Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank


Summary

Background Understanding the genetic basis of airflow obstruction and smoking behaviour is key to determining the pathophysiology of chronic obstructive pulmonary disease (COPD). We used UK Biobank data to study the genetic causes of smoking behaviour and lung health.

Methods We sampled individuals of European ancestry from UK Biobank, from the middle and extremes of the forced expiratory volume in 1 s (FEV1) distribution among heavy smokers (mean 35 pack-years) and never-smokers. We developed a custom array for UK Biobank to provide optimum genome-wide coverage of common and low-frequency variants, dense coverage of genomic regions already implicated in lung health and disease, and to assay rare coding variants relevant to the UK population. We investigated whether there were shared genetic causes between different phenotypes defined by extremes of FEV1. We also looked for novel variants associated with extremes of FEV1, and smoking behaviour and assessed regions of the genome that had already shown evidence for a role in lung health and disease. We set genome-wide significance at p<5×10⁻⁸.

Findings UK Biobank participants were recruited from March 15, 2006, to July 7, 2010. Sample selection for the UK BiLEVE study started on Nov 22, 2012, and was completed on Dec 20, 2012. We selected 50 008 unique samples: 10 002 individuals with low FEV1, 10 000 with average FEV1, and 5002 with high FEV1 from each of the heavy smoker and never smoker groups. We noted a substantial sharing of genetic causes of low FEV1 between heavy smokers and never smokers (p=2·29×10⁻¹⁶) and between individuals with and without doctor-diagnosed asthma (p=6·06×10⁻¹¹). We discovered six novel genome-wide significant signals of association with extremes of FEV1, including signals at four novel loci (KANSL1, TSEN54, TET2, and RBM19/TBX5) and independent signals at two previously reported loci (NPNT and HLA-DQBI/HLA-DQA2). These variants also showed association with COPD, including in individuals with no history of smoking. The number of copies of a 150 kb region containing the 5′ end of KANSL1, a gene that is important for epigenetic gene regulation, was associated with extremes of FEV1. We also discovered five new genome-wide significant signals for smoking behaviour, including a variant in NCAM1 (chromosome 11) and a variant on chromosome 2 (between TEX41 and PABPCIP2) that has a trans effect on expression of NCAM1 in brain tissue.

Interpretation By sampling from the extremes of the lung function distribution in UK Biobank, we identified novel genetic causes of lung function and smoking behaviour. These results provide new insight into the specific mechanisms underlying airflow obstruction, COPD, and tobacco addiction, and show substantial shared genetic architecture underlying airflow obstruction across individuals, irrespective of smoking behaviour and other airway disease.

Funding Medical Research Council.

Introduction Chronic obstructive pulmonary disease (COPD) is a global public health concern and is currently the third leading cause of death worldwide. Smoking and indoor air pollution are major environmental risk factors for development of COPD, but heritability studies also suggest a strong genetic component in smoking behaviour and in risk of COPD. Spirometry, particularly measurements of forced expiratory volume in 1 s (FEV1) and forced vital capacity (FVC), is used to measure airflow obstruction and helps in the diagnosis and grading of severity of COPD. Previous large genome-wide
Evidence before this study
UK Biobank had completed its recruitment, including its baseline phenotyping and biobanking of samples, before our study began. The DNA had not yet been extracted from the biobanked samples, and the spirometry data quality had not yet been analysed across all UK Biobank participants. We searched for evidence of other large biobanks with spirometry data, including the P3G Catalogue. We did not identify any other biobank with spirometry data and DNA as large as UK Biobank. Evidence regarding the global burden of disease due to smoking or chronic obstructive pulmonary disease (COPD) was obtained from the WHO Global Health Risks Report and a systematic analysis of the Global Burden of Disease Study 2010. Tobacco smoking accounted for about 5·1 million deaths globally in 2004, because of recent increases in smoking prevalence in developing countries, the full global effect of smoking is yet to occur. COPD is the third leading cause of death globally. For previous evidence of genetic associations, we gave the highest ranking to associations reaching genome-wide significance in genome-wide association studies, a lower ranking to associations not reaching genome-wide significance in genome-wide association studies, and the lowest ranking to associations reported in candidate gene studies. To assess evidence of loci associated with lung function, COPD, and smoking behaviour, we queried the Catalog of Published Genome-Wide Association Studies. We used this evidence to report our known findings for genetic variants shown to be associated with forced expiratory volume in 1 s (FEV1; eight loci) and smoking behaviour (seven loci). We report candidate gene associations only for variants for which we found genome-wide evidence of association.

