Supplemental Methods

DNA Samples

DNA human samples were collected by from 1,819 samples. We excluded the replicates used for validation (n=84), samples that did not attain control quality cut-offs (n=87) and the *in vitro* methylated DNAs used as whole genome positive marker for CpG methylation (IVD; n = 20). Thus, we finally analyzed 1,628 human samples. A criterion to define the quality of a sample is explained below. A detailed list of all the samples included in the study is displayed in **Supplemental Table 1**. All patients provided informed consent and the study was conducted under the approval of the corresponding Institutional Review Boards. For primary malignancies, fresh-frozen tissue samples were macrodissected to obtain a 90–95% purity of non-necrotic tumor and noninvolved adjacent non.neoplastic tissue. DNA methylation analyses for a subset of the hematological neoplasms (Martin-Subero et al. 2009a; Martin-Subero et al. 2009b) (GEO Expression Omnibus), lupus (Javierre et al. 2010) (GEO Accession number: GSE19033) and stem cells (Aranda et al. 2009) (GEO) have been previously reported. In order to assess the quality of the dataset, we computed the Pearson correlation coefficient of all pairs of methylation profiles; almost all replicate pairs had values close to 1. For subsequent analyses, we combined replicates by averaging the CpG methylation profiles of all records for a sample.

DNA methylation analysis using universal BeadArrays

Microarray-based DNA methylation profiling was performed on all samples with the GoldenGate Methylation Cancer Panel I (Illumina, Inc.). The panel was developed to assay 1,505 CpG sites selected from 807 genes, which include oncogenes and tumor suppressor genes, previously reported differentially methylated or differentially expressed genes, imprinted genes, genes involved in various signaling pathways, and those responsible for DNA repair, cell cycle control, metastasis, differentiation and apoptosis. The DNA methylation analyses were performed in the Human Genotyping Unit–CEGEN of the Spanish National Cancer Research Centre (Madrid, Spain), except for 8% of cases (127 hematological malignancies) where the analysis was developed at

the Illumina Headquarters (San Diego, CA). No significant inter-laboratory variation was observed.

DNA methylation assay was performed as previously described by Bibikova et al. in 2006 (Bibikova et al. 2006). Briefly, four probes were designed for each CpG site: two allele-specific oligos (ASOs) and two locus-specific oligos (LSOs). Each ASO-LSO oligo pair corresponded to either the methylated or unmethylated state of the CpG site. Bisulfite conversion of DNA samples was done using the EZ DNA methylation kit (Zymo Research, Orange, CA). After bisulfite treatment, the remaining assay steps were identical to the GoldenGate genotyping assay (Fan et al. 2003) using Illumina-supplied reagents and conditions. The array hybridization was conducted under a temperature gradient program, and arrays were imaged using a BeadArray Reader (Illumina Inc.). Image processing and intensity data extraction software were performed as described previously (Galinsky 2003b; Galinsky 2003a). Each methylation data point is represented by fluorescent signals from the M (methylated) and U (unmethylated) alleles. Background intensity computed from a set of negative controls was subtracted from each analytical data point. The ratio of fluorescent signals was then computed from the two alleles according to the following formula:

$$Beta = \frac{Max(M,0)}{Max(U,0) + Max(M,0) + 100}$$

Beta is a quantitative measure of DNA methylation levels of specific CpGs, and ranges from 0 for complete unmethylation to 1 for complete methylation. DNA methylation Beta values and p-values (measure of quality) for the 1,628 samples are available on the website: http://ubio.bioinfo.cnio.es/biotools/Human_DNA_Methylomes/ (user: data; password: 10HUMAN54). The sequence data from this study have been submitted to the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under accession no. GSE28094.

Analysis of differentially methylated probes

We used different methods of analysis depending on (1) the number of groups compared, and (2) when comparing two groups, the number of samples in the “case” and “control” groups.

We used elastic net methods to compare several groups of samples. The probes were selected by elastic net classifiers, trained with 10-fold cross-validation using misclassification loss. This approach was designed for applications, in which the number of features (probes) greatly exceeds the number of analyzed samples. These methods have recently been introduced to the Bioinformatics community and have been applied in SNP and gene expression datasets.

We used the Kruskal-Wallis test with the Benjamini-Hochberg algorithm to calculate the false discovery rate when we compared two groups with a large number of samples. Note that all methods were applied after a prefiltering step, as suggested by (Martin-Subero et al. 2009b), and only probes with mean methylation group differences of at least 0.25 were considered.

