Cytosine Methylation Profiles as a Molecular Marker in Non–Small Cell Lung Cancer

Mathias Ehrich,[1] John K. Field,[2] Triantafillos Liloglou,[2] George Xinarianos,[2] Paul Oeth,[1] Matthew R. Nelson,[1] Charles R. Cantor,[1] and Dirk van den Boom[1]

[1]SEQUENOM, Inc., San Diego, California and [2]The University of Liverpool Cancer Research Centre, Roy Castle Lung Cancer Research Program, Liverpool, United Kingdom

Abstract

Aberrant promoter methylation is frequently observed in different types of lung cancer. Epigenetic modifications are believed to occur before the clinical onset of the disease and hence hold a great promise as early detection markers. Extensive analysis of DNA methylation has been impeded by methods that are either too labor intensive to allow large-scale studies or not sufficiently quantitative to measure subtle changes in the degree of methylation. We used a novel quantitative DNA methylation analysis technology to complete a large-scale cytosine methylation profiling study involving 47 gene promoter regions in 96 lung cancer patients. Each individual contributed a lung cancer specimen and corresponding adjacent normal tissue. The study identified six genes with statistically significant differences in methylation between normal and tumor tissue ($P < 10^{-6}$). We explored the quantitative methylation data using an unsupervised hierarchical clustering algorithm. The data analysis revealed that methylation patterns differentiate normal from tumor tissue. For validation of our approach, we divided the samples to train a classifier and test its performance. We were able to distinguish normal from lung cancer tissue with >95% sensitivity and specificity. These results show that quantitative cytosine methylation profiling can be used to identify molecular classification markers in lung cancer. (Cancer Res 2006; 66(22): 10911-8)

Introduction

Lung cancer accounts for 30% of all cancer deaths in industrialized nations and remains the leading cause of cancer mortality (1). Most patients with non–small cell lung cancer (NSCLC) remain symptom free until later stages and present with advanced disease at the time of diagnosis. Like in many other neoplastic diseases, the survival rate is critically influenced by the progression of the tumor. Whereas the 5-year survival rate for patients with a stage I tumor is ~70%, it decreases to 30% for stage IIIa (2). There is a need for improved clinical stratification methods that can identify patients with early-stage disease and identify those with high risk of recurrence (3). Conventional methods, including spiral computed tomography, sputum cytology, histopathology, or tumor-node-metastasis classification, have thus far failed to overcome limitations in early detection and risk assessment. On the contrary, a variety of novel molecular methods, such as detection of K-ras (4) and p53 mutation status (5, 6), microsatellite instability (7), protein profiling (8, 9), and especially gene expression profiling (10, 11), have shown very promising results.

Other potential molecular markers in lung cancer are epigenetic changes of the DNA (12–14). Alterations in DNA methylation and related chromatin changes have been reported as an early event in carcinogenesis and hence hold the promise of being useful as one of the earliest detection markers available (15). To date, researchers in the field have focused on detection of hypermethylated DNA as a marker for tumor progression using methylation-specific PCR (MSP; ref. 16). MSP is an easy to use method with very high sensitivity, but it suffers from limited versatility. The method only allows assessment of the presence or absence of methylation at the CpG sites enclosed in the PCR primer hybridization site. Consequently, tissues (tumors) with different fractions of methylated DNA cannot be differentiated; relative changes in the amount of methylated DNA usually remain invisible. Different methods, such as semiquantitative real-time PCR or bisulfite sequencing, are now being used to obtain more quantitative results. Current methods are limited by restricted CpG coverage per assay, poor quantitative resolution, or a combination of both. Hence, large-scale studies that evaluate quantitative methylation for multiple CpG sites in various gene regions and a large number of samples are rare.

A novel technology has been introduced recently that aims to overcome these shortcomings and allows large-scale cytosine methylation profiling (17). Here, we used this technology to quantify the degree of cytosine methylation at 47 genes in tumor and adjacent normal tissue from 96 lung cancer patients. We evaluate the feasibility of this approach to reveal practical markers for lung cancer research.

We selected 96 patients with NSCLC and a history of smoking. From each patient, we collected one specimen from the primary tumor and one specimen from adjacent normal tissue, resulting in a total of 192 samples. The patient collection consists of 34 females and 62 males, ages 41 to 87 years (median, 67 years). In this collection, 50 patients were diagnosed with squamous cell carcinoma, 43 with adenocarcinoma, 1 with large cell carcinoma, 1 with atypical carcinoma, and 1 with not further classified NSCLC. At the time of diagnosis, 45 patients had stage I disease, 27 patients had stage II disease, 20 patients had stage IIIa disease, 3 had stage IIIb disease, and 1 was undetermined. For analysis purposes, this patient collection was randomly divided into separate training and test sets, matched for age, sex, histology, and disease stage. These groups are summarized in Table 1.

We analyzed 47 preselected genes for promoter methylation. The genes were either chosen based on their biological relevance in cell

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Mathias Ehrich, Molecular Biology, SEQUENOM, Inc., 3595 John Hopkins Court, San Diego, CA 92121. Phone: 858-202-9068; Fax: 858-202-9084; E-mail: mehrich@sequenom.com.

doi:10.1158/0008-5472.CAN-06-0400

Table 1. Clinical features and test statistic for all samples after they have been separated into training and test set

