A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters

Article *in* Proceedings of the National Academy of Sciences · February 2006

DOI: 10.1073/pnas.0510310103 · Source: PubMed

A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters

Serge Saxonov*†, Paul Berg†‡, and Douglas L. Brutlag*†§

*BioMedical Informatics Program and †Department of Biochemistry, Stanford University, Stanford, CA 94305

Contributed by Paul Berg, December 2, 2005

A striking feature of the human genome is the dearth of CpG dinucleotides (CpGs) interrupted occasionally by CpG islands (CGIs), regions with relatively high content of the dinucleotide. CGIs are generally associated with promoters; genes, whose promoters are especially rich in CpG sequences, tend to be expressed in most tissues. However, all working definitions of what constitutes a CGI rely on ad hoc thresholds. Here we adopt a direct and comprehensive survey to identify the locations of all CpGs in the human genome and find that promoters segregate naturally into two classes by CpG content. Seventy-two percent of promoters belong to the class with high CpG content (HCG), and 28% are in the class whose CpG content is characteristic of the overall genome (low CpG content). The enrichment of CpGs in the HCG class is symmetric and peaks around the core promoter. The broad-based expression of the HCG promoters is not a consequence of a correlation with CpG content because within the HCG class the breadth of expression is independent of the CpG content. The overall depletion of CpGs throughout the genome is thought to be a consequence of the methylation of some germ-line CpGs and their susceptibility to mutation. A comparison of the frequencies of inferred deamination mutations at CpG and GpC dinucleotides in the two classes of promoters using SNPs in human–chimpanzee sequence alignments shows that CpGs mutate at a lower frequency in the HCG promoters, suggesting that CpGs in the HCG class are hypomethylated in the germ line.

CpG islands | DNA methylation | epigenetics | gene expression

In vertebrates, the postreplication addition of methyl groups to the 5-position of cytosine in certain CpG dinucleotides and the maintenance of a particular genomic pattern of methylated CpGs provides an epigenetic means for differential regulation of gene expression (1–7). Indeed, the pattern of methylation often varies between cell types and different conditions, changes throughout development, and is abnormal in many disease states (5–10). A prevalent view holds that the state of CpG methylation regulates and stabilizes chromatin structure, perhaps regulating accessibility of the transcription machinery to regions of DNA (6, 9–11). Thus, whereas methylated CpGs restrict transcription, unmethylated CpGs in the vicinity of a gene allow that gene to be expressed.

The abundance of CpG dinucleotides in human DNA is much lower than expected based on the GC content (12–14), which results from the inherent mutability of methylated cytosine. Whereas the product of cytosine deamination, uracil, is readily recognized as aberrant and is repaired (4, 12, 15), the deamination product of methylated cytosine is thymine, leading to transition mutations in the next round of replication. Consequently, methylated CpGs in the germ line are likely to be lost over time (16–19). The resulting dearth of methylated CpGs is not uniform; typically, regions several hundreds of base pairs long contain an elevated number of CpGs and are referred to as CpG islands (CGIs) (13, 14, 20). Ostensibly, CGIs are retained because their CpGs are hypomethylated in the germ line, but some can arise through circumstances unrelated to methylation, such as strong selection or as a result of the prevalence of CpGs in some repeats (2, 21, 22).

Because no objective standard exists for defining a CGI, the prevailing approach is to rely on ad hoc thresholds of length, CpG fraction, and GC content (20, 22, 23). Despite the absence of a satisfactory definition, CGIs have been intensively studied. On the experimental front, CGIs have conventionally been targets for interrogation when probing the methylation status of the genome (24–28). Computationally, it has been observed that CGIs are imperfectly associated with promoters, leading to their use in promoter prediction (29, 30). Based on the threshold-based definitions, promoters with higher levels of CpGs are presumed to be associated with widely expressed genes. However, any study that attempts to analyze CGI-related properties of promoters is faced with the dual difficulty of defining what constitutes a CGI and what constitutes a CGI–promoter association.

As a prelude to determining the genome-wide pattern of CpG methylation, we have surveyed the pattern of CpGs over the human genome (31) and have calculated the prevalence of CpGs with respect to various gene-related features as annotated by the RefSeq database (32). By foregoing the use of threshold-based definitions of CGIs, we were able to uncover the existence and catalog the membership of two classes of promoters based on their CpG content: 72% of promoters with high CpG concentrations (HCG) and 28% of promoters whose CpG content was characteristic of the overall genome [low CpG concentration (LCG)]. By cataloging the promoters of the two classes, we were also able to analyze the differences in CpG distributions, mutation rates, and expression profiles.

