Original Article

Biomarkers for non-small cell lung cancer risk using multi-omics approaches: a nested case-control study

Youngmin Han[1,2]^, Keum Ji Jung[1,2], Seong Gyu Choi[1], Yeun Soo Yang[1,2], Kwangbae Lee[3], Sun Ha Jee[1,4]^

[1]Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Republic of Korea; [2]Department of Epidemiology and Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Republic of Korea; [3]Korea Medical Institute, Seoul, Republic of Korea; [4]Graduate School of Transdisciplinary Health Sciences, Yonsei University, Seoul, Republic of Korea

Contributions: (I) Conception and design: Y Han, KJ Jung, SH Jee; (II) Administrative support: SH Jee, K Lee; (III) Provision of study materials or patients: SH Jee; (IV) Collection and assembly of data: Y Han, KJ Jung, YS Yang; (V) Data analysis and interpretation: Y Han, SG Choi; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Sun Ha Jee, PhD. Graduate School of Transdisciplinary Health Sciences, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea; Institute for Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Republic of Korea. Email: jsunha@yuhs.ac.

Background: Lung cancer poses a major public health challenge, accounting for the highest cancer-related mortality worldwide. This study aimed to identify non-invasive biomarkers for the early detection of non-small cell lung cancer (NSCLC) risk.

Methods: We randomly selected 150 incident NSCLC cases during follow-up from the Korean Cancer Prevention Study-II. Controls (n=150) were matched to cases by age, gender, and the time of blood collection. Non-targeted metabolite screening by ultra-high-performance liquid chromatography (UHPLC)/mass spectrometry (MS) was conducted on the pre-diagnostic biological samples. The 11 reported lung cancer-associated single-nucleotide polymorphisms (SNPs) in Koreans were extracted from DNA genotyping data of the study population. Metabolite markers related to NSCLC risk were identified through clustering using hierarchical density-based spatial clustering of applications with noise. The associations between smoking, dietary factors, and NSCLC were also examined.

Results: Six discriminative serum metabolites were identified as having an association with NSCLC incidence. Notably, the relationship between specific metabolite levels and NSCLC risk differed by rs7086803 genotype. Smoking status and occupational exposures appear to influence specific metabolite profiles, while dietary vegetable intake may modulate the risk of NSCLC among smokers.

Conclusions: The meaningful biomarkers revealed in the current research could be used to enhance the predictive ability for NSCLC risk. Furthermore, we suggest that the protective role of dietary vegetables against NSCLC may be attenuated or absent in smokers.

Keywords: Non-small cell lung cancer (NSCLC); hierarchical density-based spatial clustering of applications with noise (HDBSCAN); predictive biomarkers; multi-omics; dietary vegetables

Submitted May 20, 2025. Accepted for publication Jul 31, 2025. Published online Sep 25, 2025.
doi: 10.21037/tlcr-2025-603
View this article at: https://dx.doi.org/10.21037/tlcr-2025-603

^ ORCID: Youngmin Han, 0000-0001-5517-3396; Sun Ha Jee, 0000-0001-9519-3068.

Introduction

Lung cancer is a leading cause of death worldwide, resulting in significant public health concerns. The Global Cancer Observatory reported approximately 2.5 million new cases (12.4% of total new cases) and 1.8 million deaths (18.7% of all cancer deaths) from lung cancer in 2022 (1). The results are in line with the 2020 statistics, where lung cancer ranks as the top cause of cancer mortality (1). In Republic of Korea, lung cancer is a predominant cancer, accounting for the highest cancer-related mortality, with a reported mortality rate of 35 per 100,000 in 2022 (2).

Highlight box

Key findings

- Six serum metabolites were identified as significant biomarkers for the non-small cell lung cancer (NSCLC) risk.
- The association between metabolite levels and NSCLC risk was influenced by the genotype of rs7086803, highlighting a gene-metabolite interaction.
- Dietary vegetable intake showed a significant correlation with metabolite levels and appeared to modify NSCLC risk, especially among smokers.

What is known and what is new?

- Non-invasive biomarkers can aid in the early diagnosis of NSCLC, but reliable predictive markers are still under investigation.
- Smoking is a well-established major risk factor for NSCLC, and diet, especially vegetable intake, has been suggested to have a protective role.
- Six serum metabolites linked to NSCLC risk through untargeted metabolomics in a Korean cohort based on hierarchical density-based spatial clustering.
- Genotype-dependent associations between metabolites and NSCLC risk were observed.
- Evidence that dietary vegetables influence metabolite profiles associated with NSCLC risk, particularly modulating risk in smokers.

What is the implication, and what should change now?

- These metabolite biomarkers could be integrated into risk prediction models for earlier, non-invasive detection of NSCLC, improving screening and preventive strategies.
- Personalized risk assessments considering genetic background (e.g., rs7086803 genotype) and lifestyle factors like diet and smoking should be developed for more effective NSCLC prevention.
- Public health messages and interventions promoting vegetable intake may need to consider smoking status, as smokers may not receive the same protective benefits.
- Further studies are warranted to validate these biomarkers and to understand the mechanisms behind gene-diet interactions in NSCLC risk.

Non-small cell lung cancer (NSCLC) is the most common form of lung cancer, accounting for approximately 85% of all cases (3). One of the defining features of NSCLC is its therapeutic diversity: a wide range of molecular targeted therapies, such as epidermal growth factor receptor (EGFR) tyrosine kinase inhibitors and anaplastic lymphoma kinase (ALK) inhibitors, are available for patients with specific genetic alterations (4). Moreover, immune checkpoint inhibitors have demonstrated significant clinical efficacy in many NSCLC subtypes, particularly in those with high programmed death ligand-1 (PD-L1) expression (5). Nevertheless, NSCLC is often diagnosed at a later stage due to a lack of early symptoms, which limits curative treatment options (6).

Multi-omics strategy integrating genomics, metabolomics, transcriptomics, and microbiome has been emphasizing the value of discovering promising biomarkers based on understanding the complex biological mechanisms of disease (7,8). Such combined techniques hold great promise for enhancing detection accuracy and implementing precision medicine. An approach comprehensively analyzing tissue- and serum metabolites from NSCLC patients alongside 16S ribosomal RNA sequencing for the gut microbiome revealed a dysregulated metabolic axis associated with NSCLC pathogenesis (9). Another omics research confirmed that key metabolites and genes contributing to hemostasis, angiogenesis, and cell proliferation were predominant in both subtypes of NSCLC, suggesting adenosine diphosphate as a potential therapeutic target for lung cancer metastasis (10). Diagnostic accuracy was remarkably improved by combining multi-omics data, including cell-free DNA liquid biopsy markers by next-generation sequencing with machine learning techniques, showcasing a promising early lung cancer detection approach in a recent Korean study (11). However, studies applying multi-omics technology to pre-disease samples other than lung cancer patients are still limited.

Machine learning approaches offer powerful tools for optimizing high-dimensional multi-omics data. In particular, Uniform Manifold Approximation and Projection (UMAP), a non-linear dimensionality reduction technique that preserves both local and global data structures. It has demonstrated superior performance over principal component analysis and t-distributed Stochastic Neighborhood Embedding (t-SNE) in resolving NSCLC subtypes in transcriptomic studies (12,13). Furthermore, hierarchical density-based spatial clustering of applications with noise (HDBSCAN), a density-based clustering

algorithm that detects clusters of varying shapes and densities without requiring predefined cluster numbers, has shown strong potential. Cook *et al.* (14) reported that it effectively identified irregularly shaped clusters and distinguished subtle biological signals in NSCLC datasets.

