Supplementary Methods

Methyl DNA immunoprecipitation (mDIP):

Purified sonicated (500 – 2000 bp average size) or mononucleosomal DNA (10 g) obtained from EBV-transformed human lymphoblast cell lines, primary lymphocytes from fresh blood samples, the colon carcinoma cell line Caco-2, the prostate carcinoma cell line PC3 or freshly dissected colon carcinomas is denatured by heating and immunoprecipitated with 20 l of cell supernatant containing an anti-5-methylcytidine monoclonal antibody (Mayer et al., 2000; Reynaud et al., 1992). This antibody is commercially distributed by EMD Biosciences (U.S.A.), Serotec (U.K.) and Eurogentech (Belgium). Input and bound fractions were separated using protein A Sepharose beads and spinX columns (Maruyama et al., 2002) and extracted by phenol-chloroform and ethanol precipitation. The DNA was then resuspended in 100 l double distilled water, dialyzed and subjected to semi-quantitative PCR from specific gene regions using 2 concentrations (1 or 3 l) of DNA in the presence of ^{32}Pα–dCTP (Amersham). Since we usually precipitated < 1% of the DNA, PCR of the bound fraction was compared to 1/100 dilutions of the input DNA. Human lymphoblast carrier DNA was added to some experiments in order to provide an internal control for the efficiency of precipitation. This was accomplished by analyzing promoter enrichment of the human control genes (methylated) (Hashimshony et al., 2003) as compared to *APRT* or *GAPDH* (unmethylated). It should be noted that mDIP enrichment is dependent on the density of methylated CpG residues. For this reason, the *HSVtk* gene appears highly enriched, while the *CRYAA* promoter, which *has* a CpG density of 3% is only enriched 5-8 fold. PCR primer sequences are available upon request.

To test the sensitivity of this assay, a plasmid containing the human *KOC* gene was methylated *in vitro* with either *Hpa*II methylase, *Hha*I methylase, or both; cut with *Ita*I and *Hin*fI; mixed with sonicated mouse DNA; and subjected to mDIP and PCR using specific primers. Because this region contains one *Hpa*II and two *Hha*I sites, we were able to carry out mDIP on the same DNA fragment harboring either zero, one, two or three methyl groups. The mDIP

results of each precipitation were normalized by comparison with another fragment in the same plasmid that has no sites for either *Hpa*II or *Hha*I.

mDIP microarray analysis:

Sonicated chromatin was immunoprecipitated with anti-5mC antibody and the bound fraction isolated by protein A column chromatography. Bound or input DNA fractions were labeled with cy-3 (green) or cy-5 (red) nucleotides using a random primed Klenow Polymerase reaction in an overnight incubation at 37°C (Ren et al., 2002). This resulted in a 10-20 fold amplification. Labeled samples were then hybridized to a microarray containing a nominal library of ~13,000 promoter sequences (~1 kb around the start of transcription) designed according to the April 2001 release of the human genome and remapped according to the August 2003 release by electronic PCR (Odom et al., 2004). A dataset reporting the location of each promoter relative to its transcriptional start site is downloadable from the supporting website (Supplementary Table 4). For our analysis we considered only promoters that were mapped in close proximity (+/- 1.5 kb) to the transcription start site (9,426 promoters), and 69% of these contain a CpG island. It should be noted that this high throughput approach may also be used with standard CpG island microarrays.

DNA methylation was determined directly by mixing bound and input DNA labeled with different fluorescent probes in equal quantities. For comparing the methylation pattern of different cell types, we mixed the labeled bound fractions prior to hybridization. The ratio of the two fluorescent probes gives an indication of the degree of methylation when displayed as a scatter plot. Differentially methylated genes ($P < 0.001$) appear above the upper red line. Many precipitation experiments were repeated for two or three times while swapping the dyes. As a control we co-hybridized input DNA samples from two cell types (Caco-2 and lymphocytes), and all of the sequences were found to be within the normal range of ratios, indicating that the differential hybridization observed with bound DNA cannot be due to gene amplification. Moreover, a comparison between normal colon and lymphocytes showed very few detectable differences in their promoter methylation. Indeed, only one of the 367 genes found methylated in cancer

was found to be differentially methylated in normal colon, thus strengthening the claim that the methylation observed in our assay is, in fact, tumor specific.

Ten selected methylated gene promoters were subject to whole-population bisulfite analysis (Engemann et al., 2001) without cloning individual molecules. This analysis showed that 162/164 individual CpG sites were more than 90% methylated in Caco-2 cells. In contrast, almost all of these sites (132) were less than 10% methylated in lymphoblasts, while 32 sites were found to be slightly methylated (10-30%) in these normal cells. 30 CpG sites were also examined in DNA from normal colon and the results were similar to those obtained from lymphoblasts. It should be noted that it is important to know the sex of each sample. In competitive hybridization between bound fractions from male vs. female lymphocytes, we detected about 50 methylated genes, with over 50% of these being derived from the X chromosome. In light of this finding, all of our studies were carried out using DNA from males. Analysis of tumors from female subjects would require using normal female cell DNA as a control.

