Identifying chronic obstructive pulmonary disease subtypes using multi-trait genetics

Andrey Ziyatdinov,[a,b] Brian D. Hobbs,[b,c] Samir Kanaan-Izquierdo,[e,f,g] Matthew Moll,[b,c] Phuwanat Sakornsakolpat,[b,h] Nick Shrine,[k] Jing Chen,[k] Kijoung Song,[i] Russell P. Bowler,[j] Peter J. Castaldi,[b] Martin D. Tobin,[k,l] Peter Kraft,[a] Edwin K. Silverman,[b,c] Hanna Julienne,[d] Michael H. Cho,[b,c,m] and Hugues Aschard[a,d,m,*]

[a]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[b]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA
[c]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA
[d]Institut Pasteur, Université Paris Cité, Department of Computational Biology, Paris F-75015, France
[e]Centre de Recerca en Enginyeria Biomèdica, Universitat Politècnica de Catalunya, Barcelona 08028, Spain
[f]CIBER of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Barcelona, Catalonia, Spain
[g]Institut de Recerca Sant Joan de Deu, Esplugues de Llobregat, Spain
[h]Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand
[i]Human Genetics, GlaxoSmithKline, Collegeville, PA, USA
[j]Division of Pulmonary and Critical Care, Dept. Med, National Jewish Health, Denver, CO, USA
[k]Department of Health Sciences, University of Leicester, Leicester, UK
[l]National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK

eBioMedicine
2025;113: 105609
Published Online xxx
https://doi.org/10.1016/j.ebiom.2025.105609

Summary

Background Chronic Obstructive Pulmonary Disease (COPD) has a broad spectrum of clinical characteristics. The aetiology of these differences is not well understood. The objective of this study is to assess whether respiratory genetic variants cluster by phenotype and associate with COPD heterogeneity.

Methods We clustered genome-wide association studies of COPD, lung function, and asthma and phenotypes from the UK Biobank using non-negative matrix factorization. We constructed cluster-specific genetic risk scores and tested these scores for association with phenotypes in non-Hispanic white subjects in the COPDGene study.

Findings We identified three clusters from 482 variants and 44 traits from genetic associations in 379,337 UK Biobank participants. Variants from asthma, COPD, and lung function were found in all three clusters. Clusters displayed varying effects on white blood cell counts, height, and body mass index (BMI)-related phenotypes in the UK Biobank. In the COPDGene cohort, cluster-specific genetic risk scores were associated with differences in steroid use, BMI, lymphocyte counts, and chronic bronchitis, as well as variations in gene and protein expression.

Interpretation Our results suggest that multi-phenotype analysis of obstructive lung disease-related risk variants may identify genetically driven phenotypic patterns in COPD.

Funding MHC was supported by R01HL149861, R01HL135142, R01HL137927, R01HL147148, and R01HL089856. HA and HJ were supported by ANR-20-CE36-0009-02 and ANR-16-CONV-0005. The COPDGene study (NCT00608764) is supported by grants from the NHLBI (U01HL089897 and U01HL089856), by NIH contract 75N92023D00011, and by the COPD Foundation through contributions made to an Industry Advisory Committee that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer and Sunovion.

Keywords: COPD; Genetic epidemiology; Multitrait analysis; Pathways

Introduction

Chronic Obstructive Pulmonary Disease (COPD) is a phenotypically heterogeneous disease.[1] Some have hypothesized that COPD is a syndrome constituted of multiple disease subtypes involving different biological mechanisms.[2,3] Understanding the molecular basis

*Corresponding author. Institut Pasteur, Université Paris Cité, Department of Computational Biology, Paris F-75015, France.
E-mail address: hugues.aschard@pasteur.fr (H. Aschard).
[m]Jointly supervised the project.

Research in context

Evidence before this study

Chronic Obstructive Pulmonary Disease (COPD) is a phenotypically heterogeneous disease that varies significantly in symptomatic and physiologic presentation. Many studies explored the biological basis for this heterogeneity through the classification of patients based on their molecular and phenotypic characteristics. However, this approach is severely hampered by the variability of phenotypes and biomarkers with time, treatment, and disease progression.

Added value of this study

This work investigates whether a genetically driven approach, based on shared genetics between phenotypes and biomarkers, can better inform the heterogeneity of obstructive lung diseases. By using germline genetic markers, which are present from birth and do not vary with time, our work addresses the existing limitation of biomarkers. Furthermore, the proposed approach takes advantage of the availability of genome-wide association study results from many human traits typically not available in a single cohort.

Implications of all the available evidence

Our study demonstrates that examining the shared genetics across phenotypes can help resolve some of the heterogeneity underlying obstructive lung diseases. We identified three distinct clusters of lung-associated variants displaying different associations with biologic and clinical phenotypes: one related to eosinophils and inflammatory biomarkers; a second related to lower body mass and greater emphysema; and a third with height. Overall, it suggests that genetic risk scores built out of these clusters of variants can help identifying patients that may benefit from more specific treatments.

underlying this heterogeneity in COPD can advance our knowledge of COPD aetiology and improve patient treatment. However, efforts in learning specific disease mechanisms and potential COPD subtypes using molecular markers have been hampered by multiple issues, including variability in measurement across time and conditions, variability across targeted tissues, and reverse causation (e.g., when the trait is influenced by disease treatment).

Genetic variants identified in genome-wide association studies (GWAS) are present from birth and do not vary with time, disease, treatment, or disease course. Thanks to these properties, they can be used to assess subtypes driven by biological mechanisms, while circumventing potential reverse causation effects. Moreover, GWAS have now been conducted on a vast number of human traits and diseases. These studies showed that genetic variants for one trait or disease often also have effects on other traits, a phenomenon known as pleiotropy. Building on these two features, several methods have been proposed for inferring genetically driven disease subtypes.[4–7] These methods leverage the relationships between disease-associated variants and other phenotypes to construct clusters of variants based on similarity in their multitrait association pattern. The variants within each group can be further characterized and might ultimately be used to classify individuals. COPD is a strong candidate for such disease subtype inference, as we and others have demonstrated that genetic variants associated with COPD have substantial pleiotropic effects.[1,8]

In this work, we examined genetic variants identified in GWAS of moderate-to-severe COPD; spirometry, as COPD is defined by decrements in lung function; and asthma, a disease with extensive clinical and physiological overlap with COPD. We applied a Bayesian method previously used in type 2 diabetes, another complex disease characterized by disease heterogeneity,[5] to cluster these variants based on their association with a broad range of traits measured in the UK Biobank, a large population cohort with hundreds of phenotypic measures available. We then used individual-level genetic and phenotypic data from the COPDGene[9] study, a cross-sectional cohort of COPD cases and controls with deep pulmonary phenotyping and extensive clinical data, to investigate potential subtypes based on the identified clusters.

