

國立臺灣大學公共衛生學院健康數據拓析統計研究所

碩士論文



Graduate Institute of Health Data Analytics and Statistics

College of Public Health

National Taiwan University

Master's Thesis

DNA 甲基化在吸菸與血脂性狀之間的中介作用

DNA methylation may mediate the association between
smoking and lipid traits

楊正毅

Cheng-Yi Yang

指導教授：林苑俞 博士

Advisor: Wan-Yu Lin, Ph.D.

中華民國 115 年 7 月

July 2026



誌謝

時光飛逝，碩士班生涯轉眼間到了尾聲。回首這段撰寫論文與探索學術的歷程，充實卻也充滿挑戰。這份論文的順利完成，絕非憑一己之力所能達成，心中充滿了無盡的感激。

首先，我要感謝我的指導教授林菀俞老師。在論文寫作期間，老師不時給予我悉心指導與耐心引導。每當我在研究陷入瓶頸時，老師總能簡要且精準的點出關鍵問題、指引出清晰的研究方向，並引導我用更嚴謹、客觀的思維去思考。老師對學術的熱情與以及對細節的要求奠定了我研究的基礎，令我受益良多。對此，我要再次對老師致上我最高的敬意、謝意。

也感謝口試委員陳弘昕老師與馮嫻臻老師（依照姓氏筆畫排序）在百忙之中抽空審閱論文，並於口試期間提出許多寶貴且切中的建言。老師們精闢的見解與建設性的建議，補足了我研究中的盲點，使本論文的架構與論述得以更加完善與嚴謹。

最後，我要將最深摯的感謝獻給我的家人與朋友們。感謝你們無條件的包容、陪伴與鼓勵，給予我最堅定的情感後盾，讓我能無後顧之憂地專心完成學業。

中文摘要



吸菸是心血管疾病 (cardiovascular disease, CVD) 已知的重要危險因子，並與脂質特徵相關，例如高密度脂蛋白膽固醇 (high-density lipoprotein cholesterol, HDL-C) 與三酸甘油酯 (triglycerides, TG)。DNA 甲基化可能是解釋這些關聯背後的表觀遺傳途徑之一。據此，本研究旨在探討 DNA 甲基化在吸菸與脂質特徵之間關聯中可能扮演的中介角色。

本研究使用臺灣人體生物資料庫中 2,463 名參與者的全基因組 DNA 甲基化資料進行高維度中介分析。首先，使用調整潛在混雜因子的線性回歸模型，評估吸菸狀態與脂質特徵之間的關聯。接著，針對具有顯著關聯的吸菸—脂質特徵組合，應用高維度中介檢定 (high-dimensional mediation testing, HDMT) 以辨識可能作為中介因子的 CpG 位點 ($FDR < 0.05$)。為估計整體中介比例，本研究針對辨識出的 CpG 位點進行主成分分析，並透過基於主成分建立的中介模型估計潛在中介因子的中介比例。為進一步篩選出具獨立中介效應的位點，本研究使用 HIMA2 建立多中介因子模型以同時考慮多個中介位點。

以目前吸菸作為暴露變項時，在 846,232 個 CpG 位點中，分別有 109 個與 65 個 CpG 位點被辨識為吸菸—HDL-C 與吸菸— $\log(TG)$ 關聯中的顯著中介因子。DNA 甲基化估計分別可解釋吸菸與 HDL-C 及 $\log(TG)$ 關聯中約 65% 與 7% 的中介比例。多重中介因子模型進一步辨識出 14 個與 HDL-C 相關以及 1 個與 TG 相關的獨立中介 CpG 位點。

本研究結果為 DNA 甲基化參與吸菸與脂質特徵之間的關聯的可能性提供了統計支持。然而，受到橫斷面研究之故有侷限，仍待後續的縱向研究與功能性研究，以闡明其潛在的生物學脈絡。

關鍵字：吸菸、DNA 甲基化、脂質特徵、表觀遺傳學、高維度中介分析、臺灣人體生物資料庫

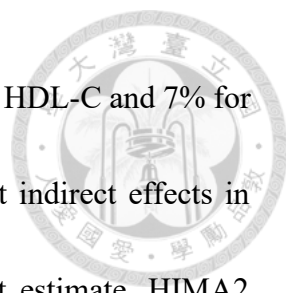
Abstract



Cigarette smoking is a well-established risk factor for cardiovascular disease (CVD) and is associated with lipid traits such as high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG). DNA methylation may represent an epigenetic mechanism underlying these associations. This study aimed to explore the potential mediating role of DNA methylation in the association between smoking and lipid traits.

High-dimensional mediation analysis was performed using genome-wide DNA methylation data (Illumina EPIC Array) from 2,463 participants in the Taiwan Biobank (TWB). Linear regression models adjusted for potential confounders were used to assess the associations between smoking status and lipid traits. High-dimensional mediation testing (HDMT) was then applied to identify CpG sites that potentially act as mediators ($FDR < 0.05$) of the significantly associated smoking–lipid pairs. To estimate the overall mediation proportion, principal component analysis (PCA) was performed on the identified CpG sites, and the mediation proportion explained by the top principal components was estimated. In addition, HIMA2 was applied to identify independent mediating CpG sites.

Using current smoking as the exposure, among 846,232 CpG sites, 109 and 65 CpG sites were identified by HDMT as potential mediators of the smoking–HDL-C and smoking–log(TG) associations, respectively. The net PCA-based mediation estimates



across significant principal components were approximately 65% for HDL-C and 7% for log(TG). For log(TG), six principal components showed significant indirect effects in opposing directions, resulting in substantial cancellation in the net estimate. HIMA2 identified 14 joint-model mediation signals for HDL-C and identified cg00574958 in CPT1A for log(TG).

These findings provide statistical evidence supporting a potential role of DNA methylation in the association between smoking and lipid traits. Further longitudinal and functional studies are needed to clarify the underlying biological mechanisms.

Keywords: smoking; DNA methylation; lipid traits; epigenetics; High-dimensional mediation analysis; Taiwan Biobank

目次

誌謝.....	ii
中文摘要.....	iii
Abstract	iv
目次.....	vi
圖次.....	vii
表次.....	viii
Introduction.....	1
Materials and Methods.....	3
The Taiwan Biobank data	3
Quality Control for methylation data	3
Smoking factors and lipid traits	5
Covariates.....	6
Association analysis	6
Mediation analysis	6
Functional enrichment analysis.....	10
Results.....	11
Sample characteristics	11
Associations between smoking and lipid traits	12
Mediation analysis	12
Functional enrichment analysis.....	14
Discussion	15
Conclusion.....	21
Tables	22
Figures.....	24
References.....	33

圖次



Figure 1. Adjusted associations between current smoking and lipid traits.....	24
Figure 2. Adjusted associations between former smoking and lipid traits.	25
Figure 3. Manhattan Plot of HDMT Results for the Current Smoking–HDL-C Association	26
Figure 4. Manhattan Plot of HDMT Results for the Current Smoking–log(TG) Association	27
Figure 5. Estimated Mediation Effects of HDMT-identified CpGs for the Current Smoking–HDL-C Association.....	28
Figure 6. Estimated Mediation Effects of HDMT-identified CpGs for the Current Smoking–log(TG) Association.....	29
Figure 7. Estimated Mediation Effects of HIMA2-identified CpGs for the Current Smoking–HDL-C Association.....	30
Figure 8. GO enrichment analysis for the current smoking–HDL-C association	31
Figure 9. GO enrichment analysis for the current smoking–log(TG) association	32

表次



Table 1. Basic characteristics of the 2,463 Taiwan Biobank participants	
stratified by smoking status	22

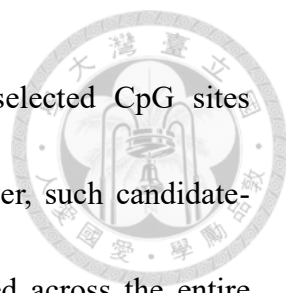


Introduction

Cigarette smoking substantially elevates the risk of cardiovascular disease (CVD)[1-3]. One potential mechanism linking smoking to cardiovascular risk is its adverse effects on lipid metabolism. Smoking's association with unfavorable lipid profiles has been shown in previous research [4, 5]. Because these lipid traits are closely related to cardiometabolic health, elucidating the biological mechanisms underlying smoking-associated lipid alterations is of considerable interest.

DNA methylation represents a dynamic epigenetic mark modulated by external factors, including smoking exposure [6]. Across numerous epigenome-wide association studies, robust smoking-induced methylation alterations have been documented, particularly at CpGs annotated to genes such as AHRR, F2RL3, and GPR15 [7, 8]. In addition, accumulating evidence has demonstrated that DNA methylation patterns are closely linked to lipid profiles across populations[9, 10]. These findings point to DNA methylation as a potential epigenetic mediator through which smoking exposure influences lipid profiles.