Added value of this study
We describe, to our knowledge, the first genetic study using the UK Biobank resource and show the quality of the phenotype and genotype data. Additionally, we describe an advance in imputation quality afforded by the use of a newly designed genotyping array used in conjunction with the largest reference panel available so far. A slightly modified version of this array is being used to genotype the remaining samples in UK Biobank. As evidence of the usefulness of these data, we describe novel insights into the genetic architecture of airflow obstruction and smoking. Specifically, we show that there are shared genetic causes of airflow obstruction between smokers and non-smokers, consistent with the limited evidence for gene-smoking interactions described so far. We show that the genetic determinants of low FEV1 in individuals without asthma are also informative in individuals with asthma. We report new loci associated with extremes of FEV1 and COPD, including evidence that a genomic region of complex structural variation has an effect on lung function and airflow obstruction in the general population. Our novel signals implicate epigenetic mechanisms as contributors to lung health. These findings, taken together with previous findings, will help define pathways underlying predisposition to development of COPD and smoking behaviours. A full understanding of the biological mechanisms underlying these genetic associations will improve our understanding of the pathophysiology of COPD and smoking behaviour, and potentially give rise to novel therapeutic strategies for the management of airway disease and prevention of nicotine addiction.

Implications of all the available evidence
This study has improved our understanding of the genetic and molecular basis of smoking behaviour and lung function and provided potential targets for therapeutic intervention. It has also shown the value of sampling from the extremes using a large biobank such as UK Biobank. A similar approach could be adopted for genetic studies of other health-related traits in UK Biobank, using either new genetic assays or the extensive genome-wide data that we and UK Biobank have generated.

Research in context
This article describes the first genome-wide association study of FEV1 in population-based samples. The study was conducted using the UK Biobank resource, which contains spirometry data and DNA as large as any other biobank. The study identifies 12 genetic variants associated with FEV1, including eight previously reported in genome-wide association studies. The study also identifies 12 genetic variants associated with smoking behaviour. The findings provide potential targets for therapeutic intervention and improve our understanding of the pathophysiology of COPD and smoking behaviour.

Implications of all the available evidence
This study has improved our understanding of the genetic and molecular basis of smoking behaviour and lung function and provided potential targets for therapeutic intervention. It has also shown the value of sampling from the extremes using a large biobank such as UK Biobank. A similar approach could be adopted for genetic studies of other health-related traits in UK Biobank, using either new genetic assays or the extensive genome-wide data that we and UK Biobank have generated.

Articles

Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK (A Ramasamy DPhil), Department of Medical and Molecular Genetics, King’s College London, London, UK (A Ramasamy), Jenner Institute (A Ramasamy), Department of Statistics (J O’Conell DPhil), E Frangou MSc[Res], O Delaneau PhD, Prof F Marchini DPhil, Wellcome Trust Centre for Human Genetics (J O’Conell), C Freeman PhD, D Petkova PhD, Prof M McCarthy MD, Prof F Marchini, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (E Frangou), Oxford Centre for Diabetics, Endocrinology and Metabolism (Prof M McCarthy), and Respiratory Medicine (I Favier FMedSci), University of Oxford, Oxford, UK; Estonian Genome Center, University of Tartu, Tartu, Estonia (R Mag PhD, E Mihalov MSc, E Reinmaa MSc), Institute of Environmental Medicine, Karolinska Institutet and Sachs’ Children’s Hospital, Stockholm, Sweden (E Melén MD), Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland (O Delaneau); William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University London, London, UK (Prof P Deloukas PhD); Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdullah University, Jeddah, Saudi Arabia (Prof P Deloukas); UK Small Area Health Statistics Unit, MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK (A L Hansell); Imperial College Healthcare NHS Trust, St Mary’s Hospital, Paddington, London, UK (A L Hansell); Institute of Infection, Immunity and Inflammation, University of Glasgow, Glasgow, UK (Prof N C Thomson MD); Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK (E Zeggini PhD); Department of Biostatistics, University of Liverpool, Liverpool, UK

Association studies (GWAS) of general population cohorts have identified 32 common genetic variants (minor allele frequency [MAF] >5%) associated with lung function,12 of which have also shown association with airflow obstruction and risk of COPD.10,11 However, these findings only explain a small proportion of the phenotypic variance (~1·5% for FEV1).17–19 Further insight into the genetic factors affecting lung function, smoking behaviour, and COPD could lead to new approaches for smoking cessation and prevention and treatment of COPD.

UK Biobank is the largest European biobank available at present and represents an extensive resource from which association studies (GWAS) of general population cohorts have identified 32 common genetic variants (minor allele frequency [MAF] >5%) associated with lung function,12 of which have also shown association with airflow obstruction and risk of COPD.10,11 However, these findings only explain a small proportion of the phenotypic variance (~1·5% for FEV1).17–19 Further insight into the genetic factors affecting lung function, smoking behaviour, and COPD could lead to new approaches for smoking cessation and prevention and treatment of COPD.

