**Identification of genetically predicted DNA methylation markers associated with non-****small cell lung cancer risk with** **34,964 cases and 448,579 controls**

**Authors：**Xiaoyu Zhao 1,†, Meiqi Yang 1, †, Jingyi Fan 1,2,3,†, Mei Wang 1, Yifan Wang 1, Na Qin 1,2,Meng Zhu 1,2,4, Yue Jiang 1,2, Olga Y Gorlova 5,6, Ivan P Gorlov 5,6, Demetrius Albanes 7, Stephen Lam 8, Adonina Tardón 9, Chu Chen 10, Gary E Goodman 11, Stig E Bojesen 12,13, Maria Teresa Landi 14, Mattias Johansson 15, Angela Risch 16,17,18, H-Erich Wichmann 19, Heike Bickeböller 20, David C Christiani 21, Gad Rennert 22, Susanne M Arnold 23, Paul Brennan 15, John K Field 24, Sanjay Shete 25, Loïc Le Marchand 26, Olle Melander 27,28, Hans Brunnström 27, Geoffrey Liu 29, Rayjean J Hung 30, Angeline S Andrew 31, Lambertus A Kiemeney 32, Shanbeh Zienolddiny 33, Kjell Grankvist 34, Mikael Johansson 35, Neil E Caporaso 36, Penella J Woll 37, Philip Lazarus 38, Matthew B Schabath 39, Melinda C Aldrich 40, Alpa V Patel 41, Hongxia Ma 1,2, Guangfu Jin 1,2, Zhibin Hu 1,2, Christopher I Amos 42, Hongbing Shen 1,2,\*, Juncheng Dai 1,2,\*

**Affiliations：**

1 Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China.

2 Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine and China International Cooperation Center for Environment and Human Health, Gusu School, Nanjing Medical University, Nanjing 211166, China.

3 Health Management Center, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Gusu School, Nanjing Medical University, Suzhou 215008, China.

4 Department of Thoracic Surgery, Jiangsu Key Laboratory of Molecular and Translational Cancer Research, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research, The Affiliated Cancer Hospital of Nanjing Medical University, Nanjing 210009, China.

5 Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire.

6 Department of Medicine, Epidemiology Section, Institute for Clinical and Translational Research, Baylor Medical College, Houston, Texas.

7 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland.

8 Department of Integrative Oncology, British Columbia Cancer Agency, Vancouver, British Columbia, Canada.

9 Department of Public Health IUOPA, University of Oviedo, ISPA and CIBERESP, Oviedo, Spain.

10 Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington.

11 Public Health Sciences Division, Swedish Cancer Institute, Seattle, Washington.

12 Department of Clinical Biochemistry, Copenhagen University Hospital, Copenhagen, Denmark.

13 Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.

14 National Cancer Institute, Bethesda, Maryland.

15 Genetic Epidemiology Group, International Agency for Research on Cancer, Lyon, France.

16 University of Salzburg, Department of Biosciences, Allergy-Cancer-BioNano Research Centre, Salzburg, Austria.

17 Division of Epigenomics and Cancer Risk Factors, DKFZ-German Cancer Research Center, Heidelberg, Heidelberg, Germany.

18 Translational Lung Research Center Heidelberg (TLRC-H), German Center for Lung Research (DZL), Heidelberg, Germany.

Institute of Epidemiology, Helmholtz Center Munich, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany.

19 Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany.

20 Department of Genetic Epidemiology, University Medical Center Goettingen, Goettingen, Germany.

21 Departments of Environmental Health and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts.

22 Technion Faculty of Medicine, Carmel Medical Center, Haifa, Israel.

23 Markey Cancer Center, University of Kentucky, Lexington, Kentucky.

24 Molecular and Clinical Cancer Medicine, Roy Castle Lung Cancer Research Programme, The University of Liverpool Institute of Translational Medicine, Liverpool, United Kingdom.

25 Department of Epidemiology, The University of Texas, MD Anderson Cancer Center, Houston, Texas.

26 Epidemiology Program, University of Hawai'i Cancer Center, Honolulu, Hawai'i.

27 Clinical Sciences, Lund University, Lund, Sweden.

28 Department of Internal Medicine, Skåne University Hospital, Malmö, Sweden.

29 Princess Margaret Cancer Centre, Toronto, Ontario, Canada.

30 Prosseman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada.

31 Department of Neurology, Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire.

32 Department for Health Evidence, Radboud University Medical Center, Nijmegen, the Netherlands.

33 National Institute of Occupational Health (STAMI), Oslo, Norway.

34 Department of Medical Biosciences, Umeå University, Umea, Sweden.

35 Department of Radiation Sciences, Umeå University, Umea, Sweden.

36 Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland.

37 Academic Unit of Clinical Oncology, University of Sheffield, Sheffield, United Kingdom.

38 Department of Pharmaceutical Sciences, College of Pharmacy and Pharmaceutical Sciences, Washington State University, Spokane, Washington.

39 Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida.

40 Department of Medicine (Division of Genetic Medicine), Vanderbilt University Medical Center, Nashville, Tennessee.

41 Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA, USA.

42 Baylor College of Medicine, Institute for Clinical and Translational Research, Houston, TX, USA.

† These authors contributed equally to this work.

**Corresponding to:** Juncheng Dai, Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Rd., Nanjing 211166, China. E-mail: djc@njmu.edu.cn; Hongbing Shen, Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Rd., Nanjing 211166, China. Tel.: +86-25-8686-8439, Fax: +86-25-8686-8437, Email: hbshen@njmu.edu.cn.

**Word count:** 261 for abstract; 4,258 for text; **Tables:** 2; **Figure:** 3; **Supplementary figures:** 1; **Supplementary tables:** 17

**Abstract**

**Background:** Although several genetic studies have explored the associations between genetic variations and lung cancer risk, the epigenetic consequences of DNA methylation in lung cancer development are largely unknown. Here, we investigated the genetically predicted DNA methylation markers associated with non-small cell lung cancer (NSCLC) risk based on a large-scale two-stage case-control design.

**Methods:** We established the genetic prediction models for methylation levels based on genetic and methylation data of 1,595 subjects from the Framingham Heart Study. Eligible models were applied to a fixed-effect meta-analysis of screening datasets with 27,120 NSCLC cases and 27,355 controls to identified the related methylation markers, which were then replicated in independent datasets with 7,844 lung cancer cases and 421,224 controls. We also performed a multi-omics functional annotation for the identified CpGs by integrating genomics, epigenomics and transcriptomics and further investigated the potential regulation pathways.

**Results:** We initially identified 39 CpGs at ten loci associated with NSCLC risk (Bonferroni-corrected *P* ≤ 1.67 × 10−6). Among them, 16 CpGs remaining significant in the validation stage (Bonferroni-corrected *P* ≤ 1.28 × 10−3), including 4 novel CpGs (cg22795331, cg05012158, cg06752398, and cg19720302). Multi-omics functional annotation showed 9 of 16 CpGs were potentially functional biomarkers for NSCLC risk. We also observed 35 genes within a 1-Mb window of 12 CpGs, including *TERT*, *TRIP13*, *JAML*, *CHRNA5*, *PSMA4*, and *KPNA2* *et.al*,might involve in regulatory pathways of NSCLC risk.

**Conclusion:** We identified 16 DNA methylation markers associated with NSCLC risk. Changes of the methylation level at these CpGs may influence the development of NSCLC by regulating the expression of genes nearby.

**Keywords:** DNA methylation; genetic prediction; gene expression; non-small cell lung cancer risk; association study**Introduction**

Lung cancer is the second most commonly diagnosed cancer and the top cause of cancer death worldwide (1). It is estimated that nearly 2.21 million new lung cancer cases and 1.80 million new lung cancer deaths occurred in 2020, accounting for 11.4% and 18.0% of total cancer respectively (1). In China, lung cancer is the leading type of cancer with the highest morbidity and mortality (2). Besides, non-small cell lung cancer (NSCLC) accounts for about 85% of total lung cancer cases and mainly includes adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) two subtypes (3). The development of lung cancer involves the interplay between environmental and genetic risk factors (4-6). Over the past decade, several important genetic loci were identified for lung cancer risk by genome-wide association studies (GWASs) (7), while contrast to the traditional genetic mechanisms, epigenetics has also been found to play a critical role in lung cancer pathogenesis, including DNA methylation (one kind of the key epigenetic alterations) (8).

