Genetic analysis of lung cancer reveals novel susceptibility loci and germline impact on somatic mutation burden

Aurélie AG Gabriel, PhD *[1] and Joshua R Atkins, PhD*[1], Ricardo CC Penha, PhD [1], Karl Smith-Byrne, PhD [1], Valerie Gaborieau, DipHE [1], Catherine Voegele, PhD [1], Behnoush Abedi-Ardekani, MD[1], Maja Milojevic, PhD [1], Robert Olaso, PhD [2], Vincent Meyer, PhD [2], Anne Boland, PhD [2], Jean François Deleuze, PhD [2], David Zaridze, PhD, MD [3], Anush Mukeriya, PhD [3], Beata Swiatkowska, PhD [4], Vladimir Janout, PhD, MD [5], Miriam Schejbalová, PhD [6], Dana Mates, PhD [7], Jelena Stojšić, PhD [8], Miodrag Ognjanovic, PhD [9], the ILCCO consortium, John S Witte, PhD [10], Sara R Rashkin, PhD [10,11], Linda Kachuri, PhD [10], Rayjean J Hung, PhD [12], Siddhartha Kar, PhD, MD [13,14], Paul Brennan, PhD [1], Anne-Sophie Sertier, PhD [15], Anthony Ferrari, PhD [15], Alain Viari, PhD [15,16], Mattias Johansson, PhD [1], Christopher I Amos, PhD [17], Matthieu Foll, PhD [1], James D McKay, PhD [1]

*** Authors contributed equally to this presented work**

1. International Agency for Research on Cancer/World Health Organization (IARC/WHO), Genomic Epidemiology branch, Lyon, France
2. Université Paris-Saclay, CEA, Centre National de Recherche en Génomique Humaine, 91057, Evry, France
3. Russian N.N. Blokhin Cancer Research Centre, Moscow, The Russian Federation
4. Nofer Institute of Occupational Medicine, Department of Environmental Epidemiology, Lodz, Poland
5. Faculty of Health Sciences, Palacky University, Olomouc, Czech Republic
6. Charles University 1st Faculty of Medicine, Prague Czech Republic
7. National Institute of Public Health, Bucharest, Romania
8. Department of Thoracic Pathology, Service of Pathology, University Clinical Centre of Serbia, 11000 Belgrade, Serbia
9. International Organisation for Cancer Prevention and Research, Belgrade, Serbia
10. Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA, USA
11. Department of Hematology, St. Jude Children's Research Hospital, Memphis, TN, USA
12. Lunenfeld Tanenbaum Research Institute, Sinai Health
13. MRC Integrative Epidemiology Unit, University of Bristol, UK
14. Population Health Sciences, Bristol Medical School, University of Bristol, UK
15. Synergie Lyon Cancer, Plateforme de bioinformatique 'Gilles Thomas' Centre Léon Bérard, 28 promenade Lea et Napoleon Bullukian, 69008 Lyon, France
16. INRIA Grenoble-Rhône-Alpes, 655 Avenue de l'Europe, Montbonnot-Saint-Martin 38330, France
17. Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA

Correspondence
JD McKay email: mckayj@iarc.fr

ABSTRACT

Germline genetic variants are involved in lung cancer (LC) susceptibility. Previous genome-wide association studies (GWAS) have implicated genes involved in smoking propensity and DNA repair but further work is required to identify additional LC susceptibility variants and to investigate LC disease development dynamics.

We have undertaken a family history-based genome-wide association (GWAx) study of LC, analysing 48,843 European cases with a parent/sibling with LC compared to 195,387 controls from the UK Biobank. This was meta-analysed with previously described LC GWAS results. We performed Polygenic Risk Scores (PRS) analyses and further evaluated the PRS influence on the somatic environment in exome (N=736) and genome sequencing (N=61) profiled cohorts.

Eight novel loci were identified including DNA repair genes (*CHEK1*, *MDM4*), metabolic genes (*CYP1A1*) and variants that were also associated with smoking propensity, such as both subunits of the neuronal α4β2 nicotinic acetylcholine receptor (*CHRNA4* and *CHRNB2)*. PRS analysis demonstrated that variants related to eQTLs and/or smoking propensity are enriched for susceptibility variants, including variants below genome-wide significant threshold. PRS of LC variants related to smoking propensity were associated with somatic mutation burden in two case cohorts, with individuals with higher polygenic genetic risk having increased numbers of somatic mutations in their lung tumours.

This study has expanded the number of susceptibility loci linked with LC and provided insights into the molecular mechanisms by which these susceptibility variants contribute to the development of lung cancer.

Keywords: Lung Cancer, GWAx, GWAS, Smoking, α4β2 nicotinic acetylcholine receptor, Mutational Signatures
Abstract 232
Word count: 2965

INTRODUCTION

Lung cancer (LC) is the most common cause of cancer-related deaths worldwide. While most LC risk is attributable to exposure to tobacco smoke, a genetic basis for LC susceptibility was initially identified from familial aggregation studies after accounting for personal smoking habits [1–3], segregation based analyses[4] and twin studies.[5] Genome-wide association studies (GWAS) have identified multiple lung cancer susceptibility loci in genes related to propensity to smoke tobacco (*CHRNA5*, *CHRNA3*, *CHRNB4*, *CYP2A6*)[6–8], DNA repair (*CHEK2*, *BRCA2* , *ATM*)[9–11] and genes related to telomere regulation (*TERT*, *RTEL*, *OBFC1*)[12,13] as well as many loci where the target genes are less obvious.[13]

While traditional GWAS approaches continue to expand in size, novel analytical approaches can leverage existing data from large, genotyped biorepositories to identify additional loci. An example is the genome-wide association by proxy (GWAx) approach, which considers 1st degree relatives previously diagnosed with the given trait of interest as "proxy" cases and individuals without relatives with that given trait as "proxy" controls.[14,15] In the current study, we undertook a GWAx of lung cancer in the UK Biobank and combined this with the largest GWAS of LC undertaken to date.[13] Furthermore, we constructed Polygenic Risk Scores (PRS) with variants related to LC and used these scores to investigate the influence of these germline susceptibility variants on the somatic mutation burden in two independent cohorts.

MATERIALS AND METHODS

Cohorts

A detailed description of each dataset (Transdisciplinary Research for Cancer in Lung (TRICL, the traditional LC GWAS),[13] UK Biobank (GWAx and a subset left out of the GWAx, forming the germline PRS test set)[16], The Cancer Genome Atlas (TCGA, somatic mutations and signatures analysis (https://www.cancer.gov/tcga)) and the GeniLuc cohort (somatic mutation and signatures replication cohort (unpublished)) can be found in the supplementary information.

