Published in final edited form as:
Cancer. 2024 March 15; 130(6): 913–926. doi:10.1002/cncr.35130.

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

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

Correspondence Juncheng Dai and Hongbing Shen, Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, 101 Longmian Rd, Nanjing 211166, China. djc@njmu.edu.cn and hbshen@njmu.edu.cn.

The first three authors contributed equally to this work.

AUTHOR CONTRIBUTIONS

Xiaoyu Zhao: Analysis and interpretation of data and writing, review, and/or revision of the manuscript. **Meiqi Yang:** Analysis and interpretation of data and writing, review, and/or revision of the manuscript. **Jingyi Fan:** Analysis and interpretation of data and writing, review, and/or revision of the manuscript. **Mei Wang:** Analysis and interpretation of data and acquisition of data. **Yifan Wang:** Analysis and interpretation of data and acquisition of data. **Na Qin:** Analysis and interpretation of data and writing, review, and/or revision of the manuscript. **Meng Zhu:** Analysis and interpretation of data and writing, review, and/or revision of the manuscript. **Yue Jiang:** Writing, review, and/or revision of the manuscript. **Olga Y. Gorlova:** Acquisition of data. **Ivan P. Gorlov:** Acquisition of data. **Demetrius Albanes:** Acquisition of data. **Stephen Lam:** Acquisition of data. **Adonina Tardón:** Acquisition of data. **Chu Chen:** Acquisition of data. **Gary E. Goodman:** Acquisition of data. **Stig E. Bojesen:** Acquisition of data. **Maria Teresa Landi:** Acquisition of data. **Mattias Johansson:** Acquisition of data. **Angela Risch:** Acquisition of data. **H.-Erick Wichmann:** Acquisition of data. **Heike Bickeböller:** Acquisition of data. **David C. Christiani:** Acquisition of data. **Gad Rennert:** Acquisition of data. **Susanne M. Arnold:** Acquisition of data. **Paul Brennan:** Acquisition of data. **John K. Field:** Acquisition of data. **Sanjay Shete:** Acquisition of data. **Loïc Le Marchand:** Acquisition of data. **Geoffrey Liu:** Acquisition of data. **Rayjean J. Hung:** Acquisition of data. **Angeline S. Andrew:** Acquisition of data. **Lambertus A. Kiemeney:** Acquisition of data and writing, review, and/or revision of the manuscript. **Shanbeh Zienolddiny:** Acquisition of data. **Kiell Grankvist:** Acquisition of data. **Mikael Johansson:** Acquisition of data. **Neil E. Caporaso:** Acquisition of data. **Penella J. Woll:** Acquisition of data. **Philip Lazarus:** Acquisition of data. **Matthew B. Schabath:** Acquisition of data. **Melinda C. Aldrich:** Acquisition of data. **Alpa V. Patel:** Acquisition of data. **Michael P.A Davies:** Acquisition of data. **Hongxia Ma:** Writing, review, and/or revision of the manuscript. **Guangfu Jin:** Writing, review, and/or revision of the manuscript. **Zhibin Hu:** Writing, review, and/or revision of the manuscript. **Christopher I. Amos:** Acquisition of data and writing, review, and/or revision of the manuscript. **Hongbing Shen:** Acquisition of data; writing, review, and/or revision of the manuscript; study conception, design, and acquisition of funding; and analysis and interpretation of data. **Juncheng Dai:** Analysis and interpretation of data; writing, review, and/or revision of the manuscript; and study conception, design, and acquisition of funding. All authors read and approved the final manuscript.

CONFLICT OF INTEREST STATEMENT

Alpa Patel is employed by the American Cancer Society, which receives grants from private and corporate foundations, including foundations associated with companies in the health sector, for research outside the submitted work. The remaining authors declare no conflicts of interest. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Hongxia Ma, PhD[1,3], **Guangfu Jin, PhD**[1,3], **Zhibin Hu, PhD**[1,3], **Christopher I. Amos, PhD**[42], **Hongbing Shen, PhD**[1,3], **Juncheng Dai, PhD**[1,3]

[1] Department of Epidemiology, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China

[2] Department of Statistics, The First Affiliated Hospital of Nanjing Medical University, Nanjing, China

[3] 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, China

[4] Health Management Center, Gusu School, The Affiliated Suzhou Hospital of Nanjing Medical University, Suzhou Municipal Hospital, Suzhou, China

[5] 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, China

[6] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire, USA

[7] Department of Medicine, Epidemiology Section, Institute for Clinical and Translational Research, Baylor Medical College, Houston, Texas, USA

[8] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA

[9] Department of Integrative Oncology, British Columbia Cancer Agency, Vancouver, British Columbia, Canada

[10] Department of Public Health IUOPA, University of Oviedo, ISPA and CIBERESP, Oviedo, Spain

[11] Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[12] Public Health Sciences Division, Swedish Cancer Institute, Seattle, Washington, USA

[13] Department of Clinical Biochemistry, Copenhagen University Hospital, Copenhagen, Denmark

[14] Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

[15] National Cancer Institute, Bethesda, Maryland, USA

[16] Genetic Epidemiology Group, International Agency for Research on Cancer, Lyon, France

[17] Department of Biosciences, Allergy-Cancer-BioNano Research Centre, University of Salzburg, Salzburg, Austria

[18] Division of Epigenomics and Cancer Risk Factors, DKFZ-German Cancer Research Center, Heidelberg, Heidelberg, Germany

[19] Translational Lung Research Center Heidelberg (TLRC-H), German Center for Lung Research (DZL), Heidelberg, Germany

[20] Institute of Epidemiology, Helmholtz Center Munich, Neuherberg, Germany

[21]Department of Genetic Epidemiology, University Medical Center Goettingen, Goettingen, Germany

[22]Departments of Environmental Health and Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

[23]Technion Faculty of Medicine, Carmel Medical Center, Haifa, Israel

[24]Markey Cancer Center, University of Kentucky, Lexington, Kentucky, USA

[25]Molecular and Clinical Cancer Medicine, Roy Castle Lung Cancer Research Programme, The University of Liverpool Institute of Translational Medicine, Liverpool, UK

[26]Department of Epidemiology, The University of Texas, MD Anderson Cancer Center, Houston, Texas, USA

[27]Epidemiology Program, University of Hawai'i Cancer Center, Honolulu, Hawaii, USA

[28]Princess Margaret Cancer Centre, Toronto, Ontario, Canada

[29]Prosseman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Ontario, Canada

[30]Department of Neurology, Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire, USA

[31]Department for Health Evidence, Radboud University Medical Center, Nijmegen, the Netherlands

[32]National Institute of Occupational Health (STAMI), Oslo, Norway

[33]Department of Medical Biosciences, Umeå University, Umea, Sweden

[34]Department of Radiation Sciences, Umeå University, Umea, Sweden

[35]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA

[36]Academic Unit of Clinical Oncology, University of Sheffield, Sheffield, UK

[37]Department of Pharmaceutical Sciences, College of Pharmacy and Pharmaceutical Sciences, Washington State University, Spokane, Washington, USA

[38]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA

[39]Department of Medicine (Division of Genetic Medicine), Vanderbilt University Medical Center, Nashville, Tennessee, USA

[40]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, Georgia, USA

[41]Institute of Translational Medicine, University of Liverpool, Liverpool, UK

[42]Baylor College of Medicine, Institute for Clinical and Translational Research, Houston, Texas, USA

Abstract

Background: Although the associations between genetic variations and lung cancer risk have been explored, the epigenetic consequences of DNA methylation in lung cancer development are largely unknown. Here, the genetically predicted DNA methylation markers associated with non–small cell lung cancer (NSCLC) risk by a two-stage case-control design were investigated.