Here, we designed a nested case-control study based on the Korean Cancer Prevention Study (KCPS)-II, utilizing multi-omics tools and machine learning approaches. Genotyping and non-targeted metabolite screening were conducted on the pre-diagnostic biological samples. A comprehensive understanding of the NSCLC pathogenesis is expected to facilitate the identification of relevant biomarkers for risk prediction. We anticipated that well-defined molecular architecture could improve diagnostic accuracy and advance a personalized medicine approach for NSCLC. We present this article in accordance with the REMARK reporting checklist (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-2025-603/rc).

Methods

Study population

Study subjects were drawn from the KCPS-II cohort. Briefly, KCPS-II enrollment began in April 2004 and involved participants from 18 health promotion centers across Republic of Korea. Data were collected from hospital records, mortality registries, and the National Cancer Center registry during the follow-up period. Written informed consent was obtained from all participants for inclusion in the cohort and for using their data and samples in future research. For the current research, individuals aged 30 to 70 years were randomly selected from the KCPS-II cohort, specifically those with serum stored in suitable condition for metabolomics (Figure S1). Participants who were cancer-free at enrollment but developed NSCLC during the follow-up period were defined as the case group [NSCLC incidence (n=150) *vs.* control (n=150)]. When classifying cancer incidence cases by subtype, there were 50 male cases of adenocarcinoma (ADN), 50 female cases of ADN, and 50 male cases of squamous cell carcinoma (SqCC). For these groups, controls who did not develop lung cancer during the follow-up period were matched using age, gender, and blood collection point at a 1:1 ratio. The study excluded female SqCC cases due to an insufficient number meeting the criteria. Of the total subjects, genotyping was completed for 292 participants, and only 60 individuals [NSCLC incidence (n=34) *vs.* control (n=26)] responded to the dietary questionnaire (Figure S1).

The sample size for this study was calculated using G*Power v 3.1.9.7 (Franz Faul). An effect size of 0.5, a statistical power of 0.95, a two-sided significance level of 0.05, and an allocation ratio of 1:1 were set. Based on these parameters, the minimum required sample size was approximately 130 subjects per group. Considering potential outliers and other practical aspects of the study, we included 150 subjects in each group to ensure adequate statistical power and reliability of the results.

The study was approved by the Institutional Review Board at the Yonsei University Health System (IRB No. 4-2022-1136) under the Declaration of Helsinki and its subsequent amendments.

Data collection

Clinical variables were obtained by anthropometric measurements and human specimens analysis in the hospital laboratory using a COBAS INTEGRA 800 and a 7600 Analyzer (Hitachi, Tokyo, Japan). Details on the examination methods are previously described (15).

Each participant answered a self-administered questionnaire regarding sociodemographic characteristics and health habits. Smoking status was collected in three categories (never-, ex-, or current smoker) and occupational factors were collected in nine groups: (I) professionals; (II) clerical and administrative staff; (III) professional-technical and administrative staff; (IV) sales and marketing workers; (V) service workers; (VI) production and manual labor workers; (VII) agricultural, forestry, and fishery workers; (VIII) homemakers; and (IX) others (Table S1). For analysis, some were grouped into broader categories: professionals, clerical and administrative staff, and professional-technical and administrative staff were combined as professional and office workers (group 1); sales and marketing workers and service workers as sales and service workers (group 2); and production and manual labor workers together with agricultural, forestry, and fishery workers as production and agricultural workers (group 3). Due to the limited sample sizes of production and manual labor workers (n=2), and considering that, although the nature of occupational environmental exposures may differ, agricultural, forestry, and fishery workers experience greater exposure levels compared to predominantly office-based groups, these categories were combined into group 3. Homemakers and others were retained as groups 4 and 5. Nutritional

data were obtained from a brief dietary questionnaire that assessed the intake frequency of 7 food groups (16). Intake amounts per day were estimated based on the list of food exchanges for Koreans (16). Some of these estimates were used in the present research.

Metabolome analysis

Non-targeted metabolomics using ultra-high-performance liquid chromatography (UHPLC)-mass spectrometry (MS)

Prepared pre-diagnostic serum samples were precipitated with cold acetonitrile (Wako Pure Chemical Industries, Osaka, Japan) (1:4, v/v) and centrifuged for 15 min (13,000 rpm, 4 ℃). Without heating, a separate supernatant was dried in a vacuum concentrator (HyperVAC-MAX, Hanil Scientific Inc., Gimpo, Republic of Korea). Next, 200 µL of 10% methanol (J.T. Baker® Chemicals; Avantor Performance Materials, Inc., Radnor, PA, USA) was added for reconstitution and filtered with a 0.45 µm polyvinylidene difluoride syringe filter. We used L-Leucine-1-^{13}C (Sigma-Aldrich, Saint Louis, MO, USA) as an internal standard (ISTD). The quality control (QC) sample was prepared following the exact steps by combining all serum samples.

Serum samples were analyzed using a Thermo UHPLC system (Ultimate 3000 BioRS; Dionex, Thermo Fisher Scientific, Bremen, Germany) equipped with an Acquity UHPLC-BEH-C18 column (Waters, Milford, MA, USA). The column temperature was set to 50 °C throughout the analysis. For the separation of compounds in samples, a gradient was created for 20 minutes with two mobile phases [A, 0.1% formic acid in liquid chromatography-mass spectrometry (LC-MS) grade water (Thermo Fisher Scientific, Fair Lawn, NJ, USA); B, 0.1% formic acid in LC-MS grade ACN (Thermo Fisher Scientific, Fair Lawn, NJ, USA)]. Orbitrap Exploris 240 (Thermo Fisher Scientific, Waltham, MA, USA) was combined with the UHPLC system for data detection. On MS, positive electrospray ionization mode (ESI +) with 3500 V of positive ion, 2500 V of negative ion, 50 (arbitrary units) of a flow rate of nitrogen sheath gas, and 10 (arbitrary units) of a flow rate of auxiliary gas was performed. Full scan-ddms2 mode was employed with a 70% replacement frequency (RF) lens, a standard automatic gain control (AGC) target, and 30% higher-energy collisional dissociation (HCD) collision energy. Data were collected within a scan range of 80–1,000 m/z.

QC samples were measured every 10th serum sample to monitor sensitivity and reproducibility, followed by blank measurements. Additionally, reliable analysis was performed using both intra-assay and inter-assay evaluations.

Identification of metabolites

Compound Discoverer 3.2 software (Thermo Fisher Scientific, San Jose, CA, USA) processed raw spectra, conducting alignment and normalization based on the spectra of QCs. Less than 80% of the features observed in all QC samples were eliminated. Processed features were identified based on online databases ChemSpider (http://www.chemspider.com), LIPID MAPS (https://www.lipidmaps.org), mzCloud (https://www.mzcloud.org), and Kyoto Encyclopedia of Genes and Genomes (KEGG; https://www.genome.jp/kegg).

Genotyping

DNA genotyping was conducted using the KORV1.0-96 Array (Affymetrix, Santa Clara, CA, USA) provided by the K-CHIP consortium, along with the Affymetrix Genomewide Human SNP Array 5.0 (Affymetrix Inc.). For QC, markers and individuals with a missing rate exceeding 5% were excluded. In addition, single-nucleotide polymorphisms (SNPs) with a minor allele frequency below 0.05 or significant deviation from Hardy-Weinberg equilibrium (P<1.0e–6) were filtered out. At last, information on 11 lung cancer-associated SNPs (rs4488809, rs2736100, rs9387478, rs3817963, rs2395185, rs2179920, rs7741164, rs72658409, rs7086803, rs11610143, rs7216064) reported in Koreans (17) was extracted from our genotyping data.

