strategies are still lacking. Here, we systematically demonstrate the transfer of classic phylogenetic methods such as maximum likelihood based on substitution models, parsimony, and distance-based to interval-scaled epigenetic data. Using a great apes blood data set, we demonstrate that DNA methylation is evolutionarily conserved at the level of individual CpGs in promotors, enhancers, and genic regions. Our analysis also reveals that this epigenomic conservation is significantly correlated with its transcription factor binding density. Binding sites for transcription factors involved in neuron differentiation and components of AP-1 evolve at a significantly higher rate at methylation than at the nucleotide level. Moreover, our models suggest an accelerated epigenomic evolution at binding sites of BRCA1, chromobox homolog protein 2, and factors of the polycomb repressor 2 complex in humans. For most genomic regions, the methylation-based reconstruction of phylogenetic trees is at par with sequence-based reconstruction. Most strikingly, phylogenetic reconstruction using methylation rates in enhancer regions was ineffective independently of the chosen model. We identify a set of phylogenetically uninformative CpG sites enriched in enhancers controlling immune-related genes.**

***Key words:*** **methylation, human, great apes, epigenomics, phylogenetics, polycomb repressor 2.**

Article

## Introduction

Sequence-based methods for phylogenetic tree reconstruction developed more than half a century ago (Sokal and Michener 1958; Fitch and Margoliash 1967; Jukes and Cantor 1969; Fitch 1971) have laid the methodological foundation for much of the progress in evolutionary genetics. In addition to determining sequences of speciation events, they have associated genotypes with phenotypes and investigated critical selection pressures (Pennacchio et al. 2006; Kosiol et al. 2008; Ge et al. 2013; Gaya-Vidal and Alba 2014; Roux et al. 2014; Reichwald et al. 2015; Webb et al. 2015; Sahm et al. 2018; Cui et al. 2019).

It has become increasingly clear that heritable information does not solely consist of the sequence of the four nucleobases adenine, cytosine, guanine, and thymine (Boffelli and Martin 2012; Burggren 2016; Yi 2017; Lind and Spagopoulou 2018). Although the functional analysis of chemical DNA modifications and its associated proteins is gaining pace, little is known about the long-term evolution and conservation of epigenomic signals. Among the most important of these epigenetic modifications are DNA methylation and post-translational modifications of histones (Chen et al. 2017; Michalak et al. 2019). DNA methylation marks, for instance, are copied to newly synthesized DNA strands by DNA methylation transferases targeted to the replication foci (Leonhardt et al. 1992; Vertino et al. 2002; Kar et al. 2012). Importantly, epigenetic information may not only be passed on from cell to cell in the soma but also through the germline from generation to generation (Verhoeven et al. 2016; Perez and Lehner 2019).

However, it is not necessary to invoke these findings to take an interest in the phylogenetic information conveyed by the epigenome. The evolution of epigenomic readers and writers themselves ultimately affects their function and changes in the epigenomic landscape may thus be understood as a consequence of this very process. Also, the DNA sequence context codetermines the epigenomically encoded information (Xiao et al. 2014; Lowdon et al. 2016). For instance, transcription factor binding sites (TFBS) critical for complexes involving epigenomic writers and readers implicitly link DNA sequence and methylation signal. Conversely, it is also known that the presence or absence of epigenomic markers can influence DNA sequence evolution (Xia et al. 2012; Makova and Hardison 2015). Thus, changes in epigenomic markers, their respective local sequence context, and in the composition of epigenomic readers and writers influence each other and are therefore difficult to disentangle.

Although it is clear that epigenetic marks in principle have a major influence on gene expression and cell identity, it is still largely open which marks have which functional significance where in the genome (Barrero et al. 2010; Kim and Costello

2017; Xia et al. 2020). As in many other examples (Bergmiller et al. 2012; Luo et al. 2015; Arun et al. 2016), it is plausible that the degree of conservation would be a strong indicator for functional relevance. Furthermore, it could contribute to the elucidation of the molecular causes of phenotypic differences. To date, comparatively few studies have compared the epigenome of different species, for example, to identify pairwise differentially methylated regions (Molaro et al. 2011; Zeng et al. 2012; Mendizabal et al. 2016; Böck et al. 2018). To learn more about the long-term evolution of epigenomic marks, it appears necessary to develop new models allowing systematic evolutionary analyses—as is the case at the genetic level (Xiao et al. 2014; Lowdon et al. 2016).

Obviously, the central question is whether the evolutionary information of individual marks is sufficient to allow meaningful analyses. In the light of their tissue-specificity and responsiveness to environmental stimuli, it remains to be established which epigenomic marks are conserved well enough over longer evolutionary distances allowing the reconstruction of phylogenetic relationships.

That correct tree topologies can, in principle, be reconstructed from methylation data was shown by Martin et al. (2011) using blood samples from several primate species and a simple distance-based tree method to reconstruct a single tree from the methylome. Also, Qu et al., whose primary focus was on hypomethylated regions, reconstructed a single tree from the whole methylome using a broader species set and a sophisticated time-continuous Markov chain model. Their results indicated faster epigenomic evolution in rodent than in primate sperm (Qu et al. 2018). The work of Hernando-Herraez et al. was mainly concerned with differentially methylated regions resulting from pairwise comparisons but also demonstrated, using a simple hierarchical clustering approach and great apes blood data, that genomic regions that showed incomplete lineage sorting on the nucleotide level recapitulated this on methylation level.

Building on these results, we systematically investigate different models of epigenomic evolution to facilitate insights into functions of single genes or pathways. Specifically, we transfer tree reconstruction such as Maximum Parsimony, Maximum Likelihood (based on substitution models), and distance-based methods to the level of DNA methylation. Subsequently, models are applied to simulated data as well as publicly available real data to analyze their ability to reconstruct correct phylogenetic trees based on DNA methylation information. Substitution models arguably promise the greatest potential regarding functionally relevant analyses such as positive selection or accelerated epigenetic evolution in comparison to the genetic level. To this end, we evaluate different evolutionary scenarios for various genomic features (e.g., enhancers, gene bodies).

## Materials and Methods

### Real Data Set

To test the methods we developed, we used publicly available whole-genome bisulfite sequencing data from blood samples of four primate species: *Homo sapiens* (hereafter, human), *Pan troglodytes* (chimpanzee), *Gorilla gorilla* (gorilla), and *Pongo abelii* (orangutan) (PRJNA286277; Hernando-Herraez et al. 2015). The data set consisted of 286–324 million read pairs per species. With a length of 90 base pairs per read, this amounts to 17-fold to 25-fold genome coverage, assuming a genome size of 3 Giga base pairs per species (supplementary table S1, Supplementary Material online). Supplementary fig. S1, Supplementary Material online summarizes the processing of the real data; details are given in the supplementary material, Supplementary Material online.

### Identification of Orthologous Defined Regions and Annotations

Our study distinguishes between four classes of genomic regions (fig. 1A). The gene body class, reflecting the translated part of the genome, is differentiated from regulatory regions embedded in enhancer and two promoter-related regions reflecting potential differences in selection pressures of methylation in coding and noncoding sequences. Specifically, we consider promoter-near regions 2000 base pairs upstream and downstream of the TIS, respectively. In practice, the 2000-Up-TIS class covers the promotor and untranslated regions. Coordinates of the translation initiation and end sites of the human protein-coding genes ($n = 19{,}374$) were obtained from the Uniprot track of the *University of California, Santa Cruz* (UCSC) table browser for the genome version hg38 (Karolchik et al. 2004). Coordinates of enhancers were obtained from UCSCs geneHancer track retaining only "double elite" enhancers with known interactions ($n = 25{,}572$). Coordinates of 9,679 genes and 20,327 enhancers were successfully translated to the three nonhuman primates using the UCSC liftover tool (Kent et al. 2010). For practical reasons, analysis was restricted to elements with a maximum length of 100,000/50,000 base pairs, respectively (fig. 1A). For functional analyses, the UCSC hg38 *Transcription Factor ChIP-seq Clusters* track, aggregating TFBS from more than 1,200 experiments in human samples for 340 TFs, was incorporated into this study.

### CpG Alignment

For each region instance, the respective four orthologous sequences were aligned using Clustalw2, version 2.0.10 (Larkin et al. 2007). Since the alignment of effectively nonhomologous bases in poorly conserved regions may lead to false phylogenetic inference (Jordan and Goldman 2012), we used trimAl, version v1.4.rev15 with the parameter "-strictplus" removing unreliable alignment columns (Capella-Gutiérrez et al. 2009). In addition, only those alignments were considered for further analysis for which at least 25 evaluable CpGs remained. The methylation fractions previously determined in the individual species were then mapped to the CpGs of the alignment using the known coordinates. This led to MFA, forming the empirical data basis of this work (fig. 1A).