Results

Although CpGs occur ≈25% as often over the whole human genome as would be expected based on the GC content, their presence is elevated relative to this background level in exons and upstream regions of genes (Table 1). At any given distance from the transcription start site (TSS), exons are similarly enriched for CpGs compared to introns. We infer that the retention and enrichment of CpGs in exons stems from coding constraints, which strongly limit the range of acceptable mutations, because noncoding exons closely resemble introns in their CpG content (Fig. 1*A*). Furthermore, our analysis of the CpG occurrence with respect to the coding frame is consistent with

Conflict of interest statement: No conflicts declared.

Freely available online through the PNAS open access option.

Abbreviations: CGI, CpG island; TSS, transcription start site; LCG, low CpG concentration; HCG, high CpG concentration.

‡To whom correspondence may be addressed at: Department of Biochemistry, Beckman Center B400, MC 5307, Stanford University, Stanford, CA 94304-5307. E-mail: pberg@cmgm.stanford.edu.

§To whom correspondence may be addressed at: Department of Biochemistry, Beckman Center B403, 279 Campus Drive, MC 5307, Palo Alto, CA 94305-5307. E-mail: brutlag@stanford.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0510310103

Table 1. Overview of CpG distribution in the human genome

Subset	Length, Mb	GC content	Observed CpG fraction	Normalized CpG fraction
Whole genome	3.1*	0.38	0.009	0.25
1 kb upstream regions	15	0.53	0.042	0.60
1 kb downstream regions	15	0.45	0.013	0.26
Transcription units	930	0.42	0.011	0.26
Exons	45	0.50	0.028	0.45
Introns	880	0.41	0.010	0.24

Length refers to the total length of DNA examined.
*Length given in gigabases.

this claim (Table 4, which is published as supporting information on the PNAS web site). In addition to their prevalence in exons, CpGs are also relatively enriched around the TSS. In fact, the enrichment pattern peaks sharply close to the core promoter 15 bp upstream of the TSS and extends symmetrically to ≈2 kb from the TSS (Fig. 1*B*). Within individual promoters, CpGs tend to come in clusters (data not shown), implying that the enrichment pattern reflects an average across many CpG islands, which tend to appear close to the core promoter and show no preference for being upstream or downstream.

Two Promoter Classes. Considering only the average pattern of CpG occurrence around the TSS conceals the existence of two distinct promoter classes. The distribution of promoters' normalized CpG content is bimodal and can be approximated by a mixture of two Gaussian curves with means of 0.23 and 0.61 normalized CpG content and relative abundances of 28% and 72%, respectively (Fig. 2*A*). It is unlikely that the bimodality can be explained by AT-rich and GC-rich isochores, because the distribution of GC content is distinctly unimodal (Fig. 2*B*). Taking the intersection of the Gaussian curves as a decision boundary, we assign a promoter to class LCG if the normalized CpG content of the 3 kb centered at the TSS is <0.35, and we assign a promoter to class HCG otherwise. This partitioning allocates 4,575 promoters (along with the corresponding genes) to the LCG class and 11,305 promoters to the HCG class, although there is minor cross-contamination because of the overlap between the curves. Reexamining the pattern of CpG occurrence around the TSS, there is a striking difference between the two classes. Whereas HCG promoters exhibit a prominent peak in the frequency of CpG centered some 15 bp upstream of the TSS, the CpG frequency for LCG promoters is relatively flat except for a small increase near the TSS (Fig. 2 *C* and *D*, lower curves). The most straightforward explanation for this qualitative difference between the classes is that all of the HCG promoters contain CGIs, and all of the LCG promoters lack them.

Estimation of CpG Mutation Rates. As previously discussed, elevated levels of CpGs can be due to the presence of CpG-rich repeats, general selection pressure, or methylation-related CpG-specific effects. To investigate the proximate cause of the difference in CpG content between the two classes, we analyzed mutation frequencies by using SNPs in human–chimpanzee sequence alignments. SNPs represent sites of recent mutations in the human genome, and the aligned chimpanzee sequence can be used to infer which alleles are ancestral (33). To distinguish the effects of methylation from the effects of selection, we examined the frequencies of deamination mutations at the CpG dinucleotides (CpG to TpG or CpA) and those at the GpC dinucleotides (GpC to GpT or ApC). Although negative selection should act indiscriminately on the two dinucleotides, changes related to methylation should only affect mutation frequencies at the CpGs. The last two rows of Table 2 show that CpGs mutate at a lower frequency in the HCG promoters than they do in the LCG promoters, whereas mutation frequencies of GpCs differ only modestly.

