ABSTRACT

De novo methylation of CpG islands is seen in many cancers, but the general rules governing this process are not known. By analyzing DNA from tumors, as well as normal tissues, and by utilizing a range of published data, we have identified a universal set of tumor targets, each with its own “coefficient” of methylation that is largely correlated with its inherent relative ability to recruit polycomb. This pattern is initially formed by a slow process of de novo methylation that occurs during aging and then undergoes expansion early in tumorigenesis where, we hypothesize, it may act as an inhibitor of development-associated gene activation.
INTRODUCTION

Many studies have begun to shed light on the normal genome-wide pattern of DNA methylation during development (1-3). It is now clear that following erasure in the early embryo there is a wave of de novo methylation that occurs at about the time of implantation, generating a bimodal pattern in which most CpGs become methylated while CpG-island-like sequences are protected (4, 5). This pattern is then maintained through cell divisions in the adult, and only a minority of these islands gets methylated in somatic cells during normal development (1, 3).

It is well known that CpG islands get aberrantly hypermethylated in cancer (6) and from genome-wide analyses it is clear that a large number of islands are involved (7-9). Nonetheless, in these studies genes are usually classified as being methylated or not methylated in the tumor as compared to the source tissue using an arbitrary threshold for DNA modification without taking into consideration the actual levels of methylation at individual CpG islands in different tissue types. For this reason, there is a need to analyze this data in a more holistic manner that better reflects this process in vivo.

Although it has been shown that a large percentage of tumor-methylated sequences are polycomb target genes (10-12) that may already be repressed in the parent cell type (13, 14), little is known about the dynamics of this process during tumorigenesis. Based on a relatively small number of sequences, several studies noted that the pattern of DNA methylation in tumors may be similar to that associated with aging (15), but the dynamic...
genome-wide relationship between aging-related and cancer-related methylation has not been elucidated. Similarly, the rules that govern tissue-of-origin specificity in cancer, as well as the impact of this modification on gene expression and tumor physiology, are not well understood.

We have attempted to take a closer look at genome-wide DNA methylation from a developmental perspective. By combining this with a re-analysis of previously published data from a variety of different tumor types, we have been able to derive general rules that characterize the origin and general behavior of DNA methylation in cancer biology.
MATERIALS AND METHODS

Methylation analysis

Human DNA samples were purchased from Biochain or Bioserve (Supplementary Table S1). Human genomic DNA (10 μg) was sonicated to an average fragment size of 300–1000 bp, precipitated with 400 mM NaCl, 2 volumes of ethanol and 1 μl glycogen. 1.5 μg were set aside as the Input fraction. DNA was denatured and anti-5-methylcytidine monoclonal antibody (10 μl for 5 μg) (16, 17) was added and incubated on a rotator at 4°C overnight. 40 μl Dynabeads (Sheep anti-Mouse IgG) were prewashed with 0.1% BSA/PBS and added to the DNA. The DNA was then washed 3 times and Ab-bound DNA resuspended and extracted with proteinase K, phenol-chloroform and ethanol precipitation. Purified DNA was checked for enrichment (Bound/Input) using Real time PCR on specific gene regions known to be methylated.

The Input and Bound DNAs were labeled and hybridized on human (244 K) CpG island microarrays or human DNA methylation microarrays (244 K) (Agilent Technologies) as described previously (3). We used feature extraction software (Agilent) to obtain background-subtracted intensity values for the two fluorescence dyes on each individual array feature and for carrying out linear normalization and calculation of the log ratio (Cy5/Cy3). Data was analyzed as previously described (3). Briefly, probe log ratio signals were transformed into Z-scores according to their Tm and an Island Methylation Score (IMS) was then calculated by averaging the island corresponding probes’ Z-scores. We defined a set of "background" CpG islands by selecting
those with an IMS of less than 0.5 in all 10 fetal tissue samples (Brain, Liver, Skeletal Muscle and Colon, Supplementary Table S1). Microarray data is available at the Gene Expression Omnibus (GEO) repository under accession GSE38142.

