RESEARCH PAPER

Epigenetics 9:4, 483–491; April 2014; © 2014 Landes Bioscience

Fundamental diversity of human CpG islands at multiple biological levels

Jia Zeng, Hema K Nagrajan, and Soojin V Yi*

School of Biology; Georgia Institute of Technology; Atlanta, GA USA

Keywords: CpG Islands, DNA methylation, whole genome methylation maps, gene ontology, cancer, aging

CpG islands (CGIs) are commonly used as genomic markers to study the patterns and regulatory consequences of DNA methylation. Interestingly, recent studies reveal a substantial diversity among CGIs: long and short CGIs, for example, exhibit contrasting patterns of gene expression complexity and nucleosome occupancy. Evolutionary origins of CGIs are also highly heterogeneous. In order to systematically evaluate potential diversities among CGIs and ultimately to illuminate the link between diversity of CGIs and their epigenetic variation, we analyzed the nucleotide-resolution DNA methylation maps (methylomes) of multiple cellular origins. We discover novel "clusters" of CGIs according to their patterns of DNA methylation; the stably hypomethylated CGI cluster (cluster I), sperm-hypomethylated CGI cluster (cluster II), and variably methylated CGI cluster (cluster III). These epigenomic CGI clusters are strikingly distinct at multiple biological features including genomic, evolutionary, and functional characteristics. At the genomic level, the stably hypomethylated CGI cluster tends to be longer and harbors many more CpG dinucleotides than those in other clusters. They are also frequently associated with promoters, while CGI clusters II and III mostly reside in intragenic or intergenic regions and exhibit highly tissue-specific DNA methylation. Functional ontology terms and transcriptional profiles co-vary with CGI clusters, indicating that the regulatory functions of CGIs are tightly linked to their heterogeneity. Finally, CGIs associated with distinctive biological processes, such as diseases, aging, and imprinting, occur disproportionately across CGI clusters. These new findings provide an effective means to combine existing knowledge on CGIs into a genomic context while bringing new insights that elucidate the significance of DNA methylation across different biological conditions and demography.

Introduction

The human genome exhibits high levels of DNA methylation,[1-5] a pattern referred to as "global DNA methylation."[6] However, contrary to this general pattern, there are numerous genomic regions characterized by low levels of DNA methylation. These regions, called "CpGs Islands (CGIs)," were originally discovered as sequences with an unusually high frequency of unmethylated (hypomethylated) CpG dinucleotides.[7-9] With the advent of genome sequencing, several computational algorithms have been developed to identify CGIs from the genomic sequences.[10,11] A key feature of these computational algorithms is a metric to quantify the observed frequency of CpG dinucleotides normalized by the G+C content, commonly referred to as "CpG O/E."[1,12] Genomic regions exhibiting particularly high CpG O/E, among other characteristics, are generally considered good CGI candidates (e.g., ref. 11).

Even though CGIs are generally characterized by their unmethylated status, some of them undergo DNA methylation in a tissue- or developmental stage-specific manner (e.g., refs. 13–16). Aberrant methylation at some CGIs is implicated with disease, particularly cancer.[17,18] Interestingly, recent studies have begun to unfold intriguing functional heterogeneity among CGIs. For example, long and short CGIs exhibit different regulatory activities such as gene expression complexity[19] as well as nucleosome depletion patterns.[20] A recent evolutionary study determined that while the majority of CGIs may actively avoid DNA methylation, some CGIs are likely to maintain high CpG contents via methylation-independent processes such as biased gene conversion.[21] These findings begin to shed light on the potential diversity among CGIs. At the same time, they highlight many unanswered and critical questions: for example, are all CGIs similarly hypomethylated? Do all CGIs exhibit tissue and developmental stage specific variation in DNA methylation? Alternatively, is there a group of CGIs that tends to exhibit variable patterns of DNA methylation? How are these variations in DNA methylation related to regulatory functions of CGIs? Do methylation profiles of CGIs differ according to their evolutionary mechanisms?

Here we utilized recently generated whole genome DNA methylation maps (methylomes) with distinct cellular origins[22-25] to investigate these pressing questions. Because these whole genome methylation maps provide information on nearly every CpG dinucleotide in the genome, they are superior to other assays in

*Correspondence to: Soojin Yi; Email: soojinyi@gatech.edu
Submitted: 10/09/2013; Revised: 12/19/2013; Accepted: 12/23/2013; Published Online: 01/13/2014
http://dx.doi.org/10.4161/epi.27654

Table 1. Human DNA methylome data sets used in this study

Reference	Tissue	Gender	Data set ID	Coverage	Total Mapped CpGs (millions)
22	Embryonic stem cells	F	GSE19418	9.1×	26.2
22	Neonatal foreskin fibroblasts	M	GSE19418	9.8×	26.4
25	Prefrontal cortex of brain	M	GSE37202	11.4×	26.8
23	Peripheral Blood Mononuclear Cells	M	GSE17972	10×	27.1
24	Sperm	M	GSE30340	16×	28.2
Additional low-coverage methylomes					
36	Placenta	F	GSE39775	1.6×	23.9
36	Cerebellum	M	GSE39775	0.3×	8.3
36	Kidney	M	GSE39775	0.5×	8.9

analyzing detailed patterns of DNA methylation variation.[5,26,27] We show that CGIs can be classified into several distinctive clusters based upon their patterns of DNA methylation variability. These epigenetically identified clusters of CGIs exhibit highly significant differences in genomic and evolutionary features and are characterized by distinctive functions and transcriptional profiles. Our results thus reveal fundamental diversity in human CGIs, and may provide new insights into epigenomic surveys of human health and aging.

Results

Patterns of CGIs DNA methylation in whole genome methylomes

We analyzed comprehensive whole genome nucleotide-resolution DNA methylation maps (referred to as "methylomes" henceforth) from five distinct human tissue samples. These data include the prefrontal cortex region of the brain from Zeng et al.,[25] embryonic stem cells and neonatal foreskin fibroblasts from Laurent et al.,[22] peripheral-blood mononuclear cells from Li et al.,[23] and sperm from Molaro et al.[24] These five methylomes offer similarly comprehensive, high coverage information across the whole genome (**Table 1**). We calculated the methylome-specific DNA methylation of 26.7 million CpG dinucleotides (88.7% of all CpG dinucleotides in the human genome) and of 25 131 CGIs (89% of all annotated CGIs in the UCSC genome browser). Comparisons to methylation data from other methods indicate that the data we use offer a superior resolution for examining the detailed variation of DNA methylation in CGIs (**Fig. S1**).