Methods

Selection of genetic variants associated with COPD, lung function, and asthma

Genetic variants relevant to COPD were identified from three obstructive lung disease-related GWAS[1,8,10]: COPD itself,[1] spirometric lung function phenotypes[8] and asthma.[10,11] We relied on the first two datasets' contemporaneous (2019) published genome-wide association studies. For asthma, we conducted a custom meta-analysis of asthma results from UK Biobank[11] and the GABRIEL consortium.[10] Our list included 164, 279, and 45 variants for COPD, lung function, and asthma, respectively (see Supplementary Methods). All variants were reported as genome-wide significant (or conditionally significant, after adjusting for nearby variants, in the case of COPD). After removing 6 duplicates, we assessed a total of 482 variants. For the primary analysis of clustered variants, as the causal variants were not known, we used all 482 variants. However, for analyses that required independent variants, we removed correlated ones by performing linkage disequilibrium pruning, filtering out variants correlated across the four subsets (e.g., variants selected in the asthma GWAS that might be correlated with variants

selected from the COPD GWAS) using SNPclip (ldlink.nci.nih.gov/?tab=snpclip) and an r^2 threshold of 0.1 in 1000 Genomes European ancestry subjects,[12] and resulting in 377 independent variants.

Selection and analysis of traits from UK Biobank

To assess the effect of these respiratory genetic variants on other phenotypes, we relied on genetic association studies of phenotypes in unrelated European ancestry participants from UK Biobank.[13] Although a large number of phenotypes have been studied, most of these do not have any association signals, and some are redundant with the GWAS included here. To ensure a robust inference of pleiotropic effects, we applied a stringent filtering from an initial set of 2409 phenotypes to remove non-informative and low-quality GWAS. First, we excluded traits directly related to lung function, COPD, and asthma (e.g., respiratory diagnosis codes and different measures of asthma). Second, we removed attributes with small sample sizes (effect sample size, or N_{eff} < 200,000) to ensure that the included GWASs carried a similar amount of information. Third, we removed traits that did not show any genome-wide significant association ($P < 5 \times 10^{-8}$) for at least one of the 482 selected variants. Fourth, we removed highly correlated phenotypes based on Pearson correlations of Z-scores, as implemented in the R package caret[14] (r^2 threshold = 0.8). Finally, we oriented all Z-scores to COPD risk-increasing alleles and divided Z-scores by the square root of the effective sample size, thus, converting Z-scores to standardized effect sizes. As an additional requirement for implementing the clustering approach (see next section), we split each row of the Z-score matrix into two meta-trait Z-scores, positive and negative. That resulted in doubling the number of traits in the downstream clustering analysis (Fig. 1).

Clustering by nonnegative matrix factorization

We clustered variants using a previously described Non-negative Matrix Factorization (NMF) method, which has shown promising results in disease subtype learning in diabetes.[5] As non-negative matrix factorization requires positive values, we construct one matrix of positive Z-scores, and another comprised of negative Z-scores (thus, 2T instead of T traits). NMF decomposes the $2T \times S$ matrix of Z-scores, where S the number of variants into a lower-dimensional representation $Z \approx WH^T$ with weight matrices W and H of sizes $2T \times K$ and $S \times K$, respectively, and where K is a latent dimension to be learned from the data. More specifically, we applied a probabilistic Bayesian model of NMF[15] that iteratively learns the weights in W and H. The approach includes four key steps: (i) reducing a reconstruction error through the β-divergence function; (ii) adding a K_λ-length vector of relevance weights λ as an auxiliary variable (K_λ = 32, by default); (iii) using half-normal priors on weights (L2-norm regularization); and (iv) satisfying the non-negative constraints on weights (W > 0, H > 0). Altogether, this Bayesian implementation of NMF automatically learns the latent dimensionality K and avoids ambiguity compared to other NMF algorithms.[15] In practice, the number of learned dimensions K is obtained by taking non-zero entries in the K_λ-length of relevance weights λ. Further details are available in Udler et al.[5]

Building cluster-specific genetic risk scores

After identification of clusters, we derived weighted genetic risk scores (GRS) for each of the K clusters inferred from the NMF using the variants weight from the matrix H. More specifically, for a cluster i and variants g_j with j = 1...S, the weighted cluster-specific GRS are calculated as $GRS_i = \sum_S H_{ji} g_j$. For comparison purposes, we also defined a baseline GRS defined as the unweighted sum of genotypes: $GRS_0 = \sum_S g_j$. Note that because the $H_{j=1...K}$ does not sum to 1, the sum of the GRS_i is not equal to GRS_0.

COPDGene dataset

The characteristics of various genetic risk scores were examined using individual-level data from the

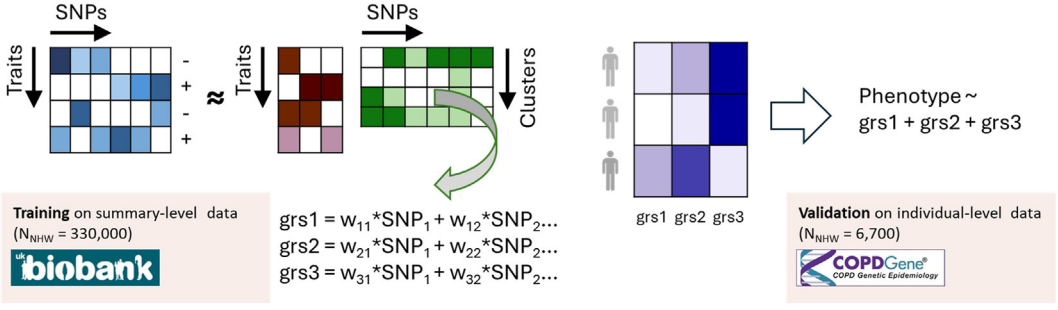

Fig. 1: Study overview. We identified a set of 482 genetic variants associated with obstructive lung disorder and extracted association z-scores between those variants and 44 phenotypes from genome-wide association studies (GWAS) conducted in the UK biobank. We factorized this association matrix into two matrices of traits and variants weights using the NMF approach, resulting in the clustering of the variants in three groups. Those clusters were then used to derive three component genetic risk scores (grs), which were tested for association with phenotypes in the COPDGene cohort. The distribution of the grs across COPD cases was ultimately used to examine potential COPD disease axes.