Statistical analysis

All statistical analyses were conducted by SPSS 26 (IBM Corp, Armonk, NY, USA) and R 4.1.3. We performed independent *t*-tests to evaluate the differences in clinical/biochemical variables between the two groups. The skewed variables were logarithmically transformed. Even after logarithmic transformation, continuous variables with a non-normal distribution were tested using a Mann-Whitney *U* test. A Chi-squared test was used to measure nominal variables. The data are expressed as the mean ± standard error, and two-tailed P values less than 0.05 were considered to indicate significance. The normalized metabolite data were extracted from Compound Discoverer 3.2 for multivariate analysis. Linear regression analysis was performed with metabolites as the outcome variable

and smoking status and occupational factors as predictors. In case of smoking status, the current-smoker group was considered as the reference group. When using occupational data, the reference group was professional and office workers (group 1), as they are considered to have lower exposure to environmental hazards compared to other occupational groups, with the largest sample size among the clearly defined groups. The moderation effect analysis was conducted with age adjustment to assess how nutritional factors affect NSCLC risk [independent variable (X), nutritional intake amount per day; moderator (M), smoking status; interaction term, X*M; dependent variable (Y), NSCLC incidence].

Dimensionality reduction with UMAP and clustering by HDBSCAN

Clinical data (age, gender, smoking status, and alcohol consumption status), genetic data, and metabolomic profiles were integrated into a unified dataset. Categorical variables were label-encoded, and continuous variables were Z-score normalized to address scale discrepancies. The dataset was partitioned into a training set (70%, n=210) and a test set (30%, n=90) for external validation. UMAP was trained on the training set to generate a two-dimensional embedding space, and then applied to transform the test set, ensuring consistent low-dimensional representations across both datasets.

HDBSCAN was applied to group similar data points from the UMAP-reduced data into clusters. The model was trained on the training set with parameters set to allow out-of-sample cluster assignments. For the test set, each sample was assigned to the most suitable cluster based on the density distribution learned from the training set, ensuring consistent subgroup identification across both sets.

The clustering was evaluated by comparing the assigned cluster labels with the ground-truth NSCLC incidence labels. A confusion matrix was constructed, and performance metrics, including accuracy, sensitivity, specificity, and precision, were calculated to assess the clustering's ability to distinguish between cancer and non-cancer samples.

Identification of metabolite biomarkers

To explore metabolic patterns associated with NSCLC, the mean value of each metabolite was calculated across the entire dataset and within each cluster. Metabolites showing significant deviations from the overall dataset mean in specific clusters were identified and analyzed to determine their association with NSCLC incidence.

Results

Characteristics of the study population

Table 1 summarizes the baseline clinical and biochemical characteristics of the total subjects.

The mean age of the NSCLC incidence group was 55.05 years, and that of the control group was 55.04 years, with no statistical difference. Some clinical variables showed significant differences between groups. The level of white blood cells was significantly higher in the NSCLC incidence group (P=0.02), whereas bilirubin was lower in the NSCLC incidence group compared to the control group (P<0.001). In addition, the proportion of current smokers was considerably higher in the NSCLC group than in the control group (P=0.02), and there was no significant difference in the occupational factors (Table S1).

Clustering and performance evaluation

Figure 1 presents the results of clustering performed using UMAP and HDBSCAN. Both training and test sets were classified into four distinct clusters. Each cluster exhibited notable differences in the NSCLC ratio (Table S2) and several metabolite levels.

Cluster 1 had the lowest NSCLC rate (2.27%) and was predominantly composed of healthy individuals, representing a low-risk group. Cluster 2, similar to Cluster 1, also primarily represented control subjects with only a few male patients with ADN (n=2 in the training set). In contrast, cluster 0 exclusively comprised subjects with NSCLC incidence, and cluster 3 was uniquely characterized by the classification of only men with ADN lung cancer.

The clustering results were subsequently evaluated against actual NSCLC diagnoses to assess predictive performance (Table S3). In the training set, 104 of 105 controls were correctly classified as controls, with only one false positive. Among 105 true NSCLC incidence cases, 101 were correctly identified as true positives, with four false negatives. In the test set, after excluding three outliers, evaluation of 87 subjects showed that all 43 controls were correctly predicted with no false positives. Of the 44 NSCLC incidence cases, 43 were correctly identified, with one false negative.

Performance metrics underscored the robustness of the

Table 1 Baseline characteristics of subjects

Characteristics	NSCLC incidence (n=150)	Control (n=150)	P
Age (years)	55.05±0.69	55.04±0.69	>0.99
Body mass index (kg/m^2)	23.82±0.23	24.28±0.20	0.53
Waist circumference (cm)	83.11±0.70	83.73±0.73	0.54
Systolic blood pressure (mmHg)	121.85±1.15	124.59±1.35	0.16
Diastolic blood pressure (mmHg)	77.52±0.92	77.2±0.95	0.95
Glucose (mg/dL)	96.48±1.70	96.31±1.59	0.91
Total cholesterol (mg/dL)	190.05±2.85	192.58±2.86	0.53
Triglyceride (mg/dL)	146.55±7.61	139.2±5.97	0.56
HDL-cholesterol (mg/dL)	51.08±1.16	51.18±0.98	0.24
LDL-cholesterol (mg/dL)	114.88±2.63	115.57±2.50	0.30
Albumin (g/dL)	4.51±0.03	4.55±0.02	0.13
White blood cell (10^3/μL)	6.43±0.17	5.78±0.14	0.02*
AST (IU/L)	24.24±1.49	24.99±1.53	0.96
ALT (IU/L)	24.65±1.04	24.58±1.21	1
GGT (IU/L)	43.59±4.30	36.77±3.04	0.91
ALP (IU/L)	118.7±5.53	114.6±4.62	0.48
Bilirubin (mg/dL)	0.81±0.03	0.94±0.03	<0.001***
Blood urea nitrogen (mg/dL)	14.3±0.30	14.47±0.32	0.58
Creatinine (mg/dL)	0.98±0.02	0.98±0.02	0.49
Uric acid (mg/dL)	5.36±0.12	5.55±0.13	0.27

Data are presented as mean ± SE. Continuous variables were tested by independent *t*-test. P was derived from an independent *t*-test (lung cancer occurrence *vs.* control). *, P<0.05; ***, P<0.001. ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; GGT, γ-glutamyltransferase; HDL, high-density lipoprotein; LDL, low-density lipoprotein; NSCLC, non-small cell lung cancer; SE, standard error.

clustering approach. In the training set, sensitivity was 0.962 [95% confidence interval (CI): 0.906–0.985], specificity was 0.990 (95% CI: 0.948–0.998), precision was 0.990 (95% CI: 0.948–0.998), and accuracy was 0.976 (95% CI: 0.945–0.990). In the test set, sensitivity was 0.977 (95% CI: 0.882–0.996), specificity was 1.000 (95% CI: 0.918–1.000), precision was 1.000 (95% CI: 0.918–1.000), and accuracy was 0.989 (95% CI: 0.938–0.998) (Table S3).

Comparison of metabolite profiles across clusters

A total of 2,654 features were analyzed, of which 494 metabolites were identified. Among these, the following metabolites showed statistically significant differences in specific clusters compared to the overall mean value of each metabolite across all subjects.

