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

Gabriel AAG *[1], Atkins JR *[1], Penha RCC[1], Smith-Byrne K[1], Gaborieau V[1], Voegele C[1], Abedi-Ardekani B[1], Milojevic M[1], Olaso R[2], Meyer V[2], Boland A[2], Deleuze JF[2], Zaridze D[3], Mukeriya A[3], Swiatkowska B[4], Janout V[5], Schejbalova M[6], Mates D[7], Stojsic J[8], Ognjanovic M[9], the ILCCO consortium, Witte JS[10], SRashkin SR[10,11], Kachuri L[10], Hung R[12], Kar S[13,14], Brennan P[1], Sertier A[15], Ferrari A[15], Viari A[15,16], Johansson M[1], Amos C[17], Foll M[1], McKay JD[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. Center for Applied Bioinformatics, St. Jude Children's Research Hospital, Memphis, TN, USA
12, Lunenfeld Tanenbaum Research Institute, Sinai Health System
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

Large international efforts are describing how germline variants influence susceptibility to lung cancer. We have undertaken a genome-wide association by proxy (GWAx) study of lung cancer in 48,843 proxy "cases" with a parent/sibling with lung cancer to 195,387 proxy controls without a family history of any cancer from the UK Biobank and meta-analysed the results with previously described GWA study results. 21 loci achieved genome-wide statistical significance, including 8 novel loci including expression quantitative trait loci (eQTLs) in DNA repair genes (*CHEK1*, *MDM4*) and metabolic genes (*CYP1A1*). This study also discovered loci associated with propensity to smoke, such as both subunits of a key element in nicotine response, the neuronal α4β2 nicotinic acetylcholine receptor. Polygenic risk scores (PRS) analysis of variants below genome-wide significant threshold in an independent lung cancer population demonstrated that variants related to eQTLs and/or smoking propensity are enriched for susceptibility variants. PRS of lung cancer variants related to propensity to smoke were associated with somatic mutation burden in matched tumours from the same patients, with individuals with higher polygenic genetic risk having increased mutation burden in two case cohorts. This study has expanded the number of susceptibility loci linked with lung cancer and provided insights into how the molecular mechanisms by which these susceptibility variants contribute to the development of lung cancer.

Tags: Lung Cancer, GWAx, GWAS, Smoking, Cancer, *CHRNA4*, *CHRNB2*, Mutational Burden, *CHEK1*, *MDM4*, *CYP1A1*, α4β2 nicotinic acetylcholine receptor, Mutational Signatures, *RP11-10O17.1*
Word count: 5480

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, only about 15% of people that smoke develop LC suggesting individual differences in susceptibility. A genetic basis for LC susceptibility was initially identified from the familial aggregation of LC after accounting for personal smoking habits (Ooi et al., 1986; Schwartz et al., 1996; Tokuhata and Lilienfeld, 1963) and segregation analyses (Sellers et al., 1990). Twin studies' have estimated an 18% excess familial risk of LC (Mucci et al., 2016). A subsequent large genome-wide association study (GWAS) in 29,266 cases and 56,450 controls identified 18 susceptibility loci (McKay et al., 2017) with heritability explained by common variation estimated at about 9% (Jiang et al., 2019). While traditional GWAS approaches continue to recruit patients and generate additional data, novel methodologies, such as genome-wide association by proxy (GWAx), can leverage existing data from large biorepositories by identifying proxy cases who have a 1st degree relative previously diagnosed with the given trait of interest (Liu et al., 2017). This method has proven successful in Alzheimer's disease, coronary artery disease and type 2 diabetes (Jansen et al., 2019; Liu et al., 2017).

In the current study, we undertook a family history GWAx of 48,843 proxy "cases" with a parent/sibling with lung cancer to 195,387 proxy controls without a family history of any cancer from the UK Biobank. We meta-analysed those results with the current largest GWAS in LC which has allowed for new loci to be identified. Furthermore, we constructed Polygenic Risk Scores (PRS) with variants related to LC, associated with propensity to smoke and gene expression characteristics. These PRS were then used to investigate the influence of these germline susceptibility on the somatic environment, including tumour mutation burden and mutation signatures that have been linked with tobacco smoking.

RESULTS

The 8 novel susceptibility loci

The meta-analysis between the GWAx and the GWAS was performed using METASOFT under a fixed effect model (Figure 1). Further information regarding the GWAx approach can be found in Extended Data 1. After clumping genetic variants (retaining variants with lowest P value of any given correlated pair when $R^2 > 0.1$, kb=10,000), 65 variants achieved a P-value of less than $5x10^{-8}$ across 21 distinct genomic loci defined by cytoband (Supplementary Table 2). At previously described lung cancer susceptibility loci, in addition to the common sentinel variant, we also identified 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) (Supplementary Table 2). Conversely, at 13q13.1, where a rare lung cancer susceptibility allele has previously been linked with lung cancer (rs11571833, K3326X *BRCA2*), we identified a common allele associated with lung cancer (rs11571734, MAF = 0.28).

