The protective effect of club cell secretory protein (CC-16) on COPD risk and progression: a Mendelian randomisation study

SUPPLEMENTARY METHODS

AUTHORS:

Stephen Milne[1,2,3]*, Xuan Li[1]*, Ana I Hernandez Cordero[1], Chen Xi Yang[1], Michael H Cho[4], Terri H Beaty[5], Ingo Ruczinski[6], Nadia N Hansel[7], Yohan Bossé[8], Corry-Anke Brandsma[9], Don D Sin[1,2], Ma'en Obeidat[1]
*equal contributions to the manuscript

1. Centre for Heart Lung Innovation, St Paul's Hospital and University of British Columbia, Vancouver, BC, Canada
2. Division of Respiratory Medicine, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
3. Faculty of Medicine and Health, University of Sydney, Sydney, New South Wales, Australia
4. Channing Division of Network Medicine and Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, USA
5. Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
6. Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA
7. Pulmonary and Critical Care Medicine, School of Medicine, Johns Hopkins University, Baltimore, MD, USA
8. Institut universitaire de cardiologie et de pneumologie de Québec, Department of Molecular Medicine, Laval University, Quebec City, Canada
9. University of Groningen Department of Pathology and Medical Biology, University Medical Centre Groningen, Groningen, The Netherlands

CORRESPONDING AUTHOR:

Dr Stephen Milne
UBC Centre for Heart Lung Innovation
Rm 166, St Paul's Hospital
1081 Burrard Street,
Vancouver, BC, V6Z 1Y6
CANADA

E: Stephen.milne@hli.ubc.ca

T: +1 604 806 8346

Milne S, *et al. Thorax* 2020; 75:934–943. doi: 10.1136/thoraxjnl-2019-214487

CONTENTS

Milne S, *et al. Thorax* 2020; 75:934–943. doi: 10.1136/thoraxjnl-2019-214487

1. DATASETS ANALYSED FOR SERUM CC-16 GENOME-WIDE ASSOCIATION STUDY

Description of study populations

Lung Health Study

The details of the LHS have been previously published.[1-3] The original LHS ("LHS-I") was a longitudinal study examining the effects of a smoking cessation intervention (including counselling and nicotine replacement therapy) and regular inhaled bronchodilator (ipratropium bromide) on the rate of change in lung function (post-bronchodilator FEV1) in smokers aged 35-60 years with mild-moderate COPD. COPD was defined by post-bronchodilator FEV1 55-90 percent predicted and FEV1/FVC ratio <0.7. A total of 5,887 participants were enrolled in LHS-I.[1] Interviews and spirometry were performed annually for 5 years. An extension of the original LHS, named LHS-III, involved an additional study visit 11 years after enrolment.[2]

In LHS, there was an increase in FEV1 between screening and Year 1, which has been attributed to the effects of smoking cessation.[1] For the remainder of the observation period, there was a general decline in FEV1 over time. Therefore, for the purposes of our analysis, data from Year 1 to 11 were used in the analysis of change in FEV1 over time.

Smoking status was assessed at each visit by self-report, and verified by exhaled carbon monoxide (eCO) and salivary cotinine (sCot) analysis. Data on smoking status was only available to us from LHS-I; we therefore assigned smoking status based on the baseline to year 5 data. We labelled participants as "continuous smokers" (eCO/sCot-verified smoking at all study visits), "sustained quitters" (eCO/sCot-verified non-smoking, plus self-reported number of cigarettes = 0, at every study visit) or "intermittent quitters" (eCO/sCot-verified non-smoking at some but not all study visits, or self-reported smoking despite negative eCO/sCot at some but not all visits). We then used these smoking status labels as covariates in the GWAS and pQTL versus lung function decline models.