Genome-wide association by using a family history and genetic correlations analysis

The UK Biobank resource was accessed under project number 15825. The sample selection process and variant filtering from the UK Biobank is detailed in Supplementary Table 1 along with the UK biobank data ID fields. We performed a traditional GWAS on the family history status (GWAx) of LC and adjusted the betas and standard error as described previously.[14] Individuals diagnosed with lung cancer directly were not included in the GWAx. All samples were reported as having a European ancestry (confirmed with ancestry inferred by genetic profile) with non-European individuals excluded from the study due to different genetic architecture. Consent and ethics were approved for all cohorts used. Genetic correlation analysis was undertaken using the LDSC package.[17] Summary statistics were obtained from the LC traditional GWAS which has been previously published elsewhere. [13] Each summary statistic file from the Sequencing Consortium of Alcohol and Nicotine (GSCAN) consortium (with UK Biobank samples removed) along with the traditional LC GWAS and the GWAx were tested for genetic similarity using LDSC regression. [18]

Both the LC family history GWAx and the LC GWAS were meta-analysed using a fixed effect model. LD clumping was undertaken using PLINK ($R^2 < 0.1$ and 10,000 kb). eQTL analysis was performed using GTEx version 8 data for both lung and brain tissues containing all variant and gene pairs. Coffee intake and forced vital capacity (FVC) summary statistics were obtained from the Benjamin Neale UK Biobank work (http://www.nealelab.is/uk-biobank/). To estimate the colocalization between genetic associations of two traits at a given locus, we calculated the Bayesian posterior probability (PP4) for colocalisation of two datasets for the H_4 (one shared variant across both traits),[19] by firstly calculating the log bayes factor for each SNP in each dataset, then the PP4 was calculated by the COLOC package in python (https://github.com/anthony-aylward/coloc).

Mutation burden analysis

The somatic mutations from the TCGA samples were retrieved from the study of Ellrott *et al 2018* [20] excluding individuals flagged for QC issues (see supplementary methods for further QC details). Germline genotypes were derived from Affymetrix 6.0 arrays.

For the GeniLuc cohort, 61 lung cancer patients were identified from central and eastern Europe as described previously.[13] Subsequent to histopathological review, to ensure appropriate tumour purity, DNA was extracted from normal material (blood) and the lung tumour resection. Whole genome sequencing (WGS) was undertaken using PCR free whole genome library preparation and sequenced to a depth of 30X for the paired tumour normal for each patient using an Illumina HiSeq X 5 DNA sequencer at the National Center of Human Genomic Research (CNRGH) laboratory in Paris, France. Raw sequencing data was processed by inhouse Nextflow pipelines (https://github.com/IARCbioinfo). Somatic mutations were defined using Mutect2 and germline calls using Strelka2.[21,22] For germline genotypes from WGS tissue, PRS SNPs were extracted from VCF files and put in the PLINK BED format.[23] Individual PRS scores were generated using PRsice2 from the normal calls.[24]

Mutational signatures computation

In order to compute mutational signatures, mutational matrices for each mutation type (Single Base Substitution (SBS), Doublet Base Substitution (DBS), and small Insertion and Deletion (ID)) were generated using SigProfilerMatrixGenerator (v1.1.20) with default parameters.[25] Mutational signatures were then extracted with SigProfilerExtractor (v1.0.17) from the TCGA-WES (LUAD and LUSC) samples and GENILUC-WGS lung cancer cohorts, separately, using the default options.[26] SigProfilerExtractor extracted *de novo* signatures for each context (SBS96, DBS78, and ID83) and the optimum number of *de novo* signatures (suggested solution method) were decomposed into COSMIC (version 3.1) reference signatures. Previously reported smoking tobacco-related signatures, SBS4, DBS2, and ID3 (ID83A and ID83B), and the absolute mutation counts for each COSMIC signature per sample were assessed.

Statistical analysis

For the GWAx analysis, association testing was performed using a logistic regression model using the --glm function in PLINK 2.0 on European ancestry individuals. Each model was adjusted by age at recruitment, sex, array type, and the first 5 principal components that define genetic ancestry (PCs) to account for population structure.
The meta-analysis was done using METASOFT using a fixed-effects model based on an inverse-variance-weighted effect size.[27] Germline PRS analysis in the UK Biobank samples was performed using a logistic regression model after standardising raw PRS scores. Covariates that were used in the model included sex, array type, age of recruitment and the first 5 principal components from genetic inferred ancestry. Odd ratios for PRS are given as a one unit increase per a standard deviation in score. For the analysis of PRS associations with mutational signatures, a model diagnostic was used to compare a linear model, negative binomial model and a Quasi-Poisson model due to frequent zero-inflation for mutational signatures. Covariates included in the models were age, gender, the 5 first principal components resulting from Eigenstrat and tumour purity. In TCGA, a categorical variable indicating the cohort type was included in the model as appropriate.

RESULTS

The 8 novel susceptibility loci

The family history GWAS (GWAx) on 48,843 self-reported "family history lung cancer cases" and 197,029 "controls" (Supplementary Table 1) identified five loci (5p15.33, 6p21.32, 12p13.33, 13q13.1 and 15q25.1) that had previously been discovered from the traditional GWAS (Supplementary Table 2 and Supplementary Figure 2) and LDSC confirmed a strong relationship between both GWAx and the GWAS (r_g = 1, se = 0.066, p = 4.0 x 10^{-52}) supporting the utility of this approach (see supplementary material document for further details).
Meta-analysis between the GWAx and the traditional LC GWAS [13] identified 65 variants that achieved a P-value of less than $5x10^{-8}$ across 21 distinct genomic loci defined by cytoband (Figure 1), after LD clumping genetic variants (Supplementary Table 2). At previously described lung cancer susceptibility loci, the meta-analysis also identified independent (R^2 < 0.1) low-frequency (MAF < 0.05) variants associated with lung cancer at 5p15.33 (rs35812074), 19q13.2 (rs1801272), 15q25.1 (rs2229961, rs8192479, rs151118057) and at 12p13.33 (rs7487683) in addition to previously described common genetic variants (Supplementary Table 2). At 13q13.1, where a rarer lung cancer susceptibility allele has previously been described (rs11571833, K3326X *BRCA2*, MAF = 0.01), a common susceptibility allele was noted (rs11571734, MAF = 0.28).

Eleven lung cancer susceptibility variants at eight loci have not previously associated with lung cancer at genome wide (GW) significance (Table 1). Of these, the lung cancer susceptibility variants at 1q21.3-rs78062588, 6p22.2-rs7766641 and 20q13.33-rs11697662 were also associated at GW significance with traits related to propensity to smoke tobacco (Supplementary Table 2). The sentinel variants at 1q21.3-rs78062588 and 20q13.33-rs11697662 are

eQTLs for the nicotinic acetylcholine receptors (nAChRs) subunits *CHRNB2* and *CHRNA4* (Supplementary Table 2 and Supplementary Table 3). At 6p22.2, LC susceptibility loci were noted (Supplementary Table 2), typified by two sentinel variants, rs6913550 and rs7766641. rs7766641 was also associated with propensity to smoke, whereas curiously rs6913550 was not (Supplementary Table 2).