Nine lung cancer susceptibility variants at 8 loci that had not previously been described at genome wide (GW) significance were identified. Of these, the lung cancer susceptibility variants at 1q21.3-rs78062588, 6p22.2-rs7766641 and 20q13.33- rs11697662 have been also associated at GW significance with traits related to propensity to smoke tobacco by the Sequencing Consortium of Alcohol and Nicotine use (GSCAN) (Supplementary Table 2). Variants at 20q13.33 have been previous linked lung adenocarcinoma (*RTEL1*) (McKay *et al.*, 2017), however, a novel variant rs11697662 located telomeric to these variants was observed. rs11697662 is not in LD to the *RTEL1* variants and associated evenly across different histological subtypes of LC, implying it is independent of the variants previously reported at this locus (Supplementary Table 2). The sentinel variants at 1q21.3-rs78062588 and 20q13.33-rs11697662 are reported to be eQTLs for the nicotinic acetylcholine receptors (nAChRs) subunits *CHRNB2* and *CHRNA4* by the GTEx consortium (Supplementary Table 2 and Supplementary Table 3). While rs78062588 is a *CHRNB2* eQTL exclusively to brain tissue (Cerebellum, Putamen, Cerebellar Hemisphere and Caudate), rs11697662 has a multi-tissue *CHRNA4* eQTL effect (Nucleus accumbens, Liver, Putamen, Testis and cervical c-1 spinal cord). Interestingly, both variants have eQTLs identified in the Putamen tissue. Colocalisation probability, as estimated by COLOC (Giambartolomei et al., 2014), between LC and brain tissues eQTLs was confirmed (Bayesian posterior probability (PP) test for colocalisation in Putamen, *CHRNB2*/LC PP = 98.67%, *CHNRA4*/LC PP = 96.48%) (Figure 2.A and B). At 6p22.2, we identified multiple susceptibility variants telomeric to the MHC region (Supplementary Table 2). These are typified by two sentinel variants, rs6913550 and rs7766641. These variants are ~7 Mbp from a previously described LC susceptibility loci, and curiously rs7766641 was also associated with propensity to smoke, whereas 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 clearly associated with propensity to smoke (Supplementary Table 2). The 11q24.2 sentinel variant rs61612408 was associated with the expression of the *CHEK1* gene in multiple tissues including lung epithelia (colocalization between *CHEK1* lung eQTL and LC PP = 91.1%), with the allele associated with increased expression correlating with decreased risk of lung cancer (Figure 2.C). The association with rs61612408 appeared more prominent in lung squamous cell carcinomas compared to other lung cancer histological subtypes (Table 1). At 15q24, the sentinel variant rs12441817 is located near the enzymic genes, *CYP1A1* and *CYP1A2*. The 15q24-*CYP1A1* locus has been associated with multiple traits, including coffee consumption (Sulem et al., 2011) and forced vital capacity (FEV), although there was evidence for colocalisation with the association with lung cancer only for the later (colocalization between coffee consumption and LC PP = 0.0003%, colocalization between FEV and LC PP = 97.05%)(Supplementary Figure 4). There was evidence that rs12441817 influenced *CYP1A1* expression in the nucleus accumbens (colocalisation PP=70.26%) (Supplementary Figure 5), we also report an eQTL effect of rs12441817 with the processed pseudogene *RP11-10O17.1* in lung tissue (colocalization between eQTL_ *P11-10O17.1* and LC PP = 95.25%) (Figure 2.D). At 1q32.1, we identified rs4252707, an intronic variant within the *MDM4* gene (also known as *MDMX*) (Table 1). 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* (lung) reported in GTEx but no evidence for colocalization (*C1QTNF4* PP = 0.06%) was observed implying this may be an independent signal.

Exploration of sub-genome-wide significant variants and integrative multi-trait polygenic risk score construction.
As noted above, many of the variants that achieved GW significance also tended to be associated with smoking related traits and/or eQTL's, and often multiple of these correlated traits (Supplementary Table 2.). To explore sub-GW significant variants with similar characteristics, we obtained GSCAN association statistics for smoking-related traits (age of first cigarette, cigarettes per a day, smoking cessation and smoking initiation. Taken from Liu et al., 2019 (see Methods) and, separately, eQTLs across 49 different tissues from GTEx for 7,562,169 variants from the GWAS-GWAx. We used partial least squares regression (PLS) to identify components within these correlated association statistics related to lung cancer association statistics. The first component explained 1.13% and 0.53% of variance in lung cancer association statistics, respectively, and tended to be related to smoking traits and multi-tissue eQTL's and (Supplementary Figure 6). After LD pruning on lung cancer (as described above), we subsequently ranked all variants by their loading on the first PLS component, aggregated them in bins of 100 variants and plotted these ranked bins by the bin-specific mean Z stat for the lung cancer association (Figure 3.A). Variant bins from the highest values of PLS components tended to have elevated mean association statistics relative to other variants, implying that these variants are enriched for association with lung cancer, more so for bins of eQTLs than smoking propensity (Figure 3.A bottom). We developed polygenic risk scores (smPRS and eQTLPRS) of sub-GW significant variants associated with lung cancer and ranked highly by the PLS (see Methods) and tested them in an independent 1,666 lung cancer cases and 6,664 matched controls from the UK Biobank not included in the GWAx described above. The smPRS and eQTLPRS were significantly associated with lung cancer risk (smPRS p= $7x10^{-5}$, eQTLPRS p= $2x10^{-8}$), consistent with the notion that these variants are enriched for lung cancer susceptibility variants (Figure 3B). Sub-genome-wide significant variants were located near candidate genes that appear to similar biological functions as those noted above, for example additional nicotinic acetylcholine receptors (nAChRs) subunits (*CHNRA6)* and Dopamine beta-hydroxylase (*DBH)* related to smoking propensity (Supplementary Table 5) and DNA repair genes like *ERCC2*, *RAD51C*, *XRCC3* and *CASP8* in the eQTLs (Supplementary Table 6).

When combining with the GW significant variants from the GWAx + GWAS meta-analysis (and pruning for LD), the PRS was robustly associated with lung cancer in this independent series of lung cancer cases and controls (smPRS: OR per standard deviation = 1.216, 95% CI: 1.15-1.29, P = $2.4 x10^{-11}$; eQTLPRS: OR = 1.277, 95% CI: 1.2-1.35, P = $1.45 x 10^{-16}$, combined OR = 1.304, 95% CI 1.23-1.38, P = $3.95 x10^{-19}$), which was only modestly attenuated when adjusting for smoking status (Figure 3B).

PRS germline influences on mutational burden and mutational signatures
Next, we projected the smPRS and eQTLPRS (GW and sub GWAS variants combined) into the TCGA cohort of 8,346 tumours. Both PRS' were associated with lung cancer (N=791) when comparing with other forms of cancer (N=7,555) (smPRS: OR = 1.185, P = $1.12 x10^{-5}$; eQTLPRS: OR = 1.263, P = $2.24 x10^{-9}$; combined: OR = 1.256, P = $3.90 x10^{-9}$). This association was more marked when excluding other tobacco related cancers from the comparison group from the TCGA dataset (Supplementary Table 4).

We then evaluated the association between the two PRS and somatic mutational burden in the 736 lung cancer patients from the TCGA where somatic and germline data overlapped and passed QC metrics (see Methods). While there was little evidence for association involving the eQTLPRS (Supplementary Figure 9), we found that the smPRS was associated with tumour mutational burden (TMB) ($P = 1.23x10^{-3}$, 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 ($P = 9.73 \times 10^{-5}$), ID3A ($P = 1.78 \times 10^{-3}$), ID3B ($P = 3.77 \times 10^{-2}$) and DBS2 ($P = 3.05 \times 10^{-3}$)) (Figure 4.A and Supplementary Figure 7). This association was observed more prominently in LUAD samples (Figure 4.A). Of the individual variants, the 15q25 *CHRNA5* lung cancer sentinel variant, rs72740955, had the most striking effect (Supplementary Table 3 and Supplementary Figure 9), nevertheless, the associations remained significant after genome-wide variants for lung cancer were excluded from the smPRS (Figure 4.A). We sought to validate the association between the smPRS and somatic mutation burden within an independent cohort of 61 lung cancer patients whose germline and matched tumour samples have undergone whole genome sequencing (GENILUC cohort). The association between smPRS polygenic load and increased somatic mutation burden was replicated ($P = 0.034$) and similarly for the association with mutation signatures attributed to tobacco smoking (SBS4: $P = 0.023$, ID3: $p = 0.054$, DBS2: $P = 0.035$, Supplementary Figure 9). 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 7).

