nature genetics

Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome

Michael Weber[1], Ines Hellmann[2,3], Michael B Stadler[1], Liliana Ramos[4], Svante Pääbo[2], Michael Rebhan[1] & Dirk Schübeler[1]

To gain insight into the function of DNA methylation at *cis*-regulatory regions and its impact on gene expression, we measured methylation, RNA polymerase occupancy and histone modifications at 16,000 promoters in primary human somatic and germline cells. We find CpG-poor promoters hypermethylated in somatic cells, which does not preclude their activity. This methylation is present in male gametes and results in evolutionary loss of CpG dinucleotides, as measured by divergence between humans and primates. In contrast, strong CpG island promoters are mostly unmethylated, even when inactive. Weak CpG island promoters are distinct, as they are preferential targets for *de novo* methylation in somatic cells. Notably, most germline-specific genes are methylated in somatic cells, suggesting additional functional selection. These results show that promoter sequence and gene function are major predictors of promoter methylation states. Moreover, we observe that inactive unmethylated CpG island promoters show elevated levels of dimethylation of Lys4 of histone H3, suggesting that this chromatin mark may protect DNA from methylation.

Cytosine methylation is the only covalent DNA modification described in mammals. Genetic studies have established that this epigenetic mark is required for embryonic development[1], genomic imprinting[2] and X-chromosome inactivation[3], and alterations in DNA methylation are linked to many human diseases, including cancer[4].

In mammals, methylation is restricted to CpG dinucleotides, which are largely depleted from the genome except at short genomic regions called CpG islands, which commonly represent promoters[5]. Cytosine methylation can interfere with transcription factor binding, yet repression seems to occur largely indirectly, via recruitment of methyl-CpG binding domain (MBD) proteins that induce chromatin changes[6]. Consequently, the strength of repression could depend on the local concentration of CpGs within the promoter. Indeed, it is established that methylation of CpG-rich promoters is incompatible with gene activity, yet no conclusive picture has emerged for promoters containing low amounts of CpGs[7,8]. Equally uncertain is the contribution of promoter DNA methylation to tissue-specific gene expression, which predicts a dynamic reprogramming during development[9]. Most CpG island promoters remain unmethylated even in cell types that do not express the gene[10]. However, changes in DNA methylation linked to tissue-specific gene expression have been seen sporadically on CpG-rich promoters[11,12], although other studies failed to show such a connection based on the analysis of a small set of genes[13,14]. This inconclusive picture is a consequence of the limited number of genes analyzed and is further complicated by potential artifacts resulting from studying immortalized cell lines, which accumulate aberrant methylation in culture[15].

Genomic depletion of CpG dinucleotides in mammals is thought to reflect inherent mutability of methylated cytosines as observed in bacteria[16] and *in vitro*[17]. Moreover, deamination of an unmethylated cytosine creates a uracil that is easily recognized by the base excision repair machinery, yet deamination of a methylated cytosine creates a thymine, leading to a potential C to T transition. Notably, two enzymes (thymine DNA glycosylase (TDG) and MBD4) have been reported to selectively remove thymine from a T:G mismatch in the context of CpG dinucleotides[18,19], thus questioning if C to T transitions are mandatory. In light of these repair pathways, the evolutionary dynamics of CpGs could depend on positive or negative selection for CpGs rather than methylation in the germline. However, current estimates are mostly derived indirectly from sequence rather than actual measurement of DNA methylation[20,21].

To test models on the genomic distribution of DNA methylation and its impact on gene activity and sequence evolution, we generated an epigenomic map of DNA methylation, RNA polymerase II occupancy and chromatin state for 16,000 promoters in human primary somatic and germline cells. We find that both methylation frequency and its silencing potential are related to a gene's promoter sequence and the function of its product, and we propose that weak CpG islands are predisposed to *de novo* methylation during differentiation.

[1]Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, CH-4058 Basel, Switzerland. [2]Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, D-04103 Leipzig, Germany. [3]University of Copenhagen, Universitetsparken 15, Copenhagen ∅, Denmark, 2100. [4]Department of Obstetrics and Gynaecology, Radboud University Nijmegen Medical Center, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands. Correspondence should be addressed to D.S. (dirk@fmi.ch).

Received 31 October 2006; accepted 29 January 2007; published online 4 March 2007; doi:10.1038/ng1990

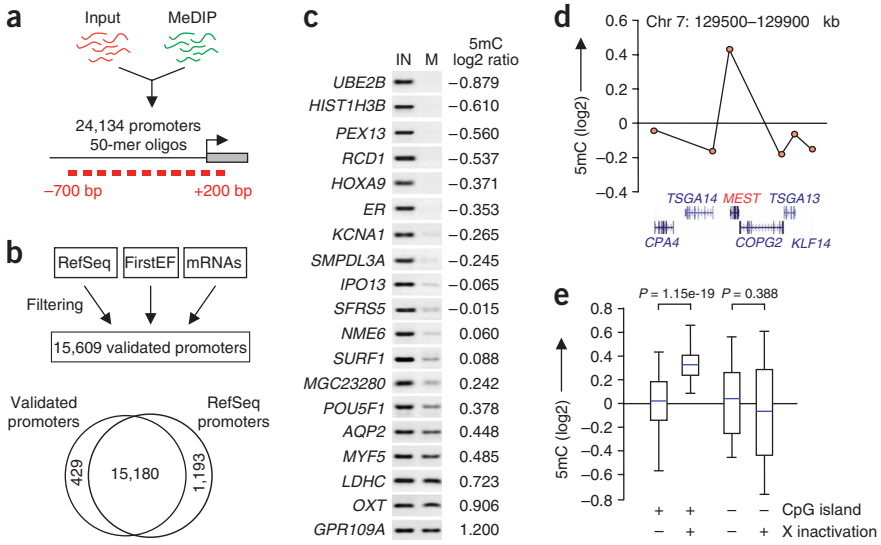

Figure 1 Defining the promoter methylome in human primary fibroblasts. (**a**) Input DNA and 5-methylcytosine (5mC)-enriched MeDIP samples were cohybridized to a high-density oligonucleotide microarray representing human promoters. Promoter methylation levels are calculated as the average of oligonucleotide ratios (5mC bound over input) between –700 bp and +200 bp relative to the transcription start site (**Supplementary Fig. 1**). (**b**) To remove potentially falsely annotated promoters, we filtered them based on RefSeq, FirstEF and mRNA annotations (see Methods). The Venn diagram illustrates that the validated promoters largely overlap with promoters of RefSeq genes. (**c**) Validation of microarray results. Randomly selected promoters were amplified by PCR from input (IN) and MeDIP-enriched (M) fractions. In each case, the PCR reflects the enrichment measured on the microarray (given as a log2 ratio). (**d**) Microarray detection of DNA hypermethylation on the promoter of the imprinted *MEST* gene, as previously described[49]. The dots mark the methylation level (log2 ratio) of RefSeq gene promoters shown below the graph. (**e**) Promoter DNA methylation on the X chromosome. Promoter sequences were matched to published X inactivation expression data[45]. Box plots show promoter methylation levels for genes subjected to (+) or escaping (–) X-inactivation, depending if promoters contain a CpG island. Only CpG island promoters of genes that undergo X inactivation show hypermethylation. Here and in all figures, the blue line marks the median, lower and upper limits of the box mark the 25th and 75th percentiles, and lower and upper horizontal lines mark the 10th and 90th percentiles. *P* values were calculated using a *t*-test.

RESULTS

Profiling promoter DNA methylation in the human genome

To determine the methylation status for a comprehensive set of human promoters, we enriched methylated DNA from human primary fibroblasts using methylated DNA immunoprecipitation (MeDIP) methodology[22] and combined it with microarray detection. The chosen array represents 24,134 putative human promoters, each covered by 15 oligonucleotides spanning from 1.3 kb upstream to 0.2 kb downstream of the transcription start site (**Fig. 1a**). To eliminate potentially falsely assigned promoters that might represent intergenic regions, we used experimental and computational evidence from various sources (see Methods) to generate a subset of 15,609 high-confidence promoters (**Fig. 1b**). These promoters largely overlapped with start sites of Ref Seq genes (**Fig. 1b**) and were used in all further analyses. In addition, measurements were limited to oligonucleotides from 700 bp upstream to 200 bp downstream of the transcription start site, to reduce noise caused by distal oligonucleotides residing in upstream intergenic regions (**Supplementary Fig. 1** online). The measurements for each promoter proved to be highly reproducible between biological replicates (*R* ranging from 0.91 to 0.95; see **Supplementary Fig. 2** online and Methods), from which we calculated a mean value.

