Introduction

The CpG dinucleotide is unique for its ability to carry both genetic and epigenetic information in the genome of a differentiated mammalian cell. Variation in DNA methylation, facilitated by this two-base pair motif, influences gene expression, and thereby enables tissue-specific function. However, this dinucleotide is substantially depleted, to one-fifth of the expected level, due to the hypermutability (~11-fold) of cytosines when methylated. Yet, a minority of CpGs is retained against this strong tide of loss by a variable combination of: evasion of methylation in the germline, functional importance or chance. These are predominantly in CpG dense regions. Additionally, new CpGs are created by mutation through base substitution and as a by-product of the increase of GC in regions of biased gene conversion (BGC). A high density region of unmethylated CpGs can recruit CpG binding proteins, such as Cfp1 and KDM2A, which modify histone tails. Thomson et al. have shown that the experimental inclusion of a cluster of unmethylated CpGs is sufficient to establish domains of H3K4me3. This histone modification leads to genomic three-dimensional structure change and the acquisition of permissive chromatin regions within the expanse of repressed genome. CpG clusters, termed CpG Islands (CGIs), co-locate with 60–70% of human gene promoters, often those of housekeeping genes that are hypomethylated in the germ line, but also 40% that are tissue-specific. Methylation of CGI promoters acts as a durable silencing mechanism. However, the majority of CGIs are unmethylated in differentiated cells independently of their transcriptional activity. The methylation state of CGI is strongly correlated with its CpG content, with high density CGIs being predominately constitutively unmethylated and “weak” low density islands the preferred target for tissue-specific methylation. CpG gain that shifts an island from weak to strong status therefore affects its dynamic ability for methylation change.

CpGs located in the lower density regions surrounding islands, termed CpG shores (~2 kb up- or down-stream), identify significant tissue-, cancer- and reprogramming-specific methylation variability. Therefore, shore accretion and island erosion by subtle modulation in CpG density within these regions may have a disproportionate influence on the methylation levels and locations of these flanking regions. Additionally, an increase in methylation variance has been proposed to have an evolutionary important role, as well as being a potential influence on disease susceptibility.

Keywords: epigenetics, epigenomics, CpG islands, gene regulation, evolution, human disease

Abbreviations: BGC, biased gene conversion; CGI, CpG islands; MeDIP, methylated DNA immunoprecipitation; PBC, positive beacon clusters; PBMC, peripheral blood mononuclear cells

Regulatory change has long been hypothesized to drive the delineation of the human phenotype from other closely related primates. Here we provide evidence that CpG dinucleotides play a special role in this process. CpGs enable epigenome variability via DNA methylation, and this epigenetic mark functions as a regulatory mechanism. Therefore, species-specific CpGs may influence species-specific regulation. We report non-polymorphic species-specific CpG dinucleotides (termed “CpG beacons”) as a distinct genomic feature associated with CpG island (CGI) evolution, human traits and disease. Using an inter-primate comparison, we identified 21 extreme CpG beacon clusters (≥ 20/kb peaks, empirical p < 1.0 × 10^-3) in humans, which include associations with four monogenic developmental and neurological disease related genes (Benjamini-Hochberg corrected p = 6.03 × 10^-3). We also demonstrate that beacon-mediated CpG density gain in CGIs correlates with reduced methylation in these species in orthologous CGIs over time, via human, chimpanzee and macaque MeDIP-seq. Therefore mapping into both the genomic and epigenomic space the identified CpG beacon clusters define points of interaction where a substantial two-way interaction between genetic sequence and epigenetic state has occurred. Taken together, our data support a model for CpG beacons to contribute to CGI evolution from genesis to tissue-specific to constitutively active CGIs.
The genetic loss and gain of CpG dinucleotides over evolutionary time will impact upon the epigenome. Genome-wide variation in GC content at the megabase scale led to the formation of isochores before mammalian radiation with an increase in CpGs occurring ~90 million years ago (MYA). A subsequent clock-like loss of CpGs, due to the time-dependent substitution rate of cytosine deamination, has led to roughly similarly numbered, but differing sets of CpGs in primates. The mutability of individual CpGs can be determined by accounting for the influence of surrounding CpG density, as well as by sequence context and nucleosome position. Regions of CpGs that remain hyperconserved have been found to co-locate with polycomb repressive complex 2 binding domains and developmental genes. On the other hand, GC increase is influenced by primate recombination rates. So much so, that regions of extreme substitutional divergence in the human genome co-locate with recombination-associated BGC. This process therefore negates or obscures any potential evidence of weak selection. Cohen et al. have recently shown that CGIs can evolve without the requirement of selective pressure, although a possible subtle influence on CpGs via gene body methylation may exist.