At 1q32.1, 11p11.2, 11q24.2 and 15q24, the sentinel variants (rs4252707, rs72905558, rs61612408, rs12441817, respectively) were not associated with propensity to smoke (Supplementary Table 2). 11q24.2-rs61612408 was associated with the expression of the *CHEK1* gene in multiple tissues including lung epithelia (colocalisation between *CHEK1* lung eQTL and LC: PP4 = 91.1%), with the allele associated with increased expression correlating with decreased risk of lung cancer (Figure 2.C). The association with 11q24.2-rs61612408 appeared to be more prominent in lung squamous cell carcinomas (Table 1). 15q24-rs12441817 is located near the *CYP1A1* and *CYP1A2* enzymatic genes. This locus has been associated with coffee consumption and forced vital capacity (FVC),[28,29] although there was colocalisation between variants associated with lung cancer only for FVC (colocalisation between coffee consumption and LC: PP4 = 0.0003%, colocalisation between FVC and LC: PP4 = 97.05%) (Supplementary Figure 3). There was evidence that rs12441817 influenced *CYP1A1* expression in the nucleus accumbens (colocalisation PP4 = 70.26%) (Supplementary Figure 4) and an eQTL effect with the processed pseudogene *RP11-10O17.1* in lung tissue (colocalisation between eQTL *RP11-10O17.1* and LC PP = 95.25%) (Figure 2.D). At 4q13.2-rs185666783 the candidate genes remain ambiguous (AC104806.2 and RNU6-699P) and the association with lung cancer appeared most prominent in lung adenocarcinoma. At 11p11.2, rs72905558 was associated with expression of *C1QTNF4* in lung tissue reported in GTEx but there was no evidence for colocalization between variants related to *C1QTNF4* expression and lung cancer (*C1QTNF4* PP4 = 0.06%).

Exploration of subgenome-wide significant variants and integrative multi-trait polygenic risk score construction.

The variants that achieved GW significance also tended to be associated with propensity to smoke and/or an eQTL (Supplementary Table 2). We therefore used partial least squares regression (PLS) to identify subgenome wide significant genetic variants with similar propensity to smoke and/or an eQTL features (represented by PLS components) (see Supplementary Material for details). We constructed bins of variants ranked by these PLS components and represented them as a function of the mean LC association statistic calculated within each bin (Figure 3.A). The bins that were ranked highly by a smoking and/or an eQTL components were observed to have elevated mean LC association statistics relative to most other bins, implying that the variants within these bins are enriched for LC susceptibility alleles (Figure 3.A). Interestingly, this enrichment was more marked for the eQTL PLS component (Figure 3.A). We constructed two polygenic risk scores, smPRS and eQTLPRS, based on the top 100 and 1,000 ranking SNPs from the smoking and eQTLs PLS analyses, respectively (see Methods section and Supplementary Material), with number of variants guided by the degree of enrichment observed (Figure 3.A) and tested them in an independent cohort of 1,666 lung cancer cases and 6,664 matched controls from the UK Biobank. The PRS were robustly associated with lung cancer in this independent series (smPRS: OR per standard deviation = 1.246, 95% CI: 1.176-1.32, P = 8.5×10^{-14}; eQTLPRS: OR = 1.349, 95% CI: 1.27-1.43, P = 7.29×10^{-17}, combined (both smPRS and eQTLPRS combined after adjusting for variant overlap), OR = 1.366, 95% CI 1.288-1.448, P = 1.44×10^{-25}). These risk estimates were only modestly attenuated when excluding the GWAS significant variants and adjusting for smoking status, again implying that these variants are enriched for LC-susceptibility alleles (Figure 3.B).

PRS germline influences on mutational burden and mutational signatures

We evaluated the association of the smPRS and eQTLPRS with somatic mutational burden in the 736 TCGA lung cancer patients where somatic and germline data overlapped and passed QC metrics (see Methods). There was little evidence for association involving the eQTLPRS and mutation burden (Supplementary Figure 8), however the smPRS was associated with tumour mutational burden (TMB) (P = 1.23×10^{-3}, Figure 4.A), with evidence of a trend between increasing polygenic load and somatic mutation burden (Figure 4.A). The smPRS was similarly associated with burden of mutational signatures attributed to tobacco smoke (SBS4 (P = 9.73×10^{-5}), ID3A (P = 1.78×10^{-3}), ID3B (P = 3.77×10^{-2}) and DBS2 (P = 3.05×10^{-3})) (Figure 4.A and Supplementary Figure 6). These associations were observed more prominently in patients with LUAD (Figure 4.A). The 15q25 *CHRNA5* lung cancer sentinel variant, rs72740955, had the most striking effect (Supplementary Table 2 and Supplementary Figure 7) but the associations remained significant after excluding genome-wide variants for lung cancer (Figure 4.A). The associations between the smPRS and somatic mutation burden (P = 0.034) and with mutation signatures attributed to tobacco smoking (SBS4: P = 0.023, ID3: p = 0.054, DBS2: P = 0.035, Supplementary Figure 8) were similarly observed in an independent cohort of 61 lung cancer patients whose germline and matched tumour samples have undergone WGS,

replicating this finding. We additionally projected the smPRS into other cancer types in TCGA cohorts and the association with TMB was also observed in the esophageal carcinoma (ESCA) cohort (Supplementary Table 6).

DISCUSSION

This study identified 21 lung cancer susceptibility loci, including eight novel loci by combining large, genotyped biobank data and traditional genome-wide association studies. Three of eight novel loci were also associated with propensity to smoke. This included brain eQTLs variants for both subunits of the neuronal nAChRs α4β2 receptor. Variants in LD with the α4 subunit (rs2373500) have been described in nicotine dependency and lung cancer risk, albeit not at GW significance for lung cancer,[18] while the β2 receptor and lung cancer risk has not been described. The neuronal nAChRs α4β2 receptor is the most abundant nAChR subtype within the human brain and important within the dopaminergic signalling pathway. The α4β2 receptor has a key role in nicotine dependence behaviours[30] and a major target in nicotine addiction intervention.[31,32] The third novel locus related to lung cancer and propensity to smoke is telomeric to the MHC region, where the target candidate gene(s) is less obvious. The MHC region was among the first susceptibility loci to be associated with lung cancer.[6,33–35] However, rs7766641, is not in LD with these previously described variants ($R^2 < 0.001$) and associated with the number of cigarettes smoked per day, implying that these are distinct associations.

This meta-analysis also identified additional lung cancer susceptibility loci that appear to be independent of smoking propensity. This included variants at 15q24 near *CYP1A1*, *CYP1A2* and *CYP11A1* that participate in the metabolism of many different xenobiotics and some endogenous substrates. Variants at the 15q24 *CYP1A1* / *CYP1A2* locus have been linked with multiple traits, notably other forms of propensity (coffee consumption) and forced vital capacity (FVC).[28,29] The coffee consumption/FVC variants at this locus appear distinct and colocalisation analysis implicates FVC more in the lung cancer association, which also seems aetiologically more plausible. For tissue expression, rs12441817 colocalised with lung tissue expression of the processed pseudogene *RP11-10O17.1* (Figure 2.D) although how this pseudogene relates to lung cancer susceptibility is unclear.

An additional novel lung cancer susceptibility variant, rs61612408, was a lung tissue eQTL for the DNA repair gene *CHEK1* (Figure 2.C). Similar to previously described variants associated near *CHEK2* and *BRCA2*, the association between rs61612408 and lung cancer appears more prominent in lung squamous cell carcinomas[9,10]. We additionally noted the variant impacting the *MDM4* gene, which is an important p53 regulator. This variant was previously associated with non-glioblastoma tumours[36] and more recently squamous cell carcinomas of the lung and head / neck,[37] although here we noted weak evidence for association in lung adenocarcinoma (Table 1). At 11p11.2 colocalisation analysis showed little evidence for involvement with genes *C1QTNF4* (lung) and *MTCH2* (brain-cortex), suggesting that these signals are unlikely to explain the lung cancer association, one other candidate is potentially *PTPRO* which is hypermethylated in several cancers including lung.[38,39] At 4q13.2, the finding remains ambiguous, but from histological subtypes analysis performed from the previous reported GWAS study, it appears that this signal is mostly found in lung adenocarcinoma.