Discussion

This study has leveraged large, genotyped biobank data and GWAS studies to identify novel susceptibility loci for lung cancer. The similarity in the general genetic architecture from the family history GWAx and traditional GWAS approach supported the combining of these approaches (see Extended Data 1) and the meta-analysis identifying 21 loci that achieved genome wide significance, including eight loci that have not been described at GW significance.

Three of eight novel loci were also associated with propensity to smoke and included brain eQTLs variants for both the α4 and β2 subunits of the neuronal nAChRs α4β2 receptor. Variants in LD with the former (rs2373500) have been previously described in nicotine dependency and lung cancer risk, albeit not at genome wide significance for lung cancer (Liu et al., 2019), 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. This receptor is important within the dopaminergic signalling pathway where these α4β2 receptors have a key role in nicotine dependence behaviours and are also the principal target in nicotine addiction intervention (Gonzales et al., 2006; Walsh et al., 2018) . The deletion of the α4β2 nACHRs within dopaminergic neurons results in changes in nicotine-related behaviours in animal studies (McGranahan et al., 2011) . Both variants have expression effects in the putamen, notable as functional magnetic resonance imaging studies observe greater activation of the putamen region in people that smoke when processing smoking-related cues (Lin et al., 2020) . Expression changes in these nAChRs subunits may modulate the reward mechanisms in variants carriers and lead to changes in smoking propensity and lung cancer risk between individuals.

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 (Broderick et al., 2009; Ferreiro-Iglesias et al., 2018; Wang et al., 2008). However, the variants described here, rs7766641, is not in LD with these previously described variants ($R^2 < 0.001$) and rs7766641 is strongly 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. At 15q24, we identified rs12441817 located nearby genes *CYP1A1*, *CYP1A2* and *CYP11A1*. The cytochrome *CYP1A1* is related to P450 enzymes, which primary function is to participate in the metabolism of many different xenobiotics and some endogenous substrates. rs12441817 is located centromeric (~3MB) to the well described lung cancer susceptibility locus at *CHNRA3-5*. However, there is very low linkage disequilibrium between these loci and rs12441817 is not associated with any smoking trait, implying that the rs12441817 association is statistically and biologically independent to that of the 15q25 locus. Variants at the 15q24 *CYP1A1* / *CYP1A2* locus has been linked with multiple traits, notably other forms of propensity (coffee and alcohol consumption) and forced vital capacity (FEV), although colocalization appears to implicate the later as more likely to be involved in this association. 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 *CHEK1* gene (Figure 2.C). This association between rs61612408 and lung cancer appears more prominent in lung squamous cell carcinomas, similar to previously described variants associated near DNA repair genes *CHEK2* and *BRCA2* (Table 1)(Brennan et al., 2007; Wang et al., 2014). We additionally noted the variant impacting *MDM4* gene, an important p53 regulator. This variant was previously associated with non-glioblastoma tumours (Melin et al., 2017) and most recently squamous cell carcinomas of the lung and head and neck (Lesseur, 2020), although here we noted weak evidence for association in lung adenocarcinoma (Table 1). Additionally, our finding at 11p11.2 was reported in GTEx as an eQTL for C1QTNF4 (lung) and MTCH2 (brain-cortex) but colocalization analysis showed little evidence that these signals were the same. Interestingly, the nearest gene PTPRO, is a Tyrosine Phosphatase receptor, which hypermethylation of promotor has been observed across several cancers including lung (Du and Grandis, 2015; Motiwala et al., 2004). At 4q13.2, the finding remains ambiguous, but from histological subtypes analysis performed from the previous reported GWAS study, it appears this signal is mostly in lung adenocarcinoma.

We additionally sought to use the shared genetic aetiology between lung cancer susceptibility and smoking related traits and functional annotations (eQTL) to explore variants that did not achieve genome-wide significant. We used the partial least squared (PLS) method to select variants related to these traits and PRS analysis demonstrated that such variants are indeed enriched for susceptibility alleles. The sub-genome-wide significant variants included in the PRS were located near relevant candidate genes like *CHNRA6* and *DBH* (Supplementary Table 5) and eQTLs for *ERCC2*, *RAD51C*, *XRCC3* and *CASP8* (Supplementary Table 6), although the role of these individual variants remains to be confirmed. It is also notable that when combining both sub-genome-wide PRS lists (smoking PRS and

eQTL PRS) with GW significant results, the combined score for lung cancer is now an OR 1.304 per a SD increase in score. This is an improvement on previous PRS predictions (OR 1.17 and 1.26) (Hung et al., 2021; Kachuri et al., 2020), despite the conservative clumping approach ($R2 < 0.1$) employed. This suggests that integrating functional annotations may be of interest for PRS, particularly in rarer traits where assembling very large sample sizes may be problematic.

Lastly, the analysis with the smoking PRS demonstrated an association between a person's genetic polygenic risk load and mutation burden. This association was observed 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, which is observable as mutation burden or burden of somatic mutational signatures related to tobacco consumption in the carriers' tumours.

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 also provided insights into the carcinogenic mechanisms.

METHODS

UK Biobank, Genome-wide association by proxy analysis and genetic correlations analysis.

The UK Biobank resource has been described by Bycroft et al., 2018 and was accessed under project number 15825. Summary statistics were obtained from the GWAS which has been previously published elsewhere (McKay et al., 2017). Sample selection are detailed in Supplementary Table 1 along with the UK biobank data ID fields. Association testing was performed using a logistic regression model using the --glm function in PLINK 2.0 on white British ancestry. Each model was adjusted by age at recruitment, sex, and the first 10 PCs and array type (50,000 samples produced on the UK BiLEVE array; the rest of samples were performed on the UK Biobank Axiom array). The GWAx method has previously been described by Liu *et al* (Liu et al., 2017). We performed a traditional GWAS on the family history status of LC and then adjusted the betas and standard error. Individuals that self-reported more than one affected first-degree relative (for example, sibling and a parent) were only included once. Individuals diagnosed with LC were extracted from the UK Biobank and formed an independent LC cohort which consisted of 1,666 LC cases (other cases were excluded by failure to meet QC) and 6,664 controls. If a LC case also reported a family history of lung cancer, they were excluded from the GWAx analysis. For the genetic correlation analysis, the LDSC package was used (Bulik-Sullivan et al., 2015a, 2015b) . This was conducted on the defined HapMap SNPs as described in the LDSC documentation. Each summary statistic file from the GSCAN GWAS study (UK Biobank removed) along with the meta-analysis and GWAS were converted to the sumstats format using the munge_sumstats.py. Genetic correlation was performed across each trait using the ldsc.py script. Both the family history GWAx and GWAS were meta-analysed. METASOFT was used to perform this meta-analysis using a fixed-effects model based on an inverse-variance-weighted effect size (Han and Eskin, 2011). The summary effect size estimates, the standard errors and p-values were taken for further analysis. Manhattan plots, colocalisation plots, bin plots and the forest plot were generated with Plotly. LD clumping was undertaken using PLINK ($R^2 < 0.1$ and 10,000 kb). The eQTL analysis was performed using GTEx version 8 data, with all pairs data downloaded and processed. To quantify the colocalisation, we calculated the Bayesian posterior probability for colocalization of two datasets for the H_4 (association with trait 1 and trait 2, one shared SNP), by firstly calculating the log bayes factor for each SNP in each dataset, then the posterior probability was calculated by the COLOC package in python (https://github.com/anthony-aylward/coloc). GWAS summary statistics for coffee intake and forced vital capacity were from taken the Ben Neale UK Biobank work (http://www.nealelab.is/uk-biobank/).