Cytosine deamination is consequently the predominant single nucleotide mutational force, occurring at one order of magnitude higher in the genome than other single base substitutions. Conversely, a highly localized BGC-mediated increase of CpGs occurs, associated with recombination. To discover the locations of potential species-specific regulatory modulation, due to CpG dinucleotide change, we identified a subset of human CpGs that were only present, either uniquely maintained or gained, in that lineage. While there are approximately ~40 million genetic differences between human and chimpanzee, the vast majority of these differences occur, associated with recombination. 35-39 This process therefore negates or obscures any potential evidence of weak selection. BGC can lead to the formation of CGIs and, furthermore, regions of CGIs that remain quasi-genomic set (referred to as h1c1o1: human, chimpanzee and other primate) to 79.99% of the human genome. This contained 25,100,205 or ~88.95% of the total haploid human CpGs. Each of these remaining CpGs was then interrogated with the requirement that at its precise position none of the other primates had a CpG dinucleotide present. Furthermore the chimpanzee sequence and the closest nearest other primate present in the alignment block (96.6% Gorilla) were required to have aligned sequence at this position i.e. was N or -. This led to an initial human-specific subset of 1,820,319 CpGs. These CpGs were then conservatively filtered for polymorphism utilizing 1,000 genomes data removing any CpG with any evidence of variation, as a SNP, or within a copy number or structural variant, leading to a final estimate of 1,192,484 human-specific CpGs.

CpG beacons. We define “beacons” as species-specific non-polymorphic DNA motifs able to carry both genetic and epigenetic information. According to the above analysis, we estimated the number of CpG beacons to be ~1.19 million in the human genome. In the future a definitive set will be able to be established following mass whole genome sequencing in a large number of these primates. However this current calculation will already be enriched for “true” human CpG beacons that can facilitate unique species-specific epigenomic variation. A user interface to view the human CpG beacons in the UCSC genome browser in the context of existing annotation is available at www2.cancer.ucl.ac.uk/medicalgenomics/humanCpGBeacons/trackList.php.

The density distribution of the human beacons in 1 kb windows was estimated, which showed more than half were singletons, ~2% were ≥ 5 beacons/kb, and 0.03% were ≥ 20 beacons/kb. To assess the significance of this long tail with higher density, we performed 1,000 permutations by choosing a set of random beacons from the CpG locations in the h1c1o1 genomic set. This simulation never exceeded the number of peaks that are ≥ 20 CpG beacons/kb in the observed genome set (peaks ≥ 20 CpG beacons/kb: simulation peaks range = 0–7, simulation average = 1.527 peak per genome, observed peaks = 21, empirical p < 1 × 10⁻⁹).

Extreme CpG beacon clusters. Taking this ≥ 20 CpG beacons/kb as an initial threshold (which reflects an increase of ≥ 4% in CpG density per kb in human compared with the other primates) we identified 21 extreme genomic outliers of human CpG beacon density (see Table S1). Beacon density distribution is displayed across the genome in Figure 1. This initial observation revealed that the third highest peak on Chromosome 20 was identified to be co-expressed by Cajal-Retzius neurons, with Reelin, a secreted glycoprotein that
Transmission (SYT1); as well as total cerebral brain volume in a radiological examination (CDH4) (see Table S1). To take into consideration any possible gene size bias, we also performed an analysis using the regional based binomial test included in the GREAT gene enrichment analysis tool65 (using default cut-offs but reducing the maximum distal extension from 1 Mb to 100 kb, see methods) for these 20 CpG beacon clusters, excluding the known HARIA result. This was significant for only three categories of biological process with an FDR Q ≤ 0.15: which included the categories cognition (binomial raw p = 1.43 × 10⁻⁵, FDR Q = 1.26 × 10⁻¹), and behavior (binomial raw p = 2.25 × 10⁻⁵, FDR Q = 9.87 × 10⁻²). Furthermore, as a negative control, we also calculated the locations of the Chimpanzee CpG beacon clusters that exceed the 20/kb threshold and these identified no genes implicated with developmental delay or mental retardation (see Table S2) and as well was non-significant with GREAT analysis (via liftOver to human).