We implemented a specific strategy for determining differentially methylated probes in cases where two sample groups (cases and controls) were compared, and the control group was relatively small. This strategy does not include a prefiltering step, and is based on a heuristic approach, described briefly below. With this algorithm, a very small number of control (healthy) samples are compared with a larger group of case (disease) samples. We defined a probe P as unmethylated in a set of control samples, when the mean methylation value for this probe was < 0.25. Similarly, P was taken to be methylated if the average methylation value was > 0.75. We report P as hypermethylated in the case samples if and only if P was unmethylated in the control samples and the beta value of P was > 0.75 in at least 10% of the case samples. Likewise, the set of hypomethylated probes are those probes P that were methylated in the control group and had methylation values < 0.25 in at least 10% of the samples in case group. Another situation in which standard statistical methods are inapplicable is when the methylation profiles of two very small groups of samples (controls and cases) are compared. We applied a heuristic approach very similar to the previous one. We first classified a probe in the control group as unmethylated if the all methylation values for this probe among samples in the group were < 0.5 and the mean values were < 0.25. Alternatively, a probe was considered to be methylated in the control group if the observed values for all samples were > 0.5 and the mean value was > 0.75. The criteria for case group membership were stricter: unmethylated probes

were those in which the observed methylation value in all samples was < 0.25; the methylation values for all samples in a methylated probe were > 0.75. The set of differentially methylated probes consists of all probes that were methylated in the control group but unmethylated in the case (hypomethylated probes), as well as all probes that were unmethylated in the controls but methylated in the cases (hypermethylated probes).

In all settings, in which the methylation profiles of two groups were compared, we characterized the differentially methylated probes as being hypomethylated or hypermethylated with respect to the control groups, using the Kruskal-Wallis test with the Benjamini-Hochberg algorithm or heuristic methods. Associations between differentially methylated probes and CGI or non CGI location were compared using Fisher's exact test. In addition to Fisher's exact test, we calculated permutation-based p-values to account for interdependencies between the methylation states of different CpGs. Briefly, we performed Fisher's exact test in 10^4 random reassignments of the studied samples and calculated the proportion of resulting p-values that is lower than or equal to the originally obtained one.

For normal primary tissues we classified the probes as consistently unmethylated and consistently methylated. The consistently unmethylated group consisted of all probes that < 0.25 methylation in at least 99% of the samples. All probes with > 0.75 methylation in at least 99% of the samples formed the group of consistently methylated probes. The top-scoring genes with tissue-specific DNA methylation were defined as genes with methylation values > 0.75 in each tissue type.

We identified CpGs with age-specific methylation as the probes that (1) show an absolute correlation with age of at least 0.5 and (2) show a predictive power measured by a p-value < 0.01 after FDR correction. P-values were computed by applying analysis of variance (ANOVA) on a linear model with the methylation state of a CpG dinucleotide as its single predictor and age as the outcome.

Hierarchical cluster analysis and graphical representations

Statistical analyses were done and graphs produced with R (version 2.1.0) and Excel (Microsoft). Hierarchical clustering and heatmaps often contained tissue-,

cancer- or disease-specific probes calculated by Kruskal-Wallis test and elastic nets with misclassification. The Manhattan distance was used as the appropriate metric. A methylated CoG was always represented in red and an unmethylated CpG in green The track legend that accompanies the heatmaps represents the CpG location as inside or outside a CpG island (in red and blue, respectively).

The deviation plot depicts the variability of methylation values for set samples. Probes are ordered on the x-axis and are ranked with respect to their median methylation, as visualized by a curve. The yellow area enclosed within a grey border depicts the 5th and 95th percentile among the methylation values for each probe. Additional information about the probes is presented color-coded below the x-axis; CpG island- and non-CpG island- (CGI- and non-CGI-) associated probes are marked in red, and blue, respectively. The amount of variation in the methylation profiles can be quantified as the relative area of deviation (yellow bars) in a deviation plot, which is a number between 0 and 1. An area of zero indicates no variation, whereas the value of 1 depicts that all possible degrees of methylation are observed for every probe. The Wilcoxon test was used to calculate p-values for the association between methylation variability and CGI overlap. The variability of a probe was estimated as the difference between the 5th and 95th percentile of the methylation values of this probe. The differences between two deviation plots were measured, taking into account the median and variation in methylation. For this purpose, we first equalized the number of samples used in both plots, and then performed a paired Wilcoxon test using the values of the visualized sequences.

