A Novel Epigenetic Signature for Early Diagnosis in Lung Cancer

Angel Díaz-Lagares, Jesus Mendez-Gonzalez, David Hervas, Maria Saigi, Maria J. Pajares, Diana García, Ana B. Crujerias, Rubén Pio, Luis M. Montuenga, Javier Zulueta, Ernest Nadal, Antoni Rosell, Manel Esteller, and Juan Sandoval

Abstract

Purpose: Lung cancer remains as the leading cause of cancer-related death worldwide, mainly due to late diagnosis. Cytology is the gold-standard method for lung cancer diagnosis in minimally invasive respiratory samples, despite its low sensitivity. We aimed to identify epigenetic biomarkers with clinical utility for cancer diagnosis in minimally/noninvasive specimens to improve accuracy of current technologies.

Experimental Design: The identification of novel epigenetic biomarkers in stage I lung tumors was accomplished using an integrative genome-wide restrictive analysis of two different large public databases. DNA methylation levels for the selected biomarkers were validated by pyrosequencing in paraffin-embedded tissues and minimally invasive and noninvasive respiratory samples in independent cohorts.

Results: We identified nine cancer-specific hypermethylated genes in early-stage lung primary tumors. Four of these genes presented consistent CpG island hypermethylation compared with nonmalignant lung and were associated with transcriptional silencing. A diagnostic signature was built using multivariate logistic regression model based on the combination of four genes: BCAT1, CDO1, TRIM58, and ZNF177. Clinical diagnostic value was also validated in multiple independent cohorts and yielded a remarkable diagnostic accuracy in all cohorts tested. Calibrated and cross-validated epigenetic model predicts with high accuracy the probability to detect cancer in minimally and noninvasive samples. We demonstrated that this epigenetic signature achieved higher diagnostic efficacy in bronchial fluids as compared with conventional cytology for lung cancer diagnosis.

Conclusions: Minimally invasive epigenetic biomarkers have emerged as promising tools for cancer diagnosis. The herein obtained epigenetic model in combination with current diagnostic protocols may improve early diagnosis and outcome of lung cancer patients.

Introduction

Lung cancer is the main cause of death from cancer worldwide (1). Several factors are associated with the poor outcome of lung cancer patients. One of them is, despite recent advances, the scarcity of effective therapies achieving durable responses. Another, and even more important, factor is late diagnosis, as most lung tumors are detected at advanced stages of the disease (2). This is crucial, taking into account that survival rates drop substantially from early to late stages.

In this context, the data reported by early observational studies and by the randomized National Lung Screening Trial (NLST) have shown that lung cancer screening with low-dose helical computed tomography (LDCT) is able to reduce lung cancer mortality, as significantly more cases can be detected in earlier stages (3). Last year, the United States Preventive Services Task Force recommended lung cancer screening with LDCT for people at high risk (4). However, the implementation of LDCT screening has been slow, mainly due to concerns about its cost and harm (5).

Other studies have shown that early detection of lung cancer is feasible through liquid biopsy approaches using circulating tumor DNA (ctDNA). In this context, the detection of novel epigenetic biomarkers in bronchial fluids as compared with conventional cytology for lung cancer diagnosis.
Translational Relevance

Lung cancer is the leading cause of cancer mortality worldwide. Patient outcome is closely linked to tumor stage at diagnosis and unfortunately, most lung cancer patients are diagnosed at late stages when a curative treatment is no longer possible. Using an integrative genome-wide experimental method whereby hundreds of stage I patients from two independent lung cancer datasets were examined, we identified an epigenetic four-gene model with diagnostic value for detecting lung cancer. This DNA methylation signature was validated with a gene locus–specific technique in three minimally and noninvasive independent cohorts. The combination of this highly sensitive and specific epigenetic model with standard clinical markers may help to improve lung cancer diagnosis and therefore decrease current mortality rates.