We additionally sought to use the shared genetic aetiology between lung cancer susceptibility and smoking related traits and gene expression annotations (eQTL) to explore variants that did not achieve GW significance. We used the partial least squared (PLS) method to select variants related to these traits for the PRS analysis and demonstrated that such variants are indeed enriched for susceptibility alleles. While the role of these individual variants remains to be confirmed, these sub-GW significant variants were located near relevant candidate genes (smoking traits like *CHNRA6*, *DBH* and eQTLs for *ERCC2*, *RAD51C*, *XRCC3* and *CASP8*). Combining both sub-GW PRS lists (smoking PRS and eQTL PRS) with GW significant results reached an OR of 1.36 per standard deviation unit increase in score improving on previous PRS predictions (OR 1.17 and 1.26),[40,41] despite the conservative clumping approach ($R2 < 0.1$) employed. This suggests that integrating functional annotations may be of interest for PRS.

Lastly, the analysis of the smoking PRS demonstrated an association between a person's genetic risk load and mutation burden, and or burden of tobacco-related somatic mutational signatures, within two independent case cohorts and using different sequencing methods (exome sequencing and whole genome sequencing). These associations appear consistent with the notion that genetic variants influence an individual's smoking behaviour, which in turn, influences their carcinogenic exposure and consequently, their somatic mutation burden.

In conclusion, this work has increased the number of variants associated with lung cancer susceptibility, with the identification of novel susceptibility loci. PRS analysis highlighted that many additional variants remain to be discovered and provided insights into the carcinogenic mechanisms.

Funding
This work was supported by the Institut National du Cancer (INCa) (GeniLuc 2017-1-TABAC-03-CIRC-1 - [TABAC 17□022], NIH/NCI, Integral NIH 5U19CA203654-03, Cancer Research UK [grant number C18281/A29019], the France Génomique National infrastructure, funded as part of the « Investissements d'Avenir » program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09). Christopher Amos is a Research Scholar of the Cancer Prevention Institute of Texas and supported by RR170048

Conflicts of Interest Statement
The authors have no conflicts of interest regarding the present study

Acknowledgements
We would like to acknowledge the TCGA Research Network (https://www.cancer.gov/tcga) and the contribution of specimen donors and research groups involved in this resource. We also would like to acknowledge the GTEx project and the supporting bodies (https://commonfund.nih.gov/GTEx), specimen donors and research groups. Additionally, we would like to acknowledge the work carried about by the Benjamin Neale lab for their work on the UK Biobank (http://www.nealelab.is/uk-biobank/).
The ILLCO consortium is listed in the supplementary text with affiliations.

Disclaimer: Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, James McKay (mckayj@iarc.fr)

Data and Code Availability

The code generated during this study are available at GitHub (https://github.com/IARC-genetics/GWAx_lung_cancer). The polygenic risk scores variants used in this study are available within the supplementary tables. Summary statistics from the UK Biobank lung cancer family history summary statistics will be made available on GWAS Catalog. Summary statistics from the meta-analysis (McKay et al. 2017 and the UK Biobank lung cancer family history) are not publicly available due to controlled access of Oncoarray consortium data. Oncoarray data can be accessed by the database of Genotypes and Phenotypes (dbGaP) under accession phs000876.v1.p1

CRediT author Statement

Aurélie AG Gabriel: Conceptualization, Formal analysis, Data curation, Methodology, Investigation, Software, Validation, Writing – Draft, review and editing, Visualization **Joshua R Atkins**: Conceptualization, Formal analysis, Data curation, Methodology, Investigation, Software, Validation, Writing – Draft, review and editing, Visualization **Ricardo CC Penha**: Formal analysis, Data curation, Validation, Investigation, Writing-Review & Editing. **Karl Smith-Byrne**: Formal analysis, Writing-Review & Editing **Valerie Gaborieau**: Formal analysis, Data curation **Catherine Voegele**: Formal analysis, Data curation **Behnoush Abedi-Ardekani**: Formal analysis, Data curation **Maja Milojevic**: Formal analysis, Data curation **Robert Olaso**: Data curation **Vincent Meyer**: Data curation **Anne Boland**: Data curation **Jean François**: Data curation **David Zaridze**: Resources **Anush Mukeriya**: Resources **Beata Swiatkowska**: Resources **Vladimir Janout**: Resources **Miriam Schejbalová**: Resources **Dana Mates**: Resources **Jelena Stojšić**: Resources **Miodrag Ognjanovic**: Resources **the ILCCO consortium**: Resources, Funding acquisition **John S Witte**: Data curation **Sara R Rashkin**: Data curation **Linda Kachuri**: Writing - Review and Editing **Rayjean J Hung**: Writing - Review and Editing **Siddhartha Kar**: Writing - Review and Editing **Paul Brennan**: Resources **Anne-Sophie Sertier**: Data curation **Anthony Ferrari**: Data curation **Alain Viari**: Data curation **Mattias Johansson**: Writing - Review and Editing **Christopher I Amos**: Conceptualization, Writing - Review and Editing, Funding acquisition, Resources **Matthieu Foll**: Conceptualization, Supervision, Methodology, Writing – Draft, review and editing **James D McKay**: Conceptualization, Supervision, Data curation, Methodology, Writing – Draft, review and editing, Funding acquisition

REFERENCES

1. Tokuhata GK, Lilienfeld AM. Familial aggregation of lung cancer in humans. *J Natl Cancer Inst*. 1963;30:289-312. https://www.ncbi.nlm.nih.gov/pubmed/13985327

2. Schwartz AG, Yang P, Swanson GM. Familial risk of lung cancer among nonsmokers and their relatives. *Am J Epidemiol*. 1996;144(6):554-562. doi:10.1093/oxfordjournals.aje.a008965

3. Ooi WL, Elston RC, Chen VW, Bailey-Wilson JE, Rothschild H. Increased familial risk for lung cancer. *J Natl Cancer Inst*. 1986;76(2):217-222. doi:10.1093/jnci/76.2.217

4. Sellers TA, Bailey-Wilson JE, Elston RC, et al. Evidence for mendelian inheritance in the pathogenesis of lung cancer. *J Natl Cancer Inst*. 1990;82(15):1272-1279. doi:10.1093/jnci/82.15.1272

5. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA*. 2016;315(1):68-76. doi:10.1001/jama.2015.17703

6. Hung RJ, McKay JD, Gaborieau V, et al. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature*. 2008;452(7187):633-637. doi:10.1038/nature06885

7. Amos CI, Wu X, Broderick P, et al. Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat Genet*. 2008;40(5):616-622. doi:10.1038/ng.109

8. Patel YM, Park SL, Han Y, et al. Novel association of genetic markers affecting CYP2A6 activity and lung cancer risk. *Cancer Res*. 2016;76(19):5768-5776. doi:10.1158/0008-5472.CAN-16-0446

