Expanded genetic landscape of chronic obstructive pulmonary disease reveals heterogeneous cell type and phenotype associations

Phuwanat Sakornsakolpat[1,2,48], Dmitry Prokopenko[1,48], Maxime Lamontagne[3], Nicola F. Reeve[4], Anna L. Guyatt[4], Victoria E. Jackson[4], Nick Shrine[4], Dandi Qiao[1], Traci M. Bartz[5,6,7], Deog Kyeom Kim[8], Mi Kyeong Lee[9], Jeanne C. Latourelle[10], Xingnan Li[11], Jarrett D. Morrow[1], Ma'en Obeidat[12], Annah B. Wyss[13], Xiaobo Zhou[1], Per Bakke[14], R Graham Barr[15], Terri H. Beaty[16], Steven A. Belinsky[17], Guy G. Brusselle[18,19,20], James D. Crapo[21], Kim de Jong[22,23], Dawn L. DeMeo[1,24], Tasha E. Fingerlin[25,26], Sina A. Gharib[27], Amund Gulsvik[14], Ian P. Hall[28], John E. Hokanson[29], Woo Jin Kim[9], David A. Lomas[30], Stephanie J. London[13], Deborah A. Meyers[11], George T. O'Connor[31,32], Stephen I. Rennard[33,34], David A. Schwartz[25,35,36], Pawel Sliwinski[37], David Sparrow[38], David P. Strachan[39], Ruth Tal-Singer[40], Yohannes Tesfaigzi[17], Jørgen Vestbo[41], Judith M. Vonk[22,23], Jae-Joon Yim[42], Yohan Bossé[3,43], Ani Manichaikul[44,45], Lies Lahousse[46,18], Edwin K. Silverman[1,24], H. Marike Boezen[22,23], Louise V. Wain[4], Martin D. Tobin[4,47], Brian D. Hobbs[1,24,49], Michael H. Cho[1,24,49], and International COPD Genetics Consortium

1 Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA
2 Department of Medicine, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand
3 Institut universitaire de cardiologie et de pneumologie de Québec, Québec, Canada
4 Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, UK
5 Cardiovascular Health Research Unit, University of Washington, Seattle, WA, USA
6 Department of Medicine, University of Washington, Seattle, WA, USA
7 Department of Biostatistics, University of Washington, Seattle, WA, USA
8 Seoul National University College of Medicine, SMG-SNU Boramae Medical Center, Seoul, South Korea
9 Department of Internal Medicine and Environmental Health Center, School of Medicine, Kangwon National University, Chuncheon, South Korea
10 Department of Neurology, Boston University School of Medicine, Boston, MA, USA
11 Department of Medicine, University of Arizona, Tuscson, AZ
12 The University of British Columbia Center for Heart Lung Innovation, St Paul's Hospital, Vancouver, BC, Canada
13 Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC, USA
14 Department of Clinical Science, University of Bergen, Bergen, Norway
15 Department of Medicine, College of Physicians and Surgeons and Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA
16 Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, USA
17 Lovelace Respiratory Research Institute, Albuquerque, NM, USA
18 Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands
19 Department of Respiratory Medicine, Ghent University Hospital, Ghent, Belgium
20 Department of Respiratory Medicine, Erasmus Medical Center, Rotterdam, the Netherlands
21 Department of Medicine, Division of Pulmonary and Critical Care Medicine, National Jewish Health, Denver, CO, USA
22 University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, the Netherlands
23 University of Groningen, University Medical Center Groningen, Groningen Research Institute for Asthma and COPD (GRIAC), Groningen, the Netherlands

24 Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA
25 Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA
26 Department of Biostatistics and Informatics, University of Colorado Denver, Aurora, CO, USA
27 Computational Medicine Core, Center for Lung Biology, UW Medicine Sleep Center, Department of Medicine, University of Washington, Seattle, WA, USA
28 Division of Respiratory Medicine, Queen's Medical Centre, University of Nottingham, Nottingham, UK
29 Department of Epidemiology, University of Colorado Anschutz Medical Campus, Aurora, CO, USA
30 University College London, London, UK
31 The National Heart, Lung, and Blood Institute's Framingham Heart Study, Framingham, MA, USA
32 Pulmonary Center, Department of Medicine, Boston University School of Medicine, Boston, MA, USA
33 Pulmonary, Critical Care, Sleep and Allergy Division, Department of Internal Medicine, University of Nebraska Medical Center, Omaha, NE, USA
34 Clinical Discovery Unit, AstraZeneca, Cambridge, UK
35 Department of Medicine, School of Medicine, University of Colorado Denver, Aurora, CO, USA
36 Department of Immunology, School of Medicine, University of Colorado Denver, Aurora, CO, USA
37 2nd Department of Respiratory Medicine, Institute of Tuberculosis and Lung Diseases, Warsaw, Poland
38 VA Boston Healthcare System and Department of Medicine, Boston University School of Medicine, Boston, MA, USA
39 Population Health Research Institute, St. George's University of London, London, UK
40 GSK R&D, King of Prussia, PA, USA
41 School of Biological Sciences, University of Manchester, Manchester, UK
42 Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, Seoul National University College of Medicine, Seoul, South Korea
43 Department of Molecular Medicine, Laval University, Québec, Canada
44 Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
45 Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA
46 Department of Bioanalysis, Ghent University, Ghent, Belgium
47 National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester, UK
48 These authors contributed equally
49 These authors jointly supervised the work

Correspondence should be addressed to M.H.C. (remhc@channing.harvard.edu).

Summary

Chronic obstructive pulmonary disease (COPD) is the leading cause of respiratory mortality worldwide. Genetic risk loci provide novel insights into disease pathogenesis. To broaden COPD genetic risk loci discovery and identify cell type and phenotype associations we performed a genome-wide association study in 35,735 cases and 222,076 controls from the UK Biobank and additional studies from the International COPD Genetics Consortium. We identified 82 loci with P value < $5x10^{-8}$; 47 were previously described in association with either COPD or population-based lung function. Of the remaining 35 novel loci, 13 were associated with lung function in 79,055 individuals from the SpiroMeta consortium. Using gene expression and regulation data, we identified enrichment for loci in lung tissue, smooth muscle and alveolar type II cells. We found 9 shared genomic regions between COPD and asthma and 5 between COPD and pulmonary fibrosis. COPD genetic risk loci clustered into groups of quantitative

imaging features and comorbidity associations**.** Our analyses provide further support to the genetic susceptibility and heterogeneity of COPD.

Background

Chronic obstructive pulmonary disease (COPD) is a disease of enormous and growing global burden[1], ranked third as a global cause of death by the World Health Organization in 2016[2]. Environmental risk factors, predominately cigarette smoking, account for a large fraction of disease risk, but there is considerable variability in COPD susceptibility among individuals with similar smoking exposure. Studies in families and in populations demonstrate that genetic factors account for a substantial fraction of disease susceptibility. Similar to other adult-onset complex diseases, common variants likely account for the majority of population risk[3,4]. Our previous efforts identified 22 genome-wide significant loci[5]. Expanding the number of risk loci can lead to novel disease pathogenesis insights not only through discovery of novel biology[6,7] but also through informing more global insights such as functional links between loci and cell-type and phenotype identification driving COPD genetic risk[5].

We performed a genome-wide association study including previously described studies from the International COPD Genetics Consortium (ICGC) with additional subjects from UK Biobank[8], a population-based study of several hundred thousand subjects with lung function and cigarette smoking assessment. We determined, through bioinformatic and computational analysis, the likely set of variants, genes, cell types, and biologic pathways implicated by these associations. Finally, we assessed our genetic findings for relevance to COPD-specific, respiratory, and other phenotypes.

Results

Genome-wide association study of COPD

We included a total of 257,811 individuals from 25 studies in the analysis, including studies from International COPD Genetics Consortium and UK Biobank. We defined COPD based on pre-bronchodilator spirometry using pre-bronchodilator spirometry according to modified Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria for moderate to very severe airflow limitation[9], resulting in 35,735 cases and 222,076 controls (**Supplementary Tables**

See the Excel file.

Supplementary Table 1). We tested association of COPD and 6,224,355 variants in a meta-analysis of 25 studies using a fixed-effects model. We found no evidence of confounding by population substructure using linkage disequilibrium score regression (LDSC) intercept (1·0377, s.e. 0·0094).