**Bisulfite Analysis**

Bisulfite conversion of genomic DNA was carried out using the EZ DNA Methylation-Direct Kit (Zymo Research) according to the manufacturer’s instructions. PCR primers were designed using Methyl Primer Express® Software v1.0 (Applied Biosystems). Barcodes and adaptors were added to the primers and deep-sequenced using the Ion Torrent (Life Technologies).
RESULTS

Methylation patterns in cancer

The global bimodal methylation pattern of the human genome is established at the time of implantation and then maintained throughout development and adult life (18). Thus, in order to objectively study de novo methylation in cancer and its relationship to aging, we first identified all of the unmethylated CpG islands formed in the embryo and then determined their modification patterns in normal tissues and tumors. This approach allowed us to obtain an overall picture of tumor methylation as opposed to other studies which generally focused on specific cancers and their tissues of origin.

We have adopted mDIP microarray technology to analyze the global methylation state of CpG islands (2, 3, 13). This method has the advantage that it provides an excellent validated (3) measure of average methylation levels over each island. We first analyzed a variety of different fetal tissues (Supplementary Table S1) and identified over 13,000 CpG islands that have constitutively low Island Methylation Scores (IMS) of methylation (Materials and Methods) in all ten samples examined. In this way, we defined a background set of sequences that were originally set up as unmethylated at implantation. We then used mDIP to compile the levels of methylation for these same islands in individual colorectal cancers (CRC) and ordered them from the most down to the least modified sequences. As can be seen in Figure 1A, there is a gradient of DNA modification with a large number of CpG islands showing excess methylation levels over background. All of the individual CRC samples are characterized by DNA methylation at
approximately the same CpG island sites. Analysis of other tumor types such as liver or lung indicated that while the overall levels of methylation were lower, the same set of target CpG islands modified in CRC are significantly methylated in these other tumors, as well.

To better visualize this phenomenon, we selected the top 1,524 CRC islands (IMS > 0.6) and mapped their methylation levels as they project onto other tumors using our own data for lung, as well as public data for breast (19) and pancreatic cancer. These distribution curves demonstrate that this same set of islands is modified in other tumors, as well. Furthermore, it appears that the set of highest methylation-ranked CpG islands in a wide range of other tumors is always found to be significantly enriched (P < 10^{-300}) with the same set of ~1,500 CRC islands (Fig. 1B and Supplementary Table S2), as determined by the minimal hypergeometric test (mHG) (20, 21). These results suggest that there is a universal set of sequences commonly modified in many tumors.

Source of de novo methylation

Several studies have addressed the question of how DNA methylation arises during tumorigenesis (22-24). To obtain some indication about the source of tumor-related methylation, we used the same CRC islands (Fig. 1) and mapped their profile (IMS) in normal colon DNA from age-ranked samples. Strikingly, these methylated islands form a well-defined population whose methylation levels in colon are significantly higher than the average set of constitutively unmethylated sequences (background islands) and increase as
a function of aging (Fig. 2 and Supplementary Fig. S1A). It thus appears that these CpG islands are already specifically modified in normal colon tissue, and the same seems to be true for other cell types, as well. In contrast, no significant excess methylation was detected for ES cell DNA which represents the pattern present at implantation (Supplementary Fig. S2).

It should be noted that the overall levels of this excess DNA methylation are relatively low in normal tissues, and despite the recent identification of some age-dependent sequences (25), it is unlikely that the full modification picture would have been detected. It is only because we specifically pinpointed the cancer-associated target population that we could distinguish this phenomenon. Nonetheless, analysis of age-related methylation from our data (colon) as well as others (T-cells) (26) indicates that CRC islands are significantly enriched (mHG $P < 10^{-300}$) in sequences that exhibit age-related methylation level shifts (from fetal to 25 y.o. to 65 y.o., Fig. 2 and Supplementary Methods). Thus, CRC islands are overmethylated as a function of aging in normal tissues with the highest degree being seen in colon DNA. Furthermore, hypermethylation in tumors is a reflection of these excess methylation patterns and is always at a higher level than that present in the control tissue of origin (Supplementary Fig. S3).

