Identification and validation of the methylation biomarkers of Non-small cell lung cancer (NSCLC)

Shicheng Guo[1,5], Fengyang Yan[1], Jibin Xu[2], Yang Bao[3], Ji Zhu[2], Xiaotian Wang[1], Junjie Wu[2], Yi Li[1], Weilin Pu[1], Yan Liu[4], Zhengwen Jiang[4], Momiao Xiong[5], Li Jin[1,6], Jiucun Wang[1,6]

[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, Shanghai 200433, China
[2] Department of Cardiothoracic Surgery, Changhai Hospital of Shanghai, Shanghai, China
[3] Yangzhou No.1 People's Hospital, Yangzhou, China
[4] Center for Genetic & Genomic Analysis, Genesky Biotechnologies Inc., Shanghai
[5] Human Genetics Center, University of Texas School of Public Health, Houston, Texas
[6] 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

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

Corresponding authors:
Li Jin, National Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China, Phone: +86-21-55664885, Fax: +86-21-55664885, E-mail: 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: +86-21-55665499, Fax: +86-21-556648845, E-mail: jcwang@fudan.edu.cn.

Abstract:

DNA methylation was suggested as the promising biomarker for lung cancer diagnosis. However, it is a great challenge to search for the optimized combination of the methylation biomarkers to obtain the maximum diagnosis performance. In this study, we developed a panel of DNA methylation biomarkers and validated their diagnostic efficiency for non-small cell lung cancer (NSCLC) in a large Chinese Han NSCLC retrospective cohort. Three high-throughput DNA methylation microarray dataset (458 samples) were collected in the discovery stage. After normalization, batch effect elimination and integration, significant differentially methylated genes and the best combination of the biomarkers were determined by leave-one-out SVM (support vector machine) feature selection operation. Then candidate promoters were examined by MSD-SNuPET (methylation status determined single nucleotide primer extension technique) in an independent set of 150 pairwise NSCLC/normal tissues. Four statistical models with 5-fold cross-validation were used to evaluate the performance of the discriminatory algorithms. The sensitivity, specificity and accuracy were 86.3%, 95.7% and 91% respectively in Bayes tree model. The logistic regression model incorporated five gene methylation signature at AGTR1, GALR1, SLC5A8, ZMYND10 and NTSR1, adjusted with age, gender and smoking showed robust performance in which the sensitivity , specificity, accuracy, area under the curve (AUC) were 78%, 97%, 87%, and 0.91, respectively. In summary, high throughput DNA methylation microarray dataset followed by batch effect elimination can be a good method to discover optimized DNA methylation diagnostic panels. Methylation profiles of AGTR1, GALR1, SLC5A8, ZMYND10 and NTSR1, could be an effective methylation-based assay for the NSCLC diagnosis.

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 is 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 come up 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 most stable biomarker for heritable disease, since the status of the SNPs can be detected with almost 100% accuracy and unchanged all over the life. It is specific and powerful for single gene caused disease. However, for the complex disease, 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], comparing with SNP/mutation [4], CNVs [8] and gene/microRNA expression [9]. Hundreds of aberrant DNA methylation changes in the early stage of NSCLC have been identified in the past decades [10, 11]. However, despite several diagnostic panels have been developed [12], 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 majority of these studies could not be replicated.

In the present study, we first systematically integrated three independent high-throughput DNA methylation datasets from GEO [13] and TCGA project (**Supplementary Table 1**). Optimized DNA methylation combination was established through feature selection procedure after preliminary normalization and batch effect elimination among the datasets to maximum the NSCLC prediction performance. Five gene methylation statuses at *AGTR1*, *GALR1*, SLC5A8, *ZMYND10* and *NTSR1* were identified to be the most powerful combination for the NSCLC prediction. Then, to further evaluate their performance for diagnosis, we designed a novel methylation status determined single nucleotide primer extension technique (MSD-SNuPET) for the simultaneous quantification of methylation at these five methylated loci. These five significant differentially methylated genes were used to validate the results in 150 pairs of NSCLC and normal tissues from Chinese Han population with MSD-SNuPET.