9. Brennan P, McKay J, Moore L, et al. Uncommon CHEK2 mis-sense variant and reduced risk of tobacco-related cancers: case control study. *Hum Mol Genet*. 2007;16(15):1794-1801. doi:10.1093/hmg/ddm127

10. Wang Y, McKay JD, Rafnar T, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet*. 2014;46(7):736-741. doi:10.1038/ng.3002

11. Ji X, Mukherjee S, Landi MT, et al. Protein-altering germline mutations implicate novel genes related to lung cancer development. *Nat Commun*. 2020;11(1):2220. doi:10.1038/s41467-020-15905-6

12. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet*. 2008;40(12):1404-1406. doi:10.1038/ng.254

13. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet*. 2017;49(7):1126-1132. doi:10.1038/ng.3892

14. Liu JZ, Erlich Y, Pickrell JK. Case-control association mapping by proxy using family history of disease. *Nat Genet*. 2017;49(3):325-331. doi:10.1038/ng.3766

15. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*. 2019;51(3):404-413. doi:10.1038/s41588-018-0311-9

16. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203-209. doi:10.1038/s41586-018-0579-z

17. Bulik-Sullivan BK, Loh P-R, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291-295. doi:10.1038/ng.3211

18. Liu M, Jiang Y, Wedow R, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. 2019;51(2):237-244. doi:10.1038/s41588-018-0307-5

19. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet*. 2014;10(5):e1004383. doi:10.1371/journal.pgen.1004383

20. Ellrott K, Bailey MH, Saksena G, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *cels*. 2018;6(3):271-281.e7. doi:10.1016/j.cels.2018.03.002

21. Benjamin D, Sato T, Cibulskis K, Getz G, Stewart C, Lichtenstein L. Calling Somatic SNVs and Indels with Mutect2. *bioRxiv*. Published online December 2, 2019. doi:10.1101/861054

22. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*. 2018;15(8):591-594. doi:10.1038/s41592-018-0051-x

23. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795

24. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019;8(7). doi:10.1093/gigascience/giz082

25. Bergstrom EN, Huang MN, Mahto U, et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics*. 2019;20(1):685. doi:10.1186/s12864-019-6041-2

26. Ashiqul Islam SM, Wu Y, Díaz-Gay M, et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cold Spring Harbor Laboratory*. Published online December 13, 2020:2020.12.13.422570. doi:10.1101/2020.12.13.422570

27. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet*. 2011;88(5):586-598. doi:10.1016/j.ajhg.2011.04.014

28. Sulem P, Gudbjartsson DF, Geller F, et al. Sequence variants at CYP1A1-CYP1A2 and AHR associate with coffee consumption. *Hum Mol Genet*. 2011;20(10):2071-2077. doi:10.1093/hmg/ddr086

29. Elsworth B, Lyon M, Alexander T, et al. The MRC IEU OpenGWAS data infrastructure. *bioRxiv*. Published online August 10, 2020. doi:10.1101/2020.08.10.244293

30. McGranahan TM, Patzlaff NE, Grady SR, Heinemann SF, Booker TK. A4β2 nicotinic acetylcholine receptors on dopaminergic neurons mediate nicotine reward and anxiety relief. *J Neurosci*. 2011;31(30):10891-10902. doi:10.1523/JNEUROSCI.0937-11.2011

31. Walsh RM Jr, Roh S-H, Gharpure A, Morales-Perez CL, Teng J, Hibbs RE. Structural principles of distinct assemblies of the human α4β2 nicotinic receptor. *Nature*. 2018;557(7704):261-265. doi:10.1038/s41586-018-0081-7

32. Gonzales D, Rennard SI, Nides M, et al. Varenicline, an alpha4beta2 nicotinic acetylcholine receptor partial agonist, vs sustained-release bupropion and placebo for smoking cessation: a randomized controlled trial. *JAMA*. 2006;296(1):47-55. doi:10.1001/jama.296.1.47

33. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet*. 2008;40(12):1407-1409. doi:10.1038/ng.273

34. Ferreiro-Iglesias A, Lesseur C, McKay J, et al. Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. *Nat Commun*. 2018;9(1):3927. doi:10.1038/s41467-018-05890-2

35. Broderick P, Wang Y, Vijayakrishnan J, et al. Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. *Cancer Res*. 2009;69(16):6633-6641. doi:10.1158/0008-5472.CAN-09-0680

36. Melin BS, Barnholtz-Sloan JS, Wrensch MR, et al. Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nat Genet*. 2017;49(5):789-794. doi:10.1038/ng.3823

37. Lesseur C, Ferreiro-Iglesias A, McKay JD, et al. Genome-wide association meta-analysis identifies pleiotropic risk loci for aerodigestive squamous cell cancers. *PLoS Genet*. 2021;17(3):e1009254. doi:10.1371/journal.pgen.1009254

38. Du Y, Grandis JR. Receptor-type protein tyrosine phosphatases in cancer. *Chin J Cancer*. 2015;34(2):61-69. doi:10.5732/cjc.014.10146

39. Motiwala T, Kutay H, Ghoshal K, et al. Protein tyrosine phosphatase receptor-type O (PTPRO) exhibits characteristics of a candidate tumor suppressor in human lung cancer. *Proc Natl Acad Sci U S A*. 2004;101(38):13844-13849. doi:10.1073/pnas.0405451101

40. Kachuri L, Graff RE, Smith-Byrne K, et al. Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. *Nat Commun*. 2020;11(1):6084. doi:10.1038/s41467-020-19600-4

41. Hung RJ, Warkentin MT, Brhane Y, et al. Assessing lung cancer absolute risk trajectory based on a polygenic risk model. *Cancer Res*. Published online January 20, 2021:canres.1237.2020. doi:10.1158/0008-5472.CAN-20-1237

Figure legends

Figure 1: Manhattan plot of the meta-analysis of genome-wide by proxy (GWAx) with genome-wide association study (GWAS) into lung cancer.
The Manhattan plot displays the results of the meta-analysis of the GWAx (48,843 proxy cases and 197,029 proxy controls without a family history of any cancer) and the GWAS (29,266 cases and 56,450 controls) with the new novel loci highlighted in black with the likely candidate gene name presented. This meta-analysis discovered 65 novel loci across 21 cytoband regions. The x-axis is the chromosome position across the autosomal chromosomes, with the Y-axis containing the association level displayed as the -log10(P-value), derived by a multivariate logistics regression model. The red dotted line displays the genome-wide significance threshold ($5x10^{-8}$)

Figure 2: Brain and lung eQTLs discovered within the eight novel loci
Co-localisation between lung cancer (x axis) and *CHRNB2* putamen expression (**A**), *CHRNA4* putamen expression (**B**), *CHEK1* lung expression (**C**) and *RP11-10O17.1* lung gene expression(**D**) (y axis). Each variant and eQTL status were compared using COLOC for colocalisation to confirm that the lung cancer SNP was the same SNP driving the eQTL effect in both brain and lung tissues, the Bayesian posterior probability (PP4) of each gene was tested, *CHRNB2* (PP4=98.67%), *CHNRA4* (96.48%), *CHEK1* (91.1%) and *RP11-10O17* (95.25%)

Figure 3: Germline polygenic risk score construction using smoking and eQTL related SNPs and performance testing within the UK Biobank lung cancer cohort.
(**A**) The mean lung cancer association statistics calculated by variant bins (100 variants per bin) ranked by component. Variants (clumped on LD based on lung cancer P values) were ranked based on PLS component for smoking propensity (Component1_smoking, top), and eQTLs (Component1_eQTL, bottom) (x axis) and plotted against the mean lung cancer Z statistics calculated across variants in each bin (y axis). Values that exceed 3 SDs from the mean are noted in red (NbinsSmoking =9, NbinsQTL = 37) and are those that have the highest values of the PLS component. (**B**) A Forest plot of the performance for the constructed PRS in comparison to just using the 65 genome-wide significant (GWS) independent loci as a baseline using the model LC ~ PRS + array + sex + array of recruitment + first 5 PCs. The top panel contains the smoking PRS and the eQTL PRS list without containing any of the 65 GW loci within each list. The middle panel contains the model with smoking status (previous, current, never) added. The bottom panel contains the full lists without adjusting for smoking status. The combined PRS contains all the 65 loci plus both the smoking and eQTL lists.