Notably, several metabolites were significantly elevated in cluster 0. Specifically, triethylene glycol monobutyl ether (TEGBE), hydroxypropionylcarnitine, tyramine glucuronide, indoline, and sphinganine showed markedly higher levels than the overall mean values (all P<0.001), as visualized in *Figure 1*. Conversely, cis-5-tetradecenoylcarnitine and L-acetylcarnitine were significantly lower in cluster 0 (P<0.001).

In cluster 3, TEGBE, cis-5-tetradecenoylcarnitine, Cer(d18:1/16:0), and tyramine glucuronide were all markedly higher (P<0.001), along with hydroxypropionylcarnitine (P=0.01) and L-carnitine (P=0.02), which also showed a significant elevation. On the other hand, Cer(d18:0/14:0) was significantly lower (P<0.001). Given that clusters 0 and 3

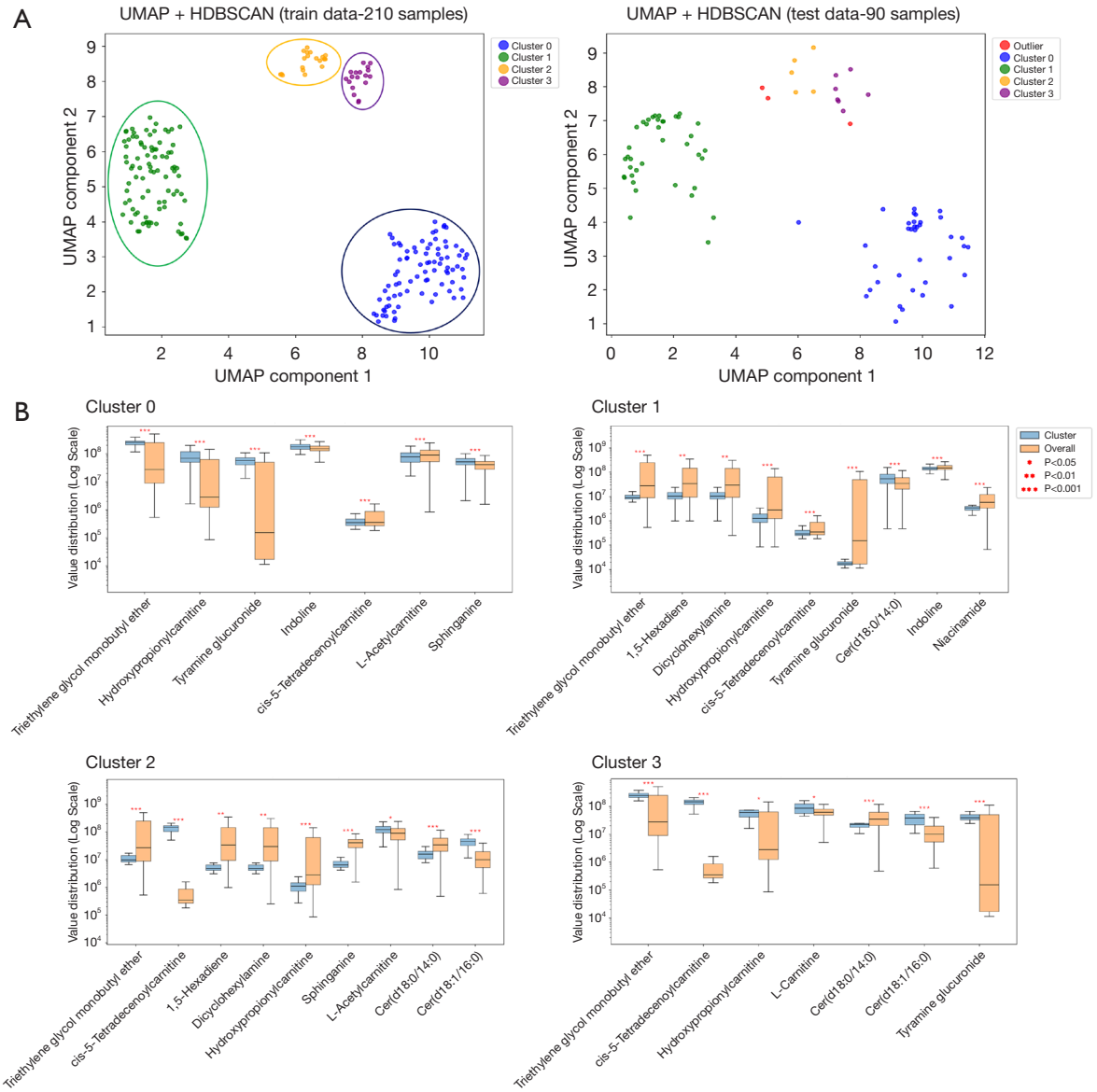

Figure 1 Clustering and comparison of metabolite profiles across clusters. (A) Clustering analysis was performed based on clinical data (age, gender, smoking, and alcohol consumption status), genetic data, and metabolomic profiles. The dataset was partitioned into a training set (70%, n=210) and a test set (30%, n=90) for external validation. (B) Metabolites with statistically significant differences based on *t*-tests comparing the overall mean (controls and NSCLC cases combined, orange bar) to the mean within each cluster (blue bar). Metabolite concentrations were log-transformed. *P<0.05, **P<0.01, ***P<0.001. HDBSCAN, hierarchical density-based spatial clustering of applications with noise; NSCLC, non-small cell lung cancer; UMAP, Uniform Manifold Approximation and Projection.

were composed solely of individuals who developed NSCLC, the exclusively listed metabolites in these clusters are likely to be strongly associated with cancer onset.

In cluster 1, most metabolites (TEGBE, hydroxypropionylcarnitine, cis-5-tetradecenoylcarnitine, tyramine glucuronide, indoline, and niacinamide) were significantly reduced compared to the overall mean value (P<0.001). Two structurally related compounds, 1,5-hexadiene and dicyclohexylamine, were also significantly decreased (P=0.003), suggesting consistent metabolic downregulation across related pathways. Only Cer(d18:0/14:0) was significantly elevated in this cluster (P<0.001).

Table 2 Frequency of 11 single-nucleotide polymorphisms in study participants

SNP	Location	Mapped gene	EA	EAF[a]	EAF[b]	P	P[†]
rs4488809	3:189638472	*TP63*	C	0.497	0.521	0.83	0.65
rs2736100	5:1286401	*TERT*	A	0.65	0.634	0.73	0.86
rs9387478	6:117465017	*DCBLD1*	C	0.563	0.722	0.17	0.50
rs3817963	6:32400310	*TSBP1-AS1*, *BTNL2*	T	0.7	0.75	0.07	0.54
rs2395185	6:32465390	*HLA-DRB9*	G	0.637	0.690	0.12	0.88
rs2179920	6:33091097	*HLA-DPB1*, *HLA-DPA2*	C	0.887	0.933	0.15	0.59
rs7741164	6:41525674	*FOXP4-AS1*	G	0.687	0.725	0.47	0.22
rs72658409	9:22160088	*CDKN2B-AS1*, *DMRTA1*	C	0.96	0.951	0.58	–
rs7086803	10:112738717	*VTI1A*	G	0.723	0.799	0.04*	0.03*
rs11610143	12:51955287	*ACVR1B*	C	0.687	0.637	0.30	0.12
rs7216064	17:67902693	*BPTF*	A	0.677	0.627	0.47	0.30

*, P<0.05. EAF[a], effect allele frequency in the case group. EAF[b], effect allele frequency in the control group. The P derived from the Chi-square test between the NSCLC incidence and control groups. The P[†] was derived from the Chi-squared test, which compares lung cancer occurrence and control groups using a genetic variable categorized as allele carrier and non-carrier. EA, effect allele; SNP, single-nucleotide polymorphism.