		Training set, n = 49 (%)	Test set, n = 43 (%)	Test statistic
Gender	(Male)	65 (67)	53 (62)	χ^2 = 0.58; df = 1; P = 0.448
Histology	Adenocarcinomas	44 (45)	39 (45)	χ^2 = 6.34; df = 4; P = 0.175
	Atypical carcinoid	0 (0)	2 (2)	
	Large cell lung carcinoma	2 (2)	0 (0)	
	NSCLC	0 (0)	2 (2)	
	Squamous cell carcinomas	51 (53)	43 (50)	
T status	1	6 (6)	2 (2)	χ^2 = 4.43; df = 3; P = 0.219
	2	80 (84)	70 (81)	
	3	6 (6)	12 (14)	
	4	3 (3)	2 (2)	
N status	0	51 (54)	47 (55)	χ^2 = 1.92; df = 2; P = 0.383
	1	29 (31)	31 (36)	
	2	15 (16)	8 (9)	
Stage	1	45 (47)	43 (50)	χ^2 = 0.55; df = 3; P = 0.908
	2	27 (28)	21 (24)	
	3a	20 (21)	20 (23)	
	3b	3 (3)	2 (2)	
Follow-up in months	(min/mean/max)	0.32/4.43/21.23	0.39/2.32/11.70	F = 0.31; df = 1,121; P = 0.578
Fate	(Death)	12 (17)	16 (27)	χ^2 = 2.11; df = 1; P = 0.147
Histologic differentiation	Poor	29 (31)	24 (29)	χ^2 = 14.23; df = 4; P = 0.0066
	Moderate/poor	4 (4)	18 (21)	
	Moderate	51 (54)	38 (45)	
	Well/moderate	1 (1)	0 (0)	
	Well	10 (11)	4 (5)	

Abbreviation: *df*, degrees of freedom.

adhesion and cell interaction or they have been shown to change expression levels during cancer development and progression. For each of the genes, we selected a single CpG island and preferentially focused on those CpG islands located in the promoter and 5′-untranslated region (5′-UTR). The selected target regions contained a total of 1,426 CpG positions, listed in Supplementary Table S1.

For each sample in our collection, 2 μg of genomic DNA were isolated from frozen tissue specimens using a standard phenol/chloroform protocol. The DNA was prepared for methylation analysis using a commercially available bisulphite conversion kit (see Materials and Methods for details). The bisulphite-treated DNA was then used for PCR amplification (independent of methylation status).

We measured DNA methylation using a novel technique that combines base-specific cleavage of single-stranded nucleic acids with MALDI-TOF mass spectrometry (MS) analysis of the cleavage products (17). In brief, the method starts with PCR amplification of the target region from bisulphite-treated DNA, which is followed by *in vitro* transcription to generate a single-stranded RNA molecule. The RNA strand is then cleaved base specifically in individual reactions either after U or C, determined by the usage of noncleavable nucleotides (18). The cleavage reaction driven to completion and the resulting cleavage products represent a well-defined substring of the analyzed target region, which is only dependent on the sequence context and not dependent on the reaction conditions. The cleavage products are then analyzed using MALDI-TOF MS. For analysis of DNA methylation, we examine the methylation-dependent C/T sequence changes introduced by bisulfite treatment. Those C/T changes are reflected as G/A changes on the reverse strand and hence result in a mass difference of 16 kDa for each CpG site enclosed in the cleavage products generated from the RNA transcript. The mass signals representing nonmethylated DNA and those representing methylated DNA build signal pairs, which are representative for the CpG sites within the analyzed sequence substring. The intensities of the are compared, and the relative amount of methylated DNA can be calculated from this ratio. The method yields quantitative results for each of these sequence defined analytic units, which contain either one individual CpG site or an aggregate of subsequent CpG sites. We refer to these analytic units as "CpG units."

Materials and Methods

Bisulfite treatment. Bisulfite treatment of genomic DNA was done with a commercial kit from Zymo Research Corp. (Orange, CA) that combines bisulfite conversion and DNA clean up. The kit follows a protocol from Paulin et al., 1998 (19). Briefly, in this protocol, 2 μg of genomic DNA are denatured by the addition of denatured by the addition of 3 mol/L sodium hydroxide and incubated for 15 minutes at 37°C. A 6.24 mol/L urea/2 mol/L sodium metabisulfite (4 mol/L bisulfite) solution is prepared and added with 10 mmol/L hydroquinone to the denatured DNA. The corresponding final concentrations are 5.36, 3.44, and 0.5 mmol/L, respectively. This reaction mix is repeatedly heated between 55°C for 15 minutes and 95°C for 30 seconds in a PCR machine (MJ Tetrad) for 20 cycles. Finally, a DNA purification and cleaning step is done.

PCR and *in vitro* transcription. The target regions were amplified using the primer pairs described in Supplementary Table S1. The PCRs were carried out in a total volume of 5 μL using 1 pmol of each primer, 40 μmol/L

deoxynucleotide triphosphate (dNTP), 0.1 units HotStar Taq DNA polymerase (Qiagen, Valencia, CA), 1.5 mmol/L $MgCl_2$, and buffer supplied with the enzyme (final concentration, 1×). The reaction mix was pre-activated for 15 minutes at 95°C. The reactions were amplified in 45 cycles of 95°C for 20 seconds, 62°C for 30 seconds, and 72°C for 30 seconds followed by 72°C for 3 minutes. Unincorporated dNTPs were dephosphorylated by adding 1.7 μL H_2O and 0.3 units shrimp alkaline phosphatase (SAP; SEQUENOM, Inc., San Diego, CA). The reaction was incubated at 37°C for 20 minutes and SAP was then heat inactivated for 10 minutes at 85°C.

Typically, 2 μL of the PCR were directly used as template in a 6.5 μL transcription reaction. Twenty units T7 R&DNA polymerase (Epicentre, Madison, WI) were used to incorporate either dCTP or dTTP in the transcripts. Ribonucleotides were used at 1 mmol/L and the dNTP substrate at 2.5 mmol/L; other components in the reaction were as recommended by the supplier. In the same step, the *in vitro* transcription RNase A (SEQUENOM) was added to cleave the *in vitro* transcript. The mixture was then further diluted with H_2O to a final volume of 27 μL. Conditioning of the phosphate backbone before MALDI-TOF MS was achieved by the addition of 6 mg Clean Resin (SEQUENOM). Further experimental details have been described elsewhere (18).