Materials and Methods

Study design and pipeline description

Public high-throughput microarray databases which include GEO and ArrayExpress were searched to collect NSCLC related DNA methylation microarray data. Non-small cell lung cancer and/or Methylation were taken as the key words in the retrieving procedure. Although larger number studies have been conducted in NSCLC biomarker research, only two GSE records were retrieved, including GSE16559 and GSE28094. GSE16559 with 57 NSCLC and 52 normal tissue samples was to discover aberrant DNA methylation in lung adenocarcinoma and mesothelioma. GSE28094 with 33 NSCLC and 3 normal tissue samples was designed to make the DNA Methylation fingerprint with 1,628 human samples in different tissues and status. Both of these two datasets were based on *Illumina GoldenGate* platform which includes 371 genes with 1,536 loci. Additional, TCGA project is another comprehensive study. It included 262 NSCLC and 51 normal tissue samples. Infinium methylation 27K with 14, 495 genes and 27,578 loci were used to make the DNA methylation profiling. The common number of DNA methylation genes shared by two methylation microarray platforms was 107 genes (112 probes). Eventually, DNA methylation profiling data of 458 NSCLC

associated samples (352 NSCLC and 106 normal tissue) were obtained from the above three public datasets. These data will be taken as the primary data in the biomarker discovery stage (**Supplementary Table 1**).

When the microarray is provided as fluorescent signals, gene methylation level was calculated with the fluorescent signals of methylation and un-methylation alleles by the traditional function of beta $= \frac{\max(M,0)}{\max(M,0)+\max(U,0)}$, 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 batch effect elimination tools, *ComBat*, to eliminate the batch effects that exist in independent datasets [14]. In present study, we use the principal component analysis (PCA) to visualize the extension of the elimination of batch effect by observe 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 Wilcoxon 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 above panel (methylation genes combination) would be detected in 150 NSCLC and normal tissues from 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 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 [15] and AJCC Cancer Staging Manual, 7th Edition [16], 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 our previously described [17, 18]. Methylation status determined 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 1 μl 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 [19]. DNA sequencing was conducted with 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, standard calibration curve could be fitted as quadratic regression model: $y = \beta_0 x^2 + \beta_1 x$, in which β_0 and β_1 is optimized parameters. x indicates the ratio of H and T alleles (H_C/H_T). In present study, one technique and biological control were set. Reference site was a C site that was not in CpG site, therefore low methylation signal should be detected and non-significant association should be detected between cancer and normal. Methylation status of *LINE-1* was taken as biological control since we are clear that it is hypo-methylation in cancer tissues.

Statistical analysis and machine learning

We select methylated genes for classification by ranking genes with P-values for testing diffenretial methylation between tumor and normal tissue samples. We use three test statistics: student t-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. 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.

All statistical analyses were conducted in R [20]. Protein-protein interaction networks were constructed by *String 9.0* to show the function network of the genes in our study [21].