We identified 82 loci (defined using 2-Mb windows) at genome-wide significance ($P < 5 \times 10^{-8}$) (**Figure 2**). Forty-seven of 82 loci were previously described as genome-wide significant in COPD[5,10–12] or lung function[13,14,23,15–22] (**Supplementary Table 2**), leaving 35 novel loci (**Table 1**). We then sought to replicate these novel loci. Given the strong genetic correlation between population-based lung function and COPD, we tested the lead variant at each for association with FEV_1 or FEV_1/FVC in 79,055 individuals from SpiroMeta. We identified 13 loci - *C1orf87, DENND2D, DDX1, SLMAP, BTC, FGF18, CITED2, ITGB8,*

STN1, ARNTL, SERP2, DTWD1, and *ADAMTSL3* that replicated using a Bonferroni correction for a one-sided P < 0·05/35; **Table 1**). Although not meeting the strict Bonferroni threshold, additional 14 novel loci were nominally significant in SpiroMeta (consistent direction of effect and one-sided P < 0·05): *ASAP2, EML4, VGLL4, ADCY5, HSPA4, CCDC69, RREB1, ID4, IER3, RFX6, MFHAS1, COL15A1, TEPP,* and *THRA* (**Table 1**). In our overall meta-analysis, all 82 of the genome-wide significant loci showed consistent direction of effect with either FEV_1 or FEV_1/FVC ratio in SpiroMeta (**Table 1** and **Supplementary Table 2**). We note that 9 of our 35 novel loci were recently described in a contemporaneous analysis of lung function in UK Biobank[24]. None of the novel loci appeared to be due to cigarette smoking (**Supplementary Results**). Including all 82 genome-wide significant variants, we explain up to 7·0 % of the phenotypic variance in liability scale, using a 10% prevalence of COPD, and acknowledging that these effects are likely overestimated in the discovery sample. This represents a 48% increase in COPD phenotypic variance explained by genetic loci compared to the 4·7% explained by 22 loci reported in a recent GWAS of COPD[5].

Identification of secondary association signals

We then used approximate conditional and joint analysis to find secondary signals at each of the 82 genome-wide significant loci. We found 82 secondary signals at 50 loci, resulting in a total of 164 independent associations in 82 loci (**Supplementary Table 3**). Of 50 loci containing secondary associations, 33 were at loci previously described for COPD or lung function, and 6 at Bonferroni-replicated novel loci. Of 82 secondary associations, 20 reached genome-wide significance ($P < 5 \times 10^{-8}$) (**Supplementary Table 3**). Of 61 novel (not previously described in COPD or lung function) independent associations, 21 reached a region-wise Bonferroni-corrected threshold (one-sided P < 0.05/novel independent association(s) in each locus) in unconditioned associations from SpiroMeta (**Methods** and **Supplementary Table 3**).

Tissue and specific cell types

In determining the tissue in which COPD genetic variants function to increase COPD risk, lung is the obvious tissue to consider; however, COPD is a systemic disease[25,26] and within the lung, the specific cell-types underlying disease pathogenesis are largely unknown. Furthermore, available databases often include cell types (e.g. smooth muscle) from non-lung organs (e.g. the gastrointestinal tract). To identify putative causal tissues and cell types, we assessed the enrichment of our genome-wide significant COPD loci in integrated genome annotations at the single tissue level[27], tissue-specific epigenomic marks[28], and genome-wide gene expression patterns[29]. Lung tissue showed the most significant enrichment (OR 9·25, $P=1{\cdot}36 \times 10^{-9}$), as previously described, though significant enrichment was also seen in heart (OR 6·85, $P=3{\cdot}83 \times 10^{-8}$) and the gastrointestinal (GI) tract (OR 5·53, $P=6{\cdot}45 \times 10^{-11}$). In an analysis of enriched epigenomic marks, the most significant enrichment was in fetal lung and GI smooth muscle DNase hypersensitivity sites (DHS) ($P= 6{\cdot}75 \times 10^{-8}$) and H3K4me1 ($P= 7{\cdot}31 \times 10^{-7}$) (**Supplementary Table 4**). To further identify lung-specific cell types, we tested whether 47 known and 13 novel COPD-associated loci contain genes specifically expressed in data sets from single-cell RNA-seq. We found enrichment of alveolar type II and basal cells (P=0·016 and P=0·042, respectively) using single-cell RNA-seq gene expression data from respiratory cell types[30] (**Supplementary Figure 3** and **Supplementary Table 5** for individual locus results).

Fine-mapping of associated loci

To identify the most likely causal variants, we performed fine mapping using Bayesian credible sets[31]. Including 160 potential primary and secondary association signals (excluding four variants in the major histocompatibility complex (MHC) region), 61 independent signals had a 99% credible set with fewer than 50 variants; 34 signals had credible sets with fewer than 20 variants (**Supplementary Figure 4**). Only the association signal at *NPNT* (4q24) could be fine-mapped to single variant; however, in 17 other loci, a single variant had posterior probability of driving association (PPA) greater than 60% (**Supplementary Table 6**). Most sets included variants that overlapped genic enhancers of lung-related cell types (e.g., fetal lung fibroblasts, fetal lung, and adult lung fibroblasts) and were predicted to alter transcription binding motifs (**Supplementary Table 6**). Of 61 credible sets with fewer than 50 variants, eight sets contained at least one deleterious variant. These deleterious variants included 1) missense variants affecting *TNS1, RIN3, GPR126, ADAM19, ATP13A2, BTC*, and *CRLF3*; and 2) a splice donor variant affecting a lincRNA - AP003059.2.

Candidate target genes

In most cases, the closest gene to a lead SNP will not be the gene most likely to be the causal or effector gene of disease-associated variants[32–34]. Thus, to identify the potential effector ('target') genes underlying these genetic associations, we integrated additional molecular information including gene expression; gene regulation (open chromatin and methylation data), chromatin interaction, co-regulation of gene expression with gene sets and coding variant data (**Methods** and **Figure 3a**).

At 82 loci, 472 genes were implicated by analysis of least one dataset; 106 genes were implicated by gene expression (Bonferroni corrected at locus level), and an additional 50 genes by >= 2 other datasets (methylation, chromatin interaction, open chromatin regions, similarity in gene sets or deleterious coding variants (**Figure 3a**)), for a total of 156 genes meeting more strigent criteria. Excluding loci in the extended MHC region, the median numbers of potentially implicated genes per locus was four with a maximum of 17 genes (7q22.1 and 17q21.1). The median distance of implicated genes to top associated variants was 346 Kb, restricting to genes +/-1 Mb of top associated variants. Among 82 loci, 60 loci (73%) included the nearest gene. We identified 20 genes which were region-wise Bonferroni significant in exome sequencing data. Two genes (*ADAM19* and *ADAMTSL3*) were implicated by five datasets (**Figure 3b**) and another two (*EML4* and *RIN3*) were implicated by four datasets. A summary of all genes implicated using these approaches in **Supplementary Table 7**.

Associated pathways

To gain further functional insight of associated genetic loci, we performed gene-set enrichment analysis using DEPICT. Among 165 enriched gene sets at FDR < 5%, 44% of them were related to the developmental process term, with lung development $P = 1.02 \times 10^{-6}$; significant sub-terms included lung alveolus development (nominal P= 0·0003) and lung morphogenesis (nominal P= 0·0005). We also found enrichment of extracellular matrix -related pathways including laminin binding, integrin binding, mesenchyme development, cell-matrix adhesion, and actin filament bundles. Additional pathways of note included histone deacetylase binding, Wnt receptor signaling pathway, SMAD binding, the MAPK cascade, and transmembrane receptor protein serine/threonine kinase signaling pathway. Full enrichment analysis results including the top genes for each DEPICT gene set are shown in **Supplementary Table 8**.

Phenotypic effects of known and novel associations for COPD

To characterize the phenotypic effects of 82 genome-wide significant loci, we performed a phenome-wide association analysis within the deeply phenotyped COPDGene study[35]. We tested the overall structure of associated phenotypes by using hierarchical clustering across scaled Z scores of associations. We identified two clusters of associated variants (**Supplementary Figure 5**). As these two clusters appeared to predominantly differentiate among imaging features, we repeated variant clustering limited to quantitative computed tomography (CT) imaging features. We found two clusters of variants, differentiated by association with quantitative emphysema, emphysema distribution, gas trapping, and airway phenotypes (**Figure 4a**). We then evaluated the association of the 82 genome-wide significant variants in a prior GWAS of emphysema and airway quantitative CT features[36] (**Supplementary Table 10**).

We also examined all genome-wide significant loci in the NHGRI-EBI GWAS Catalog[37] (**Supplementary Figure 6 and Supplementary Table 11**) and looked for variants in linkage disequilibrium ($r^2 > 0{\cdot}2$) with the lead GWAS variant. Many variants associated with anthropometric measures including height and body mass index (BMI), measurements on blood cells (red and white cells), and cancers. COPD is well known for having common comorbidities, such as coronary artery disease (CAD), type 2 diabetes mellitus (T2D), bone density, and lung cancer. Of these diseases, we only found evidence of modest overall genetic correlation (using linkage disequilibrium score regression) between COPD and lung cancer (**Supplementary Results**). However, at individual loci, and using more stringent linkage disequilibrium ($r^2 > 0{\cdot}6$), we found evidence of shared risk factors for COPD, including a genome-wide significant variant near *PABPC4* associated with T2D, four variants with CAD (near *CFDP1, DMWD, STN1*, and *TNS1*), and a variant near *SPPL2C* with bone density (**Figure 4b**).

