Corrections

**BIOCHEMISTRY**


The authors note that Fig. 6 appeared incorrectly. There was an error in the alignment of the molecular mass markers, and minor adjustments have been made to the assignment of caspase-12 bands. The authors also note that the source of the anti-caspase-12 antibody for Fig. 6 was Sigma (clone 14F7). This error does not affect the conclusions of the article. The corrected figure and its corresponding legend appear below.

**CHEMISTRY**


The authors note that they incorrectly assigned the structure of the reaction product reported in Scheme 4. The published structure represents the γ-aminated adduct 8, when it should instead be the α-analogue arising from an α-site selective pathway. As a result of this, Scheme 4 and its related comments should be removed from the article.

On page 20642, left column, within the Abstract, lines 16–18, “Finally, we describe the extension of the dienamine catalysis-induced vinylogous nucleophilicity to the asymmetric γ-amination of cyclohexene carbaldehyde” should be removed from the article.

On page 20645, right column, third full paragraph, lines 1–8, to page 20646, left column, first paragraph, lines 1–2, “Finally, to explore the potential of the chiral primary amine-induced vinylogous nucleophilicity, we wondered whether this unique reactivity concept may be translated to an aldehyde derivative adorned with a six-membered ring scaffold, reminiscent of the β-substituted cyclohexane framework. Although the vinylogous Michael addition of 1-cyclohexene-1-carboxaldehyde 7 to nitrostyrene 2a did not proceed at all, the combination with tert-butylazodicarboxylate under the catalysis of A furnished the γ-amination product 8 with perfect regio- and enantioselectivity (Scheme 4)” should be removed from the article.

These errors do not affect the conclusions of the article of the vinylogous Michael addition of cyclic enones to nitroalkenes. The ability of primary amine catalysis to address the synthetic issue connected with the enantioselective carbon–carbon bond formation gamma to a carbonyl group, promoting vinylogous nucleophilicity upon selective activation of unmodified cyclic unsaturated ketones, is fully supported by the separated results presented in Tables 1, 2, and 3, and Schemes 2 and 3.

The authors note the following: “The mating frequencies reported in Table 1 of this paper did not follow a multinomial distribution, making the statistical analysis inapplicable. This problem was obviated by considering only the first matings in each experimental unit and computing odds ratios. After submitting the paper, we continued to perform experiments identical in design to those we reported. In the table below, we combined the results of those additional replicate experiments with those already reported. From the new analysis, we now find that experiment 4, in which flies were infected with a mixture of Lactobacillus spp., assortative mating was not restored. Otherwise, the conclusions of the article were not changed by our reanalysis. We acknowledge the statistical advice of Dan Yekutieli and thank Tal Lahav for calculating the odds ratios and their 95% confidence intervals, and for performing the chi-squared tests presented in the corrected Table 1.”

The corrected Table 1 appears below.

Table 1. The role of bacteria in diet-induced mating preference of D. melanogaster

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Fib treatment*</th>
<th>N†</th>
<th>OR‡</th>
<th>95% CI</th>
<th>P value‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Starch-grown × CMY-grown</td>
<td>18</td>
<td>3.21</td>
<td>2.14–4.81</td>
<td>1.8 × 10⁻⁸</td>
</tr>
<tr>
<td>2</td>
<td>Experiment 1 after antibiotics</td>
<td>11</td>
<td>1.04</td>
<td>0.63–1.71</td>
<td>0.9888</td>
</tr>
<tr>
<td>3</td>
<td>Experiment 2 after infection of starch-grown flies with homologous bacteria§</td>
<td>6</td>
<td>2.68</td>
<td>1.40–5.11</td>
<td>0.0477</td>
</tr>
<tr>
<td>4</td>
<td>Experiment 3 with Lactobacillus spp. replacing homologous bacteria</td>
<td>4</td>
<td>1.76</td>
<td>0.74–4.19</td>
<td>0.2912</td>
</tr>
<tr>
<td>5</td>
<td>Experiment 3 with Lactobacillus plantarum replacing homologous bacteria</td>
<td>7</td>
<td>2.14</td>
<td>1.35–3.39</td>
<td>0.0019</td>
</tr>
<tr>
<td>6</td>
<td>Infection control (no added bacteria)</td>
<td>4</td>
<td>1.26</td>
<td>0.53–3.00</td>
<td>0.7712</td>
</tr>
</tbody>
</table>

*After all treatments, the flies were grown for one generation in CMY medium before performing the mating preference test.

†N is the number of replicate experiments.

‡Cochran-Mantel-Haenszel Odds Ratio and P value are from the Cochran-Mantel-Haenszel Chi-squared test (34, 35).

§Antibiotic-treated starch- and CMY-grown flies were infected with bacteria isolated from their respective growth medium (before antibiotic treatment).


