A novel epigenetic signature for early diagnosis in lung cancer

Angel Diaz-Lagares [1#], Jesus Mendez-Gonzalez [1#], David Hervas [2], Maria Saigi [3], Maria J. Pajares [4,5], Diana Garcia [1,&], Ana Belen Crujerias [8], Ruben Pio [4,6], Luis M. Montuenga [4,5], Javier Zulueta [7], Ernest Nadal [3], Antoni Rosell [9], Manel Esteller [1,10,$], Juan Sandoval-del Amor [1,&,$,*].

1 Cancer Epigenetics and Biology Program (PEBC), Bellvitge Biomedical Research Institute (IDIBELL), Avda Gran Via 199-203, L'Hospitalet, Catalonia, Spain

2 Biostatistics Unit, Medical Research Institute La Fe, Av. Fernando Abril Martorell 106, Valencia, Spain,

3 Department of Medical Oncology, Catalan Institute of Oncology, Avda Gran Via 199-203, L'Hospitalet, Barcelona, Spain

4 IDISNA and Program in Solid Tumors and Biomarkers, Centre for Applied Medical Research (CIMA), University of Navarra, Av. de Pío XII, 55, Pamplona, Spain.

5 Department of Histology and Pathology, School of Medicine, University of Navarra, Av. de Pío XII, 55, 31008 Pamplona, Spain.

6 Department of Biochemistry and Genetics, School of Science, University of Navarra, Av. de Pío XII, 55, 31008 Pamplona, Spain.

7 Pulmonary Department, Clínica Universidad de Navarra, Av. de Pío XII, 36, Pamplona, Spain

8 Laboratory of Molecular and Cellular Endocrinology, Health Research Institute of Santiago (IDIS), University Hospital of Santiago (XXIS/SERGAS), Santiago de Compostela University (USC), Travesia da Choupana S/N, Santiago de Compostela, Spain. CIBER of Physiopathology of Obesity and Nutrition (CIBERobn), Madrid, Spain.

9 Pneumology Department, Hospital University Bellvitge, IDIBELL, Avda Gran Via 199-203, L'Hospitalet, Barcelona and CIBER of Respiratory diseases (CIBERes), Madrid, Spain.

10 Catalan Institution for Research and Advanced Studies (ICREA); Passeig de Lluís Companys, 23, Barcelona, Catalonia, Spain

& Current address: Laboratory of Personalized Medicine, Epigenomics Unit, Medical Research Institute La Fe, Av. Fernando Abril Martorell 106, Valencia, Spain

# Both authors should be considered as first authors
$ Both authors should be considered as last authors

*** Corresponding author**
Juan Sandoval
E-mail: juan.sandoval@uv.es (JS)
Laboratory of Personalized Medicine,
Epigenomics Unit,
Medical Research Institute La Fe, Av. Fernando Abril Martorell 106

46026 Valencia (Spain)
Phone: 931246709

*** 2º Corresponding author**
Manel Esteller
E-mail:jsandoval@idibell.cat
Epigenetics and Biology Program
(PEBC), Bellvitge Biomedical Research Institute (IDIBELL), 3rd Floor
Hospital Duran i Reynals, Av.Gran Via de L'Hospitalet 199, 203
08908 L'Hospitalet de Llobregat, Barcelona (Spain).
Phone: 932607140

Running Title: Epigenetic signature for lung cancer diagnosis

Keywords: Epigenetics, diagnosis, lung cancer

Disclosure of Potential Conflicts of Interest: No potential conflicts of interest were disclosed.

Statement of 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 non-invasive 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.

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/non-invasive 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 non-invasive 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 to non-malignant 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 non-invasive samples. We demonstrated that this epigenetic signature achieved higher diagnostic efficacy in bronchial fluids as compared with conventional cytology for lung cancer diagnosis.

CONCLUSION: 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, since 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 (USPSTF) issued the recommendation to implement annual lung cancer screening for smokers with the inclusion criteria of the NLST. Nevertheless, there are still a significant number of open questions and areas for optimization of the different aspects related to this screening strategy. For example, there is a need for risk models and markers to improve the screening cost-benefit ratio by better selecting the screened population. Moreover, although the CT-based imaging is a very sensitive technique, its specificity is low, and it yields a large proportion of cases with indeterminate nodules, which may require further follow up or invasive procedures, which may turn out to be futile in the frequent case of these nodules being benign. Biomarkers for the correct classification of the indeterminate nodules and as an adjunct to the diagnostic procedure are a clear unmet clinical need (4,5).

Epigenetic biomarkers, mainly DNA methylation, have emerged as one of the most promising approaches to improve cancer diagnosis and present several advantages as compared to other markers, such as gene expression or genetic signatures. DNA methylation alterations are covalent modifications that are remarkably stable and often occur early during carcinogenesis.

Additionally, DNA methylation can be detected by a wide range of sensitive and cost-efficient techniques even in samples with low tumor purity. This epigenetic modification can also be detected in different biological fluids which represents a promising tool for minimally and non-invasive cancer detection (6). In recent years, different epigenetic candidates have been proposed, but none has reached the clinic yet, mainly due to the lack of large validation studies or the use of analytical methods difficult to standardize. Additionally, most studies were performed by single candidate-gene hypothesis-driven (7-11), although incipient genome-wide approaches are also appearing (12). Nowadays, high throughput epigenomic studies, that permit an unbiased data-driven research, have become a great tool for systematically dissecting the role of epigenetic variation in cancer with the potential of identifying novel and more robust biomarkers (13).

Bronchoscopic examination and pathological assessment of cytological specimen is the most currently used diagnostic method. However, almost half of the cases remain occult, especially in peripherally located tumors (14). This leads to additional invasive procedures, such as surgical lung biopsy or transthoracic needle biopsy associated with significant morbidity (15). The implementation of molecular biomarkers, including epigenetic and gene expression classifiers, in bronchial aspirates or sputum represents a promising approach to improve the accuracy of minimally and non-invasive neoplasm diagnosis (16,17). This kind of biomarkers can also be used to develop clinical tools such as nomograms, which allow calculating the probability of a clinical event. These predictive models can increase the individualized risk assessment compared with risk groups leading to a more personalized medicine (18).

Here, we have identified and validated a signature of DNA methylation biomarkers already present in early stage lung cancer and globally absent in normal tissue. For this purpose, we used two different datasets: the CURELUNG FP7 Consortium and the Cancer Genome Atlas (TCGA). Subsequently, we tested by pyrosequencing the selected biomarkers in several independent case-control datasets (formalin-fixed paraffin-embedded tissues, bronchial aspirates, bronchioalveolar lavages and sputum samples). This study provides a novel epigenetic predictive model that may help to improve lung cancer diagnosis.

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) (19) or from the TCGA data repository (lung adenocarcinoma LUAD or Lung squamous cell carcinoma LUSC). 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 iv) 98 sputums were obtained from Catalan Institute of Oncology and Bellvitge University Hospital, Barcelona. iii) 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 clinic-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 non-invasive 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 each sample of tumor tissue, subsequent sections were stained with hematoxylin and eosin for histological confirmation of the presence (>50%) of tumor cells. Unstained tissue sections were deparaffinized, and DNA was extracted using the same protocol as for minimally invasive specimens. Extracted DNA was checked for integrity and quantity with 1.3% agarose gel electrophoresis and picogreen quantification, respectively. Bisulfite conversion of 500 ng of DNA for each sample was performed according to the manufacturer's recommendation.