Identification of loci overlapping with asthma and pulmonary fibrosis

Based on our previous identification of genetic overlap of COPD with asthma, and COPD with pulmonary fibrosis, we also examined loci for specific overlap with these two diseases. In asthma, we noted an $r^2 > 0.2$ with one of our variants, and previously reported variants and *ID2, ZBTB38, C5orf56, MICA, AGER, HLA-DQB1, ITGB8, CLEC16A,* and *THRA.* In pulmonary fibrosis, in addition to our previously described overlap between *FAM13A*, *DSP*, and 17q21, we noted *ZKSCAN1* and *STN1* (**Supplementary Table 11**). To more closely examine overlap, applied a Bayesian method (gwas-pw) of COPD associations from our current GWAS with previous GWASs of asthma (limited to those of European ancestry) and pulmonary fibrosis[38,39]. To mitigate the effect of including asthma cases in the GWAS of COPD, we excluded individuals with self-reported asthma from the UK Biobank (**Methods** and **Supplementary Results**). We identified 14 shared genome segments (posterior probability > 70%), nine with asthma and five with pulmonary fibrosis (**Figure 4c** and **Supplementary Table 9**). Of nine segments shared with asthma, five segments reside within the MHC region (6p21-22). Non-MHC segments included loci near *ADAM19*, *ARMC2*, *ELAVL2*, and *STAT6*. Of five segments shared with pulmonary fibrosis, two segments were identified including loci near *ZKSCAN1*, *STN1* (formerly known as *OBFC1),* in addition to the three segments identified in the previous study[5] (*FAM13A*, *DSP,* and the 17q21 locus, here *CRHR1*). For all overlapping loci between COPD and asthma, overlapping variants had the same direction of effect (i.e., increasing risk for both COPD and asthma). Conversely, shared variants between COPD and pulmonary fibrosis all had an opposite effect (i.e., increasing risk for COPD but protective for pulmonary fibrosis).

Discussion

Genetic factors play an important role in COPD susceptibility. We examined genetic risk of COPD in a genome-wide association study including a total of 35,735 cases and 222,076 controls. We identified 82 genome-wide significant loci for COPD, of which 47 were previously identified in genome-wide association studies of COPD, or population-based lung function. Of 35 loci not previously described, 13 replicated in an independent study of population-based lung function. Our results identify important disease pathways and may further explain the clinical heterogeneity seen in COPD.

Our study supports the role of early life events in the risk of COPD. Gene set enrichment analysis on our putative causal genes identified lung morphogenesis and lung alveolar development, the canonical Wnt receptor signaling pathway[40,41], the MAPK cascade, Ras protein signal transduction, and the nerve growth factor receptor signaling pathway. Further, the importance of gene regulation at fetal stages was highlighted through enrichment of heritability in epigenomic marks of various fetal tissues, with fetal lung showed the strongest signals. Our findings are consistent with recent epidemiologic studies demonstrating that a substantial portion of the risk of COPD may be develop in early life: genetic variants may set initial lung function[42] and patterns of growth[42–44].

We also identified several genes and pathways of interest not primarily related to lung development, some of which have been previously identified in studies of lung function[13], including mesenchyme development and extracellular matrix, cilia structure, elastin-associated microfibrils, and retinoic acid receptor beta[45–47]. We used several data sources to attempt to assign causal gene at each locus, identifying 156 genes at 82 loci that were supported by either gene expression or a combination of at least 2 other data sources. One of our genes with the most support was *ADAMTSL3.* In addition to a role in development, this gene plays a role in cell-matrix interactions or in assembly of specific extracellular matrices[48]. Another novel finding was an association with the chitinase acidic (*CHIA*) gene at 1p13.3, which encodes a protein that degrades chitin[49] and exhibits lung-specific expression[50,51]. *CHIA* variants have been associated with FEV_1[52], asthma[53–56], and acid mammalian chitinase activity[55,57]. Its role in airway inflammation was demonstrated in an animal model of asthma[58]. Interleukin 17 receptor D (*IL17RD*) at 3p14.3 encodes a membrane protein belonging to the interleukin-17 receptor (IL-17R) protein family[49]. The gene product affects fibroblast growth factor signaling, inhibiting or stimulating growth through MAPK/ERK signaling[49]. It also interacts with TNF receptor 2 (TNFR2) to activate NF-κB[59]. Integrin subunit beta 8 (*ITGB8*) at 7p21.1 is a member of the integrin beta chain family and ITGB8 protein expression protein is increased in COPD[60–62]. This locus was also recently described in a separate study of allergic disease and asthma[63]. The *ITGB8* gene and encodes a single-pass type I membrane protein that binds to an alpha subunit to form an integrin complex[49]. The complex mediates cell-cell and cell-extracellular matrix interactions and plays a role in human airway epithelial proliferation[49] and repair[64].

In addition to identifying the effector gene, the effector cell type is of critical important for functional studies. We identified an overall enrichment of epigenomic marks in lung tissue and smooth muscle (also identified in studies of lung function[21]); this latter association is driven by non-respiratory cell types. Within a set of four lung cell types identified by single cell RNA-Seq, we identified enrichment for alveolar type 2 (AT2) cells, which recently have been shown to have regenerative properties[65]. The lung is comprised at least 40 different resident cell types[66]; thus, while our findings suggest cell types for further functional studies, they also highlight the need for profiling of additional lung cell types.

Characterization of genome-wide significant associations revealed heterogenous effect to COPD-related phenotypes and other biological processes. Within the well-phenotyped COPDGene cohort, we identified variable effects of these variants on computed tomography (CT) features, smoking status and intensity, diffusing capacity of carbon monoxide, asthma, and inflammatory biomarkers. Clustering these variants found differential effects on emphysema and airway phenotypes, a well-described source of heterogeneity in COPD[67–69]. Analyzing over hundreds of diseases/traits in GWAS Catalog, we identified overlapping associations with various diseases/traits in multiple organ systems, comorbidities such as coronary artery disease, bone mineral density, and type 2 diabetes mellitus. Together, the identification of variable COPD risk loci associations with sub-phenotypes and other diseases[70,71] may have potential for more nuanced approaches to therapy for COPD.

We performed additional specific analysis in two diseases that overlap with COPD, asthma and pulmonary fibrosis. Extending from previously reported genetic correlations[5], we identified 14 genetic loci shared between these pairs of diseases. Our analysis is the first to identify evidence for shared genetic segments between asthma and COPD. We identified multiple overlapping loci with asthma at the MHC region (6p21-22) and four other loci. The locus near *DDX1* was previously suggested to be associated with both COPD and asthma in a combined meta-analysis[72], and is involved in NFkB pathway[73], leading to a possible shared inflammatory mechanism between these diseases. In addition to three previously reported overlapping loci for COPD and IPF (*FAM13A* and *DSP,* both genome-wide significant, and 17q21, identified through Bayesian overlap analysis), we identified two loci near *ZKSCAN1* (7q22.1) and *STN1* (previously known as *OBFC1*, 10q24). The top associated variant near *STN1* is in linkage disequilibrium with a CAD-associated variant (rs12765878, r^2=0·61). This locus has also been associated with leukocyte telomere length[74], and lends additional support to a role of telomere maintenance as a common risk factor for these three diseases. Overall, our phenotype, gene-, and pathway- analyses illustrate the utility of both searching for enrichment of genetic signal overall and performing a more detailed identification of the effects of individual variants or groups of variants.

While our study is the largest genome-wide study of COPD, individuals meeting criteria for COPD in the UK Biobank may be different from other smoker-enriched studies, especially for smoking history. In addition, our use of population-based lung function for replication, along with pre-bronchodilator spirometry, could bias our findings against variants that are only associated with more severe forms of COPD. However, we observed concordant effect size estimates when including or excluding asthmatics. Thus, together with prior analyses[5], our findings suggest that bias due to COPD case misclassification is likely small. However, we cannot rule out a role for studying more severe disease. We note that the alpha-1 locus (*SERPINA1*) was identified as genome-wide significant in smaller studies of emphysema and in severe COPD in smokers. In the current study, the association of the PIZ allele had $P = 2{\cdot}2 \times 10^{-5}$ using moderate-to-severe cases, and a lower P-value ($1{\cdot}4 \times 10^{-6}$) in severe ($FEV_1 < 50\%$ predicted) cases despite a smaller sample size, an effect we have previously described[75]. Thus, despite the strong overlap of COPD with quantitative spirometry, new loci may be identified through studies of sufficiently large subsets of COPD patients with severe COPD or more specific and homogenous COPD phenotypes. Given suggestive evidence for replication using a related (but not identical) phenotype for additional novel loci beyond the 13 meeting Bonferroni, we chose to include all loci significant in discovery in subsequent analyses, recognizing that we likely included some false positive associations. Our study focused on relatively common variants, predominantly in individuals of European ancestry; more detailed studies of

rare variants, HLA regions, and other ethnicities are warranted, but broader multi-ethnic analyses are limited by the number of cases in currently available cohorts.