Polygenic Risk Score analysis

While the lung cancer meta-analysis revealed multiple significant lung cancer-associated loci, we supplemented these with relevant SNPs that did not pass the genome-wide significance threshold for further PRS analyses. In this study, we chose to select additional SNPs based on their association with smoking related traits and their effect on genes expression levels, respectively. Firstly, summary statistics of previously published GWAS on smoking related traits have been gathered. We chose the traits studied by the GWAS Sequencing Consortium of Alcohol and Nicotine (GSCAN) consortium in a large dataset gathering up to 1.2 million samples (Liu et al., 2019). This study explored four smoking traits (cigarettes per day, smoking cessation, smoking initiation and age of initiation) as well as drinking consumption (drinks per week). It has also been shown by Jiang *et al*. that there is a shared heritability, probably mostly driven by smoking, between lung cancer and head and neck cancer (Jiang et al., 2019) . Therefore, the summary statistics of the GWAS on head and neck cancer performed by Lesseur *et al.* was also considered (Lesseur et al., 2016). In order to identify SNPs associated with both lung cancer and smoking-related traits, the partial least square (PLS) model was used considering the Z scores for all the meta-analysis SNPs and for each smoking-related trait as explanatory variables and the lung cancer Z scores as the response variable (smoking PLS). Secondly, eQTLs summary statistics obtained from the GTEx data version 8 were used to perform a second PLS analysis (eQTL PLS). For each SNP in the meta-analysis, the eQTL T-values (the top value for each gene) for each tissue were considered as explanatory variables and the lung cancer Z scores as the response variable. Each first principal component of these two PLS analyses were used to rank the lung cancer GWAS SNPs. After LD clumping of genome wide lung cancer association statistics ($R^2 < 0.1$ and 10000 kb as per above), all SNPs having a high lung cancer Z-statistic (above the 99^{th} percentile) were then ranked based on the PLS component. The degree of enrichment for lung cancer related SNPs varied between the smoking and eQTL PLS, hence the top 100 and 1000 for the sm-PRS and eQTL-PRS were selected, respectively. PRS were constructed and weighted using association statistics from the lung cancer GWAx meta-analysis (see equation below) and projected into independent lung cancer cohorts.

$$\sum_{i=1}^{n} \beta * Xi$$

with Xi being the individual's SNP genotypes, and β the effect from the lung cancer meta-analysis, n = 100 for the smPRS and n = 1000 for the eQTLPRS.

Genotyping and imputation of TCGA data

TCGA access was obtained through TCGA project 2731 via the database of Genotypes and Phenotypes (dbGaP). The TCGA raw intensities CEL files associated with blood and/or normal tissue samples were downloaded from the GDC Legacy Archive portal using gdc-client. Prior to genotyping, a quality control function (apt-geno-qc) from Affymetrix Power Tools (APT 2.10.2.2) was applied to the dataset and samples with a contrast QC above 0.4 were selected for genotyping. The largest TCGA cohort, BRCA, was genotyped to define a list of probes with good performance (for example, probe-level genotyping call rates of at least 97%). This probe list was then used with the apt-probeset-genotype program and Birdseed (v2) algorithm to genotype the TCGA samples. Samples with low genotyping call rate of < 97% were removed and a second round of genotyping was performed on each cohort. Sex check (using *plink*) was performed and compared to the clinical data. Sample's relatedness (using *plink genome* with min parameter fixed at 0.185) was also tested. Among the 9,855 samples with reported sex, 46 samples had discordance between reported and imputed sex and 17 pairs of relatives were identified. These samples were not considered in the PRS analyses. Heterozygosity and genotyping missing rates *(--het* and *--missing* options respectively) were also checked in order to select samples with a genotyping missing rate lower than 3% and to ensure that the heterozygosity rate was homogeneous across samples.

Population stratification was performed using admixture (version 1.3) (Alexander et al., 2009) based on HapMap2 data. SNPs were defined by Yu et al., 2008. Among 12,898 SNPs, 11,630 were in common between the HapMap and the TCGA datasets. Around 99% of the samples reported as "WHITE" by the TCGA clinical data were predicted Europeans. Eigenstrat (without outlier removal) was used on the European samples to account for further population stratification; the first 5 principal components (PCs) were kept for downstream analyses.

Considering all TCGA samples, SNPs with genotyping call rate and MAF lower than 97% and 1% respectively were removed. In each ancestry group, we applied the *HRC-1000G-check-bim.pl* script from the McCarthy tools (https://www.well.ox.ac.uk/%7Ewrayner/tools/#Checking), removing SNPs with unmatched positions and/or alleles, duplicated SNPs, as well as SNPs with allele frequencies (AF) differing from reported AF in the same population in the 1000 Genome dataset (more than 20% difference). SNPs with a genotyping call rate below 97% and showing strong deviation (P-value < 10^{-8}) from the Hardy Weinberg equilibrium (*hwe* plink option) in any of the ancestry groups were excluded. Finally, ambiguous SNPs were not considered.

Phasing and imputation were performed on each chromosome by 20Mb chunks. Phasing was performed using Eagle (v2.4.1, flanking regions of 5Mb) (Das et al., 2016; Loh et al., 2016) with the 1000 Genomes phase 3 reference panel. Minimac4 (v1.0.1) (Das et al., 2016) was used to perform the imputation with a window size of 500kb.

Mutation burden and mutational signatures computation

The somatic mutations of the TCGA samples were retrieved from the study of Ellrott et al., 2018 This study curated and reported the mutations called by at least two variant callers out of seven distinct variant callers. In order to remove duplicated samples per patient, the variants flagged "nonpreferedpair" were removed. Samples who were whole genome amplified or flagged "gapfiller" were excluded from the analyses. The total number of mutations for each sample was computed based on the set of filtered variants. Finally, the proportion of artifacts or germline variants were computed to identify and remove potential low-quality samples (proportion higher than 10%).

