**Aberrant methylation of *CDH13* can be a diagnostic biomarker for lung adenocarcinoma**

Weilin Pu1†, Xin Geng2†, Sidi Chen1, Lixing Tan1, Yulong Tan2, An Wang3, Zhouyi Lu3, Shicheng Guo4\*, Xiaofeng Chen2\*, Jiucun Wang1\*

1. State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China

2. Department of Cardiothoracic Surgery, Huashan Hospital, Fudan University, Shanghai 200032, China

3. Department of Chest Surgery, Shanghai Pulmonary Hospital, Shanghai 200433, China

4.Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, MC0412, La Jolla, CA 92093-0412

**Running title**: Diagnostic Role of *CDH13* Promoter Methylation in NSCLC

**Corresponding authors:**

Shicheng Guo, Department of Bioengineering, University of California at San Diego, 9500 Gilman Drive, MC0412, La Jolla, CA 92093-0412. Telephone: 281-685-5882, Fax: 858-534-5722, Email: [scguo@ucsd.edu](mailto:scguo@ucsd.edu)

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](mailto:jcwang@fudan.edu.cn).

Xiaofeng Chen, Department of Cardiothoracic Surgery, Huashan Hospital, Fudan University, Shanghai 200032, China, Phone: 52888299, E-mail: [cxf3166@126.com](mailto:cxf3166@126.com)

## Abstract

**Background**: Aberrant methylation of CpG islands in tumor cells in promoter regions is a critical event in non-small cell lung carcinoma (NSCLC) tumorigenesis and a potential diagnostic biomarker for NSCLC patients. The present study systemically and quantitatively reviewed the diagnostic ability of *CDH13* methylation in NSCLC as well as in its subsets. Eligible studies were identified through searching PubMed, Web of Science, Cochrane Library and Embase. The pooled odds of *CDH13* promoter methylation in lung cancer tissues versus normal controls were calculated by meta-analysis method. Simultaneously, four independent DNA methylation datasets of NSCLC from TCGA and GEO database were downloaded and analyzed to validate the results from meta-analysis. **Results**: thirteen studies, including 1850 samples were included in this meta-analysis. The pooled odds ratio of *CDH13* promoter methylation in cancer tissue was 6.42 (95% CI: 4.59 to 8.97, P < 0.00001) compared with that in controls under fixed-effect model. In validation stage, 126 paired samples from TCGA were analyzed and 5 out of the 6 CpG sites in the CpG island of *CDH13* were significantly hypermethylated in lung adenocarcinoma tissues but none of the 6 CpG sites was hypermethylated in squamous cell carcinoma tissues. Concordantly, the results from other three datasets, which were subsequently obtained from GEO database consisting of 568 tumors and 256 normal tissues, also consisted with those from TCGA dataset. **Conclusion:** The pooled data showed that the methylation status of the *CDH13* promoter is strongly associated with lung adenocarcinoma. The *CDH13* methylation status could be a promising diagnostic biomarker for diagnosis of lung adenocarcinoma.

## Key words

CDH13, DNA methylation, Non-small cell lung cancer, NSCLC, Diagnosis，Adenocarcinoma, Biomarker

## Background

Lung cancer is a complicated disease involving genetic and epigenetic variation, and is the leading cause of cancer death all over the world[[1](#_ENREF_1)]. Lung cancer is often silent in its early stages and difficult to diagnose in the early stages when treatment would be much more effective. Non-small cell lung cancer (NSCLC) comprises the majority of lung cancer and has an increasing incidence and mortality in the last two decades in China and in the world. The overall five-year survival rates for late stage III and IV of NSCLC patients were just 5%-14% and 1% respectively, however, the rate could come up to 63% for the early stage IA if treated with surgery properly [[2](#_ENREF_2), [3](#_ENREF_3)]

DNA methylation is one of the key epigenetic modifications in eukaryote, which regulates genes and microRNAs expression [[4](#_ENREF_4)], gene alternative splicing [[5](#_ENREF_5)], playing important role in the developing of cancer. Moreover, with the advantages of stable chemical property, detection ability in remote patient media, quantitative signal, relatively low cost in detection, DNA methylation has been regarded as a promising non-invasive biomarker to detect lung cancer in the early stage [[6](#_ENREF_6)].

The *CDH13* (cadherin 13) gene, an atypical member of the cadherin superfamily, was isolated recently and has been mapped to 16q24 [[5](#_ENREF_5)], which was devoid of a transmembrane domain and anchored to the exterior surface of the plasma membrane via a glycosylphosphatidylinositol anchor [[7](#_ENREF_7)]. Evidence showed that promoter methylation, which inhibits *CDH13* gene expression, is mediated by DNA methyltransferases Dnmt3A [[8](#_ENREF_8)]. Researchers have reported that hyper-methylation and loss of function of *CDH13* was detected in in breast [[9](#_ENREF_9)] and lung cancers [[10-12](#_ENREF_10)], in pituitary adenomas[[13](#_ENREF_13)], diffuse large B cell lymphoma [[14](#_ENREF_14)], and nasopharyngeal carcinoma[[15](#_ENREF_15)]. Furthermore, *CDH13* gene has been suggested as an early detecting marker for lung cancers[[16](#_ENREF_16)].

