RESEARCH ARTICLE

# REVISED Meta-analysis of exome array data identifies six novel genetic loci for lung function [version 3; referees: 2 approved]

Victoria E. Jackson [1], Jeanne C. Latourelle[2], Louise V. Wain [1,3], Albert V. Smith[4,5], Megan L. Grove[6], Traci M. Bartz[7], Ma'en Obeidat [8], Michael A. Province[9], Wei Gao[10], Beenish Qaiser[11], David J. Porteous [12], Patricia A. Cassano [13,14], Tarunveer S. Ahluwalia [15,16], Niels Grarup [15], Jin Li[17], Elisabeth Altmaier[18], Jonathan Marten [19], Sarah E. Harris [20,21], Ani Manichaikul[22], Tess D. Pottinger [23,24], Ruifang Li-Gao[25], Allan Lind-Thomsen[26], Anubha Mahajan [27], Lies Lahousse [28,29], Medea Imboden[30,31], Alexander Teumer [32], Bram Prins[33], Leo-Pekka Lyytikäinen [34,35], Gudny Eiriksdottir[5], Nora Franceschini[36], Colleen M. Sitlani [37], Jennifer A. Brody [37], Yohan Bossé [38], Wim Timens [39,40], Aldi Kraja[9], Anu Loukola[11], Wenbo Tang[13,41], Yongmei Liu[42], Jette Bork-Jensen[15], Johanne M. Justesen [16], Allan Linneberg [43-45], Leslie A. Lange[46], Rajesh Rawal[18], Stefan Karrasch[47,48], Jennifer E. Huffman[19], Blair H. Smith [49], Gail Davies[20,50], Kristin M. Burkart[23], Josyf C. Mychaleckyj [22], Tobias N. Bonten[51,52], Stefan Enroth [26], Lars Lind[53], Guy G. Brusselle[28,54,55], Ashish Kumar[27,30,31,56], Beate Stubbe[57], Understanding Society Scientific Group, Mika Kähönen [58,59], Annah B. Wyss[60], Bruce M. Psaty[61,62], Susan R. Heckbert[63], Ke Hao[64,65], Taina Rantanen [66], Stephen B. Kritchevsky [67], Kurt Lohman[42], Tea Skaaby[43], Charlotta Pisinger[43], Torben Hansen[15], Holger Schulz[47,68], Ozren Polasek[69], Archie Campbell [12], John M. Starr[20,70], Stephen S. Rich[22], Dennis O. Mook-Kanamori[25,52], Åsa Johansson[26], Erik Ingelsson[71,72], André G. Uitterlinden[54,73], Stefan Weiss [74,75], Olli T. Raitakari[76,77], Vilmundur Gudnason[4,5], Kari E. North[78], Sina A. Gharib [79], Don D. Sin [8,80], Kent D. Taylor [81], George T. O'Connor[82,83], Jaakko Kaprio [11,84,85], Tamara B. Harris[86], Oluf Pederson[15], Henrik Vestergaard [15,16], James G. Wilson[87], Konstantin Strauch[88,89], Caroline Hayward [19], Shona Kerr [19], Ian J. Deary[20,50],

R. Graham Barr[23,90], Renée de Mutsert[25], Ulf Gyllensten[26], Andrew P. Morris [27,91], M. Arfan Ikram[54,92,93], Nicole Probst-Hensch[30,31], Sven Gläser[57,94], Eleftheria Zeggini[33], Terho Lehtimäki[34,35], David P. Strachan[95], Josée Dupuis[10], Alanna C. Morrison[6], Ian P. Hall [96], Martin D. Tobin [1,3], Stephanie J. London [60]

[1]Department of Health Sciences, University of Leicester, Leicester, UK
[2]Department of Neurology, Boston University School of Medicine, Boston, MA, USA
[3]National Institute for Health Research, Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester, UK
[4]Icelandic Heart Association, 201 Kopavogur, Iceland
[5]University of Iceland, 101 Reykjavik, Iceland
[6]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, 77030, USA
[7]Cardiovascular Health Research Unit, Departments of Medicine and Biostatistics, University of Washington, Seattle, WA, 98101, USA
[8]The University of British Columbia Centre for Heart Lung Innovation, St Paul's Hospital, Vancouver, BC, Canada
[9]Department of Genetics, Washington University School of Medicine, St. Louis, MO, USA
[10]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA
[11]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, FI-00014, Helsinki, Finland
[12]Centre for Genomic & Experimental Medicine, MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK
[13]Division of Nutritional Sciences, Cornell University, Ithaca, NY, USA
[14]Department of Healthcare Policy and Research, Division of Biostatistics and Epidemiology, Weill Cornell Medical College, New York City, NY, USA
[15]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[16]Steno Diabetes Center Copenhagen, Gentofte, 2820, Denmark
[17]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Palo Alto, CA, USA
[18]Research Unit of Molecular Epidemiology, Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany
[19]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh , EH4 2XU, UK
[20]Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh, EH8 9JZ, UK
[21]Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh, EH4 2XU, UK
[22]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA
[23]Department of Medicine, College of Physicians and Surgeons, Columbia University, New York, NY, USA
[24]Department of Preventive Medicine - Division of Health and Biomedical Informatics, Northwestern University - Feinberg School of Medicine, Chicago, IL, USA
[25]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, 2333 ZA, Netherlands
[26]Department of Immunology, Genetics, and Pathology, Biomedical Center, SciLifeLab Uppsala, Uppsala University, SE-75108 Uppsala, Sweden
[27]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK
[28]Respiratory Medicine, Ghent University Hospital, Ghent, BE9000, Belgium
[29]Bioanalysis, Ghent University, Ghent, BE9000, Belgium
[30]Swiss Tropical and Public Health Institute, Basel, Switzerland
[31]University of Basel, Basel, Switzerland
[32]Institute for Community Medicine, University Medicine Greifswald, 17475 Greifswald, Germany
[33]Human Genetics, Wellcome Trust Sanger Institute, Hinxton, CB10 1SA, UK
[34]Department of Clinical Chemistry, Fimlab Laboratories, Tampere 33520, Finland
[35]Department of Clinical Chemistry, Faculty of Medicine and Life Sciences, University of Tampere, Tampere 33014, Finland
[36]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, NC 27514, USA
[37]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, 98101, USA
[38]Institut universitaire de cardiologie et de pneumologie de Québec, Department of Molecular Medicine, Laval University, Québec, Canada

[39]Department of Pathology and Medical Biology, University Medical Center Groningen, University of Groningen, NL9713 GZ, Netherlands
[40]Groningen Research Institute for Asthma and COPD, University Medical Center Groningen, University of Groningen, Groningen, Netherlands
[41]Boehringer Ingelheim , Danbury, CT, USA
[42]Wake Forest School of Medicine, Winston-Salem, North Carolina, USA
[43]Centre for Clinical Research and Prevention, Bispebjerg and Frederiksberg Hospital, The Capital Region, Copenhagen, Denmark
[44]Department of Clinical Experimental Research, Rigshospitalet, 2600 Glostrup, Denmark
[45]Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark
[46]Department of Medicine, Division of Bioinformatics and Personalized Medicine, University of Colorado Denver, Aurora, CO, USA
[47]Institute of Epidemiology I, Helmholtz Zentrum München, German Research Center for Environmental Health, 85764 Neuherberg, Germany
[48]Institute and Outpatient Clinic for Occupational, Social and Environmental Medicine, Ludwig-Maximilians-Universität, Munich, Germany
[49]Division of Population Health Sciences, Ninewells Hospital and Medical School, University of Dundee, Dundee, DD1 9SY, UK
[50]Department of Psychology, University of Edinburgh, Edinburgh, EH8 9JZ, UK
[51]Department of Pulmonology, Leiden University Medical Center, Leiden, 2333 ZA, Netherlands
[52]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, 2333 ZA, Netherlands
[53]Department of Medical Sciences, Uppsala University Hospital, Uppsala, Sweden
[54]Epidemiology, Erasmus Medical Center, Rotterdam, 3000CA, Netherlands
[55]Respiratory Medicine, Erasmus Medical Center, Rotterdam, 3000CA, Netherlands
[56]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[57]Internal Medicine B, University Medicine Greifswald, Greifswald, 17475, Germany
[58]Department of Clinical Physiology, Tampere University Hospital, Tampere, 33521, Finland
[59]Department of Clinical Physiology, Faculty of Medicine and Life Sciences, University of Tampere, Tampere, 33014, Finland
[60]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Dept of Health and Human Services, Research Triangle Park, NC, 27709, USA
[61]Cardiovascular Health Research Unit, Departments of Epidemiology, Medicine and Health Services, University of Washington, Seattle, WA, 98101, USA
[62]Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA
[63]Cardiovascular Health Research Unit, Department of Epidemiology, University of Washington, Seattle, WA, 98101, USA
[64]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, 10029-6574, USA
[65]Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, 10029-6574, USA
[66]Department of Health Sciences, University of Jyväskylä, Jyväskylä, FI-40014, Finland
[67]Sticht Center on Aging, Wake Forest School of Medicine, Winston-Salem, NC, USA
[68]Comprehensive Pneumology Center Munich (CPC-M), Member of the German Center for Lung Research, Munich, Germany
[69]Faculty of Medicine, University of Split, Split, Croatia
[70]Alzheimer Scotland Research Centre, University of Edinburgh, Edinburgh, EH8 9JZ, UK
[71]Department of Medical Sciences, Molecular Epidemiology and Science for Life Laboratory, Uppsala University, Uppsala, Sweden
[72]Department of Medicine, Division of Cardiovascular Medicine, Stanford University School of Medicine, Stanford, CA, 94305, USA
[73]Internal Medicine, Erasmus Medical Center, Rotterdam, 3000CA, Netherlands
[74]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, 17475, Germany
[75]DZHK (German Centre for Cardiovascular Research), partner site: Greifswald, Greifswald, Germany
[76]Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, 20521, Finland
[77]Research Centre of Applied and Preventative Cardiovascular Medicine, University of Turku, Turku, 20014, Finland
[78]Department of Epidemiology and Carolina Center for Genome Science, University of North Carolina, Chapel Hill, NC, 27514, USA
[79]Computational Medicine Core, Center for Lung Biology, UW Medicine Sleep Center, Department of Medicine, University of Washington, Seattle, WA, 98109, USA
[80]Respiratory Division, Department of Medicine, University of British Columbia, Vancouver, BC, Canada
[81]Institute for Translational Genomics and Population Sciences and Department of Pediatrics, Los Angeles Biomedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA, 90502, USA
[82]Pulmonary Center, Department of Medicine, Boston University School of Medicine, Boston, MA, 02118, USA
[83]National Heart, Lung and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA, 01702, USA
[84]Department of Health, University of Helsinki, Helsinki, FI-00014, Finland
[85]Department of Public Health, National Institute for Health and Welfare, Helsinki, FI-00271, Finland
[86]National Institute on Aging, National Institutes of Health, Bethesda, MD, 20892, USA
[87]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, 39216, USA

[88]Institute of Genetic Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, 85764, Germany
[89]Chair of Genetic Epidemiology, IBE, Faculty of Medicine, LMU Munich, Munich, 81377, Germany
[90]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, 10032, USA
[91]Department of Biostatistics, University of Liverpool, Liverpool, L69 3GL, UK
[92]Radiology, Erasmus Medical Center, Rotterdam, 3000CA, Netherlands
[93]Neurology, Erasmus Medical Center, Rotterdam, 3000CA, Netherlands
[94]Department of Internal Medicine - Pulmonary Diseases, Vivantes Klinikum Spandau Berlin, Berlin, 13585, Germany
[95]Population Health Research Institute, St George's, University of London, London, SW17 0RE, UK
[96]NIHR Nottingham Biomedical Research Centre and Division of Respiratory Medicine, University of Nottingham, Nottingham, NG7 2UH, UK

**v3** **First published:** 12 Jan 2018, **3**:4 (doi: 10.12688/wellcomeopenres.12583.1)
**Second version:** 21 Jun 2018, **3**:4 (doi: 10.12688/wellcomeopenres.12583.2)
**Latest published:** 07 Aug 2018, **3**:4 (doi: 10.12688/wellcomeopenres.12583.3)

## Abstract
**Background:** Over 90 regions of the genome have been associated with lung function to date, many of which have also been implicated in chronic obstructive pulmonary disease.
**Methods:** We carried out meta-analyses of exome array data and three lung function measures: forced expiratory volume in one second ($FEV_1$), forced vital capacity (FVC) and the ratio of $FEV_1$ to FVC ($FEV_1$/FVC). These analyses by the SpiroMeta and CHARGE consortia included 60,749 individuals of European ancestry from 23 studies, and 7,721 individuals of African Ancestry from 5 studies in the discovery stage, with follow-up in up to 111,556 independent individuals.
**Results:** We identified significant ($P<2{\cdot}8x10^{-7}$) associations with six SNPs: a nonsynonymous variant in *RPAP1*, which is predicted to be damaging, three intronic SNPs (*SEC24C, CASC17* and *UQCC1*) and two intergenic SNPs near to *LY86* and *FGF10.* Expression quantitative trait loci analyses found evidence for regulation of gene expression at three signals and implicated several genes, including *TYRO3* and *PLAU*.
**Conclusions:** Further interrogation of these loci could provide greater understanding of the determinants of lung function and pulmonary disease.

