Cell

Primate CpG Islands Are Maintained by Heterogeneous Evolutionary Regimes Involving Minimal Selection

Netta Mendelson Cohen,[1] Ephraim Kenigsberg,[1] and Amos Tanay[1,*]
[1]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot 76100, Israel
*Correspondence: amos.tanay@weizmann.ac.il
DOI 10.1016/j.cell.2011.04.024

SUMMARY

Mammalian CpG islands are key epigenomic elements that were first characterized experimentally as genomic fractions with low levels of DNA methylation. Currently, CpG islands are defined based on their genomic sequences alone. Here, we develop evolutionary models to show that several distinct evolutionary processes generate and maintain CpG islands. One central evolutionary regime resulting in enriched CpG content is driven by low levels of DNA methylation and consequentially low rates of CpG deamination. Another major force forming CpG islands is biased gene conversion that stabilizes constitutively methylated CpG islands by balancing rapid deamination with CpG fixation. Importantly, evolutionary analysis and population genetics data suggest that selection for high CpG content is not a significant factor contributing to conservation of CpGs in differentially methylated regions. The heterogeneous, but not selective, origins of CpG islands have direct implications for the understanding of DNA methylation patterns in healthy and diseased cells.

INTRODUCTION

Twenty-five years ago, a seminal paper by Bird and his colleagues revealed that a significant fraction of the mouse genome is rich in unmethylated CpG dinucleotides. This fraction was estimated to form about 30,000 genomic CpG islands (Bird et al., 1985). Later, the original experimental notion of CpG islands was replaced with a criterion based on the CpG content of the DNA sequence (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002). It was demonstrated that the experimental and computational definitions largely overlap and correlate with other important genomic elements, specifically transcription start sites (TSS). CpG islands became key genomic features in epigenetic research, and according to the prevailing paradigm, the role of DNA methylation can be explored by characterizing their methylation state. Recently, comprehensive mapping of DNA methylation in various cell types has confirmed the lack of methylation in the majority of CpG islands, but also uncovered numerous cases of differentially methylated, or even constitutively methylated regions that are defined as CpG islands based on their sequence content (Dindot et al., 2009; Doi et al., 2009; Lister et al., 2009). The interpretation of these data, and of massive epigenetic profiles that are currently being collected, necessitates re-evaluation of the question of CpG island evolutionary origins. Are DNA methylation patterns and CpG densities evolutionary conserved? If so, what evolutionary forces conserve them? Are CpGs evolving under selective pressure similar to that acting on protein-coding sequences or transcription factor binding sites?

In their original study, Bird and his colleagues observed that lack of methylation and high CpG content may be evolutionarily coupled. The main mechanism proposed was the increased mutability of 5-methyl-cytosines (5mC), possibly due to inaccurate mismatch repair of deaminated 5mCs (i.e., Uracils) that introduce Thymines upon replication (Bird, 1980). In vertebrates, methylated cytosines are almost always found in the context of CpG dinucleotides. The result is increased CpG mutability, which causes methylated regions to lose CpGs rapidly. Since the rate of CpG-gaining substitutions is not increased in these regions, their sequences are converging to an evolutionary equilibrium at low CpG content (Figure 1A). In contrast, unmethylated CpG islands can sustain higher CpG content since they are not prone to hypermutability. This elegant evolutionary rationalization for CpG islands is essentially neutral—it does not assume any function for the CpGs in CpG islands, and proposes a mechanism that does not involve purifying selection against CpG loss (we denote it here as Bird's hypodeamination regime, Figure 1B). In implicit contrast (but not necessarily in contradiction) with this idea, CpG islands are often assumed to function as developmental switches, which provide the cell with a form of epigenetic memory by generating cell-type-specific hyper- and hypomethylation patterns (Baylin and Herman, 2000; Doi et al., 2009; Gal-Yam et al., 2008; Irizarry et al., 2009; Keshet et al., 2006; Reik, 2007; Straussman et al., 2009; Weber et al., 2005). Differentially methylated regions are hypothesized to function by attracting or preventing binding of specific factors in a methylation-dependent fashion (Bartke et al., 2010; Illingworth et al., 2010; Jorgensen and Bird, 2002; Kim et al., 2007). If the CpGs in CpG islands encode epigenetic switches, one may hypothesize that selection

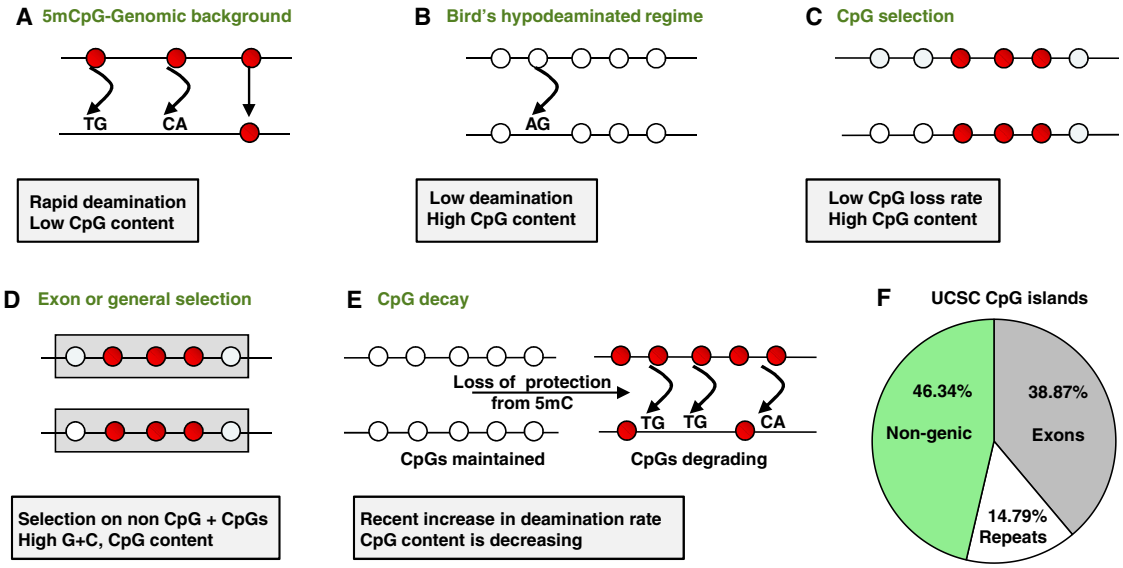

Figure 1. Models of Evolutionary Dynamics of CpG Islands

Schematic depictions of evolving CpGs (ovals) are shown, indicating methylated CpGs as filled ovals. (A) Genomic background. The genome's CpGs are typically methylated, and this is coupled with rapid deamination (CG-to-TG or CG-to-CA mutations) that leads to low stationary CpG content.

(B) Bird's regime. Regions of the genome that are not methylated do not deaminate rapidly, and therefore maintain higher CpG content. Nondeamination CpG-losing substitutions (e.g., CG to AG) in this regime are not selected against and occur at their normal rates.

(C) CpG selection. Regions in which CpGs are functional may evolve under a regime selecting against CpG loss. This would result in higher than average CpG content. CpG loss rates should be low for both deaminations and nondeaminations. Such regions may be either methylated (as shown here) or unmethylated.

(D) Exon selection. In regions under a strong selective constraint, like exons, CpGs may accumulate as part of a general slowdown in evolutionary rates and without indicating a special function for CpGs.

(E) CpG decay. Sequences evolving under regimes (B)–(D) may lose their CpG constraint (due to duplication and divergence, or by other means) and gradually converge to regime A. Depending on how recently the constraint was lost, it may be difficult to discern such pseudo-CpG islands from active CpG islands using the genomic sequence alone.

(F) UCSC CpG islands. In the current set of UCSC CpG islands, a large fraction of the elements represent exons. Many other CpG islands are heavily repetitive, which complicate their evolutionary analysis considerably. See also Figure S1.

is working to slow down the loss of CpGs within them, giving rise to a selective evolutionary process different from Bird's original regime. A selective regime can be distinguished from the hypodeamination neutral regime, since it would reduce rates of nondeamination CpG-losing substitutions (e.g., CG → AG) (compared to general substitutions, Figure 1C). Importantly, selection may contribute to the emergence of CpG islands even if it does not select for CpGs directly. Notable examples are exons (Figure 1D), which conserve their G/C-rich protein-coding sequences in general, and may therefore have higher CpG content than most of the genome. The above evolutionary processes (Bird's hypodeamination regime, CpG selection and general selection) are all expected to stabilize the CpG content in affected regions, but the genome may also contain CpG-rich sequences that are not stable and are losing their CpGs continuously (Figure 1E). CpG-losing dynamics may be initiated following loss of some constraint (lack of methylation, selection) that originally stabilized the CpG island.