COPDGene study. The COPDGene study (NCT00608764, www.copdgene.org) recruited 10,198 non-Hispanic white (NHW) or African-American (AA) participants, aged 45–80 years old, with at least 10 pack-years of smoking and no diagnosed lung disease other than COPD or asthma.[9] IRB approval was obtained at all study centers, and all study participants provided written informed consent. Illumina (San Diego, CA) performed genotyping on the HumanOmniExpress array, and imputation to HRC 1.1 was performed using the Michigan Imputation Server. COPDGene subjects were extensively phenotyped, with data collected using questionnaires, spirometry, and inspiratory and expiratory CT scans at baseline. Subjects were invited to participate in follow-up visits, including spirometry and CT scans, and a subset had cell counts and biomarkers.[16] In this study, we considered a total of 240 traits tested in NHW, as all summary statistics were generated from European ancestry cohorts. All traits were tested for association with the cluster-specific GRS described in the previous section. Models used complete case analysis and adjusted for relevant covariates such as age, sex, pack-years of smoking and smoking status, principal components of ancestry (as previously described[1]), scanner and centre as appropriate.

Statistical tests and models fit in COPDGene

We fit linear and logistic regression models for 240 quantitative and binary traits respectively modelling the effect of the four genetic risk scores ($GRS_{0...3}$) using individual-level genetic and phenotypic data from the COPDGene cohort. We first estimated the marginal effect of each GRS_i using a standard univariate model: $f(Y) \sim GRS_i + cov$, where cov represents trait specific covariates. We then assessed the impact of decomposing GRS_0 into the $GRS_{1...3}$ using two approaches. First, we compared the above marginal model against a conditional model including GRS_0: $Y \sim GRS_0 + GRS_i + cov$ using a likelihood ratio test (LRT). Second, we assessed the overall contribution of all three $GRS_{1...3}$ by comparing the marginal model for GRS_0 against a joint model, including all three cluster-specific GRS_i: $Y \sim cov + GRS_0 + \sum_{i=1...3} GRS_i$, again using a LRT. This test of heterogeneity, referred to further as P_{het}, quantifies the improvement in model fitting when decomposing GRS_0 into K GRS_i. To examine the relative contribution of the $GRS_{1...3}$ to heterogeneity, we also extracted effect estimate from this joint model. For individual GRS significance derived from the marginal, conditional, and joint model, we use the stringent Bonferroni correction threshold of 7×10^{-5}, accounting for 723 tests conducted for each model. For P_{het}, we applied a Benjamini and Hochberg[17] correction and reported the Q-value with False Discovery Rate (FDR) < 0.1. Note that both Bonferroni and FDR correction can be conservative as they do not account for the correlation between the traits tested.

Association of GRS with gene expression, protein biomarkers, and clinical outcomes in COPDGene

We tested whether cluster-specific GRSs were associated with changes in gene expression using RNA-sequencing from peripheral blood taken at the 5-year follow-up visit in COPDGene. We used limma/voom,[18] and adjusted all analyses for age, sex, smoking status, ancestry principal components, and batch. We performed a similar analysis with SomaScan plasma proteomics data, adjusting for age, sex, smoking status, and ancestry principal components. Besides investigating associations of the GRSs with phenotypes and omics data in COPDGene, we also sought to determine whether the cluster-specific GRS could identify individuals with higher or lower risk of lung-related phenotypes. As an example of an application, we used the top and bottom decile of subjects based on high scores on GRS_1, and low scores on GRS_2 and GRS_3, and first examined the association of eosinophils and steroid use, adjusting for age, sex, ancestry-based principal components, and GOLD stage. Finally, we investigated heterogeneity in medication use conditional on the same percentile of GRSs, using information on COPD exacerbations and medication treatments available in COPDGene.

Ethics statement

Ethical approval was granted for each of the original studies used in the present study and no new ethical review board approval was required. Individual-level data from the UK Biobank were accessed through the UK Biobank Application #20915.

Role of funders

The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the funding bodies, which had no role in the design of the study and collection, analysis, and interpretation of data or in writing of the manuscript.

Results

Overview of the multi-trait genetic approach

To address our objective of identifying potentially distinct COPD subtypes, we first created a matrix of 482 COPD-, lung-function, and asthma-associated genetic variants; and 44 phenotypes from the UK Biobank (Supplementary Table S1). We clustered these variants using a previously described non-negative matrix factorization (NMF) algorithm, identifying three clusters of genetic variants. Second, we created three genetic risk scores from these clusters and investigated their link with COPD-related phenotypes using individual-level data from participants of COPDGene (Fig. 1).

Variant characteristics and cluster inference

We first cross-examined the COPD genetic association with association at the three other phenotypes (asthma,

FEV_1 and FEV_1/FVC), using a subset of 377 independent variants out of the 482 variants available. Plots of effects of variants on respiratory traits are shown in Supplementary Fig. S1. We then extracted the association between the larger set of 482 variants and traits measured in the UK Biobank cohort. We filtered and harmonized a total of 2409 UK Biobank phenotypes, selecting outcomes displaying evidence of association with variants, removing phenotypes with low sample sizes, and removing redundancy due to high phenotypic correlation (see **Methods**). After this processing, the association matrix included 44 phenotypes with Z-score for all 482 variants. No clear pattern emerged from a visual inspection of this matrix when using a standard hierarchical clustering algorithm (Supplementary Fig. S2).