The global burden of COPD is increasing. Our work finds a substantial number of new loci for COPD, and uses multiple lines of supportive evidence to identify potential genes and pathways for both existing and novel loci. Further investigation of genetic overlap and phenotypic effects finds new shared loci for asthma and idiopathic pulmonary fibrosis, and suggests heterogeneity across COPD-associated loci. Together, these insights provide multiple new avenues for investigation for this deadly disease.

Methods

Study populations

The UK Biobank is a population-based cohort consisting of 502,682 individuals[8]. To determine lung function, we used measures of forced expiratory volume in 1 second (FEV_1) and forced vital capacity (FVC) derived from the spirometry blow volume-time series data, subjected to additional quality control based on ATS/ERS criteria[76] (**Supplementary Methods**). We defined COPD in European-ancestry subjects using two pre-bronchodilator measurements of lung function according to modified Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria for moderate to very severe airflow limitation[9]: FEV_1 less than 80% of predicted value (using reference equations from Hankinson et al.[77]), and the ratio between FEV_1 and FVC less than 0.7. Genotyping was performed using Axiom UK BiLEVE array and Axiom Biobank array (Affymetrix, Santa Clara, California, USA) and imputed to the Haplotype Reference Consortium (HRC) version 1.1 panel[78].

We invited participants in the prior International COPD Genetics Consortium (ICGC) COPD genome-wide association study to provide case-control association results (with the exception of the 1958 British Birth Cohort, to avoid overlapping samples with the replication sample). ICGC cohorts performed case-control association analysis based on pre-bronchodilator measurements of FEV_1 and FEV_1/FVC, and cases were identified using modified GOLD criteria, as above. Studies were imputed to 1000 Genomes reference panels. Detailed cohort descriptions and cohort-specific methods have been previously published[5] (**Supplementary Methods**).

Based on the strong genetic overlap of lung function and COPD[5], we performed lookups of select significant variants for FEV_1 and FEV_1/FVC in the SpiroMeta consortium meta-analysis[24]. Briefly, SpiroMeta comprised of a total of 79,055 individuals from 22 studies imputed to either the 1000 Genomes Project Phase 1 reference panel (13 studies) or the HRC (9 studies). Each study performed linear regression adjusting for age, age^2, sex, and height, using rank-based inverse normal transforms, adjusting for population substructure using principal components or linear mixed models, and performing separate analyses for ever- and never- smokers or using a covariate for smoking (for studies of related subjects). Genomic control was applied to individual studies, and results were combined using fixed-effects meta-analysis[24].

Genome-wide association analysis

In UK Biobank, we performed logistic regression of COPD, adjusting for age, sex, genotyping array, smoking pack-years, ever smoking status, and principal components of genetic ancestry. Association analysis was done using PLINK 2.0 alpha[79] (downloaded on December 11, 2017) with Firth-fallback settings, using Firth regression when quasi-complete separation or regular-logistic-regression convergence failure occurred. We performed a fixed-effects meta-analysis of all ICGC cohorts and UK

Biobank using METAL (version 2010-08-01)[80]. We assessed population substructure and cryptic relatedness by linkage disequilibrium (LD) score regression intercept[81]. We defined a genetic locus using a 2-Mb window (+/-1 Mb) around a lead variant, with conditional analyses as described below.

To maximize our power to identify existing and discover new loci, we examined all loci at the genome-wide significance value of $P < 5 \times 10^{-8}$. We first characterized loci as being previously described (evidence of prior association with lung function[13,14,23,15–22] or COPD[5,10–12]) or novel. We defined previously reported signals if they were in the same LD block in Europeans[82] and in at least moderate LD ($r^2 >= 0{\cdot}2$). For novel loci we attempted replication through association of each lead variant with either FEV_1 or FEV_1/FVC ratio in SpiroMeta, using one-sided p values with Bonferroni correction for the number of novel loci examined. Novel loci failing to meet a Bonferroni-corrected p value were assessed for nominal significance (one-sided $p < 0{\cdot}05$) or directional consistence with FEV_1 and FEV1/FVC ratio in SpiroMeta.

Cigarette smoking is the major environmental risk factor for COPD and genetic loci associated with cigarette smoking have been reported[5,83]. While we adjusted for cigarette smoking in our analysis, we further examined these effects by additionally testing for association of each locus with cigarette smoking and by looking at two separate analyses of ever- and never- smokers in UK Biobank.

Identification of independent associations at genome-wide significant loci

We identified specific independent associations at genome-wide significant loci using GCTA-COJO[84]. This method utilizes an approximate conditional and joint analysis approach requiring summary statistics and representative LD information. As the UK Biobank provided the predominant sample, we used 10,000 randomly drawn unrelated individuals from this discovery dataset as a LD reference sample. We scaled genome-wide significance to a 2-Mb region, resulting in a locus-wide significant threshold of 8×10^{-5}, or 2×10^{-6} for variants in the major histocompatibility complex (MHC) region (chr6:28477797-33448354 in hg19[85]). We created regional association plots via LocusZoom using 1000 Genomes EUR reference data (Nov2014 release)[86].

Identification and prioritization of tissues and cell types, candidate variants, genes, and pathways

Identification of enriched tissues and specific cell types

We used LD Score Regression (LDSC)[87] to estimate the enrichment of functional annotation in disease heritability. We utilized LDSC baseline models (e.g., conserved region, promoter flanking region), tissue-specific annotations from the Roadmap Epigenomics Program[28] and GenoSkyline[27]. We also used SNPsea[29] to estimate the enrichment of specific cell types from genome-wide significant associations using gene expression data (**Supplementary Methods**). For SNPsea, we used a single-cell RNA-seq dataset from the study of idiopathic pulmonary fibrosis (540 cells from 6 IPF lung samples and 3 control tissues, available at Gene Expression Omnibus as GSE86618)[30] (**Supplementary Methods**).

Fine-mapping of independent association signals at genome-wide significant loci

We used Bayesian fine-mapping at each locus to identify the credible set: the set of variants with a 99% probability of containing a causal variant. Briefly, for each genome-wide significant loci we calculated approximate Bayes factors[31] of association. We then selected variants in each locus, so that their cumulative posterior probability was equal or greater than 0·99 using an unscaled variance. At loci with multiple independent associations, we used statistics from approximate conditional analysis with GCTA software on each index variant adjusting for other independent variants in the loci. Otherwise, we used

unconditioned statistics from our meta-analysis. We characterized variant effects in credible sets using variant annotations from Ensembl Variant Effect Predictor[88].

Identification of target genes

We used several computational approaches with corresponding available datasets to identify target genes in genome-wide significant loci. We used two methods that utilized gene expression data: 1) S-PrediXcan and 2) DEPICT. We used S-PrediXcan[89] to identify genes with genetically regulated expression associated with COPD. We used data from the Lung-eQTL consortium[90,91] (1,038 lung tissue samples) as an eQTL and gene expression reference database. S-PrediXcan is the extension of PrediXcan[92] that test for association between a trait and imputed gene expression using summary statistics. Here, we performed S-PrediXcan using models for protein-coding genes +/- 1 Mb from top-associated variants at genome-wide significant loci. We used DEPICT (Data-driven Expression Prioritized Integration for Complex Traits) [93] to prioritize genes from 'reconstituted' gene sets.

We also used additional information on gene regulation, including epigenetic data: 1) regulatory fine mapping, 2) mQTL, and 3) chromosome conformation capture. We used regulatory fine mapping (regfm[94]) to overlap 99% credible interval (CI) variants at each GWAS locus with open chromatin regions based on DNAse hypersensitivity sites (DHS). DHS cluster accessibility state was then associated with gene expression levels (for 13,771 genes) from 22 tissues in the Roadmap Epigenomics Project[95]. Using both the 99% CI and DHS overlap, as well as the DHS state and transcript level association, regfm calculates a posterior probability of association of each gene +/- 1 Mb of the lead SNP at each GWAS locus. We also searched for overlapping methylation quantitative trait loci (mQTL) data from lung tissue, as recently described[96]. To determine whether these signals co-localized (rather than being related due to linkage disequilibrium), we performed colocalization analysis between our GWAS and mQTL in genome-wide significant loci using eCAVIAR[97] (eQTL and GWAS CAusal Variants Identification in Associated Regions, **Supplementary Methods**). We also sought information from publicly available chromosome conformation capture data[98]. We obtained statistics of high-confidence chromatin interaction in fetal lung fibroblast cell line (IMR90) and human lung tissue[98] through HUGIn[99] (Hi-C Unifying Genomic Interrogator). We anchored long range chromatin interactions on top associated variants and computed statistical significance of contact at each locus. We retained only the strongest associations (i.e., smallest P value) for each cell line/primary cell in the analysis.