Based on candidate strategy, early studies have identified some methylation markers potentially associated with lung cancer, such as hypermethylation at promoters of *RASSF1*, *CDKN2A*, *MGMT*, *APC*, and *DAPK et.al* (8). Recent emerging epigenome-wide association studies (EWAS) also revealed several new methylation markers (e.g. cg05575921-*AHRR* and cg03636183-*F2RL3*), however, more new findings were hindered by the limited sample size (9-11). Furthermore, limited by selection bias, potential confounding, and reverse causation, the causal role of DNA methylation may be inconsistent with results from observational studies (12).

DNA methylation is impacted by both environmental factors and genetic factors (13, 14). Previous studies have identified multiple DNA methylation quantitative trait loci (meQTL) (15, 16), suggesting DNA methylation at some CpGs could be predicted by genetic variants. This strategy is based on the random assortment of alleles during gamete formation, and thus could avoid the effects of biases and reverse causation commonly encountered in conventional epidemiological studies to some degree. Yang *et.al* developed new statistical models to predict DNA methylation in white blood cells via multiple genetic variants in a reference dataset, and apply them to the summary data of large GWASs to investigate the association between genetically predicted DNA methylation and disease risk (17-20).

To date, GWASs have identified 45 risk loci for lung cancer in different ethnic populations (7), which are replicated in large-scale and multi-center studies in European (21) and Chinese populations (22). Herein, we will adopt these prediction models to identify new lung-cancer-associated methylation markers by using the data from 27,120 NSCLC cases and 27,355 healthy controls and further replicated in an independent dataset with 7,844 lung cancer cases and 421,224 controls. The findings of this study provide important insight into the epigenetic susceptibility mechanisms of non-small cell lung cancer.

**Material and Methods**

**Study design and participants**

The overall design was exhibited in **Figure. 1**. First, we trained the DNA methylation prediction models by using data from 1,595 Framingham Heart Study (FHS) participants and then performed an external validation in 883 subjects of Women’s Health Initiative (WHI). After that, we selected the methylation prediction models with qualified performance to assess the association between genetically predicted methylation markers and NSCLC risk, based on a large-scale GWAS including 27,120 NSCLC cases and 27,355 controls (13,327 cases and 13,328 controls of Chinese descent as well as 13,793 cases and 14,027 controls of European descent) (22).

For those methylation markers identified from the screening stage, we conducted a validation in a total of 7,844 lung cancer cases and 421,224 controls from UK Biobank (UKB) and Female Lung Cancer Consortium in Asia (FLCCA) datasets. Therein, GWAS of malignant neoplasm of bronchus and lung (ICD-10 C34) containing 3,048 lung cases and 417,483 controls of European descent was from Pan-ancestry genetic analysis of the UKB, an open database of over 16,000 GWAS conducted across a very broad range of phenotypes in UK Biobank (<https://pan.ukbb.broadinstitute.org>). Another dataset was performed on 8,537 never-smoking Asia women (4,796 lung cases and 3,741 controls) from FLCCA (23). Basic information and clinical features of participants for these datasets were shown in **Supplementary Table 1**.

Then, we conducted a multi-omics functional annotation for the identified CpG sites by integrative analyses of epigenomics, genomics and transcriptomics and additionally investigated the potential regulatory pathways. The genomics data included cis-meQTL identified from 4,170 European individuals in the FHS cohort (24). The epigenomics and transcriptomics data were from The Cancer Genome Atlas (TCGA), including DNA methylation data of 907 NSCLC cases and gene expression data of 1,119 NSCLC cases (849 cases with matched two omics data). In addition, gene expression data of 108 lung tumor-adjacent tissue pairs from TCGA were included as well.

**DNA methylation prediction models training and refining**

Here, 1,595 unrelated European subjects with matched genetic and DNA methylation data in Framingham Heart Study (FHS) were used to construct DNA methylation prediction models (accession numbers in dbGaP: phs000342 and phs000724). The detailed information about datasets and data process has been described elsewhere (17-20) and shown in **Supplementary Methods** (available online). For each CpG site, we used genetic variants flanking a 2-Mb window to build a statistical model by the elastic net method (α = 0.5) in the “glmnet” package of R (25) to predict DNA methylation residuals. An internal validation for each model was performed using tenfold cross-validation. The *RFHS2* values, the square of correlation coefficient between measured and predicted methylation levels, were calculated to estimate the prediction performance of models.

Using the data from 883 genetically unrelated female participants of European descent derived from the Women's Health Initiative (WHI) (available from dbGaP: phs001335, phs000675, and phs000315), we performed an external validation for the built methylation predictive models. The pipeline of data process was the same as that for the FHS data. The *RWHI2* values were calculated by Spearman’s correlation test. Furthermore, we selected the models with satisfactory prediction performance according to the criteria below: 1) with a *RFHS2* ≥ 0.01 (≥ 10% correlation between predicted and measured methylation levels) in FHS; 2) with a *RWHI2* ≥ 0.01 in WHI; 3) probes with no SNPs overlapped, considering that SNPs on the probes might have a potential impact on the methylation level estimation (26, 27).

**Association analyses between predicted methylation and NSCLC risk**

At the screening stage, we used S-PrediXcan (28) to evaluate the associations between genetically predicted methylation levels and NSCLC risk. In brief, the association indicator Z-score was estimated by the formula below:

$$Z\_{m}≈\sum\_{ s\in Model\_{m}}^{}w\_{sm} \frac{\hat{σ}\_{s}}{\hat{σ}\_{m}} \frac{\hat{β}\_{s}}{se\left(\hat{β}\_{s}\right)}$$

In the formula, $w\_{sm}$ is the weight of SNP s in the prediction of the CpG m. $\hat{σ}\_{s}$ and $\hat{σ}\_{m}$ are the estimated variances of SNP s and CpG m. $\hat{β}\_{s}$ and $se\left(\hat{β}\_{s}\right)$ are the GWAS regression coefficients and standard error of $\hat{β}\_{s}$. Respectively, we used summary data from two GWASs that had been generated from 27,820 European individuals and 26,655 Chinese individuals (22) to estimate the associations between genetically predicted methylation levels with NSCLC risk. To consider the population heterogeneity and identify the shared methylation markers, we conducted a fixed-effect meta-analysis of two populations using the META v1.7 (29), while *P* ≤ 0.05 for Cochran’s Q statistic was set to indicate a high degree of heterogeneity. We further filtered out these CpGs with heterogeneity or inconsistent directions of effect size in two populations. Finally, we used a Bonferroni-corrected test to screen the statistically significant CpG sites of the total sample (*P* ≤ 1.67 × 10–6, 0.05/29,894).

At the validation stage, we replicated the CpGs identified in the screening stage using 7,844 lung cancer cases and 421,224 controls derived from two independent datasets of Pan-UKB and FLCCA. The same strategy of meta-analysis was performed for the genetically predicted associations of the two validation populations. After that, a Bonferroni-corrected test was used to determine the CpG sites passed the validation (*P* ≤ 1.28 × 10–3, 0.05/39).

For replicated CpG sites, we assessed whether the observed associations were independent of lung cancer susceptibility variants identified in previous GWASs (7, 21, 22). Briefly, we used genome-wide complex trait analysis-conditional and joint analysis (GCTA-COJO) (30) to reevaluate the betas and standard errors of lung cancer by adjusting the closest GWAS-identified risk variants, and then re-run the S-PrediXcan analyses. Additionally, we conducted the subgroup analyses by histological type (squamous cell carcinoma and adenocarcinoma), smoking status (smoker and nonsmoker) and gender to explore the difference between subgroups. Heterogeneity across subgroups was estimated by Cochran’s Q test and *P* ≤ 0.05 was considered as statistical significance.

**Systematically** **multi-****omics** **functional annotation of** **epigenomics, genomics and transcriptomics for NSCLC-related CpG sites**

We performed multi-omics functional annotations based on epigenomics, genomics and transcriptomics data for the CpG sites that passed the validation stage. The types and sources of related annotation information were described in **Supplementary Table 2**.

For the epigenomics level, using ANNOVAR (31), the CpGs were annotated by the corresponding closest gene and the genomic region as well as 8 functional categories including exonic, intronic, intergenic, upstream, 5’-UTR, ncRNA intronic, ncRNA exonic and downstream. Moreover, we used an extended annotation of the Illumina 450K platform available at GEO (GPL18809) (27) for a supplement. The chromatin interactions, topologically associated domains (TAD), transcription factor binding sites (TFBS), and histone mark (H3K27ac, H3k4me1, H3K4me3, and H3K9ac) were further annotated.

