Exome-wide analysis of rare coding variation identifies novel associations with COPD and airflow limitation in MOCS3, IFIT3 and SERPINA12


ABSTRACT

Background Several regions of the genome have shown to be associated with COPD in genome-wide association studies of common variants.

Objective To determine rare and potentially functional single nucleotide polymorphisms (SNPs) associated with the risk of COPD and severity of airflow limitation.

Methods 3226 current or former smokers of European ancestry with lung function measures indicative of Global Initiative for Chronic Obstructive Lung Disease (GOLD) 2 COPD or worse were genotyped using an exome array. An analysis of risk of COPD was carried out using ever smoking controls (n=4784). Associations with % predicted FEV1 were tested in cases. We followed-up signals of interest (p<10−5) in independent samples from a subset of the UK Biobank population and also undertook a more powerful discovery study by meta-analysing the exome array data and UK Biobank data for variants represented on both arrays.

Results Among the associated variants were two in regions previously unreported for COPD; a low frequency non-synonymous SNP in MOCS3 (rs7269297, PDiscovery=3.08×10−6, Preplication=0.019) and a rare SNP in IFIT3, which emerged in the meta-analysis (rs140549288, PMeta=8.56×10−6). In the meta-analysis of % predicted FEV1 in cases, the strongest association was shown for a splice variant in a previously unreported region, SERPINA12 (rs140198372, PMeta=5.72×10−6). We also confirmed previously reported associations with COPD risk at MMP12, HHIP, GPR126 and CHRNA5. No associations in novel regions reached a stringent exome-wide significance threshold (p<3.7×10−5).

Conclusions This study identified several associations with the risk of COPD and severity of airflow limitation, including novel regions MOCS3, IFIT3 and SERPINA12, which warrant further study.

INTRODUCTION

COPD is a major public health concern, being a leading cause of morbidity and mortality worldwide.1 The Global Initiative for Chronic Obstructive Lung Disease (GOLD) recommends that the impact of COPD on an individual patient should be assessed by considering breathlessness, symptoms and exacerbation risk, in combination with the severity of airflow limitation, which can be graded using %predicted FEV1.2 Approximately 16–2% of COPD cases can be attributed to α1-antitrypsin (AAT) deficiency, a rare inherited disorder, caused by mutations within the SERPINA1 gene.3 4 For the remainder of COPD cases, cigarette smoking is recognised as the most
significant risk factor\(^3\); however, there is also a genetic component, with several genomic regions showing association with COPD risk or airflow limitation to date, including CHRNA3/5, HHR\(^6\), HTR4, GSTD\(_D3\), TNFSI\(_6\), MMP\(_{12}\) \(^8\) and FAM13A.\(^9\) COPD diagnosis is confirmed using measures of lung function, so it is likely that the genetic determinants of COPD and lung function will overlap. Indeed, many loci identified in large genome-wide association studies (GWAS) of FEV\(_1\) and the ratio of FEV\(_1\) to forced vital capacity (FEV\(_1\)/FVC) in general population samples\(^10\–12\) have subsequently been shown to be associated with COPD or airflow limitation.\(^4\)

Despite the successes in identifying genes associated with lung function and COPD, these known loci only explain a small proportion of the expected heritability.\(^13\) Large GWAS undertaken to date have generally focused on common variants (typically >5\% minor allele frequency (MAF))\(^3\) \(^6\)–\(^14\)–\(^16\); one hypothesis is that some of the so-called ‘missing heritability’ might be accounted for by variants of lower frequencies. In this study, we set out to investigate the role of low frequency, functional variants in COPD, and to confirm the role of single nucleotide polymorphisms (SNPs) previously showing association with lung function. It is hypothesised that rare variants are more likely than common variants to have deleterious effects; identifying such SNPs could lead to greater understanding of the pathways and biological mechanisms underlying airflow obstruction and COPD, and could translate to novel targets for treatment.

We genotyped cases with a history of smoking and airflow limitation, indicative of GOLD 2 COPD or worse, and control samples using an exome chip array to which we had added custom content comprising 2585 SNPs tagging regions which had shown suggestive association (p<2.21 x 10\(^{-5}\)) with lung function in a previous large genome-wide HapMap-imputed study.\(^13\) The exome chip genotyping array design contains mostly non-synonymous, splice or stop codon altering variants that are likely to affect protein structure and function, with the majority of variants being low frequency (MAF \(1\%–5\%\)) or rare (MAF <1\%).