In the past decades, a large number of DNA methylation based cancer diagnostic biomarkers have been identified in NSCLC. The diagnostic or risk associations for several of them have been quantitatively evaluated such as SHOX2 [[17](#_ENREF_17)], APC [[18](#_ENREF_18)], RASSF1A [[19](#_ENREF_19)], FHIT [[20](#_ENREF_20)], MGMT [[21](#_ENREF_21)], RUNX3 [[22](#_ENREF_22)], RARbeta [[23](#_ENREF_23)], E-cadherin [[24](#_ENREF_24)] and P16 [[25](#_ENREF_25)]. However, the quantitative assessment of the diagnostic performance of the methylation status of CDH13 in NSCLC was not investigated. In this article, we conducted a meta-analysis of the sensitivity and specificity of *CDH13* methylation on NSCLC diagnosis. The studies included were homogeneous according to the low I−squared value. We also found that The Cancer Genome Atlas project (TCGA) and Gene Expression Omnibus (GEO) had collected hundreds of whole genome DNA methylation microarray datasets of NSCLC samples with comprehensive clinical and demographic information, providing additional resources that may be without publication bias[[26](#_ENREF_26)]. Therefore, in our work, we innovatively integrated these microarray data with the data from published articles to evaluate the diagnostic ability of the *CDH13* methylation test in NSCLC comprehensively.

## Results

### **Study characteristics**

The electronic search strategy identified 365 potentially relevant articles (PubMed, 73; Web of science, 177; Embase, 115; Cochrane Library, 0), which were further screened for inclusion on the basis of their titles, abstracts, full texts, or a combination of these terms. The electronic search was supplemented from reference lists of relevant articles including reviews. Finally, 13 studies with data on the relationship between *CDH13* gene promoter methylation and NSCLC were pooled for analysis (Table 1) [[12](#_ENREF_12), [16](#_ENREF_16), [27-37](#_ENREF_27)]. All these articles were written in English. In total, 1206 lung cancer tissues/serum and 644 normal counterpart tissues/serum were collected (Some articles studied with serum and some studies with plasma, to simplify, we use serum instead of serum/plasma). The age of the subjects in the 13 studies ranged from 26 to 87 years, with mean or median ranging from 59 to 70 years. As for the study aim, 5 articles were especially aiming at diagnosis, while the others were for prognosis, survival research, and so on. Among 13 studies, the proportions of stage I samples differed from 9.52 to 68.57%, and the percentage of male individuals in the NSCLC samples has a range of 52 to 80%. For the experimental methods to explore *CDH13* promoter methylation status, 8 of 13 inclusions used methylation-specific polymerase chain reaction (MSP), while others used quantitative MSP (qMSP, such as Methylight, Prosequencing, and so on). Four kinds of methylation detection primers or probes were found to be utilized for most of the 13 studies (Table S1).

### **Meta-analysis**

The ORs for *CDH13* methylation in cancer samples compared with that in normal controls were 6.42 (95% CI: 4.59 to 8.97, z = 10.87, P < 0.00001) in random effects model pooled using DerSimonian and Laird method, and 6.42 (95% CI: 4.59 to 8.97, z = 10.87, P < 0.00001) in fixed effect model using Inverse Variance method, demonstrating a statistically significant increasing in likelihood of methylation in lung cancer tissues comparing to controls. A homogeneity analysis revealed that the variation among them was not significant (I2 = 0%, tau2 = 0) (Figure 1).

Subgroup analyses were conducted for different subtypes, which included sample types (tissue or serum), age, counterpart categories, proportion of early stage, aim of the study (for diagnosis or non-diagnosis), proportion of adenocarcinoma (Ad%) and other possible confounding factors (Table 2). Significant differences were found only between the ORs of the diagnosis (4.16, 95% CI: 2.43 to 7.12) and non-diagnosis (8.46, 95% CI: 5.51 to 13.00) subgroup (P = 0.043) (Figure 2A). Both tissue and serum groups showed significant association between *CDH13* methylation and NSCLC (OR = 6.14 and 6.37, respectively; P = 0.93) (Figure 2B) which suggested that *CDH13* methylation can be taken as a potential biomarker for NSCLC diagnosis using either tissue or serum samples. No significant difference was found between subgroups of MSP and qMSP (OR = 6.46 and 6.29, respectively; P = 0.94), which suggested that the two methods were equivalent in methylation detection (Figure 2C). In addition, there were no significant differences between the subgroups of ratio of male to female, counterpart categories, proportion of adenocarcinoma, the primer set as well as other factors (Table 2).

Because of the significant differences between the diagnosis and non-diagnosis subgroup, we conducted further research in the non-diagnosis subgroup. However, when we focused on the studies not aiming at diagnosis, the OR was found significantly reduced in subgroups of the proportion of adenocarcinoma <65% (7.30, 95% CI: 4.43 to 12.02) than >=65% subgroup (12.85, 95% CI: 5.55 to 29.78) (P = 0.26) (Figure 2D).

### **Summary receiver operating characteristic curve for diagnostic capacity of CDH13 methylation**

Pooled sensitivity and specificity were 0.400 (95%CI: 0.317 to 0.488) and 0.906 (95%CI: 0.840 to 0.946) for all the studies based on the presupposition of the fixed effect model. The sensitivity of the tissue group was higher than that of the serum group, 0.405 (0.316 to 0.500) versus 0.356 (0.247 to 0.484), while the specificity of the tissue group was lower than that of the serum group, 0.900 (0.823 to 0.945) versus 0.937 (0.803 to 0.982), which suggested that reduced sensitivity but increased specificity can be expected in the serums when compared with the cancer tissues.

Although sensitivity and specificity were two of the most important features of a diagnostic test, in some occasions, pooling sensitivity or specificity could be misleading as mentioned in the Methods section. Therefore, we constructed the summery receiver operating characteristic (SROC) curve to depict the stability and accuracy of the methylation test’s diagnostic ability. The area under the curve (AUC) of the SROC was 0.690, suggesting a fair ability for NSCLC diagnosis (Figure 2E).

