Published in final edited form as:
Nat Genet. 2017 July ; 49(7): 1126–1132. doi:10.1038/ng.3892.

Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes

A full list of authors and affiliations appears at the end of the article.

Summary

While several lung cancer susceptibility loci have been identified, much of lung cancer heritability remains unexplained. Here, 14,803 cases and 12,262 controls of European descent were genotyped on the OncoArray and combined with existing data for an aggregated GWAS analysis of lung cancer on 29,266 patients and 56,450 controls. We identified 18 susceptibility loci achieving genome wide significance, including 10 novel loci. The novel loci highlighted the striking

Corresponding author: Christopher I. Amos Christopher.I.Amos@dartmouth.edu.
*these authors have equal contributions

Author contributions.
Drafted the Paper: JDM, RJH, CIA
Project Coordination: CIA, RJH, JDM, RJH, DCC, NEC, SC, PBr, MTL
Performed the Statistical Analysis: CIA, JDM, RJH, YH, XZ, RCT, XuJ, XX, YL, JBy, KAP, DCQ, MNT, YBr, DZh
eQTL analysis of candidate variants: JDM, YBo, RCT, MTL, BZ, LSu, MTL, ML
Genomic annotation of candidate variants: DCQ, GCT, JBe, RFT
Assessed impact of candidate variants on nicotine addiction JDM, TR, TET, GWR, KS, DBH, LJB, RJH, SPIRO, LK
Assessed impact of candidate variants on Telomere length: JDM, RJH, KAP, AD, LK
Assessed impact of candidate variants on lung function: MDTo, MSA, LVW, LK
Sample collection and development of the epidemiological studies RJH, TR, TET, GWR, DCC, NEC, MJ, GL, SEB, XW, LLM, DA, HBi, MCA, WSB, ATa, GR, MDT, JKF, LAK, PL, AH, SLa, MBS, ASA, HS, YCH, JMY, PAB, ACP, YY, ND, LSu, RZ, YB, NL, JSJ, AMe, WS, CAH, LRW, AF-S, GF-T, HFMH, JHK, JD, ZH, MPAD, MWM, HBr, JoM, OM, DCM, KO, ATr, RT, JAD, MPB, CC, GEG, AC, FT, PW, IB, H-EW, JuM, TRM, ARi, ARo, KG, MJ, FAS, M-ST, SMA, EBH, CB, IH, VJ, MK, JL, AMu, SO, TMO, GS, BS, DZa, PBa, VS, SHZ, EJD, LMB, W-PK, Y-TG, RSH, JMcL, VLS, PJ, ML, DCN, MO, WT, LSo, MSA, MDTe, MRS, NCG, SML, FG, EOJ, AK, CP, RJH, JDM, MLT
Genetic sharing analysis: RCT, SLi, XiJ, JDM, RJH

Data Availability
The datasets generated during the current study are available at the URL: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001273.v1.p1. Prior meta-analyses of genome-wide association studies contributing to this study are available at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000876.v1.p1.

The Oncoarray data deposited at dbGAP includes data excluded from the analyses presented in this paper to avoid overlap with prior studies. Readers interested in obtaining a copy of the original data can do so by completing a proposal request form that is located at http://oncoarray.dartmouth.edu)
Cluster plots of all SNPs on the Oncoarray are located at http://oncoarray.dartmouth.edu

URLs
OncoArray: http://epi.grants.cancer.gov/oncoarray/)
http://oncoarray.dartmouth.edu)
Fastpop http://sourceforge.net/projects/fastpop/)
PLINK: http://zzz.bwh.harvard.edu/plink/)
IMPUTE2: http://mathgen.stats.ox.ac.uk/impute/impute_v2.html)
SHAPEIN: https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html)
GTEx: http://www.gtexportal.org/home/)
BRAINEAC: http://braineac.org)
TAG: https://www.med.unc.edu/pgc/downloads

heterogeneity in genetic susceptibility across lung cancer histological subtypes, with four loci associated with lung cancer overall and six with lung adenocarcinoma. Gene expression quantitative trait analysis (eQTL) in 1,425 normal lung tissues highlighted *RNASET2*, *SECISBP2L* and *NRG1* as candidate genes. Other loci include genes such as a cholinergic nicotinic receptor, *CHRNA2*, and the telomere-related genes, *OFBC1* and *RTEL1*. Further exploration of the target genes will continue to provide new insights into the etiology of lung cancer.

Lung cancer continues to be the leading cause of cancer mortality worldwide[1]. Although tobacco smoking is the main risk factor, the heritability of lung cancer has been estimated at 18%[2]. Genome-wide association studies (GWAS) have previously identified several lung cancer susceptibility loci including *CHRNA3/5, TERT, HLA, BRCA2, CHEK2*[3,4], but most of its heritability remains unexplained. With the goal of conducting a comprehensive characterization of common lung cancer genetic susceptibility loci, we undertook additional genotyping of lung cancer cases and controls using the OncoArray[5] genotyping platform, which queried 517,482 SNPs chosen for fine mapping of susceptibility to common cancers as well as for *de novo* discovery (Supplementary Table 1, and Online Methods). All participants gave an informed consent, and each study obtained local ethics committee approval. After quality control filters (Online Methods), a total of 14,803 cases and 12,262 controls of European ancestry were retained and underwent imputation techniques to infer additional genotypes for genetic variants included in the 1000 Genomes Project data (Online Methods). Logistic regression was then used to assess the association between variants (n=10,439,017 SNPs) and lung cancer risk, as well as by predominant histological types and by smoking behavior (Online Methods). Fixed-effects models (Online Methods) were used to combine the OncoArray results with previously published lung cancer GWAS[3,4,6], allowing for analysis of 29,266 patients and 56,450 controls of European descent (Table 1). There were no signs of genomic inflation overall or for any subtypes (Supplementary Figure 1) indicating little evidence for confounding by cryptic population structure (Online Methods). All findings with a P-value less than 1×10^{-5} are reported in Supplementary Table 2. As shown in Figure 1, the genetic architecture of lung cancer varies markedly among histological subtypes, with striking differences between lung adenocarcinoma and squamous cell carcinoma. Manhattan plots for small cell carcinoma (SCLC), ever and never smoking are displayed in Supplementary Figure 2. The array heritability estimates were comparable among histological subsets, but squamous cell carcinoma appeared to share more genetic architecture with small cell carcinoma (SCLC) than with adenocarcinoma (Supplementary Table 3).

Table 2 presents summary results of all loci with sentinel variants (defined as the variant with the lowest P-value at each locus) that reached genome-wide significance (P-value < 5×10^{-8}) for lung cancer overall and by histological subtypes. Sentinel variants stratified by new and previous genotyping and additional statistical significance assessed based on the number of effective tests, Approximate Bayes Factors, and Bayesian False Discovery Probability are presented in Supplementary Tables 4 and 5, respectively. Repeat genotyping of 12% of the OncoArray genotyped samples confirmed the fidelity of the genotyping or imputation for the risk loci, and showed excellent concordance of imputation for SNPs with

MAF>0.05 (Online Methods, Supplementary note). Among the 18 loci that reached GWAS significance, 10 had not reached significance in a genome-wide scan (Figure 1). Of these, four novel loci were associated with lung cancer overall, and six with adenocarcinoma.

To decipher the association between these 18 loci and lung cancer risk, we further investigated their association with gene expression level in normal lung tissues (n=1,425) (Supplementary Table 6, Supplementary Figure 3), genomic annotations (Supplementary Table 7), smoking propensity (cigarettes smoked per day (n=91,046) and Fagerström Test for Nicotine Dependence metrics (n=17,074)) (Table 2). Previous studies have shown shared risk for lung cancer and COPD through inflammation and ROS pathways[7]; therefore, we also assessed the association between sentinel SNPs and reduced lung capacity through spirometry measurements (forced expiratory volume in 1 second [FEV1], forced vital capacity [FVC], n =30,199) (Table 2 and Online Methods).

Variants at 4 novel loci (1p31.1, 6q27, 8p21, 15q21.1) were associated with lung cancer risk overall, with little evidence for heterogeneity among subtypes (Supplementary Figure 4). The 1p31.1 locus, recently identified in a pathway-based analysis of the TRICL data[8], represented by rs71658797 (Odds Ratio [OR]=1.14, 95% Confidence Interval [CI] 1.09–1.18, P-value=3.25×10^{-11}), is located near *FUBP1/DNAJB4* (Supplementary Figure 4). At 6q27, rs6920364 was associated with lung cancer risk with an OR of 1.07 (95% CI 1.04–1.09, P-value=2.9×10^{-8}) with little heterogeneity found by smoking status (Supplementary Figure 4). This locus is predicted to regulate *RNASET2* (Supplementary Figure 5, Supplementary Table 7). We identified rs6920364 as a lung cis-eQTL for *RNASET2*, an extracellular ribonuclease, in all five cohorts tested (Supplementary Table 6), with increased lung cancer risk correlating with increased *RNASET2* expression (Figure 2). Variants correlated with rs6920364 (r^2>0.88) have been noted in GWAS of Crohn's disease and inflammatory bowel disease[9–13].

