Cytosine methylation profiling of cancer cell lines

Mathias Ehrich*1, Julia Turner*, Peter Gibbs¹, Lara Lipton², Mara Giovannetti², Charles Cantor*¹, and Dirk van den Boom*¹

*Sequenom, Inc., 3595 John Hopkins Court, San Diego, CA 92109; and ¹Ludwig Institute for Cancer Research, P.O. Box 2008, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia

Contributed by Charles Cantor, December 27, 2007 (sent for review December 14, 2007)

DNA-methylation changes in human cancer are complex and vary between the different types of cancer. Capturing this epigenetic variability in an atlas of DNA-methylation changes will be beneficial for basic research as well as translational medicine. Hypothesis-free approaches that interrogate methylation patterns genomewide have already generated promising results. However, these methods are still limited by their quantitative accuracy and the number of CpG sites that can be assessed individually. Here, we use a unique approach to measure quantitative methylation patterns in a set of >400 candidate genes. In this high-resolution study, we employed a cell-line model consisting of 59 cancer cell lines provided by the National Cancer Institute and six healthy control tissues for discovery of methylation differences in cancer-related genes. To assess the effect of cell culturing, we validated the results from colon cancer cell lines by using clinical colon cancer specimens. Our results show that a large proportion of genes (78 of 400 genes) are epigenetically altered in cancer. Although most genes show methylation changes in only one tumor type (35 genes), we also found a set of genes that changed in many different forms of cancer (seven genes). This dataset can easily be expanded to develop a more comprehensive and ultimately complete map of quantitative methylation changes. Our methylation data also provide an ideal starting point for further translational research where the results can be combined with existing large-scale datasets to develop an approach that integrates epigenetic, transcriptional, and mutational findings.

Results

We studied DNA methylation by gene-specific amplification of bisulfite-treated DNA followed by in vitro transcription and matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) analysis (9). The amplification regions were primarily designed to cover CpG islands (CGIs) overlapping with the 5’ UTR of the target genes. When no CGI was annotated by the UCSC genome browser in close proximity, we used the sequence directly surrounding the 5’ UTR for primer design, and if no CpG dinucleotides were found in this region, we used the next upstream flanking CGI in a 10-kb window.

The initial methylation data were filtered to exclude poor quality measurements. Poor quality was defined as amplicons with data available for <75% of all samples. These regions were excluded from further analysis. The filtered dataset contained 531 amplification regions, representing 430 genes. For excluded amplicons, PCR was identified as the leading cause of reaction failure.

All autosomal chromosomes and the X chromosome are represented in the current gene set. The median amplicon length was 413 bp (range = 171–683 bp) and the median CpG content per amplicon was 33 CpG/amplicon (range = 6–81 CpG per amplicon). For each sample, a total of 11,723 CpG sites were analyzed. The sample with the most data was a colon cancer cell line, and the sample with the least data was a breast cancer cell line. The cancer cell lines are compiled by the National Cancer Institute (NCI) as the NCI-60 panel, which has been widely used for in vitro anticancer drug testing.

Over the years, the NCI-60 set has become one of the best characterized cell line sets available. These have also been analyzed by using a variety of methods including transcriptional profiling (refs. 4 and 5, and see http://dtp.nci.nih.gov/index.html), spectral karyotyping (6), and proteomic profiling (7). In addition, cytotoxicity profiling has been documented for >100,000 chemical compounds by the NCI’s Developmental Therapeutics Program (DTP) (http://dtp.nci.nih.gov/index.html and ref. 8).

Here, we report the results of a large-scale DNA methylation profiling study, which includes the quantitative analysis of >500 genomic target regions representing >400 genes in 59 cell lines with confirmation of a subset of targets in 48 colorectal cancer/normal tissue pairs. The resulting data provide a comprehensive panel of cancer-related DNA methylation changes and can be integrated with previous datasets on mutational, transcriptional, and proteomic profiles to obtain a more comprehensive understanding of neoplastic transformations.