Replication of mutational signatures within the GeniLuc cohort

61 lung cancer cases were identified from central and eastern Europe as described previously (McKay et al., 2017) (Supplementary Note). Recruitment included collection of normal material in the form of a blood sample and a resection of the patient's tumour. Subsequent to histopathological review, to ensure appropriate tumour purity, DNA was extracted from normal material (blood) and the tumour resection. Whole genome sequencing was undertaken using PCR free whole genome library preparation and DNA sequencing undertaken to a depth of 30X for the paired tumour normal for each patient using an Illumina HiSeq X 5 DNA sequencer at the CEA laboratory in Paris France. Raw sequencing data was processed by inhouse nextflow pipelines (https://github.com/IARCbioinfo). Somatic mutations were defined using Mutect2 (Benjamin et al., 2019) and germline calls using Strelka2 (Kim et al., 2018). Reads were aligned to hg38, as such, SNPs in the PRS panels were lifted over to hg38. For normal tissue, PRS SNPs

were extracted from VCF files and put in the PLINK BED format (Purcell et al., 2007) . The PRS score generation was performed using PRsice2 (Choi and O'Reilly, 2019) from the normal calls. For somatic tissue processing, mutational signatures were computed as described below.

In order to compute mutational signatures, mutational matrices for each mutation type (SBS, DBS, and ID) were generated using SigProfilerMatrixGenerator (v1.1.20) (Bergstrom et al., 2019) with default parameters. Mutational signatures were then extracted with SigProfilerExtractor (v1.0.17) (Ashiqul Islam et al., 2020) from the TCGA-WES (LUAD and LUSC) samples and GENILUC-WGS lung cancer cohorts, separately, using the default options. 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. Of those, we focused on the previously reported smoking tobacco-related signatures, SBS4, DBS2, and ID3 (ID83A and ID83B), and the absolute mutation counts for each COSMIC signature in each sample was assessed.

The smoking-related SNPs and the eQTLs SNPs were considered respectively to compute PRS and test their association, in the TCGA samples, with the total number of mutations in the tumours as well as with tobacco-related signatures previously mentioned. Quasi-Poisson regressions were used when considering the total number of mutations and tobacco-related signatures. Covariates included age, gender and the 5 first principal components resulting from Eigenstrat. The samples purity variable, available in the panCancer Atlas supplemental data, was transformed in a categorical variable (purity less than or equal to 30%, purity between 30 and 70% included and purity above 70%) and was added to the covariates. When the LUAD and LUSC cohorts were both considered, a categorical variable indicating the cohort was included in the model. Covariates that were used in the germline PRS UK Biobank analysis included sex, array type, age of recruitment and the first 10 principal components.

Main body of text Figure legends descriptions

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 (**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 10 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 of 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 (panel A) 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.*

Supplementary Figures
Supplementary Figure 1: Meta-analysis of genome-wide by proxy (GWAx) with genome-wide association study (GWAS) into lung cancer
Supplementary Figure 2: Genomic inflation and quantile–quantile plot across studies that were meta-analysed
Supplementary Figure 3: Visual validation of genome-wide by proxy method by Manhattan plot compared to the lung cancer genome-wide association study.
Supplementary Figure 4: Z-statistic plots for variants associated with traits at 15q24(CYP1A1) compared to lung cancer
Supplementary Figure 5: CYPA1A expression in the nucleus accumbens
Supplementary Figure 6: 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 7: The smPRS and eQTLPRS associations with mutational signatures related to smoking attributed to tobacco
Supplementary Figure 8: Mutational burden in lung tumours across rs72740955 genotype categories

Supplementary Tables

Acknowledgements and 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. 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. 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)

Materials availability

No new unique reagents were generated in this study

Data and Code Availability

The code generated during this study are available at github (https://github.com/IARC-genetics/GWAx_lung_cancer). Polygenic risk scores used in this study are available within the supplementary tables. Summary statistics from the UK Biobank lung cancer family history summary statistics will be 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

Experimental model and subject details

As this study was a meta-analysis the datasets have been previously published with consent and ethics statements (Mckay et al 2017, Bycroft et al 2018). In the replication cohort for the somatic PRS, the GeniLuc cohort has not previously been publish. Ethnics for the GeniLuc cohort was reviewed by the IARC ethnics board (IARC IRB 12-05) and approved on 28th of April 2016.

Competing Interests

The authors have no competing interests in regard to the present study

REFERENCES

Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Res. *19*, 1655–1664.

Ashiqul Islam, S.M., Wu, Y., Díaz-Gay, M., Bergstrom, E.N., He, Y., Barnes, M., Vella, M., Wang, J., Teague, J.W., Clapham, P., et al. (2020). Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor.

Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., and Lichtenstein, L. (2019). Calling Somatic SNVs and Indels with Mutect2 (bioRxiv).

Bergstrom, E.N., Huang, M.N., Mahto, U., Barnes, M., Stratton, M.R., Rozen, S.G., and Alexandrov, L.B. (2019). SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. BMC Genomics *20*, 685.

Brennan, P., McKay, J., Moore, L., Zaridze, D., Mukeria, A., Szeszenia-Dabrowska, N., Lissowska, J., Rudnai, P., Fabianova, E., Mates, D., et al. (2007). Uncommon CHEK2 mis-sense variant and reduced risk of tobacco-related cancers: case control study. Hum. Mol. Genet. *16*, 1794–1801.

Broderick, P., Wang, Y., Vijayakrishnan, J., Matakidou, A., Spitz, M.R., Eisen, T., Amos, C.I., and Houlston, R.S. (2009). Deciphering the impact of common genetic variation on lung cancer risk: a genome-wide association study. Cancer Res. *69*, 6633–6641.

Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., ReproGen Consortium, Psychiatric Genomics Consortium, Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Case Control Consortium 3, Duncan, L., et al. (2015a). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241.

Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015b). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature *562*, 203–209.

Choi, S.W., and O'Reilly, P.F. (2019). PRSice-2: Polygenic Risk Score software for biobank-scale data. Gigascience *8*.

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284.

Du, Y., and Grandis, J.R. (2015). Receptor-type protein tyrosine phosphatases in cancer. Chin. J. Cancer *34*, 61–69.

Ellrott, K., Bailey, M.H., Saksena, G., Covington, K.R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K.E., McLellan, M., et al. (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. Cels *6*, 271-281.e7.

Ferreiro-Iglesias, A., Lesseur, C., McKay, J., Hung, R.J., Han, Y., Zong, X., Christiani, D., Johansson, M., Xiao, X., Li, Y., et al. (2018). Fine mapping of MHC region in lung cancer highlights independent susceptibility loci by ethnicity. Nat. Commun. *9*, 3927.

