Supplementary note.

Study Populations

ATBC. The Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC), is a randomized primary prevention trial including 29,133 male current smokers enrolled in Finland between 1985 and 1993. The purpose of the study was to determine whether selected vitamin supplements would prevent lung and other cancers. The 50- to 69-year-old participants were randomized in a factorial design to take a pill daily for five to eight years that contained one of the following: 50 milligrams (mg) of dl-alpha-tocopheryl acetate, 20 mg of all-trans-beta-carotene, both, or a placebo. The trial ended in April 1993. However, in order to evaluate the long-term effects of these vitamins on cancer incidence, and overall and cause-specific mortality, participants were followed after the trial ended using data from the national registries in Finland. The researchers acquired additional data for cancer incidence and mortality related to specific causes through December 2012 and for total mortality through December 2013 (see http://atbcstudy.cancer.gov).

CAPUA. The CAncer de PUlmon en Asturias (CAPUA study) is a hospital-based case-control study conducted in Asturias, Spain by the University of Oviedo. Lung cancer cases were recruited in three main hospitals of Asturias, following an identical protocol from 2002 to 2012. Eligible cases were incident cases of histologically confirmed lung cancer between 30 and 85 years of age and residents in the geographical area of each participating hospital. Controls were selected from patients admitted to those hospitals with diagnoses unrelated to the exposures of interest and individually matched by ethnicity, gender, age (± 5 years) and hospital. Epidemiologic data were collected personally through computer-assisted questionnaires by trained interviewers during the first hospital admission. Structured questionnaires collected information on sociodemographic characteristics, recent and prior tobacco use, environmental exposure (air pollution and passive smoking), diet, personal and family history of cancer, and occupational history from each participant. Peripheral blood samples (or mouthwash samples when they refused to donate blood) were collected from all participants. Coding of histology was based on 2001 WHO/IASLC. Genomic DNA was extracted based on standard protocol.

CARET. The Carotene and Retinol Efficacy Trial (CARET) was a randomized, double-blind, placebo-controlled trial of the cancer prevention efficacy and safety of a daily combination of 30 mg of beta-carotene and 25,000 IU of retinyl palmitate in 18,314 persons at high risk for lung cancer. CARET began in 1985, and the intervention was halted in January 1996, 21 months ahead of schedule, with the twin conclusions for definitive evidence of no benefit and substantial evidence of a harmful effect of the intervention on both lung cancer incidence and total mortality. CARET continued to follow and collect endpoints on their participants through 2005. Pathology reports and medical records were reviewed to confirm cancer endpoints, and death certificates obtained to capture cause of death. During the active intervention phase of CARET, serum, plasma, whole blood, and lung tissue specimens were collected on participants. These biospecimens make up the CARET Biorepository. For the OncoArray Project, CARET provided DNA extracted from whole blood of lung cancer cases and controls matched on age at baseline (+/- 4 years), sex, race, baseline smoking status, history of occupational asbestos exposure (asbestos vs heavy smoker), and year of enrollment (2-year intervals).

The Canadian screening study. It includes the nested case-control samples from 3 screening programs: **IELCAP-Toronto:** Ever smokers of more than 10 pack-years age 50 and above were eligible for the I-ELCAP screening program since 2003, and a total of 4782 individuals have been enrolled in the Greater Toronto Area. Participants were administered a LDCT scan along with a standard study questionnaire at baseline. Blood samples were systematically collected at baseline since 2006. Participants who had an abnormality in a CT scan were followed up every 1 to 2 years. The screening program was organized by the Princess Margaret Hospital. **PanCan:** Ever smokers between the ages of 50-75 with no previous history of invasive cancer are eligible to participate in the study. The study was carried out across Canada in Vancouver, Calgary, Hamilton, Toronto, Ottawa, Quebec, Halifax, and St. John's. A total of 2537 smokers have been screened from 2008 to 2011. All study participants completed a detailed questionnaire, spirometry, collection of blood specimens for biomarker measurement and LDCT at baseline. All participants are followed for a minimum of 3 years. On yearly follow up, an updated shorter questionnaire is administered, blood is collected and CT scans are performed. Blood samples are available from all 2537 individuals. **BCCA Screening Program:** Between 1990 to 2007, 4274 smokers above 40 years old who had smoked 20 pack-years or more were enrolled at BCCA. Upon enrollment, subjects completed a questionnaire for their lifestyle and medical history. Baseline spirometry was conducted using a flow-sensitive spirometer in accordance with the American Thoracic Society recommendations. Since 2000, a LDCT was obtained in 2440 individuals. The participants were followed prospectively to determine whether they developed lung cancer. A total of 9759 individuals participated in the CT screening program in Canada from these 3 programs. The samples included in this project are based on a subset of nested lung cancer case-control pairs based on 1:2 ratio.

Copenhagen study. Cases were diagnosed histologically from 2012 and onwards with treatable lung cancer. Controls were participants of the Copenhagen General Population Study without lung cancer.

EAGLE. The Environment And Genetics in Lung cancer Etiology (EAGLE) study is a large population-based case-control study conducted in the Lombardy region of Italy between 2002 and 2005. Details of the study are previously described[1]. Briefly, primary lung cancer cases (n = 2098) were identified from 13 hospitals, which covered approximately 80% of incident lung cancer cases in the catchment area consisting of 216 municipalities. Case response rate was 86.6%. Approximately 95% of cases were confirmed pathologically or cytologically, and the remaining 5% were confirmed based on clinical history and imaging. Detailed histologic classification was recorded for all cases. Controls (n = 2120) were randomly selected from the Lombardy Regional Health Service database and were frequency-matched to cases based on sex, 5-year age group, and area of residence. Overall participation rate was 72.4%. The study protocol was approved by the Institutional Review Board of the US National Cancer Institute and the involved institutions in Italy. Informed consent was obtained for all subjects prior to study participation.

EPIC. The European Prospective Investigation into Cancer and Nutrition (EPIC) study. The European Prospective Investigation into Cancer and Nutrition (EPIC) study is a multi-center cohort study involving 521,000 study participants from 10 European countries [2,3]. The current study involved EPIC participants from 7 countries (Greece, Netherlands, UK, France, Germany, Spain, and Italy), including 1223 incident lung cancer cases and 1249 smoking matched controls.

Liverpool. The Liverpool Lung Project (LLP) is a case-control and cohort study which has recruited over 11,500 individuals since 1996 from the Liverpool region in the UK. Detailed epidemiological and clinical data is collected with associated specimens (i.e. tumour tissue, blood, plasma, sputum, bronchial lavage and oral brushings). The participants have completed a detailed lifestyle questionnaire at recruitment, with repeat questionnaires at intervals; updated data on clinical outcome and hospital events are collected through the Health and Social Care Information Centre (including Office of National Statistics mortality data, Cancer Registry and Health Episode Statistics). The project is registered on the UK National Institute for Health Research (NIHR) lung cancer portfolio and has all the required ethical approvals and sponsorship arrangements in place. The lung tumours were reviewed by the reference pathologist.

German Lung Cancer Study. The German Lung Cancer Study was made up of three independent German studies. Two are detailed below as they were included in OncoArray genotyping. OncoArray genotyping was carried out at the Genome Analysis Center of the Helmholtz Zentrum Muenchen. After quality control 1041 cases (LUCY-study: n=806, Heidelberg lung cancer case-control study: n=235) entered the data analysis for OncoArray. Also available were n=601 relative controls (first degree relatives), but these were removed during the QC process described below.

a. LUCY-study (Helmholtz Zentrum Muenchen, PIs Bickeböller, Brüske)
LUCY (LUng Cancer in the Young) is a multicenter study with 31 recruiting hospitals in Germany. The study is conducted by the Institute of Epidemiology, Helmholtz Zentrum Muenchen, and the Department of Genetic Epidemiology, Medical School, University of Göttingen). The LUCY-study provides access to a nationwide, population based family and a case-control sample (control population KORA, described below) of lung cancer patients aged 50 years or younger at diagnosis. Detailed epidemiologic data have been collected including data on medical history, education, family history of cancer and smoking exposure by phase assessment. Blood samples are taken and DNA and lymphoblastoid cell lines are prepared of all cases and controls and of parts of the relatives. Phenotype data of 847 young patients with primary lung cancer and 5524 relatives have been collected.

b. Heidelberg Lung Cancer Case-control Study (German Cancer Research Center, PI Risch)
As part of an ongoing, hospital-based, case-control study, the DKFZ has recruited over 2000 LC cases at and in collaboration with the Thoraxklinik Heidelberg, including 300 LC cases with onset of disease at the age of 50. Approximately 750 hospital-based controls have also been recruited. Data on occupational exposure, tobacco smoking, educational status, and for a subgroup also on family history of lung cancer, assessed by a self-administered questionnaire is available. Blood samples have been taken, and DNA has been extracted.

Kentucky Lung Cancer Research Initiative. The Kentucky Lung Cancer Research Initiative is a study conducted by the Markey Cancer Center Cancer Center and the University of Kentucky using a population-based, case-control framework to study the extraordinarily high rates of lung cancer in Southeastern, Appalachian Kentucky. Cancer cases were recruited from the Kentucky Cancer Registry at the time of diagnosis and controls were recruited from a random digit dialing process from the same region. Study accrual began in January 5, 2012 and completed on September 5, 2014 and 520 subjects were recruited in a 4:1 ratio of controls: cases from Appalachian Kentucky. Of the 520 subjects recruited, 231 are included in the OncoArray analysis, including all 93 cancer cases, and 123 controls. Newly diagnosed lung cancer cases and controls underwent blood, toenail

(for trace element analysis), urine, buffy coat, water, soil, and radon collection, residence GPS mapping, as well as an extensive epidemiologic, occupational, and health history questionnaire. Clinical Trials.gov Identifier: NCT01648166.

The Harvard Lung Cancer Study (HLCS). HLCS is a case-control study based at Mass General Hospital (MGH) in Boston, Massachusetts from 1992 to 2004. Details of the study were described previously[4]. Briefly, eligible cases included any person over the age of 18 years with a diagnosis of primary lung cancer that was further confirmed by an MGH lung pathologist. Controls were recruited from the friends or spouses of cancer patients or the friends or spouses of other surgery patients in the same hospital. Potential controls were excluded from participation if they had a diagnosis of any cancer (other than non-melanoma skin cancer). Interviewer-administered questionnaires, a modified version of the standardized American Thoracic Society respiratory questionnaire, collected information on demographics, medical history, family history of cancer, smoking history, and a detailed work history, including job titles and tasks. Genome-wide genotype data were first generated using Illumina Human 610-Quad BeadChips and then imputed by MACH against the 1000 Genome Project dataset (http://browser.1000genomes.org/index.html). The Institutional Review Board of MGH and the Human Subjects Committee of the Harvard School of Public Health approved the study.

Israel study (NICCC-LCA). The NICCC-LCA study is an ongoing case-control study of newly diagnosed lung cancer cases of any histology and population age/sex/ethnicity-matched "healthy" controls. All participants undergo face-to-face interviews, provide a venous blood sample (separated into DNA, Sera, lymphocytes) after signing an IRB-approved form. Histology reports, FFPE blocks and clinical follow-up are available for most cancer cases.