UK Biobank is the largest European biobank available at present and represents an extensive resource from which association studies (GWAS) of general population cohorts have identified 32 common genetic variants (minor allele frequency [MAF] >5%) associated with lung function,12 of which have also shown association with airflow obstruction and risk of COPD.10,11 However, these findings only explain a small proportion of the phenotypic variance (~1·5% for FEV1).17–19 Further insight into the genetic factors affecting lung function, smoking behaviour, and COPD could lead to new approaches for smoking cessation and prevention and treatment of COPD.
Methods

Study design
We defined case and control groups by selecting individuals from the middle and extremes of the FEV1 distribution among both heavy smokers (mean 35 pack-years) and never smokers. We developed a custom array to provide optimum genome-wide coverage of common and low frequency (MAF 1–5%) coding variants and rare (MAF <1%) coding variants relevant to the UK population; this platform also provided dense coverage of genomic regions implicated in lung health and disease. Spirometry data in UK Biobank were obtained using a Vitalograph Pneumotrac 6800 (Buckingham, UK) on at least two occasions. Sampling was undertaken such that equal numbers of males and females were selected in total and the numbers of individuals selected from each age–sex band were proportional to the number of individuals in the band being sampled (appendix pp 3–5). One consequence of this approach is that we enriched our sample for non-smoking individuals with airflow obstruction.

To assess whether the novel regions that we identified as associated with FEV1 extremes are also associated with COPD, we defined individuals fulfilling spirometric criteria for the Global Initiative for Chronic Obstructive Lung Disease (GOLD) Stage 2+ COPD (FEV1:FVC ratio criteria for the Global Initiative for Chronic Obstructive smoking individuals with airflow obstruction. To identify whether any individual variants had a significantly different effect on the risk of airflow obstruction in heavy smokers compared with never smokers, we tested for interaction with smoking behaviour, from the middle and extremes of the FEV1 distribution we defined case and control groups by selecting individuals with FEV1 strata as controls. Post-bronchodilator spirometry was withheld before spirometry.

Project Phase 122 and UK10K Project 23,24 reference panel quality control (appendix pp 8–15), we imputed non-genotyped variants using a combined 1000 Genomes Project Phase 12 and UK10K Project p15–16, the data were used to finalise the design of the UK Biobank array, which is being used for genome-wide genotyping and imputation of the remaining UK Biobank participants.

Using data from previously published studies of whole-genome gene expression and genome-wide genotyping,25–28 we assessed whether variants at associated loci (identified as described in the Statistical analysis) regulated levels of mRNA. These expression quantitative trait loci (eQTL) studies included non-tumour lung tissue, blood, and, for variants associated with smoking behaviour, brain. For genes close to peaks of novel signals or genes implicated through eQTL, we assessed differential expression in the lungs of individuals with and without COPD and differential expression in the pseudoglandular and canalicular stages of development of the fetal lung.30,31 Additionally, we generated RNA sequencing data to discover novel transscripts of these genes in human bronchial epithelical cells. We tested all genome-wide meta-analysis p values for enrichment in biological pathways defined in publicly available databases. All functional analyses are described in detail in the appendix (pp 21–23).

Statistical analysis
Case-control comparisons of low FEV1 versus high FEV1, low FEV1 versus average FEV1, and high FEV1 versus average FEV1, were done within each of the heavy and never smokers subsets separately (appendix p 17). To identify whether any individual variants had a significantly different effect on the risk of airflow obstruction in heavy smokers compared with never smokers, we tested for interaction with smoking (appendix p 17). We calculated the proportion of the variance in FEV1, explained by genetic variants (appendix p 17). We compared heavy versus never smokers to identify loci associated with smoking behaviour. Association testing was done using a Score test (and Firth test for variants with minor allele count <400)32 with imputed marker doses, adjusting for pack-years in smokers and ten principal components. Full genome-wide association results are available via UK Biobank (appendix p 17). For genome-wide association analyses, we set genome-wide significance as p<5×10⁻⁸ and suggestive significance as 5 × 10⁻⁶<p<5 × 10⁻⁷. For other analyses, we used a Bonferroni correction for multiple testing. The appendix (p 17) describes quality control after association testing. For the lead single nucleotide polymorphism (SNP) at each of our novel signals of association with FEV1,extremes, we tested for association with COPD risk using the aforementioned definition (appendix p 18). We did a meta-analysis across smoking strata using inverse variance weighting. We assessed evidence for polygenic architecture of FEV1-defined...
For this analysis, we created discovery and target subpopulations, each of which comprised cases and control groups created by randomly splitting the low FEV$_1$ and average FEV$_1$ groups (appendix pp 18–20). Variants of MAF of at least 1% associated with low FEV$_1$ below given p value.
thresholds in the discovery population were incorporated into an aggregate score, and the association with the aggregate score was tested in the independent target population. A similar approach (appendix pp 18–20), in each case using independent discovery and target populations, was used to test for a shared polygenic component between high FEV1 and low FEV1, low FEV1, in heavy smokers and in never smokers, and low FEV1, in participants who did and those who did not report a history of doctor-diagnosed asthma. To show the reliability of the doctor diagnosis of asthma variable, we showed association with asthma at ten previously reported genome-wide significant loci (appendix pp 21, 29).

Role of the funding source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all the data in the UK BiLEVE study and had final responsibility for the decision to submit for publication.

