Whole exome sequencing analysis in severe chronic obstructive pulmonary disease

Dandi Qiao[1*], Asher Ameli[1,3], Dmitry Prokopenko[1], Han Chen[4,5], Alvin T. Kho[6], Margaret M. Parker[1], Jarrett Morrow[1], Brian D. Hobbs[1,2], Yanhong Liu[7], Terri H. Beaty[8], James D. Crapo[9], Kathleen C. Barnes[10], Deborah A. Nickerson[11], Michael Bamshad[12], Craig P Hersh[1,2], David A. Lomas[13], Alvar Agusti[14], Barry J. Make[9], Peter M.A. Calverley[15], Claudio F. Donner[16], Emiel F. Wouters[17], Jørgen Vestbo[18], Peter D. Paré[19], Robert D. Levy[19], Stephen I. Rennard[20, 21], Ruth Tal-Singer[22], Margaret R. Spitz[7], Amitabh Sharma[1], Ingo Ruczinski[23], Christoph Lange[24], Edwin K. Silverman[1,2], Michael H. Cho[1,2*]; NHLBI Exome Sequencing Project, University of Washington Center for Mendelian Genomics, Lung GO, COPDGene Investigators

[1] Channing Division of Network Medicine and [2]Division of Pulmonary and Critical Care Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, United States of America

[3] Department of Physics, Northeastern University, Boston, Massachusetts 02115, United States of America

[4] Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas 77030, United States of America

[5] Center for Precision Health, School of Public Health and School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas 77030, United States of America

[6] Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts 02115, United States of America

[7] Dan L. Duncan Comprehensive Cancer Center, Department of Medicine, Baylor College of Medicine, Houston, Texas 77030, United States of America

[8] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland 21205, United States of America

[9] National Jewish Health, Denver, Colorado 80206, United States of America

[10] Division of Allergy and Clinical Immunology, Department of Medicine, Johns Hopkins University, Baltimore, Maryland 21224, United States of America

[11] Department of Genome Sciences, University of Washington, Seattle, Washington 98195, United States of America

[12] Division of Genetic Medicine, Department of Pediatrics, University of Washington and Seattle Children's Hospital, Seattle, Washington 98195, United States of America

[13] University College London, London WC1E 6BT, United Kingdom

[14] Respiratory Institute, Hospital Clinic, IDIBAPS, University of Barcelona, CIBERES, Barcelona 08007, Spain

[15] University of Liverpool, Liverpool L69 3BX, United Kingdom

[16] Mondo Medico di I.F.I.M. srl, Multidisciplinary and Rehabilitation Outpatient Clinic, Borgomanero, Novara 28021, Italy

[17] Department of Respiratory Medicine, Maastricht University Medical Center, 6202 AZ Maastricht, Netherland

[18] University of Manchester, Manchester M13 9PL, United Kingdom

[19] Respiratory Division, Department of Medicine, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada

[20] University of Nebraska Medical Center, Omaha, Nebraska 68198, United States of America

[21] AstraZeneca, Cambridge CB2 0RE, United Kingdom

[22] GSK Research and Development, King Of Prussia, Pennsylvania 19426, United States of America

[23] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, United States of America

[24] Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, United States of America

*** Corresponding author**

Address: Channing Division of Network Medicine, 181 Longwood Avenue, Boston, MA 02115

Fax: 617 525 0958

Phone: 617 525 2113

Email: redaq@channing.harvard.edu

ABSTRACT

Chronic obstructive pulmonary disease (COPD), one of the leading causes of death worldwide, is substantially influenced by genetic factors. Alpha-1 antitrypsin deficiency demonstrates that rare coding variants of large effect can influence COPD susceptibility. To identify additional rare coding variants in patients with severe COPD, we conducted whole exome sequencing analysis in 2,543 subjects from two family-based studies (Boston Early-Onset COPD Study and International COPD Genetics Network) and one case-control study (COPDGene). Applying a gene-based segregation test in the family-based data, we identified significant segregation of rare loss of function variants in *TBC1D10A* and *RFPL1* ($P < 2x10\text{-}6$),

but were unable to find similar variants in the case-control study. In single variant, gene-based, and pathway association analyses, we were unable to find significant findings that replicated or were significant in meta-analysis. However, we found that the top results in the two datasets were in proximity to each other in the protein-protein interaction network (p=0.014), suggesting enrichment of these results for similar biological processes. A network of these association results and their neighbors was significantly enriched in the transforming growth factor beta-receptor binding and cilia-related pathways. Finally, in a more detailed examination of candidate genes, we identified individuals with putative high-risk variants, including patients harboring homozygous mutations in genes associated with cutis laxa and Niemann-Pick Disease Type C. Our results likely reflect heterogeneity of genetic risk for COPD along with limitations of statistical power and functional annotation, and highlight the potential of network analysis to gain insight into genetic association studies.

Introduction

Chronic obstructive pulmonary disease (COPD) is a heterogeneous and complex disease, with a significant genetic component to its susceptibility(1). Using genome-wide association analysis (GWAS), a number of COPD susceptibility loci have been identified, including *FAM13A (2)*, *HHIP (3, 4)*, and *CHRNA3/CHRNA5/IREB2 (5-7)*. Complementary studies have identified more than a hundred loci associated with lung function, many of which likely also affect risk of COPD(8, 9). However, identified loci only explain 5-10% of the heritability of COPD or quantitative measures of lung function traits (8, 10). GWAS effectively tests common variants, but the well-known examples of alpha-1 antitrypsin deficiency(11), cutis laxa (12-16), and the more recently described association between telomere-related genes (17-19) indicate that, as has been shown for other diseases(20-24), rare coding variants also contribute to COPD risk.

We previously analyzed exome sequencing data of 49 families with severe, early-onset COPD, and, although we found several candidate genes, none showed convincing evidence of replication (25, 26). We

further showed using simulations that genetic heterogeneity may be a major contributor to this failure to replicate(26). In this study, we applied additional sequencing and analytic strategies to increase the sample size and the power of the analysis. We applied a recently developed family-based method, GEne-based SEgregation (GESE) (25), to a larger family-based dataset enriched for severe COPD, and also performed single-variant, as well as set-based tests using SKAT-O for both genes and pathways in the family-based and in an additional case-control study. We tested for enrichment of our results in gene expression and monogenic models of disease, and examined the overlap between case-control and family-based results using network analysis. Finally, we investigated a set of candidate genes identified in previous genetic studies, including Mendelian syndromes, for potentially deleterious rare variants.

Results

Gene-based segregation test on the ICGN and Boston EOCOPD pedigrees

Baseline characteristics of the studied subjects are shown in **Table 1**. Additional information on the probands can be found in **Supplementary Table S1**. To identify causal variants in our exome sequencing data with the characteristics of Mendelian variants for COPD (e.g. alpha-1 antitrypsin deficiency), we applied our recently described gene-based segregation (GESE) test (25) to the family-based data. We focused on ultra-rare (MAF < 0.1%) predicted loss-of-function variants. Two genes were significant after Bonferroni correction for the total of 18268 genes: *RFPL1* (p = 1.60e-06) and *TBC1D10A* (p=1.10e-06). *RFPL1* segregated in 4 families, including two singleton families and two families with affected sibling pairs of severe COPD. *TBC1D10A* segregated in a parent-offspring pair and a singleton family. *TBC1D10A* is intolerant to loss-of-function variants (ExAC intolerance probability = 0.98 (27)). The top 10 genes from this analysis are shown in **Table 2**. All 10 of these genes are expressed in the adult lung (see Methods, enrichment p = 0.17), and the expression of 5 out of 9 of those genes was associated with forced expiratory volume in 1 second (FEV_1) % predicted, a measure of COPD severity, in our lung tissue data (enrichment p = 0.024). We further sought supportive evidence for association of these genes in the COPDGene case-control dataset. However, no subjects harbored loss of function variants in these genes. We additionally

tested for evidence of higher burden of rare (MAF < 0.1%), non-synonymous variants in the cases, and did not find convincing evidence of association (*RFPL1* p=0.576; *TBC1D10A* p=0.081).

