8204–8214 *Nucleic Acids Research, 2015, Vol. 43, No. 17* *Published online 13 July 2015*
doi: 10.1093/nar/gkv693

The interplay between DNA methylation and sequence divergence in recent human evolution

Irene Hernando-Herraez[1,*,†], Holger Heyn[2,†], Marcos Fernandez-Callejo[3], Enrique Vidal[2], Hugo Fernandez-Bellon[4], Javier Prado-Martinez[1], Andrew J. Sharp[5], Manel Esteller[2,6,*] and Tomas Marques-Bonet[1,3,6,*]

[1]Institute of Evolutionary Biology (UPF-CSIC), PRBB, 08003 Barcelona, Spain, [2]Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), 08908 L'Hospitalet de Llobregat, Barcelona, Spain, [3]Centro Nacional de Análisis Genómico (CNAG), Parc Científic de Barcelona, Barcelona 08028, Spain, [4]Parc Zoològic de Barcelona, Parc de la Ciutadella s/n, Barcelona 08003, Spain, [5]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai School, New York, NY 10029, USA and [6]Catalan Institution of Research and Advanced Studies (ICREA), Barcelona, Spain

Received March 16, 2015; Revised June 17, 2015; Accepted June 25, 2015

ABSTRACT

Despite the increasing knowledge about DNA methylation, the understanding of human epigenome evolution is in its infancy. Using whole genome bisulfite sequencing we identified hundreds of differentially methylated regions (DMRs) in humans compared to non-human primates and estimated that ∼25% of these regions were detectable throughout several human tissues. Human DMRs were enriched for specific histone modifications and the majority were located distal to transcription start sites, highlighting the importance of regions outside the direct regulatory context. We also found a significant excess of endogenous retrovirus elements in human-specific hypomethylated.

We reported for the first time a close interplay between inter-species genetic and epigenetic variation in regions of incomplete lineage sorting, transcription factor binding sites and human differentially hypermethylated regions. Specifically, we observed an excess of human-specific substitutions in transcription factor binding sites located within human DMRs, suggesting that alteration of regulatory motifs underlies some human-specific methylation patterns. We also found that the acquisition of DNA hypermethylation in the human lineage is frequently coupled with a rapid evolution at nucleotide level in the neighborhood of these CpG sites. Taken together, our results reveal new insights into the mechanistic basis of human-specific DNA methylation patterns and the interpretation of inter-species non-coding variation.

INTRODUCTION

A major aim of molecular biology is to understand the mechanisms that drive specific phenotypes. Humans and great apes differ in numerous morphological and cognitive aspects. However, their coding sequences are highly similar and most of the differences are located in non-coding regions (1), making it a challenge to define clear genotype-phenotype associations. It has been proposed that human specific traits originate from gene regulatory differences rather than from changes in the primary genetic sequence (2). The characterization of regulatory domains is therefore a promising strategy to unveil regions of relevance for human evolution and to understand the implications of non-coding variation.

DNA methylation is a key regulatory mechanism of the genome (3). It is present in many taxa and, in mammals, it plays an essential role in numerous biological processes ranging from cell differentiation to susceptibility to complex diseases (4,5). From a mechanistic perspective, DNA methylation has been described as an intermediate regulatory event, mediating the effect of genetic variability on phenotype formation (6). However, the mechanisms by which the DNA methylation profile is generated are poorly understood. DNA methylation function is highly dependent on its location. In promoters, for example, it tends to confer gene repression while in gene bodies it is associated with transcriptional activation (3,7). DNA methylation levels also depend on the underlying genetic sequence and the occu-

*To whom correspondence should be addressed. Tel: +34 93 3160887; Email: tomas.marques@upf.edu
Correspondence may also be addressed to Irene Hernando-Herraez. Tel: +34 93 3160803; Email: irene.hernando@upf.edu
Correspondence may also be addressed to Manel Esteller. Tel: +34 932607140; Email: mesteller@idibell.cat
†These authors contributed equally to the paper as first authors.

pancy of DNA binding factors (8,9). There is therefore no generic rule that can be applied to all biological situations, indicating the high complexity of the DNA methylation regulatory network.

In recent years, due to the development of genome-wide techniques that allow us to analyze DNA methylation profiles in multiple organisms, the field of comparative epigenomics has started to emerge. Exciting questions about how DNA methylation patterns vary through time and how this variation is linked to genome evolution can now be addressed. It has been shown that the global pattern of DNA methylation between close species, such as human and chimpanzee, is similar (10). Nonetheless, there is a special interest in the study of local changes as mechanisms of species evolution, specially of human evolution (11). Previous studies have identified several differentially methylated regions between human and primates using different techniques (12–19). Interestingly, many of these regions have been associated not only with tissue-specific functions, but also with developmental and neurological mechanisms (13,15,19). A key question that arises is how the epigenetic variability is generated and transmitted across generations. The best studied mechanism is the dependence of DNA methylation levels on the genetic sequence. A growing number of studies in humans have shown an association between a nucleotide variant and a state of methylation (6,20). However the relationship between the genetic and the epigenetic sequence has not been explored when studying different species and despite recent advances in the field, many unanswered questions remain: How do DNA methylation patterns diverge across different genomic features? What are the processes driving such differences? Is this epigenetic variation associated with a higher rate of nucleotide substitution?

To further investigate these questions, we determined blood DNA methylation patterns in human, chimpanzee, gorilla and orangutan samples using whole genome bisulfite sequencing. Because this technique is not dependent on predefined sequences or methylation-dependent restriction enzymes, it is superior to other assays in analyzing patterns of DNA methylation (10,12,13). We identified hundreds of human regions that differ in the DNA methylation pattern compared to the rest of great apes. These regions were enriched for specific histone modifications and they were located distal to transcription start sites. Furthermore, we found that DNA methylation variation and the underlying genetic code show close physical dependencies.

MATERIALS AND METHODS

The bisulfite sequencing data discussed in this publication have been deposited to the NCBI SRA database under the accession number SRP059313.

Sample collection

Human and non-human research has been approved by the ethical committee of the European Research Union. Human donors gave written informed consent to take part in the study. We obtained methylation data from peripheral whole blood DNA extracted from a human, a chimpanzee, a gorilla and an orangutan sample (Supplementary Table S1). Furthermore we obtained DNA methylation data from a human CD19+ sample. All samples were obtained from healthy donors and DNA was extracted using phenol:chloroform:isoamylalcohol (Sigma). CD19+ sample was separated using the CD19+ cell Isolation kit II (Miltenyi Biotec) following the manufacturer's instructions. DNA methylation data from additional samples (CD4+ and solid tissues) were obtained from previous publications, (21) and (GSE46698), respectively. Monocyte and neutrophil data were retrieved from Blueprint portal (http://dcc.blueprint-epigenome.eu/#/md/data).

