Untargeted metabolomics reveals multiple metabolites influencing smoking-related DNA methylation

Yunfeng Huang¹, Qin Hui¹, Douglas I Walker², Karan Uppal², Jack Goldberg³, Dean P Jones², Viola Vaccarino¹ & Yan V Sun^{*,1,4}

¹Department of Epidemiology, Emory University Rollins School of Public Health, Atlanta, GA 30322, USA

²Department of Medicine, Division of Pulmonary, Allergy & Critical Care Medicine, Emory University School of Medicine, Atlanta, GA 30322, USA

³Vietnam Era Twin Registry, VA Epidemiologic Research & Information Center & Department of Epidemiology, University of Washington School of Public Health, Seattle, WA 98195, USA

⁴Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

* Author for correspondence: yan.v.sun@emory.edu

Aim: We conducted a joint metabolomic–epigenomic study to identify patterns of epigenetic associations with smoking-related metabolites. **Patients & methods:** We performed an untargeted metabolome-wide association study of smoking and epigenome-wide association studies of smoking-related metabolites among 180 male twins. We examined the patterns of epigenetic association linked to smoking-related metabolites using hierarchical clustering. **Results:** Among 12 annotated smoking-related metabolites identified from a metabolome-wide association study, we observed significant hypomethylation associated with increased level of N-acetylpyrrolidine, cotinine, 5-hydroxycotinine and nicotine and hypermethylation associated with increased level of 8-oxoguanine. Hierarchical clustering revealed common and unique epigenetic–metabolic associations related to smoking. **Conclusion:** Our study suggested that a joint metabolome–epigenome approach can reveal additional details in molecular responses to the environmental exposure to understand disease risk.

First draft submitted: 13 August 2017; Accepted for publication: 27 November 2017; Published online: 12 March 2018

Keywords: epigenetic epidemiology • epigenetics • epigenome • EWAS • exposome • metabolome • metabolome-wide association study • methylome • metabolome-wide association study • twin

Tobacco smoking is a leading public health concern affecting more than 1 billion people worldwide [1] and causing an estimated 5.7 million deaths per year [2]. Despite a global effort to reduce tobacco smoking, it remains the major cause of chronic diseases accounting for 6.9% of years of life lost and 5.5% of disability adjusted life-years in 2010 worldwide [3]. There are over 4500 identified chemicals in tobacco, and many of them contribute to disease risk through different biological pathways [4]. The metabolomics approach measures hundreds of endogenous and exogenous (xenobiotic) small molecules simultaneously, providing an opportunity to identify biomarkers of exposure to cigarette smoke and markers that reflect host-related metabolic adaptations. Studies have used targeted and untargeted metabolites to identify smoking-related biomarkers [5–8]. Metabolites involved in caffeine, vitamin, steroid, amino acid and carbohydrate pathways were reported to be associated with cigarette smoking [8,9]. Smoking-related changes in human serum metabolites are also reversible after smoking cessation, consistent with the known cardiovascular risk reduction [7].

DNA methylation occurs at the cytosine bases of eukaryotic DNA, which are converted to 5-methylcytosine by DNA methyltransferase enzymes. As one of the main forms of epigenetic modifications, DNA methylation plays a crucial role in regulating gene expression by modifying the access to promoters where transcription factors should bind [10]. Recent studies showed that DNA methylation also affects smoking-related pathways and smoking-induced disease [11]. Several replicable smoking-related loci in the candidate genes as well as global methylation differences have been reported by earlier epigenetic studies [12]. Following the introduction of epigenome-wide

CpG microarrays (e.g., Infinium HumanMethylation BeadChips), a large number of smoking-related CpG sites and genes (e.g., *F2RL3* and *AHRR*) have been discovered via epigenome-wide association studies (EWASs) [13–16]. Furthermore, the broad effect of tobacco smoking on human cells and tissues has been suggested as reversible, and smoking cessation may reverse DNA methylation to the state of never smokers, albeit with variable speed [17]. However, although much has been learned about smoking-related epigenetics, limited knowledge exists of which smoking-related chemicals and pathways induce epigenetic modifications.

Metabolic traits represent intermediate phenotypes linking genetic and environmental factors to end points of complex disorders [18]. Epigenetic regulation of metabolic processes via DNA methylation and gene expression may play a major role in biological processes; however, few studies have examined the association between DNA methylation and metabolomics changes. In 2014, Petersen *et al.* conducted the first EWAS examining the association between 649 blood metabolic traits and 457,004 CpG sites from 1814 participants of the Kooperative Gesundheitsforschung in der Region Augsburg (KORA) population study [19]. They suggested that the association between CpG methylation and metabolite concentrations can be explained by genetic confounders and by nongenetic external (environmental) factors.

Tobacco smoking, as one of the most prevalent environmental risk factors, has enormous adverse effects on human health. EWAS have shown strong associations between smoking exposure and DNA methylation. However, the biochemical mechanism of smoking-induced DNA methylation change remains unclear. The advancement of metabolomics and epigenomics allows for a combined study to investigate the effect of smoking on many metabolic and epigenetic pathways at the population level. Our study explored the epigenome-wide associations between smoking-related metabolites and DNA methylation and identified patterns of epigenetic association with numerous smoking-related metabolites. In addition, the twin design allows us to identify effects attributed to unshared environmental factors by controlling for genetic confounding.

