CpG Island Methylation in Human Lymphocytes Is Highly Correlated with DNA Sequence, Repeats, and Predicted DNA Structure

Christoph Bock[1*], Martina Paulsen[2], Sascha Tierling[2], Thomas Mikeska[2], Thomas Lengauer[1], Jörn Walter[2]

1 Max-Planck-Institut für Informatik, Saarbrücken, Germany, 2 Universität des Saarlandes, Genetik/Epigenetik, Saarbrücken, Germany

CpG island methylation plays an important role in epigenetic gene control during mammalian development and is frequently altered in disease situations such as cancer. The majority of CpG islands is normally unmethylated, but a sizeable fraction is prone to become methylated in various cell types and pathological situations. The goal of this study is to show that a computational epigenetics approach can discriminate between CpG islands that are prone to methylation from those that remain unmethylated. We develop a bioinformatics scoring and prediction method on the basis of a set of 1,184 DNA attributes, which refer to sequence, repeats, predicted structure, CpG islands, genes, predicted binding sites, conservation, and single nucleotide polymorphisms. These attributes are scored on 132 CpG islands across the entire human Chromosome 21, whose methylation status was previously established for normal human lymphocytes. Our results show that three groups of DNA attributes, namely certain sequence patterns, specific DNA repeats, and a particular DNA structure, are each highly correlated with CpG island methylation (correlation coefficients of 0.64, 0.66, and 0.49, respectively). We predicted, and subsequently experimentally examined 12 CpG islands from human Chromosome 21 with unknown methylation patterns and found more than 90% of our predictions to be correct. In addition, we applied our prediction method to analyzing Human Epigenome Project methylation data on human Chromosome 6 and again observed high prediction accuracy. In summary, our results suggest that DNA composition of CpG islands (sequence, repeats, and structure) plays a significant role in predisposing CpG islands for DNA methylation. This finding may have a strong impact on our understanding of changes in CpG island methylation in development and disease.

Citation: Bock C, Paulsen M, Tierling S, Mikeska T, Lengauer T, et al. (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. PLoS Genet 2(3): e26.

Introduction

DNA methylation is a frequent biochemical modification of eukaryotic DNA [1–6]. In humans, it affects the C5 position of cytosines that belong to CpG dinucleotides (i.e., a cytosine directly followed by a guanine).

CpG dinucleotides are distributed unevenly across the human genome. In non-coding DNA, CpG dinucleotides are 4-fold underrepresented compared to the frequency of the other dinucleotides [7,8], with the remarkable exception of so-called CpG islands. There, CpGs are approximately as frequent as one would expect from single base pair frequencies. For practical reasons, CpG islands are usually defined as sequence stretches that fulfill three conditions [9]: (i) GC content above 50%, (ii) ratio of observed versus expected number of CpG dinucleotides above 0.6, and (iii) more than n base pairs in length (we use $n = 400$ in accordance with the source of our dataset [10]). CpG islands rarely exceed 5,000 base pairs and are often associated with functional elements. In particular, CpG islands overlap with the promoter regions of 50% to 60% of human genes, including most housekeeping genes [11,12].

As DNA methylation in the human genome is largely confined to CpG dinucleotides, it is not surprising that the distribution of DNA methylation along the genome is closely intertwined with CpG frequencies. The classical view is that almost all dispersed CpG dinucleotides in the human genome are methylated by default, whereas CpG dinucleotides inside CpG island promoters are typically unmethylated in normal (i.e., non-neoplasic, non-senescent) tissue [1]. However, exceptions have been known for a long time, such as de novo methylation during cell differentiation [13], imprinting [3], and X-chromosome inactivation [14]. Strong biallelic DNA methylation of CpG island promoters is associated with stable silencing of neighboring or associated genes and constitutes a frequent event in cancer progression [15].

Initial chromosome-wide and genome-wide studies of CpG island methylation indicate that a sizeable fraction of CpG islands is methylated in normal tissue [10,16]. However, little is known about the mechanisms that lead to methylation of certain CpG islands while leaving others unmethylated, and it is unclear whether these two groups can be identified by characteristic attributes. Inspired by recent exploratory

Editor: Wolf Reik, The Babraham Institute, United Kingdom

Received December 7, 2005; **Accepted** January 20, 2006; **Published** March 3, 2006

A previous version of this article appeared as an Early Online Release on January 20, 2006 (DOI: 10.1371/journal.pgen.0020026.eor).

DOI: 10.1371/journal.pgen.0020026

Abbreviations: HEP, Human Epigenome Project; kb, kilobase; SNP, single nucleotide polymorphism; SVM, support vector machine

* To whom correspondence should be addressed. E-mail: cbock@mpi-inf.mpg.de

Synopsis

DNA methylation is the only epigenetic mechanism in eukaryotes that is known to directly modify the DNA. It plays an important role for gene control during development and cell differentiation, and it is a promising therapeutic target in cancer research. While a genome-wide picture of DNA methylation patterns is currently emerging, we have only fragmentary knowledge about the linkage between DNA methylation and other genomic attributes such as DNA sequence and structure, repetitive elements, or sequence conservation. The authors fill this gap by reporting on a comprehensive bioinformatical analysis of DNA methylation on human Chromosome 21—and in part, extending to other regions of the human genome. They report new associations that will help elucidate the functions of DNA methylation along the human genome. Furthermore, the authors show that their findings can be applied to predicting DNA methylation patterns from genome sequence. Such predictions have the potential of speeding up genome-wide epigenetic profiling: It may be possible to focus experimental resources on a few selected areas while bioinformatics procedures are applied to the bulk of the genome.

results pointing towards a significant role of local DNA sequences in predetermining DNA methylation at the nucleotide level (i.e., CpG dinucleotides instead of CpG islands) [17,18], as well as for aberrant methylation [19], we performed a comprehensive analysis of the association of DNA-related features and normal CpG island methylation on human Chromosome 21. Our results show that DNA sequence patterns, repeat frequencies, and predicted DNA structure are highly correlated with CpG island methylation. We successfully used this association to predict the methylation status for new CpG islands.

Results

In this study, we explore the relationship between DNA methylation and various DNA-related features at the biologically functional CpG island level, using computational epigenetics methodology. Based on a dataset published by Yamada et al., comprising all CpG islands on the non-repetitive parts of human Chromosome 21 [10] and a compiled list of 1,184 DNA-related attributes, we quantify the correlation between CpG island methylation and eight attribute classes: (1) DNA sequence properties and patterns, (2) repeat frequency and distribution, (3) CpG island frequency and distribution, (4) predicted DNA structure, (5) gene and exon distribution, (6) predicted transcription factor binding sites, (7) evolutionary conservation, and (8) single nucleotide polymorphisms (SNPs). We identify the attributes that are most predictive in distinguishing between methylated and unmethylated CpG islands and we show that it is possible to predict CpG island methylation from DNA-related features with high accuracy. Finally, we validate our results both experimentally on Chromosome 21 and bioinformatically on data from the Human Epigenome Project (HEP) [20].

Identification of DNA-Related Attributes That Distinguish Methylated CpG Islands from Their Unmethylated Counterparts

As a first step towards understanding the relationship between DNA-related attributes and CpG island methylation, we statistically compared the distributions between methylated and unmethylated CpG islands for all attributes in our list (see Dataset S1 for the full list of p-values and Materials and Methods for an overview of attribute definitions). Using a conservative significance threshold, 41 attributes showed significant differences (Table 1).