Giambartolomei, C., Vukcevic, D., Schadt, E.E., Franke, L., Hingorani, A.D., Wallace, C., and Plagnol, V. (2014). Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. *10*, e1004383.

Gonzales, D., Rennard, S.I., Nides, M., Oncken, C., Azoulay, S., Billing, C.B., Watsky, E.J., Gong, J., Williams, K.E., Reeves, K.R., et al. (2006). Varenicline, an alpha4beta2 nicotinic acetylcholine receptor partial agonist, vs sustained-release bupropion and placebo for smoking cessation: a randomized controlled trial. JAMA *296*, 47–55.

Han, B., and Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am. J. Hum. Genet. *88*, 586–598.

Hung, R.J., Warkentin, M.T., Brhane, Y., Chatterjee, N., Christiani, D.C., Landi, M.T., Caporaso, N.E., Liu, G., Johansson, M., Albanes, D., et al. (2021). Assessing lung cancer absolute risk trajectory based on a polygenic risk model. Cancer Res. canres.1237.2020.

Jansen, I.E., Savage, J.E., Watanabe, K., Bryois, J., Williams, D.M., Steinberg, S., Sealock, J., Karlsson, I.K., Hägg, S., Athanasiu, L., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. Nat. Genet. *51*, 404–413.

Jiang, X., Finucane, H.K., Schumacher, F.R., Schmit, S.L., Tyrer, J.P., Han, Y., Michailidou, K., Lesseur, C., Kuchenbaecker, K.B., Dennis, J., et al. (2019). Shared heritability and functional enrichment across six solid cancers. Nat. Commun. *10*, 431.

Kachuri, L., Graff, R.E., Smith-Byrne, K., Meyers, T.J., Rashkin, S.R., Ziv, E., Witte, J.S., and Johansson, M. (2020). Pan-cancer analysis demonstrates that integrating polygenic risk scores with modifiable risk factors improves risk prediction. Nat. Commun. *11*, 6084.

Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., et al. (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods *15*, 591–594.

Lesseur, C. et al (2020). Genome-wide association meta-analysis identifies pleiotropic risk loci foraerodigestive squamous cell cancers. PLoS Genet(in Press).

Lesseur, C., Diergaarde, B., Olshan, A.F., Wünsch-Filho, V., Ness, A.R., Liu, G., Lacko, M., Eluf-Neto, J., Franceschi, S., Lagiou, P., et al. (2016). Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer. Nat. Genet. *48*, 1544–1550.

Lin, X., Deng, J., Shi, L., Wang, Q., Li, P., Li, H., Liu, J., Que, J., Chang, S., Bao, Y., et al. (2020). Neural substrates of smoking and reward cue reactivity in smokers: a meta-analysis of fMRI studies. Transl. Psychiatry *10*, 97.

Liu, J.Z., Erlich, Y., and Pickrell, J.K. (2017). Case-control association mapping by proxy using family history of disease. Nat. Genet. *49*, 325–331.

Liu, M., Jiang, Y., Wedow, R., Li, Y., Brazel, D.M., Chen, F., Datta, G., Davila-Velderrain, J., McGuire, D., Tian, C., et al. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat. Genet. *51*, 237–244.

Loh, P.-R., Palamara, P.F., and Price, A.L. (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. *48*, 811–816.

McGranahan, T.M., Patzlaff, N.E., Grady, S.R., Heinemann, S.F., and Booker, T.K. (2011). A4β2 nicotinic acetylcholine receptors on dopaminergic neurons mediate nicotine reward and anxiety relief. J. Neurosci. *31*, 10891–10902.

McKay, J.D., Hung, R.J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D.C., Caporaso, N.E., Johansson, M., Xiao, X., Li, Y., et al. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nat. Genet. *49*, 1126–1132.

Melin, B.S., Barnholtz-Sloan, J.S., Wrensch, M.R., Johansen, C., Il'yasova, D., Kinnersley, B., Ostrom, Q.T., Labreche, K., Chen, Y., Armstrong, G., et al. (2017). Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. Nat. Genet. *49*, 789–794.

Motiwala, T., Kutay, H., Ghoshal, K., Bai, S., Seimiya, H., Tsuruo, T., Suster, S., Morrison, C., and Jacob, S.T. (2004). 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. *101*, 13844–13849.

Mucci, L.A., Hjelmborg, J.B., Harris, J.R., Czene, K., Havelick, D.J., Scheike, T., Graff, R.E., Holst, K., Möller, S., Unger, R.H., et al. (2016). Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA *315*, 68–76.

Ooi, W.L., Elston, R.C., Chen, V.W., Bailey-Wilson, J.E., and Rothschild, H. (1986). Increased familial risk for lung cancer. J. Natl. Cancer Inst. *76*, 217–222.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Schwartz, A.G., Yang, P., and Swanson, G.M. (1996). Familial risk of lung cancer among nonsmokers and their relatives. Am. J. Epidemiol. *144*, 554–562.

Sellers, T.A., Bailey-Wilson, J.E., Elston, R.C., Wilson, A.F., Elston, G.Z., Ooi, W.L., and Rothschild, H. (1990). Evidence for mendelian inheritance in the pathogenesis of lung cancer. J. Natl. Cancer Inst. *82*, 1272–1279.

Sulem, P., Gudbjartsson, D.F., Geller, F., Prokopenko, I., Feenstra, B., Aben, K.K.H., Franke, B., den Heijer, M., Kovacs, P., Stumvoll, M., et al. (2011). Sequence variants at CYP1A1-CYP1A2 and AHR associate with coffee consumption. Hum. Mol. Genet. *20*, 2071–2077.

Tokuhata, G.K., and Lilienfeld, A.M. (1963). Familial aggregation of lung cancer in humans. J. Natl. Cancer Inst. *30*, 289–312.

Walsh, R.M., Jr, Roh, S.-H., Gharpure, A., Morales-Perez, C.L., Teng, J., and Hibbs, R.E. (2018). Structural principles of distinct assemblies of the human α4β2 nicotinic receptor. Nature *557*, 261–265.

Wang, Y., Broderick, P., Webb, E., Wu, X., Vijayakrishnan, J., Matakidou, A., Qureshi, M., Dong, Q., Gu, X., Chen, W.V., et al. (2008). Common 5p15.33 and 6p21.33 variants influence lung cancer risk. Nat. Genet. *40*, 1407–1409.

Wang, Y., McKay, J.D., Rafnar, T., Wang, Z., Timofeeva, M.N., Broderick, P., Zong, X., Laplana, M., Wei, Y., Han, Y., et al. (2014). Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat. Genet. *46*, 736–741.

Yu, K., Wang, Z., Li, Q., Wacholder, S., Hunter, D.J., Hoover, R.N., Chanock, S., and Thomas, G. (2008). Population substructure and control selection in genome-wide association studies. PLoS One *3*, e2551.