### Evaluation Strategy

The quality of individual tree reconstruction methods under different parametrizations was measured with increasing input sizes using two benchmarks: 1) the ability to correctly reconstruct the known primate tree topology and 2) the

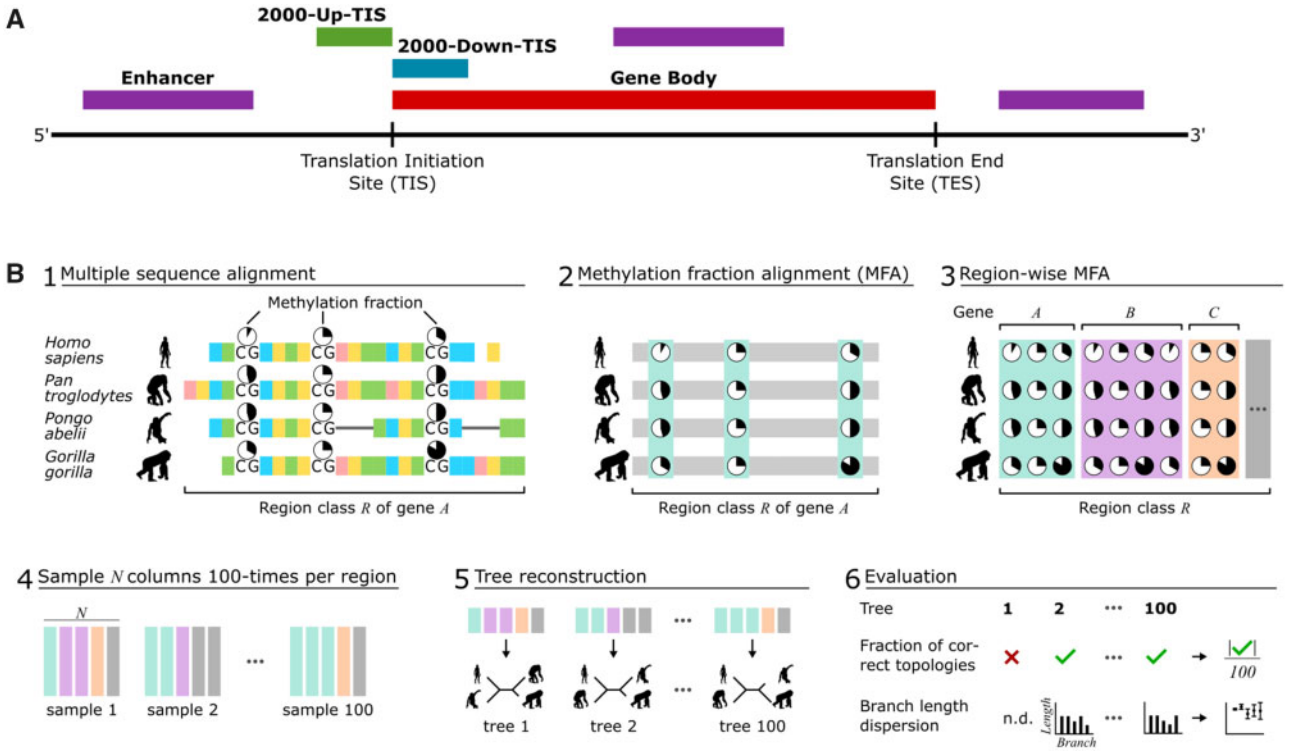


FIG. 1. (*A*) Classes of genomic regions examined in this work. The region classes were defined using the Uniprot annotation of all human protein-coding genes (2000-Up-TIS, 2000-Down-TIS, Gene body, $n = 19{,}374$) or geneHancer annotation (Enhancer, $n = 25{,}572$). The corresponding genome regions in chimpanzee, gorilla, and orangutan were acquired based on a genome alignment strategy. (*B*) Workflow and evaluation strategy. 1) For each region defined in (*A*) of each gene an MSA of the four species studied in this work was created and the methylation fractions measured from blood samples mapped to the alignment. 2) From this, the MFA was extracted. 3) We merged these gene-wise alignments to region-wise alignments. 4) From the region-wise MFAs, we sampled 100 times *N* columns. 5) From each of the 100 drawings, we reconstructed a phylogenetic tree. 6) As evaluation criteria, we used, on the one hand, the proportion of trees that correspond to the known great ape topology. On the other hand, we quantified the dispersions of the branch lengths. The procedure from 4) to 6) was performed for different values of *N* to consider the evaluation criteria as a function of the amount of input data.

degree of dispersion of the branch lengths thus determined. For these purposes, all methylation alignments of a region class were first concatenated (pooling). From this pool, *N* alignment columns were drawn for each combination of the used tree reconstruction approaches (see below) and $N \in \{100,\ 200, \ldots, 1{,}000,\ 2{,}000,\ \ldots, 10{,}000\}$. The procedure was repeated 100 times. Subsequently, it was determined how many of the 100 repetitions led to the correct tree topology (fig. 1*B*). To measure dispersion in terms of standard deviation of branch lengths from the mean length, the correct topology was fixed. Briefly, we expect to observe that the proportion of correct topologies should increase with increasing *N* and the dispersion of branch lengths should decrease.

## Tree Reconstruction Methods

We used three main strategies to reconstruct phylogenetic trees from the aligned methylation fractions: 1) Markov process-based maximum likelihood, 2) parsimony, and 3) distance-based methods. Here, we particularly focused on maximum likelihood and investigated different models, parameterizations, and assumptions based on this method. For simplicity, in the following one combination is termed the *default method* and all alternative methods and parametrizations are compared against it. Figure 2 summarizes the different tree reconstruction methods used. The methods developed for tree reconstruction from methylation (or more generally interval scaled) data were implemented in R and are available on https://github.com/Hoffmann-Lab/PhyloEpiGenomics (last accessed August 3, 2021).

## Maximum Likelihood

We propose substitution models for the analysis of DNA methylation fractions analogous to the frequently used continuous-time Markov process models at nucleotide (e.g., Jukes and Cantor 1969; Kimura 1980; Hasegawa et al. 1985), codon (e.g., Goldman and Yang 1994), and amino acid (e.g., Kishino et al. 1990) levels. In contrast to nucleotides, codons, and amino acids measured on a nominal scale, methylation fractions are measured on an interval scale. To address this difference, we introduce a discretization function

$$d : [0, 1] \rightarrow \{1, \ldots, n\}$$

with $n \in \mathbb{N}$ being the number of model methylation states and $\forall\ x > y :\ d(x) \geq d(y)$.

Let a model *M* be defined by the pair $M = (\pi, Q)$ consisting of an initial (and equilibrium) distribution $\pi \in [0, 1]^n$ and an $n \times n$ transition probability rate matrix $Q = (Q_{s,t})$ with $1 \leq s \neq t \leq n$. As usual, the transition probability

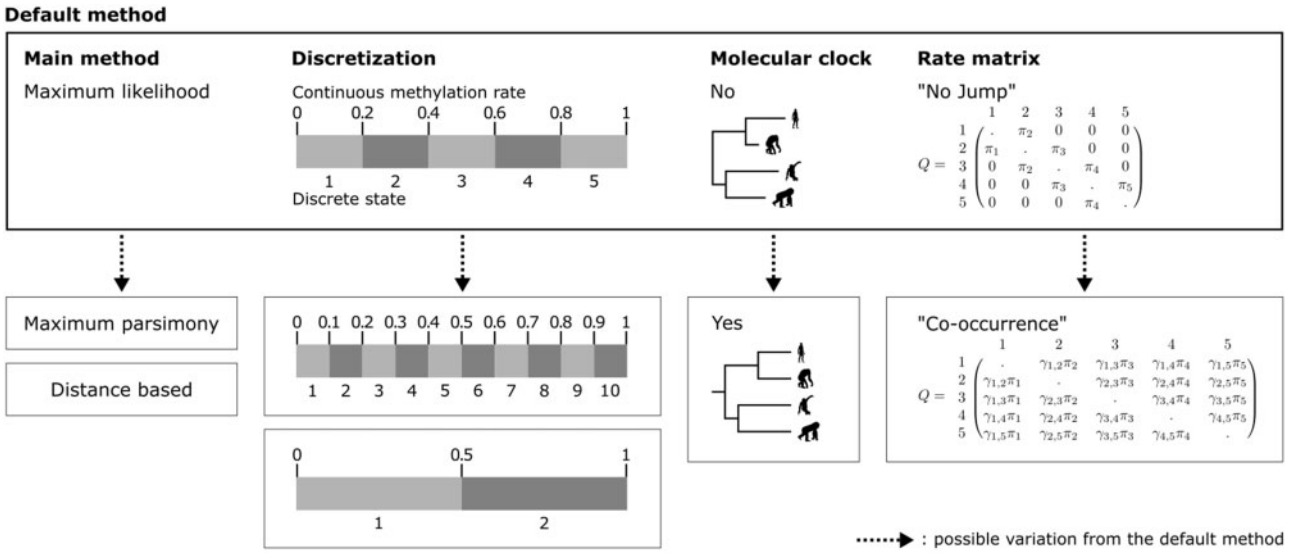


FIG. 2. Applied tree reconstruction methods. Most of the analyses in this work were performed with a maximum likelihood tree reconstruction method based on a Markov model of the evolution of methylation fractions. The model is based on a discretization of the floating-point methylation fractions into five states. No molecular clock is assumed. The design of the transition rate matrix Q assumes that a methylation fraction status cannot evolve into a nonadjacent status within short time spans (*No Jump Model*). This means that when the methylation fraction changes from a state A to a nonadjacent state B during evolution, all states between A and B have been passed through. The combination of the tree reconstruction method and the properties of the models described is called the *default method* in the context of this work for simplification. To better assess the effects of the assumptions of the default method, we compared this method with variations of itself: different number of states, molecular clock, or a transition rate matrix that allows direct change to distant states. In addition, we have also applied two tree reconstruction methods that differ fundamentally from maximum likelihood. For this, either evolutionary distances based on the model of the default method were determined and then neighbor-joining was applied or a parsimony approach was used.

matrix $P(\tau)$ for a given branch length $\tau$ is determined by numerically finding a solution to

$$P(\tau) = e^{Q*\tau}.$$

The likelihood of a given phylogenetic tree can then be determined by Felsenstein's method assuming independent evolution of CpG sites (Felsenstein 1981). Branch lengths of a given tree topology are estimated by maximizing the likelihood and the optimal topology is that with the highest likelihood. For likelihood maximization, we use an optimized version of the *Broyden–Fletcher–Goldfarb–Shanno* method (Broyden 1970; Byrd et al. 1995).

In this paper, we propose two flavors of evolutionary models that we call *No Jump Model* and *Co-occurrence Model*. The No Jump Model assumes that the methylation fractions change smoothly during evolution, that is within short time intervals the methylation state of a CpG site can only change to one of the neighboring states (see fig. 2). In contrast, the Co-occurrence Model also allows transitions to distant states in short time intervals. Here, the transition probabilities between two states are made dependent on how often these two states could be observed empirically within an alignment column.

To formally define these models, let $X = (X_{i,j})$ be a given $k$ x $l$ matrix with the rows corresponding to $k$ homologous CpG-sites, the columns corresponding to $l$ indices of the species examined, and each entry $X_{i,j} \in [0,1]$ being the measured methylation fraction of the $i$th CpG-site in the species with the index $j$; $1 \leq i \leq k$, $1 \leq j \leq l$. Here, $X$ will either be drawn from real data, that is the concatenated alignments of a region class, or from simulated data (see below).