## Keywords
Lung function, respiratory, exome array, GWAS, COPD

**Open Peer Review**

**Referee Status:** ✔ ✔

| | Invited Referees 1 | 2 |
|---|---|---|
| REVISED **version 3** published 07 Aug 2018 | | |
| REVISED **version 2** published 21 Jun 2018 | ✔ report | |
| **version 1** published 12 Jan 2018 | ? report | ✔ report |

1 **Robin Beaumont**, University of Exeter, UK
**Rachel M. Freathy**, University of Exeter, UK

2 **Lisa Strug**, Hospital for Sick Children, Canada
**Naim Panjwani**, The Hospital for Sick Children, Canada

**Discuss this article**

Comments (0)

**Corresponding authors:** Martin D. Tobin (martin.tobin@le.ac.uk), Stephanie J. London (london2@niehs.nih.gov)

**Author roles: Jackson VE**: Formal Analysis, Writing – Original Draft Preparation; **Latourelle JC**: Formal Analysis, Writing – Review & Editing; **Wain LV**: Formal Analysis, Supervision, Writing – Review & Editing; **Smith AV**: Data Curation, Formal Analysis, Writing – Review & Editing; **Grove ML**: Data Curation, Writing – Review & Editing; **Bartz TM**: Formal Analysis, Writing – Review & Editing; **Obeidat M**: Formal Analysis, Writing – Review & Editing; **Province MA**: Conceptualization, Data Curation, Writing – Review & Editing; **Gao W**: Formal Analysis, Writing – Review & Editing; **Qaiser B**: Formal Analysis, Writing – Review & Editing; **Porteous DJ**: Data Curation; **Cassano PA**: Data Curation, Formal Analysis, Writing – Review & Editing; **Ahluwalia TS**: Conceptualization, Data Curation, Writing – Review & Editing; **Grarup N**: Conceptualization, Data Curation, Writing – Review & Editing; **Li J**: Data Curation, Formal Analysis, Writing – Review & Editing; **Altmaier E**: Formal Analysis, Writing – Review & Editing; **Marten J**: Formal Analysis, Writing – Review & Editing; **Harris SE**: Data Curation, Formal Analysis, Writing – Review & Editing; **Manichaikul A**: Data Curation, Formal Analysis, Writing – Review & Editing; **Pottinger TD**: Data Curation, Formal Analysis, Writing – Review & Editing; **Li-Gao R**: Data Curation, Formal Analysis, Writing – Review & Editing; **Lind-Thomsen A**: Data Curation, Formal Analysis, Writing – Review & Editing; **Mahajan A**: Formal Analysis, Writing – Review & Editing; **Lahousse L**: Conceptualization, Data Curation, Formal Analysis, Writing – Review & Editing; **Imboden M**: Data Curation, Formal Analysis, Writing – Review & Editing; **Teumer A**: Data Curation, Formal Analysis, Writing – Review & Editing; **Prins B**: Data Curation, Formal Analysis, Writing – Review & Editing; **Lyytikäinen LP**: Data Curation, Formal Analysis, Writing – Review & Editing; **Eiriksdottir G**: Conceptualization, Data Curation, Writing – Review & Editing; **Franceschini N**: Formal Analysis, Writing – Review & Editing; **Sitlani CM**: Formal Analysis, Writing – Review & Editing; **Brody JA**: Data Curation, Formal Analysis, Writing – Review & Editing; **Bossé Y**: Data Curation, Writing – Review & Editing; **Timens W**: Data Curation, Writing – Review & Editing; **Kraja A**: Data Curation, Formal Analysis, Writing – Review & Editing; **Loukola A**: Data Curation, Writing – Review & Editing; **Tang W**: Data Curation, Formal Analysis, Writing – Review & Editing; **Liu Y**: Data Curation, Formal Analysis, Writing – Review & Editing; **Bork-Jensen J**: Conceptualization, Data Curation, Writing – Review & Editing; **Justesen JM**: Formal Analysis, Writing – Review & Editing; **Linneberg A**: Conceptualization, Writing – Review & Editing; **Lange LA**: Data Curation, Writing – Review & Editing; **Rawal R**: Data Curation, Writing – Review & Editing; **Karrasch S**: Data Curation, Writing – Review & Editing; **Huffman JE**: Formal Analysis, Writing – Review & Editing; **Smith BH**: Data Curation, Writing – Review & Editing; **Davies G**: Data Curation, Writing – Review & Editing; **Burkart KM**: Conceptualization, Writing – Review & Editing; **Mychaleckyj JC**: Data Curation, Writing – Review & Editing; **Bonten TN**: Data Curation, Writing – Review & Editing; **Enroth S**: Data Curation, Formal Analysis, Writing – Review & Editing; **Lind L**: Data Curation, Writing – Review & Editing; **Brusselle GG**: Conceptualization, Data Curation, Writing – Review & Editing; **Kumar A**: Data Curation, Formal Analysis, Writing – Review & Editing; **Stubbe B**: Conceptualization, Data Curation, Writing – Review & Editing; **Kähönen M**: Conceptualization, Data Curation, Writing – Review & Editing; **Wyss AB**: Conceptualization, Formal Analysis, Writing – Review & Editing; **Psaty BM**: Conceptualization, Data Curation, Writing – Review & Editing; **Heckbert SR**: Data Curation, Writing – Review & Editing; **Hao K**: Data Curation, Writing – Review & Editing; **Rantanen T**: Conceptualization, Data Curation, Writing – Review & Editing; **Kritchevsky SB**: Conceptualization, Data Curation, Writing – Review & Editing; **Lohman K**: Data Curation, Formal Analysis, Writing – Review & Editing; **Skaaby T**: Conceptualization, Writing – Review & Editing; **Pisinger C**: Conceptualization, Data Curation, Writing – Review & Editing; **Hansen T**: Conceptualization, Data Curation, Formal Analysis, Writing – Review & Editing; **Schulz H**: Conceptualization, Writing – Review & Editing; **Polasek O**: Conceptualization, Data Curation, Writing – Review & Editing; **Campbell A**: Data Curation, Writing – Review & Editing; **Starr JM**: Data Curation, Writing – Review & Editing; **Rich SS**: Conceptualization, Data Curation, Writing – Review & Editing; **Mook-Kanamori DO**: Conceptualization, Data Curation, Writing – Review & Editing; **Johansson Å**: Data Curation, Writing – Review & Editing; **Ingelsson E**: Data Curation, Writing – Review & Editing; **Uitterlinden AG**: Conceptualization, Data Curation, Writing – Review & Editing; **Weiss S**: Data Curation, Formal Analysis, Writing – Review & Editing; **Raitakari OT**: Conceptualization, Data Curation, Writing – Review & Editing; **Gudnason V**: Conceptualization, Formal Analysis, Writing – Review & Editing; **North KE**: Data Curation, Writing – Review & Editing; **Gharib SA**: Writing – Review & Editing; **Sin DD**: Data Curation, Writing – Review & Editing; **Taylor KD**: Data Curation, Writing – Review & Editing; **O'Connor GT**: Data Curation, Writing – Review & Editing; **Kaprio J**: Conceptualization, Data Curation, Writing – Review & Editing; **Harris TB**: Conceptualization, Data Curation, Writing – Review & Editing; **Pederson O**: Data Curation, Formal Analysis, Writing – Review & Editing; **Vestergaard H**: Data Curation, Formal Analysis, Writing – Review & Editing; **Wilson JG**: Data Curation, Writing – Review & Editing; **Strauch K**: Data Curation, Writing – Review & Editing; **Hayward C**: Conceptualization, Data Curation, Formal Analysis, Writing – Review & Editing; **Kerr S**: Data Curation, Writing – Review & Editing; **Deary IJ**: Data Curation, Writing – Review & Editing; **Barr RG**: Conceptualization, Data Curation, Writing – Review & Editing; **de Mutsert R**: Conceptualization, Data Curation, Writing – Review & Editing; **Gyllensten U**: Conceptualization, Data Curation, Writing – Review & Editing; **Morris AP**: Data Curation, Formal Analysis, Writing – Review & Editing; **Ikram MA**: Conceptualization, Writing – Review & Editing; **Probst-Hensch N**: Conceptualization, Data Curation, Formal Analysis, Writing – Review & Editing; **Gläser S**: Conceptualization, Data Curation, Writing – Review & Editing; **Zeggini E**: Conceptualization, Writing – Review & Editing; **Lehtimäki T**: Conceptualization, Data Curation, Writing – Review & Editing; **Strachan DP**: Conceptualization, Data Curation, Writing – Review & Editing; **Dupuis J**: Formal Analysis, Supervision, Writing – Review & Editing; **Morrison AC**: Formal Analysis, Writing – Review & Editing; **Hall IP**: Conceptualization, Formal Analysis, Supervision, Writing – Review & Editing; **Tobin MD**: Conceptualization, Formal Analysis, Supervision, Writing – Review & Editing; **London SJ**: Conceptualization, Formal Analysis, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** MDT has been supported by MRC fellowships G0501942 and G0902313. MDT and LVW are supported by the MRC (MR/N011317/1). IPH is supported by the MRC (G1000861). ALW and SJL are supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (ZIA ES 043012). We acknowledge use of phenotype and genotype data from the British 1958 Birth Cohort DNA collection, funded by the Medical Researanch Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. APM was a Wellcome Trust Senior Fellow in Basic Biomedical Science (grant number WT098017) and was also supported by Wellcome Trust grant WT064890. EI is supported by the Swedish Research Council (2012-1397), Knut och Alice Wallenberg Foundation (2013.0126) and the Swedish Heart-Lung Foundation (20140422). JK is supported by Academy of Finland Center of Excellence in Complex Disease Genetics grants 213506, 129680 and Academy of Finland grants 265240, 263278. The Finnish Twin Cohort is supported by the Welcome Trust Sanger Institute, UK. The Lothian Birth Cohort is supported by Age UK (The Disconnected Mind Project), the UK Medical Research Council (MR/K026992/1) and The Royal Society of Edinburgh. ÅJ is supported by the Swedish Society for Medical Research (SSMF), The Kjell och Märta Beijers Foundation, The Marcus Borgström Foundation, The Åke Wiberg foundation and The Vleugels Foundation. UG is supported by Swedish Medical Research Council grants K2007-66X-20270-01-3 and 2011-2354 and European Commission FP6 (LSHG-CT-2006-01947). SHIP is part of the Community Medicine

Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research, the Ministry of Cultural Affairs as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine (GANI_MED)' funded by the Federal Ministry of Education and Research, and the German Asthma and COPD Network (COSYCONET) (grant no.01ZZ9603, 01ZZ0103, 01ZZ0403, 03IS2061A, BMBF 01GI0883). ExomeChip data have been supported by the Federal Ministry of Education and Research (grant no. 03Z1CN22) and the Federal State of Mecklenburg-West Pomerania. The University of Greifswald is a member of the Caché Campus program of the InterSystems GmbH. UKHLS is supported by grants WT098051 (Wellcome Trust) and ES/H029745/1 (Economic and Social Research Council). Y.B. holds a Canada Research Chair in Genomics of Heart and Lung Diseases. Lies Lahousse is a Postdoctoral Fellow of the Research Foundation - Flanders (FWO grant G035014N). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Scientific Research (NOW), the Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. Genotyping in the Rotterdam study was supported by Netherlands Organization for Scientific Research (NOW grants 175.010.2005.011 ; 911-03-305 012), the Research Institute for Diseases in the Elderly (RIDE2 grants 014-93-015) and Netherlands Genomics Initiative (NGI)/Netherlands Consortium for Healthy Aging (NCHA grant050-060-810). MESA/MESA SHARe is supported by HHS (HHSN268201500003I), NIH/NHLBI (contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169) and HIH/NCATS (contracts UL1-TR-000040, UL1-TR-001079, UL1-TR-001881, DK063491). MESA SHARe is funded by NIH/NHLBI contract N02-HL-64278, MESA Air is funded by US EPA (RD831697) and MESA Spirometry funded by NIH/NHLBI (R01-HL077612). SSR and BMP are supported by NIH/NHLBI grant rare variants and NHLBI traits in deeply phenotyped cohorts (R01-HL120393). Cardiovascular Health Study: This CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, HHSN268200960009C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and NHLBI grants U01HL080295, R01HL068986, R01HL087652, R01HL105756, R01HL103612, R01HL120393, and R01HL130114 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 and R01HL085251 from the National Institute on Aging (NIA). The provision of genotyping data was suprovidedpported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR001881, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Atherosclerosis Risk in Communities (ARIC) study is carried out as a collaborative study supported by the National Heart, Lung, and Blood Institute (NHLBI) contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). Funding support for "Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium" was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). DOMK received funding from the Dutch Science Organisation (ZonMW-VENI Grant 916.14.023). The genotyping in the NEO study was supported by the Centre National de Génotypage (Paris, France), headed by Jean-François Deleuze. The NEO study is supported by the participating Departments, the Division and the Board of Directors of the Leiden University Medical Center, and by the Leiden University, Research Profile Area Vascular and Regenerative Medicine. SAPALDIA was supported by the Swiss National Science Foundation (grants no 33CS30-148470/1, 33CSCO-134276/1, 33CSCO-108796, , 324730_135673, 3247BO-104283, 3247BO-104288, 3247BO-104284, 3247-065896, 3100-059302, 3200-052720, 3200-042532, 4026-028099, PMPuIDP3_129021/1, PMPDP3_141671/1), the Federal Office for the Environment, the Federal Office of Public Health, the Federal Office of Roads and Transport, the canton's government of Aargau, Basel-Stadt, Basel-Land, Geneva, Luzern, Ticino, Valais, and Zürich, the Swiss Lung League, the canton's Lung League of Basel Stadt/ Basel Landschaft, Geneva, Ticino, Valais, Graubünden and Zurich, Stiftung ehemals Bündner Heilstätten, SUVA, Freiwillige Akademische Gesellschaft, UBS Wealth Foundation, Talecris Biotherapeutics GmbH, Abbott Diagnostics, European Commission 018996 (GABRIEL), Wellcome Trust WT 084703MA. The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk). Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, University of Edinburgh, Scotland and was funded by the Medical Research Council UK.. The Croatia KORCULA study was supported by the Ministry of Science, Education and Sport in the Republic of Croatia (108-1080315-0302). JD, JCL, WG and GTOC are supported by NIH/NHLBI Contract HHSN268201500001I. Genotyping, quality control and calling of the Illumina HumanExome BeadChip in the Framingham Heart Study was supported by funding from the National Heart, Lung and Blood Institute Division of Intramural Research (Daniel Levy and Christopher J. O'Donnell, Principle Investigators). The AGES study is supported by the NIH (N01-AG012100), the Iceland Parliament (Alþingi) and the Icelandic Heart Association. HABC was supported by NIA contracts N01AG62101, N01AG62103, and N01AG62106; NIA grant R01-AG028050, and NINR grant R01- NR012459 and was supported in part by the Intramural Research Program of the NIH, National Institute on Aging. The HABC genome-wide association study was funded by NIA grant 1R01AG032098- 01A1 and genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268200782096C. We thank the Jackson Heart Study (JHS) participants and staff for their contributions to this work. The JHS is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities. JGW is supported by U54GM115428 from the National Institute of General Medical Sciences.
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*



**How to cite this article:** Jackson VE, Latourelle JC, Wain LV *et al.* **Meta-analysis of exome array data identifies six novel genetic loci for lung function [version 3; referees: 2 approved]** Wellcome Open Research 2018, **3**:4 (doi: 10.12688/wellcomeopenres.12583.3)

**First published:** 12 Jan 2018, **3**:4 (doi: 10.12688/wellcomeopenres.12583.1)

**REVISED** **Amendments from Version 2**

We have added a further limitation to the discussion of the paper outlining a recently highlighted issue regarding the trait transformation undertaken in our replication analyses. We show through sensitivity analyses that our results are not affected by this issue (Supplementary Figure 4), but note that future studies should avoid such a transformation.