Library preparation

We spiked genomic DNA (1 or 2 μg) with unmethylated λ DNA (5 ng of λ DNA per μg of genomic DNA) (Promega). We sheared DNA by sonication to 50–500 bp with a Covaris E220 and selected 150–300 bp fragments using AMPure XP beads (Agencourt Bioscience Corp.). We constructed genomic DNA libraries using the TruSeq Sample Preparation kit (Illumina Inc.) following Illumina's standard protocol. After adaptor ligation, we treated DNA with sodium bisulfite using the EpiTect Bisulfite kit (Qiagen) following the manufacturer's instructions for formalin-fixed and paraffin-embedded (FFPE) tissue samples. We performed two rounds of conversion to achieve >99% conversion. We enriched adaptor-ligated DNA through seven cycles of polymerase chain reaction (PCR) using the PfuTurboCx Hotstart DNA polymerase (Stratagene). We monitored library quality using the Agilent 2100 BioAnalyzer (Agilent) and determined the concentration of viable sequencing fragments (molecules carrying adapters at both extremities) by quantitative PCR using the Library Quantification Kit from KAPA Biosystems. We performed paired-end DNA sequencing (two reads of 100 bp each) using the Illumina Hi-Seq 2000. Sequencing quality was assessed using the Illumina Sequencing Analysis Viewer and FastQC software. We ensured the raw reads used in subsequent analyses were within the standard parameters set by the Illumina protocol. Positional quality along the reads was confirmed to be QC>30, and we excluded biases toward specific motifs or GC-enriched regions in the PCR amplification or hybridization.

Mapping and annotation

Paired-end sequencing reads (100 bp) were mapped to the *in silico* bisulfite–converted human (hg19), chimpanzee (panTro4) (1), gorilla (gorGor3) (22) and orangutan (ponAbe2) (23) references genomes using Bismark v0.7.8 (24) not allowing multiple alignments. We also removed potential PCR duplicates using Bismark's deduplicate_bismark program. Custom Perl scripts were used to summarize the methylation levels of individual cytosines based on frequency of mapped reads.

To facilitate an unbiased comparison of the four genomes we used the Enredo-Pecan-Orthus (EPO) whole-genome multiple alignments of human, chimpanzee, gorilla and orangutan [Ensemble Compara.6_primates_EPO] (25). We identified 8,952,000 CpG positions shared among the four

species in autosomal chromosomes, this data set was used for further analysis.

Global methylome analysis

We used 5,946,947 CpG sites presenting a read coverage between 4X and 30X in all species to perform global methylome comparisons according to their genomic annotation. Promoter regions were defined as ±2 kb interval of the transcription start site. CpG island and repeat families were annotated using human UCSC Genome Browser tracks (26). We used incomplete lineage sorting (ILS) coordinates previously described (22). To assess significance of the ILS clustering a permutation test was performed as follows: based on the number of CpG sites in ILS regions, [221 908 for ((CG), H)) regions and 142,231 for ((HG)C))], we randomly sampled from the total set of shared CpG sites across species (5,946,947 CpG sites) and determined the species clustering. This process was repeated 10,000 times to create the null distribution. The *P*-value corresponded to the number of times the ILS clustering appeared within the null distribution divided by the number of permutations (n = 10,000). We also calculated the number of CpG sites in which the absolute methylation difference between the closest species (CG or HG) was smaller than between comparisons with the third species and then compared it to the null distribution.

Identification of differentially methylated regions

Methylation values and number of reads in each position were used to identify hypomethylated regions (HMRs) using each reference genome coordinates by using a two-state Hidden Markov model (15). The algorithm was developed to assess the methylation profile in humans and chimpanzees by dividing the methylome into regions of hypermethylation and hypomethylation. Non-human HMRs coordinates were converted hg19 coordinates using the EPO alignments. To call hypomethylated DMRs we first intersected a species HMRs with the other three methylomes and performed inter-species comparisons. To call hypermethylated DMRs, we intersected three species HMRs and then compared the methylation patterns to the methylome of the species of interest.

In order to define a species-specific differentially methylated region (DMR), we required a stringent threshold of > 0.3 in mean CpG methylation difference and a minimum of 5 CpGs (coverage between 4X and 30X) in all species. On the one hand, we observed no differences between genome-wide and DMR read coverage distribution (Supplementary Figure S1). On the other hand, since methylation values can be interpreted as the percentage of CpG methylation at a given site, a difference of 0.3 in CpG methylation indicates that there has been a change of methylation in 30% of the molecules tested. The proportion of cells present in the blood, being predominately neutrophils and lymphocytes, has comparable proportions in chimpanzee, gorilla and orangutan (27,28) (Supplementary Table S2), being the highest differences between human and gorilla neutrophils (20%). Therefore, as our DMR analysis required a mean CpG methylation difference > 0.3, changes in blood cell fractions have unlikely affected our results.

Pyrosequencing

Specific sets of primers for PCR amplification and sequencing were designed using a specific software pack (PyroMark assay design version 2.0.01.15). PCR was performed under standard conditions with biotinylated primers and the PyroMark Vacuum Prep Tool (Biotage, Sweden) was used to prepare single-stranded PCR products according to manufacturer's instructions. Pyrosequencing reactions and methylation quantification were performed in a PyroMark Q96 System version 2.0.6 (Qiagen) using appropriate reagents and recommended protocols.

Genomic divergence and TFBS

We computed lineage specific nucleotide substitutions by extracting EPO multi-alignments blocks of human DMRs and flanking regions. Flanking regions were chosen with length equal to DMRs and located from 1 to 5 kb upstream and downstream of DMRs. We then calculated the number of lineage specific nucleotides and divided by the amount of nucleotides present in the four species. Insertions and deletions were not taken into account in this analysis. Transcription factor binding sites coordinates were previously identified (29) and human specific substitutions were also calculated using EPO multi-alignments blocks.

Histone modification enrichment

The genomic distribution shown in Figure 4A, was performed considering the human hg19 RefGene annotation using PAVIS (30). We used processed ChIP-seq data previously published (31). To determine enrichment and significance of a particular modification, we generated 100 control sets sized-matched of the human hypo- and hypermethylated DMRs independently. To generate this control data set we also took into account chromosome location, CpG density and length. Next, we determined the proportion of each histone codification overlapping the human DMRs and the control data sets. The ratio of the two is reported as enrichment shown in Figure 4D. The *P*-value corresponded to the number of times that the DMRs proportions appeared in control data set distribution, divided by the number of sets (n = 100). Similarly, to determine the significance of DMRs location we calculated the proportion of DMRs ± 30 kb around TSS (RefSeq genes) and compared to the control data set distribution. The *P*-value corresponded to the number of times that the DMRs proportion appeared in control data set distribution, divided by the number of sets (n = 100).

RESULTS

We performed whole genome bisulfite sequencing of whole blood derived DNA from a human (*Homo sapiens*), a chimpanzee (*Pan troglodytes*), a western gorilla (*Gorilla gorilla*) and a Sumatran orangutan (*Pongo abelii*) individual. A total of ∼ 1.6 billion 100 bp Illumina paired-end reads were uniquely aligned to their respective reference genomes (hg19, panTro4 (1), gorGor3 (22), ponAbe2 (23)) using Bismark (24). To facilitate an unbiased comparison between the four species, we performed all inter-species comparisons

based on 6-primate EPO (25) restricting our analysis to 8 952 000 CpG sites conserved between the four species (see Materials and Methods). Compared to previous studies, this approach allowed us to reliably analyze a greater proportion of the species epigenomes. The read coverage in this subset of CpG sites averaged 10X in human, 12X in chimpanzee, 12X in gorilla and 13X in orangutan (Supplementary Figure S2).