In this study, we carried out discovery case–control analyses (COPD cases vs controls) and analyses of %predicted FEV\(_1\) in cases, as a measure of severity of airflow limitation. Replication was undertaken using a subset of the UK Biobank Lung Exome Variant Evaluation (BiLEVE) study, a collection of 48931 individuals from large genome-wide association studies (GWAS) of FEV\(_1\) and the ratio of FEV\(_1\) to forced vital capacity (FEV\(_1\)/FVC) in general population samples\(^10\–12\) which includes rare variants in COPD, and to confirm the role of single nucleotide polymorphisms (SNPs) previously showing association with lung function. It is hypothesised that rare variants are more likely than common variants to have deleterious effects; identifying such SNPs could lead to greater understanding of the pathways and biological mechanisms underlying airflow obstruction and COPD, and could translate to novel targets for treatment.

**Genotyping**

All 3487 cases and 1032 GS:SFHS controls were genotyped together using the Illumina Human Exome BeadChip with additional custom content for regions which have previously shown modest association with lung function (description of custom content design in online supplementary methods). The remaining discovery analyses control samples were genotyped separately using the Illumina Human Exome BeadChip.

The UK BiLEVE samples were genotyped using the Affymetrix UK BiLEVE array, which includes rare variants selected from the same sequencing project as the Illumina Human Exome BeadChip alongside additional content.\(^16\) Of the 807 411 SNPs included on the Affymetrix UK BiLEVE array, 74 891 were also present on the Illumina Human Exome BeadChip; this subset of SNPs, which were directly genotyped on both arrays, was selected for the discovery meta-analysis.

**Quality control of genotype data**

**Discovery exome analysis**

Genotypes were called using Illumina’s Gencall algorithm in Genomestudio\(^18\) with refinement of rare variants with missing calls undertaken using zCall.\(^19\) Standard quality control (QC) filters were applied, in accordance with the Exome-chip Quality Control SOP V5, as developed within the UK exome chip consortium\(^20\) and are fully described in online supplementary methods. In brief, SNPs were excluded if they had low call rate (\(<99\%)\) or deviated from Hardy Weinberg Equilibrium (p<10\(^{-8}\)) and samples were excluded if they were duplicates, sex mismatches, heterozygosity outliers (\(>3\ SD\) from mean), had an excess of singleton SNPs, or were ancestral outliers. Clusterplots for all SNPs of interest were inspected, to ensure accuracy of genotype calling.

**UK BiLEVE data**

The QC procedure of the UK BiLEVE genotype data is described elsewhere.\(^16\)
## Table 1  Clinical characteristics of samples passing genotype QC