**See referee reports**

## Introduction

Measures of lung function act as predictors of mortality and morbidity and form the basis for the diagnosis of several diseases, most notably chronic obstructive pulmonary disease (COPD), one of the leading causes of death globally[1]. Environmental factors, including smoking and exposure to air pollution play a significant role in lung function; however there has also been shown to be a genetic component, with estimates of the narrow sense heritability ranging between 39–66%[2–5]. Genome-wide association studies (GWAS) of lung function have identified associations between single nucleotide polymorphisms (SNPs) and lung function at over 150 independent loci to date[6–14]. Associations have also been identified in GWAS of COPD[15–19]; however, the identification of disease associated SNPs has been restricted by limited sample sizes. Many signals first identified in powerful studies of quantitative lung function traits, have been found to be associated with risk of COPD, highlighting the potential clinical usefulness of comprehensive identification of lung function associated SNPs[13].

Low frequency (minor allele frequency (MAF) 1–5%) and rare (MAF<1%) variants have been largely underexplored by GWAS to date. Exome arrays have been designed to facilitate the investigation of these low frequency and rare variants, predominately within coding regions, in large sample sizes. Alongside a core content of rare coding SNPs, the exome array additionally includes common variation, including tags for previously identified GWAS hits, ancestry informative SNPs, a grid of markers for estimating identity by descent and a random selection of synonymous SNPs[20].

An earlier version of this article can be found on bioRxiv (https://doi.org/10.1101/164426)

## Results

We carried out a meta-analysis of exome array data and three lung function measures: forced expiratory volume in one second ($FEV_1$), forced vital capacity (FVC) and the ratio of $FEV_1$ to FVC ($FEV_1$/FVC). These analyses included 68,470 individuals from the SpiroMeta and CHARGE consortia in a discovery analysis, with follow-up in an independent sample of up to 111,556 individuals. All studies are listed with their study-specific sample characteristics in Table 1, with full study descriptions, including details of spirometry and other measurements described in the Supplementary Note. The genotype calling procedures implemented by each study (Supplementary Table 1) and quality control of genotype data are described in the Supplementary Methods. We have undertaken both single variant analyses, and gene-based associations, which test for the joint effect of several rare variants in a gene (see *Methods* for details).

### Meta-analyses of single variant associations

We first evaluated single variant associations between $FEV_1$, FVC and $FEV_1$/FVC and the 179,215 SNPs that passed study level quality control and were polymorphic in both consortia. These analyses identified 34 SNPs in regions not previously associated with lung function, showing association with at least one trait at overall $P<10^{-5}$, and showing association with consistent direction and $P<0{\cdot}05$ in both consortia (full results in Supplementary Table 2, quantile-quantile and Manhattan plots shown in Supplementary Figure 1). We followed up these SNP associations in a replication analysis comprising 3 studies with 111,556 individuals. Combining the results from the discovery and replication stages in a meta-analysis identified six SNPs in total that were independent to known signals and met the pre-defined significance threshold ($P<2{\cdot}8\times10^{-7}$) overall in, or near to *FGF10*, *LY86*, *SEC24C*, *RPAP1, CASC17* and *UQCC1* (Table 2, Supplementary Figure 2). A SNP near to the *CASC17* signal (rs11654749, $r^2=0{\cdot}3$ with rs1859962) has previously been associated with $FEV_1$ in a genome-wide analysis of gene-smoking interactions, although this association was not replicated at the time[21]; the present analysis provides the first evidence for independent replication of this signal. A seventh signal was also identified in *LCT* (Table 2, Supplementary Figure 2); whilst this locus has not previously been implicated in lung function, this SNP is known to vary in frequency across European populations[22], and we cannot rule out that this association is not an artefact of population structure. Our discovery analysis furthermore identified associations ($P<10^{-5}$) in 25 regions previously associated with one or more of $FEV_1$, FVC and $FEV_1$/FVC (Supplementary Table 3).

Generally, the observed effect of the SNPs at the novel signals were similar in ever and never smokers; the exception was rs1448044 near *FGF10*, which showed a significant association with FVC only in ever smokers in our discovery analysis (ever smokers $P=1{\cdot}49\times10^{-6}$; never smokers $P=0{\cdot}695$, Supplementary Table 4 and Supplementary Figure 3). In the replication analysis, however, this association was observed in both ever and never smokers (ever smokers $P=3{\cdot}14\times10^{-5}$; never smokers $P=1{\cdot}40\times10^{-4}$, Supplementary Table 5). For rs1200345 (*RPAP1*) and rs1859962 (*CASC17*), associations were most statistically significant in the analyses restricted to individuals of European Ancestry (Supplementary Table 4 and Supplementary Figure 3), as was the association with rs2322659 (*LCT*), giving further support that this association may be due to population stratification.

### Meta-analyses of gene-based associations

We undertook Weighted Sum Tests (WST)[23] and Sequence Kernel Association tests (SKAT)[24] to assess the joint effects of multiple low frequency variants within genes on lung function traits. In our discovery analyses of all 68,470 individuals, we tested up to 14,380 genes that had at least two variants with MAF<5% and met the inclusion criteria (exonic or loss of function [LOF], see *Methods* for definitions) in both consortia. The SKAT analyses identified 16 genes associated ($P<0{\cdot}05$ in both consortia and overall $P<10^{-4}$) with $FEV_1$, FVC or $FEV_1$/FVC (Supplementary Table 6), whilst the WST analyses identified 12 genes

**Table 1. Sample characteristics of 11 SpiroMeta and 12 CHARGE studies contributing to the discovery analyses and three studies contributing to the replication analyses.**

| **Discovery studies** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **SpiroMeta studies** | **Total sample** | **n (%) Male** | **Ever smokers, n (%)** | **Age, mean (SD)** | **$FEV_1$, litres. mean (SD)** | **FVC, litres. mean (SD)** | **$FEV_1$/FVC, mean (SD)** |
| 1958 British Birth Cohort (B58C) | 5270 | 2961 (56·2%) | 2866 (53·3%) | 44·00 (0·00) | 3·35 (0·79) | 4·29 (1·03) | 0·788 (0·09) |
| Generation Scotland (GS:SFHS) | 8164 | 3413 (41·8%) | 3806 (46·6%) | 51·59 (13·33) | 2·78 (0·87) | 3·91 (1·01) | 0·710 (0·12) |
| Cooperative Health Research in the Region of Augsburg (KORA F4) | 1447 | 701 (48·5%) | 900 (62·2%) | 54·82 (9·66) | 3·24 (0·85) | 4·20 (1·04) | 0·771 (0·07) |
| CROATIA-Korcula cohort (KORCULA) | 791 | 296 (36·8%) | 418 (52·0%) | 55·56 (13·69) | 2·72 (0·83) | 3·29 (0·95) | 0·829 (0·10) |
| Lothian Birth Cohort 1936 (LBC1936) | 974 | 501 (50·6%) | 554 (55·9%) | 69·55 (0·84) | 2·38 (0·67) | 3·04 (0·87) | 0·787 (0·10) |
| Study of Health in Pomerania (SHIP) | 1681 | 831 (49·4%) | 955 (56·8%) | 52·25 (13·43) | 3·29 (0·88) | 3·88 (1·03) | 0·848 (0·07) |
| Northern Swedish Population Health Study (NSPHS) | 880 | 407 (46·3%) | 122 (13·9%) | 49·13 (19·96) | 2·93 (0·90) | 3·53 (1·06) | 0·831 (0·09) |
| Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) | 836 | 413 (49·4%) | 426 (51·0%) | 70·20 (0·17) | 2·44 (0·68) | 3·20 (0·87) | 0·76 (0·10) |
| Swiss study on Air Pollution and Lung Disease in adults (SAPALDIA) | 2707 | 1379 (50·9%) | 1399 (51·7%) | 40·86 (10·92) | 3·65 (0·83) | 4·62 (1·04) | 0·794 (0·07) |
| The Cardiovascular Risk in Young Finns Study (YFS) | 434 | 198 (47·3%) | 186 (44·4%) | 38·88 (5·07) | 3·73 (0·75) | 4·68 (0·99) | 0·800 (0·06) |
| Finnish Twin Cohort (FTC) | 214 | 0 (0%) | 0 (0%) | 68·73 (3·31) | 2·18 (0·47) | 2·79 (0·58) | 0·786 (0·08) |
| **Total** | **23,398** | | | | | | |
| **CHARGE studies (European Ancestry)** | **Total sample** | **n (%) Male** | **Ever smokers, n (%)** | **Age, mean (SD)** | **$FEV_1$, litres. mean (SD)** | **FVC, litres. mean (SD)** | **$FEV_1$/FVC, mean (SD)** |
| AGES-Reykjavik study (AGES) | 1566 | 649 (41·4%) | 900 (57·5%) | 76·1 (5·62) | 2·13 (0·70) | 2·87 (0·86) | 0·744 (0·09) |
| Atherosclerosis Risk in Communities Study (ARIC) | 10,680 | 5015 (47·0%) | 631 (59·1%) | 54·3 (5·70) | 2·94 (0·77) | 3·98 (0·98) | 0·738 (0·07) |
| Cardiovascular Health Study (CHS) | 3967 | 1737 (43·8%) | 2089 (52·7%) | 72·8 (5·55) | 2·11 (0·66) | 3·00 (0·86) | 0·702 (0·10) |
| NHLBI Family Heart Study (FAMHS) | 1651 | 718 (43·5%) | 698 (42·3) | 53·5 (12·60) | 2·91 (0·853) | 3·89 (1·05) | 0·746 (0·08) |
| Framingham Heart Study (FHS) | 7113 | 3241 (45·5%) | 3780 (53·1) | 50·7 (14·12) | 3·10 (0·925) | 4·09 (1·12) | 0·755 (0·08) |
| Health Aging and Body Composition Study (HABC) | 1457 | 786 (53·2%) | 831 (56·5%) | 73·7 (2·83) | 2·31 (0·66) | 3·11 (0·81) | 0·741 (0·08) |
| Health2006 Study | 2714 | 1217 (44·8%) | 1577 (58·1%) | 49·4 (13·04) | 3·13 (0·82) | 3·99 (0·99) | 0·784 (0·07) |
| Health2008 Study | 687 | 297 (43·2%) | 384 (55·9%) | 46·7 (8·22) | 3·27 (0·79) | 4·13 (0·97) | 0·791 (0·06) |
| Inter99 Study (without pack-years) | 1115 | 549 (49·2%) | 1115 (100%) | 47·2 (7·76) | 3·26 (0·71) | 4·12 (0·92) | 0·796 (0·07) |
| Inter99 Study (with pack-years) | 4179 | 2027 (48·5%) | 2307 (55·2%) | 45·8 (7·95) | 3·21 (0·76) | 4·10 (0·97) | 0·788 (0·08) |
| Multi-Ethnic Study of Atherosclerosis (MESA) | 1323 | 654 (49·4%) | 751 (56·8%) | 66·0 (9·8) | 2·57 (0·76) | 3·51 (0·10) | 0·733 (0·08) |
| The Rotterdam Study (RS) | 546 | 299 (54·8%) | 382 (70·0%) | 79·4 (5·00) | 2·27 (0·68) | 3·03 (0·86) | 0·750 (0·08) |
| **Total** | **36,998** | | | | | | |

| **Discovery studies** | | | | | | | |
|---|---|---|---|---|---|---|---|
| **CHARGE studies (African Ancestry)** | **Total Sample** | **n (%) Male** | **Ever smokers, n (%)** | **Age, mean (SD)** | **$FEV_1$, litres. mean (SD)** | **FVC, litres. mean (SD)** | **$FEV_1$/FVC, mean (SD)** |
| Atherosclerosis Risk in Communities Study (ARIC) | 3180 | 1183 (37·2%) | 1680 (59·1%) | 53·6 (5·83) | 2·48 (0·65) | 3·25 (0·82) | 0·765 (0·08) |
| Cardiovascular Health Study (CHS) | 624 | 232 (37·2%) | 340 (54·4%) | 73·2 (5·49) | 1·76 (0·58) | 2·48 (0·80) | 0·717 (0·11) |
| Health Aging and Body Composition Study (HABC) | 943 | 433 (45·9%) | 543 (57·6%) | 73·4 (2·90) | 1·96 (0·57) | 2·61 (0·71) | 0·749 (0·09) |
| Jackson Heart Study (JHS) | 2143 | 793 (36·8%) | 688 (31·9%) | 52·8 (12·6) | 2·43 (0·72) | 3·02 (0·86) | 0·807 (0·09) |
| Multi-Ethnic Study of Atherosclerosis (MESA) | 861 | 404 (46·9%) | 467 (54·2%) | 65·6 (9·6) | 2·19 (0·66) | 2·92 (0·86) | 0·756 (0·09) |
| **Total** | **7721** | | | | | | |
| **Replication studies** | | | | | | | |
| **Study name** | **Total Sample** | **n (%) Male** | **Ever smokers, n (%)** | **Age, mean (SD)** | **$FEV_1$, litres. mean (SD)** | **FVC, litres. mean (SD)** | **$FEV_1$/FVC, mean (SD)** |
| UK Biobank | 98,657 | 45,166 (45·8%) | 56,404 (57·2%) | 56·7 (7·92) | 2·75 (0·80) | 3·67 (0·98) | 0·75 (0·07) |
| UK Household Longitudinal Study (UKHLS) | 7443 | 3293 (44·2%) | 4509 (60·5%) | 53·10 (15·94) | 2·89 (0·90) | 3·83 (1·08) | 0·753 (0·09) |
| Netherlands Epidemiology of Obesity study (NEO) | 5456 | 2672 (48·0%) | 3674 (66·0%) | 55·9 (5·9) | 3·26 (0·80) | 4·26 (1·02) | 0·77 (0·07) |
| **Total** | **111,556** | | | | | | |

(Supplementary Table 7). There was one gene (*LY6G6D*) that was identified in both analyses. These genes were followed up in UK Biobank, with two genes, *GPR126* and *LTBP4,* showing evidence of replication in the exonic SKAT analysis ($P<3{\cdot}5\times10^{-6}$); however conditional analyses in UK Biobank showed that both these associations were driven by single SNPs, that were identified in the single variant association analyses and have been previously reported in GWAS of these traits (Supplementary Table 6 and Supplementary Table 7).