Data prefiltering

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 non-tumor lung tissue samples (Fig. 1).

Data filtering for hypomethylated CpGs in non-tumoral 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 non-tumor 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 non-tumor cases; 1) Average and median of beta values lower than 0.1 and 2) the percentil 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 non-tumor samples

Differentially methylated CpGs between tumor and non-tumor 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 non-tumor lung tissue samples: second group) were compared. The following three measures were calculated:

1) Differences in average beta-values between groups higher than a set threshold.

   (MD = |μT – μNT|>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 non-tumor group. Difference between quartile 1 in tumor group and quartile 3 in non-tumor should be higher that a selected cut-off. P25 (tumor)-P75 (nom-tumor)>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 histological subtypes.

a) Comparison: Adenocarcinoma (ADC) group (n:181) vs non-tumor group (n:25). Identified 29 differentially methylated CpGs (DMCpGs) specific for ADC (Supplementary table1A).
b) Comparison: Squamous cell carcinoma (SCC) group (n:56) vs non-tumour group (n:25). Identified 78 DMCpGs specific for SCC (Supplementary table 1B).
c) Comparison: Lung cancer (ADC and SCC)) (n:237) vs non-tumor group (n:25). Identified 24 significant DMCpGs when both groups are analyzed together (Supplementary table 1C).

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, Venn diagram analysis output final 12 common CpGs corresponding to 9 genes ranked by averaged z-score (Supplementary Fig. 1 and supplementary table 2).

Pyrosequencing

Pyrosequencing analyses to determine CpG methylation status were developed as previously described (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 2A). DNA was converted using the EZ DNA Methylation Gold (ZYMO RESEARCH) bisulfite conversion kit following the manufacturer's recommendations and used as a template for subsequent PCR step. PCR was performed under standard conditions with primers biotinylated to convert the PCR product to single-stranded DNA templates. We used the Vacuum Prep Tool (Biotage, Sweden) 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 bisulfite 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, standard deviation, 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 non-parametric 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 1000 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 vs. 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’s criterion. Additionally, 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 non-tumoral matched lung tissues from the CURELUNG FP7 publicly available dataset (19). Differentially DNA methylated CpGs (DMCpGs) were identified by genome-wide DNA methylation analysis. In this cohort (Table 1A), lung ADC (n=181, 76.3%) was the most frequent histological subtype, followed by SCC (n=56, 23.6%). In order to obtain highly cancer-specific biomarkers, we focused our analysis in those regions deeply hypomethylated in non-tumoral tissues. After data filtering and analysis with restrictive criteria (Supplementary Fig. 1, Supplementary Table 1 and 2), 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, Fig. 1A and Supplementary Table 2B).

To confirm these results, we evaluated the DMCpGs of the 5 selected biomarkers in an independent cohort (350 stage I NSCLC patients; 62 non-tumoral 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 (Δ_{BCAT1}: 59%; Δ_{CDO1}: 40%; Δ_{TRIM58}: 50%; Δ_{ZNF177}: 46%; Δ_{CRYGD}:

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. 2A). 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 histological 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 ADCs and SCCs separately (Supplementary Fig. 2B). Moreover, promoter hypermethylation of multiple consecutive CpGs is recognized as an important mechanism by which genes may be silenced in both physiologically and pathological 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. 3). 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 non-invasive samples as biological fluids. Therefore, an independent cohort of FFPE primary tumors (122 stage I NSCLC and 79 non-malignant 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 to non-tumoral controls (Fig. 2A). Next, receiver operating characteristics (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 (AUC_{BCAT1}=0.94, AUC_{CDO1}=0.84, AUC_{TRIM58}=0.97 and AUC_{ZNF177}=0.94), suggesting a great accuracy of these biomarkers for NSCLC diagnosis (Fig. 2B). Similarly, when samples were classified based on histological subtypes (ADC and SCC), we observed for all the biomarkers significant differences in methylation status (Supplementary Fig. 4A) and AUCs close to 1.0 (Supplementary Fig. 4B and 4C). 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 (BAS) and bronchioalveolar lavages (BAL)*

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 3B). Combination of *BCAT1, CDO1, 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. 5A). 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. 5B and materials and methods). Evaluation of the full range of predictions of the model shows that shifting the cut-off to POC=30% would yield a sensitivity of 100% and a specificity of 65.4% and shifting the cut-off to POC=80% would yield a sensitivity of 71.4% and a specificity of 92.3%. Sensitivity and specificity at the optimal cut-off (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 non-tumoral 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 3). Of note, the majority of them (16 of 24) with a predicted probability of cancer higher than 80%. Also three of them were classified as borderline non-tumor, with a predicted probability of cancer between 40% and 50%. In these three doubtful cases, clinical patient manage would require further additional studies. This led us to propose our epigenetic signature as a useful clinical diagnostic tool in BAS specimens, especially in doubtful cases.

Additionally, we evaluated DNA methylation levels in BAL from patients with lung cancer (n=82) as compared to non-malignant lung diseases (n=29) (Table 1E). The methylation levels of those four markers were significantly higher in BAL fluid from cancer patients than non-cancer patients (Fig. 4A). AUCs were significant for all four genes with the following values AUC_{BCAT1}=0.80, AUC_{CDO1}=0.65, AUC_{TRIM58}=0.72 and AUC_{ZNF177}=0.66 (Fig. 4B). Combination of the four genes in a logistic regression model achieved a significant AUC of 0.85 (95% CI [0.78, 0.93] $p < 0.001$), with an optimism-corrected value of 0.83 (Fig. 4C). Evaluation of the full range of predictions of the model is also shown (Fig. 4D). As in the case with BAS specimens, our epigenetic signature with diagnostic value may be highly valuable for doubtful patients with negative cytology.

Validation of epigenetic biomarkers in non-invasive sputum samples

Finally, the methylation level of these 4 markers was examined in additional non-invasive 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 AUC_{BCAT1}=0.92, AUC_{CDO1}=0.67, AUC_{TRIM58}=0.67 and AUC_{ZNF177}=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 non-invasive 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 computed tomography (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 non-invasive 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) (19) with high-throughput data coming from Infinium 450k arrays. Only stage I cancer cases were selected in order to identify the molecular changes associated to earlier steps of cancer evolution. We developed an integrative approach in order 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 non-invasive cohorts to define a consistent epigenetic model useful for early lung cancer diagnosis valid for both major histological subtypes. This signature yielded a notably high specificity, one of the Achilles heels of LDCT and other methylation biomarkers (27,28) and also improved sensitivity, which is generally limited when using cytology for early lung cancer diagnosis.