Methods: The genetic prediction models for methylation levels based on genetic and methylation data of 1595 subjects from the Framingham Heart Study were established. The prediction models were applied to a fixed-effect meta-analysis of screening data sets with 27,120 NSCLC cases and 27,355 controls to identify the methylation markers, which were then replicated in independent data sets with 7844 lung cancer cases and 421,224 controls. Also performed was a multi-omics functional annotation for the identified CpGs by integrating genomics, epigenomics, and transcriptomics and investigation of the potential regulation pathways.

Results: Of the 29,894 CpG sites passing the quality control, 39 CpGs associated with NSCLC risk (Bonferroni-corrected $p \leq 1.67 \times 10^{-6}$) were originally identified. Of these, 16 CpGs remained significant in the validation stage (Bonferroni-corrected $p \leq 1.28 \times 10^{-3}$), including four novel CpGs. Multi-omics functional annotation showed nine of 16 CpGs were potentially functional biomarkers for NSCLC risk. Thirty-five genes within a 1-Mb window of 12 CpGs that might be involved in regulatory pathways of NSCLC risk were identified.

Conclusions: Sixteen promising DNA methylation markers associated with NSCLC were identified. Changes of the methylation level at these CpGs might influence the development of NSCLC by regulating the expression of genes nearby.

- The epigenetic consequences of DNA methylation in lung cancer development are still largely unknown.
- This study used summary data of large-scale genome-wide association studies to investigate the associations between genetically predicted levels of methylation biomarkers and non–small cell lung cancer risk at the first time.
- This study looked at how well larotrectinib worked in adult patients with sarcomas caused by TRK fusion proteins.
- These findings will provide a unique insight into the epigenetic susceptibility mechanisms of lung cancer.

Keywords

association study; DNA methylation; gene expression; genetic prediction; non–small cell lung cancer risk

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] Non–small cell lung cancer (NSCLC) accounts for approximately 85% of total lung cancer cases and mainly includes adenocarcinoma (LUAD) and squamous cell carcinoma as subtypes.[3] The development of lung cancer involves the interplay between environmental and genetic risk factors. Over the past decade, more than 45 genetic loci were identified for lung cancer risk by genome-wide association studies (GWASs).[4-6] Epigenetics including DNA methylation has also been found to play a critical role in lung cancer pathogenesis.

Based on candidate strategy, early studies have identified some methylation markers potentially associated with lung cancer risk, such as hypermethylation at promoters of *RASSF1, CDKN2A, MGMT, APC*, and *DAPK*.[7] Recent emerging epigenome-wide association studies also revealed several new methylation markers (e.g., cg05575921-*AHRR*, cg03636183-*F2RL3*); however, more new findings were hindered by the limited sample size.[8-10] Furthermore, because of selection bias, potential confounding, and reverse causation, the causal association of DNA methylation may be inconsistent with results from observational studies.[11]

DNA methylation is impacted by both environmental factors and genetic factors. Previous studies have identified multiple DNA methylation quantitative trait loci (meQTL),[12,13] 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. Yang et al developed new statistical models to predict DNA methylation via multiple genetic variants in a reference data set and applied them to the summary data of GWASs to investigate the association between genetically predicted DNA methylation and disease risk.[14-17]

Here, we will adopt the prediction method to identify new lung cancer-associated methylation markers based on 34,964 cases and 448,579 controls. The findings will contribute to reveal the epigenetic susceptibility mechanisms of NSCLC.

MATERIALS AND METHODS

Study design and participants

The overall design is exhibited in Figure 1. First, we trained the DNA methylation prediction models by using data from 1595 Framingham Heart Study (FHS) participants and then refined in 883 subjects of Women’s Health Initiative (WHI). After that, we selected the prediction models with qualified performance to assess the association between genetically predicted methylation markers and NSCLC risk, based on summary data of GWASs 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).[6] For those identified methylation markers, we validated in external data sets with 7844 lung cancer cases and 421,224 controls from the UK Biobank (https://pan.ukbb.broadinstitute.org) and Female Lung Cancer Consortium in Asia (FLCCA).[18] Basic information and clinical features of participants for these data sets are shown in Table S1. The Biobank Japan summary data (4050 lung cancer cases and 208,403 controls) was used as an independent replication. Besides, we conducted a multi-omics functional annotation for the identified CpG sites by integrative analyses of epigenomics, genomics, and transcriptomics data obtained from a previous study[19] or The Cancer Genome Atlas, and finally investigated the potential regulatory pathways.

DNA methylation prediction models training and refining

Here, 1595 unrelated European subjects with matched genetic and DNA methylation data in the FHS were used to construct DNA methylation prediction models (dbGaP: phs000342 and phs000724). The detailed information about data sets and data process have been described elsewhere[14-17] and are shown in Supporting Information S1. 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[20] to predict DNA methylation residuals. An internal validation for each model was performed using 10-fold cross-validation. The ${R_{FHS}}^2$ 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 WHI (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 ${R_{WHI}}^2$ values were calculated by Spearman's correlation test. Furthermore, we selected the models with satisfactory prediction performance according to these criteria: (1) with a ${R_{FHS}}^2 \geq 0.01$ (≥10% correlation between predicted and measured methylation levels) in FHS; (2) with a ${R_{WHI}}^2 \geq 0.01$ in WHI; and (3) probes with no single-nucleotide polymorphism (SNPs) overlapped, considering that SNPs on the probes might have a potential impact on the methylation level estimation.[21]

Association analyses between predicted methylation and NSCLC risk

We used S-PrediXcan[22] to evaluate the associations between genetically predicted methylation levels and NSCLC risk. In brief, the association indicator Z-score was estimated by this formula:

$$Z_m \approx \sum_{s \in \mathrm{Model}_m} w_{sm} \frac{\hat{\sigma}_s}{\hat{\sigma}_m} \frac{\hat{\beta}_s}{se\left(\hat{\beta}_s\right)}$$

In the formula, w_{sm} is the weight of SNP$_s$ in the prediction of the CpG $_m$. $\hat{\sigma}_s$ and $\hat{\sigma}_m$ are the estimated variances of SNP$_s$ and CpG $_m$. $\hat{\beta}_s$ and $se(\hat{\beta}_s)$ are the GWAS regression coefficients and standard error of $\hat{\beta}_s$. We used summary data from 2 GWASs that had been generated from 27,820 European individuals and 26,655 Chinese individuals[6] to estimate the associations between genetically predicted methylation levels with NSCLC risk. Considering the population heterogeneity, we conducted a fixed-effect meta-analysis of two populations using META v1.7 to identify the shared methylation markers; $p \leq .05$ for Cochran's Q statistic indicated a high degree of heterogeneity. We further filtered out those

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 ($p \leq 1.67 \times 10^{-6}$; 0.05/29,894). At the validation stage, we replicated the 39 CpGs by summary data of Pan-UK Biobank and FLCCA. The same strategy of meta-analysis was performed, and the Bonferroni-corrected test was again used to determine the passing CpG sites ($p \leq 1.28 \times 10^{-3}$; 0.05/39).