The application of the NMF algorithm to this matrix identified three clusters with associated weight matrices W and H, reflecting the contribution of traits and genetic variants, respectively. To characterize cluster compositions in trait dimension, we operated on normalized trait weights (unit sum of cluster weights in columns of W) and identified the top normalized trait weights for each cluster (Fig. 2). We oriented all weights such that positive and negative weights of normalized traits reflect increasing and decreasing risk of COPD, respectively. Cluster 1 displayed high positive weights for wheeze, eosinophil percentage, and neutrophil counts. Cluster 2 had negative weights for traits linked to body composition and obesity (hip circumference, body fat, and BMI). Cluster 3 displayed positive weights for height, grip strength, and birth weight, and negative weights for blood cell counts.

To characterize cluster compositions by variants, we derived variant weights (unit sum of variant weights in rows of H) and assigned variants with weights >50% to the corresponding cluster. Approximately 78% of all variants match that criterion, with 156, 148, and 78 variants selected for clusters 1, 2, and 3, respectively. The assignment of variants across clusters is illustrated in the alluvial plot in Supplementary Fig. S3. We compared the origin of these variants –i.e., whether they were selected from COPD, lung function or asthma GWAS– against the expected from a random assignment (out of the 482 variants, 34%, 57%, and 10% were selected from the COPD, lung function, and asthma GWAS, respectively). Cluster 1 variants display a small enrichment for asthma variants and a reduced representation of lung function variants (33%, 51%, and 16% variants from the COPD, lung function, and asthma sets, respectively). The overrepresentation of asthma variants in this cluster is consistent with the composition of traits (Fig. 2), where wheeze and eosinophil percentage have the largest weights. Conversely, in cluster 2, lung function variants were slightly overrepresented, and asthma variants underrepresented (34%, 64%, and 2% variants from the COPD, lung function and asthma sets, respectively). Cluster 3 did not display specific enrichment (35% 59%, and 8% variants from the COPD, lung function and asthma sets, respectively).

Clinical features of inferred clusters of variants in COPDGene

To determine whether the inferred clusters of variants were related to COPD phenotypes, we constructed three cluster-specific weighted genetic risk scores (GRS_1, GRS_2 and GRS_3), and an unweighted genetic risk score (GRS_0) including all 482 variants, that we applied to individual-level genotypes from COPDGene (**Methods** and Fig. 1). We first tested the marginal association between each of the four GRSs (GRS_0, GRS_1, GRS_2 and

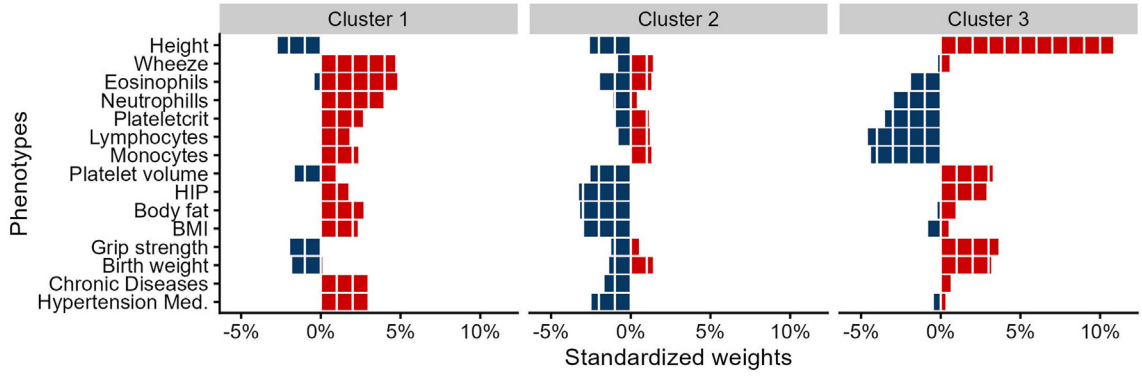

Fig. 2: Distribution of trait weights across the variant clusters in UK Biobank. Distribution of traits weight extracted from the Non-negative Matrix Factorization (NMF) analysis for the top 15 traits with the largest contribution to the three clusters. For comparison purposes, weights were normalized to have a sum of one within each cluster (X axis, in percentage). Red bars correspond to weights derived from positive z-scores, blue bars to weights derived from negative z-scores, reflect increasing and decreasing disease risk, respectively. Cluster 1 displayed high positive weights for inflammation-related phenotypes and blood cell counts. Cluster 2 had negative weights for traits linked to body composition and obesity. Cluster 3 displayed a large positive weight for height, and negative weights for blood cell counts.

GRS_3) and 240 features and outcomes measured in COPDGene (Supplementary Table S2). As expected, GRS_0, which includes all variants, displayed the strongest association with most phenotypes and was close to the best association from $GRS_{1..3}$ (Supplementary Fig. S4 and Table S3, except for height peak expiratory flow (PEF), and steroid treatment. Fig. 3 presents the effect estimates and standard errors of all four GRS for selected outcomes representing different phenotypic groups (lung function, anthropometric measurements, imaging, etc). Focusing on nominally significant signals, GRS_1 was associated with the highest eosinophils, highest self-reported steroid treatment, and the lowest 6-min walk distance. Cluster 2 was associated with the lowest FEV_1/FVC ratio and highest emphysema fraction. Cluster 3 was associated with higher height and FEV_1/FVC, and the largest risk of coronary disease.

For each trait, we then performed a test of heterogeneity comparing a joint model including all four $GRS_{0...3}$ against a baseline model including only GRS_0.