Results

UK Biobank participants were recruited from March 15, 2006, to July 7, 2010. Sample selection for the UK BiLEVE study started on Nov 22, 2012, and was completed on Dec 20, 2012. We initially selected 50008 unique samples representing the extremes and middle of the percent predicted FEV1 distributions; this comprised 10002 individuals with low percent predicted FEV1, 10000 individuals with average percent predicted FEV1, and 5002 individuals with high percent predicted FEV1, from each of the heavy smoker and never smoker groups (figure 1). Within this dataset, 48931 unrelated individuals passed quality control and were included in subsequent analyses.

We undertook genome-wide genotyping of 807,411 variants. After filtering, genome-wide imputation using the 1000 Genomes Project Phase 1 and UK10K Project reference panel resulted in 42795484 variants. Our final dataset for analysis, after excluding variants with information quality less than 0.5 or minor allele count less than three, comprised 28509962 imputed or genotyped variants in 48931 unrelated individuals (table 1 and appendix pp 16 and 95).27

Using independent discovery and target subpopulations to generate and test risk scores, we found that the association of low FEV1, versus average FEV1, with the risk score became stronger for increasingly liberal p value thresholds of association in the discovery population (p=6.24×10⁻¹⁶ for a p value threshold of 0.5). This finding suggests a polygenic component to low FEV1, in which many variants of individually small effect size contribute to the risk of low FEV1. We found substantial sharing of genetic causes across thousands of genetic variants between low FEV1, in heavy smokers and low FEV1, in never smokers (p=2.29×10⁻²²; p value threshold <0.5; figure 2; appendix p 30). Similarly, we found substantially overlapping genetic causes for low FEV1, in participants reporting a history of doctor-diagnosed asthma and low FEV1, in those without asthma (p=6.06×10⁻¹¹; p value threshold <0.5; figure 2; appendix p 31). Finally, overlapping genetic causes were shown for high FEV1, and low FEV1, (p=1.64×10⁻²²; p value threshold <0.5; figure 2; appendix p 30).

In addition to detecting signals of association previously reported by studies of quantitative lung function (appendix p 32–37), in our case-control analysis of FEV1, extremes we identified six novel signals of association (p<5×10⁻⁹) with low FEV1, versus high FEV1, (table 2; figure 3; appendix pp 38–53, 96–98, 102). The sentinel SNPs at five of these six signals, in or near TET2, NPNT, HLA-DQB1/HLA-DQA2, KANSL1, and TSEN54, were common (MAF ≥5%) and showed a stronger association with low FEV1, in never smokers than heavy smokers. The sentinel SNP at an intergenic signal between RBM19 and TBX5 was a rare variant (MAF=0.13%) that showed strongest association with low FEV1, in heavy smokers. The lead SNPs at each of these loci showed association with COPD (table 2; appendix pp 30, 31).
Table 2: Novel genome-wide significant signals of association with extremes of forced expiratory volume in 1 s (FEV₁) smoking behaviour

For variants that showed association with extremes of FEV₁, in either heavy smokers or never smokers, the results from both the never smokers and heavy smokers are presented. For variants that had genome-wide significant evidence of association with smoking behaviour (and not for extremes of FEV₁), association with COPD and effect on FEV₁ were not assessed. Chromosome and position relate to National Center for Biotechnology Information build 37 (hg19). *Indication of certainty of imputation for this variant; an imputation info score of 1 suggests a variant imputed with the highest certainty or a directly genotyped variant. †In samples included in the comparison. ‡Analysis of Global Initiative for Chronic Obstructive Lung Disease stage 2+ COPD cases versus controls (heavy smokers: 5803 cases, never smokers: 3761 cases genotyped variant. ¶Novel signals of association within previously reported loci. ||Directly genotyped.

For variants that showed association with extremes of FEV₁, in either heavy smokers or never smokers, the results from both the never smokers and heavy smokers are presented. For variants that had genome-wide significant evidence of association with smoking behaviour (and not for extremes of FEV₁), association with COPD and effect on FEV₁ were not assessed. Chromosome and position relate to National Center for Biotechnology Information build 37 (hg19). *Indication of certainty of imputation for this variant; an imputation info score of 1 suggests a variant imputed with the highest certainty or a directly genotyped variant. †In samples included in the comparison. ‡Analysis of Global Initiative for Chronic Obstructive Lung Disease stage 2+ COPD cases versus controls (heavy smokers: 5803 cases, never smokers: 3761 cases genotyped variant. ¶Novel signals of association within previously reported loci. ||Directly genotyped.