MS measurements. The cleavage reactions (15 nL) were robotically dispensed onto silicon chips preloaded with matrix (SpectroCHIP, SEQUENOM). Mass spectra were collected using a MassARRAY mass spectrometer (SEQUENOM). Spectra were analyzed using proprietary peak picking and spectra interpretation tools.

A description of the regions used for methylation analysis in NSCLC can be found as Supplementary Table S1.

Expression analysis. Gene expression levels were assayed for 48 paired normal/tumor samples, consistent with the samples used for methylation analysis, using real-competitive PCR in conjunction with quantitative primer extension measurements via MassARRAY (20, 21). Exact conditions for this methodology are published online[3] in the data normalization using multiplexed gene panels for quantitative gene expression analysis with MassARRAY application note, with the exception that cDNA samples unique to this study were diluted 1:10 in DNase-free water. Target genes (genes with clear methylation patterns) and internal control (genes used for normalization) were designed into separate multiplexed assays using MassARRAY QGE assay design software (SEQUENOM) based on the transcript sequences found at the Ensembl genome browser.[4] The target gene panel consisted of HUGO IDs: *SERPINB5, AQP1, CDH13, CDH5, CDKN2A, DAPK1*, and *MGP1*. The internal control gene panel consisted of HUGO IDs: *ACTB, GAPD, RPL13A, SDHA, TBP, UBC, YWHAZ, B2M, HMBS*, and *HPTR1*. Normalization was conducted using the six most stable internal control genes, for this sample set: *GAPD, RPL13A, SDHA, TBP, UBC*, and *YWHAZ*, identified using geNorm software. We then calculated the average expression of these six internal control genes for each sample. The averages were used to calculate a correction factor for baseline expression. Every expression value was corrected by its sample-specific correction factor. A pair-wise comparison of expression values shows the expected high correlation between internal control genes (Supplementary Fig. S1).

Statistical methods. We used the Wilcoxon signed-rank test, a nonparametric counterpart of the paired *t* test, to compare methylation levels between normal and tumor samples and to identify sites with statistically significant differences. The two-way hierarchical cluster analysis clustered the 96 tissue samples and 76 most variable CpG fragments (variance, >0.02) based on pair-wise Euclidean distances and the complete linkage clustering algorithm (22). The method first establishes a measure for the strength of a connection between two samples (called distance). Then, the samples get reorganized according to their relationship to each other. The algorithm "clusters" samples with a high degree of similarity into groups. The resulting dendrogram is used to visualize the results. The method presented in this article clusters CpG units along the x axis and samples along the y axis. The procedure was carried out using the heatmap.2 function of the "gplots" package using the R statistical environment (23). The tree-based classifier for the classification of tumor and normal samples was found using the J48 classification algorithm in the statistical package Weka (24). A complete four-node tree was pruned to the two-node tree that resulted in the lowest 10-fold cross validation error (25).

Results

The 1,426 CpG positions analyzed in this study comprised 757 CpG units. Among these units, 59 did not yield successful measurements. Fifty percent of the CpG units gave successful measurements for more than two thirds of all samples and 177 CpG units had good results for >90% of the samples. A total of 20 CpG units were invariant in this sample collection, being completely unmethylated in all tested samples. We found that 25% of all CpG units had an intersample variance >0.012. The majority of CpG units (n = 491) were methylated to very low degree, with average methylation below 10%, and only four CpG units had mean methylation levels above 90%. In general, normal and tumor samples showed similar levels of methylation. Differences in mean methylation levels were generally small and only 30 CpG units showed a difference >10% (Supplementary Fig. S2). We excluded nine samples with poor DNA quality, resulting in low quality measurements for >90% of the CpG units. Before conducting further analyses on these data, we removed CpG units that had >25% of data missing (n = 563) or had very low levels of intersample variability (variance, <0.02; n = 164) in the training set. The final training data set consisted of 30 CpG units from 15 genes measured in 97 samples. It is noteworthy that, although >90% off all CpG units did not pass these data quality and informativeness filters, ~30% of all genes examined are represented by one or more CpG units.

We carried out an unsupervised two-way hierarchical clustering of the CpG unit methylation and the combined tumor and normal tissues in the training set to explore any natural groupings in this data set (Fig. 1*A*). This reveals three visible clusters of samples, consisting of the following: (*a*) 9 tumor samples, (*b*) 1 normal and 34 tumor samples, and (*c*) 47 normal and 6 tumor samples. The clustering of CpG units reveals two primary clusters, separating the predominantly hypermethylated and hypomethylated units. Five genes had multiple CpG units included in this analysis. For *SERPINB5, MGMT, MGP*, and *TNA*, the corresponding CpG units tended to cluster together, showing similar intragenic methylation patterns. However, the six units corresponding to *SDK2* were divided evenly between the two clusters.

We repeated the clustering with the test set, which showed results very similar to those observed in the training set (Fig. 1*B*). The clustering of CpG units was nearly identical to those observed in the training set, with an identical representation in the two main clusters. The sample clustering also resulted in a similar discrimination between tumor and normal samples in three clusters: (*a*) 7 tumor samples, (*b*) 1 normal and 30 tumor samples, and (*c*) 42 normal and 6 tumor samples.

The patterns we observed in the cluster analyses show that methylation patterns of normal lung tissues are notably different from those observed in tumor tissues. To evaluate the predictive ability of these 30 CpG unit measures, we applied a statistical learning algorithm, using our training set to select a model and the test set to validate the model performance. For a classifier, we chose the decision tree–based method C4.5 (26), implemented as the so-called "J48" algorithm in the Weka data mining package (24).

[3] http://www.sequenom.com/customer_support/scientific_applicationnotes.php.
[4] http://www.ensembl.org/Homo_sapiens/index.html.