Single-gene controls confirmed that the array measurements accurately reflected the enrichment in the MeDIP procedure (**Fig. 1c**). Among genes with high promoter DNA methylation, we detected a number of imprinted genes previously shown to have allele-specific promoter methylation (**Fig. 1d**). In agreement with the link between promoter DNA methylation and X chromosome inactivation in females[3], we also observed that promoter DNA methylation was higher on the X chromosome than on autosomes (**Supplementary Fig. 3** online). This reflects CpG island promoter methylation of genes that undergo X inactivation; genes that escape X inactivation were indistinguishable from autosomal genes (**Fig. 1e**). Notably, non–CpG island promoters did not show differential DNA methylation in relation to their X inactivation status (**Fig. 1e**), suggesting that their inactivation was not reflected in changes in DNA methylation (see below).

Promoter classes in relation to CpG frequency

Approximately 70% of human genes are linked to promoter CpG islands, whereas the remaining promoters tend to be depleted in CpGs[21]. This is evident in our set of 15,609 promoters, which had two distinct populations with high and low CpG frequency (**Fig. 2**). However, both populations showed a substantial overlap corresponding to promoters with intermediate CpG frequency. We hypothesized that these might differ from low and high CpG promoters in their regulation by DNA methylation. Therefore, we defined three classes of promoters based on CpG ratio, GC content and length of CpG-rich region (see Methods for details). High-CpG promoters (HCPs) and low-CpG promoters (LCPs) form two nonoverlapping populations that represent strong CpG islands and clear non–CpG island promoters, respectively (**Fig. 2**). Promoters with intermediate CpG content (ICPs) contain many promoters that are close to the CpG island criteria introduced in ref. 23, and 91% of them (compared with 8% of LCPs and 100% of HCPs) fulfill the less-stringent CpG island criteria defined in ref. 24; therefore, ICPs will also be referred to as 'weak' CpG islands.

To estimate differences in expression patterns between the three classes, we matched the promoters with a set of 2,018 housekeeping genes defined from public expression data (see Methods). These housekeeping genes are unevenly distributed in the classes, as they are 1.2-fold overrepresented in the HCP class, 1.2-fold underrepresented in ICPs and 2.3-fold underrepresented in LCPs (χ^2 test: $P = 4.6 \times 10^{-37}$). This agrees with previous reports showing that CpG island promoters are more frequently, but not exclusively, associated with housekeeping genes[21].

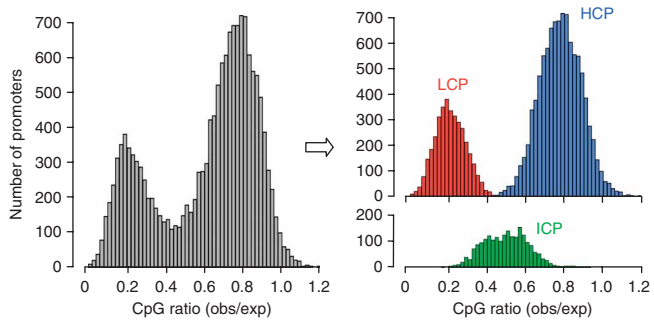

Figure 2 Promoter classification based on CpG representation. The gray histogram represents the distribution of observed versus expected CpG frequencies for all 15,609 promoters analyzed, showing a bimodal distribution of CpG-rich and CpG-poor promoters. To separate two nonoverlapping populations, lower- and higher-stringency criteria were used to define the low (LCPs, red, $n = 3,627$) and high (HCPs, blue, $n = 9,928$) CpG content groups, as well as a smaller group with intermediate CpG content (ICPs, green, $n = 2,054$) (see Methods for details on calculations).

Differential methylation of promoter classes in somatic cells

Figure 3a shows the DNA methylation levels in primary fibroblasts for all autosomal promoters in the three classes relative to their CpG content. In the case of HCPs, most promoters showed MeDIP enrichments close to the median, whereas a small subset of promoters showed strong enrichment (**Fig. 3a**). Bisulfite genomic sequencing confirmed that the least-enriched HCPs were free of methylated cytosines, whereas those with enrichments around the median contained a few methylated cytosines, which, owing to the high CpG content, translates into a low percentage of methylation (for example, for *CASP2*, 4 out of 61 CpGs were methylated (7% methylation); **Fig. 3b** and **Supplementary Fig. 4**). HCPs with MeDIP enrichment above 0.4 were strongly methylated (**Fig. 3b** and **Supplementary Fig. 4**), and these represent 3% (292 out of 9,527) of autosomal HCPs. Therefore, as predicted from previous work[20], CpG islands remain mostly free of DNA methylation even in terminally differentiated cells, yet 3% of HCPs show high methylation.

Weak CpG islands showed a markedly higher frequency of DNA methylation (**Fig. 3a**): 21% (385 out of 1,841) of autosomal ICPs showed high methylation values (log2 ratio >0.4) indicative of complete methylation, as confirmed by bisulfite genomic sequencing (**Fig. 3b**). LCPs showed a different pattern of DNA methylation: we observed a positive correlation between promoter enrichment and CpG content (**Fig. 3a**). This dependency can be reconciled if most LCPs show a high rate of CpG methylation, and consequently their enrichment becomes a function of their number of CpGs. Indeed, bisulfite genomic sequencing on randomly chosen promoters showed that most LCPs were methylated (**Fig. 3b** and **Supplementary Fig. 5** online). Thus, low enrichment in the LCP class does not reflect an unmethylated state but rather the low abundance of substrate to be recognized by the 5-methylcytosine (5mC) antibody. Similar to HCPs, modest enrichments around the median represent few methylated CpGs, yet in LCPs this translates into a high relative methylation level owing to low CpG content (for example, 4.5 out of 5 CpGs (90%) were methylated in *EHF*; **Fig. 3b**). We conclude that LCPs are overall methylated, HCPs are almost exclusively unmethylated and ICPs show a high frequency of methylation. Consequently, LCPs and ICPs are largely overrepresented among hypermethylated promoters (**Fig. 3c**).

Polymerase occupancy in relation to DNA methylation

Next, we determined the activity of all promoters by measuring RNA polymerase II occupancy using chromatin immunoprecipitation