The current results highlight the relevance of DNA methylation changes in the natural history of lung cancer. CpG island hypermethylation of *MGMT* and *GSTP1* has already proven useful for the chemotherapy response prediction in gliomas (29-31) and the screening of prostate cancer, respectively (32,33). DNA methylation biomarkers have been proposed as promising candidates for early diagnosis (20,21) for several reasons: they are covalent and stable marks and they occur as early events in carcinogenesis, even in pre-tumoral stages such as adenomatous hyperplasia of the lung (34). Great efforts have been undertaken in identifying suitable DNA methylation markers to improve lung cancer diagnosis. However, only one biomarker —*SHOX2* methylation— has been commercialized to date (35,36), 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 biological implications in early stages of this pathological process (37). To our knowledge, there is a recent study addressing this issue with a different approach, taking benefit of the TCGA database: Wrangle et al. recently identified a three-gene panel (*CDO1, HOXA9* and *TAC1*) for detecting NSCLC (12). They focused on re-expressed genes after treatment with demethylating agents and used TCGA as the only database incorporating all-stage tumors, not only stage 1, among other differences. Interestingly, despite using different strategies, *CDO1* methylation was common for both studies. On the other hand, a study combining microRNA and gene expression arrays in three lung squamous cell carcinoma patients has also identified methylation-deregulated *CDO1* (38). *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 (39). For the other genes, *BCAT1* (Branched Chain Amino-Acid Transaminase 1) is a cytosolic enzyme that promotes cell proliferation though aminoacid catabolism (40) and high frequency of methylation on *BCAT1* promoter in colorectal cancer has been reported (41). *ZNF177* is a zinc finger transcription factor that has been reported to be methylation-silenced in gastric cancer cell lines (42). *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 (43). 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 non-tumoral 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 et al., the methylation status was assessed by using the MSP technique, in a smaller number of FFPE non-tumoral 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 (44). Moreover, it is a robust and quantitative method able to detect multiple CpGs not only in FFPE tissues but also in minimally and non-invasive samples as biological 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 to 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 (45,46) 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 (47). 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 (48). A clinician could take different actions according to the (un)certainty of the predictions, maybe performing additional tests in 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 (49).

In summary, we have identified and independently validated a powerful epigenetic signature diagnosis of lung cancer in minimally and non-invasive 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 of 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 (50) could enhance their diagnostic value for the management of suspicious lung nodules in the clinic or within a program of lung cancer screening.

Grant support

This work was supported by the “Miguel Servet” (CP00055) grant, FIS grant (PI08/1042) from the FEDER, FSE and Carlos III Health Institute (ISCIII) and PR185/13 research grant from the Mutua Madrileña Foundation. Angel Diaz-Lagares was supported by a Río Hortega (CM14/00067), Ana Belen Crujeiras and Jesus Mendez-Gonzalez are “Sara Borrell” researchers (C09/00365) and (CD13/00335) from the “ISCIII”. Ernest Nadal is supported by a Juan Rodés fellowship from the ISCIII (JR13/0002). Luis M. Montuenga, Ruben Pio and Maria J Pajares work was supported by the “Instituto de Salud Carlos III” projects PI13/00806 and RD12/0036/0040. Manel Esteller is an ICREA Research Professor and Juan Sandoval is a “Miguel Servet” researcher (CP13/00055). Juan Sandoval, Luis M Montuenga, Javier J Zulueta, Ernest Nadal and Manel Esteller are supported by the “Red temática de investigación cooperativa en cancer” (RTICCS); groups: (RD 12/0036/0062), (RD12/0036/0040), (RD12/0036/0039), (RD12/0036/0045).

Acknowledgements

We thank Ana Remírez, Ana Moreno, Carles Arribas, Laia Seto, Lola González, Alejandro Fernández and Cesar García for technical support.

Authors' Contributions

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

Development of methodology: J Mendez-Gonzalez, A Diaz-Lagares, AB Crujeiras, J Sandoval

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): M Saigi, MJ Pajares, Montuenga LM, Pio R, Rosell A, D Garcia, JJ Zulueta.

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

Writing, review, and/or revision of the manuscript: J Sandoval, A Diaz-Lagares, M Esteller, J Mendez-Gonzalez, LM Montuenga.

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

Study supervision: J Sandoval, M Esteller

Bibliography

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2013; 65:5-29.

2. International Early Lung Cancer Action Program Investigators, Henschke CI, Yankelevitz DF, Libby DM, Pasmantier MW, Smith JP, et al. Survival of patients with stage I lung cancer detected on CT screening. N Engl J Med 2006;355:1763-71.

3. Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365:395-409.

4. Massion PP, Walker RC. Indeterminate pulmonary nodules: risk for having or for developing lung cancer? Cancer Prev Res (Phila). 2014;7:1173-8.

5. Massion PP. Biomarkers to the rescue in a lung nodule epidemic. J Clin Oncol. 2014;32:725-6.

6. Heyn H, Esteller M. DNA methylation profiling in the clinic: applications and challenges. Nat Rev Genet 2012;13:679-92.

7. Belinsky SA, Klinge DM, Dekker JD, Smith MW, Bocklage TJ, Gilliland FD, et al. Gene promoter methylation in plasma and sputum increases with lung cancer risk. Clin. Cancer Res 2005;11:6505–11.

8. Topaloglu O, Hoque MO, Tokumaru Y, Lee J, Ratovitski E, Sidransky D, et al. Detection of promoter hypermethylation of multiple genes in the tumor and bronchoalveolar lavage of patients with lung cancer. Clin. Cancer Res 2004;10:2284–8.

9. Geng J, Sun J, Lin Q, Gu J, Zhao Y, Zhang H, et al. Methylation status of NEYROG2 and NID2 improves the diagnosis of stage I NSCLC. Oncol Lett 2012;3:901–6.

10. Schmidt B, Liebenberg V, Dietrich D, Schlegel T, Kneip C, Seegebarth A, et al. DNA methylation is a biomarker for the diagnosis of lung cancer based on bronchial aspirates. BMC Cancer 2010;10:600

11. Nikolaidis G, Raji OY, Markopoulou S, Gosney JR, Bryan J, Warburton C, et al. DNA methylation biomarkers offer improved diagnostic efficiency in lung cancer. Cancer Res 2012;72:5692-701.

12. Wrangle J, Machida EO, Danilova L, Hulbert A, Franco N, Zhang W, et al. Functional identification of cancer-specific methylation of CDO1, HOXA9, and TAC1 for the diagnosis of lung cancer. Clin Cancer Res 2014;20:1856-64.

13. Sandoval J, Esteller M. Cancer epigenomics: beyond genomics. Curr Opin Genet Dev 2012;22:50-5.

14. Griffin JP, Zaman MK, Niell HB, Tolley EA, Cole FH Jr, Weiman DS. Diagnosis of lung cancer: a bronchoscopist's perspective. J Bronchology Interv Pulmonol. 2012;19:12-8.

15. Kaarteenaho R. The current position of surgical lung biopsy in the diagnosis of idiopathic pulmonary fibrosis. Respir Res 2013;14:43.

16 Liloglou T1, Bediaga NG, Brown BR, Field JK, Davies MP. Epigenetic biomarkers in lung cancer. Cancer Lett. 2014;342:200-12.

17. Silvestri GA, Vachani A, Whitney D, Elashoff M, Porta-Smith K, Ferguson JS, et al. A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. N Engl J Med 2015;373:243-51.