The 8p21 locus has been suggested as a lung cancer susceptibility locus by pathway analysis[14] and now confirmed at GWAS significance level. It is a complex locus represented by sentinel variant rs11780471 associated with lung cancer (OR=0.87, 95% CI 0.83–0.91, P-value=1.69×10^{-8}) (Supplementary Figure 4), but this region contained additional uncorrelated variants (pairwise r^2< 0.10) associated with lung cancer (Supplementary Table 8). Multivariate analysis was consistent with multiple susceptibility alleles at this locus (Supplementary Table 8). In contrast to lung tissue (Figure 3A, Supplementary Table 6, Supplementary Figure 3), we noted that the alleles associated with lung cancer tended to be associated with cerebellum expression of *CHRNA2*, a member of the cholinergic nicotinic receptor (Figure 3B). The *CHRNA2* rs11780471 cis-eQTL effect in the brain was limited to the cerebellum (Figure 3C), a region not traditionally linked with addictive behavior, but where an emerging role is suggested[15]. We therefore investigated rs11780471 in the context of smoking behavior (Supplementary Methods). Unlike the well-described 15q25.1 (rs55781567) *CHRNA5* locus (Table 2), rs11780471 was not associated with number of cigarettes smoked per day or the FTND metrics (Figure 3D). Nevertheless, lung cancer risk allele carriers of rs11780471 tended to be smokers and initiated smoking at earlier ages (Figure 3D), implying that this variant's association with lung cancer could potentially be mediated via influencing aspects of smoking behavior. Another potentially relevant gene in this region is *EPHX2*, a xenobiotic metabolism gene. Although the sentinel variant is not an eQTL for *EPHX2* in lung tissues, other associated variants in the region are (e.g. rs146729428, p-value of 1.77×10^{-7} (Supplementary Table 2) and 5×10^{-4} for lung cancer risk and eQTL, respectively). A potential synergistic role of both *EPHX2* and *CHRNA2* on lung cancer etiology cannot be excluded.

The genetic locus at 15q21 (rs66759488) was shown to be associated with lung cancer (OR=1.07, 95% CI 1.04–1.10, p=2.83×10^{-8}) overall and across lung cancer histologies (Supplementary Figure 4). Genomic annotation suggests that genetic variants correlated with rs66759488 may influence the *SEMA6D* gene (Supplementary Table 7), but there was no clear eQTL effect (Supplementary Table 6), and this variant did not appear to have a major influence on smoking propensity or lung function (Table 2).

For specific lung cancer histology subtypes, we identified 6 novel loci associated with lung adenocarcinoma (15q21, 8p12, 10q24, 20q13.33, 11q23.3 and 9p21.3) (Table 2). The locus at 15q21 (rs77468143, OR=0.86, 95% CI 0.82–0.89, p=1.15×10^{-16}) is predicted to target *SECISBP2L* (Supplementary Figure 5), and expression analysis indicated rs77468143 to be a cis-eQTL for *SECISBP2L* in lung tissue in all eQTL cohorts tested (Supplementary Table 6). The genetic risk allele appears to correlate with decreased expression levels of *SECISBP2L* (Figure 2, Supplementary Figure 5), an observation that is consistent with *SECISBP2L* being down regulated in lung cancers[16]. rs77468143 was nominally associated with lung function (Table 2), potentially implicating inflammation of lung as part of the mechanism at this locus.

At 8p12, expression analysis indicated that the alleles associated with lung adenocarcinoma (represented by the sentinel variant rs4236709 (Table 2)), also appear to be a lung cis-eQTL for the *NRG1* gene (Supplementary Table 6, Supplementary Figure 5). This region also contains putative regulatory regions (Supplementary Figure 5). Somatic translocations of *NRG1* are infrequently observed in lung adenocarcinomas[17]. While somatic translocations at 8p12 generally take place in never smokers and are linked with ectopic activation of NRG1, rs4236709 was associated with lung cancer in both ever and never smokers (Supplementary Figure 4) and its genetic risk correlated with decreased *NRG1* expression (Figure 2). Interestingly, 6q22.1 variants located near *ROS1*, another gene somatically translocated in lung adenocarcinoma and for which nearby germline variants were associated with never smoking lung adenocarcinoma in Asian women[18], were associated with lung adenocarcinoma at borderline genome wide significance (rs9387479; OR=0.92, 95% CI 0.89–0.95, p=6.57×10^{-8}) (Supplementary Table 2).

Three of the sentinel variants associated with lung adenocarcinoma are located near genes related to telomere regulation; rs7902587 (10q24) and rs41309931 (20q13.33) near *OBFC1* and *RTEL1*, respectively, and rs2853677 near *TERT* as previously noted[19,20]. The variants at 10q24 associated with lung adenocarcinoma also appear to be associated with telomere length (Supplementary Figure 6). By contrast, and consistent with observations with 20q13.33 variants associated with glioma[21], the variants associated with telomere length at 20q13.33 were not necessarily those associated with lung adenocarcinoma (Supplementary Figure 6). Nevertheless, more generally the variants associated by GWAS with longer telomere length[22] appear linked with risk of lung adenocarcinoma[23] and glioma[21,24], a finding consistent with our expanded analysis here (Supplementary Figure 6).

We additionally identified a complex locus at 11q23.3. The sentinel variant rs1056562 (OR=1.11, 95% CI 1.07–1.14, $p=2.7\times10^{-10}$) is more prominently associated with lung adenocarcinoma (Supplementary Figure 4). rs1056562 was correlated with expression of two genes at this locus, *AMICA1* and *MPZL3* (Supplementary Table 6). However, there did not appear to be a consistent relationship between the alleles related with *AMICA1* and *MPZL3* gene expression and those with lung adenocarcinoma (Figure 2, Supplementary Table 9), suggesting that expression of these genes alone is unlikely to mediate this association.

At 9p21.3 we identified rs885518 that appeared to be associated with lung adenocarcinoma (OR=1.17, 95% CI 1.11–1.23, $p=6.8\times10^{-10}$). 9p21.3 is a region containing *CDNK2A* and variants associated with multiple cancer types, including lung cancer. Nevertheless, rs885518 is located approximately 200kb centromeric the previously described variants (Supplementary Figure 4) and shows little evidence for LD (all pairwise $r^2 < 0.01$) with rs1333040, a variant previously associated with lung squamous cell carcinoma[3] and rs62560775, another variant suggested to be associated with lung adenocarcinoma[25] that we confirm to genome significance here. Intriguingly, these variants appear to confer predominant associations with different lung cancer histologies suggesting that they are independent associations (Supplementary Figure 7).

Aside from the clear smoking-related effects on lung cancer risk through the *CHRNA5* and *CYP2A6* regions and association with *CHRNA2* noted above, the rest of variants we have identified do not appear to clearly influence smoking behaviors (Table 2), implying that these associations are likely mediated by other mechanisms. Nevertheless, there is shared genetic architecture between smoking behavior and lung cancer risk, consistent with the notion that genetic variants do influence lung cancer risk also through behavioral mechanisms (Supplementary Figure 8).

In conclusion, the genetic susceptibility alleles we describe here explain approximately 12.3% of the familial relative risk previously reported in family cancer databases[26,27], out of which 3.5% was accounted for by the novel loci. Our findings emphasize striking heterogeneity across histological subtypes of lung cancer. We expect that further exploration of the related target genes of these susceptibility loci, as well as validation and identification of new loci, will continue to provide insights into the etiology of lung cancer.

Online methods

This work is conducted based on the collaboration of Transdisciplinary Research of Cancer in Lung of the International Lung Cancer Consortium (TRICL-ILCCO) and the Lung Cancer Cohort Consortium (LC3). The participating studies are individually described in the Supplementary Note.

OncoArray genotyping

Genotyping was completed at the Center for Inherited Disease Research (CIDR), the Beijing Genome Institute, the HelmholtzCenter Munich (HMGU), Copenhagen University Hospital, and the University of Cambridge. Quality control steps follow the approach described previously for the OncoArray[5] (Supplementary Note).