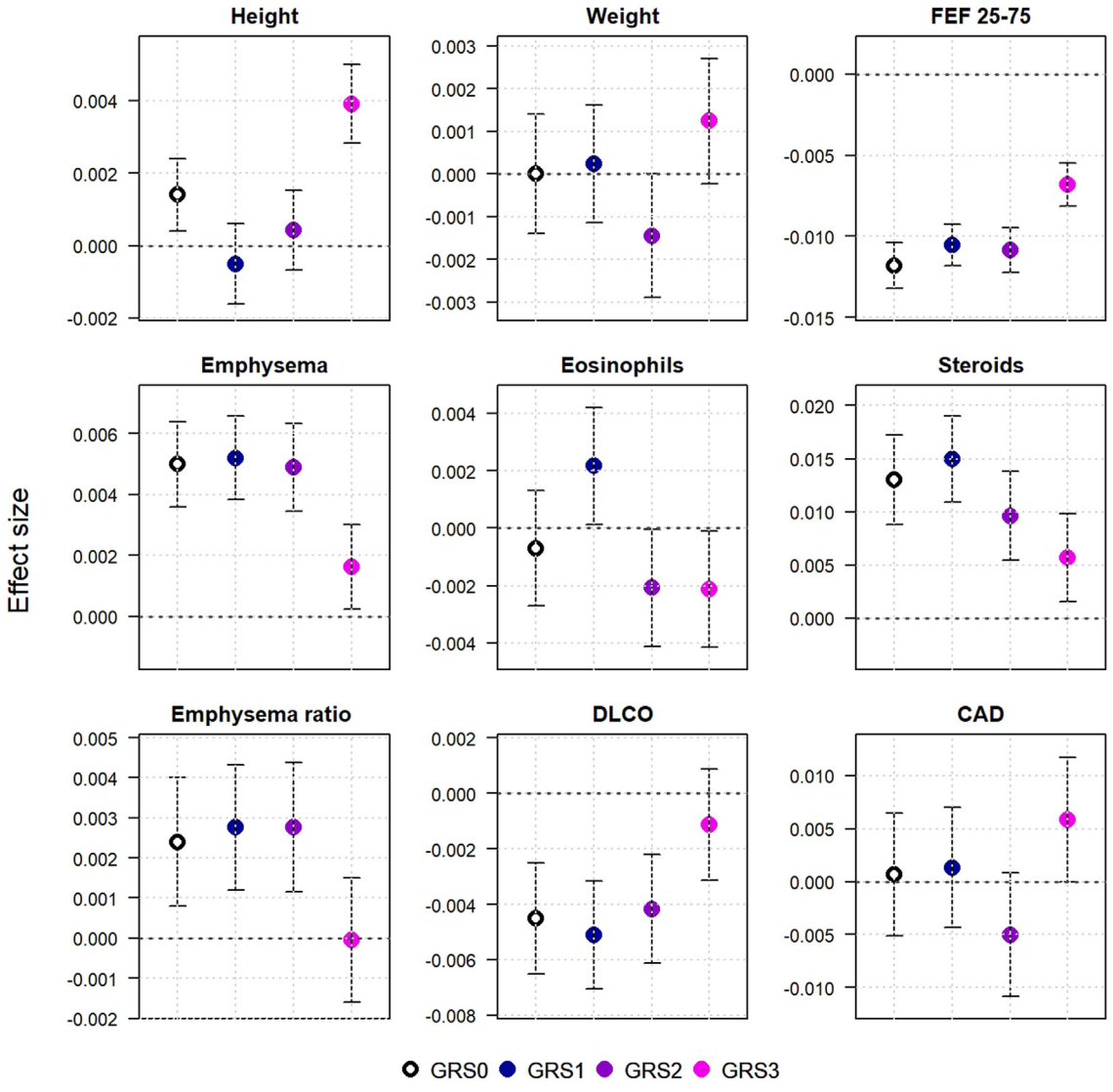

Fig. 3: Effects of GRSs on selected traits in the validation COPDGene dataset. Point estimates (effect size) and 95% confidence intervals for association between COPDgene phenotypes and cluster-specific GRSs ($\boldsymbol{GRS_{1..3}}$; from dark blue to pink) and unweighted GRS ($\boldsymbol{GRS_0}$; black). For comparison purposes, all GRS were re-scaled to a unit variance. We selected traits representing different COPD phenotypic groups: demographics (height, weight), lung function (forced expiratory flow at 25–75% of forced vital capacity (FEF 25–75) and diffusing capacity for carbon monoxide (DLCO), imaging (visual emphysema score (emphysema) and upper third/lower third emphysema ratio (emphysema ratio)), asthma-related traits (eosinophil count, steroid treatment), and comorbidities (coronary artery disease (CAD)).

Overall, 47 traits out of 240 showed a significant effect of including cluster-specific GRS_1, GRS_2 and GRS_3 in addition to GRS_0 (column P_{het}, Supplementary Table S3). Fig. 4 presents the relative contribution of the three GRSs of this conditional model for these 47 traits. GRS_1 and GRS_2 were significant for most phenotypes, suggesting a complementary contribution. We also found many of the same trait associations as in the