Materials and Methods

Study design and participants

This is a collaborative and retrospective study including data from publicly available datasets, formalin-fixed paraffin-embedded (FFPE) tissues, bronchial aspirates/lavages, and sputum samples obtained from lung cancer patients and cancer-free individuals, as they arrived to the laboratory and passed the technical quality checks. Genome-wide DNA methylation data for the discovery cohort (Infinium 450K array) was downloaded from our previous published lung cancer dataset deposited at the Gene expression omnibus (GSE39279; ref. 19) or from the TCGA data repository (lung adenocarcinoma LUAD or Lung squamous cell carcinoma LIISC). The Biologic validation of the selected methylation biomarkers was conducted by pyrosequencing in four independent cohorts. Lung validating cohorts were obtained from different institutions in Spain. (i) A total of 201 FFPE samples were obtained from Health Institute Carlos III (ISCIII), Madrid and Centre for Applied Medical Research/Hospital of the University of Navarre, (CIMA/CUN) Pamplona. Regarding minimally invasive samples, (ii) 80 bronchial aspirates and (iii) 98 sputums were obtained from Catalan Institute of Oncology and Bellvitge University Hospital, Barcelona. (iv) 111 Bronchioalveolar lavages came from CIMA, Pamplona and Hospital of Talavera de la Reina, Talavera de la Reina. All DNA extractions form different specimens were developed and run by the same technicians to avoid interlaboratory variation.

The study was approved by the corresponding Institutional review board and patients signed up the informed consent to participate. The main clinical characteristics of the different cohorts are described in Table 1 or have been also described previously, as is the case for the discovery cohort (Table 1).

Procedures

Preparation of lung specimens. DNA was extracted from minimally and noninvasive specimens using a standard phenol chloroform extraction method. DNA from FFPE tissue blocks was extracted from two sequential unstained sections, each 10-µm thick. For
Table 1. Clinical characteristics of the invasive [Discovery (A), Validation (B) and Paraffin (C)] and minimally invasive [BAS (D), BALs (E) and Sputum (F)] cohorts

<table>
<thead>
<tr>
<th>Patients</th>
<th>A. Discovery cohort</th>
<th>B. TCGA validation cohort</th>
<th>C. Paraffin cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tumor (n = 237)</td>
<td>Nonmortal (n = 25)</td>
<td>Tumor (n = 350)</td>
</tr>
<tr>
<td>Age (years)</td>
<td>68 (38–90)</td>
<td>63.5 (39–86)</td>
<td>66.9 (33–90)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male 131 (55.5%)</td>
<td>20 (80%)</td>
<td>190 (54.1%)</td>
</tr>
<tr>
<td></td>
<td>Female 106 (44.4%)</td>
<td>5 (20%)</td>
<td>160 (45.9%)</td>
</tr>
<tr>
<td>Smoking history</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current or former smoker</td>
<td>190 (80.1%)</td>
<td>24 (96%)</td>
<td>313 (89.4%)</td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>25 (10.5%)</td>
<td>1 (4%)</td>
<td>32 (9.1%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>22 (9.4%)</td>
<td>0 (0%)</td>
<td>5 (1.5%)</td>
</tr>
<tr>
<td>Stage</td>
<td>I 237 (100%)</td>
<td>350 (100%)</td>
<td>122 (100%)</td>
</tr>
<tr>
<td>Histology</td>
<td>Adenocarcinoma</td>
<td>181 (76.4%)</td>
<td>217 (62.1%)</td>
</tr>
<tr>
<td></td>
<td>Squamous cell carcinoma</td>
<td>56 (23.6%)</td>
<td>133 (37.9%)</td>
</tr>
<tr>
<td></td>
<td>Pack-years 40 (0–180)</td>
<td>54.4 (0–184)</td>
<td>46.7 (1–94.5)</td>
</tr>
<tr>
<td></td>
<td>35.9 (0–130)</td>
<td>46.7 (0–141)</td>
<td></td>
</tr>
<tr>
<td>Data are average (range) or number (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patients</th>
<th>D. BAS cohort</th>
<th>E. BALs cohort</th>
<th>F. Sputum cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>65.0 (47–83)</td>
<td>64.0 (35–87)</td>
<td>62.1 (38–83)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male 46 (90.2%)</td>
<td>16 (55.2%)</td>
<td>66 (80.4%)</td>
</tr>
<tr>
<td></td>
<td>Female 5 (9.8%)</td>
<td>10 (34.5%)</td>
<td>16 (19.6%)</td>
</tr>
<tr>
<td>Smoking history</td>
<td></td>
<td>0 (0%)</td>
<td>1 (3.5%)</td>
</tr>
<tr>
<td>Current or former smoker</td>
<td>45 (88.2%)</td>
<td>16 (55.2%)</td>
<td>42 (51.2%)</td>
</tr>
<tr>
<td>Nonsmoker</td>
<td>4 (7.8%)</td>
<td>8 (27.6%)</td>
<td>39 (47.5%)</td>
</tr>
<tr>
<td>Unknown</td>
<td>2 (4.0%)</td>
<td>5 (17.2%)</td>
<td>1 (1.3%)</td>
</tr>
<tr>
<td>Stage</td>
<td>I 5 (9.8%)</td>
<td>17 (20.7%)</td>
<td>12 (16.7%)</td>
</tr>
<tr>
<td></td>
<td>II 6 (11.8%)</td>
<td>8 (9.8%)</td>
<td>18 (22.0%)</td>
</tr>
<tr>
<td></td>
<td>III 21 (41.2%)</td>
<td>20 (24.4%)</td>
<td>23 (32.0%)</td>
</tr>
<tr>
<td></td>
<td>IV 18 (35.3%)</td>
<td>18 (22.0%)</td>
<td>19 (26.4%)</td>
</tr>
<tr>
<td>Histology</td>
<td>Adenocarcinoma</td>
<td>17 (33.3%)</td>
<td>25 (50.5%)</td>
</tr>
<tr>
<td></td>
<td>Squamous cell carcinoma</td>
<td>19 (37.3%)</td>
<td>40 (48.8%)</td>
</tr>
<tr>
<td></td>
<td>Large cell carcinoma</td>
<td>2 (4.0%)</td>
<td>2 (2.4%)</td>
</tr>
<tr>
<td></td>
<td>Small cell carcinoma</td>
<td>2 (4.0%)</td>
<td>12 (14.6%)</td>
</tr>
<tr>
<td></td>
<td>NSCLC (NOS)</td>
<td>81 (21.4%)</td>
<td>3 (3.7%)</td>
</tr>
<tr>
<td></td>
<td>Pack-years 49.6 (0–120)</td>
<td>32.4 (0–100)</td>
<td>45.5 (0–120)</td>
</tr>
<tr>
<td></td>
<td>45.9 (0–120)</td>
<td>24.1 (0–114)</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Data are average (range) or number (%).
Abbreviations: NSCLC (NOS), not otherwise specified.