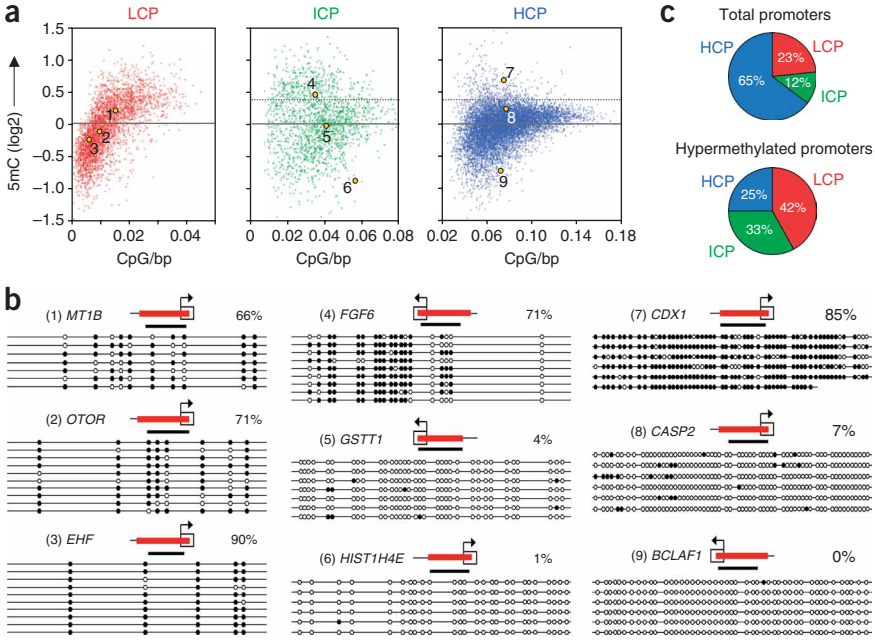

Figure 3 Frequency of DNA methylation in promoter classes. (**a**) The scatter plots show the DNA methylation levels for all promoters relative to their CpG content (CpG/bp) for the three promoter classes. Each spot represents one promoter. The dashed line marks the threshold (log2 ratio = 0.4) above which promoters in ICP and HCP classes are classified as hypermethylated based on bisulfite sequencing (**Fig. 3b** and **Supplementary Fig. 4**). A similar threshold does not apply to LCPs, as in this class, MeDIP enrichment can be limited by the low number of CpGs even in the methylated state (see **Fig. 3b** and main text). Numbered promoters refer to the bisulfite controls shown in **b**. (**b**) Bisulfite sequencing controls for a subset of promoters in each class. The red line indicates the region covered by the oligonucleotides on the microarray, and the black line the region amplified for bisulfite sequencing. CpGs are represented as open dots (if unmethylated) or filled dots (if methylated). The percentage of CpG methylation is indicated for each promoter. Additional bisulfite controls are shown in **Supplementary Figures 4** and **5**. (**c**) Pie charts showing the relative frequency of classes among total promoters and hypermethylated promoters (defined by log2 ratio > 0.4). LCPs and ICPs are largely overrepresented among hypermethylated promoters (χ^2 test: $P = 8 \times 10^{-258}$). Note that the percentage of LCPs among hypermethylated promoters is underestimated, as many fully methylated LCPs do not contain sufficient CpGs to pass the 0.4 enrichment threshold (see text).

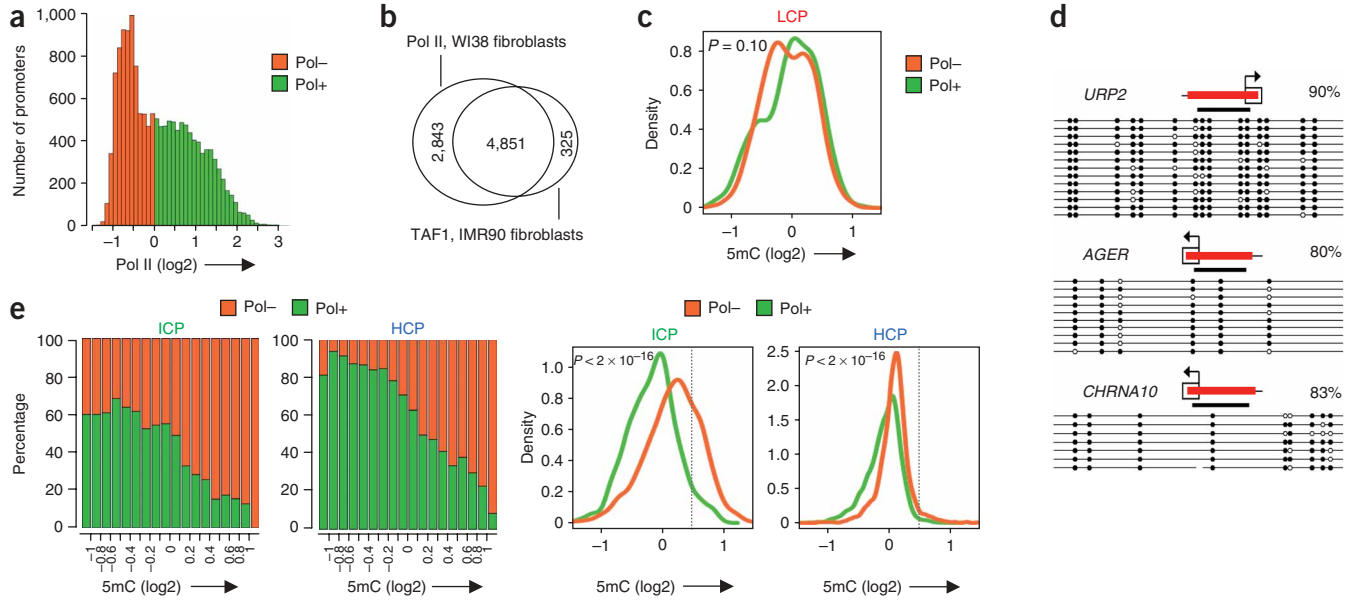

Figure 4 Functional consequence of DNA methylation on promoter activity depends on CpG content. (**a**) Density histogram representing the promoter enrichments for RNA polymerase II (Pol II). Active promoters (marked in green) are defined as having a log2 ratio >0. (**b**) The Venn diagram compares active promoters identified in this study with TAF1/Pol II binding sites identified in unrelated primary fibroblasts[25]. Of the TAF1/Pol II sites present on our array, 94% (4,851 out of 5,176) are also scored as active. Notably, we also identify additional active promoters that presumably use initiation factors other than TAF1 (ref. 50). (**c**) Density plot comparing the distribution of DNA methylation values for active (green) and inactive (orange) LCPs, which show no significant differences. The *P* value was calculated using a Wilcoxon test. (**d**) Bisulfite genomic sequencing on selected active LCPs, confirming that these are hypermethylated (**Supplementary Fig. 5**). (**e**) The left panels show the percentage of active and inactive promoters relative to increasing DNA methylation for the ICP and HCP classes. The percentage of active promoters decreases with increasing methylation levels, showing that promoter activity and hypermethylation are incompatible for ICPs and HCPs. Right panels show density plots comparing the distribution of DNA methylation values for active and inactive promoters. The vertical dashed line marks the threshold for hypermethylation (log2 ratio = 0.4). These plots illustrate the high frequency of DNA methylation among inactive ICPs, whereas most inactive HCPs remain unmethylated. *P* values were calculated using a Wilcoxon test.

(ChIP). The enrichment profile showed a bimodal distribution, which we used to define a set of polymerase-bound and presumably active promoters (**Fig. 4a**). A comparison with TAF-1 and polymerase II–bound promoters identified in unrelated human fibroblasts[25] showed marked similarity. Of those promoters identified in ref. 25 that are represented on our microarray, 94% were scored active in our data set (**Fig. 4b**). The frequency of activity varies between promoter classes, with 66% of HCPs being active compared with 41% of ICPs and 11% of LCPs. This reflects the enrichment of housekeeping genes in CpG island promoters and the higher abundance of rarely expressed tissue-specific genes in non-CpG island promoters, as demonstrated above.

Low CpG promoters showed no significant correlation between gene activity and the abundance of methylated cytosines, suggesting that active LCPs are not preferentially unmethylated. Indeed, the distribution of DNA methylation values for active and inactive LCPs was not different (**Fig. 4c**). Bisulfite sequencing on a number of active LCPs confirmed their methylated state (**Fig. 4d** and **Supplementary Fig. 5**). We confirmed that these methylated promoters are sites of transcriptional initiation by showing that polymerase binding is biased toward the predicted start sites (**Supplementary Fig. 5**). Notably, the promoter of the highly expressed *FGF7* gene was hypomethylated in primary fibroblasts (**Supplementary Fig. 5**), opening the possibility that a subset of LCPs is unmethylated when active. We conclude that the majority of low CpG promoters are methylated in the inactive as well as in the active state, implying that low concentrations of methylated cytosines do not preclude gene activity.

In contrast to LCPs, the activity of ICPs and HCPs was negatively correlated with their DNA methylation status. The percentage of active genes decreased to low levels for promoters showing elevated DNA methylation (**Fig. 4e**), indicating that DNA methylation of ICPs and HCPs is largely incompatible with their activity. However, inactive ICPs and HCPs differed in their frequency of DNA methylation. Whereas the vast majority of inactive HCPs remained unmethylated, a much higher proportion of inactive ICPs was hypermethylated (**Fig. 4e**). Thus, HCPs remain unmethylated even when inactive, whereas inactive ICPs are frequently methylated, implying that they are less protected against *de novo* methylation.