MEDICAL SCIENCES


The authors note that the author name Vadim Kapulkin should instead appear as Wadim Jan Kapulkin. The corrected author line appears below. The online version has been corrected.

Virginia Fonte, Wadim Jan Kapulkin, Andrew Taft, Amy Fluet, David Friedman, and Christopher D. Link

www.pnas.org/cgi/doi/10.1073/pnas.1302545110

SYSTEMS BIOLOGY


The authors note that, within the corresponding author footnote on page 9715, the email address “colaneri@niehs.nih.gov” should instead appear as “acolaneri_2000@yahoo.com.ar”.

www.pnas.org/cgi/doi/10.1073/pnas.1302473110
Expanded methyl-sensitive cut counting reveals hypomethylation as an epigenetic state that highlights functional sequences of the genome

Alejandro Colaneri1,a, Nickolas Staffa2, David C. Fargo3, Yuan Gao4, Tianyuan Wang5, Shyamal D. Peddada6, and Lutz Birnbaumer1,a

1Laboratory of Neurobiology, 2Library and Information Services, and 3Biostatistics Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC 27709; and 4Division of Genomics, Epigenomics, and Bioinformatics, The Lieber Institute for Brain Development, and Neuroregeneration and Stem Cell Program, Institute for Cell Engineering, The Johns Hopkins University, Baltimore, MD 21205

Contributed by Lutz Birnbaumer, April 13, 2011 (sent for review March 3, 2011)

Methyl-sensitive cut counting (MSCC) with the Hpall methylation-sensitive restriction enzyme is a cost-effective method to pinpoint unmethylated CpGs at single base-pair resolution. However, it has the drawback of addressing only CpGs in the context of the CCGG site, leaving out the remainder of the possible 16 XCGX tetranucleotides in which CpGs are found. We expanded MSCC to include three additional enzymes to address a total of 5 of the 16 XCGX combinations. This allowed us to survey methylation at about one-third of all a mammalian genome’s CpGs. Applied to mouse liver DNA, we correctly confirmed data reported with other methods showing hypomethylation to be concentrated at promoters and in CGI island (CGIs) versus non-CGI island (nCGIs) integrative locations and the genome’s CpGs are mostly methylated. Grouping unmethylated CpGs, characterized by high MSCC scores (7% false discovery rate), we found a large number of unmethylated regions not qualifying as CGIs located in intergenic and intronic regions, which are highly enriched in functional DNA sequences (open regulatory annotation database) as well as in noncoding yet highly conserved mammalian sequences thought to be important but with as yet unknown function. About 50% of MSCC-defined unmethylated regions do not overlap algorithm-defined CGIs and offer a novel search space in which new functionalities of DNA may be found in health and disease.

methylation of DNA at position five of the cytosine ring is a widespread modification in the vertebrate genomes (1). A family of DNA methyltransferases whose primary targets are the cytosines located at CpG dinucleotides catalyzes this chemical modification (2). Many CpGs are not distributed at random, because a significant proportion of them have coalesced into what has been called CpG islands (CGIs), where they are mostly hypomethylated, whereas CpGs outside of CGIs are mostly methylated (1, 3–5). CGIs are identified with computer algorithms that search for shared distinctive properties; traditionally CpG and (G + C) richness (5–8). A different approach selects CGIs between clusters of CpGs whose maximum inter-CpG distances are below a threshold (e.g., median genomic inter-CpG distance) (9). The filtering criteria used by all these programs seek to optimize the possibility that the selected CGIs are not the product of chance but the result of evolutionary processes. The most widely accepted explanation for the origin of CGIs is based in the tendency of 5-methylcytosines to undergo spontaneous deamination to uracil producing C-to-T mutations. This process drives a nonselective purge of CpGs from the broadly methylated genomic sequences with no evolutionary constraints (10, 11) (SI Appendix, Tables S1, S2, and S3). However, this purge did not occur in regions rich in regulatory elements that have been protected from being methylated. According to this simplified and generally accepted hypothesis, the functionality of a CGI is measured by the probability of finding it unmethylated. For example, the program called CpGCluster uses a statistical criterion (P value) to select for clusters with low probability of having been formed by chance (9). This means that these loci have retained their CpG density during evolution, presumably because of their prevalence in an unmethylated state. CGIs, including CpGCluster CGIs with the lowest P values, are more frequently found overlapping promoters (12), which supports connections between evolutionary origin, unmethylated state, and functionality. The idea that active promoters protect their CpGs from being methylated is supported by site-specific mutagenesis experiments. For example, mutations that prevent the transcription factor Sp1 from recognizing and binding its target sequences in a particular CGI remove the protection of that CGI from DNA methylation (13).