At Year 5, the LHS investigators collected blood samples on 89% of the participants. A description of CC-16 measurement is given below. Genotyping in the LHS was undertaken using the Illumina Human660W-Quad v.1_A BeadChip (Illumina, San Diego, CA, USA), as previously described.[4]

Study protocols in the LHS were approved by the institutional review boards at each trial center, and written informed consent was obtained from each participant

Milne S, *et al. Thorax* 2020; 75:934–943. doi: 10.1136/thoraxjnl-2019-214487

Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE):

The details of the ECLIPSE cohort have been previously published.[5] ECLIPSE is a longitudinal study of current or former smokers with ≥10 pack year exposure aged 40-75 years, followed over 3 years. For the purposes of our analysis, we analysed only Caucasian COPD cases defined by post-bronchodilator FEV1 <80 percent predicted and FEV1/FVC ratio ≤0.7. Interviews and spirometry were performed at baseline, 3 months, 6 months and every 6 months after. An increase in mean FEV1 was observed between enrolment and Visit 1 (3 months). Therefore, we used only data from Visit 1 to Visit 7 (3 years) to determine change in FEV1 over time.

Smoking status was assessed at each study visit by interview, and verified by eCO. However, only smoking status from the baseline visit was available to us. We labelled participants as “smoker” (positive self-reported smoking or positive eCO) or “non-smoker” (negative self-reported smoking verified by negative eCO, or self-reported status missing but negative eCO). Subjects without an eCO status were considered missing data.

Blood was collected at baseline for biomarker analysis and genotyping. A description of CC-16 measurement is given below. Genotyping in the ECLIPSE cohort was undertaken using the HumanHap 550 V3 (Illumina) and quality control was performed using BeadStudio, as previously described.[6, 7] The present analysis is based on the use ECLIPSE study data downloaded from the dbGaP web portal, under study accession #phs001252.v1.p1 (available: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001252.v1.p1).

Human research ethics approval was granted by the institutional review boards at each of the participating centers.

Serum CC-16 measurement in the studies

Measurement of serum CC-16 levels in the LHS and ECLIPSE cohorts has been described previously.[8, 9] Briefly, blood samples were centrifuged and separated, and the serum frozen and stored until the time of analysis. CC-16 concentrations in thawed serum samples were measured using a commercially-available, sandwich enzyme linked immunosorbent assay (ELISA) (BioVendor, Heidelberg, Germany). Concentrations were determined by reference to a standard curve of known CC-16 concentrations.

In the LHS, subjects with serum CC-16 concentration below the lower limit of quantitation (LLQ) of 0.65 ng/mL were assigned a value equal to half the LLQ i.e 0.325 ng/mL.[9] For the present analysis, all subjects with CC-16 level equal to 0.325 ng/mL (n=240) were excluded due to their effect on the distribution of CC-16 concentrations. In ECLIPSE, subjects with serum CC-16 concentration below the LLQ were excluded; according to the study investigators, this equated to less than 1% of the cohort.[8] We further excluded outlier subjects from each cohort with serum CC-16 concentration >40 ng/mL (LHS, n=1, ECLIPSE, n=1). In order to better approximate a normal distribution, we transformed serum CC-16 concentrations by their natural logarithm prior their use in our analysis.

Effect of serum CC-16 level on COPD progression

To determine the relationship between CC-16 and COPD progression, we used a linear mixed-effects regression model. For individual i at time j:

$$FEV1_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})time_{ij} + \beta_2 lnCC16_i + \beta_3 lnCC16_i \times time_{ij} + \beta_4 Age_i + \beta_5 Age_i \times time_{ij} + \beta_6 Sex_i + \beta_7 Sex_i \times time_{ij} + \beta_8 BMI_i + \beta_9 BMI_i \times time_{ij} + \beta_9 Smoking_i + \beta_{10} Smoking_i \times time_{ij} + \beta_{11} Baseline\ FEV1_i + \beta_{12} Baseline\ FEV1_i \times time_{ij} + e_{ij}$$

where β represents the fixed effects, b represents the random effects, and e_{ij} represents the random errors. β_3 is therefore the estimated effect of lnCC-16 levels on FEV_1 change over time. Each covariate's interaction with time was included in the model to adjust for its effect on FEV_1 over time.

