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

Louise V Wain*, Nick Shrine*, Suzanne Miller, Victoria E Jackson, Joanna Ntalla, Maria Soler Artigas, Charlotte K Billington, Abdul Kader Kheirallah, Richard Allen, James P Cook, Kelly Probert, Ma’en Obeidat, Yohan Bosse, Ke Hao, Dirkje S Postma, Peter D Paré, Adaikalavan Ramasamy, UK Brain Expression Consortium (UKBEC)†, Reedik Magi, Evelin Mihailov, Eva Reinmaa, Erik Melén, Jared O’Connell, Andrew P Morris, Jonathan Marchini, David P Strachan*, Martin D Tobin*, Ian P Hall*

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 optimal 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 $p<5 \times 10^{-8}$.

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 ($p=2.29 \times 10^{-19}$) and between individuals with and without doctor-diagnosed asthma ($p=6.06 \times 10^{-11}$). 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-DQβ1/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 5ʹ 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 PABPC1P2) 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.

Copyright © Wain et al. Open Access article distributed under the terms of CC BY.

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...
Articles

Research in context

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 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, adopting either new genetic assays or the extensive genome-wide data that we and UK Biobank have generated.
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 <0·7 and percent predicted FEV₁<80%) as COPD cases. 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 <0·7 and percent predicted FEV₁<80%) as COPD cases. This platform also provided dense coverage of genomic regions implicated in lung health and disease. 

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) with imputed marker doses, adjusting for pack-years in the analysis. We used a Bonferroni correction for multiple testing. The appendix (p 17) describes quality control (appendix pp 8–15), we imputed non-genotyped variants using a combined 1000 Genomes Project Phase 1 and UK10K Project reference panel (appendix pp 15–16). 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, we discovered variants at associated loci (identified as described in the Statistical analysis) that were highly associated with FEV₁ extremes. These variants were also 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. Additionally, we generated RNA sequencing data to discover novel transcripts of these genes in human bronchial epithelial cells. We tested all genome-wide meta-analysis 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) with imputed marker doses, adjusting for pack-years in the analysis. We used a Bonferroni correction for multiple testing. The appendix (p 17) describes quality control (appendix pp 8–15), we imputed non-genotyped variants using a combined 1000 Genomes Project Phase 1 and UK10K Project reference panel (appendix pp 15–16). 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, we discovered variants at associated loci (identified as described in the Statistical analysis) that were highly associated with FEV₁ extremes. These variants were also 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. Additionally, we generated RNA sequencing data to discover novel transcripts of these genes in human bronchial epithelial cells. We tested all genome-wide meta-analysis values for enrichment in biological pathways defined in publicly available databases. All functional analyses are described in detail in the appendix (pp 21–23).
phenotypes (appendix p 18). For this analysis, we created discovery and target subpopulations, each of which comprised cases and control groups created by randomly splitting the low FEV1, and average FEV1, groups (appendix pp 18–20). Variants of MAF of at least 1% associated with low FEV1 below given p value thresholds in the discovery

**Figure 1: Sample selection strategy**
ATS=American Thoracic Society. ERS=European Respiratory Society. FEV1=forced expiratory volume in 1 s. FVC=forced vital capacity. *See appendix (pp 3–5) for more details of sample selection.
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 50,008 unique samples representing the extremes and middle of the percent predicted FEV₁ distributions; this comprised 10,002 individuals with low percent predicted FEV₁, 10,000 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, 48,931 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 42,795,484 variants. Our final dataset for analysis, after excluding variants with information quality less than 0.5 or minor allele count less than three, comprised 28,509,962 imputed or genotyped variants in 48,931 unrelated individuals (table 1 and appendix pp 16 and 95).23

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 (NPNT), was a rare variant (MAF=0.13%) that showed strongest association with low FEV₁ versus high FEV₁ (p=1.64×10⁻²²; p value threshold <0.5; figure 2; appendix pp 38–53). The 26 previously reported SNPs 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 p 55). The 26 previously reported SNPs