Despite the potential evolutionary heterogeneity of the genomic CpG repertoire, many of the current attempts to understand the role of DNA methylation in the regulation of development and cancer are based on an approach that analyzes all CpG-rich regions in the genome uniformly, or based on stratified CpG content (for example, regions with high, intermediate or low CpG content [Meissner et al., 2008; Weber et al., 2007]). In this work we introduce a comprehensive model for the study of the evolution of primate CpGs and use it to characterize the origins of CpG-rich sequences in the human genome. We reveal that the current working set of CpG islands (Figure 1F) must be expanded and reclassified to describe several radically different evolutionary regimes. Our proposed classification includes the classical unmethylated CpG islands, CpG islands in exons, constitutively methylated CpG islands driven by increased G/C content in biased gene conversion hotspots, and pseudo CpG islands that deteriorate throughout the primates' genome evolution. The detailed evolutionary model allows us to characterize the forces that give rise to these classes of CpG islands and to conclude that purifying selection on CpG content is unlikely to be globally involved in maintaining CpG rich regions in the human genome. In particular we demonstrate that the evolutionary dynamics in tissue-specific differentially methylated regions (TDMRs) are not different from those observed in unmethylated CpG islands globally. We propose a revised genomic framework for the understanding of DNA methylation in primate genomes (see http://compgenomics.weizmann.ac.il/tanay/?page_id=196 for a list of genomic intervals and their classification), which we

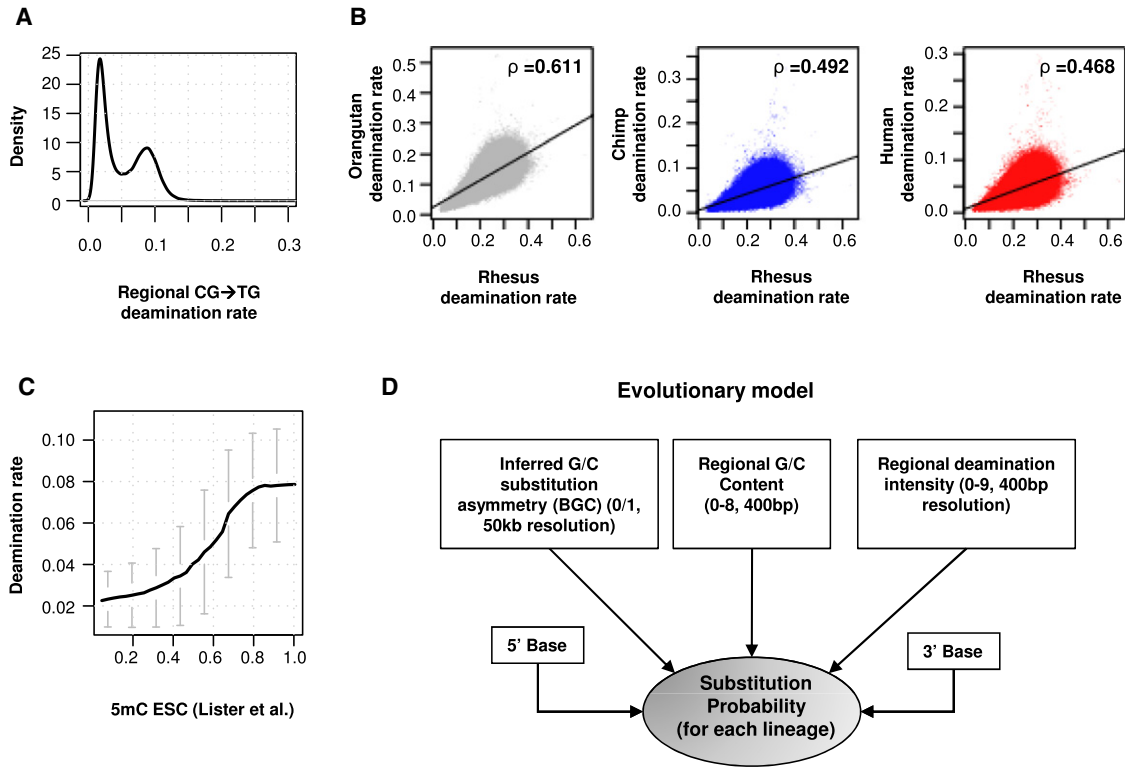

Figure 2. Modeling Methylation-Dependent CpG Evolution

(A) Bimodality of CpG deamination rates. Shown is the distribution of CpG deamination rates in regions with CpG content exceeding 3%, using a sliding window of 2 kb and collecting information from all of the lineages in the primate phylogeny used in this study (Figure S1), except for the Marmoset outgroup lineage.

(B) Phylogenetic universality of regional CpG deamination rates. Shown are inferred rates of CpG deamination in the Rhesus lineage (x axis) compared to the rates in three other lineages. Each point represents the behavior of a 20 kb region. The data suggest that CpG deamination is highly variable between genomic regions (as shown by the range of deamination rates), but that it is conserved and scales uniformly between lineages (as shown by the correlation between lineage rates). It is therefore possible to model the variation in CpG deamination intensity jointly across all lineages.

(C) Methylation and deamination are highly correlated. Shown are average deamination rates and their standard deviations collected from all lineages (y axis), calculated in regions with different levels of average embryonic stem cell DNA methylation (x axis, computed as the number of methylated CpGs divided by the total number of CpGs for 2 kb windows).

(D) A parameter-rich evolutionary model for CpGs. We modeled the evolutionary dynamics throughout the genome using lineage-specific substitution matrices that depend on several factors. The 3′ and 5′ nucleotides determine the flanking context, which in particular defines the CpG context given a 5′ Cytosine or 3′ Guanine. The regional G/C content is defined using bins of 400 bp. We discretize the G/C content to 9 levels and allow different substitution matrices to be learned for each G/C content level. The regional CpG deamination intensity is inferred for each region directly from the data and ensures the substitution matrices take into account the variability in CpG deamination rate. This variability is correlated with, but cannot be predicted from, the regional G/C content. Finally, the regional G/C substitution asymmetry parameter is also inferred directly from the data and prevents systematic biases in regions that are subject to biased gene conversion ([Dreszer et al., 2007], see text). Loci that are evolving under the influence of biased gene conversion are modeled using specific substitution parameters to correct for potential under-estimation of their ancestral G/C content. See also Figure S2.

believe will eliminate much of the confusion that currently confounds the interpretation of emerging genome-wide epigenomic profiles.

RESULTS

Distinct Regimes of CpG Evolution

Using a new parameter-rich evolutionary model (Figure S1A available online, Experimental Procedures), we inferred ancestral sequences and regional evolutionary substitution rates from alignments of five primate genomes (Figure S1B). The model was designed to carefully control for context-dependent variations in substitution rates, in particular variation in CpG deamination rates (Figure 2A, Figure S2A), without which evolutionary inference on CpG dinucleotides is highly biased. The model successfully inferred regional deamination rates and indicated that these are quantitatively conserved across the different primate lineages (Figure 2B, Figure S2B) and remarkably well correlated with DNA methylation levels in human embryonic stem cells (hESCs) (Figure 2C). Analysis of artificial alignments simulated from our evolutionary model confirmed that our learning and inference algorithms are robust (Figure S2C) and demonstrated a good fit of inferred and simulated evolutionary statistics even for counts of infrequent events (e.g., substitutions on the overall rare CpGs). To comprehensively study the evolutionary dynamics of CpG-rich regions in the human genome,

we focused on all genomic, nonexonic, nonrepetitive DNA with at least 3% CpG content. We excluded repetitive regions, since evolutionary analysis of these regions is not reliable, and modeled exonic regions separately from intergenic regions. For each CpG-rich region, we estimated the overall CpG and G/C content, inferred rates of CpG gain and deamination, and inferred rates of G/C gain and loss (Figure 2D, Figures 3A–3D). We supplemented these evolutionary parameters with high resolution data on DNA methylation levels in hESCs and fibroblasts (Lister et al., 2009). As shown in Figure 3E (see also Figures S3A and S3B), clustering analysis based on these parameters reveals several distinct evolutionary regimes that contribute to the formation of high CpG content regions in the genome. One large cluster (denoted hypodeaminated islands and encompassing 8.43 Mbp in total) represents classical unmethylated CpG islands that exhibit low deamination rates, with variable CpG and G/C content, and generally slower than average non-CpG nucleotide divergence. Another class of high CpG content regions (BGC [biased gene-conversion] islands, 4.37 Mbp in total) is evolving under a different regime, exhibiting more rapid deamination and higher methylation levels (an additional 4.54 Mbp cluster includes regions with ambiguous classification mostly due to insufficient evolutionary data). For reference, exonic CpG islands (which we modeled and analyzed separately) are subject to another distinct regime, showing variable methylation levels and overall low divergence rates (for both CpGs and non-CpGs), as expected from the general functional constraint preserving their sequences. Taken together, the data characterize the well supported class of unmethylated and hypodeaminated CpG islands, which is compatible with the scheme of Figure 1B and may or may not be affected by CpG selection as in Figure 1C (see below). Notwithstanding this class, a surprisingly substantial fraction of the genome's CpG content is methylated and evolving dynamically, showing distinct sequence content (Figure S3C) and genomic properties (Figure S3D). Interestingly, the evolutionary dynamics in methylated CpG islands are continuously challenging their CpG content through rapid deamination – it was therefore unclear if the CpG content of these elements is evolutionarily stable, and if so, what mechanisms compensate for the observed rapid CpG loss.