MDACC study. Lung cancer cases and frequency-matched controls were ascertained from a large ongoing case-control study at The University of Texas MD Anderson Cancer Center (UTMDACC) since 1991. Detailed study description was provided previously[5]. In brief, cases were newly-diagnosed and histologically confirmed lung cancer patients recruited from UTMDACC. Controls were healthy individuals without a history of cancer (except for non-melanoma skin cancer) and recruited from the Kelsey-Seybold Clinics, the largest private multispecialty physician group in the Houston metropolitan area. Controls were frequency-matched to cases on age (±5 years), sex, and race/ethnicity. After providing written informed consent, each study participants completed an in-person interview by staff interviewers to collect information on demographics, smoking status, etc. Blood samples were also drawn from all the study participants. This study was approved by institutional review boards of UTMDACC and Kelsey-Seybold Clinics.

The Malmö Diet and Cancer Study (MDCS). The Malmö Diet and Cancer Study (MDCS) is a population-based prospective cohort study that between 1991 and 1996 recruited men and women aged 44 to 74 years of age living in Malmö, Sweden. The main goal of the MDCS is to study the impact of diet on cancer incidence and mortality. It consists of a baseline examination including dietary assessment, a self-administered questionnaire, anthropometric measurements and collection of blood samples. A total of 165 incident lung cancer cases and 174 individually smoking-matched controls were available for this analysis.

The Multiethnic Cohort (MEC). The MEC Study includes 215,251 men and women aged 45-74 years at recruitment, primarily from five ethnic/racial groups – African Americans and Latinos mostly recruited from CA (mainly from Los Angeles County) and Japanese Americans, Native Hawaiians and whites (mostly recruited

from HI). The cohort was assembled in 1993-1996 by mailing a self-administered questionnaire to persons identified primarily through driver's license files. The baseline questionnaire obtained information on demographics, anthropometry, smoking history, medical and reproductive histories, family history of cancer, diet and physical activity. Incident cancer cases are identified by regular linkage with the State of California Cancer Registry and the Hawaii Tumor Registry, both members of the SEER Program of the NCI. In 2001-2006, a prospective biorepository was assembled by collecting a pre-diagnostic blood specimen from 67,594 surviving MEC members. At the time of blood collection, a short questionnaire was administered that included information on smoking during the previous 15 days. For this study, cases were all lung cancer cases incident to blood draw and diagnosed before December 2012. For each case, a control was selected among unaffected MEC participants who were alive at time of the case's diagnosis and matched on study site, sex, race/ethnicity, age (age at diagnosis for cases; age at blood collection for controls), and date of blood collection.

The Mount-Sinai Hospital-Princess Margaret Study (MSH-PMH). MSH-PMH was conducted in the greater Toronto area from 2008 to 2013. Lung cancer cases were recruited at the hospitals in the network of the University of Toronto. Controls were selected randomly from individuals registered in the family medicine clinics databases and were frequency matched with cases on age and sex. All subjects were interviewed, and information on lifestyle risk factors, occupational history and medical and family history was collected using a standard questionnaire. Tumors were centrally reviewed by the reference pathologist (a member of the International Association for the Study of Lung Cancer (IASLC) committee) and a second pathologist in the University Health Network. If the reviews conflicted, a consensus was arrived at after discussion. Coding of histology was based on 2001 WHO/IASLC. Genomic DNA was extracted based on standard protocol.

The New England Lung Cancer Study (NELCS). NELCS is a population-based case-control study of lung cancer among residents of Northern and Central New Hampshire counties and the bordering region of Vermont. Cases with histologically confirmed primary incident lung cancer were identified from 2005 to 2007 using the New Hampshire State Cancer Registry and the Dartmouth-Hitchcock Medical Center (DHMC) Tumor Registry. Control participants were identified using a commercial database and matched to lung cancer cases within 5-year age groups, sex and county. Genomic DNA was isolated from blood or buccal specimens provided by consenting participants. The study complied with requirements of the Dartmouth College's Committee for Protection of Human Subjects

The Nijmegen Lung Cancer Study. The Netherlands patients with lung cancer were identified through the population-based cancer registry of the Netherlands Comprehensive Cancer Organisation in Nijmegen, the Netherlands. Patients who were diagnosed in one of three hospitals (Radboud university medical center and Canisius Wilhelmina Hospital in Nijmegen and Rijnstate Hospital in Arnhem) since 1989 and who were still alive at April 15th, 2008 were recruited for a study on gene-environment interactions in lung cancer. 458 patients gave informed consent and donated a blood sample. This case series was expanded with 94 patients to a total of 552 by linking three other studies to the population-based cancer registry in order to identify new occurrences of lung cancer among the participants of these other studies. All three other studies (i.e., POLYGENE, the Nijmegen Biomedical Study, and the Radboudumc Urology Outpatient Clinic Epidemiology Study) were initiated to study genetic risk factors for disease and participants to these studies gave general informed consent for DNA-related research and linkage with disease registries. Information on histology, stage of disease, and age at diagnoses was obtained through the cancer registry. Lifestyle information was collected

through a structured questionnaire and whole blood for DNA isolation was collected by the regional thrombosis services.

The cancer-free controls (46% males) were selected from participants of the "Nijmegen Biomedical Study" (NBS;), an age- and sex-stratified random sample of the general population of the municipality of Nijmegen, The Netherlands. All participants provided extensive lifestyle information by structured questionnaires and blood samples for DNA isolation, serum and plasma. All controls are of self-reported European descent. The study protocols of the NBS were approved by the Institutional Review Board of the Radboudumc and all study subjects signed a written informed consent form.

Norway. Early-stage NSCLC cases and healthy controls at the time of enrollment were Caucasians of Norwegian origin and were recruited from the same geographical region (Western Norway). The patients were enrolled in the study, whenever practically feasible among patients admitted for lung cancer at the Haukeland University Hospital in Bergen, Norway. The informed written consents covering analysis of molecular and genetic markers was signed by the patients prior to surgery. Only patients with histologically confirmed early-stage NSCLC were included in our study. The subjects included in this project are a subgroup recruited into the project "lung cancer genetics" at NIOH. The controls were recruited from the same geographical region of Western Norway and frequency-matched with cases on cumulative smoking dose (pack-years). Pack-years smoked [(20 cigarettes per day) x years smoked] were calculated to indicate the cumulative smoking dose. The Cases and controls were interviewed using similar questionnaires and were categorized as never smokers, ex-smokers or current smokers. Never smokers are subjects indicating having smoked less than 100 cigarettes in their life time. Ex-smokers were defined as those having quitted at least 1 year before sampling, and current smokers were those indicating that they were smokers at the time of sampling. The project has been approved by the Regional Committee for Medical and Health Research Ethics in Southern Norway in accordance with the WMA Declaration of Helsinki. The ethical approval covered access to the NSCLC databank.

Northern Sweden Health and Disease Study Cohort (NSHDS). The Northern Sweden Health and Disease Study (NSHDS) encompasses several prospective cohorts, the current study involving study participants from the Västerbotten Intervention Project (VIP), a sub-cohort within NSHDS. VIP is an ongoing prospective cohort and intervention study intended for health promotion of the general population of the Västerbotten County in northern Sweden (12). VIP was initiated in 1985 and all residents in the Västerbotten County were invited to participate by attending a health check-up at 40, 50 and 60 years of age. Participants were asked to complete a self-administered questionnaire including various demographic factors such as education, smoking habits, physical activity and diet. In addition, height and weight were measured and participants were asked to donate a fasting blood sample for future research. A total of 243 incident lung cancer cases and 266 individually smoking-matched controls were available for this analysis.

PLCO. The PLCO study, a randomized trial aimed at evaluating the efficacy of screening in reducing cancer mortality, recruited approximately 155,000 men and women age 55 to 74 years from 1992 to 20014. Screening for lung cancer among participants in the intervention arm included a chest x-ray at baseline followed by either three annual x-rays (for current or former smokers at enrollment) or two annual x-rays (for never smokers); participants in the control arm received routine health care. Screening-arm participants provided data on sociodemographic factors, smoking behavior, anthropometric characteristics, medical history, and family history of cancer, as well as blood samples annually for the first 6 years of the study (baseline [T0] and T1

through T5). Lung cancers were ascertained through annual questionnaires mailed to the participants, and positive reports were followed up by abstracting medical records or death certificates. Follow-up in the trial as of July 2009 was 96.7%. Patients were excluded because of missing baseline questionnaire, previous history of any cancer, diagnosis of multiple cancers during follow-up, missing smoking information at baseline, missing consent for utilization of biologic specimens for etiologic studies, or unavailability/insufficient quantity of serum or DNA specimens.

Resolucent. The Resource for the Study of Lung Cancer Epidemiology in North Trent is an ongoing study conducted in Sheffield from 2006 and due to complete recruitment in 2016. The study recruited pathologically confirmed lung cancer cases diagnosed at age 60 years or younger and family matched controls. Lung cancer cases diagnosed at ages older than 60 years were recruited if they reported a family history of lung cancer. The cases and matched controls were recruited through several major cancer treatment centres, however, the majority were recruited in North Trent. All participants completed a detailed lifestyle questionnaire which included questions about occupational exposures, education, medical history and family history of cancer and lung disease. Participants also donated blood samples for DNA extraction. The ReSoLuCENT study has been funded by the Sheffield Hospitals Charity, Sheffield ECMC and Weston Park Hospital Cancer Charity.

The IARC L2 study. Lung cancer cases and controls were recruited through a multicentric case-control study coordinated by the International Agency for Research on Cancer in Russia, Poland, Serbia, Czech Republic, and Romania from 2005 to 2013. Cases were incident cancer patients collected from general hospitals. Controls were recruited from individuals visiting general hospitals and out-patient clinics for disorders unrelated to lung cancer and/or its associated risk factors, or from the general population. Information on lifestyle risk factors, medical and family history was collected from subjects by interview using a standard questionnaire. All study participants provided written informed consent. The current study included 1,133 lung cancer cases and 1,117 controls genotyped on the OncoArray.

Total Lung Cancer (TLC) Study. Molecular Epidemiology of Lung Cancer Survival: The Total Lung Cancer (TLC) Study is a hospital-based study that included 458 lung cancer patients recruited for Moffitt Cancer Center's Total Cancer Care™ protocol between April 2006 and August 2010. Total Cancer Care™ is a multi-institutional observational study of cancer patients that prospectively collects self-reported demographic and clinical data, medical record information and blood samples for research purposes. All patients used in this cohort were recruited from the Thoracic Oncology Clinic at the Moffitt Cancer Center.

The Washington State University Lung Cancer Study. This study is a hospital case-control study of 511 subjects with newly-diagnosed (within 1 year of diagnosis) lung cancer and 820 race-, sex- and age-matched controls. Lung cancer cases were recruited from lung cancer clinics within the H. Lee Moffitt Cancer Center while controls were recruited from the Lifetime Cancer Screening Center, a H. Lee Moffitt Cancer Center affiliate. None of the controls were diagnosed with any form of cancer at the time of screening. Detailed questionnaire data and oral buccal cells were collected for all subjects.

