Blood-Transcriptomics Based Machine Learning Prediction of Emphysema in Smokers

Rahul Suryadevara[1], Robin Lu[1], Zhonghui Xu[1], Andrew Gregory[1], Aria Masoomi[2], Seth Berman[1], Jeong H. Yun[1,3], Aabida Saferali[1], Craig P. Hersh[1,3], Edwin K. Silverman[1,3], Jennifer Dy[2], Peter J. Castaldi[1,4], Adel Boueiz[1,3], for the COPDGene Investigators.

[1]Channing Division of Network Medicine, BWH; [2]Department of Electrical and Computer Engineering, Northeastern University;
[3]Pulmonary and Critical Care Medicine, BWH; [4]Division of General Medicine and Primary Care, BWH | Boston, MA.

STUDY OBJECTIVES

- Identify whole-blood genes and proteins associated with CT-quantified emphysema.
- Develop multi-omic biomarker risk score of emphysema using machine learning.

METHODS

Clinical, Transcriptomic, and Proteomic Data Collection

- 3,386 current/former smokers from the COPDGene Study, a longitudinal study investigating the genetic factors underlying COPD.
- 19,177 blood RNA transcripts generated from RNA Sequencing.
- 4,979 plasma proteins quantified using SomaLogic SomaScan.
- Data was log transformed and split into 70:30 train and test datasets.

Differential Gene Expression and Protein Association Analyses

- *Phenotype*: Adjusted Perc15 density (Hounsfield units at the 15th percentile of CT density histogram at total lung capacity, corrected for the inspiratory depth)
- *Covariates*: Age, sex, race, current smoking, pack-years of smoking, FEV_1, CBC (WBC + Platelets), library batch (transcriptomic) or clinical center (proteomic), scanner model
- Multiple testing adjustments: False discovery rate (FDR) 10%
- *Methods*: Voom/limma (genes); linear regression (proteins); Gene Ontology (GO) enrichment analysis

Prediction Analysis

- *Method:* Elastic net
- *Predictors*: Genes and proteins that reached statistical significance in the association analysis, separately or in combination with candidate CBC predictors (neutrophils, lymphocytes, monocytes, eosinophils, platelets).
- All predictors were scaled and their importance scores were defined by the absolute values of their coefficients in the regression models.
- *Outcome*: Emphysema (Adjusted Perc15 density)
- We classified subjects into tertiles of emphysema severity and assessed the accuracy of the models in predicting those at highest and lowest risk tertiles for emphysema in the testing sample by evaluating the areas under the receiver-operator-characteristic-curves (AUROC).

CONCLUSION

- This study highlighted novel candidate genes and pathways which may provide greater insight into emphysema pathophysiology.
- Emphysema transcriptomic and proteomic risk scores could help identify those who have low probability of emphysema (eliminate radiation) or those at high risk for emphysema (enrich clinical trials).

FUTURE DIRECTION

Investigate transcriptomic signatures from lung tissue data and study alternative splicing mechanisms of emphysema.

FUNDING

NHLBI K08HL141601, K01HL157613, K08HL146972, R01HL125583, R01HL147326, R01 HL147326, U01 HL089897, and U01 HL089856.

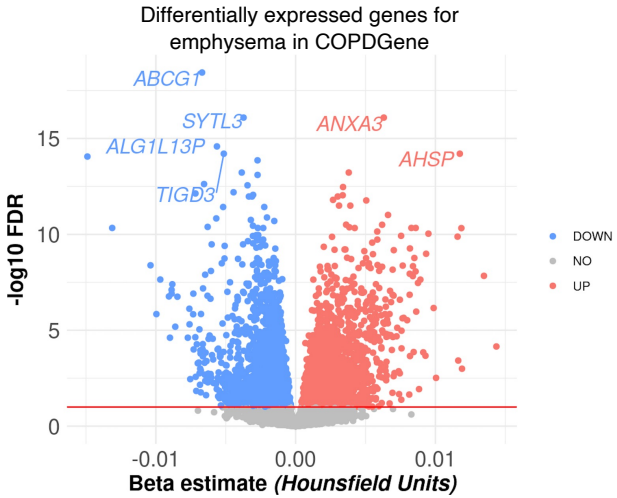
RESULTS

Baseline Characteristics

Median [IQR] and Count (Proportions)