However, although smoking has been shown to be associated with both DNA methylation and lipid traits, it remains unclear whether DNA methylation serves as a functional mediator between smoking exposure and lipid alterations. Previous mediation



studies involving DNA methylation have often focused on preselected CpG sites identified from prior associations or biological knowledge. However, such candidate-based approaches may fail to identify mediation signals distributed across the entire genome. Given the large number of correlated CpG sites measured in genome-wide methylation data, an epigenome-wide mediation analysis requires statistical methods specifically designed for high-dimensional mediator settings. Therefore, we applied high-dimensional mediation analysis to systematically evaluate potential CpG mediators of the associations between smoking and lipid traits in Taiwan Biobank participants.

Materials and Methods



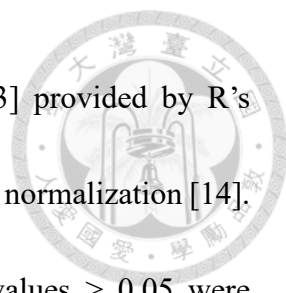
The Taiwan Biobank data

Since its inception in October 2012, the Taiwan Biobank has registered nearly 200,000 adult volunteers from the general community [11]. Community members in Taiwan aged 30 to 70 were enrolled, provided they had not been previously diagnosed with any form of cancer. After providing written informed consent, participants completed lifestyle questionnaires, underwent physical examinations, and provided blood and urine samples following at least 6 hours of fasting.

Between 2016 and 2021, a subset of 2,474 individuals was randomly sampled from the TWB cohort for DNA methylation quantification stratified by county population size and national sex ratios to ensure population representativeness. Epigenome-wide DNA methylation was profiled using the Illumina Infinium MethylationEPIC BeadChip, which covers 865,859 CpG sites.

Quality Control for methylation data

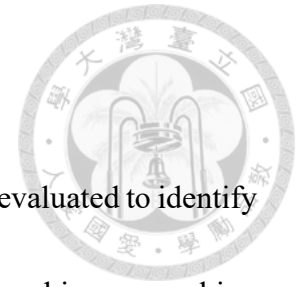
Following DNA methylation profiling via the Illumina Infinium MethylationEPIC BeadChip, individual CpGs were annotated to genes using ‘IlluminaHumanMethylationEPICanno.ilm10b4.hg19’ R package (version 0.6.0) . Raw intensity (.idat) files were processed according to previously described procedures [12].



Briefly, the normal-exponential out-of-band (NOOB) approach [13] provided by R's minfi package was performed for background correction and dye-bias normalization [14]. After NOOB normalization, methylation levels with detection p-values > 0.05 were considered unreliable [15] and treated as missing values, and thus were imputed using the mean methylation level of the corresponding CpG site across all samples. Additionally, CpG sites on sex chromosomes were removed due to X-chromosome inactivation[16]. Because the M-value method has better statistical properties [17], methylation levels were transformed into M-values for all subsequent statistical analyses.

Due to discordance between self-reported sex and methylation-predicted sex, 6 samples were excluded. 5 participants were excluded due to missing data in the required covariates. After these QC steps, a total of 2,463 eligible participants remained for the downstream association analyses.

Smoking factors and lipid traits



Smoking-related variables and cardiometabolic lipid traits were evaluated to identify exposure–outcome pairs for subsequent mediation analysis. Two binary smoking variables were considered potential exposure variables: current smoking and former smoking. Participants were categorized as current smokers if they actively smoked at enrollment, whereas former smokers were defined as those with a history of smoking who had ceased smoking. Participants who were neither current nor former smokers were classified as never smokers.

Lipid traits included high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (T-CHO), and triglycerides (TG), which were treated as candidate outcome variables. Because the distribution of triglyceride (TG) levels was right-skewed, TG values were log-transformed to reduce skewness and ensure conformity with linear-model assumptions[18].

To identify smoking – lipid trait combinations relevant for mediation analysis, associations between smoking variables and each lipid trait were evaluated using linear regression models. Only exposure – outcome pairs showing statistically significant associations in these preliminary analyses were retained for subsequent mediation analysis. Based on this screening step, only the associations of current smoking with HDL-C and log(TG) were significant. Therefore, downstream mediation analyses

focused on these two exposure-outcome pairs.



Covariates

To control for potential confounding effects, all statistical models were adjusted for demographic and lifestyle covariates, including age [19], sex [20], BMI [21], drinking status [22], regular exercise status [23], educational level [24], and 11 estimated blood cell type proportions. To account for cellular heterogeneity in whole blood samples, estimation of proportions of 12 immune cell types were calculated using the Houseman constrained projection method [25, 26] provided by R's EpiDISH package[27]. 11 of 12 estimated immune cell types were included as covariates. (Neutrophil was excluded due to multicollinearity). These covariates were included in every statistical model, including the preliminary association analyses and subsequent mediation analyses.

Association analysis

Prior to mediation analysis, association analyses were performed to evaluate the associations between smoking status and lipid traits. Subsequent mediation analyses were restricted to exposure–outcome pairs that demonstrated statistically significant associations.

Mediation analysis

For each exposure – outcome pair with a significant association, we used two



regression models to estimate the mediation effect of the i -th CpG:

$$M_i = \alpha_{0,i} + \alpha_{X,i}X + \alpha_{C,i}\mathbf{C} + \epsilon_i \quad (1)$$

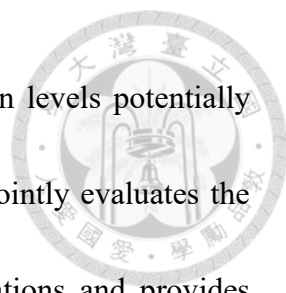
$$Y = \beta_{0,i} + \beta_{X,i}X + \beta_{M,i}M_i + \beta_{C,i}\mathbf{C} + \delta_{Y,i} \quad (2)$$

where X denotes smoking status, M_i indicates the DNA methylation status (M-value) for the i -th CpG, Y denotes the lipid trait, and $\mathbf{C}$ represents the set of covariates. In this framework, $\alpha_{X,i}$ captures the effect of the exposure on the i -th methylation level of CpG, and $\beta_{M,i}$ captures the effect of the methylation level of i -th CpG site on the outcome conditional on the exposure. The mediation effect for each CpG site is characterized by the product $\alpha_{X,i} \times \beta_{M,i}$. The corresponding hypothesis testing problem can be formulated as a composite null hypothesis:

$$H_{0,i}: \alpha_{X,i} \times \beta_{M,i} = 0$$

$$H_{a,i}: \alpha_{X,i} \times \beta_{M,i} \neq 0$$

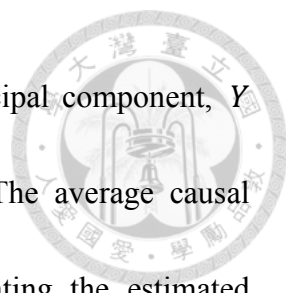
This type of hypothesis testing is commonly referred to as evaluating mediation effects under a composite null, where the null hypothesis holds if either $\alpha_{X,i} = 0$ or $\beta_{M,i} = 0$. In high-dimensional settings, where candidate mediators substantially outnumber the sample size, traditional mediation approaches (e.g., Sobel test or joint significance test) may suffer from low power of test and severe multiple testing burden [28]. Therefore, high-dimensional mediation methods are required to appropriately control false positives while maintaining statistical power. In this study, high-dimensional mediation testing



(HDMT) [29] was applied to identify CpG sites whose methylation levels potentially mediate the association between smoking and lipid traits. HDMT jointly evaluates the evidence of the exposure–mediator and mediator–outcome associations and provides valid inference under high-dimensional settings. To account for multiple testing across genome-wide CpG loci, false discovery rate (FDR) [30] correction was applied via the Benjamini–Hochberg method. CpG sites were considered potential mediators if their Benjamini–Hochberg adjusted P-values were below 0.05. To estimate the cumulative mediation proportion of the total effect attributable to DNA methylation in smoking-induced lipid variations, we applied a PCA-based dimensionality reduction framework. Principal components (PCs) were derived from the methylation M-values of selected CpG sites. CpG sites identified by HDMT were included in PCA to reduce dimensionality and summarize correlated methylation patterns. The number of PCs retained for subsequent mediation analysis was determined by visual inspection of the scree plot. The incremental variance explained by additional components substantially leveled off after approximately PC10; therefore, the first 10 PCs were retained. For each retained PC, mediation analysis was performed via the mediate function implemented in the mediation R package [31]. The mediation analysis was based on the following models:

$$PC_k = \alpha_{0,k} + \alpha_{X,k}X + \alpha_{C,k}C + \epsilon_k$$

$$Y = \beta_{0,k} + \beta_{X,k}X + \beta_{PC,k}PC_k + \beta_{C,k}C + \delta_k$$



where X denotes smoking status, PC_k represents the k -th principal component, Y denotes the lipid trait, and C represents the set of covariates. The average causal mediation effect (ACME) was calculated for each PC, representing the estimated mediation effect of smoking on the lipid trait for the corresponding methylation component. For each PC, the mediation proportion was determined by dividing the estimated ACME by the total effect. The total mediation effect was then estimated by summing the mediation effects across the significant PCs. To account for potential confounding, all regression models incorporated the identical set of covariates specified in the primary analysis. Statistical inference was conducted using nonparametric bootstrapping to obtain confidence intervals and p-values. In addition, HIMA2 [32], a high-dimensional mediation analysis method based on de-biased Lasso, was applied as a complementary approach to identify a sparse set of potential mediators. Unlike HDMT, which evaluates mediation effects for each CpG site individually, HIMA2 performs variable selection in a joint model and estimates the effects of multiple mediators simultaneously. By incorporating a de-biased Lasso procedure, HIMA2 provides asymptotically unbiased estimates and valid statistical inference for both exposure–mediator (α) and mediator–outcome (β) effects in high-dimensional settings.