We also quantified FEV_1 change as the slope estimate of FEV1~time. This model is less robust for longitudinal series with few time points (e.g. ECLIPSE) and was therefore not used for any analysis, only for demonstration purposes (Supplementary Results Figure S1).

2. DATASET ANALYSED FOR ASSOCIATION WITH COPD: ICGC/UK BIOBANK META-ANALYSIS

A total of 35,735 COPD cases and 222,076 non-COPD controls within the International COPD Genetics Consortium (ICGC) and UK Biobank cohorts were included in a recent genome-wide

Milne S, *et al. Thorax* 2020; 75:934–943. doi: 10.1136/thoraxjnl-2019-214487

association meta-analysis for the presence of COPD.[10] For our study, we used the resulting summary statistics of this meta-analysis.

Description of cohorts included in the meta-analysis

The ICGC is an international collaboration of COPD cohort, case-control and general population studies with spirometry and genotype data available; the full description of the individual studies was previously described[11]. Case-control association analyses (based on prebronchodilator spirometry measurements, with COPD cases defined based on Global Initiative for Obstructive Lung Disease (GOLD) criteria [12]) were performed by each member of ICGC, and the results made available to the consortium; cohort-specific methods have been previously described.[11] Genotype imputation of the ICGC cohorts was performed using the 1000 Genomes reference panel.[13] Human research approval for was obtained for each cohort in the ICGC, as previously described.[11]

The UK Biobank cohort is part of the UK Biobank project,[14] which involved deep phenotyping and genotyping of a total of 502,682 individuals. The majority of individuals within this cohort are white Europeans. COPD cases within the UK Biobank cohort were also defined according to GOLD criteria[12] based on prebronchodilator spirometry. Genotyping of the UK Biobank cohort was performed using the Affymetrix Axiom UK BiLEVE and UK Biobank array, and genotypes were imputed to the Haplotype Reference Consortium version 1.1 panel.[15] Human research ethics approval for the UK Biobank project was granted by the North West Multi-centre Research Ethics Committee (MREC). Oversight for the ethical conduct of the UK Biobank project is performed by the UK Biobank Ethics Advisory Committee, for which the terms of reference can be found at https://www.ukbiobank.ac.uk/ethics/.

Meta-analysis

The authors performed logistic regression for COPD case/control status on the UK Biobank data; this analysis was adjusted for sex, age, genotyping array, smoking exposure (pack-years), ever-smoking status, and genetic principal components.[10] The authors then combined the association study results from the UK Biobank and each of the 22 ICGC studies by fixed-effects meta-analysis. Statistically significant results were defined at the genome-wide significant threshold of $p<5x10^{-8}$.

3. GENOTYPE IMPUTATION AND QUALITY CONTROL

Genotypes in the LHS and ECLIPSE were imputed to the Haplotype Reference Consortium panel version 1.1 using the Michigan Imputation Server.[16] After imputation, we removed 218,380 SNPs with duplicated position information. For our GWASs, we only kept SNPs with minor allele frequency (MAF) > 0.01 and imputation quality (r^2) ≥ 0.7 (7,426,986 SNPs in LHS, and 7,471,599 SNPs in ECLIPSE).

4. GENOME WIDE ASSOCIATION STUDIES FOR SERUM CC-16 LEVEL

In both LHS and ECLIPSE, we determined the effects of each SNP on serum CC-16 level using an additive genetic model. For individual i:

$$lnCC16_i = \beta_0 + \beta_1 SNP_i + \beta_2 Age_i + \beta_3 Sex_i + \beta_4 BMI_i + \beta_5 Smoking_i + \beta_6 PC1_i + \beta_7 PC2_i + \beta_8 PC3_i + \beta_9 PC4_i + \beta_{10} PC5_i + e_i$$

where β represents the effects and e_i represents the random errors. β_1 is therefore the estimated effect of an allele change on lnCC16 concentration. PC1-5 represent the first 5 genetic principal components.