**“Coefficient” of DNA methylation**

In light of the observation that de novo methylation targets are similar in all tissues and tumors, we asked whether the levels of DNA methylation at each CpG island is determined stochastically, or alternatively, may be targeted in a
preferred manner. One way to approach this issue is to compare the methylation patterns of tumors from the same cell type. To this end, we used the full set of background CpG island methylation scores to map individual CRC samples against another. Strikingly, this analysis yielded a high correlation coefficient, indicating that specific islands have a very similar degree of methylation in each sample (Spearman’s R = 0.75, Supplementary Fig. S4A). Similar strong correlations were obtained by comparing methylation levels of CpG islands in two different Breast tumor (19) or Glioblastoma (TCGA) samples as determined by high throughput analysis (Spearman’s R > 0.88, Supplementary Fig. S4A). These studies suggest that each island has its own propensity to undergo de novo methylation.

Previous studies have noted that it may be possible to define two separate categories of colon tumors as either positive (CIMP+) or negative (CIMP−) for the methylator phenotype, with the difference being determined by the methylation state of a fixed panel of selected CpG islands (27). To better understand these categories, we analyzed genome-wide methylation data (TCGA, Supplementary Methods) for both tumor types. Strikingly, this approach indicated that they have very similar qualitative patterns of methylation with both categories being highly enriched for the CRC islands (mHG P ≤ 10\(^{-20}\)), but in CIMP\(^+\) cells the same targets are modified to a higher degree. Indeed, our analysis shows that tumors with more above-threshold methylated marker genes indeed have a higher average CpG island methylation level (Supplementary Fig. S4B), as has also been suggested by others (28).
Having established that tumors from any particular cell type have a similar hierarchy of de novo methylation, we then devised a simple visual approach to compare different tumor cell types. To this end, we ordered the entire set of CRC islands according to their natural levels of modification in colon. On this basis, we then compared methylation patterns for these same islands in other cancer types. Strikingly, all of the tumors show a gradient of methylation almost identical to the control (colon) (Spearman’s R > 0.4, Supplementary Table S2), independent of the actual overall levels of excess methylation in each DNA sample (Fig. 3A) using our own data as well as that from the literature, and this was confirmed by demonstrating that the rank variance across all samples is very low (Supplementary Fig. S4C).

This same relationship is also observed when the Island Methylation Scores of each site are plotted as a function of their rank in normal colon (Fig. 3B). Finally, we generated a scatter plot for all background CpG islands comparing samples from CRC and other tumors (Supplementary Fig. S4D). These data clearly show that the patterns of methylation are very similar even though the overall intensity of modification is higher in the CRC samples. Taken together, these results suggest that every island has its own inherent methylation propensity independent of the source of DNA or the measurement platform.

Histone modification and de novo DNA methylation
Based on ChIP data from ES cells, it has been suggested that the polycomb complex may play a role in recruiting DNA methylases to specific CpG islands in cancer (10-12), but this relationship has never been tested by mapping polycomb sites in the actual tissue that gives rise to the tumor. We took advantage of H3K27me3 ChIP-Seq data for normal colon mucosa (29), and plotted the degree of DNA methylation (IMS) at all CpG islands as a function of their polycomb-catalyzed H3K27me3 density (Fig. 4A). Strikingly, we detected a monotonically increasing relationship between these two parameters, strongly suggesting that Polycomb components already present in the primary tissue may serve as a biochemical predictor for the amount of de novo DNA methylation ultimately seen in tumors. Indeed, abnormal methylation appears to take place mainly in the 35% of the genome that has above-normal background levels of H3K27me3, while CpG islands that lack this histone mark (< 3.5) appear to be free of DNA modification (Fig. 4A). Indeed, over 88% of the CRC islands are polycomb positive according to this criterion (P < 10^{-300}, hypergeometric test).

Many studies have shown an inverse correlation between DNA methylation and H3K4me3 (2, 30) and it is thought that this histone modification may actually protect the underlying DNA from becoming methylated (31, 32). To test whether this is also true for de novo methylation in cancer, we analyzed all H3K27me3-positive CpG islands in normal colon and plotted their CRC methylation levels as a function of H3K4me3 concentrations (Fig. 4B). These data show that the presence of H3K4me3 has a marked negative effect on the degree of local de novo methylation in tumors. Indeed, at every level of
H3K27me3, sites with high concentrations of H3K4me3 are less methylated than those with lower concentrations (Supplementary Fig. S5A).