Unfortunately, this finding is not sufficient to establish the existence of a CpG-specific effect, because it can, in principle, be explained by a difference in general selection. One would expect that mutation rates of CpGs would be more strongly affected than those of GpCs, because many CpGs have been purged from the genome, making it more likely that the remaining ones are under stronger selection. Therefore, when examining regions conserved by evolution, the frequency of CpG mutations would be expected to be dampened to a higher extent than for GpC mutations. Consequently, in addition to examining the promoter regions of the two classes, we also examined the mutation patterns in regions downstream of the transcription start sites. Because methylation is unlikely to be a factor in sequences that are distant from the TSS, any differences in mutation frequen-

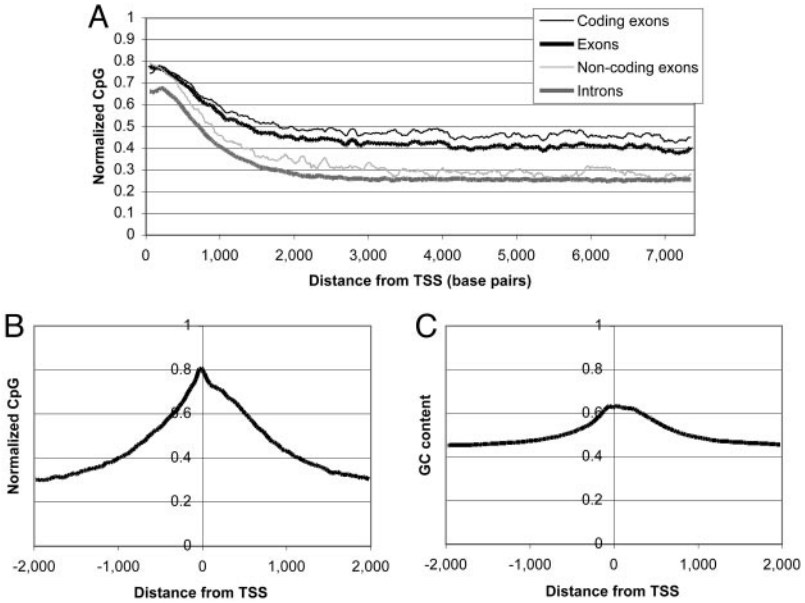


Fig. 1. Patterns of CpG occurrence with respect to gene features. The measures were made on overlapping segments aligned with respect to the TSS and identified by the distance of the midpoint from the TSS. The analysis included all (15,880) RefSeq genes for which the TSS was annotated differently from the start of the coding region. (*A*) To compare CpG presence in exons and introns as well as coding and noncoding sequences, the normalized CpG fraction was computed on overlapping 99-bp segments downstream of the TSS. Sequences were filtered according to whether they were in introns or exons; exons were further split into coding and noncoding (3′ and 5′ UTRs) sets. Exons carry a consistently higher level of CpGs than introns; the difference between the coding and noncoding exonic sequence shows that the CpG content of noncoding exons is only slightly above that of introns, suggesting the culpability of the coding potential in maintaining the higher CpG levels in exons. (*B* and *C*) Patterns of CpG occurrence (*B*) and GC content (*C*) around transcription start sites. Normalized CpG fraction and GC content were computed in 50-bp overlapping segments across 4-kb regions centered at the TSS.