Results

Public dataset collection, Batch effect elimination and candidate gene selection

NSCLC related public DNA methylation microarrays were searched through Gene Expression Omnibus (GEO), ArrayExpress and TCGA project. In total, 3 independent NSCLC datasets were created with a total of 458 microarrays which included 352 NSCLC and 106 normal tissues **(Figure 1 and Supplementary Table 1)**. Batch effect was significantly existed among the datasets which was showed in the first and second principle components. We observed that the samples were clustered mainly by studies rather than tumor and normal tissue samples **(Figure 2A)**. *ComBat*, an empirical Bayes method, were used to eliminate the batch effects after quantile normalization to three datasets. As a result, batch effect was largely removed by *Combat* **(Figure 2B)**. In addition, as the hierarchical cluster analysis shown, biological information was highly preserved after batch effect elimination (**Supplementary Figure 2**). SVM were used to conduct feature selection and assess the prediction abilities with leaving-one-out cross-validation. The accuracy of the SVM for classifying NSCLC 98.98%, in the test set. Among the 112 shared probes, five CpG sites (*NTSR1*, *SLC5A8*, *GALR1*, *AGTR1* and *ZMYND10*) were selected in the feature selection stage. We found these five genes were significantly differentially methylated between tumor and normal tissue samples. In detail, meta-analysis of the DNA methylation microarrays showed that *NTSR1* (P-value = 5.4×10^{-15}), *SLC5A8* (P-value = 5.9×10^{-9}), *GALR1* (P-value = 9.9×10^{-10}) and *AGTR1* (P-value = 6.7×10^{-5}) were significantly hypermethylated in NSCLC while *ZMYND10* (P-values = 6.2×10^{-20}) was significantly hypomethylated in NSCLC **(Supplementary Figure 3)**. These results suggested that the selected five predictors would be potential biomarkers for the NSCLC diagnosis. To further evaluate their performance for diagnosis of NSCLC, we developed a panel of these five DNA methylation biomarkers and validate their diagnostic efficiency in 150 paired NSCLC and normal tissue samples in China.

Methylation status validation with MSD-SNuPET

In order to validate the result from Meta-analysis, methylation status of the above 5 genes were detected with MSD-SNuPET in 150 pairs NSCLC and adjacent normal tissues. The characteristics of patients were showed in **Table 1**. Consistent with the microarray data, the absolute DNA methylation percentage of these five genes were significantly differentially methylated between NSCLC and normal tissues (**Table2**, **Figure 2C-2I**). Logistic regression analysis showed that hypermethylated *NTSR1*, *SLC5A8*, *GALR1*, *AGTR1* and hypomethylated *ZMYND10* were significantly associated with the NSCLC risk adjusted for age, gender and smoking status after FDR multiple test correction with the P-value of 5.9×10^{-7}, 7.8×10^{-9}, 2.3×10^{-6}, 1.3×10^{-6}, and 5.2×10^{-8}, respectively **(Table 2)**. MSD-SNuPET results showed the methylation of *LINE-1* was significantly lower in NSCLC than normal tissue (*t*-test, P-value=6.03×10^{-14}). Additionally, DNA methylation of *LINE-1* was significantly associated with gender (R^2=0.18, P-value=0.0087), which was highly consistent with the previous reports about the methylation status of this gene [22, 23], suggesting the highly credibility of MSD-SNuPET. The prediction ability for each gene separately was also evaluated by logistic regression. Moderate prediction ability were identified, in which sensitivity ranges from 44.3%

to 73.15%; specificity ranges from 79.59% to 94.56% and AUC ranges from 0.67 to 0.80 (Table 2). Correlation analysis showed that there was no co-methylation among the five genes. In addition, no significant association was observed between any of the 5 genes with age, smoking, TNM stage, lung cancer differentiation and lung cancer subtype (Ad or Sc) in our study. However, significant association between gender and *SLC5A8* (P-value=0.0001), *ZMYND10* (P-value=0.045) were identified which might indicate specific biological mechanism of *SLC5A8* and *ZMYND10* in the tumorigenesis of NSCLC. Protein-protein interaction networks from *String 9.0* showed that there were comprehensive networks for both *NTSR1* and *GALR1*. Majority of these genes were cancer related genes, which has been reported to play important roles in cancer initiation, progress or therapy, such as *S100A9*, *NGF*, *TAC1*, *CCK*, *FPR2*, *ADRA1B*, and *CCL21* in the gene-gene interaction networks (**Supplementary Figure 4**).

