A trans-omics gene-smoking interaction study of lung cancer based on consortium data

Ning Xie[1*], Xiaowen Xu[1*], Yanru Wang[1*], Aoxuan Wang[1*], Xiang Wang[1*], Xuan Wang[1*], Mengshen Zhao[1], Jiacheng Zhou[1], Yongyue Wei[2,3], Manel Esteller[4,5,6,7], Zhibin Hu[8,9,10,11†], Hongbing Shen[8,9,10†], Rayjean J. Hung[12†], Christopher I. Amos[13†], Yi Li[14†], David C Christiani[15,16†], Feng Chen[1,9,10‡], Yang Zhao[1,9‡], Ruyang Zhang[1,9,17,18‡]

[1]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, Jiangsu 211166, China

[2]Center for Public Health and Epidemic Preparedness & Response, Peking University, Beijing, 100191, China

[3]Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, 100191, China

[4]Josep Carreras Leukaemia Research Institute, Barcelona, Catalonia, 08021, Spain

[5]Centro de Investigacion Biomedica en Red Cancer, Madrid, 28029, Spain

[6]Institucio Catalana de Recerca i Estudis Avanc¸ats, Barcelona, Catalonia, 08010, Spain

[7]Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona, Barcelona, Catalonia, 08007, Spain

[8]Department of Epidemiology, School of Public Health, Key Laboratory of Public Health Safety and Emergency Prevention and Control Technology of Higher Education Institutions in Jiangsu Province, Nanjing Medical University, Nanjing, Jiangsu 211166, China

[9]China International Cooperation Center (CICC) for Environment and Human Health, Nanjing Medical University, Nanjing, Jiangsu 211166, China

[10]Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Cancer Center, Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, Jiangsu 211166, China

[11]State Key Laboratory of Reproductive Medicine, Nanjing Medical University, Nanjing, Jiangsu 211166, China

[12]Lunenfeld-Tanenbaum Research Institute, Sinai Health, and University of Toronto, Toronto, ON, Canada M5G 1X5

[13]The Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX 77030, USA

[14]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA

[15]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

[16]Pulmonary and Critical Care Division, Department of Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA

[17]Changzhou Medical Center, Nanjing Medical University, Changzhou, Jiangsu 213164, China

[18]Information Center, The Second People's Hospital of Changzhou, the Third Affiliated Hospital of Nanjing Medical University, Changzhou, Jiangsu, China, 213164

*These authors contributed equally as co-first authors of this work.

†These authors contributed equally as co-senior authors of this work.

‡These authors are corresponding authors.

Corresponding authors:

Ruyang Zhang, PhD

Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, SPH Building Room 406, 101 Longmian Avenue, Nanjing, Jiangsu 211166, China.

Tel: +86-025-86868436; E-mail: zhangruyang@njmu.edu.cn

Yang Zhao, PhD

Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, SPH Building Room 402, 101 Longmian Avenue, Nanjing, Jiangsu 211166, China.

Tel: +86-025-86868438; E-mail: zhaoyang@njmu.edu.cn

Feng Chen, PhD

Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing

Medical University, SPH Building Room 412, 101 Longmian Avenue, Nanjing, Jiangsu 211166, China.

Tel: +86-025-86868414; E-mail: fengchen@njmu.edu.cn

Author contributions:

R.Z., Y.Z., F.C., D.C.C., Y.L., C.I.A. and R.J.H. were responsible for study conception and study design; N.X., X.X., Y. Wang, A.W., Xuan Wang and M.Z. contributed to data analyses and interpretation; Xiang Wang and Xuan Wang validated all analytical procedures and results; J.Z. and Y. Wei developed the online platform; N.X., X.X. and R.Z. wrote the manuscript; Y. Wei, M.E., Z.H. and H.S. critically reviewed the manuscript and results. All authors read the manuscript and agreed on the final version.

Funding:

This study was supported by the National Natural Science Foundation of China (82220108002 to F.C., 82273737 to R.Z., 82373690 to Y.Z., 82473728 to Y.W.W.), Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0520000 and 2024ZD0520003 to R.Z.), the US National Institutes of Health (CA209414, CA249096, CA092824, and ES000002 to D.C.C., CA 249096 and CA209414 to Y.L.), Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). R.Z. was partially supported by the Outstanding Young Teachers Training Program of Nanjing Medical

University.

Running title: Trans-omics gene-smoking interaction in lung cancer

Descriptor number: 6.6 Gene-Environment Interaction

Word count (text): 4571 Words.

At a Glance Commentary

Scientific Knowledge on the Subject:

Smoking is the well-recognized major environmental exposure for lung cancer, however, only a fraction of smokers develop the disease. This discrepancy points to a specific type of individual susceptibility, where genetic variant and environmental exposure synergistically or antagonistically affect lung cancer risk. Unraveling these gene-smoking interactions is critical for precise intervention of targeted smokers having very high risk of lung cancer, moving beyond a common way to understanding personal vulnerability.

What This Study Adds to the Field:

We launch a free online platform, LungCancer-xWAS-GxE, through conducting the first trans-omics gene-smoking interaction study of lung cancer by integrating consortium-scale individual genotype data and with alliance-based summary-level molecular quantitative trait loci (xQTL) data, involving DNA methylation, gene expression, protein, and metabolite. We release 0.27 million gene-smoking interaction signals, pinpoint eight key biomarkers that modify effect of smoking on lung cancer risk, and weave them into a novel Molecular Modifying Score (MMS). The MMS effectively stratifies lung cancer risk among smokers adjusted for age, sex and pack-year of smoking, providing a practical tool to precisely target and intervene the smokers at high risk of lung cancer.

Artificial Intelligence Disclaimer: No artificial intelligence tools were used in writing this manuscript.

This article has an online data supplement, which is accessible at the Supplements tab.

Abstract

Rationale: Genetically predicted molecular traits provide a cost-effective approach for identifying biomarkers and uncovering underlying biological mechanisms. We extended this framework to investigate gene-smoking interactions in lung cancer susceptibility.

Objectives: To identify trans-omics gene-smoking interactions affecting lung cancer risk and to assess how biomarkers modify effect of smoking.

Methods: We conducted the first trans-omics gene-smoking interaction study of lung cancer by integrating consortium-scale individual genotype data (27,737 cases *vs* 449,910 non-cases) from the International Lung Cancer OncoArray Consortium (ILCCO-OncoArray), Transdisciplinary Research Into Cancer of the Lung (TRICL), Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO), and the UK Biobank (UKB) with alliance-based summary-level molecular quantitative trait loci (xQTL) data, involving DNA methylation, gene expression, protein, and metabolite. Based on the identified biomarkers, we developed a molecular modifying score (MMS) to delineate gene-smoking interaction patterns and stratify high-risk smokers of lung cancer.

Measurements and Main Results: Eight biomarkers showing significant interactions with smoking were identified through a two-phase analytic strategy, comprising CpG sites in the nicotinic acetylcholine receptor region and gene *RP11-326C3.14*. The MMS, constructed by integrating these biomarkers with their effect estimates derived from meta-analysis of all available datasets, effectively stratified lung cancer risk among smokers. Trans-omics integrative analysis revealed functional relationships across molecular layers, particularly implicating the *NELFE* gene in smoking-related carcinogenesis pathways.

Conclusions: The xWAS framework enables systematic discovery of trans-omics gene-environment interactions. The MMS effectively delineates the patterns of the interaction effects and facilitates risk stratification. Additionally, we launched a free online platform, LungCancer-xWAS-GxE (http://bigdata.njmu.edu.cn/LungCancer-xWAS-GxE/).

Word count (abstract): 249

Key words: gene-smoking interaction, genome-wide association study, lung cancer, trans-omics, xWAS.

Introduction

Lung cancer is one of the most prevalent and deadly malignancies worldwide, accounting for approximately 1.8 million deaths annually(1). It is reported that mortality burden of lung cancer could rise to 3 million deaths by 2040, largely driven by the ongoing smoking epidemic(2). Tobacco smoke contains numerous carcinogens that cause various forms of DNA damage(3), exerting both cytotoxic and carcinogenic effects on bronchial and lung epithelial cells(4-6). Notably, only about 10% of smokers develop lung cancer in their lifetime, suggesting substantial inter-individual variability in susceptibility to tobacco-induced lung cancer across populations with different genetic predispositions(7, 8). This variability can be attributed to a potential role for gene-smoking interactions, where the joint effects of genetic variants and smoking exposure differ from the sum of their independent effects(9).

The critical importance of gene-environment (G×E) interactions in the development of complex diseases has gained increasing recognition. Numerous studies have demonstrated that genetic factors could modulate the effects of environmental exposure(10), which help explain part of the "missing heritability"(11). Therefore, investigating G×E interaction is essential for advancing precision prevention strategies and for identifying high-risk populations.

The advent of high-density genotyping has facilitated genome-wide association studies (GWASs), enabling the identification of genetic variants and G×E interactions associated with disease risk on a genome-wide scale(12-15). Our previous study identified significant gene-smoking interactions associated with lung cancer susceptibility(8, 16, 17), while our other

studies have reported that DNA methylation sites could modify lung cancer prognosis(18-20).

Given the high cost and logistical challenges of large-scale trans-omics profiling, integrating existing GWAS data with molecular quantitative trait loci (xQTLs) data provides a promising and cost-effective approach for biomarkers discovery and for elucidating underlying biological mechanisms within the trans-omics association study (xWAS) framework(21). While proteome-wide association study (PWAS)(22) and transcriptome-wide association study (TWAS)(23, 24) have successfully identified molecular phenotypes associated with complex traits, few studies have comprehensively evaluated G×E interactions across multiple omics layers.

Therefore, we conducted the first xWAS of lung cancer by integrating alliance-based summary-level xQTL data with the consortium-scale individual-level genotype data, and extended it to G×E interaction analysis. Through a trans-omics integrative approach encompassing epigenome-wide association study (EWAS), TWAS, PWAS and metabolome-wide association study (MWAS), we aimed to investigate the roles of biomarkers in modifying the relationship between smoking and lung cancer risk. Using a two-phase study design (discovery and replication), we exclusively evaluated the gene-smoking interactions across multi-omics molecular biomarkers, including DNA methylation, gene expression, protein and metabolite. Based on the significant findings, we further quantified the modifying effects of molecular

biomarkers on the association between smoking and lung cancer.

Methods

The overall principle and design flowchart of the study are presented in Figure 1. First, we tested gene-smoking interaction terms, defined as the product of the "imputed" molecular trait and smoking exposure (measured in pack-years) for lung cancer susceptibility. The "imputed" molecular phenotypes were computed by integrating consortium-scale individual-level genotype data with summary-level xQTL data. We implemented a rigorous two-phase analytical design to validate the statistically significant interactions. Next, we additionally assessed the alternative smoking exposure measured in smoking status and conducted a series of subgroup analyses for sensitivity analyses. Finally, a molecular modifying score (MMS) was developed, and we conducted an integrative analysis to assess various biomarker results from different omics layers.

- ***Consortium-scale individual GWAS data***

Individual-level genotype data were obtained from four datasets, including the International Lung Cancer OncoArray Consortium (ILCCO-OncoArray)(25), Transdisciplinary Research Into Cancer of the Lung (TRICL)(26), Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)(27), and the UK Biobank (UKB)(28). Detailed information of each study and the quality control (QC) procedures applied are provided in the online data supplement. Only genetically unrelated individuals of European ancestry were retained for

analysis. In cohort studies, only participants without any cancer diagnosis at baseline were considered eligible and incident lung cancers during active follow-up (median follow-up time: 6.4 years) were defined as cases. All cohorts obtained informed consent from participants, and ethical approval was granted by relevant institutional review boards.

- ***Alliance-based summary-level QTL data***

Summary-level QTL data were profiled across four molecular layers: DNA methylation (meQTL), gene expression (eQTL), protein (pQTL), and metabolite (metabQTL). Specifically, meQTL data were derived from the GoDMC consortium (*n*=27,750)(29), eQTL from eQTLGen (*n*=31,684)(30), pQTL from Ferkingstad et al. (*n*=35,559)(31), and metabQTL from Shin et al. (*n*=7,824)(32). All data were collected from whole blood, peripheral blood, or cord blood specimens. Detailed information is provided in the online data supplement.