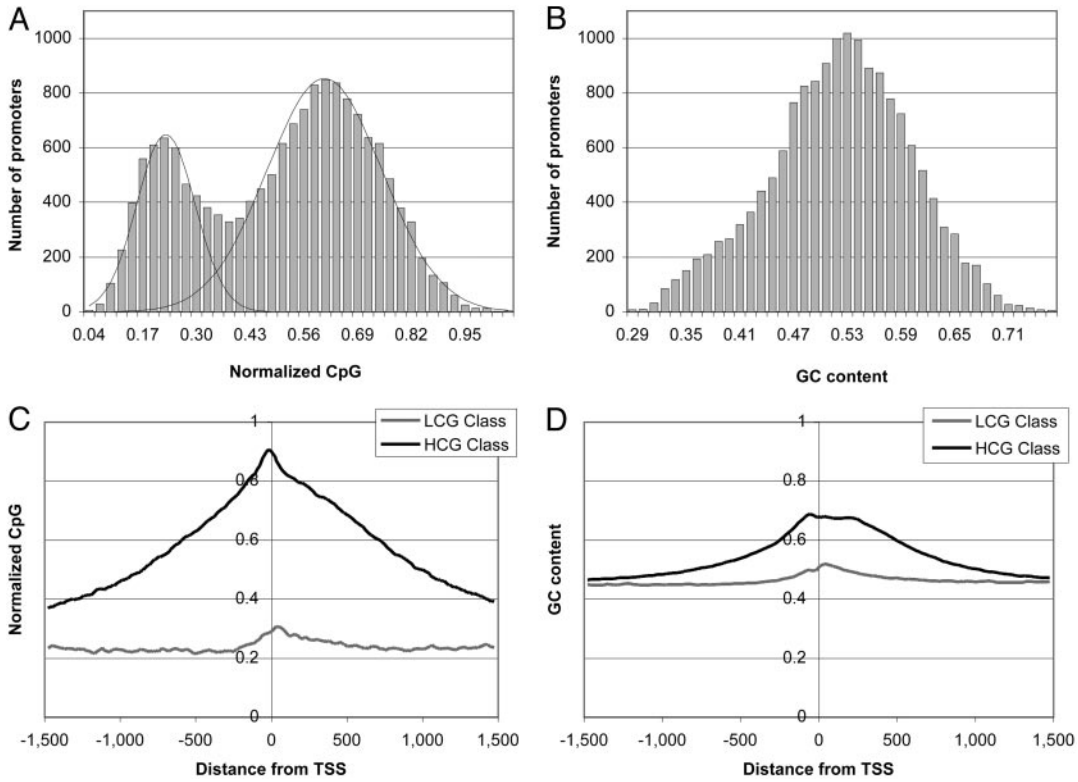


Fig. 2. Distribution of promoters with respect to CpG properties. (*A* and *B*) Histograms of normalized CpG fractions (*A*) and GC content (*B*) of 3-kb regions around TSSs. The *y* axis counts the number of promoters with the given CpG or GC content in the 3 kb centered at each promoter's TSS. Two Gaussian curves were fitted to the distribution in *A* with means of 0.23 and 0.61, σ values of 0.07 and 0.14, and weights of 4,430 and 11,450, respectively. The intersection of the two curves, at 0.35, is the decision boundary we used to separate promoters and their genes into classes LCG and HCG. See Table 6, which is published as supporting information on the PNAS web site, for a full listing of the TSSs in the two classes, along with their RefSeq IDs and chromosome locations. (*C* and *D*) Plotting the normalized CpG fraction (*C*) and GC content (*D*) separately for the two classes.

cies in such sequences should be due to differences in selection pressure. For the downstream analysis, we examined mutations in introns and the three coding phases of exons (phase 0, phase 1, and phase 2 refer to mutations that are in the first, second, and third positions of a codon, respectively). As expected, frequencies of mutations varied in accordance with the amount of selection on the sequences being considered. For both CpGs and GpCs, mutations were more prevalent in introns and in phase 2 (wobble) exonic positions, compared with phase 0 and 1 exonic positions (Table 2).

Observations of mutation frequencies in downstream introns and exons provide a basis from which to reexamine the differences between the LCG and HCG classes. The frequency of GpC mutations, which we can view as an inverse indicator of general selection, is only slightly higher in the LCG promoters compared with the HCG promoters, whereas for both classes it is close to the corresponding frequency in introns and at wobble positions. Most importantly, the HCG class appears to be an outlier because the frequency of CpG mutations is the lowest of any of the regions examined and the GpC mutation frequency is consistent with HCG promoters being under only very modest selection. Taken together, the evidence argues for a CpG-

Table 2. Frequencies of deamination mutations at CpG and GpC dinucleotides in exons, introns, and promoters

Gene regions	GpC→GpT mutation frequency*	CpG→TpG mutation frequency*	Ratio (CpG frequency/GpC frequency)
Downstream exons, phase 0	0.42 ± 0.06	2.30 ± 0.04	5.5
Downstream exons, phase 1	0.39 ± 0.06	2.78 ± 0.04	7.2
Downstream exons, phase 2	0.72 ± 0.04	7.73 ± 0.02	10.8
Downstream introns	0.75 ± 0.00	8.31 ± 0.00	11.1
LCG promoters†	0.75 ± 0.03	7.31 ± 0.02	9.8
HCG promoters†	0.64 ± 0.02	1.62 ± 0.01	2.5

Downstream refers to all the sequences > 3 kb downstream of the TSS. Recent mutations in the human lineage were identified by compiling human SNPs that fell within the examined regions. For every SNP we determined which allele was ancestral by identifying the aligned base in the chimpanzee genome.
*For mutations XpY → X′pY′, mutation rate is presented as 1,000·(XpY → X′pY′ mutations/XpY dinucleotides).
†3-kb sequences centered at the TSS.