Single-variant association analysis in the case-control and family data

Next, we performed single variant association analysis. We tested both rare coding variants (moderate effect by SNPEff, and MAF < 5%) as well as all variants. We found no significant results (**Supplementary Table S2** and **S3**) in either our primary analysis using COPDGene as the discovery cohort (using a Bonferroni significance level of 1.32e-06 for non-synonymous variants with MAF < 5%, and 5.07e-07 for all variants), or using the family-based data (3.55e-07 for non-synonymous variants with MAF < 5% and significance level 1.86e-07 for all variants). However, top variants in the case-control analysis included rs8040868 (MAF = 0.41) and rs1051730 (MAF = 0.35) in *CHRNA3*(28) with p = 5.05e-05 and 7.39e-05 respectively, which reside at a previously described GWAS locus (**Supplementary Table S2**). Top variants in the family-based analysis included rs2232710 (MAF = 0.012; p = 4.05e-05) in *SERPINA10* (in high D' with the alpha-1 Z allele, which causes alpha-1 antitrypsin deficiency – note that severe alpha-1 antitrypsin deficiency, including ZZ homozygosity, was an exclusion criteria for these studies) and rs10507051 (MAF = 0.063; p = 1.28e-04) in *VEZT* (**Supplementary Table S3**), near a locus associated with COPD in a recent GWAS of lung function(8). We also considered whether any variants were significant in meta-analysis by combining results from the two studies (case-control status in the COPDGene data, and lung function in the BEOCOPD-ICGN data) using the Stouffer method. Meta-analysis did not identify significant variants among the rare coding variants (significance level 2.82e-06) or among all variants (significance level 7.08e-07) (**Supplementary Table S4, Supplementary Table S5**); top results overall included variants in *CHRNA3* and *SERPINA10*.

Gene-and pathway-based analyses in case-control and family data

Next, we performed gene-based analyses. In the analysis using SKAT-O and predicted deleterious variants with a MAF < 1%, we found no significant genes in the COPDGene data. The top 10 genes are shown in **Supplementary Table S6**. We found no significant enrichment of genes expressed in lung (enrichment p =

0.97) among the top 10 genes. However, four genes have expression associated with FEV_1 % predicted (p = 0.087), including the top two genes *VNN1* and *PLA1A*. *EGFL8,* the 3rd ranked gene in the list, is located near the *AGER* locus which was previously associated with risk of COPD (8, 9). In the pathway analysis using the KEGG(29) database, we found one significant pathway using the burden test, the Jak-STAT signaling pathway (p = 6.78e-05)(30). However, association with this pathway was not replicated in the family-based analysis (p = 0.54) using the burden test. Top results from family-based analyses can be found in **Supplementary Table S7**. We also conducted a meta-analysis of the COPDGene dataset and the BEOCOPD-ICGN dataset; however, no gene achieved significance (**Supplementary Table S8**).

Enrichment and network-based approach to overlap

Given our lack of significant associations using standard association tests, we sought evidence that our top case-control and family-based results were enriched for associations with COPD. We tested for enrichment of overlap of genes yielding nominal significance (i.e. $p < 0.01$) between the case-control and the family-based association results using a standard hypergeometric approach. The enrichment p-value was 1, which was consistent with our lack of overlap and meta-analysis findings.

While we did not observe overlap between the top results in the case-control analysis and the family-based analysis using a simple hypergeometric test, we were interested in studying common biological pathways shared by the two sets of top genes. Recently, network-based methods have demonstrated the ability to identify related diseases in the protein-protein interactome (31). We hypothesized that application of this method to two independent association results for COPD would a) identify whether there were overlapping association signals, despite the lack of replication; and b) identify genes or pathways of highest priority. We computed the network-based separation (31) defined as the normalized average shortest path between members from the two modules to see whether top genes from the case-control and family-based analysis were close to each other in the protein-protein interaction (PPI) network. For the analysis of rare and deleterious variants, we found genes with $p < 0.01$ from the case-control analysis and the family-based analysis had significantly overlapping neighborhoods with negative separation score (score = -2.29, p = 0.014). To explore the neighborhood of these genes and the common pathways that connect the top genes,

we added the first neighbors of the top genes in the PPI. These genes (top genes, along with all of their first-degree neighbors - a total of 522 genes) formed a largest connected component (LCC) of 513 genes, which means almost all the top genes and their first neighbors were connected. **Figure 1** shows the network module containing the largest connected component formed by the top genes from the two analyses and their first neighbors. There were 19 genes with $p < 0.01$ in the family-based data, which had 274 first-degree neighbors in the LCC network; there were 14 genes with $p < 0.01$ in the COPDGene data, which had 216 first-degree neighbors. Between the two groups of 274 and 216 first-degree neighbors, 10 overlapped, thus these genes together formed a network module of 513 genes. 14 genes at loci previously associated with COPD or lung function (out of 329 genes in the curated set, see Methods) were in this set (enrichment $p = 0.065$, Supplementary Material). Additional examination of these genes in murine models showed that the 513 genes were significantly enriched for genes associated with the respiratory system (enrichment $p = 0.045$) and were enriched for genes involved in normal murine lung development in three common inbred strains of mice (enrichment p = 1.35e-02, 1.96e-03 and 2.40e-03, respectively) (See Methods). From this result, we postulate that there is a large disease network module exists likely including a subset of these 513 genes for severe COPD, and only part of this disease module was observed using either analysis alone due to limited power. However, since the two sections of disease module share similar function and pathways, they were significantly close to each other in the PPI network.

To further explore the functions of this network, we also looked at the pathways enriched for these 513 genes using ToppFun in the ToppGene Suite(32), and found a large number of Gene-ontology pathways were significantly enriched. To examine more specific pathways, we examined GO pathways with fewer than 100 genes in total. The top two pathways meeting these criteria were "GO:0005160: transforming growth factor beta (TGFB) receptor binding" and "GO:0030991: intraciliary transport particle A". Fourteen out of fifty three genes in the TGFB receptor binding pathway were present in the network module. Multiple lines of evidence, including genetic association (which has identified *TGFB2*) and other genomic and mechanistic studies have implicated this pathway in risk to COPD (33-35). Twelve of these fourteen genes are expressed in human lung tissue (two genes have missing data). The right panel in **Figure 1** shows the

small network formed by genes in this pathway and *ACVR2B*, which was the top-ranked gene from this set of genes in the association analysis and was also the second largest hub in the network. For the "intraciliary transport particle A pathway" (36) , seven out of eight genes were in the network, which are shown in the left panel in **Figure 1**. Six out of seven genes were expressed in human lung tissue (one gene had missing data). *TULP3* was in the top ranked genes from the family-based analysis and was the largest hub in the module. *TULP3* is a known target of the Hedgehog pathway. Notably, genome-wide association studies and follow-up functional studies have identified an important role for *HHIP* in the development of COPD(7); *TULP3* has been shown to change expression after *HHIP* silencing (37). Also, *WDR35* and *IFT140* were associated with respiratory system abnormalities in mouse models (*WDR35* leads to lung hypoplasia and mutations in *IFT140* produces severely misshapen lungs). Additional top results from this GO analysis can be found in **Supplementary Table S9**. Thus, our network results highlight *ACVR2B* and *TULP3,* which may be prioritized for further examination of functional rare variants.

Evidence of association for candidate genes

A substantial proportion of rare variants identified for complex disease are located at loci that also harbor common risk variants (38, 39). In addition, several Mendelian syndromes have COPD, emphysema, or obstructive lung disease as a manifestation of disease. Therefore, in addition to looking at exome-wide results, we examined a list of the 329 curated genes (see Methods, **Supplementary Table S10**) (1, 7-9, 12-19, 40-45). This included regions identified from 105 SNPs from GWAS analyses(8, 9) and 29 Mendelian genes with manifestations that include COPD or emphysema in their resulting syndromes (**Supplementary Material**). We examined functional and rare variants with MAF < 5%, and found multiple genes to be nominally associated with COPD status or FEV_1 % predicted value, including *CHRNA5, AGER*, and *CYP2A6* (**Supplementary Table S11**). To identify whether there was any independent evidence of rare variant effects at these loci in the COPDGene cohort, we conditioned on the risk allele for the 104 SNPs identified by GWAS. Several genes were still nominally significant after conditioning on the GWAS SNPs, including *CYP2A6* (full results shown in **Table 3**); whether these rare variants have independent effects on COPD susceptibility at these loci will likely need to be addressed by additional, larger studies.