- ***Molecular trait prediction models***

The xWAS framework contains two steps to identify lung cancer associated gene-smoking interactions. First, we imputed molecular traits by computing polygenic scores for each biomarker using LDpred2(33), a Bayesian shrinkage method that selects SNPs and estimates functional weights by integrating xQTL summary statistics with linkage disequilibrium (LD)

matrix. The polygenic scores for each molecular trait were calculated as **Equation 1**.

$$\tilde{X} = \sum \beta_{G_i} G_i . \quad (1)$$

Where G_i represents the genotype of the selected variant i, and β_{G_i} denotes the corresponding effect sizes from the xQTL data. Only imputed biomarkers with significant heritability (posterior heritability $h^2 \geq 0.01$) were retained for subsequent association testing. Further methodological details are available in the online data supplement.

- ***Association testing for xWAS***

Second, we used a logistic regression model adjusted for covariates to test the gene-smoking interaction. To ensure the robustness and generalizability of the results, we implemented a two-phase analytic strategy. ILCCO-OncoArray was included in the discovery phase, while TRICL, PLCO and UKB were used for replication. To address sample overlap between ILCCO-OncoArray and TRICL (n = 3,669), the overlapping individuals were assigned to TRICL.

The association tests were conducted using logistic regression models, adjusted for age, sex, genotyping platform (see online data supplement), and the first ten genetic principal components, to assess G×E interaction (**Equation 2**). The biomarkers were standardized in the regression model prior to analysis.

$$\begin{aligned} \text{logit}(\pi) = \beta_0 + \beta_X \times \tilde{X} + \beta_E \times \text{pack-year} + \\ \beta_{XE} \times \tilde{X} \times \text{pack-year} + \sum \beta_i \times \text{covariate}_i \end{aligned} \quad (2)$$

Where π represents the individual's probability of lung cancer. Gene-smoking interactions were considered statistically significant only if they met the following criteria: (1) Benjamini

& Hochberg (BH) procedure based false discovery rate (q-FDR) ≤ 0.05 in the discovery phase, and (2) consistent effect direction with $P \leq 0.05$ in the replication phase. In order to derive robust estimates of gene-smoking interactions, a fixed-effect meta-analysis was used to combine results across all datasets. Sensitivity analyses were conducted using smoking status as an alternative environmental exposure, along with subgroup analyses stratified by histological type and sex to examine potential heterogeneity. Additionally, we launched the free LungCancer-xWAS-G×E online platform (http://bigdata.njmu.edu.cn/LungCancer-xWAS-GxE/), which stores entire results for all types of biomarkers analyzed in the current study.

• ***MMS***

To quantify how molecular biomarkers modify the effects between smoking exposure and lung cancer risk, we developed the MMS. While mathematically similar to a polygenic risk score (PRS), the MMS has distinct biological interpretation which captures interaction effects. Individuals with higher MMS values are expected to experience greater smoking effect on lung cancer risk compared to those with lower MMS. It was calculated as below (**Equation 3-5**):

$$\text{logit}(\pi) = \beta_0 + \sum \beta_i \times \text{covariate}_i + \beta_X \times \tilde{X} + \left(\beta_E + \beta_{XE} \times \tilde{X}\right) \times \text{pack-year} \quad (3)$$

$$\text{Modifying effect of } \tilde{X} = \beta_{XE} \times \tilde{X} \quad (4)$$

$$\text{MMS} = \sum_{\tilde{X} \in \Omega} \beta_{XE} \times \tilde{X}, \Omega \text{ represents the set of weakly correlated biomarkers } \tilde{X} \quad (5)$$

Here, β_E represents the population-average effects of smoking (measured in pack-years) and

$\beta_{XE} \times \tilde{X}$ denotes the individual-level effects attributable to genetically predicted molecular traits. To avoid collinearity, only the most significant biomarker was retained from any correlated pairs with Pearson's correlation coefficient $|\rho| > 0.7$.

We stratified the population into three groups based on MMS tertiles. Lung cancer risk in relation to smoking exposure was then evaluated within each group. Generalized additive models (GAMs) were fitted separately by sex to estimate age-specific lung cancer incidence at ages 50, 60, and 70 in the UKB population. Differences in predicted risk across MMS groups were interpreted as the excess risk due to distinct genetic background.

- ***Trans-omics integrative analysis***

To uncover the potential trans-omics regulatory relationships, we performed an integrative analysis by assessing correlation between genetically determined molecular biomarkers across different layers. Following the Central Dogma, which holds that heritable genetic information in DNA undergoes epigenetic modifications (e.g. DNA methylation at CpG sites) that regulate its transcription into mRNA and subsequent mRNA translation into protein, we aimed to identify trans-omics molecular paths (DNA methylation → gene expression → protein abundance → metabolites level) consisting of correlated biomarkers which yields to gene-smoking interaction. To maximize the number of molecular paths, we applied a loosen criterion to select the CpG site as a starting point, if its q-FDR ≤ 0.05 in the meta-analysis. Then, we picked up gene expressions which met below criteria: (1) The gene expression is significantly (q-FDR ≤ 0.05 and correlation coefficients $|\rho| > 0.3$) associated with the identified CpG site

which contributes to the gene-smoking interaction; (2) The gene expression and smoking exhibits significant ($P \leq 0.05$) interaction effect on lung cancer risk. Second, we chose proteins which linked to gene expression using the same criteria. Finally, metabolites linked to proteins was analyzed in a same manner.

Results

- ***Demographic and clinical characteristics of study participants***

A total of 27,737 incident lung cancer cases and 449,910 non-cases of European ancestry passed QC, with all individuals genetically confirmed to be unrelated (Figure E1). Table 1 presents demographic characteristics of lung cancer cases and non-cases (controls) across the four datasets. Lung cancer cases exhibited higher smoking pack-years and greater proportion of current smokers, as well as a lower proportion of never smokers compared to non-cases across all datasets. Furthermore, the histological subtypes of lung cancer were similarly distributed across the four studies.

- ***Acceptable performance of predicted molecular traits***

We used alliance-based xQTL summary-level data, involving 244,533 CpGs, 19,942 genes, 4,907 proteins, and 486 metabolites, to impute four types of molecular biomarkers for individuals with genotype data in ILCCO-OncoArray, TRICL, PLCO and UKB using LDpred2. Finally, 127,621 CpGs, 13,039 genes, 4,730 proteins and 298 metabolites were successfully

predicted. Figure 2 illustrates the distribution of median heritability (h^2) estimated across 30 replications for the predicted biomarkers across the four omics layers. Totally, for the entire 145,688 biomarkers, 103,317 (70.9%), 25,294 (17.4%), 5,954 (4.9%) and 1,619 (1.1%) biomarkers had $h^2 > 0.01$, $h^2 > 0.1$, $h^2 > 0.3$ and $h^2 > 0.5$, respectively. Only reliable biomarkers who met below criteria were retained for downstream association analyses: (i) $h^2 > 0.01$ in the current prediction model, and (ii) median of $h^2 > 0.01$ across 30 replications (see online data supplement). As a result, 89,073 out of 89,186 (99.9%) DNA methylation, 9,012 out of 9,418 (95.7%) genes, 4,225 out of 4,571 (92.4%) proteins, and 140 out of 142 (98.6%) metabolites were remained in the subsequent analyses. The heritability of current biomarker prediction model is highly in accordance with the median posterior heritability estimated across 30 replications (Figure E3), indicating these predicted biomarkers are robust and reliable. Since the "sparse" option was set in LDpred2, we obtained number of SNPs used to predicted each biomarker. There is a weak but significant (Pearson's $\rho = 0.04$, $P < 0.001$) correlations between heritability and the number of SNPs included in the prediction model across four omics layers. The median of the number of SNPs for predicted DNA methylation (86), gene expression (7,239), protein (415,460) and metabolite (340,348) biomarkers exhibited significant difference among four layers (Kruskal-Wallis test: $P < 0.001$). As shown in Figure E3, more pQTLs and metabQTLs were utilized to predict proteins and metabolites, which likely results

in higher heritability of them.

- ***Significant gene-smoking interactions identified by the two-phase study***

In the discovery phase, we identified 53 CpG sites, 14 gene expressions, and 4 proteins significantly associated with smoking-related lung cancer risk (q-FDR ≤ 0.05; Table E1), while no metabolites reached significance threshold. In the replication phase, seven CpG sites and one gene were successfully replicated ($P \leq 0.05$) with consistent directions of effect (Table 2). Among the significant CpG sites, four were located in the chromosome 15q25.1 region, which harbors the nicotinic acetylcholine receptors gene cluster (*CHRNA3/5-CHRNB4-IREB2*) (e.g., cg13714459$_{IREB2}$: $P_{\text{discovery}} = 7.83\times10^{-9}$, q-FDR$_{\text{discovery}} = 3.49\times10^{-4}$ and $P_{\text{replication}} = 8.02\times10^{-3}$). Two CpG sites are located in the *RDM1* gene at 17q12 (e.g., cg02662658$_{RDM1}$: $P_{\text{discovery}} = 1.91\times10^{-6}$, q-FDR$_{\text{discovery}} = 1.08\times10^{-2}$ and $P_{\text{replication}} = 2.55\times10^{-2}$), and one CpG site, cg15034267, is located within the body of *GTPBP1* gene within 22q13.1 region (cg15034267$_{GTPBP1}$: $P_{\text{discovery}} = 1.08\times10^{-6}$, q-FDR$_{\text{discovery}} = 8.31\times10^{-3}$ and $P_{\text{replication}} = 4.38\times10^{-2}$). One gene expression locus, *RP11-326C3.14*, located at 11p15.5, was observed significant in our two-phase design ($P_{\text{discovery}} = 7.21\times10^{-6}$, q-FDR$_{\text{discovery}} = 1.63\times10^{-2}$ and $P_{\text{replication}} = 5.17\times10^{-3}$). Pairwise correlations between the identified biomarkers are shown in Figure E4.

To enhance the robustness of our effect estimates, we conducted a comprehensive meta-analysis integrating results from both discovery and replication phases. All gene-smoking interactions identified in the two-phase analysis remained significant in the combined dataset (Figure 3). Moreover, the meta-analysis identified additional 138 biomarkers exhibiting

significant interactions with smoking (q-FDR ≤ 0.05, Table E2). Quantile-quantile (Q-Q) plots for the meta-analysis results are shown in Figure E5. The most significant signal remained within the nicotinic acetylcholine receptors region (e.g., cg13714459$_{IREB2}$, $P_{meta} = 3.73\times10^{-8}$).

- ***The influence patterns of significant gene-smoking interactions***

The patterns of 8 gene-smoking interactions identified through the two-phase analytic strategy were exhibited in the combined data in Figure E6. The odds ratio (*OR*) for smoking pack-years in relation to lung cancer risk in the overall population was 1.032 (95%CI: 1.031–1.033). Higher levels of five biomarkers (cg13714459, c919941054, cg20622131, cg15034267 and *RP11-326C3.14*) were associated with an increased carcinogenic effect of smoking, while elevated levels of cg13561554, cg02662658 and cg26989927 were found to attenuate the smoking-related risk of lung cancer, highlighting biologically distinct modification patterns.

To explore the effects of pack-years on lung cancer across different levels of the eight significant biomarkers, each biomarker was categorized into a three-level factor (low, medium and high) based on tertiles in the combined dataset. The forest plot revealed that the effect of pack-years varied across subpopulations with different epigenetic or transcriptional profiles (Figure E7). For example, among individuals with lower methylation levels of cg13714459, the association between 10 smoking pack-years and lung cancer risk showed an *OR* of 1.324, whereas in those with higher level of cg13714459, *OR* increased to 1.398. Significant

heterogeneity was observed across three subpopulations ($P_{heterogenity} = 1.75\times10^{-10}$).