Figure 1A shows the mean DNA methylation levels (as measured by fractional DNA methylation, Materials and Methods) from the whole genome and from CGIs. At the whole genome level, the five methylomes exhibit heavy (63–81%) DNA methylation, corresponding to those reported in the original studies of these methylomes.[22-25] Consistent with many previous reports (e.g., refs. 28–31), CGIs exhibit significantly reduced methylation compared with the genomic background (Mann-Whitney test, $P < 10^{-15}$, **Fig. 1A**). Notably, CGIs in sperm show the most pronounced pattern of hypomethylation, even though at the whole genome level sperm is not the most heavily methylated (**Fig. 1A**). Previously reported sperm-specific hypo-DNA methylation compared with somatic cells (e.g., refs. 2, 32, and 33) may have captured the particularly strong hypomethylation of sperm CGIs. **Figure 1B** illustrates the distribution of mean methylation levels of CGIs from the five methylomes. The majority of CGIs are hypomethylated (methylation level < 20%). However, substantial numbers of CGIs are hypermethylated (methylation level > 80%) (**Fig. 1B and C**). Interestingly, there is a strong negative correlation between CpG island length and average methylation level across methylomes: longer CGIs tend to be more markedly hypomethylated (Spearman's $\rho = -0.38$, $P < 10^{-16}$, **Fig. 1D**).

Distinctive clusters of human CGIs based on DNA methylation patterns

We employed a hierarchical clustering approach (Materials and Methods) to group CGIs according to their similarities of DNA methylation across the five methylomes. The resulting heatmap of DNA methylation variation across CGIs reveals several intriguing patterns (**Fig. 2**). Strikingly, CGIs form several distinct clusters according to their methylome-specific DNA methylation patterns (**Fig. 2**). As expected, many CGIs exhibit sparse levels of DNA methylation in all five methylomes. These CGIs are designated as "Cluster I" (**Fig. 2A**). It is notable that many CGIs in this cluster still exhibit high levels of methylation variability (**Fig. 2A**). The remaining CGIs are differentially methylated across methylomes. Among these, approximately half of the CGIs are notably hypomethylated in sperm, yet exhibit highly variable patterns of methylation in somatic tissues and embryonic stem cells (Cluster II, **Fig. 2A**). Some of these CGIs may correspond to CpG-rich sequences that are known to be specifically hypomethylated in sperm (e.g., refs. 33–35). The remaining CGIs tend to exhibit relatively high levels of DNA methylation across the examined methylomes (Cluster III in **Fig. 2A**).

We can further divide Clusters II and III into sub-clusters. For example, Cluster II can be subdivided into those CGIs that exhibit sparse methylation in sperm but relatively heavy (yet considerably variable) methylation in somatic cells (sub-cluster IIa), and those exhibiting sparse methylation in sperm and highly variable (ranging between hyper- and hypo-) methylation in somatic cells (sub-cluster IIb) (**Fig. 2B**). Cluster III includes a distinctive sub-cluster of CGIs that exhibit heavy methylation in all tissues

(sub-cluster IIIa), compared with those that show variable methylation across tissues (sub-cluster IIIb, **Fig. 2C**). The presence of these multiple CGI clusters remains when the X chromosome is analyzed separately (**Fig. S2**). We also examined a larger number of tissues, including three additional methylomes of placenta, kidney, and cerebellum.[36] These additional methylomes consist of markedly lower sequencing coverage and/or fewer CpG sites compared with the five comprehensive methylomes (**Table 1**). Despite such difference in sequence coverage and quantity, clustering analyses using these eight methylomes clearly demonstrate the presence of three distinctive CGI clusters (**Fig. S3**).

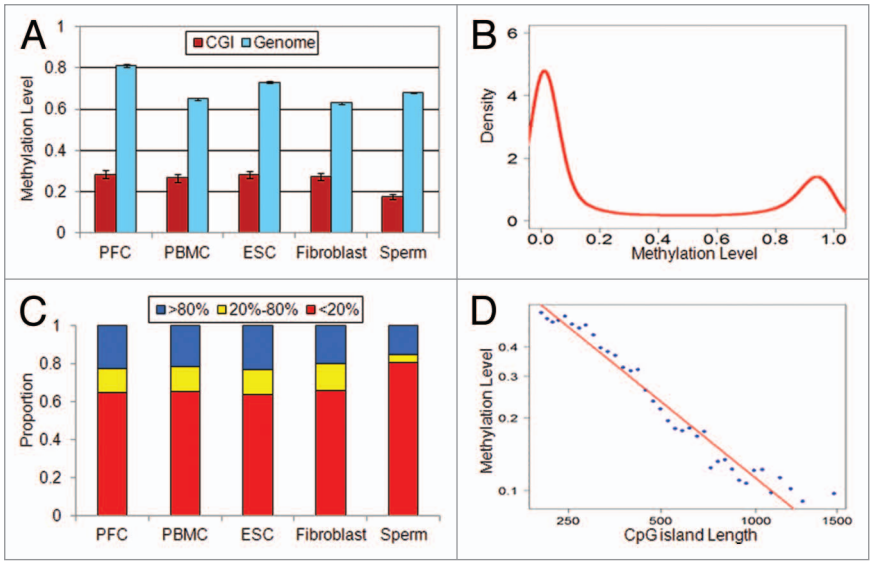

Figure 1. Overview of CGI methylation in 5 human tissues. (**A**) Mean methylation levels of all CpGs in the genomic background (n = ~26.7 million CpGs, blue bars) vs. those in CpG island context (n = ~2.1 million CpGs belonging to CpG islands, red bars). (**B**) Density distribution of mean DNA methylation levels of CGIs across the five human methylomes. (**C**) Distribution of CGIs that are lowly methylated (< 20% mean fractional methylation levels), intermediately methylated (20–80% mean fractional methylation levels), and highly methylated (>80% mean fractional methylation levels) across the five methylomes examined. (**D**) Correlation of CpG island length and methylation level. A regression of log transformed CpG island length vs. log transformed average methylation level from 5 human methylomes, divided into 40 bins, shows a high negative correlation (R^2 = 0.96 for binned data).

Epigenomically identified CpG Island Clusters are distinct at genomic level

Intriguingly, these CGI clusters, which have been identified solely based on patterns of DNA methylation variation, differ significantly in several genomic characteristics. The Cluster I CGIs tend to be the longest, which is consistent with our observation that longer CGIs tend to be less methylated (**Fig. 1D**). They are also the most GC-rich, exhibit the highest CpG O/Es, and harbor the largest numbers of CpG dinucleotides compared with CGIs in other clusters (**Fig. 3A–C**). On the other hand, Cluster III CGIs are distinctively shorter than those in other clusters, as well as exhibiting lower GC contents and lower CpG O/Es. Notably, these CGIs consist of a strikingly low number of CpG dinucleotides compared with those in other clusters (**Fig. 3D**). CGIs in the Cluster II generally exhibit genomic characteristics that are intermediate of the other two clusters. These differences are not due to a bias in mapping: CGIs in the three clusters show similarly high mapping coverages (results not shown). Autosomal and X-linked CGIs also exhibit a heterogeneous distribution: CGIs on the X chromosome are slightly yet significantly enriched in Cluster I, while deficient in Cluster III (**Table 2**).