The reduced level of sphinganine was shown in cluster 2 (P<0.001). A pattern of TEGBE (P<0.001), 1,5-hexadiene (P=0.008), dicyclohexylamine (P=0.009), and hydroxypropionylcarnitine (P<0.001) of cluster 2 was similar to cluster 1. In contrast, the marked elevation of cis-5-tetradecenoylcarnitine (P<0.001), L-acetylcarnitine (P=0.02), and Cer(d18:1/16:0) (P<0.001), and reduction of Cer(d18:0/14:0) (P<0.001) were similar patterns with cluster 0 or 3. The results may reflect pre-diagnostic metabolic shifts in at-risk individuals. Nevertheless, given the limited number of cases, additional validation is required.

For the next step, six metabolites (TEGBE, hydroxypropionylcarnitine, tyramine glucuronide, indoline, L-acetylcarnitine, and sphinganine) were selected based on their consistent and distinct patterns observed across unsupervised clustering results. TEGBE and hydroxypropionylcarnitine were consistently elevated in clusters predominantly composed of NSCLC incidence cases (0 and 3), while being relatively suppressed in clusters that consisted mostly of controls (1 and 2). The level of tyramine glucuronide was higher than the mean of total subjects in both case-dominant clusters, whereas among the control-dominant clusters, a significant decrease was observed only in cluster 1. The other three metabolites exhibited differences in only one of the case-dominant clusters and one of the control-dominant clusters, with opposite trends. The concentration of selected metabolites was log-transformed for further analysis to reduce skewness and approximate normality.

Association between genetic variants and metabolites

Of the total subjects, DNA genotyping was completed for 292 participants. Results on the 11 SNP alleles and frequencies for each subtype are presented in *Table 2*. Among 11 SNPs, only rs7086803 showed significant allele frequency differences between case and control groups. The control group exhibited a considerably greater frequency for the G allele at rs7086803 than the NSCLC incidence group (frequency of control =0.799, frequency of case =0.723, P from Chi-square test =0.04). The significance was maintained when using genetic variables categorized as allele carriers and non-carriers (P from Chi-square test =0.02).

Figure 2 describes the result of logistic regression analysis, which investigated the association between metabolite levels and NSCLC risk while adjusting for potential confounders, including age, gender, and smoking status. To explore the potential interaction between genetic variation and metabolite levels, a binary categorical variable (G allele carrier at rs7086803 *vs.* non-carrier at rs7086803) was utilized. An interaction term (G allele carrier status at rs7086803 × each metabolite) was included in the model to evaluate whether the relationship between metabolite levels and cancer risk differed by rs7086803 genotype.

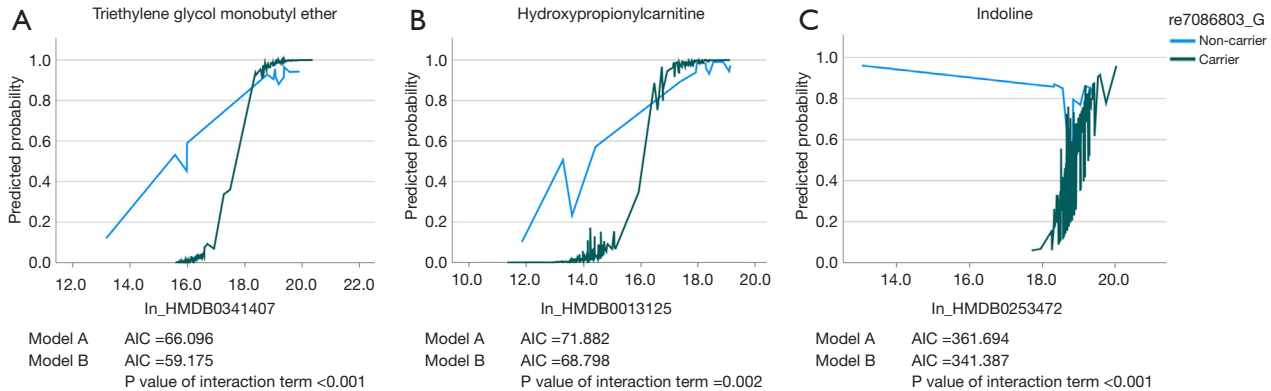

Figure 2 Interaction between metabolite and rs7086803 genotype. Predicted probabilities derived from the logistic regression model are plotted against each metabolite level. (A) Triethylene glycol monobutyl ether. (B) Hydroxypropionylcarnitine. (C) Indoline. Lines are shown separately for the G allele non-carrier of rs7086803 (blue line) and allele-carrier (green line). Model A: age, gender, smoking status, allele status of rs7086803, and each metabolite. Model B: age, gender, smoking status, rs7086803, each metabolite, and allele status of rs7086803 × each metabolite. Metabolite concentrations were log-transformed. AIC, Akaike information criterion.

As a result, significant interaction effects were observed between rs7086803 and three metabolites (TEGBE, hydroxypropionylcarnitine, and indoline). For example, in the predicted probability plot (*Figure 2*), the difference in predicted outcome probabilities between the two groups becomes more pronounced at approximately a TEGBE level of 17. Notably, the expected probability of the outcome increased sharply with higher indoline levels in individuals carrying the G allele of rs7086803, whereas no such trend was observed in non-carriers. This indicates that as serum metabolite level increases, the probability of the outcome rises more steeply in the individuals who carry the G allele at rs7086803. Furthermore, the model including the interaction term (model B in *Figure 2*) yielded a lower Akaike information criterion (AIC) compared to the main-effects-only model (model A in *Figure 2*), suggesting improved model fit despite the added complexity. This supports the relevance of the interaction between genotype and metabolite concentration in predicting disease risk.

Associations between selected metabolites and occupational, smoking status, and nutritional variables

To explore how six significant metabolites are associated with occupational and smoking status, we performed linear regression analyses in crude and age- and gender-adjusted models (*Figure 3*). Compared to current smokers, non-smokers had significantly lower levels of TEGBE (crude model, β =−1.084, 95% CI =−1.652 to −0.516, P<0.001; adjusted model, β =−0.637, 95% CI =−1.090 to −0.184, P=0.006) and tyramine glucuronide (crude model, β =−0.720, 95% CI =−1.303 to −0.138, P=0.016; adjusted model, β =−2.554, 95% CI =−3.920 to −1.187, P<0.001). The level of hydroxypropionylcarnitine was also significantly lower in non-smokers than in current smokers, as indicated by the adjusted model (β =−1.225, 95% CI =−1.956 to −0.494, P=0.001). The ex-smokers showed no significant difference compared to current smokers in any metabolites. Regarding occupational data, group 3 workers showed lower indoline level than group 1 (crude model, β =−0.335, 95% CI =−0.618 to −0.052, P=0.02; adjusted model, β =−0.332, 95% CI =−0.616 to −0.048, P=0.02). No significant associations were observed for the other metabolites.

Figure S2 depicts the results conducted on participants who responded to the nutritional questionnaire [NSCLC incidence (n=34) and control (n=26)]. Although the independent effects of vegetable intake amount (g/day) and smoking status were not statistically significant, the interaction between vegetable intake and smoking status was significant (β =0.009, P=0.03). The results suggest that smoking status influences the relationship between vegetable intake and NSCLC incidence. Specifically, vegetable intake may have a lesser effect on NSCLC in never-smokers, while it may have a more substantial impact on ex- and current smokers. Additionally, vegetable intake positively correlated with tyramine glucuronide (r =0.27, P=0.04). Negative association between grain intake and sphinganine (r =−0.27, P=0.04) was also observed.