Inactive CpG islands reside in active chromatin

To gain insight into potential mechanisms preventing DNA methylation of CpG island promoters, we tested if they are associated with an established mark of transcriptionally active chromatin: dimethylation of Lys4 of histone H3 (H3K4)[26]. Active promoters show overall higher levels of dimethylated H3K4 than inactive promoters (**Fig. 5a**), confirming previous work in higher eukaryotes[25,27], but we were surprised to find that inactive promoters formed two populations with different levels of dimethylated H3K4 (**Fig. 5a**) that mirrored their DNA methylation status. Inactive HCPs, which remain largely hypomethylated, showed elevated H3K4 dimethylation compared with inactive LCPs and most ICPs (**Fig. 5b**). The rarely methylated HCPs show no enrichment of dimethylated H3K4, but they form too small of a group to be visible in the density plot. Among inactive ICPs, only unmethylated promoters showed enrichment of dimethylated H3K4 similar to HCPs, whereas hypermethylated ones showed no enrichment (**Fig. 5c**). We conclude that CpG-rich promoters that are protected from DNA methylation are associated with elevated levels of

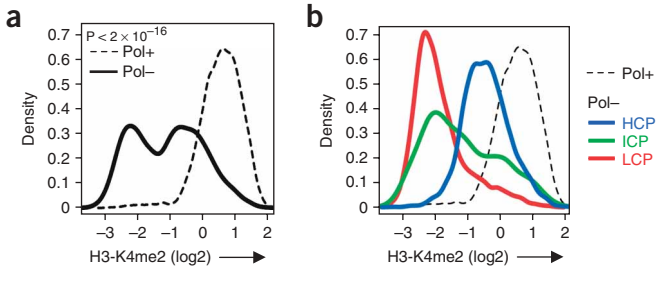

Figure 5 Elevated levels of H3K4 dimethylation mark inactive CpG islands. (**a**) Density plots comparing H3K4 dimethylation profiles for active and inactive promoters of all classes. Active promoters show uniform high H3K4 dimethylation, whereas inactive promoters show both intermediate and low levels evident as two separate peaks. The *P* value was calculated using a Wilcoxon test. (**b**) Profiles of H3K4 dimethylation for inactive promoters in each promoter class. This shows that promoters with an intermediate level of H3K4 dimethylation represent mainly HCPs and a subset of ICPs. The H3K4 dimethylation profile for active promoters is shown as a dashed line for comparison. (**c**) The box plot represents the distribution of H3K4 dimethylation values for inactive ICPs and HCPs that are hypomethylated (5mC log2 ratio <0.4) or hypermethylated (5mC log2 ratio >0.4). This demonstrates that only hypomethylated promoters show elevated H3K4 dimethylation, whereas hypermethylated promoters show no enrichment of H3K4 dimethylation. The *P* value was calculated using a *t*-test.

dimethylated H3K4 in the absence of transcription. This shows that a chromatin state can predict the DNA methylation state of inactive CpG-rich promoters and opens the possibility that chromatin structure is functionally involved in protecting CpG-rich promoters from DNA methylation.

Dynamic DNA methylation between soma and germline

To establish if the observed promoter methylation profiles are unique to somatic cells, we determined the promoter methylome in mature sperm, the product of the male germline. The MeDIP experiments proved to be highly reproducible when comparing sperm samples from the same ($R = 0.95$) or genetically unrelated donors ($R = 0.91$, **Supplementary Fig. 2**). The LCP class showed high similarity in DNA methylation patterns between fibroblasts and sperm (**Fig. 6a** and **Supplementary Fig. 6** online): 79% (373 out of 472) of the hypermethylated promoters from fibroblasts were also highly enriched in sperm (**Supplementary Fig. 6**). Similar to fibroblasts, methylation enrichment of LCPs in sperm increased with CpG content (**Supplementary Fig. 6**), indicating that constitutive methylation in this class was present in both somatic cells and gametes. In contrast, hypermethylation of ICPs and HCPs detected in fibroblasts was mostly absent in germ cells (**Fig. 6a** and **Supplementary Fig. 6**). Among HCPs and ICPs that were hypermethylated in the somatic sample, 86% (236 out of 276) and 49% (184 out of 373), respectively, were unmethylated in sperm (**Supplementary Fig. 6**). Thus, most hypermethylation of CpG-rich promoters in fibroblasts seems to be somatically acquired, indicating that a defined subset of CpG islands becomes *de novo* methylated during development. Notably, the frequency of this acquisition is higher in ICPs, suggesting that weak CpG islands are more prone to methylation during differentiation.

Evolutionary impact of CpG methylation

CpG depletion in the human genome is thought to reflect a higher mutation rate of methylated cytosines[16] in the germline. Using the promoter methylome of the sperm sample, we tested if promoter hypermethylation in germ cells was manifested in an increased rate of CpG loss. To infer rates of ongoing CpG loss and gain in the human lineage, we used the AMBIORE package[28] to perform three-way alignments of the human, chimpanzee and rhesus macaque genomes (using rhesus as an outgroup to assess the directionality of CpG mutations). This demonstrated that CpG loss was considerably higher for LCPs than for ICPs and HCPs, whereas CpG gain and non-CpG divergence was indistinguishable (**Fig. 6b** and data not shown). Given that LCPs were mostly methylated in sperm, this favors the model that DNA methylation induces CpG depletion in these promoters. To further relate CpG loss with DNA methylation, we divided the ICP class based on their methylation status in sperm and observed that CpG loss was higher for methylated promoters than for the unmethylated promoters (**Fig. 6c**). Therefore, within the same promoter class, promoter DNA methylation in the product of the male germline was associated with an increased evolutionary loss of CpGs. Notably, ICPs seem to lose CpG noticeably faster than HCPs even when unmethylated in sperm, which could reflect either temporary methylation in the germline or an inherent selection for CpG loss at ICPs.

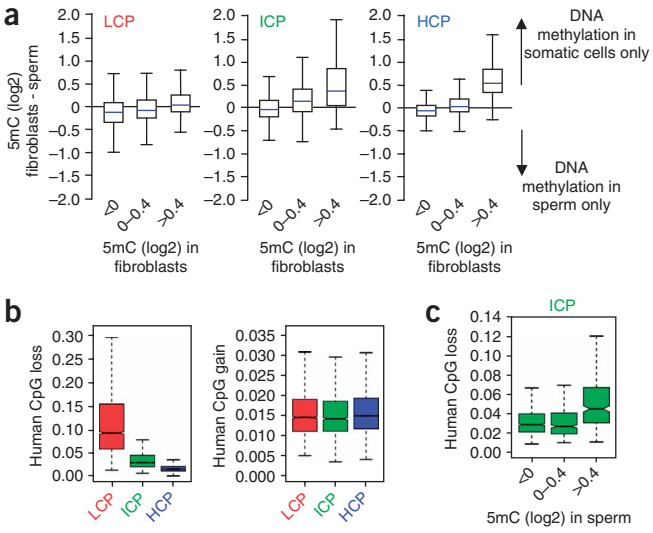

Figure 6 Promoter DNA methylation in the germline is associated with CpG loss. (**a**) Comparison of DNA methylation of autosomal promoters in human primary fibroblasts and sperm. In each class, promoters were grouped in three bins based on their DNA methylation level in fibroblasts. For promoters in each bin, we subtracted the methylation measurement in sperm from that in fibroblasts. A positive value reflects higher methylation in somatic cells than in the germline. The box plots illustrate that methylation of LCPs is very similar between fibroblasts and sperm, whereas hypermethylation of ICPs and HCPs detected in fibroblasts (log2 ratio >0.4) is largely specific to somatic cells. (**b**) Comparison of human, chimpanzee and rhesus sequence was used to define CpG loss and CpG gain in the human lineage (see main text and Methods). CpG loss and gain are shown for each promoter class, illustrating the higher rate of CpG loss in the constitutively methylated LCP group compared with ICPs and HCPs. (**c**) CpG loss for ICPs, sorted according to methylation status in sperm (hypermethylation: 5mC log2 ratio >0.4; hypomethylation: 5mC log2 ratio <0.4). This illustrates the link between DNA methylation in the germline and a higher rate of ongoing CpG depletion.