### Functional characterization of novel loci

In order to gain further insight into the six loci identified in our analyses of single variant associations (excluding *LCT*), we employed functional annotation and assessed whether identified SNPs in these regions were associated with gene expression levels. One of the identified novel SNPs was nonsynonymous, three intronic and two were intergenic. We found evidence that three of the SNPs may be involved in cis-acting regulation of the expression of several genes in multiple tissues (Supplementary Table 8).

SNP rs1200345 in *RPAP1* is a nonysynomous variant, predicted to be deleterious by both SIFT (deleterious) and Polyphen (possibly damaging) (Supplementary Table 9); *RPAP1* is ubiquitously expressed, with high levels of protein detected in the lung (Supplementary Table 10). SNP rs1200345 or proxies ($r^2>0{\cdot}8$) were also found to be amongst the most strongly associated SNPs with expression levels of *RPAP1* in several tissues, including lung, and with a further six genes in lung tissue (Supplementary Table 8), including *TYRO3,* one of the TAM family of receptor tyrosine kinases. *TYRO3* regulates several processes including cell survival, migration and differentiation and is highly expressed in lung macrophages (Supplementary Table 10). Evidence of association with gene expression was found at two more of the novel signals (sentinel SNPs rs3849969 and rs6088813), implicating a further 16 genes. Of note, in blood expression quantitative trait loci (eQTL) databases, a proxy of a SNP in complete linkage disequilibrium ($r^2=1$) with rs3849969 (rs3812637) was an eQTL for plasminogen activator, urokinase (*PLAU*).

## Discussion

We undertook an analysis of 68,470 individuals from 23 studies with data from the exome array and three lung function traits, following up the most significant single SNP and gene-based associations in an independent sample of up to 111,556 individuals. There were six SNPs which reached $P<10^{-5}$ in the discovery stage meta-analysis of single variant associations, and subsequently met the Bonferroni corrected significance threshold for independent replication ($P<1{\cdot}47\times10^{-3}$, corrected for 34 SNPs being tested). In the combined analyses of our discovery and replication analyses, these six SNPs met the exome chip-wide significance threshold ($P<2{\cdot}8\times10^{-7}$). One of the SNPs is in a region that has previously been implicated in lung function (near *KCJN2/SOX9*)[21], whilst the remaining five SNPs, although all common, have not previously been identified in other GWAS of lung function. In a recent 1000 Genomes imputed analysis of lung function (which includes some of the studies contributing to

**Table 2. Novel loci associated with lung function traits.** Results are shown for variant in novel loci associated ($P<2{\cdot}7\times10^{-7}$) with lung function traits in a two stage meta-analysis consisting of up to 68,470 individuals from the SpiroMeta and CHARGE Consortia in the discovery analyses, with follow-up in up to 111,556 individuals from UK Biobank, UKHLS and NEO. For each SNP, the result for the trait-smoking-ancestry combination which resulted in the most statistically significant association is given. The results for these SNPs and all three traits are shown in Supplementary Table 12. Beta values from SpiroMeta ($\beta_{Sp}$) reflect effect-size estimates on an inverse-normal transformed scale after adjustments for age, age$^2$, sex, height and ancestry principal components, and stratified by ever smoking status (Analysis of All individuals only). Beta values from CHARGE ($\beta_{CH}$) reflect effect-size estimates on an untransformed scale (litres for $FEV_1$ and FVC; ratio for $FEV_1$/FVC). Samples sizes (N), Z-statistics (Z) and two-sided P-values (P) are given for the combined discovery analysis and the replication analysis. Two-sided P-values are also given for the full two-stage combined analyses (discovery + replication).

| | | | | | | | | Consortium results | | Combined discovery meta-analysis | | | Replication | | | Two-stage combined |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **SNP** | **Chr:Pos** | **(Nearest) gene(s)** | **Trait** | **Smoking** | **Ancestry** | **Effect/other allele** | **Effect allele frequency (Discovery)** | $\beta_{CH}$ | $\beta_{Sp}$ | $N_{disc}$ | $Z_{disc}$ | $P_{disc}$ | $N_{rep}$ | $Z_{rep}$ | $P_{rep}$ | $P_{meta}$ |
| rs2322659 | 2:136555659 | LCT (nonsynonymous) | FVC | All Individuals | EA Only | T/C | 23·5% | 27·34 | 0·032 | 55,591 | 5·597 | $2{\cdot}18\times10^{-8}$ | 12,899 | 2·286 | 0·0223 | $1{\cdot}70\times10^{-9}$ |
| rs1448044 | 5:44296986 | *FGF10*(dist=8111), *NNT*(dist=591,318) | FVC | Ever Smokers | EA+AA | A/G | 35·6% | 18·63 | 0·057 | 30,966 | 4·813 | $1{\cdot}49\times10^{-6}$ | 64,400 | 4·805 | $1{\cdot}55\times10^{-6}$ | $2{\cdot}22\times10^{-11}$ |
| rs1294421 | 6:6743149 | *LY86*(dist=87,933), *RREB1*(dist=364,681) | $FEV_1$/FVC | All Individuals | EA+AA | T/G | 36·8% | -0·222 | -0·038 | 68,099 | -5·479 | $4{\cdot}27\times10^{-8}$ | 111,556 | -8·171 | $3{\cdot}06\times10^{-16}$ | $9{\cdot}74\times10^{-23}$ |
| rs3849969 | 10: 75525999 | *SEC24C* (intronic) | FEV1 | All Individuals | EA+AA | T/C | 29·4% | 13·10 | 0·036 | 68,116 | 4·767 | $1{\cdot}87\times10^{-6}$ | 111,556 | 5·042 | $4{\cdot}60\times10^{-7}$ | $4{\cdot}99\times10^{-12}$ |
| rs1200345 | 15: 41819716 | *RPAP1* (nonsynonymous) | FEV1/FVC | All Individuals | EA only | C/T | 48·8% | -0·217 | -0·025 | 60,381 | -4·586 | $4{\cdot}51\times10^{-6}$ | 111,556 | -5·725 | $1{\cdot}03\times10^{-8}$ | $2{\cdot}33\times10^{-13}$ |
| rs1859962 | 17: 69108753 | *CASC17* (intronic) | $FEV_1$ | All Individuals | EA only | G/T | 48·2% | 15·39 | 0·026 | 60,395 | 4·876 | $1{\cdot}08\times10^{-6}$ | 111,554 | 4·612 | $3{\cdot}99\times10^{-6}$ | $4{\cdot}10\times10^{-11}$ |
| rs6088813 | 20: 33975181 | *UQCC1* (intronic) | FVC | All Individuals | EA+AA | C/A | 36·7% | -16·16 | -0·023 | 68,115 | -4·634 | $3{\cdot}58\times10^{-6}$ | 111,556 | -7·688 | $1{\cdot}50\times10^{-14}$ | $4{\cdot}90\times10^{-19}$ |

the present discovery analysis), all of these SNPs showed at least suggestive association ($2{\cdot}97\times10^{-3}>P>1{\cdot}28\times10^{-5}$) with one or more lung function trait, but none reached the required threshold ($P<5\times10^{-6}$) to be taken forward for replication in that analysis[12].

We further identified a seventh association with rs2322659 in *LCT* (MAF=23·5%; combined discovery + replication $P=1{\cdot}70\times10^{-9}$). Given SNPs in this region are known to vary in frequency across European populations, we cannot dismiss the possibility that this association may be confounded by population stratification; hence we do not report this signal as a novel lung function locus. For SNPs at 7 loci that have been shown to have differences in allele frequency between individuals from different regions of the UK[25], and subsequently European populations (including the *LCT* locus), we undertook a look-up of associations with lung function in our discovery analyses. and subsequently across European populations[26]. Aside from the association between the *LCT* locus and FVC, no significant associations were observed between SNPs at these loci and any lung function trait, in either the analyses restricted to European Ancestry (EA) individuals, or in the analysis of EA and African Ancestry (AA) individuals combined (Supplementary Table 11); this suggests population structure was generally accounted for adequately in our analyses.

One of the novel signals was with a nonsynonymous SNP, rs1200345 in *RPAP1*, (MAF=48·8%; $P=2{\cdot}33\times10^{-13}$), which is predicted to be deleterious. This SNP and proxies with $r^2>0{\cdot}8$ were also associated with expression in lung tissue of seven genes, including *RPAP1* and the TAM receptor *TYRO3*. TAM receptors play a role in the inhibition of Toll-like receptors (TLRs)-mediated innate immune response by initiating the transcription of cytokine signalling genes (SOCS-1 and 3), which limit cytokine overproduction and inflammation[27,28]. It has been shown that influenza viruses H5N1 and H7N9 can cause downregulation of Tyro3, resulting in an increased inflammatory cytokine response[28].

Three further signals were with intronic SNPs in *SEC24C* (MAF=29·4%; $P=4{\cdot}99\times10^{-12}$), *CASC17* (MAF=48·2%; $P=4{\cdot}10\times10^{-11}$), and *UQCC1* (MAF=36·7%; $P=4{\cdot}90\times10^{-19}$). Two of these intronic SNPs have previously been implicated in GWAS of other traits: rs1859962 in *CASC17* with prostate cancer[29] and rs6088813 in *UQCC1* with height[30]. The *CASC17* locus, near *KCNJ2/SOX9* has also previously been implicated in lung function, showing significant association with $FEV_1$ in a genome-wide analysis of gene-smoking interactions; however, this association was not formally replicated[21]. Whilst the individuals utilised in the discovery stage of this analysis overlap with those included in this previous interaction analysis, the replication stage of the present study provides the first evidence of replication for this signal in independent cohorts. In the present analysis, there was no evidence that the results differed by smoking status.

SNPs rs6088813 in *UQCC1* and rs3849969 in *SEC24C* were identified as eQTLs for multiple genes. Whilst our eQTL analysis did not include formal tests of colocalisation, a SNP in complete linkage disequilibrium with rs3849969 (rs3812637, $r^2=1$) was associated with expression of *PLAU* in blood. The plasminogen activator, urokinase (PLAU) plays a role in fibrinolysis and immunity, and with its receptor (PLAUR) is involved in degradation of the extra cellular matrix, cell migration, cell adhesion and cell proliferation[31]. A study of preterm infants with respiratory distress syndrome, a condition characterised by intra-alveolar fibrin deposition, found *PLAU* and its inhibitor *SERPINE1* to be expressed in the alveolar epithelium, and an increased ratio of SERPINE1 to PLAU was associated with severity of disease[32]. Studies in mice have also shown that increased expression of *Plau* may be protective against lung injury, by reducing fibrosis[33]. *PLAU* has also been found to be upregulated in lung epithelial cells subjected to cyclic strain[34] and in patients with COPD and lung cancer, PLAU was found to be expressed in alveolar macrophages and epithelial cells[31].

The final two signals were with common intergenic SNPs close to *LY86* (MAF=36·8%; $P=9{\cdot}74\times10^{-23}$) and *FGF10* (MAF=35·6%; $P=2{\cdot}22\times10^{-11}$). *LY86* (lymphocyte antigen 86) interacts with the Toll-like receptor signalling pathway, to form a heterodimer, when bound with RP105[35]. The sentinel SNP in the present analysis (rs1294421) has previously shown association with waist-hip ratio[36], whilst an intronic SNP within *LY86* (rs7440529, $r^2=0{\cdot}005$ with rs1294421) has been implicated in asthma in two studies of individuals of Han Chinese ancestry[37,38]. FGF10 is a member of the fibroblast growth factor family of proteins, and is involved in a range of biological processes, including embryonic development and morphogenesis, cell growth and repair, tumor growth and invasion. Specifically, the FGF10 signalling pathway is thought to play an criticial role in the development of the lung and in lung epithelial renewal[39]. A deficiency in *Fgf10* has been demonstrated to lead to a fatal disruption of branching morphogenesis during lung development in mice[40].

Our discovery analyses included individuals of both EA and AA. Two of the identified six novel signals showed inconsistent effects in the AA and EA individuals. For these SNPs, the associations in AA individuals were not statistically significant, and we report associations from the analysis restricted to EA individuals only. For the remaining four SNPs similar effects were observed in both the EA and AA individuals (Supplementary Figure 3). We also examined the effects of the novel SNPs in ever smokers and never smokers separately and found these to be broadly similar, with the exception of rs1448044 in *FGF10,* which in the discovery analysis showed significant association with FVC in ever smokers, whilst showing no association in never smokers (P=0·695). However, in our replication stage analyses, similar effects were seen in both ever and never smokers for this SNP, and the combined analysis of discovery and replication stages for this SNP, including both ever and never smokers, met the exome chip-wide significance level overall ($P=4{\cdot}22\times10^{-9}$). We also considered whether this signal could be driven by smoking behaviour in our discovery stage as our primary analyses in SpiroMeta did not adjust for smoking quantity. We undertook a look-up of this SNP in the publicly available results of a GWAS of several smoking behaviour traits[41]; there was only weak evidence that this SNP was associated with ever versus never smoking (P=0·039), and no evidence for association with amount smoked (cigarettes per day, P=0·10).

Through the use of the exome array, we aimed to identify associations with low frequency and rare functional variants, thereby potentially uncovering some of the missing heritability of lung function. However, whilst our discovery analyses identified single SNP associations with 23 low frequency variants (Supplementary Table 2), we did not replicate any of these findings. Eleven of these 23 SNPs we were unable to follow-up in our replication studies, due to them either being not genotyped, or monomorphic. Overall, limited statistical power is likely to explain our lack of convincing single variant associations with rare variants, in particular if those variants exhibit only modest effects[42]. We additionally investigated the joint effects of low frequency and rare variants within genes, on lung function trait, by employing SKAT and WST gene-based tests. These analyses identified associations with a number of genes that could not attributed to the effect of a single SNP. Replication of these gene-based signals proved difficult however, as again a number of SNPs included in the discovery stage of these analyses were monomorphic, or had not been not genotyped in the replication studies. This lead to a disparity in the gene unit being tested in our discovery and replication samples; hence interpretation of these results was not clear-cut. In the end, we were able to replicate only findings with common SNPs. This finding is in line with several other studies of complex traits and exome array data that have been unable to report robust associations with low frequency variants[43–45] and it is clear that future studies will require increasingly larger sample sizes in order to fully evaluate the effect of variants across the allele frequency spectrum. The identification of common SNPs remains important, however, as such findings have the potential to highlight drug targets[46], and these variants collectively could have utility in risk prediction.