The Vanderbilt Lung Cancer Study (BioVU). The Vanderbilt Lung Cancer Study is a case-control study nested within the Vanderbilt University Medical Center biobank, BioVU. BioVU is a biorepository of DNA extracted from blood drawn from patients seeking routine clinical care at Vanderbilt University Medical Center and linked

to de-identified electronic health records for research purposes. Lung cancer cases and controls were identified from BioVU participants in February 2014. Lung cancer cases were identified from the Vanderbilt tumor registry. All specimens undergo pathologic review for determination of morphology. Coding of histology was based on SEER Program Coding Guidelines. Controls were randomly selected from BioVU participants, excluding cancer patients, and were matched to cases on age (+/- 5 years), sex, and race. Relevant covariates were identified from electronic health records using natural language processing. Genomic DNA was extracted based on a standard protocol.

Genotype Quality Control

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[6]. Briefly, genotype definition was undertaken using Genome Studio and jointly clustered data from 57,775 individuals and 533,631 SNPs. This included 44,591 samples (43,398 individuals and 1,193 QC duplicate samples) associated with this study of lung cancer, 12,901 individuals from other unrelated OncoArray studies and 283 HapMap control individuals of European, African, Chinese and Japanese origin. The v2c cluster file, available from the OncoArray wiki, was used for clustering (http://epi.grants.cancer.gov/oncoarray/). Genotype clustering was carried out at Dartmouth, with the exception of Copenhagen and a small portion of the Harvard samples for which clustering was conducted at the University of Cambridge and the Center for Inherited Disease Research, respectively.

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 were used to exclude underperforming individuals (DNAs) and genotyping assays (judged by success rate, genotype distributions deviated from that expected by Hardy Weinberg equilibrium). 7,633 individuals were excluded as they were already included in the previous GWAS studies as part of the final meta-analysis (re-genotyped for fine- mapping projects unrelated to this manuscript). Samples were subjected to genotype calling rate and individual calling rate check. 1,708 individuals were removed for call rate less than 95%, and 16,149 SNPs with call rates of less than 95% were removed. After filtering, there were 517,482 SNPs available for analysis. We applied the standard OncoArray consortium filter[6] for removing SNPs if they showed departure from Hardy-Weinberg equilibrium in the controls (P-value<1 x 10^{-7}) or cases (P-value <1 x 10^{-12}) but no SNPs failed these stringent criteria after those removed for lower call rate were excluded.

IBD was calculated between each pair of samples in the data using PLINK to detect unexpected duplicates and relatedness. 340 unexpected duplicated samples were removed with proportion IBD greater than 0.95 (one triplet was detected) and of these 29 were individuals who had been seen at the same center but at different times. For related individuals, we formed clusters of related individuals using the pair-wise PLINK IBD statistics and selected the individual with the highest genotyping success rate within each cluster (irrespective of case/control status for the relative pairs). An additional 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. In PLINK inference, F(homozygosity rate)>= 0.8 is inferred as male; F <= 0.3 is inferred as female; F value between 0.3 and 0.8 were recorded as 0 (undefined). By using these strategies, 306 samples were removed as not being concordant.

We used the program FastPop (http://sourceforge.net/projects/fastpop/)[7] to identify continental ancestry and 5,406 samples were removed as non-Caucasian ancestry (estimated European ancestry probability < 0.8). 27,065 samples were included in final association analysis including 14,803 lung cancer cases and 12,262 controls.

Supplementary Note Table 1. Comparison of results for samples that exclude second degree relatives (0.2< Proportion IBD<0.45) to primary results.

						Excluding Second Degree Relatives			Including Second Degree Relatives		
	rs_number	chr	position	ref allele	effect	EAF Weighted	Odds Ratio	P-value	EAF	Odds Ratio	P-value
Overall	rs71658797	1	77967507	T	A	0.10	1.14	3.25E-11	0.10	2.72E-11	1.14
Overall	rs6920364	6	167376466	G	C	0.46	1.07	1.29E-08	0.46	1.56E-08	1.07
Overall	rs11780471	8	27344719	G	A	0.06	0.87	1.69E-08	0.06	1.83E-08	0.87
Overall	rs11571833	13	32972626	A	T	0.01	1.6	6.12E-16	0.01	1.10E-15	1.6
Overall	rs66759488	15	47577451	G	A	0.36	1.07	2.83E-08	0.36	2.17E-08	1.07
Overall	rs55781567	15	78857986	C	G	0.37	1.3	3.08E-103	0.37	8.51E-104	1.3
Overall	rs56113850	19	41353107	C	T	0.44	0.88	5.02E-19	0.44	1.75E-19	0.88
Adeno	rs13080835	3	189357199	G	T	0.49	0.9	7.45E-12	0.49	2.95E-11	0.9
Adeno	rs7705526	5	1285974	C	A	0.34	1.25	3.80E-35	0.34	2.35E-35	1.25
Adeno	rs4236709	8	32410110	A	G	0.22	1.13	1.28E-10	0.22	1.80E-10	1.13
Adeno	rs885518	9	21830157	A	G	0.1	1.17	9.96E-10	0.1	7.97E-10	1.17
Adeno	rs11591710	10	105687632	A	C	0.14	1.16	6.30E-11	0.14	5.80E-11	1.16
Adeno	rs1056562	11	118125625	C	T	0.47	1.11	2.76E-10	0.47	3.68E-10	1.11
Adeno	rs77468143	15	49376624	T	G	0.25	0.86	1.69E-16	0.25	1.02E-16	0.86
Adeno	rs41309931	20	62326579	G	T	0.12	1.17	1.31E-09	0.12	1.88E-09	1.17
Squam	rs116822326	6	31434111	A	G	0.16	1.25	3.83E-19	0.15	2.73E-19	1.25
Squam	rs7953330	12	998819	G	C	0.32	0.86	7.26E-13	0.32	7.36E-13	0.86
Squam	rs17879961	22	29121087	A	G	0.005	0.41	5.70E-13	0.005	5.49E-13	0.41
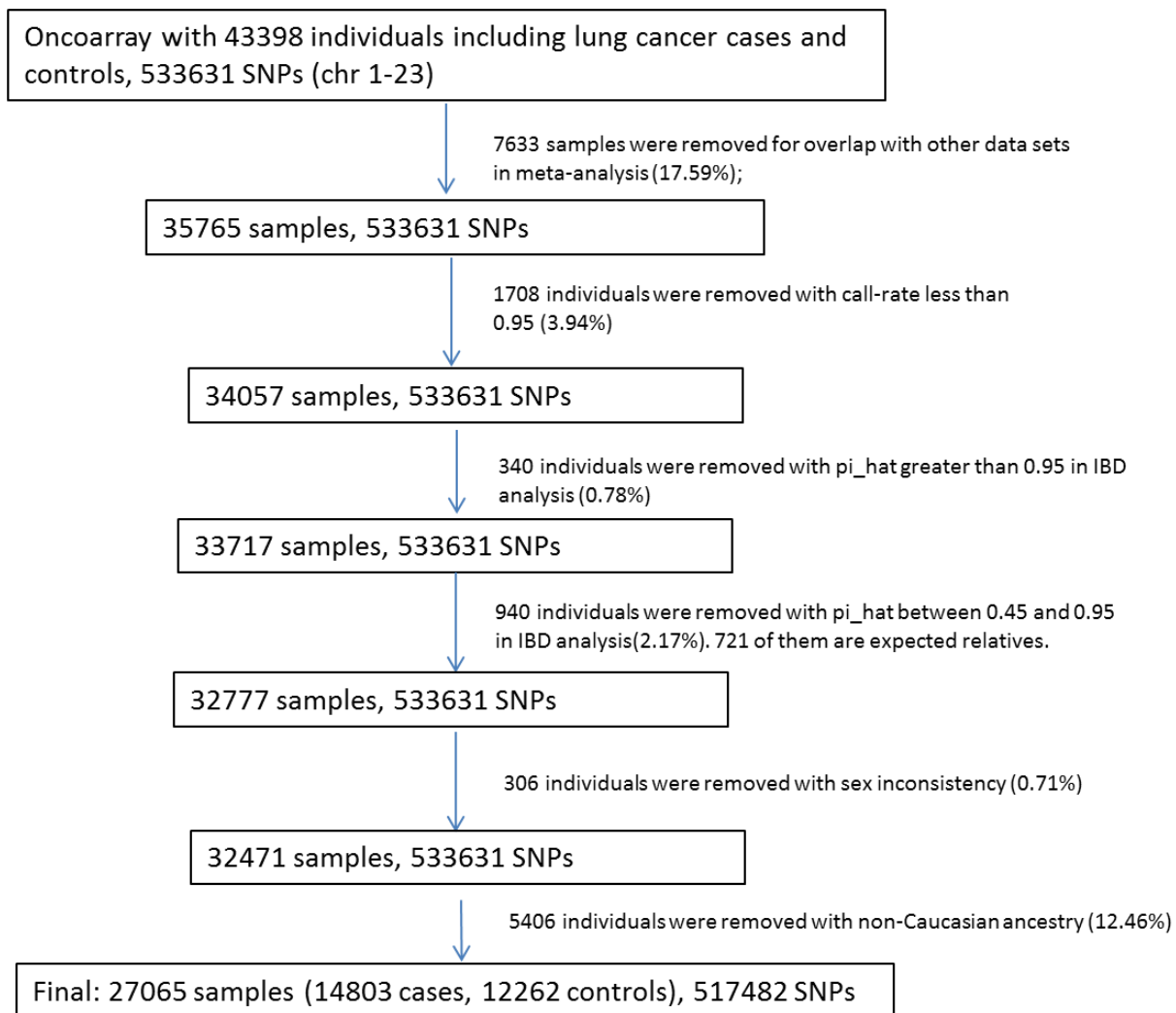
Supplementary Note Figure 1. Summary of the quality control procedures and the numbers of individual and variants excluded across the different quality control steps.

Imputation and Statistical Analysis

Imputation analysis: The OncoArray whole genome data were imputed in a two-stage procedure using SHAPEIT (shapeit.v2.r790.Ubuntu_12.04.4.static) to derive phased genotypes, and IMPUTEv231 (impute_v2.3.2_x86_64_static) to perform imputation of the phased data. We used the default parameters used to derive phased genotypes with SHAPEIT, increasing: - the number of burn-in iterations used by the algorithm to reach a good starting point to 10 ("--burn 10"), - the number of pruning iterations used by the algorithm to find a parsimonious graph for each individual to 10 ("--prune 10"), - and the number of iterations used by the algorithm to compute transition probabilities in the haplotype graphs to 50 ("--main 50"). We performed imputation with IMPUTEv2 using ~5Mb non- overlapping intervals for the whole genome. The flag "-use_prephased_g" was provided to indicate that pre-phased haplotypes were being used. In addition we excluded from imputation the 1000 Genomes variants whose minor allele frequency in Europeans and East Asians was lower than 0.001. The missing genotypes at typed SNPs were replaced with imputed genotypes using the option "-pgs_miss". The number of reference haplotypes to use as templates when imputing missing genotypes was increased to 800 ("-k_hap 800"), and the buffer region was increased to 500kb ("-buffer 500"). For the fine mapping regions we also imputed the non-genotyped data with IMPUTEv2 but without prephasing in SHAPEIT in order to improve imputation accuracy. For this we also increased: - the default number of Markov chain Monte Carlo (MCMC) iterations (including burn-in) to 50 ("-iter 50"), - the number of MCMC iterations to

discard as burn-in to 15 ("-burnin 15"), - and the number of haplotypes to use as templates when phasing observed genotypes to 100 ("-k 100"). We included for imputation only the more common variant if more than one variant yielded a match at the same position.