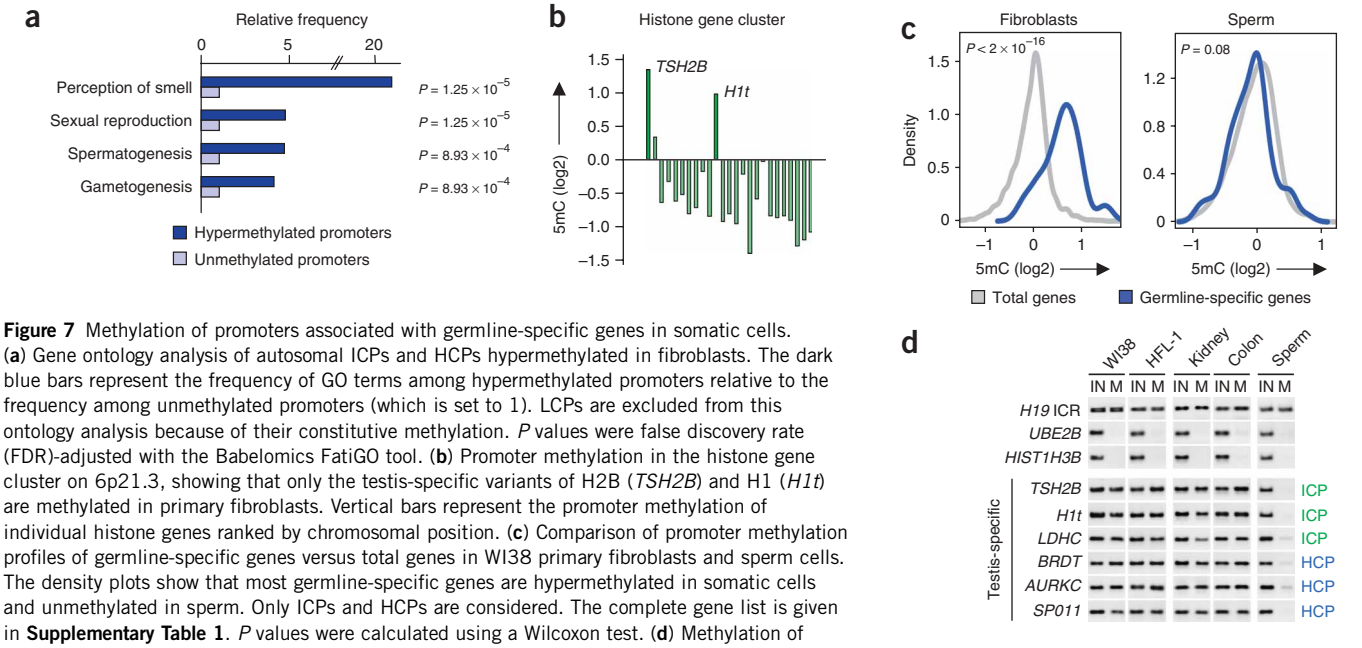

Figure 7 Methylation of promoters associated with germline-specific genes in somatic cells. (**a**) Gene ontology analysis of autosomal ICPs and HCPs hypermethylated in fibroblasts. The dark blue bars represent the frequency of GO terms among hypermethylated promoters relative to the frequency among unmethylated promoters (which is set to 1). LCPs are excluded from this ontology analysis because of their constitutive methylation. *P* values were false discovery rate (FDR)-adjusted with the Babelomics FatiGO tool. (**b**) Promoter methylation in the histone gene cluster on 6p21.3, showing that only the testis-specific variants of H2B (*TSH2B*) and H1 (*H1t*) are methylated in primary fibroblasts. Vertical bars represent the promoter methylation of individual histone genes ranked by chromosomal position. (**c**) Comparison of promoter methylation profiles of germline-specific genes versus total genes in WI38 primary fibroblasts and sperm cells. The density plots show that most germline-specific genes are hypermethylated in somatic cells and unmethylated in sperm. Only ICPs and HCPs are considered. The complete gene list is given in **Supplementary Table 1**. *P* values were calculated using a Wilcoxon test. (**d**) Methylation of germline-specific genes in other somatic tissues. Candidate promoters were PCR amplified from input (IN) and MeDIP-enriched (M) fractions from WI38 and HFL-1 primary fibroblasts, primary kidney and colon samples and sperm cells. Germline-specific promoters are methylated in all somatic tissue samples tested. The promoter class of the tested genes is indicated on the right. The imprinted *H19* ICR serves as positive control for methylation, and the housekeeping genes *UBE2B* and *HIST1H3B* as unmethylated negative controls.

Promoter methylation of germline-specific genes in soma

Finally, to gain insights into the biological roles of DNA methylation in somatic cells, we asked if methylated CpG-rich promoters in primary fibroblasts regulate genes involved in specific biological processes. Of the rarely hypermethylated HCPs, 17% are linked to genes showing a testis-specific expression, according to GNF SymAtlas[29] (which does not provide expression data for the human female germline), including well-studied genes expressed in both male and female germline, such as *DAZL*, *SPO11*, *SOX30*, *BRDT*, *ALF*, *TPTE* or *REC8* (refs. 30,31). To confirm this observation, we analyzed Gene Ontology annotations for methylated autosomal ICPs and HCPs and observed a significant enrichment for ontology terms related to generation of gametes (**Fig. 7a**). The only other enriched GO category in the methylated fraction refers to perception of smell and reflects DNA methylation of a small subgroup of olfactory receptor genes that contain CpG-rich promoters (data not shown). This unique methylation of germline-specific genes is illustrated by the histone gene cluster, where the testis-specific histone variants *HIST1H2BA* (known as *TSH2B*) and *HIST1H1T* (known as *H1t*) show high promoter DNA methylation, as reported in rodents[32,33] (**Fig. 7b**). We confirmed this observation by PCR (**Fig. 7**) and bisulfite sequencing (**Supplementary Fig. 4**). Notably, the majority of described germline-specific genes (**Supplementary Table 1** online) showed hypermethylation (**Fig. 7c**), indicating that this process happens quantitatively in this class of genes. This methylation of germline-specific genes was absent in mature sperm (**Fig. 7c,d**), suggesting that it is established after fertilization during somatic development. Moreover, it was not unique to the particular cells we studied, as we observed it in genetically unrelated male fibroblasts and primary samples from kidney and colon (**Fig. 7d**). We conclude that somatic cells show a systematic methylation of promoters for germline-specific genes, including strong CpG islands that are otherwise protected from DNA methylation.

DISCUSSION

Previous models of the distribution and function of DNA methylation at *cis*-regulatory regions have been deduced from small data sets or inferred indirectly from DNA sequence. Moreover, the impact of DNA methylation on transcription was determined using approaches such as transient transfections[7,8] or genomic targeting of random integration sites[34], which do not necessarily recapitulate the endogenous chromosomal situation. In each case, low sampling numbers limited the potential to generalize findings, especially when exceptions occur at low frequencies. Thus, our comprehensive analysis of DNA methylation, polymerase occupancy and chromatin state of 15,609 promoters provides a useful framework to derive quantitative and predictive models of promoter DNA methylation (**Fig. 8**).