### Table 1: Absolute and percent predicted forced expiratory volume in 1 s in each subgroup in heavy and never smokers

|                      | Heavy smokers | Never smokers |
|----------------------|---------------|--------------|
|                      | Number of individuals | Absolute FEV₁ (L) | Predicted FEV₁ (%) | Number of individuals | Absolute FEV₁ (L) | Predicted FEV₁ (%) |
| Low FEV₁             | 9748          | 1.92 (0.55) | 65.6% (11.8) | 9745 | 2.05 (0.54) | 69.3% (10.0) |
| Average FEV₁         | 9803          | 2.68 (0.56) | 90.6% (3.9)  | 9827 | 2.92 (0.57) | 98.7% (1.3)  |
| High FEV₁            | 4906          | 3.49 (0.72) | 118.0% (8.1) | 4902 | 3.93 (0.73) | 130.3% (8.3) |

Data are mean (SD), unless otherwise specified. See appendix (pp 3–5) for details of sample selection. FEV₁=forced expiratory volume in 1 s.

Figure 2: Polygenic component of low forced expiratory volume in 1 s and shared polygenic component of different phenotypes defined by forced expiratory volume in 1 s, smoking, and doctor diagnosis of asthma. The p value in the target population shown above the bars is for the p value threshold <0.5. The sample sizes differed between the comparisons; details of these and the assumptions used in the analyses are described in the appendix (pp 18–20). FEV₁=forced expiratory volume in 1 s.

www.thelancet.com/respiratory   Vol 3   October 2015 773
**Table 2: Novel genome-wide significant signals of association with extremes of forced expiratory volume in 1 s or smoking behaviour**

For variants that showed association with extremes of FEV\textsubscript{1}, in either heavy smokers or never smokers, the results from both the never smokers and heavy smokers are presented. p values and 95% CIs for signals of association with extremes of FEV\textsubscript{1}, are genomic control corrected for the result in the smoking status stratum for which the signal was most highly significant, and uncorrected for the alternate smoking status stratum. All smoking behaviour results are genomic control corrected. Genomic control λs are given in the appendix (pp 113–116). For variants that had genome-wide significant evidence of association for (hg19).

| Locus | Non-coded/ coded allele (minor allele) | Imputation info score\* | Smoking status | OR (95% CI) p value | Association with COPD‡ | Effect on FEV\textsubscript{1}† |
|-------|---------------------------------------|-------------------------|----------------|---------------------|-------------------------|-----------------------------|
| Chr12:114743533 | RRM19/TREX5 | T/C (T) | 0.737 | Never | 0.002 (60) | 0.97 (0.57–1.67) | 0.90 | 1.16 (0.54–2.51) | 0.71 | 0.101 (0.118) | 0.39 |
| ... | ... | ... | ... | Heavy | 0.001 (39) | 11.73 (5.03–27.32) | 1.16 × 10\textsuperscript{4} | 6.44 (2.89–14.37) | 5.40 × 10\textsuperscript{4} | -0.728 (0.151) | 1.31 × 10\textsuperscript{4} |

For variants that showed association with extremes of FEV\textsubscript{1}, low vs high FEV\textsubscript{1},

**Genome-wide significant in heavy smokers**

rs434712979¶ [¶], Chr4:106875905  
NPMT G/A (A) 1.000  
Never 0.268 (7842)  
1.27 (1.20–1.34)  
9.62 × 10\textsuperscript{−16}  
1.36 (1.27–1.46)  
2.10 × 10\textsuperscript{−16}  
0.009  
0.39 |

**Genome-wide significant in never smokers**

rs9274600¶, Chr3:32635592  
HLA-DQB1/ HLA-DQA2 A/G (G) 0.962  
Heavy 0.472 (13 838)  
1.18 (1.13–1.25)  
1.26 × 10\textsuperscript{−10}  
1.24 (1.16–1.32)  
1.95 × 10\textsuperscript{−10}  
0.009  
0.037 |

rs2532349, Chr17:44339473  
KANG1 A/G (G) 0.976  
Never 0.242 (7088)  
1.22 (1.15–1.29)  
1.66 × 10\textsuperscript{−10}  
1.24 (1.16–1.34)  
3.97 × 10\textsuperscript{−10}  
0.011  
3.22 × 10\textsuperscript{−10} |