### **Bias analysis and robust estimation of pooled OR**

A funnel plot of methylation status of lung cancer tissues versus normal tissues showed no significant publication bias (Harbord test, t = 1.11, P = 0.29) and no study exceeded the 95% confidence limits (Figure S1). In order to eliminate the effect of publication bias, trim and fill analysis was performed with the fixed effect model. The adjusted pooled OR were 5.58 (95% CI: 4.05 to 7.67) in the fixed effect model and 5.64 (95% CI: 3.84 to 8.27) in the random effects model. Both results demonstrate a significantly positive association between *CDH13* methylation and NSCLC (Figure S2). In sensitivity analysis to determine the effect of omitting a single study on the overall effect, the overall ORs were between 6.11 (95% CI: 4.33 to 8.61) and 7.20 (95% CI: 5.00 to 10.37) in the fixed effect model, which suggested that combined OR is consistent and reliable (Figure S3).

### **Validation by independent TCGA and GEO lung cancer dataset**

Data from The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) was used to verify the findings from the meta-analysis. In TCGA dataset, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) methylation datasets were obtained for further analysis. There were 6 CpG sites sharing the same CGI (CpG Island) with the primers in the papers collected for meta-analysis, and the overall methylation status of the six CpG sites could be used to represent the methylation status of *CDH13* gene. Surprisingly, the result from LUAD dataset and LUSC dataset differed and differential methylation profiles were shown between these two subtypes (Figure 3). In LUAD methylation dataset, 5 out of the 6 CpG sites showed significantly differential methylation level between cancer tissues and paired adjacent normal tissues according to the criteria (See Method). While in LUSC dataset, though 5 of the 6 CpG sites had p-values less than 0.05 after multiple correction, the absolute mean difference was less than 0.1 for all and thus couldn’t be considered as significantly methylated as well (Table 3).

In order to draw a more robust conclusion, GEO dataset GSE39279 as well as GSE52401 were then downloaded and combined from Gene Expression Omnibus. There were 322 lung adenocarcinomas and 122 lung squamous cell carcinomas and 244 normal tissues in the combined dataset, which is of sufficient sample size to be an independent validation database. We performed the same analysis as before and obtained the result consistent with the TCGA dataset (Figure S4). Due to the large sample size, p-values of all the CpG sites in each dataset were less than 0.05 after multiple corrections. However, In LUAD dataset, 5 out of the 6 CpG sites showed absolute mean difference above 0.1 and thus could be regarded as significantly methylated while none of the CpG sites passed the criteria in LUSC dataset (Table S2). Moreover, another independent GEO dataset GSE56044 with 83 lung adenocarcinoma and 23 lung squamous cell carcinoma tissues and 12 adjacent normal tissues was also downloaded for further validation (Figure S5). Unsurprisingly, the result was almost the same with the previous two datasets. According to our criteria, 6 out of 7 CpG sites were significantly hypermethylated in LUAD dataset while none of the 7 CpG sites was differentially methylated in LUSC dataset (Table S3). Further, we then combined the TCGA and GEO datasets for evaluation of the diagnosis ability of CDH13 methylation status. After quality control, four CpG sites within the CpG island were selected. The AUCs of logistic regression models based on the CpG site was 0.83-0.94 for Ad and 0.60-0.75 for SqC, which showed that the diagnosis ability of CDH13 methylation status is much better in Ads than in SqCs (Table S4).

### **Gene Expression data with TCGA RNA-Seq dataset**

DNA methylation is one of the key regulators for gene expression. It is found that DNA methylation change especially in promoter region would affect gene expression level. As a result, we downloaded the level 3 RNA-Seq dataset from TCGA. Per million mapped reads (RPKM) was used as the measurement for gene expression. After calculating the fold change and p-value with multiple correction, significantly differential expression was shown in LUAD (P = 5.03×10-8, fold change = 0.44) but not shown in LUSC (P = 0.97, fold change = 1.14) samples when compared with normal tissues, which was also concordant with the result from the microarray data analysis (Figure 3).

## Discussion

The *CDH13* gene has been reported hyper-methylated in many types of cancers, such as breast [10] and lung cancers [11-13], in pituitary adenomas [14], diffuse large B cell lymphoma [15], and nasopharyngeal carcinoma [16]. However, the diagnostic role of the methylation status of the *CDH13* gene in lung cancer lacks quantitative assessment. We therefore performed an integrated analysis to quantify the ability for the *CDH13* promoter methylation test in NSCLC diagnosis, and a significant association was identified between *CDH13* methylation and NSCLC (OR = 6.42, 95% CI: 4.59 to 8.97, P < 0.00001). Four imputed studies were filled when trim and fill tests were performed to eliminate the influence of publication bias on the fixed effect model, and the overall OR (5.58, 95% CI: 4.05 to 7.67) was still significant, although it was slightly smaller than that in the crude meta-analysis, indicating the existence of a strong association between *CDH13* promoter methylation and lung cancer.

