Abnormal methylation of seven genes and their associations with clinical characteristics in early stage non-small cell lung cancer

Shicheng Guo1,5, Fengyang Yan1, Jibin Xu2, Yang Bao3, Ji Zhu4, Xiaotian Wang1, Junjie Wu1,5, Yi Li1, Weilin Pu1, Yan Liu6, Zhengwen Jiang6, Yanyun Ma1, Xiaofeng Chen7, Momiao Xiong8, Li Jin1,9, Jiucun Wang1,9

2016年9月28日

1 State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center for Genetics and Development, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University Jiangwan Campus, Shanghai 200438, China

2 Department of Cardiothoracic Surgery, Changzheng Hospital of Shanghai, Shanghai, China

3 Yangzhou No.1 People's Hospital, Yangzhou, China

4 Department of Cardiothoracic Surgery, Changhai Hospital of Shanghai, Shanghai, China

5 Department of Pneumology, Changhai Hospital of Shanghai, Shanghai, China

6 Center for Genetic & Genomic Analysis, Genesky Biotechnologies Inc., Shanghai, China

7 Department of Cardiothoracic Surgery, Huashan Hospital, Fudan University, Shanghai, China

8 Human Genetics Center, The University of Texas School of Public Health, Houston, Texas

9 Fudan-Taizhou Institute of Health Sciences, 1 Yaocheng Road, Taizhou, Jiangsu 225300, China

**Run title**: Novel panel of DNA methylation biomarkers for NSCLC diagnosis

Figures counts: 2

Table counts: 3

Abstract counts: 202

Mainbody counts: 3671

**Email**:

Shicheng Guo [Shicheng.Guo@hotmail.com](mailto:Shicheng.Guo@hotmail.com)

Fengyang Yan [yanfengyang1990@gmail.com](mailto:yanfengyang1990@gmail.com)

Jibin Xu [jibinx@yahoo.com](mailto:jibinx@yahoo.com)

Yang Bao [baoyangdr@gmail.com](mailto:baoyangdr@gmail.com)

Ji Zhu [luckypig222@163.com](mailto:luckypig222@163.com)

Xiaotian Wang [wangxiaot@gmail.com](mailto:wangxiaot@gmail.com)

Junjie Wu [wjjxcc@126.com](mailto:wjjxcc@126.com)

Yi Li [liyidadi19891124@gmail.com](mailto:liyidadi19891124@gmail.com)

Weilin Pu [puweilin@gmail.com](mailto:puweilin@gmail.com)

Yan Liu [liuyan@geneskies.com](mailto:liuyan@geneskies.com)

Zhengwen Jiang [zhengwenj@geneskies.com](mailto:zhengwenj@geneskies.com)

Yanyun Ma [mayymail@gmail.com](mailto:mayymail@gmail.com)

Xiaofeng Chen [cfx3166@126.com](mailto:cfx3166@126.com)

Momiao Xiong [momiao.xiong@gmail.com](mailto:momiao.xiong@gmail.com)

Li Jin [lijin.fudan@gmail.com](mailto:lijin.fudan@gmail.com)

Jiucun Wang [jcwang@fudan.edu.cn](mailto:jcwang@fudan.edu.cn)

**Corresponding authors**:

Li Jin, National Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200438, China, Phone: +86-21-51630607, Fax: +86-21-51630607, E-mail: [lijin.fudan@gmail.com](mailto:lijin.fudan@gmail.com)

Jiucun Wang, National Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China, Phone: +86-21-51630606, Fax: +86-21-51630607, E-mail: [jcwang@fudan.edu.cn](mailto:jcwang@fudan.edu.cn).

**Abstract**:

To identify novel abnormally methylated genes in early stage non-small cell lung cancer (NSCLC), we analyzed the methylation status of 13 genes (ALX1, BCL2, FOXL2, HPP1, MYF6, OC2, PDGFRA, PHOX2A, PITX2, RARB, SIX6, SMPD3 and SOX1) in cancer tissues from 101 cases of stage I NSCLC patients and lung tissues from 30 cases of non-cancerous lung disease controls, using methylation-specific PCR (MSP). The methylation frequencies (29.70-64.36%) of 7 genes (MYF6, SIX6, SOX1, RARB, BCL2, PHOX2A and FOLX2) in stage I NSCLC were significantly higher compared with those in non-cancerous lung disease controls (P<0.05). The co-methylation of SIX6 and SOX1, or the co-methyaltion of SIX6, RARB and SOX1 was associated with adenosquamous carcinoma (ADC), and the co-methylation of BCL2, RARB and SIX6 was associated with smoking. A panel of 4 genes (MYF6, SIX6, BCL2 and RARB) may offer a sensitivity of 93.07% and a specificity of 83.33% in the diagnosis of stage I NSCLC. Furthermore, we also detected the expression of 8 pathological markers (VEGF, HER-2, P53, P21, EGFR, CHGA, SYN and EMA) in cancer tissues of stage I NSCLC by immunohistochemistry, and found that high expression levels of p53 and CHGA were associated with the methylation of BCL2 (P=0.025) and PHOX2A (P=0.023), respectively. In this study, among the 7 genes which demonstrated hypermethylation in stage I NSCLC compared with non-cancerous lung diseases, 5 genes (MYF6, SIX6, PHOX2A, FOLX2 and SOX1) were found for the first time to be abonormally methylated in NSCLC. Further study of these genes shed light on the carcinogenesis of NSCLC.