A global view of great Ape methylomes

Overall, the four species exhibited similar levels of DNA methylation with average levels of 74% in human, 71% in chimpanzee, 71% gorilla and 70% in orangutan samples (Figure 1A). These findings are comparable to levels reported in previous studies analyzing blood methylomes (21,32). To investigate the epigenetic divergence between species, we retained 5,946,947 CpG sites that had between 4X and 30X coverage in all species and performed correlation analysis in different regions of the genome. Here, the correlation coefficients of DNA methylation levels between species were in agreement with species phylogeny, the highest being in human-chimpanzee comparisons and the lowest in all comparisons involving orangutan (Figure 1B). From a genomic perspective, DNA methylation values correlated notably in promoters and CpG island regions and to a lesser extent at repeat loci (Figure 1B). Among the major repeat families, *Alu* elements presented the lowest correlation coefficients between species (Figure 1B). Only high confidence reads mapping uniquely to orthologous regions were considered, wherein no major differences in coverage were observed (Supplementary Figures S3 and S4). However, due to the exclusion of multi-mapping read information and the high frequency of C>T mutations resulting from CpG deamination in repetitive elements, further studies are required to confirm the significance of these findings.

To gain insights into the relationship between nucleotide sequence and DNA methylation levels, we analyzed the DNA methylation patterns of regions whose sequence genealogy differs from the species phylogeny, known as regions of incomplete lineage sorting (ILS) (22). Specifically, we studied regions where humans are more closely related to gorillas than to chimpanzees, represented as ((H,G)C)O) and regions where chimpanzees are more closely related to gorillas than to humans, represented as ((C,G)H)O). These regions are typically small (average length 473 bp) and contained 364,139 CpG sites conserved among all four species (Supplementary Figure S5). Interestingly, hierarchical clustering and correlation analyses showed incomplete lineage sorting also at DNA methylation levels (Figure 1C) ($P < 0.0001$). Furthermore we estimated that 25% and 33% of CpG sites within ((C,G)H)O) and ((H,G)C)O) regions respectively, are in concordance with the ILS pattern ($P < 0.0001$). These results suggest a physical interplay between the methylation levels of a substantial fraction of CpG sites and the genetic sequence. While environmental heterogeneity can contribute to epigenetic variation (33) and therefore, it is a common confounding factor when comparing epigenomes of different species, the study of ILS regions allowed us to overcome this limitation. In addition, although 80% of the analyzed regions are not located either at promoters or within coding sequences, it has been shown that nearby genes present higher expression divergence between human and chimpanzee (22).

Species-specific DNA methylation patterns

We then focused our study on species-specific regions, which present a DNA methylation pattern exclusive to a single species. Therefore, we first identified hypomethylated regions (HMRs) throughout the genomes using a Hidden Markov Model (15). This algorithm has been previously applied on human and chimpanzee DNA methylomes to detect putative regulatory regions (15,34,35). We identified 28 835 (34.9 Mb) HMRs in human, 29 257 (33.6 Mb) in chimpanzee, 30 782 (36.5 Mb) in gorilla and 27 349 (33.1 Mb) regions in orangutan DNA methylomes. These hypomethylated regions were similar in size and methylation levels in all species (Supplementary Figure S6) and harbored ∼15% of the CpG sites tested. Interestingly, an average of 72% (24.9 Mb) of HMRs were shared among all species and were mostly located in or close to human CpG islands (42.6% in CpG islands and 52.6% in CpG shores).

The resulting hypomethylated blocks were used to perform DNA methylation inter-species comparisons (see Materials and Methods). Due to the epigenomic differences between blood cell types (36), we required a stringent threshold of > 0.3 in mean CpG methylation difference (at least 30% methylation difference) to define a species-specific differentially methylated region (DMR) (see Materials and Methods). Moreover, this threshold allowed us to identify potential variant regions with higher biological impact. We defined two categories of DMRs: hypomethylated regions (in which one species is uniquely hypomethylated) (Supplementary Table S3) and hypermethylated regions (in which one species is uniquely hypermethylated) (Supplementary Table S4). Overall, we identified 360 hypomethylated DMRs in human (1.2% of HMRs), 340 in chimpanzee (1.1% of HMRs), 845 in gorilla (2.7% of HMRs) and 1015 in orangutans (4.2% of HMRs) (Figure 2A). Further, we determined 210 DMRs specifically hypermethylated in human, 124 in chimpanzee, 167 in gorilla and 698 in orangutans (Figure 2A). One limitation of the method when calling a species hypermethylated DMRs is the intersection of multiple HMRs (from the other species) what resulted in smaller and fewer hypermethylated DMRs. Interestingly, species-specific hypomethylated regions were smaller in size (Wilcoxon test; $P < 0.01$) and had lower CpG density (Wilcoxon test; $P < 0.01$) compared to HMRs common to the analyzed species. Due to the methodological limitations when calling hypermethylated DMRs, we could not assess size differences in this data set.

Importantly, we validated the results in an independent cohort of 48 individuals (31 humans, 5 chimpanzees, 6 gorillas and 6 orangutans), confirming differential DNA methylation in 88% (14/16) of randomly selected human DMRs, underlining the reliability of the genome-wide screening approach (Supplementary Figure S7). An interesting example is represented by an intergentic regions (Figure 2B) specifically hypomethylated in human whole blood compared to chimpanzee, gorilla and orangutan. Interestingly, we also observed hypomethylation at this region when com-