Using this data and employing machine learning methods (WEKA) (33), we devised a predictor for DNA methylation levels in cancer as a function of local histone modification density in the source tissue (Supplementary Methods), and then cross-validated this function in a separate test sample (Fig. 4C and Supplementary Fig. S5B). This analysis indicates that the patterns of H3K4me3 and H3K27me3 in colon are excellent predictors for abnormal DNA methylation in colon tumors (AUC=0.88) as well as in aging (AUC=0.81).

Despite previous studies showing a correlation between de novo methylation in cancer and the distribution of polycomb in ES cells (10, 12), the histone modification profile in primary colon tissue is a much better predictor of abnormal methylation (Supplementary Fig. S5C).

**Tumor-specific methylation**

While the overall pattern of DNA methylation appears to be universal, some de novo methylation may be tumor-type specific, as has been previously suggested (34, 35). To determine whether this tissue specificity can also be explained by the molecular rules governing cancer methylation, we selected 30 CpG islands that have a high methylation score (IMS > 0.5) in lung cancer, but appear to be relatively undermethylated (IMS < 0.3) in CRC, and analyzed the patterns of histone modification in their normal tissues (29) (Supplementary Methods). In lung, these sequences have a level of H3K4me3 similar to common CpG islands that are methylated in both tumor
types. In colon, however, the concentration of H3K4me3 at these sites is much higher, on average, than what is seen at most CpG islands that undergo de novo methylation in colon cancer (Fig. 5A). Similar results were obtained for tissue-specific de novo methylation in breast and liver tumors. This effect cannot be attributed to differences in levels of H3K27me3 at these sites (data not shown). These results suggest that tissue-specific differences are strongly correlated with local packaging of H3K4me3, in keeping with the principles derived by our predictor (Fig. 4).

In addition to the large number of constitutively unmethylated background islands used in our studies, there are many CpG islands that undergo tissue-specific de novo methylation during normal development (1, 3). Since these methyl footprints are unique to each individual cell type, we reasoned that all tumors originating from a particular tissue should bear these same distinctive marks, thus providing a type of barcode for identifying tumor origin. To test this idea, we compared DNA methylation patterns from colon tumors and hepatomas.

We identified approximately 20 CpG islands specifically methylated in normal colon, and all were found to be modified in colorectal cancer. In contrast, these CpG islands were unmethylated in hepatomas. Conversely, we identified a set of CpG islands methylated in normal liver and these continue to carry this mark in primary hepatomas (Fig. 5B). Using this same strategy, we were also able to distinguish between lung and liver tumor samples (Supplementary Fig. S6). While this form of lineage tracing is not necessarily
perfect for all individual CpG island markers, having a large set of specific indicators for each cell type makes it relatively easy to identify the tissue of origin for any tumor. Furthermore, analysis of a small number of metastases (Fig. 5B and Supplementary Fig. S6) shows that they retain these differential patterns, suggesting that this information can be used for identifying their tissue-of-origin in cases where it is unclear from the morphology alone.
DISCUSSION

All tumors are characterized by changes in DNA methylation that include both widespread demethylation as well as de novo modification (6). Although this phenomenon was first observed over 30 years ago, little is known about the molecular rules that govern these processes. We used genome-wide technologies to weave together a large amount of data on de novo methylation with the aim of generating a general picture of how this takes place in vivo. Previous studies envisaged de novo DNA methylation as a local event specifically associated with cell transformation, and as a result concentrated on methylation patterns in the tumor as compared to the source tissue. In this paper, we have taken a more holistic approach which views DNA methylation from an organism-wide developmental perspective. This manner reveals a more global process in which abnormal DNA modification comes about as a natural consequence of aging in a wide variety of cell types, with cancer methylation being an outgrowth of this process. Furthermore, there appears to be a universal hierarchy of methylation sites in the genome, with every CpG island having a different propensity for methylation that, despite some variation, is largely independent of cell type or overall level of modification.