Table 2: Novel genome-wide significant signals of association with extremes of forced expiratory volume in 1 s or smoking behaviour

<table>
<thead>
<tr>
<th>Locus</th>
<th>Non-coded/ coded allele (minor allele)</th>
<th>Imputation info score*</th>
<th>Smoking status</th>
<th>MAF (MAC)†</th>
<th>p value (genomic control corrected)</th>
<th>Association with COPD‡</th>
<th>Effect on FEV₁ (95% CI)</th>
<th>p value</th>
<th>Beta (SE)$</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chr1:114743533</td>
<td>RBM19/7BKS</td>
<td>T/C (T)</td>
<td>0.737</td>
<td>Never</td>
<td>0.002 (60)</td>
<td>0.97 (0.57–1.67)</td>
<td>0.90 (0.54–2.51)</td>
<td>1.16 (0.54–2.51)</td>
<td>0.71 (0.54–2.51)</td>
<td>0.010 (0.118)</td>
</tr>
<tr>
<td>Chr6:12635592</td>
<td>HLA-DQB1/ HLA-DQA2</td>
<td>A/G (G)</td>
<td>0.962</td>
<td>Never</td>
<td>0.472 (13 838)</td>
<td>1.18 (1.11–1.25)</td>
<td>1.10 × 10⁻⁸ (1.11–1.25)</td>
<td>1.24 (1.16–1.32)</td>
<td>1.95 × 10⁻⁹ (1.88–1.34)</td>
<td>0.056 (0.009)</td>
</tr>
<tr>
<td>Chr12:114743533</td>
<td>KANSL1</td>
<td>A/G (G)</td>
<td>0.976</td>
<td>Never</td>
<td>0.242 (7 088)</td>
<td>1.22 (1.15–1.29)</td>
<td>1.66 × 10⁻³ (1.16–1.34)</td>
<td>1.24 (1.02–1.14)</td>
<td>8.58 × 10⁻⁹ (1.02–1.14)</td>
<td>0.009 (0.009)</td>
</tr>
<tr>
<td>Chr7:12635592</td>
<td>NPNT</td>
<td>G/A (A)</td>
<td>1.000</td>
<td>Never</td>
<td>0.268 (7 842)</td>
<td>1.27 (1.20–1.34)</td>
<td>9.62 × 10⁻¹⁰ (1.27–1.34)</td>
<td>1.36 (1.27–1.34)</td>
<td>2.10 × 10⁻⁹ (1.27–1.34)</td>
<td>0.007 (0.009)</td>
</tr>
<tr>
<td>Chr1:99445471</td>
<td>LPPR5</td>
<td>C/T (T)</td>
<td>1.000</td>
<td>NA</td>
<td>0.096 (2 379)</td>
<td>1.05 (1.00–1.10)</td>
<td>0.83 × 10⁻⁸ (1.00–1.10)</td>
<td>1.24 (1.02–1.14)</td>
<td>8.58 × 10⁻⁹ (1.02–1.14)</td>
<td>0.009 (0.009)</td>
</tr>
<tr>
<td>Chr12:114743533</td>
<td>TSEN54</td>
<td>C/A (C)</td>
<td>0.997</td>
<td>Never</td>
<td>0.291 (8 538)</td>
<td>1.18 (1.11–1.25)</td>
<td>1.18 × 10⁻⁸ (1.11–1.25)</td>
<td>1.22 (1.14–1.31)</td>
<td>4.56 × 10⁻⁹ (1.14–1.31)</td>
<td>0.005 (0.005)</td>
</tr>
<tr>
<td>Chr10:106137033</td>
<td>TET2</td>
<td>G/A (G)</td>
<td>0.998</td>
<td>Never</td>
<td>0.345 (10 117)</td>
<td>1.17 (1.11–1.23)</td>
<td>1.31 × 10⁻⁸ (1.11–1.23)</td>
<td>1.70 (1.10–1.25)</td>
<td>1.64 × 10⁻⁹ (1.10–1.25)</td>
<td>0.056 (0.009)</td>
</tr>
<tr>
<td>Chr11:114743533</td>
<td>TCONS12</td>
<td>T/C (T)</td>
<td>0.998</td>
<td>Never</td>
<td>0.356 (10 440)</td>
<td>1.07 (1.02–1.13)</td>
<td>8.01 × 10⁻³ (1.02–1.13)</td>
<td>1.09 (1.03–1.16)</td>
<td>2.92 × 10⁻³ (1.03–1.16)</td>
<td>0.003 (0.009)</td>
</tr>
</tbody>
</table>

For variants that showed association with extremes of FEV₁, in either heavy smokers or never smokers, the results from both the never smokers and heavy smokers are presented. For variants that had genome-wide significant evidence of association with smoking behaviour (and not for extremes of FEV₁), association with COPD and effect on FEV₁ were not assessed. Chromosome and position relate to National Center for Biotechnology Information build 37 (hg19). *Indication of certainty of imputation for this variant; an imputation info score of 1 suggests a variant imputed with the highest certainty or a directly genotyped variant. †In samples included in the comparison. ‡Analysis of Global Initiative for Chronic Obstructive Lung Disease stage 2+ COPD cases versus controls (heavy smokers: 5803 cases, never smokers: 3761 cases genotyped variant. ¶Novel signals of association within previously reported loci. ||Directly genotyped.
The second (rs2047409) was 552 kb from the signals at INTS12, GSTCD, and NPNT and was localised to TET2. The signal of association for rs34712979 was strongest in never smokers \( (p=9.62 \times 10^{-16}; \text{table 2}) \) and was weakly correlated (linkage disequilibrium \( r^2=0.31 \)) with another SNP in NPNT—rs6856422—which has also been identified as a novel secondary signal of association at 4q24 by an independent concurrent study of lung function in the general population.35 When rs6856422 was included as a covariate in our analysis, the signal for rs34712979
The novel signal at TET2 (rs2047409) also showed strongest association in never smokers (p=1·31×10–⁸). This included as a covariate; rs2047409, p=9·8 × 10 –⁹). TET2 encodes tet methylcytosine dioxygenase 2, which has a role in myelopoiesis, and SNPs in TET2 have shown association with height.11 TET2 was differentially expressed during fetal lung development (appendix pp 58–59).

