Online Methods

Custom GoldenGate Array Design

We selected 151 of the cDMRs identified in Irizarry et al.[4], regions consistently differentially methylated in all 13 colon cancers studied by comprehensive high-throughput array based methylation (CHARM) analysis. Probes were designed around CpGs that showed consistent differences in CHARM, while passing Illumina's quality control metrics[7]. The resulting probes covered 139 regions, with 1-7 probes per region. The majority of the probes were in CpG island shores (66%), defined as less than 2 kb away from the edge of a canonically defined high-CpG density island[4]. The remainder of the probes were either inside CpG islands (11%) or >2 kb distant (23%).

Sample Preparation

Cryogenically stored freshly frozen samples were obtained from the Cooperative Human Tissue Network (NCI, Bethesda, MD), the National Wilms Tumor Study tissue bank (Edmonton, Alberta, Canada) and the Johns Hopkins Hospital, under an IRB-approved waiver of consent. In total 290 samples were assayed, including cancers from colon (10), lung (24), breast (27), thyroid (36), and kidney (Wilms') (25), with matched normal tissues to 111 of these 122 cancers, along with 30 colon premalignant adenomas, 18 normal colon, and 9 normal breast samples. Two small sections were taken from each sample (~0.5cmx0.5cmx0.2cm); one for DNA purification and one for histopathology. Histopathology samples were submitted to the immunohistochemistry lab at Johns Hopkins Hospital for processing. Normal and cancer samples were matched from the same patient and the same tissue whenever available.

A board-certified oncology pathologist validated classification of all samples independently and blindly. The pathologist also quantified specific cellular subtypes and p53 status for tumor and normal specimens from colon and kidney. Supplementary Figure 3 summarizes the histological analysis of the colon and kidney normal and tumor samples demonstrating that normal samples are typically more heterogeneous in cellular composition than the tumor samples.

DNA purification was done using either the DNeasy Blood and Tissue Kit (Qiagen) or the MasterPure Complete DNA Purification Kit (Epicenter). DNA concentration and purity was assayed using a Nanodrop spectrophotometer. Tumor and normal status and tissue type were balanced by plate to avoid batch effects. Methylated and unmethylated controls (Zymo Research), along with sample cross-plate controls, were included on plates. Samples were bisulfite treated using the EZ-96 Methylation Gold kit (Zymo Research), and hybridization was performed by the Center for Inherited Disease Research of Johns Hopkins University.

Custom Illumina methylation array processing and analysis

We quantile normalized[27] separately the raw intensity data from the Cy5 and Cy3 channels representing methylated and unmethylated DNA, and methylation level was calculated as the ratio of the Cy5 intensity over the sum of the intensities from both channels. To control for array quality, arrays for which the average of the median log intensities from the two channels was small (<7), or for which the median absolute deviation of the overall methylation signal was small (<1.9) were removed from the dataset. We ruled out batch effects following the procedures described by Leek et al. [28].Differences in methylation variability were measured and tested using an F-test. Differences in mean methylation levels were measured and tested using a t-test. Significance was taken as 0.01.

We regressed age out of the methylation data (further detail available in Supplemental Note) by fitting a linear regression model and repeated the analysis (Figure 1) using age-corrected measurements and obtained almost identical results (Supplementary Fig. 4). We also performed array-based copy number analysis using on 5 colon tumors and normals, alongside 5 Wilms' tumor samples and normal kidney samples. We used intensities from the Cy3 channel (total DNA) that were corrected for spatial effects, quantile normalized, and corrected for sequence effects[29](more details available in Supplemental Note). The resulting log ratios and estimated copy number segments are shown in Supplementary Figure 5.

Whole genome bisulfite sequencing

Bisulfite sequencing libraries were prepared using the approach previously described by Bormann Chung et al.[30](see Supplemental Note). Corresponding colon cancer and normal mucosa samples were sequenced simultaneously in adjacent flow cells on the SOLiD 3+ platform yielding 50 base pair reads.