The human CpG beacon clusters represent regions of potential regulatory modulation or change to the nearby genes that is human-specific. This correlation, and not causation within these regions, is of interest particularly as these monogenic disease genes have shown that genetic mutation within them is not lethal but carries significant developmental and neurological pathology. These important genes could therefore be the plausible targets of significant regulatory change between human and other closely related primates due to the similarity of their proteomes.66

Biased-gene conversion overlap. The human extreme beacon clusters showed very strong overlap with the top 200 regions of BGC identified by Drezzer et al. (57.1% of ≥ 20 CpG beacons/kb clusters, χ² p < 2.20 × 10⁻⁵),66 thus implicating localized GC
rise, which is thought to be a neutral process, with a consequent increase in CpGs. Therefore this implies the CpG beacon clusters associate with a recombination driven CpG increase in human, as opposed to regions of high CpG mutability in other primates. Moreover, we also identified the majority of these clusters in telomeric regions (52.3% in terminal chromosome bands), which are known to have elevated rates of recombination in males, with hotspots associated with BGC. A 15% greater divergence in terminal ends of chromosomes was identified in the chimpanzee sequencing project.

Cohen et al. recently reclassified CGIs using evolutionary modeling into those that were classical hypo deaminated islands, with ~80% of these 10 kb from a transcription start site (TSS) and with strong overlap with H3K4me3, and those that had arisen as a by-product of BGC that were typically constitutively hypermethylated. However, on examination of the available sperm methylome data via MedIP-seq, which includes data for 18 of these extreme beacon cluster regions that co-located with CGI, these were found to be predominately hypomethylated. The average methylation level was 26.38%, therefore aiding the retention of CpGs by reduced mutability, enabling potential regulation by methylation to occur.

Positive CpG beacon clusters. To differentiate between specific CpG increase, as opposed to generalized regions of GC rise, we controlled by the concurrent formation of the exact inverse dinucleotide GpC; which lacks methylation ability. We defined Positive CpG Beacon Clusters (PBCs) as regions where CpG beacons outweighed their local human-specific GpC content. BGC increases regional GC content and therefore passively CpGs, but if CpGs are methylated in the germ line their continual loss will eventually lead to the acquired GpCs outweighing CpGs over time. We calculated this via a sliding window analysis with a window size of 1 kb and slide of 100 bp across the genome (see Fig. 2). The vast majority of the extreme beacon clusters were genomic outliers of PBC score, i.e. EHMT1 and CDH4 and all except two possessed positive scores (see Table S1). These two extreme negative scores were identified in loci known for extensive and continual gene conversion, the olfactory receptors, with PBC score peaks of -23/kb and -16/kb for OR2T3 and OR2T12, respectively.

Extreme CpG beacon clusters appear to be strongly driven by BGC; therefore, PBCs indicate regions where the gained CpGs beacons are not as hypermutable as would be expected, likely due to a loss of methylation in germ line. By retaining from the 20 clusters only those with at least a +5 PBC score, more significant p values in both biological category enrichments of cognition and behavior were obtained (binomial p = 7.19 × 10⁻⁶ and 9.41 × 10⁻⁶, Q FDR value = 6.3 × 10⁻² and 4.1 × 10⁻², respectively). To explore the potential of this CpG beacon-specific increase genome-wide, we identified all the PBC ≥ +5 loci comprising 2,601 regions, that account for ~0.1% of the human genome. IPA analysis of associated PBC genes showed significant results for a large number of common disease categories (P_BH < 1 × 10⁻²⁰) (data not shown), although this result will be biased disproportionally with larger gene regions. Examining these PBC loci with GREAT (genomic regions enrichment of annotation tool) analysis, which corrects for this issue of potential genomic space available to input signal, we identified a number of significant results for potential human phenotypes and traits (see Table S3, FDR Q < 0.05), such as cortical gyral simplification (binomial FDR Q-value = 1.94