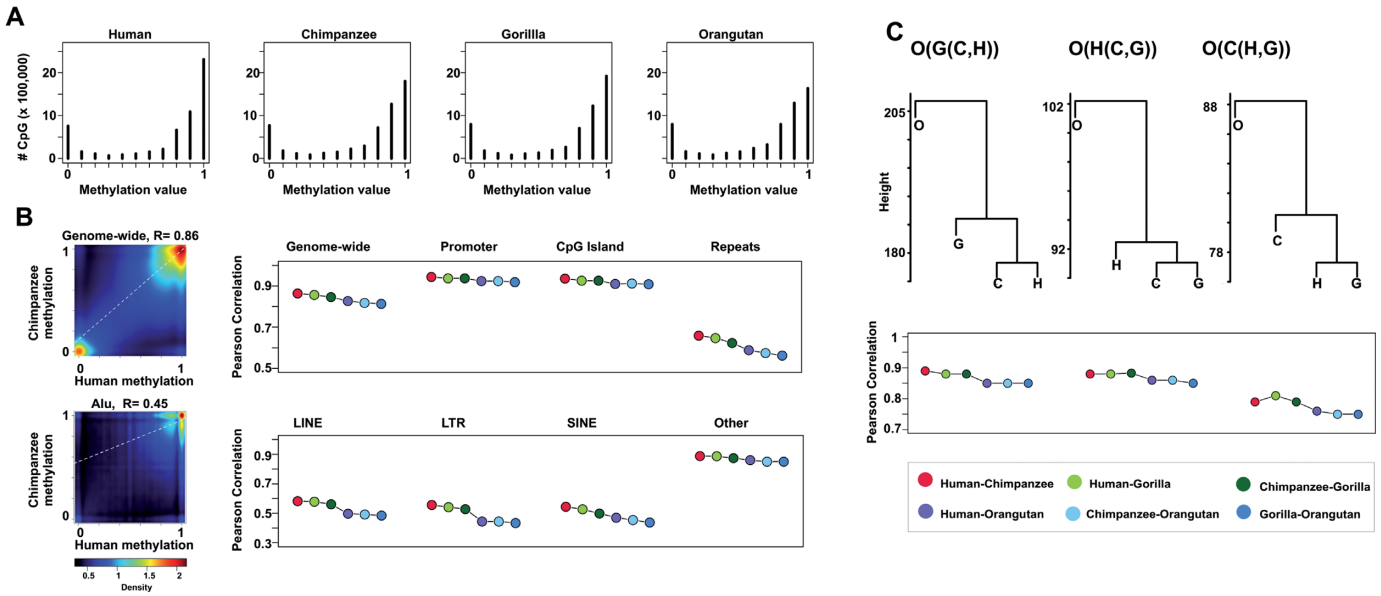

Figure 1. Global DNA methylation patterns. (**A**) DNA methylation profile of 5 946 947 CpG sites shared among the four species. (**B**) Pairwise-correlation analysis in different regions of the genome (right). Genome-wide $n = 5\,946\,947$, promoter $n = 1\,466\,948$, CpG Island $n = 740\,153$, repeats $n = 2\,310\,842$, LINE $n = 433\,317$, LTR $n = 385\,009$, SINE 141 380, *Alu* 1 160 930, other $n = 190\,206$. Density scatterplot of DNA methylation levels between human and chimpanzee genome-wide and in *Alu* elements (left), R indicates the Pearson correlation coefficient. (**C**) Hierarchical cluster tree and pairwise-correlation analysis based of methylation data from incomplete lineage sorting regions. O(G(C,H) $n = 922\,701$, O(H(C,G)) $n = 221\,908$, O(C(H,G)) $n = 142\,231$.

paring with published human methylomes (lymphoid (21) and myeloid cell types, and three other solid tissue types: brain, placenta and liver (37)]), indicating that this pattern is independent of cell types and likely to be conserved during development. Additional information (31) suggests that this region acts as strong enhancer (lymphocyte and lung fibroblast), weak enhancer (mammary epithelial cells) and weak promoter (myoblasts, umbilical vein endothelial cells, embryonic stem cells and keratinocytes). An illustrative example of a human specific hypermethylated DMR which is conserved across both human hematopoietic cell types and solid tissues, is represented by the last exon and 3′ UTR of the *SEMA6C* gene (Figure 2C). This gene encodes a member of the semaphorin family involved in axonal growth and synaptic connectivity maintenance (38). This regions is annotated as weakly transcribed region (lymphocyte, myoblast, lung fribroblast and umbilical vein endothelial cells,) and as inactive promoter (mammary epithelial cells, embryonic stem cells and keratinocytes)(31). Overall, we determined a strong correlation in the DNA methylation profile of human DMRs between the human whole blood sample and major hematopoietic cell types (Pearson's correlation test, $r^2 > 0.8$; Figure 2D). Herein, 66% of human hypomethylated DMRs and 64% of human hypermethylated DMRs were also detectable in the sorted blood cell types (mean difference < 0.3 between human whole blood and all cell types). In addition, 20% and 36% of human hypo- and hypermethylated DMRs, respectively, were detectable in all human tissues including solid cell types (brain, placenta, liver) (Figure 2D).

Genomic divergence in differentially methylated regions

We further investigated the relationship between epigenetic and genetic evolution. First, we evaluated evolutionary conservation within human DMRs and flanking regions using the PhastCons score (Cons 46-Way) from all species (vertebrate) and two subsets (primate and placental mammal) (39,40). Interestingly, human hypo- and hypermethylated DMRs are more evolutionary conserved than the flanking regions (Supplementary Figure S8 and Figure 3). The conservation in deep phylogenies suggests that these regions may have important biological functions. Second, we aimed to determine the association between changes in DNA methylation and changes in the underlying genetic sequence. Therefore, we used the EPO multi-alignments blocks (25) to calculate lineage specific nucleotide substitutions that occurred in human DMRs and at their flanking regions (see Materials and Methods). We observed that human hypermethylated DMRs accumulated nucleotide substitutions in the same branch where the DNA methylation change occurred, clearly suggesting the epigenetic evolution to be coupled with nucleotide changes in these regions (Wilcoxon test, $P < 0.05$; Figure 3A). Moreover, due to the fact that hypermethylated cytosines deaminate spontaneously, we hypothesized a decrease in the number of CpG sites in the hypermethylated species as result of C>T mutations. However, no significant differences in CpG density were observed between species (Figure 3B) and no increase at C>T mutation was observed when classifying the human specific substitutions (Figure 3C), demonstrating that the increase of nucleotide substitutions is not due to cytosine deamination. Surprisingly, we instead observed an increase in the frequency of C>G mutations within human hypermethylated DMRs. Previous studies has pointed to oxida-