Finally, we searched for signals from non-synonymous variants. We identified coding variants present in the credible set in the GWAS with a high posterior probability. We also searched for rare coding variants, based on exome sequencing results in the COPDGene, Boston Early-Onset COPD, and International COPD Genetics Network studies. In brief, we performed exome sequencing on 485 severe COPD cases and 504 smoking resistant controls from the COPDGene study and 1,554 subjects ascertained through 631 probands with severe COPD from the Boston Early-Onset COPD study (BEOCOPD) and the ICGN study. We performed single-variant analyses using Firth and efficient resampling methods (SKAT R package[100]) for the COPDGene data (case-control) and generalized linear mixed models (GMMAT) for the BEOCOPD-ICGN data (using lung function). Gene-based analyses were conducted using Burden, SKAT, and SKAT-O tests with asymptotic and efficient resampling methods (SKAT package) combined with Fisher's method for the COPDGene data, and using SKAT-O tests (MONSTER) for the BEOCOPD-ICGN data. Two variant-filtering criteria were considered: deleterious variants (predicted by FATHMM) with MAF < 0·01, and functional variants (moderate effect predicted by SNPEff) with MAF < 0·05. Gene-based

segregation test (GESE) was also applied to the ultra-rare (MAF < 0·1%) and loss-of-function variants in the BEOCOPD-ICGN data on the severe COPD affection status.

For each dataset described above, we used Bonferroni-corrected P values, or a fixed posterior probability threshold to determine target genes at each locus. To correct for possible number of genes in each locus, we obtained a list of protein-coding genes +/-1 Mb from a top associated variant from BioMart[88]. For each locus, we used a 5% Bonferroni-corrected threshold (i.e., P < 0·05 divided by number of genes at that locus) to determine significance for 4 data types: gene expression data, chromatin conformation capture data, co-regulation of gene expression, and exome sequencing results. For two remaining datasets, we used a fixed posterior probability (of gene association with a GWAS locus) threshold of 0·1 for regfm and eCAVIAR. We considered genes that were implicated by gene expression or >= 2 combination of other datasets (e.g., methylation and chromatin conformation capture data) as target genes.

Identification of pathways

To identify enriched pathways in COPD-associated loci, we performed gene-set enrichment analysis using the "reconstituted" genes sets from DEPICT, as described above[93]. We defined significant gene sets using false discovery rate (FDR) < 5%.

Effects on COPD-related and other phenotypes

COPD is a complex and heterogeneous disorder, comprised of different biologic processes and specific phenotypic effects. In addition, many loci discovered by GWAS have pleiotropic effects. To identify these effects, we performed analyses of a) identification of overlapping genetic loci between related disorders (asthma and pulmonary fibrosis) b) genetic association studies of our genome-wide significant findings using COPD-related phenotypes, including a cluster analysis to identify groups of variants that may be acting via similar mechanisms; c) look up of top variants in prior COPD-related quantitative computed tomography (CT) imaging feature GWAS, and d) look up of associations with other diseases/traits using GWAS Catalog.

To identify overlapping loci between COPD and other respiratory disorders, we used gwas-pw[101] to perform pairwise analysis of GWAS. This method searches for shared genomic segments[82] using adaptive significance threshold, allowing detection of sub genome-wide significant loci. We identified shared segments or variants using posterior probability of colocalizing greater than 0·9[101]. We obtained GWAS summary statistics from two previous studies of pulmonary fibrosis[39] and GWAS of asthma in Europeans[38]. To assess heterogenous effects of COPD susceptibility loci on COPD-related features (phenotypes), we evaluated associations of our genome-wide significant SNPs with 121 detailed phenotypes (e.g., lung function, computed tomography-derived metrics, biomarkers, and comorbidities) available in 6,760 COPDGene non-Hispanic whites. We calculated Z-scores for each SNP-phenotype combination relative to the COPD risk allele to create a SNP by phenotype Z-score matrix. We tested each COPD-related phenotype with at least one nominally significant association with one of our genome-wide significant COPD SNPs, leaving us with 107 phenotypes. We then oriented all Z-scores to be positive (based on sign of median Z score) in association with each phenotype to avoid clustering based on direction of association. To avoid clustering phenotypes only by strength of association with SNPs, we scaled Z-scores within each phenotype by subtracting mean Z-scores and dividing by the standard deviation of Z-scores within each phenotype. We then scaled Z-scores across SNPs to circumvent clustering of SNPs according only to relative strength of association with phenotypes. We

then performed hierarchical clustering of the scaled Z-scores of associations between SNPs and phenotypes to identify clusters of SNPs and phenotypes for all 107 phenotypes as well as in the subset of 26 quantitative imaging phenotypes. We identified optimal number of clusters using the Calinski index[102,103]. To identify features that independently predict cluster membership, we fitted a logistic regression model via penalized maximum likelihood using the glmnet package[104]. We determined optimal regularization parameters using 10-fold cross validation. We further examined top variant associations with COPD-related traits through a look-up of top variants in a prior GWAS of 12,031 subjects with quantitative emphysema and airway CT features[105]. To examine overlap of our COPD results with other traits, we downloaded genome-wide significant associations from the GWAS Catalog[37,106] ($P < 5 \times 10^{-8}$). Between a pair of COPD- and trait- associated variants within the same LD block in Europeans[82], we computed the LD using the European ancestry panel and considered the overlap if variants were in at least in moderate LD ($r^2 >= 0{\cdot}2$).

Figures

Figure 1 Study design

COPD, chronic obstructive pulmonary disease; FEV_1, force expiratory volume in one second; FVC, forced vital capacity.

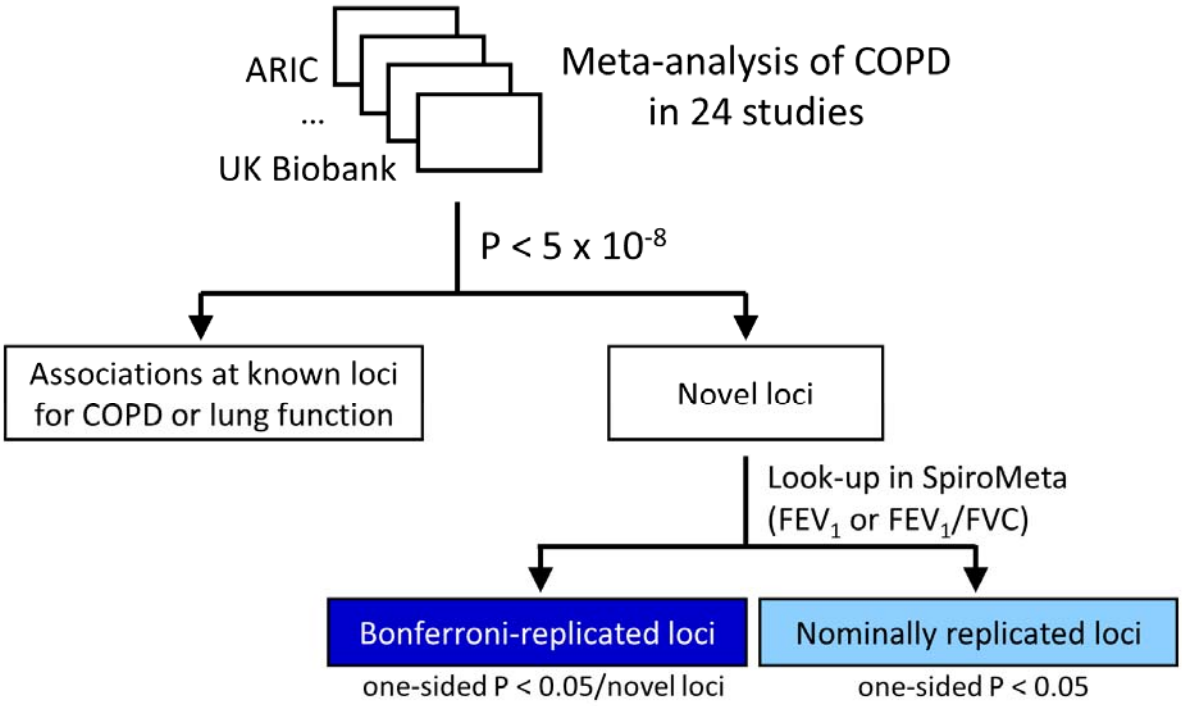


Figure 2 Manhattan plot

Loci are labeled with the closest gene to the lead variant. Colors indicates variants at novel loci which replicated using Bonferroni-corrected threshold in SpiroMeta (dark blue, one sided P < 0.05/35) and nominally significant threshold (light blue, one-sided P < 0.05).