Research efforts focused on improving the prediction of locations of CGIs aim to identify functionally relevant epigenetic loci in development and disease; as a consequence, CGIs still constitute the framework on which the majority of researchers base their high-throughput methylation analyses. However, the filtering criteria used by these programs frequently fail to identify a large percentage of subsequences having the potential to encode regulatory functions that can be disrupted or activated by changes in methylation. Inspection of the mammalian genome shows it to be divided into two classes of subsequences. In one class (85% of the genome), CpGs are sparse (one every 250 bp). The other class (15% of the genome) concentrates half of the total genomic CpGs at an average inter-CpG distance of 40 bp. At this density, methylation has been shown to have a deleterious effect on the functionality of the DNA elements (14). These relative CpG-rich subsequences accommodate the totality of the CGIs, regardless of the algorithms used to define them, and overlap with 95% of the RefSeq-defined transcription start site (TSS) regions (~3 Kb to +2 Kb from the TSS).

Our goal was to design a method that allows us to identify targets of methylation-mediated epigenetic processes throughout the genome without having to select a priori candidate subsequences. We developed a high-throughput sequencing-based DNA methylation analysis, which consists of an expanded (more comprehensive) version of the methyl-sensitive cut counting assay (MSCC) (15, 16). Applied to mouse, our method identifies the methylation status of 6 million CpGs (one-third of all existing CpGs) and covers 58% of the CpG-rich subsequences.

We found that a surprising proportion (50%) of our unmethylated regions (UMRs) do not meet the traditional CGI cri-

Author contributions: A.C. designed research; A.C. performed research; N.S., D.C.F., Y.G., S.D.P., and A.C. and L.B. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

To whom correspondence may be addressed. E-mail: colaneria@niehs.nih.gov or birnbau1@niehs.nih.gov.

This article contains supporting information online at www.pnas.orglookup/suppmaterials/pnas.1105713108.

PNAS June 7, 2011 vol. 108 no. 23 9715–9720

www.pnas.org/cgi/doi/10.1073/pnas.1105713108

PNAS | June 7, 2011 | vol. 108 | no. 23 | 9715–9720

PNAS | June 7, 2011 | vol. 108 | no. 23 | 9715–9720
teria. Interestingly, most of these non-CGI UMRs are located in noncoding DNA outside of promoters and are more enriched in experimentally determined regulatory sequences than CGI-like UMRs (UMRs that overlap predicted CGIs). At least 10% of the UMRs identified through MSCC contain literature-published liver-specific and liver-related regulatory sequences.

Results and Discussion

Coverage. CpG tag libraries prepared from DNA digested with four methylation-sensitive restriction enzymes (4Enz; SI Appendix, SI Materials and Methods) were sequenced by Illumina Genome Analyzer reporting the sequences returned by individual sequencing channels of their flow cells. The data analyzed here were obtained from sequencing three CpG tag libraries, yielding three datasets that we called Slax1, Slax2, and Slax3 (SI Appendix, Table S4). The majority of the data analyzed below is from the Slax3 dataset returned to us from five sequencing channels, which, when pooled, rendered 29 million reads that mapped to unique nonrandom CpGs (SI Appendix, Tables S4 and S5).

The number of reads per identified CpG (MSCC score, see below) was found to vary from site to site because of disparities in the level of methylation, but a minimum coverage is required to be able to perceive these differences. We used the reads recovered from lambda phage (lambda) CpGs to gauge the coverage and followed the behavior of the 1,202 unmethylated lambda CpGs addressable by 4Enz during the preparation of our CpG tag libraries. A read frequency map of the 22,728 reads that mapped to the lambda genome is shown in Fig. 1A. We identified 92% (1,107 hits) of the 4Enz CpGs with at least 1 read and an average reads-per-hit ratio of 20.5 ± 16.1 (mean ± SD). Thus, although the measures come from equally unmethylated sites, not all were identified with similar frequency. The read values were scattered between a minimum of 0 (8%) and a maximum of 99. For the mouse genome, whose methylation status is unknown, it is necessary to discriminate between highly methylated CpGs and CpGs that were poorly covered. Because we aligned the reads returned by each of the five channels separately, we identified the CpGs according to the number of channels in which they were identified, revealing a relation between the experimental coverage and the number of channels in which each CpG was identified. The hits were grouped into three classes: one- to three-channel hits, four-channel hits, and five-channel hits (Fig. 1 B and C). We compared the number of CpGs identified in the mouse and lambda genomes according to these criteria. At a sequencing depth of about 30 million mapped reads from five channels, for lambda (100% unmethylated), the majority of the CpGs (61%) were identified by reads that came from all five channels and only a small portion (16%) from one to three channels. The proportion of lambda one- to three-channel hits plus the proportion of not identified CpGs estimate the failure rate of the method. In contrast to lambda, we found only 24% of mouse hits in five channels. The majority of which (63%) were in one to three channels at an average of 2.5 reads per hit, which reflects the widely methylated status of the genome and the sensitivity of the method to detect CpGs even when they are heavily methylated (Fig. 2). The box-and-whisker plots in Fig. 1 B and C show that CpGs located in the same class were identified with a similar level of reads. Interestingly, the MSCC scores (reads per hit) recorded for mouse or lambda five-channel hits show a similar distribution of values (median MSCC score of 24.5 for lambda and 23.2 for mouse). The distributions of MSCC scores obtained from the mouse and lambda genomes are compared in Fig. 1D, which reflect the difference between a completely unmethylated genome and one with different levels of hypomethylation. For the mouse genome, we obtained higher frequencies of CpGs identified with lower numbers of reads. However, when we compared lambda CpGs against those lo-