Hypodeaminated CpG Islands

The largest class of CpG-rich regions is characterized by slow CpG deamination rates and represents genomic regions with low levels of methylation. This set (Table S1) is the most natural genomic analog to the original (experimental) notion of CpG islands (Bird et al., 1985). As shown in Figure 4A, the chromosomal distribution of these elements is generally uniform. Moreover, analysis of the location of these islands reveals that 78.2% of them are present within 10 kb of an annotated transcription start site (TSS) (Figure 4B). Furthermore, comparison of the hypodeaminated islands to available data on chromatin structure in hESCs (data from GSM466734 and GSM469971), highlights the correspondence between these islands and the chromatin marks H3K4me3 and H3K27me3 (Figure 4C). A remarkably high 80% of the hypodeaminated islands in the 1 kb around a known TSS overlap with H3K4me3 marked domains. On the other hand, 76% of the islands that are over 1.5 kb from a TSS overlap with H3K27me3 marked domains. The correlation with histone methylation patterns (Edwards et al., 2010; Tanay et al., 2007) is distinctive for hypodeaminated CpG islands, as it is not observed for other CpG-rich regions, suggesting CpG richness is not sufficient for creating H3K27me3 or H3K4me3 domains.

Methylated Biased Gene Conversion CpG Islands

In marked contrast to the class of hypodeaminated CpG islands, a different class of CpG-rich regions exhibits rapid deamination rates and high methylation levels (Table S1). The chromosomal distribution of these elements (Figure 5A, Figure S4) reveals a nonuniform behavior, with clusters at sub-telomeric regions (and a few other hotspots, e.g., within chromosomes 2, 9, and 11). These elements are mostly located far away from known TSSs (Figure 5B). Detailed analysis of the evolutionary dynamics in this class reveals high rates of CpG deamination that are balanced by high rates of CpG-gaining substitutions. More generally, rapid gain of G/C nucleotides is observed in these regions, in contexts other than CpG dinucleotides (Figure S2A). This G/C substitution asymmetry, which is commonly attributed to biased gene conversion (Brown and Jiricny, 1987; Duret and Galtier, 2009; Eyre-Walker, 1993; Galtier et al., 2001), is driving increased G/C content (Dreszer et al., 2007) and thereby (indirectly) increased CpG content. The CpG-islands thus generated are evolutionarily distinct from the classical hypodeaminated CpG islands. Hence, the BGC dynamic leads to evolutionarily stable constitutively methylated CpG-dense regions. This is further supported by meta-analysis of multiple DNA methylation profiles in human and rhesus (Figure 5C). Importantly, 1,723 UCSC CpG islands (12% of the UCSC CpG islands that are not repetitive or exonic) are evolving solely due to BGC and not the classical hypodeamination dynamic. An additional 734 UCSC CpG islands (5%) are shown to combine the BGC and hypodeamination regimes. This heterogeneity in the current definition of CpG islands shows that their classification based solely on G/C content and CpG ratio may be misleading. A large number of CpG islands that are constitutively methylated and uncoupled to transcription start sites should therefore be evaluated as a specific class that lacks the distinctive epigenetic properties typically associated with CpG islands. The evolutionary justification for the CpG content and high levels of methylation in this class is readily found in the underlying substitution process.

Conservation and Decay of Regional CpG Content

The variable rates of CpG deamination and CpG gaining substitutions across the genome suggest that the overall CpG content of particular regions may have changed since the divergence of the human and rhesus lineages. The model we used to infer CpG substitution rates (Table S2) was specifically designed to ensure that the analysis would be robust for nonstationary regimes in which the net regional CpG content increases or decreases along the lineages. For example, we calculated substitution rates that were specific for each phylogenetic lineage, and considered the variability in deamination rates so that the ancestral state of diverged CpGs could be accurately estimated (Experimental Procedures). Comparison of the inferred change in CpG content

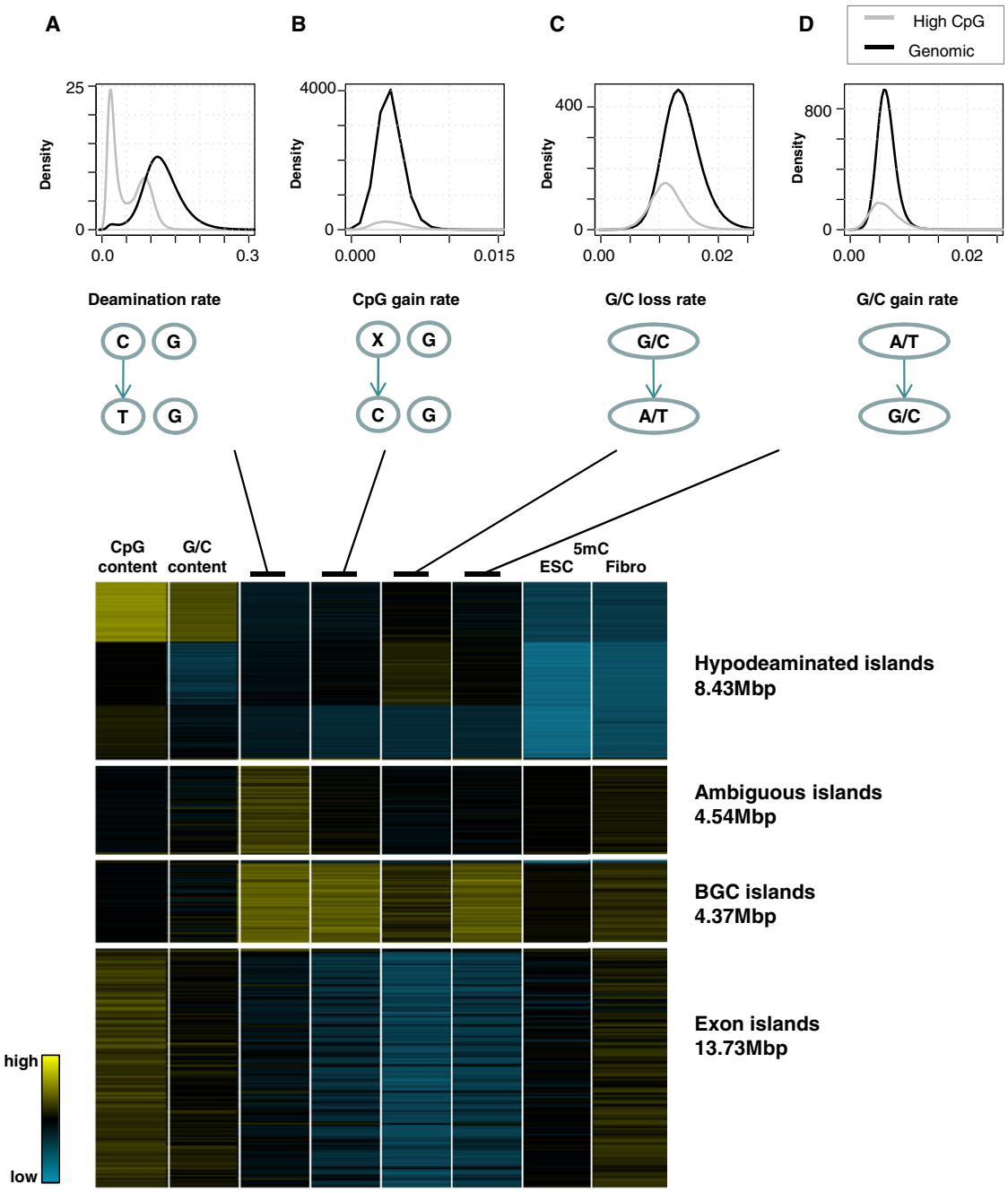

Figure 3. Evolutionary Classification of CpG Islands

(A–D) Statistics on ancestral sequences and substitution rates throughout the human genome were inferred based on the model outlined in Figure 2. The overall G/C and CpG content, rates of CpG deamination and CpG-gaining substitutions, as well as rates of G/C nucleotide gain and loss were smoothed using windows of 2 kb. All regions with CpG content exceeding 3% were clustered given evolutionary statistics and the average DNA methylation levels in ESCs and fibroblasts (Lister et al., 2009). (Similar results are obtained when omitting DNA methylation data, Figure S3B.) Shown is a color-coded clustergram (blue = low, yellow = high) in which rows represent genomic windows and columns depict the evolutionary dynamics and methylation patterns within them. Global distributions of the evolutionary parameters are shown above, depicting the behavior for regions included in the cluster analysis (high CpG content - gray) and the rest of the nonexonic, nonrepetitive genome (Genomic background - black). The resulting clusters reveal two broad classes. The first class (upper) includes regions with low levels of DNA methylation and low rates of CpG deamination and corresponds to Bird's original notion of CpG islands. The second class (denoted biased gene conversion (BGC) islands) includes regions with high levels of DNA methylation and high rates of CpG deamination. For reference we also depict exon CpG islands, which show medium to high levels of methylation, but also low rates of background G/C gain or loss substitutions. The evolution of these islands reflects a generic (non CpG-specific) selective constraint. See Figure S3A for quantitative parameter distributions in each cluster. See also Figure S3.