Data filtering. DNA methylation status of 450,000 CpG sites by using the Infinium 450K Methylation array was available at (19). Methylation score of each CpG was represented as beta (β) value and were previously normalized for color bias adjustment, background level adjustment and quantile normalization across arrays. Probes and sample filtering involved a two-step process for removing SNPs and unreliable betas with high detection P value P > 0.001. Sex chromosome probes were also removed. After the filtering, the remaining 409,219 CpGs were considered valid for the study. Stage I patients were selected coming up to 237 lung tumor patients and 25 histologically nontumor lung tissue samples (Fig. 1).

Data filtering for hypomethylated CpGs in nontumoral tissues. The choice of region to be studied is one of the critical challenges to establishing a DNA methylation biomarker that is clinically useful. The investigated region should ideally fulfill the following criteria: first, the region should be unmethylated in nontumor cases and methylated in lung cancer cases; and second, the methylation levels of this region should clearly allow the classification of a sample as non-cancerous or cancerous (20,21). We set thresholds to select homogeneous unmethylated CpGs in nontumor cases; (i) Average and median of β values lower than 0.1 and (ii) the percentile 90 for beta values of control donors lower than 0.2. Using these restrictive thresholds, we obtained 133,444 filtered CpGs.
Differentially methylated CpG identification between tumor and nontumor samples. Differentially methylated CpGs between tumor and nontumor groups were identified using the following procedure: for each probe/CpG, the sets of methylation β values T (belonging to the tumor samples: first group) and NT (belonging to the nontumor lung tissue samples: second group) were compared. The following three measures were calculated:

1. Differences in average β values between groups higher than a set threshold. \( \text{MD} = |\mu_T - \mu_{NT}| > 0.20 \)
2. Multiple testing correction P value with 95% of confidence to assign significant differentially methylated sites. False discovery rate (FDR) with adjusted P value < 0.05.
3. To maximize differences between tumor and nontumor group. Difference between quartile 1 in tumor group and quartile 3 in nontumor should be higher that a selected cut-off. P25 (tumor)-P75 (nontumor) > 0.1

To identify early-stage cancer-related epigenetic markers with diagnostic value for both subtypes together, these three criteria (1,2,3) were evaluated in three distinct comparisons based on histologic subtypes.

a) Comparison: Adenocarcinoma (ADC) group \( (n = 181) \) versus nontumor group \( (n = 25) \). Identified 29 differentially methylated CpGs (DMCpG) specific for ADC (Supplementary Table S1A).

b) Comparison: Squamous cell carcinoma (SCC) group \( (n = 56) \) versus nontumor group \( (n = 25) \). Identified 78 DMCpGs specific for SCC (Supplementary Table S1B).

c) Comparison: Lung cancer (ADC and SCC \( n = 237) \) versus nontumor group \( (n = 25) \). Identified 24 significant DMCpGs when both groups were analyzed together (Supplementary Table S1C).