For the genomics level, we assessed the cis-meQTL of the corresponding CpGs whether were overlapped with expression quantitative trait loci (eQTL) in the Genotype-Tissue Expression (GTEx) v8 database. Six bioinformatics predictive algorithms, including Combined Annotation Dependent Depletion (CADD), Functional Analysis through Hidden Markov Models (FATHMM), LRT, MutationTaster, PolyPhen-2, and Sorting Tolerant from Intolerant (SIFT), were used for the evaluation that these cis-meQTL whether were detrimental missense variants (32).

For transcriptomics level, we identified the methylation-related protein-encoding genes within a 1-Mb range (±500 kb) of each CpG site by estimating Spearman correlation coefficients, and a FDR (false discovery rate) corrected threshold of *P* ≤ 0.05 was used. For these methylation-related genes, we determined whether they were driver genes of lung cancer (32) or lung-cancer-associated genes (33). Finally, results from two transcriptome-wide association studies (TWASs) of lung cancer (34, 35) were also used as annotation resources for the association between methylation-related genes and NSCLC risk.

To estimate the functional importance of these identified CpGs with NSCLC risk, a functional score system was constructed. One score was given if CpG was met the corresponding criterion of each indicator in three of the omics (**Supplementary Table 3, Supplementary Methods**, available online). Altogether, functional score in the epigenomics level ranged from 0 to 10 (omics score: one score given if score ≥ 5), in the genomics level from 0 to 3 (omics score: one score given if score ≥ 2), and in the transcriptomics level from 0 to 7 (omics score: one score given if score ≥ 4). We classified the CpGs into three levels based on the omics scores: level A (3 scores), level B (2 scores), and level C (0 or 1 score), indicating the functional importance from high to low.

**Integrative analysis of multi-omics data to investigate the potential** **regulatory pathways**

Based on gene expression data of 108 lung cancer tissues-adjacent tissues pairs from TCGA, we conducted the differential expression analyses for those methylation-related genes flanking each valid CpGs. The tumor/adjacent tissues ratio of each pair (relative genes expression level) was calculated by log2-transformed gene expression of tumor and adjacent tissues, and the ratio greater than 0 was considered as up-regulation. Once over half of lung cancer tissues-adjacent tissues pairs showed the up-regulation, the corresponding gene was considered positively relate to NSCLC. A Wilcoxon rank sum test was used and FDR-corrected threshold of *P* ≤ 0.05 was statistically significant. Finally, we further integrated the association between genetically predicted methylation and NSCLC risk, the correlation between DNA methylation and gene expression in lung cancer tissues from TCGA, and differential genes between lung cancer tissues and adjacent normal tissues to elucidate the putative pathways through which DNA methylation affects the development of NSCLC.

**Results**

**DNA methylation prediction models**

Based on individual-level genotyping and DNA methylation data from the FHS cohort, DNA methylation prediction models for a total of 223,959 CpG were established, of which 81,352 models with a predictive performance (*RFHS2*) of at least 0.01, suggesting at least a 10% correlation between predicted and measured methylation levels, were retained. To further test the reliability of these models, the genetic data from WHI were applied to conduct an independent validation, and then 70,330 models (86.45%) showed the predictive performance no less than 0.01 (*RWHI2*≥ 0.01). We observed the points representing a predictive performance of 70,330 CpGs fluctuated around the line Y = X, suggesting a high correlation between *RFHS2* and *RWHI2* (Pearson’s correlation r = 0.95, *P* ≤ 0.0001, **Supplementary Figure 1**). After that, methylation probes of 7,284 had SNPs within the binding site, which required to be excluded. Totally, there were 63,046 CpGs with 751,458 SNPs remaining for the downstream analyses.

**Associations of genetically predicted methylation with NSCLC risk**

At the screening stage, we did a meta-analysis for predicted associations of 62,981 CpGs available in 27,120 NSCLC cases and 27,355 controls. After removing the CpGs with heterogeneity *P* ≤ 0.05 (n=7,626) and those without consistent directions of effect size (n=25,371) in the fixed-effects meta-analysis, a total of 29,894 CpGs remained. We observed 39 CpGs located in 10 loci were significantly associated with NSCLC risk after Bonferroni correction (*P* ≤ 1.67 × 10−6, 0.05/29,894) in the screening stage (**Figure 2; Supplementary Table 4**).

At the validation stage, we replicated the 39 CpGs using independent datasets with 7,844 lung cancer cases and 421,224 controls. As shown in **Supplementary Table 5**, 25 CpGs with the same effect direction were at *P* less than 0.05, 16 of which met the Bonferroni correction (*P* ≤ 1.28 × 10−3, 0.05/39). Four of the replicated 16 CpGs (cg22795331, cg05012158, cg06752398, and cg19720302) were firstly reported methylated loci associated with NSCLC risk, 12 ones were located in susceptibility regions reported previously (**Figure 2 and Table 1**). Among the 16 CpGs, a positive association of 3 CpGs with NSCLC risk were detected (5p15.33: cg07493874 and cg27028750; 15q25.1: cg06752398), while a higher predicted methylation level was associated with a decreased risk of NSCLC at the other 13 CpGs (**Table 1**). We did not observe any of 16 valid CpGs remaining significant (*P* ≤ 1.67 × 10−6) after adjusting GWAS-identified lung cancer susceptibility variants (**Supplementary Table 6**). Additionally, the respective results of methylation markers derived from two populations were shown in **Supplementary Table 7 and 8**.

Moreover, subgroup analyses by histological type, smoking status, and gender were performed to estimate the potential heterogeneity across subgroups (**Supplementary Table 9**). We found 3 of 16 valid CpGs (cg07507801, cg22795331, and cg18468235) showed stronger associations in lung adenocarcinoma (*P*-value of heterogeneity test, *P*-het:3.08 × 10−2, 1.42 × 10−4 and 6.73 × 10−3, respectively). Interestingly, we also observed stronger associations of cg08285415 (*P* = 7.41 × 10−13, *P*-het = 2.57 × 10−5), cg05012158 (*P* = 2.40 × 10−13, *P*-het = 8.53 × 10−3) and cg06752398 (*P* = 1.90 × 10−20, *P*-het = 4.88 × 10−9) in smokers, while non-significant in those without smoking history (*P* > 0.05). Moreover, cg06752398 was more associated with NSCLC risk in male participants as well (*P* = 6.31 × 10−13, *P*-het = 2.26 × 10−9).

**Systematically multi-omics functional annotation for identified lung-cancer-associated CpG sites**

We integrated the evidence of epigenomics, genomics and transcriptomics and adopted a scoring strategy to systematicallyassess the functional importance of the 16 CpGs identified in all samples. As the heatmap of the functional annotation shown, we observed 5 CpGs at the “Level A” association, including cg11624060, cg26209169, and cg10441424 in 5p15.33, cg18468235 in 11q23.3 and cg19720302 in 17q24.2, 4 ones at “Level B” and 7 ones at “Level C” (**Figure 3**). In detail, the physical locations of the cg11624060, cg26209169, and cg10441424 were very close and located in ~1.8 kb downstream of *CLPTM1L* and ~20.9 kb upstream of *TERT*. Moreover, we observed the predicted enhancer signals of *TERT* and *SLC6A3* and active promoter and enhancer-related histone markers for them. Similarly, cg19720302 was shown to be a methylation marker within *C17orf58*, *BPTF,* and *AMZ2* enhancer regions (**Supplementary Table 10**). For meQTL of CpGs in 5p15.33, they were also identified as the eQTL of *CLPTM1L*, *SLC6A3,* or *NDUFS6* (**Supplementary Table 11 and 12**). Besides, three meQTLs of cg18468235 (rs2298831-C, rs17121881-T, and rs9332745-G) were predicted as the detrimental mutations for *JAML* and *KMT2A*; while rs139709271-TCC, meQTL of cg19720302, was a frameshift mutation for *BPTF* (**Supplementary Table 13**). Using the DNA methylation and gene expression data from lung cancer tissues in TCGA, we performed the Spearman rank correlation test to indentify the genes correlated with methylation level of CpG at a 1-Mb window, and most of the CpG sites in 5p15.33 were correlated with the expression level of *CLPTM1L*, *TERT*, and *SLC6A3* , of which TERT was a known driver gene for cancer(**Supplementary Table 14**). Besides, methylation-related gene (*JAML*, *IREB2* and *PSMA4*) of cg18468235, cg08285415 and cg05012158 were respectively shown the consistent directions of associations across CpG, gene expression and lung cancer (**Supplementary Table 15**).