Our studies point to a combined role for both polycomb (H3K27me3) and H3K4me3 as a pre-set tissue chromatin code that correlates with the subsequent pattern of de novo methylation formed in the tumor. By relating to these histone modification marks as continuous, as opposed to “all or none” variables, it is possible to derive a formula that can actually predict the degree
of methylation at each site (Supplementary Fig. S5B). Our data also show that the combined histone modification signature in the somatic tissues from which the tumors are derived, as opposed to ES cells, represents the most accurate indicator of DNA methylation, dispelling the previously suggested (10, 12) concept that tumor methylation is merely a reflection of stem-cell chromatin patterns (Supplementary Fig. S5C).

Understanding the dynamics of age-dependent de novo methylation may also provide insight into the origins of tumor-associated modification patterns (36). Although the overall levels of methylation in normal tissue are much lower than those observed in tumors (Supplementary Fig. S3), their modification target pattern is almost identical (Fig. 3B). To generate a better picture for the distribution of DNA methylation in a tissue context, we carried out single-molecule bisulfite sequencing of CpG islands using high-throughput Ion-Torrent technology (Fig. 6 and Supplementary Fig. S1A). From this data, it is clear that cells in the normal colon have an unexpectedly wide distribution of methylation levels at individual CpG islands, with some molecules being completely methylated, while others show only minimal modification.

Considering that the methylation ranking rules in total DNA must also apply in individual cells, it is likely that some cells in the colon have high levels of methylation at many CpG islands, while others are less affected. This result provides a new perspective on DNA methylation in cancer.

On the basis of these findings, one can suggest two completely different possibilities to explain tumor methylation. On one hand, tumors may originate
preferentially in cells that are prone to modification (37), perhaps due to a high level of Dnmts (38). This would imply that the “de novo” methylation seen in tumors actually originates during the process of aging, with those cells having the most extensive methylation being selected for tumor formation. This is consistent with the relatively high degree of methylation observed in normal tissue adjacent to tumors (39, 40), as well as in pre-cancerous polyps, and in keeping with experiments showing that early inhibition of DNA methylation can prevent tumorigenesis in mice (41). In this scheme, the modification pattern is mainly generated through an instructive model, with selection taking place at the cellular level. The classical alternative model is that tumors can initiate in any cell of the colon with further de novo methylation then taking place subsequently as a result of the transformation process itself.

Many recent studies have shown that chromatin modifier genes are often disrupted in a wide variety of different tumors (42), suggesting that the basic mutation process characteristic of cancer may be responsible for altering epigenetic states in these cells. While this model may provide an explanation for many chromatin changes observed in specific tumors, de novo DNA methylation appears to be mainly generated by primary mechanisms that may actually begin in normal tissues prior to the onset of tumorigenesis and its accompanying enhanced mutability.

Many studies have demonstrated that de novo methylation is associated with gene repression in the tumor, implying that this modification plays a role by inhibiting specific genes that are initially expressed in the normal tissue.
Methylation of well-known tumor suppressor genes represents a key example of this proposed regulatory pathway but other genes that affect tumor-associated traits may also be involved (6, 43). In keeping with this concept, removal of DNA methylation by treatment with 5azaC appears to bring about the reversion of cancer cell lines to a more normal phenotype (44).

While this may represent one aspect of methylation, the extensive pattern of DNA modification observed in the tumors analyzed in this study strongly suggests that this modification process could have a more multi-faceted effect on tumorigenesis. Interestingly, over 90% of de novo methylated genes appear to be already repressed in normal tissue (13, 14), and this has been confirmed in this study, as well (mHG P < 10^{-55}, Supplementary Fig. S7A). This raises the intriguing possibility that DNA methylation may actually play a role in preventing the activation of target genes in the tumor (45) rather than repressing genes that were previously turned on in the normal tissue. This idea would also be consistent with the mechanistic strategy of de novo methylation as it occurs during normal development in vivo (46).