Sensitivity, specificity and accuracy of the diagnosis panel

Several classification methods including logistic regression model, random forest, support vector machine (SVM), and Bayes tree were used to construct effective diagnosis models for cancer prediction. No significant unbalances were found in the train and test dataset, which suggested the prediction models were creditable and stable. Five-fold cross validation was used to evaluate the performance of the classifiers. As a result, Bayes tree was the most powerful model for diagnosis of NSCLC, whose sensitivity (Sen), specificity (Spe) and classification accuracy (Acc) were 86%, 96% and 91% (Table 3), respectively. Other classification methods had similar performance and the worst classifier was the logistic regression. However, even the logistic regression model incorporated the same five genes above mentioned to construct a methylation panel, and the sensitivity, specificity, classification accuracy, and area under the curve (AUC) could reach 78%, 97%, 87%, and 0.906 (95% CI: 0.89-0.91) after adjusted with age, gender and smoking. The logistic regression still showed the potential diagnostic significance of the five methylated genes. In addition, prediction abilities between smoking and non-smoking, adenocarcinoma and squamous cell carcinoma, early stage (I and II) and late stage (III and IV), and well or moderate and poor differentiation population were assessed under Bayes tree model. We found there is no significant differential performance between smoking (Acc=92.1%, 95% CI: 90.6%-93.6%) and non-smoking (Acc=0.939, 95% CI: 0.935-0.943), adenocarcinoma (Acc=0.82, 95% CI: 0.72-0.92) and squamous cell carcinoma (Acc=0.94, 95% CI: 0.87-0.95), early stage (Acc=0.87, 95% CI: 0.75-0.87) and late stage (Acc=0.92, 95% CI: 0.82-0.92), while a significant difference (permutation test, P<10-10) was found between well or moderate (Acc=0.9, 95% CI:0.83-0.91) and low differentiation population (Acc=0.73, 95% CI: 0.5-0.74), which suggested further research should be considered in the future.

Discussion

NSCLC early diagnosis and corresponding surgical intervention are taken as the most effective method to increase the survival time and to decrease the mortality of NSCLC death. Since the global change of DNA methylation occurred in the beginning of the carcinogenesis, DNA methylation has been considered as the most powerful biomarker for early detection, even screening [24]. In present study, two stage biomarker discovery pipeline was applied to

optimize the combination of DNA methylation biomarkers for NSCLC diagnosis. The optimal biomarker combination was identified using 107 genes in a large discovery dataset. A novel DNA methylation diagnosis panel of five genes (*NTSR1*, *SLC5A8*, *GALR1*, *AGTR1* and *ZMYND10*) was identified. The DNA methylation diagnosis panel was then validated in another independent NSCLC study. A multi-loci DNA methylation detection method (MSD-SNuPET), was conducted to determine the absolute quantitative methylation level of the five genes in 150 pairs of NSCLC and adjacent normal tissues from Chinese Han population. In the validation stage, Bayes tree model shows highest sensitivity, specificity and accuracy for NSCLC diagnosis based on the five genes which is potential for clinical application.

It is important that five candidate biomarkers have been investigated widely in cancer research. Neurotensin receptor-1 (*NTSR1*) is a G-protein coupled receptor (*GPCR*). It has been widely reported to be associated with carcinogenesis, cancer progression [25] and prognosis [26, 27]. Previous evidence showed the potential use of the *NTSR1* as a biomarker for cancer progression and as a component of personalized medicine in selective cancers [28] which is consistent with our present result. *GALR1*, galanin receptor subtype 2, suppresses cell proliferation in several cancers such as head and neck [29, 30], oral squamous cell carcinoma [31]. Gene expression inactivation of *GALR1* can be caused by promoter hypermethylation [29]. Meanwhile, *GALR1* has also been a subtype determining gene in breast cancer which suggests the potential powerful role in cancer diagnosis. *SLC5A8* (solute carrier family 5, member 8) is a tumor suppressor gene and is usually suppressed in colon, gastric cancers [32-34]. *ZMYND10* (Zinc finger, MYND-type containing 10) has recently been identified as a candidate tumor suppressor gene due to the occurrence of missense mutations and loss of its expression in lung cancer.