Statistical analysis of imputed data: Association analysis was performed using SNPTEST (v2.5.2), genotype probabilities generated from IMPUTEv2 (impute_v2.3.2_x86_64_static) was used as input data. We used two main parameter options: -frequentist 1 and -method score.

Genotyping concordance between OncoArray genotyping and alternate genetic platforms.
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.

First, we considered general performance of the OncoArray genotyping and imputation 5,742 (~ 12%) of the OncoArray genotyped individuals were also genotyped on an Affymetrix array of 414,504 genetic variants [8]. Of these 414,504 genetic markers, 113,363 variants overlapped with markers considered or imputed from the Oncoarray and passed quality control criteria in both experiments. Concordance rate was based on agreement between Oncoarray genotyping and imputation genotypic probabilities, and we considered the general concordance and also concordance between the rare alleles only. Results were further stratified by minor allele frequency. Supplementary Note Table 2 below describes the   concordance between genotypes indicating acceptable concordance rates for common alleles, but concordance deceases   with lowering minor allele frequency.

Supplementary Note Table 2. Genotype concordance between imputed variants and variants genotyped on an Affymetrix genotyping array.

	AA	AB	BB
AA	a	b	c
AB	d	e	f
BB	g	h	i

A =rare allele

Variant concordance table for minor alleles (top) and both alleles (bottom)							
	MAF	MIN	MAX	MEAN	MEDIAN	STD	N
Rare alleles only	0-0.005	0.00	1.00	0.27	0.20	0.27	24,347
Comparing cells (left):	0.005-0.02	0.00	1.00	0.53	0.54	0.30	22,613
(a+e)/	0.02-0.05	0.00	1.00	0.72	0.79	0.25	16,885
(a+e+d+b+h+f+c+g)	0.05-0.5	0.00	1.00	0.88	0.96	0.18	49,518
All alleles	0.0-005	0.00	1.00	0.94	1.00	0.23	24,347
Comparing cells (left):	0.005-0.02	0.00	1.00	0.97	0.99	0.11	22,613
(a+e+i)	0.02-0.05	0.00	1.00	0.97	0.99	0.08	16,885
/all_counts	0.05-0.5	0.00	1.00	0.95	0.99	0.11	49,518

Second, we considered the 18 loci that achieved genome wide significance. For these loci, the sentinel variant (or correlated proxy variant) was also genotyped on the custom Affymetrix Axiom exome array described above. We therefore considered the concordance between the OncoArray genotypes and the Affymetric array for these 18 variants in the 5,742 individuals where genotyping was available for both platforms as described above.
Supplementary note Table 3 below describes the concordance between OncoArray (imputation) genotypes and the validation genotypes. rs56113850 (19q13.2) located near *CYP2A6* was found to have acceptable, but not ideal concordance rates, but this variant lies within a region of chromosome 19 that includes many CNV variations caused by nearby and   highly homologous pseudogene *CYP2A7*.

Supplementary Note Table 3. Concordance table between 18 loci associated with lung cancer and validation genotypes.

Locus	Marker	MAF	Subset	OR	Pvalue	Axiom Assay ID	OncoArray or imputed	Rare-Allele	Common-Allele
1p31.1	rs71658797	0.10	Overall	1.14	3.25E-11	AX-11322427	OncoArray	0.96	0.98
3q28	rs13314271	0.49	Adeno	0.90	7.49E-12	AX-14237578	Imputed	1.00	1.00
5p15.33	rs7705526	0.34	Adeno	1.25	3.80E-35	AX-35104637	OncoArray	0.99	1.00
6p21	rs116822326	0.16	Squam	1.25	3.83E-19	AX-11433140	Imputed	1.00	1.00
6q27	rs429083	0.47	Overall	1.07	3.29E-08	AX-41911885	Imputed	1.00	1.00
8p12	rs2439312	0.22	Adeno	1.13	3.29E-10	AX-15945005	OncoArray	1.00	1.00
8p21.1	rs11780471	0.06	Overall	0.87	1.69E-08	AX-11175013	OncoArray	0.99	1.00
9p21.3	rs885518	0.10	Adeno	1.17	9.96E-10	AX-16192161	Imputed	0.98	1.00
10q24.3	rs10786775	0.10	Adeno	1.18	2.07E-10	AX-14628844	Imputed	1.00	1.00
11q23.3	rs1056562	0.47	Adeno	1.11	2.76E-10	AX-12393712	OncoArray	1.00	1.00
12p13.33	rs10849605	0.49	Squam	0.88	2.10E-11	AX-39529951	OncoArray	1.00	1.00
13q13.1	rs11571833	0.01	Overall	1.60	6.12E-16	AX-83400600	Imputed	0.96	1.00
15q21.1	rs7173058	0.37	Overall	1.07	6.88E-08	AX-12916278	Imputed	0.98	0.99
15q21.1	rs2413932	0.26	Adeno	0.87	8.22E-14	AX-88897348	OncoArray	1.00	1.00
15q25.1	rs55853698	0.37	Overall	1.30	3.15E-102	AX-12969704	Imputed	1.00	1.00
19q13.2	rs56113850	0.44	Overall	0.88	5.02E-19	AX-88899684	OncoArray	0.80	0.86
20q13.33	rs4809324	0.11	Adeno	1.13	1.01E-06	AX-40592777	OncoArray	0.99	1.00
22q12.1	rs17879961	0.005	Squam	0.41	5.70E-13	AX-83154278	OncoArray	1.00	1.00

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 [9]. SNPs in linkage disequilibrium with lead SNPs were identified using data from the 1000 Genomes Project. The influence of candidate causal variants on transcription factor binding sites was determined using the ENCODE-Motifs resource [10] obtained from the HaploReg v4.1 database [11]. To investigate functional elements enriched across the region encompassing the strongest candidate causal SNPs, we analysed chromatin biofeatures from the Encyclopedia of DNA Elements Project [12] and Roadmap Epigenomics Projects [13] namely: Chromatin State Segmentation by Hidden Markov Models (chromHMM), DNase I hypersensitive, transcription factor and histone modification (H3K4me, H3K9ac, and H3K27ac) ChIP-seq in various lung cell types (A549 lung carcinoma and IMR90 fetal lung fibroblasts cell lines, fetal lung and adult lung primary tissue, and normal human lung fibroblast (NHLF) primary cultured cells). To identify the SNPs most likely to be functional we used RegulomeDB [14]. To identify putative target genes, we examined potential functional chromatin interactions between distal regulatory elements and the promoters at the risk regions using Hi-C data generated in IMR90 lung fibroblasts [15]. Annotation of putative *cis*-regulatory regions and predicted target genes used the Integrated Method for Predicting Enhancer Targets (IM-PET) [16], the Predicting Specific Tissue Interactions of Genes and Enhancers (PreSTIGE) algorithms [17], and

data from Hnisz [18]. Intersections between candidate causal variants and regulatory elements were identified using Galaxy [19], BedTools v2.24 [20] and HaploReg v4.1, and visualised in the UCSC Genome Browser [21]. Publicly available eQTL databases including Gene-Tissue Expression (GTEx Consortium[22]) version 6, multiple tissues including lung) and Westra [23] (blood), were queried for all candidate causal variants. All results are presented in Supplementary Table 5 and are hyperlinked to a custom session within the UCSC Genome Browser to allow the reader to explore the annotation for each loci.

eQTL Analysis.

Microarray eQTL analysis. 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. Imputation was performed with IMPUTE2 (impute_v2.2.2_x86_64_dynamic), and the reference set from the 1000 genomes project (ALL_1000G_phaseIntegrated_v3). Association tests were carried with PLINK version 1.9. eQTL analyses were performed in the three cohorts separately adjusting for age, gender and smoking status and results combined into a meta-analysis.

NCI RNAseq eQTL analysis.

RNA was extracted from lung tissue samples within the Environment and Genetics in Lung cancer Etiology (EAGLE) study, a large population-based case-control study conducted in the Lombardy region of Italy between April 2002 and February 2005. The study protocol was approved by the Institutional Review Board of the US National Cancer Institute and the involved institutions in Italy. Informed consent was obtained for all subjects prior to study participation.

Lung tissue samples were snap-frozen in liquid nitrogen within 20 minutes of surgical resection of Stage I to IIIA lung adenocarcinoma tumors. Cases were newly diagnosed primary cancers, verified by tissue pathology. Surgeons and pathologists were together in the surgery room at the time of resection and sample collection to ensure correct sampling of tissue from the tumor, the area adjacent to the tumor and an additional area distant from the tumor (~1-5 cm), without adversely affecting the participant. The precise site of tissue sampling was indicated on a lung drawing and the pathologists classified the samples as tumor, adjacent lung tissue and distant non-involved lung tissue.

For the current 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. For each subject, usually more than one non-malignant lung tissue sample was collected and at least one sample was examined by a pathologist to confirm the absence of tumor nuclei.

Transcriptome sequencing of 90 non-tumor samples was performed on the Illumina HiSeq2000/2500 platform with 100-bp paired-end reads. Sequence reads were mapped by the Mapsplice algorithm with default parameters based on the hg19 reference genome. We used the generic annotation file (version TCGA.hg19.June2011.gaf) to annotate genes and exons. The read counts mapped to the targeted genes were estimated by the RSEM algorithm and adjusted by the gene size and the total number of reads per sample.

Genotyping was undertaken using Illumina bead arrays as described previously. Imputation was conducted by using IMPUTE2 software version 2.2.2 and version 3 of the 1000 Genomes Project Phase 1 data as the reference set. First, the genomic coordinates were lifted over from NCBI human genome build 36 to build 37 using the UCSC lift over tool. Second, the strand of the inference data was aligned with the 1000 Genomes data by simple allele state comparison or allele frequency matching for A/T and G/C SNPs. A pre-phasing strategy

with SHAPEIT software version 1 was adopted to improve the imputation performance. The phased haplotypes from SHAPEIT were input directly into the IMPUTE2 program.

Normalized read counts of the gene were quantile transformed and were regarded as quantitative traits, which were regressed against the genotype or the imputed dosage of the effect allele for the targeted locus for the eQTL analysis. The regression models were adjusted by gender, age, pack-years of smoking, and the top five principal components based on ancestry-informative SNPs.