- ***Additional signals observed in sensitivity analysis by smoking status***

In the sensitivity analysis, we reassessed the eight identified interactions using smoking status instead of pack-years as the environmental exposure. While the results remained consistent with those from the primary analysis in the discovery phase, three biomarkers (cg02662658, cg26989927 and cg15034267) failed to reach significance in the replication phase. This discrepancy may be due to information loss caused by categorizing a continuous variable (pack-years of smoking) into a binary variable (smoking status), which might reduce statistical power in hypothesis testing. Nevertheless, all biomarkers showed significant interactions with smoking status in the combined dataset (Table E5).

Using the same two-phase analytical strategy, we examined interactions between all types of biomarkers and smoking status (Figure E8-9). Ten CpG probes sites showing statistically significant interactions with smoking status (Table E6). Four probes (cg13714459, cg19941054, cg20622131 and cg13561554) replicated findings from our primary analysis. Besides, six new biomarkers were additionally identified. The results of meta-analysis were provided in Table E7.

- ***General signals observed by subgroup analysis stratified by subject characteristics***

The gene-smoking interaction effects of most biomarkers were similar across different sex and

histological type strata, suggesting that these interactions are general, and not limited to specific subpopulations (Table E8). In contrast, the modifying effects of cg26989927$_{RDM1}$, cg02662658$_{RDM1}$ and *RP11-326C3.14* were observed exclusively in lung adenocarcinoma (LUAD), suggesting a more specific role in LUAD pathogenesis (Figure E10-17). In addition, significant (q-FDR ≤ 0.05) gene-smoking interactions in subgroups by sex or histological type are shown in Manhattan plots (Figure E18-23).

- ***MMS empowers the ability of identifying smokers at high risk of lung cancer***

The MMS quantifies the extent to which genetics exacerbate the effect of smoking on lung cancer risk. Five relatively independent biomarkers (cg13714459, cg13561554, cg26989927, cg15034267 and *RP11-326C3.14*) were selected for MMS construction, with the weights determined by coefficients obtained from the meta-analysis across four datasets. The proposed MMS could effectively stratify the population with given age and sex (Figure 4). Individuals in the high MMS group consistently exhibited a higher risk of lung cancer. As pack-years increased, the lung cancer risk differences among the MMS groups became more pronounced. It is interesting that, given the same age (e.g., 50), smoking pack-years (e.g., 40) and MMS group (e.g., high), female lung cancer incidence is slightly higher than that of male, which has been observed in many studies(34, 35). Anyway, females had a lower smoking prevalence (42.8% *vs* 56.5%, $P < 0.001$) and fewer pack-years (23.4 *vs* 31.7, $P < 0.001$) compared to males, inducing lower lung cancer incidence overall.

MMS is not a score predicting individual lung cancer risk. Instead, MMS can be understood as

an extra smoking effect attributed to individual susceptibility. Since MMS is not designed to improve lung cancer prediction performance, the increase of predictive power (e.g., AUC) by MMS is slight in the overall population ($AUC_{covariate}$ = 0.8392 *vs* $AUC_{covariate+MMS\times pack\text{-}year}$ = 0.8416). Anyway, we still observed that the increase of overall discrimination ability of age, sex and pack-year of smoking from low to high MMS group (Figure E24).

- ***Trans-omics integrative analysis***

Trans-omics integrative analysis revealed molecular pathways comprising correlated biomarkers across omics layers that contribute to gene-smoking interactions in lung cancer susceptibility (Figure 5). Bands connecting adjacent layers represent significant inter-omics correlations. Consistent with genetic principles, the strongest correlations occurred between biomarkers in close genomic proximity (*cis*-acting), while fewer associations were observed between biomarkers on different chromosomes or distant regions of the same chromosome (*trans*-acting). Specifically, we identified 55 DNA methylation probes correlating with 26 gene expression levels, all occurring in *cis*. Notably, only one gene expression biomarker, *NELFE*, exhibited *trans*-acting associations with three downstream protein levels. Among these proteins, *ANAPC7* showed correlations with 19 metabolites. Given the localization of the *NELFE* gene within the major histocompatibility complex (MHC) class III region on chromosome 6, the observed downstream trans-omics associations likely reflect functional biological relationships rather than direct genomic proximity.

Discussion

Under the xWAS framework, we systematically evaluated gene-smoking interactions in lung cancer using trans-omics data encompassing epigenomics, transcriptomics, proteomics, and metabolomics. Through a rigorous two-phase analytical strategy, we identified eight biomarkers, including seven CpG sites and one gene expression. Building on these findings, we innovatively developed the MMS that delineate how genetic predisposition modifies the effect of pack-years on lung cancer risk. The MMS enables effective population risk stratification while accounting for gene-smoking interaction.

The primary criterion for assigning studies to the discovery and replication phases was the balance of statistical power, with an effective sample size ratio of approximately 1:1. The effective sample size of a study for dichotomous outcome is calculated as $N_{\text{eff}} = 4/(1/N_{\text{Cases}} + 1/N_{\text{Controls}})$, which accounts for the number of cases and controls(36). Based on this formula, ILCCO-OncoArray contributed 30,272 effective sample size, TRICL 10,434, PLCO 7,002, and UKB 17,257. To achieve a balance between two phases, we therefore allocated the TRICL to the replication phase, together with PLCO and UKB. We acknowledge the theoretical difference between case-control design (TRICL) and cohort design (PLCO/UKB). The rationale for combining the coefficients derived from logistic models using samples in case-control and cohort study designs is based on a methodological paper by Prentice and Pyke (1979)(37), which demonstrates that the coefficient derived from logistic regression model is invariant to retrospective sampling in the case-control design, and is equivalent to the estimate from prospective sampling in the cohort design. Despite potential selection bias, Cochran or meta-analysis guideline suggests that multiple results can be pooled

when there is no heterogeneity across different studies(38). Hence, we performed heterogeneity tests across the datasets in the replication phase to empirically assess the feasibility of combining the TRICL case-control study with the PLCO/UKB cohort studies. The well-established criterion of heterogeneity in omics study is that Q-test $P \leq 0.10$ (39, 40) or $I^2 > 75\%$ suggested by Winkler published in *Nature Protocols*(41). Our results (see Table E9) showed non-significant (q-FDR > 0.2 and $I^2 < 75\%$) heterogeneity between TRICL case-control study and PLCO/UKB cohort studies after multiple testing corrections. Therefore, this empirical evidence supports combining multisource results in a meta-analysis, which is an acceptable approach in genomic epidemiology(42, 43).

While the utility of predicted molecular biomarkers for revealing disease mechanisms has been increasingly demonstrated under xWAS framework(21, 44, 45), which effectively addresses the limited availability of individual-level multi-omics data, our study extends this paradigm by investigating biomarkers-smoking interactions. These interactions not only help explain missing heritability in lung carcinogenesis but also reveal potential molecular pathways underlying tobacco-related lung cancer development(46, 47). As shown in Figure 1, the association between the actual molecular trait X and the disease Y could be affected by unobserved confounders U, such as diet and education(48). In this study, we used genetically determined molecular traits $\tilde{X}$, imputed from genotype data and xQTL summary statistics, to represent X. Since the unobserved confounders U are independent of $\tilde{X}$, they are not a concern anymore. Notably, causal path ① or ② indicates the possible mechanisms of interplay between genetic and smoking factors, leading to significant differences in tobacco-

related susceptibility.

Meanwhile, we compared the heritability distributions of predicted biomarkers with several previous studies (21, 49). The heritability distribution may depend on the type of the molecular biomarker(50). These existing evidences also exhibit large variation of the heritability across different omics layers. Anyway, even though there is prediction error and some biomarkers have low heritability, it does not cause inflation of type I error in xWAS framework according to error-in-variables theory(51, 52). Therefore, it does not compromise our conclusions.

Our two-phase study identified several novel and biologically plausible biomarkers, with the most significant signals being DNA methylation probes in the 15q13.3 region, which encompasses *IREB2* and *CHRNA3/5*. This region, encoding nicotinic acetylcholine receptors, was reported to show consistent association with lung cancer in previous studies across diverse populations(53, 54). In addition, two CpG sites, cg02662658, and cg26989927, were identified within the 17q12 region of the *RDM1* gene. *RDM1* demonstrates overexpression at both mRNA and protein levels in human lung tumors, particularly in lung adenocarcinoma. Experimental evidence from *in vitro* and *in vivo* studies suggests that *RDM1* functions as an oncogene in lung adenocarcinoma and contributes to cell survival and proliferation in NSCLC, likely through its role in DNA repair capabilities(55, 56). Although *RDM1* has not been implicated in lung cancer susceptibility by GWAS to date, our findings provide supporting evidence at the methylation level for the role of *RDM1*. Furthermore, we identified a CpG site, cg15034267, located within the body of the *GTPBP1* gene on chromosome 22(22q13.1). The GTP-binding protein (GTPBP) family, characterized by GTPase activity, has been shown to play a critical role in cancer(57), although evidence from *in vivo* and *in vitro* studies on its carcinogenic mechanisms remains

limited. Notably, our study offers comprehensive evidence that smoking interacts with the carcinogenic effects of these genes, revealing their potential biological mechanisms in lung cancer development.

The MMS developed in this study provides insights into the interaction between genetic predisposition and smoking-related lung cancer risk. Our findings show that females have a slightly higher lung cancer risk than males after adjusting for smoking pack-years and age. This is consistent with growing evidence suggesting that females are more susceptible to smoking-related pulmonary diseases(34). Unlike traditional polygenetic risk score, which assume a static genetic risk on the outcome, our MMS demonstrates that genetic predisposition interacts with modifiable lifestyle factors, such as smoking. This highlights the potential for personalized interventions that account for both inherited and environmental influences(58). Our findings underscore the importance of incorporating both genetic information and lifestyle modifications for more effective cancer prevention, such as smoking cessation(59).

Our study had several strengths. First, we integrated consortium-scale individual-level GWAS data of lung cancer with alliance-based summary-level xQTL data to detect gene-smoking interactions, making this the largest trans-omics G×E study of lung cancer to date. Second, we applied a two-phase analytic strategy to ensure the robustness of the identified biomarkers. Biomarkers were selected in a discovery phase using the BH procedure at q-FDR ≤ 0.05 and further tested in a replication phase at the nominal level 0.05, the overall FDR of the entire study is actually controlled at level $0.05\times0.05\times\Pi_0$ (where Π_0 is the proportion of true null hypotheses)(60, 61). The two-phase criterion is more conservative than the single-phase threshold where FDR is controlled at $0.05\Pi_0$. Therefore, the nominal significance level for the

replication phase was widely applied in omics studies(62-64). As a secondary analysis, we conducted a meta-analysis of all four studies to maximize statistical power and ensure the largest sample size for detecting G×E interaction(65-67). Indeed, additional significant signals were identified beyond those found in the two-phase design. Importantly, all 8 gene-smoking interactions identified in the replication-based analysis remained statistically significant, confirming their robustness. Finally, we developed the MMS, which enhances the potential for individualized lung cancer prevention strategies. This integrative and rigorous approach provides novel insights into the complex interplay between genetic predisposition and smoking, laying a solid foundation for future research on lung cancer susceptibility and targeted interventions.

However, we acknowledge several limitations. First, due to the potential horizontal pleiotropy of genetic variants, we could only identify associations rather than causality. Currently, no effective tools exist to fully account for pleiotropy in G×E interaction studies. Second, the xWAS framework has inherent limitations. Its performance is constrained by the heritability of molecular biomarker, which measures the reliability of the genetic prediction. As a results, biomarkers with $h^2 < 0.01$ were excluded due to limited statistical power, in accordance with criterion used by existing studies. Third, both the xQTL and GWAS datasets were derived from individuals of European ancestry, which limits the generalizability of our findings to other populations. In future, it is anticipated to confirm our findings in populations of diverse ancestries. Fourth, since all lung cancer GWAS data was derived from blood sample, we therefore only used QTL data derived from blood sample for the sake of consistency. We aim to produce blood sample based multi-omics biomarkers, which are distinct from single- or

cross-tissue somatic mutations that drive tumorigenesis. Though these identified G×E interactions may be diverse among different lung cancers driven by different somatic mutations (e.g., *KRAS* and *BRAF*), there is no somatic mutation data of participants in our study for analysis. However, we performed stratified analyses by histological subtype for these 8 interactions in the combined set. Four interactions exhibited significance in both LUAD and LUSC subgroup analyses, while four interactions exclusively contributed to LUAD risk, indicating both genetic similarity and heterogeneity exist between LUAD and LUSC. Finally, the statistical power of G×E study under xWAS framework is relatively lower compared to traditional study of main effect of biomarker(23, 68). Therefore, we observed fewer signals in both discovery and validation phases. Though another additional replication is helpful to confirm the significances of these signals, we have already utilized all available consortium-scale GWAS data and alliance-based xQTL data. Additionally, potential selection bias may be introduced in case-control studies and can never be entirely ruled out in observational research. Therefore, *in vivo* and *in vitro* experiments are warranted to verify whether and how these biomarkers modify the smoking effect on lung tumorigenesis.