In order to validate the result from the meta-analysis, we downloaded four independent datasets from TCGA and GEO database. And unexpectedly, the methylation profile in the two subsets of lung cancer differed dramatically. All of the datasets from TCGA and GEO showed significant hypermethylation in the promoter CpG sites of lung adenocarcinoma tissues when compared with normal tissues. However, none of the CpG sites was significantly differential methylated between lung squamous cell carcinoma tissues and normal tissues. Furthermore, we also conducted logistic regression model to evaluate the diagnosis ability of CDH13 methylation status in the lung adenocarcinoma and lung squamous cell tissues as well. Unsurprisingly, the AUCs of the former (AUC: 0.828-0.936) was much better than the latter (AUC: 0.596-0.744). Moreover, the expression data from TCGA level 3 RNA-Seq data were also concordant with this conclusion. The expression level of *CDH13* was significantly lower in lung adenocarcinoma tissues but not in lung squamous cell carcinoma tissues when compared with normal tissues. This result was partially confirmed when we focused on the studies not aiming at diagnosis, the OR was found to be largely reduced in subgroups of the proportion of adenocarcinoma <65% (7.30, 95% CI: 4.43 to 12.02) than >=65% subgroup (12.85, 95% CI: 5.55 to 29.78).

In summary, different results were drawn from meta-analysis and microarray dataset for the association of *CDH13* and lung squamous cell carcinoma. As for the relatively inconsistent result, firstly, qMSP is the semi-quantitative method especially being used in low dose methylation. Secondly, the sparseness of CpG sites in HumanMethylation 450K array (HM450K) may be a key factor. For the microarray is based on the designed probes to detect the methylation of selected CpG sites thus couldn’t cover the whole CpG island region of *CDH13* and therefore might be misleading[[38](#_ENREF_38)]. Besides, no method comparsion between qMSP and HM450K has been conducted to our knowledge. As a result, more comprehensive and advanced methods like WGBS (whole genome bisulfite sequencing) and RRBS (restricted region bisulfite sequencing) are needed to draw a more robust conclusion [[39-41](#_ENREF_39)].

To summarize, according to the previous results drawn from meta-analysis and microarray data analysis, *CDH13* may be a powerful potential biomarker for the diagnosis of lung adenocarcinoma while the association of *CDH13* methylation and lung squamous cell carcinoma needs more data to draw a robust conclusion. In addition, due to the different methylation profile of CDH13 in NSCLC subtypes, CDH13 methylation test could also be a promising biomarker to distinguish the lung adenocarcinoma and lung squamous cell carcinoma which might provide evidence for accurate chemotherapy and targeted therapy.

## Conclusion

This integrated analysis of the pooled data provides strong evidence that the methylation status of the *CDH13* promoter is significantly associated with lung adenocarcinoma. The aberrant *CDH13* methylation could be a promising diagnostic biomarker for non-invasive lung adenocarcinoma detection.

## Methods

### **Search strategy, selection of studies and data extraction**

This pooled study involved searching a range of computerized databases, including PubMed, Cochrane Library, OVID Medline and Web of Science for articles published in English by December 2014. The study used a subject and text word strategy with (CDH13 OR CDHH OR P105 OR H-cadherin OR Cdht OR T-cadherin OR Tcad OR CH211-122A20.1 OR BOS\_16969 OR cdhh) AND (lung or non-small) as the primary search terms. Wildcard character of star, dollar or some other truncations were applied according to the rules of the databases to allow effective article collection.

Two independent reviewers (Geng, Guo) screened the titles and abstracts derived from the literature search to identify relevant studies. The following types of studies were excluded: animal and cell experiments, case reports, reviews or meta-analyses and studies of non-case-control studies or studies with insufficient data or those proving inaccessible after making contact with the authors. The remaining articles were further examined to see if they met the inclusion criteria: 1) the patients had to be diagnosed with NSCLC (Ad and Sc), 2) the studies had to contain *CDH13* gene promoter methylation data from tissue, blood or serum, 3) the studies had to be case-control studies which included tissue-tissue, blood-blood or serum-serum in case and controls respectively. The reference sections of all retrieved articles were searched to identify further relevant articles. Potentially relevant papers were obtained and the full text articles were screened for inclusion by two independent reviewers (Geng, Guo). Disagreements were resolved by discussion with LXT, SDC, and LJ. Quality of original publications was evaluated with the rule of QUADAS and the all the publication have high QUADAS which is greater than 10. Included studies were summarized in data extraction forms. Authors were contacted when relevant data were missing. The name of the first author, year of publication, sample size, age (mean or median), gender proportion (male/female, M2F), the proportion of TNM stage I samples (proportion of early stage of NSCLC samples), the percentage of adenocarcinoma (Ad%), publication aim (for diagnosis or not), analyzing multiple genes or not (one or more genes detected simultaneously in studies design), control type (autogenous or heterogeneous counterpart) and methylation status of the *CDH13* promoter in human NSCLC and normal or control tissues were extracted.

### **Meta-analysis and SROC analysis**

Data were analyzed and visualized mainly using R Software (R version 3.1.0) including meta, metefor and mada packages. The strength of association was expressed as pooled odds ratio (OR) with corresponding 95% confidence intervals (95% CI). Data were extracted from the original studies and recalculated if necessary. Heterogeneity was tested using the I2 statistic with values over 50% and Chi-squared test with P ≤ 0.1 indicating strong heterogeneity between the studies [[18](#_ENREF_18)]. Tau-squared (τ2) was used to determine how much heterogeneity was explained by subgroup differences. The data were pooled using the DerSimonian and Laird random effects model (I2 > 50%, P ≤ 0.1) or fixed effect model (I2 < 50%) according to heterogeneity statistic I2 [[42](#_ENREF_42)]. A two-sided P ≤ 0.05 was considered significant without special annotation. With a lack of heterogeneity among included studies, the pooled odds ratio estimates were calculated using the fixed effect model [[43](#_ENREF_43)]. Otherwise, the random-effects model was used [[44](#_ENREF_44)]. Sensitivity analyses were performed to assess the contributions of single studies to the final results with the abandonment of one article each time. Publication bias was analyzed by funnel plot with mixed-effects version of the Harbord test. If bias was suspected, the conventional meta-trim method was used to re-estimate the effect size.