For replicated CpG sites, we assessed whether the observed associations were independent of lung cancer susceptibility variants identified in previous GWASs.[4-6] Briefly, we used genome-wide complex trait analysis-conditional and joint analysis[23] to reevaluate the betas and standard errors of lung cancer by adjusting the closest GWAS-identified risk variants, and then reran 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 $p \leq .05$ was statistical threshold. Finally, given the potential ethnicity heterogeneity of model application, an external replication was conducted for those shared CpGs of combined populations and Asian-specific CpGs by GWAS summary data from the Biobank Japan.[24]

Systematic multi-omics functional annotation

We performed multi-omics functional annotations based on epigenomics, genomics, and transcriptomics data for the CpGs passing the validation. The types and sources of related annotation information are described in Table S2. For the epigenomics level, we used ANNOVAR to annotate the closest genes and regions of the identified CpGs; an extended annotation obtained from the Illumina 450K platform (GEO: GPL18809) was as a supplement. Moreover, the chromatin interactions, topologically associated domains, transcription factor binding sites, and histone mark were further annotated. For the genomics level, we assessed whether the corresponding cis-meQTL was overlapped with the expression quantitative trait loci (eQTL) in the Genotype-Tissue Expression. Six bioinformatic-predictive algorithms (Supporting Information S1) were used for evaluation of detrimental missense variants among these cis-meQTL.[25] For transcriptomics level, we identified the methylation-related protein-encoding genes within a 1-Mb range of each CpG site by Spearman correlation coefficients (false-discovery-rate [FDR] corrected $p \leq .05$). We assessed these methylation-related genes were of lung cancer–driver genes,[25] lung cancer–associated genes, or consistent with findings from transcriptome-wide association studies in lung cancer.[26,27]

To estimate the functional importance of these identified CpGs with NSCLC risk, a functional score system was constructed. One score was given if CpG met the corresponding criterion of each indicator (Table S3, Supporting Information S1). Altogether, functional score ranged from 0 to 10 in the epigenomics level (1 omics score given if score ≥ 5), from 0 to 3 in the genomics level (1 omics score given if score ≥ 2), and from 0 to 7 in the transcriptomics level (1 omics 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 for potential regulatory pathways

Based on gene expression of 108 tumor-adjacent tissues pairs from lung cancer in The Cancer Genome Atlas, we conducted the differential expression analyses for those methylation-related genes. The number and percentage of upregulation pairs were calculated by log2-transformed data of tumor and adjacent tissues. A Wilcoxon rank-sum test was used and FDR-corrected threshold of $p \leq .05$ was statistically significant. Finally, we integrated the association between genetically predicted methylation and NSCLC risk, the correlation between DNA methylation and gene expression, and the relationship of differential expression between lung cancer tissues and adjacent normal tissues to elucidate the putative pathways through which DNA methylation affects the development of NSCLC.

This study was approved by the institutional review board of Nanjing Medical University. All data in this study were derived from previous studies, which were approved by the local internal review board or ethics committee.

RESULTS

DNA methylation prediction models

Based on individual-level genotyping and DNA methylation data from the FHS cohort, DNA methylation prediction models for 223,959 CpG were established, of which 81,352 models with a predictive performance (${R_{FHS}}^2$) of at least 0.01 were retained. Among these, 70,330 models (86.45%) with good repeatability were observed in the WHI cohort (${R_{WHI}}^2 \geq 0.01$), suggesting a high correlation between two cohorts (Pearson’s correlation $r = 0.95$, $p \leq .0001$; Figure S1). Besides, methylation probes of 7284 had SNPs within the binding site, which were excluded. Totally, there were 63,046 CpGs remaining for the downstream analyses.

Association of genetically predicted methylation with NSCLC risk

At the screening stage, we did a fixed-effects 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 $p \leq .05$ ($n = 7626$) and those without consistent effect directions ($n = 25,371$), a total of 29,894 CpGs remained. We observed that 39 CpGs located in 10 loci were significantly associated with NSCLC risk (Bonferroni correction $p \leq 1.67 \times 10^{-6}$, 0.05/29,894) (Figure 2 and Table S4).

At the validation stage, we replicated the 39 CpGs using summary data of 7844 lung cancer cases and 421,224 controls. As shown in Table S5, 25 CpGs with the same effect direction were at $p < .05$, 16 of which met the Bonferroni correction ($p \leq 1.28 \times 10^{-3}$, 0.05/39). Four of the replicated 16 CpGs (cg22795331, cg05012158, cg06752398, and cg19720302) were the first reported methylated loci associated with NSCLC risk and 12 were located in susceptibility regions reported previously (Figure 2 and Table 1). A positive association of 3 CpGs with NSCLC risk was detected (cg07493874, cg27028750, and cg06752398), whereas the other 13 CpGs were negatively associated with NSCLC (Table 1). However, we did not observe any of the 16 valid CpGs remaining significant ($p \leq 1.67 \times 10^{-6}$) after adjusting GWAS-identified lung cancer susceptibility variants (Table S6). Additionally, the respective results of methylation markers derived from two populations were also exhibited

(Tables S7 and S8). Briefly, methylation markers of European descent were mainly located in the 5p15.33, 6p22.1, 6p21.33, and 15q25.1 regions. Of these, 5p15.33 was shared with the Chinese population, whereas the other markers in 2p23.1, 6p21.32, 11q23.3, 17q24.2, and 20q11.23 showed a racial difference. Finally, we observed 19 of 39 shared CpGs of combined populations (including 10 of 16 valid CpGs mentioned previously) and 12 of 15 Asian-specific CpGs consistent with the Z score direction of the upstream analyses ($p \leq .05$), especially in the 5p15.33 locus (Tables S9 and S10).

In subgroup analyses by histological type, smoking status, and gender (Table S11), we found that three of 16 valid CpGs (cg07507801, cg22795331, and cg18468235) showed the stronger associations in lung adenocarcinoma (p-het: 3.08×10^{-2}; 1.42×10^{-4}; and 6.73×10^{-3}). Interestingly, we found the obvious associations of cg08285415 ($p = 7.41 \times 10^{-13}$), cg05012158 ($p = 2.40 \times 10^{-13}$), and cg06752398 ($p = 1.90 \times 10^{-20}$) in smokers, whereas this was nonsignificant in nonsmokers. Moreover, cg06752398 had a stronger association in male participants (p-het $= 2.26 \times 10^{-9}$).