Differentially methylated CpGs were selected using an integrative approach to rank the Infinium probes based on their methylation status and the fulfillment of all the criteria (1,2,3) and in the three different comparisons (a, b, c). Finally, 12 CpGs corresponding to 9 genes were common in the three comparisons and ranked by averaged z-score (Supplementary Fig. S1; Supplementary Table S2).

Pyrosequencing

Pyrosequencing analyses to determine CpG methylation status were developed as described previously (19) to validate the results obtained from the arrays. Briefly, a set of primers for PCR amplification and sequencing were designed using a specific software pack (PyroMark assay design version 2.0.01.15). Primer sequences were designed to hybridize with CpG-free sites to ensure methylation-independent amplification (Supplementary Table S2A). DNA was converted using the EZ DNA Methylation Gold (ZYMOS RESEARCH) bisulfite conversion kit following the manufacturer's recommendations and used as a template for subsequent PCR step. PCR was performed under standard conditions.
conditions with primers biotinylated to convert the PCR product to single-stranded DNA templates. We used the Vacuum Prep Tool (Biotage) to prepare single-stranded PCR products according to manufacturer's instructions. PCR products were observed at 2% agarose gels before pyrosequencing. Pyrosequencing reactions and methylation quantification were performed in a PyroMark Q24 System version 2.0.6 (Qiagen) using appropriate reagents and protocols, and the methylation value was obtained from the average of the CpG dinucleotides included in the sequence analyzed, with a minimum of 3 valid CpGs per primer. Only those average methylation values within the region analyzed with coefficient of variation lower than 1 were accepted as valid. Controls to assess correct bisulfit conversion of the DNA were included in each run, as well as sequencing controls to ensure the fidelity of the measurements.

Statistical analysis

Data were summarized by mean, SD, median, and first and third quartiles in the case of continuous variables and by relative and absolute frequencies in the case of categorical variables. Differences in expression values and methylation levels among groups were assessed using the nonparametric Wilcoxon rank sum test. Receiver Operating Characteristic (ROC) curves were used to assess the predictive capacity of each marker. Area under the curve (AUC) was computed for each ROC curve, and 95% confidence intervals (CI) were also estimated by bootstrapping with 1,000 iterations. A predictive model for each sample type was built including all selected markers in a multivariable logistic regression model. ROC curves and AUC were also computed for the predictive models. Calibration of the models was assessed by plotting predicted versus observed values obtained by bootstrap resampling of the original data. Internal validation of the models was performed using 10-fold crossvalidation. The final predictive models were represented in nomograms to facilitate their use by clinicians. Sensitivity and specificity were estimated at the optimal cut-off point according to Youden criterion. In addition, the sensitivity and specificity curves were estimated for the whole range of predictions of the model to allow for personalized decisions in different clinical scenarios. Globally, a two-tailed P value of less than 0.05 was considered to indicate statistical significance. P values were adjusted for multiple comparisons using the FDR procedure by Benjamini and Hochberg. All statistical analyses were performed using R software (version 3.2.0) and the pROC R-package (version 1.7.3).

Results

Identification and validation of cancer-related methylated genes

The discovery cohort consisted of 237 stage I non–small cell lung primary tumors (NSCLC) and 25 nontumoral matched lung tissues from the CURELUNG FF publicly available dataset (19). DMCpGs were identified by genome-wide DNA methylation analysis. In this cohort (Table 1A), lung ADC (n = 181, 76.3%) was the most frequent histologic subtype, followed by SCC (n = 56, 23.6%). To obtain highly cancer-specific biomarkers, we focused our analysis in those regions deeply hypomethylated in nontumoral tissues. After data filtering and analysis with restrictive criteria (Supplementary Fig. S1; Supplementary Tables S1 and S2), we obtained 12 significant DMCpGs common to both subtypes of NSCLC corresponding to 9 different genes. In cancer cells, hypermethylation in CpG islands (CGI) is a principal epigenetic mechanism for gene regulation that has been proposed as a relevant biomarker with diagnostic value (22). Therefore, the top 5 hypermethylated CGI-containing genes were selected as candidate biomarkers for further validation in NSCLC: BCAT1, CDO1, TRIM58, ZNF177, and CRYGD (for extended explanation, see Materials and Methods, Fig. 1A, and Supplementary Table S2B).