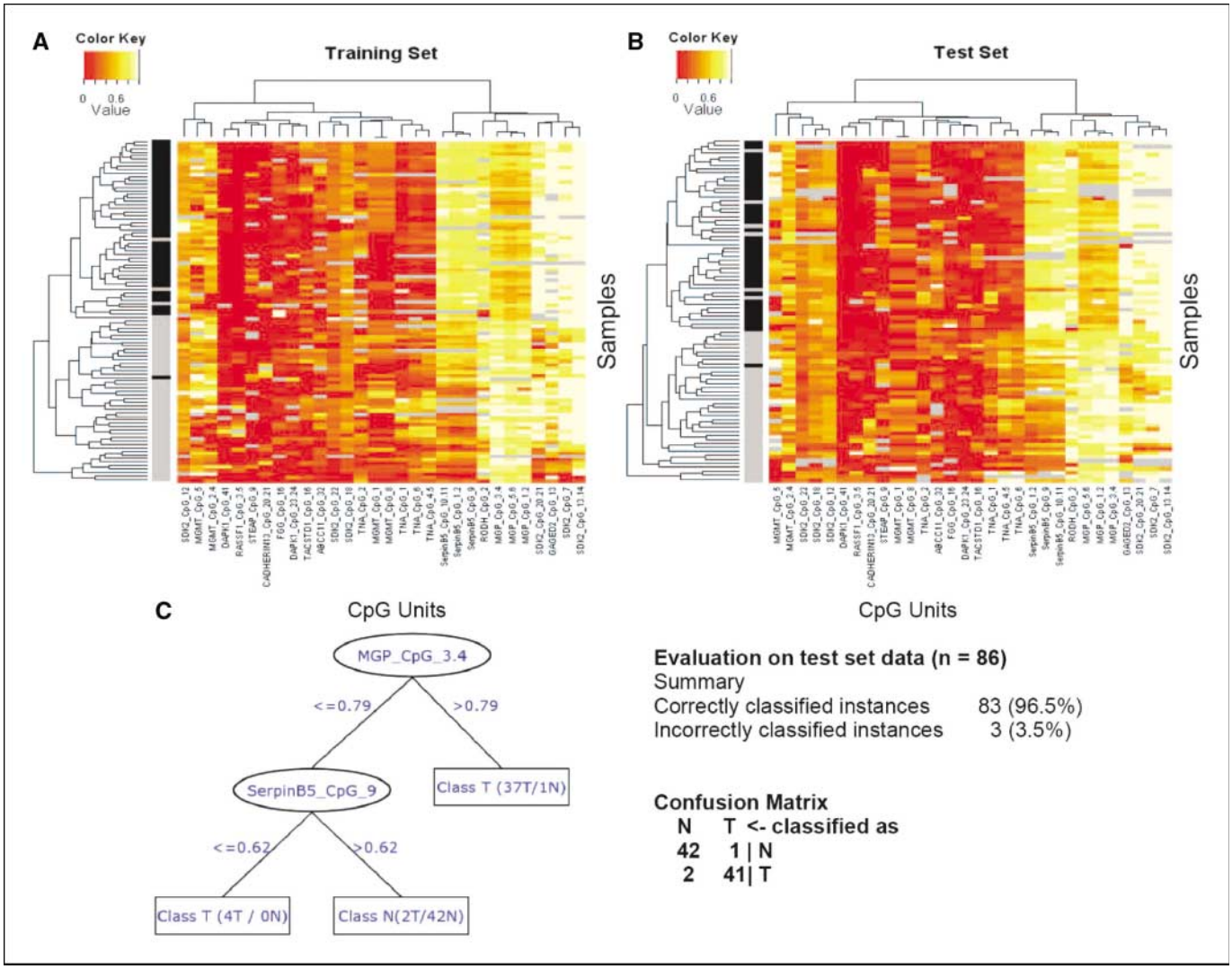

Figure 1. Results of a two-way hierarchical cluster analysis of the relative methylation of 30 CpG units (*columns*) measured on 88 tissue samples derived from a training set comprising 44 lung cancer patients (*A*) and 79 tissue samples from a test set comprising 40 lung cancer patients (*B*). *Left vertical axis,* tissue samples; *black bars,* normal samples; *gray bars,* tumor samples. *Bottom horizontal axis,* CpG units as the gene and the fragment number within the gene. The relative methylation of each fragment within each sample is presented in the central image plot with values ranging from zero (*red*) to one (*yellow*; see color key). Missing values (*gray*). *C,* classification tree as a result of a statistical learning algorithm optimized for discrimination of tissue samples into normal and tumor tissue. The model was built using the training data and performance characteristics are given for evaluation of the model using the test data.

This algorithm identified a pruned three-node tree, including CpG units from MGP and SERPINB5 as the optimal classifier and achieved >95% sensitivity and specificity when applied to the test set (Fig. 1*C*). We also evaluated several further classification methods (random forest, support vector machines, linear model transformation, naive bayes, and recursive partitioning) and found that all methods result in a predictive accuracy >90%. Here, we focused on the results from the "J48" method because the decision tree–based method allows clear interpretation of the resulting model.

In addition to the selection of a predictive model, we examined which of the genes contained CpG units where methylation differed significantly between tissue types (Table 2). We found that multiple CpG units within *MGP, SERPINB5, GAGED2, TNA, RASSF1*, and *SDK2* showed highly statistically significant associations with tissue type ($P < 10^{-6}$). Note that *AQP1* showed only one significant CpG unit and hence was excluded from the list.