**Key words**: Non-small cell lung cancer, DNA methylation, Biomarker, Batch effect elimination, Diagnosis

**Background**

  Lung cancer, a complex disease involving both genetic and epigenetic changes, is the leading cause of cancer deaths worldwide [1]. About 80% of primary lung cancers are non-small cell lung carcinoma (NSCLC) that are characterized by a long asymptomatic latency and poor prognosis. While the overall 5-year survival rates for late stage III and IV of NSCLC patients were just 5%-14% and 1% respectively, the rate could increase to 50% for the early stage of the NSCLC patients who are typically treated with surgery [2]. Many imaging and cytology-based strategies have been employed in NSCLC diagnosis; however none of them have yet been proven completely effective in reducing the mortality. The advances in molecular profiling of NSCLC over the past decade have made a paradigm shift in its diagnosis and treatment.

  Among all the genetic variations, single nucleotides polymorphisms (SNPs) have been considered as the most stable biomarker for heritable disease, since the status of the SNPs can be detected with almost 100% accuracy and unchanged during the entire life. It is specific and powerful for single gene caused disease. However, for complex diseases, such as cancers, the prediction power of SNPs is limited. The plethora of studies have shown that AUCs of the prediction model based on significant SNPs can confer only 0.54-0.55 for non-small cell lung cancer [3] and 0.54-0.60 for thyroid cancer [4], which has been considered as one of highest familial risk carcinomas among all kinds of cancers. Molecular biomarkers such as mRNA, microRNA and protein for NSCLC diagnosis have been developed and investigated in the past decades. However, their accuracy for diagnosis of NSCLC is far from reaching clinical implementations, in which >90% sensitivity and specificity of diagnosis should be guaranteed.

  DNA methylation that is one of the most important mechanism involved in genes and MicroRNAs expression regulation [5], gene alternative splicing [6], play important roles in the early stage of cancer. Because it is stable and easily detected qualitatively or quantitatively, DNA methylation was taken as the most promising diagnostic marker for the early detection of cancer[7, 8], are compared with SNP/mutation [4], CNVs [9] and gene/microRNA expression [10]. Hundreds of aberrant DNA methylation changes in the early stage of NSCLC have been identified in the past decades [11, 12]. However, despite several diagnostic panels have been developed [13], these studies on DNA methylation in NSCLC were still limited by their small sample size, low number of selected genes and qualitative rather than quantitative DNA methylation. These limitations would cause low reproducibility of the assay and explain why the majority of these studies could not be replicated.

  In this study, cancer tissues from 101 patients with stage I NSCLC and lung tissues from 30 patients with non‑cancerous lung diseases, were detected for the methylation status of 13 genes: PITX2 (paired-like homeodomain 2), RARB (retinoic acid receptor, β), OC2 (one cut homeobox 2), MYF6 (myogenic factor 6), PDGFRA (platelet-derived growth factor receptor, α polypeptide), SOX1 [SRY (sex determining region Y)-box 1], ALX1 (ALX homeobox 1), SIX6 (SIX homeobox 6), PHOX2A (paired-like homeobox 2a), FOXL2 (forkhead box L2), SMPD3 (sphingomyelin phosphodiesterase 3), BCL2 (B-cell CLL/lymphoma 2) and HPP1 (hyperpigmentation, progressive, 1). The methyaltion frequencies of 7 genes: MYF6, SIX6, SOX1, RARB, BCL2, PHOX2A and FOLX2 were significantly higher in stage I NSCLC than in non-cancerous lung diseases.