Author contributions: M.E. designed research; M.E. and J.T. performed research; M.E., P.G., L.L., and M.G. contributed new reagents/analytic tools; M.E. analyzed data; and M.E., P.G., L.L., M.G., C.C., and D.v.d.B. wrote the paper.

Conflict of interest statement: M.E., J.T., C.C., and D.v.d.B. are shareholders and full-time employees of Sequenom, Inc.

Freely available online through the PNAS open access option.

1To whom correspondence may be addressed. E-mail: mehrich@sequenom.com or ccantor@sequenom.com.

This article contains supporting information online at www.pnas.org/cgi/content/full/0712251105/DC1.

© 2008 by The National Academy of Sciences of the USA.
values in tissue samples, 76% of all amplicons showed mean methylation levels between 15% and 85%. In cancer and normal samples. CpG density of amplicons was calculated as the fraction of CpG nucleotides within the total amplicon sequence. The mean methylation value for each amplicon was generated by using all individual CpG sites’ methylation values. Amplicons with >10% CpG content are likely to have lower methylation values in normal tissues. In cancer cell lines, DNA methylation is observed more frequently in these amplicons. (c) Amplicons were binned based on their average methylation values. Each bin contained amplicons within a 5% range of methylation values. Bins from 15% to 85% average methylation contain more amplicons in the set of cancer cell lines. (d) Histogram of methylation differences. For each amplicon, we calculated the difference in mean methylation between the group of normal samples and the group of tumor cell lines. Positive values translate to hypermethylation in cancer cell lines, whereas negative values indicate that the mean methylation was higher in the group of normal samples. The distribution of methylation differences is skewed toward hypermethylation. (e) DNA methylation in relation to the closest 5’ UTR. The distance from the 5’ UTR was calculated for every individual CpG site. Each data point contains 1,770 individual methylation values for the cancer cell lines and 180 values for the normal samples. It is necessary to adjust the number of data points in each group because of the difference in sample numbers in the two sets. A window of 1 kbp around the 5’ UTR shows low methylation values in the normal and the cancer cell-line samples. Methylation values in the cancer cell-line samples are generally elevated.

Stability. All bisulfite-based methylation analysis methods suffer from a considerable amount of measurement variability introduced by the chemical treatment of genomic DNA. To assess the degree of this inherent variability, we previously dissected the method into its four components and measured the variability for each step in the process (10). The results demonstrated that the greatest source of process-dependent variability is the bisulfite conversion reaction (SD = 10–15%). To determine whether the previous results are applicable to the model system used in this study, we performed duplicate measurements of four control DNAs in 96 amplicons and observed sufficient data stability ($R^2 = 0.98$, SI Fig. 5). We also were interested in evaluating the effect of primer design on the quantitative measurements. We designed two different but overlapping amplification regions for the ERBB2 gene. The quantitative values from both reactions were almost identical and showed a high correlation ($R^2 = 0.96$; see also Fig. 1a).

Methylation Data Distribution for Cell Lines. The analyzed target regions were primarily designed within annotated CGIs, which are frequently unmethylated in normal tissue (2). In this part of the analysis, the methylation values, which were measured for each sample individually, were averaged for the group of normal samples and the group of cancer cell-line samples. Our data are in agreement with previous findings for normal tissue samples and confirm that CGIs with a CpG density >10% are generally unmethylated (<15% methylation) (11) (Fig. 1b). In the normal tissue samples, 76% of all amplicons showed mean methylation values <15% methylation; only 5% of amplicons showed methylation levels >85% methylation, and the remaining 19% of amplicons showed methylation levels between 15% and 85%. In the group of cancer cell lines, we observed a shift toward medium (between 20% and 80%) but not high (>80%) methylation levels.