To confirm these results, we evaluated the DMCpGs of the 5 selected biomarkers in an independent cohort (350 stage I NSCLC patients; 62 nontumoral lung samples) from TCGA public database. The clinical characteristics of this cohort (Table 1B) resembled the previous discovery cohort, including 217 (62.1%) ADCs and 133 (37.9%) SCCs. As expected, the methylation levels of the 5 selected genes were similar to the discovery cohort with difference in median values for each gene (ΔBCAT1: 59%; ΔCDO1: 40%; ΔTRIM58: 50%; ΔZNF177: 46%; ΔCRYGD: 40%) and all with P values lower than 0.001 (Fig. 1B). In addition, no significant differences were found between ADCs and SCCs (Supplementary Fig. S2A). These data confirmed our previous results, suggesting that the methylation of the 5 selected biomarkers is a common feature for both NSCLC subtypes despite their differences at histologic and molecular level.

Epigenetic silencing of the cancer-specific hypermethylated genes in lung cancer primary tumors

Gene expression analysis from the TCGA cohort samples showed a significantly decreased expression in BCAT1, CDO1, TRIM58, and ZNF177 (Fig. 1C). However, no expression values were detected for CRYGD and this gene was discarded for future analysis. Interestingly, expression results were also obtained for ADGs and SCCs separately (Supplementary Fig. S2B). Moreover, promoter hypermethylation of multiple consecutive CpGs is recognized as an important mechanism by which genes may be silenced in both physiologically and pathologic conditions (23).

This mechanism for gene silencing has also been shown to play a relevant functional role in the development and progression of many common human tumors (24). In this regard, analyzing the CURELUNG and TCGA datasets, we observed a similar methylation pattern between the significant DMCpGs of the selected biomarkers and their surrounding CpGs (Supplementary Fig. S3). These results reinforced the role of DNA methylation in the functional regulation of BCAT1, CDO1, TRIM58, and ZNF177. Importantly, the data obtained suggest that the methylation values of these four genes represent an epigenetic signature that may be relevant in early steps of lung carcinogenesis.

Diagnostic utility of the epigenetic signature to detect lung cancer in primary tumors

Once the epigenetic signature was established (BCAT1, CDO1, TRIM58, and ZNF177), we evaluated the ability of each individual biomarker of the four-gene panel to detect lung cancer in primary tumors by using pyrosequencing. This technique is a suitable approach in a clinical setting because it represents a quantitative and reproducible method able to detect multiple CpGs not only in FFPE tissues but also in minimally and noninvasive samples as biologic fluids. Therefore, an independent cohort of FFPE primary tumors (122 stage I NSCLC and 79 nonmalignant lung samples) was recruited and DNA methylation levels for all selected genes were determined by pyrosequencing. Clinical characteristics for
this cohort are described in Table 1C. The four biomarkers had significantly higher levels of DNA methylation in tumor samples as compared with nontumoral controls (Fig. 2A). Next, ROC analysis was performed to assess the diagnostic value of each individual biomarker to detect lung cancer. Importantly, all the genes of the signature showed significant areas under the ROC curve (AUC) greater than 0.8 (AUCBCAT1 = 0.94, AUCCD01 = 0.84, AUCTRIM58 = 0.97 and AUCZNF177 = 0.94), suggesting a great accuracy of these biomarkers for NSCLC diagnosis (Fig. 2B). Similarly, when samples were classified on the basis of histologic subtypes (ADC and SCC), we observed for all the biomarkers significant differences in methylation status (Supplementary Fig. S4A) and AUCs close to 1.0 (Supplementary Fig. S4B and S4C). These results confirmed the diagnostic value of evaluating DNA methylation levels by locus-specific PCR based techniques, such as pyrosequencing.