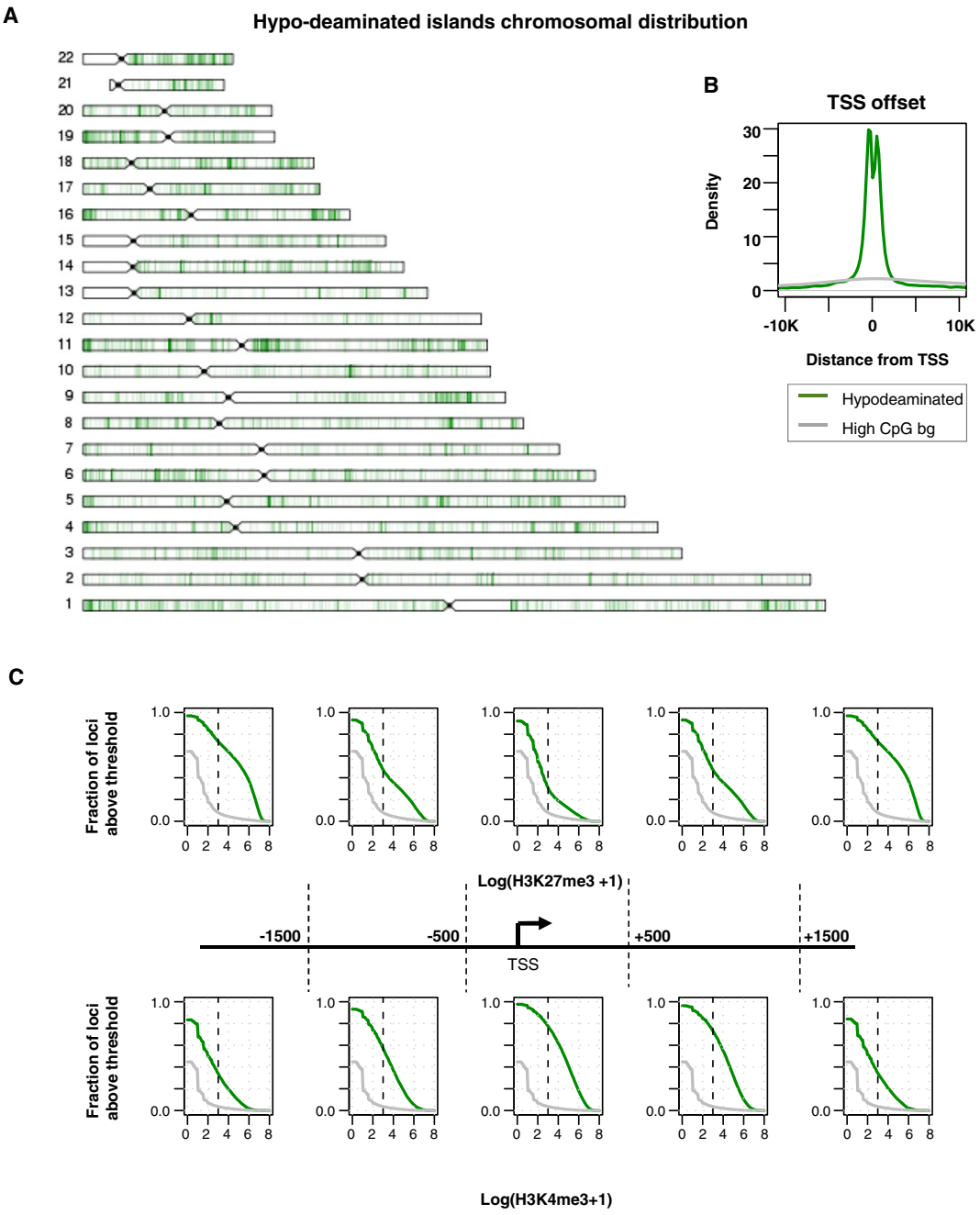

Figure 4. Hypodeaminated CpG Islands

(A) Chromosomal distribution. Shown is the chromosomal layout of CpG-rich loci that were classified as hypodeaminated islands, having low levels of DNA methylation and low deamination rates.

(B) TSS distribution. The distribution of distance from the nearest annotated TSS was computed for the set of hypodeaminated islands (green) and the remaining CpG-rich loci (gray). Over 78% of the hypodeaminated islands are associated with an annotated TSS.

(C) Histone methylation overlap with hypodeaminated CpG islands. Shown are cumulative distributions of histone 3 lysine 27 trimethylation (H3K27me3, upper) and histone 3 lysine 4 trimethylation (H3K4me3, lower) ChIP-seq coverage in human embryonic stem cells, for loci classified as hypomethylated CpG islands (green) and the remaining CpG-rich loci (gray). The analysis was done separately for loci upstream and downstream of annotated TSSs, revealing remarkable correlation between hypodeaminated islands and H3K4me3 at TSSs and H3K27me3 further away from TSSs.

in the independent lineages leading to the human and rhesus genomes reveals that the overall CpG content in hypodeamination CpG islands slightly decreased on average, and that biased-gene-conversion CpG islands frequently lost CpG content (Figure 5D). In particular, 15% of the biased gene conversion CpG islands (but only 0.28% of the hypodeaminated islands) show

significant CpG loss (over 15% decrease in their CpG content in both human and rhesus lineages; Experimental Procedures). A detailed screen for genomic regions with a significant indication of CpG loss in the human genome revealed a total of 1.73Mb, overlapping 619 of the UCSC CpG islands. The detailed list of these elements (which we denote pseudo-CpG islands) is available in Table S3.

No Global Signatures of CpG Selection on DMRs

A functional group of clustered CpGs is expected to create a specific evolutionary signature of selection. For example, the evolutionary dynamics at the *H19* and *GTL2/DLK1* imprinting control regions (ICRs) indicate that these well characterized functional epigenetic elements are evolving under remarkable mutational pressure caused by high absolute methylation-coupled rapid deamination (Figures 6A and 6B). In contrast to this pressure, the rate of CpG loss through nondeamination substitutions in these regions is lower than expected, suggesting that CpG-loss events are selected against. Purifying selection (in addition to potential compensatory gain of CpGs [Schulz et al., 2010]) may therefore help stabilize the CpG content in ICRs. Unlike the *H19* and *GTL2/DLK1* ICRs, which are methylated in the male germ line, ICRs that are methylated in the female germ line show lower deamination rates. Nevertheless, the rate of CpG loss through nondeamination substitutions is lower than expected in the maternal ICRs as well, suggesting that purifying selection is working to conserve CpGs in both paternal and maternal ICRs (Figure S4, Figure 6C). In summary, ICRs are shown to couple a known functional role for DNA methylation with a specific evolutionary signature of selection, providing working examples to test similar behaviors in other epigenetic hotspots.

A large number of tissue-specific differentially methylated regions (TDMRs) were recently characterized by comparing DNA methylation profiles among different tissues and cell lines (Cohen et al., 2009; Doi et al., 2009; Irizarry et al., 2009; Ji et al., 2010; Kim et al., 2010; Rakyan et al., 2008). TDMRs are defined based on the collective behavior (hypo- or hypermethylation) of a group of spatially clustered CpGs. The correlations of their methylation level with the regional transcriptional state and histone methylation patterns are well documented. Nevertheless, the active regulatory role of methylation in TDMRs is unclear. It can be assumed that if TDMRs (or a substantial fraction of them) are actively functional, the evolutionary dynamics of their CpGs should provide indications for a selective signature. Examples of the evolutionary dynamics at two characterized DMRs are shown in Figures 6D and 6E. Analysis of a group of 16,379 previously characterized TDMRs (Doi et al., 2009) encompassing 1.16 Mb of nonrepetitive, nonexonic high-CpG content DNA, reveals that TDMRs are mostly observed in hypodeaminated islands (82.1% versus 5.6% in BGC islands). As shown before (Doi et al., 2009), TDMRs are enriched at the margins of CpG islands and both their G/C and CpG content are lower than that of the immediately surrounding regions (Figure S5). Consistent with this, the rate of CpG deamination in TDMRs is higher than that of the bulk islands. Interestingly, nondeamination CpG-losing substitution rates are indistinguishable in TDMRs and adjacent CpG islands (Figure 6F), and are consistent with the substitution rates in non-CpG contexts. Similar dynamics are observed for additional TDMR sets, generated by diverse experimental techniques and species (Figure 6F). These data support the hypothesis that a nonselective regime maintains CpG content at hypodeaminated islands in general and in TDMRs specifically.

DMR Polymorphisms Show No Evidence of CpG-Specific Selection

As we know from population genetics theory, allele frequencies at polymorphic CpG sites can distinguish between maintenance of CpG islands by selection and stabilization of CpG islands through mere hypodeamination. As demonstrated by evolutionary simulations (Figures 7A and 7B, Figure S6A and S6B), both low level of deamination and selection for minimal CpG content will result in high steady state CpG content. However, in the selective regime, polymorphic CpGs are expected to have significantly lower allele frequencies (average heterozygosity), and thus a higher frequency of low heterozygosity alleles than that observed in G/C dinucleotides. Analysis of the distribution of heterozygosities at human single nucleotide polymorphisms (SNPs) in hypodeaminated CpG islands (Figure 7C, also compare to BGC islands in Figure S6C) reveals a slightly higher frequency of rare alleles in G/C dinucleotides, an opposite trend to that expected under a CpG selective regime. Moreover, analysis of SNPs in TDMRs shows no evidence for a specific selective constraint on CpG polymorphic sites compared to general G/C SNPs (Figures 7D and 7E). These data suggest that the selective pressure on CpGs in TDMRs is not stronger on average than the selective pressure on any other G/C dinucleotide, nor is it stronger than the selective pressure on non-TDMR CpGs. This observation holds even when studying mouse TMDRs that are mapped onto conserved human genome CpG islands (Figures S6D–S6F). Taken together, both substitution dynamics and population genetics consistently suggest that TDMRs may deaminate more rapidly, but are otherwise evolutionary similar to the CpG islands that contain them. The evolutionary conservation of TDMRs can be explained by the variation in methylation-coupled deamination rate alone, without CpG-specific selection. It remains to be seen if this lack of evidence for selection indicates lack of function for TDMR methylation, or if nonselective CpG island maintenance suffices to preserve epigenetic function.