We also looked closely at the 29 genes causing Mendelian syndromes including emphysema or obstructive lung disease as part of their syndrome. To determine whether there was enrichment in these genes in our dataset, we performed a burden test including only variants with MAF < 0.1% in ExAC and predicted deleterious by FATHMM, SIFT, and CADD (>15). We found that the burden-based tests gave a p = 0.80 in the COPDGene case-control study, and a p = 0.018 for the family-based EOCOPD and ICGN data. Thus, we observed some significant accumulation of deleterious variants in these genes in the family-based data, suggesting that ultra-rare variants in these Mendelian genes contributing to lung function may be related to severe COPD risk in our family-based datasets.

To examine these variants individually, we intersected variants in these genes with Clinvar, using an annotation of significance level 4 (likely pathogenic) and above, and additionally included variants in published reports associated with respiratory disease in *TERT* (17-19, 41-43). We found 47 of these variants in our datasets. These variants are listed in **Supplementary Table S12** along with their counts among cases and controls separately. Given the strong evidence of pathogenicity for variants in *SERPINA1* and telomere-related genes, these findings are shown in **Table 4**. We also assessed the carriers of these rare variants using a recessive model of inheritance, and those variants with homozygous genotypes present in any dataset are listed in **Table 5**. Among our findings for Mendelian genes were two previously identified cases from COPDGene with heterozygous *TERT* variants (19), and evidence for an increased burden (cases > controls) for the *SERPINA1* Z and PI P (Lowell) (rs121912714) (**Table 4**). For recessive variants, we identified rs140130028, a splice-donor variant in *NPC2*, which is a gene for Niemann-Pick disease type C2, a disease previously associated with emphysema (46) (**Table 5**). One pair of sibs with severe COPD in the ICGN study was homozygous for this variant; two of their half-siblings carried one copy, one with severe COPD. None of these subjects had known Neimann-Pick disease. Also, variant rs61748181 in *TERT* was present as homozygous in 7 unrelated cases in the datasets (**Table 5**). While this association did not reach candidate-wide significance (p = 0.167), this variant was experimentally demonstrated to induce telomere dysfunction (47) and predicted to be disease causing by Mutation Taster (48). For variants not annotated by Clinvar or annotated with a significance level of 3 (uncertain significance) or below, we filtered based on

MAF < 0.1% in ExAC v0.3 non-Finnish Europeans and predicted deleterious effects by FATHMM, SIFT, or CADD (>15). There were in total 346 such variants in our datasets. One of these variants occurred in homozygous form in a proband with severe COPD in the BEOCOPD study. This variant is an ultra-rare splice-acceptor variant in *ATP6V0A2* (novel in ExAC database) (**Table 4**), a Mendelian gene for cutis laxa. A chest CT scan of this subject showed severe emphysema, however, no phenotypic information related to dermatological characteristics was available. In addition, 66 variants were predicted to be deleterious by all three annotations: FATHMM, SIFT and CADD (>15), and had supportive evidence in our datasets (with greater counts in cases than in controls, **Supplementary Table S13**). Multiple variants have supportive evidence in both case-control and family-based datasets. For example, rs141310608 in *EFEMP2* is present in 2 cases in COPDGene study and 2 cases in BEOCOPD-ICGN study, while none in controls. Also, there are multiple ultra-rare variants in *COL3A1* are carried by cases and none by controls. We have also listed the variants that are predicted to be deleterious by all annotations, but are present in more controls than cases; these variants are less likely to be high penetrance COPD susceptibility variants (**Supplementary Table S14**). We also applied a more liberal filtering criteria (MAF < 0.05, CADD > 10 or predicted to be deleterious by SIFT and FATHMM) for *TERT, RTEL1*, *CFTR* and *SERPINA1*. Detailed information about these genes can be found in Supplementary **Tables S15, S16** and **S17** respectively.

Discussion

COPD is a common and heterogeneous disease; under the common-disease-common-variants hypothesis, we expect that multiple common variants should contribute to a large proportion of COPD risk. However, even though a number of COPD GWAS loci have been discovered through large-scale collaborative efforts, most of the estimated heritability remains unexplained. Examples such as alpha-1 antitrypsin deficiency, cutis laxa, and more recently, telomeropathies are associated with COPD and emphysema (17-19). These results motivated us to search the entire exome for large effect variants that could represent a Mendelian subtype of COPD, in the hope of finding new treatment strategies for a subset of the patients. In this study, we examined multiple cohorts representing the largest exome sequencing study of COPD to date ascertained under an extreme phenotype approach (where samples were enriched for sever COPD and normal controls

heavily exposed to smoking but with normal pulmonary function), to screen through the entire exome to identify rare coding variants controlling risk to COPD. Results failed to identify new genes, pathways, or variants consistently significant across all of our analyses, suggesting that single-variant or single-gene effects of a contribution as large as alpha-1 antitrypsin deficiency are unlikely to exist (26). Yet, a network-based analysis identified a significant relationship between the two modules formed by the top results of the two analyses. These two sets of top genes, along with their first neighbors in the protein-protein interaction network, form a well-connected network component. This largest connected component was significantly enriched in genes involved in fetal lung development in mouse models (49). Additionally, this module sheds light on related functions or pathways where such rare variants may be contributing to risk to COPD. For example, multiple studies have suggested the transforming growth factor beta pathway is associated with COPD(50), and the *TGFB2* locus was associated with COPD in genome-wide association studies (GWAS) (51). Our study identified *ACVR2B* as a potential candidate; of interest, *ACVR1B*, an activin receptor which interacts with *ACVR2B* (52) was identified in a network-informed genetic association study of COPD (53) and in a integrative analysis of emphysema distribution (54).

Our finding lends further support to the transforming growth factor beta pathway and also suggests that rare variants related to *ACVR2B* may contribute to COPD risk. Similarly, the identification of *TULP3* lends further support to the identification of *HHIP* as a causal gene at this GWAS locus and the importance of the hedgehog pathway in the development of COPD. The identification of cilia-related pathways is intriguing given the importance of cilia to lung function (55), including reports from a smaller exome study of resistant smokers (56) and reports of shortened cilia in smokers and in COPD patients (36).

Finally, we identified subjects carrying homozygous genotypes of rare and deleterious variants in Mendelian genes for cutis laxa and Niemann-Pick disease, which are themselves intriguing candidates for causing severe COPD. These findings illustrate the potential relevance of using filtering-based technique for identifying syndromic forms of COPD. While we do not have enough power to individually test these or other individual rare variants here, our results may provide support for future studies in these recognized candidate genes.

COPD is known to be a highly heterogeneous disease, with varying contributions of emphysema and small airway disease. We did not examine specific subsets of COPD, as detailed phenotyping was not available in all cohorts. Multiple analysis methods are available for rare variant analysis ((57)), and the optimal methods are still not clear. Our sequencing of a large number of affected individuals in families was appropriate for methods such as GESE, which leverages a large reference dataset (ExAC); an alternative approach using association would require large scale exome harmonization of controls with normal lung function, preferably with heavy cigarette smoke exposure. Our results highlight the importance of integration with other types of data (e.g., gene expression, protein-protein interaction) to better understand the results from one data type. However, our analysis does not attempt to identify the confidence of individual genes in this network; we cannot rule out the possibility that this network includes many genes that are false positives, and our pathway analysis should be considered descriptive and exploratory. Additional investigation, including genetic studies, integration of multi-omics data, and careful functional studies will be needed to further infer biological mechanisms and potential disease causality for our identified genes.