In conclusion, we perform the first trans-omics gene-environment interaction study to uncover biomarkers that modify smoking effect and quantify the extent to which genetics exacerbate the smoking effect on lung cancer risk. Furthermore, we launched an open-access online

platform, LungCancer-xWAS-GxE to share the signals across different omics layers with the community, which was available at http://bigdata.njmu.edu.cn/LungCancer-xWAS-GxE/.

Data availability

Access URLs for four lung cancer GWAS individual data are as following：

- ILCCO-OncoArray data are available from https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001273.v3.p2.
- TRICL data are available from https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001681.v1.p1.
- PLCO data are available from https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi-study_id=phs001286.v2.p2.
- UK Biobank data are available from https://www.ukbiobank.ac.uk/.

Access URLs for four types of xQTL summary statistics as following:

- Methylation QTL statistics are available from http://mqtldb.godmc.org.uk/.
- Expression QTL statistics are available from https://www.eqtlgen.org.
- Protein QTL statistics are available from https://www.decode.com/.
- Metabolite QTL statistics are available from http://metabolomics.helmholtz-muenchen.de/gwas.

The following are the URLs of the software used in this article:

- PLINK 2.0: https://www.cog-genomics.org/plink/.
- TOPMed imputation server: https://imputation.biodatacatalyst.nhlbi.nih.gov/#!.
- R v4.3.0: https://www.r-project.org/.
- KING: https://www.kingrelatedness.com/manual.shtml.
- LiftOver: https://genome.ucsc.edu/cgi-bin/hgLiftOver.
- R package *bigsnpr*: https://github.com/privefl/bigsnpr.
- R package *metafor*: https://cran.r-project.org/web/packages/metafor/index.html.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (82220108002 to F.C., 82273737 to R.Z., 82373690 to Y.Z., 82473728 to Y.W.W.), Noncommunicable Chronic Diseases-National Science and Technology Major Project (2024ZD0520000 and 2024ZD0520003 to R.Z.), the US National Institutes of Health (CA209414, CA249096, CA092824, and ES000002 to D.C.C., CA 249096 and CA209414 to Y.L.), Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). R.Z. was partially supported by the Outstanding Young Teachers Training Program of Nanjing Medical University. We thank the patients and investigators who participated in ILCCO-OncoArray, TRICL, PLCO and UKB for providing the GWAS data. We also thank all the QTL cohorts in the present work for making the statistics publicly available and are grateful to all the investigators and participants who contributed to those studies, including GoDMC for meQTL, FHS, LIFE Heart, LIFE Adult, NTR-NESDA, Young Finns Study, BEST, Fehrmann and Lifelines Deep for eQTL, the Icelandic Cancer Project and deCODE for pQTL, KORA F4 and TwinsUK for metabQTL.

Competing interests

The authors declare no conflict of interest.

References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 2024; 74: 229-263.

2. Wéber A, Morgan E, Vignat J, Laversanne M, Pizzato M, Rumgay H, Singh D, Nagy P, Kenessey I, Soerjomataram I, Bray F. Lung cancer mortality in the wake of the changing smoking epidemic: a descriptive study of the global burden in 2020 and 2040. *BMJ Open* 2023; 13: e065303.

3. Stavrides JC. Lung carcinogenesis: Pivotal role of metals in tobacco smoke. *Free Radical Biology and Medicine* 2006; 41: 1017-1030.

4. Sonnenfeld G, Griffith RB, Hudgens RW. The effect of smoke generation and manipulation variables on the cytotoxicity of mainstream and sidestream cigarette smoke to monolayer cultures of L-929 cells. *Arch Toxicol* 1985; 58: 120-122.

5. Witschi H, Espiritu I, Maronpot RR, Pinkerton KE, Jones AD. The carcinogenic potential of the gas phase of environmental tobacco smoke. *Carcinogenesis* 1997; 18: 2035-2042.

6. Pouli AE, Hatzinikolaou DG, Piperi C, Stavridou A, Psallidopoulos MC, Stavrides JC. The cytotoxic effect of volatile organic compounds of the gas phase of cigarette smoke on lung epithelial cells. *Free Radical Biology and Medicine* 2003; 34: 345-355.

7. Doll R, Peto R. The causes of cancer: quantitative estimates of avoidable risks of cancer in the United States today. *J Natl Cancer Inst* 1981; 66: 1191-1308.

8. Zhang R, Chu M, Zhao Y, Wu C, Guo H, Shi Y, Dai J, Wei Y, Jin G, Ma H, Dong J, Yi H, Bai J, Gong J, Sun C, Zhu M, Wu T, Hu Z, Lin D, Shen H, Chen F. A genome-wide gene–environment interaction analysis for tobacco smoke and lung cancer susceptibility. *Carcinogenesis* 2014; 35: 1528-

1535.

9. Spitz MR, Wei Q, Li G, Wu X. Genetic Susceptibility to Tobacco Carcinogenesis: Environmental Carcinogenesis. *Cancer Investigation* 1999; 17: 645-659.

10. Zhou F, Ren J, Lu X, Ma S, Wu C. Gene-Environment Interaction: A Variable Selection Perspective. *Methods Mol Biol* 2021; 2212: 191-223.

11. Maher B. Personal genomes: The case of the missing heritability. *Nature* 2008; 456: 18-21.

12. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genetics in Medicine* 2002; 4: 45-61.

13. Wu C, Li S, Cui Y. Genetic association studies: an information content perspective. *Curr Genomics* 2012; 13: 566-573.

14. Cornelis MC, Tchetgen EJT, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am J Epidemiol* 2012; 175: 191-202.

15. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* 2009; 169: 219-226.

16. Hu Z, Wu C, Shi Y, Guo H, Zhao X, Yin Z, Yang L, Dai J, Hu L, Tan W, Li Z, Deng Q, Wang J, Wu W, Jin G, Jiang Y, Yu D, Zhou G, Chen H, Guan P, Chen Y, Shu Y, Xu L, Liu X, Liu L, Xu P, Han B, Bai C, Zhao Y, Zhang H, Yan Y, Ma H, Chen J, Chu M, Lu F, Zhang Z, Chen F, Wang X, Jin L, Lu J, Zhou B, Lu D, Wu T, Lin D, Shen H. A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. *Nature Genetics* 2011; 43: 792-796.

17. Deng Q, Guo H, Dai J, Yang L, Wu C, Wang Q, Hu Z, Yang M, Liu L, Yu D, Hu D, Hong X, Qiu

F, Yang H, Wang T, Tan W, Chu M, Feng J, Teng K, Gong J, Sun C, Hu X, Zhang K, Lu J, Lin D, Shen H, Wu T. Imputation-based association analyses identify new lung cancer susceptibility variants in CDK6 and SH3RF1 and their interactions with smoking in Chinese populations. *Carcinogenesis* 2013; 34: 2010-2016.

18. Zhang R, Lai L, Dong X, He J, You D, Chen C, Lin L, Zhu Y, Huang H, Shen S, Wei L, Chen X, Guo Y, Liu L, Su L, Shafer A, Moran S, Fleischer T, Bjaanaes MM, Karlsson A, Planck M, Staaf J, Helland Å, Esteller M, Wei Y, Chen F, Christiani DC. SIPA1L3 methylation modifies the benefit of smoking cessation on lung adenocarcinoma survival: an epigenomic-smoking interaction analysis. *Mol Oncol* 2019; 13: 1235-1248.

19. Ji X, Lin L, Fan J, Li Y, Wei Y, Shen S, Su L, Shafer A, Bjaanæs MM, Karlsson A, Planck M, Staaf J, Helland Å, Esteller M, Zhang R, Chen F, Christiani DC. Epigenome-wide three-way interaction study identifies a complex pattern between TRIM27, KIAA0226, and smoking associated with overall survival of early-stage NSCLC. *Mol Oncol* 2022; 16: 717-731.

20. Zhang R, Chen C, Dong X, Shen S, Lai L, He J, You D, Lin L, Zhu Y, Huang H, Chen J, Wei L, Chen X, Li Y, Guo Y, Duan W, Liu L, Su L, Shafer A, Fleischer T, Moksnes Bjaanæs M, Karlsson A, Planck M, Wang R, Staaf J, Helland Å, Esteller M, Wei Y, Chen F, Christiani DC. Independent Validation of Early-Stage Non-Small Cell Lung Cancer Prognostic Scores Incorporating Epigenetic and Transcriptional Biomarkers With Gene-Gene Interactions and Main Effects. *Chest* 2020; 158: 808-819.

21. Xu Y, Ritchie SC, Liang Y, Timmers PRHJ, Pietzner M, Lannelongue L, Lambert SA, Tahir UA, May-Wilson S, Foguet C, Johansson Å, Surendran P, Nath AP, Persyn E, Peters JE, Oliver-Williams C, Deng S, Prins B, Luan Ja, Bomba L, Soranzo N, Di Angelantonio E, Pirastu N, Tai ES, van Dam

RM, Parkinson H, Davenport EE, Paul DS, Yau C, Gerszten RE, Mälarstig A, Danesh J, Sim X, Langenberg C, Wilson JF, Butterworth AS, Inouye M. An atlas of genetic scores to predict multi-omic traits. *Nature* 2023; 616: 123-131.

22. Brandes N, Linial N, Linial M. PWAS: proteome-wide association study—linking genes and phenotypes by functional variation in proteins. *Genome Biol* 2020; 21: 173.

23. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, Jansen R, de Geus EJC, Boomsma DI, Wright FA, Sullivan PF, Nikkola E, Alvarez M, Civelek M, Lusis AJ, Lehtimäki T, Raitoharju E, Kähönen M, Seppälä I, Raitakari OT, Kuusisto J, Laakso M, Price AL, Pajukanta P, Pasaniuc B. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016; 48: 245-252.

24. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Consortium GT, Nicolae DL, Cox NJ, Im HK. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015; 47: 1091-1098.

25. McKay JD, Hung RJ, Han Y, Zong X, Carreras-Torres R, Christiani DC, Caporaso NE, Johansson M, Xiao X, Li Y, Byun J, Dunning A, Pooley KA, Qian DC, Ji X, Liu G, Timofeeva MN, Bojesen SE, Wu X, Le Marchand L, Albanes D, Bickeböller H, Aldrich MC, Bush WS, Tardon A, Rennert G, Teare MD, Field JK, Kiemeney LA, Lazarus P, Haugen A, Lam S, Schabath MB, Andrew AS, Shen H, Hong Y-C, Yuan J-M, Bertazzi PA, Pesatori AC, Ye Y, Diao N, Su L, Zhang R, Brhane Y, Leighl N, Johansen JS, Mellemgaard A, Saliba W, Haiman CA, Wilkens LR, Fernandez-Somoano A, Fernandez-Tardon G, van der Heijden HFM, Kim JH, Dai J, Hu Z, Davies MPA, Marcus MW, Brunnström H, Manjer J, Melander O, Muller DC, Overvad K, Trichopoulou A, Tumino R, Doherty JA, Barnett MP, Chen C, Goodman GE, Cox A, Taylor F, Woll P, Brüske I, Wichmann HE, Manz J,

Muley TR, Risch A, Rosenberger A, Grankvist K, Johansson M, Shepherd FA, Tsao M-S, Arnold SM, Haura EB, Bolca C, Holcatova I, Janout V, Kontic M, Lissowska J, Mukeria A, Ognjanovic S, Orlowski TM, Scelo G, Swiatkowska B, Zaridze D, Bakke P, Skaug V, Zienolddiny S, Duell EJ, Butler LM, Koh W-P, Gao Y-T, Houlston RS, McLaughlin J, Stevens VL, Joubert P, Lamontagne M, Nickle DC, Obeidat Me, Timens W, Zhu B, Song L, Kachuri L, Artigas MS, Tobin MD, Wain LV, Rafnar T, Thorgeirsson TE, Reginsson GW, Stefansson K, Hancock DB, Bierut LJ, Spitz MR, Gaddis NC, Lutz SM, Gu F, Johnson EO, Kamal A, Pikielny C, Zhu D, Lindströem S, Jiang X, Tyndale RF, Chenevix-Trench G, Beesley J, Bossé Y, Chanock S, Brennan P, Landi MT, Amos CI. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* 2017; 49: 1126-1132.

26. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, Zong X, Laplana M, Wei Y, Han Y, Lloyd A, Delahaye-Sourdeix M, Chubb D, Gaborieau V, Wheeler W, Chatterjee N, Thorleifsson G, Sulem P, Liu G, Kaaks R, Henrion M, Kinnersley B, Vallee M, LeCalvez-Kelm F, Stevens VL, Gapstur SM, Chen WV, Zaridze D, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Krokan HE, Gabrielsen ME, Skorpen F, Vatten L, Njolstad I, Chen C, Goodman G, Benhamou S, Vooder T, Valk K, Nelis M, Metspalu A, Lener M, Lubinski J, Johansson M, Vineis P, Agudo A, Clavel-Chapelon F, Bueno-de-Mesquita HB, Trichopoulos D, Khaw KT, Johansson M, Weiderpass E, Tjonneland A, Riboli E, Lathrop M, Scelo G, Albanes D, Caporaso NE, Ye Y, Gu J, Wu X, Spitz MR, Dienemann H, Rosenberger A, Su L, Matakidou A, Eisen T, Stefansson K, Risch A, Chanock SJ, Christiani DC, Hung RJ, Brennan P, Landi MT, Houlston RS, Amos CI. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* 2014; 46: 736-741.

27. Machiela MJ, Huang WY, Wong W, Berndt SI, Sampson J, De Almeida J, Abubakar M, Hislop J, Chen KL, Dagnall C, Diaz-Mayoral N, Ferrell M, Furr M, Gonzalez A, Hicks B, Hubbard AK, Hutchinson A, Jiang K, Jones K, Liu J, Loftfield E, Loukissas J, Mabie J, Merkle S, Miller E, Minasian LM, Nordgren E, Park B, Pinsky P, Riley T, Sandoval L, Saxena N, Vogt A, Wang J, Williams C, Wright P, Yeager M, Zhu B, Zhu C, Chanock SJ, Garcia-Closas M, Freedman ND. GWAS Explorer: an open-source tool to explore, visualize, and access GWAS summary statistics in the PLCO Atlas. *Sci Data* 2023; 10: 25.

28. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, Collins R. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015; 12: e1001779.

29. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, Carnero-Montoro E, Lawson DJ, Burrows K, Suderman M, Bretherick AD, Richardson TG, Klughammer J, Iotchkova V, Sharp G, Al Khleifat A, Shatunov A, Iacoangeli A, McArdle WL, Ho KM, Kumar A, Soderhall C, Soriano-Tarraga C, Giralt-Steinhauer E, Kazmi N, Mason D, McRae AF, Corcoran DL, Sugden K, Kasela S, Cardona A, Day FR, Cugliari G, Viberti C, Guarrera S, Lerro M, Gupta R, Bollepalli S, Mandaviya P, Zeng Y, Clarke TK, Walker RM, Schmoll V, Czamara D, Ruiz-Arenas C, Rezwan FI, Marioni RE, Lin T, Awaloff Y, Germain M, Aissi D, Zwamborn R, van Eijk K, Dekker A, van Dongen J, Hottenga JJ, Willemsen G, Xu CJ, Barturen G, Catala-Moll F, Kerick M, Wang C, Melton P, Elliott HR, Shin J, Bernard M, Yet I, Smart M, Gorrie-Stone T, Consortium B, Shaw C, Al Chalabi A, Ring SM, Pershagen G, Melen E, Jimenez-Conde J, Roquer J, Lawlor DA, Wright J, Martin NG, Montgomery GW, Moffitt TE, Poulton R, Esko T, Milani L, Metspalu A, Perry JRB, Ong KK,

Wareham NJ, Matullo G, Sacerdote C, Panico S, Caspi A, Arseneault L, Gagnon F, Ollikainen M, Kaprio J, Felix JF, Rivadeneira F, Tiemeier H, van IMH, Uitterlinden AG, Jaddoe VWV, Haley C, McIntosh AM, Evans KL, Murray A, Raikkonen K, Lahti J, Nohr EA, Sorensen TIA, Hansen T, Morgen CS, Binder EB, Lucae S, Gonzalez JR, Bustamante M, Sunyer J, Holloway JW, Karmaus W, Zhang H, Deary IJ, Wray NR, Starr JM, Beekman M, van Heemst D, Slagboom PE, Morange PE, Tregouet DA, Veldink JH, Davies GE, de Geus EJC, Boomsma DI, Vonk JM, Brunekreef B, Koppelman GH, Alarcon-Riquelme ME, Huang RC, Pennell CE, van Meurs J, Ikram MA, Hughes AD, Tillin T, Chaturvedi N, Pausova Z, Paus T, Spector TD, Kumari M, Schalkwyk LC, Visscher PM, Davey Smith G, Bock C, Gaunt TR, Bell JT, Heijmans BT, Mill J, Relton CL. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet* 2021; 53: 1311-1321.

30. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, Kirsten H, Saha A, Kreuzhuber R, Yazar S, Brugge H, Oelen R, de Vries DH, van der Wijst MGP, Kasela S, Pervjakova N, Alves I, Favé MJ, Agbessi M, Christiansen MW, Jansen R, Seppälä I, Tong L, Teumer A, Schramm K, Hemani G, Verlouw J, Yaghootkar H, Sönmez Flitman R, Brown A, Kukushkina V, Kalnapenkis A, Rüeger S, Porcu E, Kronberg J, Kettunen J, Lee B, Zhang F, Qi T, Hernandez JA, Arindrarto W, Beutner F, Dmitrieva J, Elansary M, Fairfax BP, Georges M, Heijmans BT, Hewitt AW, Kähönen M, Kim Y, Knight JC, Kovacs P, Krohn K, Li S, Loeffler M, Marigorta UM, Mei H, Momozawa Y, Müller-Nurasyid M, Nauck M, Nivard MG, Penninx B, Pritchard JK, Raitakari OT, Rotzschke O, Slagboom EP, Stehouwer CDA, Stumvoll M, Sullivan P, t Hoen PAC, Thiery J, Tönjes A, van Dongen J, van Iterson M, Veldink JH, Völker U, Warmerdam R, Wijmenga C, Swertz M, Andiappan A, Montgomery GW, Ripatti S, Perola M, Kutalik Z, Dermitzakis E, Bergmann S,

Frayling T, van Meurs J, Prokisch H, Ahsan H, Pierce BL, Lehtimäki T, Boomsma DI, Psaty BM, Gharib SA, Awadalla P, Milani L, Ouwehand WH, Downes K, Stegle O, Battle A, Visscher PM, Yang J, Scholz M, Powell J, Gibson G, Esko T, Franke L. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* 2021; 53: 1300-1310.

31. Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, Gunnarsdottir K, Helgason A, Oddsson A, Halldorsson BV, Jensson BO, Zink F, Halldorsson GH, Masson G, Arnadottir GA, Katrinardottir H, Juliusson K, Magnusson MK, Magnusson OT, Fridriksdottir R, Saevarsdottir S, Gudjonsson SA, Stacey SN, Rognvaldsson S, Eiriksdottir T, Olafsdottir TA, Steinthorsdottir V, Tragante V, Ulfarsson MO, Stefansson H, Jonsdottir I, Holm H, Rafnar T, Melsted P, Saemundsdottir J, Norddahl GL, Lund SH, Gudbjartsson DF, Thorsteinsdottir U, Stefansson K. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet* 2021; 53: 1712-1721.

32. Shin SY, Fauman EB, Petersen AK, Krumsiek J, Santos R, Huang J, Arnold M, Erte I, Forgetta V, Yang TP, Walter K, Menni C, Chen L, Vasquez L, Valdes AM, Hyde CL, Wang V, Ziemek D, Roberts P, Xi L, Grundberg E, Multiple Tissue Human Expression Resource C, Waldenberger M, Richards JB, Mohney RP, Milburn MV, John SL, Trimmer J, Theis FJ, Overington JP, Suhre K, Brosnan MJ, Gieger C, Kastenmuller G, Spector TD, Soranzo N. An atlas of genetic influences on human blood metabolites. *Nat Genet* 2014; 46: 543-550.

33. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics* 2021; 36: 5424-5431.

34. Menson KE, Coleman SRM. Smoking and pulmonary health in women: A narrative review and

behavioral health perspective. *Preventive Medicine* 2024; 185: 108029.

35. International Early Lung Cancer Action Program I, Henschke CI, Yip R, Miettinen OS. Women's susceptibility to tobacco carcinogens and survival after diagnosis of lung cancer. *JAMA* 2006; 296: 180-184.

36. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; 26: 2190-2191.

37. Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; 66: 403-411.

38. Higgins J. Cochrane handbook for systematic reviews of interventions. *Cochrane Collaboration and John Wiley & Sons Ltd* 2008.

39. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Bmj* 2003; 327: 557-560.

40. Dickersin K, Berlin JA. Meta-analysis: State-of-the-Science. *Epidemiologic Reviews* 1992; 14: 154-176.

41. Winkler TW, Day FR, Croteau-Chonka DC, Wood AR, Locke AE, Mägi R, Ferreira T, Fall T, Graff M, Justice AE, Luan Ja, Gustafsson S, Randall JC, Vedantam S, Workalemahu T, Kilpeläinen TO, Scherag A, Esko T, Kutalik Z, Heid IM, Loos RJF, The Genetic Investigation of Anthropometric Traits C. Quality control and conduct of genome-wide association meta-analyses. *Nature Protocols* 2014; 9: 1192-1212.

42. Shen S, Li Z, Jiang Y, Duan W, Li H, Du S, Esteller M, Shen H, Hu Z, Zhao Y, Christiani DC, Chen F. A Large-Scale Exome-Wide Association Study Identifies Novel Germline Mutations in Lung Cancer. *Am J Respir Crit Care Med* 2023; 208: 280-289.

43. Zhang R, Shen S, Wei Y, Zhu Y, Li Y, Chen J, Guan J, Pan Z, Wang Y, Zhu M, Xie J, Xiao X, Zhu D, Li Y, Albanes D, Landi MT, Caporaso NE, Lam S, Tardon A, Chen C, Bojesen SE, Johansson M, Risch A, Bickeböller H, Wichmann HE, Rennert G, Arnold S, Brennan P, McKay JD, Field JK, Shete SS, Le Marchand L, Liu G, Andrew AS, Kiemeney LA, Zienolddiny-Narui S, Behndig A, Johansson M, Cox A, Lazarus P, Schabath MB, Aldrich MC, Dai J, Ma H, Zhao Y, Hu Z, Hung RJ, Amos CI, Shen H, Chen F, Christiani DC. A Large-Scale Genome-Wide Gene-Gene Interaction Study of Lung Cancer Susceptibility in Europeans With a Trans-Ethnic Validation in Asians. *J Thorac Oncol* 2022; 17: 974-990.