Patients & methods

Study population

Our peripheral blood-based metabolomic and DNA methylomic data were obtained from samples of the Emory Twin Study (ETS). The ETS consists of 307 middle-aged male monozygotic and dizygotic twin pairs from the Vietnam Era Twin Registry [20] who were born between 1946 and 1956 [21,22]. All twins were examined in pairs at the Emory University General Clinical Research Center between 2002 and 2010. Zygosity information was determined by DNA analysis. All twins involved in this study were male and Caucasians. The ETS was approved by the Emory Institutional Review Board, and all participants signed an informed consent.

Phenotypic measurements

Twins were fed the same diet the night before the assessments and instructed to refrain from smoking. All measurements were performed in the morning after an overnight fast, and both twin pairs were tested at the same time. All medications were held for about 24 h prior to testing. All biochemical assays for each twin pair were processed in the same analytical run. A medical history and a physical examination were obtained from all twins. Weight and height were measured and used to calculate BMI. Cigarette smoking was classified into current smoker (any number of cigarettes) versus never or past smoker. Venous blood samples were drawn for the collection of plasma and peripheral blood leukocytes (PBL). Plasma and PBL samples were stored at -80°C until the biomedical or molecular assays.

DNA methylation methods

Genomic DNA were extracted from PBL samples, examined for quality and quantity and standardized for the DNA methylation assay. Total amount of 0.5 μg genomic DNA was epityped using the Illumina HumanMethylation450 BeadChip (450 K) in two batches of 142 and 78 twins, respectively. Genomic DNA was first bisulfite converted, then whole-genome amplified, enzymatically fragmented and purified. Samples were then hybridized in batches of 12 to each BeadChip. Each DNAm site was quantified using β -values. More detailed procedures of DNA methylation measurements were reported previously [23]. Methylation sites were excluded from analyses if they overlapped with SNPs, or were not uniquely mapped to the reference genome [24]. More detailed information on quality control steps for DNA methylation data can be found in Figure 1.

Figure 1. Summary of data quality control for study population and DNA methylation.

ETS: Emory Twin Study.

Metabolomics measurements

High-resolution metabolomics profiling of monozygotic and dizygotic male twins was completed using liquid chromatography with high-resolution mass spectrometry. Plasma samples were randomized into blocks of 20 prior that included pooled quality control samples and prepared for analysis by treating 65 μ l of plasma with two volumes of acetonitrile containing a mixture of eight stable isotopic standards [25,26]. All plasma samples were analyzed in triplicate by dual C_{18} chromatography (C_{18} , Higgins Analytical, Inc., Mountain View, CA, USA) with acetonitrile/formic acid gradient. The high-resolution mass spectrometer (Q-Exactive, Thermo Fisher Scientific, Waltham, MA, USA) was operated at 70,000 resolution defined by full width at half maximum (FWHM) and data were collected using both positive and negative electrospray ionization (ESI). Detailed information on sample preparation, chromatography conditions and mass spectrometer operation can be found elsewhere [27]. Raw data were extracted using adaptive processing by apLCMS [28] in combination with systematic variation in parameter settings, statistical data filtering and data merger by xMSanalyzer [29]. Detected mass-to-charge (m/z) ratio features were defined by accurate mass m/z , retention time and ion intensity. Positive ESI and negative ESI data were analyzed and summarized separately.

Metabolome-wide association study of current smoking status

Relative abundance for about 20,000 metabolic features in plasma samples was collected by high-resolution metabolomics, evaluated for quality control and filtered according to coefficient of variation and percentage of missing values prior to analysis. The association analysis included 20,035 metabolic features. Potential metabolite identities were first determined by performing an online search (10 p.p.m. mass accuracy) against the METLIN database [30], and the Human Metabolomics Database [31]. In the metabolome-wide association study analyses, the abundance of metabolomics features was logarithm transformed and modeled as the dependent variable; linear mixed models were implemented to explore the association between current smoking status (current vs noncurrent

smoking) and metabolomics features. Age, BMI and alcohol drinking were included as covariates, potential correlation within twin pairs was accounted for by including a random intercept [32]. A subset of metabolites associated with current smoking status, including cotinine, hydroxycotinine, nicotine, caffeine and ornithine were confirmed by comparison of ion dissociation patterns and retention time to authentic reference standards (level 1 confirmation) or online databases (level 2) [33].

Epigenome-wide association analysis

To identify the association between smoking-related metabolites and DNA methylation, we modeled the β -value as the dependent variable, with the plasma level of smoking-related metabolites as primary independent variables in the association analysis. Linear mixed models were implemented in a multiple regression framework, and two random intercepts were included to account for: potential correlation among DNA samples processed on the same 450 K methylation array (12 samples on one array) and potential correlation within twin pairs [23,32]. Age in years, BMI and calculated proportion of PBL subtypes were included in the models as potential confounders. A reference-based method [34] was implemented to infer PBL subtypes including B cells, granulocytes, monocytes, natural killer cells, CD4⁺ and CD8⁺ T cells. Ten epigenetic principle components were calculated using a method developed by Barfield *et al.* [35] and included in the model as covariates to account for population stratification in the EWAS analysis. A total of 139,183 CpG sites (with no missing data) that lie within ≤ 50 bp of an SNP variant identified in the 1000 Genomes Project (Phase I) with MAF > 0.01 were included in the calculation for principle components. Among 218 twins with DNA methylation data, 37 of them were excluded due to lack of metabolomic data, and one was excluded due to missing BMI (Figure 1).