Recent studies have shown a decrease of epigenetic marking in a 1-kb window around the transcription start site. In active Drosophila melanogaster promoters, histone occupancy is decreased, and in normal human tissue samples, DNA methylation is reduced within this core region (11, 12). To further investigate this relationship, we mapped the distance from the 5’ UTR for each measured CpG in the dataset (700,000 data points). CpG methylation in normal samples showed the expected core window of unmethylated CpG sites within 1 kb around the 5’ UTR (Fig. 1). In cancer cell lines, methylation averages are generally elevated, but the same symmetrical methylation decrease is observed.
variations (8, 13). It was reported that gene-expression patterns from gene-expression studies or clustering based on copy-number alterations in breast cancer samples do not show strong similarities to each other. Methylation patterns of ALL samples and one sample derived from large-cell immunoblastic lymphoma cluster together. Methylation patterns of normal samples are minimal, and, despite their different tissues of origin, they cluster tightly together. The cancer cell-line samples are characterized by consistent low methylation levels. The normal samples are likely to comprise DNA from non-tissue-specific patterns (5).

Confirmation of Cell-Line Results in Colon Cancer Samples. It remains unclear whether the observed methylation differences are a consequence of the manipulation of cell lines during in vitro growth or whether they represent cancer-specific characteristics. Accordingly, our model system bears the risk of overinterpreting the detected methylation differences. To explore the validity of our findings, we chose the colon cancer cell-line models for confirmation in clinical samples. A set of 50 genes was selected that showed significant differential methylation ($\Delta M > 20\%$, $P < 0.001$, two-sided t test) in the colon cancer cell lines. To assess the specificity of our finding, we also selected 14 genes that did not show any cell line methylation differences. We investigated the methylation status of these genes in 48 matched sample pairs of colon cancer tissue and adjacent normal colon tissue. The majority of patients were male (M) = 30, female (F) = 18; the median age at diagnosis was 65 years (range 46–83). Forty-four patients had experienced local or distant cancer recurrence, and all stages (I–IV) were evenly represented. The analysis of methylation differences between the normal and cancer tissue samples confirmed the previous cell-line findings for the majority of genes. In the set of differentially methylated genes, we found 42 of 50 (84%) genes to be significantly differentially methylated in the clinical tissue samples. Additionally, all 14 genes that did not show a methylation difference were still not differentially methylated in the clinical samples (SI Table 2).

We next used the methylation patterns to characterize relationships among the colon cancer samples and to explore potential associations to their clinical features. None of the clinical features showed a strong correlation to the resulting colon cancer methylation groups (SI Fig. 7a). We explored the diversity among the colon cancer samples and their tissue counterparts by using hierarchical clustering (Fig. 3). As expected, the normal tissues grouped with the normal colon tissue samples, and the colon cancer cell lines grouped with the colon cancer tissue samples. However, the segregation of normal and colon cancer tissue samples was not perfect. A subset (n = 10) of colon cancer samples is found in the group of normal tissue samples (Fig. 3). Although the colon cancer samples cluster together with the colon cancer cell-line samples, there are obvious differences in their methylation patterns. The cell-line samples tend to show extreme methylation values. The majority of CpG sites are either not methylated or they are fully methylated. The colon cancer tissue samples show much more heterogeneity and a much broader spectrum of methylation values. In general, methylation differences between normal and cancer were larger in the cell-line models compared with the clinical tissue samples. We find that although we can confirm methylation differences, which have been identified between normal samples and colon cancer cell lines, the magnitude of the effect is much smaller in clinical tissue specimens. This effect might be caused by heterogeneity observed. Thus, these results confirm previous findings and expand their applicability to cancer cell lines (see also SI Text and SI Fig. 6).

Methylation-Based Cell-Line Clustering. To examine relationships among cell lines and CpG sites, we performed an unsupervised two-dimensional hierarchical clustering analysis, which provides an unbiased view on these relationships (Fig. 2).