GTEx eQTL analysis.

eQTL summary statistics based on RNAseq analysis were obtained for eQTL summary statistics from the GTEx data portal http://www.gtexportal.org/home/ [22]. This data included 278 individuals with data from lung tissue. For the 8p21 locus, we additionally assessed the rs11780471 variant in Brain tissues in the GTEx and the BRAINEAC consortium (http://braineac.org) for analysis of the *CHRNA2* gene.

Smoking propensity.
Oncoarray: Cigarettes Per Day (CPD) analysis.

We performed an association study of CPD (as the dependent variable) using a linear regression model with genotype probabilities. FlashPCA was run for principal component analysis (PCA) and the first three eigenvalues were extracted for association analysis as described above. Analysis was performed using SNPTEST (v2.5.2) and 8,120 ever-smoking controls from Oncoarray data were used on CPD analysis.

DeCODE: CPD.

Ever smokers were recruited in the years 1997–2014 as part of various Icelandic studies. The information on smoking quantity (SQ) was available from a standardized smoking questionnaire that asks: 'How many cigarettes per day do/did you smoke on average (on most days)?' Current smokers answered regarding their current consumption and former smokers referred to their consumption in the past. In cases where multiple records were available we used the maximum value for analysis. The SQ was categorized into four levels, (1–10, 11–20, 21–30 and 31+ CPD). A total of 40,882 subjects (34 200 chip-typed) were included in the analysis.

Long-range phasing of all chip-genotyped individuals was performed with methods described previously [24] and imputed into chip-typed individuals and their close relatives using methods [24] based on IMPUTE [25]. A generalized form of linear regression was used to test the correlation between the variants tested and quantitative traits (CPD and FNTD see below) in Iceland [24,26,27]. The generalized form assumes that the smoking behavior of related individuals is correlated proportional to the kinship between them rather than assuming that the smoking phenotypes of all individuals are independent. In order to account for the relatedness and stratification within the case and control sample sets we applied the method of genomic control based on genotyped markers using a subset of about 300,000 common variants. Quoted P-values were adjusted accordingly.

UK Biobank: CPD.

The UK Biobank[28] is a prospective cohort that recruited more than 500,000 men and women aged 40-96 years between 2006 and 2010, gathered anthropometric measures, health and life style parameters and collected biological samples. We analyzed 118,730 individuals from UK Biobank dataset with European background who had valid genetic data and smoking behavior parameters from the recruitment questionnaire or interview. A total of 42,044 ever smokers had valid number of smoked cigarettes per day (between 1 and 150 cigarettes per day). Genome-wide association analysis with CPD was performed based on linear regression adjusted for age, sex, genotyping array, and principal components for population stratification.

CPD Meta-analysis.

METAL software[29] was used for the meta-analysis of CPD genetic association parameters from the three different sample sources.

Fagerström Test for Nicotine Dependence (FTND)

We used FTND data collected among current and former smokers in four studies [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 participants were ever smokers defined by the FTND as having mild (FTND score 0–3 or low-level smoking, N=9,137), moderate (score 4–6, N=4,881), or severe (score 7–10, N=3,056) dependence. In the deCODE Genetics study only, 4,313 low-intensity smokers (10 or fewer cigarettes-per-day) with no FTND data available were added into the mild category, as we previously found a high concordance rate between the category of 10 or fewer cigarettes-per-day and FTND scores of 3 or less, and this enabled us to increase sample size with little phenotype misclassification [26]. All participants were aged 18 years old or older. Males comprised 53.9% of the total sample size. In each study, as done before [26], additive SNP genotype dosage was tested for association with categorical nicotine dependence (mild, moderate, and severe) with a linear regression model adjusting for age, sex, and study-specific covariates (DSM-IV-defined alcohol dependence and cocaine dependence for SAGE* and COPD severity for COPDGene). Principal components were also included to minimize bias due to population stratification, selecting the number needed to account for >75% of the variability in nicotine dependence. The study-specific SNP association results were combined using fixed effects, inverse variance-weighted meta-analysis with genomic control applied.

8p21 variant (rs11780471) and smoking behavior.

We further undertook a focused analysis of the 8p21 variant rs11780471 and various smoking behavior under log additive genetic model, based on the individual level data from UKBiobank and deCODE (using covariates as described above) and OncoArray controls for CPD (as described above). We additionally consider summary statistics for the rs11780471 variants from the TAG consortium (Tobacco And Genetics consortium (https://www.med.unc.edu/pgc/downloads)[30]. METAL software[29] was used to combine genetic association parameters across the cohorts statistics into a meta-analysis. This included:

Smoking: Logistic regression was applied to estimate the association between rs11780471 and the likelihood of being ever versus never smokers, based on data from UK Biobank (n=118,730), the TAG Consortium (n=74,035), and deCODE (n=40,882).

Smoking cessation: Logistic regression was applied to estimate the association between rs11780471 and the likelihood of being a former smoker versus current smokers based on data from UK Biobank (n=55,312), deCODE (n=40,882) and the TAG Consortium (n=41,278) samples.

Difficulty not smoking for 1 day: Logistic regression was applied to estimate the association between rs11780471 and being able to refrain from smoking for 1 day. This parameter was initially described in 4 categories: very easy, fairly easy, fairly difficult and very difficult. For our analysis, we used dichotomized response (as difficult versus easy) based on valid UK Biobank answers (N= 11,231)

Time from waking until first cigarette smoked: Logistic regression was applied to estimate the association between rs11780471 and the time from waking until first cigarette smoked for the day. This parameter was initially registered within the following categories: less than 5 min, between 5-15 min, between 30min-1h,

between 1-2h, and longer than 2h. Valid UK Biobank answers (N= 11,211) were classified as before versus after 30 min.

Smoked CPD: Linear regression was applied to evaluate the association between variants and number of cigarettes smoked per day within ever smokers from UK Biobank (n= 42,044), deCODE (n=40,882), and TRICL-ILCCO OncoArray (n=8,120) samples. Not all sentinel variants associated with lung cancer described in Table2 were analysed by TAG, hence TAG estimates are not included in Table 2. However, in the particular case of rs11780471, it was analyzed by TAG consortium data (n=38,181) and therefore TAG was included in the CPD analysis of this particular variant.

Age of smoking initiation: Linear regression was applied to evaluate the association between rs11780471 and Age of smoking initiation (log transformed) within ever smokers from UK Biobank (n=43,699) and TAG Consortium (n=41,278) data.

Nicotine addiction (ADDNico) index; from deCODE (n=40,882) sample.

FTND index. As described above.

Lung Function (FEV1, FVC)
The lung function *in silico* look up was conducted in SpiroMeta consortium, which included 38,199 European ancestry individuals. The genomewide associations between genetic variants and forced expiratory volume in 1 s (FEV1), forced vital capacity (FVC) and FEV1/FVC with 1000 Genomes Project (phase 1)-imputed genotypes and followed up top associations in 54,550 Europeans. The results included in this manuscript are based on the Stage 1 GWAS with 38,199 individuals. Covariates including age, sex, height and principal components for population structure were adjusted in the analysis and the analysis was undertaken separately for ever smokers and never smokers, and then the results were meta-analysed [31].

Telomere Length (TL).
Sentinel genetic variants associated with telomere length were those described by Codd et al[32], (rs10936599-*TERC,* rs2736100-*TERT*, rs7675998-*NAF1*, rs9420907-*OBFC1*, rs8105767-*ZNF208*, rs755017-*RTEL1*, rs11125529-*ACYP2*, rs2967374-*MPHOSPH6*). Telomere lengths in 6,766 individuals from the UK Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH) study controls were measured from lymphocyte derived DNA using a real-time PCR methodology as described in Pooley et al., 2013[33]. Twelve percent of samples had been run in duplicate and the coefficient of variation (%CV) of the ΔCt variable was 6%. Genotypes were derived from the iCOAGs genotyping beadchip and linear regression was applied to evaluate the association between TL and genetic variants under a log additive genetic model as described previously[34].

Consortium investigators
SpiroMeta: Eva Albrecht, John Beilby, Harry Campbell, Alexessander Couto Alves, Ian J Deary, Stefan Enroth, Christian Gieger, Sven Gläser, Harald Grallert, Ulf Gyllensten, Ian P Hall, Sarah E Harris, Anna-Liisa Hartikainen, Caroline Hayward, Joachim Heinrich, Markku Heliövaara, Lynne Hocking, Momoko Horikoshi, Jing Hua Zhao, Jennifer E Huffman, Jennie Hui, Nina Hutr Kähönen, Medea Imboden, Alan L James, Marjo-Riitta Jarvelin, Åsa Johansson, Peter K Joshi, Mika Kähönen, Stefan Karrasch, Abdul Kader Kheirallah , Beate Koch, Ivana Kolcic, Ashish Kumar, Terho Lehtimäki, Lorna M Lopez, Leo-Pekka Lyytikäinen, Jonathan Marten, Wendy L McArdle, Suzanne Miller, Arthur W Musk, Pau Navarro, Ioanna Ntalla, Sandosh Padmanabhan, Ozren Polasek, Nicole M Probst-Hensch, Olli T Raitakari, Rajesh Rawal, Janina Ried, Samuli Ripatti, Igor Rudan, Ian Sayers, Holger Schulz, Generation Scotland, Robert A Scott, María Soler Artigas, John M Starr, David P Strachan, Ida Surakka, Alexander Teumer, Martin D Tobin, Holly Trochet, Veronique Vitart, Henry Völzke, Louise V Wain, Nicholas J Wareham, Sarah H Wild, James F Wilson, Alan F Wright, Tatijana Zemunik

References

1.   Landi, M.T. *et al.* Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. *BMC Public Health* **8**, 203 (2008).
2.   Riboli, E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann Oncol* **3**, 783-91 (1992).
3.   Lips, E.H. *et al.* Association between a 15q25 gene variant, smoking quantity and tobacco-related cancers among 17 000 individuals. *Int J Epidemiol* **39**, 563-77 (2010).
4.   Spitz, M. R. *et al.* A Risk Model for Prediction of Lung Cancer. *J Natl Cancer Inst* **99 (9)**, 715-726 (2007).
5.   Thurston, Sally W. *et al.* Modeling Lung Cancer Risk in Case-Control Studies Using a New Dose Metric of Smoking. Cancer epidemiology, biomarkers & prevention : *a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **14.10**, 2296–2302 (2005).
6.   Amos, C.I. *et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* **26**, 126-135 (2017).
7.   Li, Y. *et al.* FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. *BMC Bioinformatics* **17**, 122 (2016).
8.   Kachuri, L. *et al.* Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. *Carcinogenesis* **37**, 96-105 (2016).
9.   Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-5 (2012).
10.   Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976-87 (2014).
11.   Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* **40**, D930-4 (2012).
12.   ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, e1001046 (2011).
13.   Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
14.   Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).
15.   Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
16.   He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A* **111**, E2191-9 (2014).
17.   Corradin, O. *et al.* Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* **24**, 1-13 (2014).
18.   Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
19.   Blankenberg, D. *et al.* Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* **Chapter 19**, Unit 19.10.1-21 (2010).
20.   Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11.12.1-34 (2014).
21.   Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996-1006 (2002).
22.   GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
23.   Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-43 (2013).
24.   Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435-44 (2015).
25.   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* **39**, 906-13 (2007).