44. Yin X, Bose D, Kwon A, Hanks SC, Jackson AU, Stringham HM, Welch R, Oravilahti A, Fernandes Silva L, FinnGen, Locke AE, Fuchsberger C, Service SK, Erdos MR, Bonnycastle LL, Kuusisto J, Stitziel NO, Hall IM, Morrison J, Ripatti S, Palotie A, Freimer NB, Collins FS, Mohlke KL, Scott LJ, Fauman EB, Burant C, Boehnke M, Laakso M, Wen X. Integrating transcriptomics, metabolomics, and GWAS helps reveal molecular mechanisms for metabolite levels and disease risk. *American Journal of Human Genetics* 2022; 109: 1727-1741.

45. Zhao X, Yang M, Fan J, Wang M, Wang Y, Qin N, Zhu M, Jiang Y, Gorlova OY, Gorlov IP, Albanes D, Lam S, Tardón A, Chen C, Goodman GE, Bojesen SE, Landi MT, Johansson M, Risch A, Wichmann HE, Bickeböller H, Christiani DC, Rennert G, Arnold SM, Brennan P, Field JK, Shete S, Le Marchand L, Liu G, Hung RJ, Andrew AS, Kiemeney LA, Zienolddiny S, Grankvist K, Johansson M, Caporaso NE, Woll PJ, Lazarus P, Schabath MB, Aldrich MC, Patel AV, Davies MPA, Ma H, Jin G, Hu Z, Amos CI, Shen H, Dai J. Identification of genetically predicted DNA methylation markers associated with non-small cell lung cancer risk among 34,964 cases and 448,579 controls. *Cancer* 2024; 130: 913-926.

46. Xiang R, Fang L, Liu S, Macleod IM, Liu Z, Breen EJ, Gao Y, Liu GE, Tenesa A, Cattle GC, Mason BA, Chamberlain AJ, Wray NR, Goddard ME. Gene expression and RNA splicing explain large proportions of the heritability for complex traits in cattle. *Cell Genom* 2023; 3: 100385.

47. Kim MS, Shim I, Fahed AC, Do R, Park W-Y, Natarajan P, Khera AV, Won H-H. Association of genetic risk, lifestyle, and their interaction with obesity and obesity-related morbidities. *Cell Metabolism* 2024; 36: 1494-1503.e1493.

48. Zhao S, Crouse W, Qian S, Luo K, Stephens M, He X. Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. *Nat Genet* 2024; 56: 336-347.

49. Carrasco-Zanini J, Wheeler E, Uluvar B, Kerrison N, Koprulu M, Wareham NJ, Pietzner M, Langenberg C. Mapping biological influences on the human plasma proteome beyond the genome. *Nat Metab* 2024; 6: 2010-2023.

50. Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. *Nat Rev Genet* 2013; 14: 139-149.

51. Liang Y, Nyasimi F, Im HK. Pervasive polygenicity of complex traits inflates false positive rates in transcriptome-wide association studies. *bioRxiv* 2024: 2023.2010.2017.562831.

52. Fuller WA. Measurement error models. John Wiley & Sons; 2009.

53. Liu Q, Han H, Wang M, Yao Y, Wen L, Jiang K, Ma Y, Fan R, Chen J, Su K, Yang Z, Cui W, Yuan W, Jiang X, Li J, Payne TJ, Wang J, Li MD. Association and cis-mQTL analysis of variants in CHRNA3-A5, CHRNA7, CHRNB2, and CHRNB4 in relation to nicotine dependence in a Chinese Han population. *Transl Psychiatry* 2018; 8: 83.

54. Hancock DB, Wang J-C, Gaddis NC, Levy JL, Saccone NL, Stitzel JA, Goate A, Bierut LJ,

Johnson EO. A multiancestry study identifies novel genetic associations with CHRNA5 methylation in human brain and risk of nicotine dependence. *Hum Mol Genet* 2015; 24: 5940-5954.

55. Tong L, Liu J, Yan W, Cao W, Shen S, Li K, Li L, Niu G. RDM1 plays an oncogenic role in human lung adenocarcinoma cells. *Sci Rep* 2018; 8: 11525.

56. Xu G, Du J, Wang F, Zhang F, Hu R, Sun D, Shen J. RAD52 motif‑containing protein 1 promotes non‑small cell lung cancer cell proliferation and survival via cell cycle regulation. *Oncol Rep* 2018; 40: 833-840.

57. Hu Y, Chen L, Tang Q, Wei W, Cao Y, Xie J, Ji J. Pan-cancer analysis revealed the significance of the GTPBP family in cancer. *Aging (Albany NY)* 2022; 14: 2558-2573.

58. Dar-Nimrod I, Heine SJ. Genetic essentialism: on the deceptive determinism of DNA. *Psychol Bull* 2011; 137: 800-818.

59. Bloss CS, Schork NJ, Topol EJ. Effect of direct-to-consumer genomewide profiling to assess disease risk. *N Engl J Med* 2011; 364: 524-534.

60. Benjamini Y, Yekutieli D. Quantitative trait Loci analysis using the false discovery rate. *Genetics* 2005; 171: 783-790.

61. Zehetmayer S, Posch M. False discovery rate control in two-stage designs. *BMC Bioinformatics* 2012; 13: 81.

62. Wingo TS, Liu Y, Gerasimov ES, Gockley J, Logsdon BA, Duong DM, Dammer EB, Lori A, Kim PJ, Ressler KJ, Beach TG, Reiman EM, Epstein MP, De Jager PL, Lah JJ, Bennett DA, Seyfried NT, Levey AI, Wingo AP. Brain proteome-wide association study implicates novel proteins in depression pathogenesis. *Nature Neuroscience* 2021; 24: 810-817.

63. Zhao B, Lu Q, Cheng Y, Belcher JM, Siew ED, Leaf DE, Body SC, Fox AA, Waikar SS, Collard

CD, Thiessen-Philbrook H, Ikizler TA, Ware LB, Edelstein CL, Garg AX, Choi M, Schaub JA, Zhao H, Lifton RP, Parikh CR. A Genome-Wide Association Study to Identify Single-Nucleotide Polymorphisms for Acute Kidney Injury. *Am J Respir Crit Care Med* 2017; 195: 482-490.

64. Suryadevara R, Gregory A, Lu R, Xu Z, Masoomi A, Lutz SM, Berman S, Yun JH, Saferali A, Ryu MH, Moll M, Sin DD, Hersh CP, Silverman EK, Dy J, Pratte KA, Bowler RP, Castaldi PJ, Boueiz A. Blood-based Transcriptomic and Proteomic Biomarkers of Emphysema. *Am J Respir Crit Care Med* 2024; 209: 273-287.

65. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006; 38: 209-213.

66. Lambert JC, Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, Bellenguez C, DeStafano AL, Bis JC, Beecham GW, Grenier-Boley B, Russo G, Thorton-Wells TA, Jones N, Smith AV, Chouraki V, Thomas C, Ikram MA, Zelenika D, Vardarajan BN, Kamatani Y, Lin CF, Gerrish A, Schmidt H, Kunkle B, Dunstan ML, Ruiz A, Bihoreau MT, Choi SH, Reitz C, Pasquier F, Cruchaga C, Craig D, Amin N, Berr C, Lopez OL, De Jager PL, Deramecourt V, Johnston JA, Evans D, Lovestone S, Letenneur L, Morón FJ, Rubinsztein DC, Eiriksdottir G, Sleegers K, Goate AM, Fiévet N, Huentelman MW, Gill M, Brown K, Kamboh MI, Keller L, Barberger-Gateau P, McGuiness B, Larson EB, Green R, Myers AJ, Dufouil C, Todd S, Wallon D, Love S, Rogaeva E, Gallacher J, St George-Hyslop P, Clarimon J, Lleo A, Bayer A, Tsuang DW, Yu L, Tsolaki M, Bossù P, Spalletta G, Proitsi P, Collinge J, Sorbi S, Sanchez-Garcia F, Fox NC, Hardy J, Deniz Naranjo MC, Bosco P, Clarke R, Brayne C, Galimberti D, Mancuso M, Matthews F, Moebus S, Mecocci P, Del Zompo M, Maier W, Hampel H, Pilotto A, Bullido M, Panza F, Caffarra P, Nacmias B, Gilbert JR, Mayhaus M, Lannefelt L, Hakonarson H, Pichler S, Carrasquillo MM, Ingelsson M, Beekly D, Alvarez V, Zou F,

Valladares O, Younkin SG, Coto E, Hamilton-Nelson KL, Gu W, Razquin C, Pastor P, Mateo I, Owen MJ, Faber KM, Jonsson PV, Combarros O, O'Donovan MC, Cantwell LB, Soininen H, Blacker D, Mead S, Mosley TH, Jr., Bennett DA, Harris TB, Fratiglioni L, Holmes C, de Bruijn RF, Passmore P, Montine TJ, Bettens K, Rotter JI, Brice A, Morgan K, Foroud TM, Kukull WA, Hannequin D, Powell JF, Nalls MA, Ritchie K, Lunetta KL, Kauwe JS, Boerwinkle E, Riemenschneider M, Boada M, Hiltuenen M, Martin ER, Schmidt R, Rujescu D, Wang LS, Dartigues JF, Mayeux R, Tzourio C, Hofman A, Nöthen MM, Graff C, Psaty BM, Jones L, Haines JL, Holmans PA, Lathrop M, Pericak-Vance MA, Launer LJ, Farrer LA, van Duijn CM, Van Broeckhoven C, Moskvina V, Seshadri S, Williams J, Schellenberg GD, Amouyel P. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 2013; 45: 1452-1458.

67. Oldham JM, Allen RJ, Lorenzo-Salazar JM, Molyneaux PL, Ma SF, Joseph C, Kim JS, Guillen-Guio B, Hernández-Beeftink T, Kropski JA, Huang Y, Lee CT, Adegunsoye A, Pugashetti JV, Linderholm AL, Vo V, Strek ME, Jou J, Muñoz-Barrera A, Rubio-Rodriguez LA, Hubbard R, Hirani N, Whyte MKB, Hart S, Nicholson AG, Lancaster L, Parfrey H, Rassl D, Wallace W, Valenzi E, Zhang Y, Mychaleckyj J, Stockwell A, Kaminski N, Wolters PJ, Molina-Molina M, Banovich NE, Fahy WA, Martinez FJ, Hall IP, Tobin MD, Maher TM, Blackwell TS, Yaspan BL, Jenkins RG, Flores C, Wain LV, Noth I. PCSK6 and Survival in Idiopathic Pulmonary Fibrosis. *Am J Respir Crit Care Med* 2023; 207: 1515-1524.

68. Van der Auwera S, Peyrot WJ, Milaneschi Y, Hertel J, Baune B, Breen G, Byrne E, Dunn EC, Fisher H, Homuth G, Levinson D, Lewis C, Mills N, Mullins N, Nauck M, Pistis G, Preisig M, Rietschel M, Ripke S, Sullivan P, Teumer A, Völzke H, Major Depressive Disorder Working Group of the Psychiatric Genomics C, Boomsma DI, Wray NR, Penninx B, Grabe H. Genome-wide gene-

environment interaction in depression: A systematic evaluation of candidate genes: The childhood trauma working-group of PGC-MDD. *Am J Med Genet B Neuropsychiatr Genet* 2018; 177: 40-49.

Figure Legends

Figure 1. The principle of xWAS framework and overall design flowchart.

(A) The causal diagram illustrates the principle of standard xWAS framework. G represents SNPs used to build a prediction model. X represents unobserved molecular traits. $\tilde{X}$ represents predicted molecular biomarkers using genotype and xQTL data. In xWASs, association tests are conducted between $\tilde{X}$ and outcomes *Y*, thereby avoiding confounding by U. **(B)** The causal diagram of G×E analysis within the xWAS framework. The interaction between $\tilde{X}$ and *E* is considered, reflecting how *E* may interact with molecular biomarker. Both causal paths ① and ② may contribute to the modifying effects. **(C)** The flowchart of the overall study design.

Figure 2. Distribution of heritability of molecular biomarkers estimated by LDpred2.

The heritability of molecular biomarkers (h^2) estimated by LDpred2 is used to measure the accuracy of prediction models. Pie charts reflect the number of molecular biomarkers in a particular h^2 range.