**Figure 2.** Human positive CpG beacon scores calculated across the genome in 1 kb windows with 100 bp slide. Extreme positive or negative loci are indicated.
methylation in CGI across all three species in the Ensembl CGI set (see Fig. S7A, \(p < 2.2 \times 10^{-16}\)) and as well the Wu et al. CGIs that are proposed to have improved trans-species CGI prediction (Fig. S7B, \(p < 2.2 \times 10^{-16}\)).

We then investigated whether the influence on methylation change was still apparent in CpG density that changed over time in orthologous CGI between these species. We identified the orthologous CGI set between human and chimpanzee (Ensembl \(n = 13,999\), Wu et al. \(n = 34,053\)), and human and macaque (Ensembl \(n = 4,654\), Wu et al. \(n = 19,200\)) and chimpanzee and macaque (Ensembl \(n = 4,004\), Wu et al. \(n = 18,747\)). For example, the orthologous \(DPP10\) CpG beacons extreme cluster CGI, showed average methylation (RPM) of 0.51 and 4.80 in human and chimpanzee respectively, but not enough CpG density in macaque for a CGI to be defined even by the Wu et al. methodology. The subset of these orthologous islands that were \(\geq 20\%\) CpG density in one species and \(\leq 19\%\) in the other was then obtained. A significant difference was identified in the Ensembl set for human vs. chimpanzee CGI (Wilcoxon \(p = 1.954 \times 10^{-12}\); data not shown) and in the larger Wu et al. set these groupings showed a small but significant reduction in methylation \([\text{expressed as average reads per million (RPM)}]\) consistently in the higher density CGI group across all species comparisons (see Fig. 4, all \(p\) Wilcoxon \(< 2.2 \times 10^{-16}\)).

Correlation between CpG density and CGI hypomethylation. While specific genetic methylation-determining regions (MDRs\(^7\)) have been identified within CGIs, a correlation with CpG density and hypomethylation has also previously been recognized.\(^10\) Therefore, CpG beacon clusters will lead to species-specific CpG density increases which may be associated with increased CGI hypomethylation and formation of permissive chromatin.\(^76\) A CpG density of \(-20\%\) CpGs (or 10\% methylatable cytosines) was proposed by Eckhardt et al.\(^39\) as a threshold beyond which CGI are highly likely to be constitutively unmethylated across all differentiated tissues. Examining the available data from two bisulphite sequencing experiments from Li et al.\(^77\) in peripheral blood cells and Lister et al.\(^2\) from fibroblasts, we find methylation within CGIs is strongly correlated in these sets (\(r^2 = 0.84\)), despite being confounded by different experiments, tissues and cell line effects. Furthermore, the same significant trend of reduced methylation when CGIs were categorized into subgroups of all, \(\geq 15\%\), \(\geq 20\%\), \(\geq 25\%\) CpG density is seen using both the Ensembl CGI definition and an alternate CGI set by Wu et al. identified via hidden Markov models.\(^78\) (Kruskal-Wallis \(p < 2.2 \times 10^{-16}\)) (Fig. 3; Fig. S6A–C).

Next, we generated peripheral blood cell methylome data by MeDIP-seq of pooled samples from chimpanzee and rhesus macaque as well as pooled human samples. Examination of these data also supported the inverse correlation of CpG density with methylation in CGI across all three species in the Ensembl CGI set (see Fig. S7A, \(p < 2.2 \times 10^{-16}\)) and as well the Wu et al. CGIs that are proposed to have improved trans-species CGI prediction (Fig. S7B, \(p < 2.2 \times 10^{-16}\)).