Compared with traditional SNP association studies, methylation-associated research might be involved with different methylation-definition thresholds. In these cases, traditional weighted averages (pooled sensitivity and specificity) would not reflect the overall accuracy of the test, because the extremes of threshold criteria could skew the distribution, known as the threshold effect [[45](#_ENREF_45)]. Thus, SROC analysis was applied to meta-analysis of diagnostic tests [[45](#_ENREF_45), [46](#_ENREF_46)]. The SROC curve shows the performance of the diagnostic ability of *CDH13* methylation to NSCLC. Each study produces values for sensitivity, specificity and therefore true positive rate (TPR) and false positive rate (FPR), and the plots were placed over the TPR and FPR points to form a smooth curve. A linear regression model was selected to fit the SROC curve where sensitivity and (1-specificity) are transformed into complex logarithmic variables. The exact AUC for the SROC function was used to assess the accuracy of the test.

### **TCGA and GEO data extraction and analysis**

TCGA DNA methylation datasets which included 23 lung adenocarcinoma and 40 lung squamous cell carcinoma tissues as well as 63 paired adjacent tissues, were collected from TCGA project [http://cancergenome.nih.gov/]. And GEO datasets including GSE39279 and GSE52401 and GSE56044 were downloaded from Gene Expression Omnibus [http://www.ncbi.nlm.nih.gov/geo/], including a sum of 568 NSCLC tissues and 256 adjacent or normal lung tissues. Illumina HumanMethylation450K Beadchip was used to detect the methylation level for all of the above datasets. The estimation of methylation for each CG probe was calculated between methylated (M) and unmethylated (U) alleles. Specifically:

beta =

Both M and U represent mean signal intensities for about 30 replicates on the array. Beta value of the CpG sites are used as the measurement of methylation. CpG site would be immediately omitted if there was one missing samples or more in the dataset. CpG sites of CDH13 gene in TCGA dataset and GEO datasets were not completely the same due to the quality control procedure previously mentioned. 6 or 7 CpG sites located in the same CpG island with the primers mentioned in the meta-analysis were the signatures for the methylation status of *CDH13* (Table S1). Wilcoxon rank sum test along with logistic regression were conducted and generated a p-value for each comparison. Multiple comparison of the differential methylation was conducted with Benjamini and Hochberg at 5% FDR as the threshold. The diagnosis model was conducted using logistic regression along with 5 fold cross-validation. The statistical analysis was performed using R version 3.1.0[[47](#_ENREF_47)].

### **RNA-Seq data extraction and analysis**

Level 3 RNA-Seq dataset was obtained from TCGA database, which includes 114 lung adenocarcinoma and 104 lung squamous cell carcinoma as well as 218 normal tissues. Per million mapped reads (RPKM) was regarded as the measurement for gene expression. We assessed the significance of the differential gene expression by comparing the tumor tissues with paired adjacent normal tissues using Wilcoxon signed-rank test. For identification of differentially expression genes, p-value0.05 and fold change 2.0 or was set as the criteria. All the data analysis was conducted with open-source R software (version 3.1.0)

## Competing interests

The authors declare that they have no competing interests.

## Authors’ contributions

JW, SG and XC contributed to the conception, design and final approval of the submitted version. XG, SG, ZL, AW, YT, JW, WP contributed to the meta-analysis and interpretation of data, WP, SG, XG, SC, LT, YX contributed to TCGA and GEO NSCLC data analysis. All authors read and approved the final manuscript.

## Acknowledgements

This research was partially supported by the grants from the 111 Project (B13016). The computations involved in this study were supported by Fudan University High-End Computing Center.

## Reference:

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA: a cancer journal for clinicians. 2015; 65: 5-29.

2. Hankey BF, Ries LA, Edwards BK. 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: 1117-21.

3. van Rens MT, de la Riviere AB, Elbers HR, van Den Bosch JM. Prognostic assessment of 2,361 patients who underwent pulmonary resection for non-small cell lung cancer, stage I, II, and IIIA. Chest. 2000; 117: 374-9.

4. He Y, Cui Y, Wang W, Gu J, Guo S, Ma K, 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: 841-53.

5. Flores K, Wolschin F, Corneveaux JJ, Allen AN, Huentelman MJ, Amdam GV. Genome-wide association between DNA methylation and alternative splicing in an invertebrate. Bmc Genomics. 2012; 13.

6. Gokul G, Khosla S. DNA methylation and cancer. Sub-cellular biochemistry. 2013; 61: 597-625.

7. Hulpiau P, van Roy F. Molecular evolution of the cadherin superfamily. The international journal of biochemistry & cell biology. 2009; 41: 349-69.

8. Shamay M, Krithivas A, Zhang J, Hayward SD. Recruitment of the de novo DNA methyltransferase Dnmt3a by Kaposi's sarcoma-associated herpesvirus LANA. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103: 14554-9.

9. Riener MO, Nikolopoulos E, Herr A, Wild PJ, Hausmann M, Wiech T, et al. Microarray comparative genomic hybridization analysis of tubular breast carcinoma shows recurrent loss of the CDH13 locus on 16q. Human pathology. 2008; 39: 1621-9.

10. Sato M, Mori Y, Sakurada A, Fujimura S, Horii A. The H-cadherin (CDH13) gene is inactivated in human lung cancer. Hum Genet. 1998; 103: 96-101.