All the results in present study were based on quantitative signals of the DNA methylation. We also conducted the analyses which were based on discrete DNA methylation signals in which beta values < 0.3 was defined as the un-methylated CpGs; beta values > 0.8 were defined as the full methylated CpGs and beta values between 0.3-0.8 defines semi-methylated CpGs [35, 36]. In this condition, five genes were still significantly differential methylation between NSCLC and normal tissues. No significant changes were found in classification sensitivity, specificity and accuracy. Also, the sensitivity, specificity and AUC of diagnosis with one gene added to the model each time were summarized in the Supplementary Figure 5 in which we can found the sensitivity and AUC were gradually increased step by step.

Lung cancer diagnosis is a challenging problem. In order to discover a potential panel of DNA methylation-based biomarkers for diagnosis of NSCLC, We should perform a genome-wide search for an optional combination of tens or hundreds of loci from genome-wide DNA methylation profile. Integration analysis to inter-platform genome-wide DNA methylation datastes with appreciated data normalization and batch effect elimination could provide optimal biomarker combination in a large sample population to obtain maximum diagnosis efficiency. With this approach, we identified a 5 gene signature

including *AGTR1*, *GALR1*, *SLC5A8*, *ZMYND10* and *NTSR1*, which could provide highly diagnosis sensitivity and specificity.

Conclusion

Integrated analysis of multiple-platform high throughput DNA methylation microarray datasets followed by batch effect elimination can be taken as 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.

Competing interests

ZJ and YL are the founder and employee of Genesky Biotechnologies, respectively.

Authors' contributions

SG and JW, 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 statistical analysis. 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 National High-Tech Research and Development Program (2012AA021802), National Science Foundation of China (NSFC, 81172228), 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.

Reference

1. Siegel, R., D. Naishadham, and A. Jemal, *Cancer statistics, 2012.* CA Cancer J Clin, 2012. **62**(1): p. 10-29.
2. Hankey, B.F., L.A. Ries, and B.K. Edwards, *The surveillance, epidemiology, and end results program: a national resource.* Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 1999. **8**(12): p. 1117-21.
3. Li, H., et al., *Prediction of lung cancer risk in a Chinese population using a multifactorial genetic model.* BMC medical genetics, 2012. **13**: p. 118.
4. Guo, S., et al., *Significant SNPs have limited prediction ability for thyroid cancer.* CANCER MEDICINE, 2014.
5. He, Y., et al., *Hypomethylation of the hsa-miR-191 locus causes high expression of hsa-mir-191 and promotes the epithelial-to-mesenchymal transition in hepatocellular carcinoma.* Neoplasia, 2011. **13**(9): p. 841-53.
6. Flores, K., et al., *Genome-wide association between DNA methylation and alternative splicing in an invertebrate.* BMC Genomics, 2012. **13**: p. 480.
7. Laird, P.W., *The power and the promise of DNA methylation markers.* Nature reviews. Cancer, 2003. **3**(4): p. 253-66.
8. Jiang, F., et al., *A panel of sputum-based genomic marker for early detection of lung cancer.* Cancer