Validation of the epigenetic signature for lung cancer diagnosis using minimally invasive respiratory samples: bronchial aspirates and bronchioalveolar lavages

One of the most important aspects for early diagnostics is to identify markers associated with cancer using minimally invasive methods for sample collection (25). In line, we collected an independent cohort of BAS from patients diagnosed with lung cancer ($n = 51$) and cancer-free patients ($n = 29$; Table 1D). This cohort included different lung cancer subtypes, especially ADC and SCC. We compared by pyrosequencing the median methylation levels and generated ROC curves to assess the performance of each marker independently. Airways fluids from lung cancer patients presented significant differences in DNA methylation levels and high AUCs for all four genes (Fig. 3A and B). Combination of BCAT1, CD01, TRIM58, and ZNF177 in a logistic regression model yielded a significant AUC of 0.91 [95% CI (0.83–0.98) $P < 0.001$, Fig. 3C]. Calibration of the model showed no evident deviations from the ideal identity slope (data not shown). Internal validation of the AUC estimate for this model yielded optimism corrected AUC of 0.90, showing high generalization of the predictive capacity of the model for future samples. There were also no evident differences in prediction accuracy among early and late tumor stages.

A visual representation of the methylation profile for the genes included in the model is provided as a heatmap (Supplementary Fig. S5A). A nomogram based on the results of this model is proposed as a predictive tool for clinical diagnostic use. Results of the nomogram provide an individual probability (0%–100%) for suffering lung cancer for each patient (Supplementary Fig. S5B and Materials and Methods). Evaluation of the full range of predictions of the model shows that shifting the cutoff to POC = 30% would yield a sensitivity of 100% and a specificity of 65.4% and shifting the cutoff to POC = 80% would yield a sensitivity of 71.4% and a specificity of 92.3%. Sensitivity and specificity at the optimal cutoff (probability of cancer; POC = 63%) were 84.6% and 81.0%, respectively (Fig. 3D). It is important to point out that current protocols for lung cancer diagnosis are based mainly in bronchioalveolar cytology and further lung biopsy. There are cases where the cytology is doubtful or inconclusive. Moreover, there are a notable number of cases where cytology and biopsy are negative for cancer cells, but there is high
suspicion of cancer. Our results not only improve the overall prediction accuracy of BAS cytology in this cohort (sensitivity = 43.8%, specificity = 100%), but also permit a flexible and personalized approach for the clinicians in every possible scenario by simply adapting the cut-off value of the probabilistic model. In this sense, in our cohort 24 of 51 tumor samples were misinterpreted as nontumoral by the cytology test. However, using our predictive epigenetic model, 19 out of the 24 false negative cytologies (79%) would have been considered as positive setting our threshold at 50% probability of cancer (Supplementary Table S3). Of note, the majority of them (16/24) with a predicted probability of cancer higher than 80%. Also three of them were classified as borderline nontumor, with a predicted probability of cancer between 40% and 50%. In these three doubtful cases, clinical patient manage would require a closer follow up. This led us to propose our epigenetic signature as a useful clinical diagnostic tool in BAS specimens, especially in doubtful cases.

**Validation of epigenetic biomarkers in noninvasive sputum samples**

Finally, the methylation level of these 4 markers was examined in additional noninvasive samples. Sputums samples from 72 lung cancer patients and 26 cancer-free individuals were considered for evaluation (Table 1F). Methylation levels were significantly higher in individuals with lung cancer for all the genes tested, except for CDO1 (Fig. 5A). Individual AUC values were AUCBCAT1 = 0.92, AUCCDO1 = 0.67, AUCTRIM58 = 0.67, and AUCCZNF177 = 0.69 (Fig. 5B). The multivariable logistic regression model yielded an AUC value of 0.93 [95% CI (0.86–1.0), P < 0.001; Fig. 5C]. Sensitivity and specificity for the different threshold values of the model are depicted (Fig. 5D). This result suggests that our markers may be of high value to detect lung cancer even in noninvasive specimens as sputum.

**Discussion**

Lung cancer is the leading cause of cancer-related death worldwide with 1.3 million deaths annually, following data from the World Health Organization (WHO) in 2011. Late diagnosis in lung cancer is one of the main reasons that explain the extremely
high mortality of this disease. On one hand, screening by means of low-dose helical CT (LDCT) has shown to reduce mortality in a large randomized trial (26); however, the positive predictive value is still low. On the other hand, low sensitivity associated with minimally invasive cytologies is also a current hurdle for the accurate diagnosis of lung cancer. Thus, lung cancer diagnosis using minimally and noninvasive strategies is a major challenge to improve survival and its refinement is urgently needed to ameliorate the overall mortality figures for lung cancer worldwide.