Genotype quality control

After removing the 1,193 expected duplicates, QC procedures for the 43,398 individuals are summarized in Supplementary Note Figure 1. Standard quality control procedures (detailed in the Supplementary Note) were used to exclude underperforming individuals (number of DNAs=1,708) and genotyping assays (judged by success rate, genotype distributions deviated from that expected by Hardy Weinberg equilibrium, number of variants=16,149). After filtering, there were 517,482 SNPs available for analysis.

Identity by Descent (IBD) was calculated between each pair of samples in the data using PLINK to detect unexpected duplicates and relatedness. Details are described in Supplementary Note. 340 unexpected duplicated samples (proportion IBD>0.95) and 940 individuals were removed as related samples with proportion IBD between 0.45 and 0.95. Of these, 721 of them were expected first degree relatives. In total, 0.56% of the total samples were removed as unexpected duplicates or relatives in the QC analysis. We additionally considered the potential that more distant familial relationships could have impacted the results. However, further restriction to proportion IBD > 0.2 identified 139 second degree relatives and excluding these had minimal impact on the association results (Supplementary Note Table 1).

Complete genotype data for X chromosomes were used to verify reported sex by using PLINK sex inference and a support vector machine procedure resulting in 306 non-concordant samples being removed (Supplementary Note).

We used the program FastPop (http://sourceforge.net/projects/fastpop/)[28] was used to identify 5,406 individuals of non-European ancestry (Supplementary Note) resulting in a final association analysis including 14,803 lung cancer cases and 12,262 controls.

We confirmed the fidelity genotyping (directly and imputed) of the OncoArray platform by considering concordance of these genotypes relative to genotypes obtained from analogous genotyping platform (Supplementary Note).

Imputation analysis

A detailed description of the imputation procedures used by the OncoArray consortium and in this Lung Oncoarray project, has been described previously.[5] Briefly, the reference Dataset was the 1000 Genomes Project (GP) Phase 3 (Haplotype release date October 2014). The forward alignment of SNPs genotyped on the Oncoarray was confirmed by blasting the sequences used for defining SNPs against the 1000 Genomes. Any ambiguous SNPs were subjected to a frequency comparison to 1000 Genomes variants. Allele frequencies were calculated from a large collection of control samples from Europeans (from 108,000 samples) and Asians (11,000 samples). A difference statistic is calculated by the formula: (|

p1-p2|- 0.01)^2/((p1+p2)(2-p1-p2)) where p1 and p2 are the frequencies our dataset and in the 1000 genomes respectively[5]. A cutoff of 0.008 in Europeans and 0.012 in Asians is needed to pass. SNPs where the frequency would match if the alleles were flipped were excluded from imputation, but not from the association analyses.[5] AT/GC SNPs were not present in previously genotyped lower density arrays. Because all imputation was performed to the same standard all SNPs had the same orientation at the time of imputation. The OncoArray whole genome data were imputed in a two-stage procedure using SHAPEIT to derive phased genotypes, and IMPUTEv2[29] to perform imputation of the phased data. We included for imputation only the more common variant if more than one variant yielded a match at the same position. The detailed parameter settings are in the Supplementary Note.

Meta-analysis of lung cancer GWAS

FlashPCA[30] was run for principal component analysis (PCA) to infer genetic ancestry by genotype. The regression model assumed an additive genetic model and included the first three eigenvalues from FlashPCA as covariates. For imputed data of smaller sample size, which was enrolled in our analysis later, we changed the method score to EM algorithm to accommodate smaller sample size.

We combined imputed genotypes from 14,803 cases and 12,262 controls from the OncoArray series with 14,436 cases and 44,188 controls samples undertaken by the previous lung cancer GWAS[3,4,6], including studies of IARC, MDACC, SLRI, ICR, Harvard, NCI, Germany and deCODE as described previously[3,4,6], and we ensured that there were no overlap between the ATBC, EAGLE and CARET studies included in both the previous GWAS and current OncoArray dataset by comparing the identity tags (IDs) of all study participants.

In addition to lung cancer, analyses by histological strata (adenocarcinoma, squamous cell carcinoma, small cell carcinoma (SCLC) and smoking status (Ever/Never) was assessed where data were available. Results from analyses defined by Ever and Never smoking strata did not identify any novel variants.

We conducted the fixed effects meta-analysis with the inverse variance weighting and random effects meta-analysis from the DerSimonian-Laird method[31]. All meta-analysis and calculations were performed using SAS version 9.4 (SAS Institute Inc., Cary, NC, USA). As the same referent panel was used for all studies, all SNPs showed the same forward alignment profiles. We excluded poorly imputed SNPs defined by imputation quality $R^2 < 0.3$ or Info < 0.4 for each meta-analysis component and SNPs with a Minor allele frequency (MAF) >0.01 (except for *CHEK2* rs17879961 and *BRCA2* rs11571833 which we have validated extensively previously[4]. We generated the index of heterogeneity(I^2) and P-value of Cochran's Q statistic to assess heterogeneity in meta-analyses and considered only variants with little evidence for heterogeneity in effect between the studies (P-value of Cochran's Q statistic >0.05). SNPs were retained for study provided the average imputation R-square was at least 0.4. For SNPs in the 0.4–0.8 range that reached genome wide significance results were evaluated for consistency with neighboring SNPs to assure a reliable inference. Due to the smaller sample size and fewer sites contributing in the strata of

Never Smokers and SCLC, we additionally required variants to be present in each of the meta-analysis components to be retained for these 2 stratified analyses.

Conditional analysis was undertaken using SNPTEST where individual level data was available and GCTA[32] packages for the previous lung cancer GWAS, with the LD estimates obtained from individuals of European origin for the later. Results were combined using fixed effects inverse variance weighted meta-analysis as described above[33].

Assessing Statistical Significance

Genome wide statistical significance was considered at P-values of 5×10^{-8} or lower, but we also presented significance per alternative criteria following Bonferroni correction for the number of effective tests or Bayesian False Discovery Probability (BFDP) described below.

To evaluate the effective number of tests we used the Li and Ji (2005)[33] method which performs an initial step of filtering out SNPs with MAF<0.01 (imputation is less reliable for these and power is also limited for most odds ratios). Among the 4,751,148 markers with that MAF there were 1,182,363 effective tests.

The BFDP combines significance level, study power, and cost of false discovery and non-discovery into consideration. The detailed procedures of this method are described in Wakefield, 2007[34]. Essentially, the approximate Bayes Factor (ABF) which BFDP uses reflects how much the prior odds change in the light of the observed data (i.e. relative probability of the observed estimates under the null versus alternative hypothesis). Given the nature of GWA studies, we applied a flat prior for all variants at prior probability of 10^{-6} and 10^{-8} to demonstrate the range of BFDP.

Annotation of susceptibility loci

We combined multiple sources of *in silico* functional annotation from public databases to help identify potential functional SNPs and target genes, based on previous observations that cancer susceptibility alleles are enriched in *cis*-regulatory elements and alter transcriptional activity. The details are described in the Supplementary Note.

eQTL analysis of lung cancer sentinel variants

To investigate the association between the sentinel variants and mRNA expression, we used three different eQTL datasets: (i) Microarray eQTL study: The lung tissues for eQTL analyses were from patients who underwent lung surgery at three academic sites, Laval University, University of British Columbia (UBC), and University of Groningen. Whole-genome gene expression profiling in the lung was performed on a custom Affymetrix array (GPL10379). Microarray pre-processing and quality controls were described previously. Genotyping was carried on the Illumina Human 1M-Duo BeadChip array. Genotypes and gene expression levels were available for 409, 287 and 342 patients at Laval, UBC, and Groningen, respectively. (ii) NCI RNAseq eQTL study: RNA was extracted from lung tissue samples within the Environment and Genetics in Lung cancer Etiology (EAGLE) study. RNAseq was carried out on 90 lung tissue sampled from an area distant from the tumor (defined here as “non-malignant lung tissue”) to minimize the potential for local cancer field effects. Transcriptome sequencing of 90 non-tumor samples was performed on the Illumina HiSeq2000/2500 platform with 100-bp paired-end reads. Genotyping was undertaken using Illumina bead arrays as described previously. (iii) GTEx: eQTL summary statistics based on RNAseq analysis were obtained for eQTL summary statistics from the GTEx data portal http://www.gtexportal.org/home/[35]. This data included 278 individuals with data from lung tissue. Details of these three eQTL studies are included in the Supplementary Note.