**Smoking behaviour: heavy vs never smokers**

rs4466874, Chr11:112861434  
NCAM1 T/C (C) 0.998  
NA 0.385 (37 709)  
1.10 (1.07–1.13)  
3.22 × 10\textsuperscript{−10}  
NA 0.156 (0.105) |

rs10193706, Chr2:146316219  
TEX41/ PABPC1P2 A/C (A) 0.983  
NA 0.473 (85 120)  
1.09 (1.06–1.12)  
1.10 × 10\textsuperscript{−10}  
NA 0.069 (0.040) |

rs143125561, rs57342388, Chr20:31162590  
NOL4/C/CACGG C/G (C) 0.983  
NA 0.233 (22 820)  
1.10 (1.07–1.13)  
4.65 × 10\textsuperscript{−10}  
NA 0.059 (0.035) |

rs617846511, Chr9:9445471  
LPPRS C/T (T) 1.000  
NA 0.170 (16 609)  
1.10 (1.07–1.14)  
2.89 × 10\textsuperscript{−10}  
NA 0.056 (0.035) |

rs10807199, Chr6:38901867  
DNAH8 C/T (T) 1.000  
NA 0.473 (28 286)  
1.08 (1.05–1.11)  
3.17 × 10\textsuperscript{−10}  
NA 0.023 (0.009) |

(associated with FEV\textsubscript{1}, FEV\textsubscript{1}/FVC ratio, or both)\textsuperscript{15–19} explained 2·33% of the variance of FEV\textsubscript{1} in our data; adding in the SNPs representing our six novel signals of association with FEV\textsubscript{1} extremes, we explained 3·63% of the variance of FEV\textsubscript{1} (appendix p 17).

Although association with lung function at 4q24 is well established,\textsuperscript{5} we report two further independent signals of association at this locus (table 2; appendix pp 56–57, 99–101). The first (rs34712979) was localised to NPNT but independent of the previously reported signal, which

Articles

www.thelancet.com/respiratory Vol 3 October 2015
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 × 10⁻¹⁶; table 2) and was weakly correlated (linkage disequilibrium r²=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. Figure 3: Manhattan plots for low versus high forced expiratory volume in 1 s in never smokers and heavy smokers and for heavy versus never smokers p values are from a Score test and have genomic control applied unless the MAC was less than 400 and Score test p<1.00 × 10⁻⁶, in which case p values are from a Firth test with no genomic control. Novel loci are underlined and novel signals at previously reported loci are shown in bold. The dashed red line shows the threshold for genome-wide significance (p<5 × 10⁻⁸). Variants with suggestive evidence of association (p<5 × 10⁻⁷) are coloured red. Quantile–quantile plots for these analyses are shown in the appendix (pp 113–116). FEV1=forced expiratory volume in 1 s. MAC=minor allele count.

spanned INTS12, GSTCD, and NPNT (appendix p 102). 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 × 10⁻¹⁶; table 2) and was weakly correlated (linkage disequilibrium r²=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. When rs6856422 was included as
a covariate in our analysis, the signal for rs34712979 (p=9·62×10−10) was only slightly attenuated (p=4·66×10−11). The novel signal at TET2 (rs2047409) also showed strongest association in never smokers (p=1·31×10−8). rs2047409 is separated from the previously reported association of rs10516526 with FEV1, (GSTCD) by a recombination hotspot and is statistically independent (rs10516526 included as a covariate; rs2047409, p=9·8×10−7). TET2 encodes tet methylcytosine dioxygenase 2, which has a role in myelopoiesis, and SNPs in TET2 have shown association with height. 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. The signal we report was strongest in never smokers (rs9274600, p=1·26×10−10; table 2; appendix p 102). With an imputed proxy (rs17843604) of the asthma-associated SNP rs9273349—included as a covariate in the analysis of rs9274600—the signal for rs9274600 was attenuated (p=5·66×10−4), confirming that rs9274600 and rs9273349 are correlated. After exclusion of individuals with doctor-diagnosed asthma, the odds ratio for rs9274600 decreased from 5·0 to 1·8 (95% CI 1·11–1·25) to 1·14 (0·81–1·20), but remained significant (p=3·25×10−6; appendix p 103). This signal is independent of nearby signals reported for lung function, including rs7764819, which is 45 kb from rs9274600 (association for rs9274600 conditioned on rs7764819; p=6·71×10−11).

We identified a rare SNP that was associated with FEV1, extremes in heavy smokers only, after adjusting for pack-years of smoking (p=1·16×10−4; table 2). This intergenic SNP on chromosome 12 (chr12:114743533, MAF=0·13%) also showed weak evidence of association with smoking behaviour (p=6·12×10−4; appendix pp 60–62). We noted evidence for change in expression levels with increasing fetal lung age (p=0·04) for one or more probes after adjustment for multiple testing for the nearby gene TBX5 (appendix pp 58–59).