In summary, in an exome sequencing study of COPD, we were unable to identify exome-wide significant associations, but through network analysis we identified candidate genes in related pathways and a disease module driven by rare variants. Our study is consistent with a potential contribution of multiple, heterogeneous rare variants in COPD, and demonstrates the insight that network-based methods can offer.

Materials and Methods

The COPDGene study

The COPDGene study is a multi-center epidemiologic and genetic study of 10,192 current or ex-smokers, which has been previously described (58). COPDGene subjects were sequenced in two sets. The first set sequenced as part of the NHLBI Exome Sequencing Project (ESP; named COPDGene ESP) included severe COPD cases with GOLD (Global Initiative for Chronic Obstructive Lung Disease) Grade 3 or 4 (post-bronchodilator FEV1 < 50% predicted and FEV1/FVC < 0.70), and age < 65 years old, with substantial emphysema (> 15% at -950 HU) by quantitative chest CT scan. Controls were selected to be resistant

smokers with frequency-matched pack-years of cigarette smoking, normal lung function (FEV_1 > 80% predicted and FEV_1/FVC > 70%), age > 65 years old, and no significant emphysema (< 5% at -950 HU). The second set sequenced at Baylor (named COPDGene Baylor) included severe COPD cases (GOLD Grade 3 or 4) with no age requirement. Controls were selected to be resistant smokers with normal lung function with age > 55.

The Boston Early-Onset COPD study and the International COPD Genetics Network study

The family-based data contained samples selected from the Boston Early-Onset COPD study (BEOCOPD) (59) and the International COPD Genetics Network (ICGN) study(45). Probands from BEOCOPD were selected to be physician-diagnosed COPD cases with $FEV_1 \leq 40\%$ predicted, age ≤ 53. All first-degree relatives, older second-degree relatives and additional affected family members were enrolled. Probands in the ICGN study were subjects with known COPD and were required to have FEV_1 < 60% predicted, FEV1/FVC < 90% predicted at age 45-65, pack-years ≥ 5, and have at least one eligible sibling. An initial set of 49 pedigrees selected from the Boston Early-Onset COPD Study were described and analyzed previously (26). To this sample we added 147 families from BEOCOPD and 462 families from the ICGN study. The COPDGene, BECOPD, and ICGN studies all excluded subjects with severe alpha-1 antitrypsin deficiency.

Exome sequencing

We sequenced all subjects using Nimblegen capture and Illumina platforms. The COPDGene ESP, BEOCOPD, and ICGN subjects were all sequenced at the University of Washington, using Nimblegen V2 exome capture; COPDGene Baylor samples used VChrome capture. Alignment, variant calling, and quality control were performed using bwa, GATK, and in-house pipelines respectively. As COPDGene ESP and COPDGene Baylor used slightly different capture platforms, calling was performed on these datasets separately. All BEOCOPD and ICGN subjects were called together (joint calling) and went through the same quality control steps together to provide the final family-based data (named BEOCOPD-ICGN) for

analysis. Baseline characteristics of the subjects in each of the cleaned datasets are shown in **Table 1** and our overall study design is shown in **Figure 2**. More details can be found in the **Supplementary Material.**

Analysis strategy

Loss of function variants using the Gene-based Segregation test (GESE)

We first performed the gene-based segregation test (GESE)(25) on loss of function variants (defined by SnpEff (60)) with MAF < 0.1% in the family-based BEOCOPD-ICGN data using COPD affection status as the outcome. We included only the most severe COPD subjects (GOLD spirometry grade 3 or 4) and resistant smoking control subjects (normal spirometry, age > 40 yr, with at least 5 pack-years of cigarette smoking). This analysis took advantage of the unique properties of a family-based strategy, including having multiple copies of rare variants, and assumes a Mendelian model with a few rare variants with very large effects. We sought supportive evidence for identified causal genes in COPDGene dataset by attempting to identify similarly deleterious variants.

Association analyses

Second, we performed single variant, gene-based, and pathway-based association analyses. For all association analyses, we used Bonferroni correction based on the number of genes, pathways, or variants tested. For the COPDGene case-control data, COPD affection status was used as the outcome, which was adjusted for pack-years, gender, age, and ancestry-based principal components (PCs) in the COPDGene Baylor data, and the top PCs alone in the COPDGene ESP data due to the selection criteria, and as performed previously (Supplementary Material). For the family-based data, due to the low number of controls with normal lung function, but a wider range of FEV_1 available through family members, we analyzed FEV_1 (forced expiratory volume in one second), a lung function measure highly correlated with COPD (9) instead of COPD affection status itself. The outcome in the family-based association tests was the rank of the residuals from regressing raw post-bronchodilator FEV_1 value on height, pack-years, sex, age, top 5 genetic ancestry PCs and batch indicator variable.

Single-variant association analysis

For single-variant analyses, we applied the Stouffer method to meta-analyze the results from the hybrid method in SKATBinary_Single function (SKAT package) in the COPDGene case-control data, since the two cohorts selected from the COPDGene study were sequenced and called separately. The hybrid method in SKATBinary_Single function selects the most appropriate approach to compute p-values for each variant. For single variant analysis in family-based data, we applied the variant-based generalized linear mixed model association test (GMMAT(61)). In addition to using COPDGene as discovery and BEOCOPD-ICGN as replication, we also examined using BEOCOPD-ICGN and both datasets as discovery by meta-analyzing the results from the COPDGene case-control data and the BEOCOPD-ICGN data using the Stouffer method. For single variant analyses, we tested all variants, and also the subset with moderate effect with MAF < 5%.

Gene- and pathway-based association analysis

For both the gene-based and pathway-based analyses, we applied SKAT-O tests. In the COPDGene case-control datasets, we applied the hybrid method in the SKATBinary function, implemented in the SKAT package to each of the datasets, and meta-analyzed the two datasets using Fisher's method. For the BEOCOPD-ICGN family-based data, we applied MONSTER(62), which is a generalized version of SKAT-O for family-based studies. We also meta-analyzed all results (case-control and family-based results) using Fisher's method. Our primary gene- and pathway-based association analyses focused on deleterious variants defined using FATHMM (57, 63) with MAF < 1% in the association analysis. In one study of amyotrophic lateral sclerosis (ALS), FATHMM was found to give the best power to identify known causal genes for ALS in gene-based association tests(57). Our secondary analyses included association testing on functional variants with moderate effects (defined by SnpEff (60)) with MAF < 5%. This is a less stringent filtering criterion on the variants to prevent missing signals in this set of variants. Pathways were defined using KEGG pathways(29) and the c2 collection of curated gene sets from the Molecular Signatures Database (MsIGdb) in GSEA(64).

Identification of enrichment in gene expression

To help determine whether the identified genes were relevant for our phenotypes, we used publicly available FPKM (per kilobase of gene model per million mapped reads) results from gene expression data from the Lung Genomics Research Consortium (LGRC, http://www.lung-genomics.org) to identify whether any gene was expressed in the lung (using a cutoff of 0.5(26)). We also used the results of differential expression for lung function and COPD case-control status in an independent set of lung tissue from severe COPD subjects and controls (65). In addition, enrichment for genes associated with respiratory system in mouse was carried out using a curated set of genes associated with respiratory phenotype in the Mouse Genome Database (http://www.informatics.jax.org/marker) (66). Gene expression information in human and normal murine lung development for three common inbred strains of mice were obtained from the GEO dataset (GSE14334 and GSE74243), and genes involved in fetal lung development were obtained using methods described in (49).

Network-based analysis

Finally, we applied the network-based separation measure defined in (31) to examine how closely connected the top genes from the two independent analyses are in the protein-protein interaction (PPI) network. This measure has been shown to predict pathobiological similarity of two sets of disease genes(31). In our application here, since the two outcomes analyzed for the COPDGene and BEOCOPD-ICGN dataset are highly correlated, genes that are causal for these outcomes should have much shorter network-based distance. Therefore, a significant result tells us that at least a subset of the top genes from the two analyses are topologically overlapping and exert some effect on risk of COPD.