We then investigated whether the influence on methylation change was still apparent in CpG density that changed over time in orthologous CGI between these species. We identified the orthologous CGI set between human and chimpanzee (Ensembl \(n = 13,999\), Wu et al. \(n = 34,053\)), and human and macaque (Ensembl \(n = 4,654\), Wu et al. \(n = 19,200\)) and chimpanzee and macaque (Ensembl \(n = 4,004\), Wu et al. \(n = 18,747\)). For example, the orthologous \(DPP10\) CpG beacons extreme cluster CGI, showed average methylation (RPM) of 0.51 and 4.80 in human and chimpanzee respectively, but not enough CpG density in macaque for a CGI to be defined even by the Wu et al. methodology. The subset of these orthologous islands that were \(\geq 20\%\) CpG density in one species and \(\leq 19\%\) in the other was then obtained. A significant difference was identified in the Ensembl set for human vs. chimpanzee CGI (Wilcoxon \(p = 1.954 \times 10^{-12}\); data not shown) and in the larger Wu et al. set these groupings showed a small but significant reduction in methylation \([\text{expressed as average reads per million (RPM)}]\) consistently in the higher density CGI group across all species comparisons (see Fig. 4, all \(p\) Wilcoxon \(< 2.2 \times 10^{-16}\)).
modification can lead to disparate effects depending on genomic location; repressive in CGI, activating in gene bodies and splicing influence in CTCF binding sites. Changes in developmental timing are significant in species-specific differences and the epigenetic modulation of intragenic islands may direct developmentally critical isoforms. Thus, this epigenomic extra layer of control enables additional axes to the adaptive landscape and aids in the evolution of complex phenotypes. Cytosine methylation has also been suggested to be significant in karyotype evolution. Even simply focusing on human higher cognitive functioning, notwithstanding all the other phenotypic differences, levels of brain tissue-specific imprinting, distinctive neuronal DNA methylation profiles and potential role in synaptic plasticity, as well as pathogenic Methyl Binding Domain gene mutations in post-natal brain development disorders, all postulate that the gain and loss of CpG may be fundamental in the human-specific phenotype.

We therefore identified a subset of species-specific CpGs by inter-primate comparison, impartial to mechanistic cause, which we have termed CpG beacons. Focusing initially on extreme human CpG beacon clusters, we showed they are enriched for neurological disease genes and, additionally, co-locate with the evolutionary accelerated HAR1A nc-RNA gene. A strong correlation between accelerated genomic loci and bias toward increased GC content was observed previously, due to the effects of recombination. Fine scale recombination hotspots show high diversity between human and chimpanzee, as they are short-lived relative to divergence times, potentially strongly influenced by the variation in zinc finger binding of PRDM9. GC-coupled CpG increase due to recombination has been suggested to have played a considerable role in CGI formation and thus may be a strong driver in the formation of CpG beacon clusters and thus species-specific regulation.

However, on top of the localized strong effect of BGC on increased GC, multiple subtle substitutions have been shown to have a morphological evolution effect altering the timing and level of expression. We looked for potential CpG-specific signatures by identifying where human-specific CpG exceeded human-specific GpC, defining Positive CpG Beacon Clusters, which identified potential human traits that may have arisen during human evolution.

Recent comparative methylome analyses have revealed species-specific differences. Molaro et al. examined chimpanzee and human sperm and supported the link between genome and epigenome, by identifying strong CpG decay correlated with methylation over brief evolutionary periods. They also found extensive species-specific methylation differences in SVA repeats, with significantly increased numbers of orthologous hypomethylated SVAs within humans. Interestingly, this is the same subtype in which we identified high enrichment of positive beacon clusters, which could be driving this hypomethylation. Additionally SVAs
In conclusion, the CpG dinucleotide is vital for regulation and not only transmits genomic data but also enables epigenomic variation. Thus, genomic change in this dynamic dinucleotide required for DNA methylation, influencing CGI methylation, gene body methylation, imprinting and splicing, is fundamental to understanding our evolutionary acquired traits and vulnerabilities to disease.

Methods

CpG beacon identification. CpG loss is time- not replication-dependent; therefore, there are almost equal counts of CpG in human and chimpanzee. Recent estimates from whole genome sequencing for mutation rate is $1 \times 10^{-8}$ per generation with the CpG dinucleotide approximately ten times this. Due to the rapid turnover in regulatory elements, most are too weakly conserved to mouse to distinguish which will particularly be the case with the highly mutable CpGs. By utilizing the additive collective divergence of multiple primates a CpG's human-specific state was attributed. The comparatively inferior sequence quality of these individual non-human primate sequences was balanced by the combined multispecies comparison vs. human. Assignment of ancestral state by use of chimpanzee alone was found to have an error rate of 0.65% utilizing macaque as a...
second out-group. Furthermore, most SNPs have been calculated to be < 1 million years of age, compared with the minimum divergence time of chimpanzee and human which is estimated at 5 million years.