Subsequently, the number of each methylation state $s$ in each species $a$ is counted using the function $o_a : \{1, \ldots, n\} \rightarrow \mathbb{N}_0$ with $1 \leq a \leq l$

$$o_a(s) = \sum_{r=1}^{k} \delta_{s,d(X_{r,a})}$$

with $\delta_{p,q}$ being the Kronecker delta $\delta_{p,q} = \begin{cases} 0 \text{ if } p \neq q \\ 1 \text{ if } p = q \end{cases}$.

Based on this, we define the equilibrium frequency $\pi = (\pi_s)$ for both, the No Jump and the Co-occurrence Model as

$$\pi_s = \frac{\sum_{a=1}^{l} o_a(s)}{\sum_{u=1}^{n} \sum_{a=1}^{l} o_a(u)}.$$

Accordingly, for the No Jump Model, we define $Q = (Q_{s,t})$ as

$$Q_{s,t} = \begin{cases} 0 \text{ if } |s-t| > 1 \\ \pi_t \text{ else} \end{cases}$$

with $1 \leq s \neq t \leq n$. For the Co-occurrence Model, the number of co-occurrences of all combinations of methylation states $s$ and $t$ and species $a$ and $b$ is counted using a function

$co_{a,b} : \{1, \ldots, n\}^2 \rightarrow \mathbb{N}_0$ with $1 \leq a \leq l$, $1 \leq b \leq l$, $a \neq b$

$$co_{a,b}(s,t) = \sum_{r=1}^{k} (\delta_{s,d(X_{r,a})} * \delta_{t,d(X_{r,b})} + \delta_{t,d(X_{r,a})} * \delta_{s,d(X_{r,b})}).$$

Let, $\gamma_{s,t} = \gamma_{t,s} = \frac{\sum_{a=1}^{l-1} \sum_{b=a+1}^{l} co_{a,b}(s,t)}{\sum_{u=1}^{n-1} \sum_{v=u+1}^{n} \sum_{a=1}^{l-1} \sum_{b=a+1}^{l} co_{a,b}(u,v)}$

be the fraction observed co-occurrences of the states $s$ and $t$ across the alignment.

Finally, we define $Q = (Q_{s,t})$ for the Co-occurrence Model as

$$Q_{s,t} = \gamma_{s,t} * \pi_t$$

with $1 \leq s \neq t \leq n$. As usual, for both models $Q_{s,s}$ are fixed such that

$$Q_{s,s} = -\sum_{t \neq s} Q_{s,t}.$$

For this study, we tested the models with different parameters and technical settings. Detailed parametrizations for the No Jump and Co-occurrence models are shown in supplementary figures S2 and S3, Supplementary Material online. Specifically, we tested models with different state numbers $n$, and we used the models both with and without the assumption of a molecular clock.

### Maximum Parsimony

The basic idea of Fitch's Maximum Parsimony algorithm (Fitch 1971) is to find a set of sequence states at the inner nodes minimizing the number of necessary changes along the edges of the tree. We adapted the algorithm to work on the interval scale. Supplementary figures S4–S6, Supplementary Material online illustrate the differences between the algorithm and our modification.

In the bottom-up postorder tree traversal, an interval $I(m)$ is assigned to each node $m$. Leaves are initialized with

$$I(l) = [\text{observed number at } l, \text{observed number at } l]$$

.

Then, for each inner node $m$ with children $x$ and $y$, $I(m)$ is determined by

$$I(m) = \begin{cases} I(x) \cap I(y), & \text{if } I(x) \cap I(y) \neq \varnothing \\ [min(\max(I(x)), \max(I(y))), \max(\min(I(x)), \min(I(y)))], & \text{else} \end{cases}$$

.

In the top-down preorder tree traversal, each node $m$ is assigned a number $i(m) \in I(m)$. For the root $r$, $i(r)$ is chosen as an arbitrary number within $I(r)$; for all other inner nodes with parent node state $p$

$$i(m) = \begin{cases} p, & \text{if } p \in I(m) \\ \arg\min_{x \in I(m)} |p - x|, & \text{else} \end{cases}$$

.

As with the original algorithm, the optimal tree topology is the one that explains the sequence alignment with the lowest number of necessary changes overall.

### Distance-Based Methods

Distance matrices were determined based on the *No Jump Model* described above. Each distance matrix $T = (T_{a,b})$ with $1 \leq a \leq l$, $1 \leq b \leq l$ was determined by maximizing the likelihood for the pairwise distances in the usual way

$$T_{a,b} = T_{b,a} = \arg\max_{\tau \in \mathbb{R}^{*+}} \prod_{1 \leq i \leq k} (\pi_{d(X_{i,a})} * P_{a,b}(\tau))$$

for $a \neq b$; and $T_{a,a} = 0$ for $1 \leq a \leq l$. For each distance matrix $T$, the corresponding phylogenetic tree was reconstructed using the Neighbor-Joining algorithm (Saitou and Nei 1987).

### Simulations

Artificial alignments were generated based on the known tree topology and divergence times of the great apes (Locke et al. 2011). For each artificial alignment column, the tree was traversed in preorder. For each node $m$, a state was drawn from $(1, \ldots, n)$ using a probability vector $\varphi(m)$. For the root $r$, $\varphi(r)$ was set to the equilibrium distribution $\varphi(r) = \pi$. For every other node $m$ with the incoming edge $e$ with length $\tau_e$ and the parent node in state $s$, $\varphi(m)$ was set to the $s$-th row of the matrix $P(\tau_e)$ (supplementary fig. S7, Supplementary Material online). The simulated alignment column then results from the states assigned to the leaves of this tree. This simulation approach has been widely used at nucleotide and codon level (e.g., Rambaut and Grassly 1997; Yang 1997; Fletcher and Yang 2009).

This general scheme has been extended for the different concrete analyses. For the comparison of real and simulated data, noise of different orders of magnitude was added to the simulated data. For the analyses of long-branch attraction and resolution limits for reconstruction depending on the branch length the tree used for the simulation was changed (see supplementary methods, Supplementary Material online for details, also for the analysis of site-specificity that was conducted on real data).

## Results and Discussion

On nucleotide data, maximum likelihood-based Markov models are critical tools for formal testing of evolutionary hypothesis. This includes the detection of sequences under positive selection on certain branches of a phylogeny. In order to address such questions in the epigenome, we first focused on the evaluation of maximum likelihood reconstructions in this work. Specifically, we investigated different flavors of this methodology. A discretization of CpGs according to their methylation levels is at the heart of the default method and derivatives thereof. Depending on the methylation, in our models, a single CpG can assume one of two, five (default method), or ten states

### CpG Methylation Varies Widely between Regions but Little between Species

To test the functionality of the developed methods, we used a public whole-genome bisulfite sequencing data set generated from blood samples of four great apes: human, chimpanzee,

gorilla, and orangutan (Hernando-Herraez et al. 2015) as well as to simulated data. On a genome-wide level, methylation patterns are extremely similar in all species and apparently dominated by the functional aspects of the four selected classes of regions (figs. 1 and 2; see supplementary results, Supplementary Material online for details and fig. 1A for region definition). Also, the average similarity of the methylation fractions varies more between the regions under consideration than between species pairs—reflecting the distinct functional roles of the regions. Interspecies similarity ranges from 87.2–90.0% in the gene body to 94.6–95.4% in the 2000-Up-TIS region (table 1).

To model a null-hypothesis for the analysis of region-specific similarities, we repeatedly uniformly ($n$=1,000) drew random pairs of methylation rates from the class-specific background distributions and calculated expected differences. Based on this, we found that the level of region-specific similarity between species pairs is significantly higher than expected ($P < 10_{-3}$, table 1). These initial, descriptive results suggest that the methylation data contain evolutionarily conserved, phylogenetically analyzable signals. Despite a high degree of similarity probably dominated by region-specific biological functions, differences at this rather coarse-grained level already reflect the phylogeny of the great apes.

## Tree Reconstruction Works Best for Translation Initiation Site-Downstream and Gene Body Classes

To get a more detailed view, we compared the classes of regions in terms of their potential to reconstruct correct tree topology using the methylation data. With the default method, the 2000-Down-TIS region contains the highest phylogenetic signal (fig. 3C). Notably, this class has by far the highest proportion of hemi-methylation (0.2–0.8). It seems plausible that CpGs neither fully methylated nor demethylated across the tissue also have the technical property of being phylogenetically most informative. It is well documented that an increase in methylation downstream of the transcription start site and thus potentially overlapping with translated regions may have a potent suppressive effect on gene expression (confer fig. 3 of Ehrlich and Lacey [2013] and Appanah et al. [2007]). At the same time, the region overlaps with proximal parts of the gene body where a hypermethylation is frequently associated with strong expression (Zemach et al. 2010). In conjunction with the high level of global similarity established above, these results provide further evidence that specific phylogenetic information is embedded in this regulatorily relevant region. In addition to informative signals in regulatory regions, also interspecies expression differences may contribute to the surprisingly high rate of correct reconstructions. The observation that tree reconstruction based on the gene body works second best supports this notion.

**Table 1.** Average similarity of methylation fractions between species in percent.

| | 2000-Up-TIS | 2000-Down-TIS | Gene Body | Enhancer |
|---|---|---|---|---|
| **Human–chimpanzee** | 95.4 | 93.5 | 90.0 | 92.9 |
| **Human–gorilla** | 94.9 | 92.7 | 87.9 | 92.0 |
| **Human–orangutan** | 94.6 | 92.6 | 88.7 | 91.8 |
| **Chimpanzee–gorilla** | 95.0 | 92.8 | 87.6 | 92.1 |
| **Chimpanzee–orangutan** | 94.7 | 92.7 | 88.7 | 91.9 |
| **Gorilla–orangutan** | 94.6 | 92.5 | 87.2 | 91.3 |

With the notable exception of enhancers, the proportion of correctly reconstructed tree topologies converges clearly toward 1 in all region classes, depending on the amount of available data (fig. 3C). The result is similar concerning the second benchmark, that is whether the branch lengths we determined converge with increasing data volume. This is the case for all branches of the great apes in all regions studied (fig. 3D).