**Results**

**Methylation frequencies of the 7 genes differ significantly between stage**

  Methylation frequencies of the 7 genes differ significantly between stage I NSCLC and non-cancerous controls. Among the 13 genes, the methylation frequencies of 7 genes (MYF6, SIX6, SOX1, RARB, BCL2, PHOX2A and FOLX2) had significant difference between the group of stage I NSCLC and the group of non-cancerous lung diseases (Table III). ROC curves were constructed for each of the 7 genes to classify stage I NSCLC and non-cancerous lung disease. The AUC of the ROC curve for MYF6 was 0.704 (P<0.0001; 95% CI, 0.613-0.795), which was the largest among the 7 genes. The sensitivity and specificity of MYF6 were 64.36 and 93.33%, respectively, in the diagnosis of stage I NSCLC. The AUC of the ROC curves for the other 6 genes (SIX6, SOX1, RARB, BCL2, PHOX2A and FOLX2) ranged from 0.573 to 0.667; the sensitivity of each gene ranged from 29.70 to 51.49% and the specificity ranged from 73.33 to 93.33%, if they were used separately to diagnose stage I NSCLC. The methylation frequencies of the other 6 genes (ALX1, PDGFRA, PITX2, HPP1, OC2 and SMPD3) had no significant difference between tumors and controls, ranging from 24.75 to 59.41% in stage I NSCLC, and from 56.67 to 90.00% in non-cancerous lung diseases (Table III).

**Correlation between the methylation status of the 7 genes and the clinical characteristics of NSCLC**

  The correlations between the methylation status of the 7 genes, individually or combined, with each of the clinical characteristics was primarily assessed by univariate analysis, and the result was displayed in the form of a forest plot (Fig. 1). The methylation status of each of the 7 genes individually had no association with histological types, degree of differentiation or smoking. Next, the correlation between the co-methylation of two genes with the clinical characteristics was assessed in 21 possible pairs of genes. The co-methylation of SIX6 and SOX1 was negatively associated with adenocarcinoma (ADC) with an odds ratio (OR) of 0.24 (95% CI, 0.06-0.90). The co-methylation of BCL2 and RARB was associated with smoking (OR, 9.52; 95% CI, 1.20-75.49). We also analyzed the correlations between co-methylation of three genes with the clinical characteristics. The co-methylation of SIX6, RARB and SOX1 occurred less frequently in adenocarcinoma (ADC) than in squamous cell carcinoma (OR, 0.45; 95% CI, 0.25-0.76). The co-methylation of MYF6, SIX6 and FOLX2 (OR, 2.60; 95% CI, 1.11-6.66) or the co-methylation of SIX6, BCL2 and RARB(OR, 2.37; 95% CI, 1.27-4.44) were associated with smoking (Fig. 1).

   To verify these correlations, multivariate regression models were established. These indicated that the co-methylation of SIX6 and SOX1, as well as the co-methylation of SIX6, RARB and SOX1, was negatively associated with ADC; the latter association being more significant (SIX6 and SOX1: OR,0.243; 95% CI, 0.06-0.98; P=0.045; SIX6, RARB and SOX1: OR, 0.008; 95% CI, 0.001-0.149; P=0.007). The association between the co-methylation of SIX6, BCL2 and RARB and smoking has also been validated (OR, 3.09; 95% CI, 1.20-7.95; P=0.019). However, the association beween smoking and the co-methylation of BCL2 and RARB, or the co-methylation of MYF6, SIX6 and FOLX2, had no statistical significance (P>0.05; Table IV).

**A panel of 4 genes for the diagnosis of stage I NSCLC**

  Combination of several markers is a common strategy to improve diagnostic sensitivity in studies of clinical biomarkers. In this study, the most outstanding gene for the diagnosis of stage I NSCLC, MYF6, was found to be methylated in 65 of the 101 cases of patients with stage I NSCLC, displaying a sensitivity of 64.36%; and the methylation of MYF6 was also found in 2 of the 30 cases of patients with non-cancerous lung diseases, displaying a specificity of 93.3%. In the 36 cases of stage I NSCLC patients without MYF6 methylation, the methylation frequency of SIX6 was 41.67% (15/36), the highest among the 6 genes other than MYF6. Therefore, we made the first combination of MYF6 and SIX6 for the diagnosis of stage I NSCLC. The sensitivity was improved to 79.21%, while the specificity was dropped to 90.00%. However, the AUC of the ROC curve for the combination of MYF6 and SIX6 was 0.774 (P<0.0001; 95% CI, 0.681-0.866), higher than MYF6 alone, which meant that the combination of MYF6 and SIX6 was superior to MYF6 alone in diagnostic power. The methylation of BCL2 was detected in 8 of the 21 cases without methylation of either MYF6 or SIX6; more frequently than the other 4 genes, thus we made the second combination to form a 3-gene panel (MYF6, SIX6 and BCL2). The sensitivity, specificity and AUC were 87.13%, 86.67% and 0.812 (P<0.0001; 95% CI, 0.717-0.906), respectively. Using this method we analyzed a total of 6 panels of genes. The AUC of the 4-gene panel (MYF6, SIX6, BCL2 and RARB) was the largest among the them, and thus made it the best combination of markers in this study. The sensitivity, specificity and AUC of the 4-gene panel were 93.07%, 86.67% and 0.874 (P<0.0001; 95% CI, 0.787-0.960), respectively (Table V).