We detected a signal of association with FEV1 extremes within the HLA region on chromosome 6 that was correlated with a previously reported signal of association with asthma.37 The signal we report was strongest in never smokers (rs9274600, p=1·18 × 10–⁸; table 2). This signal was strongest in never smokers (rs9274600, p=1·31 × 10–⁸). rs2047409 is included as a covariate; rs2047409, p=9·8 × 10 –⁹). The novel signal at TET2 (rs92743349)—included as a covariate—with asthma-associated SNP rs927334937—included as a covariate. The odds ratio for rs9274600 decreased from 1·18 (95% CI 1·11–1·25) to 1·14 (1·08–1·20), but remained significant (p=3·25×10–⁶; appendix p 103). This signal is independent of nearby signals reported for lung function,5,7,8 including rs7764819,38 which is 45 kb from rs9274600 (association for rs9274600 conditioned on TTEN54 trans eQTL for NCAM1 (chromosome 11), and rs10193706, an intergenic SNP on chromosome 2 downstream of the 300 kb duplication region found only in individuals who do not carry the inversion) was associated with extremes of FEV1 (p=2·40 × 10–⁶; appendix p 71). The sentinel SNP rs2532349 lies within this region.

A second signal of association with FEV1, extremes on chromosome 17 (17q25.1) was within TSEN54 and occurred only in never smokers (rs7218675; p=1·18 × 10–⁸; table 2). TSEN54 encodes a subunit of the tRNA splicing endonuclease complex, and rs7218675 was associated with expression of KIAA0195, TSEN54, and GRB2 in blood and expression of GRB2 in lung tissue (appendix pp 63–66). GRB2 is a ligand of the epidermal growth factor receptor, which links signalling by epidermal growth factor with the MAPK/ERK signalling pathway, triggering cell proliferation. RNA sequencing in human bronchial epithelial cells identified a relatively abundant novel fusion transcript of TSEN54 and LILG2 (appendix p 108).

To corroborate the new signals we identified in TET2 and TSEN54, we present evidence of association with FEV1, in a previously reported study4 of 48 201 individuals (p=9·9× 10–⁵ and p=0·006, respectively; appendix pp 24, 82). We identified a further 21 loci with suggestive (5×10–⁸<p<5×10–⁷) evidence of association with FEV1, extremes (appendix pp 72–73), including six rare variants with a minor allele count less than 400. These included signals in CCDC91 and RSRC1, both of which showed genome-wide significant association with lung function in an independent concurrent study of lung function in the general population.35

By comparing heavy smokers and never smokers, we identified five novel regions of association with smoking behaviour and confirmed four previously reported loci (15q25, 7p14, DBH, and BDNF; table 2; appendix pp 32–37).5,13 The novel signals included rs4466874, in an intron of NCAM1 (chromosome 11), and rs10193706, an intergenic SNP on chromosome 2 downstream of TEX41 and upstream of PABPC1P2 (table 2; appendix pp 110–112). Uncorrelated (r²=0·0001 with rs4466874) SNPs in TTC12 and ANKK1, near to NCAM1, have also previously shown association with nicotine dependence (appendix pp 74–75). A proxy of rs10193706 on chromosome 2 (rs953246, r²=0·48) is a trans eQTL for NCAM1 on chromosome 11 in brain tissue (appendix pp 76–77). Another proxy of rs10193706 on chromosome 2 (rs12622738, r²=0·86) is a trans eQTL in the substantia
nigra for WDR61 on chromosome 15, 300 kb from the established 15q25 smoking locus (appendix pp 76–77).

We also noted novel genome-wide significant signals of association with smoking behaviour in NOL4L, LPPRS, and DNAH8 (table 2, appendix pp 110–112). A SNP in C20orf203, near to NOL4L, but independent of our sentinel variant, has previously been implicated in nicotine dependence.44 We identified secondary independent signals, which did not reach genome-wide significance, at three of the loci associated with smoking behaviour (appendix pp 56–57, 99–101),45 including a novel rare (MAF=0.09%) intergenic SNP near NCAM1. For novel signals for smoking behaviour, we did a meta-analysis of summary statistics from two previous, less powerful studies7,8 and found corroborative evidence for NCAM1, TEX41/PABPC1P2, and NOL4L (eg, for smoking initiation p=0.0003, p=0.017, and p=0.0006, respectively; appendix pp 24, 83–84). We identified a further eight loci with suggestive (5×10⁻⁸<p<5×10⁻⁷) evidence of association with smoking behaviour (appendix pp 72–73), including CHRNA4 at 20q13.33 (p=1.01×10⁻⁷).47