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

We integrated the evidence of epigenomics, genomics, and transcriptomics and adopted a scoring strategy to systematically assess the functional importance of the 16 CpGs. As the heatmap shows, 5 CpGs were at “level A,” including cg11624060, cg26209169, and cg10441424 in 5p15.33, cg18468235 in 11q23.3. and cg19720302 in 17q24.2; four at “level B”; and seven at “level C” (Figure 3). In detail, the physical locations of the cg11624060, cg26209169, and cg10441424 were very close and located ~1.8 kb downstream of *CLPTM1L* and ~20.9 kb upstream of *TERT*. We observed the predicted enhancer signals of *TERT* and promotor/enhancer-related histone markers (Table S12). The meQTL of CpGs in 5p15.33 also overlapped with eQTL of *CLPTM1L* or *NDUFS6* (Tables S13 and S14). Besides, two meQTLs of cg18468235 (rs2298831-C and rs17121881-T) were predicted as the detrimental mutations for *JAML* (Table S15). Most of the CpGs in 5p15.33 were correlated with the expression of *CLPTM1L* and *TERT*, of which *TERT* is a known driver gene for cancer (Table S16). Finally, three methylation-related genes of cg18468235, cg08285415, and cg05012158 (*JAML, IREB2*, and *PSMA4*) were shown the consistent associations directions across CpG, gene expression, and lung cancer (Table S17).

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. The expression levels of 55 genes were significantly differential between lung tumor and adjacent normal tissues (FDR-corrected $p \leq .05$) (Table S18). Then, we integrated all associations to estimate whether the DNA methylation at CpGs could affect the development of NSCLC through regulating the gene expression. There were 12 CpGs and 34 genes having the potential CpG gene–NSCLC regulatory pathways (Table S19). For example, cg11624060 (5p15.33) with a decreased NSCLC risk (Z score = −12.20, $p = 3.01 \times 10^{-34}$) was negatively associated with expression of *TERT* (Rho = −0.34, $p = 1.05 \times 10^{-22}$), *TRIP13* (Rho = −0.34, $p = 4.24 \times 10^{-23}$), and *MRPL36* (Rho = −0.36, $p = 3.89 \times 10^{-26}$). Meanwhile, these genes were respectively upregulated in 93.52% ($p = 8.47 \times 10^{-31}$), 93.52% ($p = 4.83 \times 10^{-27}$), and 90.74% ($p =$

2.71×10^{-23}) tumor-adjacent tissues pairs, constructing a potential closed loop of regulatory pathway. 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).

DISCUSSION

In this study, we initially observed 39 statistically significant CpGs and 16 of them, which were mainly located in six lung cancer susceptibility loci from previous GWASs[4,6] except for cg08285415 (15q24.3), passed the downstream validation. Given that 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. The ethnic characteristics of the distribution of methylation markers in two populations were consistent with the differences in genetics as well. Moreover, the results of two populations hinted the CpGs in 5p15.33 might be shared markers between European and Asian populations. We then retrieved 16 replicated CpGs in lung cancer risk–related publications from the EWAS Atlas,[28] and found cg22795331 in 6q22.1, cg05012158 and cg06752398 in 15q25.1, and cg19720302 in 17q24.2 are located in novel methylation regions not reported by previous methylation studies. Besides, hypomethylation at cg22795331 and cg18468235 was observed in colorectal cancer[29] and papillary thyroid carcinoma,[30] indicating a potential methylation phenomenon of multi-cancer risk.

By integrating the multi-omics results across DNA methylation, gene expression, and NSCLC, we revealed some pathways with consistent directions of association, which might be useful to expound the potential regulation mechanism. In 5p15.33 locus, *TERT*, 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 upregulation of telomerase, contributing to cancer initiation.[31] The *TRIP13* gene promotes proliferation and invasion of lung cancer cells through AKT/mTORC1/c-Myc signaling,[32] Wnt signaling, and epithelial-mesenchymal transition pathways.[33] Some researchers observed silencing *NDUFS6* significantly decreased reactive oxygen species levels in breast cancer, inhibiting the cancer-associated inflammation response.[34] Furthermore, mitochondrial ribosomal protein L36 (*MRPL36*) is essential for maintaining mitochondrial functions and significantly increases in lung squamous cell carcinoma compared with normal lung tissue,[35] playing a crucial role in energy metabolism for human cancer.[36]

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.[37] As a crucial component of epithelial gammadelta T-cell biology, *JAML* also has broader implications in tissue homeostasis and repair.[38] Protein encoding by *IL10RA* is a receptor for interleukin-10 and has been shown to mediate the immunosuppressive signal of interleukin-10, and inhibits the synthesis of proinflammatory cytokines, which may restrain lung adenocarcinoma aggressiveness.[39] In 17q24.2, overexpression of *KPNA2* flanking cg19720302 was observed in various cancers, including lung cancer.[40] It has been shown to participate in cell differentiation, proliferation, apoptosis, and immune response, and thus promote tumor formation and progression.[40] Although most of the evidence from previous functional experimental studies supported the regulation pathways we identified, there still were some inconsistencies without results. For example, *LPCAT1* (5p15.33) was reported upregulated in LUAD tissues and cell lines and promoted brain metastasis.[41]

In subgroup analyses by smoking status, we observed a significant association heterogeneity between smokers and nonsmokers 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 upregulation of the *CHRNA5* gene in NSCLC tumor tissue[42,43] and low levels of *CHRNA5* mRNA were associated with lower risk for nicotine dependence and lung cancer,[44] 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.[26,45] *PSMA4* is an important component of the 20S core proteasome complex and related to tobacco addiction (recorded in GeneCards: https://www.genecards.org/). To our knowledge, chemicals in tobacco smoke, such as Benzo[a]pyrene and N-nitrosamines, lead to DNA damage, oxidative stress, and inflammation, and increase the likelihood of lung cancer.[46,47] 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.[48] In addition, *PSMA4* has been also considered as a strong candidate mediator of lung cancer cell growth and directly affects lung cancer susceptibility through its modulation of cell proliferation and apoptosis.[48]

Considering that DNA methylation changes usually occur in the early stages of the disease and precedes pathological or imaging detection, methylation markers, as a noninvasive diagnostic tool, have a promising potential in clinical translation of lung cancer. For example, previous study observed an 8% improvement in discrimination of lung cancer by adding 6 CpGs into conventional risk prediction models.[9] Similarly, methylation changes at candidate genes could initially identify the highest risk smokers for computed tomography screening for early detection of lung cancer,[49] as well as help the detection of lung cancer and differentiation of nonmalignant diseases.[50] These evidence hint that by integrating traditional risk factors, genetic variation, methylation changes, and other biomarkers of multi-omics, prediction models with high performance will be developed to identify potential high-risk populations and for early detection. Additionally, the methylation-related genes that we identified are also worthy of further investigation to search the potentially druggable targets and develop a novel targeted therapy.