We noted a broad signal of association (~1·5 Mb) in an inversion locus at 17q21.31 (rs2532349, near KANSL1) with extremes in heavy smokers (rs2532349, p=1·66×10−10), but was also detected in heavy smokers (p=1·47×10−5). Genes in this locus, which include MAPT and CRHR1, have previously been associated with pulmonary fibrosis and inhaled corticosteroid response in asthma. SNP rs2532349 (and SNPs in strong linkage disequilibrium [r²>0·8]) was associated with mRNA expression levels of at least 15 genes in lung and blood (appendix pp 63–70). We identified differential expression for six genes at 17q21.31 during fetal lung development (and for four genes on different chromosomes regulated by transcripts at 17q21.41; appendix pp 58–39). Relatively abundant novel transcripts (ie, compared with other transcripts detected) were identified by RNA sequencing in human bronchial epithelial cells for WNT3 and LRRC37A4P; expression of both genes was associated with rs2532349 in lung and blood (appendix pp 104–109). The SNP rs2532349 (MAF=0·24%) was in linkage disequilibrium with the inversion (r²>0·9); the allele associated with low FEV1, was positively correlated with the inverted haplotype. Another signal at the asthma-associated SNP rs9274600 (association for rs9274600 conditioned on rs7218675, p=2·40×10−6; appendix p 71). The sentinel SNP rs2532349 lies within this region.

Second, we note a signal of association on chromosome 17 (17q25.1) that was within TSEN54 and occurred only in never smokers (rs7218675; p=1·18×10−4; 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 LGNL2 (appendix p 108).

To corroborate these new signals, we identified in TET2 and TSEN54, we present evidence of association with FEV1 in a previously reported study 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 (5×10−4<p<5×10−3) 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.

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). The novel signals included rs4466874, 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
behaviour (appendix pp 56–57, 99–101), including a significance, at three of the loci associated with smoking independent signals, which did not reach genome-wide association test statistics for low FEV₁ locus at the.

population by a concurrent study, and signal of association with lung function in the general CCDC91 extremes. These included association with smoking behaviour (appendix p 72–73), which there is corroborative evidence of association with smoking behaviour (appendix p 78), no gene–smoking interactions were versus high FEV₁ across heavy and never smokers, (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. 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), 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 studies 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 p 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, 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. The disruption of the gene gives rise to a novel truncated transcript, which encodes a protein missing a domain essential for key interactions with other proteins important for NSL1 function. Therefore, widespread effects on gene regulation through altered histone acetylation could underlie this association. Reduced expression of KANSL1 causes a rare multisystem disorder, 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

(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 DNAH18 (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.46 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), 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 studies 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 p 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, 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. The disruption of the gene gives rise to a novel truncated transcript, which encodes a protein missing a domain essential for key interactions with other proteins important for NSL1 function. Therefore, widespread effects on gene regulation through altered histone acetylation could underlie this association. Reduced expression of KANSL1 causes a rare multisystem disorder, 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

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. The substantia nigra plays an important part in reward and addiction, but little is known about WDR61 other than that expression can be induced by mechanical strain in mesenchymal stem cells.

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. The disruption of the gene gives rise to a novel truncated transcript, which encodes a protein missing a domain essential for key interactions with other proteins important for NSL1 function. Therefore, widespread effects on gene regulation through altered histone acetylation could underlie this association. Reduced expression of KANSL1 causes a rare multisystem disorder, 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


Articles

association with FEV, in a previously reported large study of FEV, in the general population. However, the rare SNP on chromosome 12 for which we found association with extremes of FEV, 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/PABPC1P2, and NOL4L in independent populations. 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. 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 spiroometric 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. 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, CKB, AKK, KP, MO, AR, RM, EMI, and ER undertook functional follow-up analyses. YB, KH, DSP, PDP, EMtE, JOC, EF, OD, IS, and JM provided data for follow-up of signals. NS, CF, MM, EZ, APM, M, and DPS provided statistical support and advice. LVW, IS, PD, RH, JP, 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)