One of the most striking features of abnormal methylation is that it targets many genes involved in development, morphogenesis, organogenesis or differentiation (Supplementary Fig. S7B). Inhibition of these loci by methylation may thus have a profound effect by preventing proper lineage progression, thereby promoting continued cell division and in this way contribute to tumor formation. Since targets originally packaged with H3K27me3 in normal tissue appear to lose this mark as they acquire
increased DNA methylation (14, 47), the end result is that genes normally repressed by a relatively flexible mechanism directed by polycomb binding become stably and perhaps irreversibly silenced through modification of the underlying DNA itself (45) (Supplementary Fig. S7A).

In the case of colon, tumors probably originate in proliferating cells of the crypt prior to their conversion to colonic epithelium (48). It is likely that this epithelial differentiation involves activating a defined set of genes that are initially repressed in cells of the crypt (11, 49). As we have demonstrated, many CpG islands become over-methylated in the colon as a function of aging, including the promoters of key developmental genes. We propose that this methylation, which undoubtedly occurs and is maintained in stem cells of the crypt, serves as an unwanted over-riding repression mechanism which not only prevents the activation of tumor response genes (34), but may also inhibit the normal molecular process of differentiation, thus forcing cells to remain in their proliferative mode, and, as such, begin the process of tumorigenesis. Much less is known about the formation of tumors in other tissues, but high-throughput bisulfite analysis of individual CpG islands suggest that like colon, cells in the liver or lung have a relatively non-uniform level of DNA methylation (Supplementary Fig. S1B), suggesting that the connection between age-related DNA modification and cancer may be similar in these cell types, as well.
ACKNOWLEDGMENTS AND GRANT SUPPORT

We would like to thank Ilana Keshet for her help on many aspects of this project, the NIH Roadmap Epigenomics Mapping Consortium (http://nihroadmap.nih.gov/epigenomics/) and The Cancer Genome Atlas (TCGA - http://cancergenome.nih.gov/) for use of data. The work was supported by grants from the ERC, the ISF, I-CORE, The Rosetrees Trust, Lew Sanders and Norton Herrick.
REFERENCES


FIGURE LEGENDS

Figure 1. De novo methylation in cancer

(A) Heatmap showing methylation (mDIP) of all constitutively unmethylated CpG islands in fetal tissues (“background islands”) ordered according to CRC methylation levels. The 1,524 most methylated CpG islands in CRC are shown in the right panel, together with a graph representing the average level of methylation (IMS) per array. Fetal brain (Br), fetal liver (Li), fetal skeletal muscle (Mu), fetal colon (Co), polyp (Po), colorectal cancer (CRC), hepatocellular carcinoma (HCC), lung cancer (LC), brain tumor (BT). (B) The methylation distribution of these same islands (red) is shown for CRC and lung cancer (LC) as well as for HumanMethylation450 β-value data from pancreatic adenocarcinoma (PAAD, The Cancer Genome Atlas - TCGA) and breast cancer (BrC) (19). The set of constitutively unmethylated (~13,000) and methylated (~2,500) islands is used as an internal reference (grey) in the same DNA sample. The number of samples analyzed is specified in parentheses. P values were determined using the minimal Hypergeometric test (20) (mHG, Supplementary Methods).

Figure 2. Methylation of cancer islands in normal tissues during aging

Methylation (IMS) distribution of the 1,524 CRC islands from Figure 1 (black) in DNA from fetal, young (~25 y.o.) and old (~65 y.o.) colon (A) and liver (B). A similar distribution graph is shown for methylation of these islands in CD4+ T-cells derived from a newborn, a 26 y.o. and a 103 y.o., as measured by whole-genome bisulfite sequencing (WGBS, % methylation) (26). (C) Each graph includes the constitutively unmethylated and methylated islands (grey)
for reference. The number (n) of samples analyzed is specified. P values for every graph and for the shift in methylation with aging (bold) were determined using mHG test (Supplementary Methods).