The most notable feature of the resulting sample clusters is a separation into different terminal branches of the tree of all normal tissue samples from all cancer cell-line samples. The normal samples are likely to comprise DNA from non-tissue-specific sources, like connective tissue, which might account for some of the difference between the normal samples and the cancer cell-line samples. The differences between the various normal samples are minimal, and, despite their different tissues of origin, they cluster tightly together. The cancer cell-line samples are found in two major groups. A group of 19 samples splits as the first branch in the tree. This cluster is characterized by higher methylation levels in a number of genes and includes all seven glioblastoma cell lines SNB-19 and U251 are derived from one and the NCI/ADR-RES line and the ovarian line OVCAR-8 are derived from another individual (13). Our results support these findings. Both cell-line pairs have remarkably similar methylation patterns and are located next to each other in end-terminal branches of the dendrogram. Also, in concordance with expression data from another study, our findings for SN12C (renal CA), PC3 (prostate CA), and DU-145 (prostate CA) did not show tissue-specific patterns (5).
was partially methylated when the difference in methylation values to filter out false positives. We classified a gene to be differentially methylated between each type of cancer cell line and normal tissues. We then examined the individual groups and determined the minimum difference of 20% as an additional selection criterion, though this does not completely erase the issue, we included a one for each cell line), the results will contain false positives that in specific tumor types only.

Differentially Methylated Genes. Tissue-specific DNA methylation has been observed in normal tissues (17), and several cancer-specific methylation markers have been described. The specificity of such cancer markers to a single cancer type remains controversial. Several markers have been found to be differentially methylated in multiple cancer types. These markers might be more universally involved in the progression of cancer. Here, we attempted to identify groups of genes that are differentially methylated between each type of cancer cell line and normal tissues. We then examined the individual groups and determined which genes overlap in multiple cancer types and which are found in specific tumor types only.

Because several genetic loci were tested in many separate runs (one for each cell line), the results will contain false positives that arise from multiple testing in high-dimensional datasets. Although this does not completely erase the issue, we included a minimum difference of 20% as an additional selection criterion to filter out false positives. We classified a gene to be differentially methylated when the difference in methylation values between the normal samples and the subset of cancer cell lines was >20% and the \( P \) value for a two-sided \( t \) test was <0.001.

The results for the group of leukemia samples and the group of prostate cancer cell lines should be viewed with some caution, because the group of leukemia samples represent a biologically heterogeneous group. Prostate cancer is the smallest subgroup containing only two cell-line samples, which, in addition, do not show prostate cancer-specific gene-expression signatures (5). Hence, their results are less likely to be representative.

A total of 71 genes were statistically significantly hypermethylated in at least one tumor type. A large fraction of these genes \((n=30, 42\%)\) were found only in one tumor type, and nearly 10% were found in more than five tumor types \((TSPYL, PAX8, LEP, PHOX2B, \) and \(TMPRSS2\) were found in five tumor types; \(MYOD1\) was found in six tumor types; \(PA\) was found in eight tumor types).

Seven genes were hypomethylated \((TCL1A, SLC22A2, TRPM5, IGF2, PEG3\) were found in one tumor type; 2\(KCNQ1\), \(DLK1\) were found in two tumor types). As suggested from our previous analysis depicted in SI Fig. 8, CNS neoplasms \((n=4)\) and melanomas \((n=3)\) had the highest number of hypomethylated genes. Interestingly almost all of the hypomethylated genes are known to be imprinted, which might point to a loss of imprinting in these cases (SI Table 3).

### Table 1. Fraction of PRC2 target and test for statistical significance

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Fraction PRC2 target in nonsignificant genes</th>
<th>Fraction PRC2 target in significant genes</th>
<th>( P ) value, Fisher’s exact test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blood</td>
<td>0.15</td>
<td>0.33</td>
<td>0.237</td>
</tr>
<tr>
<td>Breast</td>
<td>0.12</td>
<td>0.67</td>
<td>0.000</td>
</tr>
<tr>
<td>CNS</td>
<td>0.12</td>
<td>0.52</td>
<td>0.000</td>
</tr>
<tr>
<td>Colon</td>
<td>0.12</td>
<td>0.56</td>
<td>0.000</td>
</tr>
<tr>
<td>Lung</td>
<td>0.13</td>
<td>0.73</td>
<td>0.000</td>
</tr>
<tr>
<td>Ovarian</td>
<td>0.15</td>
<td>0.70</td>
<td>0.000</td>
</tr>
<tr>
<td>Prostate</td>
<td>0.18</td>
<td>0.71</td>
<td>0.003</td>
</tr>
<tr>
<td>Renal</td>
<td>0.12</td>
<td>0.50</td>
<td>0.000</td>
</tr>
<tr>
<td>Skin</td>
<td>0.17</td>
<td>0.44</td>
<td>0.003</td>
</tr>
<tr>
<td>Normal</td>
<td>0.19</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

NA, not available.