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	-
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

** 6p22.2 contains two independent SNPs*

*** 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*

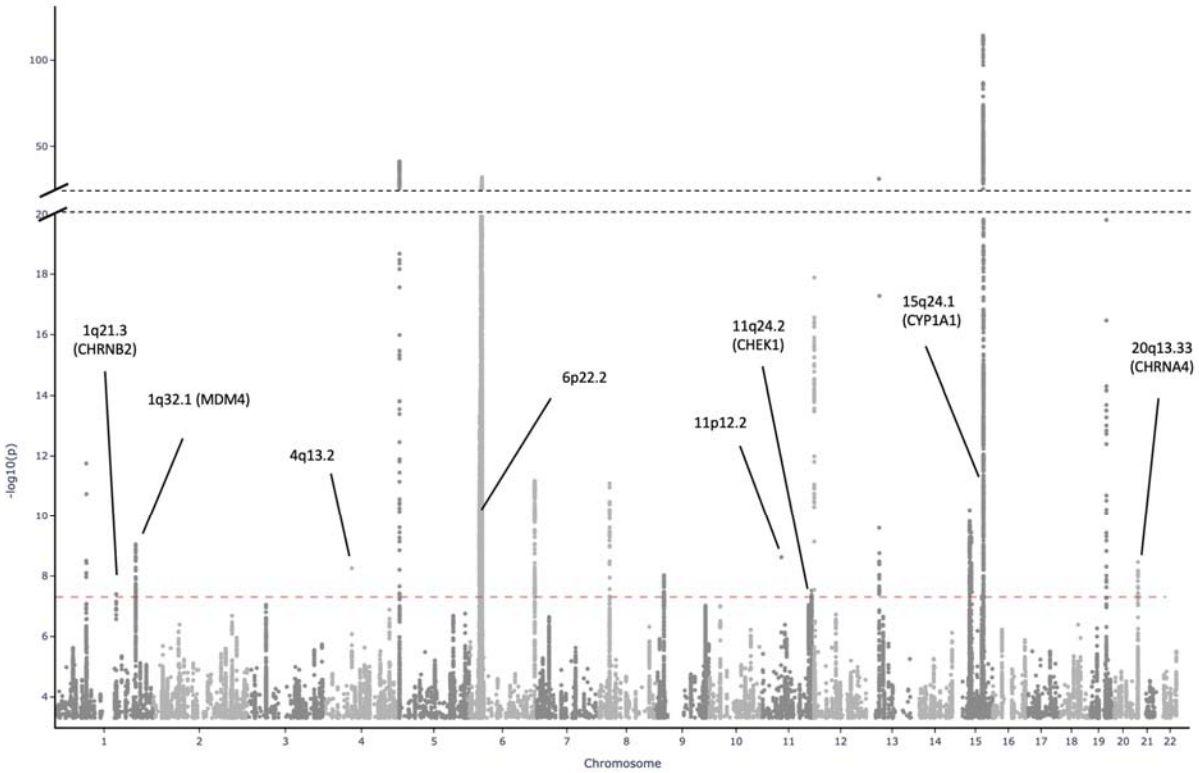

Figure 1: Manhattan plot of the meta-analysis of genome-wide by proxy (GWAx) with genome-wide association study (GWAS) into lung cancer.

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 ($5x10^{-8}$)

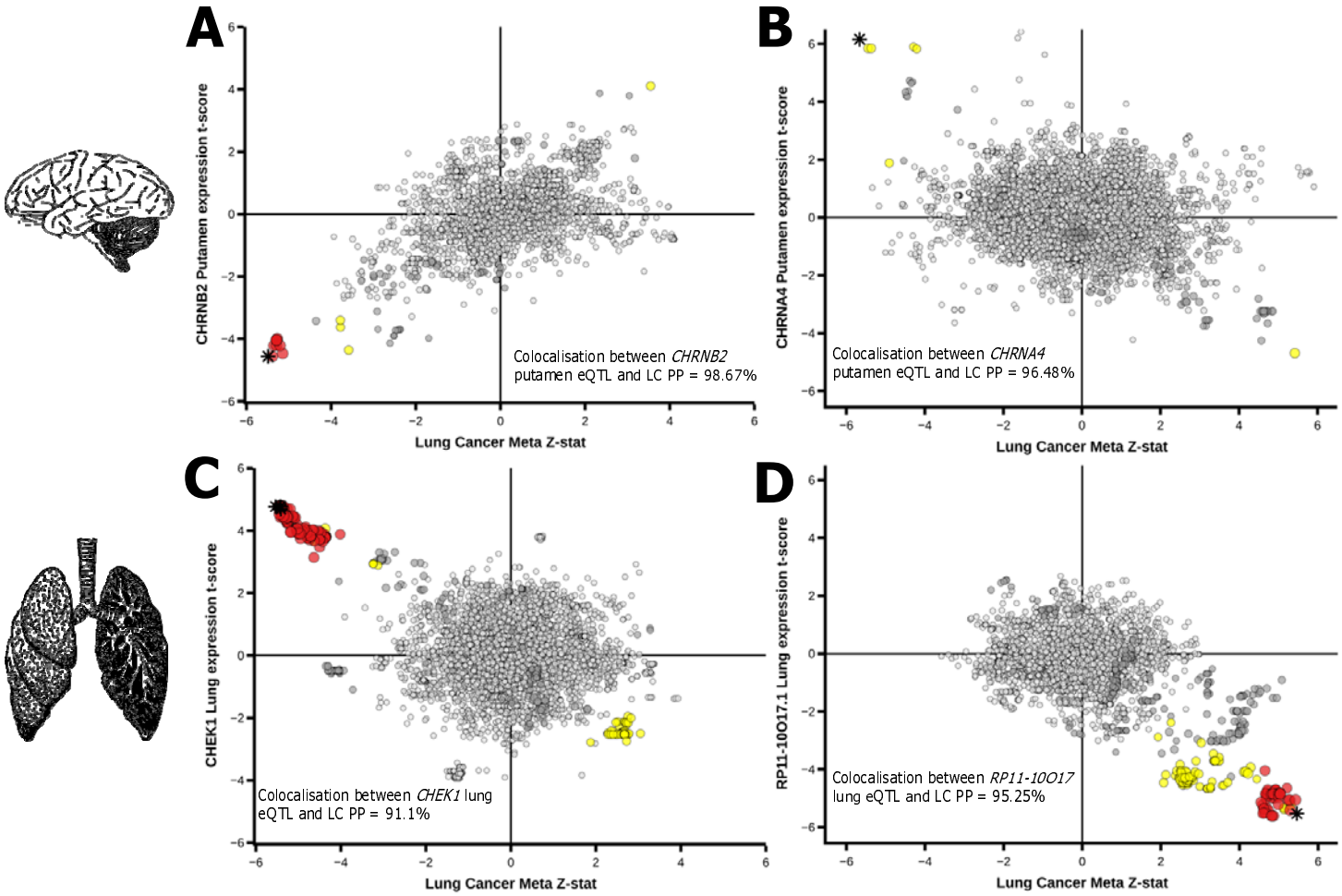

Figure 2: Brain and lung eQTLs discovered within the 8 novel loci

Co-localisation between lung cancer (x axis) and CHRNB2 putamen (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%)