**Integrative analyses of multi-omics for CpG-gene-NSCLC regulatory pathways**

To estimate the effect direction of methylation-related genes, we performed a differential expression analysis for 75 unduplicated genes by the gene expression data from 108 paired lung tumor tissues and adjacent normal tissues in TCGA. The FDR-corrected *P*-value of 55 genes was less than 0.05, indicating the expression levels of lung tumor tissues were significantly different from those of adjacent normal tissues (**Supplementary Table 16**). Then, we integrated all association results to investigate whether the DNA methylation at CpGs could affect the development of NSCLC through regulating the gene expression. Totally, there were 12 CpGs and 34 genes having the potential CpG-gene-NSCLC regulatory pathways (**Supplementary Table 17**). For example, the CpG site cg11624060 in 5p15.33 locus had a decreased NSCLC risk (Z score = −12.20, *P* = 3.01 × 10−34), and its DNA methylation level was negatively associated with expression of *TERT* (Rho = −0.34, *P* = 1.05 × 10−22), *TRIP13* (Rho = −0.34, *P* = 4.24 × 10−23) and *MRPL36* (Rho = −0.36, *P* = 3.89 × 10−26). Meanwhile, these genes were all up-regulated in lung tumor and the corresponding percentage of up-regulated pairs were 93.52% (*P* = 8.47 × 10−31), 93.52% (*P* = 4.83 × 10−27) and 90.74% (*P* = 2.71 × 10−23), respectively. The results of cg26209169 and cg10441424 were similar. Additionally, CpG sites and the genes nearby, such as cg18468235 with *JAML* and *IL10RA*, cg05012158 with *CHRNA5* and *PSMA4*, and cg19720302 with *KPNA2* and *AMZ2*, were also showing the CpG-gene-NSCLC regulatory pathways (**Table 2 and Supplementary Table 17**).

**Discussion**

In this large-scale two-stage study, we systematically assess the associations of genetically predicted DNA methylation CpGs with NSCLC risk with 34,964 cases and 448,579 controls. Sixteen DNA methylation markers of NSCLC were identified by performing a meta-analysis of genetically predicted association results, including 4 novel CpGs. After a multi-omics functional annotation, 9 CpG sites showed important functional evidence for NSCLC risk and 12 CpG sites might influence the NSCLC risk by regulating the expression of 34 genes flanking them. Overall, our study identified a couple of DNA methylation biomarkers and provided important insight into the epigenetic susceptibility mechanisms of NSCLC.

Here, we observed 39 statistically significant CpG sites at the screening stage and 16 of them passed the downstream validation, mainly located in 6 lung cancer susceptibility loci from previous GWAS (7, 22) except for cg08285415 in 15q24.3. Given in predictive associations were calculated from GWAS summary data, it is rational that the methylation loci we identified are highly overlapped with loci reported by genetics studies. Furthermore, we compared all of 16 replicated CpGs to those recorded in publications related to lung cancer risk in EWAS Atlas (36), cg22795331 in 6q22.1, cg05012158, cg06752398 in 15q25.1, and cg19720302 in 17q24.2 are novel methylation loci, while the others were located in regions reported in previous methylation studies. These findings may contribute to the understanding of the epigenetic susceptibility mechanisms of NSCLC risk, especially for the interplay of genetics and epigenetics. Besides, hypomethylation at cg22795331 and cg18468235 were respectively observed in colorectal cancer (37) and papillary thyroid carcinoma (38), indicating a potential methylation phenomenon of multi-cancer risk.

By integrating the multi-omics results across methylation level, gene expression, and NSCLC risk, we also found some consistent pathways that might be useful to expound the potential regulation mechanism. In 5p15.33 locus, *TERT* (Telomerase Reverse Transcriptase), one of the components of human telomerase, plays an important role in maintaining telomere length and activity. Nearly 90% of types of cancer have been found an up-regulation of telomerase, contributing to the cancer initiation (39). The *TRIP13* gene promotes proliferation and invasion of lung cancer cells through AKT/mTORC1/c-Myc signaling (40), Wnt signaling and epithelial-mesenchymal transition et.al pathways (41). For the gene *NDUFS6* that involves electrons transfer, some researchers observed silencing *NDUFS6* or *NDUFA11* significantly decreased reactive oxygen species levels in breast cancer, inhibiting the cancer-associated inflammation response (42). Furthermore, Mitochondrial Ribosomal Protein L36 (*MRPL36*) is essential for maintaining mitochondrial functions and significantly increases in LUSC compared with normal lung tissue (43), playing a crucial role in energy metabolism for human cancer (44).

For genes in 11q23.3, *JAML* (Junction Adhesion Molecule Like, alias *AMICA1*) expression was positively associated with infiltrating levels of diverse immune cells in LUAD (45). As a crucial component of epithelial gammadelta T cell biology, *JAML* also has broader implications in tissue homeostasis and repair (46). Protein encoding by *IL10RA* is a receptor for interleukin 10 (IL-10) and has been shown to mediate the immunosuppressive signal of IL-10, and thus inhibits the synthesis of proinflammatory cytokines. Production of IL-10 from lung tumors and immune cells promotes lung adenocarcinoma aggressiveness (47), and a positive feedback regulation between IL-10 and *EGFR* leads to lung cancer formation as well (48). In 17q24.2, overexpression of *KPNA2* flanking cg19720302 was observed in various cancers, including lung cancer (49). In addition, it has been shown to participate in cell differentiation, proliferation, apoptosis, immune response, and viral infection, and thus promote tumor formation and progression (49). Although most of the evidences from previous functional experimental studies above supported the potential regulation pathways of CpG-gene-lung cancer we identified, there still were some inconsistent without results, for example, *LPCAT1* (5p15.33) was reported up-regulated in LUAD tissues and cell lines and promoted brain metastasis (50) and defects in *AMZ2* (17q24.3) might be associated with lung tumorigenesis (51).

In subgroup analyses by smoking status, we also observed a significant association heterogeneity between smokers and non-smokers at cg08285415, cg05012158, and cg06752398. Interestingly, the nicotinic receptor subunit gene *CHRNA5* and tobacco addiction-related gene *PSMA4* were located nearby and showed a putative regulatory pathway in our study. Previous studies detected an up-regulation of the *CHRNA5* gene in NSCLC tumor tissue (52, 53) and low levels of *CHRNA5* mRNA were associated with lower risk for nicotine dependence and lung cancer (54), in agreement with our findings. However, some researchers found that lower expression of *CHRNA5* was causally linked to increased lung cancer risk using genetic instruments (34, 55). The *PSMA4* is an important component of the 20S core proteasome complex and marked by the tobacco addiction-associated gene in GeneCards (https://www.genecards.org/). To our knowledge, Benzo[a]pyrene and N-nitrosamines in tobacco smoke are two major carcinogens, leading to DNA damage, oxidative stress, and inflammation, and thus increasing the likelihood of developing lung cancer (56, 57). Therefore, it is reasonable to hypothesize that these genes may affect nicotine dependence and propensity to smoke and thus promote the initiation and growth of lung tumors, indirectly (58). In addition, *PSMA4* has been also considered as a strong candidate mediator of lung cancer cell growth, and also directly affects lung cancer susceptibility through its modulation of cell proliferation and apoptosis (58).

This is the first study to identify the genetically predicted DNA methylation markers associated with NSCLC risk using a two-stage case-control design with 34,964 cases and 448,579 controls. Besides, the prediction model approach has been proved that the performance of the results was improved obviously, compared to the single-meQTL SNP approach (17). Considering the chromatin interaction and distal regulation of gene expression, the nearest genes may not necessarily be the target genes of some CpGs. We comprehensively assess all the protein-coding genes within a 1-Mb range of each valid CpG site. Then, by multi-omics functional annotation and integrative analyses, we further explored the potential regulatory pathways for the identified CpGs, which may further contribute to the understanding of carcinogenesis mechanisms of NSCLC from the perspective of DNA methylation. However, some limitations are remaining in this study. First, the subjects used in the validation stage from FLCCA were only non-smoker females, lacking the necessary samples of smokers and males, which may lead to a decrease of validation power to some degree. Moreover, the histological types of lung cancer in the UK Biobank were undetermined, indicating a fraction of small cell lung cancer cases may be included. Furthermore, although most of the potential regulatory pathways we observed can be supported by experimental or biological evidence, the findings of integrative analyses are only data-driven evidence and would still be affected by unknown confounding factors and reverse causality. Therefore, further mechanism studies are warranted to test the authenticity behind it.

