From tobacco smoking to mutational signature: the role of epigenetic changes in human cancers

Zhishan Chen
Vanderbilt University Medical Center
Wanqing Wen
Vanderbilt University Medical Center
Qiuyin Cai
Vanderbilt University Medical Center
Jirong Long
Vanderbilt University Medical Center
Ying Wang
Zhejiang University School of Medicine
Weiqiang Lin
Zhejiang University School of Medicine
Xiao-ou Shu
Vanderbilt University Medical Center
Wei Zheng
Vanderbilt University Medical Center
Xingyi Guo (Xingyi.guo@vumc.org)
Vanderbilt University Medical Center https://orcid.org/0000-0001-5269-1294

Research article

Keywords: gene expression, methylation, tobacco smoking, mutational signature, mediation analysis

Posted Date: January 17th, 2020

DOI: https://doi.org/10.21203/rs.2.21105/v1

Abstract

Background

Tobacco smoking is associated with a unique mutational signature in the human cancer genome. It is unclear whether tobacco smoking -altered DNA methylations and gene expressions affect smoking-related mutational signature.

Methods

We evaluated the smoking-related DNA methylation sites reported from five previous studies using peripheral blood cells to identify possible target genes. Using the mediation analysis approach, we evaluated whether the association of tobacco smoking with mutational signature was mediated through altered DNA methylation and expression of these target genes.

Results

Based on data obtained from 21,108 blood samples, we identified 374 smoking-related DNA methylation sites, annotated to 248 target genes. Using data from DNA methylations, gene expressions and smoking-related mutational signature generated from ~7,700 tumor tissue samples across 26 cancer types from The Cancer Genome Atlas (TCGA), we found 11 of the 248 target genes whose expressions were associated with smoking-related mutational signature at a Bonferroni-correction $P < 0.001$. This included four for head and neck cancer, and seven for lung adenocarcinoma. In lung adenocarcinoma, our results showed that smoking increased the expression of three genes, AHRR , GPR15 , and HDGF, and decreased the expression of two genes, CAPN8 , and RPS6KA1 , which were consequently associated with increased smoking-related mutational signature. Additional evidence showed that the elevated expression of AHRR (cg14817490) and GPR15 (cg19859270), were associated with smoking-altered hypomethylations. Lastly, we showed that the elevated expression of HDGF and decreases expression of RPS6KA1, were associated with poor survival of lung cancer patients.

Conclusions

Our findings provide novel insights into the contributions of tobacco smoking to carcinogenesis through complex molecular mechanisms of the elevated mutational signature by altered DNA methylations and gene expressions.

Background

Somatic mutations are one of the most common causes of carcinogenesis in humans [1, 2]. Recent studies using data from The Cancer Genome Atlas (TCGA) have created a landscape of somatic mutations in each cancer genome, ranging from hundreds to thousands of somatic mutations across multiple cancer types [2, 3]. To explore the biological processes of somatic mutations, Alexandrov and colleagues developed a mathematical framework to deconvolute them into mutational signatures. The

approach characterized 96 mutation classifications that included six substitution types, together with a flanking base pair to the mutated base [3]. More than 30 mutational signatures have been identified across cancer types in TCGA [3, 4]. Certain mutational signatures were associated with tobacco smoking, exposure to ultraviolet (UV) light, aging, deficient mismatch repair (MMR), mutations in POLE, increased activity of the APOBEC family of cytidine deaminases, and DNA polymerase POLH [3–5]. In particular, a smoking-related mutational signature featured by predominantly C > A mutations with a transcriptional strand bias was observed in multiple human cancer types, including lung adenocarcinoma, lung small cell carcinomas, head and neck squamous, liver, larynx, oral cavity, and esophagus cancers [3, 6, 7].

Tobacco smoking is a well-known risk factor for multiple cancer types, especially lung cancer [8–10]. DNA methylation, one of the major forms of epigenetic modification, essentially plays a regulatory role in gene expression. It has been a focus of multiple studies as a potential underlying molecular mechanism for tobacco smoking-related cancers. Previous epigenome-wide association studies (EWAS) have reported thousands of DNA methylations at CpG sites associated with tobacco smoking in blood, buccal cell and tumor-adjacent normal lung tissue samples [11–18]. These epidemiological studies showed that tobacco smoking was consistently associated with DNA hypomethylated CpG sites in specific genes such as AHRR (encoding aryl-hydrocarbon receptor repressor) and GPR15 (encoding G protein-coupled receptor 15) in multiple studies [19]. In particular, Stueve and colleagues identified seven smoking-associated hypomethylated CpG sites in adjacent normal tissues from 237 lung cancer patients. Of note, five of the seven sites including a hypomethylated CpG site in AHRR had been reported by previous blood-based EWAS, which suggests that methylation biomarkers identified from blood samples might reflect methylation changes in the target tissues [15].

In our study, we evaluated the previously reported smoking-related DNA methylations from a total of 21,108 blood samples to identify candidate target genes [11–13, 17, 18]. Using data from DNA methylations, gene expressions and smoking-related mutational signature generated from approximately 7,700 tumor samples across 26 cancer types, we evaluated the associations of expression of these target genes with the smoking-related mutational signature for each cancer type. Using a mediation approach, we further evaluated whether the association of tobacco smoking with the mutational signature may be mediated through an altered expression of these target genes. Similar analyses were performed to evaluate the association of tobacco smoking with the gene expression mediated through smoking-altered DNA methylation.

Methods

Data resources

We collected the previously reported smoking-related methylations in blood samples from five previous EWAS, including Joehanes et al., 2016 (N = 15,907) [13], Zeilinger et al., 2013 (N = 2,272) [18], Besingi and Johansson, 2014 (N = 432) [12], Tsaprouni et al., 2014 (N = 920) [17], and Ambatipudi et al., 2016 (N = 940) [11]. In the discovery stage, we only used the 2,622 methylations at CpG sites reported from the

study with the largest sample size (N = 15,907). In the replication stage, we only used methylations at CpG sites where we observed consistent associations in at least one other study at an adjusted $P < 0.05$ (Fig. 1). We annotated methylation sites to their target genes based on the annotation from the Bioconductor package FDb.InfiniumMethylation.hg19 (version 2.2.0).