Figure 3. The Manhattan plot of the interaction between genetically determined molecular biomarkers and pack-year of smoking on lung cancer risk under the xWAS framework.

Red dots indicated the identified biomarkers through two-phase analytic strategy. The dots above the green dash lines represent biomarkers with q-FDR ≤ 0.05 in the combined set.

Figure 4. Predicted probability of lung cancer risk stratified by age, smoking pack-years and molecular modifying score (MMS) group.

Panels **(A)** to **(E)** show the predicted lung cancer risk at ages 50, 60, and 70 for males and females in the UK Biobank stratified by MMS group. Vertical gray dashed lines indicate the mean pack-year of smoking stratified by sex.

Figure 5. Results of trans-omics integrative analysis across four omics layers.

The bands represent the correlation between predicted biomarkers across different molecular layers. Only correlations with q-FDR ≤ 0.05 and Pearson's $|\rho| > 0.3$ are presented in the figure.

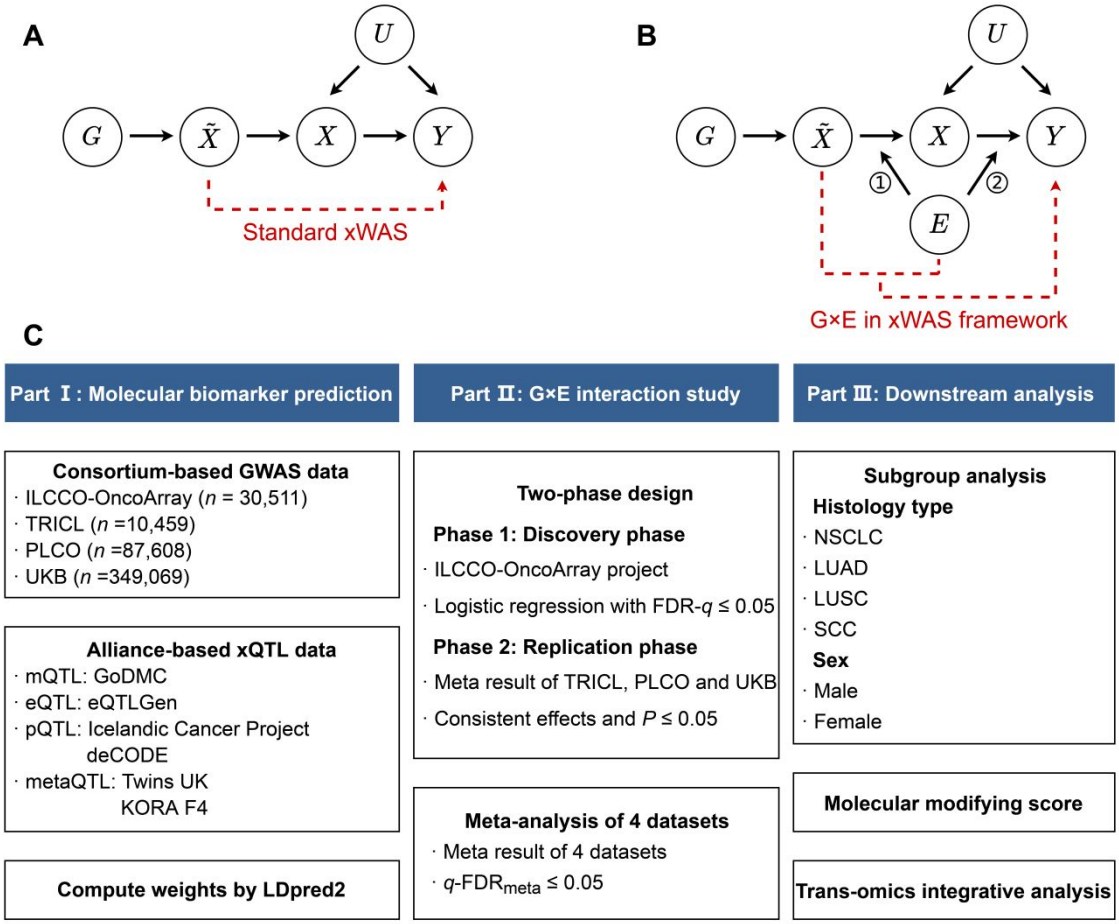

Figure 1

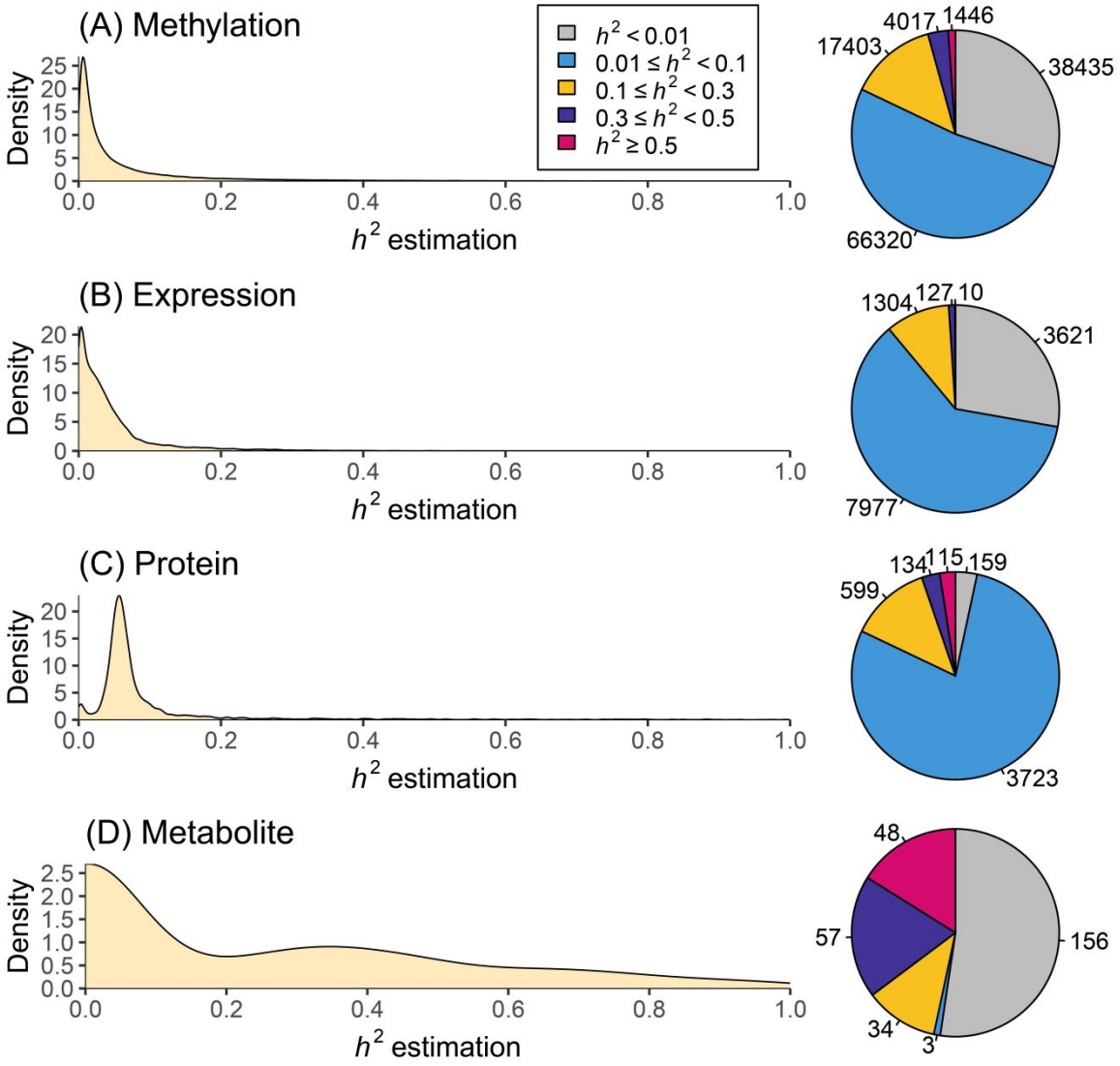

Figure 2

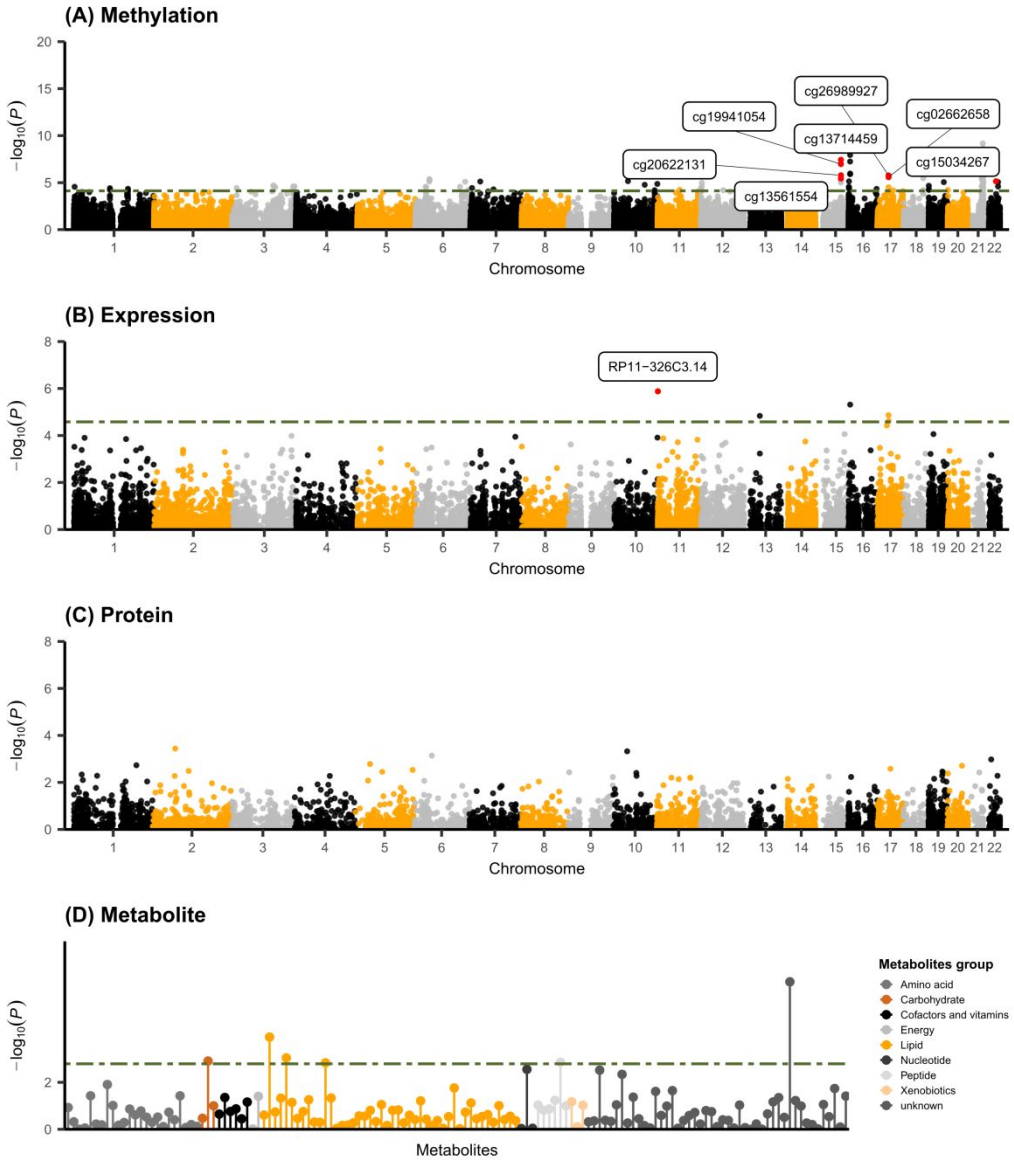

Figure 3

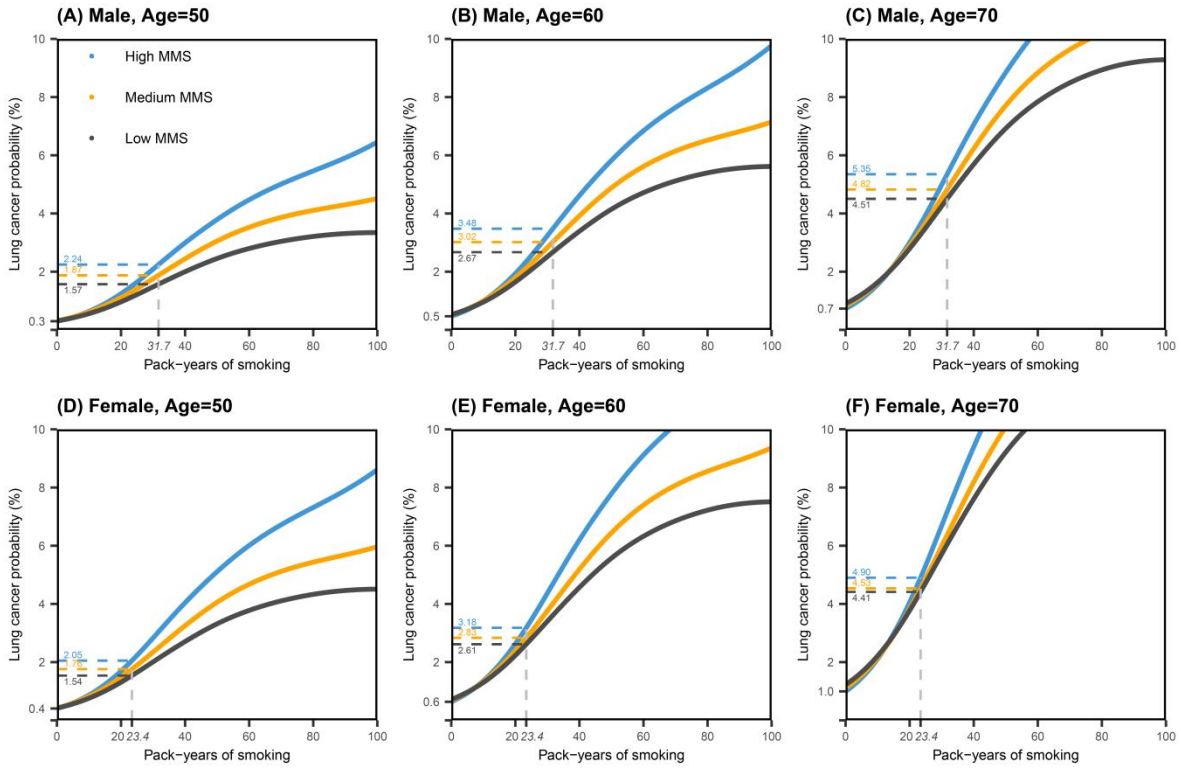

Figure 4

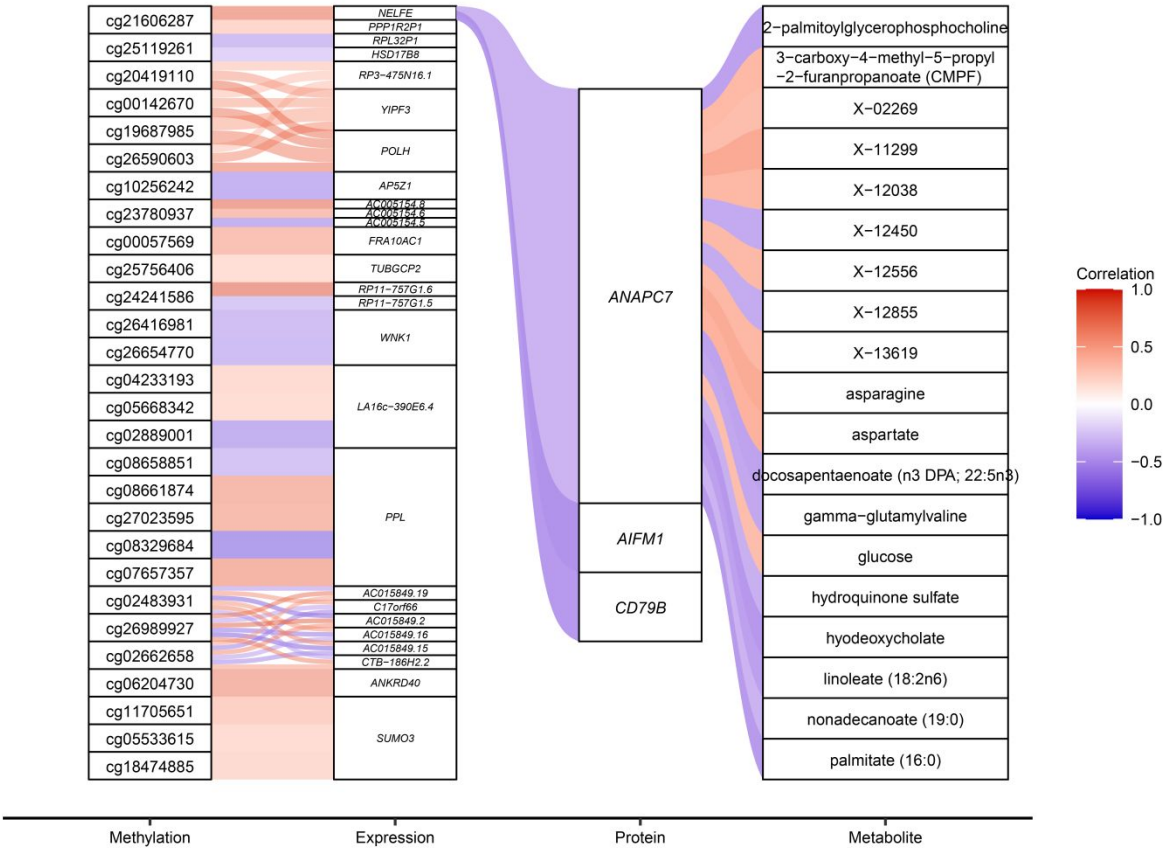

Figure 5

Table

Table 1. Demographic and clinical characteristics of lung cancer cases and non-cases in ILCCO-OncoArray, TRICL, PLCO and UK Biobank.

Characteristics	Discovery Phase		Replication Phase					
	ILCCO-OncoArray		TRICL		PLCO		UKB	
	Case	Control	Case	Control	Incident LC	Non-LC	Incident LC	Non-LC
Sample size	16606	13905	4975	5484	1787	85821	4369	344700
Age (years) *	64.1 ± 10.5	61.8 ± 10.5	60.9 ± 10.0	58.6 ± 9.3	64.1 ± 5.1	61.8 ± 5.1	69.0 ± 6.7	57.9 ± 8.4
Sex, *n* (%)								
Male	10285 (61.9%)	8308 (59.7%)	2676 (53.8%)	2941 (53.6%)	1080 (60.4%)	39749 (46.3%)	2281 (52.2%)	166580 (48.3%)
Female	6321 (38.1%)	5597 (40.3%)	2299 (46.2%)	2543 (46.4%)	707 (39.6%)	46072 (53.7%)	2088 (47.8%)	178120 (51.7%)
Smoking status, *n* (%)								
Never	1612 (9.9%)	4378 (32.3%)	499 (10.3%)	1653 (30.2%)	153 (8.7%)	42210 (50.0%)	708 (16.4%)	188011 (54.8%)
Former	6440 (39.7%)	5536 (40.8%)	1755 (36.0%)	1942 (35.5%)	92 (52.5%)	35763 (42.4%)	1961 (45.4%)	121127 (35.3%)
Current	8169 (50.4%)	3656 (26.9%)	2618 (53.7%)	1875 (34.3%)	679 (38.8%)	6448 (7.6%)	1653 (38.2%)	34100 (9.9%)
Unknown	385	335	103	14	34	1400	47	1462
Pack-year†	41.0 ± 31.5	19.9 ± 24.6	40.7 ± 28.8	25.1 ± 26.0	53.6 ± 36.3	16.0 ± 24.9	38.0 ± 23.8	22.2 ± 18.4
Histology, *n* (%)								
LUAD	6877 (41.4%)	-	2114 (42.5%)	-	710 (39.7%)	-	1272 (29.1%)	-
LUSC	4038 (24.3%)	-	1083 (21.8%)	-	339 (19.0%)	-	618 (14.1%)	-