In conclusion, we investigated the associations of genetically predicted methylation level with NSCLC risk and a total of 16 CpG sites were identified, including 4 novel CpGs. We speculated these CpGs were likely to affect the NSCLC risk via regulating the flanking genes related to cancer formation and development. The findings of this study will provide an important insight into biology and epigenetics and also shed light on the epigenetic susceptibility mechanisms for NSCLC risk.

**Acknowledgements:** We thank Dr. Yaohua Yang and Prof. Jirong Long of Vanderbilt University Medical Center for their technical support in the construction of predicted models in this research. We thank the International Lung and Cancer Consortium and Carotene and Retinol Efficacy Trial project for lung cancer, who shared the genome-wide association study data from European populations.

**Data Availability:** The data we used in models building and refining (Framingham Cohort and Women's Health Initiative) are publicly available via dbGaP (dbGaP Accession Numbers: phs000342 and phs000724 for FHS; phs000315, phs000675, and phs001335 for WHI). Further information is available from the corresponding author upon request.

**Conflict of interest:** We declare no competing interests.

**Funding:** This work was supported by the National Key Research and Development Project (2019YFC1315704), National Natural Science of China (81820108028, 81922061, 81521004, 81973123, and 82072579), Research Unit of Prospective Cohort of Cardiovascular Diseases and Cancer, Chinese Academy of Medical Sciences (2019RU038), Jiangsu outstanding fund (BK20211533).

**Authors’ Contributions:**

**Conception and design:** X. Zhao, J. Fan, J. Dai and H. Shen

**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** X. Zhao, M. Yang, J. Fan, M. Wang, Y. Wang, M. Zhu, N. Qin and J. Dai

**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** M. Wang, Y. Wang, D. Albanes, S. Lam, A. Tardón, C. Chen, G.E. Goodman, S.E. Bojesen, M.T. Landi, M. Johansson, A. Risch, H.-E. Wichmann, H. Bickeböller, D.C. Christiani, G. Rennert, S.M. Arnold, P. Brennan, J.K. Field, S. Shete, L. Le Marchand, O. Melander, H. Brunnström, G. Liu, R.J. Hung, A.S. Andrew, L.A. Kiemeney, S. Zienolddiny, K. Grankvist, M. Johansson, N.E. Caporaso, P.J. Woll, P. Lazarus, M.B. Schabath, O.Y. Gorlova, I.P. Gorlov and A.V. Patel, M.C. Aldrich, C.I. Amos and H. Shen

**Writing, review, and/or revision of the manuscript:** X. Zhao, M. Yang, Fan, M. Zhu, N. Qin, Y. Jiang, H. Ma, G. Jin, Z. Hu, C.I. Amos, J. Dai and H. Shen

**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** X. Zhao, M. Wang, Y. Wang, G.E. Goodman, M. Johansson, D.C. Christiani, J.K. Field, G. Liu, S. Zienolddiny, M. Johansson, M.B. Schabath, H. Ma, G. Jin, Z. Hu, C.I. Amos, J. Dai and H. Shen

**Study supervision:** J. Dai and H. Shen

**Reference**

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209-49.

2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66(2):115-32.

3. Alberg AJ, Brock MV, Ford JG, Samet JM, Spivack SD. Epidemiology of lung cancer: Diagnosis and management of lung cancer, 3rd ed: American College of Chest Physicians evidence-based clinical practice guidelines. Chest. 2013;143(5 Suppl):e1S-e29S.

4. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000;343(2):78-85.

5. Dai J, Shen W, Wen W, Chang J, Wang T, Chen H, et al. Estimation of heritability for nine common cancers using data from genome-wide association studies in Chinese population. Int J Cancer. 2017;140(2):329-36.

6. Sampson JN, Wheeler WA, Yeager M, Panagiotou O, Wang Z, Berndt SI, et al. Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for Thirteen Cancer Types. J Natl Cancer Inst. 2015;107(12):djv279.

7. Bossé Y, Amos CI. A Decade of GWAS Results in Lung Cancer. Cancer Epidemiol Biomarkers Prev. 2018;27(4):363-79.

8. Duruisseaux M, Esteller M. Lung cancer epigenetics: From knowledge to applications. Semin Cancer Biol. 2018;51:116-28.

9. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun. 2015;6:10192.

10. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int J Cancer. 2017;140(1):50-61.

11. Sun YQ, Richmond RC, Suderman M, Min JL, Battram T, Flatberg A, et al. Assessing the role of genome-wide DNA methylation between smoking and risk of lung cancer using repeated measurements: the HUNT study. Int J Epidemiol. 2021;50(5):1482-97.

12. Battram T, Richmond RC, Baglietto L, Haycock PC, Perduca V, Bojesen SE, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. Int J Epidemiol. 2019;48(5):1493-504.

13. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12(1):R10.

14. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. 2014;15(5):R73.

15. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61.

16. Morrow JD, Glass K, Cho MH, Hersh CP, Pinto-Plata V, Celli B, et al. Human Lung DNA Methylation Quantitative Trait Loci Colocalize with Chronic Obstructive Pulmonary Disease Genome-Wide Association Loci. Am J Respir Crit Care Med. 2018;197(10):1275-84.

17. Yang Y, Wu L, Shu XO, Cai Q, Shu X, Li B, et al. Genetically Predicted Levels of DNA Methylation Biomarkers and Breast Cancer Risk: Data From 228 951 Women of European Descent. J Natl Cancer Inst. 2020;112(3):295-304.

18. Yang Y, Wu L, Shu X, Lu Y, Shu XO, Cai Q, et al. Genetic Data from Nearly 63,000 Women of European Descent Predicts DNA Methylation Biomarkers and Epithelial Ovarian Cancer Risk. Cancer Res. 2019;79(3):505-17.

19. Wu L, Yang Y, Guo X, Shu XO, Cai Q, Shu X, et al. An integrative multi-omics analysis to identify candidate DNA methylation biomarkers related to prostate cancer risk. Nat Commun. 2020;11(1):3905.

20. Zhu J, Yang Y, Kisiel JB, Mahoney DW, Michaud DS, Guo X, et al. Integrating Genome and Methylome Data to Identify Candidate DNA Methylation Biomarkers for Pancreatic Cancer Risk. Cancer Epidemiol Biomarkers Prev. 2021;30(11):2079-87.

21. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat Genet. 2017;49(7):1126-32.

22. Dai J, Lv J, Zhu M, Wang Y, Qin N, Ma H, et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. Lancet Respir Med. 2019;7(10):881-91.

23. Wang Y, Wu W, Zhu M, Wang C, Shen W, Cheng Y, et al. Integrating expression-related SNPs into genome-wide gene- and pathway-based analyses identified novel lung cancer susceptibility genes. Int J Cancer. 2018;142(8):1602-10.

24. Huan T, Joehanes R, Song C, Peng F, Guo Y, Mendelson M, et al. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Nat Commun. 2019;10(1):4267.

25. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat Genet. 2018;50(7):968-78.

26. McRae AF, Marioni RE, Shah S, Yang J, Powell JE, Harris SE, et al. Identification of 55,000 Replicated DNA Methylation QTL. Sci Rep. 2018;8(1):17605.

27. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8(2):203-9.

28. Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018;9(1):1825.

29. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. Nat Genet. 2010;42(5):436-40.

30. Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44(4):369-75, s1-3.

31. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.

32. Qin N, Li Y, Wang C, Zhu M, Dai J, Hong T, et al. Comprehensive functional annotation of susceptibility variants identifies genetic heterogeneity between lung adenocarcinoma and squamous cell carcinoma. Front Med. 2021;15(2):275-91.

33. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016;2016.

34. Bossé Y, Li Z, Xia J, Manem V, Carreras-Torres R, Gabriel A, et al. Transcriptome-wide association study reveals candidate causal genes for lung cancer. Int J Cancer. 2020;146(7):1862-78.

35. Zhu M, Fan J, Zhang C, Xu J, Yin R, Zhang E, et al. A cross-tissue transcriptome-wide association study identifies novel susceptibility genes for lung cancer in Chinese populations. Hum Mol Genet. 2021;30(17):1666-76.