This study utilized multiple dimension datasets, including matched gene expression, DNA methylation, and clinical data that included age, gender and tobacco smoking. This was generated from 7757 samples in 26 cancer types from TCGA. The sample size for each cancer type was listed: adrenocortical carcinoma (n = 78), bladder urothelial carcinoma (n = 407), breast invasive carcinoma (n = 972), cervical squamous cell carcinoma and endocervical adenocarcinoma (n = 278), colon adenocarcinoma (n = 272), lymphoid neoplasm diffuse large B-cell lymphoma (n = 37), esophageal carcinoma (n = 181), glioblastoma multiforme (n = 161), head and neck squamous cell carcinoma (n = 495), kidney chromophobe (n = 65), kidney renal clear cell carcinoma (n = 321), kidney renal papillary cell carcinoma (n = 274), acute myeloid leukemia (n = 125), brain lower grade glioma (n = 506), liver hepatocellular carcinoma (n = 354), lung adenocarcinoma (n = 507), lung squamous cell carcinoma (n = 474), ovarian serous cystadenocarcinoma (n = 211), pancreatic adenocarcinoma (n = 171), prostate adenocarcinoma (n = 492), rectum adenocarcinoma (n = 86), sarcoma (n = 233), skin cutaneous melanoma (n = 445), stomach adenocarcinoma (n = 411), uterine corpus endometrial carcinoma (n = 173), and uterine carcinosarcoma (n = 28). All the data were downloaded from TCGA using the Broad Institute Genome Data Analysis Center (GDAC) Firehose portal (stamp data/analyses__2016_01_28) through Firebrowse. Detailed information about datasets, analyses, and data sources are described at Firebrowse (http://gdac.broadinstitute.org/).

For gene expressions, the normalized expression levels for genes in tumor tissue samples were measured by RNA-Seq by Expectation Maximization (RSEM). To create a better distribution for downstream analysis, a log2 transfer of the RSEM values was applied. We further transformed the gene expression levels across samples for each cancer type using an inverse normalizing transformation method.

For DNA methylation, the data (Level 3) from the Illumina Infinium HumanMethylation450 BeadChip array for each sample in TCGA was measured. The Beta value of the methylation levels of each of the methylation sites were transformed to M value based on the equation ,

$$\mathrm{M} = \log_2\left(\frac{Beta}{1-Beta}\right),$$

using the function beta2m from the bioconductor package lumi (version 2.32.0) for the downstream analysis.

A total of 30 somatic mutational signatures for each sample in TCGA have been characterized from mSignatureDB (http://tardis.cgu.edu.tw/msignaturedb). We downloaded the data and only analyzed the known tobacco-associated “mutational signature 4” reported in the mSignatureDB, corresponding to

tobacco-associated mutational signature in this study. We measured the enrichment score of this mutational signature for each sample (details described in our previous work [20]).

The analysis of predicted neoantigen load

We downloaded the number of neoantigen loads for each sample from TCIA and applied log2 transfer to fit it into a better distribution. Mutational neoantigens were predicted by the use of HLA typing and MHC class I/II binding capabilities. The established neoantigen prediction algorithm NetMHCcons [21] was applied to missense somatic mutations to estimate their binding affinity to the HLA alleles. A more detailed analysis of the processing has been described in previous literature [22, 23].

Statistical analysis

The distribution for relative contribution of smoking-related mutational signature to overall mutation burden is severely right-skewed. To better fit regression models, we used the ordinal semi-parametric regression models [24] to evaluate the associations of smoking-related mutational signature with tobacco smoking, gene expression and DNA methylation. The analyses were implemented in the 'orm' function from the 'rms' library of the R package [24]. To evaluate the enrichment of the significant associations for the 248 smoking-related target genes, we compared the proportion of the significant associations from samples of 248 gene randomly selected from the whole genome - this process was repeated 1000 times. To explore the mediation effects of DNA methylation on the association of tobacco smoking with smoking-related gene expression and the mediation effects of the smoking-related gene expression on the association of tobacco smoking with the smoking-related mutational signature, we conducted mediation analyses using the R package 'mediation' [25] to estimate the average direct effect (ADE) and the average causal mediation effect (ACME) of the mediators, which represent the population averages of these causal mediation and direct effects. All the analyses were adjusted for age and gender. To estimate the association between the smoking-related gene expression and overall survival of lung cancer patients, we conducted survival analysis using the Cox proportional hazards model with the adjustment of age, gender and tobacco smoking.

Results

Identifying blood-based DNA methylations associated with tobacco smoking

To identify smoking-related DNA methylations at CpG sites, we evaluated previously reported methylations in blood samples from five EWAS, including Joehanes et al., 2016 (N = 15,907), Zeilinger et al., 2013 (N = 2,272), Besingi and Johansson, 2014 (N = 432), Tsaprouni, 2014 (N = 920), and Ambatipudi et al., 2016 (N = 940) (Fig. 1A) [11–13, 17, 18]. For our discovery data, we used a total of 2,622

methylations at CpG sites reported by Joehanes et al's study, which had the largest sample size. In the replication stage, we kept only those methylations at CpG sites which showed consistent associations in at least one of the remaining four studies (at the significance level of Bonferroni-correction $P < 0.05$) (Supplementary table 1; see Materials and Methods). In the end, we identified a total of 374 smoking-related DNA methylations at CpG sites, annotated to 248 target genes (Fig. 1A; Supplementary table 2). Of the 374 DNA methylations, the majority were hypomethylated CpG sites (n = 252, 67.4%), compared to hypermethylated CpG sites (n = 122, 32.6%).

Identifying genes associated with the smoking-related mutational signature in a pan-cancer study

The smoking-related mutational signature was characterized in TCGA samples in previous studies [3, 26] (Fig. 1B). Utilizing this study, we used the relative contribution of the mutational signature to overall mutation burden, with values ranging from 0 to 1, for each sample across 26 cancer types in TCGA (see Materials and Methods). Using regression analyses, adjusting for gender and age, we observed that tobacco smoking was significantly associated with increased smoking-related mutational signature in lung adenocarcinoma ($P = 1.75 \times 10^{-9}$; Fig. 1C). In line with previous studies, we observed that the contributions of smoking-related mutational signature to the overall mutation burdens varied in different cancers, with the most enrichments being observed in lung adenocarcinoma (median of contribution: 42%) and lung carcinoma (median of contribution: 35%) (Fig. 1D). Using regression analyses, adjusting for gender and age (see Materials and Methods), we evaluated the associations between the expressions of the identified 248 smoking-related target genes and smoking-related mutational signature for each cancer type. Of these target genes, we found that 234 genes were associated with smoking-related mutational signature in 19 cancer types (at a $P < 0.05$) (Supplementary table 3), suggesting that these genes were over-represented ($P < 0.001$ for the enrichment analysis, see Materials and Methods). At a more strict threshold of a $P < 1 \times 10^{-4}$, a total of 59 genes were identified in six cancer types: breast (n = 2), colon (n = 1), head and neck (n = 24), lung adenocarcinoma (n = 28), lung carcinoma (n = 2), and melanoma (n = 2) (Fig. 1E; Supplementary table 3).