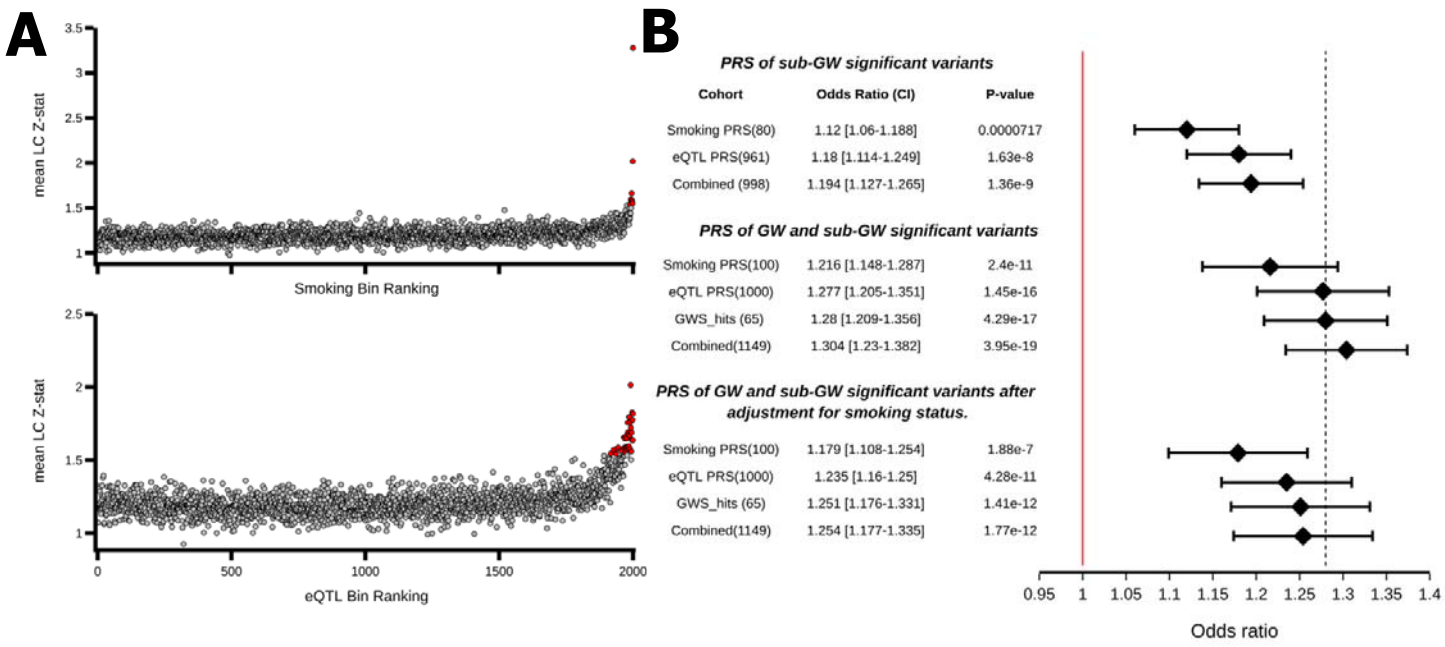

Figure 3: Germline polygenic risk score construction using smoking and eQTL related SNPs and performance testing within the UK Biobank lung cancer cohort.

(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 10 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 of the 65 loci plus both the smoking and eQTL list.

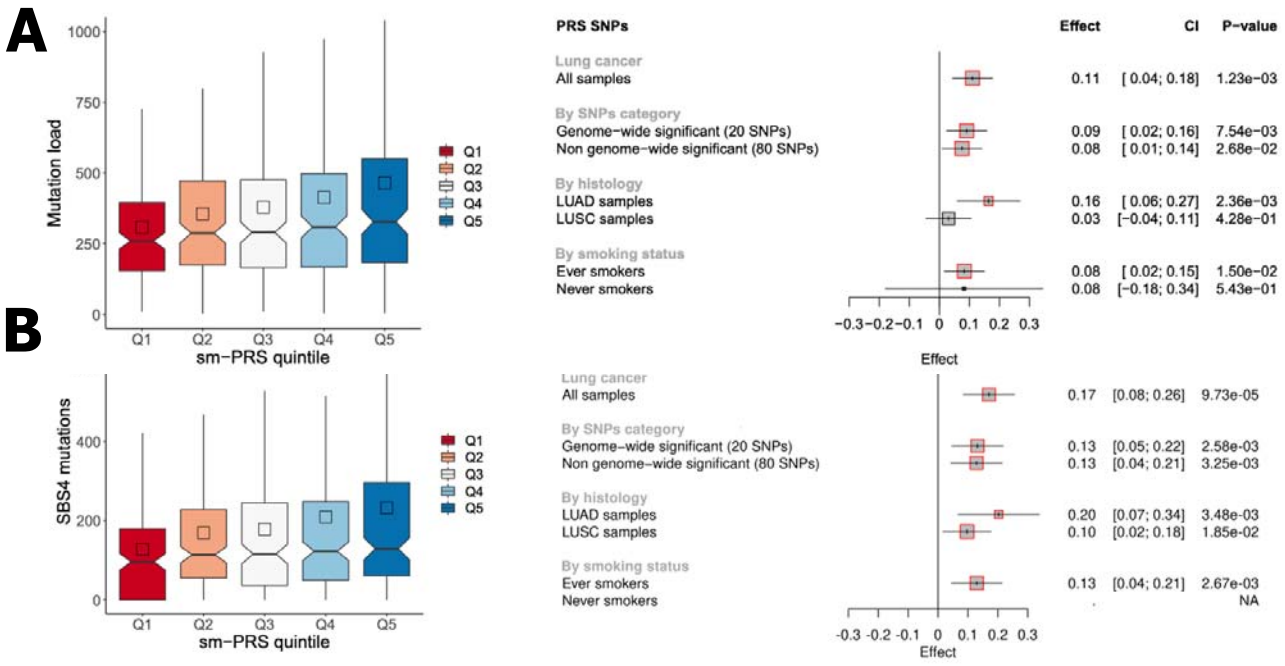

Figure 4: Polygenic risk scores for smoking (smPRS) associations with total number of mutations and mutations attributable to SBS4 in the TCGA cohort.

(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 (panel A) 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.