36. Li M, Zou D, Li Z, Gao R, Sang J, Zhang Y, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. Nucleic Acids Res. 2019;47(D1):D983-d8.

37. Zhu L, Yan F, Wang Z, Dong H, Bian C, Wang T, et al. Genome-wide DNA methylation profiling of primary colorectal laterally spreading tumors identifies disease-specific epimutations on common pathways. Int J Cancer. 2018;143(10):2488-98.

38. Beltrami CM, Dos Reis MB, Barros-Filho MC, Marchi FA, Kuasne H, Pinto CAL, et al. Integrated data analysis reveals potential drivers and pathways disrupted by DNA methylation in papillary thyroid carcinomas. Clin Epigenetics. 2017;9:45.

39. Kim NW, Piatyszek MA, Prowse KR, Harley CB, West MD, Ho PL, et al. Specific association of human telomerase activity with immortal cells and cancer. Science. 1994;266(5193):2011-5.

40. Cai W, Ni W, Jin Y, Li Y. TRIP13 promotes lung cancer cell growth and metastasis through AKT/mTORC1/c-Myc signaling. Cancer Biomark. 2021;30(2):237-48.

41. Li ZH, Lei L, Fei LR, Huang WJ, Zheng YW, Yang MQ, et al. TRIP13 promotes the proliferation and invasion of lung cancer cells via the Wnt signaling pathway and epithelial-mesenchymal transition. J Mol Histol. 2021;52(1):11-20.

42. Mao W, Xiong G, Wu Y, Wang C, St Clair D, Li JD, et al. RORα Suppresses Cancer-Associated Inflammation by Repressing Respiratory Complex I-Dependent ROS Generation. Int J Mol Sci. 2021;22(19).

43. Li W, Li X, Gao LN, You CG. Integrated Analysis of the Functions and Prognostic Values of RNA Binding Proteins in Lung Squamous Cell Carcinoma. Front Genet. 2020;11:185.

44. Piao L, Li Y, Kim SJ, Byun HS, Huang SM, Hwang SK, et al. Association of LETM1 and MRPL36 contributes to the regulation of mitochondrial ATP production and necrotic cell death. Cancer Res. 2009;69(8):3397-404.

45. Fang L, Yu W, Yu G, Zhong F, Ye B. Junctional Adhesion Molecule-Like Protein (JAML) Is Correlated with Prognosis and Immune Infiltrates in Lung Adenocarcinoma. Med Sci Monit. 2022;28:e933503.

46. Witherden DA, Verdino P, Rieder SE, Garijo O, Mills RE, Teyton L, et al. The junctional adhesion molecule JAML is a costimulatory receptor for epithelial gammadelta T cell activation. Science. 2010;329(5996):1205-10.

47. Sung WW, Wang YC, Lin PL, Cheng YW, Chen CY, Wu TC, et al. IL-10 promotes tumor aggressiveness via upregulation of CIP2A transcription in lung adenocarcinoma. Clin Cancer Res. 2013;19(15):4092-103.

48. Hsu TI, Wang YC, Hung CY, Yu CH, Su WC, Chang WC, et al. Positive feedback regulation between IL10 and EGFR promotes lung cancer formation. Oncotarget. 2016;7(15):20840-54.

49. Han Y, Wang X. The emerging roles of KPNA2 in cancer. Life Sci. 2020;241:117140.

50. Wei C, Dong X, Lu H, Tong F, Chen L, Zhang R, et al. LPCAT1 promotes brain metastasis of lung adenocarcinoma by up-regulating PI3K/AKT/MYC pathway. J Exp Clin Cancer Res. 2019;38(1):95.

51. Huang W, Wang Y, Chu W, Tseng R. Refined-mapping of the novel TSG within the 17q24.3 chromosomal region in non-small cell lung cancer samples. Oncol Lett. 2016;12(3):1975-80.

52. Falvella FS, Galvan A, Frullanti E, Spinola M, Calabrò E, Carbone A, et al. Transcription deregulation at the 15q25 locus in association with lung adenocarcinoma risk. Clin Cancer Res. 2009;15(5):1837-42.

53. Song P, Sekhon HS, Fu XW, Maier M, Jia Y, Duan J, et al. Activated cholinergic signaling provides a target in squamous cell lung carcinoma. Cancer Res. 2008;68(12):4693-700.

54. Wang JC, Cruchaga C, Saccone NL, Bertelsen S, Liu P, Budde JP, et al. Risk for nicotine dependence and lung cancer is conferred by mRNA expression levels and amino acid change in CHRNA5. Hum Mol Genet. 2009;18(16):3125-35.

55. Yao C, Joehanes R, Wilson R, Tanaka T, Ferrucci L, Kretschmer A, et al. Epigenome-wide association study of whole blood gene expression in Framingham Heart Study participants provides molecular insight into the potential role of CHRNA5 in cigarette smoking-related lung diseases. Clin Epigenetics. 2021;13(1):60.

56. Mossman BT, Lounsbury KM, Reddy SP. Oxidants and signaling by mitogen-activated protein kinases in lung epithelium. Am J Respir Cell Mol Biol. 2006;34(6):666-9.

57. Hecht SS. DNA adduct formation from tobacco-specific N-nitrosamines. Mutat Res. 1999;424(1-2):127-42.

58. Liu Y, Liu P, Wen W, James MA, Wang Y, Bailey-Wilson JE, et al. Haplotype and cell proliferation analyses of candidate lung cancer susceptibility genes on chromosome 15q24-25.1. Cancer Res. 2009;69(19):7844-50.