SCC	1751 (10.5%)	-	538 (10.8%)	-	239 (13.4%)	-	313 (7.2%)	-
Others‡	3940 (23.7%)	-	1240 (24.9%)	-	499 (27.9%)	-	2166 (49.6%)	-

All participants included in the analysis are of European ancestry, which were confirmed by the 1000 Genomes Project reference panel.

*In ILCCO-OncoArray and TRICL studies, the age for lung cancer cases refers to the age at diagnosis, while for controls, it refers to the age at blood sample collection. In the PLCO study, age for both cases and non-cases was recorded at the time of randomization. In the UKB study, age for incident cases was defined at lung cancer diagnosis after enrollment, while for non-cases, it was defined at last follow-up. Age is described as mean ± standard deviation (SD).

†The smoking histories were assessed by baseline questionnaires in PLCO and UKB studies. Pack-year of smoking is expressed as mean ± SD. The pack-years of never smoker is imputed to be zero.

‡Other histological types include those recorded as unknown in the electronic records.

Abbreviation: LC, lung cancer; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; SCC, small-cell lung cancer.

Table 2. Significant gene-smoking interactions identified by the two-phase analytic strategy.

Biomarker	CHR	h^2	Gene	Discovery Phase			Replication Phase		Meta-analysis	
				Beta (95% CI)	*P*	*q*-FDR	Beta (95% CI)	*P*	Beta (95% CI)	*P*
Methylation										
cg13714459	15	0.04	*IREB2*	0.016 (0.011, 0.022)	7.83×10^{-09}	3.49×10^{-04}	0.005 (0.001, 0.009)	8.02×10^{-03}	0.009 (0.006, 0.012)	3.73×10^{-08}
cg19941054	15	0.05	*IREB2*	0.015 (0.010, 0.020)	4.37×10^{-08}	9.73×10^{-04}	0.005 (0.001, 0.009)	7.05×10^{-03}	0.008 (0.005, 0.011)	1.05×10^{-07}
cg20622131	15	0.02	*CHRNB4-RP11-335K5.2*	0.021 (0.013, 0.028)	1.04×10^{-07}	1.54×10^{-03}	0.006 (0.000, 0.011)	4.28×10^{-02}	0.011 (0.007, 0.016)	1.64×10^{-06}
cg13561554	15	0.12	*IREB2-HYKK*	-0.007 (-0.010, -0.004)	1.12×10^{-06}	8.31×10^{-03}	-0.002 (-0.004, -0.000)	1.57×10^{-02}	-0.004 (-0.005, -0.002)	3.62×10^{-06}
cg02662658	17	0.04	*RDM1*	-0.012 (-0.017, -0.007)	1.91×10^{-06}	1.08×10^{-02}	-0.004 (-0.008, -0.001)	2.55×10^{-02}	-0.007 (-0.010, -0.004)	2.58×10^{-06}
cg26989927	17	0.03	*RDM1*	-0.014 (-0.020, -0.008)	7.87×10^{-06}	2.93×10^{-02}	-0.006 (-0.011, -0.002)	9.58×10^{-03}	-0.009 (-0.013, -0.005)	1.73×10^{-06}
cg15034267	22	0.06	*GTPBP1*	0.009 (0.005, 0.012)	1.08×10^{-06}	8.31×10^{-03}	0.003 (0.000, 0.005)	4.38×10^{-02}	0.005 (0.003, 0.007)	7.19×10^{-06}
Expression										
RP11-326C3.14	11	0.09		0.017 (0.009, 0.024)	7.21×10^{-06}	1.63×10^{-02}	0.007 (0.002, 0.012)	5.17×10^{-03}	0.010 (0.006, 0.014)	1.32×10^{-06}