Contribution of blood DNA methylation to the association between smoking and lung cancer

Arce Domingo-Relloso,[1] Roby Joehanes,[2] Zulema Rodriguez-Hernandez,[3] Karin Haack,[4] M. Daniele Fallin,[5] Jason G. Umans,[6] Lyle G. Best,[7] Tianxiao Huan,[8] Chunyu Liu,[9] Jiantao Ma,[10] Jose D. Bermudez,[11] Shelley A. Cole,[12] Dorothy A. Rhoades,[13] Daniel Levy,[14] Ana Navas-Acien,[15] and Maria Tellez-Plaza[16]

[1] ad3531@cumc.columbia.edu. Department of Biostatistics, Columbia University, USA.

[2] roby.joehanes@nih.gov. National Heart, Lung, and Blood Institute, USA.

[3] zulema.rodriguez@isciii.es. National Center for Epidemiology, Spain.

[4] khaack@txbiomed.org. Population Health Program, Texas Biomedical Research Institute, TX, USA.

[5] dfallin@jhu.edu. University of Emory, GA, USA.

[6] jgu@georgetown.edu. MedStar Health Research Institute, Washington DC, USA.

[7] lbest@restel.com. Missouri Breaks Industries and Research Inc, SD, USA.

[8] Tianxiao.Huan@umassmed.edu. University of Massachusetts Medical School, MA, USA.

[9] liuc@bu.edu. Boston University School of Public Health, MA, USA.

[10] jiantao.ma@tufts.edu. Tufts University Friedman School of Nutrition Science and Policy, MA

[11] Jose.D.Bermudez@uv.es. Department of Statistics, University of Valencia, Spain.

[12] scole@txbiomed.org. Population Health Program, Texas Biomedical Research Institute, TX, USA.

[13] Dorothy-Rhoades@ouhsc.edu. Stephenson Cancer Center, USA.

[14] levyd@nhlbi.nih.gov. National Heart, Lung, and Blood Institute, USA.

[15] an2737@cumc.columbia.edu. Department of Environmental Health Sciences, Columbia University.

[16] m.tellez@isciii.es. National Center for Epidemiology, Spain.

Abstract

In this study, we investigated the potential mediating role of blood DNA methylation (DNAm) on the association between smoking and lung cancer. We extended the novel *multimediate* algorithm for multiple mediation analysis to survival data, which helped identify joint mediated effects of DNAm in the Strong Heart Study, with validation in the Framingham Heart Study. We additionally conducted functional validation using gene expression data, and bioinformatics analyses to confirm the biological plausibility of the findings.

Keywords: Smoking, DNA methylation, causal mediation analysis.

1. Introduction

Differential patterns in blood DNA methylation (DNAm) are associated with lung cancer, the main cause of cancer death worldwide,[1] suggesting that DNAm changes may play a key role in tumorigenesis.[2] Epigenetic signatures associated with smoking are robust across ethnically diverse populations,[3] and support that DNAm might be a causal intermediary in the biological pathway linking smoking to lung cancer.[4] However, studies investigating the role of DNAm in smoking-related lung cancer are unclear.[5] In most studies of smoking, DNAm and cancer are limited by the lack of time to

incident (i.e. newly diagnosed) cancer or the lack of formal mediation analyses. In addition, DNAm positions tend to be evaluated separately as independent mediators, however, given the high correlations between DNAm sites on the genome, considering them as joint mediators is more appropriate. To date, no algorithms have been developed that are able to conduct multiple mediation analysis with correlated mediators and survival outcomes.

In this study, we investigated whether the association of current and cumulative smoking with lung cancer risk might be explained by differences in human blood DNAm. We used data from the Strong Heart Study (SHS), a cohort of US Native Americans (discovery population), and the Framingham Heart Study (FHS) (replication population). We extended the *multimediate* algorithm,[6] which is able to conduct multiple mediation analysis in presence of correlated mediators, to a time-to-event setting, which enabled the evaluation of the most impactful DMPs potentially driving lung cancer risk. In addition, we conducted validation of the findings using whole blood gene expression in a subset of FHS participants, as well as a bioinformatic pathway enrichment analysis to assess the potential biological implication of the findings.

2. Methods

The SHS is a prospective cohort study of American Indian adults.[7] Blood DNAm was measured at baseline in 2,351 participants using Illumina's MethylationEPIC BeadChip. After preprocessing, 2235 individuals and 788,368 CpGs were included in this study. Lung cancer incidence was assessed by self-report during interviews, death certificates, chart reviews and pathology reports if available.

The FHS started in 1948. DNAm from whole blood was measured using the Illumina Infinium HumanMethylation 450K BeadChip. Gene expression from paired whole blood RNA was sequenced at >×30 depth of coverage using RNA-SeQC v1.1.9. according to TOPMed RNA-Seq pipeline v2. Cancer incidence was assessed by interviews, death certificates, and/or chart reviews that included pathology reports, and crosschecked with official medical records whenever possible.