Functional enrichment analysis



To elucidate the biological processes associated with the identified DNA methylation mediators, Gene Ontology (GO) enrichment analysis was performed using the methylRRA() function of the R package methylGSA [33]. The complete set of unadjusted CpG-level P-values obtained from the HDMT analysis, rather than only CpGs reaching FDR significance, was used as input. CpGs were mapped to genes using the EPIC array annotation provided by IlluminaHumanMethylationEPICanno.ilm10b4.hg19. The methylRRA method combines multiple CpG-level P-values mapped to the same gene based on the Robust Rank Aggregation (RRA) algorithm [34], while accounting for the unequal numbers of CpG sites annotated to different genes in array-based DNA methylation data. Over-representation analysis (ORA) was conducted to identify enriched GO terms. Statistical significance was defined as a Benjamini-Hochberg adjusted P-value < 0.05 .

Results



Sample characteristics

Table 1 summarizes the baseline characteristics of the 2,463 participants from the Taiwan Biobank, stratified by smoking status. The study population comprised 1,870 (75.9%) never smokers, 310 (12.6%) former smokers, and 283 (11.5%) current smokers. Differences in lipid profiles were observed across the three groups. Mean HDL-C level decreased progressively from never smokers (55.6 ± 13.7 mg/dL) to former smokers (49.6 ± 11.9 mg/dL) and current smokers (46.2 ± 11.7 mg/dL). In contrast, current smokers exhibited substantially higher TG levels (160.8 ± 144.3 mg/dL) compared with never smokers (111.0 ± 101.7 mg/dL). LDL-C and T-CHO levels were comparable across groups.

Differences were also observed in demographic and lifestyle characteristics. Current smokers and former smokers were predominantly male (83.0% and 87.4%, respectively), whereas the proportion of males was lower among never smokers (39.3%). Current smokers also had a higher mean BMI (25.4 ± 3.7 kg/m²) and a higher prevalence of alcohol consumption (24.4%) compared to never smokers. These variables, along with education level, regular exercise, and estimated cell-type proportions, were included as covariates in the subsequent association and mediation analyses.

Associations between smoking and lipid traits



After adjusting for potential confounders, current smoking was significantly associated with two of the four evaluated lipid traits (Figure 1 and 2). Specifically, current smoking was associated with decreased levels of HDL-C ($\beta = -4.43, SE = 0.82, P = 7.72 \times 10^{-8}$) and increased log-transformed TG levels ($\beta = 0.14, SE = 0.04, P = 8.85 \times 10^{-5}$). In contrast, no evidence of association was observed for LDL-C ($P = 0.802$) or T-CHO ($P = 0.281$). Former smoking was not significantly associated with any of the lipid traits (all $P > 0.05$), although a marginal association was observed for HDL-C ($\beta = -1.32, P = 0.083$). Accordingly, subsequent mediation analyses focused on the associations of current smoking with HDL-C and log(TG).

Mediation analysis

Using high-dimensional mediation testing (HDMT) to screen 846,232 CpG sites across the epigenome, we identified 109 and 65 CpG sites as potential mediators of the smoking – HDL-C association and smoking – log(TG) association, respectively (FDR < 0.05) (Figure 3 ,4 ,5 and 6). Among the potential CpGs identified in the HDMT analysis of the smoking–HDL-C association, cg21161138, located in the AHRR gene, was estimated to have the largest mediation effect. AHRR has been widely reported as a smoking-related methylation marker [7]. In the HDMT analysis of the smoking–log(TG)

association, three of the top significant CpG sites (cg17058475, cg00574958, cg05325763) were annotated to CPT1A. DNA methylation at CPT1A has been reported to be negatively associated with TG levels [9, 35, 36]. Notably, cg17058475 and cg00574958 were also identified as potential mediators in the smoking–HDL-C association.

To address the high collinearity among the identified loci and to estimate the overall epigenetic contribution, PCA was applied to reduce the dimensionality of the HDMT-identified CpGs. By summing the estimated mediation effects of all significant principal components ($P < 0.05$), DNA methylation was estimated to account for 64.74% and 6.71% of total smoking effects on HDL-C and log(TG), respectively. Given the high-dimensional and cross-sectional design, these values provide exploratory, model-based estimates of the extent to which shared methylation patterns may account for the observed smoking–lipid associations, rather than established causal mediation proportions.

Furthermore, to identify a sparse set of independent mediators, we applied a high-dimensional penalized multivariate mediation model (HIMA2). HIMA2 identified 14 independent CpGs for the smoking – HDL-C association (Figure 7). The CpG identified by HDMT cg21161138 at AHRR remained significant in HIMA2 and had the largest estimated mediation effect. For the smoking–log(TG) association, HIMA2 identified one CpG site, cg00574958 in CPT1A. Notably, cg00574958 was also identified by HDMT.

Functional enrichment analysis



The ORA-based GO enrichment analysis for the smoking–HDL-C association identified enrichment related to viral- and interferon-related terms (Figure 8). The top GO terms were mainly related to viral processes, interferon response, and antiviral immune pathways. For the smoking–log(TG) association, the ORA-based GO enrichment analysis identified enrichment mainly in kinase-related terms (Figure 9).

Discussion



Few studies have examined mediation effects across the epigenome. Previous research investigating the association between smoking, DNA methylation, and lipid traits have primarily relied on standard epigenome-wide association studies. Typically, these studies pre-select candidate CpG sites and subsequently apply traditional single-mediator models to evaluate whether individual CpG sites mediate the association between smoking and lipid traits [37]. To the best of our knowledge, this is the first epigenome-wide study to investigate the mediating role of DNA methylation in the association between smoking and lipid traits using high-dimensional mediation analysis.

In the current study, cg21161138, located within the AHRR gene, was estimated to have the highest mediation proportion of the effect of smoking on HDL-C in both single-mediator model and multi-mediators models. The site cg21161138 is a well-established smoking-associated DNA methylation marker [38-41]. Previous studies have extensively shown that smoking exposure induces significant hypomethylation at cg21161138, and this hypomethylated state exhibits a strong connection to an elevated risk of cardiovascular diseases, subclinical atherosclerosis, and higher cardiovascular mortality [41, 42]. Hypomethylation of cg05575921, a well-known smoking biomarker that is also located on AHRR, has been shown to outperform self-reported smoking exposure in its association with CVD prevalence [43]. These findings support the interpretation that

cg21161138 may capture a smoking-related epigenetic signature that is also relevant to cardiovascular risk. Given that HDL-C is an important lipid trait related to cardiometabolic health, the prominent mediation signal observed at cg21161138 in the present study may reflect the contribution of smoking-related methylation changes to HDL-C alterations.

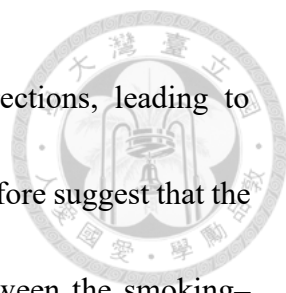
In the present study, some top HDMT-identified CpGs for the current smoking–log(TG) association were annotated to CPT1A. CPT1A has been reported as a lipid-related methylation locus, particularly in relation to TG levels [9, 35, 36]. The identification of these CpGs in the present analysis is consistent with previous epigenome-wide studies and supports the relevance of CPT1A methylation to TG-related metabolic variation. These findings suggest that the HDMT analysis captured CpG-level methylation signals with established relevance to TG regulation, providing biological context for the observed smoking–log(TG) mediation signals.

Among these loci, cg00574958 was additionally identified by HIMA2. Because HIMA2 evaluates mediators in a sparse joint framework, the concordant identification of cg00574958 indicates that its mediation signal remained distinguishable after accounting for other candidate mediators. This provides complementary statistical support for cg00574958 beyond its marginal HDMT signal. Nevertheless, given the cross-sectional design, this concordance should not be interpreted as evidence that CPT1A methylation

causally mediates the effect of smoking on TG.