The Microarray eQTL study was used as a discovery cohort. Probe sets located within 1 Mb up and downstream of lung cancer SNPs were considered for cis-eQTL analyses. We have also explored a 5 Mb interval for lung cancer-associated SNPs not acting as lung eQTL within the 1 Mb window. The top eQTL association for that sentinel variant (or if contained multiple eQTL's with P-value<0.0005 each was considered), this particular eQTL was then chosen and assessed specifically in the independent NCI and GTEx RNAseq eQTL datasets. Statistical significance was defined the eQTL surpassed a locus specific Bonferroni correlation in the discovery cohort (P-value=0.05/number of probes at that locus) and subsequently there was evidence for replication of the eQTL effect with that variant and gene within the validation cohorts (NCI/GTEx RNAseq).

Lung cancer susceptibility variants in other phenotypes

We assessed associations between sentinel genetic variant associated with lung cancer and other phenotypes, including smoking behavior Fagerström Test for Nicotine Dependence, lung function and telomere length. Additional details of these analyses for other phenotypes are described in Supplementary Note. Briefly:

Smoking behaviors—The effects of lung cancer sentinel variants and smoking behavior were assessed based on the meta-analysis across 3 studies: ever-smoking controls with intensity information from the Oncoarray studies (N=8,120), deCODE (N=40,882) and UK Biobank (N=42,044). The association with nicotine dependence was evaluated based on Fagerström Test for Nicotine Dependence (FTND) data collected in 4 studies (n=17,074): deCODE Genetics, Environment and Genetics in Lung Cancer Etiology (EAGLE), Collaborative Genetic Study of Nicotine Dependence (COGEND), and Study of Addiction: Genetics and Environment (SAGE) and among current smokers in one other study [Chronic Obstructive Pulmonary Disease Gene (COPDGene). The study-specific SNP association results were combined using fixed effects, inverse variance-weighted meta-analysis with genomic control applied. Specifically, for the 8p21 variant rs11780471, we additionally considered other aspects of smoking behavior data from UKBiobank, deCODE and OncoArray controls. We additionally included summary statistics for the rs11780471 variants from the TAG consortium (described in detail in the Supplementary Note).

Lung function—The lung function *in silico* look up was conducted in SpiroMeta consortium, which included 38,199 European ancestry individuals. The genome-wide associations between genetic variants and forced expiratory volume in 1 second (FEV1), forced vital capacity (FVC) and FEV1/FVC with 1000 Genomes Project (phase 1)-imputed genotypes in the GWAS with 38,199 individuals[36].

Telomere Length (TL)—Sentinel genetic variants associated with telomere length were those described by Codd et al[22]. Telomere lengths in 6,766 individuals from the UK Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH) study controls using a real-time PCR methodology and genotyping as described in Pooley et al., 2013[37].

Genetic heritability and correlations

Genome-wide SNP heritability and correlation estimates were obtained using association summary statistics and linkage disequilibrium (LD) information through LD Score (LDSC) regression analyses[38,39]. These analyses were restricted to HapMap3 SNPs with minor allele frequency above 5% in European populations of 1000 Genomes. Association summary statistics used for these analyses were based on lung cancer histological/smoking types (lung cancer overall, adenocarcinoma, squamous cell, small cell, ever smokers and never smokers) and smoking behavior parameters (cigarettes per day (CPD), smoking status (ever vs never smokers), and smoking cessation (current vs former smokers) from TRICL-ILCCO OncoArray consortium and Tobacco And Genetics consortium (https://www.med.unc.edu/pgc/downloads)[40].

Estimating the percentage of familial relative risks explained

The familial relative risk to a first degree relative accounted for by an individual variant (denoted as λ_i) is estimated based on relative risk per allele and allele frequency for that variant, using the method described in Hemminki et al[41,] and Bahcall[42], under the assumption of log-additive effect. Assuming the effects of all susceptibility variants combined multiplicatively and not in linkage disequilibrium, the combined effect (λ_T) can then be expressed as the product of all λ_i. The proportion of the familial relative risk attributable to the totality of the susceptibility variants can then be computed as $\log(\lambda_T)/\log(\lambda_P)$. For lung cancer, the λ_P is approximately 2.0 based on the family cancer databases[26,27]. The percentage reported is based on the 18 sentinel variants reported in Table 2. The multiple independent alleles in the same locus are not accounted for in this estimation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

James D. McKay[*,1], Rayjean J. Hung[*,2], Younghun Han[3], Xuchen Zong[2], Robert Carreras-Torres[1], David C. Christiani[4], Neil E. Caporaso[5], Mattias Johansson[1], Xiangjun Xiao[3], Yafang Li[3], Jinyoung Byun[3], Alison Dunning[6], Karen A. Pooley[6], David C. Qian[3], Xuemei Ji[3], Geoffrey Liu[2], Maria N. Timofeeva[1], Stig E. Bojesen[7,8,9], Xifeng Wu[10], Loic Le Marchand[11], Demetrios Albanes[5], Heike Bickeböller[12], Melinda C. Aldrich[13], William S. Bush[14], Adonina Tardon[15], Gad Rennert[16], M. Dawn Teare[17], John K. Field[18], Lambertus A. Kiemeney[19], Philip Lazarus[20], Aage Haugen[21], Stephen Lam[22], Matthew B. Schabath[23], Angeline S. Andrew[24], Hongbing Shen[25], Yun-Chul Hong[26], Jian-Min Yuan[27], Pier Alberto

Bertazzi[28,29], Angela C. Pesatori[29], Yuanqing Ye[10], Nancy Diao[4], Li Su[4], Ruyang Zhang[4], Yonathan Brhane[2], Natasha Leighl[30], Jakob S. Johansen[31], Anders Mellemgaard[31], Walid Saliba[16], Christopher A. Haiman[32], Lynne R. Wilkens[11], Ana Fernandez-Somoano[15], Guillermo Fernandez-Tardon[15], Henricus F.M. van der Heijden[19], Jin Hee Kim[33], Juncheng Dai[25], Zhibin Hu[25], Michael PA Davies[18], Michael W. Marcus[18], Hans Brunnström[34], Jonas Manjer[35], Olle Melander[35], David C. Muller[36], Kim Overvad[37], Antonia Trichopoulou[38], Rosario Tumino[39], Jennifer A. Doherty[24,40,41,42], Matt P. Barnett[40], Chu Chen[40], Gary E. Goodman[43], Angela Cox[44], Fiona Taylor[44], Penella Woll[44], Irene Brüske[45], H.-Erich Wichmann[45,46,47], Judith Manz[48], Thomas R. Muley[49,50], Angela Risch[48,49,50,51,52], Albert Rosenberger[12], Kjell Grankvist[53], Mikael Johansson[54], Frances A. Shepherd[55], Ming-Sound Tsao[55], Susanne M. Arnold[56], Eric B. Haura[57], Ciprian Bolca[58], Ivana Holcatova[59], Vladimir Janout[60], Milica Kontic[61], Jolanta Lissowska[62], Anush Mukeria[63], Simona Ognjanovic[64], Tadeusz M. Orlowski[65], Ghislaine Scelo[1], Beata Swiatkowska[66], David Zaridze[63], Per Bakke[67], Vidar Skaug[21], Shanbeh Zienolddiny[21], Eric J. Duell[68], Lesley M. Butler[27], Woon-Puay Koh[69], Yu-Tang Gao[70], Richard S. Houlston[71], John McLaughlin[72], Victoria L. Stevens[73], Philippe Joubert[74], Maxime Lamontagne[74], David C. Nickle[75], Ma'en Obeidat[76], Wim Timens[77], Bin Zhu[5], Lei Song[5], Linda Kachuri[2], María Soler Artigas[78,79], Martin D. Tobin[78,79], Louise V. Wain[78,79], SpiroMeta Consortium[80], Thorunn Rafnar[81], Thorgeir E. Thorgeirsson[81], Gunnar W. Reginsson[81], Kari Stefansson[81], Dana B. Hancock[82], Laura J. Bierut[83], Margaret R. Spitz[84], Nathan C. Gaddis[85], Sharon M. Lutz[86], Fangyi Gu[5], Eric O. Johnson[87], Ahsan Kamal[3], Claudio Pikielny[3], Dakai Zhu[3], Sara Lindströem[88], Xia Jiang[89], Rachel F. Tyndale[90,91], Georgia Chenevix-Trench[92], Jonathan Beesley[92], Yohan Bossé[74,93], Stephen Chanock[5], Paul Brennan[1], Maria Teresa Landi[5], and Christopher I. Amos[3]