In a genome-wide gene–smoking interaction analysis, although common SNPs on chromosomes 6 and 19 showed suggestive SNP–smoking interactions (p<5×10⁻⁷; appendix p 78), no gene–smoking interactions were detected at genome-wide significance (p<5×10⁻⁸). Three of the six variants associated with FEV1, extremes (table 2) showed weak evidence of interaction with smoking (Bonferroni correction for six tests p<0.0083; appendix pp 60–62), including common variants at the HLA-DQB1/HLA-DQA2 and TSEN54 loci and the rare variant at the RBM19/TBX5 locus. In a meta-analysis of the genome-wide association test statistics for low FEV1, versus high FEV1, across heavy and never smokers, motivated by our finding of shared genetic causes between heavy and never smokers and to increase the sample size, we identified an additional six novel genome-wide significant signals of association with FEV1, extremes. These included CCDC91, reported as a novel signal of association with lung function in the general population by a concurrent study,49 and SLMAP, for which there is corroborative evidence of association with lung function (appendix pp 79–80).50 Our pathway analysis identified a novel signal of enrichment of the histone subset of the chromatin packaging and remodelling process gene set, which was independently replicated in a concurrent GWAS of lung function in the general population (appendix p 81).51 Replication of a previously reported enrichment of the systemic lupus erythematosus pathway was also noted (appendix p 81).

**Discussion**

We describe, to our knowledge, the first genetic association analyses in UK Biobank, targeting the genetic architecture of smoking behaviour and lung function phenotypes. By sampling from the extremes of the FEV1, and smoking phenotype distributions, we identified novel associations for FEV1 and smoking behaviour. We show genome-wide evidence for shared genetic causes of low FEV1 between heavy smokers and never smokers. Furthermore, our analyses suggest that smoking is only likely to interact with a small proportion of the genetic effects we have identified on lung function—that is, smoking and genetic effects generally act separately. We also show shared genetic causes of airflow obstruction between participants who reported doctor-diagnosed asthma and those who did not.

Two of our novel signals of association with smoking behaviour implicate NCAM1; one SNP lies within an intron of NCAM1 and a second variant, located distantly on chromosome 2, is a trans eQTL for NCAM1 in brain tissue (medulla)—ie, it is associated with the level of expression of NCAM1. This second SNP is also a trans eQTL in substantia nigra tissue for another gene called WDR61, which is close to the genes CHRNA3 and CHRNA5 at 15q25—a locus strongly associated with smoking behaviour.60–61 The substantia nigra plays an important part in reward and addiction,62 but little is known about WDR61 other than that expression can be induced by mechanical strain in mesenchymal stem cells.63

We describe six new signals of association with FEV1, extremes, all of which were also associated with COPD using our definition based on spirometry. Five of these signals were most strongly associated with extremes of FEV1 (low vs high) in never smokers. The signal at 17q21.31 for extremes of FEV1 suggests a role for structural variation and epigenetic regulation in lung health. We found that the number of copies of the 5′ end of KANSL1—a gene disrupted by duplication events—is associated with FEV1, extremes. KANSL1 encodes a protein that is a key component of the NSL1 (histone acetyltransferase) complex.64 The disruption of the gene gives rise to a novel truncated transcript,65 which encodes a protein missing a domain essential for key interactions with other proteins important for NSL1 function.66 Therefore, widespread effects on gene regulation through altered histone acetylation could underlie this association. Reduced expression of KANSL1 causes a rare multisystem disorder,67,68 suggesting an essential role for KANSL1 in epigenetic regulation. In a genome-wide pathway analysis, we identified the histone gene set, further implicating a role for epigenetic regulation in lung health.

We maximised the power of our study by sampling from the extremes of a large biobank. No other similar resources of a comparable size were available for replication studies. Nevertheless, the novel genome-wide significant signals of association with FEV1 extremes in NPNT and KANSL1 in never smokers were also significantly associated in the independent set of heavy smokers. Furthermore, to corroborate the new signals we identified in TET2 and TSEN54, we present evidence of association with FEV1 in a previously reported large study.
of FEV₁, in the general population.² However, the rare SNP on chromosome 12 for which we found association with extremes of FEV₁ was exclusive to the recently released UK10K Project component of the imputation panel and has not yet been measured in suitably large studies. Our comparison of smokers and never smokers represents a powerful approach because of the restriction to heavy smokers rather than ever smokers. For novel signals for smoking behaviour, we present additional evidence of association with smoking behaviour for NCAM1, TEX41/PAPPCIP2, and NOL4L in independent populations.³⁷⁹ Although these independent datasets have limited power, they provide corroboration of key genome-wide significant findings in UK BiLEVE.