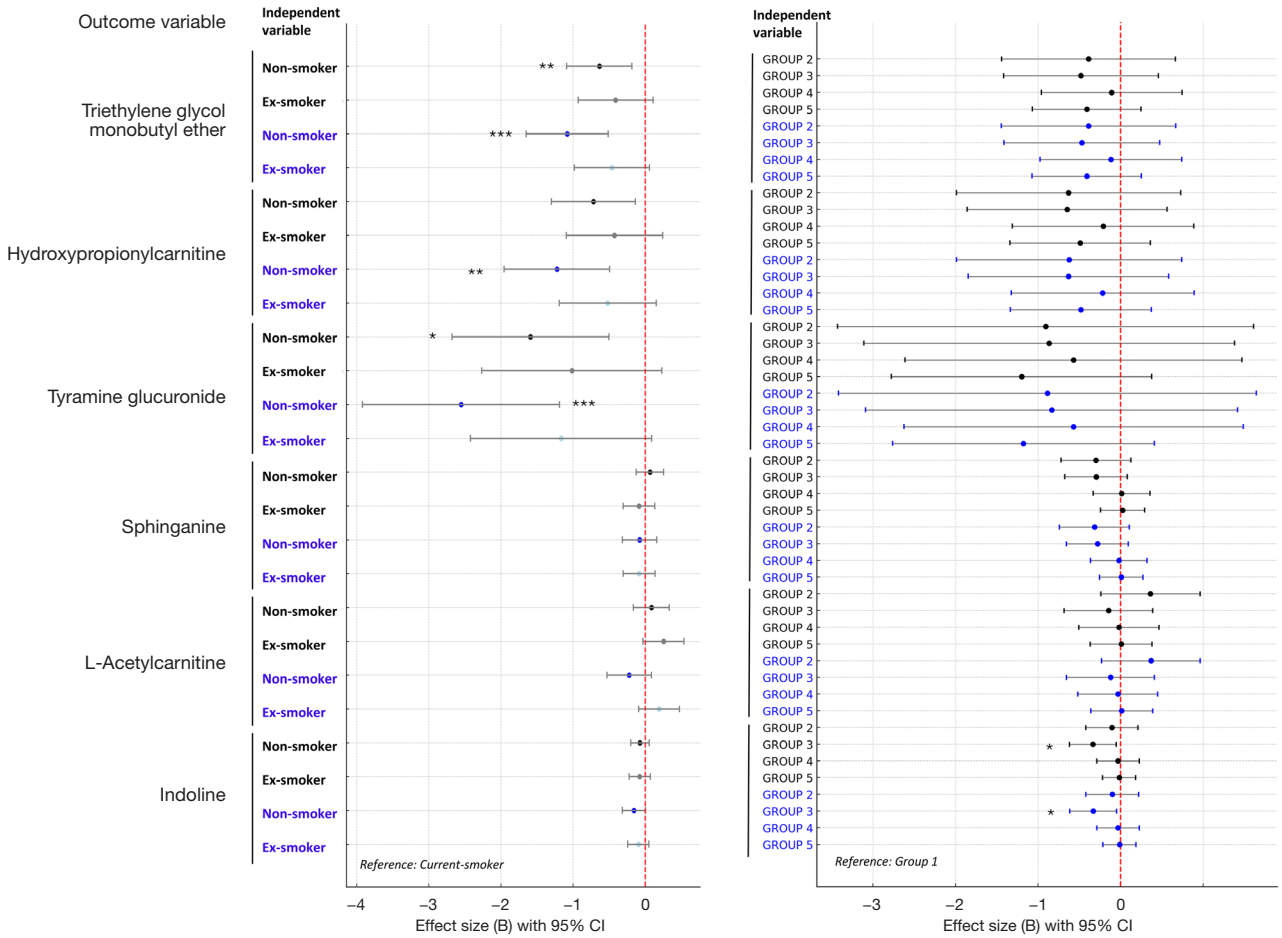

Figure 3 Associations of metabolite levels with smoking status and occupational factors. Forest plot presenting β coefficients and 95% CIs from linear regression models examining associations between exposures and metabolite levels. Smoking status is displayed on the left side, and occupational factors on the right side. Group 1, professional and office workers; group 2, sales and service workers; group 3, production and agricultural workers; group 4, homemakers; group 5, others. Metabolite concentrations were log-transformed. *P<0.05, **P<0.01, ***P<0.001. CI, confidence interval.

Discussion

This study aimed to identify biomarkers predictive of NSCLC risk and to enhance our understanding of the underlying characteristics of NSCLC. The primary strategy was to perform metabolomics on the pre-diagnostic serum to find biomarkers that could be utilized as risk prediction markers along with SNPs known to be associated with lung cancer in the Korean population. As a result, we identified six metabolites associated with NSCLC risk by clustering analysis. Among them, three metabolites were shown to have relevance to the interaction with rs7086803 in predicting the outcome. It highlights the genotype-dependent association between metabolite levels and disease risk. Moreover, both smoking status and occupational factors seemed to impact certain metabolites among the six examined, with dietary vegetable consumption potentially influencing NSCLC risk in smokers. This approach is valuable for identifying biomarkers associated with the onset of NSCLC. Further, by leveraging occupational, smoking status, and nutritional information, we explored how these metabolites relate to lifestyle factors to provide scientific insights.

At first, the integration of genomic, metabolomic, and clinical features through clustering effectively distinguishes NSCLC incidence cases from controls, achieving high predictive performance in both training and test sets. However, due to the relatively small sample size, further

validation in larger, independent cohorts is necessary to confirm the generalizability of these findings. Notably, a clear separation of metabolic profiles across the clusters was confirmed. This approach, which links unsupervised clustering-based metabolic profiles directly to biomarker selection, ensures that the selected metabolites reflect inherent disease-associated metabolic dysregulation, independent of prior labeling. By focusing on features that consistently distinguish cancer incidence-enriched clusters, the strategy enhances the potential for identifying robust biomarkers with high diagnostic sensitivity.

Metabolites exhibiting significance in the opposite direction between the case-dominant cluster and the control-dominant cluster have potential as candidate biomarkers for NSCLC risk. Several distinct metabolites appear to be biologically relevant in the context of NSCLC.

No direct association between TEGBE and lung cancer has been reported, whereas Kwon *et al.* (18) suggested its potential relevance; TEGBE may enhance pulmonary toxicity when co-administered with benzalkonium chloride, with increasing cellular oxidative stress by facilitating benzalkonium chloride uptake into lung cells, highlighting a potential risk associated with long-term co-exposure to these agents. Borgatta *et al.* (19) observed that glycol ethers are rapidly absorbed into human blood following inhalation exposure, with internal doses increasing in a nonlinear manner. These findings suggest that elevated TEGBE levels in pre-diagnostic serum may reflect significant environmental or occupational exposure, potentially contributing to lung cancer risk by cumulative pulmonary damage. While no significant association was identified between occupational factors and TEGBE, the level of TEGBE was significantly lower in non-smokers than in current smokers in the current study. Additional investigations are required to determine its mechanistic relevance in lung cancer development.