DISCUSSION

Classes of CpG-Rich Genomic Sequences

We used a new parameter-rich model of sequence evolution combined with meta-analysis of DNA methylation data to study the origin of the CpG repertoire in primate genomes. Our data reveal at least three major evolutionary modes that govern the emergence and maintenance of CpG-rich genomic regions. Most CpG islands are constitutively unmethylated and undergo slow C-to-T deamination. We have shown that the stability of CpG content in these elements can be explained solely by the neutral effect of slow deamination associated with lack of methylation, with no evidence for purifying selection on CpG densities. In contrast to the hypodeaminated CpG islands, biased gene conversion CpG islands are constitutively methylated

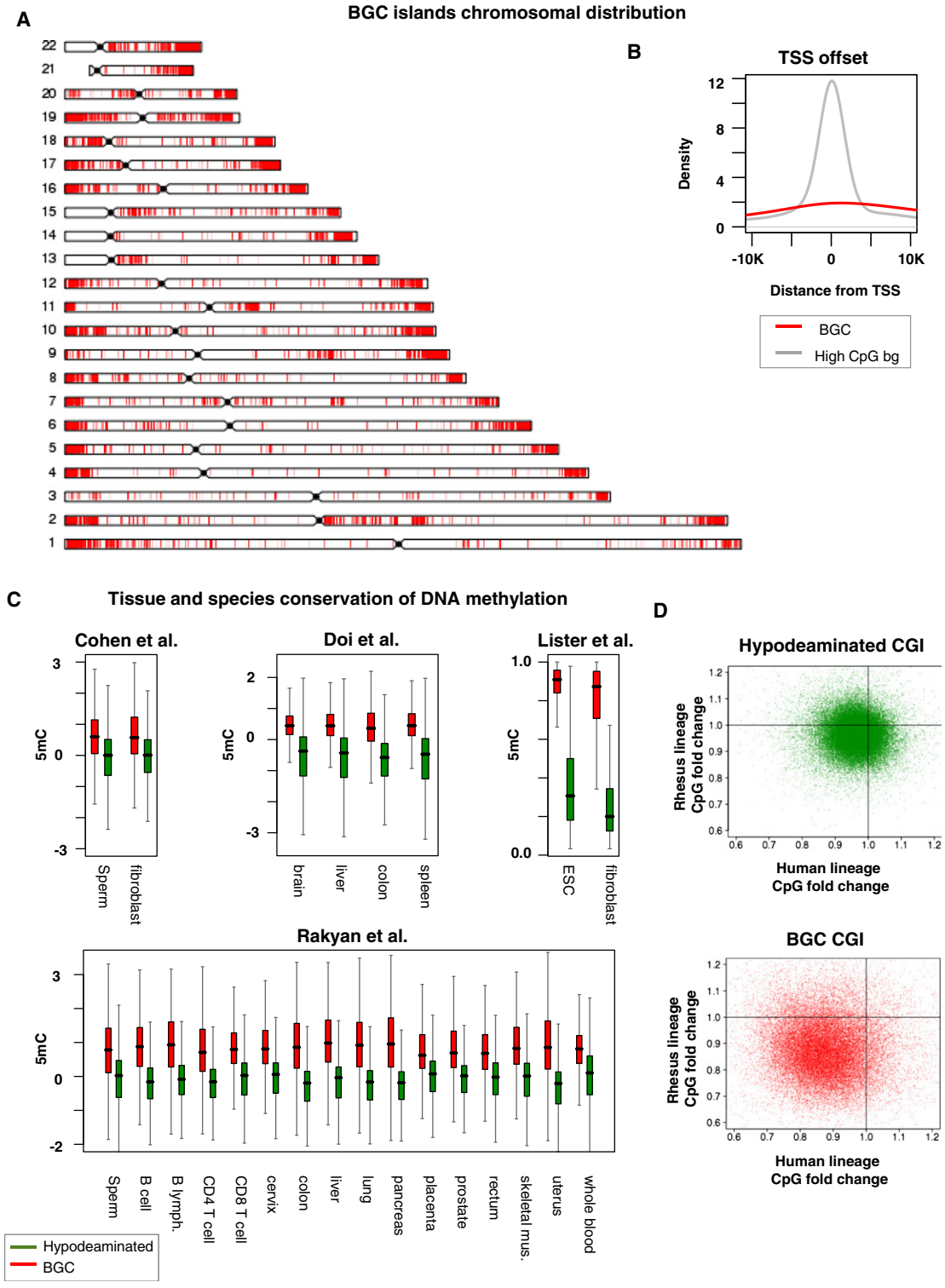

Figure 5. Constitutive Methylation at Biased Gene Conversion CpG Islands

(A) Chromosomal distribution. Shown is the chromosomal layout of CpG-rich loci that were classified as methylated and hyperdeaminated islands. The distribution is shown to be highly nonuniform, with hotspots on most sub-telomeric regions. This set was denoted as biased gene conversion (BGC) islands since it overlaps extensively with regions undergoing G/C substitution asymmetry (Dreszer et al., 2007).

(B) TSS distribution. The distribution of distances from the nearest annotated TSS was computed for the set of BGC islands (red) and the remaining CpG-rich loci (gray). BGC islands are shown to lack proximity preferences to genes.

elements that are clustered primarily in subtelomeric regions, where G/C content was hypothesized to be high due to asymmetric gene conversion (Duret and Galtier, 2009; Eyre-Walker, 1993; Galtier et al., 2001). These elements deaminate quickly, but also gain CpGs rapidly leading to high stationary CpG content. Therefore, the evolutionary origins of these CpG islands can also be accounted for without invoking selection. BGC CpG islands are not compatible with the original notion of CpG islands (unmethylated regions that are typically observed near gene promoters), and their current grouping with the classical unmethylated islands (BGC islands overlap with a total of 2,457 UCSC CpG islands), is misleading. A third regime of evolutionary dynamics in CpG islands involves elements where CpG content is decaying. These elements are typically methylated constitutively and may represent sequences that were previously protected from methylation or subject to biased gene conversion, but subsequently (through duplication or changes in *cis*) lost the mechanism(s) stabilizing CpG content. This type of process is analogous to the formation of pseudo-genes following loss of a selective constraint. These three evolutionary regimes, combined with exonic CpG islands and repetitive elements, provide a comprehensive and unbiased framework for understanding patterns of DNA methylation in the human genome.

Lack of CpG-Specific Selection in Differentially Methylated Regions

Classical CpG islands are uniformly associated with conservation due to low deamination rates and low levels of methylation. The evolutionary dynamics in these islands are typically not neutral, since these sequences are likely to encode regulatory information including transcription factor binding sites and short and long noncoding RNAs near TSSs. We have shown, using observations on both substitution rates and SNP heterozygosity, that there is no particular selective constraint on CpGs (compared to other dinucleotides) in these islands. Moreover, we could not identify such constraints in regions identified as tissue-specific DMRs, which were a-priori more likely to represent functionally important clusters of CpGs that are under selection. The evolutionary perspective on the long standing debate (Baylin and Bestor, 2002) on the functionality of DNA methylation in CpG islands may therefore have two interpretations. The simplest explanation is that DNA methylation is not functional outside aberrant (e.g., carcinogenic) contexts, and therefore selection on its genomic encoding (CpGs) is not observed. Alternatively, functional CpG islands do exist, but retain discriminatively high CpG content without the need for classical natural selection, through epigenetic control of low germ-line methylation resulting in slow mutability. In this scenario epigenetic mechanisms can fundamentally affect the evolutionary process by instructing (indirectly, but consistently) the otherwise blind mutational process to slow down at key genomic sites.

Searching for Selection on DNA Methylation Switches

One should note that selection on CpG density is still a probable driving force in a small fraction of the genome, as demonstrated for the *H19* and *GTL2/DLK1* ICRs. Dozens or hundreds of elements, each of several hundred base pairs, may conserve dense CpG clusters by selection, but the resolution of the current evolutionary data is not sufficient to identify these with high specificity. Moreover, selection on individual CpG sites, or very small groups of CpGs, is still undetectable using the current evolutionary analysis and may be prevalent if it affects only a small fraction of the CpGs in each CpG island. Single base-pair resolution data on DNA methylation profiles (Lister et al., 2009) and refined evolutionary analysis using additional primate genomes may provide more definitive answers on the selection for functional DNA methylation in specific regulatory contexts.