We find the vast majority of strong CpG island promoters (HCPs) hypomethylated on autosomes, in agreement with previous observations[10,20,35] and computational predictions[36]. Thus, even though DNA methylation is sufficient to inactivate CpG island promoters, it is not necessary, as most inactive CpG island promoters are unmethylated. The fact that transcription seems not to be required to maintain a hypomethylated state points to alternative mechanisms that protect CpG islands against *de novo* methylation. Our results imply chromatin structure as a putative pathway, as hypomethylated CpG islands show elevated levels of H3K4 dimethylation even in the absence of transcription. Dimethylation of H3K4 occurs uniformly on all CpG island promoters, arguing that it is an inherent characteristic of CpG islands. Equally notably, H3K4 dimethylation is not shared by the LCP class (**Fig. 5**), which contain as few methylated cytosines as HCPs; therefore, H3K4 dimethylation seems to require a local concentration of unmethylated CpGs. In line with this model, recruitment of H3K4 methylases by unmethylated CpGs has recently been suggested[37,38]. Moreover, the euchromatic features of CpG islands do not seem to be limited to H3K4 methylation, as a broad H3 hyperacetylation in CpG

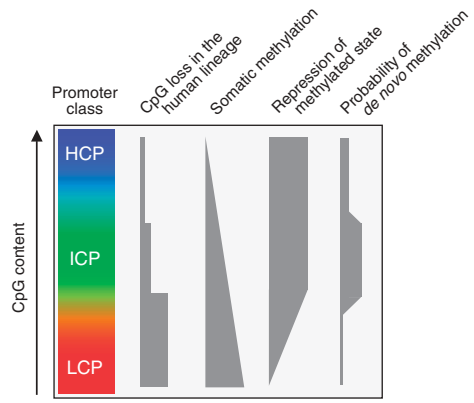

Figure 8 Regulation of promoter DNA methylation in the human genome. Schematic representation of promoter CpG content (which translates into the different classes) relative to frequency of hypermethylation, impact of methylation on sequence evolution, ability of methylated state to repress transcription and likelihood of *de novo* methylation in somatic cells. This synopsis illustrates that weak CpG islands (ICPs) are prone to regulation by DNA methylation, as they show frequent DNA methylation in somatic cells, and this methylated state precludes their activation. The width of each bar represents frequency of the event or strength in case of transcriptional repression.

islands has been reported[39]. These observations make it conceivable that an active chromatin state is involved in precluding DNA methyltransferase (DNMT) recruitment to CpG islands. However, it also raises the question of how spurious activation of such accessible promoters is prevented.

In contrast to CpG islands, promoters with low CpG content (LCPs) are predominantly methylated, in agreement with recent bisulfite sequencing results on human chromosomes[35]. We now show that this hypermethylation does not preclude gene expression. The lack of repression of low abundance of 5mC is also illustrated in the HCP class, where many active promoters contain a low percentage of methylation (4%–7%; see **Figs. 3** and **4**). This indicates that repression by DNA methylation requires high 5mC density. In light of the prevailing model of an indirect repression pathway by MBD protein, this suggests that MBD binding is not sufficient at low DNA methylation density for active repression. However, this does not exclude a role for low-density methylation in reducing transcriptional noise that could be generated by spurious initiation[40]. If it indeed occurs, such regulation might be more prominent at tissue-specific genes, which are enriched among LCPs. Of note, we also observe a low number of LCPs that are unmethylated and active, opening the possibility that at some LCPs, demethylation occurs upon gene activation.

The dynamics and role of DNA methylation in somatic cell differentiation is controversial[13]. Our data argue that dynamic DNA methylation cannot be a default repression mechanism for tissue-specific gene expression, as most inactive CpG island promoters remain unmethylated in primary cells. However, we identify several hundred CpG island promoters (4% of the total number in the studied fibroblasts) that are methylated in somatic cells but not in the germline, demonstrating that somatic methylation of CpG islands does occur at a significant rate in primary cells. Genomic imprinting is unlikely to account for most of this methylation, as alleles were found equally methylated in all six cases tested by bisulfite sequencing. Notably, this soma-specific methylation occurs more frequently at ICPs, indicating that weak CpG islands are preferential targets for *de novo* methylation in development (**Fig. 8**) and that the promoter sequence is a determinant of dynamic methylation. Preliminary data in mouse suggest that preferential targeting of weak CpG islands is a general phenomenon in mammals (F. Mohn, M. Bibel and D.S., unpublished data). One possibility is that protection from *de novo* methylation is a direct function of the local CpG density, making it more likely for weak CpG islands to become *de novo* methylated.

Targets for CpG island *de novo* methylation in somatic cells are also partly specified by the function of the linked gene, as germline-specific genes are preferentially methylated. This observation is in agreement with recent data on five genes in mouse[41,42], but we now show that it is a quantitative process, because almost all CpG island promoters of germline-specific genes are DNA methylated in somatic cells. Although it remains to be determined how DNA methylation is preferentially targeted to promoters of germline-specific genes and how this process is temporarily coordinated, we speculate that DNA methylation functions to preclude deleterious activation of meiotic genes in somatic cells. This finding predicts that the frequently observed ectopic expression of testis-specific genes in tumors entails promoter demethylation[43]. Notably, the preferential methylation of germline promoters and the increased frequency of ICPs methylation are probably independent processes, as most methylated germline-specific genes fall in the HCP class (**Supplementary Table 1**). Furthermore, germline-specific genes account only for a subgroup of somatically methylated CpG islands. The remaining targets do not represent defined ontology groups, yet we observe methylation of several tissue-specific transcription factors (for example, *CDX1*, *TFDP1*, *FHL2*, *NRF3*, *MYF5* and *RUNX3*), opening the possibility that *de novo* methylation could be used in part to prevent alternative differentiation pathways by selectively repressing lineage-specific genes.

The promoter methylome of male gametes also sheds light on the evolutionary consequences of DNA methylation. When comparing the human and chimpanzee genomes, we observe that promoters methylated in the product of the male germline show a higher rate of evolutionary CpG loss. Although the methylation state of other stages of the male and female germline remains to be tested, this finding provides evidence that the ongoing CpG depletion in the hominid lineage is DNA methylation dependent. However, a subset of ICPs (10% of total) show high methylation in sperm, but they are CpG rich. These might reflect evolutionarily recent methylation events, and consequently these promoters might have different epigenetic states between human and chimpanzee. Further work is necessary to address this possibility. At the same time, most ICPs are unmethylated in the germline, thus raising the question of why these promoters have a lower CpG content than expected. It is possible that this reflects a specific selection for intermediate CpG content promoters in mammalian genomes.

Our results demonstrate that DNA methylation is primarily a function of promoter CpG content, which results in a constitutive hypo- or hypermethylated state. On top of this stable framework, we identify a dynamic component that mediates soma-specific *de novo* methylation preferential to weak CpG islands. Although the exact mechanisms of targeting dynamic methylation are still elusive, our results suggest that in primary cells, both frequency of reprogramming and its impact on transcription are influenced by the composition of individual *cis*-regulatory regions.

METHODS

Array design and analysis. Samples were hybridized to a microarray representing promoter regions of 24,134 human genes (NimbleGen Systems,

HG17_min_promoter array). Each promoter is represented by up to 15 repeat-masked 50-mer oligonucleotides positioned on average every 100 bp from –1,300 bp to +200 bp relative to the transcription start site (TSS). Sample labeling, hybridization and data extraction were performed according to standard procedures by NimbleGen Systems. After hybridization, raw fluorescence values were extracted in a format compatible with Excel using a custom Perl script. To minimize noise coming from intergenic regions, we considered only oligonucleotides located in a window of 900 bp from the 5′-most oligonucleotide (–700 to +200 bp relative to the TSS). This filtering reduces the average number of sampled oligonucleotides per promoter to 11.05 but significantly increases the consistency of the measurements along each promoter as shown by a reduced s.d. between oligonucleotide values of the same promoter (**Supplementary Fig. 1**). Oligonucleotides showing an abnormally high input signal were ignored (>8,000 for MeDIP arrays and >15,000 for ChIP arrays, representing on average 2% of all oligonucleotides). We considered promoters with at least seven oligonucleotide measurements after this filtering. We calculated the bound-to-input ratio between Cy3 and Cy5 signals for each oligonucleotide, and final promoter values are the mean of individual oligonucleotide log2 ratios. The resulting promoter values were median normalized to log2 = 0. All data processing and analysis was performed using Excel, Spotfire DecisionSite and the R package (see URL's section below).