One of the strengths of our study design was that the genotyping platform we used allowed for fine mapping of regions already known to contain genetic variants that affect lung function. For example, we were able to identify a novel signal in NPNT that was independent of the previously reported signal of association at this locus (spanning GSTCD, INTS12, and NPNT). The independent NPNT signal captured by the genotyped variant rs34712979 was not detected in previous or concurrent studies because it was neither directly genotyped nor imputed with sufficient quality; this finding highlights a further advantage of the UK BiLEVE and UK Biobank array design.

The design of this genotyping array combined the best features of existing genome-wide platforms targeting common SNPs (MAF ≥5%) and putative functional exome chip content, plus additional content to improve imputation of low-frequency variants (MAF 1–5%). In combination with a new large UK-specific imputation reference panel (UK10K Project), these features increase the potential to discover novel signals. In our study, more than 28·5 million variants were imputed; current large meta-analyses combining data from several studies with older arrays and using equivalent quality control filters after imputing to 1000 Genomes Project Phase 1 alone typically measure about 10·6 million variants.³⁵ The genome-wide genotype data for these 50 008 individuals have been deposited in UK Biobank to be made available to other approved research projects across many disease areas. The UK BiLEVE array was used as a prototype for the array that is being used in the remaining roughly 450 000 UK Biobank participants. The UK Biobank array shares more than 95% of its content with the UK BiLEVE array. When genotyping of all UK Biobank participants is complete, UK Biobank will provide a unique resource for genome-wide studies of quantitative traits, nested case-control studies, and studies in which longitudinal outcomes can be studied.

Despite the strengths of using a large resource such as UK Biobank, this study has some limitations. In particular, there is a trade-off between obtaining the large sample sizes generally needed for genetic studies and the depth of phenotyping that is practicable in such large populations. Our spirometric definition of COPD was not based on bronchodilator reversibility testing, although we have shown previously that by limiting inclusion to individuals with GOLD stage 2+ spirometry, most of these individuals are likely to have COPD according to more rigorous criteria.¹³ Similarly, our definition of asthma was based on self-reporting of doctor-diagnosed disease. These limitations might have reduced our ability to identify some novel disease associations, although we were able to replicate many known associations using this approach.

In summary, we show the usefulness in sampling from the extremes of UK Biobank data to identify novel genetic signatures underlying phenotypes important in the development of airway disease and smoking behaviour. The ongoing genotyping, and further phenotyping, of the rest of the UK Biobank resource will facilitate further GWAS, which will undoubtedly improve our understanding of the genetic and molecular basis of common disease.

**Contributors**

LVW, DPS, MDT, and IPH conceived and designed the study; LVW, NS, VEJ, IN, MSA, RA, JPC, CF, DP, JM, DPS, MDT, and IPH analysed data. SM, CBK, AKK, KO, MO, AR, RM, EMI, and ER undertook functional follow-up analyses. YB, RH, DSP, PDP, EMR, JO'C, EF, OD, IS, and JM provided data for follow-up of signals. NX, CF, MM, EZ, APM, JM, and DPS provided statistical support and advice. LVW, IS, PD, RH, IP, ALH, NCT, DPS, MDT, and IPH provided advice on study conduct. All authors discussed the results and implications and commented on the manuscript at all stages.

**UK Brain Expression Consortium (UKBEC)**


**OxGSK Consortium**

Jason Z Liu (University of Oxford, Oxford, UK), Federica Tozzi (GlaxoSmithKline, Verona, Italy), Dawn M Waterworth (GlaxoSmithKline, London, UK), Stefano Zucchelli (GlaxoSmithKline, London, UK), Giacomo S Gori (GlaxoSmithKline, London, UK), Jana Vandrovcova (UCL Institute of Neurology, London, UK), Daniah Trabzuni (UCL Institute of Neurology, London, UK), Robert Walker (The University of Edinburgh, Edinburgh, UK), Karishma D’Sa (UCL Institute of Neurology, London, UK; and King’s College London, London, UK), and Martin Preissig (University of Lausanne, Lausanne, Switzerland; and University Hospital of Lausanne, Lausanne, Switzerland; and University Hospital of Lausanne, Lausanne, Switzerland; and University Hospital of Lausanne, Lausanne, Switzerland; and University Hospital of Lausanne, Lausanne, Switzerland; and University Hospital of Lausanne, Lausanne, Switzerland).
Articles

Québec – Santé (FRQS). EGCGUT received funding by FP7 grants (278913, 306031, 313010), Center of Excellence in Genomics (EXCEGEN), and University of Tartu (SPIGVARENG). We thank EGCGUT technical personnel, especially V Soo and S Smit. Data analyses were done in part in the High Performance Computing Center of University of Tartu. We thank Paul Jennings for assistance with generation of code to extract expression data. MO is a Postdoctoral Fellow of the Michael Smith Foundation for Health Research and the Canadian Institute for Health Research Integrated and Mentored Pulmonary and Cardiovascular Training program. This research used the ALICE and SPECTRE High Performance Computing Facilities at the University of Leicester.

References