This is the first study to identify the genetically predicted DNA methylation markers associated with NSCLC risk. To some degree, predicted models constructed by genetic instruments can control the selection bias, potential confounding, and reverse causation in traditional observational studies. Moreover, this approach has proved that the results were improved, compared with the single-meQTL SNP approach.[14] However, some limitations must be acknowledged. Through meta-analysis, we identified the shared CpGs in the two populations, but the ethnic heterogeneity of model application could not be completely ignored in this study. Although we adopted a strategy of upstream filter plus downstream multi-validation to control the effect of racial bias, we still should carefully draw that conclusion, and a further ethnicity-specific study is necessary to validate our findings. Second, the subjects used in the validation stage from FLCCA were only nonsmoker females, lacking the necessary samples of smokers and males. Furthermore, although most of the potential regulatory pathways can be supported by experimental or biological evidence, the findings are only data-driven evidence and 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 systematically assessed the associations of genetically predicted DNA methylation CpGs with NSCLC risk, and a total of 16 CpG sites were identified, including four novel CpGs. Our findings indicated that these CpGs are likely to affect the NSCLC risk via regulating the flanking genes related to cancer formation and development. The findings of this study may contribute to the understanding of the epigenetic susceptibility mechanisms of NSCLC risk, especially for the interplay of genetics and epigenetics.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We would like to express our sincere thanks to 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 Transdisciplinary Research in Cancer of the Lung and The International Lung Cancer Consortium, who shared the genome-wide association study data from European populations. We also appreciate the Biobank Japan Project for their work on the sharing of GWAS data from the Asian population. Funding support from the National Natural Science of China (82192904, 81820108028, 81922061, 81521004, 81973123, and 82072579), the Science Fund for Distinguished Young Scholars of Jiangsu Province (BK20211533), Research Unit of Prospective Cohort of Cardiovascular Diseases and Cancer, Chinese Academy of Medical Sciences (2019RU038), and National Key Research and Development Project (2019YFC1315704).

DATA AVAILABILITY STATEMENT

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.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, 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):209–249. doi:10.3322/caac.21660 [PubMed: 33538338]
2. Chen W, Zheng R, Baade PD, et al. Cancer statistics in China, 2015. CA Cancer J Clin. 2016;66(2):115–132. doi:10.3322/caac.21338 [PubMed: 26808342]
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. doi:10.1378/chest.12-2345 [PubMed: 23649439]
4. Bossé Y, Amos CI A decade of GWAS results in lung cancer. Cancer Epidemiol Biomarkers Prev. 2018;27(4):363–379. doi:10.1158/1055-9965.epi-16-0794 [PubMed: 28615365]

5. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Not Genet. 2017;49(7):1126–1132. doi:10.1038/ng.3892

6. Dai J, Lv J, Zhu M, 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):881–891. doi:10.1016/s2213-2600(19)30144-4 [PubMed: 31326317]

7. Duruisseaux M, Esteller M. Lung cancer epigenetics: from knowledge to applications. Semin Cancer Biol. 2018;51:116–128. doi:10.1016/j.semcancer.2017.09.005 [PubMed: 28919484]

8. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Not Commun. 2015;6(1):10192. doi:10.1038/ncomms10192

9. Baglietto L, Ponzi E, Haycock P, 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. doi:10.1002/ijc.30431 [PubMed: 27632354]

10. Sun YQ, Richmond RC, Suderman M, 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):1482–1497. doi:10.1093/ije/dyab044 [PubMed: 33729499]

11. Battram T, Richmond RC, Baglietto L, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. Int J Epidemiol. 2019;48(5):1493–1504. doi:10.1093/ije/dyz190 [PubMed: 31549173]

12. Gaunt TR, Shihab HA, Hemani G, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17(1):61. doi:10.1186/s13059-016-0926-z [PubMed: 27036880]

13. Morrow JD, Glass K, Cho MH, 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):1275–1284. doi:10.1164/rccm.201707-1434oc [PubMed: 29313708]

14. Yang Y, Wu L, Shu XO, 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. doi:10.1093/jnci/djz109 [PubMed: 31143935]

15. Yang Y, Wu L, Shu X, 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):505–517. doi:10.1158/0008-5472.can-18-2726 [PubMed: 30559148]

16. Wu L, Yang Y, Guo X, et al. An integrative multi-omics analysis to identify candidate DNA methylation biomarkers related to prostate cancer risk. Not Commun. 2020;11(1):3905. doi:10.1038/s41467-020-17673-9

17. Zhu J, Yang Y, Kisiel JB, et al. Integrating genome and methylome data to identify candidate DNA methylation biomarkers for pancreatic cancer risk. Cancer Epidemiol Biomarkers Prev. 2021;30(11):2079–2087. doi:10.1158/1055-9965.epi-21-0400 [PubMed: 34497089]

18. Wang Y, Wu W, Zhu M, 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):1602–1610. doi:10.1002/ijc.31182 [PubMed: 29193083]

19. Huan T, Joehanes R, Song C, et al. Genome-wide identification of DNA methylation QTLs in whole blood highlights pathways for cardiovascular disease. Not Commun. 2019;10(1):4267. doi:10.1038/s41467-019-12228-z

20. Wu L, Shi W, Long J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Not Genet. 2018;50(7):968–978. doi:10.1038/s41588-018-0132-x

21. Chen YA, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium Human-Methylation450 microarray. Epigenetics. 2013;8(2):203–209. doi:10.4161/epi.23470 [PubMed: 23314698]

22. Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Not Commun. 2018;9(1):1825. doi:10.1038/s41467-018-03621-1

23. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Not Genet. 2012;44(4):369–375, s1-3. doi:10.1038/ng.2213

24. Ishigaki K, Akiyama M, Kanai M, et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. Not Genet. 2020;52(7):669–679. doi:10.1038/s41588-020-0640-3

25. Qin N, Li Y, Wang C, et al. Comprehensive functional annotation of susceptibility variants identifies genetic heterogeneity between lung adenocarcinoma and squamous cell carcinoma. Front Med. 2021;15(2):275–291. doi:10.1007/s11684-020-0779-4 [PubMed: 32889700]

26. Bossé Y, Li Z, Xia J, et al. Transcriptome-wide association study reveals candidate causal genes for lung cancer. Int J Cancer. 2020;146(7):1862–1878. doi:10.1002/ijc.32771 [PubMed: 31696517]

27. Zhu M, Fan J, Zhang C, 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):1666–1676. doi:10.1093/hmg/ddab119 [PubMed: 33909040]

28. Li M, Zou D, Li Z, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. Nucleic Acids Res. 2019;47(D1):D983–D988. doi:10.1093/nar/gky1027 [PubMed: 30364969]

29. Zhu L, Yan F, Wang Z, 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):2488–2498. doi:10.1002/ijc.31765 [PubMed: 30183087]

30. Beltrami CM, Dos Reis MB, Barros-Filho MC, et al. Integrated data analysis reveals potential drivers and pathways disrupted by DNA methylation in papillary thyroid carcinomas. Clin Epigenet. 2017;9(1):45. doi:10.1186/s13148-017-0346-2

31. Kim NW, Piatyszek MA, Prowse KR, et al. Specific association of human telomerase activity with immortal cells and cancer. Science. 1994;266(5193):2011–2015. doi:10.1126/science.7605428 [PubMed: 7605428]

32. 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):237–248. doi:10.3233/cbm-200039 [PubMed: 33136091]