Therefore, to identify the human-specific changes we utilized the Ensembl Compara.6_primates_EPO six primates alignment\(^ {51,52}\) build 58. This includes the species Homo sapiens, Pan troglodytes, Gorilla gorilla, Pongo abelii, Macaca mulatta and Callithrix jacchus. The builds included are human GRCh37, chimpanzee Mar. 2006 (CGSC 2.1/pan Trog2), gorilla (gor Gor3); orang-utan Jul 2007 (WUSTL version Pongo_alb elit2.0.2); macaque Jan 2006 (MGSC Merged 1.0/rheMac2); and marmoset Mar 2009 (WUGSC 3.2 (GCA_000004665.1)). The Ensembl Compara.6_primates_EPO alignment blocks was then reduced in a stepwise fashion due to the requirement of: unique human sequence, i.e. does not align to greater than one location in human genome (82.71% of human genome remaining); unique chimpanzee sequence (80.54%); and contains at least one other primate (79.99%) in order to utilize the strength of the inter-primate comparison.\(^ {45}\) Within this ~80% of the human genome that was able to be aligned, an algorithm was devised to identify the location of each human CpG site within these blocks and then compared with the corresponding bases in other species. To be identified as a potential human-specific CpG, the requirement in chimpanzee sequence was that it did not match CG and did not contain N or –. All other species sequences at this position also did not match CpG and the closest other primate (gorilla in 96.64% of sites) did not contain N or –. If this was the case then it was recorded as a human-specific CpG site, which led to a set of 1,820,319 CpGs. All human cluster locations are given in build Human GRCh37 coordinates. The chimpanzee beacons were calculated using the exact same methodological approach as the human, but instead changing the focused species to Pan troglodytes.

**Polymorphic filter.** Any CpG with evidence of polymorphism from 1,000 Genomes data for SNP, CNV and Indel was then removed from the set. The December 2010 Data update Full Project Genotype Release from calls on 629 individuals from the 20100804 sequence was used. Indel data was from the February 2011 update, which were calls from Dindel on the same 629 individuals from the 20100804 sequence and alignment, and also available CNV data from 179 unrelated individuals.\(^ {109}\) This resulted in a non-polymorphic human set of 1,192,484 CpGs.

**CpG Beacon density calculation and permutations.** Initial density permutations were calculated for each CpG beacon by counting the number of CpG beacons within a region of 499 bp downstream and 500 bp upstream. Random beacon sets of the same number 1,192,484 were generated from the total set of locations of CpGs in the h1c1o1 genome set of 20,207,732 and density calculated.

**Positive CpG Beacon clusters.** Positive beacon clusters were calculated, via a sliding window of 1 kb and slide of 100 bases across the genome. The total non-polymorphic human-specific CpGs within the 1kb region was subtracted from the total CpG beacons within this region.

\[
\text{Positive CpG Beacon Score} = \frac{(\text{CpG beacons} - \text{human specific CpGs})}{1 \text{ kb}}
\]

**Gene set enrichment analyses.** Ingenuity Pathway Analysis (IPA) was performed for gene set enrichment. The location of clusters was assigned to genes if it mapped to within 100k 5’ and 50k 3’ of the transcript. The following IPA analysis settings were used: Reference set: Ingenuity Knowledge Base (Genes Only), Relationship to include: Direct and Indirect, Includes Endogenous Chemicals. Optional Analyses: My Pathways My List. Filter Summary: Consider only molecules and/or relationships where (species = Human) AND (confidence = Experimentally Observed). Benjamini-Hochberg multiple test corrected p values are shown only. The region-based binomial analysis of GREAT analysis \(2.0.1^ {65}\) was utilized to identify genome regional enrichments from the location of the extreme beacon clusters, as well as the moderate positive beacons clusters \(\geq 5\). This takes into consideration the bias of potential genomic space available compared with the traditional hypergeometric test. The parameters used were association by Basal plus extension with default values of proximal 5 kb upstream, 1 kb downstream, but a reduced distal limit to 100 kb from 1 Mb and significance assessed by the regional based Binomial test FDR Q value.