Our findings regarding hydroxypropionylcarnitine align with a previous report of altered short-chain acylcarnitines in patients with NSCLC (20). A significant elevation of hydroxypropionylcarnitine in the case-dominant cluster could reflect a broader disruption of short-chain fatty acid metabolism and mitochondrial β-oxidation. However, Zhang *et al.* (21) found that acylcarnitine (C16:2) is decreased in NSCLC tumor tissues compared to controls. Another acylcarnitine derivative, L-acetylcarnitine, identified as a significant metabolic biomarker in our study, has also been inconsistently and rarely reported in previous research. Acetylcarnitine (C2:0) levels were found to be reduced in NSCLC tumor tissues (21) but increased in the serum of lung cancer patients (22). Therefore, the two acylcarnitine derivatives detected in our research need further validation and exploration.

To date, there is no established link between tyramine glucuronide and lung cancer. Nevertheless, associations involving these compounds, their analogs, or derivatives have been reported in other cancer types; the nitrosated derivative of dicyclohexylamine exhibited genotoxicity in lymphocytes (23), and tyramine from fecal bacteria increased colorectal cancer risk by inducing DNA damage and tumorigenic processes (24), implying possible carcinogenicity as their characteristic. Increased levels of tyramine glucuronide in the NSCLC incidence group raise the possibility that these metabolites are indicative of disrupted metabolic states or early metabolic signatures of malignancy.

Sphinganine is a key intermediate in *de novo* sphingolipid metabolism and indirectly influences lung cancer cell growth and death by serving as a precursor to ceramide (25). An LC-MS-based study conducted by Huang *et al.* (26) revealed that a reduction of sphinganine alongside ceramide in the A549T human lung ADN cell line supports the notion that inhibition of the *de novo* synthesis pathway may contribute to the development of chemoresistance in lung cancer. The significant alteration in sphinganine levels observed in our pre-diagnostic samples may also be closely related to this *de novo* biosynthetic pathway.

Indoline itself has not been previously associated with lung cancer. Given that indoline-related compounds have demonstrated anticancer activity, including against lung cancer, through cell cycle arrest and apoptosis (27,28), the observed elevation in the pre-diagnostic sample of incident NSCLC cases may reflect compensatory metabolic responses, tumor-host interactions, or early metabolic dysregulation associated with tumorigenesis. Further functional studies are warranted to elucidate the role of indoline in the early stages of lung cancer development. It remains to be determined whether elevated indoline levels are causally involved in carcinogenic processes. Besides, a lower indoline level was observed in production and agricultural workers (group 3) than in professional and office workers (group 1). This occupational variation may reflect differences in environmental exposures, metabolic demands, or lifestyle factors across job categories. These findings highlight the need to further investigate how occupational contexts may influence early metabolic changes related to NSCLC risk.

Next, we explored the interplay between these metabolic alterations and critical genetic variants that potentially influence the pathogenesis of NSCLC. Among 11 SNPs previously reported to be related to lung cancer in Koreans, the G allele of rs7086803 showed significant frequency variations between the two groups of our study population. A large-scale genome-wide association studies (GWAS) reported the A allele of rs7086803 to be associated with increased lung cancer risk, particularly in never-smoking women of East Asian ancestry (29), whereas individuals of European descent did not replicate this association (30). This discrepancy may be attributable to differences in study populations, including population differences, environmental exposures, or statistical modeling approaches. A rs7086803 is associated with the *VTI1A* gene, which is involved in intracellular vesicle transport, while the precise molecular mechanism linking *VTI1A* to lung cancer remains unclear (31). These findings suggest that rs7086803 may play a significant role in genetic susceptibility to NSCLC.

Indeed, logistic regression analysis revealed that among individuals with the G allele at rs7086803, higher serum metabolite levels (TEGBE, hydroxypropionylcarnitine, and indoline) were associated with a more pronounced elevation in the likelihood of NSCLC incidence. It underscores the genotype-dependent association between metabolite levels and disease risk. In other words, the effect of the G allele was more pronounced in the presence of elevated serum metabolite levels, suggesting a gene-metabolite interaction that modulates NSCLC risk. However, given the skewed distribution of the rs7086803 genotype in the dataset (approximately 10% carriers *vs.* 90% non-carriers), the wide confidence intervals suggest uncertainty in the estimates, likely due to sparse data or distributional imbalance. Therefore, further studies with larger and more balanced sample sizes are warranted to validate these findings.

Interestingly, vegetable intake was significantly positively correlated with the levels of tyramine glucuronide in the present study, indicating potential dietary influences on specific metabolic pathways. In addition, vegetable consumption appeared to be associated with NSCLC risk among smokers, suggesting that certain dietary patterns may interact with smoking-related carcinogenic processes. In linear regression analysis, non-smokers exhibited significantly lower levels of tyramine glucuronide compared to current smokers. These findings collectively suggest that both dietary habits and smoking status potentially influence metabolic profiles relevant to NSCLC development. Although no research results have yet reported that dietary vegetable consumption is associated with lung cancer risk in smokers, there is a need to confirm the results of this study in a larger sample, and the underlying mechanism needs to be demonstrated through experimental studies.

Dietary vegetables have been reported to have benefits in preventing lung cancer (32,33). However, smoking could be a critical confounder in determining the association between vegetable intake and lung cancer risk. In the Japanese prospective study, an inverse association between cruciferous vegetable intake and lung cancer risk was noted in currently non-smoking men (34). The European Food Safety Authority Panel reported scientific evidence for tolerable upper intake levels for supplementary preformed vitamin A and β-carotene for smokers (35). Meta-analysis of the randomized controlled trials confirmed that taking supplementary β-carotene increased lung cancer risk, while not significant for other site-specific malignancies. Besides, increasing the dose of supplementary β-carotene elevated the cancer risk among smokers (36). Consequently, we suggest that dietary vegetable consumption, not only supplementary, may not have a positive effect in all cases. The interaction between specific nutrients and smoking should be considered. However, further validation is warranted due to the limited sample size.

Several limitations should be delineated. First, our data was limited to a relatively small sample size. With a larger sample size, it was possible to discover more meaningful biomarkers and associated lifestyle factors with substantial statistical power (e.g., the composition of group 3 in occupational variables, and the limited number of participants with nutritional data). An additional limitation is the inability to distinguish between clusters 1 and 2, which showed no significant differences in demographic, lifestyle, or genetic factors, likely due to the limited sample size. Furthermore, since the investigation was conducted solely on a Korean population, it cannot be generalized to other ethnic groups without further research. Next, drawing the causality and interpreting the underlying mechanisms between biomarkers was difficult in our study design. Additional experimental research is needed to elucidate the exact mechanism of pathogenesis related to the discovered associations. Lastly, validation in a sufficient number of independent Korean cohorts is necessary for clinical application.

Conclusions

Despite these limitations, this study successfully identified

meaningful biomarkers that can efficiently predict the risk of NSCLC based on metabolomics technology. Potential associations between these metabolites and smoking status, along with occupational factors, were also examined. Notably, we provided evidence that vegetable consumption may not have a positive effect in all cases from the perspective of lung cancer, and the interaction between specific nutrients and smoking should be taken into account. Overall, this study may lay the groundwork for future evidence-based biomarker discovery in NSCLC.

Acknowledgments

None.

Footnote

Reporting Checklist: The authors have completed the REMARK reporting checklist. Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-2025-603/rc

Data Sharing Statement: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-2025-603/dss

Peer Review File: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-2025-603/prf

Funding: This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF No. 2022R1A6A3A01085831) and the research program of the Korea Medical Institute.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-2025-603/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was approved by the Institutional Review Board at the Yonsei University Health System (IRB No. 4-2022-1136) under the Declaration of Helsinki and its subsequent amendments. Written informed consent was obtained from all participants.