In the end, using a Bonferroni-correction of $P < 0.001$ (corresponding to a raw $P = 5.0 \times 10^{-8}$, given 20,000 tests in a genome-wide level), we identified four genes for head and neck cancer and seven genes for lung adenocarcinoma. Specifically, for head and neck cancer, the expression levels of three genes, NFE2L2, RMND5A and SLC44A1, were associated with increased smoking-related mutational signature, while an inverse association was observed for one gene, ARRB1 (Fig. 1F, Table 1). For lung adenocarcinoma, we found that the expression levels of three genes, GPR15, HDGF, and AHHR, were associated with increased smoking-related mutational signature, while an inverse association was observed for the other four genes, NWD1, KCNQ1, CAPN8 and RPS6KA1 (Fig. 1F, Table 1). GPR15 showed the most significant association with a $P < 2.22 \times 10^{-16}$ (Table 1).

Table 1
Associations between smoking-associated mutational signature and expression of candidate genes (Bonferroni-correction P < 0.01).

Cancer type	Gene	Beta	P
head and neck (N = 495)	NFE2L2	0.54	4.1×10^{-11}
	RMND5A	0.56	2.0×10^{-10}
	SLC44A1	0.56	2.9×10^{-10}
	ARRB1	-0.46	5.1×10^{-8}
	FAM60A	0.44	5.8×10^{-8}
	RHOG	-0.43	5.9×10^{-8}
lung adenocarcinoma (N = 507)	GPR15	0.44	2.2×10^{-16}
	NWD1	-0.40	2.0×10^{-13}
	HDGF	0.42	1.9×10^{-12}
	AHRR	0.34	6.6×10^{-10}
	KCNQ1	-0.29	3.9×10^{-8}
	CAPN8	-0.27	4.4×10^{-8}
	RPS6KA1	-0.30	5.0×10^{-8}

"N" refers to sample size for each cancer type. A regression analysis was constructed to include tobacco smoking-associated mutational signature as a dependent variable and gene expression levels as the independent variable for each gene of each cancer type.

Mediation effects of the identified seven genes on the association of smoking with mutational signature in lung adenocarcinoma

For the identified seven genes for lung adenocarcinoma, we evaluated the associations between their expression and tobacco smoking (see Materials and Methods). We found that tobacco smoking was significantly associated with an increased expression of AHRR, GPR15 and HDGF with a $P = 6.9 \times 10^{-5}$, $P = 2.7 \times 10^{-7}$ and $P = 3.3 \times 10^{-4}$, respectively, and a decreased expression of CAPN8 and RPS6KA1 with a $P = 9.6 \times 10^{-4}$ and P = 0.01, respectively (Fig. 2A; Supplementary table 4). Using a mediation analysis approach, we further estimated the ACME of the expression of these genes that would be altered by

smoking on the mutational signature. We found that they showed significant mediation effects on the association of smoking with the signature (Fig. 2B, C). Specifically, we observed a significant percentage of ACME for the smoking-related gene expressions: 13.4% (95% CI: 0.046 and 0.256) with a $P = 2.0 \times 10^{-4}$ for AHRR, 9.8% (95% CI: 2.4% and 21.7%) with a $P = 2.2 \times 10^{-3}$ for CAPN8, 22.8% (95% CI: 11.3% and 39.4%) with a $P < 1 \times 10^{-4}$ for GPR15, 12.3% (95% CI: 4.7% and 24.6%) with a $P = 8.0 \times 10^{-4}$ for HDGF, and 8.6% (95% CI: 0.5 and 20.6%) with a P = 0.032 for RPS6KA1 (Fig. 2C; Table 2).

Table 2

Results from a mediation analysis of the direct effects of tobacco smoking and causal mediation (indirect) effects of gene expression that would be altered by tobacco smoking, on the mutational signature in lung adenocarcinoma ($P < 0.05$).

Gene	Effect *	Beta	95% CI		P
			Lower	Upper	
AHRR	ACME	4.5×10^{-4}	1.6×10^{-4}	8.3×10^{-4}	$< 1.0 \times 10^{-4}$
	ADE	2.9×10^{-3}	1.7×10^{-3}	4.1×10^{-3}	$< 1.0 \times 10^{-4}$
	Total Effect	3.3×10^{-3}	2.1×10^{-3}	4.5×10^{-3}	$< 1.0 \times 10^{-4}$
	Prop	13.4%	4.6%	25.6%	2.0×10^{-4}
CAPN8	ACME	3.4×10^{-4}	8.2×10^{-5}	6.8×10^{-4}	$< 1.0 \times 10^{-4}$
	ADE	3.0×10^{-3}	1.8×10^{-3}	4.2×10^{-3}	$< 1.0 \times 10^{-4}$
	Total Effect	3.3×10^{-3}	2.1×10^{-3}	4.5×10^{-3}	$< 1.0 \times 10^{-4}$
	Prop	9.8%	2.4%	21.7%	2.2×10^{-3}
GPR15	ACME	7.7×10^{-4}	3.9×10^{-4}	1.2×10^{-3}	$< 1.0 \times 10^{-4}$
	ADE	2.6×10^{-3}	1.4×10^{-3}	3.7×10^{-3}	$< 1.0 \times 10^{-4}$
	Total Effect	3.4×10^{-3}	2.2×10^{-3}	4.4×10^{-3}	$< 1.0 \times 10^{-4}$
	Prop	22.8%	11.3%	39.4%	$< 1.0 \times 10^{-4}$
HDGF	ACME	4.2×10^{-4}	1.6×10^{-4}	7.6×10^{-4}	$< 1.0 \times 10^{-4}$
	ADE	2.9×10^{-3}	1.8×10^{-3}	4.1×10^{-3}	$< 1.0 \times 10^{-4}$
	Total Effect	3.4×10^{-3}	2.2×10^{-3}	4.5×10^{-3}	$< 1.0 \times 10^{-4}$
	Prop	12.3%	4.7%	24.6%	8.0×10^{-4}
RPS6KA1	ACME	3.0×10^{-4}	1.8×10^{-5}	6.7×10^{-4}	0.040
	ADE	3.0×10^{-3}	1.9×10^{-3}	4.2×10^{-3}	$< 1.0 \times 10^{-4}$
	Total Effect	3.3×10^{-3}	2.1×10^{-3}	4.5×10^{-3}	$< 1.0 \times 10^{-4}$

"*": "ACME" refers to the average causal mediation effects. "ADE" refers to the average direct effects. "Prop" refers to the proportion of ACME related to total affects.

	Prop	8.6%	5%	20.6%	0.032

"*": "ACME" refers to the average causal mediation effects. "ADE" refers to the average direct effects. "Prop" refers to the proportion of ACME related to total affects.