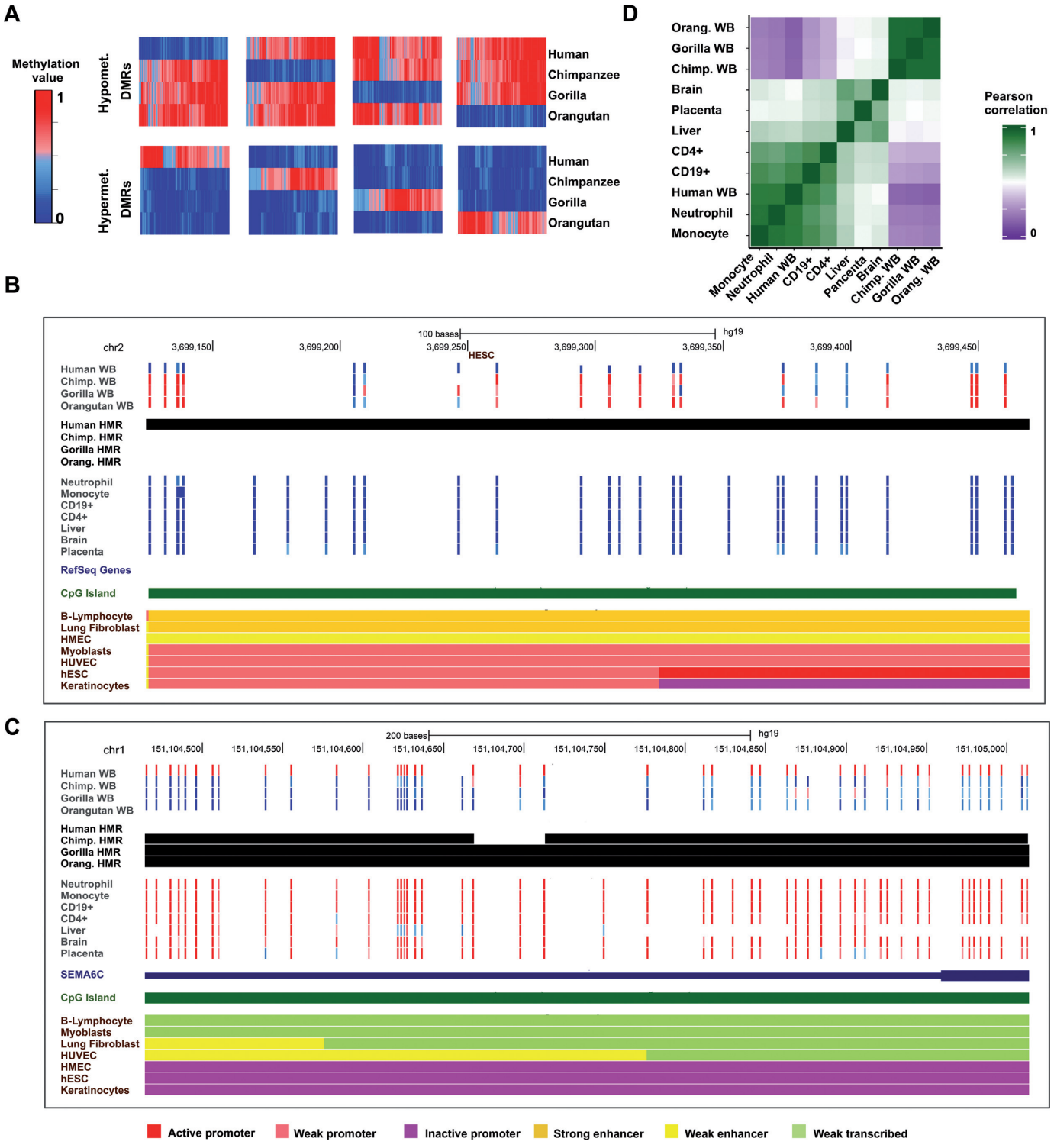

Figure 2. Differentially methylated regions. (**A**) Heat maps showing species specific hypo- (top) and hypermethylated (bottom) DRMs. Each vertical line represents the mean methylation value of a region. (**B**) Browser representation of human hypomethylated DMR (**C**) human hypermethylated DMR, within *SEMA6C*. Each vertical bar shows the methylation value of a single CpG site. Black blocks correspond to hypomethylated regions (HMRs) called by the Hidden Markov Model algorithm. Human samples: WB (whole blood), monocyte and neutrophil (myeloid lineage), CD19+ and CD4+ (lymphoid lineage), liver, brain and placenta. Non-human samples: WB: whole blood. The bottom panel displays the chromatin-state segmentation track (ChromHMM) for 7 different cell types (B-lymphocyte, lung fibroblast, HMEC: mammary epithelial cells, skeletal muscle myoblasts, HUVEC: umbilical vein endothelial cells, hESC: embryonic stem cells, epidermal keratinocytes). (**D**) Pearson correlation matrix of human hypo- and hypermethylated DMRs.

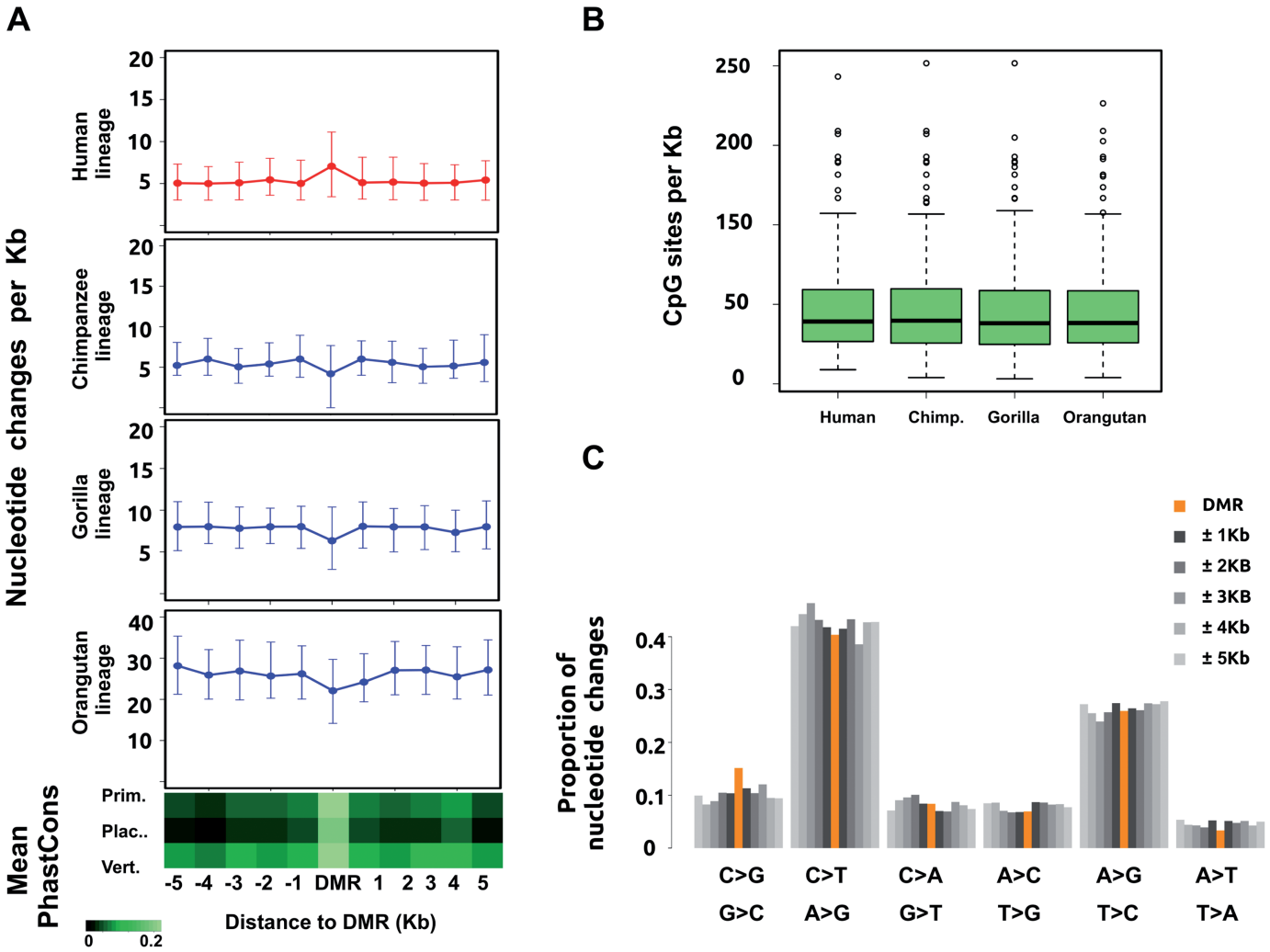

Figure 3. Nucleotide divergence at human hypermethylated DMRs. (**A**) Top, nucleotide changes of human hypermethylated DMRs estimated in each species lineage. The color plot represents the methylation state of the lineage species, red hypermethylated and blue hypomethylated. Data are represented as mean ± 2 standard deviations above and below the mean. Bottom, PhastCons score (Cons 46-Way) from all species (vertebrate) and two subsets (primate and placental mammal). (**B**) Number of CpG sites per kb in human hypermethylated DMRs estimated at each species lineage. (**C**) Classification of human-specific substitutions showing an excess of C>G mutations at human hypermethylated DMRs compared to the flanking regions.

tive conditions as the cause of this type of mutation (41), however further studies are required to interpret our finding.

In contrast to the association of the genetic and epigenetic code in hypermethylated DMRs, human specific hypomethylated DMRs did not show significant differences in the rate of nucleotide substitutions compared to their flanking regions (Supplementary Figure S8), suggesting that alternative mechanisms are implicated in the evolutionary loss of CpG methylation.