EXPERIMENTAL PROCEDURES

Overview of the Evolutionary Model

We wished to infer the evolutionary histories of CpG-rich regions in the human genome by comparative analysis of genomic sequences of Human, Chimp, Orangutan, and Rhesus (using Marmoset as an outgroup). This challenging task required accurate modeling of the remarkable heterogeneity in the rates of C-to-T deamination at CpG loci. CpG deaminations occur up to 20 times faster than other single nucleotide point mutations and depend strongly on the genomic and sequence context (Arndt, 2007; Baele et al., 2010). This can lead to highly biased estimations of the substitution rates and ancestral CpG content when using standard context independent models of molecular evolution. For example, if deamination rate is assumed too slow, the inferred CpG content of ancestral sequences will be too low, and the rate of CpG gaining substitutions may be overestimated. Assuming deamination rates too high would result in the opposite bias. As described below, we developed a new computational model for inference of ancestral sequences and estimation of substitution parameters while taking into account context-dependent substitution rates in general, and rapid CpG deamination in particular. Our model and inference algorithms were designed and implemented to allow genome-wide analysis (a total of 1.74 gbp genomic loci on five species), and the genome-wide approach guaranteed sufficient statistics for the robust estimation of a parameter rich model (Figure S1).

Basic Substitution Model

The evolutionary model relies on a factor graph (Kschischang et al., 2001) defining a joint distribution of three types of variables:

- Sequence variables - for each loci j and for each species i, denoted by S_i^j
- Context variables - for each loci j and for each lineage i, represents the distribution of nucleotides over the lineage between each species and its ancestor and is denoted by C_i^j

(C) Methylation meta-analysis. The distributions of methylation levels for hypodeaminated islands (green) and BGC islands (red) are depicted as boxplots using data from three studies on human cells (Doi et al., 2009; Lister et al., 2009; Rakyan et al., 2008) and one study on rhesus macaque cells(Cohen et al., 2009). The data include different tissues and cell lines, in various developmental stages, yet in all cases the BGC islands are methylated. Interestingly, the macaque data suggest that methylation of BGC islands is evolutionarily conserved despite the rapid CpG deamination in these regions.

(D) Conservation and decay of CpG content. CpG content in the inferred human-macauqe ancestral genome and the extant species genomes' was compared for regions classified as hypodeaminated CpG islands (green) and BGC CpG islands (red). Shown are the ratios between extant and ancestral CpG content for the human lineage (*x* axis) versus the rhesus lineages (*y* axis), reflecting more cases of CpG content decay in BGC islands than in hypodeaminated islands. See also Figure S4.

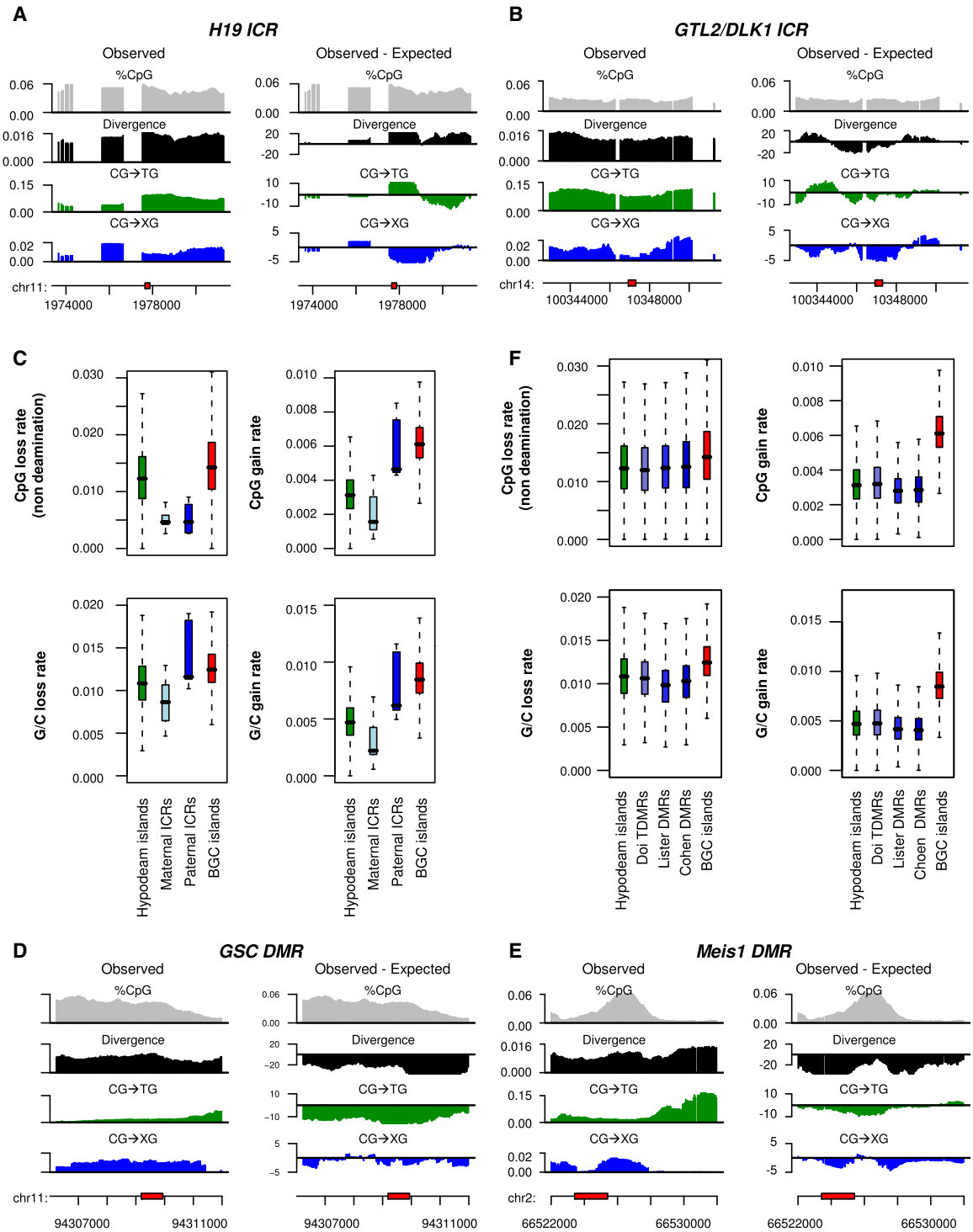

Figure 6. Substitution Dynamics at ICRs and DMRs

(A) The *H19* imprinting control region. Shown are the spatial distributions of observed evolutionary statistics (left) and differences between observed and expected statistics (right) for 6 kb around the H19 ICR (red box). The tracks depicted include data on CpG percentage (gray), overall number of substitutions (black), CpG deaminations (CG → TG/CA, green) and CpG loss through nondeamination events (CG → XG, where X is not T, blue).

(B) The GTL2/DLK1 imprinting control region. Similar to (A), showing 6 kb around the GTL2/DLK1 ICR.

(C) Low CpG loss rates at ICRs. Shown are the distributions (boxplots) of rates for four types of substitutions for 2 paternally (blue) and 7 maternally (light blue) methylated ICRs. Data for hypodeaminated islands (green) and BGC islands (red) are provided for reference. Although relatively few examples of ICRs are characterized, their evolution provides evidence for low rates of nondeamination CpG loss substitutions at both the paternal and maternal ICRs.

- Regional variables - including the G/C variable which reflects the mean value of G/C content in each region k and is denoted by G^k.

Random variables are connected via different types of factors, which assign potentials to each combination of variable values. The model uses four types of factors. First, the mutational factor $\mu_i^j(s_i^j, s_{pa(i)}^j, c_i^{j-1}, c_i^{j+1}, g^{b(j)})$, represents the conditional probability of observing a nucleotide s_i^j at loci j in species i given the nucleotide at the same locus of the ancestral species pa(i), the flanking context variables and the regional G/C content at region k=b(j). Second, the background factor $\beta^j(s_r^j, s_r^{j-1}, s_r^{j-2}, g^{b(j)})$, represents the conditional probability of observing a nucleotide s_i^j at loci j of the root species r, given the preceding two nucleotides. Third, the context factor $\delta_i^j(c_i^j, s_i^j, s_{pa(i)}^j)$ represents the conditional probability of the context variable at locus j of lineage i given the sequence variables at the end points of the lineage. Last, the GC factor $\gamma(g^k, s_r^{b(j)=k})$ represents the G/C content of region k. We note that other factorizations can be used to represent the context-dependent evolutionary process, but for our genomewide and parameter rich application, lineage segmentation (Hwang and Green, 2004) or explicit model of context dependent rate matrices (Cohn et al., 2010) were not sufficiently efficient.

Modeling Regional Variation in CpG Deamination Intensity

In order to address the variable deamination rate of CpGs, an additional regional deamination intensity variable M^k (similar to the GC variable) is considered in the mutational factor. This discrete variable takes values in the range [0..9] . The mutation factor is then parameterized such that the new variable affects only the rate of C-to-T substitutions in CpG context (CG → TG or CG→CA):

$$\mu_i'^j\left(s_i^j, s_{pa(i)}^j, c_i^{j-1}, c_i^{j+1}, g^{b(j)}, m^{b(j)}\right) = \begin{cases} \mu_i^j\left(s_i^j, s_{pa(i)}^j, c_i^{j-1}, c_i^{j+1}, g^{b(j)}\right) & \neg(XCG \to XTG, CGX \to CAX) \\ \nu_i^j\left(g^{b(j)}, m^{b(j)}\right) & XCG \to XTG, CGX \to CAX \end{cases}$$

Where ν represents the deamination probability for specific values of the G/C content variable and the deamination rate variable. We note that the M variables are defined per region and are common to all lineages. This allows more robust inference of the methylation intensity in each region, and is supported by the scaling of regional deamination rates between lineages (Figure 2B).

