Methods

Data collection. Genomic sequence, features and alignments were downloaded from the UCSC genome browser site (1), versions hg18/panTro1. The genome was classified to intergenic, intronic and exonic regions using the UCSC known genes list (use of other datasets of known genes gave similar results). Only CpGs that were not annotated as repetitive (using RepeatMasker as documented in the UCSC database) were used throughout the analysis. SNPs were downloaded from the UCSC dbSNP tables and mapped to the hg18 coordinates. Only intergenic and intronic SNPs with available average heterozygosity were analyzed. Suz12 binding domains and raw ChIP binding data (representing PRC2 occupancy) were downloaded from the ArrayExpress site (www.ebi.ac.uk/arrayexpress/) and Lee *et al*. (2) supporting data web site. Methylation status of CpGs in fractions of human chromosomes 6 and 20 were taken from the HEP project database (3).

Probabilistic modeling of CpGs divergence and the M-score model. All nonrepetitive intergenic CpGs in the human genome were partitioned into four groups: chimp-conserved, chimp-plus-deaminated (CpG->TG), chimp-minus-deaminated (Cp>CpA) and other. For each group, the dinucleotide counts at each position relative to the CpG (-200 to +200) were collected. Denote the densities of a dinucleotide **d** at relative position **i** by:

$\mathbf{p(d, i)}$	for conserved CpGs
$\mathbf{p_+(d, i)}$	for plus strand deaminated CpGs
$\mathbf{p_-(d, i)}$	for minus strand deaminated CpGs

Basically, we assume that the sequence context of conserved CpGs is characterized by the dinucleotide distribution **p(d, i),** and that the sequence context of plus (minus) strand deaminated CpGs is characterized by the dinucleotide distribution $\mathbf{p_+(d, i)}$ ($\mathbf{p_-(d, i)}$). The M-score for a CpG at position **i** inside sequence context **s** is defined by summing up log-odds:

Plus strand deamination: $\mathbf{M_+(i) = \Sigma_{-200<j<200} \log(p_+(s[i+j]s[i+j+1], j) / p(s[i+j]s[i+j+1], j))}$

Minus strand deamination: $\mathbf{M_-(i) = \Sigma_{-200<j<200} \log(p_-(s[i+j]s[i+j+1], j) / p(s[i+j]s[i+j+1], j))}$

Summing up more than 200 values provides similar results to those reported here. In principle, it is possible to transform the M-score log-odds directly into posterior deamination probabilities. Alternatively, as done here, one can use the M-score, together with additional factors (here the regional mutation rate) to construct an empirical background hypothesis for the rate of evolution of CpG distribution (see below).

Computing regional mutation rates. Regional mutation rates were computed by counting human-chimp conserved and diverged, non-repetitive, non-CpG nucleotides in windows of 20 kb. Low-quality alignments (divergence larger than 10%) were excluded from the analysis. Rates were computed separately for intronic and intergenic regions. Windows with less than 500 intergenic (intronic) nonrepetitive and aligned nucleotides were excluded from the analysis.

Computing empirical CpG divergence rates in bins of regional rate and M-score (Fig. 2B). CpG dinucleotides in the aligned human and chimp genome were grouped into two-dimensional bins according to their regional mutation rate and M-score, using mutation rate bins of size 0.002 and M-score bins of size 2. M-scores were computed from the human sequence (computing M-scores from the chimp sequence provide very similar results). For each bin, the joint distribution of human and chimp dinucleotides was assessed, and denoted by:

$\mathbf{Q_b(d_1,d_2)}$ = fraction of aligned d1 (human) and d2 (chimp) in bin b.

The joint distribution was reconstructed separately for intergenic and intronic sequences, with very similar results, intron being slightly more conserved. To gain accuracy, the two distributions (intergenic and intronic) were used separately (see below). Figure 2B represents the CpG divergence rate in intergenic bins by plotting:

$$\mathbf{1-Q_b(CG,CG)/\sum_d Q_b(CG,d).}$$

The COCAD assay. The COCAD (COntext aware CpG Analysis of Divergence) assay is a simple heuristic application of the M-score model and the $\mathbf{Q_b}$ empirical distributions. After extensive experimentation with principled maximum likelihood based models (which will be described elsewhere), the empirical approach was preferred as being conservative and robust. The empirical approach does not attempt to reconstruct the ancestral sequence or to model the irreversible deamination process explicitly. Instead, the COCAD background hypothesis assumes that CpGs are evolving independently once their M-score and regional divergence rates are given. The divergence probabilities are computed using the $\mathbf{Q_b}$ distributions and the assay is analyzing only genomic positions with a CpG in either the human or chimp genome. It is thus ignoring positions which possibly lose a CpG in both lineages, and rely on the relative proximity of the chimp and human genome to increase the probability that the vast majority of CpGs in the human-chimp ancestral genome were conserved in at least on of the species. The COCAD assay tests the neutral hypothesis in a sliding window (here of size 20 kb). In a given window, all loci bearing a CpG in either the human or chimp genomes are being considered. For each such CpG, the

observed divergence equals 1 if the CpG was not conserved between human and chimp and zero otherwise. The divergence probability for that CpG is computed by looking up the joint distribution of the bins defined by the locus's regional mutation rate and plus- and minus- strand M-scores. Note that we are heuristically averaging the estimates from the two strands m-scores and that these are typically very similar. Denote the appropriate bins as **pb** for the plus strand M-score and **mb** for the minus strand M-score. The divergence probability is defined as:

$$\mathbf{1-(Q_{pb}(CG,CG)/p_{pCG} + Q_{mb}(CG,CG)/p_{mCG})/2,}$$

where

$\mathbf{p_{pCG} = (\sum_d Q_{pb}(CG,d) + \sum_d Q_{pb}(d,CG) - Q_{pb}(CG,CG))}$ (fraction of positions with at least one CpGs in the positive m-score bin)

$\mathbf{p_{mCG} = (\sum_d Q_{mb}(CG,d) + \sum_d Q_{mb}(d,CG) - Q_{mb}(CG,CG))}$ (fraction of positions with at least one CpGs in the negative m-score bin).

The **Q** distributions are intergenic or intronic according to the genomic context and the summation is done over all **d** dinucleotides. The COCAD score equals the Z-score of the sum of observed divergences for all CpGs in the window, given the total expected divergence and assuming the variance to be the sum of individual CpG variances (**p(1-p)**, were **p** is the CpG divergence probability). To use the Z-score for normal estimation of P-values, one have to consider windows with sufficiently high expected divergence (e.g., more than six), which is almost always the case for 20kb windows and the divergence rates typical to the human-chimp lineage.

Robustness of estimated model parameters and COCAD scores. The M-score and COCAD models are based on relatively few parameters that are estimated from very large amount of data. For example, for each offset, the M-score log-odds 16 parameters (dinucleotide frequencies) are estimated based on millions of CpG loci. It is therefore easy to verify the robustness of the M-score predictions using cross-validation (data not shown). For the **Q** distribution, the robustness of the parameters depends on the number of CpGs that fall into the appropriate two-dimensional bin. To ensure the robustness of the COCAD analysis, CpGs with extreme M-score levels (>20 or <-40) or regions with very high (>0.025) or very low (<0.004) regional mutation rates were excluded from the analysis. COCAD genomic analysis is subject to extensive multiple testing (all CpGs are used as sliding window centers), and the results reported here are significant even given the very conservative Bonferroni multiple testing correction (more sophisticated strategies for multiple testing correction would suggest a more permissive COCAD score significance threshold). To further minimize false positives, we have chosen a conservative COCAD threshold of -5, giving a corrected P-value of 10^{-6}. We note that the P-values reported are correct only given the null

hypothesis, which assumes CpGs are evolving independently once their M-scores and regional divergence rates are known.

Methylation profiling. Regions within the hyperconserved domains were selected for Southern blot analysis based on high COCAD score, high Suz12 binding, and substantial numbers of McrBC and HpaII sites. High-molecular-weight DNA was subjected to two rounds of digestion with McrBC or HpaII, followed by digestion with a methylation-insensitive enzyme so that the region could be visualized as a discrete band (BamHI for *TBX5*, PvuII for *FOXA1* and XmnI for *HOXD*). The samples were also digested with an additional methylation-insensitive enzyme that did not cut within the region but reduced the background on the blots (SphI for Tbx5, XbaI for FoxA1 and HoxD). Control samples were prepared by methylating the DNA at all CpG dinucleotides with M.SssI prior to digestion. Details of the methods can be found in Rollins et al. (4). .