For initial CpG site discovery analyses, we used a false discovery rate (FDR) of 0.05 to account for multiple testing. Manhattan plots of the epigenome-wide results were created using a FDR-adjusted threshold of 0.05 for significant sites. We also produced normalized quantile–quantile plots comparing observed to expected p-values. To replicate significant results from previous reports, we compared the significant CpG sites from blood cell-based epigenetic association studies of active smoking exposure, with current smoking-associated CpG sites (current vs noncurrent smokers) determined through our data. We calculated the minimum base pair distance between the significant sites identified by previous studies of cigarette smoking and the significant sites identified by smoking-related metabolites in our study. To further explore the independent epigenetic associations, we also conducted the epigenome-wide association analysis with smoking-related metabolites controlling for current smoking status.

Clustering analyses of smoking-related metabolomic & DNA methylation markers

We summarized the t-test statistics from the association analyses with smoking-related metabolites, and 58 smoking-associated significant CpG sites that were categorized as reported $\geq$ three-times by the systematic review were included in the cross-omic analyses [14]. Hierarchical clustering was implemented to identify the patterns of DNA methylation profile among smoking-related metabolites and the patterns of metabolic association among smoking-associated CpG sites. This method groups the most similar t-statistics across all epigenetic association tests in two dimensions. The results of subset analyses with and without adjustment of current smoking were presented in separate 2D heat maps.

To understand the genetic and environmental contributions to these smoking-related metabolites and DNAm sites, we used a structural equation modeling method implemented in OpenMX [36] to partition the additive genetic (A), common environmental (C) and unique environmental (E) variance for each identified metabolomics feature and CpG site in twin samples [23]. Associations between smoking-related metabolites and the 58 smoking-associated significant CpG sites were also analyzed using twin-specific models separating between-pair and within-pair effects. Two coefficients for both the mean values of each twin pair and the distinct within-twin difference for each twin member (individual value minus the twin average) were included in this model; the mean values of a twin pair (between-twin) represent the measurements similar to the general unrelated population, but the within-twin difference represents the twin-specific effect due to unshared environment adjusted for shared genetic and environmental factors [32]. t-test statistics of within-pair effects were summarized and clustered to identify the patterns of DNA methylation profile due to unshared environmental factors.

All statistical analyses were performed in the R statistical environment version 3.1.2 [37]. R package nlme was used to implement linear mixed effect model. R package heatmap.2 was used to generate the heat maps based on 2D clustering analysis.

Table 1. Demographic information for the study sample of 180 middle-aged male twins.

Variable	n (%) or mean (SD)			
	MZ twins (n = 130)	DZ twins (n = 46)	Singletons (n = 4)	Total (n = 180)
Age	55.6 (3.3)	56.5 (3.0)	57.3 (3.3)	55.9 (3.3)
BMI	29.3 (4.9)	29.7 (3.8)	29.8 (3.0)	29.4 (4.6)
Smoker:				
– Current	46 (35.4)	11 (23.9)	0 (0.0)	57 (31.7)
– Past	54 (41.5)	24 (52.2)	3 (75.0)	81 (45.0)
– Never	30 (23.1)	11 (23.9)	1 (25.0)	42 (23.3)
Ever smoker:				
– Age started smoking	17.2 (5.5)	17.9 (3.6)	13.7(0.6)	17.3 (5.0)
– Pack years	35.7 (23.2)	32.2 (21.0)	35.1(19.3)	34.8 (22.4)

DZ: Dizygotic; MZ: Monozygotic.

Figure 2. Untargeted metabolome-wide association analysis with current smoking. The green line represents a Bonferroni corrected p-value of 0.05. m/z ratio: Mass-to-charge ratio.

Results

Among 180 male twins included in our study, the mean age was 55.9 years (55.5 for current smokers and 56.0 for noncurrent smokers); the mean BMI was 29.4 kg/m² (27.8 for current smokers and 30.1 for noncurrent smokers) and 57 (31.7%) were current smokers at the visit of blood draw (Table 1). After correction for multiple testing (Bonferroni corrected p-value <0.05), 17 out of 7508 m/z features detected in positive ionization mode and 70 out of 12,527 m/z features detected with negative ionization were significantly associated with current smoking, adjusted for age, BMI and alcohol drinking. The raw p-values were ordered by m/z ratio and stratified by ion charge (Figure 2). Using the Metlin Database, annotation of these metabolomic features were obtained by matching possible combination of metabolites and adducts within 10 p.p.m. of difference. Totally, 12 out of 30 top metabolomic features were successfully mapped to known metabolites including cotinine (two isoforms with 12C and 13C isotopes, Pearson $r = 0.996$), 5-hydroxycotinine, N-acetylpyrrolidine, nicotine, 8-oxoguanine, norcotinine, hexy glucoside, hydroxypyridine, ornithine and caffeine (two isoforms with 12C and 13C, Pearson $r = 0.942$) (Table 2). Cotinine, hydroxycotinine, caffeine and ornithine were confirmed by comparison of ion dissociation patterns and retention time to authentic reference standards (level 1 confirmation). Nicotine was

Table 2. Summary of mapped metabolites from metabolome-wide association analysis with current smoking.