Figure 4: Polygenic risk scores for smoking (smPRS) associations with total number of mutations and mutations attributable to SBS4 in the TCGA cohort.
(A) Associations with total number of mutations. (**B**) Associations with SBS4 mutations. The left panels represent the distribution of the number of mutations in the sm-PRS quintiles. The right panels correspond, respectively, to the forest plots of sm-PRS associations with total mutational burden (panel A) and SBS4 mutations (panel B). For each PRS, the association was tested: i) in all lung cancer cases when considering all SNPs in the smPRS SNPs selection, ii) in all lung cancer cases when considering different subsets of SNPs in the PRS computation, iii) stratifying by histology, iv) stratifying by smoking status. Gray squares correspond to the estimate resulting from Quasi-Poisson models. The squares are highlighted in red when the associated P-value is below 0.05.

Supplementary Figures
Supplementary Figure 1: Genomic inflation and quantile–quantile plot across studies that were meta-analysed
Supplementary Figure 2: Visual validation of genome-wide by proxy method by Manhattan plot compared to the lung cancer genome-wide association study.
Supplementary Figure 3: Z-statistic plots for variants associated with traits at 15q24(*CYP1A1*) compared to lung cancer
Supplementary Figure 4: *CYPA1A* expression in the nucleus accumbens
Supplementary Figure 5: Partial least squares of mean z-scores for lung cancer for the polygenic risk scores construction and correlation across smoking traits and eQTLs
Supplementary Figure 6: The smPRS and eQTLPRS associations with mutational signatures related to smoking attributed to tobacco
Supplementary Figure 7: Mutational burden in lung tumours across rs72740955 genotype categories

Supplementary Tables within Supplementary Materials

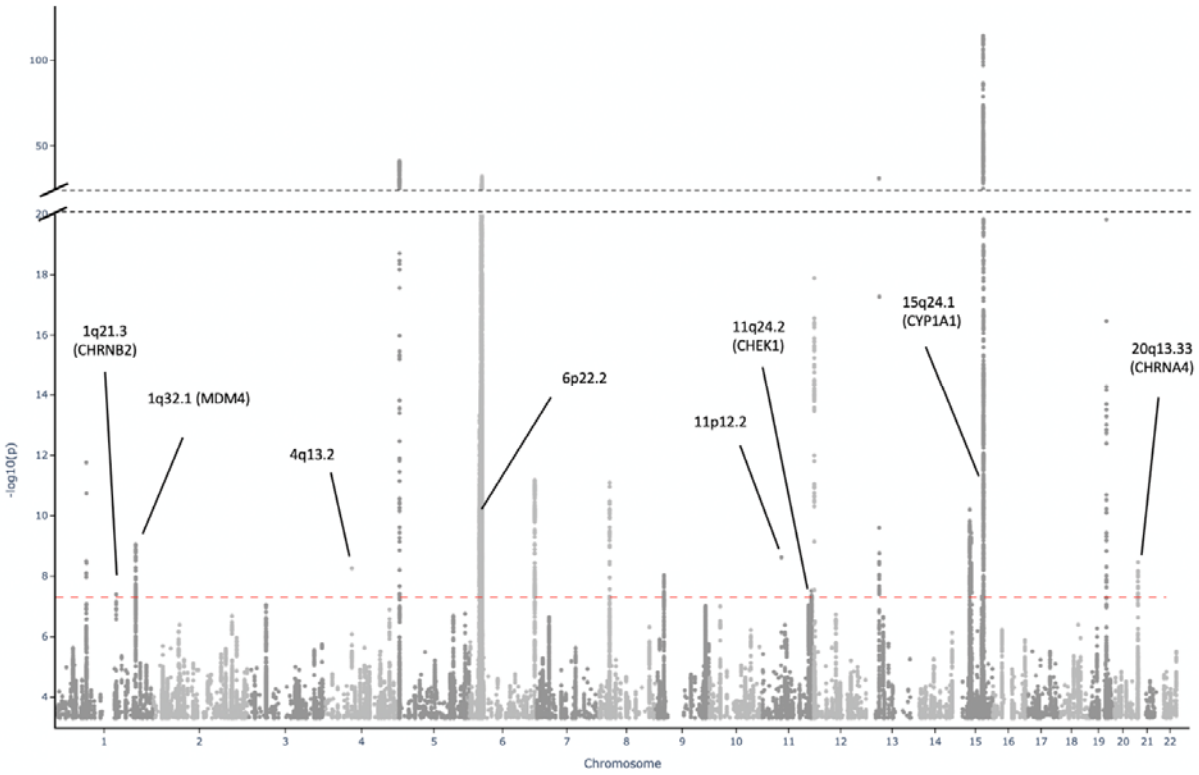

Figure 1: Manhattan plot of the meta-analysis of genome-wide by proxy (GWAx) with genome-wide association study (GWAS) into lung cancer.
The Manhattan plot displays the results of the meta-analysis of the GWAx (48,843 proxy cases and 197,029 proxy controls without a family history of any cancer) and the GWAS (29,266 cases and 56,450 controls) with the new novel loci highlighted in black with the likely candidate gene name presented. This meta-analysis discovered 65 novel loci across 21 cytoband regions. The x-axis is the chromosome position across the autosomal chromosomes, with the Y-axis containing the association level displayed as the -log10(P-value), derived by a multivariate logistics regression model. The red dotted line displays the genome-wide significance threshold ($5x10^{-8}$)