Of the significant attributes, the majority are frequencies of GC-rich and CpG-rich DNA sequence patterns, which are over-represented in unmethylated CpG islands. Non-strand-specific patterns and patterns that are strand-specific relative to the chromosomal plus-strand occur with similar frequency and composition. Several attributes that refer to repetitive DNA are more frequent in methylated CpG islands (such as segmental duplications, self chain alignments, and tandem repeats).

Interestingly, two aspects of predicted DNA structure, most prominently the average rise of the DNA helix, also show different distributions for methylated and unmethylated CpG islands (see Olson et al. [21] for an overview of DNA structure nomenclature). The role of predicted DNA structure becomes even more pronounced when considering not only the CpG island itself, but also the −20-kilobase (kb) to +20-kb sequence windows surrounding it. In that case, the predicted average rise and the predicted average twist are the second and third most significant among all attributes (Dataset S1, second worksheet). An inspection of the corresponding boxplots (Figure 1) shows that the predicted DNA rise increases on average within CpG islands compared to the genomic neighborhood, whereas the twist decreases. However, this effect is much stronger for methylated than for unmethylated CpG islands. Hence, methylated CpG islands tend to co-locate with areas of unusual predicted DNA structure.

Furthermore, it is apparent from Table 1 that a single pattern is over-represented in methylated CpG islands, namely the non-strand-specific CACC/GGTG pattern. Because this pattern contains a TpG, in contrast to the CpG-rich patterns that are frequent in unmethylated CpG islands, it is tempting to argue that this pattern may be the result of sporadic deamination of original $GG^{M}CG$ patterns (such mutations are less likely to be repaired for methylated CpGs [6]). In order to test whether differential CpG → TpG mutation rates may be a source of differential pattern frequencies between methylated and unmethylated CpG islands, we compared the palindromic pattern CGCG with the non-strand-specific pattern TGTG/CACA, which can evolve from the former pattern by two subsequent deamination mutations.

In agreement with our hypothesis, we find the CGCG pattern more frequently in unmethylated CpG islands (mean of 12.61 occurrences per kb) compared to methylated CpG islands (7.15 occurrences per kb) and the TGTG/CACA pattern more frequently in methylated CpG islands (10.92 occurrences per kb) compared to unmethylated CpG islands (2.93 occurrences per kb). In both cases, p-values were below 0.001 according to a Wilcoxon test. These results suggest that during evolution, higher rates of germline CpG → TpG mutation occurred in those CpG islands that are methylated in human lymphocytes compared to those that are unmethylated.

Finally, we analyzed the dataset for evidence of experimental bias. Because restriction enzyme digestion was used to discriminate between methylated and unmethylated CpG islands [10], incomplete digestion is a potential error source. In this case, we would expect the HpaII recognition site

Table 1. DNA-Related Attributes Differ Significantly between Methylated and Unmethylated CpG Islands

Rank	Attribute Name	Attribute Description	Attribute Class	Higher Value for	Single Test Significance
1	SAl_len	Total length of self-alignments (alignments of the human genome against itself)	(2)	Methylated CpG Islands	2.62×10^{-11}
2	SAl_no	Total number of self-alignments	(2)	Methylated CpG Islands	3.23×10^{-10}
3	Pat_CCGC	Chromosome plus-strand pattern frequency of CCGC	(1)	Unmethylated CpG Islands	5.18×10^{-10}
4	Pat_CCCC	Chromosome plus-strand pattern frequency of CCCC	(1)	Unmethylated CpG Islands	1.39×10^{-9}
5	SAl_std	Standard deviation of self-alignment lengths	(2)	Methylated CpG Islands	1.96×10^{-9}
6	Uni_AAAG	Non-strand-specific pattern frequency of AAAG/CTTT	(1)	Unmethylated CpG Islands	8.87×10^{-9}
7	fC_std	Standard deviation of C content distribution	(1)	Unmethylated CpG Islands	9.13×10^{-9}
8	Rise_avg	Average DNA structure rise (as predicted from sequence)	(4)	Methylated CpG Islands	3.82×10^{-8}
9	Pat_CGCC	Chromosome plus-strand pattern frequency of CGCC	(1)	Unmethylated CpG Islands	5.05×10^{-8}
10	Pat_AAAG	Chromosome plus-strand pattern frequency of AAAG	(1)	Unmethylated CpG Islands	7.72×10^{-8}
11	Roll_skew	Skewness of DNA structure roll distribution (as predicted from sequence)	(4)	Unmethylated CpG Islands	1.15×10^{-7}
12	Pat_CTCC	Chromosome plus-strand pattern frequency of CTCC	(1)	Unmethylated CpG Islands	1.46×10^{-7}
13	fCG_std	Standard deviation of CpG content distribution	(1)	Unmethylated CpG Islands	2.15×10^{-7}
14	Pat_TCCC	Chromosome plus-strand pattern frequency of TCCC	(1)	Unmethylated CpG Islands	2.57×10^{-7}
15	SDu_no	Total number of sequential duplications	(2)	Methylated CpG Islands	3.49×10^{-7}
16	SAl_sco	Average self-alignment score	(2)	Methylated CpG Islands	4.19×10^{-7}
17	Pat_CTTT	Chromosome plus-strand pattern frequency of CTTT	(1)	Unmethylated CpG Islands	4.23×10^{-7}
18	Uni_CGGA	Non-strand-specific pattern frequency of CGGA/TCCG	(1)	Unmethylated CpG Islands	5.15×10^{-7}
19	Uni_CCGC	Non-strand-specific pattern frequency of CCGC/GCGG	(1)	Unmethylated CpG Islands	9.08×10^{-7}
20	Pat_CGGA	Chromosome plus-strand pattern frequency of CGGA	(1)	Unmethylated CpG Islands	1.16×10^{-6}
21	Pat_GCCG	Chromosome plus-strand pattern frequency of GCCG	(1)	Unmethylated CpG Islands	1.46×10^{-6}
22	Uni_AAGG	Non-strand-specific pattern frequency of AAGG/CCTT	(1)	Unmethylated CpG Islands	1.58×10^{-6}
23	Pat_CCCG	Chromosome plus-strand pattern frequency of CCCG	(1)	Unmethylated CpG Islands	1.86×10^{-6}
24	SAl_avg	Average length of self-alignments	(2)	Methylated CpG Islands	1.91×10^{-6}
25	Pat_TCCG	Chromosome plus-strand pattern frequency of TCCG	(1)	Unmethylated CpG Islands	2.60×10^{-6}
26	Pat_CGCG	Chromosome plus-strand pattern frequency of CGCG	(1)	Unmethylated CpG Islands	2.65×10^{-6}
27	Uni_CGCG	Non-strand-specific pattern frequency of CGCG/CGCG	(1)	Unmethylated CpG Islands	2.65×10^{-6}
28	Pat_ACCC	Chromosome plus-strand pattern frequency of ACCC	(1)	Unmethylated CpG Islands	2.87×10^{-6}
29	Uni_CAAA	Non-strand-specific pattern frequency of CAAA/TTTG	(1)	Unmethylated CpG Islands	2.90×10^{-6}
30	Pat_CAAA	Chromosome plus-strand pattern frequency of CAAA	(1)	Unmethylated CpG Islands	3.01×10^{-6}
31	Uni_CGGC	Non-strand-specific pattern frequency of CGGC/GCCG	(1)	Unmethylated CpG Islands	3.46×10^{-6}
32	Pat_GCCC	Chromosome plus-strand pattern frequency of GCCC	(1)	Unmethylated CpG Islands	3.95×10^{-6}
33	Pat_GGAA	Chromosome plus-strand pattern frequency of GGAA	(1)	Unmethylated CpG Islands	5.93×10^{-6}
34	Pat_TATT	Chromosome plus-strand pattern frequency of TATT	(1)	Unmethylated CpG Islands	6.43×10^{-6}
35	Pat_CCGG	Chromosome plus-strand pattern frequency of CCGG	(1)	Unmethylated CpG Islands	7.21×10^{-6}
36	Uni_CCGG	Non-strand-specific pattern frequency of CCGG/CCGG	(1)	Unmethylated CpG Islands	7.21×10^{-6}
37	Tan_sco	Goodness of fit score of tandem repeats	(2)	Methylated CpG Islands	9.34×10^{-6}
38	Uni_CACC	Non-strand-specific pattern frequency of CACC/GGTG	(1)	Methylated CpG Islands	9.55×10^{-6}
39	Tan_avg	Average lengths of tandem repeats	(2)	Methylated CpG Islands	9.60×10^{-6}
40	RC1_Low_	Alignment score of low complexity class repeats	(2)	Unmethylated CpG Islands	1.37×10^{-5}
41	RF1_Low_	Alignment score of low complexity family repeats	(2)	Unmethylated CpG Islands	1.37×10^{-5}