Metabolites	m/z	Retention time	Estimate	SE	t-statistic	p-value	FDR-corrected p-value [†]
13C cotinine	178.105552	70.40	1.856	0.075	24.85	2.74E-41	2.06E-37
5-hydroxycotinine	193.097251	65.77	1.765	0.088	20.10	1.90E-34	7.14E-31
Cotinine	177.102216	70.16	1.766	0.089	19.86	4.54E-34	1.14E-30
N-acetylpyrrolidine	155.117834	68.18	1.685	0.097	17.34	5.68E-30	1.07E-26
Nicotine	163.122952	73.26	1.574	0.104	15.14	4.24E-26	6.36E-23
Norcotinine	163.086519	68.25	1.201	0.135	8.89	7.58E-14	7.45E-11
Hexy glucoside	287.146904	83.07	-0.964	0.141	-6.86	9.71E-10	7.29E-07
13C caffeine	196.091004	70.64	-0.704	0.152	-4.63	1.25E-05	0.005
8-oxoguanine	148.027207	69.68	-0.696	0.155	-4.49	2.21E-05	0.008
Caffeine	195.087649	70.39	-0.670	0.152	-4.42	2.86E-05	0.009
Hydroxypyridine	96.0448451	83.82	0.672	0.157	4.28	4.82E-05	0.014
Ornithine	133.097239	129.26	0.600	0.155	3.87	2.12E-04	0.049

[†]FDR-adjusted p-value on metabolome-wide level.
FDR: False discovery rate; m/z: Mass-to-charge ratio; SE: Standard error.

Table 3. Summary of epigenome-wide significant associations of smoking-related metabolites.

Metabolites	Number of epigenome-wide significant CpG sites [†]	
	Bonferroni adjusted	FDR adjusted
N-acetylpyrrolidine	16	65
Cotinine	14	45
5-hydroxycotinine	12	19
Nicotine	5	9
8-oxoguanine	3	3

[†]Adjusted for age, BMI, PBL and ten PCs.
FDR: False discovery rate; PBL: Peripheral blood leukocyte; PC: Principle component; WBC: White blood cell proportion.

confirmed by comparison to spectral databases (level 2 confirmation). All but hexy glucoside, 8-oxoguanine and caffeine were elevated among current smokers. Among top smoking-associated metabolomic features, caffeine (m/z of 196.091004 and 195.087649) was negatively associated with current smoking status. We combined the two isoforms of cotinine and caffeine as a single metabolite in the following EWAS due to high correlation and similar biological mechanisms. The isoform with a stronger epigenetic association was reported. By partitioning the genetic and environmental contributions on these smoking-related metabolites using OpenMX, a structural equation modeling method available as an R library; we estimated that the mean heritability of these metabolites was 20.4% (see Supplementary Table 1, Supplementary Material Part 1). In addition, unshared environmental effects were the main driver of smoking-related metabolomics changes (explained on average 74.7% of the variance).

In the EWAS of smoking-associated metabolites, 65 CpG sites were significantly associated with N-acetylpyrrolidine after multiple testing correction (FDR-corrected $p < 0.05$) (Table 3 & Figure 3). In addition, 45, 19, nine and three CpG sites were significantly associated with cotinine, 5-hydroxycotinine, nicotine and 8-oxoguanine, respectively (Table 3). Hypomethylation of most identified CpG sites was associated with increased level of metabolites including N-acetylpyrrolidine (58 out of 65), cotinine (41 out of 45), 5-hydroxycotinine (18 out of 19) and nicotine (nine out of nine). However, hypermethylation of all three significant CpG sites was associated with increased level of 8-oxoguanine. More detailed information of all epigenome-wide significant associations identified through smoking-related metabolites was summarized in Supplementary Material Part 3. No CpG site was significantly associated with norcotinine, hexy glucoside, hydroxypyridine and ornithine or caffeine after multiple testing correction. We identified a number of novel epigenetic associations with these smoking-related metabolites by comparing genomic locations to previously reported DNA methylation sites [14].

With additional adjustment for current smoking status, most epigenome-wide associations diminished. Among those CpG sites significantly associated with smoking-related metabolites, only 52, 16, six, five and one CpG

Figure 3. Epigenome-wide association with N-acetylpyrrolidine, adjusted for age, body mass index, peripheral blood leukocyte and ten principle components. (A) Quantile–quantile plot (red straight line: $y = x$, red curves: 95% CI for the global null hypothesis, inflation factor = 1.1). **(B)** Manhattan plot (red line: false discovery rate significance level of 0.05).

sites were marginally associated ($p < 0.05$) with N-acetylpyrrolidine, cotinine, 5-hydroxycotinine, nicotine and 8-oxoguanine after controlling for current smoking, respectively.