Statistical methods

We first conducted a screening among the CpG sites that were associated with smoking in previous work in the SHS (303 CpGs in total),[8] by using a Cox ISIS coupled with elastic-net (ISIS-ENET, as conducted by the *SIS* R package), to select CpG sites associated with time to lung cancer. Models were adjusted by age, sex, BMI, study center, cell counts (CD8T, CD4T, NK, B cells and monocytes) and five genetic PCs. We calculated natural direct, indirect and total effects based on the product of coefficients method for mediation analysis using additive hazards models.[9] Mediated effects were reported as differences in cancer cases for current vs never smokers, or differences in cancer cases per a 10 cigarette pack-years increase, attributable to blood DNAm per 100,000 person-years.

We conducted functional validation of the genes identified in the mediation analysis by doing an expression quantitative trait methylation analysis (eQTM). We fitted a linear model for CpGs that were significant in the simple mediation analysis both in the SHS and the FHS. Batch effect-corrected expression was the dependent variable, batch effect-corrected DNAm was the predictor, and the model was adjusted for sex, age, predicted blood cell fraction, five expression PCs and 10 DNAm PCs, which accounted for population. We also conducted a KEGG enrichment analysis out of the genes annotated to cis- and trans- eQTMs to explore possible biological implications of our findings. The Kappa statistic, which is used to define KEGG terms interrelations (edges) and functional groups based on shared genes between terms, was set to 0.4. The enrichment analysis was performed using Cytoscape (version.3.8.2).

In presence of correlated mediators, traditional mediation analysis methods might lead to individual relative mediated effects that add up to more than 100 %, which suggests that some pathways

are overlapping and the joint and individual effects remain unidentifiable. To address this limitation, we extended the *multimediate* algorithm,[6] which uses the counterfactual multiple mediation framework, to the survival data setting using additive hazards models. The R code is available in Github (https://github.com/AllanJe/multimediate). This algorithm[6] is able to identify individual mediated effects of several mediators simultaneously while taking into account correlations between mediators.

Oncogenic transformations can happen several years before cancer diagnosis. Thus, as an attempt to discard cases where DNAm may have been measured after oncogenic transformations started, we repeated the mediation analysis excluding individuals with cancer that was diagnosed in the first 5 follow-up years (10 lung cancer cases excluded).

3. Results

The ISIS model selected 62 Differentially Methylated Positions (DMPs) associated with lung cancer. Of those, 29 DMPs had statistically significant indirect effects in the SHS for current versus never smoking. Among those, 20 were also measured in the FHS, of which 14 were replicated in the FHS. For cumulative smoking, 20 CpGs had statistically significant indirect effects in the SHS. Among those, 14 were also measured in the FHS, of which four were replicated in the FHS. The mediation models excluding cancer cases diagnosed during the first 5 follow-up years yielded similar results.

In the eQTM analysis, at a statistical significance p-value < 10^{-4}, 17 mediating DMPs of lung cancer in common for the SHS and FHS were associated with 12 cis-eQTMs and 2415 trans-eQTMs. The large majority of the eQTM-associated transcripts (75.7 % of transcripts in trans and 83.3 % of transcripts in cis) showed gene expression downregulation. Biological pathway enrichment analysis of target genes annotated to eQTM-associated transcripts showed 54 enriched biological pathways. Figure 1 displays overlapping DMPs, eQTMs and KEGG biological pathways by the evaluated exposures and endpoints. The enriched pathways were largely related to cancer.

In multi-mediator models, in absolute terms, of 385.7 (95% CI 265.9, 509.8) incident lung cancer cases per 100,000 person-years attributable to current smoking, 223.6 (95 % 126.1, 324.5), 62.6 (95 % CI 16.8, 110.2) and 28.3 (95 % CI 11.5, 46.5) lung cancer cases were attributable to differences in DNAm in cg05575921 (*AHRR*), cg24859433 (*IER3*) and cg11902777 (*AHRR*), respectively. This corresponds to 81.3 % of the effect of smoking in lung cancer driven by DNA methylation changes.

4. Discussion

We conducted a formal mediation analysis using time-to-newly diagnosed cancer data, and found that a substantial extent of the prospective association of smoking with lung cancer was explained by differences in blood DNAm. Results were largely consistent in the FHS, including additional validation of findings with expression data, which mostly showed methylation-related downregulation of distant genes that have a plausible role on cancer biological pathways. In the multimediator model, a joint mediated effect of 81.3 % was driven by three DMPs (annotated to *AHRR* and *IER3*).

Of note, our novel *multimediate* algorithm enabled us to explore the joint mediated effects of DMPs. Although many DMPs showed individual mediated effects in the single mediation analysis, the *multimediate* algorithm identified that the mediated effect was only driven by three DMPs. This means that many DMPs were identified as mediators by the single mediation analysis just because of having high correlations with actual mediators, but when considering them jointly in the same model, their contribution to the mediated effect was not significant. This fact highlights the importance of considering a multiple mediation approach as opposed to a simple mediation one.

This study has several limitations. First, although the replication in the FHS was high for lung cancer in the current versus never smoking model, it was smaller for lung cancer in the cumulative smoking model. Differences in smoking intensity and cessation across the SHS and FHS could explain some of the non-replicated DMPs. Also, non-fatal cancer data might be incomplete in the SHS as no linkage with the cancer registry is available. Despite these limitations, however, we still found substantial replication of findings between the SHS and the FHS.