This table lists all attributes with significantly different distribution among methylated and unmethylated CpG islands, respectively, according to a Wilcoxon test with Bonferroni correction for multiple testing and an overall significance threshold of 1%. The rightmost column displays single-test *p*-values, the significance threshold after multiple testing correction is $0.01/706 = 1.42 \times 10^{-5}$. Attributes with significantly higher values in fully methylated CpG are in green. Attributes in red are significantly higher in unmethylated CpGs. Detailed information on attribute definitions is given in Table S1.

DOI: 10.1371/journal.pgen.0020026.t001

(CCGG) to behave significantly different from patterns that are never cut. However, we observe no indication of this in our attribute statistics (Dataset S1, first worksheet). Five out of ten GC-rich and CpG-containing sequence patterns have higher *p*-values than the CCGG pattern (CCGC, CGCC, GCCG, CCCG, and CGCG), while the same number of patterns has a lower *p*-value (GCGC, CGGC, GCGG, CGGG, and GGCG). We conclude that the experimental method that was applied by Yamada et al. is sufficiently unbiased for our analysis.

Quantification of the Association between DNA-Related Attributes and CpG Island Methylation

Strikingly, all attributes that were significantly different between methylated and unmethylated CpG islands (Table 1) fall into three (out of eight) attribute classes: (1) DNA sequence properties and patterns, (2) repeat frequency and distribution, and (4) predicted DNA structure. In order to investigate this observation more systematically, we calculated the group-wise correlation between CpG island methylation and each of the eight attribute classes.

In contrast to single-attribute correlation coefficients, group-wise correlations are able to capture combined effects of interacting attributes (e.g., when neither A nor B has any significant impact on methylation alone, but the combined presence of both is highly associated with a certain methylation status). Support vector machines (SVMs) have been successfully employed to detect such combined effects. Therefore, we trained a (linear) SVM to predict CpG island methylation and tested its performance on unseen data (10-fold cross-validation). Then, we calculated the correlation coefficient between the SVM's predictions on unseen data

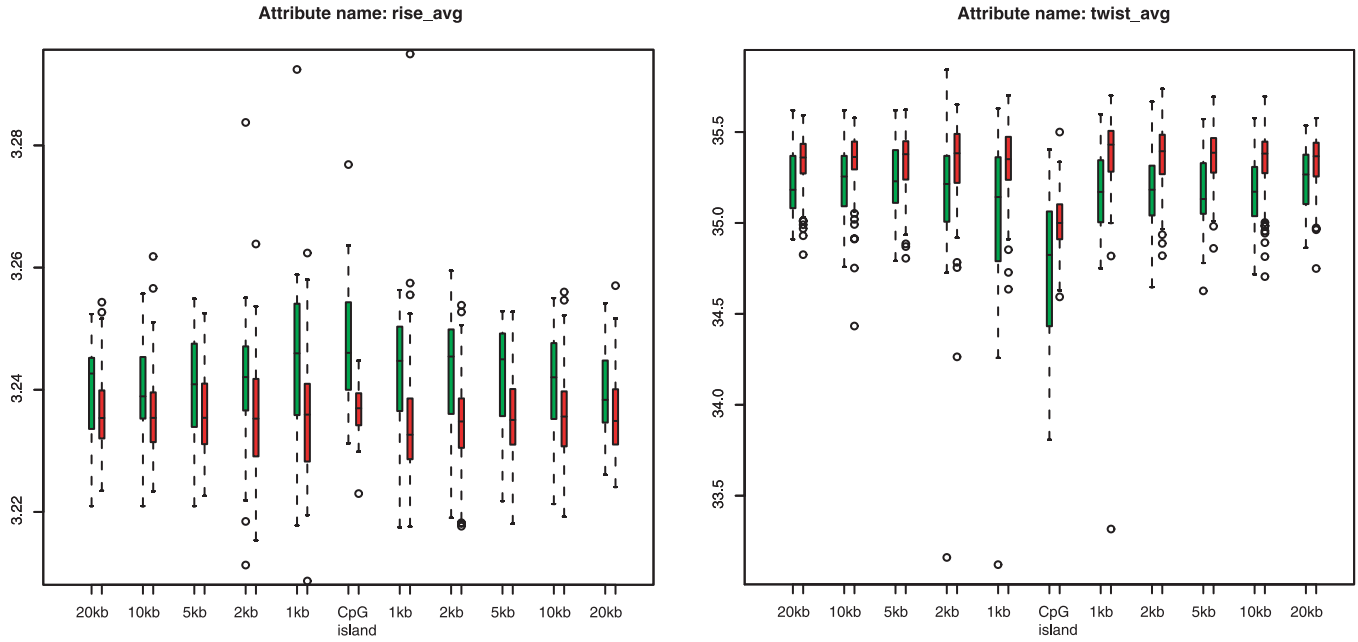

Figure 1. Predicted DNA Structure Differs in the Neighborhood of Methylated CpG Islands Compared with Their Unmethylated Counterparts

The diagram on the left shows boxplots of the predicted DNA rise distribution over the CpG island and the ten sequence windows from −20 kb to 20 kb surrounding the CpG island (averaged over all 132 CpG islands in the Chromosome 21 dataset). Green bars (left) correspond to methylated CpG islands, red bars (right) to unmethylated CpG islands. The diagram on the right shows similar information for the predicted DNA twist.

DOI: 10.1371/journal.pgen.0020026.g001

and the correct values, averaging over 20 independent cross-validation runs. This measure gives us a conservative estimate for the group-wise correlation between the attribute group and CpG island methylation—conservative because it may well be that the SVM does not capture all information on CpG island methylation that is present in the attribute group, while it is highly unlikely to predict methylation correctly over multiple runs if not enough information is contained in the attributes.