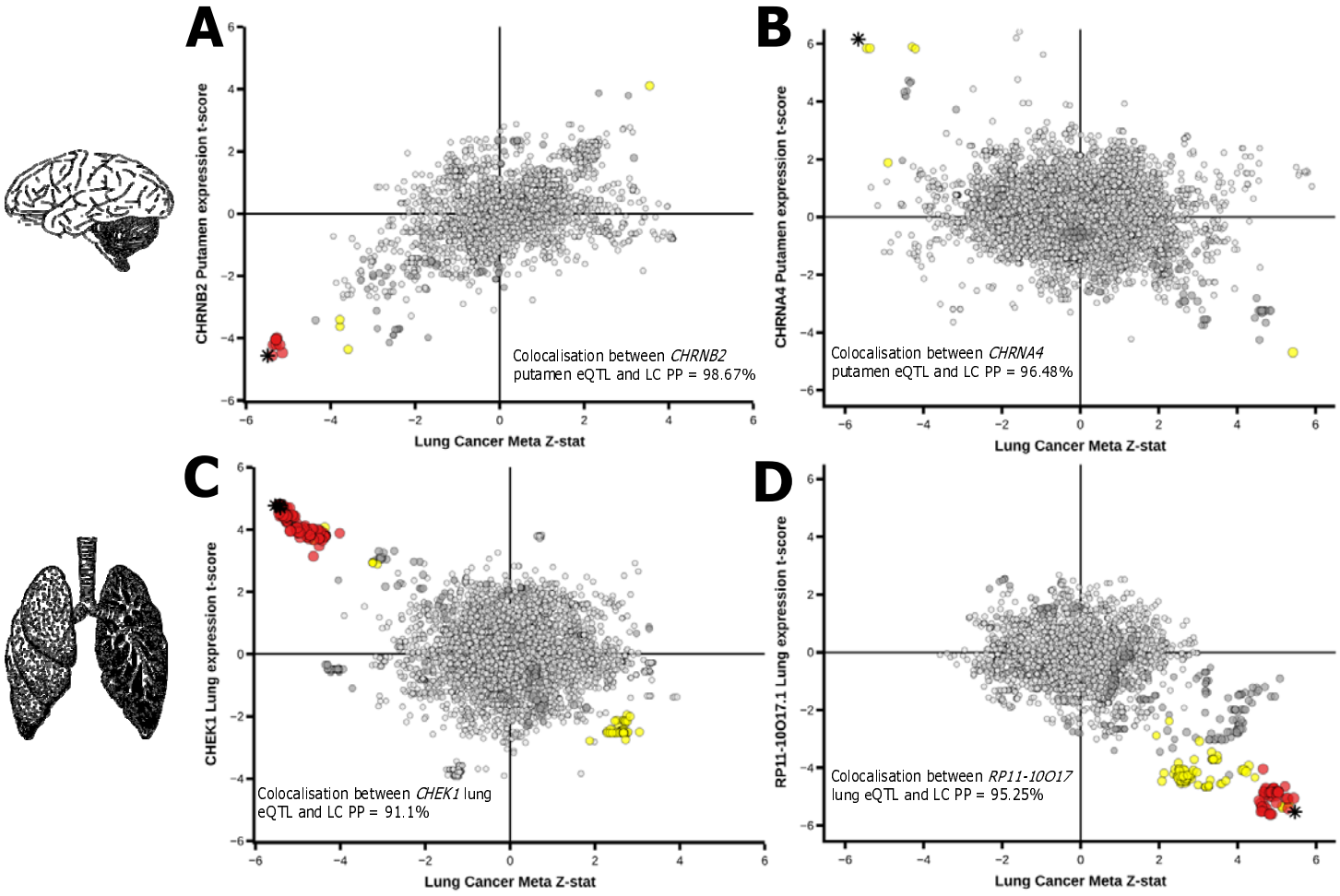

Figure 2: Brain and lung eQTLs discovered within the 8 novel loci
Co-localisation between lung cancer (x axis) and *CHRNB2* putamen expression (**A**), *CHRNA4* putamen expression (**B**), *CHEK1* lung expression (**C**) and RP11-10O17.1 lung gene expression (**D**) (y axis). Each variant and eQTL status were compared using COLOC for colocalisation to confirm that the lung cancer SNP was the same SNP driving the eQTL effect in both brain and lung tissues, the Bayesian posterior probability (PP) of each gene was tested, *CHRNB2* (PP=98.67%), *CHNRA4* (96.48%), *CHEK1* (91.1%) and *RP11-10O17* (95.25%)

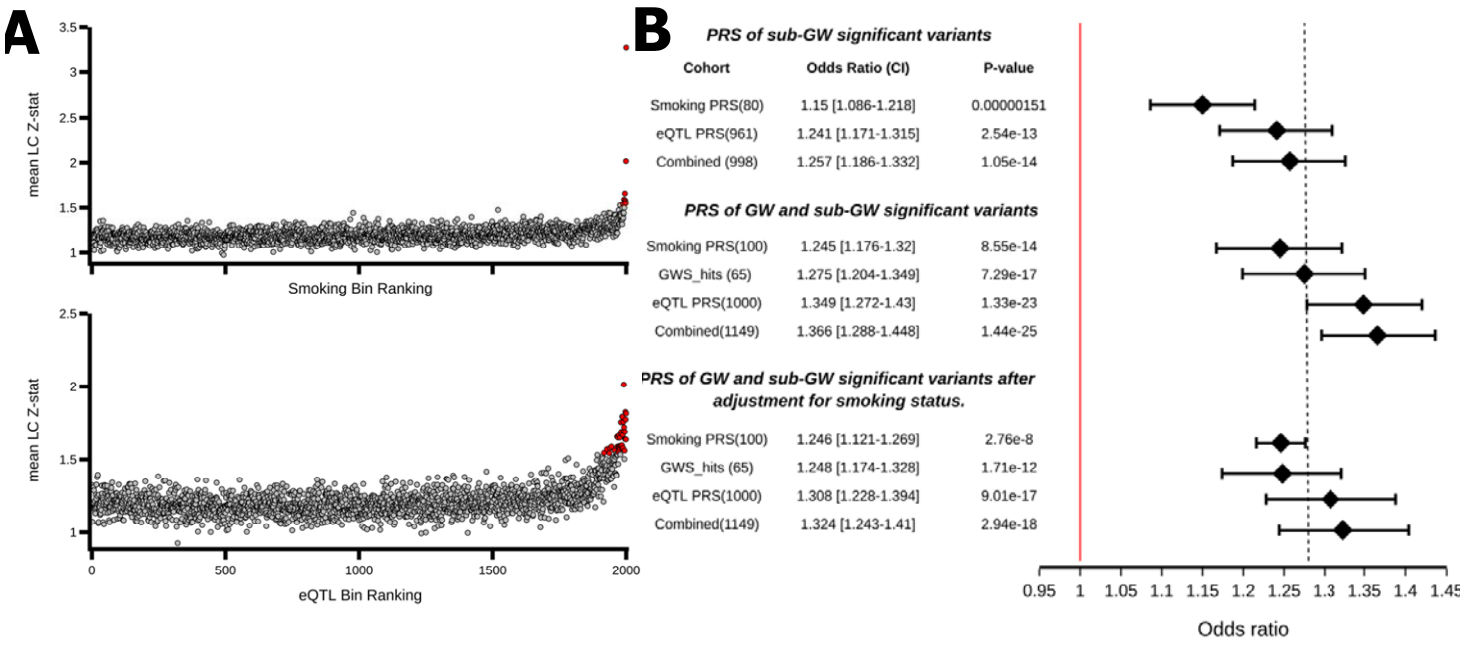

Figure 3: Germline polygenic risk score construction using smoking and eQTL related SNPs and performance testing within the UK Biobank lung cancer cohort.
(**A**) The mean lung cancer association statistics calculated by variant bins (100 variants per bin) ranked by component. Variants (clumped on LD based on lung cancer P values) were ranked based on PLS component for smoking propensity (Component1_smoking, top), and eQTLs (Component1_eQTL, bottom) (x axis) and plotted against the mean lung cancer Z statistics calculated across variants in each bin (y axis). Values that exceed 3 SDs from the mean are noted in red (NbinsSmoking = 9, NbinsQTL = 37) and are those that have the highest values of the PLS component. (**B**) A Forest plot of the performance for the constructed PRS in comparison to just using the 65 genome-wide significant (GWS) independent loci as a baseline using the model LC ~ PRS + array + sex + array of recruitment + first 5 PCs. The top panel contains the smoking PRS and the eQTL PRS list without containing any of the 65 GW loci within each list. The middle panel contains the model with smoking status (previous, current, never) added. The bottom panel contains the full lists without adjusting for smoking status. The combined PRS contains all the 65 loci plus both the smoking and eQTL list.