We then combined the results from the two cohorts in a fixed-effects meta-analysis with each SNP-serum CC-16 association weighted by 1/(standard error of the CC-16 estimate)^2 (i.e. inverse variance weighted (IVW)). For the meta-analysis, we only included the 7,312,348 overlapping SNPs in both LHS and ECLIPSE.

In order to identify independently-associated SNPs, we performed conditional association analysis[17] within each 2 Mb gene region using the Genome-wide Complex Trait Analysis (GCTA) platform version 1.92.0beta3[18] with the larger cohort (LHS) as the linkage disequilibrium (LD) reference. The relevant conditional and joint P values are shown in Supplementary Results, Table S2. From this, we retained only those SNPs with independent, genome-wide significant ($p<5x10^{-8}$) association with CC-16 level.

5. GENETIC ASSOCIATIONS WITH COPD OUTCOMES

COPD progression (FEV_1 change over time)

To determine the effects of each SNP on COPD progression, we used a mixed effects model. For individual i at time j:

$$FEV1_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i})time_{ij} + \beta_2 SNP_i + \beta_3 SNP_i \times time_{ij} + \beta_4 Age_i + \beta_5 Age_i \times time_{ij} + \beta_6 Sex_i + \beta_7 Sex_i \times time_{ij} + \beta_8 BMI_i + \beta_9 BMI_i \times time_{ij} + \beta_9 Smoking_i + \beta_{10} Smoking_i \times time_{ij} + \beta_{11} Baseline\ FEV1_i + \beta_{12} Baseline\ FEV1_i \times time_{ij} + \beta_{13} PC1_i + \beta_{14} PC2_i + \beta_{15} PC3_i + \beta_{16} PC4_i + \beta_{17} PC5_i + e_{ij}$$

where β represents the fixed effects, b represents the random effects, and e_{ij} represents the random errors. β_3 is therfore the estimated effect of an allele change on FEV_1 over time. Each covariate's interaction with time was included in the model to adjust for its effects on FEV_1 over time. PC1-5 represent the first 5 genetic principal components.

"COPD risk"

To quantify the effects of the CC-16 pQTLS on COPD risk, we used the summary statistics from the ICGC meta-analysis for COPD case status[10] as described in Section 2 above.

6. MENDELIAN RANDOMISATION ANALYSIS

To estimate the causal effect of CC-16 concentration on COPD outcomes (Analysis D in Figure 1), we related the CC-16 pQTL per-allele effects on serum CC-16 levels to their effects on COPD outcomes ("COPD risk" in the ICGC dataset, and "COPD progression" in LHS and ECLIPSE) in a multi-variable MR analysis using the MendelianRandomization v0.2.2 package in R.[19, 20] The 7 serum CC-16 pQTLs identified from the GWAS meta-analysis were used as instrumental variables (IVs). We used an inverse variance weight (IVW) MR model i.e. the estimate was weighted by 1/(standard error of the COPD outcome effects)^2 as described by Burgess et al.[21] We adjusted for LD between SNPs on the same chromosome by calculating LD using PLINK v1.9[22, 23] (503 European-descent samples from the 1000 Genomes Project Phase 3)[13] and importing the correlations using the mr_input function in the MendelianRandomisation package. The model intercept was constrained to zero. We set nominal significance at $p<0.05$.