Examination of previously identified genetic associations with COPD

To examine loci previously described to be associated with risk of COPD or lung function itself in GWAS or harboring Mendelian variants related to COPD, we curated a set of 329 genes for closer examination (**Supplementary Table S12**)(8, 9). At COPD GWAS loci, we identified all variants in a European reference population with an $r^2 > 0.8$ with the lead variant, and then expanded these borders by 100kb. For Mendelian

syndromes, we included connective tissue disorders such as cutis laxa (12-16), as well as telomere-related genes including *TERT, TERC, RTEL1,* and *NAF1 (17, 18, 41-43)*. We looked for supportive evidence of association for these genes using several methods. First, we examined the association results in both primary and secondary analyses as described above. Since 104 of the previously described lead SNPs based on GWAS of lung function or COPD were also available for the COPDGene subjects, we additionally performed conditional analyses for these genes by conditioning on the GWAS SNPs in proximity in an attempt to identify independent rare variants contributing to COPD susceptibility. For both the marginal association analyses and conditional analyses, COPD affection status was the outcome in the COPDGene case-control analyses and FEV_1 was the outcome in the family-based analyses. Finally, we examined Mendelian genes for evidence of pathogenic variants using Clinvar and other public annotation resources.

Acknowledgements

This work was supported by NHLBI R01 HL089856 (EKS), R01 HL089897 (JDC), R01 HL113264 (MHC and EKS), P01 HL105339 (EKS) and PO1 114501 (EKS), K01 HL129039 (DQ), K07 CA181480 (YL). The COPDGene project (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, GlaxoSmithKline, Novartis, Pfizer, Siemens and Sunovion. The ICGN study was funded by GlaxoSmithKline. Sequencing for the Boston Early-Onset COPD Study was provided by the University of Washington Center for Mendelian Genomics (UW CMG) and was funded by the National Human Genome Research Institute and the National Heart, Lung and Blood Institute grant 1U54HG006493.

Authors would like acknowledge Victor M Pinto-Plata (Baystate Medical Center, Springfield, MA), Nathaniel Marchetti (Temple University School of Medicine, Philadelphia, PA), Raphael Bueno(Brigham and Women's Hospital and Harvard Medical School, Boston, MA), Bartolome R Celli (Brigham and Women's Hospital and Harvard Medical School, Boston, MA), Gerald J Criner (Temple University School of Medicine, Philadelphia, PA), and Dawn Demeo (Brigham and Women's Hospital and Harvard Medical

School, Boston, MA) for providing the lung tissue samples and their support of the project. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health,

Conflict of Interest Statement

Dr. Silverman has received honoraria and consulting fees from Merck, grant support and consulting fees from GlaxoSmithKline, and honoraria from Novartis. Dr. Hersh has been a consultant for CSL Behring and Mylan. Dr. Cho has received grant support from GSK.

References

1 Hersh, C.P., Demeo, D.L. and Silverman, E.K. (2005) Silverman EK, S.S., Lomas DA, Weiss ST, editors (ed.), In *Respiratory genetics*. Hodder Arnold, New York, in press.

2 Cho, M.H., Boutaoui, N., Klanderman, B.J., Sylvia, J.S., Ziniti, J.P., Hersh, C.P., DeMeo, D.L., Hunninghake, G.M., Litonjua, A.A., Sparrow, D. *et al.* (2010) Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat. Genet.*, **42**, 200-202.

3 Zhou, X., Baron, R.M., Hardin, M., Cho, M.H., Zielinski, J., Hawrylkiewicz, I., Sliwinski, P., Hersh, C.P., Mancini, J.D., Lu, K. *et al.* (2012) Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Hum. Mol. Genet.*, **21**, 1325-1335.

4 Wilk, J.B., Chen, T.H., Gottlieb, D.J., Walter, R.E., Nagle, M.W., Brandler, B.J., Myers, R.H., Borecki, I.B., Silverman, E.K., Weiss, S.T. *et al.* (2009) A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet.*, **5**, e1000429.

5 Hardin, M., Zielinski, J., Wan, E.S., Hersh, C.P., Castaldi, P.J., Schwinder, E., Hawrylkiewicz, I., Sliwinski, P., Cho, M.H. and Silverman, E.K. (2012) CHRNA3/5, IREB2, and ADCY2 are associated with severe chronic obstructive pulmonary disease in Poland. *Am. J. Resp. Cell. Mol.*, **47**, 203-208.

6 DeMeo, D.L., Mariani, T., Bhattacharya, S., Srisuma, S., Lange, C., Litonjua, A., Bueno, R., Pillai, S.G., Lomas, D.A., Sparrow, D. *et al.* (2009) Integration of genomic and genetic approaches implicates IREB2 as a COPD susceptibility gene. *Am. J. Hum. Genet.*, **85**, 493-502.

7 Pillai, S.G., Ge, D., Zhu, G., Kong, X., Shianna, K.V., Need, A.C., Feng, S., Hersh, C.P., Bakke, P., Gulsvik, A. *et al.* (2009) A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.*, **5**, e1000421.

8 Wain, L.V., Shrine, N., Artigas, M.S., Erzurumluoglu, A.M., Noyvert, B., Bossini-Castillo, L., Obeidat, M., Henry, A.P., Portelli, M.A., Hall, R.J. *et al.* (2017) Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat. Genet.*, **49**, 416-425.
9 Hobbs, B.D., de Jong, K., Lamontagne, M., Bosse, Y., Shrine, N., Artigas, M.S., Wain, L.V., Hall, I.P., Jackson, V.E., Wyss, A.B. *et al.* (2017) Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat. Genet.*, **49**, 426-432.
10 Zhou, J.J., Cho, M.H., Castaldi, P.J., Hersh, C.P., Silverman, E.K. and Laird, N.M. (2013) Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am. J. Respir. Crit. Care Med.*, **188**, 941-947.
11 Laurell, C.B. and Eriksson, S. (2013) The electrophoretic alpha1-globulin pattern of serum in alpha1-antitrypsin deficiency. 1963. *COPD*, **10 Suppl 1**, 3-8.
12 Urban, Z., Gao, J., Pope, F.M. and Davis, E.C. (2005) Autosomal dominant cutis laxa with severe lung disease: synthesis and matrix deposition of mutant tropoelastin. *J. Invest. Dermatol.*, **124**, 1193-1199.
13 Hucthagowder, V., Sausgruber, N., Kim, K.H., Angle, B., Marmorstein, L.Y. and Urban, Z. (2006) Fibulin-4: a novel gene for an autosomal recessive cutis laxa syndrome. *Am. J. Hum. Genet.*, **78**, 1075-1080.
14 Hucthagowder, V., Morava, E., Kornak, U., Lefeber, D.J., Fischer, B., Dimopoulou, A., Aldinger, A., Choi, J., Davis, E.C., Abuelo, D.N. *et al.* (2009) Loss-of-function mutations in ATP6V0A2 impair vesicular trafficking, tropoelastin secretion and cell survival. *Hum. Mol. Genet.*, **18**, 2149-2165.
15 Hu, Q., Loeys, B.L., Coucke, P.J., De Paepe, A., Mecham, R.P., Choi, J., Davis, E.C. and Urban, Z. (2006) Fibulin-5 mutations: mechanisms of impaired elastic fiber formation in recessive cutis laxa. *Hum. Mol. Genet.*, **15**, 3379-3386.
16 Callewaert, B., Su, C.T., Van Damme, T., Vlummens, P., Malfait, F., Vanakker, O., Schulz, B., Mac Neal, M., Davis, E.C., Lee, J.G. *et al.* (2013) Comprehensive clinical and molecular analysis of 12 families with type 1 recessive cutis laxa. *Hum. Mutat.*, **34**, 111-121.
17 Stanley, S.E., Merck, S.J. and Armanios, M. (2016) Telomerase and the Genetics of Emphysema Susceptibility. Implications for Pathogenesis Paradigms and Patient Care. *Ann. Am. Thorac. Soc.*, **13**, S447-S451.
18 Stanley, S.E., Gable, D.L., Wagner, C.L., Carlile, T.M., Hanumanthu, V.S., Podlevsky, J.D., Khalil, S.E., DeZern, A.E., Rojas-Duran, M.F., Applegate, C.D. *et al.* (2016) Loss-of-function mutations in the RNA biogenesis factor NAF1 predispose to pulmonary fibrosis-emphysema. *Sci. Transl. Med.*, **8**, 351ra107.
19 Stanley, S.E., Chen, J.J., Podlevsky, J.D., Alder, J.K., Hansel, N.N., Mathias, R.A., Qi, X., Rafaels, N.M., Wise, R.A., Silverman, E.K. *et al.* (2015) Telomerase mutations in smokers with severe emphysema. *J. Clin. Invest.*, **125**, 563-570.
20 Do, R., Stitziel, N.O., Won, H.H., Jorgensen, A.B., Duga, S., Angelica Merlini, P., Kiezun, A., Farrall, M., Goel, A., Zuk, O. *et al.* (2015) Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature*, **518**, 102-106.
21 Ellinghaus, D., Zhang, H., Zeissig, S., Lipinski, S., Till, A., Jiang, T., Stade, B., Bromberg, Y., Ellinghaus, E., Keller, A. *et al.* (2013) Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology*, **145**, 339-347.