Affiliations

[1]International Agency for Research on Cancer, World Health Organization, Lyon, France [2]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, University of Toronto, Toronto, Canada [3]Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover NH [4]Department of Environmental Health, Harvard TH Chan School of Public Health, and Massachusetts General Hospital/Harvard Medical School, Boston, MA. 02115 [5]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD [6]Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, United Kingdom [7]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Denmark [8]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark [9]Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen, Denmark [10]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX USA [11]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA [12]Department of Genetic Epidemiology, University Medical Center, Georg-August-University Göttingen, Germany [13]Department of Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center

[14]Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, OH [15]University of Oviedo and CIBERESP, Faculty of Medicine, Campus del Cristo s/n, 33006 Oviedo, Spain [16]Clalit National Cancer Control Center at Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel [17]School of Health and Related Research, University of Sheffield, England, UK [18]Institute of Translational Medicine, University of Liverpool, Liverpool, United Kingdom [19]Radboud University Medical Center, Nijmegen, The Netherlands [20]Department of Pharmaceutical Sciences, College of Pharmacy, Washington State University, Spokane, Washington, USA [21]National Institute of Occupational Health, Oslo, Norway [22]British Columbia Cancer Agency, Vancouver, Canada [23]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL [24]Department of Epidemiology, Geisel School of Medicine, Hanover, NH [25]Department of Epidemiology and Biostatistics, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, Nanjing, P.R. China [26]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea [27]University of Pittsburgh Cancer Institute, Pittsburgh, USA [28]Department of Preventive Medicine, IRCCS Foundation Ca' Granda Ospedale Maggiore Policlinico, Milan, Italy [29]Department of Clinical Sciences and Community Health – DISCCO, University of Milan, Milan, Italy [30]University Health Network- The Princess Margaret Cancer Centre, Toronto, CA [31]Department of Oncology, Herlev and Gentofte Hospital, Copenhagen University Hospital, Denmark [32]Department of Preventive Medicine, Keck School of Medicine, University of Southern California Norris Comprehensive Cancer Center, Los Angeles, CA [33]Department of Integrative Bioscience & Biotechnology, Sejong University, Gwangjin-gu, Seoul, Republic of Korea [34]Dept. of Pathology, Lund University, Lund, Sweden [35]Faculty of Medicine, Lund University, Lund, Sweden [36]School of Public Health, St Mary's Campus, Imperial College London, UK [37]Section for Epidemiology, Department of Public Health, Aarhus University, Denmark [38]Hellenic Health Foundation, Athens, GR [39]Tumino. Molecular and Nutritional Epidemiology Unit CSPO (Cancer Research and Prevention Centre), Scientific Institute of Tuscany, Florence, Italy [40]Fred Hutchinson Cancer Research Center, Seattle, Washington, USA [41]Huntsman Cancer Institute, 2000 Circle of Hope, Salt Lake City, UT 84112 [42]Huntsman Cancer Institute, Department of Population Health Sciences, University of Utah, Salt Lake City, Utah, USA [43]Swedish Medical Group, Seattle, WA, USA [44]Department of Oncology, University of Sheffield, Sheffield, UK [45]Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany [46]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig Maximilians University, Munich, Germany [47]Institute of Medical Statistics and Epidemiology, Technical University Munich, Germany [48]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany [49]Thoraxklinik at University Hospital Heidelberg [50]Translational Lung Research Center Heidelberg (TLRC-H),

Heidelberg, Germany [51]German Center for Lung Research (DZL), Heidelberg, Germany [52]University of Salzburg and Cancer Cluster Salzburg, Austria [53]Department of Medical Biosciences, Umeå University, Umeå, Sweden [54]Department of Radiation Sciences, Umeå University, Umeå, Sweden [55]Princess Margaret Cancer Centre, Toronto, Canada [56]University of Kentucky, Markey Cancer Center, Lexington, Kentucky, USA [57]Department of Thoracic Oncology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA [58]Institute of Pneumology "Marius Nasta", Bucharest, Romania [59]2nd Faculty of Medicine, Charles University, Prague, Czech Republic [60]Faculty of Medicine, University of Ostrava, Czech Republic [61]Clinical Center of Serbia, Belgrade. School of Medicine, University of Belgrade [62]M. Sklodowska-Curie Cancer Center, Institute of Oncology, Warsaw, Poland [63]Department of Epidemiology and Prevention, Russian N.N.Blokhin Cancer Research Centre, Moscow, Russian Federation [64]International Organization for Cancer Prevention and Research, Belgrade, Serbia [65]Department of Surgery, National Tuberculosis and Lung Diseases Research Institute, Warsaw, Poland [66]Nofer Institute of Occupational Medicine, Department of Environmental Epidemiology, Lodz, Poland [67]Department of Clinical Science, University of Bergen, Bergen, Norway [68]Unit of Nutrition and Cancer, Catalan Institute of Oncology (ICO-IDIBELL), Barcelona, Spain [69]Duke-National University of Singapore Medical School, Singapore, Singapore [70]Department of Epidemiology, Shanghai Cancer Institute, China [71]The Institute of Cancer Research, London, England [72]Public Health Ontario, Canada [73]American Cancer Society, Inc., Atlanta, Georgia, USA [74]Institut universitaire de cardiologie et de pneumologie de Québec, Québec, Canada [75]Merck Research Laboratories, Genetics and Pharmacogenomics, Boston, MA, USA [76]The University of British Columbia Centre for Heart Lung Innovation, St Paul's Hospital, Vancouver, BC, Canada [77]University of Groningen, Groningen, University Medical Center Groningen, Department of Pathology and Medical Biology, GRIAC Research Institute, The Netherlands [78]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester LE1 7RH, UK [79]National Institute for Health Research (NIHR) Leicester Respiratory Biomedical Research Unit, Glenfield Hospital, Leicester, UK [80]SpiroMeta Consortium see Supplemental Materials for full list of participating members [81]deCODE Genetics, Amgen Inc., Reykjavik, Iceland [82]Behavioral and Urban Health Program, Behavioral Health and Criminal Justice Division, RTI International, Research Triangle Park, North Carolina, USA [83]Department of Psychiatry, Washington University School of Medicine, St. Louis, Missouri, USA [84]Duncan Cancer Center, Baylor College of Medicine, Houston, TX 77030 [85]Research Computing Division, RTI International, Research Triangle Park, North Carolina, USA [86]Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA [87]Program and Behavioral Health and Criminal Justice Division, RTI International, Research Triangle Park, North Carolina, USA [88]Department of Epidemiology, University of Washington, 1959 NE Pacific Street, Health Sciences Bldg., F-247B, Box 357236, Seattle, WA 98195 [89]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, 02115 [90]Departments of Pharmacology and Toxicology & Psychiatry, Toronto, Ontario, Canada [91]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, Ontario, Canada [92]Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia [93]Department of Molecular Medicine, Laval University, Québec, Canada

Acknowledgments

Transdisciplinary Research for Cancer in Lung (TRICL) of the International Lung Cancer Consortium (ILCCO) was supported by (U19-CA148127 and CA148127S1). The ILCCO data harmonization is supported by Cancer Care Ontario Research Chair of Population Studies to R. H. and Lunenfeld-Tanenbaum Research Institute, Sinai Health System.

The TRICL-ILCCO OncoArray was supported by in-kind genotyping by the Centre for Inherited Disease Research (26820120008i-0-26800068-1).

IARC acknowledges and thanks V. Gaborieau, M. Foll, L. Fernandez-Cuesta, P. Chopard, T. Delhomme and A. Chabrier for their technical assistance in this project.

The authors would like to thank the staff at the Respiratory Health Network Tissue Bank of the FRQS for their valuable assistance with the lung eQTL dataset at Laval University. The lung eQTL study at Laval University was supported by the Fondation de l'Institut universitaire de cardiologie et de pneumologie de Québec, the Respiratory Health Network of the FRQS, the Canadian Institutes of Health Research (MOP – 123369). Y.B. holds a Canada Research Chair in Genomics of Heart and Lung Diseases.