18. Lowrance WT, Scardino PT. Predictive models for newly diagnosed prostate cancer patients. Rev Urol 2009;11:117-26.

19. Sandoval J, Mendez-Gonzalez J, Nadal E, Chen G, Carmona FJ, Sayols S, et al. A prognostic DNA methylation signature for stage I non-small-cell lung cancer. J Clin Oncol 2013;31:4140-7.

20. Belinsky SA. Gene-promoter hypermethylation as a biomarker in lung cancer. Nat Rev Cancer 2004;4:707-17.

21. Mikeska T, Bock C, Do H, Dobrovic A. DNA methylation biomarkers in cancer: progress towards clinical implementation. Expert Rev Mol Diagn 2012;12:473-87.

22. Sandoval J, Peiro-Chova L, Pallardo FV, García-Gimenez JL. Epigenetic biomarkers in laboratory diagnostics: emerging approaches and opportunities. Expert Rev Mol Diagn 2013;13:457-71.

23. Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome. Hum Mol Genet 2007;16:R50-9.

24. Baxter E, Windloch K, Gannon F, Lee JS. Epigenetic regulation in cancer progression. Cell Biosci 2014;4:45.

25. Heitzer E, Ulz P, Geigl JB. Circulating tumor DNA as a liquid biopsy for cancer. Clin Chem 2015;61:112-23.

26. National Lung Screening Trial Research Team, Church TR, Black WC, Aberle DR, Berg CD, Clingan KL, et al. Results of initial low-dose computed tomographic screening for lung cancer. N Engl J Med 2013;368:1980-91.

27. Palmisano WA, Divine KK, Saccomanno G, Gilliland FD, Baylin SB, Herman JG, et al. Predicting lung cancer by detecting aberrant promoter methylation in sputum. Cancer Res 2000;60:5954-8.

28. Leng S, Do K, Yingling CM, Picchi MA, Wolf HJ, Kennedy TC, et al. Defining a gene promoter methylation signature in sputum for lung cancer risk assessment. Clin Cancer Res 2012;18:3387-95.

29. Esteller M, Garcia-Foncillas J, Andion E, Goodman SN, Hidalgo OF, Vanaclocha V, et al. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. N Engl J Med 2000;343:1350-4.

30. Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med 2005;352:997-1003.

31. Barault L, Amatu A, Bleeker FE, Moutinho C, Falcomatà C, Fiano V, et al. Digital PCR quantification of MGMT methylation refines prediction of clinical benefit from alkylating agents in glioblastoma and metastatic colorectal cancer. Ann Oncol 2015;26:1994-9.

32. Jerónimo C, Usadel H, Henrique R, Oliveira J, Lopes C, Nelson WG, et al. Quantitation of GSTP1 methylation in non-neoplastic prostatic tissue and organ-confined prostate adenocarcinoma. J Natl Cancer Inst 2001;93:1747-52.

33. Hoque MO, Topaloglu O, Begum S, Henrique R, Rosenbaum E, Van Criekinge W, et al. Quantitative methylation-specific polymerase chain reaction gene patterns in urine sediment distinguish prostate cancer patients from control subjects. J Clin Oncol 2005;23:6569-75.

34. Licchesi JD, Westra WH, Hooker CM, Herman JG. Promoter hypermethylation of hallmark cancer genes in atypical adenomatous hyperplasia of the lung. Clin Cancer Res 2008;14:2570-8.

35. Schmidt B, Liebenberg V, Dietrich D, Schlegel T, Kneip C, Seegebarth A, et al. SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer based on bronchial aspirates. BMC Cancer. 2010;10:600.

36. Dietrich D, Kneip C, Raji O, Liloglou T, Seegebarth A, Schlegel T, et al. Performance evaluation of the DNA methylation biomarker SHOX2 for the aid in diagnosis of lung cancer based on the analysis of bronchial aspirates. Int J Oncol 2012;40:825-32.

37. Belinsky SA. Silencing of genes by promoter hypermethylation: key event in rodent and human lung cancer. Carcinogenesis 2005;26:1481-7.

38. Kwon YJ, Lee SJ, Koh JS, Kim SH, Lee HW, Kang MC, et al. Genome-wide analysis of DNA methylation and the gene expression change in lung cancer. J Thorac Oncol 2012;7:20-33.

39. Brait M, Ling S, Nagpal JK, Chang X, Park HL, Lee J, et al. Cysteine dioxygenase 1 is a tumor suppressor gene silenced by promoter methylation in multiple human cancers. PLoS One 2012;7:e44951.

40. Tönjes M, Barbus S, Park YJ, Wang W, Schlotter M, Lindroth AM, et al. BCAT1 promotes cell proliferation through amino acid catabolism in gliomas carrying wild-type IDH1. Nat Med 2013;19:901-8.

41. Mitchell SM, Ross JP, Drew HR, Ho T, Brown GS, Saunders NF, et al. A panel of genes methylated with high frequency in colorectal cancer. BMC Cancer 2014;14:54.

42. Yamashita S, Tsujino Y, Moriguchi K, Tatematsu M, Ushijima T. Chemical genomic screening for methylation-silenced genes in gastric cancer cell lines using 5-aza-2'-deoxycytidine treatment and oligonucleotide microarray. Cancer Sci 2006;97:64-71.

43. Tao R, Li J, Xin J, Wu J, Guo J, Zhang L, *et al.* Methylation profile of single hepatocytes derived from hepatitis B virus-related hepatocellular carcinoma. PLoS One 2011;6:e19862.

44. Doyle B, O'Riain C, Appleton K. Pyrosequencing of DNA extracted from formalin-fixed paraffin-embedded tissue. Methods Mol Biol 2011;724:181-90

45. Dobler CC, Crawford AB. Bronchoscopic diagnosis of endoscopically visible lung malignancies: should cytological examinations be carried out routinely? Intern Med J 2009;39:806-11.

46. Van't Westeinde SC, van Klaveren RJ. Screening and early detection of lung cancer. Cancer J 2011;17:3-10.

47. Harrell FE. Jr. Regression Modeling Strategies 2nd Edition. Springer; 2015. p. 4-6.

48. Dawid AP. Statistical Theory: The Prequential Approach. Journal of the Royal Statistical Society. Series A (General). The 150th Anniversary of the Royal Statistical Society; 1984; p. 278-292

49. Balachandran VP, Gonen M, Smith JJ, DeMatteo RP. Nomograms in oncology: more than meets the eye. Lancet Oncol 2015;16:e173-80.

50. Wiencke JK, Bracci PM, Hsuang G, Zheng S, Hansen H, Wrensch MR, et al. A comparison of DNA methylation specific droplet digital PCR (ddPCR) and real time qPCR with flow cytometry in characterizing human T cells in peripheral blood. Epigenetics 2014;9:1360-5.

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.