John A Hardy (UCl Institute of Neurology, London, UK).
Michael E Weale (King’s College London, London, UK), Mina Ryten (UCl Institute of Neurology, London, UK, and King’s College London, London, UK), Colin Smith (The University of Edinburgh, Edinburgh, UK), Robert Walker (The University of Edinburgh, Edinburgh, UK), Juan Botia (UCl Institute of Neurology, London, UK, and King’s College London, London, UK), Jana Vandrovcova (UCl Institute of Neurology, London, UK, and King’s College London, London, UK), Karishma D’Sa (UCl Institute of Neurology, London, UK, and King’s College London, London, UK), Mar Matarin (UCl Institute of Neurology, London, UK, and King’s College London, London, UK), Sebastien Guelfi (UCl Institute of Neurology, London, UK, and King’s College London, London, UK), Daniah Trabzuni (UCl Institute of Neurology, London, UK), Akhikaltavan Ramasamy (UCl Institute of Neurology, London, UK, and King’s College London, London, UK), Vibin Varghese (King’s College London, London, UK, and King’s College London, London, UK), and Laura Waterworth (UCl Institute of Neurology, London, UK).

OxGSK Consortium

Jason Z Liu (University of Oxford, Oxford, UK), Federica Tozzi (Gliaxosmithkline, Verona, Italy), Dawn M Waterworth (Gliaxosmithkline, Upper Merion, PA, USA), Sreekumar G Pillai (Gliaxosmithkline, Upper Merion, PA, USA), Pietro Pereira Mavilla (Gliaxosmithkline, Upper Merion, PA, USA), Richard Lefkowitz (Imperial College London, London, UK), Wade Berrettini (University of Pennsylvania School of Medicine, Philadelphia, PA, USA), Christopher W Knoff (Gliaxosmithkline, Research Triangle Park, NC, USA), Xin Yuan (Gliaxosmithkline, Upper Merion, PA, USA), Gérard Waeber (University of Lausanne, Lausanne, Switzerland), and Peter Vollenweider (University of Lausanne, Lausanne, Switzerland, and University Hospital of Lausanne, Lausanne, Switzerland), Martin Preissig (University of Lausanne, Lausanne, Switzerland, and University Hospital of Lausanne, Lausanne, Switzerland, and University Hospital of Lausanne, Lausanne, Switzerland, Switzerland); and University Hospital of Lausanne, Lausanne, Switzerland, Martin Preissig (University of Lausanne, Lausanne, Switzerland, and University Hospital of Lausanne, Lausanne, Switzerland, and University Hospital of Lausanne, Lausanne, Switzerland, Switzerland); and University Hospital of Lausanne, Lausanne, Switzerland, Martin Preissig (University of Lausanne, Lausanne, Switzerland, and University Hospital of Lausanne, Lausanne, Switzerland, Switzerland); and University Hospital of Lausanne, Lausanne, Switzerland.
Hospital of Lusanne, Lausanne, Switzerland), Nicholas J wareham (Institute of Metabolic Science, Cambridge, UK), Jing Hua Zhao (Institute of Metabolic Science, Cambridge, UK), Ruth J F Loos (Institute of Metabolic Science, Cambridge, UK), Inês Barroso (Wellcome Trust Sanger Institute, Hinxton, UK), Kay-Tee Khaw (University of Cambridge, Cambridge, UK), Scott Grundy (University of Texas Southwestern Medical Center, Dallas, TX, USA), Philip Barter (The Heart Research Institute, Sydney, NSW, Australia), Robert Mahley (Glade Stone Institute of Cardiovascular Disease, University of California, San Francisco, CA, USA; and American Hospital, Istanbul, Turkey), Antero Kesaniemi (University of Oulu, Oulu, Finland), Ruth McPherson (University of Ottawa Heart Institute, Ottawa, ON, Canada), John B Vincent (University of Toronto, Toronto, ON, Canada), John Strauss (University of Toronto, Toronto, ON, Canada), James L Kennedy (University of Toronto, Toronto, ON, Canada), Anne Farmer (King's College London, London, UK), Peter McGuffin (King's College London, London, UK), Richard Day (University of Dundee, Dundee, UK), Keith Matthews (University of Dundee, Dundee, UK), Pet Bakke (University of Bergen, Bergen, Norway), Amund Gulsvik (University of Bergen, Bergen, Norway), Susanne Luace (Max Planck Institute of Psychiatry, Munich, Germany), Marcus Ising (Max Planck Institute of Psychiatry, Munich, Germany), Tanja Brueckl (Max Planck Institute of Psychiatry, Munich, Germany), Sonja Hoffmann (Max Planck Institute of Psychiatry, Munich, Germany), H-Erich Wichmann (Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; Ludwig-MaximiliansUniversität, Munich, Germany; and Klinikum Grosshadern, Munich, Germany), Rajesh Rawal (Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany), Norbert Dahnen (University of Mainz, Mainz, Germany), Claudia Lamina (Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany; and Innsbruck Medical University, Innsbruck, Austria), Ozren Polasek (University of Zagreb, Zagreb, Croatia), Lina Zagga (University of Edinburgh, Edinburgh, UK; Jennifer Huff man (MRC