References

1. Ferlay, J., et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer; Lyon, France: 2013.
2. Mucci LA, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. Jama. 2016; 315:68–76. [PubMed: 26746459]
3. Timofeeva MN, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Hum Mol Genet. 2012; 21:4980–95. [PubMed: 22899653]
4. Wang Y, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet. 2014; 46:736–41. [PubMed: 24880342]
5. Amos CI, et al. The OncoArray Consortium. A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev. 2017; 26:126–135. [PubMed: 27697780]
6. Wang Y, et al. Deciphering associations for lung cancer risk through imputation and analysis of 12,316 cases and 16,831 controls. Eur J Hum Genet. 2015; 23:1723–8. [PubMed: 25804397]
7. Durham AL, Adcock IM. The relationship between COPD and lung cancer. Lung Cancer. 2015; 90:121–7. [PubMed: 26363803]
8. Yuan H, et al. A Novel Genetic Variant in Long Non-coding RNA Gene NEXN-AS1 is Associated with Risk of Lung Cancer. Sci Rep. 2016; 6:34234. [PubMed: 27713484]
9. Barrett JC, et al. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat Genet. 2008; 40:955–62. [PubMed: 18587394]
10. Franke A, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 2010; 42:1118–25. [PubMed: 21102463]
11. Jostins L, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012; 491:119–24. [PubMed: 23128233]
12. McGovern DP, et al. Fucosyltransferase 2 (FUT2) non-secretor status is associated with Crohn's disease. Hum Mol Genet. 2010; 19:3468–76. [PubMed: 20570966]

13. Yang SK, et al. Genome-wide association study of Crohn's disease in Koreans revealed three new susceptibility loci and common attributes of genetic susceptibility across ethnic populations. Gut. 2014; 63:80–7. [PubMed: 23850713]
14. Brenner DR, et al. Hierarchical modeling identifies novel lung cancer susceptibility variants in inflammation pathways among 10,140 cases and 11,012 controls. Hum Genet. 2013; 132:579–89. [PubMed: 23370545]
15. Moulton EA, Elman I, Becerra LR, Goldstein RZ, Borsook D. The cerebellum and addiction: insights gained from neuroimaging research. Addict Biol. 2014; 19:317–31. [PubMed: 24851284]
16. Yu CT, et al. The novel protein suppressed in lung cancer down-regulated in lung cancer tissues retards cell proliferation and inhibits the oncokinase Aurora-A. J Thorac Oncol. 2011; 6:988–97. [PubMed: 21566536]
17. Fernandez-Cuesta L, et al. CD74-NRG1 fusions in lung adenocarcinoma. Cancer Discov. 2014; 4:415–22. [PubMed: 24469108]
18. Lan Q, et al. Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. Nat Genet. 2012; 44:1330–5. [PubMed: 23143601]
19. Landi MT, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am J Hum Genet. 2009; 85:679–91. [PubMed: 19836008]
20. Truong T, et al. Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. J Natl Cancer Inst. 2010; 102:959–71. [PubMed: 20548021]
21. Walsh KM, et al. Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk. Nat Genet. 2014; 46:731–5. [PubMed: 24908248]
22. Codd V, et al. Identification of seven loci affecting mean telomere length and their association with disease. Nat Genet. 2013; 45:422–7. 427e1–2. [PubMed: 23535734]
23. Zhang C, et al. Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. Hum Mol Genet. 2015; 24:5356–66. [PubMed: 26138067]
24. Walsh KM, et al. Longer genotypically-estimated leukocyte telomere length is associated with increased adult glioma risk. Oncotarget. 2015; 6:42468–77. [PubMed: 26646793]
25. Fehringer G, et al. Cross-cancer genome-wide analysis of lung, ovary, breast, prostate and colorectal cancer reveals novel pleiotropic associations. Cancer Res. 2016
26. Amundadottir LT, et al. Cancer as a complex phenotype: pattern of cancer distribution within and beyond the nuclear family. PLoS Med. 2004; 1:e65. [PubMed: 15630470]
27. Lindelof B, Eklund G. Analysis of hereditary component of cancer by use of a familial index by site. Lancet. 2001; 358:1696–1698. [PubMed: 11728548]
28. Li Y, et al. FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. BMC Bioinformatics. 2016; 17:122. [PubMed: 26961892]
29. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007; 39:906–13. [PubMed: 17572673]
30. Timofeeva MN, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Hum Mol Genet. 2012; 21:4980–95. [PubMed: 22899653]
31. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software. 2010; 36:1–48.
32. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011; 88:76–82. [PubMed: 21167468]
33. Li J, Ji L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity. 2005; 95:221–7. [PubMed: 16077740]
34. Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet. 2007; 81:208–27. [PubMed: 17668372]
35. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348:648–60. [PubMed: 25954001]

36. Soler Artigas M, et al. Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. Nat Comm. 2015; 6:8658.

37. Pooley KA, et al. Telomere length in prospective and retrospective cancer case-control studies. Cancer Res. 2010; 70:3170–6. [PubMed: 20395204]

38. Bulik-Sullivan B, et al. An atlas of genetic correlations across human diseases and traits. Nat Genet. 2015; 47:1236–41. [PubMed: 26414676]

39. Bulik-Sullivan BK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015; 47:291–5. [PubMed: 25642630]

40. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nat Genet. 2010; 42:441–7. [PubMed: 20418890]

41. Hemminki K, Bermejo JL. Relationships between familial risks of cancer and the effects of heritable genes and their SNP variants. Mutat Res. 2005; 592:6–17. [PubMed: 15990124]

42. Bahcall OG. iCOGS collection provides a collaborative model. Foreword. Nature genetics. 2013; 45:343. [PubMed: 23535721]

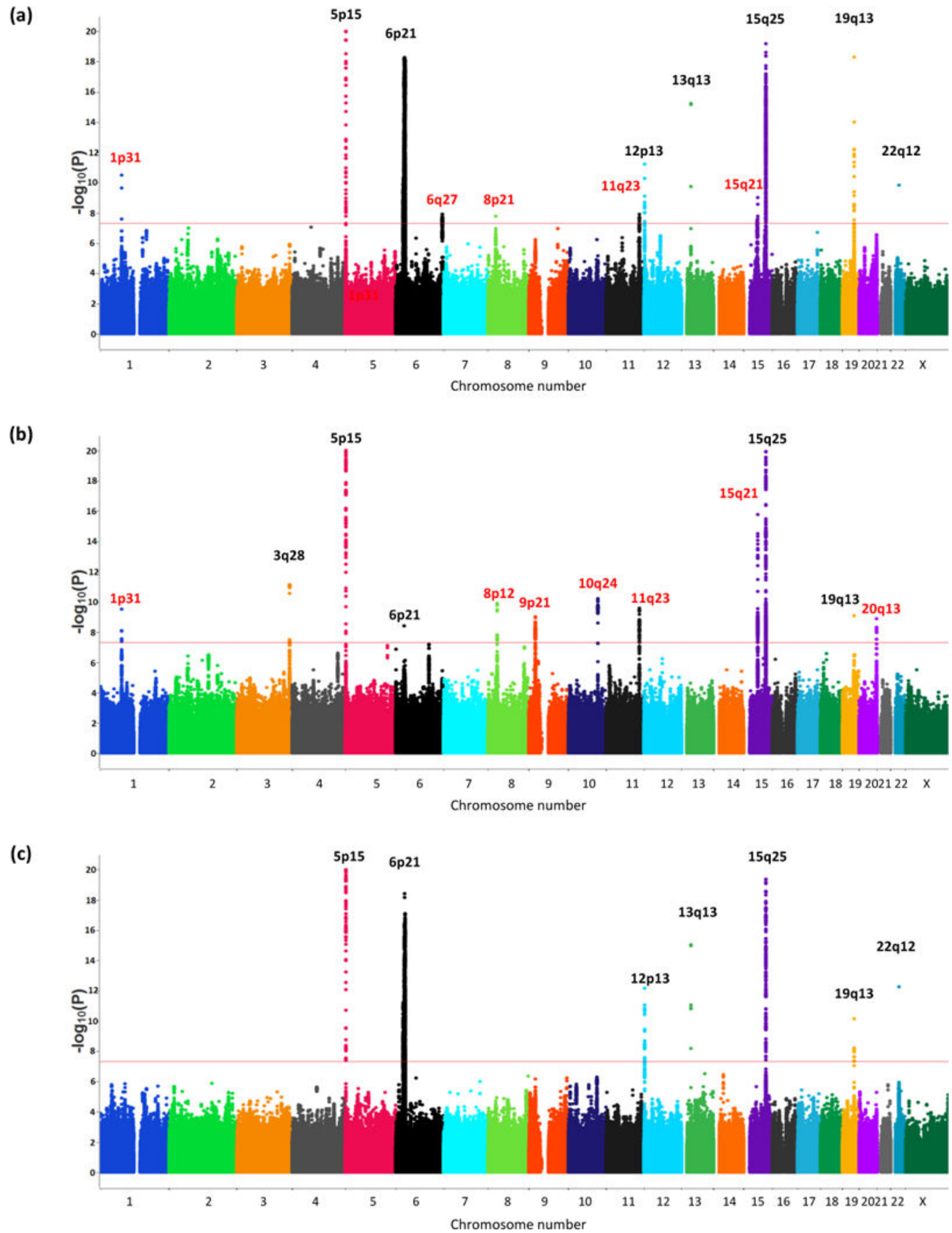


Figure 1. Manhattan plots of lung cancer risk overall and by histological subtypes