1. Hinrichs AS, Karlchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, Diekhans M, Furey TS, Harte RA, Hsu F, *et al.* (2006) *Nucleoic Acids Res* **34**: D590-D598.
2. Lee TI, Jenner RG, Boyer LA, Guenther MG, Levine SS, Kumar RM, Chevalier B, Johnstone SE, Cole MF, Isono K, *et al.* (2006) *Cell* **125** : 301-313.
3. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, Burton J, Cox TV, Davies R, Down TA*, et al.* (2006) *Nat Genet* **38:**1378-1385.
4. Rollins RA, Haghghi F, Edwards JR, Das R, Zhang MQ, Ju J, Bestor TH (2006) *Genome Res* 16:157-163.

Note 1. Evidence for the neutrality of the M-score predictions. As discussed in the main text, CpG divergence rates can be affected by variable levels of germ-line methylation or by selection. The correlation between divergence and the surrounding sequence (as expressed by the M-score) may therefore be a neutral effect (caused mostly by inconsequential changes in germ-line methylation levels) or it may express increased selection on CpGs that are embedded in specific sequence contexts. There are several lines of evidence, which suggest that the neutral effects are more dominant than the non neutral ones.

*M-scores are estimated globally***.** Most of the CpGs that are used to estimate the parameters of the M-score model are unlikely to be functional given their genomic context. Coding regions are excluded from the analysis, and the M-score parameters remain similar even if gene promoters (up to 10kb) are excluded as well.

M-scores are strongly correlated with deamination rates, but less so with other type of mutations. The increase in CpG divergence for high M-score is primarily a result of increased deamination rates (mutations to TpG or CpA). The increase in other types of mutations can be largely explained by an increase in double mutations (data not shown). This associates the M-score with changes in methylation levels (which are thought to affect only deamination rates), rather than with changes in the selective pressure on CpGs.

M-scores are correlated with in vivo methylation levels. As shown in Fig. 9, M-scores are strongly correlated with in-vivo methylation levels as measured by the HEP project. According to this result, CpGs with higher M-scores are more likely to be methylated (in particular in sperm cells). Such CpGs are thereby undergoing rapid deamination and divergence which is not necessarily related to their function.

Heterozygosity of polymorphic CpG SNPs is anti-correlated with the M-score. If the association between the M-score and the divergence rate is a consequence of selection, one would anticipate SNPs at CpGs with low M-scores to appear at low average heterozygosity and SNPs at CpGs with high M-scores to appear at neutral average heterozygosity. Analysis of average heterozygosity in 145,587 human CpG SNPs reveals that while selection plays a major role in the evolution of CpGs (Fig. 10), a positive

correlation between M-score and heterozygosity cannot be detected. In fact, when analyzing human CpG SNPs that are conserved in chimp, the M-score is negatively (rather than positively) correlated to the SNPs' average heterozygosity [$P<10^{-10}$ (Spearman)]. This correlation can be explained by having more negative selection acting on CpGs that are conserved even though their M-score based mutation rate is high. The SNP data therefore support the neutrality of the divergence rates predicted from the CpGs' sequence contexts by the M-score model. It also indicates that independently from the M-score, selection does contribute significantly to CpG evolution.

Note 2. Testing deamination vs. non-deamination divergence rates for HCGD-PRC2 domains. We used the **Q** distributions to test if hyperconservation of CpG distributions at HCGDs is primarily due to lower rates of deamination or overall decreases in CpG mutability. To that end, the COCAD test was adapted so that instead of counting all divergence events, we counted separately divergences to TpG/CpA and divergences to other dinucleotides. In both cases, we computed the Z-scores as described for the standard case. According to this analysis, Mutation of CpG to TpG/CpA is occurring 62.4% less than expected in these regions while the rate of other mutations is only 24.5% lower than expected. The decrease in the rate of non deamination mutations is still significantly lower (Z-score = -8.7, $p<10^{-80}$) and cannot be explained by the effect of double mutations. Nevertheless, the pronounced decrease in CpG divergence is primarily a result of low deamination levels. As shown in Fig. 14, the decrease in deamination rates is correlated to the M-score, which further suggest that CpGs in HCGDs/PRC2 domains are hypomethylated in the germ line.