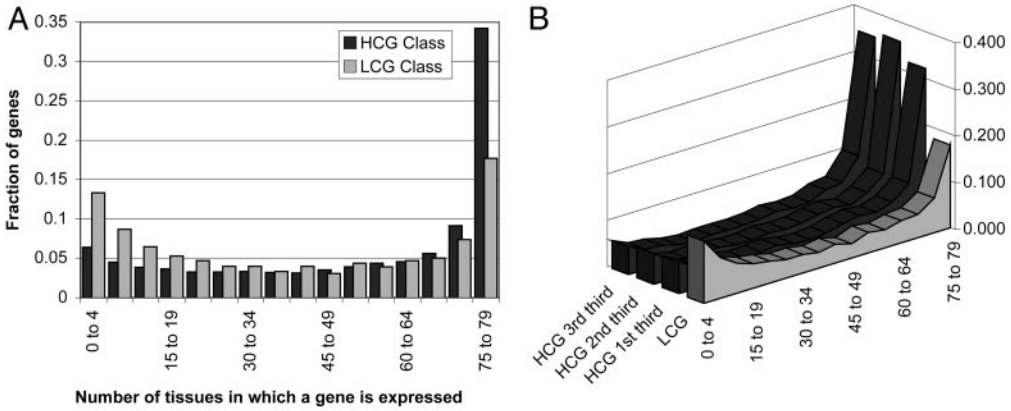
Table 3. Distributions of top-level GO terms for the LCG and the HCG classes

GO code	GO term description	Appearances LCG	Appearances HCG	*P* value
Overrepresented in class LCG				
09607	[BP]response to biotic stimulus	307	192	7.9×10^{-52}
09605	[BP]response to external stimulus	296	218	5.8×10^{-41}
07582	[BP]physiological process	603	789	1.8×10^{-26}
05615	[CC]extracellular space	116	88	1.1×10^{-15}
06950	[BP]response to stress	227	268	2.5×10^{-13}
09628	[BP]response to abiotic stimulus	108	97	2.1×10^{-11}
05886	[CC]plasma membrane	429	656	1.0×10^{-10}
05102	[MF]receptor binding	128	146	2.0×10^{-08}
30246	[MF]carbohydrate binding	30	19	1.5×10^{-05}
07267	[BP]cell-cell signaling	136	196	1.1×10^{-04}
05576	[CC]extracellular region	39	36	2.2×10^{-04}
04872	[MF]receptor activity	182	288	3.1×10^{-04}
19825	[MF]oxygen binding	14	6	5.6×10^{-04}
05623	[CC]cell	512	965	7.9×10^{-04}
08233	[MF]peptidase activity	62	76	8.7×10^{-04}
07154	[BP]cell communication	75	103	2.6×10^{-03}
07165	[BP]signal transduction	429	810	3.0×10^{-03}
05578	[CC]extracellular matrix	44	52	4.0×10^{-03}
Overrepresented in class HCG				
05634	[CC]nucleus	85	535	1.1×10^{-18}
06139	[BP]nucleo-metabolism	78	458	1.2×10^{-14}
07049	[BP]cell cycle	37	294	2.6×10^{-13}
06350	[BP]transcription	71	401	3.3×10^{-12}
06259	[BP]DNA metabolism	22	193	1.1×10^{-09}
05739	[CC]mitochondrion	16	168	1.3×10^{-09}
05575	[CC]cellular_component	74	367	3.9×10^{-09}
03723	[MF]RNA binding	22	180	1.3×10^{-08}
30528	[MF]transcription regulator activit	52	286	1.5×10^{-08}
05622	[CC]intracellular	16	140	3.4×10^{-07}
09719	[BP]response to endogenous stin	8	96	3.3×10^{-06}
05654	[CC]nucleoplasm	12	103	2.1×10^{-05}
03700	[MF]transcription factor activity	52	236	3.3×10^{-05}
03677	[MF]DNA binding	22	129	1.4×10^{-04}
05840	[CC]ribosome	1	44	2.2×10^{-04}
15031	[BP]protein transport	10	82	2.6×10^{-04}
06464	[BP]protein modification	73	277	5.8×10^{-04}
05730	[CC]nucleolus	1	39	6.6×10^{-04}
05694	[CC]chromosome	5	56	8.6×10^{-04}
08152	[BP]metabolism	285	836	1.4×10^{-03}
04672	[MF]protein kinase activity	51	200	2.6×10^{-03}
06118	[BP]electron transport	0	25	4.4×10^{-03}
05783	[CC]endoplasmic reticulum	24	113	4.9×10^{-03}

All of the terms were mapped to the goslim_generic subset, which is meant to represent the top levels of the GO hierarchy. *P* values were calculated by using the χ^2 statistic. Only terms significant at the 0.005 level are presented. Parenthesized markings stand for the three major subontologies comprising GO: CC for ''cellular component,'' BP for ''biological process,'' and MF for ''molecular function.'' Results for the full ontology (not just the goslim_generic subset) can be found in Table 4.

specific effect and not general selection as the dominant culprit for the high levels of CpGs in HCG promoters.