Here, we have searched for powerful biomarkers by using the two largest publicly available databases (FP7 Curelung and TCGA; ref. 19) with high-throughput data coming from Infinium 450k arrays. Only stage I cancer cases were selected to identify the molecular changes associated to earlier steps of cancer evolution. We developed an integrative approach to identify the most discriminative marks leading to a final epigenetic signature consisting of top four selected genes: CDO1, BCAT1, TRIM58, and ZNF177. We conducted several validation steps using minimally and noninvasive cohorts to define a consistent epigenetic model useful for early lung cancer diagnosis. However, only one biomarker, SHOX2 methylation, has been commercialized to date (10,35), although is not routinely used in the clinic.

It is noteworthy to explain that cancer-specific DNA methylation in our selected biomarkers correlated with gene silencing in lung primary tumors. This fact suggests a potential functional role with biologic implications in early stages of this pathologic process (36). To our knowledge, there is a recent study addressing this issue with a different approach, taking benefit of the TCGA database: Wrangle and colleagues recently identified a three-gene panel (CDO1, HOXA9, and TAC1) for detecting NSCLC (12). They focused on reexpressed genes after treatment with demethylating agents and used TCGA as the only database incorporating all-stage tumors, not only stage I, among other differences. Interestingly, despite using different strategies, CDO1 methylation was common for both studies. On the other hand, a study combining miRNA and gene expression arrays in three lung squamous cell carcinoma patients has also identified methylation-deregulated CDO1 (37). CDO1, cysteine dioxygenase type 1, has been postulated as a tumor suppressor gene silenced by promoter methylation in multiple human cancers, including breast, esophagus,
lung, bladder, and stomach (38). For the other genes, BCAT1 (branched chain amino-acid transaminase 1) is a cytosolic enzyme that promotes cell proliferation though amino acid catabolism (39) and high frequency of methylation on BCAT1 promoter in colorectal cancer has been reported (40). ZNF177 is a zinc finger transcription factor that has been reported to be methylation-silenced in gastric cancer cell lines (41). TRIM58, tripartite motif containing 58, is an E3 ubiquitin ligase superfamily member that has already been patented as a potential epigenetic marker for detecting neoplastic cells originating from lung tissue of NSCLC patients (PCT/EP2012/061852). Moreover, it has also been reported hypermethylation of Trim58 promoter in hepatocytes derived from hepatitis B virus–related hepatocellular carcinoma (42). It is also worthy to indicate that we were very stringent in the selection of those genes, so alternative analyses from the same dataset may identify new DNA methylation biomarkers for lung cancer diagnosis in the future.

The diagnostic value of the epigenetic signature was first validated by pyrosequencing in FFPE samples from non–small cell lung primary tumors. Results from our four-gene methylation signature presented high diagnostic accuracy and were extremely similar to those obtained from public databases. Importantly, we analyzed a total of 79 nontumoral control tissues, and DNA methylation was almost negligible in the vast amount of samples, thus confirming previous results and encouraging the potential of the selected markers. Of note, in the study by Wrangle and colleagues, the methylation status was assessed by using the MSP technique, in a smaller number of FFPE nontumoral samples (12). We chose pyrosequencing, as targeted-region validation technique, because is an affordable and quantitative method that counterbalances some weaknesses of previous and extensively used methods, due to its easy standardization and lower false positive rate (43). Moreover, it is a robust and quantitative method able to detect multiple CpGs not only in FFPE tissues but also in minimally and noninvasive samples as biologic fluids with potential use in daily basis clinical settings.

The performance of the epigenetic model in these types of specimens, such as BAS, BAL, and sputum was outstanding despite the limited number of tumoral cells compared with FFPE samples. The improvement of the diagnosis of lung cancer patients represents a major challenge. Our epigenetic model provides a balanced and flexible approach able to cater to both extreme scenarios: the high sensitivity and low specificity of low dose CT in screening programs and the high specificity and low sensitivity of cytology (44,45) in respiratory specimens routinely used for lung cancer diagnosis. Our signature improves the predictions of cytology by providing a method for continuous predictions. Cytology is a useful dichotomized classifier producing two types of predictions: 100 % positive or 0% positive (100% negative). Therefore, the final output will be either a complete success or a total failure. In contrast, our signature based in a logistic regression model, represented by a nomogram, thus being able to produce a continuous range of predictions between 100% positive and 0% positive (46). That way, not all predictions are a complete success or a total failure, uncertainty can be measured for each prediction and errors are almost always lower (47). A clinician could take different actions according to the (un)certainty of the predictions, maybe performing additional tests in