|  |
| --- |
| **Table 1. Thirty-nine DNA methylation markers from meta-analysis associated with NSCLC risk** |
| CpG | CytoBand a | Position b | Classification | Closest gene | Screening Stage | Vaidation Stage | Combined Stage | I2(%) | *P.* Het d |
| Z score | *P* value c | Z score | *P* value c | Z score | *P* value |
| cg07507801 e | 5p15.33 | 1291235 | intronic | TERT | -7.96  | 1.70E-15 | -4.32  | 1.56E-05 | -9.06  | 1.35E-19 | 0.0  | 8.42E-01 |
| cg07380026 e | 5p15.33 | 1296007 | upstream | TERT | -7.69  | 1.43E-14 | -4.24  | 2.26E-05 | -8.78  | 1.61E-18 | 0.0  | 8.56E-01 |
| cg26603275 e | 5p15.33 | 1298965 | intergenic | TERT; MIR4457 | -9.29  | 1.49E-20 | -4.31  | 1.65E-05 | -10.17  | 2.81E-24 | 36.9  | 2.08E-01 |
| cg11624060 e | 5p15.33 | 1316038 | intergenic | MIR4457; CLPTM1L | -10.99  | 4.30E-28 | -5.48  | 4.18E-08 | -12.20  | 3.01E-34 | 48.1  | 1.65E-01 |
| cg26209169 e | 5p15.33 | 1316265 | intergenic | MIR4457; CLPTM1L | -10.04  | 9.72E-24 | -5.37  | 7.71E-08 | -11.35  | 7.23E-30 | 0.0  | 3.44E-01 |
| cg10441424 e | 5p15.33 | 1316637 | intergenic | MIR4457; CLPTM1L | -8.95  | 3.54E-19 | -5.59  | 2.29E-08 | -10.55  | 5.09E-26 | 0.0  | 8.40E-01 |
| cg07493874 e | 5p15.33 | 1342172 | intronic | CLPTM1L | 11.62  | 3.14E-31 | 5.43  | 5.55E-08 | 12.71  | 4.89E-37 | 66.0  | 8.62E-02 |
| cg19915256 e | 5p15.33 | 1345677 | upstream | CLPTM1L | -9.73  | 2.18E-22 | -5.90  | 3.63E-09 | -11.38  | 5.37E-30 | 0.0  | 7.76E-01 |
| cg27028750 e | 5p15.33 | 1349422 | intergenic | CLPTM1L; LINC01511 | 10.54  | 5.64E-26 | 5.96  | 2.54E-09 | 12.10  | 1.09E-33 | 0.0  | 6.15E-01 |
| cg23266546 | 6p22.1 | 28190811 | intergenic | TOB2P1; ZSCAN9 | 5.37  | 7.71E-08 | 1.07  | 2.84E-01 | 5.06  | 4.11E-07 | 77.2  | 3.63E-02 |
| cg15671450 | 6p22.1 | 29895116 | upstream | HCG4B | 5.84  | 5.23E-09 | 0.66  | 5.11E-01 | 5.18  | 2.19E-07 | 87.0  | 5.60E-03 |
| cg06710082 | 6p22.1 | 29943408 | ncRNA\_intronic | HCG9 | -5.23  | 1.67E-07 | -1.75  | 7.97E-02 | -5.39  | 7.02E-08 | 28.3  | 2.37E-01 |
| cg16368146 | 6p22.1 | 29943426 | ncRNA\_intronic | HCG9 | -4.99  | 6.17E-07 | -1.10  | 2.71E-01 | -4.72  | 2.37E-06 | 73.7  | 5.13E-02 |
| cg24694606 | 6p22.1 | 29977957 | ncRNA\_intronic | ZNRD1ASP | -5.83  | 5.53E-09 | -2.17  | 3.03E-02 | -6.06  | 1.37E-09 | 49.2  | 1.61E-01 |
| cg01044849 | 6p22.1 | 30002723 | ncRNA\_exonic | ZNRD1ASP | 5.73  | 9.91E-09 | 3.04  | 2.34E-03 | 6.49  | 8.59E-11 | 0.0  | 9.52E-01 |
| cg27493649 | 6p22.1 | 30042987 | intronic | RNF39 | 4.98  | 6.23E-07 | 1.50  | 1.35E-01 | 5.17  | 2.32E-07 | 0.0  | 5.63E-01 |
| cg14461571 e | 6p21.33 | 30458099 | exonic | HLA-E | -5.00  | 5.64E-07 | -3.26  | 1.11E-03 | -5.97  | 2.33E-09 | 0.0  | 9.66E-01 |
| cg19110902 | 6p21.33 | 30698937 | intronic | FLOT1 | 4.98  | 6.35E-07 | 1.37  | 1.70E-01 | 4.82  | 1.46E-06 | 71.3  | 6.21E-02 |
| cg06480496 | 6p21.33 | 31430676 | upstream | HCP5 | -4.90  | 9.47E-07 | -1.29  | 1.97E-01 | -4.78  | 1.73E-06 | 64.7  | 9.25E-02 |
| cg00848392 | 6p21.33 | 31734401 | exonic | VWA7 | -5.09  | 3.60E-07 | -1.27  | 2.05E-01 | -5.00  | 5.79E-07 | 60.4  | 1.12E-01 |
| cg21042276 | 6p21.33 | 32038542 | intronic | TNXB | -5.11  | 3.14E-07 | -0.97  | 3.34E-01 | -4.72  | 2.30E-06 | 79.0  | 2.89E-02 |
| cg06871764 | 6p21.32 | 32376096 | downstream | TSBP1-AS1 | 4.99  | 6.12E-07 | 0.99  | 3.20E-01 | 4.79  | 1.68E-06 | 65.9  | 8.66E-02 |
| cg22795331 e | 6q22.1 | 117785611 | intergenic | ROS1; DCBLD1 | -5.49  | 4.08E-08 | -4.03  | 5.69E-05 | -6.79  | 1.09E-11 | 0.0  | 6.90E-01 |
| cg27642470 | 6q22.1 | 117802711 | intergenic | ROS1; DCBLD1 | 4.83  | 1.33E-06 | 2.71  | 6.64E-03 | 5.54  | 3.02E-08 | 0.0  | 8.22E-01 |
| cg23172480 | 6q22.1 | 117802787 | upstream | DCBLD1 | 4.85  | 1.23E-06 | 2.92  | 3.47E-03 | 5.66  | 1.50E-08 | 0.0  | 8.89E-01 |
| cg17808183 | 7q11.21 | 63491010 | upstream | LINC01005 | 4.82  | 1.43E-06 | 1.83  | 6.77E-02 | 5.08  | 3.84E-07 | 0.0  | 3.71E-01 |
| cg10870165 | 8p12 | 32345448 | intronic | NRG1 | 4.94  | 7.93E-07 | 1.97  | 4.88E-02 | 5.26  | 1.46E-07 | 0.0  | 4.30E-01 |
| cg18468235 e | 11q23.3 | 118066105 | intronic | JAML | -5.48  | 4.16E-08 | -3.71  | 2.10E-04 | -6.62  | 3.64E-11 | 0.0  | 9.03E-01 |
| cg15794034 | 11q23.3 | 118095776 | upstream | JAML | -5.08  | 3.80E-07 | -2.81  | 4.92E-03 | -5.79  | 6.86E-09 | 0.0  | 7.22E-01 |
| cg18051914 | 11q23.3 | 118134912 | UTR5 | MPZL2 | 5.96  | 2.59E-09 | 2.72  | 6.54E-03 | 6.54  | 6.18E-11 | 0.0  | 7.49E-01 |
| cg26426447 | 11q23.3 | 118134959 | UTR5 | MPZL2 | 5.97  | 2.43E-09 | 2.57  | 1.02E-02 | 6.46  | 1.06E-10 | 0.0  | 4.84E-01 |
| cg09033131 | 11q23.3 | 118135094 | UTR5 | MPZL2 | 5.92  | 3.12E-09 | 1.36  | 1.74E-01 | 4.20  | 2.71E-05 | 94.8  | 1.09E-05 |
| cg15376097 | 11q23.3 | 118135271 | upstream | MPZL2 | 5.98  | 2.25E-09 | 2.99  | 2.82E-03 | 6.68  | 2.34E-11 | 0.0  | 9.67E-01 |
| cg08285415 e | 15q24.3 | 78283681 | intergenic | COMMD4P1; LOC91450 | -7.58  | 3.44E-14 | -4.25  | 2.17E-05 | -8.67  | 4.23E-18 | 0.0  | 5.92E-01 |
| cg08701566 | 15q25.1 | 78911099 | intronic | CHRNA3 | -4.96  | 7.01E-07 | -1.47  | 1.42E-01 | -5.02  | 5.09E-07 | 35.0  | 2.15E-01 |
| cg05012158 e | 15q25.1 | 79051864 | exonic | ADAMTS7 | -7.43  | 1.12E-13 | -4.31  | 1.65E-05 | -8.58  | 9.56E-18 | 0.0  | 7.50E-01 |
| cg06752398 e | 15q25.1 | 79053858 | intronic | ADAMTS7 | 9.16  | 5.15E-20 | 4.54  | 5.62E-06 | 10.12  | 4.40E-24 | 51.7  | 1.50E-01 |
| cg15822222 | 15q25.1 | 79164807 | upstream | MORF4L1 | -5.46  | 4.83E-08 | -1.50  | 1.33E-01 | -5.24  | 1.64E-07 | 78.4  | 3.14E-02 |
| cg19720302 e | 17q24.2 | 65990670 | upstream | C17orf58 | -5.65  | 1.65E-08 | -3.86  | 1.13E-04 | -6.84  | 8.16E-12 | 0.0  | 8.08E-01 |

a: Cytogenic band where the variant is positioned.

b: Chromosomal position, hg19/GRCh37 build.

c: Bonferroni correction threshold for *P* value is 1.67×10−6 (0.05/29,894) in the screening stage and 1.28×10−3 (0.05/39) in the validation stage.

d: Cochran’s Q test is used to test for heterogeneity in effect sizes of CpGs across two stages (I²; heterogeneity *P* value) , and *P* ≤ 0.05 is statistically significant.

e: CpG sites pass the independent validation.

