**Supplemental Materials for**

**Blood RNA and Protein Biomarkers Are Associated with Vaping and Dual use and Prospective Health Outcomes**

Andrew Gregory1, Zhonghui Xu1, Katherine Pratte2, Seth Berman1, Robin Lu1, Rahul Suryadevara1, Robert Chase1, Jeong H. Yun1,3, Aabida Saferali1, Craig P. Hersh1,3, Edwin K. Silverman1,3, Russell P. Bowler4, Laura E. Crotty Alexander5,6, Adel Boueiz1,3\*, Peter J. Castaldi1,7\* for the COPDGene investigators.

1Channing Division of Network Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA; 2Department of Biostatistics, National Jewish Health, Denver, CO; 3Division of Pulmonary and Critical Care Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA; 4Division of Pulmonary, Critical Care and Sleep Medicine, National Jewish Health, Denver, CO; 5Division of Pulmonary, Critical Care, Physiology and Sleep Medicine, University of California San Diego (UCSD), La Jolla, CA; 6Pulmonary Critical Care Section, Veterans Affairs (VA) San Diego Healthcare System, La Jolla, CA; 7Division of General Medicine and Primary Care, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA

\* Equal contribution

**Corresponding Author:** Peter J. Castaldi, MD, MSc,Channing Division of Network Medicine, Brigham and Women’s Hospital, 181 Longwood Avenue, Boston, MA 02115, Email: peter.castaldi@channing.harvard.edu

**Table of Contents**

1. **Supplemental Methods**
2. **Supplemental Tables**
	1. **ST 1. Whole blood RNA differential expression analysis for vapers versus former cigarette users.** (see file Table\_S1toS3\_RNAseqResults.xlsx)
	2. **ST 2. Whole blood RNA differential expression analysis for dual users versus former cigarette users.** (see file Table\_S1toS3\_RNAseqResults.xlsx)
	3. **ST 3. Whole blood RNA differential expression analysis for current versus former cigarette users.** (see file Table\_S1toS3\_RNAseqResults.xlsx)
	4. **S4. Differential expression analysis for all vaper group contrasts for genes significantly associated only to vaping in the primary analyses.** (see file Table\_S4toS5\_VapingandDualUse\_RNAcontrasts.xlsx)
	5. **S5. Differential expression analysis for all dual user group contrasts for genes significantly associated only to dual use in the primary analyses.** (see file Table\_S4toS5\_VapingandDualUse\_RNAcontrasts.xlsx)
	6. **ST6. Gene set enrichment analysis for the vapers vs former cigarette users from RNA-seq differential expression analysis.** (see file Table\_S6toS8\_RNA\_GSEA.xlsx)
	7. **ST7. Gene set enrichment analysis for the dual users vs former cigarette users from RNA-seq differential expression analysis.** (see file Table\_S6toS8\_RNA\_GSEA.xlsx)
	8. **ST8. Gene set enrichment analysis for the current vs former cigarette users from RNA-seq differential expression analysis.** (see file Table\_S6toS8\_RNA\_GSEA.xlsx)
	9. **S9. Protein differential expression results by vaping/smoking group.** (see file TableS9\_protein\_DE\_results.xlsx)
	10. **S10. Differential expression analysis for all vaping contrasts for proteins significantly associated only to vaping in the primary analyses.** (see file Table\_S11toS12\_VapingandDualUse\_Proteincontrasts.xlsx)
	11. **S11. Differential expression analysis for all dual use contrasts for proteins significantly associated only to dual in the primary analyses.** (see file Table\_S11toS12\_VapingandDualUse\_Proteincontrasts.xlsx)
	12. **S12. Pairwise comparisons of vaping and smoking groups using both RNA-seq and proteomic biomarkers.** (see file Table\_S12\_LDA\_Vaping\_AllPairwise\_noSE.docx)
	13. **S13. RNA Biomarker Associations to 5-year Change in FEV1.** (see file TableS13to16\_OutcomeAssociations.xlsx)
	14. **S14. Protein Biomarker Associations to 5-year Change in FEV1.** (see file TableS13to16\_OutcomeAssociations.xlsx)
	15. **S15. RNA Biomarker Associations to Time to Event Outcomes.** (see file TableS13to16\_OutcomeAssociations.xlsx)
	16. **S16. Protein Biomarker Associations to Time to Event Outcomes.** (see file TableS13to16\_OutcomeAssociations.xlsx)
	17. **S17. Previously reported RNA and protein biomarker associations to vaping or dual use.** (see file TableS17\_PriorBiomarkers.xlsx)
3. **Supplemental Figures**
	1. **SF1. Study overview.**
	2. **SF2. Flow diagram for study subjects.**
	3. **SF3. GSEA enrichment plots for Hallmark Inflammation Pathway**

**Supplemental Methods:**

***Definitions of exacerbations, incident comorbidities, and mortality assessment***

Respiratory exacerbations were defined as the acute worsening of respiratory symptoms that required systemic antibiotics and/or steroids as determined by self-report. Respiratory disease-related health impairments and quality of life were assessed using the St George’s Respiratory Questionnaire (SGRQ). Incident self-reported cardiovascular disease (CVD) was defined as a composite endpoint of stroke, heart attack, coronary artery disease, coronary artery bypass graft surgery, peripheral artery disease, or cardiac angina. All-cause mortality was determined through longitudinal follow-up and by referencing the social security death index. A systematic adjudication process based on methods outlined in the Towards a Revolution in COPD Health (TORCH) trial was used to categorize cause-specific mortality[[15]](https://paperpile.com/c/BPlaru/rPfiE).