**Discussion**

  This study showed that 7 genes (MYF6, SIX6, SOX1, RARB, BCL2, PHOX2A and FOLX2) were frequently methylated in 101 cases of patients with stage I NSCLC, while rarely methylated in 30 patients with non-cancerous lung diseases. A panel of 4 genes (MYF6, SIX6, BCL2 and RARB) was able to diagnose stage I NSCLC from non-cancerous lung diseases with a sensitivity of 93.07% and a specificity of 86.67%.

  RARB and BCL2 have already been found to be hypermethylated in NSCLC (4,5). The protein encoded by SOX1 acts as transcription factor and plays a part in the regulation of embryonic development and in the determination of cell fate. Although the methylation of SOX1 have been found to be associated with genitourinary tumors, including cervical cancer, prostate cancer and ovarian cancer, this is the first time that methylation of SOX1 has been found in NSCLC (6-8). It has been reported that SOX1 antibodies are common in the serum of patients with small cell lung carcinoma (SCLC) and may be serve as specific serological markers (9). The methylation of 4 genes, MYF6, SIX6, PHOX2A, FOLX2, has never previously been reported in any types of cancer. MYF6 (12q21) is involved in muscle differentiation, SIX6 (14q23.1) is thought to be involved in eye development, and PHOX2A (11q13.2) is vital for development of the autonomic nervous system (10), FOXL2, as a forkhead transcription factor, may be involved in ovarian development and function. Further studies in the functions of these gene may help to reveal the mechanism of malignant transformation of non-small-cell lung cells.

  Multi-cellular issue is a great challenge in epigenetic studies. On one side, cancer tissues include cancer cells (epithelial cells), mesenchymal cells and so on. However, the proportion (at least 70% in general) of the tumor cells in cancer tissue is always much significantly higher than that of other cells. On the other side, normal tissues also include epithelial cells, mesenchymal cells and some others. In the present study, the null hypothesis is that the methylation level in the cancer tissue (mixed cells) is the same with normal tissue (mixed cells). The alternative hypothesis is that the methylation level in the cancer tissue (mixed cells) is different from normal tissue (mixed cells). We used the paired t-test to test the difference in the mean of the methylation between cancer tissue and normal tissue. The background or the noises from the adjacent non-cancer cells could be adjusted from the cancer cells when the methylation profiles of the corresponding cells were established.

  Tumorigenesis is an intricate process, involving a variety of genetic and epigenetic aberrations. Even a single tumor‑related gene may simultaneously display several types of abnormalities, and contribute to tumorigenesis through several different ways. In this study, 5 genes were for the first time found to be hypermethylated in NSCLC, and the function of those genes and how they act in the carcinogenesis of NSCLC is worth further exploration.

**Conclusion**

  Integrated analysis of multiple-platform high throughput DNA methylation microarray datasets followed by batch effect elimination is a good approach to discover diagnostic biomarker panels for NSCLC. Methylation profiles of AGTR1, GALR1, SLC5A8, ZMYND10 and NTSR1, would be an effective methylation-based assay for the NSCLC diagnosis.

**Materials and Methods**

**Study design and pipeline description**

  The clinical tissue samples from 101 stage I NSCLC patients and 30 patients with non-cancerous lung diseases used in this study were obtained from the Shanghai Chest Hospital (Shanghai, China) and the First Affiliated Hospital of Guangxi Medical University (Nanning, China). Informed consent was obtained from the patients and the study was approved by the Medical Institutional Review Boards of the two hospitals. Tumor-node-metastasis (TNM) staging/classification of the patients was performed according to the WHO classification. Table I shows the clinical patient profiles.