prevention research, 2010. **3**(12): p. 1571-8.
9. Zhu, J. and X. Yao, *Use of DNA methylation for cancer detection: promises and challenges.* The international journal of biochemistry & cell biology, 2009. **41**(1): p. 147-54.
10. Zhao, Y., et al., *Abnormal methylation of seven genes and their associations with clinical characteristics in early stage non-small cell lung cancer.* Oncology letters, 2013. **5**(4): p. 1211-1218.
11. Anglim, P.P., T.A. Alonzo, and I.A. Laird-Offringa, *DNA methylation-based biomarkers for early detection of non-small cell lung cancer: an update.* Molecular cancer, 2008. **7**: p. 81.
12. Nikolaidis, G., et al., *DNA methylation biomarkers offer improved diagnostic efficiency in lung cancer.* Cancer research, 2012. **72**(22): p. 5692-701.
13. Edgar, R., M. Domrachev, and A.E. Lash, *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository.* Nucleic Acids Research, 2002. **30**(1): p. 207-10.
14. Chen, C., et al., *Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods.* Plos One, 2011. **6**(2): p. e17238.
15. Gibbs, A.R. and F.B. Thunnissen, *Histological typing of lung and pleural tumours: third edition.* Journal of clinical pathology, 2001. **54**(7): p. 498-9.
16. Edge, S.B. and C.C. Compton, *The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM.* Annals of surgical oncology, 2010. **17**(6): p. 1471-4.
17. Zhao, Y., et al., *Methylcap-seq reveals novel DNA methylation markers for the diagnosis and recurrence prediction of bladder cancer in a Chinese population.* PLoS One, 2012. **7**(4): p. e35175.
18. Wang, X., et al., *Hypermethylation reduces expression of tumor-suppressor PLZF and regulates proliferation and apoptosis in non-small-cell lung cancers.* FASEB J, 2013. **27**(10): p. 4194-203.
19. Wang, Y.L., et al., *Confirmation of papillary thyroid cancer susceptibility loci identified by genome-wide association studies of chromosomes 14q13, 9q22, 2q35 and 8p12 in a Chinese population.* J Med Genet, 2013. **50**(10): p. 689-95.
20. Dessau, R.B. and C.B. Pipper, *["R"--project for statistical computing].* Ugeskr Laeger, 2008. **170**(5): p. 328-30.
21. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic acids research, 2011. **39**(Database issue): p. D561-8.
22. El-Maarri, O., et al., *Methylation at global LINE-1 repeats in human blood are affected by gender but not by age or natural hormone cycles.* PloS one, 2011. **6**(1): p. e16252.
23. El-Maarri, O., et al., *Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males.* Human genetics, 2007. **122**(5): p. 505-14.
24. Tsou, J.A., et al., *DNA methylation analysis: a powerful new tool for lung cancer diagnosis.* Oncogene, 2002. **21**(35): p. 5450-61.
25. Heakal, Y., et al., *Neurotensin receptor-1 inducible palmitoylation is required for efficient receptor-mediated mitogenic-signaling within structured membrane microdomains.* Cancer Biology & Therapy, 2011. **12**(5): p. 427-35.
26. Valerie, N.C., et al., *Inhibition of neurotensin receptor 1 selectively sensitizes prostate cancer to ionizing radiation.* Cancer Research, 2011. **71**(21): p. 6817-26.
27. Alifano, M., et al., *Neurotensin receptor 1 determines the outcome of non-small cell lung cancer.* Clinical Cancer Research, 2010. **16**(17): p. 4401-10.
28. Dupouy, S., et al., *The potential use of the neurotensin high affinity receptor 1 as a biomarker for cancer progression and as a component of personalized medicine in selective cancers.* Biochimie, 2011.

93(9): p. 1369-78.
29. Misawa, K., et al., *Epigenetic inactivation of galanin receptor 1 in head and neck cancer.* Clinical Cancer Research, 2008. **14**(23): p. 7604-13.
30. Kanazawa, T., et al., *Galanin receptor subtype 2 suppresses cell proliferation and induces apoptosis in p53 mutant head and neck cancer cells.* Clinical Cancer Research, 2009. **15**(7): p. 2222-30.
31. Henson, B.S., et al., *Galanin receptor 1 has anti-proliferative effects in oral squamous cell carcinoma.* Journal of Biological Chemistry, 2005. **280**(24): p. 22564-71.
32. Park, J.Y., et al., *Silencing of the candidate tumor suppressor gene solute carrier family 5 member 8 (SLC5A8) in human pancreatic cancer.* Pancreas, 2008. **36**(4): p. e32-9.
33. Ueno, M., et al., *Aberrant methylation and histone deacetylation associated with silencing of SLC5A8 in gastric cancer.* Tumour Biology, 2004. **25**(3): p. 134-40.
34. Miyauchi, S., et al., *Functional identification of SLC5A8, a tumor suppressor down-regulated in colon cancer, as a Na(+)-coupled transporter for short-chain fatty acids.* Journal of Biological Chemistry, 2004. **279**(14): p. 13293-6.
35. Ron-Bigger, S., et al., *Aberrant epigenetic silencing of tumor suppressor genes is reversed by direct reprogramming.* Stem Cells, 2010. **28**(8): p. 1349-54.
36. Richter, J., et al., *Array-based DNA methylation profiling of primary lymphomas of the central nervous system.* Bmc Cancer, 2009. **9**: p. 455.