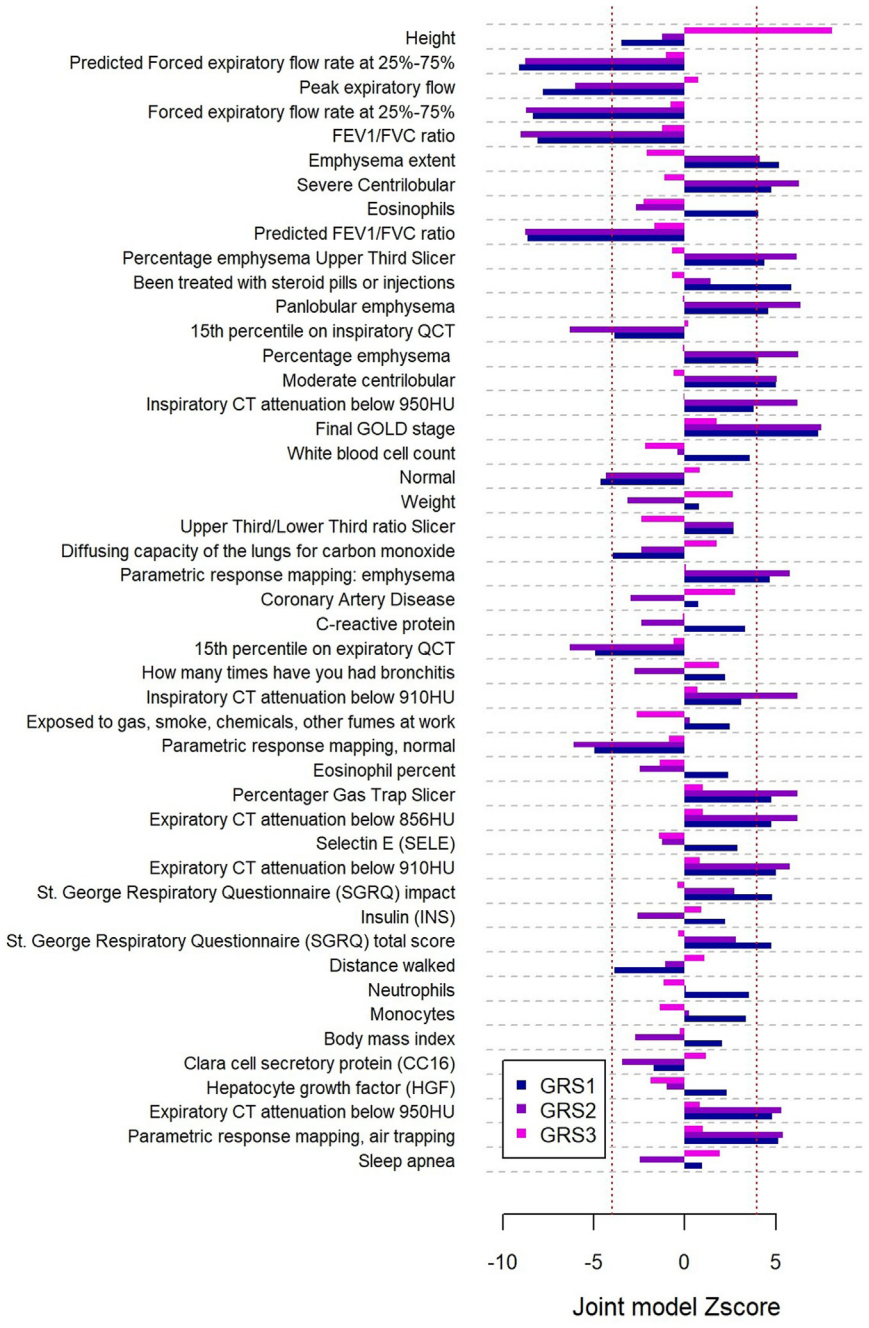

Fig. 4: Contribution of cluster-specific GRSs. Out of 240 COPDgene phenotypes tested for association with genetic risk scores, a total of 47 phenotypes showed a statistically significant (P_{het}) improvement of model fit at an FDR of 0.1 when comparing the marginal GRS_0 model against a full model including GRS_0 and all GRS_{1-3}. The barplots represent the relative contribution of GRS_1, GRS_2, and GRS_3, measured as Zscore derived from the full model, for these 47 phenotypes, highlighting which of the three GRS convey the improved fit. Phenotypes are order by P_{het}. Red dash lines indicate the stringent Bonferoni significance threshold accounting for a total of 723 tests.

UK Biobank. GRS_1 showed enrichment for significant association with blood cell counts and inflammatory biomarkers, including C-Reactive Protein and Selectin, and hepatocyte growth factor (HGF/c-MET), previously implicated in COPD pathogenesis. GRS_2 showed association with obesity and related traits including BMI, insulin, coronary-artery disease, and sleep apnoea. Finally, GRS_3 showed association with height, most clinical lung function measurements, and COPD-related phenotypes including CC16, a biomarker with previous associations with COPD. Altogether these results offer an indirect validation of the trait weights learned in the UK Biobank dataset (Fig. 2) and suggest that the derived partitioned GRS can partly capture heterogeneity of clinical features related to COPD.

Functional enrichment and GRS-stratified participants characteristics

We first conducted in silico functional enrichment analysis for variants within each cluster using FUMA[19] and a limited set of annotations (Supplementary Table S4 and Supplementary Methods). Cluster 1 showed a significant enrichment for immunity and inflammation pathways that was consistent across multiple input databases, and in agreement with all previous results. Cluster 2 harboured significant enrichment for a single annotation related to endocytosis. Cluster 3 showed enrichment for a broad range of pathways covering many biological components and cellular processes.

We then tested the association of the GRSs with peripheral blood RNA-Sequencing and SomaScan proteomics data using individual-level data from COPDGene. Data were available on 2666 subjects for gene expression, and 3687 subjects (4979 protein levels) for SomaScan proteomics.[20] At an FDR of 0.1, we found 1, 2, and 18 differentially expressed genes for GRS_1, GRS_2, and GRS_3 (Supplementary Table S5). GRS_1 was associated with ANKRD35. GRS_2 with NISCH and ILF3, and GRS_3 was associated with multiple genes in the MHC region on chromosome 6 including ZFP57, BTN3A2, HLA-A, H4C13, HLA-DQB1, and HLA-DQB2. Protein results are shown in Supplementary Table S6. No protein expression was significantly different GRS_1 and GRS_2. Top results for GRS_1 included LAIR2 and CRHBP. Top results for GRS_2 included IL-17 RC and SDF1 (stromal derived factor-1). Five proteins display highly significant differential expression with GRS_3. This included three proteins from the MHC region: BT3A3, MICA, and MICB; and FTMT, a mitochondrial ferroxidase enzyme and RGAP1 (encoded by *RACGAP1*), neither in the MHC region.

Finally, we explored whether cluster-specific GRS could identify individuals with higher or lower risk of specific clinical outcomes. Based on the results of heterogeneity testing, we examined high scores on GRS_1, and low scores on GRS_2 and GRS_3, and tested the association with eosinophils. The top decile had a significantly higher level of eosinophils, after adjusting for age, sex, ancestry-based principal components, and GOLD stage (P = 0.007 [Wald test]), and a OR 1.66 (95% CI, 1.10–2.52) fold increased requiring steroid treatment. Finally, we investigated potential heterogeneity in treatment among participants from COPDGene harbouring extreme values of the genetic risk scores (Supplementary Fig. S5 and Table S7). When comparing the characteristics of the top and bottom 5th percentiles of the three GRSs, we observed strong heterogeneity in corticosteroid, steroid, and theophylline treatments when stratifying participants by GRS_1. Groups defined by GRS_2 were marked by differences in long-acting beta-antagonist and ipratropium treatments.