22 Yu, T.W., Chahrour, M.H., Coulter, M.E., Jiralerspong, S., Okamura-Ikeda, K., Ataman, B., Schmitz-Abe, K., Harmin, D.A., Adli, M., Malik, A.N. *et al.* (2013) Using whole-exome sequencing to identify inherited causes of autism. *Neuron*, **77**, 259-273.
23 Gonzaga-Jauregui, C., Harel, T., Gambin, T., Kousi, M., Griffin, L.B., Francescatto, L., Ozes, B., Karaca, E., Jhangiani, S.N., Bainbridge, M.N. *et al.* (2015) Exome Sequence Analysis Suggests that Genetic Burden Contributes to Phenotypic Variability and Complex Neuropathy. *Cell. Rep.*, **12**, 1169-1183.
24 Lange, L.A., Hu, Y., Zhang, H., Xue, C., Schmidt, E.M., Tang, Z.Z., Bizon, C., Lange, E.M., Smith, J.D., Turner, E.H. *et al.* (2014) Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.*, **94**, 233-245.
25 Qiao, D., Lange, C., Crapo, J.D., Beaty, T.H., Laird, N.M., Won, S., Silverman, E.K. and Cho, M.H. (2016), In *American Thoracic Society International Conference 2016*, San Francisco, in press.
26 Qiao, D., Lange, C., Beaty, T.H., Crapo, J.D., Barnes, K.C., Bamshad, M., Hersh, C.P., Morrow, J., Pinto-Plata, V.M., Marchetti, N. *et al.* (2016) Exome Sequencing Analysis in Severe, Early-Onset Chronic Obstructive Pulmonary Disease. *Am. J. Resp. Crit. Care. Med.*, **193**, 1353-1363.
27 Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, **536**, 285-291.
28 Matsson, H., Soderhall, C., Einarsdottir, E., Lamontagne, M., Gudmundsson, S., Backman, H., Lindberg, A., Ronmark, E., Kere, J., Sin, D. *et al.* (2016) Targeted high-throughput sequencing of candidate genes for chronic obstructive pulmonary disease. *BMC Pulm. Med.*, **16**, 146.
29 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353-D361.
30 Yew-Booth, L., Birrell, M.A., Lau, M.S., Baker, K., Jones, V., Kilty, I. and Belvisi, M.G. (2015) JAK-STAT pathway activation in COPD. *Eur. Respir. J.*, **46**, 843-845.
31 Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J. and Barabasi, A.L. (2015) Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.
32 Chen, J., Xu, H., Aronow, B.J. and Jegga, A.G. (2007) Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics*, **8**, 392.
33 Zandvoort, A., Postma, D.S., Jonker, M.R., Noordhoek, J.A., Vos, J.T., van der Geld, Y.M. and Timens, W. (2006) Altered expression of the Smad signalling pathway: implications for COPD pathogenesis. *Eur. Respir. J.*, **28**, 533-541.
34 Baraldo, S., Bazzan, E., Turato, G., Calabrese, F., Beghe, B., Papi, A., Maestrelli, P., Fabbri, L.M., Zuin, R. and Saetta, M. (2005) Decreased expression of TGF-beta type II receptor in bronchial glands of smokers with COPD. *Thorax*, **60**, 998-1002.
35 Ezzie, M.E., Crawford, M., Cho, J.H., Orellana, R., Zhang, S., Gelinas, R., Batte, K., Yu, L., Nuovo, G., Galas, D. *et al.* (2012) Gene expression networks in COPD: microRNA and mRNA regulation. *Thorax*, **67**, 122-131.
36 Hessel, J., Heldrich, J., Fuller, J., Staudt, M.R., Radisch, S., Hollmann, C., Harvey, B.G., Kaner, R.J., Salit, J., Yee-Levin, J. *et al.* (2014) Intraflagellar transport gene expression associated with short cilia in smoking and COPD. *PLoS One*, **9**, e85453.
37 Zhou, X., Qiu, W., Sathirapongsasuti, J.F., Cho, M.H., Mancini, J.D., Lao, T., Thibault, D.M., Litonjua, A.A., Bakke, P.S., Gulsvik, A. *et al.* (2013) Gene expression analysis uncovers novel

hedgehog interacting protein (HHIP) effects in human bronchial epithelial cells. *Genomics*, **101**, 263-272.
38 Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitziel, N.O., Brody, J.A., Khetarpal, S.A., Crosby, J.R., Fornage, M. *et al.* (2014) Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.*, **94**, 223-232.
39 Beaudoin, M., Goyette, P., Boucher, G., Lo, K.S., Rivas, M.A., Stevens, C., Alikashani, A., Ladouceur, M., Ellinghaus, D., Torkvist, L. *et al.* (2013) Deep resequencing of GWAS loci identifies rare variants in CARD9, IL23R and RNF186 that are associated with ulcerative colitis. *PLoS Genet.*, **9**, e1003723.
40 Stuart, B.D., Choi, J., Zaidi, S., Xing, C., Holohan, B., Chen, R., Choi, M., Dharwadkar, P., Torres, F., Girod, C.E. *et al.* (2015) Exome sequencing links mutations in PARN and RTEL1 with familial pulmonary fibrosis and telomere shortening. *Nat. Genet.*, **47**, 512-517.
41 Diaz de Leon, A., Cronkhite, J.T., Katzenstein, A.L., Godwin, J.D., Raghu, G., Glazer, C.S., Rosenblatt, R.L., Girod, C.E., Garrity, E.R., Xing, C. *et al.* (2010) Telomere lengths, pulmonary fibrosis and telomerase (TERT) mutations. *PLoS One*, **5**, e10680.
42 Petrovski, S., Todd, J.L., Durheim, M.T., Wang, Q., Chien, J.W., Kelly, F.L., Frankel, C., Mebane, C.M., Ren, Z., Bridgers, J. *et al.* (2017) An Exome Sequencing Study to Assess the Role of Rare Genetic Variation in Pulmonary Fibrosis. *Am. J. Respir. Crit. Care. Med.*, **196**, 82-93.
43 Tsakiri, K.D., Cronkhite, J.T., Kuan, P.J., Xing, C., Raghu, G., Weissler, J.C., Rosenblatt, R.L., Shay, J.W. and Garcia, C.K. (2007) Adult-onset pulmonary fibrosis caused by mutations in telomerase. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 7552-7557.
44 Bertuch, A.A. (2016) The molecular genetics of the telomere biology disorders. *RNA Biology*, **13**, 696-706.
45 Zhu, G., Warren, L., Aponte, J., Gulsvik, A., Bakke, P., Anderson, W.H., Lomas, D.A., Silverman, E.K., Pillai, S.G. and International, C.G.N.I. (2007) The SERPINE2 gene is associated with chronic obstructive pulmonary disease in two large populations. *Am. J. Respir. Crit. Care. Med.*, **176**, 167-173.
46 Elleder, M., Houstkova, H., Zeman, J., Ledvinova, J. and Poupetova, H. (2001) Pulmonary storage with emphysema as a sign of Niemann-Pick type C2 disease (second complementation group). Report of a case. *Virchows Arch.*, **439**, 206-211.
47 Zhang, Y., Calado, R., Rao, M., Hong, J.A., Meeker, A.K., Dumitriu, B., Atay, S., McCormick, P.J., Garfield, S.H., Wangsa, D. *et al.* (2014) Telomerase variant A279T induces telomere dysfunction and inhibits non-canonical telomerase activity in esophageal carcinomas. *PLoS One*, **9**, e101010.
48 Dumanski, J.P., Rasi, C., Bjorklund, P., Davies, H., Ali, A.S., Gronberg, M., Welin, S., Sorbye, H., Gronbaek, H., Cunningham, J.L. *et al.* (2017) A MUTYH germline mutation is associated with small intestinal neuroendocrine tumors. *Endocr. Relat. Cancer*, **24**, 427-443.
49 Kho, A.T., Chhabra, D., Sharma, S., Qiu, W., Carey, V.J., Gaedigk, R., Vyhlidal, C.A., Leeder, J.S., Tantisira, K.G. and Weiss, S.T. (2016) Age, Sexual Dimorphism, and Disease Associations in the Developing Human Fetal Lung Transcriptome. *Am. J. Respir. Cell Mol. Biol.*, **54**, 814-821.
50 Beghe, B., Bazzan, E., Baraldo, S., Calabrese, F., Rea, F., Loy, M., Maestrelli, P., Zuin, R., Fabbri, L.M. and Saetta, M. (2006) Transforming growth factor-beta type II receptor in pulmonary arteries of patients with very severe COPD. *Eur. Respir. J.*, **28**, 556-562.
51 Cho, M.H., McDonald, M.L., Zhou, X., Mattheisen, M., Castaldi, P.J., Hersh, C.P., Demeo, D.L., Sylvia, J.S., Ziniti, J., Laird, N.M. *et al.* (2014) Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir. Med.*, **2**, 214-225.