## Methylation-Based Tree Reconstruction May Outcompete Nucleotide-Based Reconstruction

For comparison, we juxtaposed methylation-based reconstruction with classical nucleotide-based reconstructions. For the gene body and Up-TIS-2000 classes, the fraction of correctly inferred trees is almost identical for methylation and nucleotide data. Most strikingly, using methylation data from the Down-TIS-2000 class reconstruction clearly outcompeted nucleotide-based reconstruction. Conversely, reconstruction based on methylation data obtained from enhancer regions essentially failed (fig. 4A). In the light of high similarities of methylation fractions in this region (table 1), this result is most surprising.

To additionally quantify the performance of methylation-based reconstruction, we generated artificial methylation fraction alignments (MFAs) in complete analogy to nucleotide- or amino acid-based alignments frequently used in the assessment of reconstruction algorithms or evolutionary models (e.g., Rosenberg and Kumar 2001; Zhang et al. 2005; Shavit Grievink et al. 2010; Zaheri et al. 2014). Using these MFAs, we reconstructed phylogenetic trees and determined the fraction of correct topologies (fig. 4A). The procedure was repeated after adding different levels of artificial noise to the simulated MFAs. At a noise level between 0.1 and 0.2 standard deviations, the reconstruction performance of simulated data is at par with real data derived from gene body and Up-TIS-2000. Consistently, enhancers perform worse with values corresponding with noise between 0.2 and 0.4 standard deviations, whereas the Down-TIS-2000 class shows substantially better benchmarks with values between 0.05 and 0.1 standard deviations.

In summary, the reconstruction of phylogenetic trees from DNA methylation data seems to work well. The performance compared with nucleotide data is astonishing since methylation can be expected to be subject to frequent changes. Unlike DNA sequences, they are influenced by circadian rhythms, environmental factors, or diseases. Clearly, in practice, the success of a methylation-based reconstruction critically depends on the amount of available data, for example, the read coverages achieved in WGBS experiments, the quality of reference genomes and sequence alignments. Furthermore, the reconstruction of phylogenies in single genes is naturally limited by the number of available data

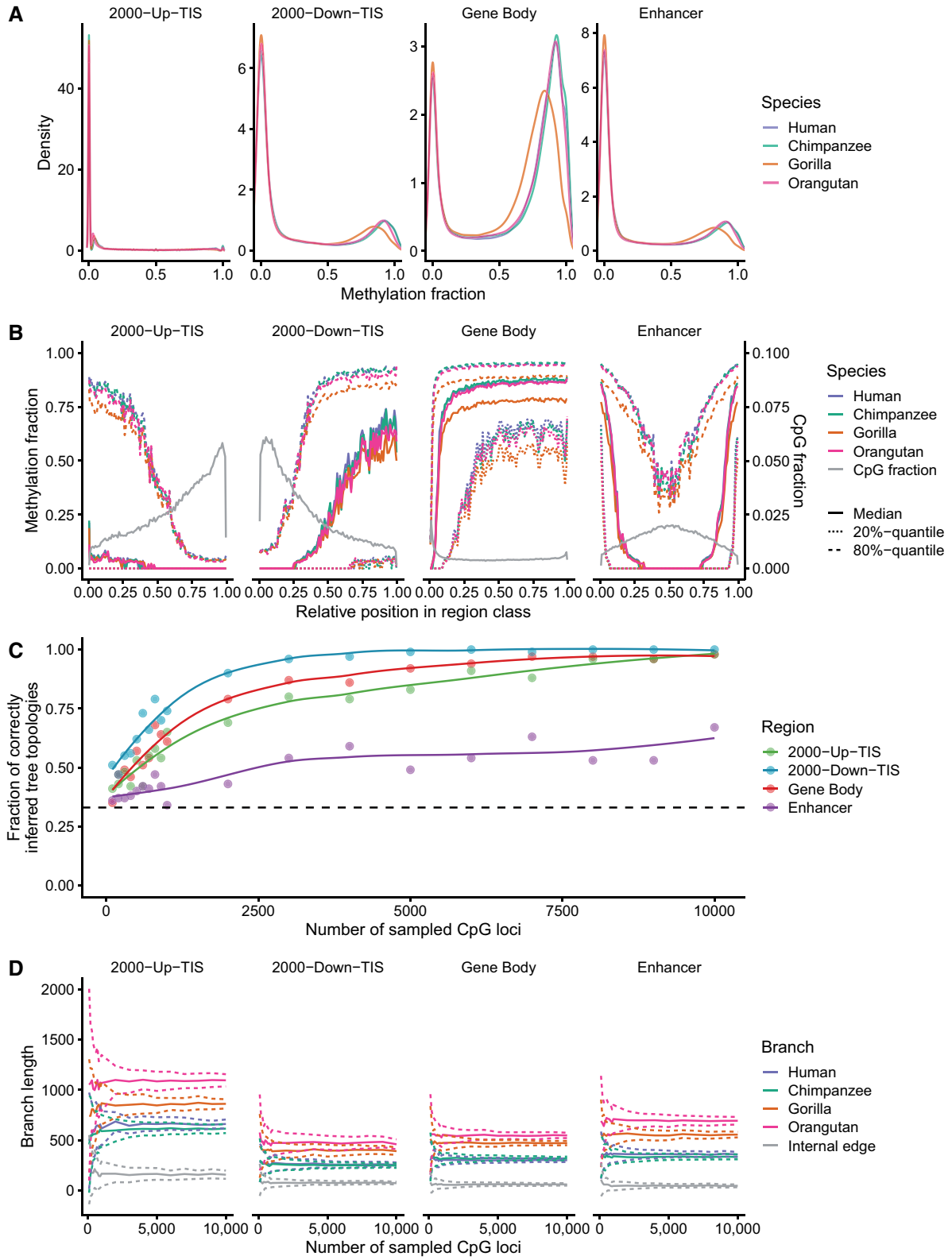


**Fig. 3.** (*A*) Distributions of methylation fractions in the regions and species examined. (*B*) Methylation and CpG fractions by relative position in region class. (*C*) Comparison of regions examined by the fraction of correctly inferred tree topologies. The total number of reconstructed trees per fixed amount of input data (i.e., Number of CpG loci) was always 100. The dashed line indicates the probability of randomly reconstructing the correct tree topology (one-third), solid lines interpolate points for visual guidance. (*D*) Dispersion of branch lengths in the regions examined. The solid lines represent the mean values of the branch lengths for 100 drawings, and the dashed lines represent the mean values ± the corresponding standard deviations. (*A*, *B*, *D*) The given numbers of CpG loci were drawn from the respective region-wise MFAs with the following total numbers of evaluable CpG loci: Up-TIS-2000–126,560; Down-TIS-2000–134,686; Gene body—441,286; enhancer—271,195.