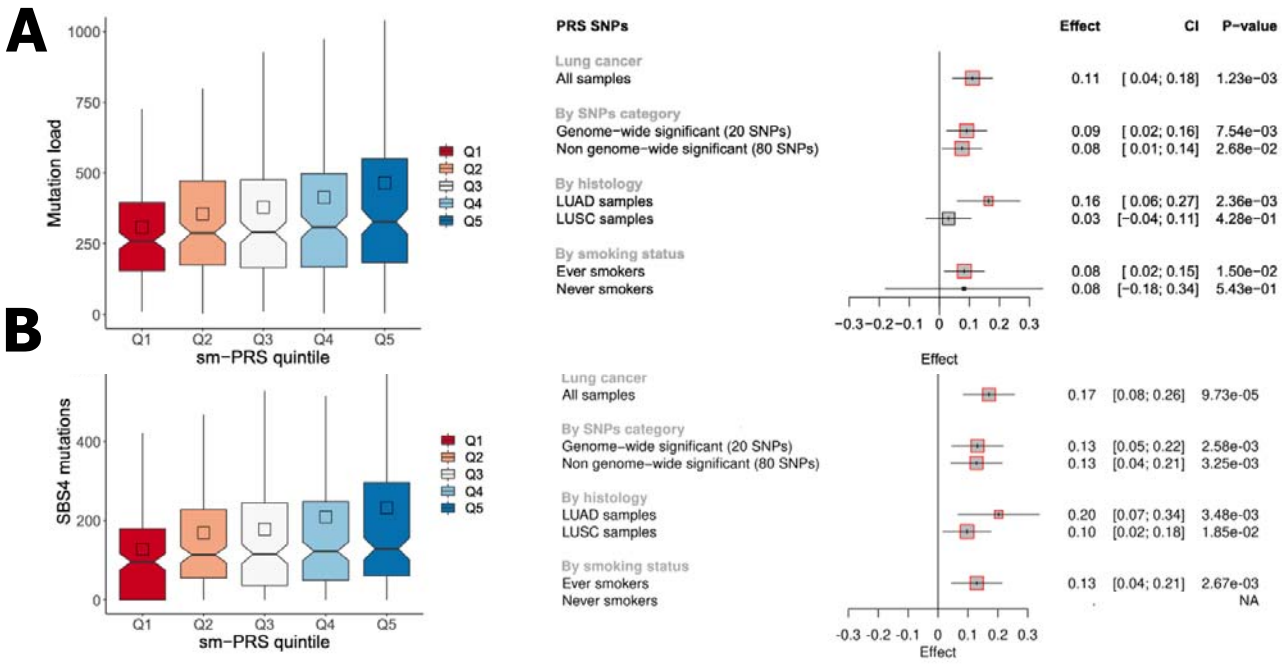

Figure 4: Polygenic risk scores for smoking (smPRS) associations with total number of mutations and mutations attributable to SBS4 in the TCGA cohort.

(**A**) Associations with total number of mutations. (**B**) Associations with SBS4 mutations. The left panels represent the distribution of the number of mutations in the sm-PRS quintile. The right panels correspond, respectively, to the forest plots of sm-PRS associations with total mutational burden (panel A) and SBS4 mutations (panel B). For each PRS, the association was tested: i) in all lung cancer cases when considering all SNPs in the sm-PRS SNPs selection, ii) in all lung cancer cases when considering different subsets of SNPs in the PRS computation, iii) stratifying by histology, iv) stratifying by smoking status. Gray squares correspond to the estimate resulting from Quasi-Poisson models. Those squares are highlighted in red when the associated P-value is below 0.05.

Table 1: The 8 novel genome-wide significant loci associated with lung cancer risk

Variant	Cytoband	chr:pos (hg19)	Ref	Alt	P-value	OR (L95%-U95%)	Likely targets (sentinel distance)	Adeno	Squam	Small cell
rs78062588	1q21.3	chr1:154566225	T	C	4.03E-08	0.904 [0.868-0.94]	ADAR(0kb), **CHRNB2**(+13.87kb)	1.15E-03	6.99E-06	1.73E-02
rs4252707	1q32.1	chr1:204508147	G	A	9.11E-10	0.931 [0.908-0.954]	MDM4(0kb)	1.42E-01	1.57E-03	-
rs7551222	1q32.1*	chr1:204599295	A	G	2.56E-08	1.057 [1.037-1.076]	MDM4	1.97E-01	1.44E-03	3.26E-02
rs185666783	4q13.2	chr4:67833774	G	C	5.56E-09	1.062 [1.042-1.083]	-	4.92E-05	3.34E-02	1.66E-04
rs7766641	6p22.2	chr6:26184102	G	A	7.05E-14	0.926 [0.906-0.946]	HIST1H2BE(0kb) – broad locus	4.79E-04	6.06E-08	2.98E-04
rs6913550	6p22.2*	chr6:26540683	C	T	4.82E-14	0.918 [0.896-0.94]	BTN1A1,HCG11,HMGN4	2.09E-02	2.26E-03	9.33E-05
rs72905558	11p11.2	chr11:48201643	A	T	2.41E-09	0.913 [0.88-0.94]	PTPRJ(+9.249kb)	1.50E-03	2.22E-01	2.51E-03
rs61612408	11q24.2	chr11:125495044	G	A	3.07E-08	0.903 [0.87-0.94]	**CHEK1**(0kb)	1.40E-01	1.15E-05	7.26E-02
rs12441817	15q24.1	chr15:75025814	T	C	4.77E-08	1.096 [1.06-1.13]	**CYP1A1**(+7.937kb), CYP1A2(-15.37kb)	2.19E-04	-	0.443
rs11697662	20q13.33	chr20:61992005	T	C	1.49E-08	1.071 [1.05-1.09]	**CHRNA4**(0kb)	6.42E-02	5.50E-06	3.40E-03
rs2281925**	20q13.33*	chr20:62376503	G	A	3.49E-09	1.091 [1.062-1.12]	RTEL1	3.05E-08	9.46E-01	5.34E-02

** 1q32.1, 6p22.2 and 20q13.33 contains two independent SNPs*

*** rs2281925 has been reported as genome-wide significant for adenocarcinoma, but not for overall lung cancer risk*

*Histological subtypes taken from McKay et al 2017, Adeno = Adenocarcinoma, Small = Small cell carcinoma, Squam = Squamous cell carcinoma, Variants that are in **bold** indicate that the SNP is also related to smoking propensity, Likely targets in **bold** are eQTLs for the given SNP in either the lung or brain tissues within GTEx*