  When the microarray is provided as fluorescent signals, the gene methylation level was calculated with the fluorescent signals of methylation and un-methylation alleles by the traditional function of beta = (maxa(M,0))/(maxa(M,0)+maxa(U,0) ), where M and U represent the signal intensities for about 30 methylated (M) and un-methylated (U) probes on the array. Background-correction was conducted with recommended methods for each platform. K-nearest neighbor imputation (KNN imputation) was performed to deal with the missing values. 112 probes were shared between these two microarray platforms. DNA methylation signals of these probes were combined for all the samples. Quantile normalization was applied to combine all the data from different studies. To further reduce biases, we use the batch effect elimination tool, ComBat, to eliminate the batch effects that exist in independent datasets [15]. In the present study, we use the principal component analysis (PCA) to visualize the extension of the elimination of batch effect by observing the batch information distribution in the two-dimension plot of principle component 1 (PC1) and principle component 2 (PC2). The data adjusted by the Combat was then used for feature selection procedure in classification and differential methylation analysis. Feature selection was conducted by random forest and SVM with leave-one cross-validation. Differential methylation analysis was conducted by Wilcox signed-rank test without normality assumption. The most powerful panel was identified and the differential methylation status was estimated. In the validation stage, the methylation status of genes from the above panel (methylation genes combination) was detected in 150 NSCLC and normal tissues from the Chinese Han population by MSD-SNuPET. Logistic regression model, random forest, support vector machine (SVM), and Bayes tree were used to classify NSCLC in the validation data with five-fold cross-validation.

**Patients, samples and DNA**

  NSCLC samples and corresponding normal lung tissues for validation study in the Chinese population were   obtained from 150 patients who underwent pulmonary resection for primary NSCLC at Changhai Hospital, Shanghai, China. The study was approved by Fudan University and Changhai Hospital and informed consents were obtained from the patients. Exclusion criteria included subjects with a family history of lung cancer, previous radiotherapy, and chemotherapy or adjuvant therapy before surgery. All tissues were immediately frozen at -80℃ after surgical resection. Histological examination and tumor-node-metastasis classification were conducted according to World Health Organization classification criteria [31] and AJCC Cancer Staging Manual, 7th Edition [32], respectively. Age, gender, smoking status, histology type, TNM stage and differentiation status were collected as the covariates when conducting association between DNA methylation and disease status. Smoking status was assigned to binary status: never and ever smoking. TNM stage was assigned to early stage (I and II) or late stage (III and IV) when it is necessary so that the sample size can be big enough to get the efficient statistic power.

**MSD-SNuPET: Methylation status dependent single nucleotide primer extension assay**

  DNA extraction and Bisulfite conversion were performed as previously described [33, 34]. Methylation status determined by the single nucleotide primer extension technique (MSD-SNuPET) was designed for the quantification of methylation at multiple methylated loci simultaneously. MSD-SNuPET was developed based on SNPshot technology to bisulfite converted CpG sites. Un-methylated cytosine would be converted to uracil when treated with bisulfite while methylated cytosine maintains as the cytosine. Therefore, methylation status detection can be detected by specific primer and PCR amplification. Primer 3.0 was used to design primer sets (called amplifying primer) which were applied to amplify genome regions including the target CpG sites. Allele-specific elongation primer was used to quantify the copy number of C and T alleles. Primer pairs were showed in Supplementary Table 2. PCR was performed in a final volume of 10 μL containing 1x HotStarTaq buffer, 3.0 mM Mg2+, 0.3 mM dNTP, 1 U HotStarTaq polymerase (Qiagen Inc. USA), 1 µl DNA template and 1µl multiple primer set. Amplifications were conducted in a GeneAmp PCR System 9700 thermal cycler (Applied Biosystems, Foster City, CA) with the following thermal cycling profile: denaturation for 2 min at 95°C, followed by 11 cycles, each consisting of 20 sec at 94°C, 40 sec at 60°C, 90 sec at 72°C, and a final extension step for 2 min at 72°C, respectively. Negative and positive controls were included in each run of PCR as described above. The products of the sequencing reactions were purified and SNaPshot analysis of single nucleotides extension for multiple loci operation was shown as in our previous works [35]. DNA sequencing was conducted with the 3730 DNA analyzer. GeneMapper 4.1 (Applied Biosystems, Co., Ltd., USA) was used to analysis the fluorescence signals that represent different alleles. DNA methylation level is positively correlated with the magnitude of the C allele (H\_C) and negative corrected with the magnitude of the T allele (H\_T) in MSD-SNuPET technique (Supplementary Figure 1). In order to quantitatively estimate the methylation level for each CpG site, standard calibration curve was established, in which synthetic DNA fragments of C and T alleles were mixed with C allele proportion at 10%, 20%, 30%, 35%, 40%, 50%, 60%, 70%, 75%, 80% and 90%, respectively. Then, a standard calibration curve could be fitted as a quadratic regression model: y = β\_0 x^2+β\_1 x, in which β\_0 and β\_1 are optimized parameters. x indicates the ratio of H and T alleles (H\_C/H\_T). In the present study, one technique and biological control were set. The reference site was a C site that was not in the CpG site, therefore a low methylation signal should be detected and non-significant association should be detected between cancer and normal samples. Methylation status of LINE-1 was taken as a biological control since we are clear that it is hypo-methylation in the cancer tissues.