We next applied our analysis to NSCLC tumor tissue attributes to explore whether the methylation patterns differ significantly between tumor types. All tumor samples were combined into one data set. The vast majority of the tumor samples in this data set were adenocarcinomas and squamous cell carcinomas (39 adenocarcinomas and 47 squamous cell carcinomas). The data were then filtered according to the previously described criteria for quality and variability. A total of 90 CpG units and 88 tumors passed our filtering criteria and were used in a two-way hierarchical cluster analysis (Fig. 2*A*). This resulted in two large and highly differentiated clusters, with four noticeable subclusters within the largest cluster. We examined the relationship between the resulting clusters and clinical tumor attributes. The five clusters show a significant association with tumor histology ($P = 8.6 \times 10^{-5}$) but not with other clinical characteristics, such as gender ($P = 0.30$), tumor stage ($P = 0.22$), or differentiation ($P = 0.15$). The most significant differences ($P < 0.01$) between adenocarcinomas and squamous cell

carcinoma promoter methylation were observed in SDK, GAGED2, ADAMTS8, TNA, PRAME, and CADHERIN13. The difference between the two histologic groups in ADAMTS8 methylation agrees with our previous observations (27).

Beer et al. (10) have reported earlier that adenocarcinomas cluster by phenotype using gene expression profiles. Hence, we were interested to evaluate whether methylation profiles will generate similar grouping effects within adenocarcinomas or squamous cell carcinomas. We divided the NSCLC tumor samples into subsets based on tumor histology and analyzed each subset separately. The adenocarcinomas showed three clusters that are strongly associated with gender (Fig. 2*B*). Eleven of 17 female samples were localized into one cluster (which included one male). This grouping is largely the result of the X-chromosomal gene *GAGED2*, which is more methylated in females, likely because of X-chromosomal inactivation. Further differentiation of this female adenocarcinoma cluster is mainly a result of different methylation levels in the MAGEE1 (X-chromosomal) and PRAME promoter region. In a similar analysis of squamous cell carcinomas, multiple distinctive clusters were also observed (Fig. 2*C*). However, these clusters were not significantly associated with any of the clinical variables evaluated.

We carried out a survival analysis based on the methylation patterns of all tumor samples. In the present sample, survival information was only available for 61 individuals. Furthermore, the data were largely right censored (46 alive and 12 dead). Hence, a robust survival analysis could not be carried out. We analyzed the relationship between patient survival and tumor stage and found that this data set fails to present the established association between survival and tumor stage ($P = 0.52$; Fig. 3*A*). Nevertheless, we used a supervised approach to search for a combination of CpG units that improve survival prediction. We evaluated each of the 377 variable CpG units for an association with survival ($P < 0.05$). The 21 CpG units (derived from 13 genes) that satisfied this criterion were subsequently included in a hierarchical cluster analysis to group patients with similar patterns of methylation (Supplementary Fig. S3). We used the first split in the dendrogram to separate patients into two groups for survival analysis, which displayed a modest association with survival ($P = 0.021$; Fig. 3*B*). Notably, only nine stage I tumors can be found in the good prognosis group.

Evaluation of the relationship between promoter methylation and gene expression was conducted on a subset of the genes using real-time competitive PCR coupled with MassARRAY (20, 21). We selected six genes for further analysis, from which some showed strong association with tissue pathology, whereas others did not (strong association: *SERPINB5* and *MGP1*; weak or no association: *AQP1, CDH13, CDKN2A*, and *DAPK1*). Gene expression analysis included the use of six internal control genes for data normalization (*GAPDH, RPL13A, SDHA, TBP, UBC*, and *YWHAZ*). A detailed technical explanation of the normalization process is beyond the scope of this article but the process has been described elsewhere (28). Methylation values of an individual CpG site in one gene were averaged per sample and a mean methylation value was calculated for every analyzed DNA sample. Expression analysis was carried out for both normal and tumor tissues, previously subjected to methylation analysis. We were therefore able to calculate the differences in expression between normal and tumor and compare these results with changes in methylation. Because the changes in expression ranged over multiple orders of magnitude, we logarithmically transformed the absolute difference (base 10). The direction of the change was preserved by multiplying the logarithmic value by -1 if the original difference was negative. Negative differences are a result of higher values in the tumor specimen. For genes that are hypermethylated and consequently down-regulated in tumors, we expect to see negative values for methylation differences combined with positive values for expression changes. The inverse is true for hypomethylated genes. Figure 4 shows the differences in methylation plotted on the x axis and the expression differences on the y axis. As expected, the largest clusters of sample pairs can be found in the space of hypermethylated/down-regulated (Fig. 4, *top, left quadrant*). However, in our data set, three genes do not show the expected relationship (*CDH13, DAPK1*, and *CDKN2a*). When establishing a correlation between DNA methylation and gene expression changes, the regression is affected by these genes. Hence, the correlation is modest but highly statistically significant (nonparametric correlation coefficient

Table 2. Summary of genes and gene function (when available) for genes, which had at more than one CpG unit with significant difference in methylation levels between normal and tumor tissue

Gene name	Description	Comment	No. CpG units in the gene with *P* below E-6
MGP	Structural component of extracellular matrix; matrix Gla protein	Extracelluular matrix structural constituent	4
SerpinB5	Serine (or cysteine) proteinase inhibitor, clade B (ovalbumin), member 5	Tumor suppressor function assumed, completely silenced in normal lung	3
GAGED2	G antigen family D 2 protein (XAGE-1 protein)	Unknown function, reported to be highly expressed in lung cancer	2
TNA	Tetranectin (plasminogen-binding protein)	Tetranectin binds to plasminogen and to isolated kringle 4; may be involved in the packaging of molecules destined for exocytosis, extracellular region, osteogenesis	3
RASSF1	Ras association (RalGDS/AF-6) domain family	Negative regulation of cell cycle, potential tumor suppressor, epigenetic inactivation in lung cancer described (Dammann R, 2000 NG)	4
SDK2	Sidekick homologue 2		3