<table>
<thead>
<tr>
<th>Sample collection</th>
<th>Sex</th>
<th>Age</th>
<th>%Predicted FEV\textsubscript{1}</th>
<th>FEV\textsubscript{1}/FVC</th>
<th>Pack-years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male, n (%)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Samples with data (n)</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>Mean (SD)</td>
<td>n</td>
</tr>
<tr>
<td>Discovery analyses airflow limitation cases (total n=3226, with pack-years n=2517)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS:SFHS</td>
<td>508</td>
<td>224 (44.1%)</td>
<td>58.9 (8.94)</td>
<td>64.84 (12.64)</td>
<td>0.580 (0.108)</td>
</tr>
<tr>
<td>British Regional Heart Study</td>
<td>425</td>
<td>425 (100%)</td>
<td>70.1 (5.46)</td>
<td>59.41 (14.66)</td>
<td>0.597 (0.084)</td>
</tr>
<tr>
<td>British Women’s Heart and Health Study</td>
<td>254</td>
<td>0 (0%)</td>
<td>69.3 (5.46)</td>
<td>64.26 (12.40)</td>
<td>0.603 (0.074)</td>
</tr>
<tr>
<td>UK COPD cohort*</td>
<td>209</td>
<td>129 (61.7%)</td>
<td>68.7 (8.11)</td>
<td>37.94 (15.29)</td>
<td>0.447 (0.119)</td>
</tr>
<tr>
<td>Hertfordshire Cohort Study</td>
<td>317</td>
<td>203 (64.0%)</td>
<td>66.1 (2.79)</td>
<td>62.89 (13.57)</td>
<td>0.589 (0.101)</td>
</tr>
<tr>
<td>COPDEBEAT*</td>
<td>87</td>
<td>62 (71.3%)</td>
<td>67.6 (8.77)</td>
<td>45.19 (16.24)</td>
<td>0.480 (0.115)</td>
</tr>
<tr>
<td>Nottingham COPD study*</td>
<td>76</td>
<td>48 (63.2%)</td>
<td>67.2 (8.97)</td>
<td>50.29 (15.04)</td>
<td>0.482 (0.111)</td>
</tr>
<tr>
<td>Nottingham smokers</td>
<td>125</td>
<td>78 (62.4%)</td>
<td>63.1 (8.60)</td>
<td>46.27 (17.65)</td>
<td>0.503 (0.125)</td>
</tr>
<tr>
<td>Gedling study</td>
<td>33</td>
<td>26 (78.8%)</td>
<td>69.0 (8.23)</td>
<td>59.67 (16.81)</td>
<td>0.593 (0.103)</td>
</tr>
<tr>
<td>English Longitudinal Study of Aging</td>
<td>166</td>
<td>75 (45.2%)</td>
<td>66.0 (8.17)</td>
<td>54.84 (17.24)</td>
<td>0.526 (0.149)</td>
</tr>
<tr>
<td>EU COPD Gene Scan*</td>
<td>277</td>
<td>155 (56.0%)</td>
<td>67.0 (8.88)</td>
<td>38.51 (14.74)</td>
<td>0.467 (0.120)</td>
</tr>
<tr>
<td>GoTARDIS Study*</td>
<td>749</td>
<td>412 (55.0%)</td>
<td>68.8 (8.97)</td>
<td>52.16 (14.14)</td>
<td>0.509 (0.110)</td>
</tr>
<tr>
<td>Discovery analyses controls (total n=4784, with pack-years n=3889)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GS:SFHS</td>
<td>961</td>
<td>552 (57.4%)</td>
<td>54.5 (8.41)</td>
<td>98.18 (10.92)</td>
<td>0.783 (0.051)</td>
</tr>
<tr>
<td>British 1958 Birth Cohort</td>
<td>1429</td>
<td>888 (62.1%)</td>
<td>44 (0)</td>
<td>100.90 (13.46)</td>
<td>0.809 (0.060)</td>
</tr>
<tr>
<td>Oxford Biobank</td>
<td>1770</td>
<td>832 (47.0%)</td>
<td>41.6 (5.77)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>GoDARTS</td>
<td>624</td>
<td>402 (64.4%)</td>
<td>59.0 (10.75)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UK Biobank Lung Exome Variant Evaluation samples (meta-analysis and replication)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Airflow limitation cases</td>
<td>4231</td>
<td>2379 (56.2%)</td>
<td>59.54 (6.86)</td>
<td>61.76 (11.8)</td>
<td>0.607 (0.076)</td>
</tr>
<tr>
<td>Controls</td>
<td>8979</td>
<td>4260 (47.4%)</td>
<td>56.19 (7.92)</td>
<td>101.40 (8.1)</td>
<td>0.773 (0.038)</td>
</tr>
</tbody>
</table>

*Sample collection is COPD case cohort.

GS:SFHS, Generation Scotland: Scottish Family Health Study; GoTARDIS, Tayside Allergy and Respiratory Disease Information System; QC, quality control.

---

**Figure 1**  Two-stage study design. Stage 1: exome discovery analyses. Stage 2: Follow-up in UK BiLEVE: A. Replication of signals; B. meta-analysis of UK COPD exome chip consortium and UK BiLEVE.
Statistical analyses
SNP associations with COPD risk were carried out using a logistic regression model, adjusting for age, sex and pack-years and assuming an additive genetic model. Associations with untransformed %predicted FEV1 in cases were tested, using a linear regression model, with adjustment for pack-years (analysis of severity of airflow limitation). Since not all samples had pack-years data available, secondary analyses were carried out without adjustment for pack-years, for both the COPD risk and severity of airflow limitation analyses, allowing the inclusion of all samples. Single variant analyses were carried out using PLINK v1.07.21 Using a Bonferroni correction for the number of tests undertaken, a significance level of p<3.7x10^-7 would be required in the exome single variant analysis to retain a type 1 error of 5%. We defined SNPs of interest as those with p<10^-5 in the discovery exome analysis; for these SNPs, we undertook replication analyses in the UK BiLEVE study to corroborate findings (see online supplementary methods). We set a Bonferroni corrected significance level for replication, for the number of SNPs in novel loci taken forward to replication (p<0.017 for analysis of COPD risk). Gene-based analyses using SKAT-O were additionally undertaken; the methods and results of these analyses are described in the online supplementary information.

Custom content single variant analyses
Custom content comprising 2585 SNPs tagging regions which had shown suggestive association (p<2.21x10^-3) with lung function in a previous large genome-wide HapMap-imputed study4,13 were also included on the array for cases and G5/SHFS controls. Additional controls from 1958BC and Busselton Health Study (BHS) with genome-wide data were also used; full methods and results of this analysis are given in the supplementary information.