**Statistical analysis and machine learning**

  We select methylated genes for classification by ranking genes with P-values for testing differential methylation between tumor and normal tissue samples. We use three test statistics: student -test, Wilcoxon rank sum test and Wilcoxon signed rank test statistic to test for differential methylation between two conditions for the normal distribution of methylation level, non-paired tumor and normal tissue samples and paired tumor and normal tissue samples, respectively. False discovery rate (FDR) correction was used for multiple test correction with the R function of p.adjust with fdr as a parameter. Euclidean distance and partitioning around medoids were used to conduct hierarchical cluster analysis. Logistic regression (Package stats), support vector machine (SVM, Package e1071), random forest based classification (Package randomForest) and Bayes tree (Package BayesTree) were used to classify the NSCLC tumor and normal tissues. The optimized prediction model was built with the best prediction accuracy in the training dataset and then the sensitivity, specificity, accuracy were obtained from logistic regression, SVM, random forest and Bayes tree model in the test dataset with previous parameters applied in the training stage. All statistical analyses were conducted in R [36]. Protein-protein interaction networks were constructed by String 9.0 to show the function network of the genes in our study [37].

**Abbreviations:**

NSCLC: non-small cell lung cancer MSD-SNuPET: methylation status determined single nucleotide primer extension technology AUC: area under the curve TCGA: the cancer genome atlas project AGTR1: angiotensin II receptor, type 1 GALR1: galanin receptor 1 NTSR1: neurotensin receptor 1 SLC5A8: solute carrier family 5, member 8 ZMYND10: zinc finger, MYND-type containing 10 LINE-1: long interspersed element-1

**Competing interests**

ZJ and YL are the founders and employees of Genesky Biotechnologies, respectively.

**Authors’ contributions**

SG, JW, and LJ contributed to the conception, design and final approval of the submitted version. SG contributed to the integrated analysis of multiple microarray datasets, batch effect elimination and the statistical analysis. FY conducted bisulfite treatment of DNA. All authors read and approved the final manuscript. JX, YB, JZ, XW, and JW collected samples and helped to draft the manuscript. YL, WP, MX helped in data cleaning and statistical analysis. YL, ZJ helped to conduct MSD-SNuPET. All authors read and approved the final manuscript.

**Acknowledgements**

We thank all participating subjects for their kind cooperation in this study. We thank Dr. Hongyan Xu (Department of Biostatistics and Epidemiology, Georgia Regents University) and Dr. Yan Sun (Department of Biomedical Informatics, School of Medicine, Emory University) for their critical review and comments. This research was supported by the Science and Technology Committee of Shanghai Municipality (11DJ1400102), National High-Tech Research and Development Program (2012AA021802), National Science Foundation of China (NSFC, 81172228, 30890034), Ministry of Science and Technology (2011BAI09B00), and 111 Project (B13016). The computations involved in this study were supported by Fudan University High-End Computing Center.