11. Sato M, Mori Y, Sakurada A, Fukushige S, Ishikawa Y, Tsuchiya E, et al. Identification of a 910-Kb region of common allelic loss in chromosome bands 16q24.1-q24.2 in human lung cancer. Gene Chromosome Canc. 1998; 22: 1-8.

12. Toyooka KO, Toyooka S, Virmani AK, Sathyanarayana UG, Euhus DM, Gilcrease M, et al. Loss of expression and aberrant methylation of the CDH13 (H-cadherin) gene in breast and lung carcinomas. Cancer research. 2001; 61: 4556-60.

13. Qian ZR, Sano T, Yoshimoto K, Asa SL, Yamada S, Mizusawa N, et al. Tumor-specific downregulation and methylation of the CDH13 (H-cadherin) and CDH1 (E-cadherin) genes correlate with aggressiveness of human pituitary adenomas. Modern Pathol. 2007; 20: 1269-77.

14. Ogama Y, Ouchida M, Yoshino T, Ito S, Takimoto H, Shiote Y, et al. Prevalent hyper-methylation of the CDH13 gene promoter in malignant B cell lymphomas. Int J Oncol. 2004; 25: 685-91.

15. Sun D, Zhang Z, Van DN, Huang GW, Ernberg I, Hu LF. Aberrant methylation of CDH13 gene in nasopharyngeal carcinoma could serve as a potential diagnostic biomarker. Oral Oncol. 2007; 43: 82-7.

16. Kim DS, Kim MJ, Lee JY, Kim YZ, Kim EJ, Park JY. Aberrant methylation of E-cadherin and H-cadherin genes in nonsmall cell lung cancer and its relation to clinicopathologic features. Cancer. 2007; 110: 2785-92.

17. Zhao QT, Guo T, Wang HE, Zhang XP, Zhang H, Wang ZK, et al. Diagnostic value of SHOX2 DNA methylation in lung cancer: a meta-analysis. Oncotargets Ther. 2015; 8: 3433-9.

18. Guo SC, Tan LX, Pu WL, Wu JJ, Xu K, Wu JH, et al. Quantitative assessment of the diagnostic role of APC promoter methylation in non-small cell lung cancer. Clin Epigenetics. 2014; 6.

19. Huang YZ, Wu W, Wu K, Xu XN, Tang WR. Association of RASSF1A Promoter Methylation with Lung Cancer Risk: a Meta-analysis. Asian Pac J Cancer P. 2014; 15: 10325-8.

20. Weilin Pu, X.G., Shicheng Guo, Yulong Tan, An Wang, Zhouyi Lu, Lixing Tan, Sidi Chen, Jiucun Wang, Xiaofeng Chen, Aberrant methylation of FHIT can be a diagnostic biomarker for NSCLC in Asian population. Submitted, 2016

21. Fang, N., et al., [A meta-analysis of Association between MGMT gene promoter methylation and non-small cell lung cancer]. Zhongguo Fei Ai Za Zhi, 2014. 17(8): p. 601-5.

22. Liang YL, He LP, Yuan H, Jin YL, Yao YS. Association between RUNX3 promoter methylation and non-small cell lung cancer: a meta-analysis. J Thorac Dis. 2014; 6: 694-705.

23. Hua F, Fang NZ, Li XB, Zhu SW, Zhang WS, Gu JD. A Meta-Analysis of the Relationship Between RAR beta Gene Promoter Methylation and Non-Small Cell Lung Cancer. PloS one. 2014; 9.

24. Zeng, Y., R. Liu, and H. Zhang, [Meta-analysis of association between E-cadherin promoter methylation and lung cancer risk]. Zhongguo Fei Ai Za Zhi, 2013. 16(7): p. 353-8.

25. Gu JD, Wen YJ, Zhu SW, Hua F, Zhao H, Xu HR, et al. Association between P-16INK4a Promoter Methylation and Non-Small Cell Lung Cancer: A Meta-Analysis. PloS one. 2013; 8.

26. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002; 30: 207-10.

27. Toyooka S, Maruyama R, Toyooka KO, McLerran D, Feng Z, Fukuyama Y, et al. Smoke exposure, histologic type and geography-related differences in the methylation profiles of non-small cell lung cancer. International journal of cancer Journal international du cancer. 2003; 103: 153-60.

28. Feng Q, Hawes SE, Stern JE, Wiens L, Lu H, Dong ZM, et al. DNA methylation in tumor and matched normal tissues from non-small cell lung cancer patients. Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology. 2008; 17: 645-54.

29. Hsu HS, Chen TP, Hung CH, Wen CK, Lin RK, Lee HC, et al. Characterization of a multiple epigenetic marker panel for lung cancer detection and risk assessment in plasma. Cancer. 2007; 110: 2019-26.

30. Jin M, Kawakami K, Fukui Y, Tsukioka S, Oda M, Watanabe G, et al. Different histological types of non-small cell lung cancer have distinct folate and DNA methylation levels. Cancer science. 2009; 100: 2325-30.

31. Tsou JA, Galler JS, Siegmund KD, Laird PW, Turla S, Cozen W, et al. Identification of a panel of sensitive and specific DNA methylation markers for lung adenocarcinoma. Molecular cancer. 2007; 6: 70.

32. Ulivi P, Zoli W, Calistri D, Fabbri F, Tesei A, Rosetti M, et al. p16INK4A and CDH13 hypermethylation in tumor and serum of non-small cell lung cancer patients. Journal of cellular physiology. 2006; 206: 611-5.