**Figure 3. Coefficient of CpG island methylation**

The set of 1,524 CRC-methylated CpG islands were ranked according to their IMS in normal colon (lane 1). (A) Methylation heatmap for DNA from different cancer samples, all ordered according to the same specific colon-rank. Each sample was individually and conditionally colored according to its relative methylation levels. Lanes 1-5 show data for tissues analyzed by mDIP in this study. Lane 1 - normal colon (n=3 samples), lane 2 - colorectal cancer (n=3), lane 3 - hepatocellular carcinoma (n=2), lane 4 - lung cancer (n=4), lane 5 - brain tumor (n=2). Published data for a large number of tumors of different cell type are shown in the succeeding lanes. See Supplementary Table S2 for details of all samples analyzed. (B) The set of 1,524 CRC islands were partitioned into 20 equal groups (percentiles) according to their ranking. Each point represents the average IMS of all the islands in that group as determined by mDIP on different tissue samples.

**Figure 4. Abnormal de novo methylation is correlated with histone methylation**

(A) Constitutively unmethylated background CpG islands were sorted according to their H3K27me3 concentration in normal colon (29), partitioned into ten equal groups (percentiles) of increasing H3K27me3 and their average IMS (±SD) in CRC then determined. (B) 4,251 CpG islands that are
considered enriched for H3K27me3 (score > 3.5) were sorted according to their H3K4me3 score in normal colon (29), grouped into bins and graphed against their average IMS (±SD) in colorectal cancer. CpG islands marked exclusively with H3K4me3 were found to be relatively unmethylated (AVG IMS =-0.77) . (C) A methylation formula (predictor) for calculating the predicted susceptibility of CpG islands to de novo methylation as a function of H3K27me3 and H3K4me3 concentrations in normal colon was derived using 70% of the data (learning set, n = 6,543 CpG islands, see Supplementary Methods). The remaining 30% of the data (test set, n = 2,803 CpG islands) was used to validate this predictor. Receiver Operating Characteristic (ROC) curves show the accuracy of this formula for predicting susceptibility to methylation in CRC, as compared to H3K27me3 levels or the concentration of CpG dinucleotides (observed/expected) in the sequences (CpGO/E) alone. This colon data was also a good predictor for other tumors, including Breast (AUC=0.88), Lung (0.92), AML (0.90) and GBM (0.86) using the Infinium methylation platform (Supplementary Table S2).

Figure 5. Tissue specific methylation

(A) CpG islands specifically methylated (IMS > 0.5) in lung cancer (LC, n=4), liver cancer (HCC, n=2) or breast cancer (19) (BrC, β-value > 0.45, n=8) as compared to CRC (IMS < 0.3) were selected. Histograms show average levels of H3K4me3 (29) (Z-scores, see Supplementary Methods) in normal lung, liver or breast (myo-epithelial cells) as compared to colon for these tumor specific sites (blue) as well as for sites commonly methylated (red) both in the test tumor and CRC. Each bar notes the average methylation value for
these CpG islands in the corresponding tumor DNA samples. P values (student’s t-test) are shown. (B) Methylation heatmap of samples from primary Hepatocellular cancer (HCC) and CRC as well as a CRC metastasis to the liver (M) showing a set of 41 CpG islands that are differentially methylated in normal colon versus liver.

**Figure 6. Molecule-by-molecule pattern of de novo methylation in the colon.**

(A) Bisulfite deep-sequencing (see Supplementary Methods) was performed on a colon sample (~65 y.o.) from an island (15 CpG sites) located at the 5’ end of the EPHA5 gene (blue – unmethylated, yellow – methylated).

(B) Histogram shows the observed distribution according to the number of CpG sites methylated per molecule as compared to what would be expected stochastically (binomial distribution) as calculated from total average methylation. P value was calculated using the Kolmogorov-Smirnov test (Supplementary Methods).
Figure 2

A  Colon  P<10^{-300}  
n = 2  P<10^{-78}

B  Liver  P<10^{-90}  
n = 3  P<10^{-9}

C  CD4+ T cells  P<10^{-300}  
n = 1  P<10^{-265}

Distribution (percent)

Fetal

Young

Old

IMS (mDIP)

IMS (mDIP)

% Methylation (WGBS)
Figure 4

A

![Graph A](image)

B

![Graph B](image)

C

![Graph C](image)
Molecular rules governing de novo methylation in cancer

Deborah Nejman, Ravid Straussman, Israel Steinfeld, et al.

Cancer Res  Published OnlineFirst January 22, 2014.