To further understand the epigenetic associations with smoking-related metabolites, we conducted a subset analysis focusing on replicated smoking-associated CpG sites [14]. Among 58 CpG sites strongly associated with cigarette smoking, 54 were replicated with significant association (FDR-corrected $p < 0.05$) with current smoking status in our twin sample (see Supplementary Material Part 4). A total of 54 CpG sites were significantly associated with N-acetylpyrrolidine (FDR-corrected $p < 0.05$). In addition, 55, 57, 50, 44, 37, 26, 25 and three CpG sites were significantly associated with cotinine, 5-hydroxycotinine, nicotine, 8-oxoguanine, norcotinine, hexy glucoside, ornithine and caffeine, respectively (FDR-corrected $p < 0.05$) (see Supplementary Material, Part 4). No significant association was found with hydroxypyridine. Conditional on current smoking status, most of the epigenetic associations with smoking-related metabolites weakened substantially. No significant (FDR-corrected $p < 0.05$) association was found with 5-hydroxycotinine, nicotine, 8-oxoguanine, norcotinine, hexy glucoside, hydroxypyridine, ornithine and caffeine. However, 23 and five out of 58 smoking-related CpG sites remained significant for N-acetylpyrrolidine and cotinine (FDR-corrected $p < 0.05$) adjusted for current smoking status. We estimated the genetic (i.e., heritability), common and unshared environmental components of each CpG sites. Most CpG sites are driven by environmental effects with on average 16.8 and 41.2% of the variance explained by common and unshared environmental factors, respectively (see Supplementary Table 2, Supplementary Material Part 1).

We clustered t-statistics of epigenetic associations between 58 smoking-related CpG sites and ten smoking-related metabolites as well as current smoking status (Figures 4–6). We identified common and unique patterns of associations through 2D hierarchical clustering. Consistent with the effect of current smoking status, increased levels of cotinine, N-acetylpyrrolidine, 5-hydroxycotinine and nicotine were associated with hypomethylation for most smoking-related CpG sites. In contrast, increased levels of hexy glucoside and 8-oxoguanine were associated with hypermethylation for these smoking-related CpG sites. Metabolites directly linked to nicotine metabolism, such as nicotine, cotinine and hydroxycotinine shared similarly epigenetic association profile as that for current smoking status (Figure 4). In contrast, other smoking-specific metabolites, such as hexy glucoside and 8-oxoguanine, revealed different patterns of epigenetic associations among smoking-related CpG sites. Adjusted for current smoking status, clustering of t-statistics has similar patterns among smoking-related metabolites compared with the patterns identified before conditioning on current smoking status, but the degree of association is substantially reduced (Figure 5). Within-pair associations between smoking-related metabolites and CpG sites showed a similar pattern in DNA methylation profile comparing to the associations identified through general population-based model (Figure 6). Slight reduction in strength of associations was observed after separating the genetic and environmental components, which suggested that the associations between smoking-related CpG sites and metabolites are mostly driven by environmental factors, not caused by genetic confounding.

Discussion

Combining metabolomic and epigenomic data, we identified metabolites in the nicotine metabolism pathway that have common associations with DNA methylation. N-acetylpyrrolidine was found to be the metabolite bearing the strongest association with DNA methylation comparing to other common recognized smoking biomarkers such as cotinine and nicotine. We also discovered unique epigenetic association patterns involving caffeine, hexy glucoside and 8-oxoguanine. One of smoking-related metabolites, hydroxypyridine, was not associated with DNA methylation from either candidate CpG or EWAS analysis. By clustering the epigenetic associations with smoking-related metabolites, we identified several genes and potential pathways that might be involved in smoking metabolism and downstream effects. Genes such as *AKT3* [38,39], *PTK2* [40], *F2RL3* [41] and *RARA* [42] are involved in cell proliferation and apoptosis evasion in the development of cancers. Hypomethylation at *F2RL3* was reported to be associated with lung cancer in human patients [43]. An enrichment analysis using DAVID Bioinformatics Resources 6.8 [44] identified pathways such as chemokine signaling (genes *AKT3*, *GNG12* and *PTK2*), glutathione metabolism (genes *ANPEP* and *SRM*) and neuroactive ligand-receptor interaction (genes *F2RL3* and *AVPR1B*), sharing association patterns with smoking-related metabolites.

Smoking has broad adverse effects on human health. Recent EWASs show consistent epigenome-wide associations with smoking across races and age groups [14–16,45]; however, the understanding of effective chemicals and biochemical mechanisms inducing smoking-related DNA methylation change remains unclear. Improvements in metabolomic measurements provide us an opportunity to capture a more comprehensive chemical profile of

Figure 5. Hierarchical clustering of t-statistics of associations between smoking-related metabolites and replicated smoking-associated CpG sites controlling for current smoking status.