Our results substantiate the observation that three classes of DNA-related attributes are distinctly associated with CpG island methylation status (Table 2, experiments 1 to 8): (1) DNA sequence properties and patterns, as well as (2) repeat frequency and distribution are correlated with CpG island methylation at medium to high rates (correlation coefficient of 0.635 and 0.657, respectively); whereas (4) predicted DNA structure falls behind (0.486). Three of the remaining attribute classes exhibit weak correlation with CpG island methylation, namely (5) gene and exon distribution (0.300), (8) SNPs (0.286), and (3) CpG island frequency and distribution (0.045). (6) Predicted transcription factor binding sites and (7) evolutionary conservation are uncorrelated with CpG island methylation (–0.021 and 0.000, respectively). Furthermore, the combination of all eight attribute classes results in a higher correlation value than any single class (0.740), indicating that at least some attribute classes capture complementary information.

To quantify the degree of complementarity and to find out which attribute classes are positively correlated with DNA methylation only due to indirect or secondary effects, we applied the following strategy. Given two attribute classes, we calculate the correlation for both classes separately and for the combination of both. If the latter is higher than any of the former, we can conclude that the attributes complement each other. Comparing DNA sequence with all other attribute classes reveals that only (2) repeat frequency and distribution and (4) predicted DNA structure give rise to an increased correlation when combined with (1) DNA sequence properties and patterns, by 18.4% and 8.3%, respectively (Table 2, experiments 10 to 16). However, among these three classes, all combinations significantly increase the correlation (Table 2, experiments 10, 12, and 17).

Therefore, we conclude that three attribute classes, namely (1) DNA sequence properties and patterns, (2) repeat frequency and distribution, and (4) predicted DNA structure are correlated with CpG island methylation on their own right (primary effect). The remaining attribute classes are either not correlated with CpG island methylation at all (class 7 evolutionary conservation and class 8 predicted transcription factor binding sites), or their correlations are secondary, explainable by their co-location with certain DNA sequence patterns alone (class 3 CpG island frequency and distribution, class 5 gene and exon distribution, and class 8 SNPs).

Prediction of CpG Island Methylation Status from DNA-Related Attributes

While the previous section was concerned with quantifying the relative contribution of different attribute classes to explaining CpG island methylation, the same methodology can be used to predict the methylation status of new CpG islands. Here we report the prediction performance of our method and we address potential limitations.

Without prior knowledge it is sensible to include all 918 non-zero attributes simultaneously in order to achieve best prediction results. In a 10-fold stratified cross-validation of a linear SVM, which we repeated 20 times with different random partitions, this setup resulted in an average

Table 2. The Predictive Power of Attribute Classes Differs Remarkably; Control Experiments Confirm the Appropriateness of the Prediction Method

ID	Attribute Set	Number of Attributes	Prediction Method	Correlation	Accuracy	TN	FN	FP	TP
1	DNA sequence properties and patterns	426	A (linear SVM)	0.635	0.884	1,994	241	66	339
2	Repeat frequency and distribution	311	A (linear SVM)	0.657	0.890	1,995	225	65	355
3	CpG island frequency and distribution	13	A (linear SVM)	0.045	0.755	1,994	532	116	48
4	Predicted DNA structure	28	A (linear SVM)	0.486	0.844	2,053	406	7	174
5	Gene and exon distribution	52	A (linear SVM)	0.300	0.806	2,044	495	16	85
6	Predicted transcription factor binding sites	68	A (linear SVM)	−0.021	0.779	2,056	580	4	0
7	Evolutionary conservation	10	A (linear SVM)	0.000	0.780	2,060	580	0	0
8	Single nucleotide polymorphisms	10	A (linear SVM)	0.286	0.804	2,030	487	30	93
9	All attributes	918	A (linear SVM)	0.740	0.915	2,027	191	33	389
10	Class 1 (sequence) and class 2 (repeats)	737	A (linear SVM)	0.752	0.919	2,037	191	23	389
11	Class 1 and class 3 (CpG islands)	439	A (linear SVM)	0.626	0.880	1,977	233	83	347
12	Class 1 and class 4 (DNA structure)	454	A (linear SVM)	0.688	0.900	2,024	229	36	351
13	Class 1 and class 5 (genes)	478	A (linear SVM)	0.614	0.877	1,980	244	80	336
14	Class 1 and class 6 (TFBS)	494	A (linear SVM)	0.655	0.890	2,007	238	53	342
15	Class 1 and class 7 (conservation)	436	A (linear SVM)	0.626	0.881	1,989	243	71	337
16	Class 1 and class 8 (SNPs)	436	A (linear SVM)	0.618	0.879	1,988	248	72	332
17	Class 2 (repeats) and class 4 (DNA structure)	339	A (linear SVM)	0.713	0.907	2,020	205	40	375
18	DNA sequence properties and patterns	426	B (RBF-kernel SVM)	0.580	0.869	2,040	327	20	253
19	DNA sequence properties and patterns	426	C (AdaBoost)	0.664	0.892	2,009	233	51	347
20	DNA sequence properties and patterns	426	D (C4.5 trees)	0.566	0.852	1,869	200	191	380
21	DNA sequence properties and patterns	426	E (linear SVM using LIBSVM in R)	0.684	0.898	2,018	226	42	354
22	Transcription start site overlap	1	Heuristic (if TSS overlap: unmethylated, otherwise: throw coin)	0.358	0.788	1,810	310	250	270
23	Empty set	0	Trivial (predict everything as unmethylated)	0.000	0.780	2,060	580	0	0

This table summarizes the prediction experiments that were performed in order to analyze the association between DNA-related attributes and CpG island methylation (1 to 17), plus several control experiments (18 to 23). Each row corresponds to one prediction experiment. The column "Attribute Set" specifies the attributes that were used for prediction, "Number of Attributes" gives the size of the attribute set, and "Prediction Method" summarizes the algorithm used (see Materials and Methods for details). The columns "TN," "FN," "FP," and "TP" give the test set results for true-negatives, false-negatives, false-positives, and true-positives over a 10-fold stratified cross-validation that was repeated 20 times. Correlation and accuracy (the remaining two columns) are calculated in the usual way [30] with the modification that, in the case of correlation, we add 0.0001 to TN, FN, FP, and TP in order to prevent the correlation from being undefined when an algorithm always predicts the same class.

TSS, transcription start site; TFBS, transcription factor binding sites.

DOI: 10.1371/journal.pgen.0020026.t002

correlation of 0.74, a test set accuracy of 91.5%, a specificity of 98.4%, and a sensitivity of 67.1% (Table 2, experiment 9).

In order to test the appropriateness of the prediction method that we used (SVM with linear kernel), we performed several control experiments employing other state-of-the-art machine learning algorithms [22], namely SVM with radial basis function kernel, AdaBoost using tree stumps, C4.5 tree generator, and a second widely used implementation of a linear SVM. The results show that performances of all methods lie within the same range (Table 2, experiments 18 to 21).

Next, we investigated how prediction accuracies vary between CpG islands that are located at different positions relative to their closest annotated gene. For this analysis, we regard a single CpG island as reliably predicted if its prediction is correct in at least 15 out of 20 randomized cross-validations, and we manually assigned each of the 132 CpG islands of the Chromosome 21 dataset to one of the following categories (Dataset S2):

Category 1: Promoter CpG islands, defined as overlapping with the transcription start site of an annotated gene: 80 cases fall into this category, of which 78 are unmethylated.

Category 2: Intragenic CpG islands, defined as overlapping introns and/or exons of an annotated gene, but not the transcription start site: 24 cases fall into this category, of which 12 are unmethylated.

Category 3: Gene-terminal CpG islands, defined as overlapping mainly the last exon and/or the 3′ UTR of an annotated gene: six cases fall into this category, of which one is unmethylated.