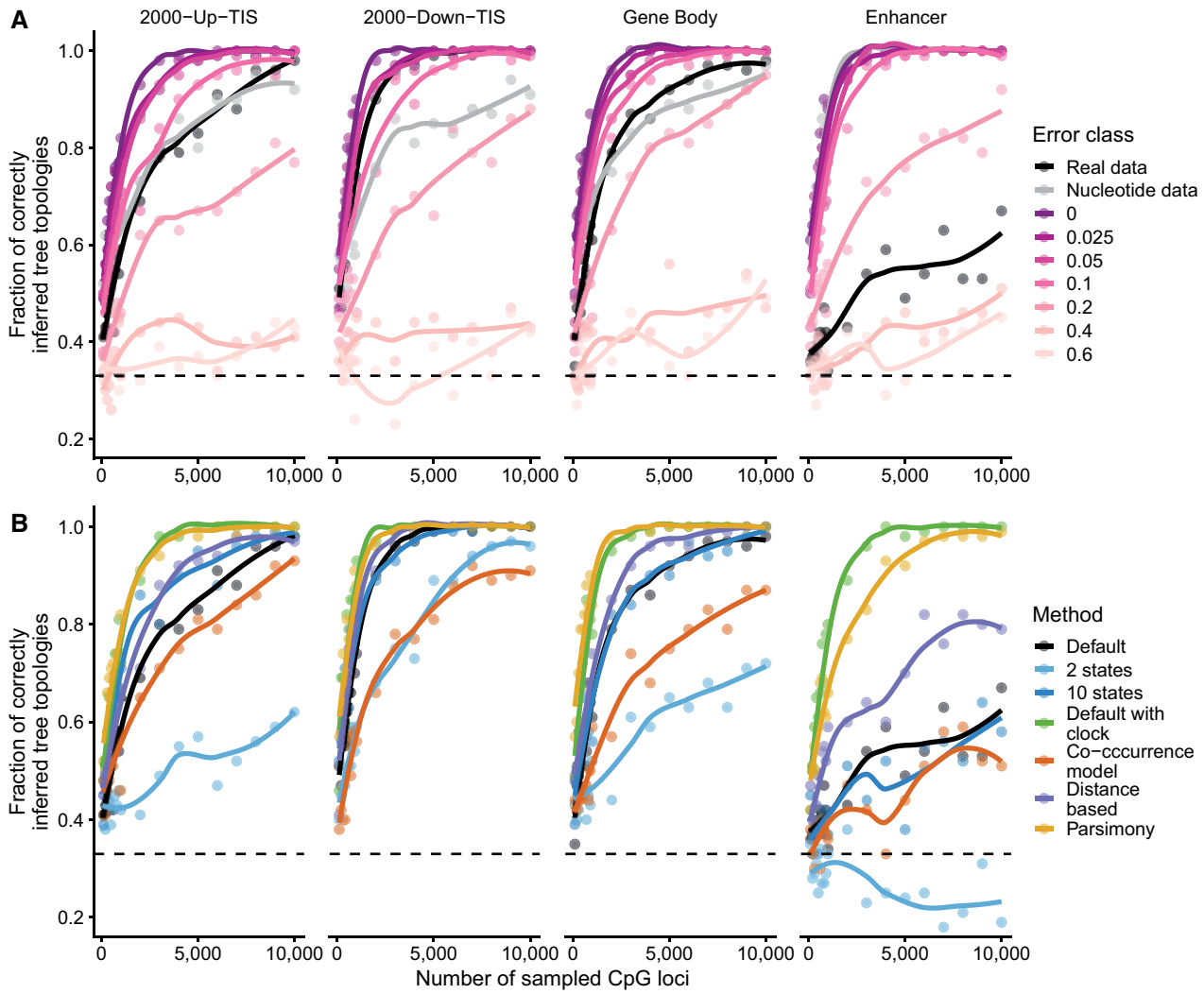


FIG. 4. (*A*) Model versus reality. The performance difference between the simulated data and the real data gives an estimate of how well the model matches reality. We first pretended that the model of the default method perfectly reflects reality: this model and the known phylogenetic tree of the great apes were used to generate artificial alignments. We then reconstructed phylogenetic trees from these alignments using the same model and applied the quality scale of fraction of correctly reconstructed tree topologies. To quantify the difference of model and reality, we additionally applied defined error terms, that is fractions of the standard deviation of the standard normal distribution, to the artificial alignments before tree reconstruction. Reconstructions based on the nucleotide sequences obtained from the alignments are given for comparison. (*B*) Comparison of the different methods (or deviations of our standard method) by fractions of correctly reconstructed tree topologies. The given numbers of CpG loci were drawn from the respective region-wise MFAs with the following total numbers of evaluable CpG loci: Up-TIS-2000−126,560; Down-TIS-2000−134,686; Gene body—441,286; enhancer—271,195. (*A*, *B*) The total number of reconstructed trees per fixed amount of input data (i.e., Number of CpG loci) was always 100. The dashed line indicates the probability of randomly reconstructing the correct tree topology (one-third), solid lines interpolate points for visual guidance.

points. Although the body of a protein-coding gene has often more than 10,000 nucleotides, the measurement of the methylome is restricted to at most a few hundred CpGs (supplementary fig. S8, Supplementary Material online). Experimental or biological noise may thus easily lead to faulty reconstructions. Unlike the genome, which is almost identical in all cells of an individual, the epigenome may reflect tissue-specific phylogenetical information. Thus, interspecies comparisons yield the potential of gaining insights into the evolution of tissues as opposed to the entire organism. The incorporation of other epigenetic data, for example, histone modifications, will be helpful to illuminate such processes.

## Failure of Tree Reconstruction at Enhancers

We then addressed the question of why tree reconstruction with methylation data from enhancers performs significantly worse than in the other regions. Our analyses show that enhancers have more phylogenetically uninformative CpG sites. These uninformative sites were found disproportionally often in enhancers that regulate the expression of IL12 pathway genes and leukocyte and lymphocyte differentiation (see supplementary results, Supplementary Material online for details). Since immunity-related genes themselves are targets of rapid evolutionary changes (Shultz and Sackton 2019), it is tempting to speculate that also sudden methylation changes

within associated regulatory elements contribute to shaping the evolution of the immune response.

## Nonbinary Models of CpG Methylation Work Best for Tree Reconstruction

As shown in figure 4B, the model selection critically impacts phylogenetic reconstructions based on methylation data. The default No Jump Model is based on the idea that within short periods of time it is only possible to switch to adjacent states (fig. 2). If a change from state A to a nonadjacent state B is to be made over a more extended period, all intermediate states between A and B must be transited. With the alternative Co-occurrence Model, we permit sudden changes to nonadjacent states. Interestingly, the Co-occurrence Model consistently performs much worse than the No Jump Model across all regions. This behavior immediately raises the question to which extent such states are also functionally relevant. In contrast, when increasing the number of states to ten typically no or only minimal improvement is observed. Naturally, the precise resolution of tissue-wide methylation fractions critically depends on the read coverage. Thus, for the data used in this study, a ten-state model likely is too fine-grained and over-specified. Once more and better data are available, however, it would be most exciting to consider models with more states or even methods able to omit such discretizations.

The presented default method builds on assumptions frequently made in the analysis of evolutionary relationships. To investigate the influence of those assumptions, we added a model with a molecular clock (1), used Neighbor-Joining as a distance-based method for tree reconstruction (2), and applied the Maximum Parsimony principle (3; fig. 2). In summary, tree reconstruction from DNA methylation data faces similar challenges as tree reconstruction from nucleotide data. For instance, under conditions of similar evolutionary rates on all branches, the aforementioned alternative methods reconstruct the correct tree topology more often than our default method, but are prone to long-branch attraction when rates diverge (figs. 4*B* and 5*A*, 5*B*, see supplementary results, Supplementary Material online for details).

## DNA Methylation Is Conserved at the Individual CpG Site Level

To investigate the degree of local conservation of methylation fractions, we determined the similarity of methylation for neighboring CpGs, defined as $1 - |X_{i,a} - X_{j,a}|$, where $X_{i,a}$ and $X_{j,b}$ are the methylation fractions at CpG-site $i$ and $j$ in species $a$, respectively. This similarity is contrasted with similarities of fractions in the spatial neighborhood of each individual species. Specifically, we define the 1-neighborhood of a CpG as the set of the next CpG upstream and the next CpG downstream (without the CpG in the middle). The 2-neighborhood consists of the next but one CpGs in both directions (without the three CpGs in the middle) and so on. Expectedly, the average similarity of methylation decreases with growing neighborhoods toward a baseline level in all classes of regions. In the 2000-Down-TIS region, for instance, a similarity level of about 93% in the 1-neighborhood, decreases to about 80% in the 25-neighborhood and remains almost constant from then on (fig. 5C, supplementary fig. S12*A*, Supplementary Material online). Strikingly, the class-dependent relationship between neighborhood and methylation similarities is practically identical in all of the investigated species. To put this observation into context, we contrasted the averaged neighborhood similarities with the similarity of methylation rates between two species at a given CpG, that is the in-between similarity. In any given class, the in-between similarity is at least as strong as the 3-neighborhood across all species. For the species pair human–chimpanzee, for instance, the in-between similarity is greater than that of the 1-neighborhood. In other words, our data indicate that, although separated by millions of years of evolution, the methylation rate of a CpG in humans is on average more similar to its orthologous CpG in chimpanzees than to that of its closest neighboring CpG (fig. 5C, supplementary fig. S12*A*, Supplementary Material online). To further investigate this, we have randomly swapped the methylation fractions of the real ape data set within defined exchange distances before reconstructing the phylogenetic trees. Even an exchange distance of one results in a significant decrease in the proportion of correct topologies, which continues to increase as the exchange distance is enlarged (fig. 5*D*, supplementary fig. S12*B*, Supplementary Material online). This clearly indicates that the methylation on the DNA strand is conserved with high local resolution. On the flipside, this result also underscores the dependency of such methylation-based studies on high-quality alignments—very much similar to studies based on the nucleotide or codon level (Jordan and Goldman 2012).

In this context, we addressed the question of whether the phylogenetic information conveyed by DNA methylation and MFAs could arise artificially as a mere consequence of the underlying multiple sequence alignments (MSAs). If the phylogenetic information of the MSA was simply carried over to the MFA, we would expect observing clear correlations between their abilities to infer correct phylogenetic trees. The observation that MSA-based reconstruction of phylogenies using enhancers is at par with other region classes whereas MFA-based reconstruction fails more frequently (fig. 4*A*, supplementary table S4, Supplementary Material online) already indicates that this is not necessarily the case. To investigate this question more systematically, however, we compared methylation-based and sequence-based reconstruction performances subject to increasing levels of local sequence conservation. Although the sequence-based reconstruction performance clearly and expectedly degrades with increased sequence similarities, the MFA-based reconstruction is largely unaffected. Notably, for locally highly sequence-conserved regions it outperforms sequence-based reconstruction and supports the notion that the 5mC signal investigated here is phylogenetically informative in its own right (supplementary fig. S13, Supplementary Material online, see supplementary results and methods, Supplementary Material online for details).