PATIENTS	A. DISCOVERY COHORT		B. TCGA VALIDATION COHORT		C. PARAFFIN COHORT	
	Tumor (n=237)	Non-tumoral (n=25)	Tumor (n=350)	Non-tumoral (n=62)	Tumor (n=122)	Non-tumoral (n=79)
Age (years)	68 (38-90)	63.5 (39-86)	66.9 (33-90)	66.9 (40-86)	63.8 (40-80)	62.5 (42-85)
Gender						
Male	131 (55.3%)	20 (80%)	190 (54.1%)	36 (58.0%)	108 (88.5%)	66 (83.5%)
Female	106 (44.7%)	5 (20%)	160 (45.9%)	26 (42.0%)	14 (11.5%)	13 (16.5%)
Smoking history						
Current or former smoker	190 (80.1%)	24 (96%)	313 (89.4%)	55 (89.4%)	70 (57.4%)	71 (89.9%)
Nonsmoker	25 (10.5%)	1 (4%)	32 (9.1%)	2 (9.1%)	36 (29.5%)	8 (10.1%)
Unknown	22 (9.4%)	0 (0%)	5 (1.5%)	5 (1.5%)	16 (13.1%)	0 (0%)
Stage						
I	237 (100%)		350 (100%)		122 (100%)	
Histology						
Adenocarcinoma	181 (76.4%)		217 (62.1%)		62 (50.8%)	
Squamous cell carcinoma	56 (23.6%)		133 (37.9%)		60 (49.2%)	
Pack-years	40 (0-180)	54.4 (0-184)	46.7 (1-94.5)	46.4 (5-192)	35.9 (0-130)	46.7 (0-141)

Data are average (range) or number (%).

PATIENTS	D. BAS COHORT		E. BALs COHORT		F. SPUTUM COHORT	
	Lung cancer patient (n=51)	Cancer-free donor (n=29)	Lung cancer patient (n=82)	Cancer-free donor (n=29)	Lung cancer patient (n=72)	Cancer-free donor (n=26)
Age (years)	65.6 (47-85)	64.0 (35-87)	62.1 (38-83)	57.5 (30-82)	65.1 (40-83)	52.7 (29-69)
Gender						
Male	46 (90.2%)	16 (55.2%)	66 (80.4%)	19 (65.5%)	62 (86.1%)	17 (65.4%)
Female	5 (9.8%)	10 (34.5%)	16 (19.6%)	9 (31.0%)	7 (9.7%)	9 (34.6%)
Unknown	0 (0%)	3 (10.3%)	0 (0%)	1 (3.5%)	3 (4.2%)	
Smoking history						
Current or former smoker	45 (88.2%)	16 (55.2%)	42 (51.2%)	15 (51.7%)	62 (86.1%)	20 (76.9%)
Nonsmoker	4 (7.8%)	8 (27.6%)	39 (47.5%)	12 (41.3%)	7 (9.7%)	6 (23.1%)
Unknown	2 (4.0%)	5 (17.2%)	1 (1.3%)	2 (7%)	3 (4.2%)	0 (0%)
Stage						
I	5 (9.8%)		17 (20.7%)		12 (16.7%)	
II	6 (11.8%)		8 (9.8%)		13 (18.0%)	
III	21 (41.2%)		20 (24.4%)		23 (32.0%)	
IV	18 (35.3%)		18 (22.0%)		19 (26.4%)	
Unknown	1 (1.9%)		19 (23.1%)		5 (6.9%)	
Histology						
Adenocarcinoma	17 (33.3%)		25 (30.5%)		38 (52.7%)	
Squamous cell carcinoma	19 (37.3%)		40 (48.8%)		24 (33.3%)	
Large cell carcinoma	2 (4.0%)		2 (2.4%)		2 (3%)	
Small cell carcinoma	2 (4.0%)		12 (14.6%)		5 (7%)	
NSCLC (NOS)	11 (21.4%)		3 (3.7%)		3 (4%)	
Pack-years	49.6 (0-120)	32.4 (0-100)	45.5 (0-120)	26.3 (0-90)	49.1 (0-120)	24.1 (0-114)

Data are average (range) or number (%).NSCLC (NOS): Non-small cell lung cancer. Not otherwise specified

Figure Legends

Figure 1. Epigenetic signature in lung primary tumor patients using genome-wide DNA methylation datasets. (A) DNA methylation levels of selected genes (Branched Chain Aminoacid Transaminase 1 -*BCAT1*-, Cysteine Dioxygenase type 1 -*CDO1*-, Tripartite Motif Containing 58 -*TRIM58*-, zinc finger protein 177 -*ZNF177*- and Crystallin, Gamma D -*CRYGD*-) in primary tumor samples from patients with lung cancer and non-tumoral specimens using our FP7 Curelung dataset. (B) Validation of DNA methylation values using public available dataset from The Cancer Genome Atlas database (TCGA). (C) Expression values for the gene candidates using the TCGA database. P values for all the analyses were calculated using the two-sided Mann–Whitney U test. NT (light grey circle dots) stands for non-tumoral and T (dark grey square dots) for tumor. *** correspond to p<0.001.

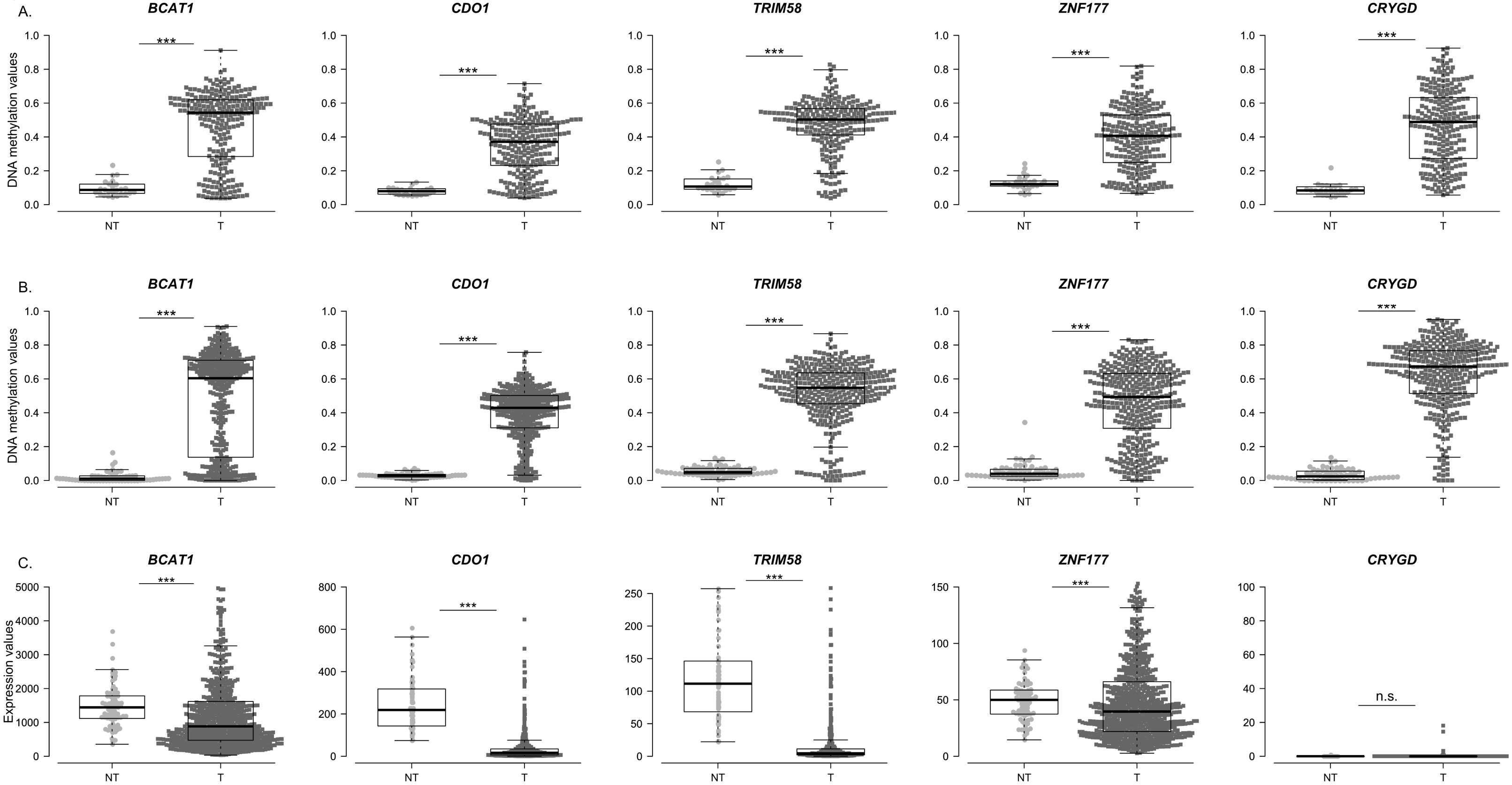
Figure 2. Epigenetic signature in paraffin samples using pyrosequencing. (A) DNA methylation levels of candidate genes in paraffin-embedded sections from patients with lung cancer and control donors. P values for all the analyses were calculated using the two-sided Mann–Whitney U test. NT (light grey circle dots) stands for non-tumoral and T (dark grey square dots) for tumor. *** correspond to p<0.001. (B) ROC curves and area under the curve (AUC) with 95% confidence intervals for the candidate genes.