Furthermore, these CGI clusters are highly heterogeneously distributed across different genomic regions. Cluster I largely consists of promoter-associated CGIs, while Clusters II and III include large numbers of intragenic and intergenic CGIs (**Fig. 3E**). The observation that CGIs in Clusters II and III tend to exhibit highly methylome-specific patterns of DNA methylation is thus consistent with the idea that intragenic and intergenic CpG sites are highly variably methylated and exhibit tissue-specific DNA methylation.[37]

DNA methylation variation supports evolutionary diversity of CGIs

Cohen et al.[21] used the genomes of humans, chimpanzees, orangutans, rhesus macaques, and marmoset (encompassing an evolutionary history dating over 35 million years ago[38]) to infer evolutionary forces underlying CGIs. They classify CGIs into three evolutionary categories: the first class, "hypodeamination" CGIs, exhibit evolutionary signatures of hypomethylation. The second class, "biased gene conversion" (BGC) CGIs, are under strong BGC pressures. The third and final class, "pseudo" CGIs, have been formed by stochastic processes. Since evolutionary patterns of substitution depends upon germline DNA methylation, it is interesting to compare the evolutionarily inferred methylation to those of CGI clusters (**Fig. 4**), contrasting two aspects: evolutionary inference vs. present methylation status, and germline vs. somatic cells. Cluster I CGIs are overrepresented in 'hypo-deamination' groups (**Fig. 4B**), as expected; evolutionary pressures for germline hypomethylation share some of the same underlying mechanisms with somatic hypomethylation.[39,40] In contrast, Clusters II and III CGIs include large numbers of BGC islands (**Fig. 4C and D**). It is counterintuitive that many Cluster II CGIs, which are hypomethylated in sperm, are classified as BGC islands; if evolutionary classification truly indicate germline DNA methylation patterns, we should expect to see most Cluster II CGIs belonging to hypodeamination islands. CGIs in Cluster III include disproportionately large numbers of "pseudo" CGIs (9%, significantly higher than 0 and 2% in CpG island ClustersI and II). Consequently, these results indicate that many genomic regions that have arisen by non-methylation related processes (BGC or chance) and have been co-opted as current CGIs, exhibiting highly variable DNA methylation across different tissues.

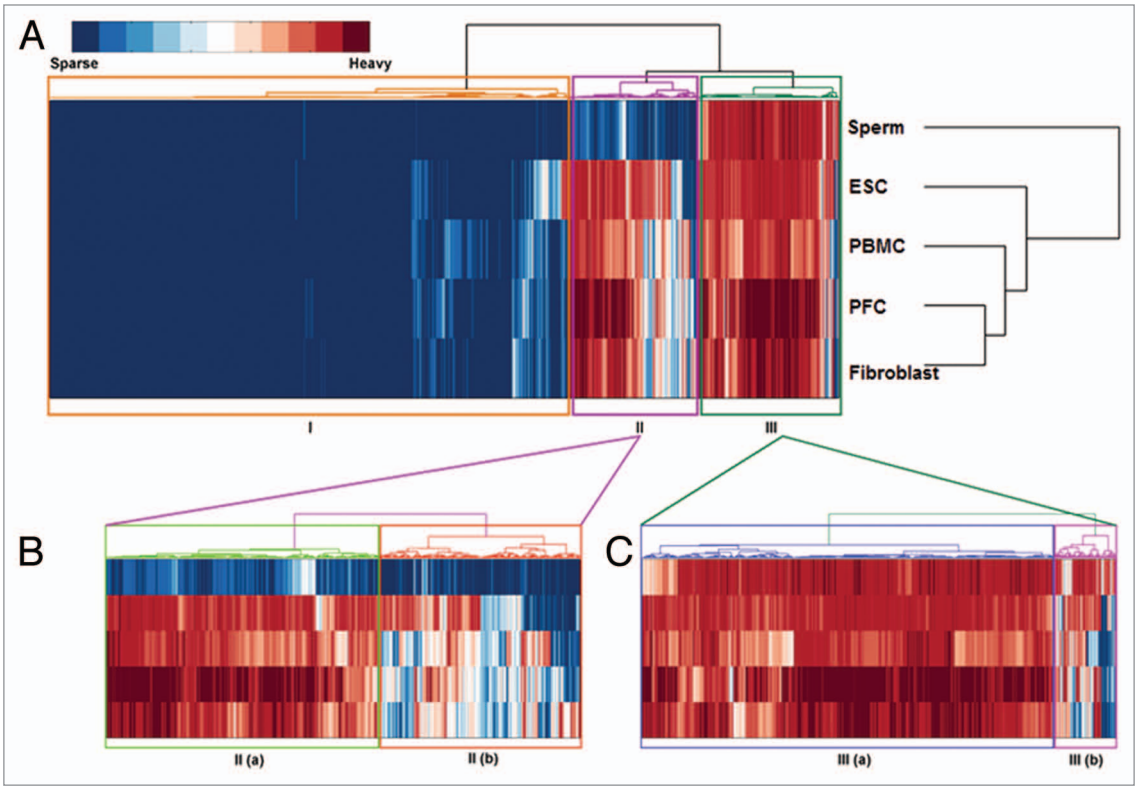

Figure 2. Hierarchical Clustering of CGIs according to their methylation levels in 5 human methylomes. The bar on top left represents relative methylation levels, where "Heavy" stands for the methylation level of 100% while "Sparse" stands for no methylation. (**A**) Three distinctive clusters are indicated. Cluster I, II, II consist of 16 549, 4108, and 4474 CGIs, respectively. ESC: embryonic stem cells; PBMC: peripheral-blood mononuclear cells; PFC: prefrontal cortex of brain. (**B**) Some CGIs are hypomethylated in the sperm methylome, but hypermethylated (yet considerably variable) in other methylomes (IIa, n = 2357) or exhibit highly variable levels of hypermethylation in other methylomes (IIb, n = 1751). (**C**) Some CGIs are generally hypermethylated in all methylomes (IIIa, n = 3885) or exhibit some level of tissue-specific hypomethylation (IIIb, n = 589).