Figure legends

Figure 1. Sketch of the study design and pipline.
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.

Tables and footnote

Table 1. Characteristics of patients

	NSCLC = 150
Age	40 (IQR = 15-65)
Gender	
Male	120
Female	30
Smoke Status	
Non-smokers (never)	41
Smokers (ever)	96
Histology	
Adenocarcinoma	53
Squamous cell carcinoma	63
others	34
Stage	
I (IA,IB)	42 (10,32)
II (IIA,IIB)	48 (16,32)
III (IIIA,IIIB)	46 (41,5)
IV	2
Differentiation	
Well/Moderate	74
Poor	30

Smokers include former and current smoker individuals. Others include adenosquamous carcinoma (ADSQ), bronchioloalveolar carcinoma, mucoepidermoid lung tumor, Sarcomatoid carcinoma. TNM Stages were assessed by the seventh edition of TNM classification criteria. Qualitative assessment of tumor differentiation was based on sum of the architecture score and cytologic atypia score (2 = well differentiated, 3 = moderately differentiated, 4 = poorly differentiated).

Table 2. Differential Methylation in NSCLCs

	NSCLC	Control	P-value[a]	$\log_{10}$(OR) (95%CI)	P-value[b]	Sen	Spe	AUC
AGTR1	12.88%	4.48%	1.06E-07	3.49 (2.08, 4.91)	1.30E-06	59.73%	79.59%	0.71
GALR1	18.31%	2.91%	6.58E-09	2.56 (1.5, 3.63)	2.30E-06	46.98%	85.03%	0.67
NTSR1	9.37%	0.56%	1.09E-09	9.02 (5.48, 12.55)	5.90E-07	44.30%	94.56%	0.70
SLC5A8	25.59%	11.66%	4.77E-12	3.80 (2.51, 5.09)	7.80E-09	52.35%	88.44%	0.67
ZMYND10	6.95%	12.82%	1.08E-07	-4.61 (-6.27, -2.95)	5.20E-08	73.15%	92.52%	0.80
LINE-1	72.10%	76.76%	2.39E-12	-10.3 (-13.5, -7.2)	1.80E-10	-	-	-
Reference	1.78%	1.83%	2.85E-01	-19.37 (-45.35, 6.62)	0.14	-	-	-

Differential methylation analysis was conducted between 150 NSCLC and adjacent normal tissues with logistic regression adjusted for gender and age. P-value[a] is the raw P-value based on logistic regression. P-value[b] represents p-value adjusted by FDR. Reference site was a C site that was not in CpG site, therefore, no or low methylated signal would be detected and non-significant association should be detected between cancer and normal tissues. Sensitivity, specificity and AUC were calculated by logistic regression prediction model without adjustment for gender, age and smoking status.

Table 3. Diagnosis accuracy, sensitivity and specificity based on several classification methods with five-fold cross-validation

	test			train		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Logistic Regression	0.791	0.993	0.891	0.775	0.969	0.871
SVM	0.897	0.977	0.937	0.855	0.941	0.897
Random Forest	0.934	0.928	0.931	0.890	0.886	0.886
Bayes Tree	0.911	0.976	0.944	0.863	0.957	0.909

AUC, 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.