**Publicly available bisulphite sequencing data.** The Lister et al. fibroblast methylome data was downloaded from http://neomorph.salk.edu/human_methylome/data.html. The Li et al. PBMC methylome data was downloaded from the NCBI Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17972). For the Lister et al. and Li et al. data sets individual CpG methylation was calculated by combining the reads from the two strands and subsequently requiring a five read minimum coverage for inclusion.

**Comparative MeDIP-seq.** DNA was extracted from peripheral blood of five chimpanzees (Pan troglodytes, three males, two females) and five rhesus macaques (Macaca mulatta, three males, two females). Samples were taken from captive individuals at Tierpark Nordhorn, Basel Zoo, Leipzig Zoo and at the German Primate Center during routine health checks and not specifically for this study. Microsatellite analysis conducted at the German Primate Center verified that respective individuals are not related. Sample collection adhered to the American Society of Primatologists (ASP) Principles for the Ethical Treatment of Non-Human Primates (www.asp.org/society/resolutions/EthicalTreatmentOfNonHumanPrimates.cfm).

To obtain averaged methylomes and reduce individual genotype effects, DNAs were pooled for each species at equal concentration for each individual. MeDIP was then executed according to Auto-MeDIP-seq protocol as described previously\(^ {10}\) and sequenced on Illumina GAIIx. This was performed with paired end reads of 36 bp with average fragment sizes of: 197 bp in human, 222 bp in chimpanzee, and 217 bp in macaque. The corresponding methylome data are available from the authors on request. A comprehensive analysis of these methylomes will be described elsewhere (Wilson G.A. et al., manuscript in preparation).
MeDIP-seq data was processed using MeDUSA (methylated DNA utility for sequence analysis). This computational pipeline performs a full analysis of MeDIP-seq data by utilizing a number of freely available software packages. Raw sequence data in fastq format were aligned to the reference genomes (Human GRCh37, panTro2 and rheMac2) using alignment software BWA (v0.5.8). with default parameters, to generate a SAM format alignment file. Aligned reads were filtered using SAMtools (v0.1.9) to remove reads that failed to form a correctly aligned pair (forward and reverse templates). Further filtering based on mapping score was also performed (read pair must contain read with mapping quality ≥ 10). Potential PCR artifacts were removed by discarding all but one read within groups of non-unique reads (i.e., reads aligned to the exact same start and stop position on the same chromosome). FastQC (www.bioinformatics.bbsrc.ac.uk/projects/fastqc) was used to determine sequence data was of acceptable quality and the Bioconductor (v2.7) package MEDIPS (v1.0.0) performed enrichment and coverage analyses. Reads per million (RPM) was calculated within regions as (reads/total reads) times 10^6 for each species (total human reads = 40,797,356, chimpanzee = 32,910,189 and macaque = 24,933,164).

**Genetic influence on the DNA methylome.** The MEDIPS package was used to approximate absolute methylation scores from relative MeDIP results. This enables regional methylation to be compared over features, i.e., CpG Islands utilizing the Genetic influence on the DNA methylome. The MEDIPS score was calculated as the excess of simulation vs. observed values. Kruskal-Wallis rank sum test was used to compare methylation in CGI density sets and Wilcoxon test for comparison between average RPM values for orthologous island sets. Chi-square calculations for enrichments were performed for PBC by bases covered of PBC vs. total bases of category and CpG beacon vs. total CpGs.

**Disclosure of Potential Conflicts of Interest**
No potential conflicts of interest were disclosed.

**Acknowledgments**
We thank the staff of the Tierpark Nordhorn, Basel Zoo, Leipzig Zoo and the German Primate Center for providing blood samples of chimpanzees and rhesus macaques. Work in the Beck lab was supported by the Wellcome Trust (084071), Royal Society Wolfson Research Merit Award (WM100023) and EU-FP7 projects HEROIC (018883), EPIGENESYS (257082) and BLUEPRINT (282510).