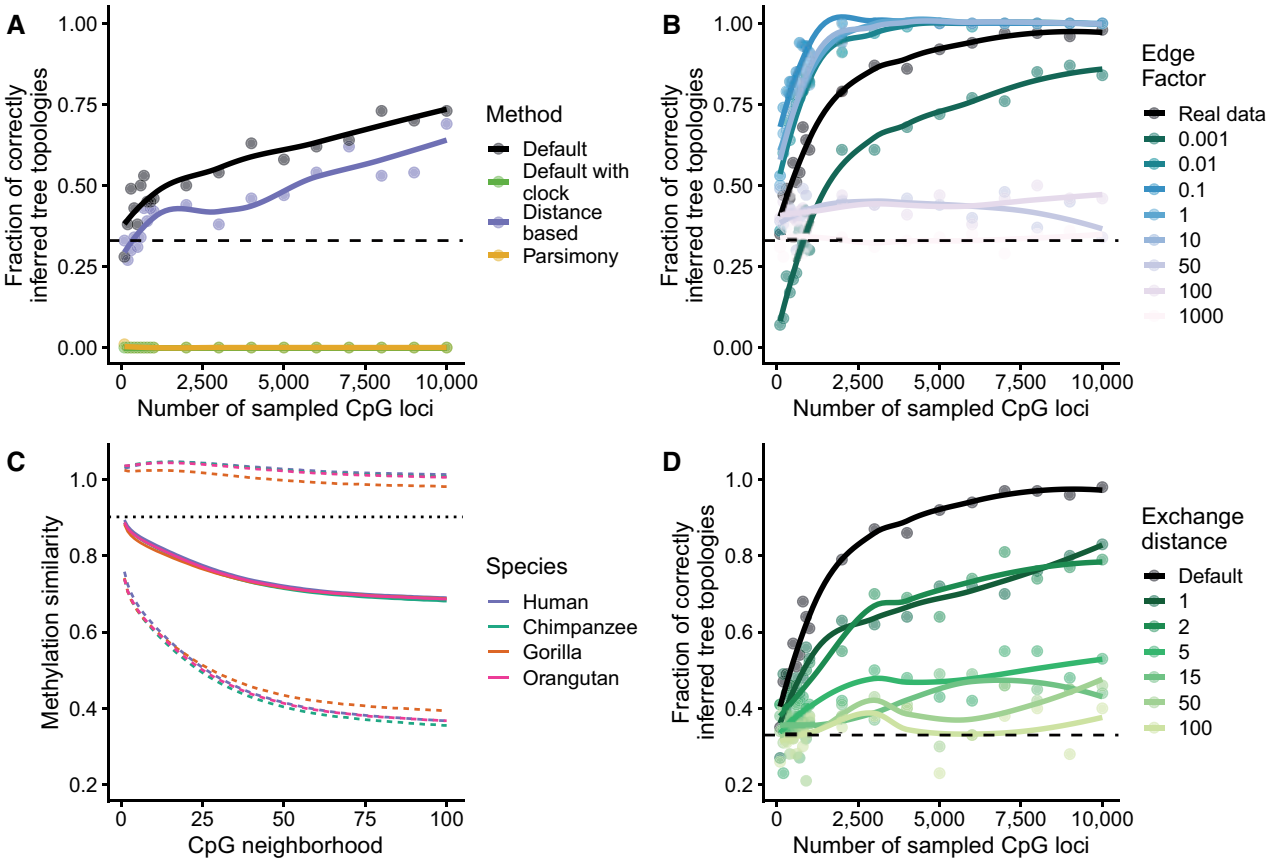


FIG. 5. (*A*) Comparison of selected methods in a long-branch attraction scenario. We generated artificial alignments using the model of our default method and a modified version of the great apes' phylogenetic tree. The tree used for alignment generation was modified in such a way that one terminal branch was given a multiple of its original length. Then, trees were reconstructed from the alignments and the scale of the fraction of correctly inferred topologies applied. (*B*) Resolution limits depending on the branch length. The same procedure as in (*A*) was followed. Here, however, the known phylogenetic tree of the great apes was used for alignment generation and all branch lengths were multiplied by fixed factors (Edge factor). (*C*) Methylation similarity as a function of CpG neighborhood. We define the x-neighborhood of a CpG as the set consisting of the x-nearest CpG upstream and the x-nearest CpG downstream. Shown are the average similarities (solid line) and standard deviations (dashed line) of the methylation fractions for all corresponding x-neighborhood pairs. For comparison, the average similarity of the orthologous human–chimpanzee methylation fractions is also shown (dotted horizontal line, see also Table 1). (*D*) Local signal resolution/alignment sensitivity. Phylogenetic trees were reconstructed from the real data set (great apes) using modified methylation fractions alignments. The alignments were modified so that methylation fractions were exchanged with a certain probability for a random methylation fraction of the same species within the given exchange distance. (*A*, *B*, *D*) The dashed line indicates the probability of randomly reconstructing the correct tree topology (one-third), solid lines interpolate points for visual guidance. (*A*–*D*) The total number of reconstructed trees per fixed amount of input data (Number of CpG loci) was always 100. The region examined here was the gene body. The total number of evaluable CpG loci in this region is 441,286.

## Transcription Factors Preferentially Bind to Epigenetically Conserved Gene Loci

Based on the observations of high interspecies similarity of CpG methylation (table 1, fig. 3*A* and *B*) and successful tree reconstruction using these data (figs. 3*C* and 4*A*), we calculated gene-wise estimates of the evolutionary rates to illustrate potential benefits of studies into epigenomic conservation. As described above, studies on single-gene level critically depend on the amount of evaluable data. With the data set at hand, the criteria for a thorough single-gene-based hypothesis testing are limited. In the context of this work, we restrict ourselves to representative examples. *CASTOR1* codes for a component of the MTORC pathways (Saxton et al. 2016) and is an example of a modestly conserved gene according to our epigenomic measure (fig. 6*A*), that is the estimated rate of evolution is almost exactly that of the region average. Upon closer inspection, however, we observe that the local methylation level is anticorrelated with the transcription factor binding density, that is the number of different binding transcription factors normalized by gene length, which we determined based on publicly available data (Karolchik et al. 2004). We also found that the rate at which a gene's methylation level evolves is globally anticorrelated with the binding density of the transcription factors ($\rho = -0.22$, $P < 2.2 \times 10^{-16}$, supplementary table S3, Supplementary Material online). As a second intuitive measure for assessing functional aspects of epigenomic conservation, we determined how much the respective gene tree

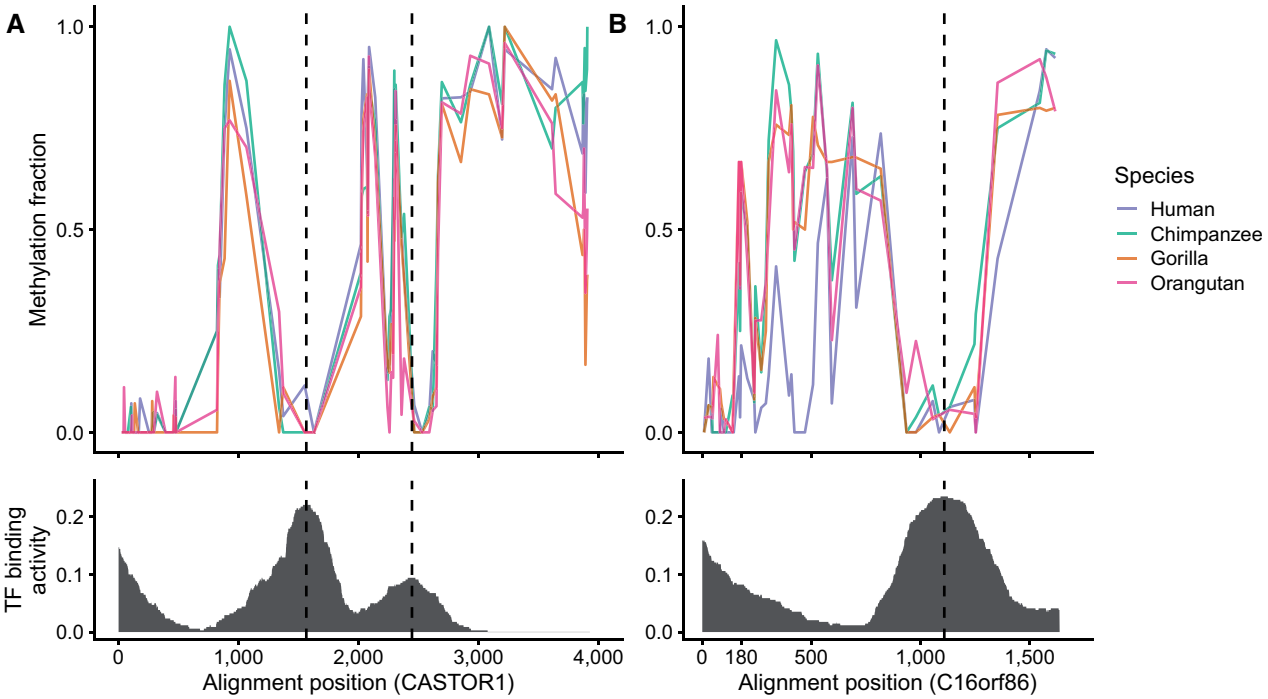


**Fig. 6.** (*A*) Measured methylation fractions and estimated local transcription factor binding density in the gene body of *CASTOR1*. The total number of evaluable CpG loci is 81. Nineteen CpG loci were skipped due to low read support (less than ten reads in at least one species). (*B*) Measured methylation fractions and estimated local transcription factor binding density in the gene body of *C16orf86*. The total number of evaluable CpG loci is 45. Four CpG loci were skipped due to low read support (less than ten reads in at least one species). Dashed vertical lines indicate the co-occurrence of high transcription factor binding activity and low methylation fraction.

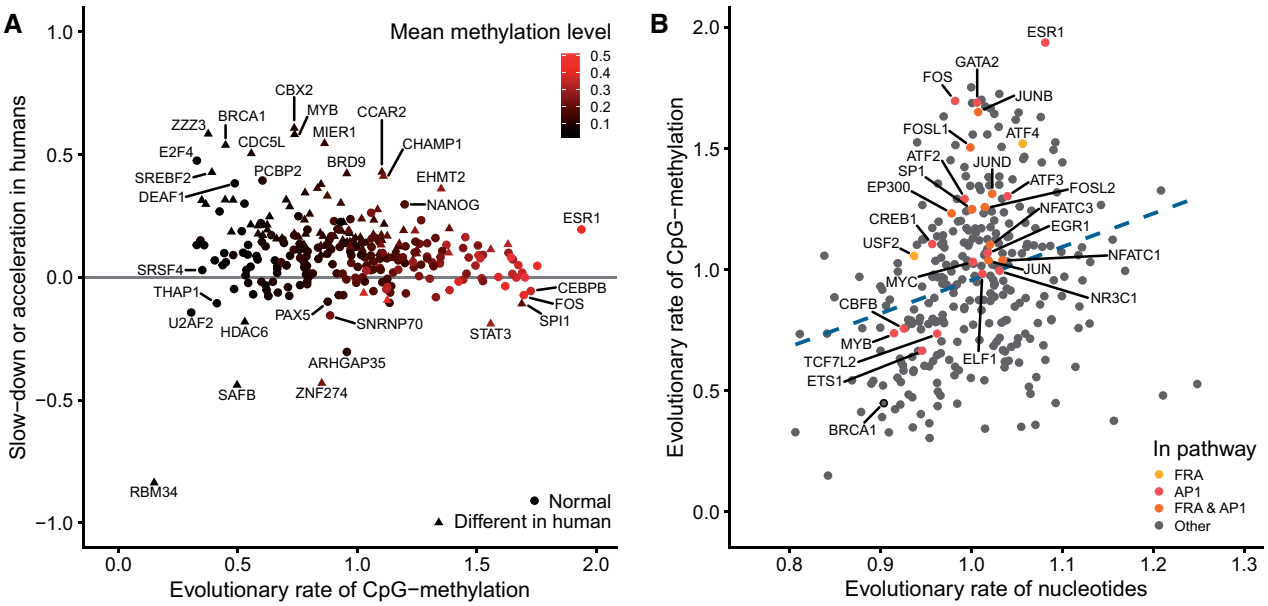


**Fig. 7.** (*A*) Estimated evolutionary rate of TFBS on CpG-methylation level in great apes versus relative slow-down/acceleration of this rate in humans. Triangles mark TFBS with statistically significant deviation of the human evolutionary rate (FDR $< 0.01$, maximum likelihood ratio test, see *Methods* chapter *Methylation at TFBS*). The color codes the mean methylation level across all evaluable CpG sites in the binding sites of the respective transcription factor. (*B*) Estimated evolutionary rate of TFBS on nucleotide versus CpG-methylation level. The dashed line indicates a simple linear regression of the two evolutionary rates. Color-coded are all transcription factors that are part of the pathways listed by the legend (Pathway Interaction Database; Schaefer et al. 2009). (*A*, *B*) Shown are those 297 out of 340 transcription factors from the UCSC hg38 *Transcription Factor ChIP-seq Clusters* track (Karolchik et al. 2004) that comprised at least 1,000 evaluable CpG sites.