**PRC2 Target Identification for Colon Cancer and All Others.** A retrospective analysis of DNA methylation by Widschwendter et al. provided evidence that genes targeted by the Polycomb repressive complex 2 are silenced in human colon cancer (18).

We were able to retrieve information about PRC2 binding sites for 440 amplicons, including 79 amplicons with more than one PRC2 -binding site. We calculated the fraction of amplicons that contain one or more PRC2 binding sites for both: The set of genes that did not show significant methylation differences and the set of genes that did show significant methylation differences in cancer cell lines versus normal tissue. Our findings show a significant \((P<0.001, \) Fisher’s exact test) enrichment for PRC2 targets in the set of significantly hypermethylated genes in six of the nine tumor types. In the group of tumor cell lines with sufficient numbers of samples (excluding leukemia and prostate), we find that only the melanoma-specific gene set is not enriched for PRC2 targets. All other tumor types are 2- to 6-fold enriched for PRC2 targets (Table 1). A graphical representation of gene–tumor associations reveals that highly connected genes also tend to be PRC2 targets (Fig. 4 and SI Fig. 9).

**Discussion**

Our dataset provides a comprehensive assessment of the methylation status for a group of cancer-related genes in the NCI 60 cell-line set. Although it does not provide a genome-wide representation it will be helpful in understanding the interaction among epigenetic, transcriptional, and mutational status.
methylation changes has to be attributed to cell-line transformation of our study. It remains unclear what fraction of observed methylation differences between colon cancer tissue samples and normal colon cancer samples were generally smaller than those observed in cell lines. To exclude overinterpretation of the results based on nonspecific promoter hypermethylation, we also analyzed a set of genes that did not show methylation differences in the cell-line model and one gene that showed hypomethylation. We confirmed indifferent methylation in 12 genes with low methylation levels (<20%) and in two genes with higher methylation levels (RUNX3 = 40% and TRPM5 = 80%). KCNQ1, a gene known to be imprinted, was hypomethylated in colon cancer cell lines and colon cancer tissues.

We further cross-compared our results with findings from another comprehensive methylation study performed on colon cancer tissue samples (16). The results from both studies are highly concordant. The few observed disparities might simply be caused by using different interrogation sites. Unfortunately, we were unable to investigate this issue, because the publication did not reveal the exact genomic location of the analyzed amplicons.

One of the most interesting results of this study was that we were able to integrate our results into a recently developed biological model of the polycomb repressive complex 2. Most functional studies have so far been performed on stem cells and association studies mainly focused on colon cancer samples or colon cancer cell lines (18, 23). Our study provides quantitative methylation data for at least seven different tumor types (excluding leukemias and prostate cancer). For every tumor type, a different but not exclusive set of genes was identified that showed higher methylation levels compared with normal samples. Interestingly, six of these seven sets of hypermethylated genes were enriched for PRC2 target genes. These findings suggest that PRC2 target methylation is a common event in cancer and is not limited to individual tumor types.

Our study also confirms the feasibility of the current method for large-scale quantitative high-resolution mapping of DNA methylation. This approach provides an ideal method for large-scale validation of genome-wide methylation studies or to expand the current dataset by analyzing additional regions.

We hope that the availability of this large-scale methylation dataset will initiate interdisciplinary research that integrates multiple available datasets. This information might help to design and execute experiments that will elucidate some of the questions that remained unsolved.