Category 4: Intergenic CpG islands, defined as not showing any overlap with an annotated human gene: 22 cases fall into this category, of which 12 are unmethylated.

Our results show that prediction accuracy is highest for promoter CpG islands, where 77 unmethylated cases and one methylated case are predicted correctly in more than 15 out of 20 runs (98% accuracy); the second methylated case is predicted correctly in seven out of 20 runs (Ensembl gene *ENSG00000197597*) and the one remaining unmethylated case is correctly predicted in only three runs (Ensembl gene *ENSG00000160207*). In categories 2, 3, and 4, the number of methylated and unmethylated CpG islands is almost balanced, thus prediction is much more difficult. Nevertheless, prediction accuracies stay high: For intragenic CpG islands, 20 cases are predicted correctly in more than 15 runs (83% accuracy). Among the gene-terminal CpG islands, four cases are predicted correctly in more than 15 runs (67% accuracy), and of all intergenic CpG islands, 18 are correctly predicted in more than 15 runs (82% accuracy).

In conclusion, our method achieves high prediction accuracy for CpG islands from all four categories. Finally, we note that the method significantly outperforms a heuristic prediction which relies on transcriptional start site overlap alone (Table 2, experiment 22), and that the very high specificity of the method (98.4%) facilitates chromosome-

wide screening for methylated CpG islands, giving rise to a low number of false-positives.

Experimental Validation by Bisulphite Sequencing

In order to further substantiate the reliability of our method, we experimentally validated its predictions for 12 CpG islands. To that end, we first predicted the methylation state of all CpG islands from Chromosome 21 that were not part of the original dataset [10]; either because they did not match the strict CpG island criteria imposed by Yamada et al. or because they (marginally) overlap with repetitive DNA. Next, we selected eight CpG islands that were predicted as unmethylated and four CpG islands that were predicted as methylated, and we experimentally determined their methylation status in human peripheral blood by bisulphite sequencing.

Hence, while keeping species (human) and chromosome (21) identical, we varied experimental technique (bisulphite sequencing instead of restriction enzyme digestion), cell type (peripheral blood instead of lymphocytes), sample origin (healthy European female instead of healthy unspecified), and—of course—the CpG island. In the selection of validation CpG islands, we did not stratify for CpG island categories (see previous section) because we wanted to assess the method's overall performance across all categories of CpG islands.

The experimental results (Table 3) show that our prediction was correct in ten out of 11 cases (p-value below 0.01). The 12th case, predicted as methylated, showed an incomplete yet significant methylation of 54%. Hence, our method can predict CpG island methylation with high accuracy on a previously unknown test set.

Comparison with HEP Dataset

The DNA methylation data from the HEP pilot study [20] gives us the opportunity to assess the generality of our method and the transferability of the predictions that we obtain from the Chromosome 21 dataset. A priori, one would not expect a high degree of transferability because the HEP data vary from the Chromosome 21 data that were used to develop the method in several important aspects. First, almost 90% of amplicons for which DNA methylation profiles were established do not fulfill CpG island properties. Second, the HEP did not analyze lymphocytes but a variety of other tissues (adipose, brain, breast, liver, lung, muscle, and prostate). Third, all analyzed sequences belong to the relatively small and exceptional major histocompatibility complex on Chromosome 6.

In order to make the HEP dataset accessible to our method, which works on CpG islands (or DNA stretches of comparable length), we calculated the average CpG dinucleotide methylation for every HEP amplicon, and we defined a threshold to distinguish methylated from unmethylated amplicons (see Materials and Methods for details). Next, we trained our method on the Chromosome 21 dataset and predicted the methylation status of all HEP amplicons, in a similar way as was done for the experimental validation in the previous section. The results show a prediction accuracy that is low but still better than random (correlation = 0.15, accuracy = 74.7%, true-negatives = 10, false-negatives = 16, false-positives = 37, true-positives = 147). Hence, there seems to be a core association between DNA-related features and CpG island methylation that is similar or identical across tissues and genomic locations. This association can be specified further by a comparison of prediction error rates. First, we observe a remarkably low false-negative rate of 10%. In other words, the characteristics that were learned to predict CpG islands as methylated in lymphocytes are to some extent transferable across tissues and genomic locations, giving rise to a low false-negative rate on the HEP dataset. Second, the false-positive rate was 8-fold higher than the corresponding false-negative rate (79%), indicating that it is difficult to transfer the DNA-related characteristics of unmethylated cases between the two datasets.

Next, we analyzed to what degree the prediction performance improves when the method is provided with a more adequate training dataset, i.e., when it is permitted to learn the characteristics that are unique to the HEP dataset. To that end, we trained and evaluated our prediction method in a cross-validation on the HEP dataset using all eight attribute

Table 3. Twelve CpG Islands Were Analyzed Experimentally to Validate Our Predictions

CpG Island Position (NCBI35)	Closest Gene	Method	Number of CpGs	Number of CpGs Analyzed	Methylation	Experimental Result	Prediction
Chr 21, 13331442–13331790	*C21orf 99*	Direct sequencing	14	11	91%	Methylated	Methylated
Chr 21, 13904631–13904830	*ANKRD21*	Direct sequencing	11	6	100%	Methylated	Methylated
Chr 21, 14676951–14678040	*STCH*	Direct sequencing	21	15	0%	Unmethylated	Unmethylated
Chr 21, 18538786–18539754	*CHODL*	Cloning and sequencing (nine clones)	26	26	7%	Unmethylated	Unmethylated
Chr 21, 26866818–26867612	*CYYR1*	Direct sequencing	21	14	0%	Unmethylated	Unmethylated
Chr 21, 29318596–29319405	*USP16*	Direct sequencing	18	13	0%	Unmethylated	Unmethylated
Chr 21, 30892864–30893090	*KRTAP6–2*	Direct sequencing	10	8	100%	Methylated	Methylated
Chr 21, 33836092–33837874	*GART*	Direct sequencing	18	14	0%	Unmethylated	Unmethylated
Chr 21, 38209756–38211197	*KCNJ6*	Direct sequencing	25	12	0%	Unmethylated	Unmethylated
Chr 21, 43461259–43461636	*CRYAA*	Cloning and sequencing (nine clones)	15	15	54%	Incomplete	Methylated
Chr 21, 45117025–45119447	*PTTG1IP*	Cloning and sequencing (five clones)	19	19	2%	Unmethylated	Unmethylated
Chr 21, 45669125–45669487	*C21orf 123*	Direct sequencing	10	7	100%	Methylated	Unmethylated

This table summarizes the results of bisulphite sequencing of 12 selected CpG islands together with our prediction that was based on all attribute sets. In nine cases, bisulphite direct sequencing produced unambiguous results. In the three remaining cases, PCR products were cloned and individual clones were sequenced in order to determine the methylation status.
DOI: 10.1371/journal.pgen.0020026.t003

classes. Taking into account all HEP amplicons, this resulted in a sharp increase in prediction performance, with a correlation of 0.47 and an accuracy of 82.4% (true-negatives = 25.7, false-negatives = 15.6, false-positives = 21.3, true-positives = 147.4, averaged over 20 independent cross-validation runs). A further performance increase was observed when we repeated the analysis on amplicons that do not deviate too strongly from the CpG island characteristic, for which the prediction method was developed. We sorted all amplicons by the ratio of observed versus expected CpG dinucleotide frequency, and ran a separate training and prediction analysis for the top, middle, and bottom 70 cases. Results show a correlation of 0.59 for the top group and 0.73 for the middle group (one third of the amplicons in the top group and none in the middle group fulfill CpG island properties). In contrast, predictions fail for the bottom group (correlation = –0.02) where unmethylated cases are rare (six out of 70), possibly because sample size is too small or because these cases behave more randomly.