deviates from the average tree of the region class (see supplementary methods, Supplementary Material online for details). *C16orf86* codes for a probably functional but so far uncharacterized protein and an example of a strong deviation (fig. 6B). According to the measure we described and the filter criteria we used, the gene shows the highest deviation from the expected tree. According to our model, this deviation can be attributed almost exclusively to an increased evolutionary speed in humans. Interestingly, however, methylation levels in humans were only locally significantly different from other species between nucleotide positions 180 and 530. These are partly coding for a fully conserved domain of the unknown function (pfam15762, DUF4691) and show a relatively high TFBS density in humans. It is tempting to speculate that in the other great apes the relatively higher methylation may lead to a lower transcription factor binding density in this region.

### Hints of Accelerated Epigenomic Evolution for Polycomb Repressive Complex 2 Protein Binding Sites

Based on our finding that transcription factor binding density appears to be negatively correlated with our measures for epigenomic evolution, we were specifically interested in characterizing individual transcription factors in terms of their evolutionary rates. To do this, methylation rates of aligned gene body CpGs overlapping with the binding site of a specific transcription factor (TF) were combined. In total, 297 out of 340 transcription factors comprised at least 1,000 CpG sites—the chosen minimum to include a TF in this analysis. The strongest deceleration on the human branch was observed for RNA-binding-protein 34, RBM34. Although little is known about the function of this gene, its binding to DNA might be influenced by the CpG-methylation.

Significantly accelerated rates (FDR $\leq$ 0.01; see Materials and Methods) were found for about a third of TFs ($n = 98$) in humans (fig. 7A). Strikingly, this set appears to enrich critical components of the polycomb repressive complex 2 (PRC2) complex, namely, YY1, EZH2, HDAC1, HDAC2, BMI1, and SUZ12 (BIOCARTA PRC2 pathway; FDR = 0.092). Binding sites of the two histone deacetylases are also components of the significantly enriched telomerase pathway (Pathway Interaction Database; FDR = 0.0026) additionally comprising accelerated TFBS of XRCC5, MAX, SP1, IRF1, MYC, NBN, NR2F2, E2F1, SAP30, SIN3A, and SIN3B. The strongest accelerations with a more than 50% increased evolutionary rate on methylation level were found for ZZZ3, chromobox homolog protein 2 (CBX2), MYB, CDC5L, MIER1, and BRCA1. The zinc-finger ZZZ3, a component of the Ada-two-A-containing (ATAC) histone acetyl-transferase complex, has recently been described to function as a reader of histones regulating ATAC-dependent promoter histone H3K9 acetylation (Mi et al. 2018). This indicates that evolutionarily driven changes of ZZZ3 binding characteristics could have a decisive impact on species-specific gene expression. The CBX2, also a reader of histone modifications, is a member of the polycomb repressive complex 1 (PRC1) (Vandamme et al. 2011). It contributes to the repression of genes by binding to H3K9me3 and H3K27me3.

One of the strongest evolutionary accelerations on the methylation level can be observed at the TFBS of BRCA1. Notably, the gene itself has been found to undergo a rapid evolution driven by positive selection altering its amino-acid composition as well its noncoding parts (Pavlicek et al. 2004; Lou et al. 2014). Involved in double-strand break repair, BRCA1 is frequently mutated in hereditary forms of human breast cancers. Thus, our data suggest that the accelerated evolution of BRCA1 in humans compared with other great apes (fig. 7A) has a measurable impact on the methylation landscape around its binding sites. At the same time, on the sequence level, BRCA1 binding sites evolution appears to be substantially decelerated (fig. 7B). This marked lack of covariance potentially supports the functional effects of changes in the BRCA1 gene itself. Notably, BRCA1's direct interaction partner, the estrogen receptor ESR1, exhibits a comparably high evolutionary rate across all primates on the methylation level as compared with the sequence of its binding sites (fig. 7A and B). The systematic comparison of sequence-based and methylation-based evolutionary rates reveals an epigenomic acceleration of binding sites for transcription factors involved in neuron differentiation (GO_NEURON_DIFFERENTIATION, FDR = 0.03) and two highly connected pathways, AP-1 (FDR = 0.03, PID_AP1_PATHWAY) and FRA (ii, FDR = 0.02, PID_FRA_PATHWAY), due to the constitutive components of the heterodimeric AP-1, including, for example, FOS, the ATF family, GATA2 and JunB (fig. 7B). The proteins belonging to the AP-1 family play a critical role in numerous cellular processes. Notably, in cooperation with other cell-type-specific factors, it is involved in the activation of cell-type-specific enhancers (Madrigal and Alasoo 2018). Hence, this result may be explained by species-specific cell compositions, systematic environmental or nutritional differences. Nevertheless, it seems possible that the enrichment of these factors may point to combinatorial changes of AP-1 composition during evolution.

## Conclusions

Here, we have systematically transferred classical methods of phylogenetics used to analyze nucleotide and amino acid sequences to the field of epigenomics. Using empirical data from great apes, we demonstrated that phylogenetic trees can be correctly reconstructed from methylation data based on the fundamental principles of maximum likelihood, parsimony, and distance-based approaches. The problems and challenges are similar in many respects to tree reconstruction from nucleotide data, for example, that parsimony and distance-based methods are prone to long-branch attraction. Nevertheless, we found that all examined regions, with the notable exception of enhancers, contain enough phylogenetic information on CpG-methylation level to outcompete reconstruction with nucleotide data. The ability of the enhancers to escape the evolutionary model can be attributed to a relatively small set of CpG sites. The enrichment of these sites in quickly evolving immune-related genes highlights the importance of the epigenome for short-term evolutionary changes.

Based on the empirical data and simulations, we showed that methylation levels are conserved at single CpG resolution. The methylation fraction of a CpG can usually be better predicted from the methylation fraction of an orthologous CpG in a species separated by millions of years of evolution than even from the methylation fraction of the closest CpG in the same species. We also found evidence that epigenetic conservation is associated with enhanced transcription factor binding density. Evolutionary rates on the nucleotide level were found, as expected, to be highly correlated with evolutionary rates on the CpG methylation level across TFBS. However, significant deviations from this general trend are observed for binding sites of transcription factors associated with neuron differentiation and components of the heterodimeric AP-1 evolving significantly faster on the methylation level. Multiple examples provide hints that epigenomic remodelers themselves could be critical components in the evolution of the human lineage. Significantly elevated evolutionary rate on methylation level in humans compared with other great apes were found at TFBS of BRCA1, CBX2, ZZZ3, MIER1, and MYB. On a global level, critical components of the polycomb repressive complex 2 and members of the telomerase pathway show an accelerated CpG methylation evolution in humans.

In the future, when data are collected for different epigenetic marks across multiple tissues, these methods should be helpful to test for accelerated or slowed epigenetic evolution affecting individual genes. Furthermore, other marks, for example, modifications of histone tails, are less codependent on the genomic context compared with CpG-methylation. Thus, such investigations may enable the identification of evolutionary effects less vulnerable to selection processes in the immediate genomic neighborhood. Analyses based on a yet-to-be-developed comprehensive model of genomic and epigenomic evolution promise new insights into mechanisms of epigenetic gene regulation and possibly the formation of phenotypes based on these mechanisms.

## Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

## Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (Grant ID 031L016D, to S.Hof.), the Joachim Herz Foundation (Add-on Fellowships for Interdisciplinary Life Science, to A.S.), and the German Research Foundation (Grand IDs 418080850, 418087534, to S.Hof.).

## Author Contributions

A.S. and S.Hof. conceived the study. A.S. performed the analyses. A.S., S.Hof., and S.Hor. interpreted the results. A.S., P.K., and S.Hof. wrote the manuscript. P.K. visualized the results. S.Hof. acquired the funding and supervised the work.

## Data Availability

The data underlying this article are available at the National Center for Biotechnology Information under the accession PRJNA286277 (Hernando-Herraez et al. 2015), in the article and in its supplementary material, Supplementary Material online.

## References

Appanah R, Dickerson DR, Goyal P, Groudine M, Lorincz MC. 2007. An unmethylated 3' promoter-proximal region is required for efficient transcription initiation. *PLoS Genet*. 3(2):e27.

Arun PVPS, Miryala SK, Chattopadhyay S, Thiyyagura K, Bawa P, Bhattacharjee M, Yellaboina S. 2016. Identification and functional analysis of essential, conserved, housekeeping and duplicated genes. *FEBS Lett*. 590(10):1428–1437.

Barrero MJ, Boué S, Izpisúa Belmonte JC. 2010. Epigenetic mechanisms that regulate cell identity. *Cell Stem Cell*. 7(5):565–570.

Bergmiller T, Ackermann M, Silander OK. 2012. Patterns of evolutionary conservation of essential genes correlate with their compensability. *PLoS Genet*. 8(6):e1002803.

Böck J, Remmele CW, Dittrich M, Müller T, Kondova I, Persengiev S, Bontrop RE, Ade CP, Kraus TFJ, Giese A, et al. 2018. Cell type and species-specific patterns in neuronal and non-neuronal methylomes of human and chimpanzee cortices. *Cereb Cortex*. 28(10):3724–3739.