7. TESTING MR ASSUMPTIONS

MR analysis is only valid if the IVs meet a number of fundamental assumptions.[24] We employed a systematic approach to test for violations of these assumptions, as outlined by Burgess et al.[25]

Testing for genetic variant associations with confounders

A condition for a valid genetic IV is that it is not associated with a variable that confounds the risk factor-outcome association. To identify potential associations with confounders, we performed a search of the NHGRI-EBI Catalog of Human Genome-Wide Association Studies (https://www.ebi.ac.uk/gwas/), a publicly-available database of existing GWAS comprised of over 4,500 publications and 180,000 genetic associations. This catalogue reports SNP-trait associations with $p<1x10^{-5}$. We searched the database for each of the serum CC-16 pQTLs used in our analysis, with a plan to further investigating any genome-wide significant ($p<5x10^{-8}$) SNP-trait associations.

Testing for weak instruments:

IVs with weak association with the risk factor (i.e. explaining little of the variance) may bias the average causal estimate because influence of confounders of the risk factor-outcome association is greater. We therefore examined the relative strength of each IV using the partial F statistic.[26] With reference to the GWAS model described above, we compared two models – one with and one without the SNP – by ANOVA. The F stastic in essence quantifies how much additional variance can be explained by the addition of the SNP to the model. SNPs that are weakly associated with CC-16 will explain only a small additional proportion of the variance compared to the rest of the model, as indicated by a low F statistic. We took the the "rule-of-thumb" of F<10 as a sign of a weak instrument[26].

Testing for heterogeneity:

The presence of significant heterogeneity suggests that the variability in the IV estimates is greater than would be expected by chance alone, and thus may indicate bias due to one or more invalid IVs.[27] We used Cochran's Q test (Heter.Stat output in the MendelianRandomization package), with nominal significance $p<0.05$ indicating heterogeneity.

Testing for reliance on individual IVs:

Multi-variable MR has superior statistical power to single-instrument MR. However, it is possible that the average causal effect is driven largely by the SNPs most strongly associated with the outcomes. To test the reliance of the multi-variable MR analysis on individual SNPs, we performed a leave-one-out sensitivity analysis by excluding each variant one at a time and re-calculating the IVW MR estimate. If the MR estimate is no longer significant, it suggests that the analysis is heavily reliant on

Milne S, *et al. Thorax* 2020; 75:934–943. doi: 10.1136/thoraxjnl-2019-214487

that SNP. To summarise the sensitivity analyses, we plotted each of the MR estimates and 95% confidence intervals.

Robust MR methods:

IVW MR analysis has superior power to other methods, but it is particularly susceptible to violations of the MR assumptions. We therefore performed a number of alternative analysis methods that are less susceptible, particularly to the effects of invalid IVs. First, we performed the weighted median test using the mr_median function in the MendelianRandomization package. This is essentially a measure of central tendancy among the individual IV estimates that is particularly robust to outlier estimates and a small number of invalid IVs (up to 50 percent of the weight). We set statistical significance for this test at $p<0.05$. Next, we performed MR-Egger analysis[28] (mr_egger function in the MendelianRandomization package) which is similar to the IVW model but with an unconstrained intercept term. This model therefore allows directional pleiotropy, the presence of which is suggested by a significant (i.e. non-zero) intercept term (nominal significance set at $p<0.05$). Finally, we performed the Mendelian Randomization Pleiotropy RESidual Sum and Outlier (MR-PRESSO) test in R.[29, 30] This test has three components: 1) a "global" test for directional pleiotropy, which compares the observed distance from the MR regression line (residual sum of squares) to the predicted distance from the regression line under the null-hypothesis of no directional pleiotropy; 2) the identification and removal of outliers that contribute to directional pleiotropy; and 3) a test for significant differences in the MR causal estimates before and after removal of outliers. Under the MR-PRESSO framework, components 2) and 3) only proceed if the global test for directional pleiotropy (1) is significant at a nominal $p<0.05$.