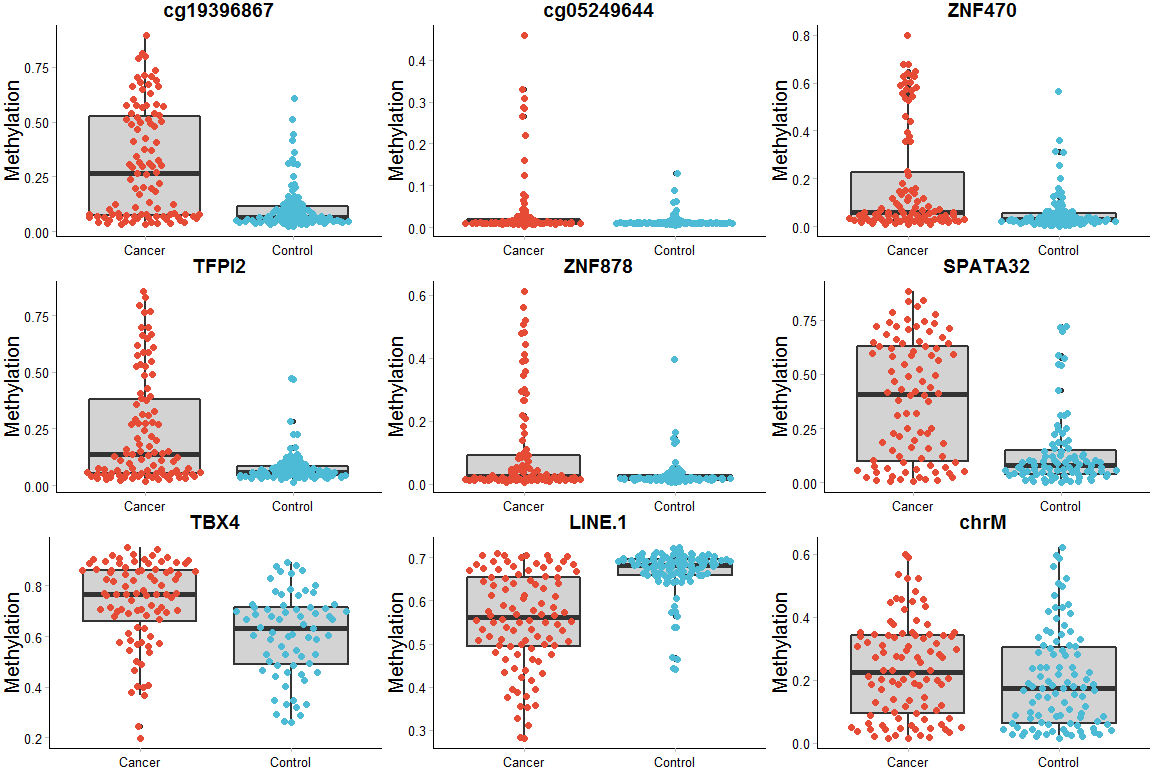
**Figure legends**

**Figure 1. Sketch of the study design and pipeline.**

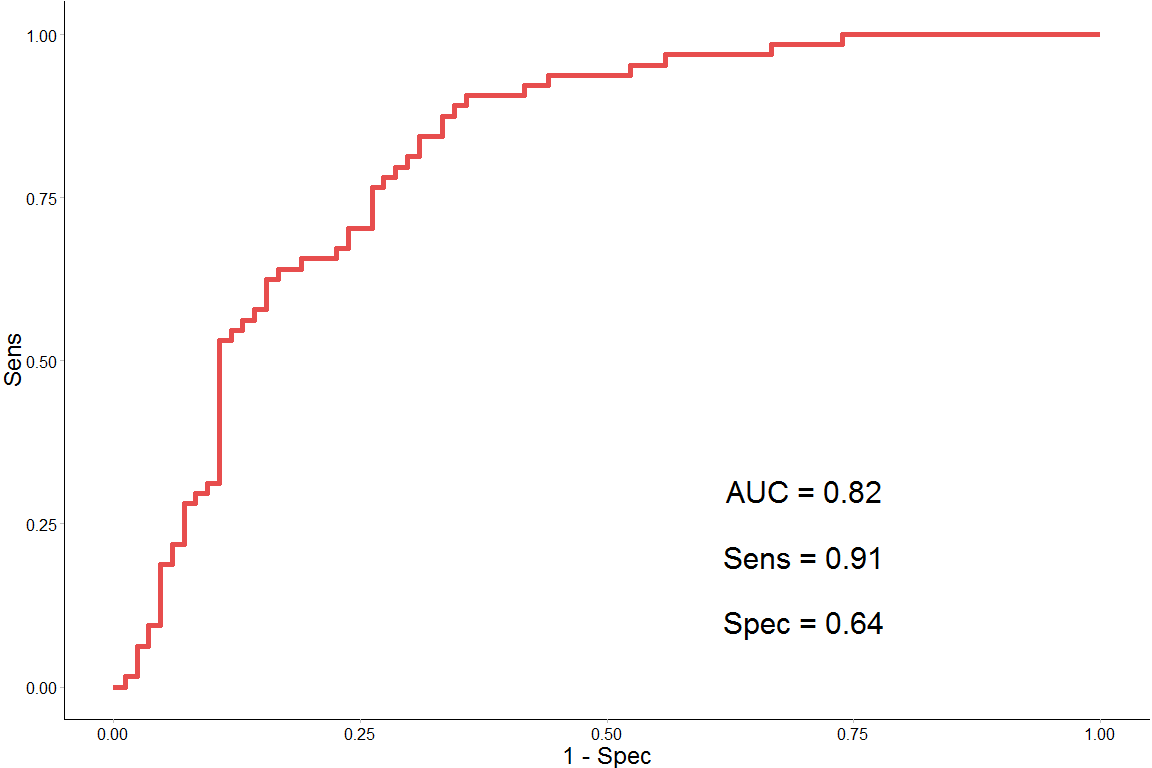
Candidate biomarkers were selected from Meta-analysis to multiple high-throughput DNA methylation microarrays. Significant or best feature combination was screened in an independent validation study of NSCLC with MSD-SNuPET technique.