In our replication analyses using UK Biobank, we applied adjustment for covariates including ancestry principal components, before undertaking inverse-normal transformations of the lung function phenotypes. Association analyses were then performed using these transformed phenotypes. It has recently been shown that such transformation has the potential to introduce correlations between principal components and phenotypes[47]; we undertook sensitivity analyses for the six reported SNPs by repeating the association analyses with phenotypes that had been transformed without prior adjustment, with covariate adjustment made as part of the SNP-trait association test. We found there to be some difference in P-values for some SNP-trait combinations; however, the six novel SNP associations we report all met the replication P-value threshold ($P<1{\cdot}47\times10^{-3}$) in the sensitivity analyses (Supplementary Figure 4). This issue may also be relevant to the gene-based tests; however no replicated novel gene-based associations were identified in this study. Future studies should avoid undertaking adjustment for principal components of ancestry prior to trait transformation, in order to avoid this potential bias.

This study has identified six common SNPs, independent to signals previously implicated in lung function. Additional interrogation of these loci could lead to greater understanding of lung function and lung disease, and could provide novel targets for therapeutic interventions.

## Methods

### Study design, cohorts and genotyping

The SpiroMeta analysis included 23,751 individuals of EA from 11 studies, and the CHARGE analysis comprised 36,998 EA individuals and a further 7,721 individuals of AA from 12 studies. Follow-up analyses were conducted in an independent sample of up to 111,556 individuals from UK Biobank (2015 interim release), the UK Household Longitudinal Study (UKHLS) and the Netherlands Epidemiology of Obesity (NEO) Study (Figure 1). All studies (excluding UK Biobank) were genotyped using either the Illumina Human Exome BeadChip v1 or the Illumina Infinium HumanCoreExome-12 v1·0 BeadChip. UK Biobank samples were genotyped using the Affymetrix Axiom UK BiLEVE or UK Biobank arrays.

### Statistical analyses

**Consortium level analyses:** Within the SpiroMeta Consortium, each study contributing to the discovery analyses calculated single-variant score statistics, along with covariance matrices describing correlations between variants, using RAREMETALWORKER[48] or rvtests[49]. For each trait, these summary statistics were generated separately in ever and never smokers. Traits were adjusted for sex, age, age$^2$ and height, and inverse normally transformed prior to association testing. For studies with unrelated individuals, SNP-trait associations were tested using linear models, with adjustments made for the first 10 ancestry principal components, whilst studies with related individuals utilised linear mixed models to account for familial relationships and underlying population structure.

Within the CHARGE Consortium, each study generated equivalent summary statistics using the R package SeqMeta[50]. For each trait, summary statistics were generated in ever and never smokers separately, and in all individuals combined. The untransformed traits were used for all analyses, adjusted for smoking status and pack-years, age, age$^2$, sex, height, height$^2$, centre/cohort. Models for FVC were additionally adjusted for weight. Linear regression models, with adjustment for principal components of ancestry were used for studies with unrelated individuals, and linear mixed models were used for family-based studies.

Within each consortium we used the score statistics and variance-covariance matrices generated by each study to construct both single variant and gene-based tests using either RAREMETAL[48] (SpiroMeta) or SeqMeta[50] (CHARGE). For single variant associations, score statistics were combined in fixed effects meta-analyses. Two gene-based tests were constructed: first, the Weighted Sum Test (WST) using Madsen Browning weightings[23], and secondly, the Sequence Kernel Association Test (SKAT)[24]. We performed the SKAT and WST tests using two subsets of SNPs: 1) including all SNPs with an overall consortium-wide MAF<5% that were annotated as splicing, stopgain, stoploss, or frameshift (loss of function [LOF] analysis), and 2) including

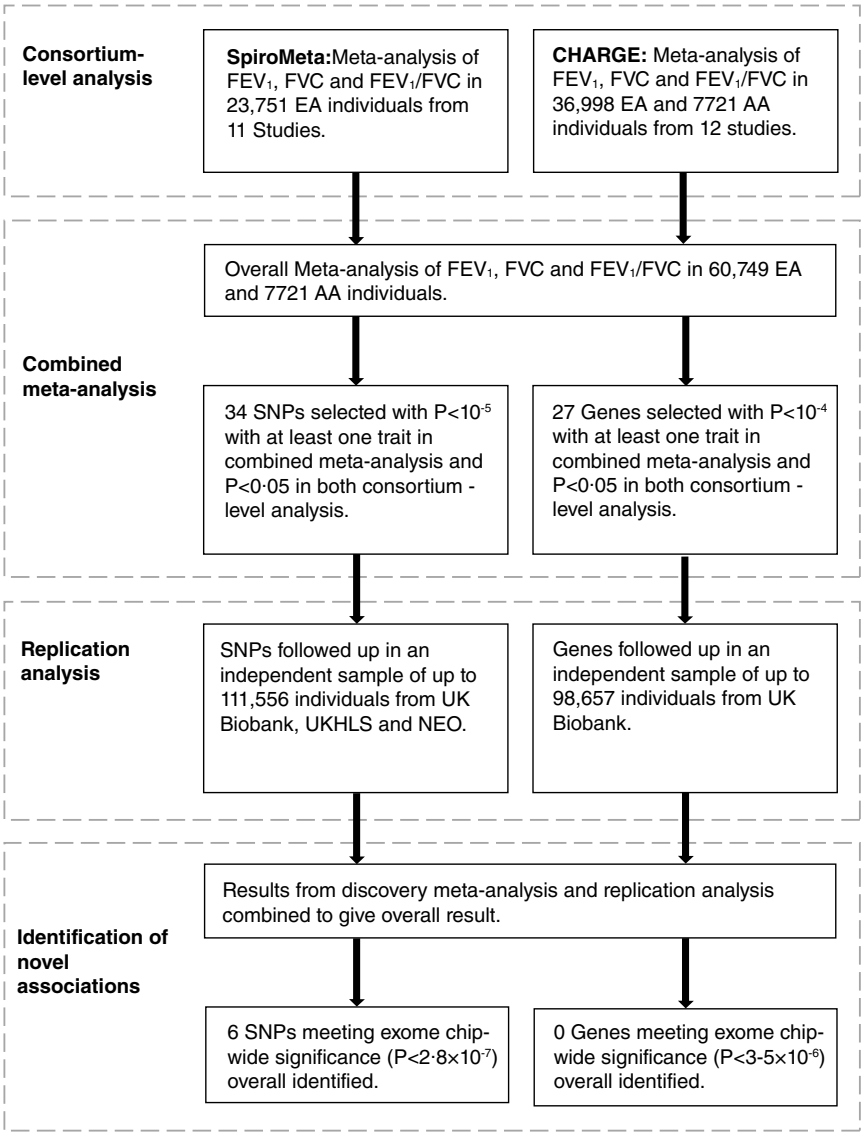


**Figure 1. Study design.**

all SNPs meeting the LOF analysis criteria in addition to all other nonsynonymous variants with consortium wide MAF<5% (exonic analysis). Variants were annotated to genes using dbNSFP v2·6[51] on the basis of the GRCh37/hg19 database.

For both single variant and gene-based associations, consortium-level results were generated for ever smokers and never smokers separately, and in all individuals combined. Within the CHARGE Consortium, results were combined separately for the EA and AA studies and also in a trans-ethnic analysis of both ancestries.

**Combined meta-analysis:** The single variant association results from the SpiroMeta and CHARGE consortia were combined as follows: The genomic inflation statistic (λ) was calculated for SNPs with consortium-wide MAF>1%; where λ had a value greater than one, genomic control adjustment was applied to the consortium level P-values. The consortium-level results were then combined using sample size weighted z-score meta-analysis. The λ was again calculated for the meta-analysis results and genomic control applied, as appropriate. λ values at the consortium and meta-analysis level are shown in Supplementary Table 13. Since we were interested in identifying low frequency and rare variants, we applied no MAF or minor allele count (MAC) filter. We identified SNPs of interest as those with an overall $P<10^{-5}$ and a consistent direction of effect and P<0·05 observed in both consortia. Rather than using a strict Bonferroni correction for defining the significance threshold, we adopted the more lenient $P<10^{-5}$ threshold in order to increase the power to detect variants with modest effect in our discovery analyses, whilst the requirement for consistency in results from the two consortia aimed to limit false positives. All SNPs meeting these thresholds were followed up in independent replication cohorts. Where we identified a SNP within 1Mb of a previously identified

lung function SNP, we deemed the SNP to represent an independent signal if it had $r^2<0{\cdot}2$ with the known SNP, and if it retained a $P<10^{-5}$, when conditional analyses were carried out with the known SNP, or a genotyped proxy, using data from the SpiroMeta Consortium, or UK Biobank. Our primary meta-analysis included all individuals; we additionally carried out analyses in smoking subgroups (ever and never smokers), and in the subgroup of individuals of European ancestry only.

For genes which contained at least 2 polymorphic SNPs in both consortia, we combined the results of the consortium level gene based tests using either z-score meta-analysis (for the WST analysis) or Fisher's Method for combining P-values (in the case of SKAT). We identified genes of interest as those with $P<0{\cdot}05$ observed in both consortia and an overall $P<10^{-4}$, thresholds again chosen to limit both false positive and false negative findings. As in the analyses of single variant associations, our primary meta-analyses included all individuals, with secondary analyses undertaken in smoking and ancestry specific subgroups.

**Replication analyses:** All SNP and gene-based associations were followed up for the trait with which they showed the most statistically significant association only. For associations identified through the smoking subgroup analyses, we followed up associations in the appropriate smoking strata; however, no ancestry stratified follow-up was undertaken as replication studies included only a sufficient number of individuals of European Ancestry.

Single variant associations in UK Biobank were tested in ever smokers and never smokers separately, and stratified by genotyping array (UK BiLEVE array or UK Biobank array) using the score test as implemented in SNPTEST v2·5b4[52]. Traits were adjusted for age, age$^2$, height, sex, ten principal components and pack-years (ever smokers only), and the adjusted traits were inverse normally transformed. Correlations between principal components and transformed phenotypes may be introduced where adjustment is made prior to transformation. In this analysis, we found any introduced correlations to have no impact on the conclusion of our replication analyses; however future studies should apply transformation of phenotypes prior to covariate adjustment, to avoid this issue. For UKHLS, analyses were undertaken analogously to the SpiroMeta discovery studies using RAREMETALWORKER, while for NEO, analyses were undertaken in the same way as was done in the CHARGE discovery studies using SeqMeta. The single variant results from all replication studies were combined using sample size weighted Z-score meta-analysis. Subsequently, we combined the results from the discovery and replication stage analyses and we report SNPs with overall exome-wide significance of $P<2{\cdot}8\times10^{-7}$ (Bonferroni corrected for the original 179,215 SNPs tested).

We followed up genes of interest ($P<10^{-4}$) using data from UK Biobank only. Summary statistics for UK Biobank were generated using RAREMETALWORKER, with gene-based tests then constructed using RAREMETAL. Finally, we combined the results from the discovery analysis with the replication results in an overall combined meta-analysis using either z-score meta-analysis (WST) or Fisher's Method (SKAT). We declared genes with overall $P<3{\cdot}5\times10^{-6}$ (Bonferroni corrected for 14,380 genes tested) in our combined meta-analysis to be statistically significant. For these statistically significant genes, we carried out additional analyses using the UK Biobank data in which we conditioned on the most significantly associated individual SNP within that gene, to determine whether this was a true gene-based signal, or whether the association could be ascribed to the single SNP (if the conditional $P<0{\cdot}01$, then association was deemed to not be driven by the single SNP).

### Characterization of findings

In order to gain further insight into the loci identified in our analyses of single variant associations, we assessed whether these regions were associated with gene expression levels in various tissues (FDR of 5%, or q-value<0·05), by querying a publically available blood eQTL database[53] and the GTEx project[54] for the sentinel SNPs, or any proxy ($r^2>0{\cdot}8$). We further assessed SNPs of interest (and proxies) within a lung eQTL resource based on non-tumour lung tissues of 1,111 individuals[55–57]. Descriptions of these resources and further details of the look-ups are provided in the Supplementary Methods. Moreover, all sentinel SNPs and proxies with $r^2>0.8$ were annotated using ENSEMBL's Variant Effect Predictor (VEP)[58]; potentially deleterious coding variants were identified as those annotated as 'deleterious' by SIFT[59] or 'probably damaging' or 'possibly damaging' by PolyPhen-2[60]. For all genes implicated through the expression data or functional annotation, we searched for evidence of protein expression in the respiratory system by querying the Human Protein Atlas[61].

## Data availability

Summary level results for all analyses are available on OSF: https://doi.org/10.17605/OSF.IO/NSDPJ[62]



This research has been conducted using the UK Biobank Resource. The genetic and phenotypic UK Biobank data are available upon application to the UK Biobank (https://www.ukbiobank.ac.uk/) to all registered health researchers. These data are from Understanding Society: The UK Household Longitudinal Study (UKHLS), which is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council. The data were collected by NatCen and the genome wide scan data were analysed by the Wellcome Trust Sanger Institute. Information on how to access the data can be found on the Understanding Society website https://www.understandingsociety.ac.uk/.