The prominent CpG-level signals identified in this study were consistent with previous methylation findings for individual components of the smoking–lipid pathway, including the well-established smoking-associated locus *AHRR* and the TG-associated locus *CPT1A*. This concordance provides biological context for the identified mediation signals, while the epigenome-wide approach allowed potential mediators to be evaluated without restriction to preselected candidate loci.

While individual CpG-level findings such as *cg21161138* provide relatively interpretable signals, the PCA-based analysis was used to evaluate the collective contribution of correlated methylation patterns. By reducing dimensionality and mitigating multicollinearity among HDMT-identified CpG sites, PCA captured shared variation across multiple CpGs that may not be adequately represented by individual-site analyses. The PCA-based mediation estimates differed substantially between HDL-C and $\log(\text{TG})$. Whereas the retained methylation components produced a relatively large net estimate for HDL-C (64.74%), the corresponding estimate for $\log(\text{TG})$ was 6.71%. This difference does not necessarily indicate that DNA methylation is biologically important for one lipid trait but not the other. Rather, PCA summarizes multiple correlated methylation signals, and the resulting net estimate depends on the magnitude and direction of the indirect effects represented by the retained components. In the $\log(\text{TG})$

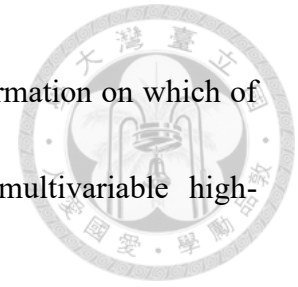


analysis, significant PCs showed indirect effects in opposing directions, leading to substantial cancellation in the summed estimate. These findings therefore suggest that the statistical patterns of methylation-related mediation may differ between the smoking–HDL-C and smoking–log(TG) associations. Given that principal components are statistical summaries of correlated CpG sites and that the present study was cross-sectional, these estimates should be interpreted as exploratory, model-based summaries rather than definitive causal mediation proportions.

The HDMT and HIMA2 analyses address complementary aspects of high-dimensional mediation. HDMT evaluates mediation evidence for each CpG in a marginal framework and is therefore suitable for epigenome-wide identification of CpG-level mediation signals. HIMA2, in contrast, applies variable selection and joint modeling to identify a sparse set of CpGs whose mediation signals remain distinguishable after other candidate mediators are considered simultaneously. Consequently, a larger number of CpGs identified by HDMT than by HIMA2 is not unexpected and does not indicate disagreement between the methods.

This distinction was particularly evident for log(TG), for which HDMT identified 65 potential mediators whereas HIMA2 identified only cg00574958. The HDMT-only CpGs may represent marginal or correlated methylation signals that were not retained under the sparse joint model, whereas cg00574958 showed evidence under both

frameworks. Thus, HIMA2 complements HDMT by providing information on which of the epigenome-wide marginal signals remain detectable in a multivariable high-dimensional mediation model.



Functional enrichment analyses suggested different enrichment patterns for the smoking–HDL-C and smoking–log(TG) associations. The enriched GO terms for HDL-C were mainly related to viral processes and immune responses. These results suggest that HDL-C-related methylation signals may be functionally linked to genes involved in host – virus interaction, antiviral defense, and innate immune signaling. This interpretation may be biologically plausible, as previous studies have suggested that smoking-related DNA methylation changes may contribute to cytokine alterations [44], and that immune activation or inflammation may be associated with reduced HDL-C levels [45, 46]. However, the existing evidence remains relatively indirect, as these studies do not directly demonstrate the full pathway linking smoking, DNA methylation, immune or inflammatory responses, and HDL-C levels. Together, these studies provide partial biological support for a possible connection between HDL-C-related methylation patterns and immune-related biological functions, although the complete pathway remains unestablished.

In contrast to the HDL-C enrichment pattern, the TG-related terms were relatively broad and did not provide a clear biological context specifically related to TG metabolism

or the smoking–log(TG) association.

Although several CpG sites showed statistically significant mediation signals after multiple-testing correction, their biological relevance should also be considered in terms of effect magnitude and biological context. In epigenome-wide analyses, modest effects across multiple CpG sites may collectively reflect broader methylation patterns, whereas individual CpG-level findings alone provide limited information about the underlying functional mechanisms. Thus, the identified CpG sites and enriched pathways support potential epigenetic links between smoking and lipid traits, while further functional and longitudinal evidence is needed to clarify their biological roles.

This study has an important limitation. Because smoking status, DNA methylation, and lipid traits were measured concurrently, the temporal relationship among these variables could not be established. Therefore, causal inference is limited, and reverse causation cannot be excluded.

Despite this limitation, our work offers several notable advantages. To the best of our knowledge, this is the first study to investigate the potential mediating role of DNA methylation in the association between smoking and lipid traits in an epigenome-wide framework. In addition, this study contributes evidence from a Taiwanese population, which remains underrepresented in the current literature.

Conclusion



In conclusion, this study identified multiple CpG sites that may mediate the association between smoking and lipid traits in the Taiwan Biobank. Among them, cg21161138 in AHRR and several CpG sites in CPT1A were notable findings. The enrichment analysis further identified immune-, antiviral-, and kinase-related GO terms. Together, these results provide statistical evidence supporting a potential role of DNA methylation in the association between smoking and lipid traits. Further longitudinal and functional studies are needed to clarify the underlying biological mechanisms.

Tables

Table 1. Basic characteristics of the 2,463 Taiwan Biobank participants stratified by smoking status

	Never smokers	Former smokers ^a	Current smokers ^b
Total	1870 (100%)	310 (100%)	283 (100%)
HDL-C (mg/dL)	55.6 (13.7)	49.6 (11.9)	46.2 (11.7)
LDL-C (mg/dL)	121.4 (32.5)	122.3 (32.6)	121.3 (31.8)
TG (mg/dL)	111.0 (101.7)	123.7 (81.8)	160.8 (144.3)
T-CHO (mg/dL)	195.5 (35.2)	193.0 (35.8)	191.9 (33.8)
BMI (kg/m ²)	24.1 (3.7)	25.3 (3.2)	25.4 (3.7)
Educational attainment ^c (ordinal score)	5.6 (0.9)	5.6 (0.9)	5.4 (0.8)
Female	1136 (60.7%)	39 (12.6%)	48 (17.0%)
Male	734 (39.3%)	271 (87.4%)	235 (83.0%)
Current drinking ^d (Yes / No)	72 (3.9%)	30 (9.7%)	69 (24.4%)
Regular exercise ^e (Yes / No)	815 (43.6%)	178 (57.4%)	95 (33.6%)

Data are presented as *n* (%) or mean (standard deviation)

^aFormer smokers were those who had a history of smoking for at least six months and had quit smoking for at least six months prior to their enrollment.

^bCurrent smokers were defined as participants who had smoked for at least six months and had not quit smoking at the time of enrollment.

^cEducational attainment was treated as an ordinal variable ranging from 1 to 7, where 1 indicated illiteracy, 2 denoted literacy without formal education, 3 through 6 represented graduates of primary, junior

high, senior high, and college/university levels, respectively, and 7 signified a Master's or higher degree.

^dDrinking status was considered positive if the participant consumed more than 150 mL of alcoholic beverages per week for at least six months and had not stopped drinking at the time of enrollment.

^eParticipants who engaged in exercise at least three times a week, with each session lasting more than 30 minutes, were defined as having a positive regular exercise status. Exercise included leisure-time activities such as badminton, tennis, swimming, yoga, jogging, cycling, mountain climbing, and dancing.

Figures

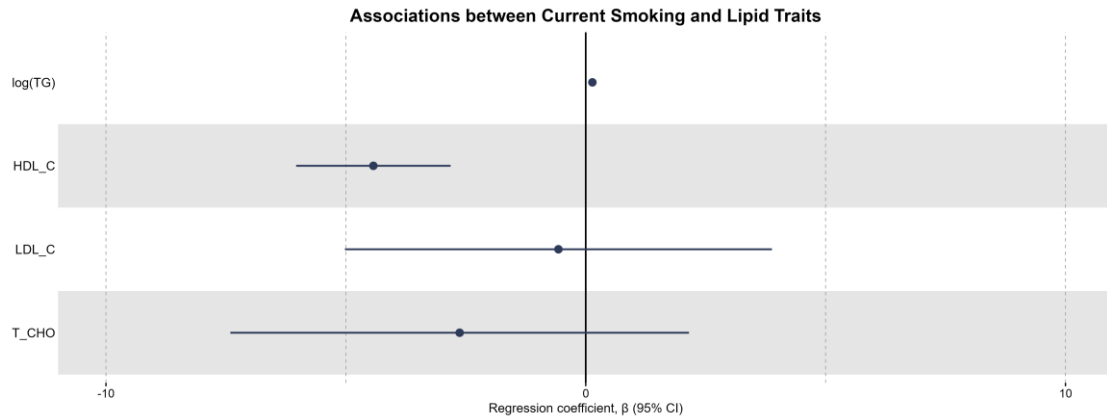


Figure 1. Adjusted associations between current smoking and lipid traits.