Capture bisulfite sequencing

To obtain accurate quantitation of absolute DNA methylation level at the single-nucleotide resolution, we used the Bisulfite Padlock Probes (BSPP) previously developed[31]. In this method, a library of long oligonucleotide sequences is designed for the regions of interest, custom designed based in part on DMRs previously found in colon cancer[4], covering ~60,000 highly curated differentially methylated regions in the human genome (620,708 CpG sites in 19.2Mb of genomic regions covered). This method was used on genomic DNA from the three tumor-normal samples already used in whole genome bisulfite sequencing.

Alignment of sequencing reads from bisulfite treated DNA

We developed a custom alignment tool for Illumina and SOLiD sequencing reads derived from bisulfite-treated DNA. The tool aligned reads with the aid of a spaced-seed index of the genome while biasing neither toward nor against methylated cytosines in CpGs. The aligner used here leaves each read as-is but penalizes neither C-to-C nor T-to-C partial alignments in CpGs. Our approach supports a broad range of spaced-seed designs and extends the BSMAP[32] approach to allow alignment of SOLiD colorspace reads as well as typical Illumina reads(see Supplementary Note for more detail).

Alignment of SOLiD sequencing reads from whole-genome bisulfite-treated DNA

The algorithm described above was used to align a total of 7.79 billion reads obtained from 8 runs of a SOLiD 3 Plus instrument against a reference sequence collection consisting of the GRCh37 human genome assembly (including mitochondrial DNA and “unplaced” contigs) plus the sequence of the spiked-in λ phage genome. Alignment results are summarized in Supplementary Table 14 and Supplementary Figure 19, more detail is provided in Supplementary Note.

Alignment of Illumina sequencing reads from captured bisulfite-treated DNA

The algorithm described above was also used to align a total of 79.3 million reads obtained from an Illumina GA II instrument against a reference sequence collection consisting of the GRCh37 human genome assembly (including mitochondrial DNA and “unplaced” contigs). Alignment results are summarized in Supplementary Table 15 and Supplementary Figure 20, more detail is provided in Supplementary Note.

Extraction of methylation evidence from alignments

After alignment, a series of scripts extracted and summarized CpG methylation evidence present in the unique alignments. The evidence was compiled into a set of per-sample, per-chromosome evidence tables. Alignments to the λ phage genome were also compiled into a separate table. Once a piece of evidence was extracted from a unique alignment, it was subjected to a filtering (Supplementary Note) determined by examining the M-bias lines (Supplementary Figures 21 and 22). Supplementary Tables 16 and 17 summarize the amount and type of evidence extracted at each stage for the whole-genome SOLiD bisulfite sequencing and Illumina capture bisulfite sequencing data respectively. Supplementary Table 18 summarizes the whole-genome SOLiD bisulfite sequencing CpG evidence coverage with respect to the GRCh37 human genome assembly for each sample.

In the case of the SOLiD reads obtained by sequencing whole-genome bisulfite-treated DNA, evidence from reads that aligned uniquely to the λ genome was used to estimate the bisulfite conversion rate for unmethylated cytosines. Supplementary Figure 23 and the final column of Supplementary Table 14 show the estimates, which all lie between 99.7% and 99.8%. To measure global prevalence of non-CpG cytosine methylation, we examined all filtered nucleotide evidence from the SOLiD reads overlapping non-CpG cytosine positions in the human reference genome. Supplementary Table 19 summarizes the results: the fraction of Cs observed overlapping non-CpG cytosines does not rise above the approximate fraction expected from unconverted cytosines. We also found that strand bias is not a concern in our data by examining the fraction of evidence reported from each strand. Results are included in Supplementary Tables 3, 7, 20 and 21.

BSmooth: Smoothing of bisulfite sequencing reads and DMR detection

The cornerstone of our analysis is a statistical approach that permits highly accurate and precise estimates of methylation values with 4x coverage data. This is achieved by first using an unbiased alignment algorithm as described previously then averaging data across neighboring CpGs and across biological samples to greatly improve precision over naïve single CpG and single sample estimates. The biological replication also permits us to parse cancer related differences from inter-individual variability (Supplementary Fig. 13).