Mediation effects of smoking-related DNA methylation on the association of smoking with gene expression in lung adenocarcinoma

In the above mediation analysis, we found that five genes, AHRR, CAPN8, GPR15, HDGF, and RPS6KA1, mediated the association between smoking and mutational signature in lung adenocarcinoma. For these, six smoking-related DNA methylations, cg11554391, cg14817490, cg21446172, cg19859270, cg00867472 and cg13092108, have been reported in blood cells [11–13, 17, 18]. We further evaluated the associations between these methylations and tobacco smoking in lung adenocarcinoma. In line with previous findings, we found that consumed tobacco smoking was significantly associated with hypomethylations at the CpG sites: cg11554391 (AHRR), cg14817490 (AHRR), and cg19859270 (GPR15) ($P < 0.05$ for all; Fig. 3A; Supplementary table 4). Next, we evaluated the association between the methylation at each CpG site and gene expression. Interestingly, our results showed that the smoking-altered hypomethylation was associated with an elevated expression for each gene ($P < 0.05$ for all): AHRR (cg11554391 or cg14817490) and GPR15 (cg19859270), indicating that these smoking-altered hypomethylations likely play an up-regulation role in their gene expression (Fig. 3B; Supplementary table 5). In particular, these hypomethylated CpG sites are located in regions with evidence of enhancer activities associated with their target genes (Supplementary Fig. 1).

Using a mediation analysis approach, we further estimated the ACME of the methylations that would be altered by smoking on gene expressions. We found that the methylations at two CpG sites, AHRR (cg14817490, $P = 0.03$) and GPR15 (cg19859270, $P < 1 \times 10^{-4}$), showed significant mediation effects on the association of smoking with gene expression (Fig. 3C, D; Table 3). Specifically, we observed a significant percentage of ACME for both smoking-related DNA methylations: 8.5% (95% CI: 8% and 24.5%) with a $P = 0.03$ for AHRR, and 15.9% (95% CI: 5.2% and 32.9%) with a $P < 1.0 \times 10^{-4}$ for GRP15 (Fig. 3D; Table 3).

Table 3

Results from a mediation analysis of the direct effects of tobacco smoking and causal mediation (indirect) effects of DNA methylations that would be altered by tobacco smoking, on the gene expression in lung adenocarcinoma ($P < 0.05$).

CpG	Effect *	Beta	95% CI		P
			Lower	Upper	
cg14817490 (AHRR)	ACME	6.5×10^{-4}	5.7×10^{-5}	1.5×10^{-3}	0.03
	ADE	6.5×10^{-3}	3.1×10^{-3}	1.0×10^{-2}	$< 1.0 \times 10^{-4}$
	Total Effect	7.2×10^{-3}	3.8×10^{-3}	1.1×10^{-2}	$< 1.0 \times 10^{-4}$
	Prop	8.5%	8%	24.5%	0.03
cg19859270 (GPR15)	ACME	1.5×10^{-3}	4.6×10^{-4}	2.9×10^{-3}	$< 1.0 \times 10^{-4}$
	ADE	7.8×10^{-3}	4.4×10^{-3}	1.1×10^{-2}	$< 1.0 \times 10^{-4}$
	Total Effect	9.3×10^{-3}	5.8×10^{-3}	1.3×10^{-2}	$< 1.0 \times 10^{-4}$
	Prop	15.9%	5.2%	32.9%	$< 1.0 \times 10^{-4}$

"*" ACME refers to the average causal mediation effects. ADE refers to the average direct effects (ADE). "Prop" refers to the proportion of ACME related to total affects.

Supplementary Data

Expression of HDGF and RPS6KA1 associated with overall survival of lung cancer patients

To explore the association between overall survival of lung cancer patients and the identified five genes that mediated the association between smoking and mutational signature in lung adenocarcinoma, we conducted the Cox regression analysis using data from TCGA (see Materials and Methods). Our results revealed that the elevated expression level of HDGF was associated with the reduced overall survival of lung cancer patients, while an opposite trend was observed for RPS6KA1 when comparing the high level of gene expression (> median) versus low level (< = median) (Hazard Ratio [HR] = 1.45 and HR = 0.61, P = 0.02 and $P = 1.5 \times 10^{-3}$ for HDGF and RPS6KA1, respectively) (Supplementary Fig. 2A, B). These findings are in line our initial results that tobacco smoking increased expression level of HDGF and decreased expression level of RPS6KA1. No significant associations with overall survival of lung cancer patients were observed for other three genes.

Discussion

In the present study, a total of 374 smoking-related methylations annotated to 248 target genes were identified using strict statistical criteria from previous EWASs in blood samples. Using data from TCGA, we identified a total of 11 candidate genes of 248 target genes whose expressions were associated with smoking-related mutational signature, including four in head and neck cancer and seven in lung adenocarcinoma. Of seven genes for lung adenocarcinoma, our results further showed that smoking increased the expression of three genes, AHRR, GPR15, and HDGF, and decreased the expression of two genes, CAPN8, and RPS6KA1. These smoking-altered gene expressions were consequently associated with increased smoking-related mutational signature. In addition, our results showed that the elevated expression of AHRR (cg14817490) and GPR15 (cg19859270), were associated with smoking-altered hypomethylations.

Our analysis focused on the identified 374 blood-based methylations associated with tobacco smoking, which have strong evidence of statistical associations from previous studies. In particular, the initial discovery of methylations associated with tobacco smoking is based on a study with the largest sample size we have found so far (N = 15,907) (see Materials and Methods) [13]. In addition to studies of blood, two studies have investigated methylations associated with tobacco smoking in buccal cells (N = 790) [16] and tumor adjacent normal lung tissue (N = 237) [15]. Notably, both studies had limited sample sizes and were insufficient in statistical power to identify smoking-related methylation sites, while they have revealed evidence that blood-based methylation biomarkers could reflect changes in their target tissues. Recently, Ma and Li performed pathway enrichment analyses based on 320 smoking-affected genes identified in blood. Their results showed that 104 of these genes were significantly enriched in pathways associated with the etiology of different cancers [27]. Consistent with these findings, two recent epidemiology studies showed that smoking-related hypomethylations in blood cells were associated with lung cancer risk [28, 29]. Thus, our study shows a connection of blood-based methylations with tobacco smoking-related mutational signature in tumor tissue. Our study not only provides an understanding of the molecular mechanisms underlying tobacco smoking carcinogenesis, but also can potentially lead to a new avenue for target intervention.