### Author contributions

Ordered alphabetically: ABW, AGE, AL, BMP, BS, CH, CP, DOMK, DPS, EZ, GGB, HS, IPH, JBJ, JK, KMB, LL, MAI, MAP, MDT, MK, NG, NMPH, OP, OTR, RdM, RGB, SBK, SG, SJL, SSR, TA,

TBH, TH, TL, TR, TS, UG contributed to study concept and designs. AC, AJ, A.Manichaikul, BHS, BMP, BS, CP, DJP, DPS, EI, GGB, GTOC, IJD, JBJ, JGW, JK, JMS, KS, LAL, LL, LL, MAP, MI, MK, NG, NMPH, OP, OTR, PAC, RdM, RGB, RR, SBK, SE, SEH, SG, SK, SK, TA, TBH, TDP, TL, TNB, TR, UG, WT, WT contributed to phenotype data acquisition and quality control. AGE, AJ, AK, AK, ALT, ALT, A.Manichaikul, APM, AT, BMP, BP, CH, DOMK, EI, GD, HV, IJD, JAB, JCM, JGW, JL, KDT, KEN, KL, L-PL, LAL, LL, MAP, MI, MLG, NMPH, OP, RGB, RLG, RR, SBK, SE, SEH, SRH, SSR, SW, TBH, TDP, TH, TL, YL contributed to genotype data acquisition and quality control. DDS, KH, WT, YB contribute to eQTL data acquisition and quality control. ABW, ACM, AK, AK, ALT, A.Mahajan, A.Manichaikul, APM, AT, BP, BQ, CH, CMS, EA, HV, IPH, JAB, JCL, JD, JEH, JL, JM, JMJ, KL, L-PL, LL, LVW, MDT, MI, MO, NF, NMPH, OP, PAC, RLG, SE, SEH, SJL, SW, TDP, TH, TMB, VEJ, WG, WT, YL contributed to data analysis. All authors contributed to writing and/or critical review of the manuscript. The 'Understanding Society Scientific Group' include the following: Understanding Society Scientific Group: Michaela Benzeval, Jonathan Burton, Nicholas Buck, Annette Jäckle, Meena Kumari, Heather Laurie, Peter Lynn, Stephen Pudney, Birgitta Rabe, Shamit Saggar, Noah Uhrig, Dieter Wolke.

## Competing interests

No competing interests were disclosed.

## Grant information

This article presents independent research funded partially by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

MDT has been supported by Medical Research Council (MRC) fellowships G0501942 and G0902313. MDT and LVW are supported by the MRC (MR/N011317/1). IPH is supported by the MRC (G1000861). ALW and SJL are supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (ZIA ES 043012). We acknowledge use of phenotype and genotype data from the British 1958 Birth Cohort DNA collection, funded by the MRC (G0000934) and the Wellcome Trust (068545). APM was a Wellcome Trust Senior Fellow in Basic Biomedical Science (098017) and was also supported by Wellcome Trust grant 064890. EI is supported by the Swedish Research Council (2012-1397), Knut och Alice Wallenberg Foundation (2013.0126) and the Swedish Heart-Lung Foundation (20140422). JK is supported by Academy of Finland Center of Excellence in Complex Disease Genetics (213506, 129680) and Academy of Finland (265240, 263278). The Finnish Twin Cohort is supported by the Welcome Trust Sanger Institute, UK. The Lothian Birth Cohort is supported by Age UK (The Disconnected Mind Project), the MRC (MR/K026992/1) and The Royal Society of Edinburgh. ÅJ is supported by the Swedish Society for Medical Research, The Kjell och Märta Beijers Foundation, The Marcus Borgström Foundation, The Åke Wiberg foundation and The Vleugels Foundation. UG is supported by Swedish Medical Research Council (K2007-66X-20270-01-3, 2011-2354) and European Commission FP6 (LSHG-CT-2006-01947). SHIP is part of the Community Medicine Research net of the University of Greifswald, Germany, which is funded by the Federal Ministry of Education and Research, the Ministry of Cultural Affairs, as well as the Social Ministry of the Federal State of Mecklenburg-West Pomerania, and the network 'Greifswald Approach to Individualized Medicine' funded by the Federal Ministry of Education and Research, and the German Asthma and COPD Network (01ZZ9603, 01ZZ0103, 01ZZ0403, 03IS2061A, BMBF 01GI0883). ExomeChip data have been supported by the Federal Ministry of Education and Research (03Z1CN22) and the Federal State of Mecklenburg-West Pomerania. The University of Greifswald is a member of the Caché Campus program of the InterSystems GmbH. UKHLS is supported by the Wellcome Trust (098051) and Economic and Social Research Council (ES/H029745/1). Y.B. holds a Canada Research Chair in Genomics of Heart and Lung Diseases. Lies Lahousse is a Postdoctoral Fellow of the Research Foundation - Flanders (G035014N). The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, the Netherlands Organization for Scientific Research (NOW), the Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. Genotyping in the Rotterdam study was supported by NOW (175.010.2005.011, 911-03-305 012), RIDE2 (014-93-015) and Netherlands Genomics Initiative/Netherlands Consortium for Healthy Aging (050-060-810). MESA/MESA SHARe is supported by the US Department of Health and Human Services (HHS) (HHSN268201500003I), NIH/National Heart, Lung and Blood Institute (NHLBI; N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169) and NIH/National Center for Advancing Translation Studies (NCATS) (UL1-TR-000040, UL1-TR-001079, UL1-TR-001881, DK063491). MESA SHARe is funded by NIH/NHLBI contract N02-HL-64278, MESA Air is funded by US Environmental Protection Agency (RD831697) and MESA Spirometry funded by NIH/NHLBI (R01-HL077612). SSR and BMP are supported by NIH/NHLBI grant rare variants and NHLBI traits in deeply phenotyped cohorts (R01-HL120393). The CHS research was supported by NHLBI (contracts: HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, HHSN268200960009C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; grants: U01HL080295, R01HL068986, R01HL087652, R01HL105756, R01HL103612, R01HL120393, R01HL130114), with additional contribution from the National Institute of Neurological Disorders and Stroke. Additional support was provided through R01AG023629 and R01HL085251 from the National Institute on Aging (NIA). The provision of genotyping data was supported in part by the NCATS, CTSI (UL1TR001881), and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DK063491) to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the

authors and does not necessarily represent the official views of the National Institutes of Health. ARIC study is carried out as a collaborative study supported by the NHLBI (contracts: HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, HHSN268201100012C). Funding support for "Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium" was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (5RC2HL102419). DOMK received funding from the Dutch Science Organisation (ZonMW-VENI Grant 916.14.023). The genotyping in the NEO study was supported by the Centre National de Génotypage (Paris, France), headed by Jean-François Deleuze. The NEO study is supported by the participating Departments, the Division and the Board of Directors of the Leiden University Medical Center, and by the Leiden University, Research Profile Area Vascular and Regenerative Medicine. SAPALDIA was supported by the Swiss National Science Foundation (33CS30-148470/1, 33CSCO-134276/1, 33CSCO-108796, 324730_135673, 3247BO-104283, 3247BO-104288, 3247BO-104284, 3247-065896, 3100-059302, 3200-052720, 3200-042532, 4026-028099, PMPDP3_129021/1, PMPDP3_141671/1), the Federal Office for the Environment, the Federal Office of Public Health, the Federal Office of Roads and Transport, the canton's government of Aargau, Basel-Stadt, Basel-Land, Geneva, Luzern, Ticino, Valais, and Zürich, the Swiss Lung League, the Canton's Lung League of Basel Stadt/Basel Landschaft, Geneva, Ticino, Valais, Graubünden and Zurich, Stiftung ehemals Bündner Heilstätten, SUVA, Freiwillige Akademische Gesellschaft, UBS Wealth Foundation, Talecris Biotherapeutics GmbH, Abbott Diagnostics, European Commission 018996 (GABRIEL), Wellcome Trust (084703). The Novo Nordisk Foundation Center for Basic Metabolic Research is an independent Research Center at the University of Copenhagen partially funded by an unrestricted donation from the Novo Nordisk Foundation (www.metabol.ku.dk). Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006]. Genotyping of the GS:SFHS samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, University of Edinburgh, Scotland, and was funded by the MRC. The Croatia KORCULA study was supported by the Ministry of Science, Education and Sport in the Republic of Croatia (108-1080315-0302). JD, JCL, WG and GTOC are supported by NIH/NHLBI (HHSN268201500001I). Genotyping, quality control and calling of the Illumina HumanExome BeadChip in the Framingham Heart Study was supported by funding from the National Heart, Lung and Blood Institute Division of Intramural Research (Daniel Levy and Christopher J. O'Donnell, Principle Investigators). The AGES study is supported by the NIH (N01-AG012100), the Iceland Parliament (Alþingi) and the Icelandic Heart Association. HABC was supported by NIA (contracts: N01AG62101, N01AG62103, N01AG62106; grant: R01-AG028050), and NINR (grant R01- NR012459), and was supported in part by the Intramural Research Program of the NIA. The HABC genome-wide association study was funded by NIA (1R01AG032098- 01A1) and genotyping services were provided by the Center for Inherited Disease Research (CIDR). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University (HHSN268200782096C). We thank the Jackson Heart Study (JHS) participants and staff for their contributions to this work. The JHS is supported by contracts HHSN268201300046C, HHSN268201300047C, HHSN268201300048C, HHSN268201300049C, HHSN268201300050C from the National Heart, Lung, and Blood Institute and the National Institute on Minority Health and Health Disparities. JGW is supported by U54GM115428 from the National Institute of General Medical Sciences.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

## Acknowledgements

The authors would like to thank the staff at the Quebec Respiratory Health Network Tissue Bank for their valuable assistance with the lung eQTL dataset at Laval University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. A full list of principal CHS investigators and institutions can be found at https://CHS-NHLBI.org. The authors thank the staff and participants of the ARIC study for their important contributions. The authors of the NEO study thank all individuals who participated in the Netherlands Epidemiology in Obesity study, all participating general practitioners for inviting eligible participants and all research nurses for collection of the data. We thank the NEO study group, Pat van Beelen, Petra Noordijk and Ingeborg de Jonge for the coordination, lab and data management of the NEO study. SAPALDIA could not have been done without the help of the study participants, technical and administrative support and the medical teams and field workers at the local study sites. Local fieldworkers: Aarau: M Broglie, M Bünter, D Gashi; Basel: R Armbruster, T Damm, U Egermann, M Gut, L Maier, A Vögelin, L Walter; Davos: D Jud, N Lutz; Geneva: M Ares, M Bennour, B Galobardes, E Namer; Lugano: B Baumberger, S Boccia Soldati, E Gehrig-Van Essen, S Ronchetto; Montana: C Bonvin, C Burrus; Payerne: S Blanc, AV Ebinger, ML Fragnière, J Jordan; Wald: R Gimmi, N Kourkoulos, U Schafroth**.** Administrative staff: N Bauer, D Baehler, C Gabriel, R Gutknecht. SAPALDIA Team: *Study directorate:* NM Probst Hensch, T Rochat, N Künzli, C Schindler, JM Gaspoz; *Scientific team:* JC Barthélémy, W Berger, R Bettschart, A Bircher, G Bolognini, O Brändli, C Brombach, M Brutsche, L Burdet, M Frey, U Frey, MW Gerbase, D Gold, E de Groot, W Karrer, R Keller, B Knöpfli, B Martin, D Miedinger, U Neu, L Nicod, M Pons, F Roche, T Rothe, E Russi, P Schmid-Grendelmeyer, A Schmidt-Trucksäss, A Turk, J Schwartz, D. Stolz, P Straehl, JM Tschopp, A von Eckardstein, E Zemp Stutz; *Scientific team at coordinating centers:* M Adam, E Boes, PO Bridevaux, D Carballo, E Corradi, I Curjuric, J Dratva, A Di Pasquale, L Grize, D Keidel, S Kriemler, A Kumar, M Imboden, N Maire, A Mehta, F Meier, H Phuleria, E Schaffner, GA Thun, A Ineichen, M Ragettli, M Ritter, T Schikowski, G Stern, M Tarantino, M Tsai, M Wanner.

This research used the ALICE and SPECTRE High Performance Computing Facilities at the University of Leicester.

## Supplementary material

Supplementary Information: File includes Supplementary Note, Supplementary Methods, Supplementary Figures and Supplementary Tables, as detailed below.

Click here to access the data.

**Supplementary Note** includes individual study descriptions.

**Supplementary Methods** includes details of study level quality control procedures and eQTL analyses.

**Supplementary Figures:**

Supplementary Figure 1 - Quantile-quantile (QQ) and Manhattan plots for consortium-wide analyses, and the combined meta-analysis.

Supplementary Figure 2 - Region Plots for novel loci.

Supplementary Figure 3 - Forest Plots for novel loci.

Supplementary Figure 4 - Trait Transformation Sensitivity Analysis

**Supplementary Tables:**

Supplementary Table 1 - Details of study specific genotyping platform, genotype calling procedure and software.

Supplementary Table 2 - Association results for all SNPs identified in single variant association discovery analyses ($P<10^{-4}$).

Supplementary Table 3 - Association results for SNPs identified in single variant association discovery analyses ($P<10^{-4}$), located in known lung function regions.

Supplementary Table 4 - Single variant association result for the seven novel signals, in smoking and ancestry subgroups.

Supplementary Table 5 - Single variant association result for rs1448044 and FVC in ever smokers and never smokers separately, and in all samples combined.

Supplementary Table 6 - Association results for all genes identified in discovery SKAT analyses (meta-analysis $P<10^{-4}$).

Supplementary Table 7 - Association results for all genes identified in discovery Weighted sum test (WST) test analyses ($P<10^{-4}$).

Supplementary Table 8 - Evidence for the role of novel variants identified in single variant association analyses as eQTLs.

Supplementary Table 9 - SIFT/Polyphen predictions for sentinel SNPs and proxies (r2>0.8).

Supplementary Table 10 - Protein and RNA expression results all implicated genes from the single variant association analyses.

Supplementary Table 11 - Look-up of association results for SNPs at 7 of the 12 loci which showed allele frequency differences between individuals from different regions in the UK.

Supplementary Table 12 - All traits results for the seven novel lung function loci.

Supplementary Table 13 - Genomic Inflation Factors: consortium and meta-analysis level.