These results indicate that our prediction method is also well-suited to predict the average methylation status for sequences that are not necessarily CpG islands, at least when a suitable training set is provided and CpG dinucleotide frequency is not too low.

Finally, because the HEP dataset contains methylation information for seven different tissues it should be possible, in principle, to detect evidence of tissue-specific methylation regulation, for example, binding site patterns of tissue-specific transcription factors. Therefore, one would expect that the prediction performance of our method was higher if trained on data from only one tissue, compared to the combination of all tissues, at least when focusing only on the most tissue-specific amplicons. However, we find no evidence for this in our dataset. Instead, prediction performances for individual tissues closely resemble the average case (unpublished data). There are several possible explanations for the method's failure to learn tissue-specific methylation information from the HEP dataset. On the one hand, tissue-specific methylation may be largely uncorrelated with the sequence-related attributes that we analyzed. On the other hand, the dataset may simply be too small. In fact, only between five and 19 out of 210 amplicons per tissue deviate from the "default" state calculated as the consensus methylation over all tissues.

Discussion

We have shown that CpG island methylation can be predicted from DNA sequence and that we may be able to enhance our understanding of the biology that controls methylation in vivo by predictive bioinformatics analysis. First, we identified DNA-related attributes that discriminate strongly between methylated and unmethylated CpG islands in human lymphocytes. Second, we quantified the correlation of CpG island methylation with eight groups of DNA-related attributes and found DNA sequence patterns, repeat frequencies, and predicted DNA structure to be the key contributors. Third, we developed a machine-learning method that can predict the methylation status of unknown CpG islands and we validated the accuracy and reliability of this method both statistically and experimentally.

Our results raise a number of questions concerning our current view of CpG island methylation. While it is apparent from the attribute statistics (Table 1) that CpG-rich patterns are over-represented in unmethylated CpG islands, we found no evidence of a simple yet accurate relationship between CpG island methylation on the one hand and CpG dinucleotide frequency, observed versus expected ratio, or the density of restriction sites such as CCGG (HpaII) and CGCG (HhaI), on the other hand. Instead, methylated and unmethylated CpG islands each seem to be characterized by a relatively complex combination of the presence or absence of certain sequence motifs and attributes of DNA structure. However, we did not find any evidence of a combinatorial "DNA sequence code" for methylation; hence, we suggest that the individual sequence and structure attributes contribute to the preferred methylation state of a CpG island independently and possibly in an additive way.

For repetitive DNA, the situation is more straightforward: CpG islands that significantly overlap with a tandem repeat or a segmental duplication are methylated in almost all cases, which is in line with the long-known fact that tandem repeats form heterochromatin [23]. Unfortunately, we could not address the influence of retrotransposons such as SINE and LINE elements on CpG island methylation since our dataset [10] only contains CpG islands that are not suppressed by RepeatMasker [24].

Based on these observations, we propose that each CpG island can be assigned a degree of methylation propensity that is encoded in its DNA. This default state is what our prediction method captures. For a relatively small number of CpG islands the default state is overruled by biological processes such as tissue differentiation, X-chromosome inactivation, or imprinting, which enforce a certain methylation state. In normal tissue and on the autosomes, these effects seem to affect only a minority of CpG islands—otherwise the high prediction accuracies that we observe could not be argued. This is consistent with the observation that only around 5% of CpG islands are differentially methylated in a tissue-specific fashion [25], and that the effect of imprinting is even more limited. However, since our prediction analysis was constrained to data from two chromosomes and few tissues, such a far-reaching interpretation of our results has to be taken with care. It will be interesting to see whether CpG islands that consistently deviate from their default methylation state due to monoallelic methylation (imprinting, X-chromosome inactivation) are characterized by a medium degree of methylation propensity or whether the underlying biological processes are so strong that basically every CpG island can become differentially methylated independently of its DNA sequence.

Besides these general observations, five more specific results are worth commenting on. First, in line with earlier observations we find almost all promoter CpG islands unmethylated, but also a significant number of intergenic CpG islands, which are often distant from any annotated gene. Little is known about the functional role of intergenic CpG islands. However, it has been observed that unmethylated CpG islands often co-localize with DNA replication origins [26], and we believe that it would be worthwhile to analyze the functional role of unmethylated intergenic CpG islands experimentally on a large sample. Methylation predictions may help to speed up and guide such an approach.

Second, we found evidence that the default methylation status of many CpG islands may be relatively stable during evolution. By comparing frequencies of the CGCG pattern to its (mutated) counterpart TGTG/CACA (the former is over-represented in unmethylated CpG islands of our dataset whereas the latter is over-represented in methylated CpG islands), we concluded that higher CpG →TpG mutation rates have applied to the CpG islands that we find methylated in human lymphocytes, than to those that we find unmethylated. Given that methylated CpG dinucleotides are more prone to CpG → TpG mutations [6], a straightforward explanation would be to postulate that the methylation status that we observe in our dataset (i) is similar to that found in the germline where mutations become fixed, and (ii) was stable over evolutionary time, so that the observed mutations could accumulate.

Third, our results show that certain aspects of DNA sequence and (predicted) DNA structure such as a high DNA rise and a low DNA twist seem to be associated with methylated CpG islands in vivo. It would be interesting to analyze how these sequence and structure attributes correlate with the in vitro recognition and methylation potential of CpG-rich sequences by mammalian DNA methyltransferases. Some reports suggest that unusual DNA structures (such as repeats and cruciform structures [27]) can lead to increased methylation activity by DNA methyltransferases. Moreover, local transitions between DNA in A-form, B-form, or Z-form may influence the methylation potential of the DNA, and it is tempting to speculate that some of our observed parameters may reflect such local differences in DNA structure formation.

Fourth, differences in error rates when training on the Chromosome 21 dataset and testing on the HEP dataset suggest that DNA-related characteristics identifying consistently methylated CpG islands are robust across tissues and genomic locations while those identifying unmethylated CpG islands are not—and have to be learned specifically for each tissue or genomic location. This interpretation is consistent with the hypothesis that most CpG islands in the human genome can become methylated, and do so if they are not preserved in the unmethylated state by specific (and tissue-dependent) influences, for example, the binding of transcription factors.

Fifth, we believe that our methodology for chromosome-wide correlation analysis and prediction is general enough to yield interesting results for other types of genomic and epigenomic data as well (such as histone modifications, replication origins, and many types of ChIP-on-Chip data). Therefore, we implemented our method as a web service which we will make accessible to interested researchers on a cooperation basis.

In conclusion, an understanding of the exact interplay between DNA-related features and CpG island methylation is likely to be of high practical and theoretical value. On the one hand, a reliable tool for predicting default CpG island methylation status from sequence would be of interest as a reference in cancer epigenetics and beyond. On the other hand, the fact that CpG island methylation is closely interwoven with certain features related to DNA sequence and structure—while minor changes such as SNPs seem to make little difference—may provide a key to uncovering the mechanisms that result in inter-individually similar and reproducible epigenetic reprogramming in the germline and the early embryo.