Data summary after alignment

The alignment algorithm described in the previous section provided two counts for each CpG: number of pieces of filtered evidence indicating presence of methylation (M) and number of pieces of filtered evidence indicating unmethylation (U). The sum of these two counts was the coverage N. We assumed that for each CpG, M followed a binomial distribution with success probability p, equal to the true methylation level, and N trials. Thus, M/N provided a naïve estimate of p with standard error $(M/N) \times (1-M/N) / \surd N$. Note that N, the coverage, ranged from 0 to 18,090 with a sample mean of 5.8-6.2 (after excluding CpGs with no coverage in all samples). We greatly improved precision by leveraging the fact that proximal CpG have similar methylation levels[33] and using a smoothing technique inspired by that used in CHARM[4].

Smoothing via local likelihood estimation

Because the data was binomially distributed, we used local likelihood estimation[34]. This approach assumes that the p(L), the methylation level at genomic location L, is a smooth function of L; in other words, that CpGs that are close have similar methylation levels[33]. We developed two related approaches; one using low-frequency smoothing for blocks and one using high-frequency smoothing for small DMRs (Supplementary Note). An example of the results from the high-frequency smoothing is provided as Supplementary Figure 24. High-coverage capture bisulfite sequencing data from selected regions confirm the highly accurate and precise estimates predicted by the statistical calculations as described in detail below. The data from the adenoma samples were smoothed in the same way.

Accounting for biological variability

We then developed a method for finding differences based on t-statistics that take into account biological variability. We started with the highly precise estimates of $p_i(L)$ for each sample i at each CpG location L. We obtained the average difference between the three tumor samples and the three normal samples referred to as d(L). To properly account for biological variability (Supplementary Fig. 13) we estimated the standard error of d(L) using the normal samples. We used only the normal samples because as we demonstrated, with independent data, cancer samples are prone to high variability (Fig. 1) and here we are concerned only with DMRs In other words, we are not assuming that the cancer samples are biological replicates. The standard error se[d(L)] was therefore estimated as $\sigma(L)*\sqrt{(2/3)}$ with σ(L) the standard deviation of the pi(L) for the three normal samples. To improve standard error estimates, we smoothed these using a running mean with a window size of 101 observations. To avoid inflated t-statistics as a result of artificially low variance, we set a threshold for the standard deviation of its 75th percentile, before computing the smoothed result. With the standard deviation in place we constructed the t-statistic t(L) = d(L)/se[d(L)].

For the high frequency analysis the t-statistic was further corrected for low frequency changes. We then defined small DMRs as contiguous CpGs within 300 bp of each other, with the t-statistics above 4.6 or below -4.6 (corresponding to the 95th quantile of the empirical distribution of the t-statistics) and all differences in the same direction. For the low frequency analysis the t-statistics cutoff was 2 and contiguous CpG were defined as within 10,000 bps from each other. These sets of regions formed our small DMRs and blocks that were subsequently filtered and merged (Supplementary Note). The final list of blocks is available as Supplementary Table 3, and of small DMRs as Supplementary Table 7.

We also computed sample-specific blocks as outlined above by comparing a single tumor sample to all three normal samples. The list of sample-specific blocks and small DMRs can be found in Supplementary Table 20 and 21. This analysis demonstrated that the tumor-specific blocks are largely contained inside the blocks from the joint analysis. We also performed a simulation analysis to verify the extent to which sample-specific blocks co-occur (Supplementary Note).

DMR Classification

Small DMRs were classified into categories based methylation profiles of the tumor and normal samples within the DMR and the two flanking regions (within 800bp). Based on these results, the DMRs that were discovered from data exploration could be classified into three types termed loss of methylation boundaries, shifting of methylation boundaries, and novel hypomethylation (Fig. 3). The algorithm used to classify DMRs is discussed in the Supplementary Note.

Comparison of bisulfite capture and whole genome bisulfite sequencing

The capture bisulfite experiment described above provided data for 474,829 CpGs (in one or more samples) from 39,262 regions. The genomic size of these regions ranged from 230 to 2,200 bp. For the analysis presented here we considered only the CpGs with coverage above 30x, which resulted in 39,285, 107,332 and 86,855 CpGs in the normal samples and 125,611, 94,320, and 104,680 in the cancer. We computed an estimate of methylation for each CpG using simply the proportion of reads showing evidence of methylation for that CpG. We did not perform any averaging across genomic regions or sample. We then compared the whole genome bisulfite sequencing data processed with BSmooth (referred to here as WGBS) to the high-coverage capture bisulfite (referred to as CAP).