Functional context of human DMRs

To further investigate the mechanistic links between genetic and epigenetic changes we studied the DNA sequence of a high-confidence set of predicted transcription factor binding sites (TFBS) located within human-specific DMRs inferred using the CENTIPEDE algorithm (29). We first identified 699 and 274 TFBS overlapping with hypo- and hypermethylated DMRs respectively, and compared these to 752 143 TFBS present in the 5 946 947 CpG sites data set (background). We then identified the proportion of binding sites whose DNA sequence is conserved between species and the proportion of binding sites containing at least one human specific change (Figure 4A). Within human-specific DMRs we observed a significant increase in the frequency of human-specific substitutions in predicted TFBS when compared with TFBS in the background set ($P < 0.001$ hypomethylated DMRs and $P = 0.008$ hypermethylated DMRs; Supplementary Figure S9), indicating a close evolutionary relationship between predicted TFBS and local DNA methylation patterns. However, it is of note that TFBS in hypermethyldated DMRs can be affected by the overall increased nucleotide substitution rate in these regions.

We also examined the presence of common repetitive elements within human-specific DMRs. Interestingly, we found that 399 CpG sites located at human hypomethylated

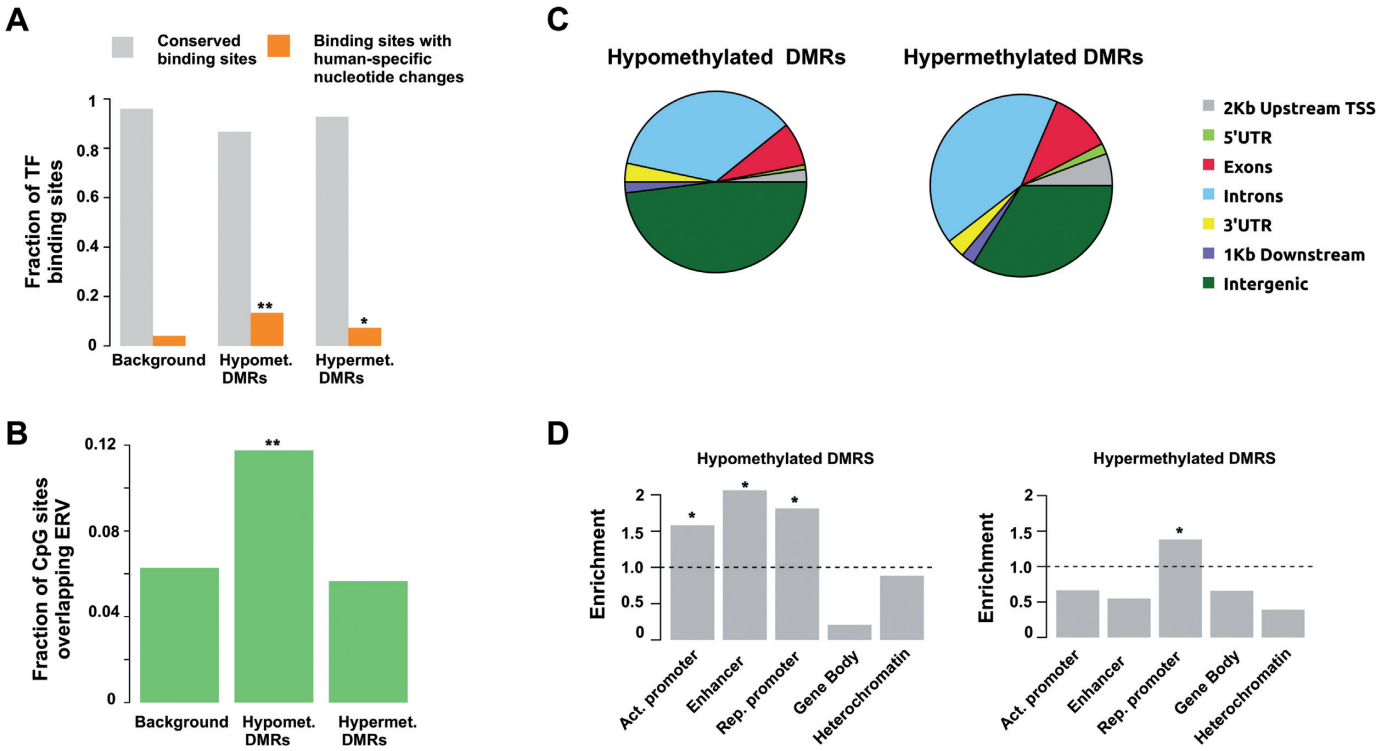

Figure 4. Characteristics of human DMRs. (**A**) Increase of human-specific substitutions in TFBS within DMRs compared with TFBS in the background set. Conserved binding sites (gray) and binding sites with human-specific changes (orange). (**B**) Fraction of CpG sites overlapping with ERV elements (**C**) Distribution of human hypo- and hypermethylated DMRs. (**D**) Histone modification enrichment at human hypo- and hyper DMRs. Active promoter: H3K9ac, enhancer: H3K4me1, repressive promoter: H3K27me3, gene body: H3K36me3 and heterochromatin: H3K9me3 **denotes $P < 0.001$ and * denotes $P < 0.01$ (permutation test).

DMRs, representing 12% of all CpG sites located within hypomethylated human DMRs, overlapped with endogenous retroviruses (ERVs) (Figure 4B). This represents a two-fold enrichment over the background ($P < 0.001$; Supplementary Figure S10). Recent studies indicate that ERV elements participate in transcriptional regulation during mammalian development (42). Hence, our results suggest that ERV methylation levels could also be an important driving force in shaping primate epigenomes.

In total, 13.6% and 21.0% of hypo- and hypermethylated human DMRs, respectively, overlapped with promoter or exonic regions (Figure 4C). To investigate the functional role of the human DMRs, we determined their co-localization with histone mark occupancy data derived from chromatin immunoprecipitation sequencing (ChIP-seq) experiments in whole blood (31). Specifically, we integrated DMRs with histone modifications data marking promoter, enhancer, gene body and heterochromatic regions (H3K9ac, H3K4me1, H3K27me3, H3K36me3 and H3K9me3). Here, we found that 52% of hypomethylated DMRs and 50% of hypermethylated DMRs co-localize with at least one histone modification. To determine a significant enrichment of a particular modification, we generated a set of random DMRs taking into account size, length, chromosomal location and CpG density. We observed that hypomethylated DMRs overlapped significantly with the regulatory histone marks H3K9ac, H3K27me3 and H3K4me1, regions with putative functions in developmental processes (43), whereas hypermethylated DMRs were significantly enriched at loci marked by H3K27me3 (Permutation test $P < 0.01$; Figure 4D). In addition, we found that human DMRs were located distal to transcription start sites (TSS) compared to the distribution of the random DMRs (Supplementary Figure S11). In particular, 44% and 37% of hypo- and hypermethylated DMRs, respectively, were located > 30 kb away from the closest TSS (random DMRs: mean hypo = 17%, mean hyper = 28%) (Permutation test $P < 0.01$ and $P = 0.01$, respectively). We therefore conclude that human DMRs are significantly enriched in regions occupied by active histone marks and are located outside the proximal gene regulatory context. Previous studies have shown that tissue regulatory events are mainly mediated by distal enhancers (32,44). Here, we propose enhancer and bivalent regions (H3K4me1 and H3K27me3) not only to be involved in the determination of cellular phenotypes, but also species phenotypes. However, further studies comparing DNA methylation maps from different tissues are required to understand tissue-specificity from an evolutionary point of view.