Age, sex, BMI, drinking status, regular exercise status, educational level, and 11 estimated blood cell type proportions were included as covariates. Current smoking shows significant association with lower HDL-C and higher TG levels ($P < 0.05$). Lipid traits were analyzed on their original scale (mg/dL), except for TG, which was log-transformed prior to analysis.

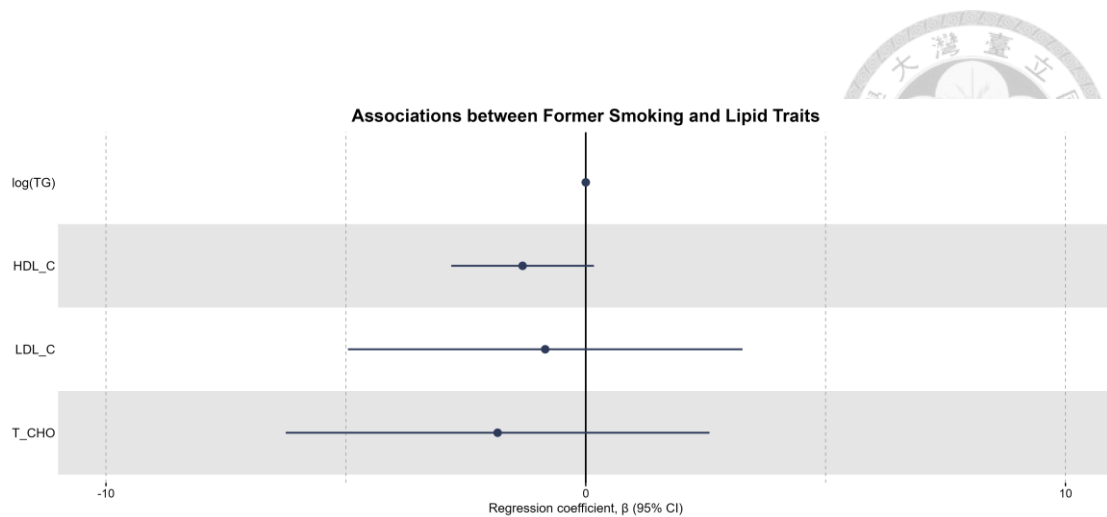


Figure 2. Adjusted associations between former smoking and lipid traits.

Age, sex, BMI, drinking status, regular exercise status, educational level, and 11 estimated blood cell type proportions were included as covariates. Former smoking shows no significant association with any lipid traits ($P > 0.05$). Lipid traits were analyzed on their original scale (mg/dL), except for TG, which was log-transformed prior to analysis.

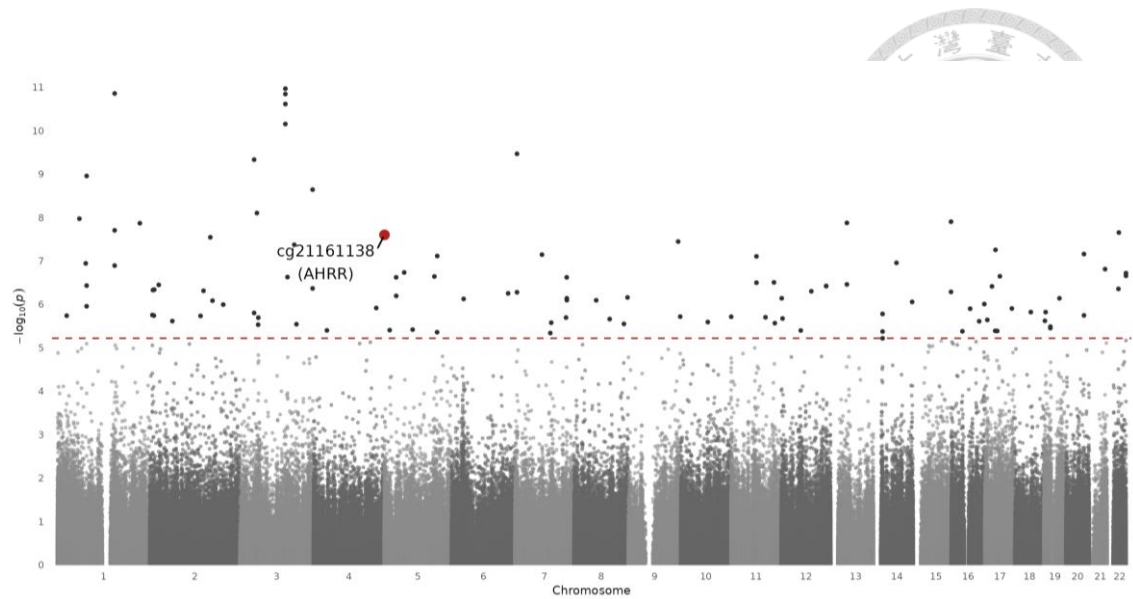


Figure 3. Manhattan Plot of HDMT Results for the Current Smoking-HDL-C Association

Manhattan plot of HDMT results for CpG-level mediation analysis of the current smoking-HDL-C association. Each point represents a CpG site arranged by chromosomal location. The y-axis corresponds to $-\log_{10}(\text{P-value})$ from the HDMT single-CpG mediation test, and the horizontal dashed threshold designates $\text{FDR} = 0.05$. The highlighted CpG, cg21161138, is annotated to AHRR.

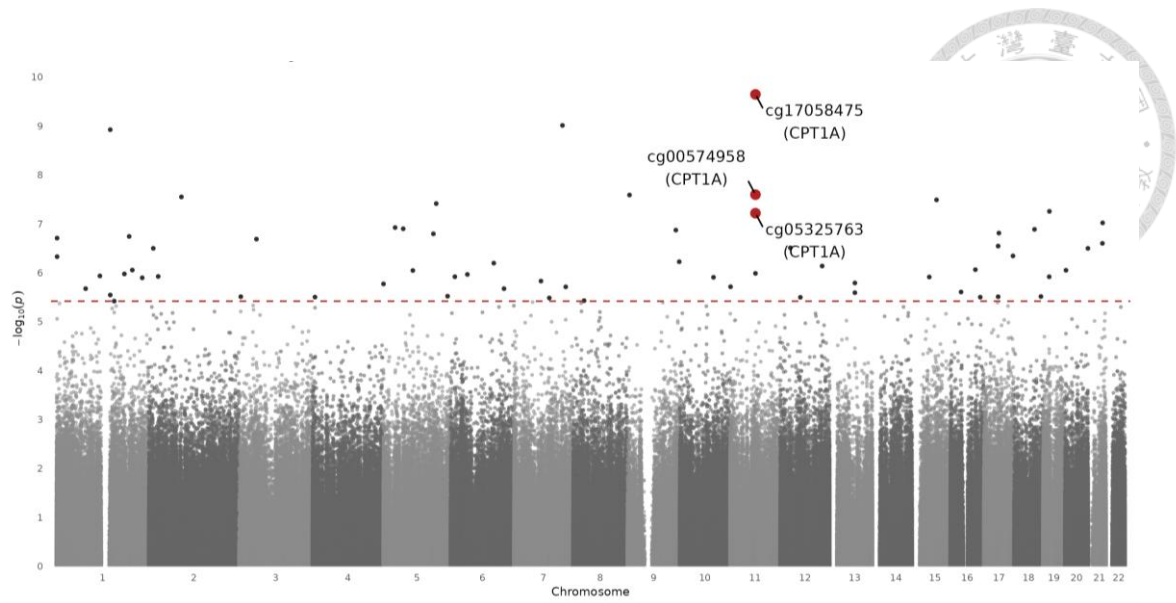


Figure 4. Manhattan Plot of HDMT Results for the Current Smoking–log(TG) Association

Manhattan plot of HDMT results for CpG-level mediation analysis of the current smoking–log(TG) association. Each point represents a CpG site arranged by chromosomal location. The y-axis corresponds to $-\log_{10}(\text{P-value})$ from the HDMT single-CpG mediation test, and the horizontal dashed threshold designates $\text{FDR} = 0.05$. The highlighted CpGs, cg17058475, cg00574958, and cg05325763, are annotated to CPT1A.