Meta-analysis with UK BiLEVE data
Single variant associations with COPD risk and severity of airflow limitation in the UK BiLEVE samples were carried out using PLINK v1.07,21 identically to the corresponding discovery analysis with pack-years adjustment. We carried out an inverse-variance–weighted meta-analysis of the union of SNPs included in the discovery exome and UK BiLEVE analyses (described in online supplementary methods).

RESULTS
Discovery exome analysis
3226 cases and 4784 controls passed all sample and SNP genotype QC and were used in the exome analysis (exclusions in online supplementary table S1). Clinical characteristics of these samples are summarised in table 1. Of the SNPs which passed all QC criteria in both cases and controls, 135 818 were polymorphic, of which 101 308 (74.6%) had a MAF<1%.

Analyses of COPD risk
We carried out pack-years adjusted analysis of COPD risk, including 2517 cases and 3889 controls, in addition to an unadjusted analysis, using all 3226 cases and 4784 controls (quantile–quantile plots shown in online supplementary figure S1). A total of four SNPs in three regions met the p<10^-5 significance threshold in the pack-years adjusted analysis, with five SNPs in four regions showing p<10^-4 in the unadjusted analysis (figure 2).

In the pack-years adjusted analysis (table 2A and figure 2A), the most significant association was for the previously reported COPD/smoking region 15q25 (sentinel SNP rs8034191 OR: 1.38, MAF=34.8%, p=2.42x10^-7). This signal was replicated in the UK BiLEVE study. Two novel signals of association with COPD risk (p<10^-3) were rs3813803 within SMPDL3B (OR: 1.37, MAF=29.2%, p=1.04x10^-6) and low frequency SNP rs7269297 within MOCS3 (OR: 0.25, MAF=1.1%, p=3.08x10^-4). There was evidence of replication, just above the Bonferroni corrected level of significance (p=0.017) for rs7269297 in the UK BiLEVE study (p=7.27x10^-3 for meta-analysis of discovery and UK BiLEVE results, table 2A).

A further two loci were associated with COPD risk in the analysis unadjusted for pack-years: rs3827522 within PRICKLE1 (OR: 0.12, MAF=0.4%, p=1.03x10^-5) and rs17368582 within MMP12 (OR: 0.712, MAF=12.2%, p=5.01x10^-6, table 2A and figure 2B); however, there was no evidence of replication of these associations with COPD risk in UK BiLEVE. rs2276109, another SNP within MMP12, (MAF=5.6%) which is strongly correlated with rs17368582 (r²=0.84), has previously been associated with COPD risk in smokers. Overall, no associations in novel regions met exome-wide significance (p<3.7x10^-5).

Analyses of severity of airflow limitation
Although no SNPs reached the p<10^-5 significance level in either the pack-years adjusted, or the unadjusted analysis (see online supplementary figures S2 and S3), six SNPs showed some evidence of association (p<10^-4) in one or both analyses (see online supplementary table S2). Of note, rs28929474, the z-allele within the SERPINA1 gene, showed modest association in the unadjusted analysis (β=-6.17%, MAF=2.0%, p=2.83x10^-5).

UK BiLEVE meta-analysis results
Analyses of COPD risk
For the 57 234 polymorphic SNPs common to both the COPD exome chip consortium samples and the UK BiLEVE study, a meta-analysis of discovery and UK BiLEVE study results was undertaken in which three regions showed association with risk of COPD (p<10^-5, figure 3, online supplementary figure S4 and table 2B). The GYPA/HHIP and GPR126 genes have previously been reported as showing association with lung function and COPD or airflow limitation risk.3 10 14 The IFIT3 region signal (rs140549288 p.Val352Leu in IFIT3, OR: 1.92, MAF=0.7%, p=7.49x10^-6) represents a novel rare variant signal of association with COPD.

Analyses of severity of airflow limitation
A total of 54 168 SNPs were included in the meta-analysis of severity of airflow limitation (see online supplementary figures S5 and S6). One SNP showed association with p<10^-3: rs140198372, a variant which alters the sequence at a site where the splicing of an intron takes place (splice site) in SERPINA12 (β=-33.51%, MAF=0.03%, p=5.72x10^-6, table 3).