Using mediation analyses, we concluded that two genes, AHRR and GPR15, significantly contributed to smoking-related mutational signature, mediated by smoking-altered methylation and gene expression in lung adenocarcinoma. These conclusions are supported by multiple layers of evidence from previous literature. It is known that the AHRR gene encodes a suppressor of the aryl hydrocarbon receptor (AHR). It is involved in the AHR signaling cascade, which plays an essential role in dioxin toxicity, including polycyclic aromatic hydrocarbons (PAHs), an important class of smoking carcinogens [30, 31]. It has been documented that the AHRR gene is associated with tobacco smoking, based on EWAS from blood, buccal cell and normal lung tissue [11–18]. In recent studies, the hypomethylated CpG sites in the AHRR gene in pre-diagnostic peripheral blood samples was reported to be associated with lung cancer risk [28, 29]. Based on in vitro experiments in lung tissue from both humans and mice, the evaluated AHRR expression has been validated by tobacco smoking-altered methylations [14]. In addition to AHRR, GPR15 encodes an orphan G-protein-coupled receptor involved in the regulation of innate immunity and T-cell trafficking in the intestinal epithelium [32, 33]. Previous studies have shown that the GPR15 gene is

associated with tobacco smoking, which is based on EWAS from blood, while no studies have reported it for lung cancer [11, 12, 16–18]. Our finding was the first to identify this smoking-related novel gene that significantly contributed to smoking-related mutational signature in lung cancer.

Our results showed three additional genes, CAPN8, HDGF and RPS6KA1, contributed to smoking-related mutational signature, mediated by gene expression altered by tobacco smoking in lung adenocarcinoma. Tobacco smoking-related methylations in these genes have been reported in the previous EWAS in blood samples. However, we did not observe that these methylations were associated with tobacco smoking in lung adenocarcinoma, although consistent association directions were observed for HDG and RPS6KA1 (Data not shown). Notably, unlike the studies in large sample size from blood studies, the statistical analysis in detecting association between DNA methylation and tobacco smoking is still challenge in tumor tissues due to possible factors, such as tumor heterogeneity, potential confounders, and limited sample size. In fact, our focus on the analysis of the reported blood-based smoking-related DNA methylation sites could identify reliably smoking-related target genes and reduce the possibility of reverse causation. Nevertheless, given the tissue-specificities of some methylations in blood, further studies with a large sample size are still needed to replicate the associations for these candidate tobacco smoking-related genes in lung adenocarcinoma. In fact, our results showed that smoking-related methylations of these genes were associated with decreased expressions of these genes ($P < 0.01$ for all), indicating that they may play a down-regulation role in their gene expression in lung adenocarcinoma (Supplementary Fig. 3). Further in vitro or in vivo functional assays are needed to validate the genes that are affected by tobacco smoking in lung cancer.

It is known that neoantigens (or neoepitopes) result from missense somatic mutations in cancer cells [34]. However, how smoking-related mutational signature contribute to neoantigen loads remain unclear. We additionally evaluated the associations between smoking-related mutation signature and predicted neoantigen loads (see Materials and Methods). We observed that smoking-related mutational signature were significantly associated with increased neoantigen loads in three cancer types, head and neck, lung adenocarcinoma, and lung carcinoma (see Materials and Methods). An inverse association was observed in melanoma ($P < 1 \times 10^{-4}$ for all; Supplementary Fig. 4A, B; Supplementary Table 6). The most significant association was observed in lung adenocarcinoma with a $P < 2.2 \times 10^{-16}$. In addition, we also observed that neoantigen loads were associated with all five identified genes ($P < 1 \times 10^{-5}$) and tobacco smoking ($P = 2.16 \times 10^{-11}$) in lung adenocarcinoma (Supplementary Fig. 4C, D). In particular, the expressions of AHRR and GPR15 had associations with an increased predicted neoantigen load with $P = 7.6 \times 10^{-10}$ and $P = 7.7 \times 10^{-7}$, respectively (Supplementary Fig. 4D). Thus, our findings may provide new clues to explore the biological and immunological mechanisms through which smoking-related mutational signature may be involved in carcinogenesis, and provide potential genomic biomarkers for the development of cancer prevention and immunotherapy.

Conclusions

Our results showed that the smoking-altered DNA methylations and gene expressions play an important role in contributing to smoking-related mutational signature in human cancers. Our results also indicated that tobacco-smoking plays an important role in clinical significance, likely affecting genes with the impact on overall survival of lung cancer patients. Our findings have provided novel insights into the contributions of tobacco smoking to carcinogenesis, especially lung cancer, through complex molecular mechanisms of the elevated mutational signature by altered DNA methylations and gene expressions.

Abbreviations

AHRR, Aryl-Hydrocarbon Receptor Repressor

CAPN8, Calpain 8

EWAS, Epigenome-Wide Association Studies

GPR15, G Protein-Coupled Receptor 15

HDGF, Heparin Binding Growth Factor

HR, Hazard Ratio

RPS6KA1, Ribosomal Protein S6 Kinase A1

TCGA, The Cancer Genome Atlas

Declarations

Acknowledgements

We thank TCGA for providing valuable data resources for the research. We thank Marshal Younger for assistance with editing and manuscript preparation. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University.

Availability of data and material

The normalized expressions of gene and DNA methylation were downloaded from the TCGA using the Broad Institute Genome Data Analysis Center (GDAC) Firehose portal through Firebrowse (stamp data/analyses__2016_01_28, http://gdac.broadinstitute.org). Somatic mutational signatures were downloaded from mSignatureDB (http://tardis.cgu.edu.tw/msignaturedb). Neoantigen data was downloaded from TCIA.

Funding

This work was supported by the research development fund from Vanderbilt University Medical Center.

Contributions

Conception and design: XG; Acquisition of data and material support: XG and ZC; Analysis and interpretation of data: XG, ZC and WW; Generation of tables/figures: ZC; Writing, review, and/or revision of the manuscript: XG, ZC, WW, QC, JL, YW, WL, XS and WZ; Study supervision: XG. All authors read and approved the final manuscript.

Competing interests

No potential conflicts of interest were disclosed.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

References

1. Garraway LA, Lander ES: Lessons from the cancer genome. *Cell* 2013, 153(1):17-37.
2. Martincorena I, Campbell PJ: Somatic mutation in cancer and normal cells. *Science* 2015, 349(6255):1483-1489.
3. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL *et al*: Signatures of mutational processes in human cancer. *Nature* 2013, 500(7463):415-+.
4. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, Stratton MR: Clock-like mutational processes in human somatic cells. *Nat Genet* 2015, 47(12):1402-1407.
5. Supek F, Lehner B: Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* 2017, 170(3):534-547 e523.
6. Alexandrov LB, Ju YS, Haase K, Van Loo P, Martincorena I, Nik-Zainal S, Totoki Y, Fujimoto A, Nakagawa H, Shibata T *et al*: Mutational signatures associated with tobacco smoking in human cancer. *Science* 2016, 354(6312):618-622.
7. Helleday T, Eshtad S, Nik-Zainal S: Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet* 2014, 15(9):585-598.
8. Gandini S, Botteri E, Iodice S, Boniol M, Lowenfels AB, Maisonneuve P, Boyle P: Tobacco smoking and cancer: A meta-analysis. *Int J Cancer* 2008, 122(1):155-164.
9. Hecht SS: Lung carcinogenesis by tobacco smoke. *Int J Cancer* 2012, 131(12):2724-2732.
10. Sasco AJ, Secretan MB, Straif K: Tobacco smoking and cancer: a brief review of recent epidemiological evidence. *Lung Cancer-J Iaslc* 2004, 45:S3-S9.

11. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, Barrdahl M, Boeing H, Aleksandrova K, Trichopoulou A *et al*: Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics-Uk* 2016, 8(5):599-618.
12. Besingi W, Johansson A: Smoke-related DNA methylation changes in the etiology of human disease. *Human Molecular Genetics* 2014, 23(9):2290-2297.
13. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan WH, Xu T, Elks CE, Aslibekyan S *et al*: Epigenetic Signatures of Cigarette Smoking. *Circ-Cardiovasc Gene* 2016, 9(5):436-447.
14. 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(5):843-851.
15. Stueve TR, Li WQ, Shi J, Marconett CN, Zhang T, Yang C, Mullen D, Yan C, Wheeler W, Hua X *et al*: Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. *Hum Mol Genet* 2017, 26(15):3014-3027.
16. Teschendorff AE, Yang Z, Wong A, Pipinikas CP, Jiao YM, Jones A, Anjum S, Hardy R, Salvesen HB, Thirlwell C *et al*: Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. *Jama Oncol* 2015, 1(4):476-485.
17. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, Vinuela A, Grundberg E, Nelson CP, Meduri E *et al*: Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics-Us* 2014, 9(10):1382-1396.
18. Zeilinger S, Kuhnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, Weidinger S, Lattka E, Adamski J, Peters A *et al*: Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS One* 2013, 8(5):e63812.
19. 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. *Clin Epigenetics* 2015, 7:113.
20. Chen Z, Wen W, Beeghly-Fadiel A, Shu XO, Diez-Obrero V, Long J, Bao J, Wang J, Liu Q, Cai Q *et al*: Identifying Putative Susceptibility Genes and Evaluating Their Associations with Somatic Mutations in Human Cancers. *Am J Hum Genet* 2019, 105(3):477-492.
21. Karosiene E, Lundegaard C, Lund O, Nielsen M: NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 2012, 64(3):177-186.
22. Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D, Hackl H, Trajanoski Z: Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep* 2017, 18(1):248-262.
23. Chen Z, Wen W, Bao J, Kuhs KL, Cai Q, Long J, Shu XO, Zheng W, Guo X: Integrative genomic analyses of APOBEC-mutational signature, expression and germline deletion of APOBEC3 genes, and immunogenicity in multiple cancer types. *BMC Med Genomics* 2019, 12(1):131.

24. Harrell FE: Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd Edition. *Springer Ser Stat* 2015:1-582.
25. Imai K, Keele L, Tingley D: A General Approach to Causal Mediation Analysis. *Psychol Methods* 2010, 15(4):309-334.
26. Huang PJ, Chiu LY, Lee CC, Yeh YM, Huang KY, Chiu CH, Tang P: mSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Res* 2018, 46(D1):D964-D970.
27. Ma Y, Li MD: Establishment of a Strong Link Between Smoking and Cancer Pathogenesis through DNA Methylation Analysis. *Sci Rep* 2017, 7(1):1811.
28. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, Chung J, Fasanelli F, Guida F, Campanella G *et al*: DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *Int J Cancer* 2017, 140(1):50-61.
29. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A *et al*: Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun* 2015, 6:10192.
30. Haarmann-Stemmann T, Bothe H, Kohli A, Sydlik U, Abel J, Fritsche E: Analysis of the transcriptional regulation and molecular function of the aryl hydrocarbon receptor repressor in human cell lines. *Drug Metab Dispos* 2007, 35(12):2262-2269.
31. Murray IA, Patterson AD, Perdew GH: Aryl hydrocarbon receptor ligands in cancer: friend and foe. *Nat Rev Cancer* 2014, 14(12):801-814.
32. Kim SV, Xiang WV, Kwak C, Yang Y, Lin XW, Ota M, Sarpel U, Rifkin DB, Xu R, Littman DR: GPR15-mediated homing controls immune homeostasis in the large intestine mucosa. *Science* 2013, 340(6139):1456-1459.
33. Koks S, Koks G: Activation of GPR15 and its involvement in the biological effects of smoking. *Exp Biol Med (Maywood)* 2017, 242(11):1207-1212.
34. Chen DS, Mellman I: Oncology Meets Immunology: The Cancer-Immunity Cycle. *Immunity* 2013, 39(1):1-10.

Supplementary Data

Supplementary Table 1: A collection of candidate blood-based methylations at CpG sites reported from five previous epigenome wide association studies.

Supplementary Table 2: A list of 374 candidate blood-based methylation CpG sites and genes identified from both discovery and replication studies, at an adjusted $P < 0.05$.

Supplementary Table 3: Associations between smoking-associated mutational signature and expression of candidate genes for each cancer type ($P < 0.05$).

Supplementary Table 4: Associations between tobacco smoking and expression of candidate genes and associations between tobacco smoking and methylation of candidate CpG sites.

Supplementary Table 5: Association between expression of candidate genes and methylation at each CpG site.

Supplementary Table 6: Associations between smoking-associated mutational signature and predicted neoantigen load for each cancer type.

Supplementary Figure 1: The epigenetic landscape of regions with methylations at three candidate CpG sites.

Supplementary Figure 2: Gene expression of *HDGF* and *RPS6KA1* associated with overall survival of lung cancer patients.

Supplementary Figure 3: Associations between gene expressions and methylations at three CpG sites.

Supplementary Figure 4: Smoking-related mutational signature contributed to neoantigen load in multiple cancer types.