Differences in Annotation and Expression Between the Two Classes. Evidence from other studies suggests that CGIs are more frequently associated with ''house-keeping'' genes than with tissue-specific genes (21, 34, 35). Our analysis of Gene Ontology (GO) (36) terms associated with genes in the HCG and LCG classes is consistent with that functional relationship (Table 3; see also Table 5, which is published as supporting information on the PNAS web site). Broadly considered, house-keeping functions are significantly overrepresented in the HCG class, whereas terms associated with specific functions characteristic of more differentiated or highly regulated cells are significantly overrepresented in the LCG class. The correlation of a promoter's CpG content with the breadth of expression of its gene is also borne out by our analysis of expression profiles of genes in the two classes (Fig. 3). Using the data set from Su *et al.* (37), who measured expression levels of an extensive set of genes in 79 different tissues, we bin genes according to the number of tissues in which they are expressed. The resulting distributions are significantly different between the two classes, the most pronounced differences being at the extremes of the distributions: therefore, genes that are expressed in only a small number of tissues are overrepresented in class LCG, and genes expressed in all or almost all of the tissues are biased toward the HCG class



C

	Class LCG	Class HCG, 1st third	Class HCG, 2nd third
Class HCG, 1st third	259 (4.5e-21)	-	-
Class HCG, 2nd third	360 (1.6e-37)	79(0.49)	-
Class HCG, 3rd third	337 (1.2e-33)	89(0.22)	86(0.27)

Fig. 3. A microarray analysis of tissue distribution of genes in class LCG and class HCG. (*A*) Tissue distributions of genes in the two classes were significantly different ($P = 1.6 \times 10^{-52}$). The fraction of genes expressed in only a few tissues was higher in the LCG class, whereas the fraction of universally expressed genes was higher in the HCG class. For plotting convenience we show distributions of genes grouped in 16 larger bins of size 5. (*B*) We partitioned class HCG into thirds by CpG content. One-third of promoters had normalized CpG fractions between 0.350 and 0.563, the next third was between 0.563 and 0.683, and the last third comprised all of the promoters with normalized CpG at >0.683. The tissue distributions of genes in the three HCG partitions were similar to each other and different from class LCG. (*C*) We quantified that conclusion by measuring dissimilarities between distributions by using χ^2 values (*P* values in parentheses).

(Fig. 3*A*). Significantly, genes within the HCG class, irrespective of whether they contain the least or the highest CpG content, exhibit very similar expression profiles (Fig. 3 *B* and *C*). The implication is that, within a class, the number of tissues in which a gene is expressed is not significantly dependent on the promoter's CpG content. This point is important because it shows that the universality of a gene's expression is specifically correlated with class membership and not directly with the CpG content.

Discussion

We should note that there have been previous studies comparing genes with or without CGIs in their 5′ regions (21, 35, 38). However, all such studies classified genes according to arbitrary and limiting definitions of CGIs, definitions based on thresholds of CpG fraction, GC content, and length. Few inferences could have been made about the underlying distribution of promoters, because applying any threshold would partition a set of promoters regardless of whether they cluster into cohesive subsets. Only one study approached classifying promoters based on CpG properties from an *ab initio* perspective. Davuluri, Grosse, and Zhang (30) found a bimodal distribution of a sliding window statistic in the vicinity of TSSs and used it to generate two separate models for first exon prediction. Our results are consistent with their findings, while bringing more clarity to the nature of promoter–CGI association and establishing that there is a biologically meaningful separation of genes based on their CGI properties. Before our work, a continuous gradation of CpG content could not be ruled out because the promoters that were deemed to lack CpG islands could have been at the tail of a distribution of CpG content. We show that there are, in fact, two classes of promoters with distinct CpG sequence profiles and a natural decision boundary. Furthermore, we find that CpG-rich promoters are expressed in more tissues but only to the extent that they are more likely to be in the HCG class.

Incidentally, it may appear surprising that the GC content around promoters forms a unimodal distribution (Fig. 2*B*), because it has been previously argued that CpG islands are preferentially located in the GC-rich isochores (21), and we have found that the normalized CpG content at the promoter is weakly correlated with the GC content (data not shown). Most likely, the GC content appears unimodal because, although different between the two classes, it varies to a much smaller extent than the CpG content.

Given the difference in CpG-specific mutation rates (Table 2), CGIs in the HCG promoters are almost certainly a consequence of their methylation state rather than of a general selection or the presence of CpG-rich transposable elements. As mentioned above, the most common explanation for such CGIs is that they are a consequence of hypomethylation in the germ line. The unmethylated CpGs in active promoters would be spared the mutagenic effect seen in methylated regions of the rest of the genome. According to this view, the pattern of CGIs in the genome should reflect a weighted average of methylation patterns in the germ line for which the weight is proportional to the time spent in the particular methylation state (1). The overrepresentation of widely expressed genes in the HCG class is consistent with the supposition that these promoters are hypomethylated in the germ line. Another possible explanation for the origin of CGIs is that they represent regions where natural selection has favored retention of CpGs for use in methylation-mediated regulation. This explanation would account for why some tissue-specific genes contain promoters that are highly enriched for CpGs.