Sensitivity analyses to assess COPD case criteria
Of our 3226 COPD cases defined as described above, 1398 also had a GOLD 2 or worse COPD based on postbronchodilator spirometry. We carried out a sensitivity analysis for all SNPs identified in our discovery or meta-analyses of COPD risk, by repeating the discovery analyses including only those 1398 COPD cases which underwent reversibility testing. This analysis
showed consistent estimated effect sizes (see online supplementary table S3 and figure S7), and in particular, the ORs were not substantially attenuated for rs7269297 in MOCS3 (sensitivity analysis OR: 0.276; original discovery OR: 0.251), nor rs140549288 in IFIT3 (sensitivity analysis OR: 2.554; original discovery OR: 2.156).

Association of novel loci with smoking behaviour

Given the disparity of smoking behaviour in our cases and control samples (table 1), we further investigated whether either of the two novel COPD risk loci were associated with smoking behaviour, to ascertain whether the associations with COPD may be explained by differences in smoking. Neither of the sentinel SNPs showed significant association with heavy versus never smoking within UK BiLEVE (p=0.956 for rs7269297 and p=0.945 for rs140549288) study. We further undertook a look-up in the publically available results of a GWAS from the Tobacco and Genetics consortium for associations with rs7269297 in MOCS3 (rs140549288 was not available in data) and a number of smoking traits; however, no evidence for association with smoking behaviour was found (cigarettes per day p=0.610; ever vs never smoking p=0.172; current vs former smoking p=0.699).

DISCUSSION

We carried out analyses of exome chip variants with COPD risk and %predicted FEV1 among cases, through which we identified a number of SNPs in both known COPD regions and at novel loci that showed suggestive association (p<10^{-5}) with risk of COPD. These novel regions (region plots: online supplementary figure S8) warrant further investigation as they may provide insight into the underlying biological mechanisms of COPD and airflow limitation in smokers and could provide novel therapeutic targets. The most significant associations in both the discovery exome analysis and the meta-analysis were with SNPs in the 15q25 region, previously identified through GWAS as being associated with smoking behaviour, lung cancer, COPD and airflow obstruction. In addition, we independently replicated previously reported associations of HHIP, GPR126 and MMP12 with COPD risk.

We identified novel associations between COPD risk and low frequency or rare coding SNPs in two genes: MOCS3

Figure 2  (A) Analysis of COPD risk, with pack-years adjustment (single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) >0.05% only; SNPs with p<10^{-5} highlighted). (B) Analysis of COPD risk, without pack-years adjustment (SNPs with MAF >0.05% only; SNPs with p<10^{-5} highlighted).
Table 2  Top associations in exome discovery analyses and meta-analysis of COPD risk

(A) SNPs with p<10\(^{-5}\) in either the pack-years adjusted or unadjusted discovery analyses

<table>
<thead>
<tr>
<th>rs no.</th>
<th>CHR</th>
<th>Position</th>
<th>Coded allele</th>
<th>Gene</th>
<th>MAF (MAC)</th>
<th>Association result</th>
<th>MAF (MAC)</th>
<th>Association result</th>
<th>MAF (MAC)</th>
<th>Association result</th>
<th>OR (95% CI)</th>
<th>p Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs3813803</td>
<td>1</td>
<td>28282292</td>
<td>C</td>
<td>SMPDL3B (non-synonymous)</td>
<td>30.6% (1541)</td>
<td>28.3% (2203)</td>
<td>1.370 (1.207 to 1.554)</td>
<td>30.3% (2722)</td>
<td>28.5% (1.430)</td>
<td>2.41×10(^{-6})</td>
<td>28.7% (5269)</td>
<td>29.4% (0.911)</td>
</tr>
<tr>
<td>rs17368582</td>
<td>11</td>
<td>102738075</td>
<td>C</td>
<td>MMP12 (synonymous)</td>
<td>11.1% (561)</td>
<td>12.9% (1001)</td>
<td>0.767 (0.642 to 0.915)</td>
<td>11.1% (719)</td>
<td>12.8% (1229)</td>
<td>5.01×10(^{-6})</td>
<td>12.0% (2198)</td>
<td>12.2% (0.866)</td>
</tr>
<tr>
<td>rs3827522</td>
<td>12</td>
<td>42853871</td>
<td>A</td>
<td>PRICKLE1 (non-synonymous)</td>
<td>0.2% (11)</td>
<td>0.4% (27)</td>
<td>0.184 (0.065 to 0.448)</td>
<td>0.2% (14)</td>
<td>0.5% (46)</td>
<td>1.03×10(^{-7})</td>
<td>0.3% (21)</td>
<td>0.3% (45)</td>
</tr>
<tr>
<td>rs4896582</td>
<td>6</td>
<td>142703877</td>
<td>A</td>
<td>GPR126 (intronic)</td>
<td>38.0% (1912)</td>
<td>32.7% (2546)</td>
<td>1.374 (1.218 to 1.550)</td>
<td>37.7% (2432)</td>
<td>32.9% (3144)</td>
<td>5.75×10(^{-7})</td>
<td>39.2% (6320)</td>
<td>35.2% (1.133)</td>
</tr>
<tr>
<td>rs140549288</td>
<td>20</td>
<td>49576664</td>
<td>G</td>
<td>MOCS3 (non-synonymous)</td>
<td>0.7% (37)</td>
<td>1.4% (110)</td>
<td>0.251 (0.140 to 0.448)</td>
<td>0.8% (54)</td>
<td>1.5% (139)</td>
<td>6.85×10(^{-5})</td>
<td>1.2% (252)</td>
<td>1.4% (0.758)</td>
</tr>
</tbody>
</table>