Materials and Methods

DNA methylation data. This analysis is based on the results of a comprehensive measurement of CpG island methylation on human Chromosome 21 [10]. Briefly, Yamada et al. repeat-masked the chromosome sequence and computationally identified all non-repetitive CpG islands using standard tools and parameters (GC content above 50%, ratio of observed versus expected number of CpG dinucleotides above 0.6, more than 400 base pairs in length). Next, they designed primers for each identified CpG island and extracted corresponding DNA from samples of human peripheral blood lymphocytes. Finally, they determined the methylation status of each CpG island by methylation-specific restriction enzymes (via HpaII-McrBC-PCR). Yamada et al. validated their method by bisulphite sequencing of some CpG islands and concluded that it is highly reliable.

Their dataset comprises the methylation status of 149 CpG islands, each belonging to one of the following categories: fully methylated, unmethylated, incompletely methylated, or compositely/differentially methylated. Exploratory analysis using bisulphite sequencing indicated that the latter two classifications were not always unambiguous (unpublished data); therefore we focused on the two well-defined categories, fully methylated (31 cases) and unmethylated (103 cases). In order to minimize potential error sources, we re-mapped the boundaries of the CpGs islands that were originally used by Yamada et al. to the current human genome sequence (NCBI35) and we excluded two cases (both belonging to the fully methylated class) from the analysis because, in the light of this new mapping, the primers did not pick the intended CpG islands. Therefore, our dataset comprised 132 independent CpG islands, which are distributed relatively evenly across Chromosome 21 (see Dataset S2).

For validation, we also used data from the HEP pilot study [20]. In this study, Rakyan et al. determined the methylation status of 3,273 unique CpG dinucleotides (belonging to 253 amplicons) across seven tissues and one to eight samples per tissue by means of bisulphite direct sequencing. Out of these 253 amplicons, 210 could be mapped unambiguously to the NCBI35 genome version and had at least one measurement for each tissue. For these amplicons, we calculated average CpG dinucleotide methylation levels, both separately for individual tissue types and for all tissues combined. Those amplicons below a (arbitrarily chosen) threshold of 60% methylation were marked as unmethylated and those above this threshold were marked as methylated, resulting in a dataset of 163 "methylated" and 47 "unmethylated" amplicons.

DNA-related attributes. In order to identify DNA sequence-related attributes that are correlated with CpG island methylation, we compiled a comprehensive list of attributes that can be linked directly or indirectly to DNA sequence (the full list is given in Table S1). Most attributes take the form of frequencies or numerical scores, averaged over sequence windows and standardized to a default window size of one kb. They fall into eight biological classes, namely: (1) DNA sequence properties and patterns (428 attributes), (2) repeat frequency and distribution (494 attributes), (3) CpG island frequency and distribution (16 attributes), (4) predicted DNA structure (28 attributes), (5) gene and exon distribution (60 attributes), (6) predicted transcription factor binding sites (135 attributes), (7) evolutionary conservation (ten attributes), and (8) SNPs (13 attributes). The data for most of these attributes were collected from annotation tracks in the UCSC Genome Browser [28]. However, the attributes for class 1 were directly calculated from DNA sequence and the attributes for class 4 were calculated from DNA sequence by averaging over octamers [29] and trimers (J. Greenbaum, personal communication), respectively. We calculated these attributes for each CpG island in our dataset, both for the re-mapped CpG island itself and for 11 sequence windows around the CpG island: −20 kb to −10 kb, −10 kb to −5 kb, −5 kb to −2 kb, −2 kb to −1 kb, −1 kb to left boundary of CpG island, CpG island, right boundary of CpG island to +1 kb, +1 kb to +2 kb, +2 kb to +5 kb, +5 kb to +10 kb, +10 kb to +20 kb. Next, we removed those attributes that were zero in all cases (e.g., binding sites of rare transcription factors), giving us a list of 918 prediction attributes. To simplify the statistical analysis, we also removed attributes that were zero in most, but not all cases. For the CpG island level statistics (see next section), only the 706 attributes with non-zero values in the CpG island window of at least five methylated and five unmethylated cases were retained. For the sequence neighborhood statistics, only the 833 attributes were

retained that had non-zero values in at least five methylated and five unmethylated cases, for at least four out of the 11 sequence windows.

Statistics. We performed statistical tests in order to determine attributes that exhibit significantly different values for fully methylated CpG islands compared to unmethylated CpG islands, at two levels. First, we compared all attributes at the CpG island level using the nonparametric Wilcoxon ranksum test (Dataset S1, first worksheet). Second, we compared all attributes across the complete sequence neighborhood of −20 kb to +20 kb around the CpG island (Dataset S1, second worksheet). To that end, quadratic regression functions were fitted over the attribute values in the 11 sequence windows around the CpG island (see previous section) and we used the ANOVA statistic to assess whether separate fitting for unmethylated versus methylated cases resulted in a significantly decreased error compared to combined fitting (quadratic regression functions were chosen to capture symmetry around the CpG island).

Significance thresholds were adjusted for multiple testing using the highly conservative Bonferroni method. Technically speaking, we controlled the family-wise error rate to be less than 1%.

Prediction. Machine learning methodology was used for two tasks: (i) to quantify the correlation between CpG island methylation and several classes of DNA-related attributes, and (ii) to predict CpG island methylation from the local genomic neighborhood.

The technical procedure is similar in both cases (cross-validation) and is discussed below. However, intention and interpretation differ for the two tasks. Task (ii) is the classical prediction scenario: given a dataset of limited size, we want to train a classifier for predicting CpG island methylation on unknown data and to quantify its expected prediction performance. Therefore, we train the classifier on the full set of 918 attributes, assuming that at least some of these attributes contain information that may be useful for the classifier.

In task (i), the goal is not so much to predict new data but to understand existing data. Here, we use a classifier as a tool to quantify the relationship between an attribute class (e.g., DNA sequence properties or repeats) and CpG island methylation. The rationale behind this is simple: If a classifier can successfully and reliably predict CpG island methylation using only information from one particular attribute class, then the attributes in this class are functionally associated with CpG island methylation and the prediction performance is a measure of the degree of functional association.

The prediction experiments follow essentially the same procedure. Given the list of CpG islands or amplicons and any selection of attributes from our list, a linear SVM is repeatedly trained to predict methylation status based on a 90% subset of cases, and its performance is evaluated on the remaining 10% of unseen cases. Technically speaking, we repeat 10-fold stratified cross-validation 20 times with different random partitions and sum the results on the test set (in terms of true- negatives, false-negatives, false-positives, and true-positives). The prediction performance is measured as the correlation coefficient between the predictions and the correct values on the test set. This criterion is commonly viewed as superior to comparing prediction accuracies because it is not as strongly affected by unbalanced class distributions [30].

For most prediction experiments (prediction setup A in Table 2), we used the linear SVM implementation provided by the WEKA package [31], which is based on the sequential minimal optimization method [32]. Additionally, several control experiments were performed that use different algorithms: an SVM with radial basis function kernel (from WEKA package, prediction setup B), AdaBoost M1 with decision tree stumps as the underlying classifier (from WEKA package, prediction setup C), the C4.5 tree generator (from WEKA package, prediction setup D), and a different implementation of a linear SVM (R implementation of LIBSVM [33], prediction setup E). All algorithms were applied with their suggested standard parameters.