environmental exposures [46]. In our study, we identified and annotated 12 metabolic features that are associated with current smoking including cotinine (two isoforms with 12C and 13C isotopes), 5-hydroxycotinine, N-acetylpyrrolidine, nicotine, 8-oxoguanine, norcotinine, hexy glucoside, hydroxypyridine, ornithine and caffeine (two isoforms with 12C and 13C). Notably, the plasma level of caffeine was found to be negatively associated with smoking status. In previous literature, smoking is positively associated with coffee intake (smokers tend to drink more coffee) [47]. Thus, the negative association between circulating caffeine levels and smoking cannot be explained by a confounding effect due to coffee intake. Since the participants had restricted diet and food intake during the visit and caffeine has relatively short half-life in human body (5–6 h), this observed negative association between caffeine levels and smoking supports that current smokers have faster caffeine metabolism than noncurrent smokers. In addition, epigenetic modification, as one of the molecular mechanisms affected by environmental exposures, plays a central role in assessing the biological responses and memories. An EWAS examining the association between circulating metabolites and PBL-based DNA methylation sites suggested that the associations between CpG methylation and metabolite concentrations can be explained by genetic confounders or by nongenetic external (environmental) factors [19]. Previous epigenomic studies have analyzed self-reported smoking status as the main measure, which can be misclassified and does not measure the large number of chemicals and metabolites induced by cigarette smoking. In this study, we observed numerous metabolites strongly associated with smoking status but only a subset of metabolites were associated with DNA methylation makers. Additionally, several metabolites showed independent epigenetic associations after controlling for smoking status, which suggests their likely role of epigenetic modifiers. Therefore, performing joint metabolomic and epigenomic analyses could potentially fine-map the biological pathways linking environmental exposures and human diseases via identification of novel biomarkers and intermediate molecular changes for the exposures and their biological responses.

This study combines two high-throughput technologies to measure metabolomic (over 20,000 features) and epigenomic (over 450,000 CpG sites) profiles to investigate biological responses to smoking using peripheral blood. Comparisons were made between current versus noncurrent smokers since previous studies have shown strongest epigenetic associations of current smoking status [16]. The participants were controlled for several environmental factors (e.g., diet, medication and smoking) before fasted blood draw to minimize the transient influences of diet on the metabolome. The twin design allows us to calculate the genetic (i.e., heritability) and environmental components of each metabolites and CpG sites. It also allows us to assess the metabolomic–epigenomic associations while controlling for genetic confounding. This joint metabolomic–epigenomic approach could also be expanded to other environmental exposures to identify multiple biological pathways in response to environmental exposures, either a single chemical or a composition of multiple potent chemicals. The joint metabolomic–epigenomic approach can potentially identify and fine-map functional effectors from common or complementary pathways, which may lead to improved understanding of complex phenotypic outcomes.

Our study also has several limitations. First, we are limited by available metabolomics and epigenomic technologies to fully explore the exposome. The findings of our joint omics study is restricted not only by the number of markers being accurately measured, but also the accuracy of annotations. Technological advancements continuously improve the coverage of epigenome [48] and metabolome [46]. However, the annotation of the untargeted metabolomics is still not optimal [49] leaving a large number of metabolomic features with uncertain or unknown identities. For better interpretation, we only included 12 uniquely annotated metabolites out of 30 top smoking-related metabolomics features (40%) in our study. Thus, additional pathways involved in smoking-related epigenetic changes may have been captured in this metabolomic data. In addition, the modest sample size only allowed us to identify metabolomic and epigenomic markers with relatively large effect sizes. The study population consists of only middle-aged male Caucasian twins, which reduced the generalizability of our findings, and the cross-sectional design cannot rule out the possibility that DNA methylation level affects concentration of circulating metabolites. Therefore, the results of this study provide a proof of concept for joint omics analysis of environmental exposures and warrant future studies with a larger longitudinal sample and more comprehensive metabolomic panel.

Conclusion & future perspective

In this study, we detected smoking-related metabolic products associated with cigarette smoking and other correlates in the plasma. Many of these smoking-related metabolites shared common effects on DNA methylation profile, but they also displayed unique patterns of epigenetic associations among known smoking-related CpG sites. For example, hydroxypyridine is strongly associated with cigarette smoking, but is unlikely a direct epigenetic modifier induced by cigarettes smoking. Both epigenome and metabolome capture environmental exposures and their

biological responses. The combination of this complimentary information has the potential to greatly improve understanding of biological mechanisms mediating environmental effects at multiple stages and molecular levels. Such joint metabolome–epigenome study design is a novel approach to investigate the biological effects in response to environmental factors by profiling two molecular layers of transient and long-term influences. Although each high-throughput technology measures a large number of molecular features, none of them provided the holistic view of the complex biosystem. To better understand the exposures and biological responses at the system level, future exposome research needs to consider multi-omics designs using carefully phenotyped population samples [50].

Summary points

- We performed joint analysis of epigenomic and untargeted metabolomic data, and identified smoking-related metabolites and their epigenetic associations among twins.
- We identified epigenome-wide significant associations with five smoking-related metabolites, including multiple nicotine metabolites and 8-oxoguanine.
- The joint metabolomic and epigenomic subset analysis with hierarchical clustering among replicated smoking-associated CpG sites showed common and unique metabolic association patterns induced by smoking.
- Our study suggested that a joint metabolome–epigenome design is powerful and effective to investigate the biological effects of environmental exposures by profiling two complementary molecular layers of transient and long-term influences.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at www.futuremedicine.com/doi/full/10.2217/epi-2017-0101

Acknowledgements

The authors thank the members of the VET Registry for their continued cooperation and participation.