33. Li ZH, Lei L, Fei LR, 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. doi:10.1007/s10735-020-09919-z [PubMed: 33128167]

34. Mao W, Xiong G, Wu Y, et al. RORα suppresses cancer-associated inflammation by repressing respiratory complex i-dependent ROS generation. Int J Mol Sci. 2021;22(19):10665. doi:10.3390/ijms221910665 [PubMed: 34639006]

35. 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. doi:10.3389/fgene.2020.00185 [PubMed: 32194639]

36. Piao L, Li Y, Kim SJ, et al. Association of LETM1 and MRPL36 contributes to the regulation of mitochondrial ATP production and necrotic cell death. Cancer Res. 2009;69(8):3397–3404. doi:10.1158/0008-5472.can-08-3235 [PubMed: 19318571]

37. 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. [PubMed: 35034089]

38. Witherden DA, Verdino P, Rieder SE, et al. The junctional adhesion molecule JAML is a costimulatory receptor for epithelial gammadelta T cell activation. Science. 2010;329(5996):1205–1210. doi:10.1126/science.1192698 [PubMed: 20813954]

39. Sung WW, Wang YC, Lin PL, et al. IL-10 promotes tumor aggressiveness via upregulation of CIP2A transcription in lung adenocarcinoma. Clin Cancer Res. 2013;19(15):4092–4103. doi:10.1158/1078-0432.ccr-12-3439 [PubMed: 23743567]

40. Han Y, Wang X. The emerging roles of KPNA2 in cancer. Life Sci. 2020;241:117140. doi:10.1016/j.lfs.2019.117140 [PubMed: 31812670]

41. Wei C, Dong X, Lu H, 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. doi:10.1186/s13046-019-1092-4 [PubMed: 30791942]

42. Falvella FS, Galvan A, Frullanti E, et al. Transcription deregulation at the 15q25 locus in association with lung adenocarcinoma risk. Clin Cancer Res. 2009;15(5):1837–1842. doi:10.1158/1078-0432.ccr-08-2107 [PubMed: 19223495]

43. Song P, Sekhon HS, Fu XW, et al. Activated cholinergic signaling provides a target in squamous cell lung carcinoma. Cancer Res. 2008;68(12):4693–4700. doi:10.1158/0008-5472.can-08-0183 [PubMed: 18559515]

44. Wang JC, Cruchaga C, Saccone NL, 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):3125–3135. doi:10.1093/hmg/ddp231 [PubMed: 19443489]

45. Yao C, Joehanes R, Wilson R, 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 Epigenet. 2021;13(1):60. doi:10.1186/s13148-021-01041-5

46. 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):666–669. doi:10.1165/rcmb.2006-0047sf [PubMed: 16484683]

47. Hecht SS. DNA adduct formation from tobacco-specific N-nitrosamines. Mutat Res. 1999;424(1-2):127–142. doi:10.1016/s0027-5107(99)00014-7 [PubMed: 10064856]

48. Liu Y, Liu P, Wen W, et al. Haplotype and cell proliferation analyses of candidate lung cancer susceptibility genes on chromosome 15q24-25.1. Cancer Res. 2009;69(19):7844–7850. doi:10.1158/0008-5472.can-09-1833 [PubMed: 19789337]

49. Leng S, Do K, Yingling CM, et al. Defining a gene promoter methylation signature in sputum for lung cancer risk assessment. Clin Cancer Res. 2012;18(12):3387–3395. doi:10.1158/1078-0432.ccr-11-3049 [PubMed: 22510351]

50. Weiss G, Schlegel A, Kottwitz D, König T, Tetzner R. Validation of the SHOX2/PTGER4 DNA methylation marker panel for plasma-based discrimination between patients with malignant and nonmalignant lung disease. J Thorac Oncol. 2017;12(1):77–84. doi:10.1016/j.jtho.2016.08.123 [PubMed: 27544059]