52 De Winter, J.P., De Vries, C.J., Van Achterberg, T.A., Ameerun, R.F., Feijen, A., Sugino, H., De Waele, P., Huylebroeck, D., Verschueren, K. and Van Den Eijden-Van Raaij, A.J. (1996) Truncated activin type II receptors inhibit bioactivity by the formation of heteromeric complexes with activin type I. receptors. *Exp. Cell Res.*, **224**, 323-334.
53 McDonald, M.L., Mattheisen, M., Cho, M.H., Liu, Y.Y., Harshfield, B., Hersh, C.P., Bakke, P., Gulsvik, A., Lange, C., Beaty, T.H. *et al.* (2014) Beyond GWAS in COPD: probing the landscape between gene-set associations, genome-wide associations and protein-protein interaction networks. *Hum. Hered.*, **78**, 131-139.
54 Boueiz, A., Chase, R., Lamb, A., Lee, S., Naing, Z.Z.C., Cho, M.H., Parker, M.M., Hersh, C.P., Crapo, J.D., Stergachis, A.B. *et al.* (2017) Integrative genomics analysis identifies ACVR1B as a candidate causal gene of emphysema distribution in non-alpha 1-antitrypsin deficient smokers. *bioRxiv*, in press.
55 Tilley, A.E., Walters, M.S., Shaykhiev, R. and Crystal, R.G. (2015) Cilia dysfunction in lung disease. *Annu. Rev. Physiol.*, **77**, 379-406.
56 Wain, L.V., Sayers, I., Soler Artigas, M., Portelli, M.A., Zeggini, E., Obeidat, M., Sin, D.D., Bosse, Y., Nickle, D., Brandsma, C.A. *et al.* (2014) Whole exome re-sequencing implicates CCDC38 and cilia structure and function in resistance to smoking related airflow obstruction. *PLoS Genet.*, **10**, e1004314.
57 Kenna, K.P., van Doormaal, P.T., Dekker, A.M., Ticozzi, N., Kenna, B.J., Diekstra, F.P., van Rheenen, W., van Eijk, K.R., Jones, A.R., Keagle, P. *et al.* (2016) NEK1 variants confer susceptibility to amyotrophic lateral sclerosis. *Nat. Genet.*, **48**, 1037-1042.
58 Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K. and Crapo, J.D. (2010) Genetic epidemiology of COPD (COPDGene) study design. *COPD*, **7**, 32-43.
59 Silverman, E.K., Chapman, H.A., Drazen, J.M., Weiss, S.T., Rosner, B., Campbell, E.J., O'Donnell, W.J., Reilly, J.J., Ginns, L., Mentzer, S. *et al.* (1998) Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. *Am. J. Respir. Crit. Care Med.*, **157**, 1770-1778.
60 Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, **6**, 80-92.
61 Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedon, J.C. *et al.* (2016) Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am. J. Hum. Genet.*, **98**, 653-666.
62 Jiang, D. and McPeek, M.S. (2014) Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet. Epidemiol.*, **38**, 10-20.
63 Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N. and Gaunt, T.R. (2013) Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.*, **34**, 57-65.
64 Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545-15550.
65 Morrow, J., Qiu, W., D.L., D., Houston, I., Pinto Plata, V.M., B.R., C., Marchetti, N., Criner, G.J., Bueno, R. and Washko, G.R. (2015) Network Analysis of Gene Expression in Severe COPD Lung Tissue Samples (abstract). *Am. J. Respir. Crit. Care Med.*, **191**.

66 Blake, J.A., Eppig, J.T., Kadin, J.A., Richardson, J.E., Smith, C.L., Bult, C.J. and the Mouse Genome Database, G. (2017) Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Res.*, **45**, D723-D729.

Legends to Figures

Figure 1. Network of the two sets of top genes with p-value < 0.01 in the case-control and family-based analyses focusing on rare, deleterious variants. The nodes in red are the top genes identified in the COPDGene case-control analysis; the nodes in pink are the first neighbors of the red nodes. The nodes in blue are the top genes identified in the BEOCOPD-ICGN family-based analysis; the nodes in light blue are the first neighbors of the blue nodes. Genes that are in both sets are colored in purple. Edges connecting genes to the largest hub *TULP3* are colored in red. These genes form one large well-connected component. Larger sized nodes indicate hubs (circle) and genes reported to be associated with COPD or lung function (diamond). Hubs include *TULP3*, *VNN2, ACVR2B, KCNA5, and GRB2,* which are the top genes with the most number of degrees in this network. 14 genes out of 513 are near GWAS loci for COPD or lung function (*CHRM3, DNLZ, EFEMP1, EFEMP2, EGFL8, GANAB, GNG3, PARN, PIP4K2B, NOTCH4, RUVBL1, SEC16A, TARS, TEKT5, THSD4*). The zoomed-in panel on the left shows the genes in the intraciliary transport particle A pathway (GO:0030991). The zoomed-in panel on the right shows the genes in the transforming growth factor beta-receptor binding pathway (GO:0005160) and *ACVR2B*.

Figure 2. A flow chart of the study design. COPDGene (pink) samples were sequenced in two batches (Baylor and ESP, see Methods). The family-based studies (blue) included two cohorts. Forty-nine pedigrees of the Boston Early-Onset study samples were sequenced and analyzed previously (26); we combined these data with another subset of these BEOCOPD and additional samples from the ICGN study. All of these sequenced subjects from BEOCOPD and ICGN were called together, forming the BEOCOPD-ICGN dataset (blue). We applied the family-based GESE test to the most severe cases and resistant controls in the BEOCOPD-ICGN dataset. We also performed single-variant, gene-based and pathway-based association tests in COPDGene and the BEOCOPD-ICGN samples. A final network analysis was conducted to look at the topological relationship between the top results from the two datasets.

Tables

Table 1. Baseline characteristics.