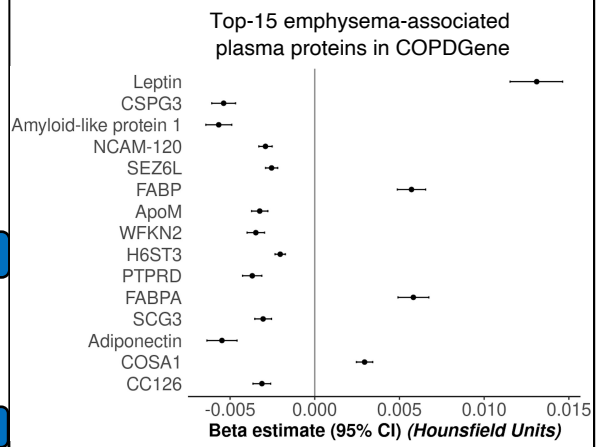
	Train (n = 2,370)	Test (n = 1,016)	P-value
Age	64.6 [57.9;71.6]	65.2 [58.3;71.8]	0.32
Sex, % male	1217 (51.35%)	506 (49.8%)	0.41
Race, % NHW	1727 (72.87%)	774 (76.18%)	0.04
BMI	28.1 [24.6;32.3]	27.9 [24.6;31.9]	0.50
Smoking pack-years	38.8 [25;53.7]	38.8 [23.8;52.5]	0.38
Current smoking	858 (36.2%)	346 (34.06%)	0.50
FEV1, % predicted	84.1 [65.7;97.7]	84.1 [65.6;97.5]	0.90
Adjusted Perc15 density	86 [72.8;101]	85.7 [71.6;101]	0.67
% Segmental airway wall thickness	49 [43.9;55.1]	48.7 [43.5;54.9]	0.67
GOLD (2, 3, 4)	718 (30.29%)	718 (31.2%)	0.89

Differential Gene Expression and Gene Ontology Enrichment Analyses



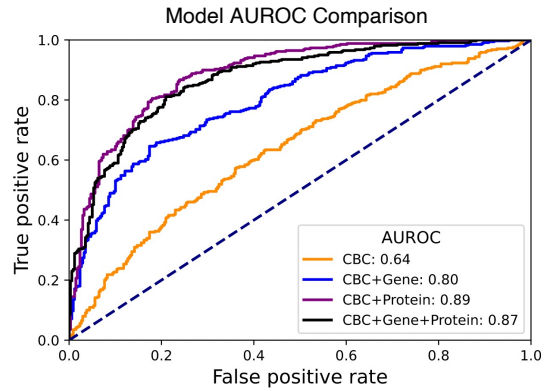
- 4,913 (26%) genes were significantly associated with emphysema (*FDR 10%).*
- 2,574 (52%) genes were upregulated (log fold change > 0).
- 2,339 (48%) genes were downregulated (log fold change < 0).
- 44 GO pathways were significantly enriched with relevance to emphysema (*P-value < 0.005*). Selected top-5 significant GO pathways were:
  - Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
  - Viral transcription
  - Neutrophil degranulation
  - Positive regulation of NF-kappaB transcription factor activity
  - Regulation of tumor necrosis factor-mediated signaling pathway

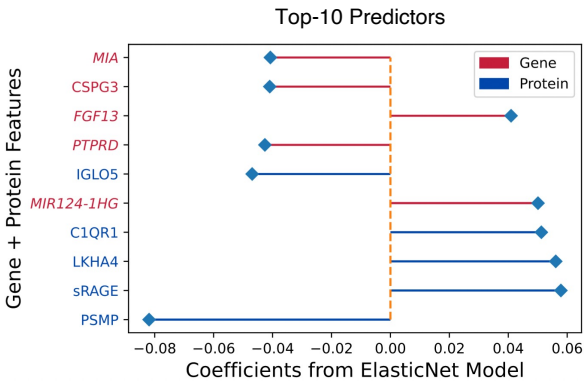
Protein Association Analysis



- 882 (18%) proteins were significantly associated with emphysema (*FDR 10%).*

Elastic Net Prediction





- Subject data types were sequentially layered to assess model performance with multi-omic data.
- There was significant (*P-value < 0.05*) incremental improvement in the prediction performance for determining emphysema severity by adding:
  - Gene to CBC
  - Protein to CBC
  - Gene + Protein to CBC

COPDGene® COPD Genetic Epidemiology