Functional diversity of CGIs reflected in DNA methylation variation

Gene ontology analysis shows that these clusters are enriched in functionally distinct genes (**Table 3**; **Table S1**). Cluster I CGIs are generally associated with genes participating in "housekeeping" functions such as transcription and RNA-processing, consistent with the idea that hypomethylation of promoter CGIs regulate housekeeping functions (e.g., refs. 28 and 41). In addition, some developmental functions, in particular neuron development, are also overrepresented in Cluster I. Cluster II CGIs are associated with genes involved in morphogenesis and cell-cell adhesion. Genes associated with Cluster III CGIs have fewer ontology terms that are significantly enriched, which include protein phosphorylation, negative-regulation pathways, and signal transduction (**Table 3**). Variable DNA methylation of Clusters II and III may regulate tissue- and developmental stage-specific functions. We then directly examined patterns of gene expression across tissues and cell types using recent RNA-seq based gene expression profiles from six distinct human tissues.[42] As expected,[28,41] genes associated with Cluster I CGIs are the most broadly expressed (tissue specificity is the lowest) compared with those associated with Cluster II and III CGIs (**Fig. 5A**). Genes associated with Cluster II exhibit the most tissue-specific patterns of gene expression (**Fig. 5A**). Cluster III CGIs are associated with genes demonstrating intermediate tissue specificity of gene expression compared with the other two clusters. The same pattern was also observed in the Novartis gene expression data (**Fig. 5B**).[43] In addition, we examined the average number of transcription factor binding sites (TFBS), normalized by length, in CGIs. Cluster I CGIs have the largest number of TFBSs while Cluster II has the least (**Fig. 5C**), consistent with the observation that ubiquitously active promoters harbor large numbers of transcription factor binding sites and many CpGs, while promoters that are tissue-specific have fewer CpGs.[44] At the same time, even tissue-specific CGIs encode large number of potential transcription factor binding sites (**Fig. 5C**).

CGIs in disease, genomic imprinting, and aging

A recent study[45] compared DNA methylation maps of over 1149 tumors with differing tissue origins and identified genes whose CpG island promoters frequently exhibit aberrant hypermethylation in cancers. Among these promoters that are prone to aberrant hypermethylation in cancers, 663 overlapped with our CpG island data. We find that 649 (97.8%) of them belonged to Cluster I. This is a significant overrepresentation even after considering that most promoter CGIs are found in Cluster I ($P = 0.03$, the Fisher exact test). The remaining 14 CGIs are from the Cluster II. A group of genes known as cancer/testis antigens (CTAs) are those that are typically expressed in testis yet

Table 2. The numbers of X-linked CGIs compared with all CGIs

	Cluster I	Cluster II	Cluster III	
X-linked CGIs	366	76	27	
All CGIs	16183	4032	4447	
				$P < 10^{-16}$*

X-linked CGIs are overrepresented in the Cluster I and underrepresented in Cluster III. *Chi-square test of heterogeneity.

aberrantly expressed in a large number of cancers.[46-51] Among the 9 CTA families[51,52] overlapping with CGIs, 7 are found in the sperm-hypomethylated Cluster II, which is a significant enrichment, as expected ($P < 0.05$, the Fisher exact test).

We also examined the association between imprinted genes and different CGI clusters. Among the monoallelically expressed human genes (http://www.geneimprint.org/, Materials and Methods), 33 overlapped with the CGIs in our data. Thirteen out of these 33 imprinted genes are found in the Cluster II, representing a significant enrichment (the expected number of imprinted genes in the Cluster II is 5, $P < 0.05$ by the Fisher exact test). In addition, we investigated whether CGIs that exhibit differential DNA methylation with respect to aging tend to be preferentially associated with specific clusters. For this purpose, we used the whole genome DNA methylation maps of three individuals of different ages (newborns, 26 y old, and a centenarian).[53] This study has identified 17 930 "aging" differentially methylated regions (DMRs),[53] 294 of them overlapping with CGIs. While these CGIs are distributed across all three clusters, they are highly significantly over-represented in the Cluster II ($P < 10^{-6}$, Fisher's exact test), and significantly underrepresented in the Clusters I and III ($P < 0.05$ and $P < 10^{-6}$, respectively, Fisher's exact test).

Discussion

We utilized recently generated nucleotide-resolution whole genome DNA methylation maps to study variation of DNA methylation across multiple methylomes with distinct cellular origins. Our analyses uncover a novel diversity in CGIs. As well known, CGIs are, on average, significantly hypomethylated compared with the genomic background (**Fig. 1A**).[8,9] However, many CGIs exhibit highly variable patterns of DNA methylation across multiple methylomes (**Fig. 2**). In fact, CGIs can be classified into several distinctive groups, or "clusters." The first cluster of CGIs is generally hypomethylated in multiple methylomes, and found in or near genes involved in essential, housekeeping pathways including transcription regulation and RNA processing. These CGIs are also associated with promoters (**Fig. 3E**). Together, these traits indicate that cluster I CGIs most closely fit the original definition of CGIs.[10,31]

In comparison, some CGIs are sparsely methylated in sperm, yet exhibit variable levels of DNA methylation in other methylomes (Cluster II in **Fig. 2A**). We tentatively refer to them as "sperm-hypomethylated" CpG island clusters, to be consistent with previous studies reporting sperm-specific hypomethylation.[2,32,33] They are enriched in cell adhesion and embryonic morphogenesis functions, and generally exhibit more tissue-specific transcription

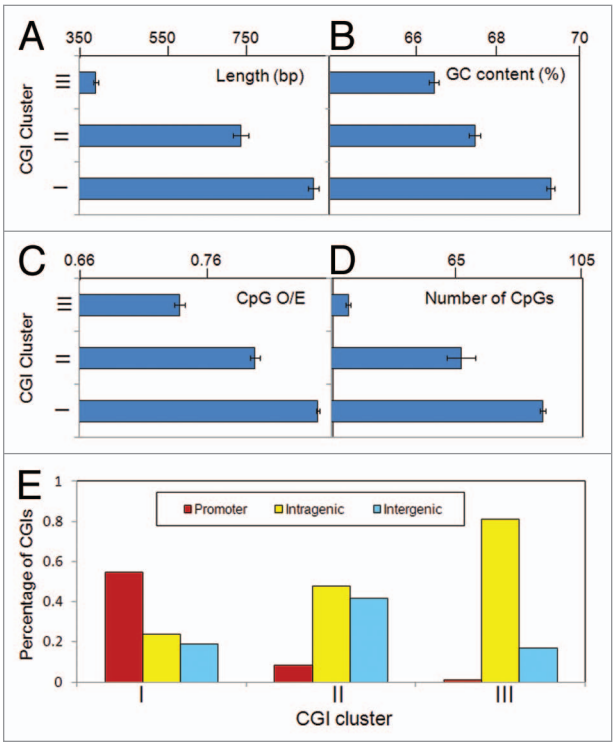

Figure 3. Contrasting genomic features of the three CpG island clusters. Significant differences are found in (**A**) lengths, (**B**) GC content, (**C**) CpG O/E, and (**D**) number of CpG dinucleotides among the three clusters. (**E**) Occurrence of promoter-, intragenic-, and intergenic-CGIs across the three CGIs clusters. All pairwise comparisons are highly significant ($P < 10^{-6}$).