Figures

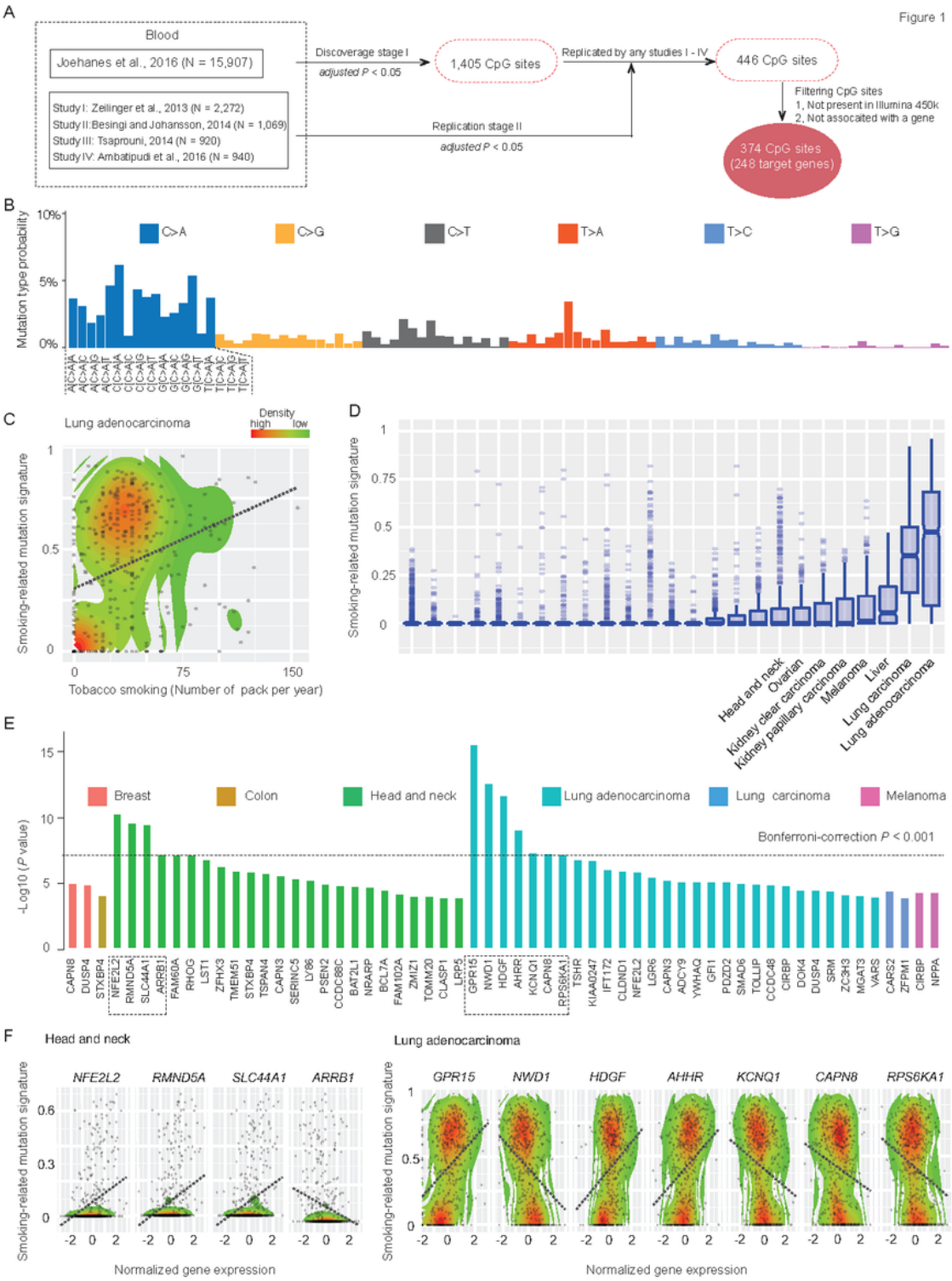

Figure 1

Identification of genes and their associations with smoking-related mutational signature. A) A flow chart to illustrate the identification of candidate smoking-related DNA methylations from the previously reported blood-based methylations in five EWAS. “N” represents the sample size for each study. B) Smoking-related mutational signature displayed according to the 96 substitution classifications characterized by six substitution types, together with a flanking base pair to the mutated base

(Alexandrov et al 2013). C) A scatter plot indicating tobacco smoking correlated with known smoking-related mutational signature in lung adenocarcinoma. The dotted line refers to association coefficient. Each point represents one sample. The x axis represents the number of packs per year for each sample, the y axis represents the contribution of smoking-related mutational signature to overall mutation burden for each sample. The color from red to green refers to a higher to lower density of samples (this note applies to all other figure legends). D) Box plots of the enrichment score of smoking-related mutational signature across 26 cancer types. E) Bar plots indicating the P value of associations between the candidate genes and smoking-related mutational signature in six cancer types. Only genes with a P value of less than 1 × 10-4 were presented. The dashed dot box highlights the genes with significant associations at a Bonferroni-correction P < 0.001. F) Scatter plots for each gene with significant associations at a Bonferroni-correction P < 0.001. From the left to the right panel, four genes in head and neck and seven genes in lung adenocarcinoma are presented.