(a) lung cancer risk overall, 29,266 cases and 56,450 controls (b) adenocarcinoma, 11,273 cases and 55,483 controls (c) squamous cell carcinoma 7,426 cases and 55,627 controls. Each locus is annotated by their cytoband locations. The X-axis represents chromosomal locations and the Y-axis represents −log10(P-value). Black denotes the previously known loci and Red denotes the new loci identified in this analysis

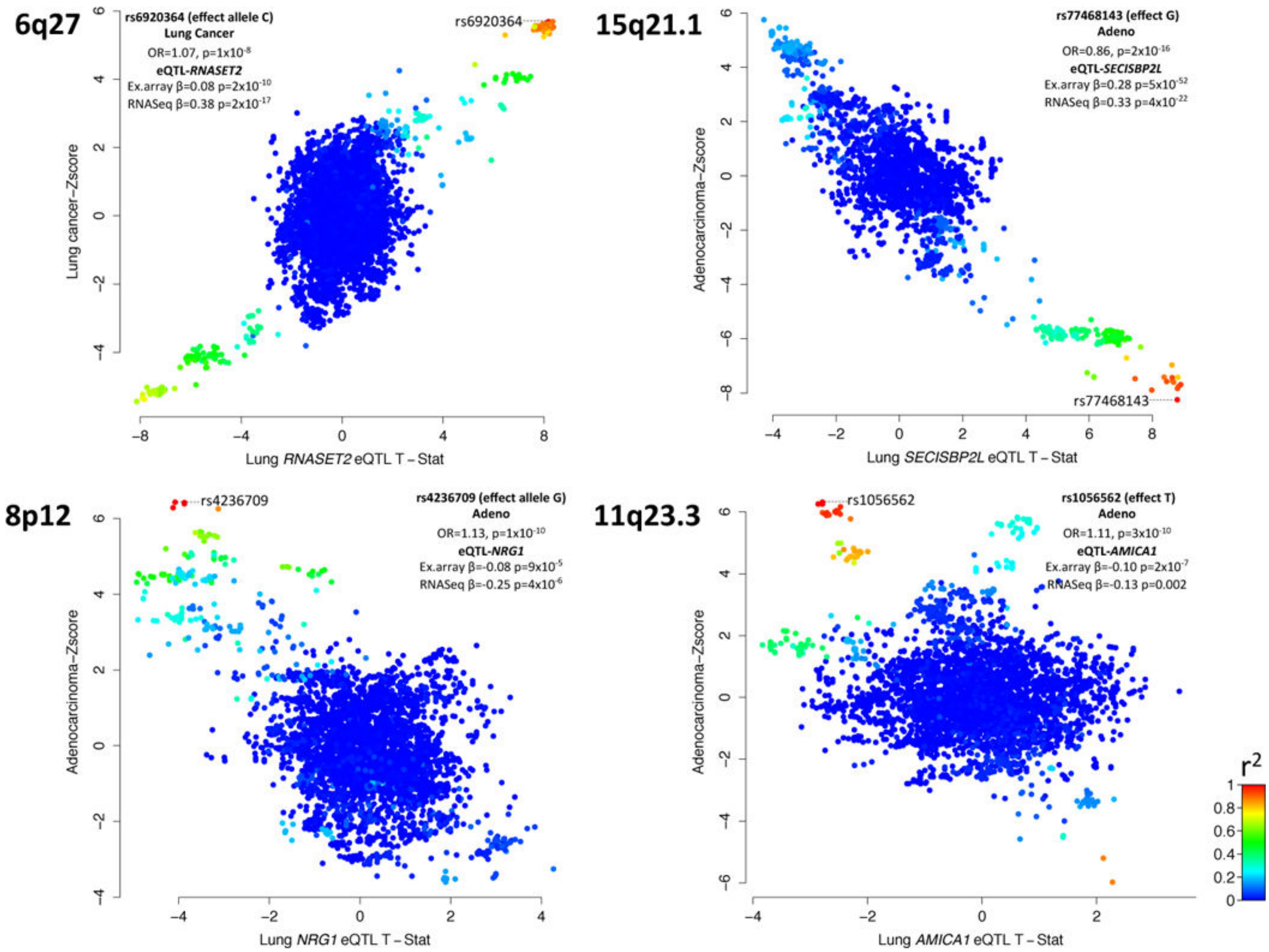
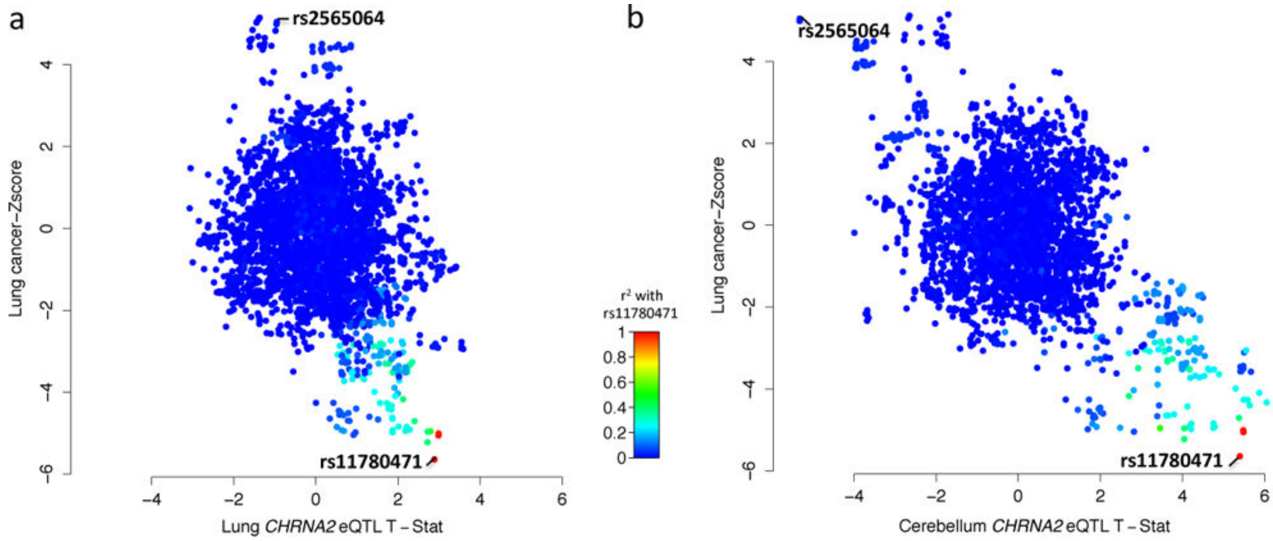


Figure 2. Scatter plots comparing variants across the 6q27, 15q21.1, 8p12 and 11q23.3 susceptibility loci and (Y-axis) their associated with lung cancer (or lung adenocarcinoma, as relevant) and (X-axis) the lung cis–eQTL (GTEx)

Each variant (dot) is colored relative the degree of linkage disequilibrium (r^2) with sentinel lung cancer variant (marked) at that locus. Indented table, association between sentinel variant and lung cancer (or histological subtype) as well as the eQTL evidence in lung epithelium for the microarray (Laval, UBC, Groningen) and RNAseq (NCI and GTEx) cohorts. At 6q27, 15q21.1 and 8p12, the variants associated with lung cancer also tend to be those that that are lung cis-eQTL's for *RNASET2*, *SECISBP2L* and *NRG1*, respectively. At 11q23.3, while the sentinel variant (rs1056562) is a lung cis-eQTL for *AMICA1*, additional variants are *AMICA1* lung cis-eQTL's, but not associated with lung adenocarcinoma and *vice versa* suggesting an alternate candidate gene may be responsible for this association or a pleiotropic effect at this locus.