Fig. 1. Analysis of reads that mapped to lambda DNA: comparison with mouse. (A) Plots of reads per hit (identified CpGs) along the lambda genome. Upper) Reads-per-hit plot of identified forward and reverse tags. (Lower) Combined forward and reverse reads per hit. (B) Box-and-whisker plots (median, quartiles, and fifth and 95th percentiles) representing the reads-per-hit distribution as a function of the number of channels in which the CpG-identifying reads were found in the lambda genome. (C) Same as in B but for CpGs detected in the mouse genome. (D and E) Frequency histograms of the reads per hit recovered for CpGs identified with 4Enz in the five sequencing channels, ch, channel. (D) Distribution of read recovery from the lambda genome is compared with that of reads recovered from the whole-mouse genome. (E) Distribution of read recovery from the lambda genome is compared with that of the reads per hit found in CpGs that were located in UMRs of the mouse genome validated by bisulfite sequencing analysis.
cluded in the newly discovered mouse UMRs, the MSCC score distributions showed no difference (Fig. 1E). By comparing different CpG tag libraries in which the same amount of lambda DNA was introduced as an internal standard, we found that certain sites systematically perform better or worse than others (SI Appendix, Fig. S2). Although this systematic bias increases the variability above that expected from the Poisson distribution, it is shown below that the level of methylation of a CpG under study is the primary parameter determining the final number of reads.

Assigning Methylation Status at Single-CpG Resolution. The 4Enz set of methylation-sensitive restriction enzymes accurately targets 6 million CpGs located in the context of five different patterns (CCGG, CACG, GCCG, CGCG, and GCGG) and collects information about their methylation state. The number of CpGs and the genomic regions that can be studied (addressable CpGs) depend not only on the number of restriction enzymes used but on the number of CpGs with unique tags (SI Appendix, Tables S1, S2, S3, S4, and S5). Although the abundance of any individual tag in the CpG tag library is expected to be inversely proportional to the methylation state of the addressed CpG (15, 16), the demonstrated local bias impairs the usefulness of the method to perceive moderate variations in levels of methylation within a genome (SI Appendix, Fig. S2). Despite this limitation, the method is sensitive to detect the small fraction of hypomethylated CpGs in the genome (Fig. 2). We used receiver operating characteristic (ROC) analysis to visualize and compare the performance of the method to classify CpGs in different categories of methylation according to their MSCC scores. ROC curves (Fig. 3) were built following the strategy described in SI Appendix, SI Materials and Methods using the MSCC scores and methylation status of a panel of 358 CpGs validated by bisulfite sequencing analysis.

Fig. 3 shows the area under the curve (AUC) plots for two situations: detection of sites <75% methylated and detection of sites <25% methylated. The AUC in Fig. 3A is 0.89 ± 0.02 and represents the probability that a randomly selected hypomethylated CpG (<75% methylation) will score a higher number of reads than a randomly selected heavily methylated CpG (>75% methylation). We found a MSCC score of 11 as the optimal cutoff for this classification (SI Appendix, Fig. S3A). A similar analysis to detect mostly unmethylated sites (<25% methylation) (Fig. 3B and SI Appendix, Fig. S3B) showed that using an MSCC score of 17 as a cutoff allows us to classify 67% of CpGs correctly as mostly unmethylated with a false discovery rate (FDR) of 7.2% (Fig. 3C). Single-CpG MSCC scores are affected by local systematic bias (SI Appendix, Fig. S2); however, CpGs with a tendency to be overestimated or underestimated are randomly distributed in the genome. When the level of methylation of a discrete region of the genome is measured by averaging the counts from the individual CpGs, the bias tends to cancel. For this reason, the larger the number of CpGs that can be covered in the MSCC analysis, the higher is the accuracy with which the method evaluates the level of unmethylation of a given region.