If CGIs are manifestations of methylation patterns, studying the properties of CGIs may yield insights into mechanisms that govern the establishment of these patterns. For instance, any proposed model for such a mechanism must account for the symmetry of CGI distribution around the core promoter. Therefore, the prevailing hypothesis involving the binding of transcription factors, such as SP1, to inhibit methylation (39–41), is probably incomplete because it is unlikely to explain the equal clustering of CpGs upstream and downstream of the core promoter. More generally, identification of the two promoter classes lays the groundwork for characterization of CGI properties and analysis of sequence elements that influence and are influenced by CGI locations and boundaries. Orthologous sequences from other mammals should be very useful in this regard

as they can help to better separate the classes and to identify CGI boundaries more precisely.

The most striking finding of our analysis is the bimodal distribution of CpG content in promoters, which should caution against excessive reliance on CGIs as gene markers. The LCG class represents a substantial fraction of known genes and is likely to be more prevalent among undiscovered genes (42–44). The discovery of the LCG class raises the question about the role of methylation in controlling the expression of LCG genes. At present, we have a paucity of experimental data because most studies of differential methylation focus on CGIs, which are absent in the LCG class. In the end, it is the state of methylation of CpGs in both HCG- and LCG-class promoters and in various physiological states that holds the key to understanding their role in molding the phenotype.

Methods

Sequence Analysis. All of the statistics were compiled for the University of California, Santa Cruz human genome assembly (hg16) from July 2003, and the corresponding gene annotations were from the National Center for Biotechnology Information RefSeq database. To determine whether false TSS predictions were skewing our results, we also analyzed annotations from cap analysis gene expression sites (RIKEN CAGE database), chromatin immunoprecipitation sites, and compiled 5′ UTR lengths. It does not appear that the essential conclusions of this work were compromised by false TSS predictions in the RefSeq database. Normalized CpG fraction was computed as (observed CpG)/(expected CpG), where expected CpG was calculated as $(\text{GC content}/2)^2$.

Analysis of Mutation Frequencies. We compiled a list of mutation locations in the human genome by relying on SNPs and inferred the ancestral alleles through comparisons with the chimpanzee genome. A compilation of human SNPs was downloaded from the National Center for Biotechnology Information and was mapped to the University of California, Santa Cruz human–chimpanzee alignments. We compiled statistics for mutations of the CpG dinucleotide to the TpG dinucleotide by collecting all of the {C, T} polymorphisms that were followed by a G and which aligned to a C in the chimpanzee genome. To account for the complementary strand, the CpG-to-CpA mutations were also included in all of the tallies. The statistics of mutations at the GpC dinucleotide were compiled in the analogous fashion. When measuring mutation rates, only nonoverlapping dinucleotides were examined (i.e., cytosines flanked by two guanines were not considered because their mutations could not be used to discriminate between mutations of GpC and CpG dinucleotides).

GO Analysis. GO terms were mapped to RefSeq genes using LocusLink annotations and RefSeq to LocusLink mappings downloaded from the National Center for Biotechnology Information web site. Only experimentally confirmed annotations were used (i.e., evidence codes IDE, IDA, IEP, IGI, IMP, IPI, ISI, and TAS).

Expression Analysis. The data were taken from an analysis of expression in 79 tissues by Su *et al.* (37); only genes (8, 272) with RefSeq identifiers were considered and each one was deemed to be expressed in a tissue if the average difference value was >200 (45). Consequently, each gene was assigned into one of 80 bins, depending on the number of tissues in which it was expressed (0–79). LCG was represented by 2,202 genes, and HCG was represented by 6,070 genes.

We thank S. Manteuil-Brutlag, B. Naughton, and I. Yeh for helpful comments on the manuscript. S.S was supported by a National Library of Medicine graduate fellowship.