***RNA-sequencing***

The extraction protocol for total blood RNA was performed either manually or with the Qiagen QIAcube extraction robot according to the company’s standard operating procedure. RNA samples with an RNA integrity number (RIN) > 6 and a concentration of ≥ 25 μg/μl were sequenced. Library preparation was performed with the Illumina TruSeq Stranded Total RNA with Ribo-Zero Globin kit (Illumina, Inc., San Diego, CA). Samples were sequenced to an average depth of 20 million 75 base pairs (bp) paired end reads on Ilumina HiSeq 2500 sequencers. Reads were aligned to the GRCh38 genome using the STAR (version 2.5.2b) aligner[(1)](https://paperpile.com/c/jsfZKy/AYowf). Gene annotations and transcript GTF were downloaded from the Biomart Ensembl database (Ensembl Genes release 94, GRCh38.p12 assembly). Quality control was performed using the FastQC and RNA-SeQC programs, and samples were included for analysis if they met the following criteria: > 10 million total reads, > 80% of reads mapped to the reference genome, XIST and Y chromosome expression was consistent with reported sex, < 10% of R1 reads in the sense orientation, Pearson correlation ≥ 0.9 with samples in the same library construction batch, and concordant genotype calls between variants called from RNA sequencing reads and DNA genotyping. Sequencing read counts were obtained from the featureCounts function in the Rsubread R package (v1.32.2). Genes were filtered to remove very low expressed genes; we limited our analysis to include only those genes which had an average counts per million (CPM) value > 1 in more than 20 subjects. The trimmed mean of M values (TMM) procedure from the edgeR R package (v3.24.3) was applied to account for the differences in sequencing depth[(2)](https://paperpile.com/c/jsfZKy/TmAsS).

***Gene set enrichment analysis***

Enrichment analyses for RNA transcripts were performed using the preranked gene set enrichment analysis method reimplemented in the FGSEA R package (v1.16.0)(3), using as gene sets the Hallmark (h.all.v7.5.1) and Immunologic signature (c7.immunesigdb.v7.5.1) pathways from the Molecular Signatures Database. Genes from each differential expression analysis were ranked by their t statistics and the enrichment scores were computed from unweighted ranks by setting the gseaParam argument to 0. The minimal and maximal size of the tested pathways were 15 and 500, respectively. The number of permutations was set to 10,000. The threshold for significant pathways from the FGSEA analyses was an adjusted (Benjamini-Hochberg method) P-value < 0.1.

***Prediction modeling for vaping status and prospective health outcomes***

Predictive models using only transcript and protein biomarkers were constructed for vaping/smoking groups and prospective health outcomes. For the vaping/smoking group prediction models we used linear discriminant analysis (LDA) machine learning for the multiclass classification into the four vaping/smoking groups using nested cross-validation. To avoid bias induced by feature selection, all genes were considered and feature selection was performed within each fold of the inner CV loop. For prediction of prospective binary outcomes, we predicted occurrence of the outcome (binary 1/0) during the follow-up period using LDA using only the transcripts or proteins that were significantly associated with vaping or dual use. For prediction of change in FEV1, the same biomarker predictors were used in an elastic net model. Model performance was assessed by cross-validation using Youden index, area under the receiver operator characteristic (AUROC), and area under the precision recall curve (AUCPR) for binary outcomes. R2 was used for change in FEV1 models. Prediction models were not made for respiratory mortality due to the low number of observed events.

**Supplemental Figures**

**SF 1. Study overview.** The objectives of the present study were to identify blood transcriptomic and proteomic biomarkers of vaping, relate them to prospective health outcomes, and investigate their ability to accurately distinguish vaping and smoking status and to predict prospective health events.



**SF2. Study flow diagram.**

****

**SF3. GSEA enrichment plots for Hallmark Inflammation Pathway.** Gene set enrichment plots for the Hallmark Inflammation pathway in the RNA-seq analyses comparing vapers (left panel), current cigarette users (middle panel), and dual users (right panel) to former cigarette users as the common control group.



**References**

1. [Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;29:15–21.](http://paperpile.com/b/jsfZKy/AYowf)

2. [Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–140.](http://paperpile.com/b/jsfZKy/TmAsS)

3. Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, Alexey Sergushichev. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. biorxiv. 2021. doi:10.1101/060012