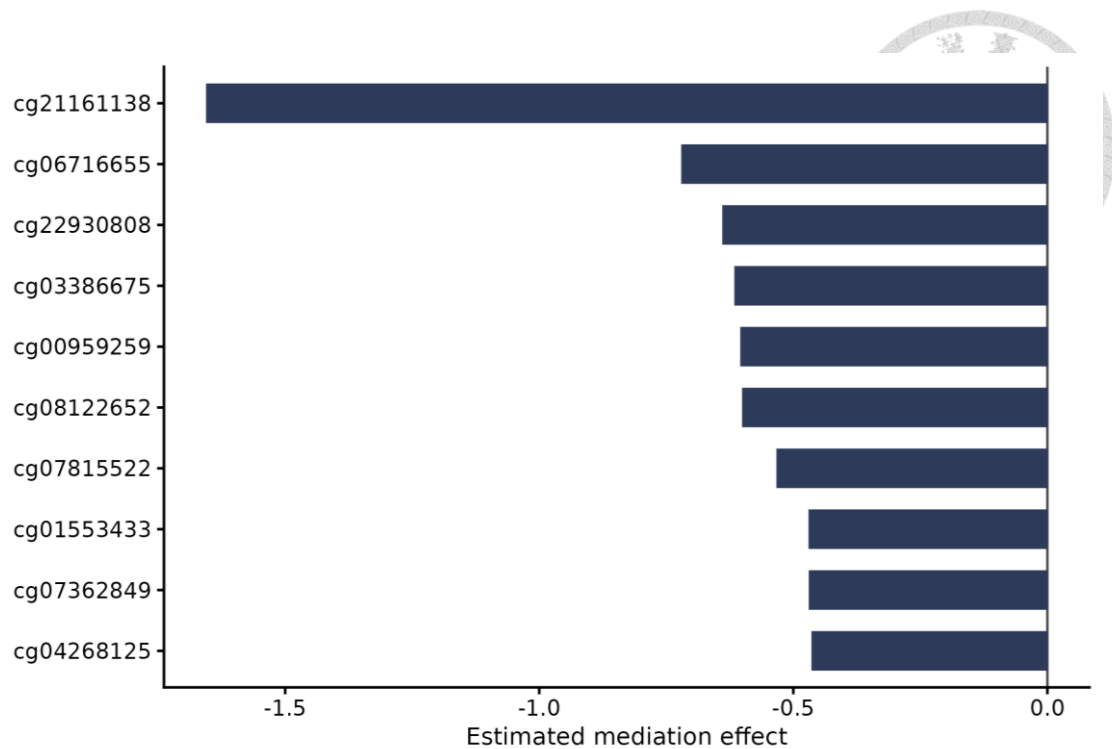


Figure 5. Estimated Mediation Effects of HDMT-identified CpGs for the Current Smoking–HDL-C Association

Effect size bar chart of HDMT-identified CpG mediators for the current smoking–HDL-C association. The top 10 CpG sites with the largest absolute estimated mediation effects are shown. Bar length represents the estimated mediation effect ($\alpha \times \beta$), and bar direction indicates whether the estimated mediation effect is positive or negative. Among these CpGs, cg21161138, annotated to AHRR, showed the largest estimated mediation effect.

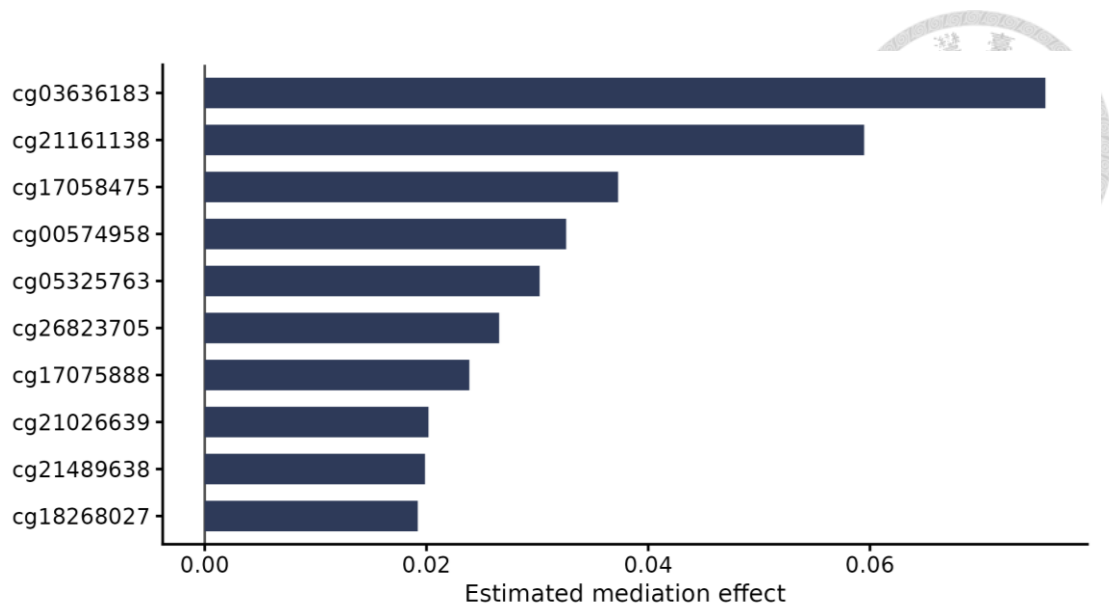


Figure 6. Estimated Mediation Effects of HDMT-identified CpGs for the Current Smoking–log(TG) Association

Effect size bar chart of HDMT-identified CpG mediators for the current smoking–log(TG) association. The top 10 CpG sites with the largest absolute estimated mediation effects are shown. Bar length represents the estimated mediation effect ($\alpha \times \beta$), and bar direction indicates whether the estimated mediation effect is positive or negative. Several CpGs with relatively large estimated mediation effects were annotated to CPT1A, including cg17058475, cg00574958, and cg05325763.

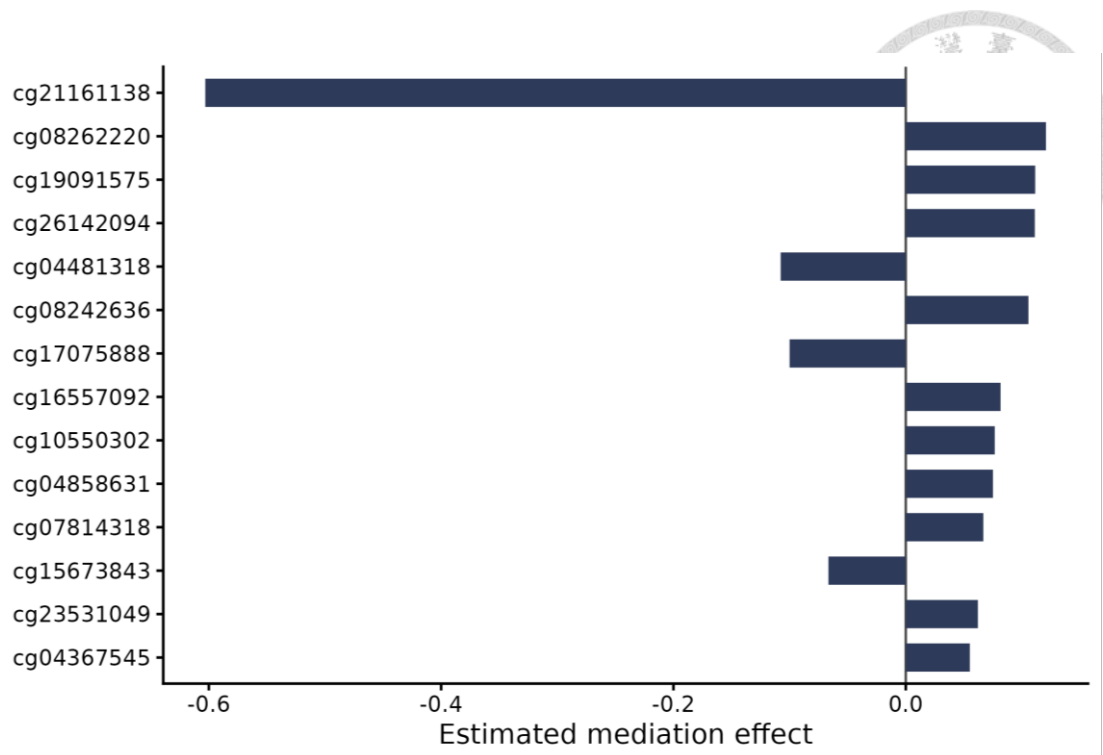


Figure 7. Estimated Mediation Effects of HIMA2-identified CpGs for the Current Smoking–HDL-C Association

Effect size bar chart of independent CpG mediators identified by HIMA2 for the current smoking–HDL-C association. Bar length represents the estimated mediation effect, and bar direction indicates whether the estimated mediation effect is positive or negative. Among the HIMA2-identified CpGs, cg21161138, annotated to AHRR, showed the largest estimated mediation effect.