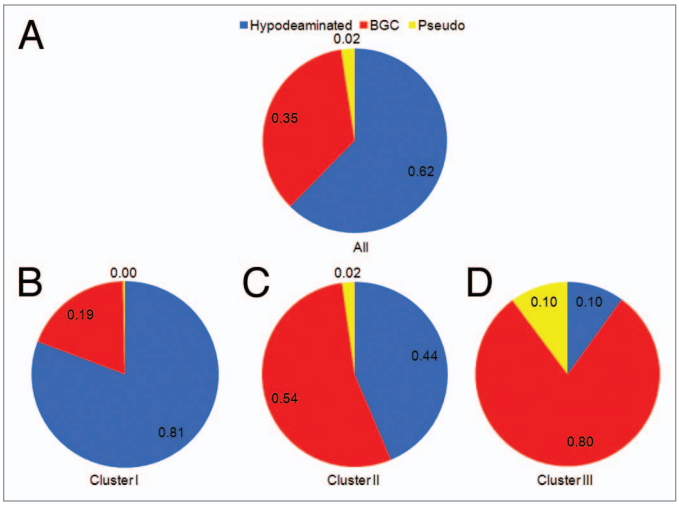

Figure 4. Evolutionary classification of CGIs. Frequencies of hypodeaminated, biased gene conversion (BGC), and pseudo CGIs in (**A**) all CGIs, (**B**) Cluster I, (**C**) Cluster II, and (**D**) Cluster III.

profiles compared with those in the Cluster I. Finally, approximately one fifth of all CGIs exhibit some degree of DNA methylation across different methylomes (Cluster III in **Fig. 2**). Even though Cluster III CGIs are generally hypermethylated in the examined methylomes, they overlap with a large number of

Table 3. Distinctive functional enrichments of specific genes according to the variable DNA methylation of CGIs

GO terms	Description	*P* values	*FDR-P* values*
Cluster I CGIs			
Sparse sperm methylation, sparse ESC and somatic cell methylation			
GO:0006350	Transcription	2.00×10^{-28}	3.90×10^{-25}
GO:0045449	Regulation of transcription	3.77×10^{-27}	7.36×10^{-24}
GO:0006396	RNA processing	1.11×10^{-17}	2.16×10^{-14}
GO:0030182	Neuron differentiation	2.49×10^{-15}	4.76×10^{-12}
GO:0051252	Regulation of RNA metabolic process	5.14×10^{-15}	9.96×10^{-12}
Cluster II CGIs			
Sparse sperm methylation, variable ESC, and somatic cell methylation			
GO:0007156	Homophilic cell adhesion	1.64×10^{-9}	3.00×10^{-6}
GO:0016339	Calcium-dependent cell-cell adhesion	9.42×10^{-9}	1.72×10^{-5}
GO:0007155	Cell adhesion	1.83×10^{-8}	3.35×10^{-5}
GO:0022610	Biological adhesion	1.98×10^{-8}	3.62×10^{-5}
GO:0048598	Embryonic morphogenesis	6.95×10^{-7}	1.27×10^{-3}
Cluster III CGIs			
Variable methylation in all five methylomes			
GO:0006468	Protein amino acid phosphorylation	6.89×10^{-6}	0.013
GO:0051056	Regulation of small GTPase mediated signal transduction	1.06×10^{-5}	0.019
GO:0031327	Negative regulation of cellular biosynthetic process	1.40×10^{-5}	0.025
GO:0007010	Cytoskeleton organization	2.62×10^{-5}	0.048
GO:0046578	Regulation of Ras protein signal transduction	2.74×10^{-5}	0.050

Additional significant GO terms for Cluster I islands are shown in **Table S1**. *FDR-corrected for multiple comparisons.

transcription factor binding sites and exhibit evolutionary hypodeamination, indicating that at least some of the CGIs in this cluster harbor true regulatory potential. For example, some intragenic and intergenic CGIs, such as those in the Clusters II and III, may overlap with cryptic promoters and enhancers that function in a highly tissue- and cell type- specific manner.[30,54]

A potential caveat is that two of the methylomes analyzed are from cell lines: embryonic stem cells harvested at passage 41 and a neonatal fibroblast cell lines at passage 13.[22] Consequently, future studies are needed to confirm these findings in a large number of primary cell populations. Nevertheless, CGI clusters identified based upon variation of DNA methylation are distinct at both evolutionary and genomic levels as well as functional levels, suggesting that these clusters represent a true heterogeneity between human CGIs at multiple levels. Long, promoter-associated CGIs are stably hypomethylated across different cell types, while short, intra- or inter-genic CGIs exhibit the most variable patterns DNA methylation. Future studies of CGIs may benefit from explicitly taking into account such heterogeneity.

For example, the majority of cancer-implicated CGIs are Cluster I CGIs, implying that methylation of these CGIs is likely to be detrimental and particularly disease-prone. It is interesting to note that an additional 14 cancer-implicated CGIs are found in Cluster II, which is hypomethylated in sperm. Similarly, the majority of cancer/testis antigen associated CGIs are found in Cluster II. In addition, "aging" CGIs are also significantly more enriched in Cluster II. The co-occurrence of aging- and tissue-specific DMRs in Cluster II suggest that these two aspects may share common molecular mechanisms. For example stochastic variation of DNA methylation across cell divisions may account for the observed variation of DNA methylation due to aging and across tissues, particularly for some short intragenic CGIs.

The distribution of the three cluster CGIs also provide practical insights into epigenetic surveys. For example, the widely used Infinium human methylation chip ("Illumina 450K chip") includes 136 000 positions in CGIs (7.2% of all CpG island CpG sites in the human genome), including 7.2, 5.8, and 10% of CpG sites belonging to Clusters I, II, and III. In light of our findings that Cluster II CGIs often exhibit the most pronounced variation of DNA methylation, this array may be less ideal than other methods to specifically investigate variation of DNA methylation. Also, this array targets a significantly higher proportion of CGIs in Cluster III, which includes a substantial number of "pseudo" evolutionary CGIs as designated by Cohen et al.[21] Consequently, some of the epigenomic variation detected by these positions may lack true regulatory meanings. Our study thus will help future studies to examine variation of DNA methylation across different biological conditions.