Figure 3. Epigenetic signature in bronchial aspirates using pyrosequencing. (A) DNA methylation levels in bronchial aspirates from patients with lung cancer and control donors. NT (light grey circle dots) stands for non-tumoral and T (dark grey square dots) for tumor. P values for all the analyses were calculated using the two-sided Mann–Whitney U test. *** corresponds to p<0.001. (B) ROC curves and areas under the curve (AUC) for the selected genes. (C) The area under the curve (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.

Figure 4. Epigenetic signature in bronchioalveolar lavages using pyrosequencing. (A) DNA methylation levels in bronchioalveolar lavages from patients with lung cancer and control donors. NT (light grey circle dots) stands for non-tumoral and T (dark grey square dots) for tumor. P values for all the analyses were calculated using the two-sided Mann–Whitney U test. *** corresponds to p<0.001; * p<0.05. (B) ROC curves and areas under the curve (AUC) for the selected genes. (C) The area under the curve (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.

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 grey circle dots) stands for non-tumoral and T (dark grey square dots) for tumor. P values for all the analyses were calculated using the two-sided Mann–Whitney U test. *** corresponds to p<0.001; ** p<0.01 and * p<0.05. (B) ROC curves and areas under the curve (AUC) for the selected genes. (C) The area under the curve (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.

Figure 1.



A.
B.
C.
BCAT1
CDO1
TRIM58
ZNF177
CRYGD
DNA methylation values
Expression values
1.0
0.8
0.6
0.4
0.2
0.0
5000
4000
3000
2000
1000
0
800
600
400
200
250
150
100
50
100
80
60
40
20
***
n.s.
NT
T

Figure 2.

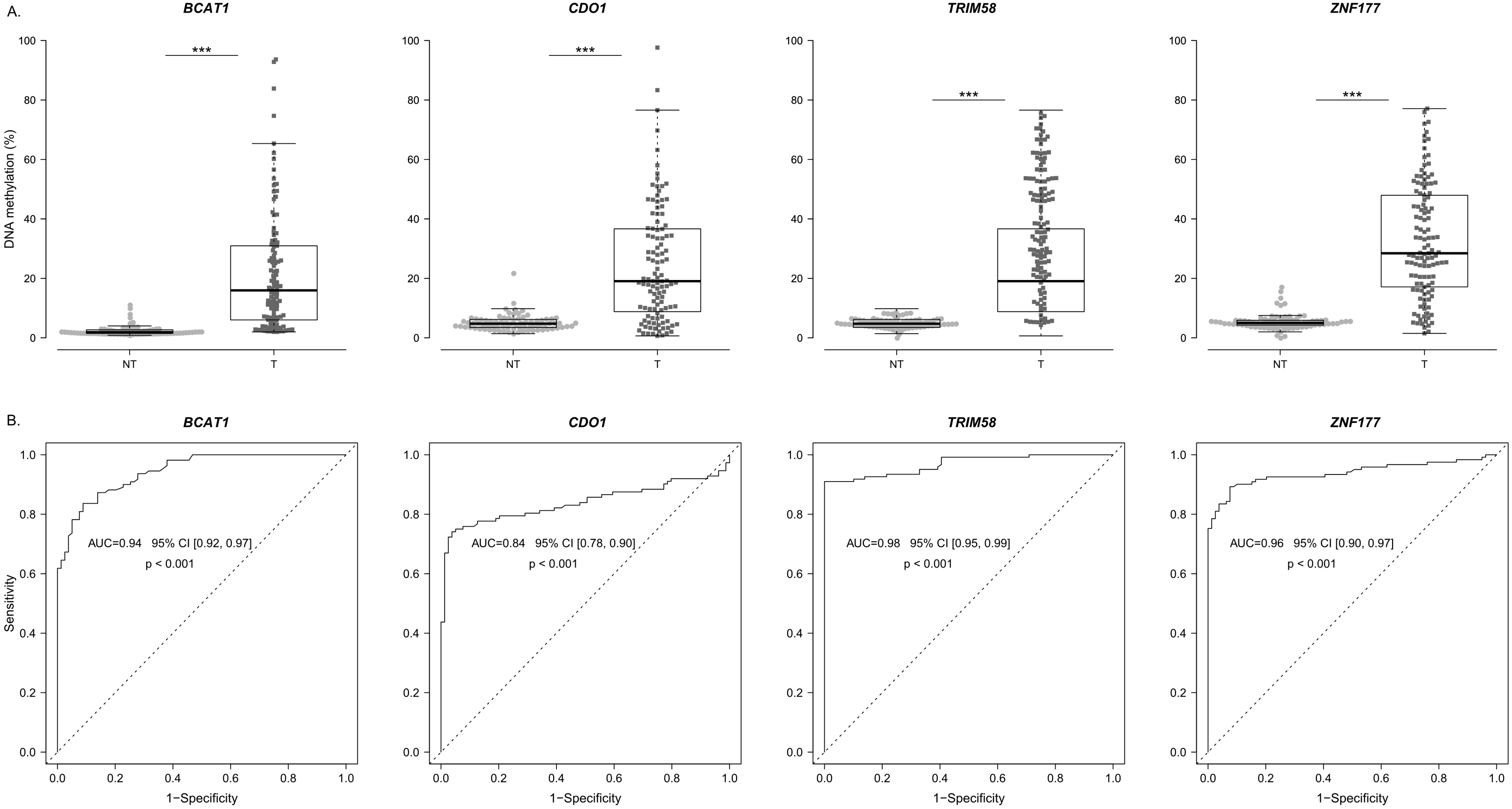

A.
BCAT1
CDO1
TRIM58
ZNF177
DNA methylation (%)
0
20
40
60
80
100
***
NT
T
B.
BCAT1
CDO1
TRIM58
ZNF177
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
AUC=0.94 95% CI [0.92, 0.97]
p < 0.001
AUC=0.84 95% CI [0.78, 0.90]
p < 0.001
AUC=0.98 95% CI [0.95, 0.99]
p < 0.001
AUC=0.96 95% CI [0.90, 0.97]
p < 0.001

Figure 3.