33. Wang Y, Zhang D, Zheng W, Luo J, Bai Y, Lu Z. Multiple gene methylation of nonsmall cell lung cancers evaluated with 3-dimensional microarray. Cancer. 2008; 112: 1325-36.

34. Zhai X, Li S-J. Methylation of RASSF1A and CDH13 Genes in Individualized Chemotherapy for Patients with Non-small Cell Lung Cancer. Asian Pacific Journal of Cancer Prevention. 2014; 15: 4925-8.

35. Zhang Y, Wang R, Song H, Huang G, Yi J, Zheng Y, et al. Methylation of multiple genes as a candidate biomarker in non-small cell lung cancer. Cancer letters. 2011; 303: 21-8.

36. 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 research. 2012; 72: 5692-701.

37. Hanabata T, Tsukuda K, Toyooka S, Yano M, Aoe M, Nagahiro I, et al. DNA methylation of multiple genes and clinicopathological relationship of non-small cell lung cancers. Oncology reports. 2004; 12: 177-80.

38. Sun ZF, Cunningham J, Slager S, Kocher JP. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. Epigenomics-Uk. 2015; 7: 813-28.

39. Bibikova M, Fan JB. Genome-wide DNA methylation profiling. Wires Syst Biol Med. 2010; 2: 210-23.

40. Bock C, Tomazou EM, Brinkman AB, Muller F, Simmer F, Gu HC, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010; 28: 1106-U196.

41. Li N, Ye MZ, Li YR, Yan ZX, Butcher LM, Sun JH, et al. Whole genome DNA methylation analysis based on high throughput sequencing technology. Methods. 2010; 52: 203-12.

42. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials. 1986; 7: 177-88.

43. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute. 1959; 22: 719-48.

44. Huizenga HM, Visser I, Dolan CV. Testing overall and moderator effects in random effects meta-regression. The British journal of mathematical and statistical psychology. 2011; 64: 1-19.

45. Midgette AS, Stukel TA, Littenberg B. A meta-analytic method for summarizing diagnostic test performances: receiver-operating-characteristic-summary point estimates. Medical decision making : an international journal of the Society for Medical Decision Making. 1993; 13: 253-7.

46. Jones CM, Athanasiou T. Summary receiver operating characteristic curve analysis techniques in the evaluation of diagnostic tests. The Annals of thoracic surgery. 2005; 79: 16-20.

47. Team R C. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012[J]. 2014.

**Figure 1. Flow chart of the literature collection procedure.**

**Figure 2. Combined estimates for the association between FHIT promoter hyper-methylation and non-small cell lung cancer (NSCLC) with forest plot.**

Author, year, country of the studies and methylated (M) and total number of the sample (T) in case and control, combined odds ratio (OR) with 95% confidence region were labeled in the left column of the figure. The DerSimonian-Laird estimator and Mantel-Haenszel method were selected to conduct combination estimation for the random effects model and fixed effect model, respectively.

**Figure 3. Subgroup meta-analysis and SROC Curve for the relationship between FHIT promoter hypermethylation and non-small cell lung cancer (NSCLC).**

(A-D) Subgroup meta-analysis based on aim, sample type, method, AD2SC in non-diagnosis, respectively. (E) Diagnostic SROC (bivariate model) for FIHT in NSCLC. The dotted circle region represents the estimated 95% confidential regions for sensitivity and specificity based on whole individual studies.

**Figure 4 CpG sites on the HM450K Beadchip across FHIT gene region and Gene expression scatterplot with paired data from TCGA dataset**

Methylation and gene expression status for FHIT gene (TCGA lung cancer dataset). A-B each represents the different methylation status of lung cancer subtypes versus normal lung tissues in different datasets. For A-B, the x-axis shows the different CpG sites in FHIT genes and the y-axis shows the beta value of each CpG site to represent the methylation level of each CpG site. C-D each represents the gene expression status of paired samples. The x-ais of the two figures shows the different types and y-axis shows the gene expression level using RPKM as measurement.

**Table 1. Characteristics of eligible studies considered in the report**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Author** | **Sample**  **Type** | **Age**a | **Stage**  **I%** | **Stage**  **(I+II) %** | **Male%** | **Patients**  **(M/T)** | **Control**  **(M/T)** | **Methods** | **Aim** | **Multiple**  **target** | **Control design** | **Ad%** | **Primer**  **set** |
| Dong et al | tissue | 63 | 0.580 | 0.761 | 0.795 | 26/88 | 7/88 | MSP | Non-Diagnosis | Multi | hom | 0.398 | 1 |
| Feng et al | tissue | 64.3 | 0.429 | 0.776 | 0.531 | 11/49 | 1/49 | qMSP | Non-Diagnosis | Multi | hom | 0.588 | 1 |
| Hanabata et al | tissue | NA | 0.686 | 0.843 | 0.629 | 26/70 | 2/30 | MSP | Non-Diagnosis | Multi | hom | 0.688 | 1 |
| Hsu et alb | tissue | 66.81 | NA | 0.651 | NA | 28/63 | 10/63 | MSP | Diagnosis | Multi | hom | 0.759 | 1 |
| Jin et al | tissue | 66.7 | NA | NA | 0.708 | 25/72 | 2/63 | qMSP | Non-Diagnosis | Multi | hom | 0.652 | 3 |
| Nikolaidis et al | tissue | 65.58 | NA | 0.771 | 0.521 | 11/48 | 0/48 | qMSP | Diagnosis | Multi | hom | 0.500 | 4 |
| Toyooka et al | tissue | NA | NA | NA | NA | 18/42 | 2/25 | MSP | Non-Diagnosis | Single | hom | 0.738 | 1 |
| Toyooka et al | tissue | 63 | NA | NA | 0.691 | 180/514 | 5/84 | MSP | Non-Diagnosis | Multi | hom | 0.606 | 1 |
| Tsou et al | tissue | NA | NA | NA | NA | 39/51 | 24/49 | qMSP | Diagnosis | Multi | both | 1 | 2 |
| Ulivi et al | serum | 70 | 0.557 | 0.721 | 0.803 | 14/61 | 0/15 | qMSP | Diagnosis | Multi | heter | 0.692 | 1 |
| Wang et al | tissue | NA | NA | NA | 0.607 | 15/28 | 3/12 | MSP | Diagnosis | Multi | hom | 0.682 | NA |
| Zhai et al | serum | 62.39 | 0.095 | 0.143 | 0.762 | 23/42 | 0/40 | MSP | Non-Diagnosis | Multi | heter | 0.762 | 1 |
| Zhang et alb | tissue | 59 | 0.321 | 0.744 | 0.744 | 38/78 | 8/78 | MSP | Non-Diagnosis | Multi | hom | 0.455 | 1 |