The average MSCC (avMSCC) score was thus used to quantify the level of hypomethylation of two regions known to differ by 50% in their methylation status as a result of imprinting (17) (SI Appendix, Fig. S4). As recorded in three independent experiments, the differentially methylated 1A domain of the Gnas locus produced an avMSCC score that was one-half of that observed for the nonimprinted 3′′ domain (SI Appendix, Fig. S4). We concluded that avMSCC scores reflect the hypomethylation status of a region accurately and reproducibly.

Distribution of Hypomethylation in the Genome. We mapped 29 million reads to 3 million unique CpG tags (SI Appendix, Table S5) and built a read frequency table (all-CpGs-all-hits frequency table; SI Appendix, Table S6) listing all addressable CpG tags (forward and reverse) and the number of times that each unique tag was identified. The MSCC threshold values derived from ROC analysis were used to assign methylation status to each CpG listed in this table, giving a distribution of hypomethylation throughout the entire genome.

The landscape that emerged from this analysis is in complete agreement with the known genome-wide mosaic pattern showing heavily methylated sites sharply separated from hypomethylated sites (Fig. 2 and SI Appendix, Fig. S4). However, the hypomethylated regions corresponded poorly with predicted CGIs. We analyzed the degree of hypomethylation in four different sets of CGIs: Gardiner-Garden and Frommer (GG&F), Takai and Jones (T&J), CpgCluster CGIs, and Epi-CGIs; with the last being based on an algorithm that combines the GG&F criterion with information gathered from epigenetic marks (6–9) (Fig. 4).

Although our method can interrogate, on average, one of every two CpGs located in CGI-like loci, not all CGIs are interrogated to the same extent (Fig. 4A and Inset). We included in our analysis those CGIs in which we can address at least one-third of the total CpGs (79% of T&J, 52% of GG&F, 56% of CpgCluster CGIs, and 60% of Epi-CGIs; SI Appendix, Table S8). Fig. 4A shows the distribution of the degree of unmethylation (avMSCC score) for 16,731 of the mouse genome’s T&J CGIs. According to the scores, the T&J set partitions into 14,993 (~90%) hypomethylated islands [MSCC score: 21.5 ± 4.6 (mean ± SD)] and 1,738 (~10%) heavily methylated islands. Bisulfite sequencing confirmed this partition of CGIs (SI Appendix, Fig. S5). The majority of T&J CGIs are located in promoter regions (SI Appendix, Table S7). This bias is probably what makes the T&J set very specific in terms of predicting UMRs; however, sensitivity is a concern for this algorithm. We performed the same analysis for the methylation status of CGIs originated with the CpgCluster algorithm (Fig. 4B). This set is three times larger than the T&J set, but 58% (39,022) of them were found to be heavily methylated, whereas only 32% (21,386) were found to be unmethylated [avMSCC score: 20.7 ± 6.4 (mean ± SD)]. Notice that whereas the T&J criterion was more effective in predicting UMRs, it missed
The average MSCC score of region of genes (3 Kb). For this analysis of methylation at CGIs, we made the CpGCluster program (9) specific and Table S9 and SI Appendix CGI-like and non-CGI UMRs as detected genome-wide in the mouse ≥ D.

We analyzed the colocalization of UMRs with predicted CGIs and classified them in non-CGI UMRs (do not overlap) and CGI-like UMRs (overlap). The result showed a poor correspondence. For example the T&J and CpGCluster CGIs fail to detect 75% and 60%, respectively, of the experimentally determined UMRs (SI Appendix, Table S9). Indeed, 52% of the UMRs could not be detected by any of the CGI-defining algorithms (Fig. 5). The high failure rate, even after combining different CGI sets, suggests that the algorithms are failing to include one or more critical features. We hypothesize that protein-binding DNA elements must be a ubiquitous feature shared by all the UMRs.

The original quantitative criteria used to define CGIs were based on a small set of sequences, likely biased by the limited size of the 1985 GenBank database (4, 7). Later programs readjusted thresholds and changed how the edges of CGIs are defined (5, 8). However, most of these new computational methods still classifying a substantial number of CGIs compared with CpGCluster (8,926 CGIs). We found that 70% of these 8,926 unmethylated loci are located in introns or intergenic regions. We extended this analysis to GG&F CGIs and Epi-CGIs (SI Appendix, Table S8). The methylation profile of the GG&F CGIs resembled that of CpGCluster CGIs (Fig. 4B); however, the number of unmethylated loci increased to 23,318 (SI Appendix, Table S8). The Epi-CGI algorithm appears to be as effective as the T&J algorithm in predicting unmethylated islands (67% of the Epi-CGIs analyzed); even though it finds the smallest number of unmethylated loci (11,191), these are less biased toward TSS regions compared with the T&J set (SI Appendix, Tables S7 and S8).