8. DATASET ANALYSED FOR LUNG TISSUE GENE EXPRESSION: THE LUNG eQTL STUDY

Description of cohorts

The Lung eQTL Study[31] was a multi-national observational study examining non-tumor lung tissue samples collected from 1,111 people and 3 different sites (University of British Columbia (UBC), Laval University, and University of Groningen). The majority of samples were collected from study participants undergoing resection of presumed lung cancers. At the UBC site, n=39 samples were obtained at autopsy, n=22 from the diseased lungs of lung transplant recipients, and n=7 from donor lungs unsuitable for transplantation. For the majority of participants, lung function tests were

performed immediately before surgery. 84% percent of the participants across all sites were current or ex-smokers.

Genotyping was performed on either blood (Laval site) or lung (UBC, Groningen sites) samples using the Illumina Human 1M-Duo BeadChip array, as previously described.[31]

Ethics approval for the Lung eQTL Study was granted by the Institut Universitaire de Cardiologie et de Pneumologie de Québec and the UBC-Providence Health Care Research Institute Ethics Boards. Lung tissue samples were collected according to the institutional review board guidelines for each the participating institutions. Written informed consent was obtained from all patients, or families of deceased patients where applicable.

9. ANALYSIS OF GENE EXPRESSION IN THE LUNG eQTL STUDY

Tissue sample processing in the Lung eQTL Study has been previously described.[31] Lung tissue gene expression (mRNA) was measured using the Affymetrix HI133 array consisting of 751 control probesets and 51,627 non-control probesets. Gene expression values were extracted using the Affymetric Power Tools software (Robust Multichip Average method).[32] After quality control filtering, normalized expression data were adjusted for age, sex and smoking status.

10.LUNG cis-eQTL DETERMINATION

We estimated the associations between SNP and mRNA expression by linear regression models assuming additive genotype effects in each site separately. We then performed a meta analysis to combine the results. We defined cis-eQTLs as within 1 Mb up- or downstream of the SNP.

11.REFERENCES

1. Anthonisen NR, Connett JE, Kiley JP, et al. Effects of smoking intervention and the use of an inhaled anticholinergic bronchodilator on the rate of decline of FEV1: The Lung Health Study. *JAMA* 1994;272(19):1497-505. doi: 10.1001/jama.1994.03520190043033
2. Anthonisen NR, Connett JE, Murray RP. Smoking and lung function of Lung Health Study participants after 11 years. *Am J Respir Crit Care Med* 2002;166(5):675-79. doi: 10.1164/rccm.2112096
3. Connett JE, Kusek JW, Bailey WC, et al. Design of the Lung Health Study: a randomized clinical trial of early intervention for chronic obstructive pulmonary disease. *Control*

Clin Trials 1993;14(2 Suppl):3S-19S. doi: 10.1016/0197-2456(93)90021-5 [published Online First: 1993/04/01]
4. Hansel NN, Ruczinski I, Rafaels N, et al. Genome-wide study identifies two loci associated with lung function decline in mild to moderate COPD. *Hum Genet* 2013;132(1):79-90. doi: 10.1007/s00439-012-1219-6
5. Vestbo J, Anderson W, Coxson HO, et al. Evaluation of COPD Longitudinally to Identify Predictive Surrogate End-points (ECLIPSE). *Eur Respir J* 2008;31(4):869-73. doi: 10.1183/09031936.00111707 [published Online First: 2008/01/25]
6. 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. doi: 10.1371/journal.pgen.1000421 [published Online First: 2009/03/21]
7. Cho MH, Boutaoui N, Klanderman BJ, et al. Variants in FAM13A are associated with chronic obstructive pulmonary disease. *Nat Genet* 2010;42(3):200-02. doi: 10.1038/ng.535 [published Online First: 2010/02/21]
8. Lomas DA, Silverman EK, Edwards LD, et al. Evaluation of serum CC-16 as a biomarker for COPD in the ECLIPSE cohort. *Thorax* 2008;63(12):1058-63. doi: 10.1136/thx.2008.102574 [published Online First: 2008/09/02]
9. Park HY, Churg A, Wright JL, et al. Club cell protein 16 and disease progression in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2013;188(12):1413-9. doi: 10.1164/rccm.201305-0892OC [published Online First: 2013/11/20]
10. Sakornsakolpat P, Prokopenko D, Lamontagne M, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. *Nat Genet* 2019;51(3):494-505. doi: 10.1038/s41588-018-0342-2 [published Online First: 2019/02/26]
11. 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. doi: 10.1038/ng.3752 [published Online First: 2017/02/07]
12. Vogelmeier CF, Criner GJ, Martinez FJ, et al. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 report: GOLD executive summary. *Eur Respir J* 2017;49(3):1700214. doi: 10.1183/13993003.00214-2017 [published Online First: 2017/02/10]
13. The 1000 Genomes Project Consortium, Auton A, Abecasis GR, et al. A global reference for human genetic variation. *Nature* 2015;526:68. doi: 10.1038/nature15393 [published Online First: 2015/09/30]
14. Sudlow C, Gallacher J, Allen N, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;12(3):e1001779. doi: 10.1371/journal.pmed.1001779 [published Online First: 2015/04/01]
15. McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 2016;48:1279. doi: 10.1038/ng.3643 [published Online First: 2016/08/22]
16. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet* 2016;48:1284. doi: 10.1038/ng.3656 [published Online First: 2016/08/29]
17. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat*