Discussion

COPD is characterized by a simple and effective diagnostic criterion based on spirometry. This diagnosis has arguably led to greater recognition of the disease, effective bronchodilator therapy, and improved outcomes. However, patients with COPD demonstrate substantial clinical heterogeneity. Identifying the molecular basis for this heterogeneity has proven challenging. One major challenge to explaining COPD heterogeneity is the long course of the disease. Most phenotypic characteristics, such as exacerbations, blood cell counts, and degree of emphysema, are affected not only by severity but also by disease course and effects of treatment. Assessing COPD heterogeneity using genetic variants offers an opportunity to assess clinical heterogeneity without these confounders.

In this work, we explored a multi-trait genetic approach based on a set of genetic variants associated with COPD, lung function, and asthma. We identified three different groups of genetic variants. Although these variants were taken from three sources (COPD, lung function, and asthma), variants did not simply segregate by their source. This finding is particularly notable for asthma, for which genetic risk appears to be enriched for immune cells and overlap with autoimmune disease, in contrast to COPD and lung function loci, for which genetic signals are enriched in regulatory regions from lung tissue.[1,8] This finding is in agreement with previous work showing substantial heterogeneity in asthma genetic association signals.[21] These results provide evidence i) that the genetic risk of COPD can be broken into at least three fairly distinct components, and ii) that the genetic profile of individuals for each of these components (as measured by the three genetic risk scores) are associated with different clinical characteristics of COPD. While these results are unlikely to directly impact current clinical decision-making, progress in genome-wide association and variant to function studies suggest that partitioned genetic risk scores may follow the progress of polygenic risk scores in having clinical impact.

These results were further supported by the analysis in COPDGene. COPDGene data was not used for the phenotypic association for clustering, and thus the consistent associations of height, body mass index, and cell counts confirm these association results. These three groups of genetic variants included one related to eosinophils and inflammatory biomarkers, a second related to lower body mass and greater emphysema; and a third with higher height, the strongest associations with lung function, and an association with coronary disease. The individual GRSs also demonstrated associations with both gene expression and protein biomarkers. Aside from the association of GRS_3 with the HLA region (likely driven by genetic variants in this region), most of these associations were relatively weak, nevertheless, they do support the hypothesis that these GRSs reflect differing biology. Variants comprising GRS_1 appear to affect ANKRD35, a paralog of ZDHHC13 that may be associated with granulocyte count.[22] For GRS_2, Nisch modified mice exhibit an emphysema-like phenotype[23] and ILF3 is reduced in bronchoalveolar lavage from patients with COPD compared with never smoking controls.[24] For GRS_3, in addition to multiple HLA associations, FTMT is involved in ferroptosis, known to be an important pathway in COPD,[25] and RGAP1 (in addition to MICB) was recently identified in a study attempting to identify drug targets for COPD by applying colocalization and Mendelian Randomization to lung function and COPD GWASs and blood protein quantitative trait loci (pQTLs) from the INTERVAL study.[26] Whether these results also translate to other cohorts and respiratory traits is unclear; and the observed association should ultimately be replicated in other cohorts for validation.

Our results also suggest that GRSs may be able to identify patients that may benefit from more specific treatments. Using a combination of GRSs consistent with higher eosinophil burden based on UK Biobank phenotypes, we identified subsets of participants from COPDGene with different eosinophil counts and history of steroid use. Our method avoids some of the confounding present in phenotypic and/or other molecular data. Genetic variants are stable biomarkers and can be used for prediction prior to the development of disease or during disease treatment. Our ability to identify clusters and associations was based on the largest genetic association studies available to date across multiple phenotypes in the UK Biobank.

This study has also some limitations. First, even with 482 variants to predict relevant subtypes or disease axes,[2] power is likely still limited.[27] Moreover, the lung disease GWAS used to select those variants included mostly individual of European ancestry (≥83%), but also some participants of other ancestries. The impact of this heterogeneity on the variant selection is expected to be modest but cannot be excluded. Second, several of these signals likely represent the same causal signal, and many associated phenotypes in the UK Biobank were highly correlated, reducing the number of analysed traits. Furthermore, the UK Biobank lacks many of the respiratory phenotypes that may be useful for COPD phenotypes (such as imaging). Third, we confirmed the relevance of our variant clustering in COPDGene, a well-characterized cohort of participants with a history of cigarette smoking. Whether these results also translate to other cohorts and respiratory traits is unclear. Fourth, our analysis was limited to non-Hispanic whites as our source data was from predominantly European ancestry; analysis in more diverse cohorts is needed. Moreover, cluster specific genetic risk scores were developed in a population-based cohort in the UK, which substantially differs from the US smoking- and COPD-case enriched COPDGene study. Fifth, our method does not include molecular data and identified target gene, as performed in the recent work by Shrine et al.[28] This study, which includes over 1000 multi-ancestry loci, created an alternative, pathway-specific genetic risk score and associated this score with phenotypes. Whether this gene-pathway-first method is better for the identification of biologically relevant pathways, and others are better for sub-phenotyping is not clear. In addition, we applied these results to a well-phenotyped COPD dataset; how our method also applies to cohorts of asthma or asthma COPD overlap is not known. Sixth, we did not include variants below significance thresholds (e.g., using pruning and thresholding or Bayesian methods) as used by most polygenic risk scores. Future work in genetic subtyping could expand in this direction, including additional sub-genome-wide significant variants, but also functional genetics data such as cell types and mechanisms, all of which should further improve our ability to identify heterogeneity from genetic profiles.

In summary, we clustered genetic variants associated with obstructive lung disease using a diverse set of phenotypes, identifying three multi-trait/multi-variant disease scores. These scores demonstrate different associations with biologic and clinical phenotypes.

Contributors

AZ, HA and MC conceived and designed the study. AZ HJ MM, KS and BDH conducted the analyses. AZ, HA and MC wrote the paper, with input from PJC, RPB, MDT, PK, EDK, SKI, NS, JC and PS. HA and MC verified the underlying data. All authors read, approved, and made changes to the manuscript's final draft.