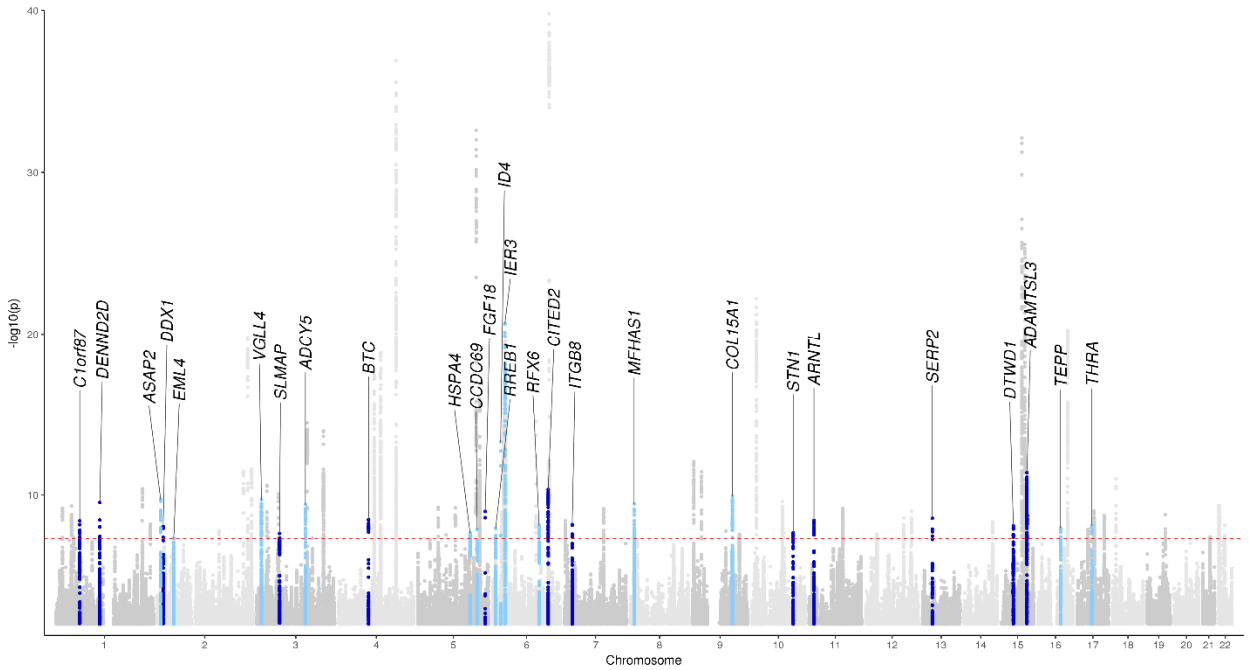
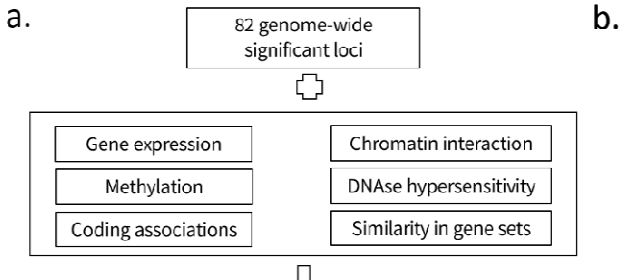


Figure 3 Identification of target genes

(a) Overview of datasets used to identify target genes at genome-wide significant loci (b) Regional association plots at *ADAMTSL3* locus showing GWAS (top), chromatin interaction in lung tissue (middle) and expression quantitative trait loci (bottom).

a.



Locus	Gene	GREx	mQTL	Cod.	Hi-C	DHS	GSet
1p13	*DRAM2*	●					
1p13	*OVGP1*		●	●	●	●	
6q24	*HECA*				●	●	
15q25	*ADAMTSL3*	●	●		●	●	●
15q25	*BNC1*	●				●	●
15q25	*HOMER2*					●	●

b.

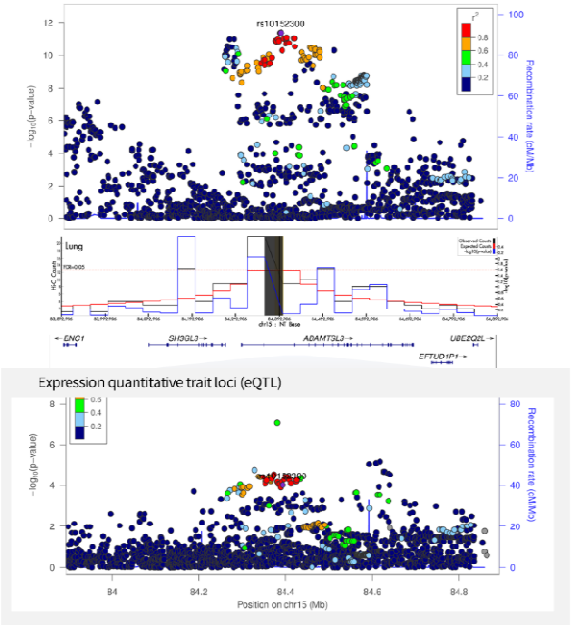
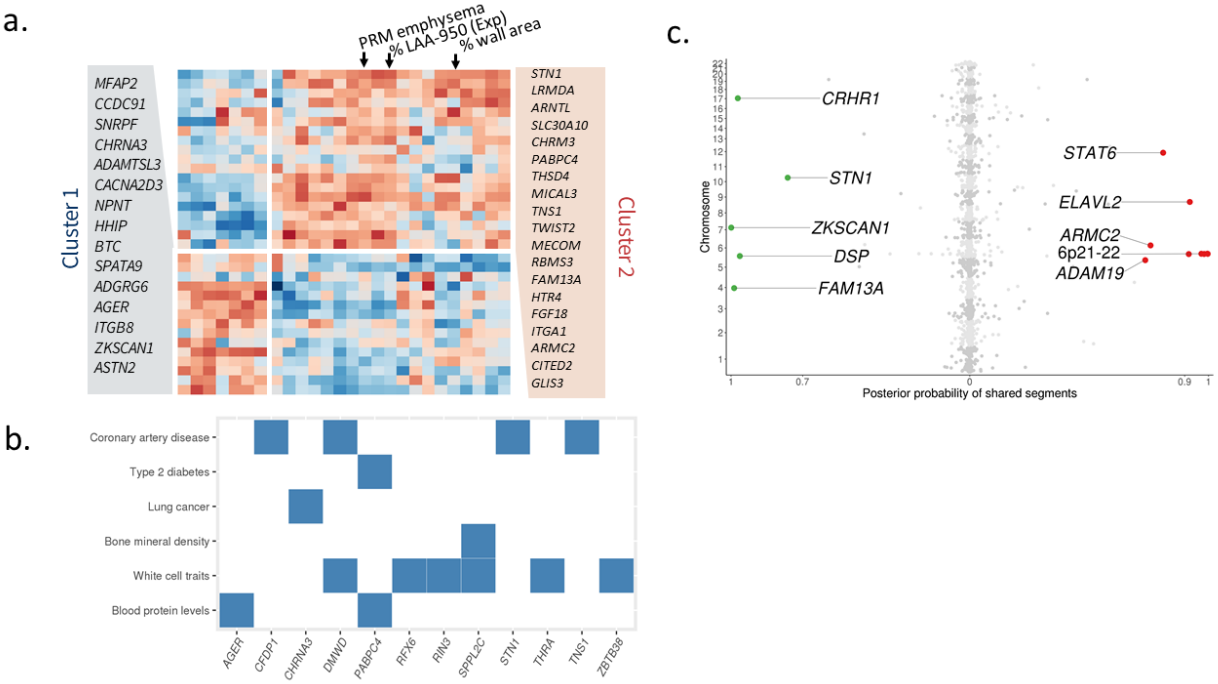


Figure 4 Effects on COPD-related and other phenotypes

(a) Heatmap of associations of 60 genome-wide significant variants (known and replicated novel associations) and imaging phenotypes in COPDGene (b) Overlapping of genome-wide significant loci of COPD and select traits from GWAS Catalog (c) Genome-wide overlapping results between COPD with

pulmonary fibrosis (left) and asthma (right). PRM emphysema, emphysema quantified by parametric response mapping; %LAA-950 (Exp), percentage of low attenuation area at -950 Hounsfield Units at expiration (related to gas trapping). GREx, Genetically regulated gene expresssion (significant associations identified by S-PrediXcan); mQTL, methylation quantitative trait loci (colocalized signals between GWAS and mQTL at posterior probability > 0·1); Cod., Coding associations (significant single variant or gene-based association tests for deleterious coding variants); Hi-C, significant chromatin interaction identified in human lung or IMR90 cell line; DHS, DNase hypersensititvity sites (using regulatory fine-mapping or regfm software); GSet, target genes identified by DEPICT using reconstituted gene sets.



Tables

Table 1 Meta-analysis results showing 35 loci novel for COPD and lung function

(see the Excel file)

Acknowledgements

Please refer to the **Supplementary Note** for full acknowledgements.