Expression data analysis

CEL files containing normal tissue gene expression data were downloaded from GEO database using the following data series:

Tissue	GEO ID
Aorta	GSE7307
Blood	GSE7307
Bone marrow	GSE3526, GSE7307
Brain	GSE3526
Breast (mammary)	GSE3526, GSE7307
Oral mucosa	GSE3526
Cerebellum	GSE3526
Cervix	GSE3526
Colon (cecum)	GSE3526, GSE7307
Endometrium	GSE3526
Esophagus	GSE3526
Heart	GSE3526
Liver	GSE3526
Lung	GSE3526
Muscle	GSE3526
Ovary	GSE3526
Prostate	GSE7307
Skin	GSE7307
Stomach	GSE3526,GSE7307
Suprarenal gland	GSE7307
Testis	GSE3526

Raw data were imported into Flexarray (version 1.4.1) and RMA normalized using Affymetrix Power Tools (32bit, version 1.12.0). Affymetrix annotation file HG-U133_Plus_2.na30.annot.csv was used to select Affymetrix probeset ID-s that corresponded to genes with tissue-specific methylation patterns. Ambiguous probesets associated with more than one gene were not included. If there were multiple probesets reporting on same gene, their intensity values were averaged to yield gene-wise expression data. Selected expression data were imported into Genesis (version 1.7.1), median-centered and gene-wise normalized. Unsupervised hierarchical clustering and heatmaps

using the expression data for the 354 genes (including the 511 tissue specific CpG sites) was carried out on the basis of Manhattan distance calculation and average linkage clustering (**Supplemental Fig. 3**). Gene expression data downloaded from GEO database and the same data series were used to define a gene as housekeeping gene. We selected genes expressed in 90% of the normal tissues included in the panel. We used the following procedure: absent-present calls were generated from 99 normal tissue samples using the "mas5calls" function in the R package "affy". 8,643 probesets were found to be present ("P") in ≥90% of the samples. For these probesets, the corresponding gene symbols were determined using the Affymetrix annotation file HG-U133 Plus2.na30.annot.csv, yielding a list of "≥90%_expressed_genes" (5,427 genes identified). "≥90%_expressed_genes" list "unmethylated_genes in normal tissues" and "other_genes" list was crossed.

A density plot of microarray-based gene expression data in colon cancer patients was also experimentally obtained (**Supplemental Fig. 6**). Expression data were obtained from 19 primary colorectal tumors for which we had obtained DNA methylation profiles. 5 µg of RNA were hybridized on the Affymetrix Human GeneChip U133 Plus 2.0 expression array (Affymetrix, Santa Clara, CA). Expression data were normalized and analyzed following the same procedures as described above.

References

Aranda P, Agirre X, Ballestar E, Andreu EJ, Roman-Gomez J, Prieto I, Martin-Subero JI, Cigudosa JC, Siebert R, Esteller M et al. 2009. Epigenetic signatures associated with different levels of differentiation potential in human stem cells. *PLoS One* **4**(11): e7809.

Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, Doucet D, Thomas NJ, Wang Y, Vollmer E et al. 2006. High-throughput DNA methylation profiling using universal bead arrays. *Genome Res* **16**(3): 383-393.

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P et al. 2003. Highly parallel SNP genotyping. *Cold Spring Harb Symp Quant Biol* **68**: 69-78.

Galinsky VL. 2003a. Automatic registration of microarray images. I. Rectangular grid. *Bioinformatics* **19**(14): 1824-1831.

-. 2003b. Automatic registration of microarray images. II. Hexagonal grid. *Bioinformatics* **19**(14): 1832-1836.

Javierre BM, Fernandez AF, Richter J, Al-Shahrour F, Martin-Subero JI, Rodriguez-Ubreva J, Berdasco M, Fraga MF, O'Hanlon TP, Rider LG et

al. 2010. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. *Genome Res* **20**(2): 170-179.

Martin-Subero JI, Ammerpohl O, Bibikova M, Wickham-Garcia E, Agirre X, Alvarez S, Bruggemann M, Bug S, Calasanz MJ, Deckert M et al. 2009a. A comprehensive microarray-based DNA methylation study of 367 hematological neoplasms. *PLoS One* **4**(9): e6986.

Martin-Subero JI, Kreuz M, Bibikova M, Bentink S, Ammerpohl O, Wickham-Garcia E, Rosolowski M, Richter J, Lopez-Serra L, Ballestar E et al. 2009b. New insights into the biology and origin of mature aggressive B-cell lymphomas by combined epigenomic, genomic, and transcriptional profiling. *Blood* **113**(11): 2488-2497.