26. Hancock, D.B. *et al.* Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. *Transl Psychiatry* **5**, e651 (2015).

27. Thorgeirsson, T.E. *et al.* A rare missense mutation in CHRNA4 associates with smoking behavior and its consequences. *Mol Psychiatry* **21**, 594-600 (2016).

28. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173-4 (2012).

29. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

30. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* **42**, 441-7 (2010).

31. Soler Artigas, M. *et al.* Sixteen new lung function signals identified through 1000 Genomes Project reference panel imputation. *Nat Commun* **6**, 8658 (2015).

32. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat Genet* **45**, 422-7, 427e1-2 (2013).

33. Pooley KA, Sandhu MS, Tyrer J, et al. Telomere length in prospective and retrospective cancer case-control studies. Cancer Res 2010;70:3170-6.

34. Pooley KA, Bojesen SE, Weischer M, et al. A genome-wide association scan (GWAS) for mean telomere length within the COGS project: identified loci show little association with hormone-related cancer risk. Human molecular genetics 2013;22:5056-64

Supplementary Figure 1. Quantile-Quantile Plots, Overall and by Histology and Smoking Status.
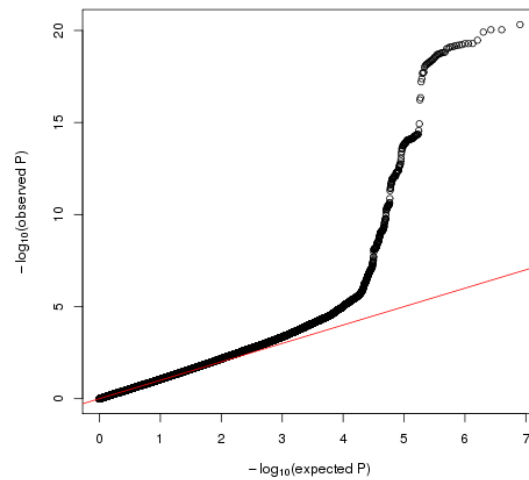
Overall: adjusted lamda=1.0035

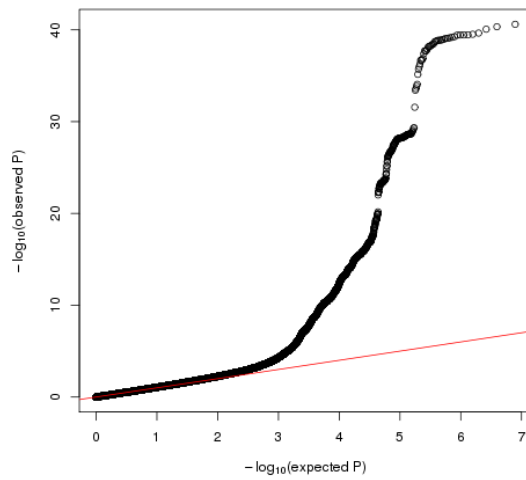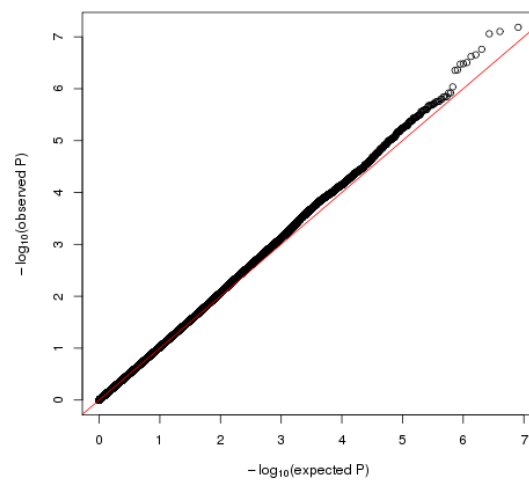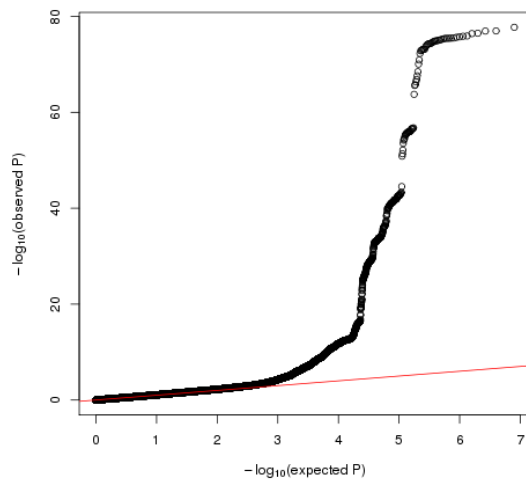Adenocarcinoma: adjusted lamda=1.0050
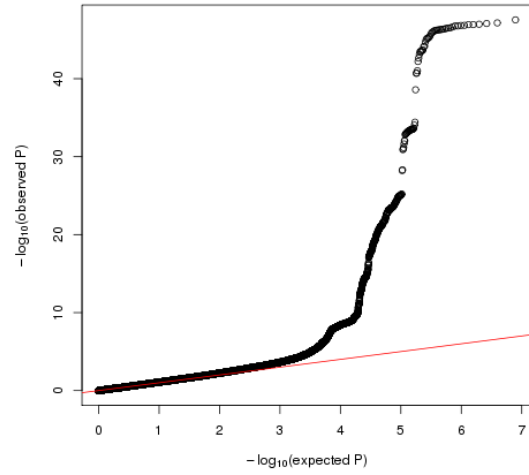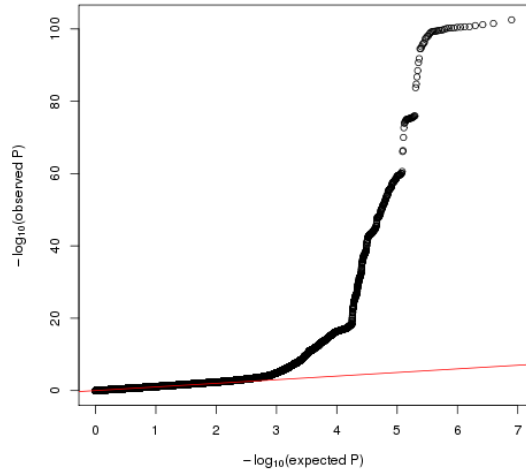
Squamous cell carcinoma: adjusted lamda=
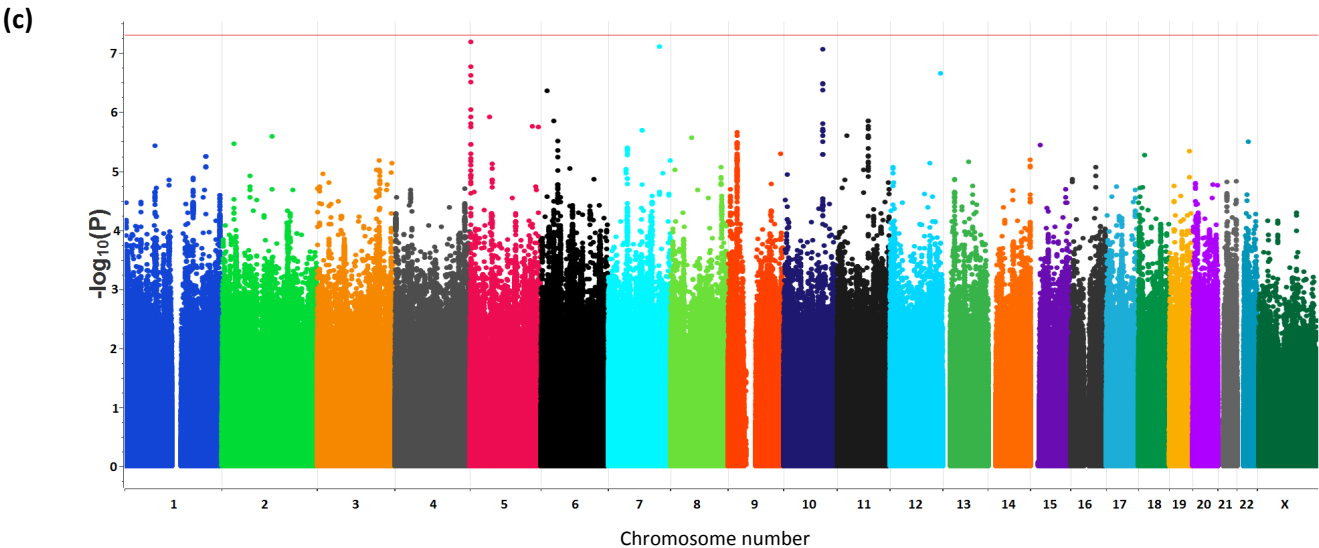1.0051