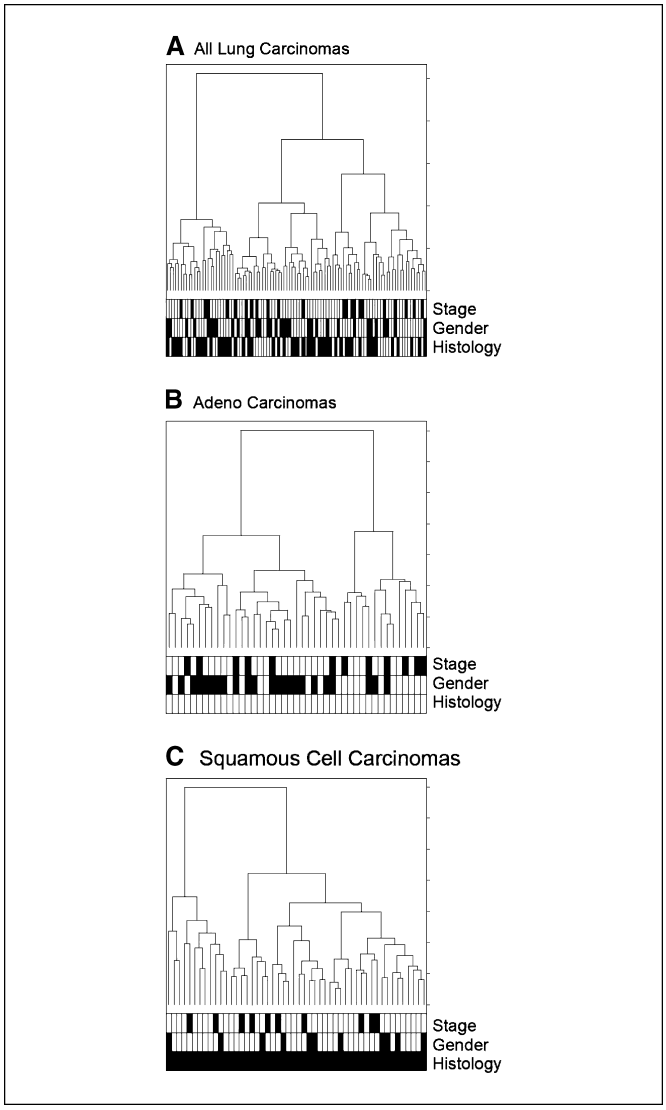

Figure 2. Dendrogram from hierarchical clustering with clinical and histologic information assigned to the individual samples. *Black boxes,* the presence of the characteristic: male sex, adenocarcinoma histology, and tumor stage greater than II. Shown are results for all samples (*A*), the subgroup of adenocarcinomas (*B*), and squamous cell carcinomas (*C*).

Spearman $\rho = -0.43$; $P = 10^{-10}$), showing a general trend toward an inverse relationship between DNA methylation and gene expression.

Discussion

We measured quantitative changes of methylation of 47 promoter regions in lung cancer and adjacent normal tissue samples and evaluated their distribution, correlation, and relationships to clinicopathologic variables using a variety of common statistical methods. Hierarchical clustering identified substantial differences in the quantitative methylation patterns of tumor tissue compared with adjacent normal tissue. Based on this observation, we used a subset of the data to train a statistical learning algorithm and achieved classification accuracies >90% when the model was applied to an independent test set. We also discovered a strong association of quantitative methylation patterns to tumor histology. In general, CpG units derived from the same gene showed highly similar methylation patterns.

We identified CpG units from the promoter regions of six genes that exhibited significantly different levels of methylation ($P < 10^{-6}$) between normal and tumor samples. Four (*SERPINB5, TNA, RASSF1*, and *GAGED2*) of the six genes have been implicated previously in tumor development. Whereas cancer-related changes in DNA methylation have been described extensively for RASSF1, methylation of SERPINB5 (maspin) has been studied less frequently. Our analysis shows that SERPINB5 is highly methylated in normal lung tissue, consistent with previous studies by Yatabe et al. (29). The lung tumor tissue analyzed in this study, however, showed hypomethylation of SERPINB5 in tumors. Interpretation of this result is unclear. Hypomethylation of SERPINB5 generally correlated with an increase in gene expression, and this agrees with Smith et al. and contradicts, at least in lung (30), its suggested role as a tumor suppressor inhibiting cell motility, invasion, angiogenesis, and metastasis *in vitro* (31, 32). In addition, Yatabe et al. have shown that SERPINB5 expression is controlled by promoter methylation and varies among the different cell types in the lung. Correspondingly, hypomethylation and therefore expression of SERPINB5 might be indicative of tumor clonality and its cell type–specific origin.

Methylation levels in the promoter regions of MGP and SDK2, not previously implicated in tumor development, were also found to be significantly different between tumor and normal tissues. Neither MGP nor any genes that are likely to be coregulated by these CpG sites have been linked to cancer (genes found within 100 kb upstream and downstream are *WBP11, DO, PDE6H, ARHGDIB* as well as three hypothetical proteins). Although unlikely, hypermethylation of the MGP region could indicate a new cancer relevant gene function besides ossification. However, it is more likely to be an effect of instable DNA methylation maintenance in cancer, which is observed more frequently.

We analyzed the change in expression for a subset of the differentially methylated genes and found that differences in methylation are strongly correlated to expression differences in three of the six examined genes ($\rho = -0.52$; $P = 10^{-8}$). Clearly, the lack of response for the remaining three genes is striking, especially because previous studies have already shown a clear relationship between expression and DNA methylation in NSCLC. The most likely explanation is of technical nature. Methylation levels for all three genes are low across all samples (mean methylation for *CDH13*, 4%; *CDKN2A*, 4%; and *DAPK1*, 2%). The used technology has a detection limit $\sim$5% methylated DNA and therefore is not suitable to reliably detect methylation changes of this scale. Furthermore, gene expression is not exclusively regulated by methylation. On the contrary, multiple factors influence genetic transcription. In addition, many genes have promoter regions that are larger than the analyzed regions; thus, CpGs in important regulatory elements may have not been analyzed in this study.