(B) SNPs with p<10\(^{-5}\) in the meta-analysis (only most statically significant SNP in each region shown)

<table>
<thead>
<tr>
<th>rs no.</th>
<th>CHR</th>
<th>Position</th>
<th>Coded allele</th>
<th>Gene</th>
<th>MAF (MAC)</th>
<th>Association result</th>
<th>MAF (MAC)</th>
<th>Association result</th>
<th>MAF (MAC)</th>
<th>Association result</th>
<th>OR (95% CI)</th>
<th>p Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1828591</td>
<td>4</td>
<td>145480780</td>
<td>A</td>
<td>GYPA/HHIP (intergenic)</td>
<td>35.6% (1794)</td>
<td>39.1% (3042)</td>
<td>0.9167 (0.814, 1.032)</td>
<td>36.6% (3088)</td>
<td>40.0% (0.819, 0.918)</td>
<td>8.68×10(^{-7})</td>
<td>0.876 (0.832)</td>
<td>0.922 (5.75×10(^{-7})</td>
</tr>
<tr>
<td>rs4896582</td>
<td>6</td>
<td>142703877</td>
<td>A</td>
<td>GPR126 (intronic)</td>
<td>29.3% (1473)</td>
<td>31.7% (2468)</td>
<td>0.8594 (0.757, 0.974)</td>
<td>28.0% (2349)</td>
<td>30.2% (0.826, 0.934)</td>
<td>3.87×10(^{-5})</td>
<td>0.875 (0.827)</td>
<td>0.925 (2.53×10(^{-6})</td>
</tr>
</tbody>
</table>

* p Values in bold significant at p<10\(^{-5}\) level.
(rs7269297, serine to alanine, MAF=1.3%, \(p_{\text{discovery}}=3.08 \times 10^{-6}\), PolyPhen prediction: benign) and \(IFIT3\) (rs140549288, valine to leucine, MAF=0.7%, \(p_{\text{meta}}=8.56 \times 10^{-6}\), PolyPhen prediction: benign). The protein encoded by \(MOCS3\) adenylates and activates molybdopterin synthase, an enzyme required to synthesise molybdenum cofactor and is expressed in bronchial epithelium and smooth muscle layer of the bronchus.\(^{26}\) \(IFIT3\) is associated with interferon-\(\alpha\) antiviral activity and has been found to be up-regulated in respiratory syncytial virus infection\(^{29}\) and in human lung epithelial cells infected with dengue virus.\(^{29}\) The SNP rs140549288 is also located within an intron of \(LIPA\); the product of this gene is involved in the hydrolysis of cholesteryl esters and triglycerides and other SNPs within this gene have previously been associated with coronary artery disease.\(^{30}\)

The z-allele within the \(SERPINA1\) gene was associated with a lower %predicted FEV\(_1\) in cases (unadjusted analysis: \(p_{\text{discovery}}=2.83 \times 10^{-5}\); as well as being a well-established cause of AAT deficiency,\(^{3,4}\) this SNP has also previously been associated with an increased annual decline in FEV\(_1\) in a general population sample\(^{31}\) and increased airflow limitation in COPD cases.\(^{32}\) In the present study, the z-allele was associated with an increased risk of COPD, although this was not statistically significant (OR: 1.27, \(p=0.252\)). The likely reason for the lack of a significant association with this known COPD locus is that some of the case collections excluded individuals with AAT deficiency,\(^{34}\) resulting in selection bias. In the meta-analysis of severity of airflow limitation, we identified a very rare SNP within another serine protease inhibitor gene, \(SERPINA12\), not previously associated with COPD (rs140198372, MAF=0.03%, \(p_{\text{meta}}=5.72 \times 10^{-6}\)). \(SERPINA12\) and \(SERPINA1\) lie 96.6 kb apart on chromosome 14 (rs140198372 and the z-allele in \(SERPINA1\) are not in linkage disequilibrium (\(r^2=9.0 \times 10^{-6}\)). \(SERPINA12\) has been associated with cardiovascular diseases, being implicated in obesity and type 2 diabetes.\(^{33}\)