Materials and Methods

Whole genome methylomes

In this study we used whole genome, nucleotide-resolution DNA methylation maps (methylomes) of humans. We focused

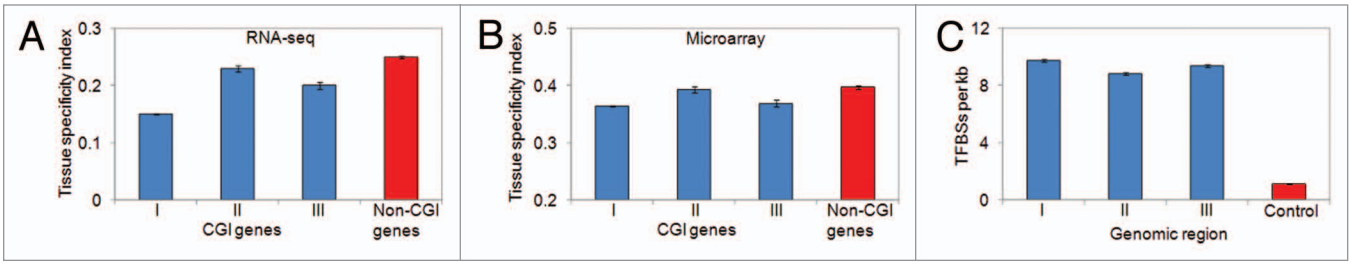

Figure 5. Contrasting expression patterns and transcription factor binding sites of the three CpG island clusters. Tissue specificity gene expression indices based upon RNA-seq (**A**) and microarray (**B**) data are shown for the CGIs genes (blue bars) and non-CGI genes (red bar). Cluster II is the most tissue-specific in both data sets, while Cluster I genes are the most broadly expressed. All CGI genes are more broadly expressed than non-CGI genes (red bars). (**C**) The mean numbers of TFBSs (per kb) for each CpG island cluster (blue bars), which are much larger than that from non-CGI genomic regions of similar lengths (control, red bar).

on analyzing normal tissues or primarily tissue derived cell lines, rather than differentiated cell lines or cancer genomes. Primary data consists of methylomes generated from embryonic stem cells (ESCs),[22] neonatal foreskin fibroblasts,[22] peripheral-blood mononuclear cells (PBMCs),[23] the prefrontal cortex region of the brain,[25] and human sperm.[24] The ESC methylome is from cell populations at passage 41, and the fibroblast methylome is from passage 13.[22] These methylomes were all generated with next-generation bisulfite sequencing technology and have similar number of mapped CpG sites, facilitating a direct comparison of CpG island methylation among tissues. As a comparison, we contrasted the whole genome methylation data from the prefrontal cortex to those generated via the reduced representation bisulfite sequencing methods as a part of the ENCODE project from the "BC_Brain_H11058N " cell line. Comparison of these data sets demonstrates that the whole genome methylation sequencing provides a superior coverage of CGIs (**Fig. S1**). We extended our analyses to three additional methylomes: placenta, kidney, and cerebellum.[36] These additional methylomes were generated using the same methods but have lower coverage (**Table 1**).

CpG island annotation and methylation

The annotations of the CpG islands used in this study were downloaded from the UCSC Genome Browser.[55] These CGIs are characterized as being at least 200 bps in length, GC content of 50% or greater, a CpG frequency (observed/expected; [o/e]) of 0.6, and having no repetitive sequences. To estimate the methylation level for each CpG island, we calculated the mean fractional methylation value for all the mapped cytosines within the CpG island. For each mapped cytosine, the fractional methylation value was calculated as: total number of "C" reads / (total number of "C" reads + total number of "T" reads), following previous studies.[5,25,26] We used the Refseq annotations based upon the current version of the human genome (hg18) to determine the overlaps between CGIs and the promoter, intragenic, and intergenic regions. Promoters were defined as regions 1.5 kb upstream and 0.5 kb downstream of the transcription start sites, similar to previous studies (e.g., refs. 25, 56, and 57). Intragenic regions were defined as those encompassing the region from the transcription start site to the transcription end site, and the rest of genomic regions were defined as intergenic regions. In total, we found that 40.2%, 38.9%, and 20.9% of CGIs overlapped with the promoter, intragenic, and intergenic regions, respectively.

Hierarchical clustering analyses

Clustering of CGIs of the five tissues methylome data was performed using a function called "clustergram" in MATLAB. It employs hierarchical clustering with Euclidean distance metric to first cluster the tissues and then cluster the CGIs. Ward linkage was employed to generate both dendograms.

Analyses of gene expression

Gene expression data from 6 human tissues (prefrontal cortex, cerebellum, heart, kidney, liver, and testis) from whole genome RNA sequencing[42] were aligned to the respective genome sequences by the TopHat program.[58] The expression levels were normalized by mean per-base read coverage with unambiguously mapping reads. The samples measured for the same tissue were averaged to represent the expression level for that specific tissue. The second data set was based on the Affymetrix human genome U133A array which was downloaded from Gene Atlas V2 (GSE1133), where the expression level was standardized by MAS5.0 algorithm.[43] We removed disease tissues, leaving only normal tissues in this data. Using these expression values, the "tissue specificity index"[59] was calculated by incorporating information on the maximum expression level among the tissues in each data set as follows:

$$T = \frac{\sum_{j=1}^{n}(1-[\log_2(E_j)/\log_2(E_{\max})])}{n-1}$$

where n is the number of tissues analyzed, E_j the expression level of the gene in the jth tissue and E_{max} the maximum expression level of the gene across the 6 tissues. The higher the tissue specificity index of a gene, the more tissue-specific its expression pattern is. To examine overlaps between CGIs and transcription factor binding sites, we downloaded the location of transcription factor binding sites conserved in the human/mouse/rat alignment from UCSC genome browser. A binding site was considered conserved across the alignment based upon the score threshold computed with the Transfac Matrix Database (v7.0).[60] We then

counted the number of transcription factor binding sites that were completely located within the CGIs.

Evolutionary substitution rates of CGIs

We used Cohen et al.[21]'s evolutionary data downloaded from the Tanay lab website (http://compgenomics.weizmann.ac.il/tanay). The data consisted of a list of bigWig tracks containing observed and expected evolutionary dynamics in 50 bp resolution, smoothed over 2 kb windows. We converted the bigWig encrypted files to bedGraph files using the UCSC utility bigWigToBedGraph. We then computed the weighted average of observed and expected rates for each CpG island region using custom perl scripts.

Discriminant and classification analyses

We performed linear discriminant analyses using the "lda" function from the package of "MASS" in R. We also performed support vector machine analyses using the "ksvm" function from the package of "kernlab" in R. For both analyses, 20% of the whole data were randomly selected as the training data set. After training the model, the predictions were made for the test data set and the accuracy was evaluated based upon the comparison between prediction and the actual label in test data set.