It is of note that the most significant methylation differences in this study were observed in genomic regions with relatively low CpG density. The University of California Santa Cruz genome browser identifies only two of the seven most significant regions revealed in this study as CpG islands. The remaining five regions are either located in the 5′-UTR or, for MGP, were selected simply because they had the highest CpG content in the entire genomic region. However, the relationship between changes in DNA methylation and gene expression is statistically significant. Our

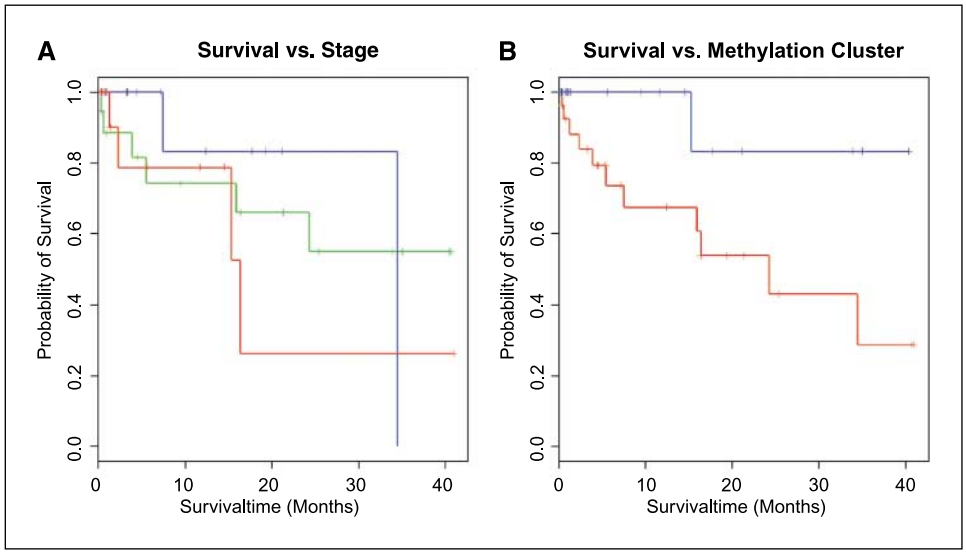

Figure 3. Kaplan-Meier survival estimates for stage-defined subgroups (*A*) and cluster-defined subgroups (*B*). *A,* stage 1 (*blue line*), stage 2 (*green line*), and stage 3 (*red line*). The difference in survival of these three groups is not statistically significant. *B,* cluster-defined poor-outcome and good-outcome prognosis groups. The survival difference is statistically significant ($P = 0.021$).

findings indicate that DNA methylation regulates gene expression outside of traditional CpG islands and suggest rethinking of the common theorem that only regions of high CpG density are involved in gene silencing.

In this study, we have examined tumor specimens that contained up to 5% to 30% stromal cells. This inevitably results in a mixture of tumor- and nontumor-related cell types in the sample. Hence, small differences in DNA methylation may not be detectable. However, the ability to quantitate methylation may make the requirement for microdissection less critical, at least for discovery of differentially methylated genes.

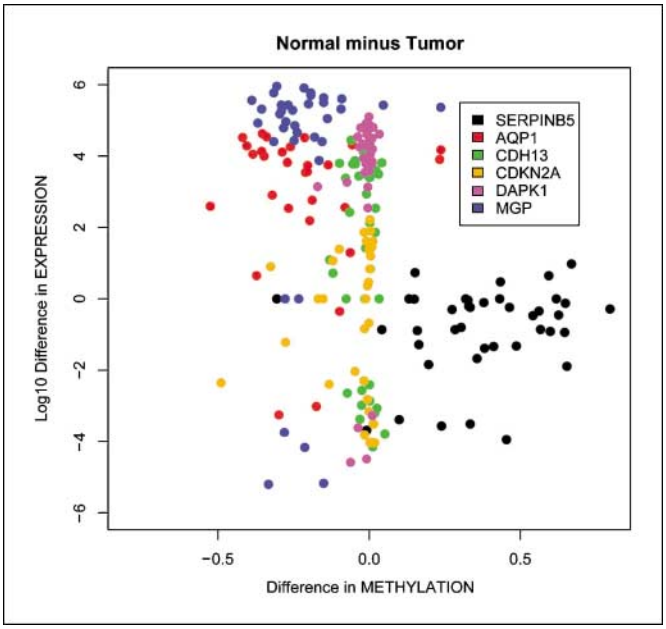

Figure 4. Scatterplot showing differences of methylation and expression levels between normal and tumor samples. For each sample pair, differences in methylation are plotted on the x axis and differences between normalized expression values are plotted on the y axis. Different colors represent different genes. *Black circles,* SerpinB5; *blue circles,* MGP1; *red circles,* AQP1; *pink circles,* DAPK1; *green circles,* CDH13; *orange circles,* CDKN4a. The expected negative correlation between methylation and expression changes is exemplified in MGP1 (*blue circles*) and AQP1 (*red circles*). An example for methylation-independent expression changes is given by DAPK1 (*pink circles*).

This study failed to identify robust CpG predictors of survival. This can partly be attributed to the fact that the survival data for the analyzed samples were insufficient to build a good model. The sample set only included 61 patients with survival data and the vast majority of samples were right censored at time of analysis. Unlike the expression profiling studies commonly done on oligonucleotide microarrays, we did not screen DNA methylation on a genome-wide scale. The set of 47 promoter regions used here was selected based on previous expression microarray data (33) in a candidate gene approach and therefore cannot be expected to be a universal clinical predictor.

The results of this study suggest that DNA methylation analysis can be used in combination with gene expression profiling to discover a clinically meaningful molecular marker set. The strength of expression profiles is obviously the number of genes that can be analyzed simultaneously. Genome-wide analysis can be done to identify genes that are differentially expressed. Once these genes are discovered, quantitative methylation analysis can be applied and a subset of methylation-regulated genes can be identified. When methylation and gene expression profiles have similar predictive value, a methylation-based test could be preferable. Although improvements have been made in the recent years and gene expression markers are now found in clinical settings, reproducibility of chip array expression profiles remains an issue. RNA is much more fragile and more prone to degradation compared with the covalent addition of methyl groups to cytosine. In our laboratory, we have observed stable methylation ratios independent of the quality of the DNA and were able to accurately analyze DNA methylation from paraffin embedded tissue samples.[5]

This study is the first to show that DNA methylation can be analyzed on a large scale and quantitative results can be used for predicting tissue pathology. The data also suggest a potential role of DNA methylation in the identification of poor and good survival groups in NSCLC.