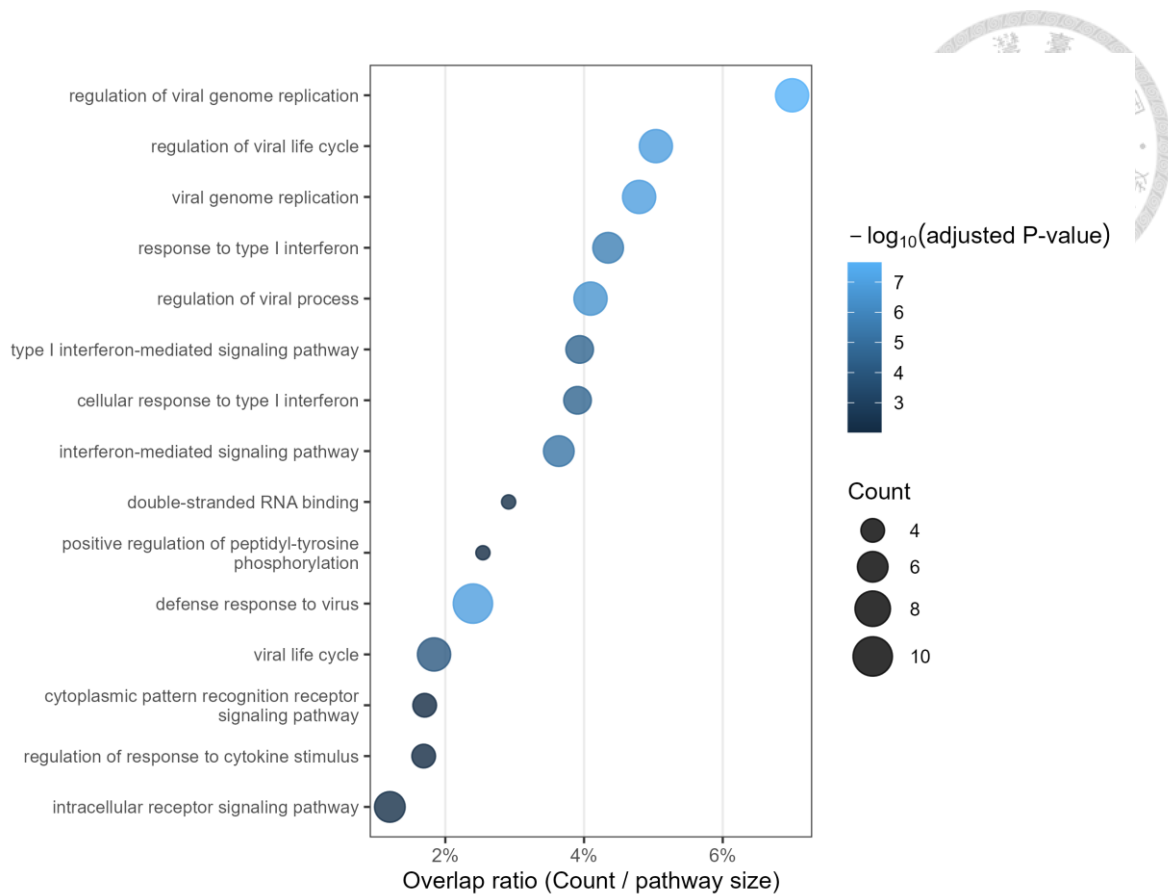


Figure 8. GO enrichment analysis for the current smoking–HDL-C association

The enrichment analysis further identified immune-, antiviral-related GO terms. The plot displays the top 15 enriched GO terms ranked by the lowest P-values. Enriched GO pathways are listed on the y-axis, whereas the x-axis indicates the overlap ratio. Node size corresponds to the count of overlapping genes, with color gradient $-\log_{10}(\text{adjusted P-value})$. The enriched terms were mainly related to viral processes, interferon response, and antiviral immune pathways.

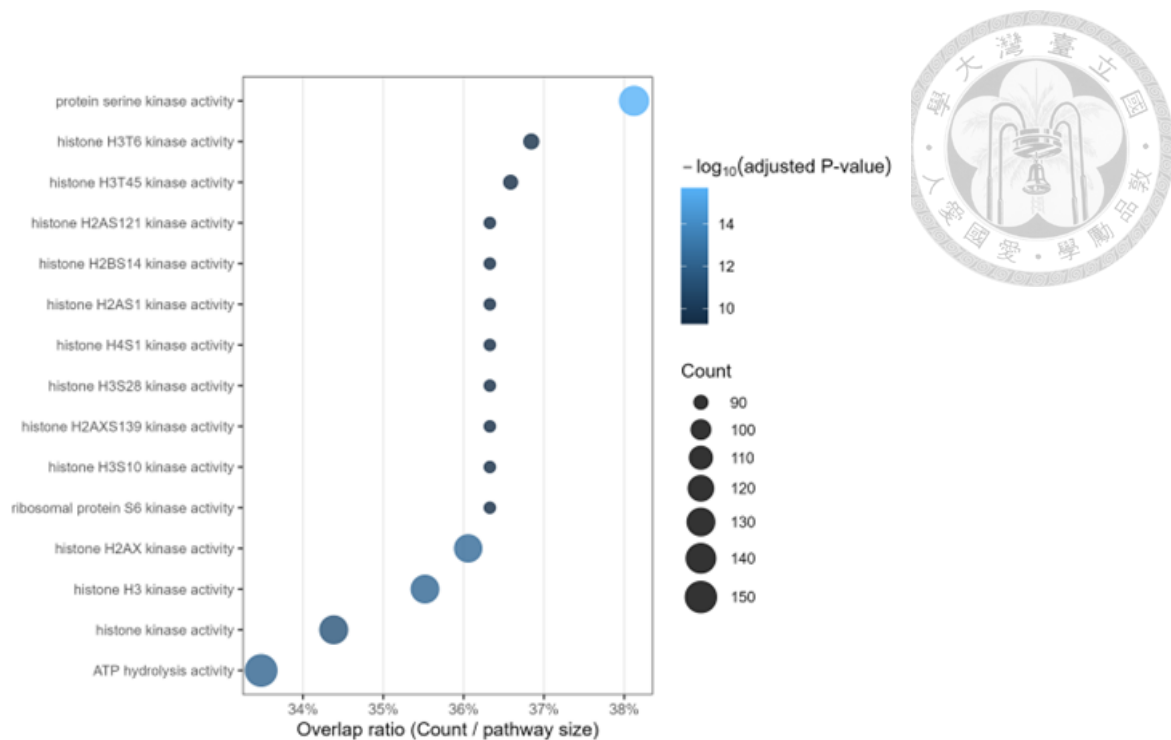


Figure 9. GO enrichment analysis for the current smoking–log(TG) association

The enrichment analysis further identified kinase-related GO terms. The plot displays the top 15 enriched GO terms ranked by the lowest P-values. Enriched GO pathways are listed on the y-axis, whereas the x-axis indicates the overlap ratio. Node size corresponds to the count of overlapping genes, with color gradient $-\log_{10}(\text{adjusted P-value})$. The enriched terms were mainly related to kinase activity, histone kinase activity, and ATP hydrolysis activity.

References



1. Ding N, Sang Y, Chen J, Ballew SH, Kalbaugh CA, Salameh MJ, Blaha MJ, Allison M, Heiss G, Selvin E: **Cigarette smoking, smoking cessation, and long-term risk of 3 major atherosclerotic diseases.** *Journal of the American College of Cardiology* 2019, **74**:498 – 507.
2. Ishida M, Sakai C, Kobayashi Y, Ishida T: **Cigarette smoking and atherosclerotic cardiovascular disease.** *Journal of atherosclerosis and thrombosis* 2024, **31**:189–200.
3. Joseph P, Leong D, McKee M, Anand SS, Schwalm J-D, Teo K, Mente A, Yusuf S: **Reducing the global burden of cardiovascular disease, part 1: the epidemiology and risk factors.** *Circulation research* 2017, **121**:677–694.
4. Nakamura M, Yamamoto Y, Imaoka W, Kuroshima T, Toragai R, Ito Y, Kanda E, Schaefer EJ, Ai M: **Relationships between smoking status, cardiovascular risk factors, and lipoproteins in a large Japanese population.** *Journal of atherosclerosis and thrombosis* 2021, **28**:942–953.
5. Opoku S, Gan Y, Fu W, Chen D, Addo-Yobo E, Trofimovitch D, Yue W, Yan F, Wang Z, Lu Z: **Prevalence and risk factors for dyslipidemia among adults in rural and urban China: findings from the China National Stroke Screening and prevention project (CNSSPP).** *BMC public health* 2019, **19**:1500.
6. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan W, Xu T, Elks CE, Aslibekyan S, et al: **Epigenetic Signatures of Cigarette Smoking.** *Circ Cardiovasc Genet* 2016, **9**:436–447.
7. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H: **DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies.** *Clinical epigenetics* 2015, **7**:113.
8. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, Belvisi MG, Brown R, Vineis P, Flanagan JM: **Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking.** *Hum Mol Genet* 2013, **22**:843–851.
9. Braun KVE, Dhana K, de Vries PS, Voortman T, van Meurs JBJ, Uitterlinden AG, Hofman A, Hu FB, Franco OH, Dehghan A, consortium B:

Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clinical Epigenetics* 2017, **9**:15.