c

***CHRNA2* eQTL rs11780471**

GTEx Region - tissue	P	Braineac Region - tissue	P
Brain cortex		**Brain cortex**	
Frontal Cortex	0.67	Frontal cortex	0.77
Cortex	0.51	Occipital cortex	0.11
Anterior cingulate	0.29	Temporal cortex	0.31
Hippocampus	0.09	Hippocampus	0.36
Hypothalamus	0.36	Thalamus	0.03
Basal Ganglia		Intralobular WM	0.02
Caudate	0.08	Putamen	0.58
Nucleus accumbens	0.62	**Basal Ganglia**	
Putamen	0.54	Substantia nigra	0.92
Cerebellum		**Cerebellum**	
Cerebellar Hemi	$3x10^{-6}$	Cerebellar cortex	$2x10^{-13}$
Cerebellum	$7x10^{-7}$	Medulla	0.03

d

rs11780471 G allele

Phenotype	Samples	OR	P
Lung cancer overall	85,716	1.15	$1.7x10^{-8}$
Ever vs never smoker	233,647	1.05	$1.1x10^{-3}$
Smoking cessation	137,472	0.96	0.07
Difficulty for 1 day	11,231	0.98	0.78
Time first cigarette	11,211	1.02	0.75

Phenotype	Samples	Beta	P
Age of initiation	84,977	-0.008	$9.0x10^{-4}$
Cigarettes per day	129,227	-0.0075	0.44
FTND	17,074	0.0151	0.40

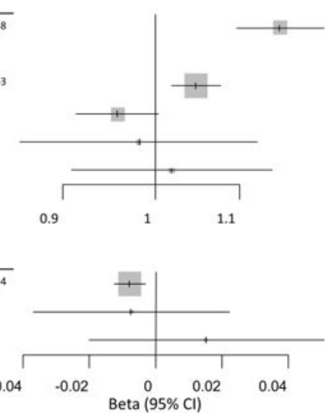


Figure 3. eQTL and smoking behavior analysis of the 8p21 lung cancer susceptibility locus. Upper panel

Scatter plots of variants across the 8p21 locus and their associated with lung cancer (Y-axis) and *CHRNA2* eQTL (X-axis) in lung epithelial tissues (panel a) and *CHRNA2* eQTL in brain cerebellum tissues (panel b). **Panel C.** eQTL association between rs11780471 across tissues from different parts of the brain from GTEx and Braineac consortia noting *CHRNA2* cis- eQTL effect appears restricted to the brain cerebellum. **Panel D.** Association between rs11780471 and smoking phenotypes, noting evidence for association between smoking status (ever vs never) and age of initiation, with lung cancer risk allele carriers (G) more likely to be ever smokers and take up smoking earlier. Fagerström Test for Nicotine Dependence (FTND) index, error bars indicate the 95% confidence intervals.

Table 1

Demographic characteristics of the participating studies after quality control filters

	Lung cancer patients		Controls	
	number	*(%)*	number	*(%)*
OncoArray studies- passed QC	14803	*(51)*	12262	*(22)*
Published GWAS studies[a]	14463	*(49)*	44188	*(78)*
Total	**29266**		**56450**	
Age				
<=50	3112	*(12)*	6032	*(12)*
>50	23025	*(88)*	44075	*(88)*
Sex				
Male	18208	*(62)*	27178	*(53)*
Female	11059	*(38)*	24069	*(47)*
Smoking status				
Never	2355	*(9)*	7504	*(31)*
Ever	23223	*(91)*	16964	*(69)*
Former	9037	*(35)*	8554	*(35)*
Current	13356	*(52)*	7477	*(31)*
Histology [c]				
Adenocarcinoma	11273	*(39)*	55483 [b]	
Squamous cell carcinoma	7426	*(25)*	55627 [b]	
Small cell carcinoma	2664	*(9)*	21444 [b]	

[a] Previous GWAS studies include IARC, MDACC, SLRI, ICR, Harvard, ATBC, CPSII, German and deCODE studies.

[b] Number of non-cancer individuals included in the corresponding histology-specific analysis.

[c] The remaining 27% includes other histological subsets, such as large cell carcinoma, non-small cell lung cancer, NOS, mixed histology, and unknown.

Table 2

The association between sentinel variants representing each lung cancer locus and lung cancer risk.

Strata	Locus*	rs number	Gene	Allele[a]	Imputed versus Oncoarray genotyped	Candidate Oncoarray Customized panel	EAF	OR	95%CI	P-value	CPD p- value	FTND p- value	FEV1 p- value	FVC p- value	FEV1/FVC p- value
Lung	1p31.1*	rs71658797	*FUBP1*	T_A	Oncoarray	*No*	0.103	1.1	1.09–1.18	3.3E-11	0.056	0.334	0.445	0.898	0.334
Lung	6q27*	rs6920364	*RNASET2*	G_C	Imputed	*eQTL*	0.456	1.1	1.05–1.10	1.3E-08	0.833	0.104	0.927	0.876	0.986
Lung	8p21.1*	rs11780471	*CHRNA2*	G_A	Imputed	*Lung*	0.060	0.9	0.83–0.91	1.7E-08	0.646	0.403	**6.9E-04**	0.055	**0.016**
Lung	13q13.1	rs11571833	*BRCA2*	A_T	Imputed	*Lung*	0.011	1.6	1.43–1.80	6.1E-16	0.890	0.312	0.601	0.667	0.237
Lung	15q21.1*	rs66759488	*SEMA6D*	G_A	imputed	*Lung*	0.362	1.1	1.05–1.10	2.8E-08	0.266	0.888	0.739	0.200	0.202
Lung	15q25.1	rs55781567	*CHRNA5*	C_G	Imputed	*Lung*	0.367	1.3	1.27–1.33	3.1E-103	**6.8E-38**	**9.7E-16**	**7.2E-03**	**0.020**	0.144
Lung	19q13.2^	rs56113850	*CYP2A6*	C_T	Oncoarray	*Lung*	0.440	0.9	0.86–0.91	5.0E-19	**8.1E-20**	**7.5E-04**	0.822	0.826	0.319
Adeno	3q28	rs13080835	*TP63*	G_T	Imputed	*Lung*	0.493	0.9	0.87–0.92	7.5E-12	0.803	0.336	0.135	0.445	0.834
Adeno	5p15.33	rs7705526	*TERT*	C_A	Oncoarray	*All*	0.342	1.3	1.21–1.29	3.8E-35	0.511	0.738	0.292	**0.038**	0.657
Adeno	8p12*	rs4236709	*NRG1*	A_G	Imputed	*eQTL*	0.218	1.1	1.09–1.18	1.3E-10	0.991	0.957	0.503	0.151	0.403
Adeno	9p21.3*	rs885518	*CDNK2A*	A_G	Imputed	*Several*	0.101	1.2	1.11–1.23	9.96E-10	0.904	0.321	0.421	0.096	0.146
Adeno	10q24.3*	rs11591710	*OBFC1*	A_C	Imputed	*Lung*	0.137	1.2	1.11–1.22	6.3E-11	0.500	0.152	**0.027**	**0.019**	0.533
Adeno	11q23.3*	rs1056562	*AMICA1*	C_T	Oncoarray	*Breast*	0.473	1.1	1.07–1.14	2.8E-10	0.717	0.538	0.449	0.718	**0.039**
Adeno	15q21.1*	rs77468143	*SECISBP2L*	T_G	Imputed	*No*	0.253	0.9	0.83–0.89	1.7E-16	0.071	0.184	**4.9E-03**	0.440	**1.4E-03**
Adeno	20q13.33*	rs41309931	*RTEL1*	G_T	Imputed	*Prost/ColR*	0.117	1.2	1.11–1.23	1.3E-09	0.146	0.939	0.964	0.657	0.284
SQC	6p21.33	rs116822326	*MHC*	A_G	Imputed	*Lung*	0.155	1.3	1.19–1.32	3.8E-19	0.392	0.774	0.132	0.498	0.103
SQC	12p13.33	rs7953330	*RAD52*	G_C	Oncoarray	*Lung*	0.315	0.9	0.83–0.90	7.3E-13	0.800	0.463	0.019	**3.3E-03**	0.424
SQC	22q12.1	rs17879961	*CHEK2*	A_G	Oncoarray	*Lung*	0.005	0.4	0.32–0.52	5.7E-13	0.441	0.360	**0.041**	**0.040**	0.805

* denote novel locus identified to GWAS significance by this study; a, reference_effect. Bolded p-values indicate significant associations with consistent direction as expected. Genome positions relative to GRCh37, EAF, effective allele frequency; OR, odds (log additive) ratio; 95%CI, 95% confidence interval. P-value, based on fixed-effect meta-analysis adjusted for age, sex and genetically derived ancestry; CPD, cigarette per day; FTND, Fagerström Test for Nicotine Dependence; FEV1, forced expiratory volume in 1 second; FVC, forced vital capacity.

Adeno, adenocarcinoma; SQC, squamous cell carcinoma.

^ marker had an acceptable, but not ideal concordance rate (see Supplementary Note)