Data Availability

CGI clusters, genomic and evolutionary features and DNA methylation levels across the examinedmethylomes are available in the **Table S2**.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This study is supported by National Science Foundation grants (MCB-0950896 and DEB-0640690). We thank Thomas Keller and Isabel Mendizabal for their comments on the manuscript.

Supplemental Materials

Supplemental materials may be found here:
www.landesbioscience.com/journals/ epigenetics/article/27654

References

1. Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 1980; 8:1499-504; PMID:6253938; http://dx.doi.org/10.1093/nar/8.7.1499
2. Ehrlich M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC, McCune RA, Gehrke C. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res 1982; 10:2709-21; PMID:7079182; http://dx.doi.org/10.1093/nar/10.8.2709
3. Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, Hetzel J, Jain J, Strauss SH, Halpern ME, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 2010; 107:8689-94; PMID:20395551; http://dx.doi.org/10.1073/pnas.1002720107
4. Gama-Sosa MA, Midgett RM, Slagel VA, Githens S, Kuo KC, Gehrke CW, Ehrlich M. Tissue-specific differences in DNA methylation in various mammals. Biochim Biophys Acta 1983; 740:212-9; PMID:6860672; http://dx.doi.org/10.1016/0167-4781(83)90079-9
5. Zemach A, McDaniel IE, Silva P, Zilberman D. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science 2010; 328:916-9; PMID:20395474; http://dx.doi.org/10.1126/science.1186366
6. Suzuki MM, Bird A. DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 2008; 9:465-76; PMID:18463664; http://dx.doi.org/10.1038/nrg2341
7. Bird AP. CpG-rich islands and the function of DNA methylation. Nature 1986; 321:209-13; PMID:2423876; http://dx.doi.org/10.1038/321209a0
8. Bird A, Taggart M, Frommer M, Miller OJ, Macleod D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 1985; 40:91-9; PMID:2981636; http://dx.doi.org/10.1016/0092-8674(85)90312-5
9. Cooper DN, Taggart MH, Bird AP. Unmethylated domains in vertebrate DNA. Nucleic Acids Res 1983; 11:647-58; PMID:6188105; http://dx.doi.org/10.1093/nar/11.3.647
10. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol 1987; 196:261-82; PMID:3656447; http://dx.doi.org/10.1016/0022-2836(87)90689-9
11. Takai D, Jones PA. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A 2002; 99:3740-5; PMID:11891299; http://dx.doi.org/10.1073/pnas.052410099
12. Yi SV, Goodisman MAD. Computational approaches for understanding the evolution of DNA methylation in animals. Epigenetics 2009; 4:551-6; PMID:20009525; http://dx.doi.org/10.4161/epi.4.8.10345
13. Edwards CA, Ferguson-Smith AC. Mechanisms regulating imprinted genes in clusters. Curr Opin Cell Biol 2007; 19:281-9; PMID:17467259; http://dx.doi.org/10.1016/j.ceb.2007.04.013
14. Pfeifer GP, Steigerwald SD, Hansen RS, Gartler SM, Riggs AD. Polymerase chain reaction-aided genomic sequencing of an X chromosome-linked CpG island: methylation patterns suggest clonal inheritance, CpG site autonomy, and an explanation of activity state stability. Proc Natl Acad Sci U S A 1990; 87:8252-6; PMID:2236038; http://dx.doi.org/10.1073/pnas.87.21.8252
15. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature 2007; 447:425-32; PMID:17522676; http://dx.doi.org/10.1038/nature05918
16. Yu D-H, Ware C, Waterland RA, Zhang J, Chen M-H, Gadkari M, Kunde-Ramamoorthy G, Nosavanh LM, Shen L. Developmentally programmed 3′ CpG island methylation confers tissue- and cell-type-specific transcriptional activation. Mol Cell Biol 2013; 33:1845-58; PMID:23459939; http://dx.doi.org/10.1128/MCB.01124-12
17. Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol 2010; 28:1057-68; PMID:20944598; http://dx.doi.org/10.1038/nbt.1685
18. Robertson KD, Wolffe AP. DNA methylation in health and disease. Nat Rev Genet 2000; 1:11-9; PMID:11262868; http://dx.doi.org/10.1038/35049533
19. Elango N, Yi SV. Functional relevance of CpG island length for regulation of gene expression. Genetics 2011; 187:1077-83; PMID:21288871; http://dx.doi.org/10.1534/genetics.110.126094
20. Fenouil R, Cauchy P, Koch F, Descostes N, Cabeza JZ, Innocenti C, Ferrier P, Spicuglia S, Gut M, Gut I, et al. CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. Genome Res 2012; 22:2399-408; PMID:23100115; http://dx.doi.org/10.1101/gr.138776.112
21. Cohen NM, Kenigsberg E, Tanay A. Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. Cell 2011; 145:773-86; PMID:21620139; http://dx.doi.org/10.1016/j.cell.2011.04.024
22. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, Low HM, Kin Sung KW, Rigoutsos I, Loring J, et al. Dynamic changes in the human methylome during differentiation. Genome Res 2010; 20:320-31; PMID:20133333; http://dx.doi.org/10.1101/gr.101907.109
23. Li Y, Zhu J, Tian G, Li N, Li Q, Ye M, Zheng H, Yu J, Wu H, Sun J, et al. The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol 2010; 8:e1000533; PMID:21085693; http://dx.doi.org/10.1371/journal.pbio.1000533
24. Molaro A, Hodges E, Fang F, Song Q, McCombie WR, Hannon GJ, Smith AD. Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell 2011; 146:1029-41; PMID:21925323; http://dx.doi.org/10.1016/j.cell.2011.08.016
25. Zeng J, Konopka G, Hunt BG, Preuss TM, Geschwind D, Yi SV. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. Am J Hum Genet 2012; 91:455-65; PMID:22922032; http://dx.doi.org/10.1016/j.ajhg.2012.07.024
26. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 2009; 462:315-22; PMID:19829295; http://dx.doi.org/10.1038/nature08514
27. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, et al. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol 2010; 28:1097-105; PMID:20852635; http://dx.doi.org/10.1038/nbt.1682
28. Antequera F. Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 2003; 60:1647-58; PMID:14504655; http://dx.doi.org/10.1007/s00018-003-3088-6
29. Cooper DN, Krawczak M. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 1989; 83:181-8; PMID:2777259; http://dx.doi.org/10.1007/BF00286715