In summary, the model joint distribution is defined by combining all factor potentials:

$$P(s,c,g,m) = \frac{1}{Z}\prod_j \beta^j\left(s_r^j, s_r^{j-1}, s_r^{j-2}, g^{b(j)}\right) \prod_{i,j} \mu_i'^j\left(s_i^j, s_{pa(i)}^j, c_i^{j-1}, c_i^{j+1}, g^{b(j)}, m^{b(j)}\right) \prod_{i,j} \delta_i^j\left(c_i^j, s_i^j, s_{pa(i)}^j\right) \prod_k \gamma\left(g^k, s_r^{b(j)=k}\right).$$

Inferring Substitution Statistics

The joint marginal distribution of all variables connected to a mutation factor $\mu_i^j(s_i^j, s_{pa(i)}^j, c_i^{j-1}, c_i^{j+1}, g^{b(j)}, m^{b(j)})$, is approximated by an extended loopy belief propagation algorithm (see supplementary methods) using the factor belief formula:

$$b_i^j\left(s_i^j, s_{pa(i)}^j, c_i^{j-1}, c_i^{j+1}, g^{b(j)}, m^{b(j)}\right) \propto \mu_i^j\left(s_i^j, s_{pa(i)}^j, c_i^{j-1}, c_i^{j+1}, g^{b(j)}, m^{b(j)}\right) \prod_{\text{var} \in N(\mu_i^j)} \prod_{\nu \in N(\text{var})} m_{\nu \to \text{var}}(x_{\text{var}}).$$

We used this approximation to collect statistics on the number of substitutions in 50 bp genomic windows. On each lineage i, we sum up all the mutational factors j in the window, collecting posterior probabilities in order to report the number of *observed* substitutions X→Y in each context LXR (i.e., LXR→LYR):

$$obs(i,l,x,r,y) = \sum_j \sum_g \sum_m b_i^j(y,x,l,r,g,m)$$

The observed number of substitutions in a window can be compared to the number expected by the model. In order to compute the latter we multiply the expected number of appearances of each context LXR by the model's substitution probability in that context:

$$\exp(i,l,x,r,y) = \sum_j \sum_g \sum_m \left(\sum_o b_i^j(o,x,l,r,g,m)\right) P(y,x,l,r,g,m)$$

Similarly, we can report the number of cases in which specific ancestral sequence is observed, i.e., the ancestral sequence LXR on lineage i is observed:

$$obs(i,l,x,r) = \sum_j \sum_g \sum_m \sum_o b_i^j(o,x,l,r,g,m)$$

Using these formulas, we can compute evolutionary statistics on different classes of substitutions, including CpG deaminations and nondeaminations (see supplementary methods for complete details).

Learning the Model in Practice: Step by Step

Primate Multiple Alignments

Multiple alignment data for the five primate species in the phylogeny: Marmoset, Rhesus, Orangutan, Chimp and Human were downloaded from UCSC. Human exonic regions were removed from the multiple alignments using the UCSC known genes annotation.

Initial Evolutionary Model

An initial evolutionary model was learned from alignments of extant sequences, as described above, using generalized EM in a context-dependent evolutionary model, but first without taking into account regional variability in CpG deamination rates.

Evolutionary Statistics

Based on the initial model, nonexonic ancestral sequences in the phylogeny were inferred and evolutionary statistics were extracted. The inferred deamination rate in each genomic window (400 bp) and on each of the lineages was recorded.

Quantification of CpG Deamination Rates across Lineages

The deamination rates observed on the rhesus lineage were quantitatively correlated with the other lineages. The rhesus deamination rate was divided into 10 bins, ranging from slow to fast deamination. For each rhesus deamination bin, average deamination rates were computed for each of the other lineages – this served to initialize the model in the next step.

Extending the Evolutionary Model with the Deamination Rate Regional Variable

Regional deamination variables were next introduced to the model. All non-CpG context parameters were initialized to the values learned for the simpler model. The CpG deamination rates were initialized for the ten values of the deamination intensity variable, using the rhesus bins as described in the previous step. Model parameters were re-optimized through the generalized EM procedure.

(D and E) Evolutionary dynamics at the GSC and MEISeis1 DMRs. Similar to (A), but including data on 5 kb around GSC and 8 kb around MEIS1 regions, both of which were characterized as tissue-specific DMRs (Doi et al., 2009; Ji et al., 2010).

(F) Neutral CpG loss rates at DMRs. Shown are the distributions (boxplots) of rates for four types of substitutions for TDMRs derived from three studies using variable techniques and two different species. Data for hypodeaminated islands (green) and BGC islands (red) are provided for reference. TDMRs behave similarly to general hypodeaminated CpG islands with respect to all evolutionary attributes. See also Figure S5.

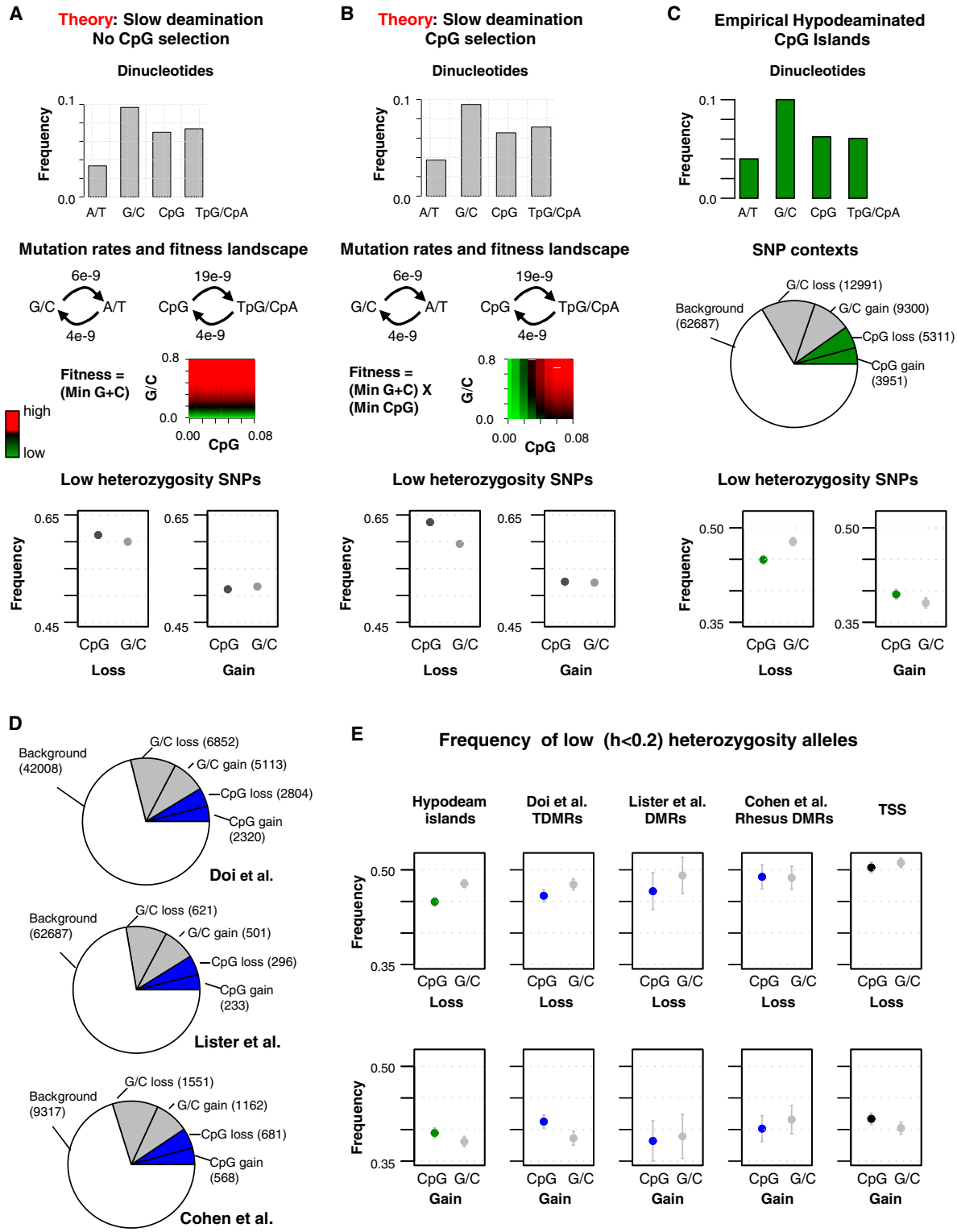

Figure 7. CpG Polymorphisms Support a Nonselective Evolutionary Regime in DMRs

(A) Evolution with selection for G/C content. A simple theoretical evolutionary model was designed to imitate the behavior of observed hypodeaminated CpG islands. The model uses mutational input favoring A/T over G/C (middle-top) and a fitness landscape selecting for some minimal G/C content (middle-low). The model also deaminates CpGs at a somewhat high (but not very high) rate characteristic of hypodeaminated CpGs. We extract stationary dinucleotide distributions for pure A/T dinucleotides (AA/AT/TA/TT), pure G/C dinucleotides (CC/GC/GG), CpGs and deamination products (TG/CA) (top) using direct

Inference and Extraction of Evolutionary Statistics from the Enhanced Model

Following the learning of the methylation dependent evolutionary model, ancestral sequences were inferred and the evolutionary statistics were re-estimated, now corrected for the variability of regional deamination rates.