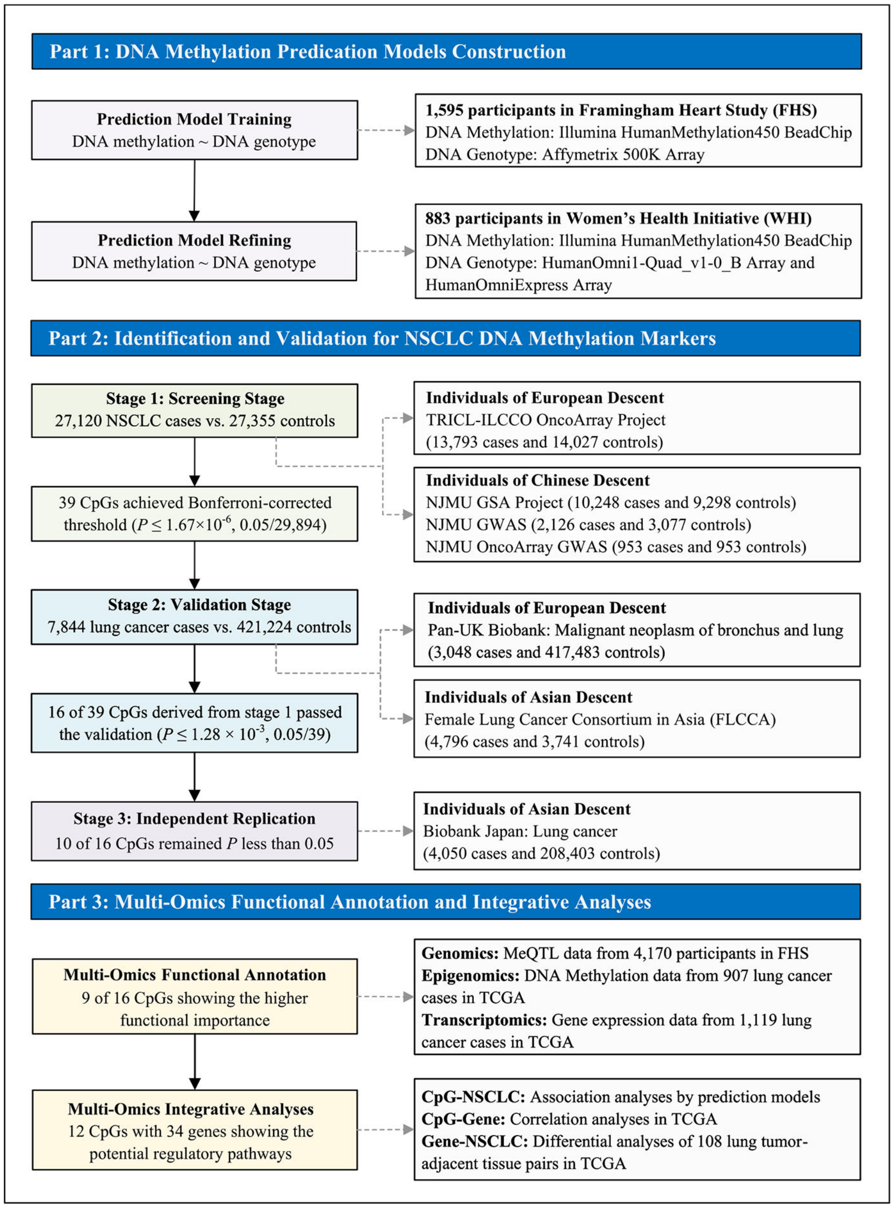

Part 1: DNA Methylation Predication Models Construction
Prediction Model Training
DNA methylation ~ DNA genotype
1,595 participants in Framingham Heart Study (FHS)
DNA Methylation: Illumina HumanMethylation450 BeadChip
DNA Genotype: Affymetrix 500K Array
Prediction Model Refining
DNA methylation ~ DNA genotype
883 participants in Women's Health Initiative (WHI)
DNA Methylation: Illumina HumanMethylation450 BeadChip
DNA Genotype: HumanOmni1-Quad_v1-0_B Array and HumanOmniExpress Array
Part 2: Identification and Validation for NSCLC DNA Methylation Markers
Stage 1: Screening Stage
27,120 NSCLC cases vs. 27,355 controls
Individuals of European Descent
TRICL-ILCCO OncoArray Project
(13,793 cases and 14,027 controls)
39 CpGs achieved Bonferroni-corrected threshold ($P \le 1.67\times10^{-6}$, 0.05/29,894)
Individuals of Chinese Descent
NJMU GSA Project (10,248 cases and 9,298 controls)
NJMU GWAS (2,126 cases and 3,077 controls)
NJMU OncoArray GWAS (953 cases and 953 controls)
Stage 2: Validation Stage
7,844 lung cancer cases vs. 421,224 controls
Individuals of European Descent
Pan-UK Biobank: Malignant neoplasm of bronchus and lung
(3,048 cases and 417,483 controls)
16 of 39 CpGs derived from stage 1 passed the validation ($P \le 1.28 \times 10^{-3}$, 0.05/39)
Individuals of Asian Descent
Female Lung Cancer Consortium in Asia (FLCCA)
(4,796 cases and 3,741 controls)
Stage 3: Independent Replication
10 of 16 CpGs remained P less than 0.05
Individuals of Asian Descent
Biobank Japan: Lung cancer
(4,050 cases and 208,403 controls)
Part 3: Multi-Omics Functional Annotation and Integrative Analyses
Multi-Omics Functional Annotation
9 of 16 CpGs showing the higher functional importance
Genomics: MeQTL data from 4,170 participants in FHS
Epigenomics: DNA Methylation data from 907 lung cancer cases in TCGA
Transcriptomics: Gene expression data from 1,119 lung cancer cases in TCGA
Multi-Omics Integrative Analyses
12 CpGs with 34 genes showing the potential regulatory pathways
CpG-NSCLC: Association analyses by prediction models
CpG-Gene: Correlation analyses in TCGA
Gene-NSCLC: Differential analyses of 108 lung tumor-adjacent tissue pairs in TCGA

FIGURE 1.

Flowchart for the study design.

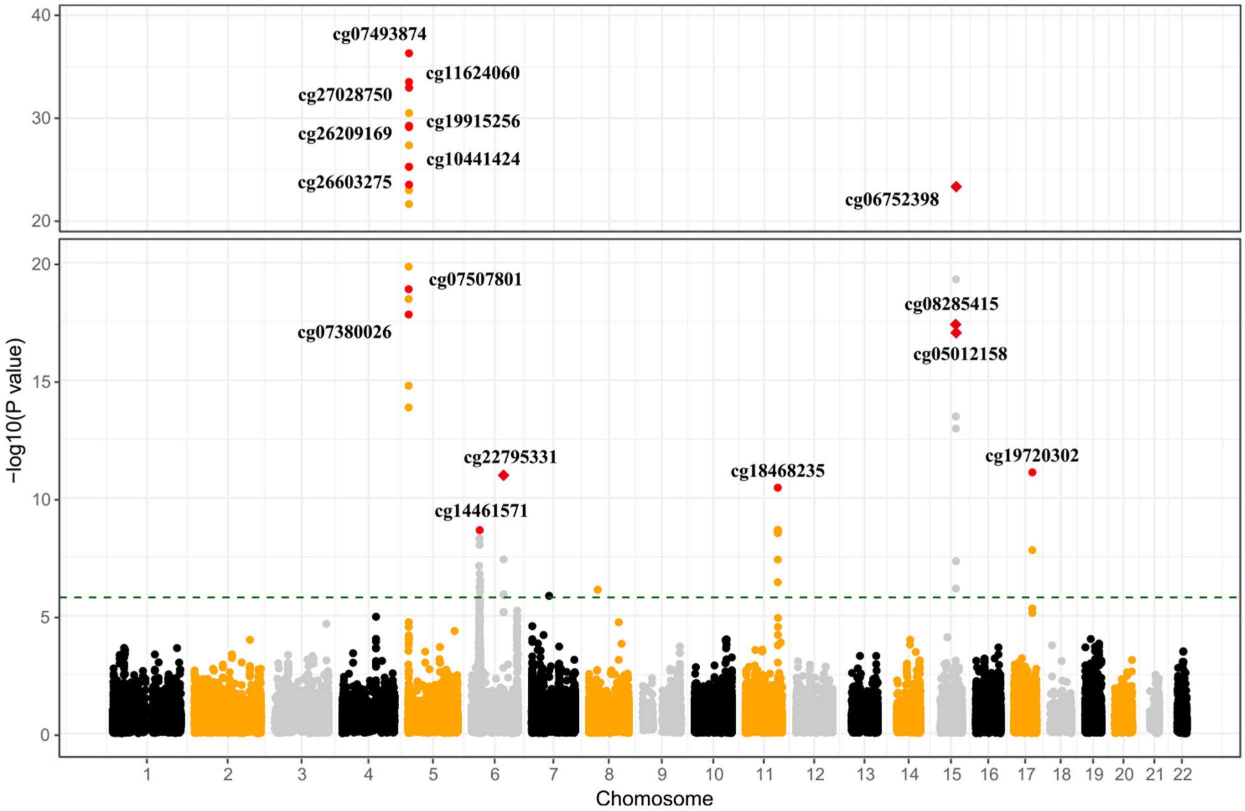
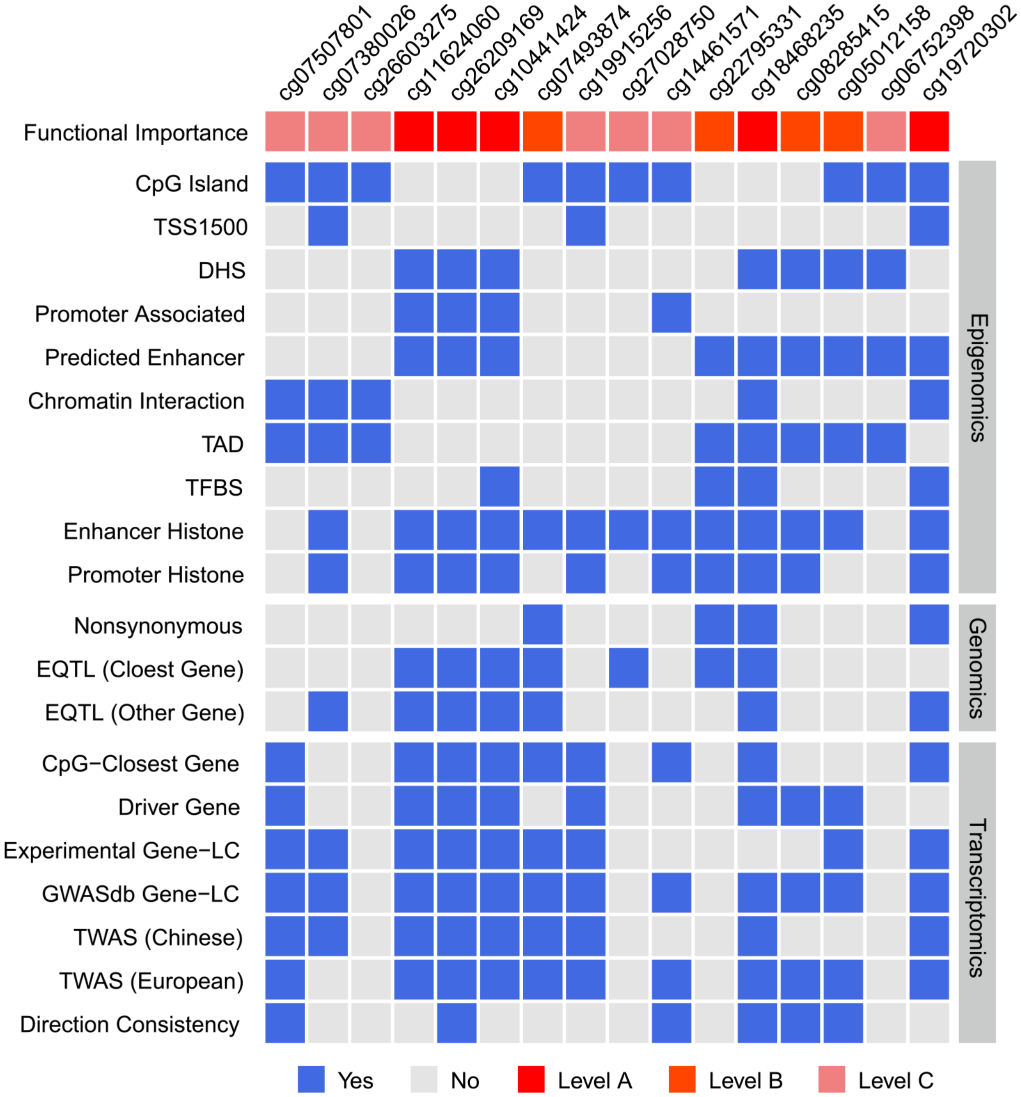