## References


1. Rabe KF, Hurd S, Anzueto A, *et al.*: **Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease: GOLD executive summary.** *Am J Respir Crit Care Med.* 2007; **176**(6): 532–55.
PubMed Abstract | Publisher Full Text
2. Palmer LJ, Knuiman MW, Divitini ML, *et al.*: **Familial aggregation and heritability of adult lung function: results from the Busselton Health Study.** *Eur Respir J.* 2001; **17**(4): 696–702.
PubMed Abstract | Publisher Full Text
3. Wilk JB, DeStefano AL, Joost O, *et al.*: **Linkage and association with pulmonary function measures on chromosome 6q27 in the Framingham Heart Study.** *Hum Mol Genet.* 2003; **12**(21): 2745–51.
PubMed Abstract | Publisher Full Text
4. Klimentidis YC, Vazquez AI, de Los Campos G, *et al.*: **Heritability of pulmonary function estimated from pedigree and whole-genome markers.** *Front Genet.* 2013; **4**: 174.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Wilk JB, Djousse L, Arnett DK, *et al.*: **Evidence for major genes influencing pulmonary function in the NHLBI Family Heart Study.** *Genet Epidemiol.* 2000; **19**(1): 81–94.
PubMed Abstract | Publisher Full Text
6. Wilk JB, Chen TH, Gottlieb DJ, *et al.*: **A Genome-Wide Association Study of Pulmonary Function Measures in the Framingham Heart Study.** *PLoS Genet.* 2009; **5**(3): e1000429.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Repapi E, Sayers I, Wain LV, *et al.*: **Genome-wide association study identifies five loci associated with lung function.** *Nat Genet.* 2010; **42**(1): 36–44.
PubMed Abstract | Publisher Full Text | Free Full Text
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**(11): 1082–90.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Hancock DB, Eijgelsheim M, Wilk JB, *et al.*: **Meta-analyses of genome-wide**

**association studies identify multiple loci associated with pulmonary function.** *Nat Genet.* 2010; **42**(1): 45–52.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Loth DW, Soler Artigas M, Gharib SA, *et al.*: **Genome-wide association analysis identifies six new loci associated with forced vital capacity.** *Nat Genet.* 2014; **46**(7): 669–77.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Wain LV, Shrine N, Miller S, *et al.*: **Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank.** *Lancet Respir Med.* 2015; **3**(10): 769–81.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Soler Artigas M, Wain LV, Miller S, *et al.*: **Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation.** *Nat Commun.* 2015; **6**: 8658.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Wain LV, Shrine N, Artigas MS, *et al.*: **Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets.** *Nat Genet.* 2017; **49**(3): 416–425.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Wyss AB, Sofer T, Lee MK, *et al.*: **Multiethnic meta-analysis identifies ancestry-specific and cross-ancestry loci for pulmonary function.** *Nat Commun.* In Press.

15. Pillai SG, Ge D, Zhu G, *et al.*: **A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci.** *PLoS Genet.* 2009; **5**(3): e1000421.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Cho MH, Boutaoui N, Klanderman BJ, *et al.*: **Variants in *FAM13A* are associated with chronic obstructive pulmonary disease.** *Nat Genet.* 2010; **42**(3): 200–2.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Cho MH, McDonald ML, Zhou X, *et al.*: **Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis.** *Lancet Respir Med.* 2014; **2**(3): 214–25.
PubMed Abstract | Publisher Full Text | Free Full Text

18. Hobbs BD, de Jong K, Lamontagne M, *et al.*: **Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis.** *Nat Genet.* 2017; **49**(3): 426–32.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Hobbs BD, Parker MM, Chen H, *et al.*: **Exome Array Analysis Identifies A Common Variant in *IL27* Associated with Chronic Obstructive Pulmonary Disease.** *Am J Respir Crit Care Med.* 2016; **194**(1): 48–57.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Abecasis GR: **Exome Chip Design Wiki.** 2013; Accessed August 30, 2013.
Reference Source

21. Hancock DB, Soler Artigas M, 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**(12): e1003098.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Campbell CD, Ogburn EL, Lunetta KL, *et al.*: **Demonstrating stratification in a European American population.** *Nat Genet.* 2005; **37**(8): 868–72.
PubMed Abstract | Publisher Full Text

23. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet.* 2009; **5**(2): e1000384.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Wu M, Lee S, Cai T, *et al.*: **Rare-variant association testing for sequencing data with the sequence kernel association test.** *Am J Hum Genet.* 2011; **89**(1): 82–93.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature.* 2007; **447**(7145): 661–78.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Heath SC, Gut IG, Brennan P, *et al.*: **Investigation of the fine structure of European populations with applications to disease association studies.** *Eur J Hum Genet.* 2008; **16**(12): 1413–29.
PubMed Abstract | Publisher Full Text

27. Grabiec AM, Hussell T: **The role of airway macrophages in apoptotic cell clearance following acute and chronic lung inflammation.** *Semin Immunopathol.* Springer; 2016; **38**(4): 409–23.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Ye S, Lowther S, Stambas J: **Inhibition of reactive oxygen species production ameliorates inflammation induced by influenza A viruses via upregulation of SOCS1 and SOCS3.** *J Virol.* 2015; **89**(5): 2672–83.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Gudmundsson J, Sulem P, Steinthorsdottir V, *et al.*: **Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes.** *Nat Genet.* 2007; **39**(8): 977–83.
PubMed Abstract | Publisher Full Text

30. Soranzo N, Rivadeneira F, Chinappen-Horsley U, *et al.*: **Meta-analysis of genome-wide scans for human adult stature identifies novel Loci and associations with measures of skeletal frame size.** *PLoS Genet.* 2009; **5**(4): e1000445.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Wang IM, Stepaniants S, Boie Y, *et al.*: **Gene expression profiling in patients with chronic obstructive pulmonary disease and lung cancer.** *Am J Respir Crit Care Med.* 2008; **177**(4): 402–11.
PubMed Abstract | Publisher Full Text

32. Cederqvist K, Sirén V, Petäjä J, *et al.*: **High concentrations of plasminogen activator inhibitor-1 in lungs of preterm infants with respiratory distress syndrome.** *Pediatrics.* 2006; **117**(4): 1226–34.
PubMed Abstract | Publisher Full Text

33. Sisson TH, Hanson KE, Subbotina N, *et al.*: **Inducible lung-specific urokinase expression reduces fibrosis and mortality after lung injury in mice.** *Am J Physiol Lung Cell Mol Physiol.* 2002; **283**(5): L1023–32.
PubMed Abstract | Publisher Full Text

34. Weber B, Bader N, Lehnich H, *et al.*: **Microarray-based gene expression profiling suggests adaptation of lung epithelial cells subjected to chronic cyclic strain.** *Cell Physiol Biochem.* 2014; **33**(5): 1452–66.
PubMed Abstract | Publisher Full Text

35. Kimoto M, Nagasawa K, Miyake K: **Role of TLR4/MD-2 and RP105/MD-1 in innate recognition of lipopolysaccharide.** *Scand J Infect Dis.* 2003; **35**(9): 568–72.
PubMed Abstract | Publisher Full Text

36. Heid IM, Jackson AU, Randall JC, *et al.*: **Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution.** *Nat Genet.* 2010; **42**(11): 949–60.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Tan JY, Luo YL, Huang X, *et al.*: **[Association of single nucleotide polymorphisms of MD-1 gene with asthma in adults of Han Nationality in Southern China].** *Zhonghua Jie He He Hu Xi Za Zhi.* 2011; **34**(2): 104–8.
PubMed Abstract

38. Lee SW, Wang JY, Hsieh YC, *et al.*: **Association of single nucleotide polymorphisms of MD-1 gene with pediatric and adult asthma in the Taiwanese population.** *J Microbiol Immunol Infect.* 2008; **41**(6): 445–9.
PubMed Abstract

39. Klar J, Blomstrand P, Brunmark C, *et al.*: **Fibroblast growth factor 10 haploinsufficiency causes chronic obstructive pulmonary disease.** *J Med Genet.* 2011; **48**(10): 705–9.
PubMed Abstract | Publisher Full Text

40. Sekine K, Ohuchi H, Fujiwara M, *et al.*: **Fgf10 is essential for limb and lung formation.** *Nat Genet.* 1999; **21**(1): 138–41.
PubMed Abstract | Publisher Full Text

41. Tobacco and Genetics Consortium: **Genome-wide meta-analyses identify multiple loci associated with smoking behavior.** *Nat Genet.* 2010; **42**(5): 441–7.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Manolio TA, Collins FS, Cox NJ, *et al.*: **Finding the missing heritability of complex diseases.** *Nature.* 2009; **461**(7265): 747–53.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Zuo X, Sun L, Yin X, *et al.*: **Whole-exome SNP array identifies 15 new susceptibility loci for psoriasis.** *Nat Commun.* 2015; **6**: 6793.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Holmen OL, Zhang H, Zhou W, *et al.*: **No large-effect low-frequency coding variation found for myocardial infarction.** *Hum Mol Genet.* 2014; **23**(17): 4721–8.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Tajuddin SM, Schick UM, Eicher JD, *et al.*: **Large-scale exome-wide association analysis identifies loci for white blood cell traits and pleiotropy with immune-mediated diseases.** *Am J Hum Genet.* 2016; **99**(1): 22–39.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Nelson MR, Tipney H, Painter JL, *et al.*: **The support of human genetic evidence for approved drug indications.** *Nat Genet.* 2015; **47**(8): 856–60.
PubMed Abstract | Publisher Full Text

47. Pain O, Dudbridge F, Ronald A: **Are your covariates under control? How normalization can re-introduce covariate effects.** *Eur J Hum Genet.* 2018.
PubMed Abstract | Publisher Full Text

48. Liu DJ, Peloso GM, Zhan X, *et al.*: **Meta-analysis of gene-level tests for rare variant association.** *Nat Genet.* 2014; **46**(2): 200–4.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Zhan X, Hu Y, Li B, *et al.*: **RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data.** *Bioinformatics.* 2016; **32**(9): 1423–6.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Lumley T, Brody J, Dupus J, *et al.*: **Meta-analysis of a rare-variant association test.** 2012.
Reference Source

51. Liu X, Jian X, Boerwinkle E: **dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations.** *Hum Mutat.* 2013; **34**(9): E2393–402.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Marchini J, Howie B, Myers S, *et al.*: **A new multipoint method for genome-wide association studies by imputation of genotypes.** *Nat Genet.* 2007; **39**(7): 906–13.
PubMed Abstract | Publisher Full Text

53. Westra HJ, Peters MJ, Esko T, *et al.*: **Systematic identification of *trans* eQTLs as putative drivers of known disease associations.** *Nat Genet.* 2013; **45**(10):

1238–43.
PubMed Abstract | Publisher Full Text | Free Full Text

54. GTEx Consortium: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet.* 2013; **45**(6): 580–5.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Hao K, Bossé Y, Nickle DC, *et al.*: **Lung eQTLs to help reveal the molecular underpinnings of asthma.** *PLoS Genet.* 2012; **8**(11): e1003029.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Lamontagne M, Couture C, Postma DS, *et al.*: **Refining susceptibility loci of chronic obstructive pulmonary disease with lung eqtls.** *PLoS One.* 2013; **8**(7): e70220.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Obeidat M, Miller S, Probert K, *et al.*: **GSTCD and INTS12 regulation and expression in the human lung.** *PLoS One.* 2013; **8**(9): e74630.
PubMed Abstract | Publisher Full Text | Free Full Text

58. McLaren W, Gil L, Hunt SE, *et al.*: **The Ensembl Variant Effect Predictor.** *Genome Biol.* 2016; **17**(1): 122.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc.* 2009; **4**(7): 1073–81.
PubMed Abstract | Publisher Full Text

60. Adzhubei IA, Schmidt S, Peshkin L, *et al.*: **A method and server for predicting damaging missense mutations.** *Nat Methods.* 2010; **7**(4): 248–9.
PubMed Abstract | Publisher Full Text | Free Full Text

61. Uhlen M, Oksvold P, Fagerberg L, *et al.*: **Towards a knowledge-based Human Protein Atlas.** *Nat Biotechnol.* 2010; **28**(12): 1248–50.
PubMed Abstract | Publisher Full Text

62. Jackson V: **SpiroMeta-CHARGE exome chip meta-analysis sumary results.** 2017.
http://www.doi.org/10.17605/OSF.IO/NSDPJ

# Open Peer Review

## Current Referee Status: ✔ ✔

**Version 2**

Referee Report 03 July 2018

**doi:**10.21956/wellcomeopenres.15514.r33384

✔ **Robin Beaumont** , **Rachel M. Freathy**
Institute of Biomedical and Clinical Science, University of Exeter, Exeter, UK

Thank you to the authors for responding to and addressing our comments. I have one further comment on the replication analysis using UK Biobank data. The sensitivity analyses which the authors carried out showed that adjusting for covariates prior to inverse-normalization does affect the results. While this does not affect the main conclusions drawn, it may affect the results of the gene-based tests, and in addition, other investigators using the methods as a guide may draw inappropriate conclusions if adjusting for principal components prior to inverse-normalizing their phenotype. Ideally, the UK Biobank analysis should be redone with the appropriate phenotype transformation, and the methods and results sections updated accordingly. However, if the authors consider that such a revision would be too extensive, given that the conclusions do not change, it would at least be helpful to note the issue as a limitation in the discussion and make it clear in the methods that adjusting for covariates (such as principal components) should be done after inverse-normalising the phenotype – so it can be used appropriately by others.

***Competing Interests:*** No competing interests were disclosed.

**We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Author Response 17 Jul 2018

**Victoria Jackson**, University of Leicester, UK

Thank you for you approval of our article, and your additional comment. As suggested, we have added a further limitation to the discussion of the paper outlining the issue regarding the trait transformation. This has also been noted in the methods. We have also included the results of the sensitivity analyses in the supplement (Supplementary Figure 4).

***Competing Interests:*** No competing interests were disclosed.

**Version 1**

Referee Report 04 April 2018

**doi:**10.21956/wellcomeopenres.13627.r30984

**Lisa Strug** [1], **Naim Panjwani** [2]
[1] Research Institute, Hospital for Sick Children, Toronto, ON, Canada
[2] The Hospital for Sick Children, Toronto, ON, Canada

The authors have performed a large genome-wide association study in subjects of European (36,998 in the discovery set and 111,556 in the replication set) and African (7,721 in the discovery set) ancestries for various lung function measures: FEV1, FVC and FEV1/FVC ratio. Both common and rare variant analyses are performed, and the effect of smoking on the associations is also assessed. The discovery set consisted of CHARGE and SpiroMeta consortia meta analysis using the Human Exome array, while the replication set consisted of genotypes on the HumanCoreExome array and the UK Biobank's custom arrays. A total of 7 novel regions were identified by the authors that met the overall (discovery+replication) Bonferroni-adjusted P-value of 2.8x10^-7 after adjustment for various covariates such as age, sex, height, and ancestry using principal components. All identified novel SNPs are of common frequency, and two of the SNPs are in high LD with missense variants predicted to be damaging.