Financial & competing interests disclosure

The United States Department of Veterans Affairs (VA) has provided financial support for the development and maintenance of the Vietnam Era Twin Registry. Numerous organizations have provided invaluable assistance, including VA Cooperative Study Program; Department of Defense; National Personnel Records Center; National Archives and Records Administration; the Internal Revenue Service; NIH; National Opinion Research Center; National Research Council, National Academy of Sciences; and the Institute for Survey Research, Temple University. This work was supported by the NIH (K24HL077506, R01 HL68630, R01 AG026255, R01 MH056120, R01 HL088726, R21 NS096455, R01 NR013520, P30 ES019776 and K24 MH076955); the National Institute of General Medical Sciences at the NIH (5K12 GM000680); the Emory University General Clinical Research Center (MO1-RR00039); and the American Heart Association (0245115N and 13GRNT17060002). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

References

Papers of special note have been highlighted as: ●● of considerable interest

1. Peto R, Lopez AD, Boreham J, Thun M, Heath C, Doll R. Mortality from smoking worldwide. *Br. Med. Bull.* 52(1), 12–21 (1996).
2. Ng M, Freeman MK, Fleming TD *et al.* Smoking prevalence and cigarette consumption in 187 countries, 1980–2012. *JAMA* 311(2), 183–192 (2014).
3. Lim SS, Vos T, Flaxman AD *et al.* A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 380(9859), 2224–2260 (2013).

4. WHO. *IARC monographs on the evaluation of carcinogenic risks to humans. Volume 88: formaldehyde, 2-butoxyethanol and 1-tert-butoxypropan-2-ol*. WHO (2006). <http://monographs.iarc.fr/ENG/Monographs/vol88/index.php>
5. Hsu P-C, Zhou B, Zhao Y *et al.* Feasibility of identifying the tobacco-related global metabolome in blood by UPLC-QTOF-MS. *J. Proteome Res.* 12(2), 679–691 (2013).
6. Wang-Sattler R, Yu Y, Mittelstrass K *et al.* Metabolic profiling reveals distinct variations linked to nicotine consumption in humans – first results from the KORA study. *PLoS ONE* 3(12), e3863 (2008).
7. Xu T, Holzapfel C, Dong X *et al.* Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC Med.* 11, 60 (2013).
8. Jones DP, Walker DI, Uppal K, Rohrbach P, Mallon COLTM, Go Y-M. Metabolic pathways and networks associated with tobacco use in military personnel. *J. Occup. Environ. Med.* 58, S111–S116 (2016).
9. Nikpay M, Goel A, Won H-H *et al.* A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* 47(10), 1121–1130 (2015).
10. Phillips T. The role of methylation in gene expression. *Nat. Educ.* 1(1), 116 (2008).
11. Bakulski KM, Fallin MD. Epigenetic epidemiology: promises for public health research. *Environ. Mol. Mutag.* 55(3), 171–183 (2014).
12. Philibert RA, Beach SRH, Brody GH. The DNA methylation signature of smoking: an archetype for the identification of biomarkers for behavioral illness. *Nebr. Symp. Motiv.* 61, 109–127 (2014).
13. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27 K discovery and replication. *Am. J. Hum. Genetics* 88(4), 450–457 (2011).
- **The first epigenome-wide association study on cigarette smoking.**
14. 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* 7, 113 (2015).
- **A thorough systematic review on epigenomic associations with smoking, which reports replicated smoking-related CpG sites for part of our joint metabolomic–epigenomic analysis.**
15. Ambatipudi S, Cuenin C, Hernandez-Vargas H *et al.* Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics* 8(5), 599–618 (2016).
16. Joehanes R, Just AC, Marioni RE *et al.* Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc. Genet.* 9(5), 436–447 (2016).
17. Zeilinger S, Kühnel B, Klopp N *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PLoS ONE* 8(5), e63812 (2013).
18. Suhre K, Gieger C. Genetic variation in metabolic phenotypes: study designs and applications. *Nat. Rev. Genet.* 13(11), 759–769 (2012).
19. Petersen AK, Zeilinger S, Kastenmuller G *et al.* Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. *Hum. Mol. Genet.* 23(2), 534–545 (2014).
20. Goldberg J, Curran B, Vitek ME, Henderson WG, Boyko EJ. The Vietnam Era Twin Registry. *Twin Res.* 5(5), 476–481 (2002).
21. Vaccarino V, Brennan M-L, Miller AH *et al.* Association of major depressive disorder with serum myeloperoxidase and other markers of inflammation: a twin study. *Biol. Psychiatry* 64(6), 476–483 (2008).
22. Vaccarino V, Lampert R, Bremner JD *et al.* Depressive symptoms and heart rate variability: evidence for a shared genetic substrate in a study of twins. *Psychosom. Med.* 70(6), 628–636 (2008).
23. Klebaner D, Huang Y, Hui Q *et al.* X chromosome-wide analysis identifies DNA methylation sites influenced by cigarette smoking. *Clin. Epigenetics* 8, 20 (2016).
24. Chen Y-A, Lemire M, Choufani S *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* 8(2), 203–209 (2013).
25. Soltow QA, Strobel FH, Mansfield KG, Wachtman L, Park Y, Jones DP. High-performance metabolic profiling with dual chromatography-Fourier-transform mass spectrometry (DC-FTMS) for study of the exposome. *Metabolomics* 9(1), 132–143 (2013).
26. Go Y-M, Walker DI, Liang Y *et al.* Reference standardization for mass spectrometry and high-resolution metabolomics applications to exposome research. *Toxicol. Sci.* 148(2), 531–543 (2015).
27. Walker DI, Pennell KD, Uppal K *et al.* Pilot metabolome-wide association study of benzo(a)pyrene in serum from military personnel. *J. Occup. Environ. Med.* 58, S44–S52 (2016).
28. Yu T, Park Y, Johnson JM, Jones DP. apLCMS–adaptive processing of high-resolution LC/MS data. *Bioinformatics* 25(15), 1930–1936 (2009).
29. Uppal K, Soltow QA, Strobel FH *et al.* xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics* 14(1), 15 (2013).
30. Smith CA, Maillie GO, Want EJ *et al.* METLIN: a metabolite mass spectral database. *Ther. Drug Monit.* 27(6), 747–751 (2005).
31. Wishart DS, Jewison T, Guo AC *et al.* HMDB 3.0 – the Human Metabolome Database in 2013. *Nucleic Acids Res.* 41(D1), D801–D807 (2012).