**Author Contributions**

**Supplemental Materials**
Supplemental materials may be found here: http://www.landesbioscience.com/journals/epigenetics/article/22127

**References**
5. Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature 1980; 287:560-1; PMID:6999365; http://dx.doi.org/10.1038/287560a0.
10. Duncan BK, Miller JH. Mutagenic deamination of cytosine residues in DNA. Nature 1980; 287:560-1; PMID:6999365; http://dx.doi.org/10.1038/287560a0.
11. Polak P, Arndt PF. Long-range bidirectional strand DNA utility for sequence analysis). This computational pipeline performs a full analysis of MeDIP-seq data by utilizing a number of freely available software packages. Raw sequence data in fastq format were aligned to the reference genomes (Human GRCh37, panTro2 and rheMac2) using alignment software BWA (v0.5.8). with default parameters, to generate a SAM format alignment file. Aligned reads were filtered using SAMtools (v0.1.9) to remove reads that failed to form a correctly aligned pair (forward and reverse templates). Further filtering based on mapping score was also performed (read pair must contain read with mapping quality ≥ 10). Potential PCR artifacts were removed by discarding all but one read within groups of non-unique reads (i.e., reads aligned to the exact same start and stop position on the same chromosome). FastQC (www.bioinformatics.bbsrc.ac.uk/projects/fastqc) was used to determine sequence data was of acceptable quality and the Bioconductor (v2.7) package MEDIPS (v1.0.0) performed enrichment and coverage analyses. Reads per million (RPM) was calculated within regions as (reads/total reads) times 10^6 for each species (total human reads = 40,797,356, chimpanzee = 32,910,189 and macaque = 24,933,164).

**Genetic influence on the DNA methylome.** The MEDIPS package was used to approximate absolute methylation scores from relative MeDIP results. This enables regional methylation to be compared over features, i.e., CpG Islands utilizing the appropriate genome sequence for each species. LiftOver was used to calculate orthologous CGI sets with overlap of at least 95% required. Greater than 20% and less than 19% orthologous sets were chosen from the orthologous island sets for each grouping with the following numbers: Hs ≥ 20 and Pt ≤ 19, Ensembl = 375, Wu = 557; Pt ≥ 20 and Hs ≤ 19, Ensembl = 150, Wu = 614; Hs ≥ 20 and Pt ≤ 19, Ensembl = 248, Wu = 605; Mm ≥ 20 and Hs ≤ 19 Ensembl = 62, Wu = 1530; Pt ≥ 20 and Mm ≤ 19, Ensembl = 256, Wu = 700; Mm ≥ 20 and Pt ≤ 19 Ensembl = 118, Wu = 1451 (Hs, Homo sapiens; Pt, Pan troglodytes; Mm, Macaca mulatta).

**Statistical**
Statistical calculations were performed in R statistical environment. Empirical p values were calculated as the excess of simulation vs. observed values. Kruskal-Wallis rank sum test was used to compare methylation in CGI density sets and Wilcoxon test for comparison between average RPM values for orthologous island sets. Chi-square calculations for enrichments were performed for PBC by bases covered of PBC vs. total bases of category and CpG beacon vs. total CpGs.

**Disclosure of Potential Conflicts of Interest**
No potential conflicts of interest were disclosed.

**Acknowledgments**
We thank the staff of the Tierpark Nordhorn, Basel Zoo, Leipzig Zoo and the German Primate Center for providing blood samples of chimpanzees and rhesus macaques. Work in the Beck lab was supported by the Wellcome Trust (084071), Royal Society Wolfson Research Merit Award (WM100023) and EU-FP7 projects HEROIC (018883), EPIGENESYS (257082) and BLUEPRINT (282510).

**Author Contributions**

**Supplemental Materials**
Supplemental materials may be found here: http://www.landesbioscience.com/journals/epigenetics/article/22127


Nekrutenko AG, Li WH. Assessment of compositional changes as a driving force of development, evolutionary adaptation, and disease. Proc Natl Acad Sci U S A 2009; 106:17577-64; PMID:20088627; http://dx.doi.org/10.1073/pnas.0906183107.