Methods

Cells Lines and Tissue Samples. The NCI-60 panel of 59 cell lines was provided by NCI Division of Cancer Treatment and Diagnosis Tumor Repository (NCI-DCTD). For individual cancer types, normal cell lines or tissue-derived genomic DNA was commercially purchased. Tissue samples from brain (catalogue nos. D1234035 and D123062), breast (D1234086), colon (D1234090), kidney (D1234142), leukocyte (D1234148), and lung (D1234152) were obtained from BioChain.

Clinical Colon Cancer Tissue. Tissue colorectal cancer and normal colon tissue samples were obtained from the Royal Melbourne Hospital (RMH) Tissue Bank as part of the Ludwig Colon Cancer Initiative biomarker project. Samples were obtained with informed consent under the approved protocol from patients having a resection for histologically confirmed colorectal cancer, and the
normal matched tissue was obtained from the same resection specimen at a site adjacent to the tumor.

Tissue samples were snap-frozen in liquid nitrogen within 30 min of collection and stored in a −80°C freezer. Matched tissue sample pairs were cut into 2-mm (see www.geneimprint.com) cube sections weighing 10–15 mg. After manual dissection, DNA was extracted from the tissue sections by using a Qiagen DNeasy blood and tissue kit. Briefly, samples were first lysed by using a Proteinase K digestion for 3 h at 56°C followed by selective binding of DNA to a membrane; final steps involving a spin-column procedure allowed for the washing and subsequent elution of DNA with precipitated DNA resuspended in a buffer solution. DNA was quantified by using a photometer, with 260/280 ratio in the range of 1.7–2.0. DNA samples were normalized to a concentration of 50 μg/ml.

Clinical data regarding patients and histopathological data for tumor samples were derived from the Australian Comprehensive Clinical Outcomes and Research Database (ACCORD) of Bio21 Molecular Medicine Informatics Model (Bio21:MMIM) (http://mmim.ssg.org.au). This is a unique resource of samples were derived from the Australian Comprehensive Clinical Outcomes/H9262.

Concentration of 50

DNA Methylation Analysis. Bisulfite treatment. Genomic DNA sodium bisulfite conversion was performed by using EZ-96 DNA methylation kit (Zymo Research). The manufacturer’s protocol was followed by using 1 μg of genomic DNA and the alternative conversion protocol (a two-temperature DNA denaturation.

Methylation analysis. Sequenom’s MassARRAY platform was used to perform quantitative methylation analysis. This system utilizes MALDI-TOF mass spectrometry in combination with RNA base specific cleavage (MassCLEAVE). A detectable pattern is then analyzed for methylation status. PCR primers were designed by using Methprimer (www.urogene.org/methprimer/). When it was feasible, amplicons were designed to cover CGIs in the same region as the 5’ UTR. For each reverse primer, an additional T7 promoter tag for in vivo transcription was added, as well as a 10-mer tag on the forward primer to adjust for melting-temperature differences. The MassCLEAVE biochemistry was performed as previously described (for details also see SI Text) (9). Mass spectra were acquired by using a MassARRAY Compact MALDI-TOF (Sequenom) and spectra’s methylation ratios were generated by the Epityper software v1.0 (Sequenom).

Statistical Methods. All statistical analyses were performed by using the R statistical environment (www.r-project.org). Distances from gene start sites have been calculated by using the RMySQL package and the SQL database version of the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway). Two-dimensional clustering has been performed by using the heatmap.2 function in the gregmisc package. Classical multidimensional scaling has been performed by using the cmdscale function, and visualization was done through the scatterplot3d function in the same-named package. Tests for statistical significance (t test, Wilcoxon test, and Fisher’s exact test) have been used with standard function in R build into the stats package.

For sequence-motif detection, we used a permutation-based method. We randomly sampled n sequences from the pool of all analyzed sequences (n is equal to the number of sequences in the low- or high-methylation group). We then counted how often every possible 6-mer (n = 4,096) is present in the sampled subset. One thousand permutations were performed for each analysis. A sequence motif was identified as being overrepresented if it occurred more often in the analyzed group of sequences than in any of the 1,000 random draws. Graphical representation of the gene-tissue relationships was performed by using the dot algorithm implemented in Graphviz.