Epigenetic events are likely to occur early in tumor progression and identification of tumor-specific methylation changes will likely influence our understanding of the disease, possibly leading to molecular markers for early detection of lung cancer.

[5] Unpublished data.

Acknowledgments

Received 1/31/2006; revised 7/21/2006; accepted 8/22/2006.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

1. Stewart BW, Kleihues P. World cancer report. Lyon: WHO; 2003. p. 352.
2. Bains MS. Surgical treatment of lung cancer. Chest 1991;100:826–37.
3. Sidransky D. Emerging molecular markers of cancer. Nat Rev Cancer 2002;2:210–9.
4. Slebos RJ, Kibbelaar RE, Dalesio O, et al. K-ras oncogene activation as a prognostic marker in adenocarcinoma of the lung. N Engl J Med 1990;323:561–5.
5. Harpole DH, Jr., Herndon JE III, Wolfe WG, Iglehart JD, Marks JR. A prognostic model of recurrence and death in stage I non-small cell lung cancer utilizing presentation, histopathology, and oncoprotein expression. Cancer Res 1995;55:51–6.
6. Horio Y, Takahashi T, Kuroishi T, et al. Prognostic significance of p53 mutations and 3p deletions in primary resected non-small cell lung cancer. Cancer Res 1993;53:1–4.
7. Mao L, Lee DJ, Tockman MS, et al. Microsatellite alterations as clonal markers for the detection of human cancer. Proc Natl Acad Sci U S A 1994;91:9871–5.
8. Yanagisawa K, Shyr Y, Xu BJ, et al. Proteomic patterns of tumour subsets in non-small-cell lung cancer. Lancet 2003;362:433–9.
9. Chen G, Gharib TG, Wang H, et al. Protein profiles associated with survival in lung adenocarcinoma. Proc Natl Acad Sci U S A 2003;100:13537–42.
10. Beer DG, Kardia SL, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med 2002;8:816–24.
11. Tomida S, Koshikawa K, Yatabe Y, et al. Gene expression-based, individualized outcome prediction for surgically treated lung cancer patients. Oncogene 2004;23:5360–70.
12. Belinsky SA. Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer 2004;4:707–17.
13. Esteller M, Sanchez-Cespedes M, Rosell R, et al. Detection of aberrant promoter hypermethylation of tumor suppressor genes in serum DNA from non-small cell lung cancer patients. Cancer Res 1999;59:67–70.
14. Palmisano WA, Divine KK, Saccomanno G, et al. Predicting lung cancer by detecting aberrant promoter methylation in sputum. Cancer Res 2000;60:5954–8.
15. Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer 2003;3:253–66.
16. Herman JG, Graff JR, Myohanen S, Nelkin BD, Baylin SB. Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc Natl Acad Sci U S A 1996;93:9821–6.
17. Ehrich M, Nelson MR, Stanssens P, et al. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. Proc Natl Acad Sci U S A 2005;102:15785–90.
18. Hartmer R, Storm N, Boecker S, et al. RNase T1 mediated base-specific cleavage and MALDI-TOF MS for high-throughput comparative sequence analysis. Nucleic Acids Res 2003;31:e47.
19. Paulin R, Grigg GW, Davey MW, Piper AA. Urea improves efficiency of bisulphite-mediated sequencing of 5′-methylcytosine in genomic DNA. Nucleic Acids Res 1998;26:5009–10.
20. Elvidge GP, Price TS, Glenny L, Ragoussis J. Development and evaluation of real competitive PCR for high-throughput quantitative applications. Anal Biochem 2005;339:231–41.
21. Ding C, Cantor CR. A high-throughput gene expression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. Proc Natl Acad Sci U S A 2003;100:3059–64.
22. Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. New York: Wiley; 1990. p. xiv, 342.
23. R Development Core Team. R: a language and environment for statistical computing, R Foundation for Statistical Computing Vienna, Austria 2003.
24. Frank IHWaE. Data mining: practical machine learning tools with Java implementations. San Francisco: Morgan Kaufmann; 2000.
25. Ripley BD. Pattern recognition and neural networks. Cambridge: Cambridge University Press; 1996.
26. Quinlan JR. C4.5: programs for machine learning. San Mateo (CA): Morgan Kaufman; 1993.
27. Dunn JR, Panutsopulos D, Shaw MW, et al. METH-2 silencing and promoter hypermethylation in NSCLC. Br J Cancer 2004;91:1149–54.
28. Vandesompele J, De Preter K, Pattyn F, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol 2002;3:RESEARCH0034.
29. Yatabe Y, Mitsudomi T, Takahashi T. Maspin expression in normal lung and non-small-cell lung cancers: cellular property-associated expression under the control of promoter DNA methylation. Oncogene 2004;23:4041–9.
30. Smith SL, Watson SG, Ratschiller D, et al. Maspin—the most commonly-expressed gene of the 18q21.3 serpin cluster in lung cancer—is strongly expressed in preneoplastic bronchial lesions. Oncogene 2003;22:8677–87.
31. Costello JF, Vertino PM. Methylation matters: a new spin on maspin. Nat Genet 2002;31:123–4.
32. Futscher BW, Oshiro MM, Wozniak RJ, et al. Role for DNA methylation in the control of cell type specific maspin expression. Nat Genet 2002;31:175–9.
33. Heighway J, Knapp T, Boyce L, et al. Expression profiling of primary non-small cell lung cancer for target identification. Oncogene 2002;21:7749–63.