**Molecular Determinants of Lung Function Decline: A Multi-Level Analysis of Gene Expression**

Zaid W. Elhusseini, Omar Rafique, Min Hyung Ryu, Peter Castaldi, Don D. Sin, Ingo Ruczinski, Craig P. Hersh

**Supplemental methods**

**COPDGene Study**

COPDGene[1] (clinicaltrial.gov identifier: NCT00608764) is a longitudinal observational study that enrolled participants aged 40–80 years with a smoking history of at least 10 pack-years. The cohort includes individuals of non-Hispanic White or African-American ancestry, both with and without chronic obstructive pulmonary disease (COPD), and excludes other known lung diseases aside from asthma. Demographic data, smoking status, medical history, post-bronchodilator spirometry, and chest CT scans were collected at each visit. The Phase 2 and 3 visits added complete blood counts and RNA-sequencing (RNA-Seq) samples. 3,819 participants in Phase 2 had both transcriptomic and phenotypic data. Of these, 3,802 had clinical data in Phase 1 and Phase 2, 2,043 had clinical data in Phase 2 and Phase 3, and 2,035 had clinical data in Phase 1 and Phase 3. A total of 437 participants had transcriptomic and clinical data in both Phase 2 and Phase 3. Ultimately, 435 participants were retained for analysis after excluding 2 individuals who each formed a unique batch, as ComBat-seq [2] requires at least two samples per batch for effective batch correction. Details of the COPDgene study were previously published[1,3]. All participants provided written informed consent, and the study was approved by the relevant institutional review boards.

**Gene Expression Filtering and Normalization**

We performed four main analyses in R (version 4.3.2). First, we restricted our analyses to subjects with both gene expression (Phase 2 or Phase 3) and phenotypic data, and we combined these datasets. For each dataset, we calculated counts per million (CPM) values for every gene, and we retained only those genes for which at least 80% of the samples had a CPM greater than or equal to 1 to ensure that lowly expressed genes were excluded. Next, we used the edgeR:TMM[4]method to normalize the retained genes, accounting for library size differences among samples. Finally, we repeated this filtering and normalization procedure in each relevant dataset (e.g., Phase 2, Phase 3, and the combined set) to ensure consistency across analyses and facilitate direct comparisons between phases.

**Gene Expression Analysis**
We constructed a design matrix for each analysis, which included the relevant FEV1 measure (cross-sectional or change in FEV1), demographic factors (age, race, gender, smoking status), and cell-type proportions. Next, we fit linear models using the edgeR:voomLmFit function -- a combination of the voom and lmFit functions in edgeR[5] -- to estimate mean-variance trends followed by eBayes to identify significantly associated genes. To annotate Ensembl gene IDs, we used EnsDb.Hsapiens.v79 and get matching gene names with Ensembl gene IDs. Finally, we reported the top differentially expressed genes (p < 0.05).

**Gene signature:**

To generate the gene signature, we identified genes associated (pvalue<0.05) with FEV1 change across three intervals (Phase 1–Phase 2, Phase 1–Phase 3, and Phase 2–Phase 3) and classified them into two sets based on the direction of log fold-change (logFC), resulting in 17 positive (GS\_Pos) and 3 negative (GS\_Neg) gene signatures. We then performed Gene Set Variation Analysis (GSVA)[6] to assign each subject a score based on these gene sets. Finally, we assessed associations with various traits using linear regression for continuous traits or logistic regression for categorical traits, adjusting for age, race, gender, smoking status, and cell-type proportions, and compiled the significant findings into a single results table.

**ECLIPSE Study Replication**

The Evaluation of COPD Longitudinally to Identify Predictive Surrogate Endpoints (ECLIPSE) (ClinicalTrials.gov identifier: NCT00292552) was a multicenter, three-year prospective study that enrolled 2,747 participants [7]. Whole-blood RNA microarray data and tested traits were available for 646 of these individuals. Participants were aged 40–75 years, had at least a 10 pack-year smoking history, no other known respiratory disease, and moderate to severe COPD (Global Initiative for Chronic Obstructive Lung Disease [GOLD] stage 2 or higher). The latter was defined as a post-bronchodilator forced expiratory volume in one second (FEV1) below 80% of predicted and an FEV1/forced vital capacity (FVC) ratio less than 0.7. Details of the ECLIPSE study were previously published [3,7]. All participants provided written informed consent, and the study was approved by the relevant institutional review boards.

To replicate our findings in ECLIPSE, we used the GSVA package [6]to compute each subject a score based on the previously selected gene sets. We then applied either linear regression or logistic regression, depending on the trait type. ECLIPSE has 2 sub-cohorts and 9 batches across samples.Each model controlled for relevant covariates (cohort, microarray batch, age, race, sex, smoking status, and cell-type proportions).

**Smoking status**

Smoking status was determined by a standardized questionnaire. Both COPDGene and ECLIPSE cohorts obtained the current smoking status by asking the subject whether they were smoking as of 1 month before the study visit when blood samples and clinical data were collected.

**Supplemental Results**

**Cross-sectional Analysis**

We performed a series of linear regression analyses, including cross-sectional analyses, FEV1 change analysis, and longitudinal analysis. In the cross-sectional analyses, the associations between gene expression and FEV1 were analyzed separately at Phases 2 and 3. We identified 740 genes in Phase 2 (Supplemental Table 2) and 225 genes in Phase 3 (Supplemental Table 3) that were associated with FEV1 at p < 0.05. Notably, only 11 genes overlapped between the two phases. This limited overlap may reflect variations in gene expression associated with COPD progression or other underlying biological changes occurring between phases. These findings suggest that cross-sectional analyses at different time points might capture distinct aspects of the disease process.

**References**

1 Regan EA, Hokanson JE, Murphy JR, *et al.* Genetic epidemiology of COPD (COPDGene) study design. *COPD*. 2010;7:32–43. doi: 10.3109/15412550903499522

2 Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom Bioinform*. 2020;2. doi: 10.1093/NARGAB/LQAA078

3 Ryu MH, Yun JH, Morrow JD, *et al.* Blood Gene Expression and Immune Cell Subtypes Associated with Chronic Obstructive Pulmonary Disease Exacerbations. *Am J Respir Crit Care Med*. 2023;208:247–55. doi: 10.1164/RCCM.202301-0085OC/SUPPL\_FILE/DISCLOSURES.PDF

4 Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11:1–9. doi: 10.1186/GB-2010-11-3-R25/FIGURES/3

5 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40. doi: 10.1093/BIOINFORMATICS/BTP616

6 Hänzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*. 2013;14:1–15. doi: 10.1186/1471-2105-14-7/FIGURES/7

7 Vestbo J, Anderson W, Coxson HO, *et al.* Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur Respir J*. 2008;31:869–73. doi: 10.1183/09031936.00111707