Defining BGC Regions

We quantified the relative conservation of G/C nucleotides in large sliding genomic windows of 50 kbp by first scaling the number of G/C losing substitutions to reflect variation in the regional conservation rate:

$$ScaledGCloss = \frac{obs(GCloss)}{\max\left(1.2, \frac{obs(GCgain)}{\exp(GCgain)}\right)}$$

The scaling ensured that we will not define regions that are generally conserved as BGC hotspots. We then quantified the G/C substitution asymmetry as:

$$Z \sim \frac{ScaledGCloss - \exp(GCloss)}{\sqrt{\exp(GCloss)}}$$

The human genome was then segmented into regions with significant G/C conservation ($Z<-4$) and regions with background behavior. Empirical G/C conservation was validated to follow closely other metrics used before to identify candidate BGC regions.

Estimating Model Parameters in BGC and Non-BGC Sequences

Two separate sets of parameters were learned independently as described above on the BGC and non-BGC fractions of the genome. Both models were initialized with the same methylation dependent evolutionary model previously learned.

Extraction of Substitution Statistics from the Final Model

Refined evolutionary statistics were inferred based on the combined BGC and non-BGC model in 50 bps windows. These final statistics provided adequate control for variability in substitution patterns due to heterogeneity in methylation and BGC intensities.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and three tables and can be found with this article online at doi:10.1016/j.cell.2011.04.024.

ACKNOWLEDGMENTS

We'd like to thank members of the Tanay group for discussions and critical reading of the manuscript. A.T. wishes to thank Tim Bestor, Peter Jones, and Einav Nili Gal-Yam for discussions. Research in A.T.'s lab was supported by the Israeli Science foundation (1372/08) and the EU EPIGENESYS NoE program.

Received: November 18, 2010
Revised: March 28, 2011
Accepted: April 26, 2011
Published: May 26, 2011

REFERENCES

Arndt, P.F. (2007). Reconstruction of ancestral nucleotide sequences and estimation of substitution frequencies in a star phylogeny. Gene *390*, 75–83.

Baele, G., Van de Peer, Y., and Vansteelandt, S. (2010). Modelling the ancestral sequence distribution and model frequencies in context-dependent models for primate noncoding sequences. BMC Evol. Biol. *10*, 244.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell *143*, 470–484.

Baylin, S., and Bestor, T.H. (2002). Altered methylation patterns in cancer cell genomes: cause or consequence? Cancer Cell *1*, 299–305.

Baylin, S.B., and Herman, J.G. (2000). DNA hypermethylation in tumorigenesis: epigenetics joins genetics. Trends Genet. *16*, 168–174.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91–99.

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. *8*, 1499–1504.

Brown, T.C., and Jiricny, J. (1987). A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. Cell *50*, 945–950.

Cohen, N.M., Dighe, V., Landan, G., Reynisdottir, S., Palsson, A., Mitalipov, S., and Tanay, A. (2009). DNA methylation programming and reprogramming in primate embryonic stem cells. Genome Res. *19*, 2193–2201.

Cohn, I., El-Hay, T., Friedman, N., and Kupferman, R. (2010). Mean Field Variational Approximation for Continuous-Time Bayesian Networks. J. Mach. Learn. Res. *11*, 2745–2783.

Dindot, S.V., Person, R., Strivens, M., Garcia, R., and Beaudet, A.L. (2009). Epigenetic profiling at mouse imprinted gene clusters reveals novel epigenetic and genetic features at differentially methylated regions. Genome Res. *19*, 1374–1383.

Doi, A., Park, I.H., Wen, B., Murakami, P., Aryee, M.J., Irizarry, R., Herb, B., Ladd-Acosta, C., Rho, J., Loewer, S., et al. (2009). Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. Nat. Genet. *41*, 1350–1353.

Dreszer, T.R., Wall, G.D., Haussler, D., and Pollard, K.S. (2007). Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res. *17*, 1420–1430.

Duret, L., and Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annu Rev Genom Hum G *10*, 285–311.

Edwards, J.R., O'Donnell, A.H., Rollins, R.A., Peckham, H.E., Lee, C., Milekic, M.H., Chanrion, B., Fu, Y., Su, T., Hibshoosh, H., et al. (2010). Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Res. *20*, 972–980.

Eyre-Walker, A. (1993). Recombination and Mamm. Genome evolution. Proc Biol Sci *252*, 237–243.

Gal-Yam, E.N., Egger, G., Iniguez, L., Holster, H., Einarsson, S., Zhang, X., Lin, J.C., Liang, G., Jones, P.A., and Tanay, A. (2008). Frequent switching of

Wright-Fisher simulations. Frequencies of low heterozygosity SNPs were computed for CpG and G/C dinucleotides (bottom, binomial confidence intervals are shown).

(B) Evolution with selection on G/C and CpG content. An extended theoretical model illustrates the predicted effect of CpG selection on CpG islands. The model is similar to that described in A, but with a fitness landscape selecting for a minimal G/C content and CpG content. The dinucleotide distribution is similar to the regime without selection on CpGs, however, the effect of CpG selection on the frequency of low heterozygosity CpG SNPs is noticeable.

(C) Heterozygosity of CpG SNPs in hypodeaminated CpG islands. Shown are dinucleotide distributions (top), distribution of SNP contexts (middle) and frequency of low heterozygosity SNPs for all hypodeaminated CpG islands.

(D) Distributions of SNP context in TDMRs. Shown are numbers of SNPs in different contexts in three sets of (potentially overlapping) TDMRs.

(E) Similar SNPs heterozygosities for CpG and non-CpG G/C loci, in TDMRs and in CpG-rich contexts. Shown are fractions of low heterozygosity SNPs for different TDMR classes (blue), compared to all hypodeaminated CpG islands (green) and regions surrounding TSSs (±300 bp, black).

Error bars represent binomial confidence intervals. See also Figure S6.

Polycomb repressive marks and DNA hypermethylation in the PC3 prostate cancer cell line. Proc. Natl. Acad. Sci. USA *105*, 12979–12984.

Galtier, N., Piganeau, G., Mouchiroud, D., and Duret, L. (2001). GC-content evolution in Mamm. Genomes: The biased gene conversion hypothesis. Genetics *159*, 907–911.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Hwang, D.G., and Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. USA *101*, 13994–14001.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R., 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 Mamm. Genome. PLoS Genet. *6*, e1001134.

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186.

Ji, H., Ehrlich, L.I., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M.J., Irizarry, R.A., Kim, K., et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature *467*, 338–342.

Jorgensen, H.F., and Bird, A. (2002). MeCP2 and other methyl-CpG binding proteins. Ment. Retard. Dev. Disabil. Res. Rev. *8*, 87–93.

Kenigsberg, E., Bar, A., Segal, E., and Tanay, A. (2010). Widespread compensatory evolution conserves DNA-encoded nucleosome organization in yeast. PLoS Comput. Biol. *6*, e1001039.

Keshet, I., Schlesinger, Y., Farkash, S., Rand, E., Hecht, M., Segal, E., Pikarski, E., Young, R.A., Niveleau, A., Cedar, H., et al. (2006). Evidence for an instructive mechanism of de novo methylation in cancer cells. Nat. Genet. *38*, 149–153.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285–290.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Kschischang, F.R., Frey, B.J., and Loeliger, H.-A. (2001). Factor Graphs and the Sum-Product Algorithm. IEEE Trans. Inf. Theory *47*, 21.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770.

Rakyan, V.K., Down, T.A., Thorne, N.P., Flicek, P., Kulesha, E., Graf, S., Tomazou, E.M., Backdahl, L., Johnson, N., Herberth, M., et al. (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). Genome Res. *18*, 1518–1529.

Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. Nature *447*, 425–432.

Schulz, R., Proudhon, C., Bestor, T.H., Woodfine, K., Lin, C.S., Lin, S.P., Prissette, M., Oakey, R.J., and Bourc'his, D. (2010). The parental nonequivalence of imprinting control regions during mammalian development and evolution. PLoS Genet. *6*, e1001214.

Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., and Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. Nat. Struct. Mol. Biol. *16*, 564–571.

Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc. Natl. Acad. Sci. USA *99*, 3740–3745.

Tanay, A., O'Donnell, A.H., Damelin, M., and Bestor, T.H. (2007). Hyperconserved CpG domains underlie Polycomb-binding sites. Proc. Natl. Acad. Sci. USA *104*, 5521–5526.

Weber, M., Davies, J.J., Wittig, D., Oakeley, E.J., Haase, M., Lam, W.L., and Schubeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat. Genet. *37*, 853–862.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. *39*, 457–466.