Small cell carcinoma: adjusted lamda=1.0147

Ever smokers: adjusted lamda = 1.0046

Never Smokers: adjusted lamda = 1.0060

Supplementary Figure 2. Manhattan plots of lung cancer risk overall and by histological subtype and smoking status. (a) small cell lung cancer (b) ever smokers (c) never smokers.
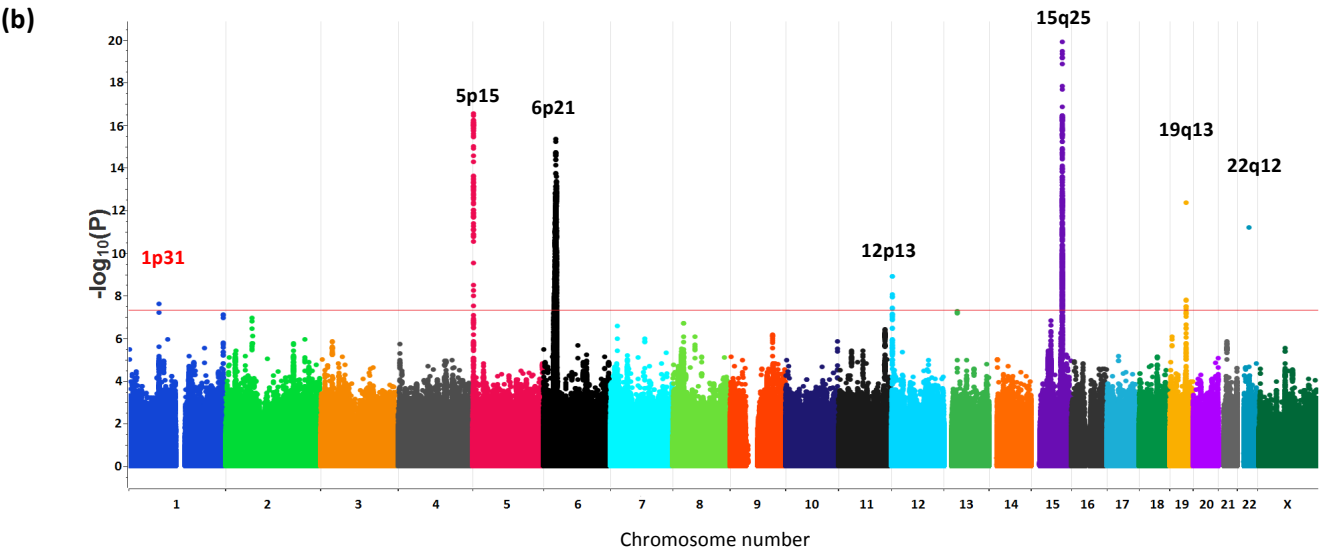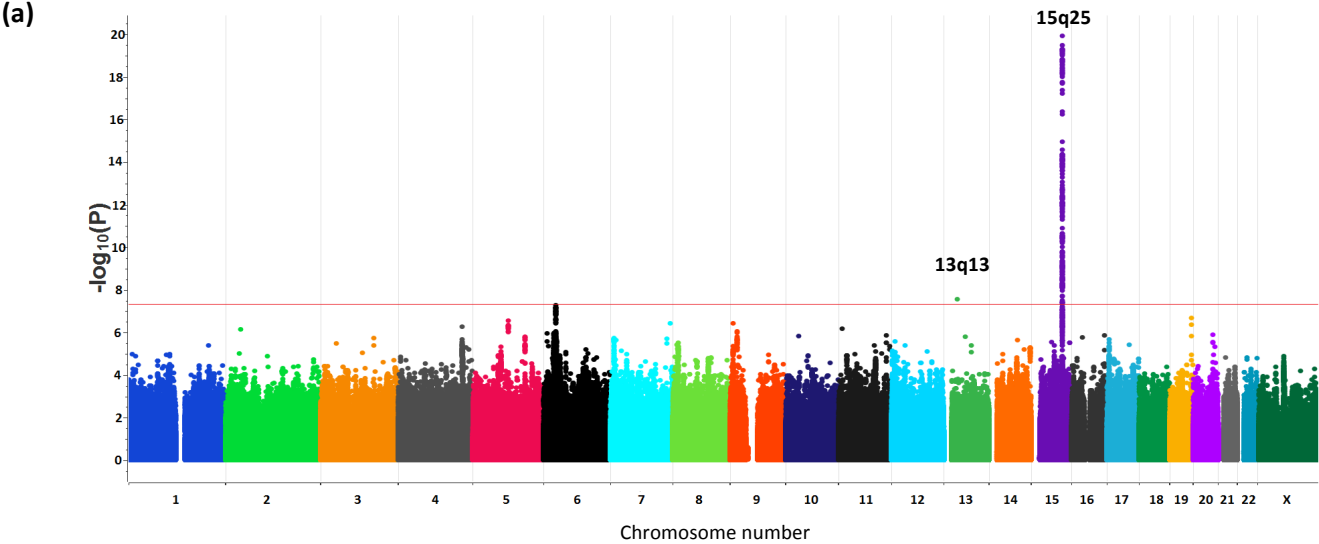
Supplementary Figure 3. Association between sentinel SNPs at the 18 lung cancer loci and lung expression levels for genes located as far as 5 Mb away. The location of genes is illustrated at the bottom of each graph. The Y-axis represented eQTL p value in –log10 scale from the meta-analysis combining Laval, UBC and Groningen. The blue dashed horizontal line represents P-value = 0.05. The red dashed horizontal line represents Bonferroni threshold (P-value=0.05/number of probe sets), if absent indicates that no variant achieved this level. In each instance the top gene associated with the sentinel SNP is illustrated in red. When multiple eQTLs exceeded the Bonferroni threshold, each variant (upper) and gene (lower) are color coded in a corresponding manner.