Experimental verification. Predictions were performed using a linear SVM that was trained on the full Chromosome 21 dataset (132 cases) and all attribute classes. Subsequently, we determined the methylation status of 12 selected CpG islands by bisulphite sequencing as follows: initially, we applied direct sequencing of the PCR product to all 12 CpG islands (Table 3). In nine cases, this produced unambiguous results (i.e., very high conversion of CpGs = unmethylated, or almost no conversion = methylated). In the three remaining cases with mixed CG/TG sequencing profiles, PCR products were cloned and individual clones were sequenced in order to determine the methylation status. Average methylation was scored from single clone sequences using the BiQ Analyzer software [34]. Details of the experimental setting and the primers that we used are reported in Protocol S1. Human peripheral blood was obtained with the written consent of the donor.

Supporting Information

Dataset S1. Attribute Statistics

This Excel table reports raw *p*-values and multiple-testing-adjusted significance thresholds for all attribute statistics.

Found at DOI: 10.1371/journal.pgen.0020026.sd001 (388 KB XLS).

Dataset S2. DNA Methylation Data

This Excel table contains a re-mapped and quality-controlled version of DNA methylation data that was originally reported by Yamada et al. [10], as it is used in this study. Furthermore, prediction accuracy and genome browser location are reported for all CpG islands.

Found at DOI: 10.1371/journal.pgen.0020026.sd002 (525 KB XLS).

Protocol S1. Experimental Validation

This PDF document gives details on the experimental protocol that was used to determine the methylation status of the validation CpG islands, including PCR primers.

Found at DOI: 10.1371/journal.pgen.0020026.sd003 (26 KB DOC).

Table S1. Overview of Prediction Attributes

This PDF document reports information on calculation, naming, and reference of all attributes that are used in this study.

Found at DOI: 10.1371/journal.pgen.0020026.st001 (89 KB DOC)

Accession Numbers

The Ensembl database (http://www.ensembl.org/Homo_sapiens/index.html) accession numbers for the genes discussed in this paper are Ensembl gene *(ENSG00000197597)* and Ensembl gene *(ENSG00000160207)*.

Acknowledgments

We would like to thank Jörg Rahnenführer for statistics consulting, Joachim Büch for technical support, Takeshi Ito, Yoichi Yamada, and Vardhman Rakyan for the provision of DNA methylation data, and Eleanor Gardiner as well as Jason Greenbaum for the provision of data on DNA structure properties. Furthermore, we acknowledge the sequencing team at the Max-Planck-Institut für Molekulare Genetik for carrying out bisulphite direct sequencing.

Author contributions. CB conceptualized the study, performed data preparation and analysis, and wrote the manuscript. ST and TM performed and evaluated the bisulphite sequencing experiments. MP and JW contributed to the interpretation of the results, and TL supervised the work. All authors contributed to the writing of the manuscript and have read and approved the manuscript.

Funding. This work was conducted within the context of the EU Network of Excellence "Biosapiens" (LSHG-CT-2003–503265), the EU Network of Excellence "The Epigenome" (LSHG-CT-2004–503433), and the BMBF program "Methylome" (01-GR-0496).

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Bird A (2002) DNA methylation patterns and epigenetic memory. Genes Dev 16: 6–21.
2. Jones PA (1999) The DNA methylation paradox. Trends Genet 15: 34–37.
3. Reik W, Santos F, Dean W (2003) Mammalian epigenomics: Reprogramming the genome for development and therapy. Theriogenology 59: 21–32.
4. Fazzari MJ, Greally JM (2004) Epigenomics: Beyond CpG islands. Nat Rev Genet 5: 446–455.
5. Freitag M, Selker EU (2005) Controlling DNA methylation: Many roads to one modification. Curr Opin Genet Dev 15: 191–199.
6. Caiafa P, Zampieri M (2005) DNA methylation and chromatin structure: The puzzling CpG islands. J Cell Biochem 94: 257–265.
7. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
8. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.

9. Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196: 261–282.
10. Yamada Y, Watanabe H, Miura F, Soejima H, Uchiyama M, et al. (2004) A comprehensive analysis of allelic methylation status of CpG islands on human Chromosome 21q. Genome Res 14: 247–266.
11. Wang Y, Leung FC (2004) An evaluation of new criteria for CpG islands in the human genome as gene markers. Bioinformatics 20: 1170–1177.
12. Larsen F, Gundersen G, Lopez R, Prydz H (1992) CpG islands as gene markers in the human genome. Genomics 13: 1095–1107.
13. Arney KL, Fisher AG (2004) Epigenetic aspects of differentiation. J Cell Sci 117: 4355–4363.
14. Heard E (2004) Recent advances in X-chromosome inactivation. Curr Opin Cell Biol 16: 247–255.
15. Feinberg AP, Tycko B (2004) The history of cancer epigenetics. Nat Rev Cancer 4: 143–153.
16. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, et al. (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37: 853–862.
17. Bhasin M, Zhang H, Reinherz EL, Reche PA (2005) Prediction of methylated CpGs in DNA sequences using a support vector machine. FEBS Lett 579: 4302–4308.
18. Handa V, Jeltsch A (2005) Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyltransferases shape the human epigenome. J Mol Biol 348: 1103–1112.
19. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM (2003) Predicting aberrant CpG island methylation. Proc Natl Acad Sci U S A 100: 12253–12258.
20. Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, et al. (2004) DNA methylation profiling of the human major histocompatibility complex: A pilot study for the Human Epigenome Project. PLoS Biol 2: e405. DOI: 10.1371/journal.pbio.0020405
21. Olson WK, Bansal M, Burley SK, Dickerson RE, Gerstein M, et al. (2001) A standard reference frame for the description of nucleic acid base-pair geometry. J Mol Biol 313: 229–237.
22. Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: Data mining, inference, and prediction. New York: Springer. 533 p.
23. Martienssen RA (2003) Maintenance of heterochromatin by RNA interference of tandem repeats. Nat Genet 35: 213–214.
24. Smit AFA, Hubley R, Green P (1996–2004) RepeatMasker Open-3.0. Available: http://www.repeatmasker.org. Accessed 1 February 2006.
25. Song F, Smith JF, Kimura MT, Morrow AD, Matsuyama T, et al. (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. Proc Natl Acad Sci U S A 102: 3336–3341.
26. Antequera F (2003) Structure, function and evolution of CpG island promoters. Cell Mol Life Sci 60: 1647–1658.
27. Chen X, Mariappan SV, Moyzis RK, Bradbury EM, Gupta G (1998) Hairpin induced slippage and hyper-methylation of the fragile X DNA triplets. J Biomol Struct Dyn 15: 745–756.
28. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, et al. (2003) The UCSC Genome Browser Database. Nucleic Acids Res 31: 51–54.
29. Gardiner EJ, Hunter CA, Packer MJ, Palmer DS, Willett P (2003) Sequence-dependent DNA structure: A database of octamer structural parameters. J Mol Biol 332: 1025–1035.
30. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: An overview. Bioinformatics 16: 412–424.
31. Witten IH, Frank E (2000) Data mining: Practical machine learning tools and techniques with Java implementations. San Francisco: Morgan Kaufmann. 371 p.
32. Platt J (1999) Fast training of support vector machines using sequential minimal optimization. In: Schölkopf B, Burges CJC, Smola AJ, editors. Advances in kernel methods: Support vector learning. Cambridge (Massachusetts): MIT Press. pp. 185–208.
33. Chang CC, Lin CJ (2005) LIBSVM: A library for support vector machines. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed 1 February 2006.
34. Bock C, Reither S, Mikeska T, Paulsen M, Walter J, et al. (2005) BiQ Analyzer: Visualization and quality control for DNA methylation data from bisulfite sequencing. Bioinformatics 21: 4067–4068.