DISCUSSION

The current study provides one of the first genome-wide comparison of genetic and epigenetic variation among humans and our closest living relatives. Previous studies have analyzed a limited proportion of the genome using array methods or methylation-dependent restriction enzymes

(6,10,12,13,17,18). Moreover the here applied use of EPO-alignments allowed us to cover a greater proportion of the epigenome in comparison to studies restricted to orthologous genes (14).

We found an overall conservation of the DNA methylation profiles between species. In concordance with previous studies (13,45,46), our results showed high conservation of DNA methylation levels specifically at CpG islands and gene promoters. Nevertheless, we found 570 regions that presented an exclusive pattern of DNA methylation in humans and contrary to expectation, these tend to be located distally to transcription start sites. This fact has to be taken into account in future evolutionary research, as to date most studies focused on the role of promoter DNA methylation and gene silencing. In this context, it was described that differences in promoter methylation underlie only 12–18% of differences in gene expression levels between humans and chimpanzees (12). However, in our genome-wide study we have observed that most human-specific changes occur outside gene promoters. Because distal regulatory elements may contribute to transcriptional activity we hypothesize that the proportion of differences in expression levels explained by DNA methylation may be higher when analyzing whole-genome data sets. In this sense, this epigenetic variation could underlie tissue differences between species. Nevertheless, we have shown that 20% and 36% of human hypo- and hypermethylated DMRs, respectively, were detectable throughout several human tissues, suggesting their conservation during development. This phenomenon could also explain the discrepancy between certain DMRs and tissue function (for example, DMRs associated to neuronal genes detected in blood; Figure 2C) and highlights the importance of developmental and cell differentiation processes in the generation of species-specific traits.

Inter-species DNA methylation divergence can be a consequence of genetic differences between species but also a consequence of sequence-independent mechanisms, such as environmental factors or stochastic events. In non-model organisms these effects are often difficult to exclude and therefore, the observed DNA methylation differences between species may be a consequence of environmental factors on that lineage. In order to homogenize the different environments, all non-human samples in this study were obtained from zoos. However as additional external factors are challenging to control for (e.g. the intake of cooked food in humans), we cannot finally exclude that different environments have partially confounded the results. Importantly, we have replicated the results in larger cohorts of donors from all species and shown a close physical relationship between the genetic and the epigenetic code in three different analyses. Firstly, we have shown ILS of DNA methylation levels in regions that do not follow the species tree, suggesting a dependence of DNA methylation state on the underlying genetic sequence. Secondly, we have determined that a substantial and significant proportion of transcription factor binding sites at human DMRs contain human-specific mutations. This suggests a mechanistic link between the modification of binding sites at the nucleotide level and alterations of DNA methylation (47). Thirdly, we found that the acquisition of DNA hypermethylation in the human lineage is frequently coupled with a rapid evolution at nucleotide level in the neighborhood of these CpG sites. This would initially suggest a loss of functionality of these regions and the subsequent accumulation of mutations. However, the observation of an enrichment for specific histone modification contradicts this hypothesis and rather point to a complex regulatory mechanism between histone modifications, DNA methylation and the underlying genetic sequence. The relationship between the species genetic background and the epigenetic code identified here also indicates that in most of these regions DNA methylation changes are not generated by stochastic events, environmental factors or cell type composition. The genetic-epigenetic association also suggests that DNA methylation patterns at these regions are a fixed feature in the human epigenome and excludes potential bias due to our limited sample sized. However, further studies are needed to accurately determine the exact genetic origin of the inter-species epigenetic variation. This will probably require larger sample sizes, the combination of individual genetic and epigenetic data sets and direct experimentation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We gratefully acknowledge Teresa Abello from Barcelona Zoo and Christina Hvilsom from Copenhagen Zoo for providing non-human samples.

FUNDING

We acknowledge support from AGAUR (Generalitat de Catlunya, Spain) and the Barcelona Zoo (Ajuntament de Barcelona) for an award to I.H.H. H.H. is a Miguel Servet (CP14/00229) researcher funded by the Spanish Institute of Health Carlos III (ISCIII).T.M.B. and M.E. are ICREA Research Professors. Funding for open access charge: European Research Council (ERC), grant EPINORC, under agreement No. 268626; MICINN Projects—SAF2011-22803 and BFU2011-28549; Cellex Foundation; European Community's Seventh Framework Programme (FP7/2007-2013), grant HEALTH-F5-2011-282510—BLUEPRINT, and the Health and Science Departments of the Generalitat de Catalunya.
Conflict of interest statement. None declared.

REFERENCES

1. Sequencing,T.C. and Consortium,A. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, **437**, 69–87.
2. King,M. and Wilson,A.C. (1975) Evolution at Two Levels. *Science*, **188**, 107–116.
3. Jones,P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, **13**, 484–492.
4. Portela,A. and Esteller,M. (2010) Epigenetic modifications and human disease. *Nat. Biotechnol.*, **28**, 1057–1068.
5. Reik,W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.
6. Heyn,H., Moran,S., Hernando-Herraez,I., Sayols,S., Gomez,A., Sandoval,J., Monk,D., Hata,K., Marques-Bonet,T., Wang,L. *et al.* (2013) DNA methylation contributes to natural human variation. *Genome Res.*, **23**, 1363–1372.