Boffelli D, Martin DI. 2012. Epigenetic inheritance: a contributor to species differentiation? *DNA Cell Biol*. 31(Suppl 1):S11–S16.

Broyden CG. 1970. The convergence of a class of double-rank minimization algorithms 1. General considerations. *IMA J Appl Math*. 6(1):76–90.

Burggren W. 2016. Epigenetic inheritance and its role in evolutionary biology: re-evaluation and new perspectives. *Biology (Basel)* 5(2):24.

Byrd RH, Lu P, Nocedal J, Zhu C. 1995. A limited memory algorithm for bound constrained optimization. *SIAM J Sci Comput*. 16(5):1190–1208.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

Chen Z, Li S, Subramaniam S, Shyy JYJ, Chien S. 2017. Epigenetic regulation: a new frontier for biomedical engineers. *Annu Rev Biomed Eng*. 19:195–219.

Cui R, Medeiros T, Willemsen D, Iasi LNM, Collier GE, Graef M, Reichard M, Valenzano DR. 2019. Relaxed selection limits lifespan by increasing mutation load. *Cell* 178:385–399.e20.

Ehrlich M, Lacey M. 2013. DNA methylation and differentiation: silencing, upregulation and modulation of gene expression. *Epigenomics* 5(5):553–568.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.

Fitch WM. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. *Syst Biol*. 20(4):406–416.

Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155(3760):279–284.

Fletcher W, Yang Z. 2009. INDELible: a flexible simulator of biological sequence evolution. *Mol Biol Evol*. 26(8):1879–1888.

Gaya-Vidal M, Alba MM. 2014. Uncovering adaptive evolution in the human lineage. *BMC Genomics* 15:599.

Ge RL, Cai Q, Shen YY, San A, Ma L, Zhang Y, Yi X, Chen Y, Yang L, Huang Y, et al. 2013. Draft genome sequence of the Tibetan antelope. *Nat Commun*. 4:1858.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*. 11(5):725–736.

Hasegawa M, Kishino H, Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 22(2):160–174.

Hernando-Herraez I, Heyn H, Fernandez-Callejo M, Vidal E, Fernandez-Bellon H, Prado-Martinez J, Sharp AJ, Esteller M, Marques-Bonet T.

2015. The interplay between DNA methylation and sequence divergence in recent human evolution. *Nucleic Acids Res.* 43(17):8204–8214.
Jordan G, Goldman N. 2012. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol.* 29(4):1125–1139.
Jukes TH, Cantor CR. 1969. Chapter 24—Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Kar S, Deb M, Sengupta D, Shilpi A, Parbin S, Torrisani J, Pradhan S, Patra S. 2012. An insight into the various regulatory mechanisms modulating human DNA methyltransferase 1 stability and function. *Epigenetics* 7(9):994–1007.
Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* 32(Database issue):D493–D496.
Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* 26(17):2204–2207.
Kim M, Costello J. 2017. DNA methylation: an epigenetic mark of cellular memory. *Exp Mol Med.* 49(4):e322.
Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol.* 16(2):111–120.
Kishino H, Miyata T, Hasegawa M. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J Mol Evol.* 31(2):151–160.
Kosiol C, Vinař T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* 4(8):e1000144.
Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al. 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23(21):2947–2948.
Leonhardt H, Page AW, Weier HU, Bestor TH. 1992. A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell* 71(5):865–873.
Lind MI, Spagopoulou F. 2018. Evolutionary consequences of epigenetic inheritance. *Heredity* 121(3):205–209.
Locke DP, Hillier LW, Warren WC, Worley KC, Nazareth LV, Muzny DM, Yang S-P, Wang Z, Chinwalla AT, Minx P, et al. 2011. Comparative and demographic analysis of orangutan genomes. *Nature* 469(7331):529–533.
Lou DI, McBee RM, Le UQ, Stone AC, Wilkerson GK, Demogines AM, Sawyer SL. 2014. Rapid evolution of BRCA1 and BRCA2in humans and other primates. *BMC Evol Biol.* 14:155.
Lowdon RF, Jang HS, Wang T. 2016. Evolution of epigenetic regulation in vertebrate genomes. *Trends Genet.* 32(5):269–283.
Luo H, Gao F, Lin Y. 2015. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. *Sci Rep.* 5:13210.
Madrigal P, Alasoo K. 2018. AP-1 takes centre stage in enhancer chromatin dynamics. *Trends Cell Biol.* 28(7):509–511.
Makova KD, Hardison RC. 2015. The effects of chromatin organization on variation in mutation rates in the genome. *Nat Rev Genet.* 16(4):213–223.
Martin DIK, Singer M, Dhahbi J, Mao G, Zhang L, Schroth GP, Pachter L, Boffelli D. 2011. Phyloepigenomic comparison of great apes reveals a correlation between somatic and germline methylation states. *Genome Res.* 21(12):2049–2057.
Mendizabal I, Shi L, Keller TE, Konopka G, Preuss TM, Hsieh T-F, Hu E, Zhang Z, Su B, Yi SV. 2016. Comparative methylome analyses identify epigenetic regulatory loci of human brain evolution. *Mol Biol Evol.* 33(11):2947–2959.
Mi W, Zhang Y, Lyu J, Wang X, Tong Q, Peng D, Xue Y, Tencer AH, Wen H, Li W, et al. 2018. The ZZ-type zinc finger of ZZZ3 modulates the ATAC complex-mediated histone acetylation and gene activation. *Nat Commun.* 9(1):3759.
Michalak EM, Burr ML, Bannister AJ, Dawson MA. 2019. The roles of DNA, RNA and histone methylation in ageing and cancer. *Nat Rev Mol Cell Biol.* 20(10):573–589.
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(6):1029–1041.
Pavlicek A, Noskov VN, Kouprina N, Barrett JC, Jurka J, Larionov V. 2004. Evolution of the tumor suppressor BRCA1 locus in primates: implications for cancer predisposition. *Hum Mol Genet.* 13(22):2737–2751.
Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, et al. 2006. In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444(7118):499–502.
Perez MF, Lehner B. 2019. Intergenerational and transgenerational epigenetic inheritance in animals. *Nat Cell Biol.* 21(2):143–151.
Qu J, Hodges E, Molaro A, Gagneux P, Dean MD, Hannon GJ, Smith AD. 2018. Evolutionary expansion of DNA hypomethylation in the mammalian germline genome. *Genome Res.* 28:145–158.
Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13:235–238.
Reichwald K, Petzold A, Koch P, Downie BR, Hartmann N, Pietsch S, Baumgart M, Chalopin D, Felder M, Bens M, et al. 2015. Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell* 163(6):1527–1538.
Rosenberg MS, Kumar S. 2001. Traditional phylogenetic reconstruction methods reconstruct shallow and deep evolutionary relationships equally well. *Mol Biol Evol.* 18(9):1823–1827.
Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. *Mol Biol Evol.* 31(7):1661–1685.
Sahm A, Bens M, Szafranski K, Holtze S, Groth M, Gorlach M, Calkhoven C, Muller C, Schwab M, Kraus J, et al. 2018. Long-lived rodents reveal signatures of positive selection in genes associated with lifespan. *PLoS Genet.* 14(3):e1007272.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* 4(4):406–425.
Saxton RA, Chantranupong L, Knockenhauer KE, Schwartz TU, Sabatini DM. 2016. Mechanism of arginine sensing by CASTOR1 upstream of mTORC1. *Nature* 536:229–233.
Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. 2009. PID: the Pathway Interaction Database. *Nucleic Acids Res.* 37(Database issue):D674–D679.
Shavit Grievink L, Penny D, Hendy MD, Holland BR. 2010. Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree. *Syst Biol.* 59(3):288–297.
Shultz AJ, Sackton TB. 2019. Immune genes are hotspots of shared positive selection across birds and mammals. *Elife* 8:e41815.
Sokal RR, Michener CD. 1958. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull.* 38:1409–1438.
Vandamme J, Volkel P, Rosnoblet C, Le Faou P, Angrand PO. 2011. Interaction proteomics analysis of polycomb proteins defines distinct PRC1 complexes in mammalian cells. *Mol Cell Proteomics.* 10:M110.002642.
Verhoeven KJF, vonHoldt BM, Sork VL. 2016. Epigenetics in ecology and evolution: what we know and what we need to know. *Mol Ecol.* 25(8):1631–1638.
Vertino PM, Sekowski JA, Coll JM, Applegren N, Han S, Hickey RJ, Malkas LH. 2002. DNMT1 is a component of a multiprotein DNA replication complex. *Cell Cycle.* 1(6):416–423.
Webb AE, Gerek ZN, Morgan CC, Walsh TA, Loscher CE, Edwards SV, O'Connell MJ. 2015. Adaptive evolution as a predictor of species-specific innate immune response. *Mol Biol Evol.* 32(7):1717–1729.

Xia B, Zhao D, Wang G, Zhang M, Lv J, Tomoiaga AS, Li Y, Wang X, Meng S, Cooke JP, et al. 2020. Machine learning uncovers cell identity regulator by histone code. *Nat Commun*. 11(1):2696.

Xia J, Han L, Zhao Z. 2012. Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. *BMC Genomics* 13(Suppl 8):S7.

Xiao S, Cao X, Zhong S. 2014. Comparative epigenomics: defining and utilizing epigenomic variations across species, time-course, and individuals. *Wiley Interdiscip Rev Syst Biol Med*. 6(5):345–352.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13(5):555–556.

Yi SV. 2017. Insights into epigenome evolution from animal and plant methylomes. *Genome Biol Evol*. 9(11):3189–3201.

Zaheri M, Dib L, Salamin N. 2014. A generalized mechanistic codon model. *Mol Biol Evol*. 31(9):2528–2541.

Zemach A, McDaniel IE, Silva P, Zilberman D. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. *Science* 328(5980):916–919.

Zeng J, Konopka G, Hunt BG, Preuss TM, Geschwind D, Yi SV. 2012. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet*. 91(3):455–465.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*. 22(12):2472–2479.