1. Reik, W., Dean, W. & Walter, J. (2001) *Science* **293,** 1089–1093.
2. Fazzari, M. J. & Greally, J. M. (2004) *Nat. Rev. Genet.* **5,** 446–455.
3. Robertson, K. D. & Wolffe, A. P. (2000) *Nat. Rev. Genet.* **1,** 11–19.
4. Singal, R. & Ginder, G. D. (1999) *Blood* **93,** 4059–4070.
5. Bird, A. (2002) *Genes Dev.* **16,** 6–21.
6. Jaenisch, R. & Bird, A. (2003) *Nat. Genet.* **33,** Suppl., 245–254.
7. Novik, K. L., Nimmrich, I., Genc, B., Maier, S., Piepenbrock, C., Olek, A. & Beck, S. (2002) *Curr. Issues Mol. Biol.* **4,** 111–128.
8. Jones, P. A. & Takai, D. (2001) *Science* **293,** 1068–1070.
9. Geiman, T. M. & Robertson, K. D. (2002) *J. Cell Biochem.* **87,** 117–125.
10. Herman, J. G. & Baylin, S. B. (2003) *N. Engl. J. Med.* **349,** 2042–2054.
11. Fahrner, J. A., Eguchi, S., Herman, J. G. & Baylin, S. B. (2002) *Cancer Res.* **62,** 7213–7218.
12. Bird, A. P. (1980) *Nucleic Acids Res.* **8,** 1499–1504.
13. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al*. (2001) *Nature* **409,** 860–921.
14. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al*. (2001) *Science* **291,** 1304–1351.
15. Duncan, B. K. & Miller, J. H. (1980) *Nature* **287,** 560–561.
16. Arndt, P. F., Burge, C. B. & Hwa, T. (2003) *J. Comput. Biol.* **10,** 313–322.
17. Arndt, P. F. & Hwa, T. (2004) *Bioinformatics* **20,** 1482–1485.
18. Lunter, G. & Hein, J. (2004) *Bioinformatics* **20,** Suppl. 1, I216–I223.
19. Sved, J. & Bird, A. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4692–4696.
20. Gardiner-Garden, M. & Frommer, M. (1987) *J. Mol. Biol.* **196,** 261–282.
21. Ponger, L., Duret, L. & Mouchiroud, D. (2001) *Genome Res.* **11,** 1854–1860.
22. Takai, D. & Jones, P. A. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 3740–3745.
23. Ponger, L. & Mouchiroud, D. (2002) *Bioinformatics* **18,** 631–633.
24. Rakyan, V. K., Hildmann, T., Novik, K. L., Lewin, J., Tost, J., Cox, A. V., Andrews, T. D., Howe, K. L., Otto, T., Olek, A., *et al*. (2004) *PloS. Biol.* **2,** e405.
25. Yamada, Y., Watanabe, H., Miura, F., Soejima, H., Uchiyama, M., Iwasaka, T., Mukai, T., Sakaki, Y. & Ito, T. (2004) *Genome Res.* **14,** 247–266.
26. Huang, T., Perry, M. & Laux, D. (1999) *Hum. Mol. Genet.* **8,** 459–470.
27. Yan, P. S., Perry, M. R., Laux, D. E., Asare, A. L., Caldwell, C. W. & Huang, T. H.-M. (2000) *Clin. Cancer Res.* **6,** 1432–1438.
28. Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H.-M. & Farnham, P. J. (2002) *Genes Dev.* **16,** 235–244.
29. Ioshikhes, I. P. & Zhang, M. Q. (2000) *Nat. Genet.* **26,** 61–63.
30. Davuluri, R. V., Grosse, I. & Zhang, M. Q. (2001) *Nat. Genet.* **29,** 412–417.
31. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. & Haussler, D. (2002) *Genome Res.* **12,** 996–1006.
32. Pruitt, K. D. & Maglott, D. R. (2001) *Nucleic Acids Res.* **29,** 137–140.
33. Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T. D., Toyoda, A., Kuroki, Y., Noguchi, H., BenKahla, A., Lehrach, H., Sudbrak, R., *et al*. (2004) *Nature* 429, 382–388.
34. Larsen, F., Gundersen, G., Lopez, R. & Prydz, H. (1992) *Genomics* **13,** 1095–1107.
35. Robinson, P. N., Bohme, U., Lopez, R., Mundlos, S. & Nurnberg, P. (2004) *Hum. Mol. Genet.* **13,** 1969–1978.
36. Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., *et al*. (2004) *Nucleic Acids Res.* **32,** 258–261.
37. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., *et al*. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 6062–6067.
38. Holmquist, G. P. (1989) *J. Mol. Evol.* **28,** 469–486.
39. Bell, A. C. & Felsenfeld, G. (2000) *Nature* **405,** 482–485.
40. Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A. & Cedar, H. (1994) *Nature* **371,** 435–438.
41. Siegfried, Z., Eden, S., Mendelsohn, M., Feng, X., Tsuberi, B. Z. & Cedar, H. (1999) *Nat. Genet.* **22,** 203–206.
42. Kapranov, P., Cawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. & Gingeras, T. R. (2002) *Science* **296,** 916–919.
43. Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., *et al*. (2004) *Science* **306,** 2242–2246.
44. Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., *et al*. (2004) *Cell* **116,** 499–509.
45. Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., *et al*. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 4465–4470.