|  |
| --- |
| **Table 2.** **Integrative analyses for potential regulatory pathways across DNA methylation, gene expression and NSCLC risk a** |
| Chr | Position | CpG | Gene | CpG vs. NSCLC risk | CpG vs. Gene expression | Gene expression vs. NSCLC risk |
| Z score b | *P* value b | Direction | Rho c | *P* value.FDR c | Direction | Up. regulation d | *P* value.FDR d | Direction |
| 5 | 1316038 | cg11624060 | TERT | -12.20  | 3.01E-34 | Negative | -0.34  | 1.05E-22 | Negative | 93.52% | 8.47E-31 | Positive |
| 5 | 1316038 | cg11624060 | TRIP13 | -12.20  | 3.01E-34 | Negative | -0.34  | 4.24E-23 | Negative | 93.52% | 4.83E-27 | Positive |
| 5 | 1316038 | cg11624060 | MRPL36 | -12.20  | 3.01E-34 | Negative | -0.36  | 3.89E-26 | Negative | 90.74% | 2.71E-23 | Positive |
| 5 | 1316038 | cg11624060 | NDUFS6 | -12.20  | 3.01E-34 | Negative | -0.34  | 1.04E-22 | Negative | 85.19% | 4.66E-23 | Positive |
| 5 | 1316038 | cg11624060 | LPCAT1 | -12.20  | 3.01E-34 | Negative | 0.22  | 5.84E-10 | Positive | 16.67% | 3.46E-16 | Negative |
| 5 | 1316264 | cg26209169 | TERT | -11.35  | 7.23E-30 | Negative | -0.30  | 7.39E-18 | Negative | 93.52% | 8.47E-31 | Positive |
| 5 | 1316264 | cg26209169 | TRIP13 | -11.35  | 7.23E-30 | Negative | -0.29  | 1.99E-17 | Negative | 93.52% | 4.83E-27 | Positive |
| 5 | 1316264 | cg26209169 | MRPL36 | -11.35  | 7.23E-30 | Negative | -0.34  | 6.98E-23 | Negative | 90.74% | 2.71E-23 | Positive |
| 5 | 1316264 | cg26209169 | NDUFS6 | -11.35  | 7.23E-30 | Negative | -0.31  | 4.81E-20 | Negative | 85.19% | 4.66E-23 | Positive |
| 5 | 1316264 | cg26209169 | LPCAT1 | -11.35  | 7.23E-30 | Negative | 0.19  | 6.27E-08 | Positive | 16.67% | 3.46E-16 | Negative |
| 5 | 1316636 | cg10441424 | TERT | -10.55  | 5.09E-26 | Negative | -0.28  | 3.90E-16 | Negative | 93.52% | 8.47E-31 | Positive |
| 5 | 1316636 | cg10441424 | TRIP13 | -10.55  | 5.09E-26 | Negative | -0.46  | 3.47E-44 | Negative | 93.52% | 4.83E-27 | Positive |
| 5 | 1316636 | cg10441424 | MRPL36 | -10.55  | 5.09E-26 | Negative | -0.43  | 5.32E-38 | Negative | 90.74% | 2.71E-23 | Positive |
| 5 | 1316636 | cg10441424 | NDUFS6 | -10.55  | 5.09E-26 | Negative | -0.34  | 7.45E-24 | Negative | 85.19% | 4.66E-23 | Positive |
| 5 | 1316636 | cg10441424 | LPCAT1 | -10.55  | 5.09E-26 | Negative | 0.58  | 4.83E-74 | Positive | 16.67% | 3.46E-16 | Negative |
| 5 | 1342172 | cg07493874 | TRIP13 | 12.71  | 4.89E-37 | Positive | 0.23  | 1.74E-11 | Positive | 93.52% | 4.83E-27 | Positive |
| 5 | 1342172 | cg07493874 | MRPL36 | 12.71  | 4.89E-37 | Positive | 0.29  | 1.17E-16 | Positive | 90.74% | 2.71E-23 | Positive |
| 5 | 1342172 | cg07493874 | NDUFS6 | 12.71  | 4.89E-37 | Positive | 0.28  | 3.06E-16 | Positive | 85.19% | 4.66E-23 | Positive |
| 5 | 1342172 | cg07493874 | CLPTM1L | 12.71  | 4.89E-37 | Positive | 0.13  | 4.95E-04 | Positive | 80.56% | 1.08E-11 | Positive |
| 5 | 1342172 | cg07493874 | BRD9 | 12.71  | 4.89E-37 | Positive | 0.25  | 1.16E-12 | Positive | 74.07% | 1.74E-07 | Positive |
| 11 | 118066105 | cg18468235 | AMICA1 | -6.62  | 3.64E-11 | Negative | 0.18  | 6.41E-07 | Positive | 1.85% | 7.02E-32 | Negative |
| 11 | 118066105 | cg18468235 | IL10RA | -6.62  | 3.64E-11 | Negative | 0.32  | 2.31E-21 | Positive | 17.59% | 1.17E-13 | Negative |
| 11 | 118066105 | cg18468235 | TMPRSS13 | -6.62  | 3.64E-11 | Negative | -0.20  | 2.41E-08 | Negative | 76.85% | 1.07E-11 | Positive |
| 11 | 118066105 | cg18468235 | ARCN1 | -6.62  | 3.64E-11 | Negative | -0.19  | 1.00E-07 | Negative | 69.44% | 1.14E-07 | Positive |
| 11 | 118066105 | cg18468235 | CD3E | -6.62  | 3.64E-11 | Negative | 0.43  | 1.03E-38 | Positive | 31.48% | 3.40E-04 | Negative |
| 15 | 78283681 | cg08285415 | SH2D7 | -8.67  | 4.23E-18 | Negative | 0.14  | 1.66E-04 | Positive | 33.33% | 6.10E-03 | Negative |
| 15 | 79051863 | cg05012158 | RASGRF1 | -8.58  | 9.56E-18 | Negative | 0.16  | 1.02E-05 | Positive | 4.63% | 8.76E-30 | Negative |
| 15 | 79051863 | cg05012158 | CHRNA5 | -8.58  | 9.56E-18 | Negative | -0.08  | 3.29E-02 | Negative | 92.59% | 1.23E-27 | Positive |
| 15 | 79051863 | cg05012158 | CTSH | -8.58  | 9.56E-18 | Negative | 0.18  | 1.84E-07 | Positive | 6.48% | 5.60E-27 | Negative |
| 15 | 79051863 | cg05012158 | PSMA4 | -8.58  | 9.56E-18 | Negative | -0.12  | 7.94E-04 | Negative | 75% | 2.24E-11 | Positive |
| 17 | 65990670 | cg19720302 | KPNA2 | -6.84  | 8.16E-12 | Negative | -0.19  | 9.40E-08 | Negative | 95.37% | 3.21E-28 | Positive |
| 17 | 65990670 | cg19720302 | NOL11 | -6.84  | 8.16E-12 | Negative | -0.33  | 3.58E-21 | Negative | 91.67% | 6.67E-24 | Positive |
| 17 | 65990670 | cg19720302 | AMZ2 | -6.84  | 8.16E-12 | Negative | -0.33  | 8.54E-22 | Negative | 87.96% | 8.36E-20 | Positive |
| 17 | 65990670 | cg19720302 | C17orf58 | -6.84  | 8.16E-12 | Negative | -0.30  | 1.29E-18 | Negative | 86.11% | 4.43E-18 | Positive |
| 17 | 65990670 | cg19720302 | SLC16A6 | -6.84  | 8.16E-12 | Negative | 0.19  | 1.84E-07 | Positive | 12.96% | 3.88E-17 | Negative |

a: For CpGs with the high functional level in multi-omics annotation, the integrative results of top 5 differential genes were selected. The complete list was shown in Supplementary Table 17.

b: Z score and *P* value were derived from the combined stage.

c: Rho and *P* value were calculated by the Spearman rank correlation test.

d: The percentage of up-regulation pair was calculated by relative expression levels of genes (indicated by log2-transformed tumor/adjacent tissues), and *P* value was calculated by Wilcoxon rank-sum test.

FDR-corrected *P* value was calculated by Benjamini-Hochberg method and *P*≤0.05 was statistically significant. All statistical tests are two-sided. Chr=chromsome. NSCLC=non-small cell lung cancer. FDR=false discovery rate.

**Titles and legends to figure**

**Figure 1. Flowchart for the study design.**

**Figure 2. Manhattan plot for thirty-nine DNA methylation markers from meta-analysis associated with NSCLC risk.** The green dotted line represents *P* = 1.67 × 10−6 (Bonferroni correction of 29,894 tests, 0.05/29,894). Each dot represents the genetically predicted DNA methylation of one specific CpG site. The x axis represents the negative logarithm of the association *P*-value, and the y axis represents the chromosome of the CpG site. The red represents the combined effect of 16 CpG sites passed the independent validation, and the diamond represents the novel CpG sites in regions not yet reported in previous lung cancer EWASs.

**Figure 3. Heatmap of multi-omics functional annotation for the identified CpG sites.** Here, we performed the functional annotation for 16 CpGs passed the validation based on evidence of epigenomics, genomics and transcriptomics level. TSS1500=transcription start site upstream 1,500bp. DHS=DNase I hypersensitivity sites. TAD=topologically associated domains. TF=transcription factor. LC=lung cancer. TWAS= transcriptome-wide association study.