7. Shukla,S., Kavak,E., Gregory,M., Imashimizu,M., Shutinoski,B., Kashlev,M., Oberdoerffer,P., Sandberg,R. and Oberdoerffer,S. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, **479**, 74–79.
8. Lienert,F., Wirbelauer,C., Som,I., Dean,A., Mohn,F. and Schübeler,D. (2011) Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.*, **43**, 1091–1097.
9. Stadler,M.B., Murr,R., Burger,L., Ivanek,R., Lienert,F., Schöler,A., van Nimwegen,E., Wirbelauer,C., Oakeley,E.J., Gaidatzis,D. *et al.* (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–495.
10. Martin,D.I.K., Singer,M., Dhahbi,J., Mao,G., Zhang,L., Schroth,G.P., Pachter,L. and Boffelli,D. (2011) Phyloepigenomic comparison of great apes reveals a correlation between somatic and germline methylation states. *Genome Res.*, **21**, 2049–2057.
11. Feinberg,A.P. and Irizarry,R. a (2010) Evolution in health and medicine Sackler colloquium: Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 1757–1764.
12. Pai,A.A., Bell,J.T., Marioni,J.C., Pritchard,J.K. and Gilad,Y. (2011) A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet.*, **7**, e1001316.
13. Hernando-Herraez,I., Prado-Martinez,J., Garg,P., Fernandez-Callejo,M., Heyn,H., Hvilsom,C., Navarro,A., Esteller,M., Sharp,A.J. and Marques-Bonet,T. (2013) Dynamics of DNA methylation in recent human and great ape evolution. *PLoS Genet.*, **9**, e1003763.
14. Zeng,J., Konopka,G., Hunt,B.G., Preuss,T.M., Geschwind,D. and Yi,S. V (2012) Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am. J. Hum. Genet.*, **91**, 455–465.
15. Molaro,A., Hodges,E., Fang,F., Song,Q., McCombie,W.R., Hannon,G.J. and Smith,A.D. (2011) Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell*, **146**, 1029–1041.
16. Farcas,R., Schneider,E., Frauenknecht,K., Kondova,I., Bontrop,R., Bohl,J., Navarro,B., Metzler,M., Zischler,H., Zechner,U. *et al.* (2009) Differences in DNA methylation patterns and expression of the CCRK gene in human and nonhuman primate cortices. *Mol. Biol. Evol.*, **26**, 1379–1389.
17. Schneider,E., El Hajj,N., Richter,S., Roche-Santiago,J., Nanda,I., Schempp,W., Riederer,P., Navarro,B., Bontrop,R.E., Kondova,I. *et al.* (2014) Widespread differences in cortex DNA methylation of the 'language gene' CNTNAP2 between humans and chimpanzees. *Epigenetics*, **9**, 533–545.
18. Wilson,G. a, Butcher,L.M., Foster,H.R., Feber,A., Roos,C., Walter,L., Woszczek,G., Beck,S. and Bell,C.G. (2014) Human-specific epigenetic variation in the immunological Leukotriene B4 Receptor (LTB4R/BLT1) implicated in common inflammatory diseases. *Genome Med.*, **6**, 3–19.
19. Gokhman,D., Lavi,E., Prüfer,K., Fraga,M.F., Riancho,J.A., Kelso,J., Pääbo,S., Meshorer,E. and Carmel,L. (2014) Reconstructing the DNA methylation maps of the Neandertal and the Denisovan. *Science*, **344**, 523–527.
20. Banovich,N.E., Lan,X., McVicker,G., van de Geijn,B., Degner,J.F., Blischak,J.D., Roux,J., Pritchard,J.K. and Gilad,Y. (2014) Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.*, **10**, e1004663.
21. Heyn,H., Li,N., Ferreira,H.J., Moran,S., Pisano,D.G., Gomez,A., Diez,J., Sanchez-mut,J. V, Setien,F., Carmona,F.J. *et al.* (2012) Distinct DNA methylomes of newborns and centenarians. *Proceedings of the National Academy of Sciences*, **109**, 10522–10527.
22. Scally,A., Dutheil,J.Y., Hillier,L.W., Jordan,G.E., Goodhead,I., Herrero,J., Hobolth,A., Lappalainen,T., Mailund,T., Marques-Bonet,T. *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**, 169–175.
23. Locke,D.P., Hillier,L.W., Warren,W.C., Worley,K.C., Nazareth,L. V, Muzny,D.M., Yang,S.-P., Wang,Z., Chinwalla,A.T., Minx,P. *et al.* (2011) Comparative and demographic analysis of orang-utan genomes. *Nature*, **469**, 529–533.
24. Krueger,F. and Andrews,S.R. (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, **27**, 1571–1572.
25. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
26. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.
27. Wakeman,L., Al-Ismail,S., Benton,A., Beddall,A., Gibbs,A., Hartnell,S., Morris,K. and Munro,R. (2007) Robust, routine haematology reference ranges for healthy adults. *Int. J. Lab. Hematol.*, **29**, 279–283.
28. Teare,J.A. (2013) (Firm) ISIS reference ranges for physiological values in captive wildlife physiological reference intervals for captive wildlife: a CD-ROM resource. *International Species Information System*. Bloomington, MN.
29. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
30. Huang,W., Loganantharaj,R., Schroeder,B., Fargo,D. and Li,L. (2013) PAVIS: a tool for Peak Annotation and Visualization. *Bioinformatics*, **29**, 3097–3099.
31. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
32. Ziller,M.J., Gu,H., Müller,F., Donaghey,J., Tsai,L.T.-Y., Kohlbacher,O., De Jager,P.L., Rosen,E.D., Bennett,D. a, Bernstein,B.E. *et al.* (2013) Charting a dynamic DNA methylation landscape of the human genome. *Nature*, **500**, 477–481.
33. Feil,R. and Fraga,M.F. (2012) Epigenetics and the environment: emerging patterns and implications. *Nat. Publ. Gr.*, **13**, 97–109.
34. Hodges,E., Molaro,A., Dos Santos,C.O., Thekkat,P., Song,Q., Uren,P.J., Park,J., Butler,J., Rafii,S., McCombie,W.R. *et al.* (2011) Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment. *Mol. Cell*, **44**, 17–28.
35. Schlesinger,F., Smith,A.D., Gingeras,T.R., Hannon,G.J. and Hodges,E. (2013) De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. *Genome research*, **23**, 1601–1614.
36. Houseman,E.A., Accomando,W.P., Koestler,D.C., Christensen,B.C., Marsit,C.J., Nelson,H.H., Wiencke,J.K. and Kelsey,K.T. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, **13**, 86.
37. Court,F., Tayama,C., Romanelli,V., Martin-Trujillo,A., Iglesias-Platas,I., Okamura,K., Sugahara,N., Simón,C., Moore,H., Harness,J.V. *et al.* (2014) Genome-wide parent-of-origin DNA methylation analysis reveals the intricacies of human imprinting and suggests a germline methylation-independent mechanism of establishment. *Genome Res.*, **24**, 554–569.
38. Yoshida,Y., Han,B., Mendelsohn,M. and Jessell,T.M. (2006) PlexinA1 signaling directs the segregation of proprioceptive sensory axons in the developing spinal cord. *Neuron*, **52**, 775–788.
39. Siepel,A. and Haussler,D. (2005) Phylogenetic hidden Markov models. *Statistical methods in molecular evolution*. Springer, NY, pp. 325–351.
40. Siepel,A., Bejerano,G., Pedersen,J.S., Hinrichs,A.S., Hou,M., Rosenbloom,K., Clawson,H., Spieth,J., Hillier,L.W., Richards,S. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.
41. Kino,K. and Sugiyama,H. (2005) UVR-induced G-C to C-G transversions from oxidative DNA damage. *Mutat. Res. Fundam. Mol. Mech. Mutagen.*, **571**, 33–42.
42. Lu,X., Sachs,F., Ramsay,L., Jacques,P.-É., Göke,J., Bourque,G. and Ng,H.-H. (2014) The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat. Struct. Mol. Biol.*, **21**, 423–425.
43. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M., Plath,K. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.

44. Bernstein,B.E., Birney,E., Dunham,I., Green,E.D., Gunter,C. and Snyder,M. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
45. Illingworth,R.S., Gruenewald-Schneider,U., Webb,S., Kerr,A.R.W., James,K.D., Turner,D.J., Smith,C., Harrison,D.J., Andrews,R. and Bird,A.P. (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet.*, **6**, e1001134.
46. Long,H.K., Sims,D., Heger,A., Blackledge,N.P., Kutter,C., Wright,M.L., Grützner,F., Odom,D.T., Patient,R., Ponting,C.P. *et al.* (2013) Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *Elife*,**2**, e00348.
47. Tsankov,A.M., Gu,H., Akopian,V., Ziller,M.J., Donaghey,J., Amit,I., Gnirke,A. and Meissner,A. (2015) Transcription factor binding dynamics during human ES cell differentiation. *Nature*, **518**, 344–349.