Some areas for improvement:

- Two rare variant tests were chosen and applied to the data as opposed to choosing a combined test (e.g. Derkach et al 2013 Genetic Epidemiology). A combined test would be more powerful.
- The authors should explain why there was an inverse normalization of the traits in SpiroMeta but not in CHARGE, and provide some sensitivity analysis.
- There appear to be very large differences in Effect Allele Frequencies between the discovery and replication samples. Do the authors have an explanation for this? This might point to local ancestry differences that could be relevant, and should be further investigated.
- The eQTL analysis could formally investigate colocalization as opposed to cross-referencing individual associated SNPs with public repositories, and there are several different methods that achieve this goal: e.g. COLOC, eCAVIAR, Sherlock, RTC or EnLoc.
- In the replication analyses section, it is stated that "Traits were adjusted for age, age^2, height, sex, ten principal components and pack-years (ever smokers only), and *inverse normally transformed.*" For clarity, the authors should be specific about whether the trait (FEV1, FVC, or FEV1/FVC) was inverse normalized first and age, age^2, sex, 10 PCs were then added as covariates in the genetic association model
- In the methods section for the rare variant testing Skat appears to be incorrectly referred to as a Fisher's combined method.
- The authors should provide the justification for their various significance criteria used in each of the analyses.
- The authors should list the MAF alongside the p-values reported in the text for clarity for the single variant analysis results

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**
Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**
Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**
Yes

**Are all the source data underlying the results available to ensure full reproducibility?**
Yes

**Are the conclusions drawn adequately supported by the results?**
Yes

***Competing Interests:*** No competing interests were disclosed.

**We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reader Comment 12 Jun 2018

**Victoria Jackson**, University of Leicester, UK

Thank you the second set of reviewers for your helpful comments. Again, we have addressed specific points below, and and made appropriate amendments to the manuscript.

*1. Two rare variant tests were chosen and applied to the data as opposed to choosing a combined test (e.g. Derkach et al 2013 Genetic Epidemiology). A combined test would be more powerful.*

We agree, a combined test would have been the preferred choice for gene-based association testing. However, in this instance, the gene-based tests were chosen due to practical reasons, as SKAT and WST, the two tests utilised, were both implemented by the meta-analysis software used by the two contributing consortia (RAREMETAL and seqMeta). Since this was a meta-analysis, and only summary statistics were available for each study, the gene-based tests we were able to utilise were restricted to those implemented by these two software packages at the time of the meta-analyses. For example, the suggested method by Derkach et al. requires permutation to calculate P-values with adequately controlled type 1 errors, which would not have been possible with the summary statistics available.

*2. The authors should explain why there was an inverse normalization of the traits in SpiroMeta but not in CHARGE, and provide some sensitivity analysis.*

As mentioned in response to the other reviewers' comments, we agree that using the raw trait in CHARGE and the transformed trait in SpiroMeta was not optimal; by the time we had made the decision to combine the results from the two consortia, all studies had already completed analyses, and reanalysis across the many cohorts would not have been feasible.

*3. There appear to be very large differences in Effect Allele Frequencies between the discovery and replication samples. Do the authors have an explanation for this? This might point to local ancestry differences that could be relevant, and should be further investigated.*

Thank you for highlighting this. There was an error with the effect allele frequencies for the replication samples in Supplementary Table 2; these have now been amended, and the allele frequencies are more consistent in the discovery and replication samples. Where there are still some differences between the discovery and replication allele frequencies, these are where the discovery meta-analysis included individuals of both European and African ancestry, whereas the replication dataset included individuals of European ancestry only.

*4. The eQTL analysis could formally investigate colocalization as opposed to cross-referencing individual associated SNPs with public repositories, and there are several different methods that achieve this goal: e.g. COLOC, eCAVIAR, Sherlock, RTC or EnLoc.*

Tests of colocalisation are more usually undertaken in dense genome-wide data, whereas the (often rare) putative causal variants included on the exome array in our study were relatively sparsely distributed. Furthermore, we did not have access to the lung eQTL data required to undertake a tests of colocalisation. We now acknowledge that the eQTL analysis did not include formal tests of colocalisation in the discussion, and in the example we highlight the variants are in complete LD.

*5. In the replication analyses section, it is stated that "Traits were adjusted for age, age$^2$, height, sex, ten principal components and pack-years (ever smokers only), and inverse normally transformed." For clarity, the authors should be specific about whether the trait ($FEV_1$, FVC, or $FEV_1$/FVC) was inverse normalized first and age, age$^2$, sex, 10 PCs were then added as covariates in the genetic association model.*

We have clarified in the methods for the replication analysis that "Traits were adjusted for age, age $^2$, height, sex, ten principal components and pack-years (ever smokers only), and the adjusted traits were inverse normally transformed."

*6. In the methods section for the rare variant testing Skat appears to be incorrectly referred to as a Fisher's combined method.*

Within each consortium we generated results for SKAT. Subsequently, we combined the SKAT results from the two consortia using Fisher's Method for combing P-values. We have clarified this in the text as "For genes which contained at least 2 polymorphic SNPs in both consortia, we combined the results of the consortium level gene based tests using either z-score meta-analysis (for the WST analysis) or Fisher's Method for combining P-values (in the case of SKAT).".

*7. The authors should provide the justification for their various significance criteria used in each of the analyses.*

Justification for the SNPs and genes taken forward to the replication stage has now been added to the methods:
"We identified SNPs of interest as those with an overall $P<10^{-5}$ and a consistent direction of effect and P<0·05 observed in both consortia. Rather than using a strict Bonferroni correction for defining

the significance threshold, we adopted the more lenient $P<10^{-5}$ threshold in order to increase the power to detect variants with modest effect in our discovery analyses, whilst the requirement for consistency in results from the two consortia aimed to limit false positives. All SNPs meeting these thresholds were followed up in independent replicatizon cohorts."
"We identified genes of interest as those with P<0·05 observed in both consortia and an overall $P<10^{-4}$, thresholds again chosen to limit both false positive and false negative findings."
The overall thresholds for the combined discovery and replication analyses were based on Bonferroni corrected thresholds, as already stated in the text.

*8. The authors should list the MAF alongside the p-values reported in the text for clarity for the single variant analysis results*

MAFs and P-values have now been added to the main text for all reported loci.

***Competing Interests:*** No competing interests were disclosed.

Referee Report 25 January 2018

**doi:**10.21956/wellcomeopenres.13627.r29790

**Rachel M. Freathy** , **Robin Beaumont**
Institute of Biomedical and Clinical Science, University of Exeter, Exeter, UK

The authors performed GWAS of FEV1, FVC and FEV1/FVC ratio at 179,215 SNPs from exome arrays. They identified 6 common frequency SNPs associated with at least one of these traits. They also identified 1 SNP in a region with known frequency differences across European populations suggesting that population structure may not have been fully accounted for in their analyses. Strengths of the study include the large sample size and comprehensive approach to assessing associations with low frequency and rare variants. We have the following concerns.

Main concerns:

1. The phenotypes seem to have been adjusted for covariates and ancestry specific principal components prior to being inverse normally transformed. This transformation has the potential to introduce correlations between principal components and the inverse normally transformed phenotype (https://www.biorxiv.org/content/early/2017/05/15/137232). Since one of the SNPs identified as being associated with the phenotype is known to vary in frequency across European populations, and the authors note that they cannot rule out the effects of population structure on the identified associations this raises concerns that some of the other associations could also be artefacts driven by failure to properly account for population stratification. It should explicitly be mentioned in the methods whether adjustments were made for ancestry specific principal components prior to inverse normal transforming the phenotype in the SpiroMeta Consortium component of the meta analysis or was included as a covariate in the phenotype - SNP association analysis.

2. Indeed, in the replication analysis in UK Biobank principal components were adjusted for prior to inverse normally transforming the data. Was genotyping chip adjusted for in this cohort (which should be done in the phenotype - SNP analysis)? The UKBiLEVE chip was enriched for smokers, which could affect association analyses unless chip is included as a covariate. In addition the

interim data release (which seems to be what is used here - please clarify in the methods whether the data comes from the interim (2015) or full (2017) data release) featured some discrepancies between the two chips, which can introduce spurious associations especially if adjustment is not made for genotyping chip.

3. Why was raw trait used in CHARGE but inverse normalised in SpiroMeta Consortium? This seems an odd choice

Minor concerns:

1. In the discussion, the authors mention that the 6 identified SNPs not attributed to population structure passed the Bonferroni significance threshold. They then mention that the SNPs ALSO pass Bonferroni corrected significance thresholds in the replication analysis. This could be misleading, since not all SNPs passed the Bonferroni threshold in the discovery only dataset.

2. The authors mention that correction was made for genomic inflation statistic (λ), but we could not find the statistics relating to this. The figures should be given in the manuscript.

**Is the work clearly and accurately presented and does it cite the current literature?**
Yes

**Is the study design appropriate and is the work technically sound?**
Partly

**Are sufficient details of methods and analysis provided to allow replication by others?**
Partly

**If applicable, is the statistical analysis and its interpretation appropriate?**
Yes

**Are all the source data underlying the results available to ensure full reproducibility?**
No source data required

**Are the conclusions drawn adequately supported by the results?**
Yes

***Competing Interests:*** No competing interests were disclosed.

**We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Reader Comment 12 Jun 2018

**Victoria Jackson**, University of Leicester, UK

We thank the reviewers for their helpful comments. We have addressed each specific comment below, and amended the manuscript correspondingly.

*1. The phenotypes seem to have been adjusted for covariates and ancestry specific principal*

*components prior to being inverse normally transformed. This transformation has the potential to introduce correlations between principal components and the inverse normally transformed phenotype (https://www.biorxiv.org/content/early/2017/05/15/137232). Since one of the SNPs identified as being associated with the phenotype is known to vary in frequency across European populations, and the authors note that they cannot rule out the effects of population structure on the identified associations this raises concerns that some of the other associations could also be artefacts driven by failure to properly account for population stratification. It should explicitly be mentioned in the methods whether adjustments were made for ancestry specific principal components prior to inverse normal transforming the phenotype in the SpiroMeta Consortium component of the meta analysis or was included as a covariate in the phenotype - SNP association analysis.*

In the SpiroMeta Consortium component of the analyses, adjustment for ancestry principal components (PCs) was not undertaken prior to transformation, rather PCs were adjusted for when fitting the SNP-trait associations. This is ambiguous in the text, and so we have amended the methods accordingly (Statistical analyses section, new wording below). Given that the adjustment for ancestry PCs was undertaken after phenotype transformation, we don't expect there to have been an introduction of correlation between the transformed trait and population structure.

"Traits were adjusted for sex, age, age$^2$ and height, and inverse normally transformed prior to association testing. For studies with unrelated individuals, SNP-trait associations were tested using linear models, with adjustments made for the first 10 ancestry principal components, whilst studies with related individuals utilised linear mixed models to account for familial relationships and underlying population structure."

*2. Indeed, in the replication analysis in UK Biobank principal components were adjusted for prior to inverse normally transforming the data. Was genotyping chip adjusted for in this cohort (which should be done in the phenotype - SNP analysis)? The UKBiLEVE chip was enriched for smokers, which could affect association analyses unless chip is included as a covariate. In addition the interim data release (which seems to be what is used here - please clarify in the methods whether the data comes from the interim (2015) or full (2017) data release) featured some discrepancies between the two chips, which can introduce spurious associations especially if adjustment is not made for genotyping chip.*

In the UK Biobank data, principal components (PCs) were adjusted for prior to transformation. As a sensitivity analysis, we have repeated the analysis for the six reported SNPs (the LCT SNP was not available in UK Biobank), transforming the phenotypes, and then adjusting for all covariates (including PCs) during the SNP-trait association test. For comparison, we have done this for all six SNPs with all three traits. Comparisons of these two analyses (not adjusted prior to transformation vs with adjustment prior to transformation) are shown here: https://doi.org/10.6084/m9.figshare.5959906. For each SNP, the P-value comparison is highlighted for the trait we report the association with, and the dashed lines indicate the Bonferroni corrected significance threshold for independent replication ($P<1{\cdot}47\times10^{-3}$). Whilst there is a difference in the P-values for some SNP-trait combinations, (more significant P-values in the analysis with covariate adjustment prior to transformation for 5 of the 6 SNPs), the SNPs all meet the replication P-value threshold in both analyses.

We have clarified in the methods (Study design, cohorts and genotyping section) that the UK Biobank data used was from the 2015 interim release. The UK Biobank analysis was stratified by

smoking status (ever and never) and also chip (UK BiLEVE array and UK Biobank array). It was not clear from the methods previously that the analysis was stratified for chip, so we have now made this clear in the methods.

We have also tested whether any of the six reported SNPs available in UK Biobank had different MAFs in the UK BiLEVE and UK Biobank samples (suggestive of a chip effect); however none showed evidence of this: https://doi.org/10.6084/m9.figshare.5959927.

*3. Why was raw trait used in CHARGE but inverse normalised in SpiroMeta Consortium? This seems an odd choice*

We agree that using the raw trait in CHARGE and the transformed trait in SpiroMeta was not ideal; however it was not planned to combine the results of these consortia from the outset. By the time we had made the decision to combine the results from the two consortia, all studies had already completed analyses and it was not feasible for contributing studies to repeat the analyses with/out the transformation, as this would have involved a substantial amount of reanalysis from contributing studies. Since the effect estimates were not on the same scale we could not do an inverse variance weighted meta-analysis; therefore we did a P-value based meta-analysis. This analysis should be valid given that appropriate analyses were done within each consortium.

*Minor concerns:*
*1. In the discussion, the authors mention that the 6 identified SNPs not attributed to population structure passed the Bonferroni significance threshold. They then mention that the SNPs ALSO pass Bonferroni corrected significance thresholds in the replication analysis. This could be misleading, since not all SNPs passed the Bonferroni threshold in the discovery only dataset.*

We have reworded this section of the discussion as follows: "There were six SNPs which reached $P<10^{-5}$ in the discovery stage meta-analysis of single variant associations, and subsequently met the Bonferroni corrected significance threshold for independent replication ($P<1{\cdot}47\times10^{3}$, corrected for 34 SNPs being tested). In the combined analyses of our discovery and replication analyses, these six SNPs met the exome chip-wide significance threshold ($P<2{\cdot}8\times10^{-7}$)."

*2. The authors mention that correction was made for genomic inflation statistic (λ), but we could not find the statistics relating to this. The figures should be given in the manuscript.*

We have added Supplementary table 13 to the supplement.

***Competing Interests:*** No competing interests were disclosed.