10. Sayols-Baixeras S, Subirana I, Lluís-Ganella C, Civeira F, Roquer J, Do AN, Absher D, Cenarro A, Muñoz D, Soriano-Tárraga C, et al: **Identification and validation of seven new loci showing differential DNA methylation related to serum lipid profile: an epigenome-wide approach. The REGICOR study.** *Hum Mol Genet* 2016, **25**:4556–4565.

11. Wei CY, Yang JH, Yeh EC, Tsai MF, Kao HJ, Lo CZ, Chang LP, Lin WJ, Hsieh FJ, Belsare S, et al: **Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese.** *NPJ Genom Med* 2021, **6**:10.

12. Lin W-Y: **Epigenome-wide mediation analysis identified Cytosine-phosphate-Guanine sites linking environmental factors with diabetes indicators.** *Epigenomics* 2026, **18**:45–54.

13. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD: **Low-level processing of Illumina Infinium DNA Methylation BeadArrays.** *Nucleic Acids Res* 2013, **41**:e90.

14. Fortin J-P, Triche Jr TJ, Hansen KD: **Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi.** *Bioinformatics* 2017, **33**:558–560.

15. Maksimovic J, Phipson B, Oshlack A: **A cross-package Bioconductor workflow for analysing methylation array data.** *F1000Research* 2016, **5**:1281.

16. Fang H, Distech CM, Berletch JB: **X Inactivation and Escape: Epigenetic and Structural Features.** *Frontiers in Cell and Developmental Biology* 2019, **7**:219.

17. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, Lin SM: **Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis.** *BMC bioinformatics* 2010, **11**:587.

18. Bland JM, Altman DG: **Transformations, means, and confidence intervals.** *Bmj* 1996, **312**:1079.

19. Feng L, Nian S, Tong Z, Zhu Y, Li Y, Zhang C, Bai X, Luo X, Wu M, Yan Z: **Age-related trends in lipid levels: a large-scale cross-sectional study of the general Chinese population.** *BMJ open* 2020, **10**:e034226.

20. Holven KB, van Lennep JR: **Sex differences in lipids: a life course approach.** *Atherosclerosis* 2023, **384**:117270.

21. Lu S, Xie Q, Kuang M, Hu C, Li X, Yang H, Sheng G, Xie G, Zou Y: **Lipid metabolism, BMI and the risk of nonalcoholic fatty liver disease in the**

general population: evidence from a mediation analysis. *Journal of translational medicine* 2023, **21**:192.

22. Vu KN, Ballantyne CM, Hoogeveen RC, Nambi V, Volcik KA, Boerwinkle E, Morrison AC: **Causal role of alcohol consumption in an improved lipid profile: the atherosclerosis risk in communities (ARIC) study.** *PloS one* 2016, **11**:e0148765.

23. Gordon B, Chen S, Durstine JL: **The effects of exercise training on the traditional lipid profile and beyond.** *Curr Sports Med Rep* 2014, **13**:253–259.

24. Tillmann T, Vaucher J, Okbay A, Pikhart H, Peasey A, Kubinova R, Pajak A, Tamosiunas A, Malyutina S, Hartwig FP: **Education and coronary heart disease: mendelian randomisation study.** *bmj* 2017, **358**:j3542.

25. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT: **DNA methylation arrays as surrogate measures of cell mixture distribution.** *BMC Bioinformatics* 2012, **13**:86.

26. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, Wiencke JK, Kelsey KT, Christensen BC: **Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling.** *Nature Communications* 2022, **13**:761.

27. Zheng SC, Breeze CE, Beck S, Dong D, Zhu T, Ma L, Ye W, Zhang G, Teschendorff AE: **EpiDISH web server: Epigenetic Dissection of Intra-Sample-Heterogeneity with online GUI.** *Bioinformatics* 2019, **36**:1950–1951.

28. Zeng P, Shao Z, Zhou X: **Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges.** *Comput Struct Biotechnol J* 2021, **19**:3209–3224.

29. Dai JY, Stanford JL, LeBlanc M: **A multiple-testing procedure for high-dimensional mediation hypotheses.** *Journal of the American Statistical Association* 2022, **117**:198 – 213.

30. Thissen D, Steinberg L, Kuang D: **Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons.** *Journal of educational and behavioral statistics* 2002, **27**:77–83.

31. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K: **mediation: R Package for Causal Mediation Analysis.** *Journal of Statistical Software* 2014, **59**:1–38.

32. Perera C, Zhang H, Zheng Y, Hou L, Qu A, Zheng C, Xie K, Liu L: **HIMA2: high-dimensional mediation analysis and its application in**

epigenome-wide DNA methylation data. *BMC bioinformatics* 2022, **23**:296.

33. Ren X, Kuan PF: **methylGSA: a Bioconductor package and Shiny app for DNA methylation data length bias adjustment in gene set testing.** *Bioinformatics* 2019, **35**:1958–1959.

34. Kolde R, Laur S, Adler P, Vilo J: **Robust rank aggregation for gene list integration and meta-analysis.** *Bioinformatics* 2012, **28**:573–580.

35. Gagnon F, Aïssi D, Carrié A, Morange PE, Tréguët DA: **Robust validation of methylation levels association at CPT1A locus with lipid plasma levels.** *J Lipid Res* 2014, **55**:1189–1191.

36. Irvin MR, Zhi D, Joehanes R, Mendelson M, Aslibekyan S, Claas SA, Thibeault KS, Patel N, Day K, Jones LW, et al: **Epigenome-wide association study of fasting blood lipids in the Genetics of Lipid-lowering Drugs and Diet Network study.** *Circulation* 2014, **130**:565–572.

37. Maas SC, Mens MM, Kühnel B, van Meurs JB, Uitterlinden AG, Peters A, Prokisch H, Herder C, Grallert H, Kunze S: **Smoking-related changes in DNA methylation and gene expression are associated with cardio-metabolic traits.** *Clinical epigenetics* 2020, **12**:157.

38. Philibert R, Beach S, Lei M, Brody G: **Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking.** *Clinical Epigenetics* 2013, **5**:19.

39. Zhang Y, Schöttker B, Florath I, Stock C, Butterbach K, Holleczeck B, Mons U, Brenner H: **Smoking-Associated DNA Methylation Biomarkers and Their Predictive Value for All-Cause and Cardiovascular Mortality.** *Environmental Health Perspectives* 2016, **124**:67–74.

40. Wen D, Shi J-W, Liu Y, He W, Qu W, Wang C, Xing H, Cao Y, Li J, Zha L: **DNA methylation analysis for smoking status prediction in the Chinese population based on the methylation-sensitive single-nucleotide primer extension method.** *Forensic science international* 2022, **339**:111412.

41. Philibert R, Moody J, Philibert W, Dogan M, Hoffman E: **The Reversion of the Epigenetic Signature of Coronary Heart Disease in Response to Smoking Cessation.** *Genes* 2023, **14**:1233.

42. Reynolds LM, Wan M, Ding J, Taylor JR, Lohman K, Su D, Bennett BD, Porter DK, Gimple R, Pittman GS, et al: **DNA Methylation of the Aryl Hydrocarbon Receptor Repressor Associations With Cigarette Smoking and Subclinical Atherosclerosis.** *Circ Cardiovasc Genet* 2015, **8**:707–716.

43. Jones GT, Williams MJA, Khashram M, Fitzgerald S, Lyons OTA, Griffin CL, Lesche S, Krysa J, Coffey S: **A DNA Methylation Marker, cg05575921 (AHRR), Outperforms Self-Reported Smoking Exposure for Its**

Association With Cardiovascular Disease Prevalence. *Nicotine Tob Res* 2026, **28**:422–428.

44. Saint-André V, Charbit B, Biton A, Rouilly V, Possémé C, Bertrand A, Rotival M, Bergstedt J, Patin E, Albert ML, et al: **Smoking changes adaptive immunity with persistent effects.** *Nature* 2024, **626**:827–835.

45. Su X, Zhang G, Cheng Y, Wang B: **New insights into the emerging effects of inflammatory response on HDL particles structure and function.** *Molecular Biology Reports* 2021, **48**:5723–5733.

46. Trakaki A, Marsche G: **Current Understanding of the Immunomodulatory Activities of High-Density Lipoproteins.** *Biomedicines* 2021, **9**:587.