FIGURE 2.

Manhattan plot for 39 DNA methylation markers from meta-analysis associated with NSCLC risk. The green dotted line represents $p = 1.67 \times 10^{-6}$ (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 epigenome-wide association studies. NSCLC indicates non–small cell lung cancer.



		cg07507801	cg07380026	cg26603275	cg11624060	cg26209169	cg10441424	cg07493874	cg19915256	cg27028750	cg14461571	cg22795331	cg18468235	cg08285415	cg05012158	cg06752398	cg19720302
	Functional Importance	Level C	Level C	Level C	Level A	Level A	Level A	Level B	Level C	Level C	Level C	Level B	Level A	Level B	Level B	Level C	Level A
Epigenomics	CpG Island	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes	Yes
	TSS1500	No	Yes	No	No	No	No	No	Yes	No	No	No	No	No	No	No	Yes
	DHS	No	No	No	Yes	Yes	Yes	No	No	No	No	No	Yes	Yes	Yes	Yes	No
	Promoter Associated	No	No	No	Yes	Yes	Yes	No	No	No	Yes	No	No	No	No	No	No
	Predicted Enhancer	No	No	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	Yes	Yes	Yes	Yes
	Chromatin Interaction	Yes	Yes	Yes	No	No	No	No	No	No	No	No	Yes	No	No	No	Yes
	TAD	Yes	Yes	Yes	No	No	No	No	No	No	No	Yes	Yes	Yes	Yes	Yes	No
	TFBS	No	No	No	No	No	Yes	No	No	No	No	Yes	Yes	No	No	No	Yes
	Enhancer Histone	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
	Promoter Histone	No	Yes	No	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	Yes	No	No	Yes
Genomics	Nonsynonymous	No	No	No	No	No	No	Yes	No	No	No	Yes	Yes	No	No	No	Yes
	EQTL (Cloest Gene)	No	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	No	No	No	No
	EQTL (Other Gene)	No	Yes	No	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	No	No	Yes
Transcriptomics	CpG−Closest Gene	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	No	No	No	Yes
	Driver Gene	Yes	No	No	Yes	Yes	Yes	No	Yes	No	No	No	Yes	Yes	Yes	No	No
	Experimental Gene−LC	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	Yes	No	Yes
	GWASdb Gene−LC	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	Yes
	TWAS (Chinese)	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	No	No	No	Yes	No	No	No	Yes
	TWAS (European)	Yes	No	No	Yes	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes	Yes	No	Yes
	Direction Consistency	Yes	No	No	No	Yes	No	No	No	No	Yes	No	Yes	Yes	Yes	No	No

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. DHS indicates DNase I hypersensitivity sites; LC, lung cancer; TAD, topologically associated domains; TF, transcription factor; TSS1500, transcription start site upstream 1500 bp; TWAS, transcriptome-wide association study.

TABLE 1

Thirty-nine DNA methylation markers from meta-analysis associated with NSCLC risk.

					Screening stage		Validation stage		Combined stage			
CpG	CytoBand[a]	Position[b]	Classification	Closest gene	Z score	p [c]	Z score	p [c]	Z score	p	I^2 (%)	p Het[d]
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

					Screening stage		Validation stage		Combined stage			
CpG	CytoBand[a]	Position[b]	Classification	Closest gene	Z score	p [c]	Z score	p [c]	Z score	p	I^2 (%)	p Het[d]
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 (I^2; heterogeneity *p* value), and $p \leq .05$ 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]

				CpG vs. NSCLC risk			CpG vs. gene expression			Gene expression vs. NSCLC risk		
Chr	Position	CpG	Gene	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

				CpG vs. NSCLC risk			CpG vs. gene expression			Gene expression vs. NSCLC risk		
Chr	Position	CpG	Gene	Z score[b]	*p* value[b]	Direction	Rho[c]	*p* value. FDR[c]	Direction	Up regulation[d]	*p* value. FDR[d]	Direction
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

Note: FDR-corrected *p* value was calculated by Benjamini-Hochberg method and $p \leq 0.05$ was statistically significant. All statistical tests are two-sided. Abbreviations: Chr, chromsome; FDR, false discovery rate; NSCLC, non–small cell lung cancer.

[a] For CpGs with the high functional level in multi-omics annotation, the integrative results of top five differential genes were selected. The complete list is shown in Table S17.

[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 upregulation 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.