One of the primary challenges associated with identifying low frequency variants associated with disease is limited statistical power, and this could explain our lack of strong statistically significant findings. Indeed, none of the reported associations in novel regions met a stringent exome-wide significance level (\(p<3.8 \times 10^{-7}\)) overall. In the present study, we would have just 54% power to detect an association with an SNP associated with COPD risk with a MAF of 1% and an OR of 2, at the \(p<3.8 \times 10^{-7}\) level. Furthermore, recent analyses undertaken by the UK10K Consortium found no evidence of low frequency SNPs having large effects, upon a series of traits.\(^{34}\) Due to the limited power to detect single variant associations of rare variants with modest effect sizes, we additionally adopted gene-based analyses using SKAT-O, a method which combines information from several rare variants (see online supplementary information). In these analyses, we only identified one gene meeting our elected significance level (\(p<10^{-5}\)); this gene-based signal in \(PRICKLE1\) was found however, to be driven by a single SNP, which was identified as being associated with COPD risk in

![Figure 3](image-url)  
Meta-analysis of COPD risk in discovery exome analysis and UK Biobank Lung Exome Variant Evaluation samples.

**Table 3** Top associations (\(p<10^{-5}\)) in meta-analysis of severity of airflow limitation

<table>
<thead>
<tr>
<th>rs no.</th>
<th>CHR</th>
<th>Position</th>
<th>Coded allele</th>
<th>Gene</th>
<th>MAF (MAC)</th>
<th>Beta (95% CI)</th>
<th>p Value</th>
<th>Severity of airflow limitation, adjusted for pack-years (n=2517)</th>
<th>UK BILEVE pack-years adjusted analysis (n=4231)</th>
<th>Meta-analysis of discovery and UK BILEVE pack-year adjusted analyses</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs140198372</td>
<td>14</td>
<td>94953832 A</td>
<td>SERPINA12 (splice site)</td>
<td>0.059% (3)</td>
<td>−29.23 (−49.50 to −8.96)</td>
<td>2.59×10^{-5}</td>
<td>0.012% (1) −38.35 (−59.88 to −16.82)</td>
<td>4.11×10^{-4}</td>
<td>−33.51 (−48.27 to −18.76)</td>
<td>5.72×10^{-6}</td>
</tr>
</tbody>
</table>

*p Values in bold significant at \(p<10^{-5}\) level.

BILEVE, Biobank Lung Exome Variant Evaluation; MAC, minor allele count; MAF, minor allele frequency.
Further identification of regions not previously implicated in COPD or lung function. We undertook case-control sensitivity analyses using our discovery samples, restricting cases to the subset of 1398 individuals taken from COPD cohorts and who had known irreversible airflow limitation. The effect estimates of our top hits did not substantially change in this sensitivity analysis, suggesting that our broader case definition, including samples that did not undergo reversibility testing, did not result in substantial misclassification bias.

A further potential source of bias in this study was the heavier smoking history in our cases compared with the control samples. For the two SNPs identified through the analyses of COPD risk, we found no evidence of association with smoking in data from the UK BiLEVE study, suggesting that the associations with COPD risk were not driven by the imbalances in smoking behaviour.

Finally, it was not possible to validate the findings of this study through additional genotyping; however for the three reported loci, consistent results were observed in both the discovery and the UK BiLEVE samples. It would not be expected to see the same false positive result in these two independent samples, therefore, strengthening the evidence for these being true associations.

In summary, we have identified potentially interesting associations with low frequency and rare SNPs and COPD risk in two regions not previously implicated in COPD or lung function. We further identified an association of %predicted FEV1 in individuals with COPD with a very rare SNP in SERPINA12. Further confirmation of these associations in larger independent collections of COPD cases and controls is needed. This study also provides further evidence that the z-allele within SERPINA1 may be related to severity of airflow limitation in COPD. While large sample sizes may be required to definitively identify novel loci, we present evidence to support the notion that the genetic contribution to COPD risk comprises polygenic contributions of rare, low frequency and common genetic variants. Future studies, alone or in combination, should aim to target the full allele frequency range to unravel the genetic architecture of COPD.