Data sharing statement

Individual-level data from the UK Biobank were accessed through the UK Biobank Application #20915. Individual-level data from COPDGene data are accessible through dbGaP Study Accession number phs000179.v6.p2 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000179.v6.p2). Genome-wide association study (GWAS) results of pulmonary phenotypes were downloaded from multiple sources, including the PheWeb catalogue,[13] UK Biobank GWAS[11] and dedicated page from outcome-specific studies.[1,8,10]

Declaration of interests

MHC has received grant support from Bayer and consulting fee from Apogee. EKS has received grant support from Northpond Laboratories and Bayer. MM has received honorarium from the NY State Thoracic Society and the ATS. PJC has received support from Bayer and Sanofi, and consulting fees from Verona Pharma. MDT has received support from Orion Pharma. The remaining authors have nothing to declare.

Acknowledgements

This research was conducted using the UK Biobank Resource under Application #20915. MHC was supported by R01HL149861, R01HL135142, R01HL137927, R01HL147148, and R01HL089856. HA and HJ were supported by ANR-20-CE36-0009-02. This work was supported by NHLBI U01 HL089897 and U01 HL089856. The COPDGene study (NCT00608764) is also supported by the COPD Foundation through contributions made to an Industry Advisory Committee comprised of AstraZeneca, Bayer Pharmaceuticals, Boehringer Ingelheim, Genentech, GlaxoSmithKline, Novartis, Pfizer and Sunovion. This research has been conducted as part of the INCEPTION program (Investissement d'Avenir grant ANR-16-CONV-0005). MDT is supported by Wellcome Trust Awards WT202849/Z/16/Z and WT225221/Z/22/Z. The research was partially supported by the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre. This research was funded in part, by the Wellcome Trust. BDH has received support from Alpha-1 Foundation and Bayer Pharmaceuticals. MM has received consulting fees from 2ndMD, TheaHealth, Sanofi, TriNetX, and Verona Pharma.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2025.105609.

References

1 Sakornsakolpat P, Prokopenko D, Lamontagne M, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet*. 2019;51(3):494–505.
2 Castaldi PJ, Boueiz A, Yun J, et al. Machine learning characterization of COPD subtypes: insights from the COPDGene study. *Chest*. 2020;157(5):1147–1157.
3 Rennard SI, Vestbo J. The many "small COPDs": COPD should be an orphan disease. *Chest*. 2008;134(3):623–627.
4 Aguirre M, Tanigawa Y, Venkataraman GR, Tibshirani R, Hastie T, Rivas MA. Polygenic risk modeling with latent trait-related genetic components. *Eur J Hum Genet*. 2021;29(7):1071–1081.
5 Udler MS, Kim J, von Grotthuss M, et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLoS Med*. 2018;15(9):e1002654.
6 Yaghootkar H, Scott RA, White CC, et al. Genetic evidence for a normal-weight "metabolically obese" phenotype linking insulin resistance, hypertension, coronary artery disease, and type 2 diabetes. *Diabetes*. 2014;63(12):4369–4377.
7 Ballard JL, O'Connor LJ. Shared components of heritability across genetically correlated traits. *Am J Hum Genet*. 2022;109(6):989–1006.
8 Shrine N, Guyatt AL, Erzurumluoglu AM, et al. New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet*. 2019;51(3):481–493.
9 Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *COPD*. 2010;7(1):32–43.
10 Demenais F, Margaritte-Jeannin P, Barnes KC, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat Genet*. 2018;50(1):42–53.
11 Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. 2018;50(9):1335–1341.
12 Genomes Project C, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68–74.
13 Gagliano Taliun SA, VandeHaar P, Boughton AP, et al. Exploring and visualizing large-scale genetic associations by using PheWeb. *Nat Genet*. 2020;52(6):550–552.
14 Kuhn M. Building predictive models in R using the caret package. *J Stat Software*. 2008;28(5):1–26.
15 Tan VY, Fevotte C. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(7):1592–1605.
16 Carolan BJ, Hughes G, Morrow J, et al. The association of plasma biomarkers with computed tomography-assessed emphysema phenotypes. *Respir Res*. 2014;15:127.
17 Benjamini Y, Hochberg Y. Controlling the False Discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B*. 1995;57(1):289–300.
18 Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47.
19 Watanabe K, Umicevic Mirkov M, de Leeuw CA, van den Heuvel MP, Posthuma D. Genetic mapping of cell type specificity for complex traits. *Nat Commun*. 2019;10(1):3222.
20 Serban KA, Pratte KA, Strange C, et al. Unique and shared systemic biomarkers for emphysema in Alpha-1 Antitrypsin deficiency and chronic obstructive pulmonary disease. *EBioMedicine*. 2022;84:104262.
21 Kim KW, Ober C. Lessons learned from GWAS of asthma. *Allergy Asthma Immunol Res*. 2019;11(2):170–187.
22 McCartney DL, Min JL, Richmond RC, et al. Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging. *Genome Biol*. 2021;22(1):194.
23 Crompton M, Purnell T, Tyrer HE, et al. A mutation in Nischarin causes otitis media via LIMK1 and NF-kappaB pathways. *PLoS Genet*. 2017;13(8):e1006969.
24 Poon J, Campos M, Foronjy RF, et al. Cigarette smoke exposure reduces leukemia inhibitory factor levels during respiratory syncytial viral infection. *Int J Chron Obstruct Pulmon Dis*. 2019;14:1305–1315.
25 Cloonan SM, Glass K, Laucho-Contreras ME, et al. Mitochondrial iron chelation ameliorates cigarette smoke-induced bronchitis and emphysema in mice. *Nat Med*. 2016;22(2):163–174.
26 Cordero AIH, Milne S, Yang CX, et al. Integrative omics reveal novel protein targets for chronic obstructive pulmonary disease biomarker Discovery. *medRxiv*. 2021, 2021.01.11.21249617.
27 Kim H, Westerman KE, Smith K, et al. High-throughput genetic clustering of type 2 diabetes loci reveals heterogeneous mechanistic pathways of metabolic disease. *medRxiv*. 2023;66(3):495–507.
28 Shrine N, Izquierdo AG, Chen J, et al. Multi-ancestry genome-wide association analyses improve resolution of genes and pathways influencing lung function and chronic obstructive pulmonary disease risk. *Nat Genet*. 2023;55(3):410–422.