References

1. World Health Organization. Global Cancer Observatory lung cancer incidence and mortality statistics. 2023. Available online: https://gco.iarc.fr/
2. Korean Statistical Information Service. Annual Report on the Cause of Death Statistics. 2023. Available online: http://kosis.kr
3. Padinharayil H, Varghese J, John MC, et al. Non-small cell lung carcinoma (NSCLC): Implications on molecular pathology and advances in early diagnostics and therapeutics. Genes Dis 2023;10:960-89.
4. Herrera-Juárez M, Serrano-Gómez C, Bote-de-Cabo H, et al. Targeted therapy for lung cancer: Beyond EGFR and ALK. Cancer 2023;129:1803-20.
5. Alexander M, Kim SY, Cheng H. Update 2020: Management of Non-Small Cell Lung Cancer. Lung 2020;198:897-907.
6. Araghi M, Mannani R, Heidarnejad Maleki A, et al. Recent advances in non-small cell lung cancer targeted therapy; an update review. Cancer Cell Int 2023;23:162.
7. Deek RA, Ma S, Lewis J, et al. Statistical and computational methods for integrating microbiome, host genomics, and metabolomics data. Elife 2024;13:e88956.
8. Babu M, Snyder M. Multi-Omics Profiling for Health. Mol Cell Proteomics 2023;22:100561.
9. Qian X, Zhang HY, Li QL, et al. Integrated microbiome, metabolome, and proteome analysis identifies a novel interplay among commensal bacteria, metabolites and candidate targets in non-small cell lung cancer. Clin Transl Med 2022;12:e947.
10. Hoang LT, Domingo-Sabugo C, Starren ES, et al. Metabolomic, transcriptomic and genetic integrative analysis reveals important roles of adenosine diphosphate in haemostasis and platelet activation in non-small-cell lung cancer. Mol Oncol 2019;13:2406-21.
11. Kwon HJ, Park UH, Goh CJ, et al. Enhancing Lung Cancer Classification through Integration of Liquid Biopsy Multi-Omics Data with Machine Learning Techniques.

Cancers (Basel) 2023;15:4556.
12. Yang Y, Sun H, Zhang Y, et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. Cell Rep 2021;36:109442.
13. Yaqoob A, Musheer Aziz R, Verma NK. Applications and techniques of machine learning in cancer classification: a systematic review. Hum-Cent Intell Syst 2023;3:588-615.
14. Cook M, Qorri B, Baskar A, et al. Small patient datasets reveal genetic drivers of non-small cell lung cancer subtypes using machine learning for hypothesis generation. Explor Med 2023;4:428-40.
15. Jee YH, Emberson J, Jung KJ, et al. Cohort Profile: The Korean Cancer Prevention Study-II (KCPS-II) Biobank. Int J Epidemiol 2018;47:385-386f.
16. Han Y, Huh R, Jung KJ, et al. Dietary modulation for the hypertension risk group in Koreans: a cross-sectional study. Nutr Metab (Lond) 2025;22:30.
17. Kim J, Park YS, Kim JH, et al. Predicting Lung Cancer in Korean Never-Smokers With Polygenic Risk Scores. Genet Epidemiol 2025;49:e22586.
18. Kwon D, Lim YM, Kwon JT, et al. Evaluation of pulmonary toxicity of benzalkonium chloride and triethylene glycol mixtures using in vitro and in vivo systems. Environ Toxicol 2019;34:561-72.
19. Borgatta M, Wild P, Hopf NB. Blood absorption toxicokinetics of glycol ethers after inhalation: A human controlled study. Sci Total Environ 2022;816:151637.
20. Shestakova KM, Moskaleva NE, Boldin AA, et al. Targeted metabolomic profiling as a tool for diagnostics of patients with non-small-cell lung cancer. Sci Rep 2023;13:11072.
21. Zhang J, Zang X, Jiao P, et al. Alterations of Ceramides, Acylcarnitines, GlyceroLPLs, and Amines in NSCLC Tissues. J Proteome Res 2024;23:4343-58.
22. Zhao F, An R, Wang L, et al. Specific Gut Microbiome and Serum Metabolome Changes in Lung Cancer Patients. Front Cell Infect Microbiol 2021;11:725284.
23. Westphal GA, Müller MM, Herting C, et al. Genotoxic effects of N-nitrosodicyclohexylamine in isolated human lymphocytes. Arch Toxicol 2001;75:118-22.
24. Glymenaki M, Curio S, Shrestha S, et al. Roux-en-Y gastric bypass-associated fecal tyramine promotes colon cancer risk via increased DNA damage, cell proliferation, and inflammation. Microbiome 2025;13:60.
25. Lin M, Li Y, Wang S, et al. Sphingolipid Metabolism and Signaling in Lung Cancer: A Potential Therapeutic Target. J Oncol 2022;2022:9099612.
26. Huang H, Tong TT, Yau LF, et al. LC-MS based sphingolipidomic study on A549 human lung adenocarcinoma cell line and its taxol-resistant strain. BMC Cancer 2018;18:799.
27. Lim HM, Park SH, Nam MJ. Induction of apoptosis in indole-3-carbinol-treated lung cancer H1299 cells via ROS level elevation. Hum Exp Toxicol 2021;40:812-25.
28. Palanivel S, Murugesan A, Subramanian K, et al. Antiproliferative and apoptotic effects of indole derivative, N-(2-hydroxy-5-nitrophenyl (4'-methylphenyl) methyl) indoline in breast cancer cells. Eur J Pharmacol 2020;881:173195.
29. Lan Q, Hsiung CA, Matsuo K, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. Nat Genet 2012;44:1330-5.
30. Hung RJ, Spitz MR, Houlston RS, et al. Lung Cancer Risk in Never-Smokers of European Descent is Associated With Genetic Variation in the 5(p)15.33 TERT-CLPTM1Ll Region. J Thorac Oncol 2019;14:1360-9.
31. Tang BL. Vesicle transport through interaction with t-SNAREs 1a (Vti1a)'s roles in neurons. Heliyon 2020;6:e04600.
32. Zheng S, Yan J, Wang J, et al. Unveiling the Effects of Cruciferous Vegetable Intake on Different Cancers: A Systematic Review and Dose-Response Meta-analysis. Nutr Rev 2025;83:842-58.
33. Luo S, Lin D, Lai S, et al. Dietary consumption trend and its correlation with global cancer burden: A quantitative and comprehensive analysis from 1990 to 2019. Nutrition 2024;117:112225.
34. Mori N, Shimazu T, Sasazuki S, et al. Cruciferous Vegetable Intake Is Inversely Associated with Lung Cancer Risk among Current Nonsmoking Men in the Japan Public Health Center (JPHC) Study. J Nutr 2017;147:841-9.
35. EFSA Panel on Nutrition, Novel Foods and Food Allergens (NDA); Turck D, Bohn T, et al. Scientific opinion on the tolerable upper intake level for preformed vitamin A and β-carotene. EFSA J 2024;22:e8814.
36. Zhang Y, Yang J, Na X, et al. Association between β-carotene supplementation and risk of cancer: a meta-analysis of randomized controlled trials. Nutr Rev 2023;81:1118-30.

Cite this article as: Han Y, Jung KJ, Choi SG, Yang YS, Lee K, Jee SH. Biomarkers for non-small cell lung cancer risk using multi-omics approaches: a nested case-control study. Transl Lung Cancer Res 2025;14(9):3645-3658. doi: 10.21037/tlcr-2025-603