Author affiliations
1Department of Health Sciences, University of Leicester, Leicester, UK
2William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK
3Division of Respiratory Medicine, University of Nottingham, Queen’s Medical Centre, Nottingham, UK
4School of Social & Community Medicine, University of Bristol, Bristol, UK
5Department of Primary Care & Population Health, UCL, London, UK
6Population Health Research Institute, St George’s, University of London, London, UK
7University College London, Farr Institute of Health Informatics, London, UK
8Cochrane Heart Group, London, UK
9Department of Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, London, UK
10ISER, University of Essex, Colchester, Essex, UK
11Department of Epidemiology and Public Health, UCL, London, UK
12Institute of Cardiovascular Science, UCL, London, UK
13School of Life Sciences, University of Nottingham, Nottingham, UK
14Division of Epidemiology and Public Health, Nottingham City Hospital, University of Nottingham, Nottingham, UK
15Cardiovascular and Diabetes Medicine, School of Medicine, University of Dundee, Dundee, UK
16Human Development & Health, Faculty of Medicine, University of Southampton, Southampton General Hospital, Southampton, UK
17NIHR Southampton Respiratory Biomedical Research Unit, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton General Hospital, Southampton, UK
18MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, Southampton, UK
19NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton General Hospital, Southampton, UK
20Victoria University, Wellington, New Zealand
21Respiratory Medicine Unit, Nuffield Department of Medicine, University of Oxford, Oxford, UK
22Institute for Lung Health, Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, UK
23National Institute for Health Research Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester, UK
24Nottingham Respiratory Research Unit, University of Nottingham, City Hospital Campus, Nottingham, UK
25Institute for Ageing and Health, Newcastle University, Campus for Ageing and Vitality, Newcastle upon Tyne, UK
26Department of Molecular Genetics and Genomics, National Heart and Lung Institute, Imperial College London, London, UK
27Freemasons’ Department of Geniatric Medicine, University of Auckland, New Zealand
28Generation Scotland, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
29Division of Population Health Sciences, University of Dundee, Dundee, UK
30Institute of Cardiovascular and Medical Sciences, University of Glasgow, Glasgow, UK
31Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK
32Department of Haematology, University of Cambridge, Cambridge, UK
33Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia

Acknowledgements This research used the ALICE and SPECTRE High Performance Computing Facilities at the University of Leicester and was supported by the National Institute for Health Research (NIHR) Leicestershire Respiratory Biomedical Research Unit. This article/paper/report presents independent research funded partially by the NIHR. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. This research has been conducted using the UK Biobank Resource.

Contributors Case collection study concept, or data acquisition and quality control: JS, IPH, DJS, RM, FW, JPC, AA, MC, CD, MK, JE, NK, SC, TGB, TMM, CNAP, RT, JW, AAS, EMD, CC, MB, BB, CB, CEB, MEJ, SGP, MFM, AWJ, MJC, BIP, BHS, SP and LH. Genotype data acquisition and QC: KES and PD. Central study design, analysis and writing of manuscript: VEJ, IN, LW, MDPH and IS.

Funding British Women’s Heart and Health Study is funded by the Department of Health grant no. 90049 and the British Heart Foundation grant no. PG09/022. British Regional Heart Study is supported by the British Heart Foundation (grant RG/ 13/16/30528). CB (COPDBEAT) received funding from the Medical Research Council UK (grant no. G0601369), CB (COPDBEAT) and AW (UKCOPD) were supported by the National Institute for Health Research (NIHR Leicestershire Biomedical Research Unit). MB (COPDBEAT) received funding from the NIHR Foundation (grant no. PDF-2013-06-052). Hertfordshire Cohort Study received support from the Medical Research Council, Arthritis Research UK, the International Osteoporosis Foundation and the British Heart Foundation; NIHR Biomedical Research Centre in Nutrition, University of Southampton; NIHR Musculoskeletal Biomedical Research Unit, University of Oxford. Generation Scotland: Scottish Family Health Study is funded by the Chief Scientist Office, Scottish Government Health Directorates, grant number CZD/16/6 and the Scottish Funding Council grant HR30036. EU COPD Gene Scan is funded by the European Union, grant no. QLG1-CT-2001-01012. English Longitudinal Study of Ageing is funded by the Institute of Aging, NIH grant no. AGT17644061. GoDARTs was funded by the Wellcome Trust grants 072960, 084726 and 104970. MDT has been supported by MRC fellowship GD902313. UK Biobank Lung Exome Variant Evaluation study was funded by a Medical Research Council scientific award to MDT, IPH, DJS and LW (MC_PC_12010).

Competing interests None declared.

Ethics approval Several (meta-analysis design).

Provenance and peer review Not commissioned; externally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/
REFERENCES