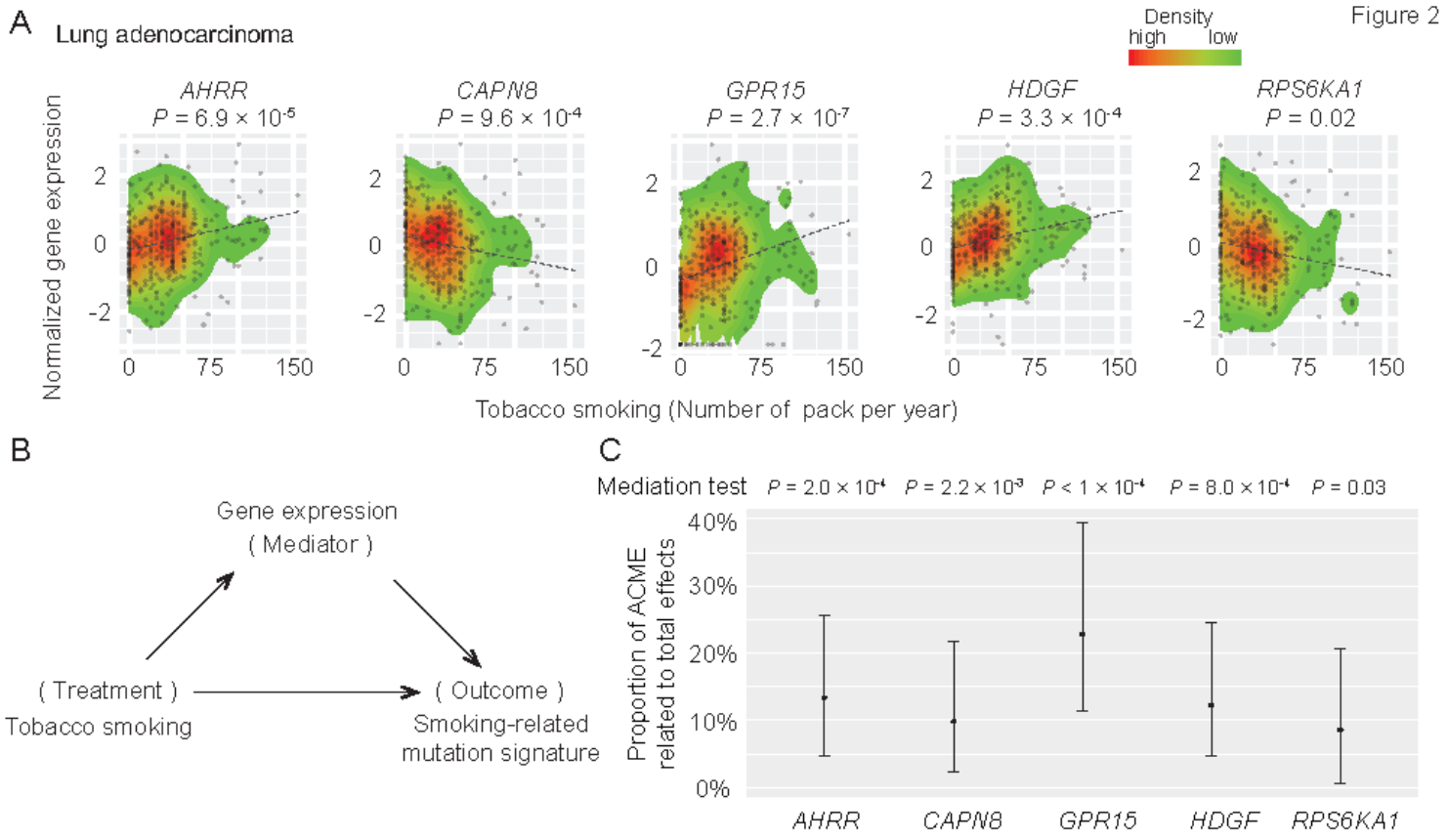

Figure 2

Mediation analysis illustrating the effect of the expression of five genes that would be altered by smoking on smoking-related mutational signature in lung adenocarcinoma. A) Scatter plots indicating the statistical significance between five candidate genes and tobacco smoking in lung adenocarcinoma. B) A diagram to illustrate a mediation analysis framework, where gene expression can be a mediator to affect smoking-related mutational signature. C) Five candidate genes are presented with significant mediation effect, at P < 0.05. “ACME” refers to the average causal mediation effects via gene expression on smoking-related mutational signature.

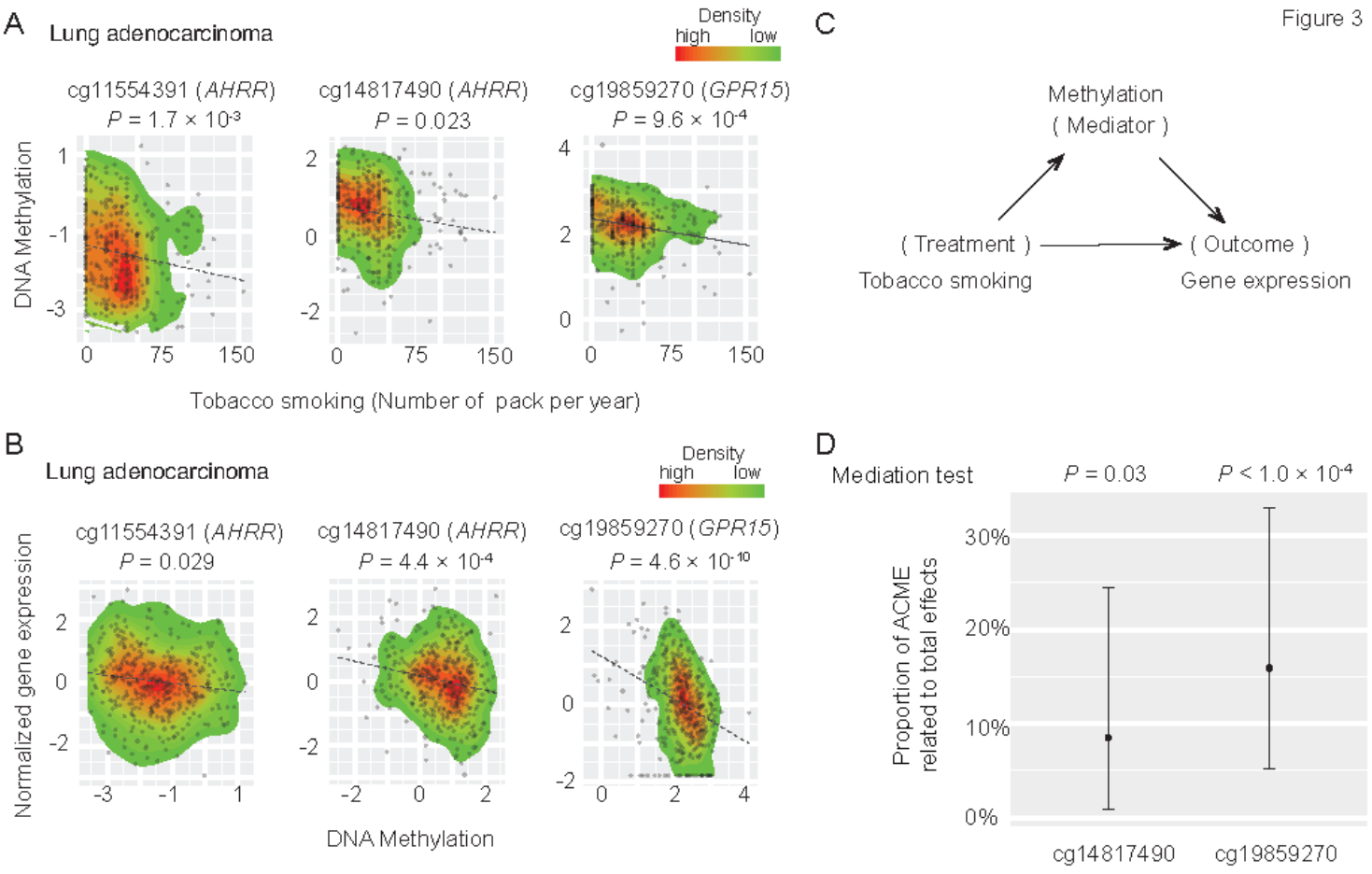

Figure 3

Mediation analysis illustrating the effect of tobacco smoking-altered methylation on gene expression in lung adenocarcinoma. A) Scatter plots indicating the statistical significance of associations between methylations at three candidate CpG sites and tobacco smoking in lung adenocarcinoma. B) Scatter plots indicating negative correlations between DNA methylation at three candidate CpG sites and gene expression in lung adenocarcinoma. C) A diagram to illustrate a mediation analysis framework, where DNA methylation can be a mediator to affect the expression of tobacco smoking-altered genes. D) Two candidate CpG sites are presented with significant mediation effects on gene expression, at $P < 0.05$. “ACME” refers to the average causal mediation effects via DNA methylation on gene expression.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables.xlsx
- SupplementaryFigure4.tif
- SupplementaryFigure2.tif
- SupplementaryFigure3.pdf

- SupplementaryFigure1.pdf