Promoter annotation. The promoter set present on the array was filtered *in silico* to remove redundant promoters, promoters spanning less than 400 bp, promoters on the Y chromosome and poorly supported promoters that might reflect intergenic DNA methylation. For this, we retrieved the following annotations from the University of California Santa Cruz (UCSC) genome browser: (i) we matched promoters with the First Exon Finder (FirstEF) predictions[44], (ii) we counted RefSeq starts in a window of 300 bp around the potential TSS (defined as 150 bp upstream of the most downstream position of an oligonucleotide) and (iii) we counted mRNA starts in a window of 150 bp around the potential TSS. Validated promoters were defined as having either (i) a RefSeq start and a FirstEF prediction, (ii) a RefSeq start and at least one mRNA, (iii) a FirstEF prediction and at least two mRNAs or (iv) at least three mRNAs. All annotations refer to the May 2004 (hg17) human genome assembly. For the X inactivation analysis, we matched X-linked promoters to the genes assayed in a recent comprehensive X inactivation profile in human cells[45]. Genes were considered to escape X inactivation if they were expressed in more than three out of nine of the somatic cell hybrids in this publication. Housekeeping genes were identified with Affymetrix gene expression data from 79 tissues[29] using the method described in ref. 46 (housekeeping genes are defined as having a normalized expression level above 200 in all tissues). Matching of these genes to the promoter set identified a total of 2,018 housekeeping promoters. For comparison with the genome-wide TAF1/RNA polymerase II data[25], we mapped the 9,328 TAF1 binding sites provided in Supplementary Table S1 of ref. 25 to our set of promoters after having relocated the promoters to the July 2003 (hg16) assembly using BLAT. TAF1 sites and promoters were considered to map if they had at least 1 bp overlap. The set of autosomal genes with germline-specific expression was generated from published literature[31,47], and their expression was systematically verified with the GNF SymAtlas[29] (see URL's section below). The analysis of gene ontology was performed by comparing the methylated autosomal ICPs and HCPs (5mC log2 ratio >0.4) with the unmethylated ones (5mC log2 ratio <0.3) using the Babelomics FatiGO tool (see URL's section below).

Definition of promoter classes. Promoters were classified in three categories to distinguish strong CpG islands, weak CpG islands and sequences with no local enrichment of CpGs. We determined the GC content and the ratio of observed versus expected CpG dinucleotides in sliding 500-bp windows with 5-bp offset. The CpG ratio was calculated using the following formula: (number of CpGs × number of bp) / (number of Cs × number of Gs). The three categories of promoters were determined as follows: HCPs (high-CpG promoters) contain a 500-bp area with CpG ratio above 0.75 and GC content above 55%; LCPs (low-CpG promoters) do not contain a 500-bp area with a CpG ratio above 0.48; and ICPs (intermediate CpG promoters) are neither HCPs nor LCPs. Thus, the ICP class contains many 'subthreshold' CpG islands (that is, CpG islands that are small (below 500 bp), have moderate CpG richness and/or have a GC content below 55%) with respect to the criteria defined in ref. 23. Consequently, 91% of ICPs fulfill the less stringent CpG island criteria defined in ref. 24. The computations above and the calculations of the number of CpGs per bp and the ratio of observed versus expected CpGs over the entire promoter were performed on the genomic sequence covered by the oligonucleotides plus 200 bp on both sides to account for the fact that DNA molecules containing flanking regions can also contribute to the hybridization signal.

Biological samples. Human female WI38 primary lung fibroblasts were obtained from the American Type Culture Collection (ATCC) and cultured in DMEM containing 10% FCS (37 °C, 5% CO_2). Primary samples from kidney and colon were obtained from M. Haase (Dresden University of Technology). Sperm samples were from two normospermic males attending Nijmegen Medical Center for routine diagnosis. Collection and cryopreservation occurred with written consent of the donors for this study. Samples were collected in sterile containers and purified by density gradient centrifugation (Pure Sperm, Nidacom) for 20 min (500*g*). This procedure was repeated twice to avoid contamination with other cell types. The purified sperm fraction was then diluted 1:1 with TEST yolk buffer medium (TYB, Irvine Scientific) and cooled in liquid nitrogen (vapor phase) for 15 min.

Methylation profiling by MeDIP. The MeDIP assay was performed on 4 µg sonicated genomic DNA (300–1,000 bp) as previously described[22]. Per array, the unamplified product of six MeDIP reactions (bound fraction) was hybridized together with sonicated input DNA. Final promoter methylation log2 ratios of bound over input signals represent the average of two or three independent experiments, including one dye swap. In each case, biological repeats showed high reproducibility ($R = 0.92$ for WI38 repeats, $R = 0.95$ and $R = 0.91$ for sperm repeats, **Supplementary Fig. 2**). For the sperm versus fibroblast comparison, we scaled individual arrays to have the same median absolute deviation using the LIMMA package in R. We defined promoters that gain methylation in fibroblasts as follows: log2 ratio in sperm <0.4 and Δlog2 ratio (fibroblast versus sperm) >0.25. Standard PCR on single genes were performed on 25 ng of input DNA and one-thirtieth of the immunoprecipitated DNA. Primer sequences are given in **Supplementary Table 2** online.

ChIP-on-chip. Six 10-cm dishes of WI38 fibroblasts grown to confluence were cross-linked in medium containing 1% formaldehyde for 10 min at room temperature, scraped off and rinsed with 10 ml 1× PBS. Pellets were resuspended in 15 ml buffer 1 (10 mM Tris (pH 8.0), 10 mM EDTA, 0.5 mM EGTA, 0.25% Triton X-100) and twice in 15 ml buffer 2 (10 mM Tris (pH 8.0), 1 mM EDTA, 0.5 mM EGTA, 200 mM NaCl). Then cells were lysed in 1 ml lysis buffer (50 mM HEPES/KOH (pH 7.5), 500 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% DOC, 0.1% SDS, protease inhibitors) and sonicated three times for 15 s (using a Branson sonicator, amplitude 70%). For the immunoprecipitation, we incubated 70 µg of chromatin overnight at 4 °C with 10 µl N-20 antibody to RNA polymerase II (Santa Cruz Biotechnology #sc-899) or 5 µl antibody to dimethylated H3K4 (Upstate #07030) and then incubated the mixture for 3 h at 4 °C with 30 µl protein A-Sepharose beads preblocked with tRNA. Beads were washed twice with 1 ml lysis buffer and once with 1 ml DOC buffer (10 mM Tris (pH 8.0), 0.25 M LiCl, 0.5% NP-40, 0.5% deoxycholate, 1 mM EDTA), and bound chromatin was eluted in 1% SDS/0.1 M $NaHCO_3$. After RNase A treatment, cross-linking was reversed by overnight incubation at 65 °C followed by proteinase K digestion. DNA was isolated by phenol/chloroform extraction followed by ethanol precipitation and resuspension in 50 µl TE. A sample of the input chromatin was treated in the same way to generate total input DNA. For the microarray analysis, we amplified 20 ng of input DNA and 40 µl ChIP DNA by ligation-mediated PCR (LMPCR) as described[48]. A set of ten genes was tested by quantitative PCR and showed similar bound-to-input ratios before and after amplification. Promoter log2 ratios are the average of three independent experiments, including one dye swap, that showed high reproducibility ($R = 0.97$ and $R = 0.95$ for RNA polymerase II repeats; $R = 0.98$ and $R = 0.99$ for dimethylated H3K4 repeats, **Supplementary Fig. 2**).

Bisulfite sequencing. Bisulfite genomic sequencing was performed as previously described[22]. Primer sequences are given in **Supplementary Table 2**.

Divergence estimates. From the UCSC genome browser, we downloaded the reciprocal best chain alignments of the chimpanzee genome (build PanTro1) and the rhesus macaque genome (build RheMac2) with the human genome (build hg17). Positions of the first and the last NimbleGen oligonucleotides on the human build hg17 plus 200 bp on each side were used as landmarks for the assayed promoters. The human sequence was kept ungapped so that we could easily create a three-way alignment. We considered only positions for which the Arachne base quality values were >20. To estimate divergence, we used the AMBIORE package[28], which is an implementation of a Bayesian Markov chain Monte Carlo allowing for context-dependent and nonreversible mutation rates. The initializing estimates were obtained from the concatenated sequences of the three species. We specified seven types of mutations with respect to their impact on CpG content: (i) non-CpG transitions, (ii) non-CpG transversions, (iii) CpG-loss transitions, (iv) CpG-loss transversions, (v) CpG-gain transitions, (vi) CpG-gain transversions and (vii) CCG→G = CGG→C. According to the recommendations for rather short sequences, we sampled 1,000 estimates after the burn-in phase. The median and the 95% confidence intervals of the 1,000 samples were determined. The 10% of the samples with the most extreme confidence intervals were removed from their respective mutation categories.