Reproducibility. The avgMSCC score of five-channel hits in our dataset (Sla2a) was 28.3. Thus, five-channel hits identify unmethylated CpGs. We found that of 199,618 HpaII five-channel hits with an avgMSCC score of 29.5, 104,701 hits mapped to 15,098 (unmethylated) T&J CGIs. In an independent experiment sequenced to a similar depth (Sla2a), we found that 90,301 HpaII five-channel hits mapped to 14,303 T&J CGIs. A common set of 14,054 T&J CGIs was identified in the two experiments (Fig. 4C), indicating that the MSCC approach is a highly reproducible tool to identify UMRs.

Genome-Wide Annotation of Experimentally Determined Hypomethylated Regions. We made the CpGCluster program (9) specific for 4Enz CpGs (GCGC, CCGG, GGCG, CCGC, and ACGT) and used it to cluster all CpGs identified in five channels. This group comprises 735,407 CpGs with a median MSCC score of 23, of which 90% scored more than 10 reads (Fig. 1C). The modified CpGCluster program creates clusters of addressed CpGs with specific inter-CpG distances (SI Appendix, Fig. S6). If all 6 million addressable CpGs were randomly distributed, the distances between neighboring sites should follow the geometrical distribution with a mean intersite separation of 311 bp. We set the distance to 300 bp to search for hypomethylated CpG clusters (w300 clusters) and found 559,901 hypomethylated CpGs grouped in 64,266 clusters (SI Appendix, Table S9). Although the five-channel CpGs included in these clusters have an avgMSCC score of 29, we also found 287,456 CpGs with an avgMSCC score of 6.9. This finding is an indication that certain clusters could have a considerable number of CpGs with high rates of methylation (SI Appendix, Fig. S7B and SI Appendix, Table S10).

Isolation and Analysis of UMRs. We showed that single CpGs with MSCC scores ≥17 can be classified as mostly unmethylated (<25% methylation), generating a low number of false-positive results (7% FDR). We reasoned that a hypomethylated region with an avgMSCC score ≥17 will have the majority of its CpGs in an unmethylated state. After calculating and applying this cutoff, we ended with a set of 46,804 UMRs that span 22 million bp of the mouse genome and include 1.3 million unmethylated CpGs (SI Appendix, Tables S9 and S11). A number of studies have shown that CpGs located in close proximity tend to share a common methylation state (15, 18, 19). The penetration of this correlation increases with the proximity of neighboring CpGs. We found that 97% of nonaddressable CpGs located in UMRs have at least 1 addressable CpG within a distance of 100 bp. The correlation between methylation states for CpGs within a distance of 100 bp has been estimated to be ~75% (18). Bisulfite sequencing analysis of randomly selected UMRs confirms the methylation phenomenon (SI Appendix, Figs. S7 and S8). We conclude that the UMRs with avgMSCC scores ≥17 constitute a set of mostly unmethylated sequences.

We analyzed the colocalization of UMRs with predicted CGIs and classified them in non-CGI UMRs (do not overlap) and CGI-like UMRs (overlap). The result showed a poor correspondence. For example the T&J and CpGCluster CGIs fail to detect 75% and 60%, respectively, of the experimentally determined UMRs (SI Appendix, Table S9). Indeed, 52% of the UMRs could not be detected by any of the CGI-defining algorithms (Fig. 5). The high failure rate, even after combining different CGI sets, suggests that the algorithms are failing to include one or more critical features. We hypothesize that protein-binding DNA elements must be a ubiquitous feature shared by all the UMRs.

The original quantitative criteria used to define CGIs were based on a small set of sequences, likely biased by the limited size of the 1985 GenBank database (4, 7). Later programs readjusted thresholds and changed how the edges of CGIs are defined (5, 8). However, most of these new computational methods still
rely on the three initially considered DNA features: Obs/Exp CpG ratio, G + C content, and length. We found these three parameters to be highly variable among the UMRs (Fig. 5). For example, many non-CGI UMRs meet the first two criteria mentioned, but all of them have sizes below the 500 bp required by the T&J algorithm (Fig. 5C, second quadrant). Also, many UMRs do not overlap CGIs but meet the size required by different algorithms (SI Appendix, Table S9). This shows that the size (base pairs) of a region rich in CpGs should not be used as a filter to select for putatively functionally important unmethylated sequences. The size of the UMRs was found to be primarily a function of their genomic location, with larger UMRs overlapping TSSs and smaller UMRs being located in intergenic regions (Fig. 5B). On the other hand, there are UMRs that overlap predicted CGIs but the CpG and G + C richness are below the CGI thresholds (Fig. 5D, quadrants 1, 3, and 4), suggesting that in addition to colocalizing, they cover substantially different sequences. The inability to predict the edges of CGIs seems not to be trivial. It has been reported that tissue- and cancer-specific differentially methylated regions map to the shores of CGIs (20). Recently, computational algorithms have incorporated hidden Markov models (HMMs) to improve the detection of CGI borders. Many of the HMM CGIs incorporated the shores of previously defined islands (5). Although this new strategy predicts a relative large number of CGIs, the majority (75%) do not overlap our experimentally determined UMRs. Our description of thousands of UMRs not qualifying as CGIs outside of promoter regions is in agreement with the previous discovery of nonpromoter UMRs subject to tissue-specific de novo methylation and suggests the existence of additional tissue-specific UMRs not evident in liver (21).