Figure 5. Epigenetic signature in sputum samples using pyrosequencing. A, DNA methylation levels in sputums from patients with lung cancer and control donors. NT (light gray circle dots) stands for nontumoral and T (dark gray square dots) for tumor. P values for all the analyses were calculated using the two-sided Mann-Whitney U test. ***, P < 0.001; **, P < 0.01; *, P < 0.05. B, ROC curves and AUC for the selected genes. C, the AUC for the combined signature using a logistic regression model. D, sensitivity and specificity profiles for the different possible cut-off values of the results from the logistic regression model.
borderline cases. In a virtual situation where our model predicts two negative samples with different probability of being positive: such as 5% and 49%, the bimodal classifier predictor (cytology) would have output only absolute responses: negative and negative. Therefore, no information about uncertainty and chances of being positive for patient 1 (very low) and patient 2 (almost 50%) would have been delivered. The combination of current diagnostic protocol with new epigenetic nomograms will be of great help for diagnosis of lung cancer and consequently improving the outcome of lung cancer patients (48).

In summary, we have identified and independently validated a powerful epigenetic signature diagnosis of lung cancer in minimally and noninvasive samples. Genome-wide DNA methylation analyses led us to identify 4 candidates that have been tested not only in publicly available datasets, but also in extensive and independent cohorts of respiratory samples. The herein identified epigenetic model, once it will be validated in intended use samples such as in patients with suspicious indeterminate lung nodules, might be extremely helpful to solve these clinical issues with current diagnostic protocols and define more precise screening programs for lung cancer. In addition, novel and more sensitive methods, currently in development, such as Methyl-Beaming or droplet digital PCR (49) could enhance their diagnostic value for the management of suspicious lung nodules in the clinic or within a program of lung cancer screening.

Disclosure of Potential Conflicts of Interest
J. Zulueta has ownership interest (including patents) in VisionGate, Inc. No potential conflicts of interest were disclosed by the other authors.

Authors’ Contributions

Conception and design: A. Diaz-Lagares, J. Mendez-Gonzalez, J. Sandoval


Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J. Mendez-Gonzalez, M. Saigi, M.J. Pajares, R. Pio, L.M. Montuenga, J. Zulueta, E. Nadal, A. Rosell, M. Esteller, J. Sandoval

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): A. Diaz-Lagares, J. Mendez-Gonzalez, D. Hervas, D. Garcia, E. Nadal

Writing, review, and/or revision of the manuscript: A. Diaz-Lagares, J. Mendez-Gonzalez, D. Hervas, L.M. Montuenga, E. Nadal, A. Rosell, J. Sandoval

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M. Saigi, M.J. Pajares, D. Garcia, R. Pio, J. Zulueta, E. Nadal, A. Rosell

Study supervision: J. Mendez-Gonzalez, J. Sandoval

Acknowledgments

The authors thank Ana Remírez, Ana Moreno, Carles Arribas, Laia Seto, Lola Gonzalez, Alejandro Fernández, and Cesar García for technical support.

Grant Support

This work was supported by the "Miguel Servet" (CP00055) grant, FIS grant (PI10/0518) from the FEDER, FSE and Carlos III Health Institute (ISCIII), and PR13/143 research grant from the Mutua Madrileña Foundation. A. Diaz-Lagares was supported by a Rio Hortega (CM14/00067). A.B. Crujeiras and J. Mendez-Gonzalez are "Sara Borrell" researchers (C909/0036/SanCD13/00335 from the "ISCIII"). E. Nadal is supported by a Juan Rodés fellowship from the ISCIII (RI1/000027). L.M. Montuenga, R. Pio, and M.J. Pajares’ work was supported by the “Instituto de Salud Carlos III” projects PI13/00806 and RD12/0036/0040. M. Esteller is an ICREA Research Professor and J. Sandoval is a "Miguel Servet" researcher (CP13/00055). J. Sandoval, L.M. Montuenga, J. Zulueta, E. Nadal, and M. Esteller are supported by the "Red temática de investigación cooperativa en cancer" (RTICC); groups: RD 12/0036/0062, RD12/0036/0040, RD12/0036/0039, and RD12/0036/0045.

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.

Received September 29, 2015; revised December 23, 2015; accepted January 16, 2016, published OnlineFirst February 3, 2016.
Epigenetic Signature for Lung Cancer Diagnosis