Author contributions

P.S. contributed to the study concept and design, data analysis, and manuscript writing. D.P., B.D.H., M.H.C. contributed to the study concept and design, data analysis, statistical support, and manuscript writing. A.B.W., K.d.J., S.J.L., D.P.S. contributed to the study concept and design and data analysis. P.B., R.G.B., J.D.C., A.G., D.A.M., G.T.O., S.I.R., D.A.S., R.T.-S., Y.T., E.K.S. contributed to the study concept and design and data collection. T.H.B., J.E.H. contributed to the study concept and design and to statistical support. I.P.H., H.M.B., L.V.W., M.D.T. contributed to the study concept and design. All authors,

including those whose initials are not listed above, contributed to the critical review and editing of the manuscript and approved the final version of the manuscript.

Competing financial interests

M.H.C., E.K.S., L.V.W., M.D.T., and I.P.H. have received grant funding from GSK. E.K.S. has received honoraria from Novartis for Continuing Medical Education Seminars and travel support from GlaxoSmithKline. I.P.H. has received grant support from BI. R.T.-S. is an employee of GSK. D.A.S. has financial support from Eleven P15. J.V. has received personal fees from GSK, Chiesi Pharmaceuticals, BI, Novartis, and AstraZeneca.

Supplementary Figures

Supplementary Figure 1 Forest plots for 82 genome-wide significant associations

Supplementary Figure 2 Regional association plots for 82 genome-wide significant associations

Supplementary Figure 3 Enrichment of genes specifically expressed in respiratory cell types

Supplementary Figure 4 Number of variants in 99% credible sets

Supplementary Figure 5 Heatmap of associations of 82 index variants and phenotypes in COPDGene

Supplementary Figure 6 Associations of index variants and traits in NHGRI-EBI GWAS Catalog

Supplementary Figure 7 Comparison of odds ratios (OR) including and excluding individuals with asthma of 82 genome-wide significant variants

Supplementary Tables

See the Excel file.

Supplementary Table 1 Cohort baseline characteristics in COPD cases and controls

Supplementary Table 2 Meta-analysis results showing 47 previously reported loci for COPD or lung function

Supplementary Table 3 Multiple independent associations within the same 2-Mb window identified using approximate conditional and joint analysis

Supplementary Table 4 Heritability enrichment in cell-type specific epigenomic mark from Roadmap Epigenomic Project

Supplementary Table 5 Genes specifically expressed in genome-wide significant loci

Supplementary Table 6 Functional annotation of variants with posterior probability of association greater than 0.6

Supplementary Table 7 Candidate target genes

Supplementary Table 8 Gene sets significantly enriched at FDR < 0.05 using DEPICT

Supplementary Table 9 Overlapping loci between COPD with asthma and pulmonary fibrosis

Supplementary Table 10 Lookup associations for quantitative computed tomography (QCT) features and cluster membership

Supplementary Table 11 Association results from GWAS catalog

Supplementary Table 12 Genetic correlation between COPD and other traits/diseases

References

1. GBD 2015 Chronic Respiratory Disease Collaborators. Global, regional, and national deaths, prevalence, disability-adjusted life years, and years lived with disability for chronic obstructive pulmonary disease and asthma, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet. Respir. Med.* **5,** 691–706 (2017).
2. *Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016*. (2018).
3. Fuchsberger, C. *et al.* The genetic architecture of type 2 diabetes. *Nature* **536,** 41–47 (2016).
4. Zhou, J. J. *et al.* Heritability of chronic obstructive pulmonary disease and related phenotypes in smokers. *Am. J. Respir. Crit. Care Med.* **188,** 941–7 (2013).
5. Hobbs, B. D. *et al.* Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis. *Nat Genet* **49,** 426–432 (2017).
6. Jiang, Z. *et al.* A Chronic Obstructive Pulmonary Disease Susceptibility Gene, FAM13A, Regulates Protein Stability of beta-Catenin. *Am J Respir Crit Care Med* **194,** 185–197 (2016).
7. Lao, T. *et al.* Hhip haploinsufficiency sensitizes mice to age-related emphysema. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E4681-7 (2016).
8. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12,** e1001779 (2015).
9. Vogelmeier, C. F. *et al.* Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. *Am. J. Respir. Crit. Care Med.* **195,** 557–582 (2017).
10. Cho, M. H. *et al.* Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med* **2,** 214–225 (2014).
11. Hobbs, B. D. *et al.* Exome Array Analysis Identifies a Common Variant in IL27 Associated with Chronic Obstructive Pulmonary Disease. *Am J Respir Crit Care Med* **194,** 48–57 (2016).
12. Cho, M. H. *et al.* A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet* **21,** 947–957 (2012).
13. Soler Artigas, M. *et al.* Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. *Nat Genet* **43,** 1082–1090 (2011).
14. Wain, L. V *et al.* Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. *Nat. Genet.* **49,** 416–425

(2017).

15. Soler Artigas, M. *et al.* Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nat Commun* **6,** 8658 (2015).

16. Wyss, A. B. *et al.* Multiethnic Meta-analysis Identifies New Loci for Pulmonary Function. *bioRxiv* (2017).

17. Loth, D. W. *et al.* Genome-wide association analysis identifies six new loci associated with forced vital capacity. *Nat Genet* **46,** 669–677 (2014).

18. Jackson, V. E. *et al.* Meta-analysis of exome array data identifies six novel genetic loci for lung function [version 1; referees: 1 approved, 1 approved with reservations]. *Wellcome Open Res.* **3,** (2018).

19. Repapi, E. *et al.* Genome-wide association study identifies five loci associated with lung function. *Nat Genet* **42,** 36–44 (2010).

20. Hancock, D. B. *et al.* Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. *Nat Genet* **42,** 45–52 (2010).

21. Wain, L. V *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* **3,** 769–781 (2015).

22. Wilk, J. B. *et al.* A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet* **5,** e1000429 (2009).

23. Lutz, S. M. *et al.* A genome-wide association study identifies risk loci for spirometric measures among smokers of European and African ancestry. *BMC Genet.* **16,** 138 (2015).

24. Shrine, N. *et al.* New genetic signals for lung function highlight pathways and pleiotropy, and chronic obstructive pulmonary disease associations across multiple ancestries. *bioRxiv* (2018).

25. Agusti, A. & Soriano, J. B. COPD as a systemic disease. *COPD* **5,** 133–8 (2008).

26. Barnes, P. J. & Celli, B. R. Systemic manifestations and comorbidities of COPD. *Eur. Respir. J.* **33,** 1165–85 (2009).

27. Lu, Q. *et al.* Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLoS Genet.* **13,** e1006933 (2017).

28. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50,** 621–629 (2018).

29. Slowikowski, K., Hu, X. & Raychaudhuri, S. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* **30,** 2496–7 (2014).

30. Xu, Y. *et al.* Single-cell RNA sequencing identifies diverse roles of epithelial cells in idiopathic pulmonary fibrosis. *JCI insight* **1,** e90558 (2016).

31. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81,** 208–27 (2007).

32. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101,** 5–22 (2017).

33. Zhou, X. *et al.* Identification of a chronic obstructive pulmonary disease genetic determinant that regulates HHIP. *Hum. Mol. Genet.* **21,** 1325–35 (2012).

34. Claussnitzer, M., Hui, C.-C. & Kellis, M. FTO Obesity Variant and Adipocyte Browning in Humans. *N. Engl. J. Med.* **374,** 192–3 (2016).

35. Regan, E. A. *et al.* Genetic epidemiology of COPD (COPDGene) study design. *COPD* **7,** 32–43 (2010).

36. Cho, M. H. *et al.* A Genome-Wide Association Study of Emphysema and Airway Quantitative Imaging Phenotypes. *Am. J. Respir. Crit. Care Med.* **192,** 559–569 (2015).

37. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).

38. Demenais, F. *et al.* Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. *Nat. Genet.* **50,** 42–53 (2018).

39. Fingerlin, T. E. *et al.* Genome-wide imputation study identifies novel HLA locus for pulmonary fibrosis and potential role for auto-immunity in fibrotic idiopathic interstitial pneumonia. *BMC Genet.* **17,** 74 (2016).

40. Skronska-Wasek, W. *et al.* Reduced Frizzled Receptor 4 Expression Prevents WNT/β-Catenin-driven Alveolar Lung Repair in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* **196,** 172–185 (2017).

41. Sakornsakolpat, P. *et al.* Integrative genomics identifies new genes associated with severe COPD and emphysema. *Respir. Res.* **19,** 46 (2018).

42. Bui, D. S. *et al.* Childhood predictors of lung function trajectories and future COPD risk: a prospective cohort study from the first to the sixth decade of life. *Lancet. Respir. Med.* (2018). doi:10.1016/S2213-2600(18)30100-0

43. McGeachie, M. J. *et al.* Patterns of Growth and Decline in Lung Function in Persistent Childhood Asthma. *N. Engl. J. Med.* **374,** 1842–1852 (2016).