30. Illingworth R, Kerr A, Desousa D, Jørgensen H, Ellis P, Stalker J, Jackson D, Clee C, Plumb R, Rogers J, et al. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biol 2008; 6:e22; PMID:18232738; http://dx.doi.org/10.1371/journal.pbio.0060022
31. Illingworth RS, Bird AP. CpG islands--'a rough guide'. FEBS Lett 2009; 583:1713-20; PMID:19376112; http://dx.doi.org/10.1016/j.febslet.2009.04.012
32. Monk M, Boubelik M, Lehnert S. Temporal and regional changes in DNA methylation in the embryonic, extraembryonic and germ cell lineages during mouse embryo development. Development 1987; 99:371-82; PMID:3653008
33. Zhang X-Y, Wang RYH, Ehrlich M. Human DNA sequences exhibiting gamete-specific hypomethylation. Nucleic Acids Res 1985; 13:4837-51; PMID:4022775; http://dx.doi.org/10.1093/nar/13.13.4837
34. Brock GJR, Charlton J, Bird A. Densely methylated sequences that are preferentially localized at telomere-proximal regions of human chromosomes. Gene 1999; 240:269-77; PMID:10580146; http://dx.doi.org/10.1016/S0378-1119(99)00442-4
35. Kochanek S, Renz D, Doerfler W. DNA methylation in the Alu sequences of diploid and haploid primary human cells. EMBO J 1993; 12:1141-51; PMID:8384552
36. Schroeder DI, Blair JD, Lott P, Yu HOK, Hong D, Crary F, et al. The human placenta methylome. Proceedings of the National Academy of Sciences 2013.
37. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M; ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012; 489:57-74; PMID:22955616; http://dx.doi.org/10.1038/nature11247
38. Peng Z, Elango N, Wildman DE, Yi SV. Primate phylogenomics: developing numerous nuclear non-coding, non-repetitive markers for ecological and phylogenetic applications and analysis of evolutionary rate variation. BMC Genomics 2009; 10:247; PMID:19470178; http://dx.doi.org/10.1186/1471-2164-10-247
39. Martin DIK, Singer M, Dhahbi J, Mao G, Zhang L, Schroth GP, Pachter L, Boffelli D. Phyloepigenomic comparison of great apes reveals a correlation between somatic and germline methylation states. Genome Res 2011; 21:2049-57; PMID:21908772; http://dx.doi.org/10.1101/gr.122721.111
40. Suzuki MM, Yoshinari A, Obara M, Takuno S, Shigenobu S, Sasakura Y, Kerr AR, Webb S, Bird A, Nakayama A. Identical sets of methylated and nonmethylated genes in Ciona intestinalis sperm and muscle cells. Epigenetics Chromatin 2013; 6:38; PMID:24279449; http://dx.doi.org/10.1186/1756-8935-6-38
41. Razin A, Cedar H. DNA methylation and gene expression. Microbiol Rev 1991; 55:451-8; PMID:1943996
42. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, et al. The evolution of gene expression levels in mammalian organs. Nature 2011; 478:343-8; PMID:22012392; http://dx.doi.org/10.1038/nature10532
43. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 2004; 101:6062-7; PMID:15075390; http://dx.doi.org/10.1073/pnas.0400782101
44. Landolin JM, Johnson DS, Trinklein ND, Aldred SF, Medina C, Shulha H, Weng Z, Myers RM. Sequence features that drive human promoter function and tissue specificity. Genome Res 2010; 20:890-8; PMID:20501695; http://dx.doi.org/10.1101/gr.100370.109
45. Sproul D, Kitchen RR, Nestor CE, Dixon JM, Sims AH, Harrison DJ, Ramsahoye BH, Meehan RR. Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. Genome Biol 2012; 13:R84; PMID:23034185; http://dx.doi.org/10.1186/gb-2012-13-10-r84
46. Whitehurst AW. Cause and consequence of cancer/testis antigen activation in cancer. Annu Rev Pharmacol Toxicol 2014; 54:251-72; PMID:24160706; http://dx.doi.org/10.1146/annurev-pharmtox-011112-140326
47. Scanlan MJ, Gure AO, Jungbluth AA, Old LJ, Chen YT. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol Rev 2002; 188:22-32; PMID:12445278; http://dx.doi.org/10.1034/j.1600-065X.2002.18803.x
48. Akers SN, Odunsi K, Karpf AR. Regulation of cancer germline antigen gene expression: implications for cancer immunotherapy. Future Oncol 2010; 6:717-32; PMID:20465387; http://dx.doi.org/10.2217/fon.10.36
49. Cheng YH, Wong EW, Cheng CY. Cancer/testis (CT) antigens, carcinogenesis and spermatogenesis. Spermatogenesis 2011; 1:209-20; PMID:22319669; http://dx.doi.org/10.4161/spmg.1.3.17990
50. Ehrlich M. DNA methylation in cancer: too much, but also too little. Oncogene 2002; 21:5400-13; PMID:12154403; http://dx.doi.org/10.1038/sj.onc.1205651
51. Kim R, Kulkarni P, Hannenhalli S. Derepression of Cancer/testis antigens in cancer is associated with distinct patterns of DNA hypomethylation. BMC Cancer 2013; 13:144; PMID:23522060; http://dx.doi.org/10.1186/1471-2407-13-144
52. Scanlan MJ, Simpson AJG, Old LJ. The cancer/testis genes: review, standardization, and commentary. Cancer Immun 2004; 4:1; PMID:14738373
53. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, Diez J, Sanchez-Mut JV, Setien F, Carmona FJ, et al. Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci U S A 2012; 109:10522-7; PMID:22689993; http://dx.doi.org/10.1073/pnas.1120658109
54. Illingworth RS, Gruenewald-Schneider U, Webb S, Kerr ARW, James KD, Turner DJ, Smith C, Harrison DJ, Andrews R, Bird AP. Orphan CpG islands identify numerous conserved promoters in the mammalian genome. PLoS Genet 2010; 6:e1001134; PMID:20885785; http://dx.doi.org/10.1371/journal.pgen.1001134
55. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, et al. The UCSC Genome Browser Database: 2008 update. Nucleic Acids Res 2008; 36:D773-9; PMID:18086701; http://dx.doi.org/10.1093/nar/gkm966
56. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, Rebhan M, Schübeler D. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 2007; 39:457-66; PMID:17334365; http://dx.doi.org/10.1038/ng1990
57. Montgomery SB, Griffith OL, Schuetz JM, Brooks-Wilson A, Jones SJM. A survey of genomic properties for the detection of regulatory polymorphisms. PLoS Comput Biol 2007; 3:e106; PMID:17559298; http://dx.doi.org/10.1371/journal.pcbi.0030106
58. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25:1105-11; http://dx.doi.org.10.1093/bioinformatics/btp120
59. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 2005; 21:650-9; PMID:15388519; http://dx.doi.org/10.1093/bioinformatics/bti042
60. Matys V, Fricke E, Geffers R, Gößling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, et al. TRANSFAC®: transcriptional regulation, from patterns to profiles. Nucleic Acids Research 2003; 31:374-8; http://dx.doi.org.10.1093/nar/gkg108