Human Genetics Unit, Edinburgh, UK), Susan Campbell (MRC Human Genetics Unit, Edinburgh, UK), Jaspal Kooner (Imperial College London, London, UK), John C Chambers (Imperial College London, London, UK), Mary Susan Burnett (Washington Hospital Center, Washington, DC, USA), Joseph M Devaney (Washington Hospital Center, Washington, DC, USA), Augusto D Pichard (Washington Hospital Center, Washington, DC, USA), Kenneth M Kent (Washington Hospital Center, Washington, DC, USA), Lowell Satler (Washington Hospital Center, Washington, DC, USA), Joseph M Lindsay (Washington Hospital Center, Washington, DC, USA), Ron Waksman (Washington Hospital Center, Washington, DC, USA), Stephen Epstein (Washington Hospital Center, Washington, DC, USA), James F Wilson (University of Edinburgh, Edinburgh, UK), Sarah H Wild (University of Edinburgh, Edinburgh, UK), Harry Campbell (University of Edinburgh, Edinburgh, UK), Veronique Vitart (MRC Human Genetics Unit, Edinburgh, UK), Muredach P Reilly (University of Pennsylvania, Philadelphia, PA, USA), Mingyou Li (University of Pennsylvania, Philadelphia, PA, USA), Liming Qu (University of Pennsylvania, Philadelphia, PA, USA), Robert Wilensky (University of Pennsylvania, Philadelphia, PA, USA), William Matti (University of Pennsylvania, Philadelphia, PA, USA), Hakon H Hakonarson (Children's Hospital of Philadelphia, Philadelphia, PA, USA), Daniel J Rader (University of Pennsylvania, Philadelphia, PA, USA), David Ellinghaus (ChristianAlbrechts-University of Kiel, Kiel, Germany), Wolfgang Lieb (ChristianAlbrechts-University of Kiel, Kiel, Germany), Andre Franke (ChristianAlbrechts-University of Kiel, Kiel, Germany), Manuela Uda (Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy), Antonio Terracciano (National Institute on Aging, Baltimore, MD, USA), Xiangjun Xiao (University of Texas MD Anderson Cancer Center, Houston, TX, USA), Philippe Durand (Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy), Paul Scheet (University of Texas MD Anderson Cancer Center, Houston, TX, USA), Philip Schlessinger (National Institute on Aging, Baltimore, MD, USA), David St Clair (University of Aberdeen, Aberdeen, UK), Dan Rujescu (Ludwig-Maximilians-Universität, Munich, Germany), Gonçalo R Abecasis (University of Michigan, Ann Arbor, MI, USA), Hans Jürgen Grabe (University of Greifswald, Greifswald, Germany), Alexander Teumer (University of Greifswald, Greifswald, Germany), Heny Völzke (University of Greifswald, Greifswald, Germany), Astrid Petersmann (University of Greifswald, Greifswald, Germany), Ulrich John (University of Greifswald, Greifswald, Germany), Igor Rudan (University of Split, Split, Croatia; and University of Edinburgh, Edinburgh, UK), Caroline Hayward (MRC Human Genetics Unit, Edinburgh, UK), Alan F Wright (MRC Human Genetics Unit, Edinburgh, UK), Ivana Kolic (University of Zagreb, Zagreb, Croatia), Benjamin J Wright (University of Leicester, Leicester, UK), John R Thompson (University of Leicester, Leicester, UK), Anthony J Balmforth (University of Leeds, Leeds, UK), Alistair S Hall (University of Leeds, Leeds, UK), Nilesh J Samani (University of Leicester, Glenfield Hospital, Leicester, UK), Carl A Anderson (Wellcome Trust Sanger Institute, Hinxton, UK), Tariq Ahmad (Peninsula College of Medicine and Dentistry, Exeter, UK), Christopher G Mathew (King's College London School of Medicine, Guy's Hospital, London, UK), Miles Parkes (Addenbrooke's Hospital, Cambridge, UK), Jack Satsangi (University of Edinburgh, Western General Hospital, Edinburgh, UK), Mark Caulfield (Barts and the London School of Medicine, Queen Mary University of London, London, UK), Patricia B Munroe (Barts and the London School of Medicine, Queen Mary University of London, London, UK), Martin Farrall (University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, UK), Anna Dominiczak (British Heart Foundation Glasgow Cardiovascular Research Centre, University of Glasgow, Western Infirmary, Glasgow, UK), Jane Worthington (Arthritis Research UK Centre UK Centre for Genetics and Genomics, University of Manchester, Manchester, UK), Wendy Thomson (Arthritis Research UK Centre UK Centre for Genetics and Genomics, University of Manchester, Manchester, UK), Steve Eyre (Arthritis Research UK Centre for Genetics and Genomics, University of Manchester, Manchester, UK), Anne Barton (Arthritis Research UK Centre for Genetics and Genomics, University of Manchester, Manchester, UK), Sophie Brousseau (Institute of Metabolic Science, Cambridge, UK), Jing Hua Zhao (University of Split, Split, Croatia; and University of Edinburgh, Edinburgh, UK), Jonathan Marchini (University of Oxford, Oxford, UK), * Members listed in appendix pp 128–131.