44. Ross, J. C. *et al.* Longitudinal Modeling of Lung Function Trajectories in Smokers with and without COPD. *Am. J. Respir. Crit. Care Med.* (2018). doi:10.1164/rccm.201707-1405OC

45. Yang, J. *et al.* Rootletin, a novel coiled-coil protein, is a structural component of the ciliary rootlet. *J. Cell Biol.* **159,** 431–40 (2002).

46. Gibson, M. A., Hughes, J. L., Fanning, J. C. & Cleary, E. G. The major antigen of elastin-associated microfibrils is a 31-kDa glycoprotein. *J. Biol. Chem.* **261,** 11429–36 (1986).

47. Massaro, G. D. *et al.* Retinoic acid receptor-beta: an endogenous inhibitor of the perinatal formation of pulmonary alveoli. *Physiol. Genomics* **4,** 51–7 (2000).

48. Hall, N. G., Klenotic, P., Anand-Apte, B. & Apte, S. S. ADAMTSL-3/punctin-2, a novel glycoprotein in extracellular matrix related to the ADAMTS family of metalloproteases. *Matrix Biol.* **22,** 501–10 (2003).

49. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733-45 (2016).

50. Saito, A., Ozaki, K., Fujiwara, T., Nakamura, Y. & Tanigami, A. Isolation and mapping of a human lung-specific gene, TSA1902, encoding a novel chitinase family member. *Gene* **239,** 325–31 (1999).

51. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* **13,** 397–406 (2014).

52. Aminuddin, F. *et al.* Genetic association between human chitinases and lung function in COPD. *Hum. Genet.* **131,** 1105–14 (2012).

53. Birben, E. *et al.* The effects of an insertion in the 5'UTR of the AMCase on gene expression and pulmonary functions. *Respir. Med.* **105,** 1160–9 (2011).

54. Chatterjee, R., Batra, J., Das, S., Sharma, S. K. & Ghosh, B. Genetic association of acidic mammalian chitinase with atopic asthma and serum total IgE levels. *J. Allergy Clin. Immunol.* **122,** 202–8, 208.e1–7 (2008).

55. Ober, C. & Chupp, G. L. The chitinase and chitinase-like proteins: a review of genetic and functional studies in asthma and immune-mediated diseases. *Curr. Opin. Allergy Clin. Immunol.* **9,** 401–8 (2009).

56. Heinzmann, A. *et al.* Joint influences of Acidic-Mammalian-Chitinase with Interleukin-4 and Toll-like receptor-10 with Interleukin-13 in the genetics of asthma. *Pediatr. Allergy Immunol.* **21,** e679-86 (2010).

57. Okawa, K. *et al.* Loss and Gain of Human Acidic Mammalian Chitinase Activity by Nonsynonymous SNPs. *Mol. Biol. Evol.* **33,** 3183–3193 (2016).

58. Zhu, Z. *et al.* Acidic mammalian chitinase in asthmatic Th2 inflammation and IL-13 pathway activation. *Science* **304,** 1678–82 (2004).

59. Yang, S. *et al.* Tumor necrosis factor receptor 2 (TNFR2)·interleukin-17 receptor D (IL-17RD) heteromerization reveals a novel mechanism for NF-κB activation. *J. Biol. Chem.* **290,** 861–71 (2015).

60. Markovics, J. A. *et al.* Interleukin-1beta induces increased transcriptional activation of the transforming growth factor-beta-activating integrin subunit beta8 through altering chromatin architecture. *J. Biol. Chem.* **286,** 36864–74 (2011).

61. Kitamura, H. *et al.* Mouse and human lung fibroblasts regulate dendritic cell trafficking, airway inflammation, and fibrosis through integrin αvβ8-mediated activation of TGF-β. *J. Clin. Invest.* **121,** 2863–75 (2011).

62. Araya, J. *et al.* Squamous metaplasia amplifies pathologic epithelial-mesenchymal interactions in COPD patients. *J. Clin. Invest.* **117,** 3551–62 (2007).

63. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49,** 1752–1757 (2017).

64. Pilewski, J. M., Latoche, J. D., Arcasoy, S. M. & Albelda, S. M. Expression of integrin cell adhesion receptors during human airway epithelial repair in vivo. *Am. J. Physiol.* **273,** L256-63 (1997).

65. Zacharias, W. J. *et al.* Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor. *Nature* **555,** 251–255 (2018).

66. Franks, T. J. *et al.* Resident cellular components of the human lung: current knowledge and goals for research on cell phenotyping and function. *Proc. Am. Thorac. Soc.* **5,** 763–6 (2008).

67. Boschetto, P. *et al.* Predominant emphysema phenotype in chronic obstructive pulmonary. *Eur. Respir. J.* **21,** 450–4 (2003).

68. Castaldi, P. J. *et al.* Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* **69,** 415–22 (2014).

69. Cerveri, I. *et al.* The rapid FEV(1) decline in chronic obstructive pulmonary disease is associated with predominant emphysema: a longitudinal study. *COPD* **10,** 55–61 (2013).

70. Hersh, C. P. *et al.* Non-emphysematous chronic obstructive pulmonary disease is associated with diabetes mellitus. *BMC Pulm. Med.* **14,** 164 (2014).

71. Higami, Y. *et al.* Increased Epicardial Adipose Tissue Is Associated with the Airway Dominant Phenotype of Chronic Obstructive Pulmonary Disease. *PLoS One* **11,** e0148794 (2016).

72. Smolonska, J. *et al.* Common genes underlying asthma and COPD? Genome-wide analysis on the Dutch hypothesis. *Eur. Respir. J.* **44,** 860–72 (2014).

73. Ishaq, M. *et al.* The DEAD-box RNA helicase DDX1 interacts with RelA and enhances nuclear factor kappaB-mediated transcription. *J. Cell. Biochem.* **106,** 296–305 (2009).

74. Levy, D. *et al.* Genome-wide association identifies OBFC1 as a locus involved in human leukocyte telomere biology. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9293–8 (2010).

75. Cho, M. H. *et al.* Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *Lancet Respir Med* **2,** 214–225 (2014).

76. Miller, M. R. *et al.* Standardisation of spirometry. *Eur. Respir. J.* **26,** 319–38 (2005).

77. Hankinson, J. L., Odencrantz, J. R. & Fedan, K. B. Spirometric reference values from a sample of the general U.S. population. *Am. J. Respir. Crit. Care Med.* **159,** 179–87 (1999).

78. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48,** 1279–83 (2016).

79. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).

80. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).

81. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–5 (2015).

82. Berisa, T. & Pickrell, J. K. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics* **32,** 283–5 (2016).

83. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42,** 441–7 (2010).

84. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44,** 369–75, S1-3 (2012).

85. Genome Refernce Consortium. Human Genome Region MHC. Available at: https://www.ncbi.nlm.nih.gov/grc/human/regions/MHC?asm=GRCh37.p13.

86. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–7 (2010).

87. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–35 (2015).

88. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4,** 1184–91 (2009).

89. Barbeira, A. N. *et al.* Integrating Predicted Transcriptome From Multiple Tissues Improves Association Detection. *bioRxiv* (2018).

90. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet.* **8,** e1003029 (2012).

91. Lamontagne, M. *et al.* Leveraging lung tissue transcriptome to uncover candidate causal genes in COPD genetic associations. *Hum. Mol. Genet.* **27,** 1819–1829 (2018).

92. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47,** 1091–8 (2015).

93. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** 5890 (2015).

94. Shooshtari, P., Huang, H. & Cotsapas, C. Integrative Genetic and Epigenetic Analysis Uncovers Regulatory Mechanisms of Autoimmune Disease. *Am. J. Hum. Genet.* **101,** 75–86 (2017).

95. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28,** 1045–8 (2010).

96. Morrow, J. D. *et al.* Human Lung DNA Methylation Quantitative Trait Loci Colocalize with COPD Genome-wide Association Loci. *Am. J. Respir. Crit. Care Med.* (2018). doi:10.1164/rccm.201707-1434OC

97. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am. J. Hum. Genet.* **99,** 1245–1260 (2016).

98. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17,** 2042–2059 (2016).

99. Martin, J. S. *et al.* HUGIn: Hi-C Unifying Genomic Interrogator. *Bioinformatics* **33,** 3793–3795 (2017).

100. Lee, S., Fuchsberger, C., Kim, S. & Scott, L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. *Biostatistics* **17,** 1–15 (2016).

101. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48,** 709–17 (2016).

102. Caliński, T. & Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **3,** 1–27 (1974).

103. Dimas, A. S. *et al.* Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes* **63,** 2158–71 (2014).

104. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33,** 1–22 (2010).

105. Cho, M. H. *et al.* A Genome-Wide Association Study of Emphysema and Airway Quantitative Imaging Phenotypes. *Am. J. Respir. Crit. Care Med.* **192,** 559–69 (2015).

106. EMBL-EBI. GWAS Catalog. Available at: https://www.ebi.ac.uk/gwas/. (Accessed: 10th April 2018)