Second, mediation analysis provides valid estimates only if the mediation assumptions such as absence of unmeasured confounding, which cannot be fully verified in practice, hold.[10] In addition, the *multimediate* algorithm is only valid in settings of non-causal correlations.[6] Experimental studies are needed to confirm the role of the identified blood DNAm signature of smoking in the association between smoking and smoking-related cancers.

Strengths of our study include replication in an independent cohort, the large sample size with methylation data from one of the largest microarrays nowadays available, the availability of information to account for numerous potential confounders and the additional validation of the results using gene expression data. In addition, we used state-of-the-art statistical methods including the novel *multimediate* algorithm for time-to-event data, which enabled the evaluation of correlated methylation sites jointly.

In conclusion, the prospective association of smoking with lung cancer in this study was largely explained by differences in few specific blood DNAm. These findings contribute to the identification of potentially novel mechanisms of lung cancer, and provide evidence in favor of DNAm as a potential biological intermediary in the association between smoking and lung cancer.

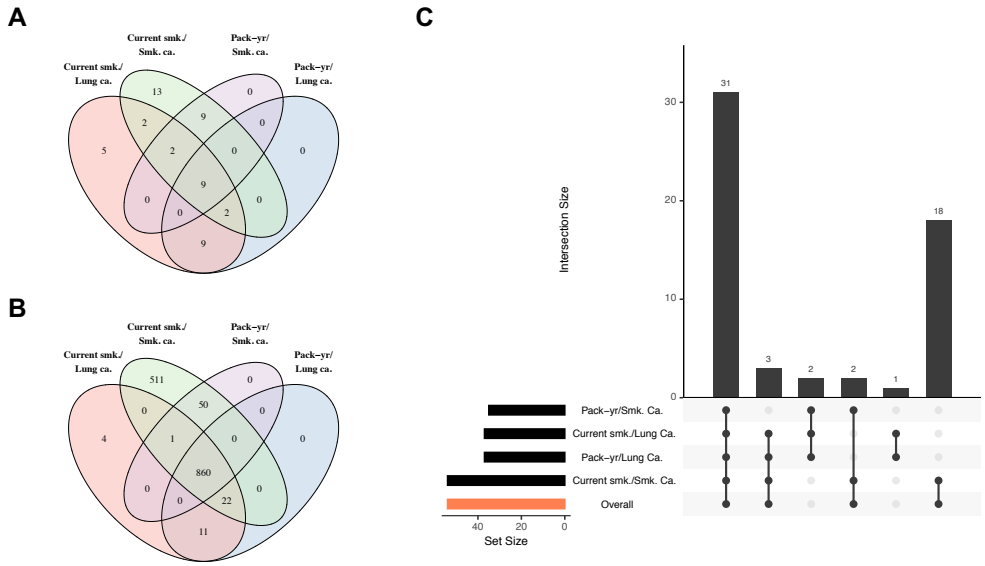

Figure 1: A) Venn diagram of CpGs with significant mediated effects both in the SHS and FHS. B) Venn diagram of genes annotated to the differentially expressed transcripts in trans in the Framingham Heart Study. C) Upset plot of the overlapping enriched KEGG pathways.

5. Bibliography

1. Bjaanæs MM, Fleischer T, Halvorsen AR, et al. Genome-wide DNA methylation analyses in lung adenocarcinomas: Association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol Oncol*. 2016;10(2):330-343.
2. Klutstein M, Nejman D, Greenfield R, Cedar H. DNA Methylation in Cancer and Aging. *Cancer Res*. 2016;76(12):3446-3450.
3. Joehanes R, Just AC, Marioni RE, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet*. 2016;9(5):436-447.
4. Fasanelli F, Baglietto L, Ponzi E, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun*. 2015;6(1):10192.
5. Herceg Z, Ambatipudi S. Smoking-associated DNA methylation changes: No smoke without fire. *Epigenomics*. 2019;11(10):1117-1119.
6. Jérolon A, Baglietto L, Birmelé E, Alarcon F, Perduca V. Causal mediation analysis in presence of multiple mediators uncausally related. *Int J Biostat*. October 2020.
7. Lee ET, Welty TK, Fabsitz R, et al. The Strong Heart Study. A study of cardiovascular disease in American Indians: design and methods. *Am J Epidemiol*. 1990;132(6):1141-1155.
8. Domingo-Relloso A, Riffo-Campos AL, Haack K, et al. Cadmium, Smoking, and Human Blood DNA Methylation Profiles in Adults from the Strong Heart Study. *Environ Health Perspect*. 2020;128(6):067005.
9. Lange T, Hansen J V. Direct and Indirect Effects in a Survival Context. *Epidemiology*. 2011;22(4):575-581.
10. Zhang Z, Zheng C, Kim C, Van Poucke S, Lin S, Lan P. Causal mediation analysis in the context of clinical research. *Ann Transl Med*. 2016;4(21):425.