Declaration of interests
We declare no competing interests.

Acknowledgments
This work was funded by a Medical Research Council (MRC) strategic award to MDT, IPH, DPS, and LVY (MC_PC_12010). This research was done using the UK Biobank resource. MDT was supported by MRC fellowships G0501942 and G0902313. IPH is funded by the Wellcome Trust (grant number 091310). JM is funded by the UK10K Consortium, derived from samples from TwinsUK and ALSPAC. A full list of the investigators who contributed to the generation of the data is available from the UK10K website. Funding for UK10K was provided by the Wellcome Trust under award WT091310. MDT was supported by the Wellcome Trust. This study makes use of data generated by the UK10K Consortium, derived from samples from TwinsUK and ALSPAC.
Articles

Joung S, Research Scholar award from the Fonds de recherche Québec – Santé (FRQS). EGCUT received financing by FP7 grants (22813, 306033, 313040). Center of Excellence in Genomics (EXCEGEN), and University of Tartu (SPVGVARENG). We thank EGCUT technical personnel, especially V Sso 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
1. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death in 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet 2012; 380: 2095–128.
2. McClellan GE, Svrzenceng M, Pedersen NL, Heller DA, Plomin R. Genetic and environmental influences on pulmonary function in aging Swedish twins. J Gerontol 1994; 49: 264–68.
3. Palmer LJ, Knudnau MW, Divitini M, et al. Familial aggregation and heritability of adult lung function: results from the Busselton Health Study. Eur Respir J 2001; 17: 696–702.
4. Zhai G, Valdes AM, Chekris L, Clement G, Strachan D. Spector TD. The interaction of genes and smoking on forced expiratory volume: a classic twin study. Chest 2007; 132: 1772–77.
5. Hancock DB, Elgebisher M, Wilk JE, et al. Meta-analyses of genome-wide association studies identify multiple loci associated with pulmonary function. Nat Genet 2010; 42: 45–52.
6. Loth DW, Artigas MS, Gharib SA, et al. Genome-wide association analysis identifies six new loci associated with forced vital capacity. Nat Genet 2014; 46: 669–77.
7. Repapi E, Sayers I, Wain LV, et al. Genome-wide association study identifies five loci associated with lung function. Nat Genet 2010; 42: 36–44.
8. Soler Artigas M, Loth DW, Wain LV, et al. Genome-wide association and large-scale follow up identifies 16 new loci influencing lung function. Nat Genet 2011; 43: 1082–90.
9. Wilk JE, Chen TH, Gottlieb DJ, et al. A genome-wide association study of pulmonary function measures in the Framingham Heart Study. PLoS Genet 2012; 8: e1002529.
10. Lamontagne M, Couture C, Postma DS, et al. Refining susceptibility loci of chronic obstructive pulmonary disease with lung eQTLs. PLoS One 2013; 8: e70220.
11. Obeidat M, Miller S, Probert K, et al. GSTCD and INTS12 regulation of gene expression in ten regions of the human brain. Nat Neurosci 2014; 17: 1438–28.
12. Stirling K, van den Berge M, Hjazi K, et al. A dynamic bronchial airway gene expression signature of chronic obstructive pulmonary disease and lung function impairment. Am J Respir Crit Care Med 2013; 187: 931–42.
13. Melen E, Kho AT, Sharma S, et al. Expression analysis of asthma candidate genes during human and murine lung development. Respir Res 2011; 12: 86.
14. Ma C, Blackwell T, Boehnke M, Scott LJ, for the GoT2D investigators. Recommended joint and meta-analysis strategies for case-control association testing of single low-cost variants. Genet Epidemiol 2013; 37: 539–50.