32. Carlin JB. Regression models for twin studies: a critical review. *Int. J. Epidemiol.* 34(5), 1089–1099 (2005).
- **Presents a thorough discussion on twin-specific regression methods to distinguish unshared environmental effects.**
33. Schymanski EL, Jeon J, Gulde R *et al.* Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environ. Sci. Technol.* 48(4), 2097–2098 (2014).
34. Houseman EA, Accomando WP, Koestler DC *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* 13(1), 86 (2012).
35. Barfield RT, Almlí LM, Kilaru V *et al.* Accounting for population stratification in DNA methylation studies. *Genet. Epidemiol.* 38(3), 231–241 (2014).
- **Presents a DNA methylation chip-based method to account for population stratification and potentially correct for global inflation in epigenome-wide association study analysis.**
36. Boker S, Neale M, Maes H *et al.* OpenMx: an open source extended structural equation modeling framework. *Psychometrika* 76(2), 306–317 (2011).
37. The R Project for Statistical Computing. www.r-project.org/
38. Stahl JM, Sharma A, Cheung M *et al.* Deregulated Akt3 activity promotes development of malignant melanoma. *Cancer Res.* 64(19), 7002–7010 (2004).
39. Cristiano BE, Chan JC, Hannan KM *et al.* A specific role for AKT3 in the genesis of ovarian cancer through modulation of G(2)-M phase transition. *Cancer Res.* 66(24), 11718–11725 (2006).
40. Tai YL, Chen LC, Shen TL. Emerging roles of focal adhesion kinase in cancer. *BioMed Res. Int.* 2015, 690690 (2015).
41. Kaufmann R, Rahn S, Pollrich K *et al.* Thrombin-mediated hepatocellular carcinoma cell migration: cooperative action via proteinase-activated receptors 1 and 4. *J. Cell. Physiol.* 211(3), 699–707 (2007).
42. Farias EF, Arapshian A, Bleiweiss JJ, Waxman S, Zelent A, Mira YLR. Retinoic acid receptor alpha2 is a growth suppressor epigenetically silenced in MCF-7 human breast cancer cells. *Cell Growth Differ.* 13(8), 335–341 (2002).
43. Zhang Y, Schöttker B, Ordóñez-Mena J *et al.* F2RL3 methylation, lung cancer incidence and mortality. *Int. J. Cancer* 137(7), 1739–1748 (2015).
44. DAVID Bioinformatics Resources 6.8. <https://david.ncifcrf.gov/>
45. Sun YV, Smith AK, Conneely KN *et al.* Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. *Hum. Genet.* 132(9), 1027–1037 (2013).
46. Walker DI, Go Y-M, Liu K, Pennell KD, Jones DP. Population screening for biological and environmental properties of the human metabolic phenotype. In: *Metabolic Phenotyping in Personalized and Public Healthcare*. Academic Press, CA, USA, 167–211 (2016).
47. Treur JL, Taylor AE, Ware JJ *et al.* Associations between smoking and caffeine consumption in two European cohorts. *Addiction (Abingdon, England)* 111(6), 1059–1068 (2016).
48. Aberg KA, McClay JL, Nerella S *et al.* Methylome-wide association study of schizophrenia. *JAMA Psychiatry* 71(3), 255 (2014).
49. Uppal K, Walker DI, Liu K, Li S, Go Y-M, Jones DP. Computational metabolomics: a framework for the million metabolome. *Chem. Res. Toxicol.* 29(12), 1956–1975 (2016).
- **Discusses the progress in exposome research and computational metabolomics, which provided both the theoretical and practical basis for our metabolome-wide association study analysis on smoking.**
50. Sun YV, Hu Y-J. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* 93, 147–190 (2016).
- **A thorough review on the designs and methods of integrative multi-omic research for complex human diseases, which enlightened our research hypothesis.**