amean or median age from articles. bwith two records since there are Tissue and serum data simultaneously in this article. M and T means methylation positive and total, respectively.

**Table 2 Subgroup analysis for the potential confounding factors with fixed effect model**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Subgroup | Number of study | OR | lower | upper | I2 | P-value |
| Overall | 13 | 6.42 | 4.59 | 8.97 | 0.00% |  |
| Age<64.3 | 4 | 7.63 | 4.60 | 12.65 | 27.5% |  |
| Age>=64.3 | 5 | 7.11 | 3.70- | 13.66 | 0.70% | 0.87 |
| Stage I<49.3% | 3 | 10.62 | 4.95 | 22.77 | 26.0% |  |
| Stage I>=49.3% | 3 | 5.78 | 2.74 | 12.18 | 0.00% | 0.26 |
| Stage I+II<75.2% | 4 | 6.73 | 3.79 | 11.96 | 37.9% |  |
| Stage I+II>=75.2% | 4 | 6.83 | 3.39 | 13.79 | 0.00% | 0.97 |
| Male %<70% | 5 | 8.01 | 4.21 | 15.27 | 0.00% |  |
| Male %>=70% | 5 | 7.89 | 4.72 | 14.20 | 20.4% | 0.96 |
| MSP | 8 | 6.46 | 4.38 | 9.53 | 0.00% |  |
| qMSP | 5 | 6.29 | 3.25 | 12.19 | 24.9% | 0.95 |
| Diagnose | 5 | 4.16 | 2.43 | 7.12 | 0.00% |  |
| Non-diagnose | 8 | 8.46 | 5.51 | 13.00 | 0.00% | **0.04** |
| Heterogeneous | 2 | 30.69 | 4.04 | 232.7 | 21.5% |  |
| Autogenous | 10 | 6.86 | 4.74 | 9.94 | 0.00% | 0.15 |
| Plasma/Serum | 4 | 6.36 | 2.84 | 14.29 | 54.0% |  |
| Tissue | 11 | 6.14 | 4.37 | 8.62 | 0.00% | 0.93 |
| Ad%<68.2% | 6 | 8.20 | 5.14 | 13.07 | 0.00% |  |
| Ad%>=68.2% | 6 | 6.06 | 3.27 | 11.26 | 6.00% | 0.45 |
| Asian | 7 | 6.43 | 4.19 | 9.85 | 19.0% |  |
| White | 6 | 6.40 | 3.73 | 11.00 | 0.00% | 0.99 |

**Table 3. Differential CDH13 methylation, odds ratio between adenocarcinoma, squamous cell carcinoma and their counterparts from TCGA dataset**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | CpG site | Position | McaM | McoM | ∆β | P-value |
| LUAD | cg08747377 | chr16:82660670 | 0.32 | 0.10 | 0.22 | 1.95E-05 |
| cg05374412 | chr16:82660727 | 0.21 | 0.03 | 0.18 | 2.27E-05 |
| cg00806490 | chr16:82660873 | 0.28 | 0.13 | 0.15 | 1.05E-04 |
| cg08856946 | chr16:82661421 | 0.19 | 0.04 | 0.15 | 2.27E-05 |
| cg09189772 | chr16:82661638 | 0.29 | 0.16 | 0.13 | 8.32E-04 |
| cg19369556 | chr16:82661725 | 0.26 | 0.22 | 0.04 | 0.292 |
| LUSC | cg08747377 | chr16:82660670 | 0.13 | 0.08 | 0.05 | 0.015 |
| cg05374412 | chr16:82660727 | 0.07 | 0.02 | 0.05 | 0.009 |
| cg00806490 | chr16:82660873 | 0.15 | 0.11 | 0.04 | 0.006 |
| cg08856946 | chr16:82661421 | 0.05 | 0.03 | 0.02 | 0.030 |
| cg09189772 | chr16:82661638 | 0.16 | 0.14 | 0.02 | 0.761 |
| cg19369556 | chr16:82661725 | 0.14 | 0.18 | 0.04 | 2.26E-05 |

McaM and McoM represent the mean of case methylation (Beta) and mean of control methylation (Beta). Methylation levels are calculated with formula: Beta = (M/M + U).

Position represents the chromosome position of each CpG site according to GRCh37/hg19.

P-valuesare calculated from Wilcoxon signed-rank test after false discovery rate (FDR adjustment).