The relation between CpG density and methylation has been studied in promoter and nonpromoter regions. Whereas promoters with high CpG density (more than one CpG every 20 bp) are found to be primary unmethylated, the methylation rate at sites where DNA outside promoters was found to increase with the CpG density until a threshold value of 0.025 (one CpG every 40 bp) was reached; beyond this threshold, the methylation rates fell sharply (19). Interestingly, we found that CGI-like UMRs and non-CGI UMRs differ in their CpG density. CGI-like UMRs have a median inter-CpG distance of 18 bp, which is just above the expected distance for the CpG dinucleotide in a DNA sequence with identical base composition (A = T = C = G). In contrast, the median distance between CpGs in the non-CGI UMRs is 40 bp (Fig. 5A, inset). Whether these differences have a relationship to the functionality of the underlying sequences has to be determined in future experiments. On the other hand, there is a clear pattern in how these two kinds of UMR partition among different noncoding compartments of the genome (Fig. 5A). Whereas 55% of CGI-like UMRs localize to TSS regions, only 12% of non-CGI UMRs do. Whereas less than 35% of CGI-like UMRs localize to intronic or intergenic sequences, more than 80% of non-CGI UMRs do. The finding that 70% of UMRs are located in exons, introns, and intergenic regions was unexpected. Introns and intergenic sequences account for almost 98% of mammalian genomes, and most of the CpGs located in these are heavily methylated (15). Mammalian DNA regulatory sequences are principally located in noncoding sequences but concentrated preferentially in the promoter regions of genes, leaving introns and intergenic regions virtually devoid of evolutionary constraints. However, the comparison of complete genomes has revealed a large number of conserved non–protein-coding DNA sequences mapping to intergenic and intronic regions, for most of which their biological function remains unknown (22).

We hypothesized that UMRs are highlighting functional sequences, which, if proven, emphasizes the usefulness of our method as a tool to identify loci at which epigenetic mechanisms could influence complex phenotypes and diseases. To provide genome-wide evidence in favor of our hypothesis, we evaluated the functional significance of our UMRs by asking if they overlap with highly conserved mammalian sequences and/or with experimentally determined protein-binding sites. To test the potential functionality of our UMRs, we used two tracks of the University of California, Santa Cruz Genome Browser (23). One is the mammal most conserved (MMC) track of conserved sequences or elements based on whole-genome alignments of different mammalian species. The other is the track for open regulatory annotation (ORegAnno), which includes literature-curated regulatory regions and transcription factor-binding sites (24). We reasoned that the sequences of these tracks could be used as probes that would allow us to follow the partitioning of regulatory sequences between the methylated regions and UMRs of the genome. We found that our UMRs are indeed enriched in both MMCs and ORegAnno sites (SI Appendix, Table S12). For ORegAnno sites, 25% populating 10% of the UMRs and non-CGI UMRs (with the majority in intronic and intergenic regions) reached enrichments of 100-fold compared with the concentration at which these regulatory elements are found in the whole genome (Fig. 6 and SI Appendix, Table S12). The likelihood that this enrichment occurred by chance is less than 4 in 100,000. Interestingly, 98% of all ORegAnno elements overlapping UMRs belong to binding sites for Esr1 and Foxa2. The Foxa 2 gene codes for the forkhead box protein A2, which is a transcriptional activator for liver-specific genes (25). Among the remaining 2% of regulatory elements, we also found tissue-related transcription-binding sites (i.e., Hnf4A). The hepatocyte nuclear factor 4a is a transcription factor found upstream in the regulation pathways of several hepatic genes (27). Surprisingly, the non-CGI UMRs produced, on average, a threefold greater enrichment in regulatory elements than the CGI-like UMRs. Relying on the way the protein-binding sites were partitioned, we believe that most of the UMRs represent regions rich in reg-

