SUPPLEMENTARY METHODS

Classifier Generation and Supervised Learning Methods

We have previously reported on the use of pattern recognition and a supervised learning strategy to develop a predictive rule based on a set of DNA sequence patterns that discriminate methylation-prone (MP) and methylation-resistant (MR) sequences (1). The training set consisted of sequences representing 4kb genomic windows encompassing nine MP and nine MR CpG islands (1). The approach consisted of 3 steps.

A general-purpose pattern recognition algorithm (described in (2,3)) was used to identify common sequence patterns within the training CpG islands. The algorithm treats multiple sequence comparisons as an evolutionary distance problem in which the goal is to derive the most likely ancestor that could have given rise to a set of sequences. This is referred to as the "minimum weight common mutated sequence" (MWCMS) model. Given a set of input sequences, the model seeks to mutate every input sequence to the same a priori unknown sequence using the operations of insertion, deletion, and substitution. Weights are assigned for each operation, and the total weight associated with all mutations is to be minimized. The method allows for non-consecutive sequence strings with various physical constraints (minimum length, distance from center, etc.) and can be set to return the longest (or some preset length) common exact-match sequence string. For this study, the criteria were set to return exact match sequence patterns 7-12 bp in length, allowing N for any base pair substitution, with control on the frequency of such substitutions. This process identified thousands of patterns of 7-12bp in the training sample.

The identified DNA patterns were then input into a machine learning environment consisting of a feature selection algorithm (involving a combinatorial decision tree analysis (4))
coupled with an optimization-based discrete support vector machine classification engine, DAMIP (described in 5-9). DAMIP has several distinct features, including the ability to classify any number of distinct groups, the ability to incorporate heterogeneous types of attributes as input, a high-dimensional data transformation that reduces noise and errors in biological data, the ability to incorporate constraints to limit the misclassification rate, a reserved-judgment region that provides a safeguard against over-training, and successive multi-stage classification capability to handle data points placed in the reserved judgment region. DAMIP has been shown to be a powerful supervised learning approach in predicting various biomedical and bio-behavior phenomena (1,10-12).

The computational design used the ‘wrapper approach’, where the feature selection algorithm was coupled directly to the DAMIP classification module. In this approach, the feature selection algorithm selects a subset of the DNA sequence patterns. An attribute vector is then formed from the frequency of occurrence of these patterns in each CpG island sequence in the training set, which then serve as the attribute input into the DAMIP classification engine. The accuracy of the resulting classifier is estimated via 10-fold cross validation. This process is repeated with the feature selection algorithm selecting a new subset of patterns, and the DAMIP engine returning the corresponding classifier and associated 10-fold cross-validation accuracy. Thus, at each step, guided by the estimated accuracy (as a measure of goodness), the machine ‘learns’ as the feature selection algorithm searches through the space of possible subsets for an improved feature set (i.e. that which provides improved cross-validation accuracy). The process continues until a pre-set maximum number of iterations is reached, or a minimum correct classification percentage is achieved. The solution at termination provides the current-best discriminatory feature set with the associated classification rule returned from the DAMIP
module. In this case, we set the maximum number of iterations at 100,000 and the minimum classification accuracy at 90%. The system returned a classification rule with a set of seven discriminatory patterns (TCCCCNCC, TTTCTNC, TCCNCCNCC, GGAGNAAG, GAGANAAG, GCCACCC, GAGGAGGNNG) that achieved an accuracy of 89% in the 10-fold cross-validation tests. We refer to this classification rule as PatMAn, for Pattern-based Methylation Analysis.

To generate the SUPER-PatMAn classifier, attribute vectors based on the frequencies of the same 7 sequence patterns noted above and SUZ12 binding status (positive for enrichment, negative for enrichment, or no data/uninformative) for each CpG island in the training set were used as direct input into the DAMIP classification engine. To facilitate direct comparison of the classification accuracy resulting from the two sets of discriminatory patterns (7 patterns, versus 7 patterns + SUZ12), the same 9MP and 9MR CpG islands were used as the training set. SUPER-PatMAn achieved an accuracy of 83% in 10-fold cross-validation tests. The PatMAn and SUPER-PatMAn classifiers were then applied to genomic sequences corresponding to CpG islands derived from the entire genome (n=37,530).

**Annotation of CpG islands for SUZ12 binding status.** SUZ12 ChIP-chip data from human embryonic stem cells (13) was obtained from [http://www.ebi.ac.uk/arrayexpress](http://www.ebi.ac.uk/arrayexpress) in the form of raw probe signal intensities for 115 arrays containing a total of 4.5x10^6 60-mers spaced at ~350 bp intervals and covering ~90% of the non-repetitive fraction of the human genome. Custom PERL programs were utilized to log2 transform probe signal intensity ratios, calculate the mean for each microarray, subtract the difference between the mean and zero from each probe value, and sort probes by chromosome and start position. SUZ12 enriched and non-enriched regions...
were identified with a modified version of the PERL-implementation of the ChIPOTle program (14). This analysis calculated chromosome-specific background standard deviation values by analyzing negative probe values since enriched regions in ChIP-chip experiments should have positive probe values. Using a sliding window (1,000bp window, 225bp step), window p-values were calculated, a multiple testing-corrected significance threshold was calculated, significant windows were identified, and adjacent significant windows were collapsed into a single enriched region. Additional criteria filters were implemented that required that a minimum of 2 probes be present per window for the window to be considered informative (although the majority had at least 3), and that 2 probes within the window have values $\geq 2*\text{background standard deviation}$. This analysis identified 4,350 SUZ12-enriched regions of average length = 1,313bp.

To annotate CpG islands for SUZ12 occupancy status, CpG island were extracted from the human genome (UCSC HG17; criteria: length $\geq 500$bp, GC content $\geq 55\%$, and CpG Obs/Exp $\geq 0.65$) and assessed for proximity to enriched and non-enriched SUZ12 regions. CpG islands that were overlapping or within 1kb of a SUZ12-enriched region were considered SUZ12-enriched. Non-enriched CpG islands were then assessed for proximity (within 1kb) to SUZ12 non-enriched regions. Remaining CpG islands that were not within 1kb of either enriched or non-enriched region had insufficient SUZ12 binding data and were labeled as uninformative. These analyses allowed for the annotation of SUZ12 binding data for 93% of CpG islands in the genome. Based on this analysis there were 3,642 SUZ12 (+) CpG islands, 31,238 SUZ12 (-) CpG islands, and the remaining 2,650 had insufficient data.

Similar analyses were performed to determine CTCF occupancy status of CpG islands. Genome-wide CTCF binding has been previously reported (15). Based on criteria established by Kim et al. (15), additional criteria filters were implemented in the ChIPOTle analysis that
required that a minimum of 4 probes per window for the window to be considered informative, and that 4 probes within the window have values $\geq 2.5$*background standard deviation.

**Functional annotation of gene ontology terms.** The Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Database (http://david.abcc.ncifcrf.gov) (16) was utilized to identify gene ontology and functional annotation terms significantly enriched among MP CpG islands. Gene symbols corresponding to 63 CpG islands methylated in at least 2 of 3 independent DNMT1-overexpressing clones was used as input. Program settings were as follows: Homo sapiens, Gene Ontology Molecular Function Term level 2, Protein Domains InterPro name, Functional Categories SP_PIR_Keywords. Data analysis and filtering included the consolidation of redundant terms with retention of that with the lowest p-value and exclusion of categories that were non-informative (e.g. disease mutations, alternative splicing), occurred at less than 5%, or had p-values $>0.01$. 
REFERENCES