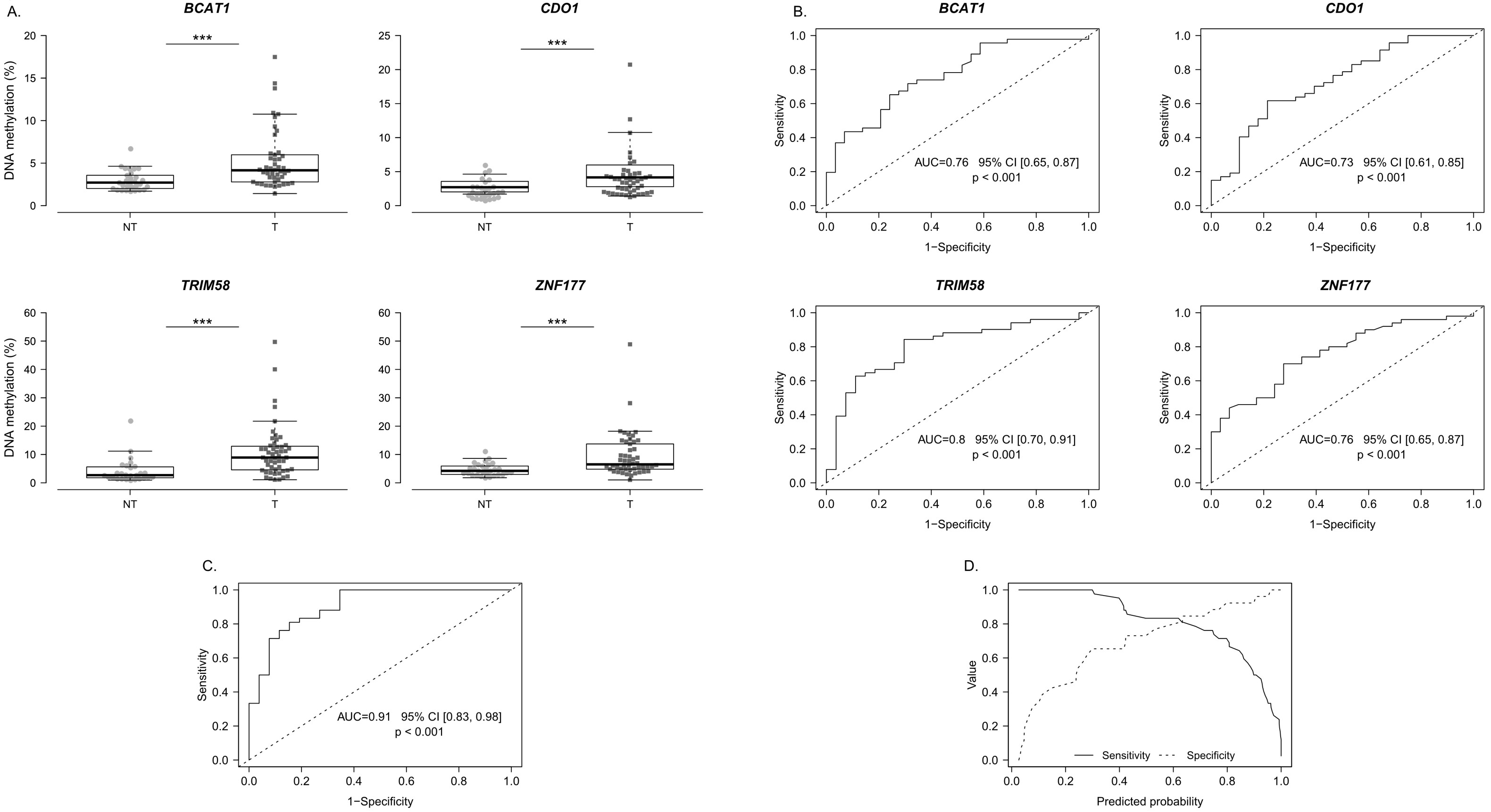

A.
BCAT1
CDO1
TRIM58
ZNF177
DNA methylation (%)
***
NT
T
B.
BCAT1
AUC=0.76 95% CI [0.65, 0.87]
p < 0.001
CDO1
AUC=0.73 95% CI [0.61, 0.85]
p < 0.001
TRIM58
AUC=0.8 95% CI [0.70, 0.91]
p < 0.001
ZNF177
AUC=0.76 95% CI [0.65, 0.87]
p < 0.001
Sensitivity
1−Specificity
C.
AUC=0.91 95% CI [0.83, 0.98]
p < 0.001
Sensitivity
1−Specificity
D.
Value
Predicted probability
Sensitivity
Specificity

Figure 4.