Fig. 6. Genomic distribution of UMRs and enrichment of annotated features. (A) Distribution of CGI-like and non-CGI UMRs in genomic regions. UMRs at TSS regions (~3 kb to +2 kb of TSS), in gene bodies (non-TSS exons, introns, and +3 kb of 3’ not-transcribed regions), and in intergenic DNA do not add up to 100 because only those UMRs with >90% overlap were considered. (B) Enrichment of ORegAnno sites and MMC sequences in the UMRs located in the indicated genomic regions. Enrichments are compared with abundance in the undiluted genome (numerical values are provided in SI Appendix, Table S12). **P < 0.005; ***P < 0.0001.
ulatory elements, with many of them being liver-related and possibly liver-specific. We used the “David Bioinformatics Functional Annotation Tool” Web application to bin UMR-containing genes according to functional annotation and analyzed the enrichments in three selected categories: biological process, molecular function, and tissue specificity (28). We took the top 5% of our UMRs, ranked according to their avgMSCC score. In each functional category, we sorted the results according to significance (smaller P values on top). For “tissue specificity,” liver is the most significantly enriched tissue (P value of 2 × 10−12). For “biological process,” genes related to system development are at the top (P value of 1 × 10−9). Finally, for “molecular function,” genes related to steroid hormone receptor activity are the most enriched class of function (P value of 1 × 10−11).

In conclusion, our screening for hypomethylated CpGs showed that CGIs are weak predictors of sensitive epigenetic loci and, in addition, revealed a large unexpected number of non-CGI UMRs with the highest enrichment in regulatory elements. The fact that 50% of the UMRs have one-half of the average CpG content of traditional CGIs prompts a rethinking of the relationship between CpG density, unmethylation, and functionality of the genomic subsequences. A mechanistic link between these variables is beginning to emerge. A recent report shows that the nonpromoter insertion of a promoterless CpG cluster is able to recruit the Cfp1 protein, which can bind to unmethylated CpGs and attract the Setd1 histone H3K4 methyltransferase complex, which, in turn, creates a new focus of H3K4me3 modification (29). There is evidence that this modification repels the methyltransferase involved in de novo methylation. However, only one-half of the cells carry the insertion-acquired methylation, indicating that CpG density, per se, is insufficient to maintain the unmethylated state. Our hypomethylation map showing thousands of UMRs not qualifying as CGIs but with the highest enrichment in DNA regulatory motifs supports the idea that methylation is the default state of CpGs, except for those that are protected from de novo methylation, which may be mediated by the action of DNA-binding proteins.

The idea that DNA-containing regulatory elements are furnished with a critical CpG density working as a local signal to recruit proteins able to create epigenetic marks that highlight functional sequences is simple and attractive. In this scenario, the unmethylated genomic regions (UMRs) could reflect the footprint of regulatory DNA-binding proteins that protected local sequences from the activity of de novo methylases.

4Enz MSCC vs. HpaII MSCC. The MSCC approach was first published in 2009 by Ball et al. (15) in a proof-of-principle report. The average reads-per-hit values (MSCC score) reported for unmethylated CCGGs was 5.0 ± 15.4, which identified 69% of the addressable sites with at least one read. In discussing the usefulness of the approach, the authors used statistical criteria to conclude that by increasing the sequencing depth, the method would allow for the identification of the UMRs of a genome. Having sequenced to a greater depth, which reached average MSCC scores for HpaII of 10.4 ± 10.0 for all CCGGs identified at least once, we sought to determine to what extent the HpaII hits are able to report on the hypomethylation measured using 4Enz five-channel hits and a window of 300 bp. Our all-channel HpaII hits allow for formation of 109,877 of 300-bp–based clusters, which overlapped with only 25,373 (54.2%) of 46,804 of our UMRs and with 7,207 (18%) of 24,399 of our non-CGI UMRs. We thus conclude that only the expanded 4Enz MSCC reports reliably on the genome’s hypomethylation.

Materials and Methods
CpG-tag libraries were prepared according to strategies similar to those outlined in ref. 15 except that four instead of one methylation-sensitive restriction enzyme was used, and the tags were retrieved with EcoP15I and not MmeI. CpG tag libraries were sequenced by Illumina Inc, using Genome Analyzer’s Solexa technology. Illumina Inc returned approximately 1 Giga-base of sequence partitioned into 36-nt long reads per library. After mapping to the mm9 Mus musculus reference genome, the data were analyzed by creating frequency histograms showing the number of times reads were found that identified any given CpG. For further details on methods and data analysis, including materials, acknowledgements, 12 tables, and 8 figures, see SI Appendix, SI Materials and Methods.

ACKNOWLEDGMENTS. This study was supported by the Intramural Research Program of the National Institutes of Health (2009 National Institutes of Health Director’s Challenge Award and Project ZI ES101643 to L.B.).