15. Purcell SM, Wray NR, Stone JL, et al. For the International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 2009; 460: 748–52.
16. Miller MR, Hankinson J, Brusasco V, et al. Standardisation of spirometry. Eur Respir J 2005; 26: 319–38.
17. Soler Artigas M, Wain LV, Miller S, et al. Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. Nat Commun (in press).
18. Langlo Allen H, Estrada K, Lettere G, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 2010; 467: 832–38.
19. Moffatt MF, Gut IG, Demenais F, et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010; 363: 3211–21.
20. Hancock DB, Artigas MS, Gharib SA, et al. Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. PLoS Genet 2012; 8: e1003098.
21. Noah I, Zhang Y, Ma SF, et al. Genetic variants associated with idiopathic pulmonary fibrosis susceptibility and mortality: a genome-wide association study. Lancet Respir Med 2013; 1: 309–17.
22. Fingerlin TE, Murphy E, Zhang W, et al. Genome-wide association study identifies multiple susceptibility loci for pulmonary fibrosis. Nat Genet 2013; 45: 613–20.
23. Tantisira KG, Lazarus R, Litonjua AA, Klanderman B, Weiss ST. Chromosome 17: association of a large inversion polymorphism with corticosteroid response in asthma. Pharmacogenet Genomics 2008; 18: 733–37.
24. Boettiger LM, Handsaker RE, Zody MC, McCarron SA. Structural haplotypes and recent evolution of the human 17q21.31 region. Nat Genet 2012; 44: 881–85.
25. Steinberg KM, Antonacci F, Sudmann PH, et al. Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat Genet 2012; 44: 872–80.
44 Gelernter J, Yu Y, Weiss R, et al. Haplotype spanning TTC12 and ANKK1, flanked by the DRD2 and NCAM1 loci, is strongly associated to nicotine dependence in two distinct American populations. *Hum Mol Genet* 2006; 15: 3498–507.

45 Wang D, Ma JZ, Li MD. Mapping and verification of susceptibility loci for smoking quantity using permutation linkage analysis. *Pharmacogenomics J* 2005; 5: 166–72.

46 Yang J, Ferreira T, Morris AP, for the Genetic Investigation of ANthropometric Traits (GIANT) Consortium, and the DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012; 44: 369–75.

47 Feng Y, Niu T, Xing H, et al. A common haplotype of the nicotine acetylcholine receptor alpha 4 subunit gene is associated with vulnerability to nicotine addiction in men. *Am J Hum Genet* 2004; 75: 112–21.

48 Faure A, Haberland U, Condé F, El Massioui N. Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *J Neurosci* 2005; 25: 2771–80.

49 Rathbone SR, Glossop JR, Gough JE, Cartmell SH. Cyclic tensile strain upon human mesenchymal stem cells in 2D and 3D culture differentially influences CCNL2, WDR61 and BAHCC1 gene expression levels. *J Mech Behav Biomed Mater* 2012; 11: 82–91.

50 Mendjan S, Taipale M, Kind J, et al. Nuclear pore components are involved in the transcriptional regulation of dosage compensation in Drosophila. *Mol Cell* 2006; 21: 811–23.

51 Li X, Wu L, Corsa CA, Kunkel S, Dou Y. Two mammalian MOF complexes regulate transcription activation by distinct mechanisms. *Mol Cell* 2009; 36: 290–301.

52 Koolen DA, Kramer JM, Neveling K, et al. Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nat Genet* 2012; 44: 639–41.

53 Zollino M, Orteschi D, Murdolo M, et al. Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nat Genet* 2012; 44: 636–38.