**Figure 2. Combat treatment and MSD-SNuPET**

Principal component analysis was applied to show the efficiency of the elimination of ComBat. Figure 2A, 2B, a total of 120 probe sets with DNA methylation values after background and quantile normalization in a set of 352 NSCLC and 106 normal samples. X and Y axes represent the first and second principal components (PC1 and PC2), respectively. Figures 2C-I were validation of the methylation status of the five candidate markers in an independent samples. Y-axis represents absolute DNA methylation percentage from MSD- SNuPET. LINE-1 and Reference were taken as the positive and negative control for MSD- SNuPET.



**Figure 3. The ROC curve based on the logistic regression with the seven genes**



**Tables and footnote**

**Table1 Characteristics of patients**

**Table2 Differential methylation in ESCA**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | McaM | McoM | Pvalue | OR | CI.upper | CI.lower | Logistic.P | Sens | Spec | AUC |
| cg19396867 | 0.31 | 0.106 | 3.8e-09 | 2.69 | 3.7 | 1.83 | 4.9e-08 | 0.59 | 0.86 | 0.75 |
| cg05249644 | 0.04 | 0.016 | 8.9e-03 | 6.44 | 14.0 | 2.10 | 3.4e-02 | 0.53 | 0.73 | 0.61 |
| ZNF470 | 0.18 | 0.059 | 5.5e-06 | 2.37 | 3.7 | 1.33 | 1.2e-04 | 0.49 | 0.82 | 0.69 |
| TFPI2 | 0.24 | 0.078 | 2.8e-06 | 3.31 | 4.9 | 2.06 | 1.3e-05 | 0.52 | 0.91 | 0.70 |
| ZNF878 | 0.10 | 0.029 | 1.0e-02 | 3.86 | 6.6 | 1.96 | 1.1e-03 | 0.38 | 0.92 | 0.61 |
| SPATA32 | 0.38 | 0.139 | 1.8e-08 | 1.99 | 2.7 | 1.34 | 4.9e-08 | 0.63 | 0.84 | 0.75 |
| TBX4 | 0.73 | 0.608 | 2.8e-06 | 1.91 | 2.9 | 1.06 | 5.0e-05 | 0.58 | 0.80 | 0.72 |
| LINE.1 | 0.56 | 0.667 | 3.7e-14 | -7.11 | -5.1 | -9.47 | 1.6e-09 | 0.71 | 0.89 | 0.82 |
| chrM | 0.23 | 0.201 | 9.7e-02 | 0.58 | 1.4 | -0.21 | 1.5e-01 | 0.63 | 0.54 | 0.57 |

Differential methylation analysis was conducted between ESCA and adjacent normal tissues. McaM: Average methylation percentage of cases. McoM: Mean Methylation percentage of controls.P-valuea is the FDR adjusted P-value which were based on wilcoxon rank sum test compared the intensity of the methylation signals between case and control. log10(OR) and P-valueb represent log-transformed Odds ratio and p-value based on logistic regression adjusted by gender, age and smoking status. Sensitivity, specificity and AUC were calculated by logistic regression prediction model without adjustment for gender, age and smoking status.

**Table3 Diagnosis accuracy, sensitivity and specificity based on several classification methods with five-fold cross-validation**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Train.Sens | Train.Spec | Train.Accu | Test.Sens | Test.Spec | Test.Accu |
| Logistic Regression | 0.69 | 0.88 | 0.77 | 0.65 | 0.83 | 0.73 |
| Random Forest | 0.74 | 0.69 | 0.72 | 0.75 | 0.69 | 0.72 |
| SVM | 0.69 | 0.87 | 0.77 | 0.67 | 0.85 | 0.75 |
| Naive Bayes | 0.57 | 0.90 | 0.72 | 0.58 | 0.90 | 0.72 |
| Neural Network | 0.76 | 0.88 | 0.81 | 0.66 | 0.77 | 0.71 |
| Linear Discriminant Analysis | 0.68 | 0.86 | 0.76 | 0.67 | 0.84 | 0.75 |
| Mixture Discriminant Analysis | 0.71 | 0.85 | 0.77 | 0.64 | 0.78 | 0.70 |
| Flexible Discriminant Analysis | 0.68 | 0.86 | 0.76 | 0.67 | 0.84 | 0.75 |

Sensitivity, specificity and classification accuracy were its mean value in 5-fold validations with 1,000 replications. SVM represents support vector machines and Kernel Methods. In the main body of the manuscript, sensitivity, specificity and accuracy were derived from training result of the classification.