Datasets	COPDGene ESP		Baylor		ICGN-BEOCOPD				
	Case	Control	Case	Control	Severe Cases[1]	Moderate Cases	Resistant controls[1]	Other Controls	Other
N	192	188	293	316	853	431	101	118	51[2]
# Females	92	103	117	146	412	199	53	68	30
# Males	100	85	176	170	441	232	48	50	21
Age, yr	58.2 (5.1)	69.5 (5.6)	68 (6.4)	61.9 (5.6)	56.1 (12.2)	59.4 (12.1)	55.5 (11.3)	37.0 (20.4)	53.9 (24.4)
Pack-years	45.0 (26.2)	45.0 (23.5)	51.0 (29.0)	50.6 (19.1)	43.8 (31.0)	37.4 (25.6)	30.8 (21.9)	0.0 (7.6)	24.0 (39.9)
FEV_1 % predicted	30.0 (15.9)	98.2 (12.8)	30.2 (15.8)	92.7 (14.3)	30.0 (17.6)	65.5 (14.1)	98.0 (15.3)	96.3 (14.6)	77.4 (10.6)
FEV1/FVC	0.33 (0.10)	0.78 (0.07)	0.35 (0.12)	0.76 (0.07)	0.33 (0.14)	0.56 (0.13)	0.76 (0.06)	0.81 (0.10)	0.71 (0.12)

N – Number of subjects

Median (IQR) is presented for age, pack-years, FEV_1 % predicted, and FEV_1/FVC ratio for each dataset.

[1] Only severe cases (GOLD level III and IV) and resistant controls (see text) were included in the GESE test of the family-based data. All subjects were included in the association analysis of the family-based data.

[2] 51 subjects in the ICGN-BEOCOPD data had lung function values not consistent with either case or control status.

Table 2. Results of the GESE analysis on the BEOCOPD-ICGN dataset.

GENE	GESE p-value	Number of segregating families
*TBC1D10A**	1.1E-06	2
*RFPL1**	1.6E-06	4
DHODH#*	6.9E-05	2
*CYP4F12**	1.0E-04	4
*ANAPC7**	1.5E-04	1
RGS5#*	1.5E-04	2
CD101#*	1.7E-04	5
KCNMB4#*	1.8E-04	1
*ARMC12**	2.1E-04	4
VPS41#*	3.9E-04	5

Variants included are loss-of-function variants with MAF < 0.1%. The third column shows the number of families each gene is segregating in (present in all the cases and not in the controls). Genes marked with * show expression in the lung (defined as at least 50% of samples with FPKM > 0.5 in the Lung Genomics Resource Consortium RNA-seq samples). Genes marked with # are differentially expressed by FEV1% predicted in lung tissue (65).

Table 3. Nominally significant gene-based results in COPDGene for 329 candidate genes after conditioning on the lead GWAS SNP.

Gene	(conditioned on) lead GWAS SNP	#SNV	SKATO (unadjusted)	SKATO (conditional)
SEC16A	rs10870202	38	9.52E-04	1.05E-03
CDC7	rs1192404	5	5.87E-03	6.78E-03
CCDC38	rs12820313	5	7.20E-03	7.66E-03
CYP2A6	rs12459249	11	6.87E-03	1.15E-02
TRIP11	rs7155279	24	1.53E-02	2.47E-02
CNGB1	rs12447804	26	3.27E-02	2.89E-02
PBLD	rs7095607	3	6.99E-02	3.43E-02
RRP15	rs10429950	2	3.75E-02	4.07E-02
TNXB	rs2070600	63	5.95E-02	4.39E-02
CYFIP2	rs10515750, rs1990950	4	4.13E-02	4.46E-02
EGFL8	rs2070600	9	8.11E-02	4.60E-02
CHRNA5	rs17486278	5	1.83E-02	1.22E-01
AGER	rs2070600	12	4.89E-03	1.36E-01

The SKAT-O tests included functional (MODERATE effect defined by SnpEff) and rare (MAF < 5%) variants in the COPDGene study.

Table 4. Selected set of likely pathogenic variants annotated by ClinVar in *SERPINA1* and telomere-related genes.

		COPDGene		BECOPD-ICGN					
GENE	SNP	Case	Cont	Case	Cont	MAF	IMPACT	CLNSIG	Disease association
RTEL1	20:62324513:T:C	0	1	.	.	6.11E-05	missense_variant	5	Telomeropathy
SERPINA1	rs28929474	29	12	60	8	1.83E-02	missense_variant	5|5	Alpha-1 antitrypsin deficiency
SERPINA1	rs17580	53	29	148	32	3.04E-02	missense_variant	5	Alpha-1 antitrypsin deficiency
SERPINA1	rs28929470	4	6	5	1	4.95E-03	missense_variant	5	Alpha-1 antitrypsin deficiency
SERPINA1	rs28931570	2	4	8	2	1.62E-03	missense_variant	4|5	Alpha-1 antitrypsin deficiency
SERPINA1	rs121912714	.	.	3	0	7.04E-04	missense_variant	4	Alpha-1 antitrypsin deficiency
TERT	rs61748181	40	33	92	16	4.97E-02	missense_variant	5|2	Telomeropathy
TERT	5:1278865	1	0	.	.	7.49E-05	missense_variant	5	Telomeropathy
TERT	5:1280427	1	0	.	.	.	missense_variant	.	Telomeropathy
TERT	rs35719940	27	22	.	.	2.11E-02	missense_variant	5|2|2	Telomeropathy
TERT	rs34094720	4	7	.	.	1.53E-02	missense_variant	5|5|5|3|2|2	Telomeropathy
TERT	rs141425941	1	1	.	.	2.68E-04	missense_variant	5	Telomeropathy
TINF2	rs142777869	2	1	1	0	7.25E-04	missense_variant	5	Telomeropathy

Case, Cont - number of alternative alleles carried by the cases and controls in each dataset. Note that in the family-based data, there are approximately 6 times more cases than controls. MAF - minor allele frequency using Non-Finnish Europeans from ExAC. IMPACT - functional impact of each variant annotated by SnpEff. CLNSIG and Disease association are annotations from ClinVar; 2 = benign, 3 = likely benign, 4 = likely pathogenic, 5= pathogenic.

Table 5. Homozygous variants in COPD-related Mendelian genes.

		COPDGene		BECOPD-ICGN					
GENE	SNP	Case	Cont	Case	Cont	MAF	IMPACT	CLNSIG	Disease association
ATP6V0A2	12:124206896*	0	0	1	0	.	Splice_acceptor_variant	.	Cutis laxa
CFTR	rs1800076	0	1	3	0	2.48E-02	missense_variant	2|2|5	Cystic fibrosis
NPC2	rs140130028	.	.	2	0	0.00551	splice_donor_variant	5	Niemann-Pick disease type C2
SERPINA1	rs17580	1	1	4	1	3.04E-02	missense_variant	5	Apha-1-antitrypsin deficiency
TERT	rs61748181	5	0	2	0	4.97E-02	missense_variant	5|2	Telomeropathy
TERT	rs35719940	1	0	.	.	2.11E-02	missense_variant	5|2|2	Telomeropathy

Variants with homozygous genotypes in 29 Mendelian genes and were annotated with significance 4 and above by ClinVar, or have potential deleterious effects (MAF < 0.1% and predicted to be deleterious by FATHMM, SIFT, and CADD (>15). Case, Cont - number of alternative alleles carried by the cases and controls in each dataset. Note that in the family-based data, there are approximately 6 times more cases than controls. MAF - minor allele frequency using Non-Finnish Europeans from ExAC. IMPACT - functional impact of each variant annotated by SnpEff. CLNSIG and Disease association are annotations from ClinVar; 2 = benign, 3 = likely benign, 4 = likely pathogenic, 5= pathogenic.

Abbreviations

COPD: Chronic obstructive pulmonary disease

GWAS: Genome-wide association study

MAF: Minor allele frequency

GESE: Gene-based segregation test

FEV_1: Forced expiratory volume in 1 second

GOLD: Global Initiative for Chronic Obstructive Lung Disease

FPKM: Fragments Per Kilobase of transcript per Million mapped reads

PPI: Protein-protein interaction network