Binding site determination and error model:
Scanned images were analyzed using GenePix (v4.1), to obtain background subtracted intensity values. We normalized the data using LOWESS normalization (Quackenbush, 2002). A whole-chip error model (Ren et al., 2000; Simon et al., 2001) was then used to calculate confidence values for each spot on each microarray, and to combine data for the replicates of each experiment to obtain a final average ratio and confidence for each promoter region. Genes were included in the set of 'methylated genes' if the *P* value in the error model was < 0.001.

Bioinformatics analysis:
For purposes of this study CpG islands containing promoters were defined as those regions having at their proximity at least 200 bp of sequences with a C+G content over 65% and an "observed CpG/expected CpG" in excess of 0.8 (Takai and Jones, 2003). The extensive literature search done for the identification of previously known cancer related hypermethylated genes was done with the aid of MILANO - an automatic literature search tool (Rubinstein and Simon, 2005). The mapping of the hypermethylated genes

to the GO biological process categories was done with the GoTree Machine (Zhang et al., 2004) without correction for multiple hypotheses. DNA sequences common to methylated promoters were detected by the use of a discriminative motif finder algorithm (Segal et al., 2003) using as input the list of genes found methylated in the mDIP microarray assay. The *P* values were calculated according to the hypergeometric distribution and we corrected for multiple hypotheses using an FDR of 1%.

For mapping the promoters to the chromosomes we used the ColoredChromosomes software (http://www.uni-essen.de/~bt0756/cc). In order to analyze clustering of the hypermethylated genes, we determined a list of 6,484 informative genes (that appear on the array and contain a CpG island) and looked for cases in which at least two adjacent genes from this list were hypermethylated. Overall, among the 367 hypermethylated genes we found 73 that are adjacent to at least one other hypermethylated gene in 37 separate clusters. In order to assess the statistical significance of this finding, we performed 100,000 computer simulations in which we randomly chose 367 genes from the list and counted the number of adjacent genes.

Expression data from normal and tumor tissues was extracted from the gene atlas project (Su et al., 2004) or other microarray information (Notterman et al., 2001; Yanai et al., 2005). Active genes are those with above-minimal expression levels as defined by Affymetrix Mas5 analysis. In general, not all the genes identified from the mDIP microarray were found in these respective databases. Studies examining the effect of Dnmt overexpression in fibroblast cells (Feltus et al., 2003), identified 12 CpG island sequences that underwent de novo methylation, 5 of which were present on our microarray. 2 of these were deemed methylated in PC3 cells by the mDIP assay.

References

Engemann S, El-Maarri O, Hajkova P, Oswald J, Walter J. Bisulfite-based methylation analysis of imprinted genes. *Methods Mol Biol* **181**:217-228 (2001).

Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM. Predicting aberrant CpG island methylation. *Proc Natl Acad Sci USA* **100**:12253-12258 (2003).

Hashimshony, T, Zhang, J, Keshet, I, Bustin, M, Cedar, H. The role of DNA methylation in setting up chromatin structure during development. *Nature Genet.* **34**: 187-192 (2003).

Maruyama R, Toyooka S, Toyooka KO, Virmani AK, Zochbauer-Muller S, Farinas AJ, Minna JD, McConnell J, Frenkel EP, Gazdar AF. Aberrant promoter methylation profile of prostate cancers and its relationship to clinicopathological features. *Clin Cancer Res* **8**:514-519 (2002).

Mayer W, Niveleau A, Walter J, Fundele R, Haaf T. Demethylation of the zygotic paternal genome. *Nature* **403**:501-502 (2000).

Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Res* **61**:3124-3130 (2001).

Odom DT, Zizlsperger N, Gordon DB, Bell GW, Rinaldi NJ, Murray HL, Volkert TL, Schreiber J, Rolfe PA, Gifford DK, Fraenkel E, Bell GI, Young RA. Control of pancreas and liver gene expression by HNF transcription factors. *Science* **303**:1378-1381 (2004).

Quackenbush J. Microarray data normalization and transformation. *Nat Genet* **32 Suppl**:496-501 (2002).

Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, Dynlacht BD. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. *Genes Dev* **16**:245-256 (2002).

Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA. Genome-wide location and function of DNA binding proteins. *Science* **290**:2306-2309 (2000).

Reynaud C, Bruno C, Boullanger P, Grange J, Barbesti S, Niveleau A. Monitoring of urinary excretion of modified nucleosides in cancer patients using a set of six monoclonal antibodies. *Cancer Lett* **61**:255-262 (1992).

Rubinstein R, Simon I. MILANO -- custom annotation of microarray results using automatic literature searches. *BMC Bioinformatics* in press (2005).

Segal E, Yelensky R, Koller D. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics* **19 Suppl 1**:i273-282 (2003).

Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS, Young RA. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**:697-708 (2001).

Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **101**:6062-6067 (2004).

Takai D, Jones PA. The CpG island searcher: a new WWW resource. *In Silico Biol* **3**:235-240 (2003).

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, Bar-Even A, Horn-Saban S, Safran M, Domany E, Lancet D, Shmueli O. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**:650-659 (2005).

Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics* **5**:16 (2004).