Genet 2012;44(4):369-75, S1-3. doi: 10.1038/ng.2213 [published Online First: 2012/03/20]
18. Yang J, Lee S, Goddard M, et al. GCTA: a tool for Genome-wide Complex Trait Analysis version 1.92.0beta3 [online software]. Available: https://cnsgenomics.com/software/gcta/ [last accessed 31 March 2019].
19. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* 2017;46(6):1734-39. doi: 10.1093/ije/dyx034
20. Yavorska OO, Burgess S. MendelianRandomization: Mendelian Randomization Package version 0.2.2 [online software package] Available: https://cran.r-project.org/web/packages/MendelianRandomization/ [last accessed 31 October 2019].
21. Burgess S, Butterworth A, Thompson SG. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet Epidemiol* 2013;37(7):658-65. doi: 10.1002/gepi.21758
22. Purcell S, Chang C. PLINK version 1.9 [online software package]. Available: www.cog-genomics.org/plink/1.9/ [last accessed 31 October 2019].
23. Chang CC, Chow CC, Tellier LC, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4(1):7. doi: 10.1186/s13742-015-0047-8 [published Online First: 2015/02/28]
24. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;32(1):1-22. doi: 10.1093/ije/dyg070 [published Online First: 2003/04/12]
25. Burgess S, Davey Smith G, Davies N, et al. Guidelines for performing Mendelian randomization investigations [version 2; peer review: 1 approved, 1 approved with reservations]. *Wellcome Open Research* 2020;4(186) doi: 10.12688/wellcomeopenres.15555.2
26. Burgess S, Thompson SG, Collaboration CCG. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol* 2011;40(3):755-64. doi: 10.1093/ije/dyr036
27. Burgess S, Bowden J, Fall T, et al. Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* 2017;28(1):30-42. doi: 10.1097/EDE.0000000000000559 [published Online First: 2016/10/18]
28. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015;44(2):512-25. doi: 10.1093/ije/dyv080 [published Online First: 2015/06/08]
29. Verbanck M, Chen C-Y, Neale B, et al. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* 2018;50(5):693-98. doi: 10.1038/s41588-018-0099-7
30. Verbanck M. MR-PRESSO version 1.0 [online software package]. Available: https://github.com/rondolab/MR-PRESSO [last accessed 13 December 2019].
31. Hao K, Bosse Y, Nickle DC, et al. Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS Genet* 2012;8(11):e1003029. doi: 10.1371/journal.pgen.1003029 [published Online First: 2012/12/05]
32. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003;4(2):249-64. doi: 10.1093/biostatistics/4.2.249 [published Online First: 2003/08/20]