To verify the accuracy of our methylation values obtained from BSmooth, Bisulfite pyrosequencing was used to verify the accuracy on the same 6 samples that were sequenced, for the small DMR regions shown in Figure 3 (primers are listed in Supplementary Table 22). Plotting these loci shows good correspondence with our smoothed methylation values (Supplementary Fig. 9).

We also performed a simulation analysis to verify the extent to which sample-specific blocks co-occur (Supplementary Note). For each chromosome, we used the observed distribution of the sample-specific blocks to estimate the distance between block starts. For each chromosome, 1,000 simulated start positions of blocks were generated according to this distribution. For illustration purposes, Supplementary Figure 10a shows the observed and simulated block start for a 20 Mb region of chromosome 1. Differences in the expected and observed distribution can be observed on boxplots in Supplementary Figure 10b.

Assays for Large Organized Chromatin K9-modifications (LOCKs)

Primary human pulmonary fibroblasts (HPF) were purchased from ScienCell Research Laboratories (San Diego, CA). Cell culture was conducted using the media and protocols recommended by ScienCell. Primary cells at the second passage were used for H3K9Me2 LOCK analysis. ChIP-on-chip experiments and microarray data analysis were performed as described earlier[12].

Hypomethylation in blocks and repeat regions

Repeat regions were identified based on the UCSC repeatMasker track[35]. Based on the repeats and/or blocks, the genome was segmented into regions both repeats and blocks, repeats but not blocks, not repeats but blocks, and neither repeats nor blocks. The methylation levels were computed as the average of the high-frequency smoothed methylation levels of all CpGs in the 4 different regions. Density estimates were computed from the same distribution. Supplementary Table 4 describes the extent to which we were able to map CpGs inside repeat elements.

Copy number analysis of sequenced samples

Estimates of copy number were based on the per-base coverage obtained after alignment. For each tumor-normal pair we computed coverage log-ratios that were then segmented using circular binary segmentation (CBS)[36]. For illustrative purposes, copy number log-ratios and the associated segmentation on chromosome 20 were depicted (Supplementary Fig. 11a). For each of these regions we computed average copy number ratios as well as average methylation ratios. These were then plotted (Supplementary Fig. 11b) and no relationship between CNV and methylation blocks was observed.

Gene expression data

We obtained expression data from the gene expression barcode (rafalab.jhsph.edu/barcode). This resource combines all the expression data from the public repositories purportedly to standardize data in a way that allows one to call a gene expressed or not expressed[26]. From this source, we used two independent colon cancer datasets (Fig 5b: GSE8671[37] and Supplementary Fig 18: GSE4183[38,39]). For the fibroblast analysis we downloaded datasets (GSE7890[40], GSE11418[41] , GSE11919[42]) and were also standardized using the gene expression barcode. Further details are provided in the Supplementary Note. To define tissue-specific genes, we downloaded 529 gene expression microarrays from NCBI GEO representing 30 different tissues for which at least 5 biological replicates were available. The GEO accession numbers for these 529 microarrays are listed in Supplementary Table 23.

Gene expression variance analysis

Because the great majority of genes exhibit increased variance in cancer samples, standard statistical inference techniques do not guide the choice of a threshold to dichotomize genes by hypervariability. To demonstrate the indisputable association between hypomethylated blocks and hypervariability of gene expression we stratified genes by their across-sample standard deviation in cancer into 10 bins and for each bin we calculated the proportion of these genes that are in hypomethylated blocks. There is a clear

(Supplementary Fig. 17), and statistically significant ($p<0.01$), direct relationship starting at about 20% and ending at 100%.

URLs

Expanded set of block plots, http://rafalab.jhsph.edu/cancer_seq/block.pdf; Expanded set of small DMR plots, http://rafalab.jhsph.edu/cancer_seq/smDMR.pdf.