Accession codes. Microarray data are accessible from the Gene Expression Omnibus (GSE6715).

URLs. Processed data can be downloaded from our project website (http://www.fmi.ch/members/dirk.schubeler/supplemental.htm). The R package can be found at http://www.r-project.org. GNF SymAtlas can be found at http://symatlas.gnf.org. The Babelomics FatiGO tool can be found at http://fatigo.bioinfo.cipf.es.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank members of the Schübeler laboratory for advice during the course of the project and comments on the manuscript; E. Oakeley for generating scripts for data reformatting, A. Peters, M. Lorincz, C. Alvarez, P. de Boer, E. Selker and M. Groudine for critical reading of the manuscript. Primary samples from kidney and colon were obtained from M. Haase (Dresden University of Technology). Work in the laboratory of D.S. is supported by the Novartis Research Foundation, the EU 6th framework program NOE 'The Epigenome' (LSHG-CT-2004-503433) and a European Molecular Biology Organization (EMBO) Young Investigator Award. I.H. is supported by an EMBO long-term fellowship (ALTF 1160-2005).

AUTHOR CONTRIBUTIONS

M.W. designed and performed experiments and analysis and wrote the manuscript. D.S. designed the study and wrote the manuscript. M.B.S. performed housekeeping annotations and wrote custom software. M.R. performed CpG classifications and promoter confidence analysis, retrieved genomic information and contributed to the writing of the manuscript. I.H. and S.P. performed divergence analysis and contributed to the writing of the manuscript. L.R. provided purified human samples.

COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Li, E., Bestor, T.H. & Jaenisch, R. Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* **69**, 915–926 (1992).
2. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).
3. Heard, E., Clerc, P. & Avner, P. X-chromosome inactivation in mammals. *Annu. Rev. Genet.* **31**, 571–610 (1997).
4. Egger, G., Liang, G., Aparicio, A. & Jones, P.A. Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429**, 457–463 (2004).
5. Ioshikhes, I.P. & Zhang, M.Q. Large-scale human promoter mapping using CpG islands. *Nat. Genet.* **26**, 61–63 (2000).
6. Klose, R.J. & Bird, A.P. Genomic DNA methylation: the mark and its mediators. *Trends Biochem. Sci.* **31**, 89–97 (2006).
7. Boyes, J. & Bird, A. Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J.* **11**, 327–333 (1992).
8. Hsieh, C.L. Dependence of transcriptional repression on CpG methylation density. *Mol. Cell. Biol.* **14**, 5487–5494 (1994).
9. Brandeis, M., Ariel, M. & Cedar, H. Dynamics of DNA methylation during development. *Bioessays* **15**, 709–713 (1993).
10. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
11. Futscher, B.W. *et al.* Role for DNA methylation in the control of cell type specific maspin expression. *Nat. Genet.* **31**, 175–179 (2002).
12. Song, F. *et al.* Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl. Acad. Sci. USA* **102**, 3336–3341 (2005).
13. Walsh, C.P. & Bestor, T.H. Cytosine methylation and mammalian development. *Genes Dev.* **13**, 26–34 (1999).
14. Warnecke, P.M. & Clark, S.J. DNA methylation profile of the mouse skeletal alpha-actin promoter during development and differentiation. *Mol. Cell. Biol.* **19**, 164–172 (1999).
15. Smiraglia, D.J. *et al.* Excessive CpG island hypermethylation in cancer cell lines versus primary human malignancies. *Hum. Mol. Genet.* **10**, 1413–1419 (2001).
16. Coulondre, C., Miller, J.H., Farabaugh, P.J. & Gilbert, W. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature* **274**, 775–780 (1978).
17. Shen, J.C., Rideout, W.M., III. & Jones, P.A. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic Acids Res.* **22**, 972–976 (1994).
18. Hendrich, B., Hardeland, U., Ng, H.H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401**, 301–304 (1999).
19. Neddermann, P. & Jiricny, J. The purification of a mismatch-specific thymine-DNA glycosylase from HeLa cells. *J. Biol. Chem.* **268**, 21218–21224 (1993).
20. Rollins, R.A. *et al.* Large-scale structure of genomic methylation patterns. *Genome Res.* **16**, 157–163 (2006).
21. Saxonov, S., Berg, P. & Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. USA* **103**, 1412–1417 (2006).
22. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.* **37**, 853–862 (2005).
23. Takai, D. & Jones, P.A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA* **99**, 3740–3745 (2002).
24. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
25. Kim, T.H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).
26. Peters, A.H. & Schubeler, D. Methylation of histones: playing memory with DNA. *Curr. Opin. Cell Biol.* **17**, 230–238 (2005).
27. Schubeler, D. *et al.* The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev.* **18**, 1263–1271 (2004).
28. Hwang, D.G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl. Acad. Sci. USA* **101**, 13994–14001 (2004).
29. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
30. Assou, S. *et al.* The human cumulus–oocyte complex gene-expression profile. *Hum. Reprod.* **21**, 1705–1719 (2006).
31. Koslowski, M. *et al.* Frequent nonrandom activation of germ-line genes in human cancer. *Cancer Res.* **64**, 5988–5993 (2004).
32. Choi, Y.C. & Chae, C.B. DNA hypomethylation and germ cell-specific expression of testis-specific H2B histone gene. *J. Biol. Chem.* **266**, 20504–20511 (1991).
33. Singal, R. *et al.* Testis-specific histone H1t gene is hypermethylated in nongerminal cells in the mouse. *Biol. Reprod.* **63**, 1237–1244 (2000).
34. Schubeler, D. *et al.* Genomic targeting of methylated DNA: influence of methylation on transcription, replication, chromatin structure, and histone acetylation. *Mol. Cell. Biol.* **20**, 9103–9112 (2000).
35. Eckhardt, F. *et al.* DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat. Genet.* **38**, 1378–1385 (2006).
36. Bock, C. *et al.* CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet* **2**, e26 (2006).
37. Ayton, P.M., Chen, E.H. & Cleary, M.L. Binding to nonmethylated CpG DNA is essential for target recognition, transactivation, and myeloid transformation by an MLL oncoprotein. *Mol. Cell. Biol.* **24**, 10470–10478 (2004).
38. Lee, J.H. & Skalnik, D.G. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J. Biol. Chem.* **280**, 41725–41731 (2005).
39. Roh, T.Y., Cuddapah, S. & Zhao, K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev.* **19**, 542–552 (2005).

40. Bird, A.P. Gene number, noise reduction and biological complexity. *Trends Genet.* **11**, 94–100 (1995).
41. Maatouk, D.M. *et al.* DNA methylation is a primary mechanism for silencing post-migratory primordial germ cell genes in both germ cell and somatic cell lineages. *Development* **133**, 3411–3418 (2006).
42. Pohlers, M. *et al.* A role for E2F6 in the restriction of male-germ-cell-specific gene expression. *Curr. Biol.* **15**, 1051–1057 (2005).
43. De Smet, C., Loriot, A. & Boon, T. Promoter-dependent mechanism leading to selective hypomethylation within the 5′ region of gene MAGE-A1 in tumor cells. *Mol. Cell. Biol.* **24**, 4781–4790 (2004).
44. Davuluri, R.V., Grosse, I. & Zhang, M.Q. Computational identification of promoters and first exons in the human genome. *Nat. Genet.* **29**, 412–417 (2001).
45. Carrel, L. & Willard, H.F. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**, 400–404 (2005).
46. Eisenberg, E. & Levanon, E.Y. Human housekeeping genes are compact. *Trends Genet.* **19**, 362–365 (2003).
47. Simpson, A.J., Caballero, O.L., Jungbluth, A., Chen, Y.T. & Old, L.J. Cancer/testis antigens, gametogenesis and cancer. *Nat. Rev. Cancer* **5**, 615–625 (2005).
48. Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proc. Natl. Acad. Sci. USA* **100**, 8164–8169 (2003).
49. Riesewijk, A.M. *et al.* Monoallelic expression of human PEG1/MEST is paralleled by parent-specific methylation in fetuses. *Genomics* **42**, 236–244 (1997).
50. Muller, F. & Tora, L. The multicoloured world of promoter recognition complexes. *EMBO J.* **23**, 2–8 (2004).