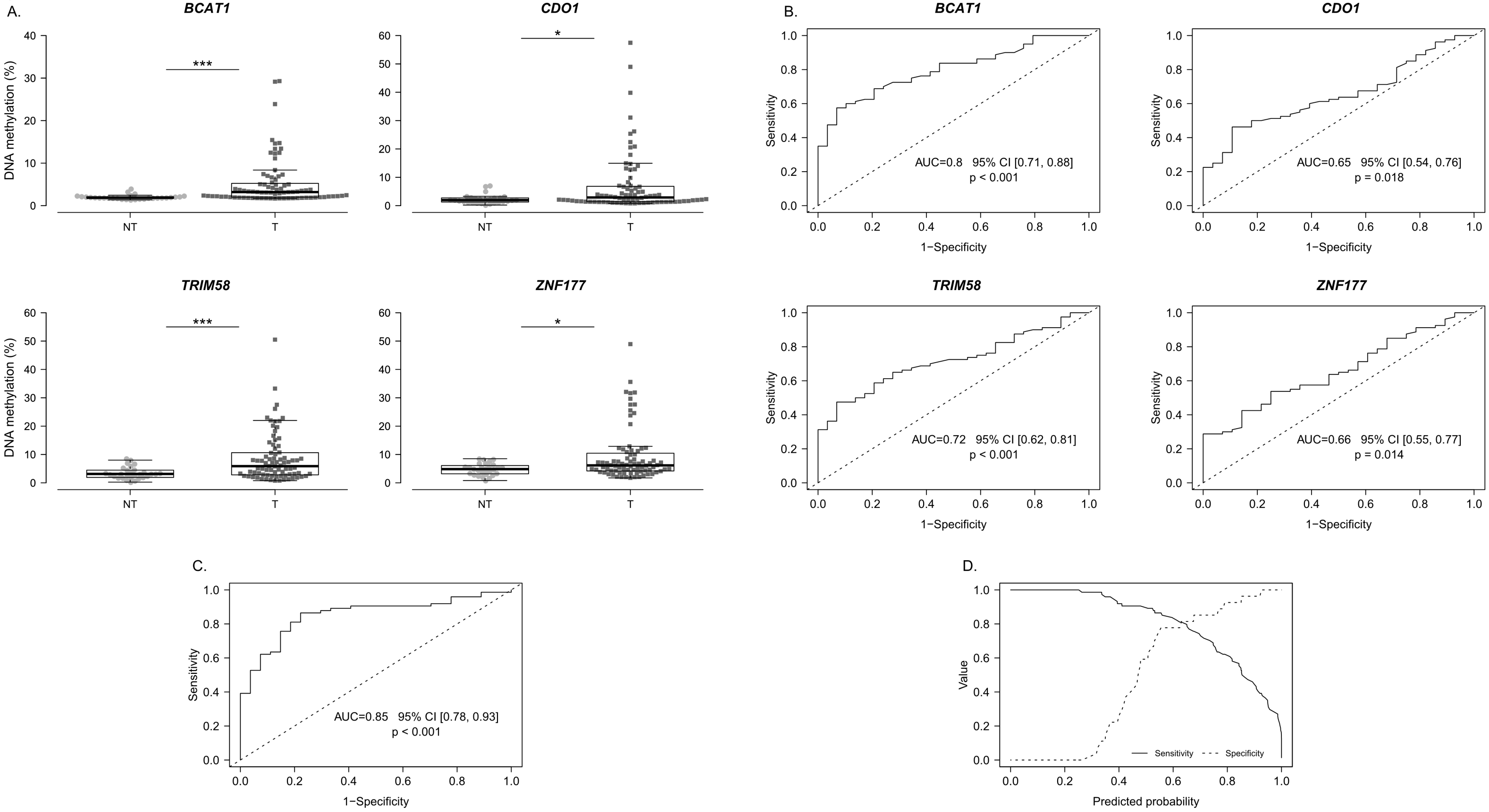
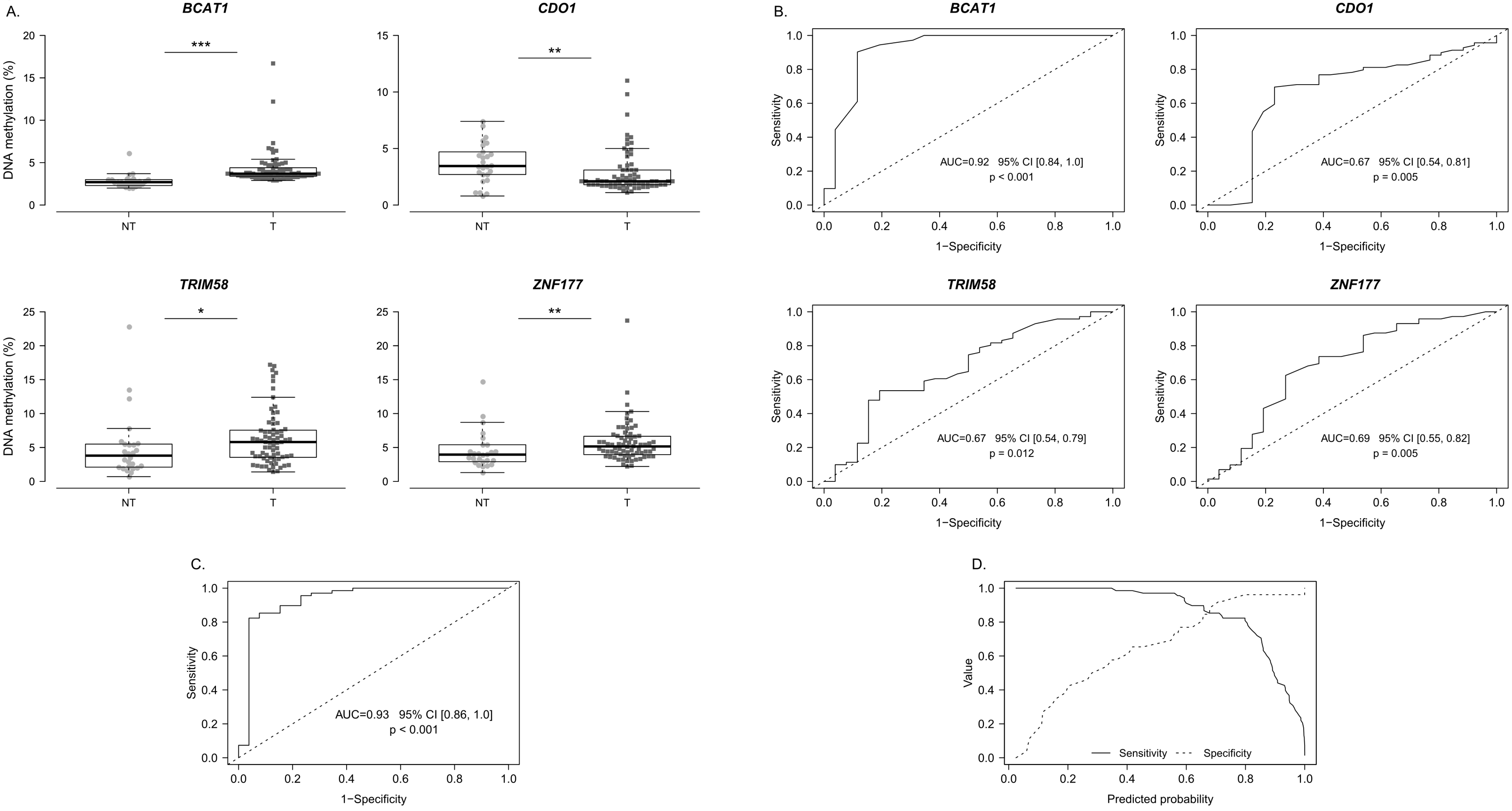

A.
BCAT1
***
CDO1
*
TRIM58
***
ZNF177
*
DNA methylation (%)
0
10
20
30
40
50
60
NT
T
B.
BCAT1
AUC=0.8 95% CI [0.71, 0.88]
p < 0.001
CDO1
AUC=0.65 95% CI [0.54, 0.76]
p = 0.018
TRIM58
AUC=0.72 95% CI [0.62, 0.81]
p < 0.001
ZNF177
AUC=0.66 95% CI [0.55, 0.77]
p = 0.014
Sensitivity
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0
C.
AUC=0.85 95% CI [0.78, 0.93]
p < 0.001
Sensitivity
1−Specificity
D.
Value
Predicted probability
Sensitivity
Specificity

Figure 5.



A.
BCAT1
CDO1
TRIM58
ZNF177
DNA methylation (%)
***
**
*
**
NT
T
0
5
10
15
20
25
B.
BCAT1
CDO1
TRIM58
ZNF177
Sensitivity
1−Specificity
AUC=0.92 95% CI [0.84, 1.0]
p < 0.001
AUC=0.67 95% CI [0.54, 0.81]
p = 0.005
AUC=0.67 95% CI [0.54, 0.79]
p = 0.012
AUC=0.69 95% CI [0.55, 0.82]
p = 0.005
0.0
0.2
0.4
0.6
0.8
1.0
C.
Sensitivity
1−Specificity
AUC=0.93 95% CI [0.86, 1.0]
p < 0.001
D.
Value
Predicted probability
Sensitivity
Specificity