1p31.1 - rs71658797
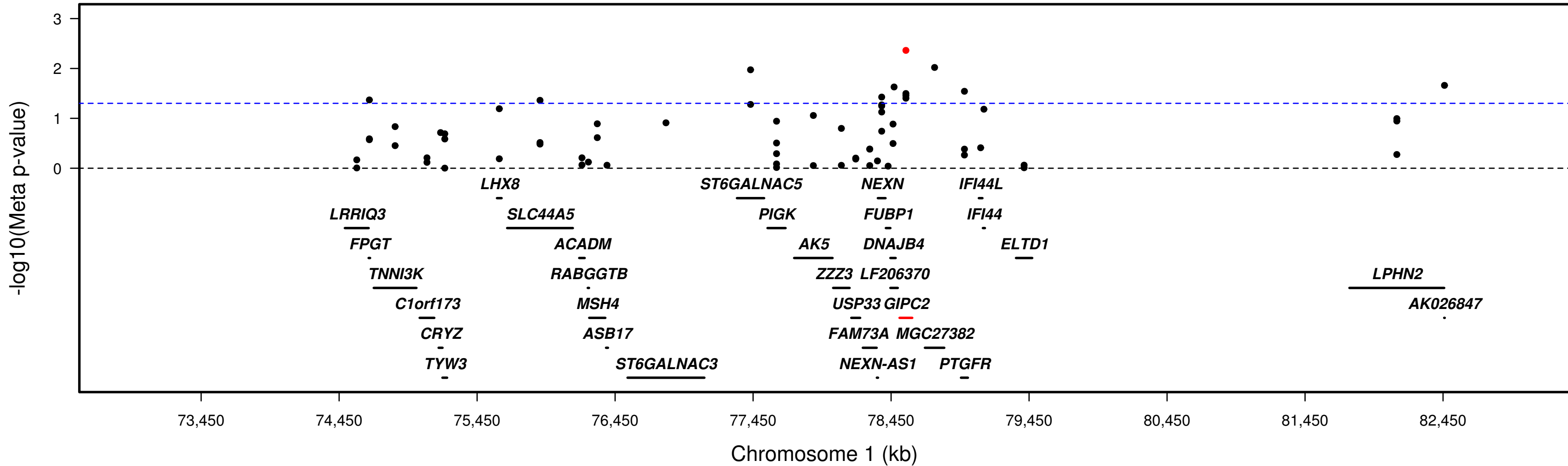
3q28 - rs13080835
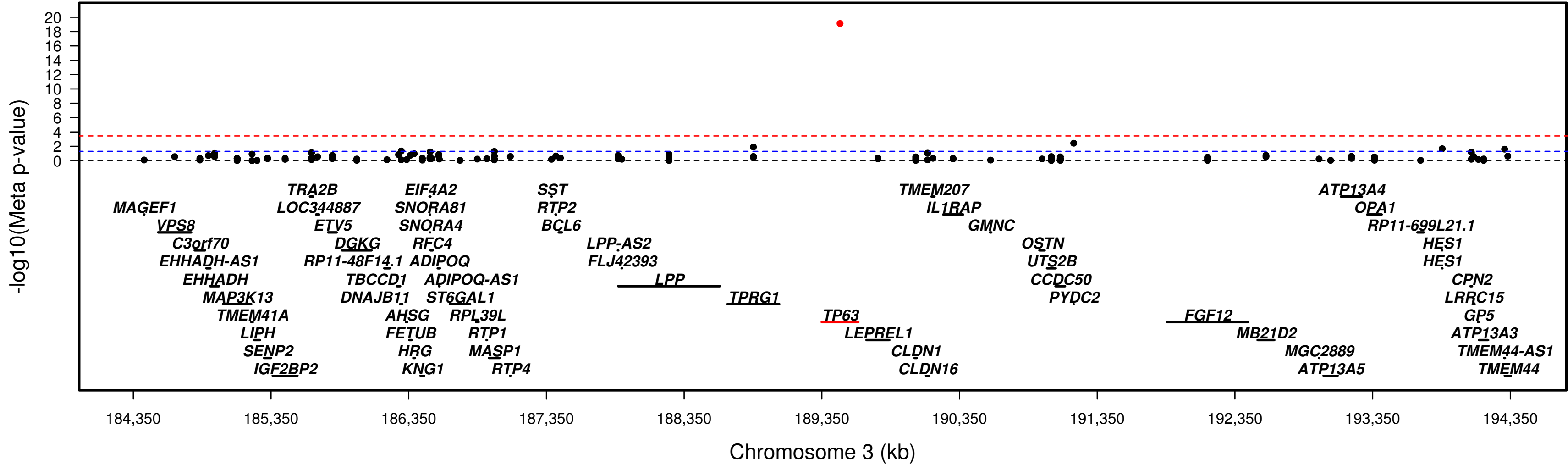
5p15.33 - rs7705526
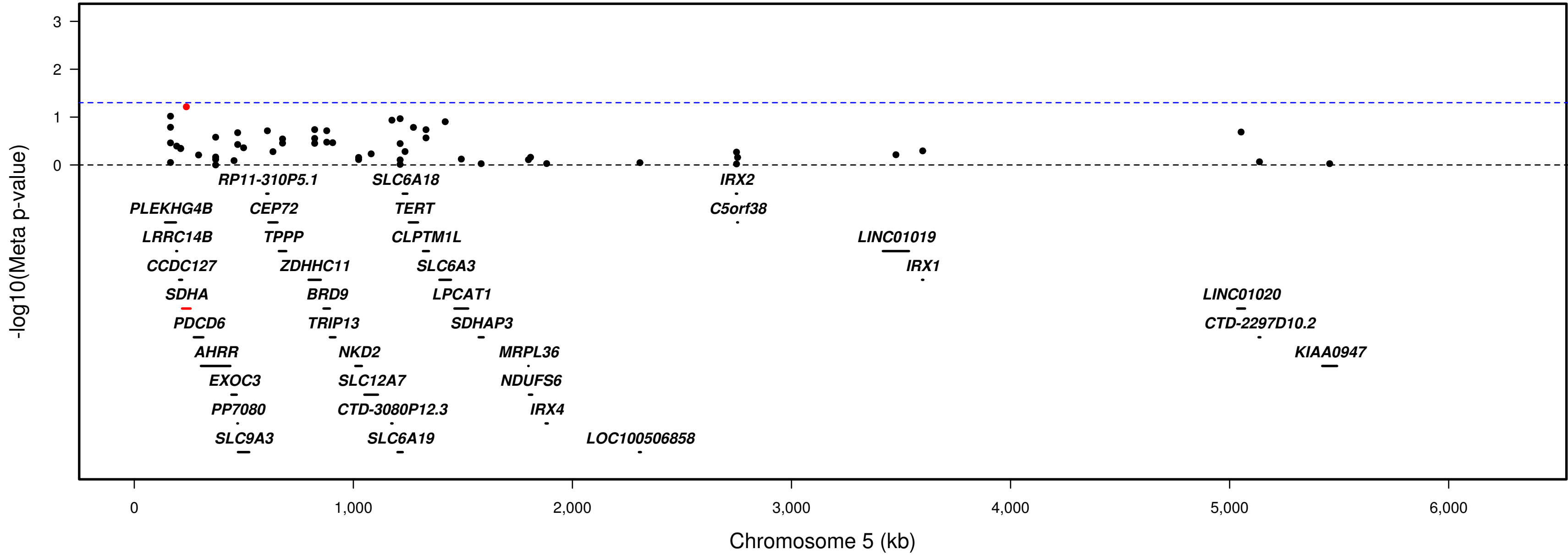
6p21.33 - rs116822326
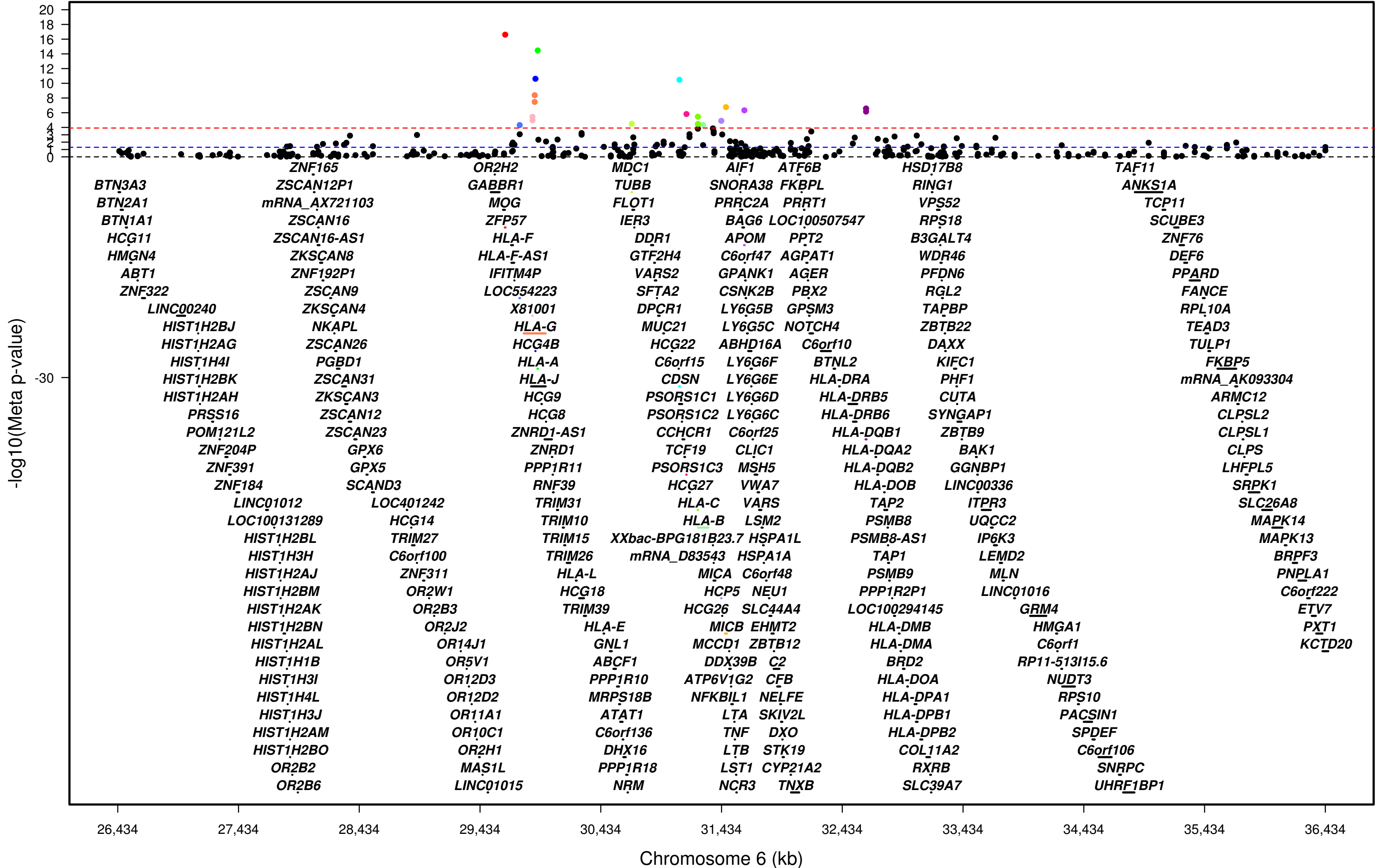
6q27 - rs6920364
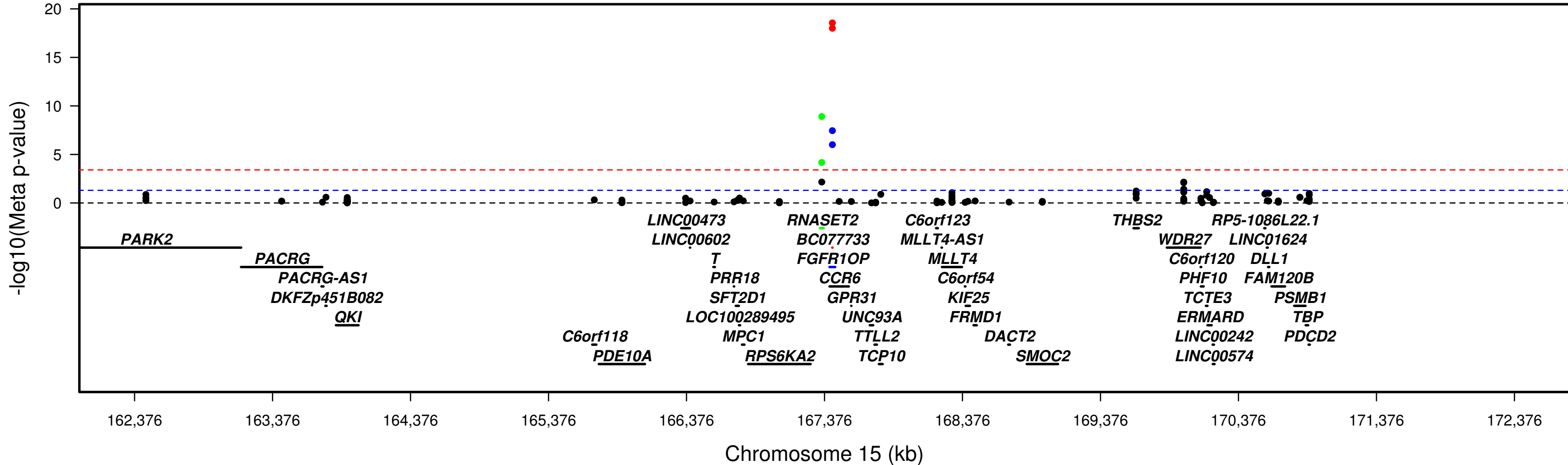
8p21.1 - rs11780471
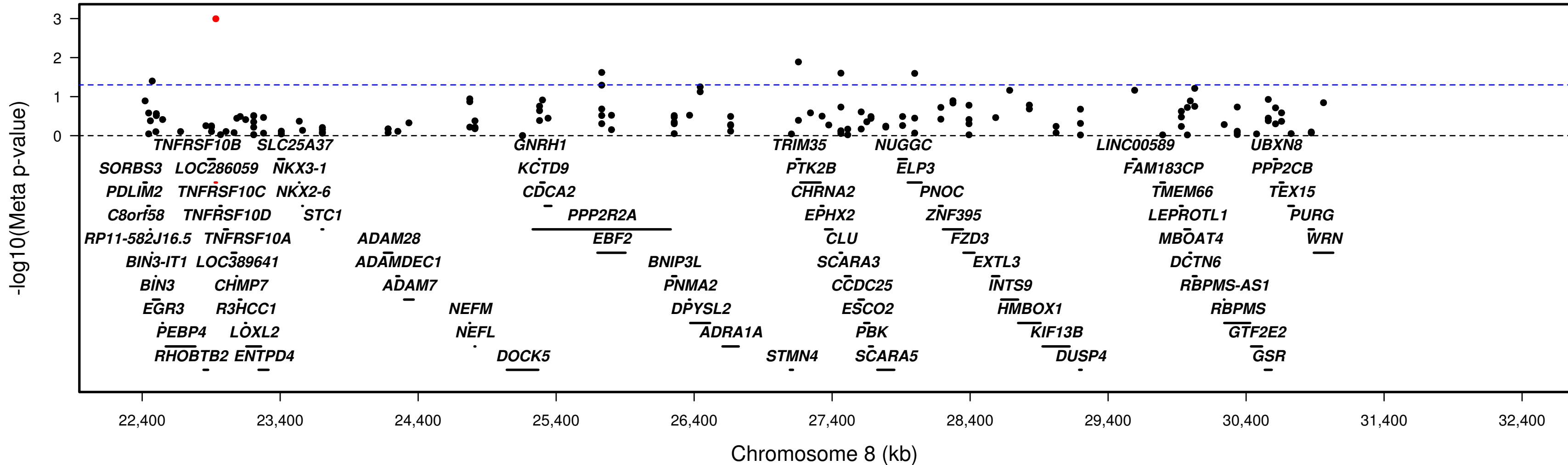
8p12 - rs4236709
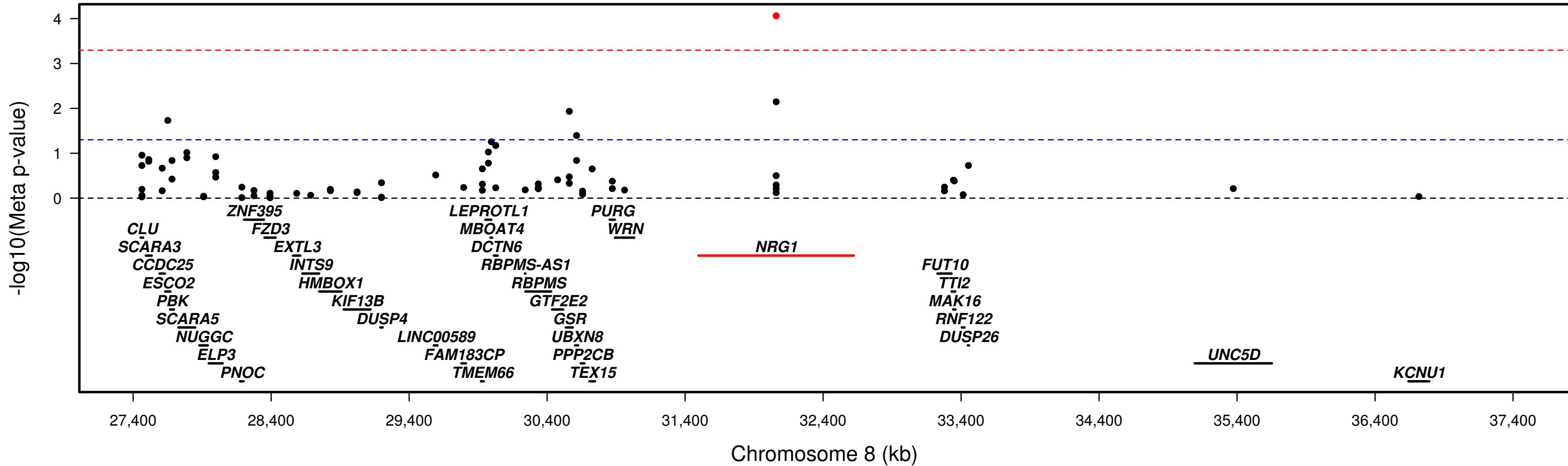
9p21.3 - rs885518
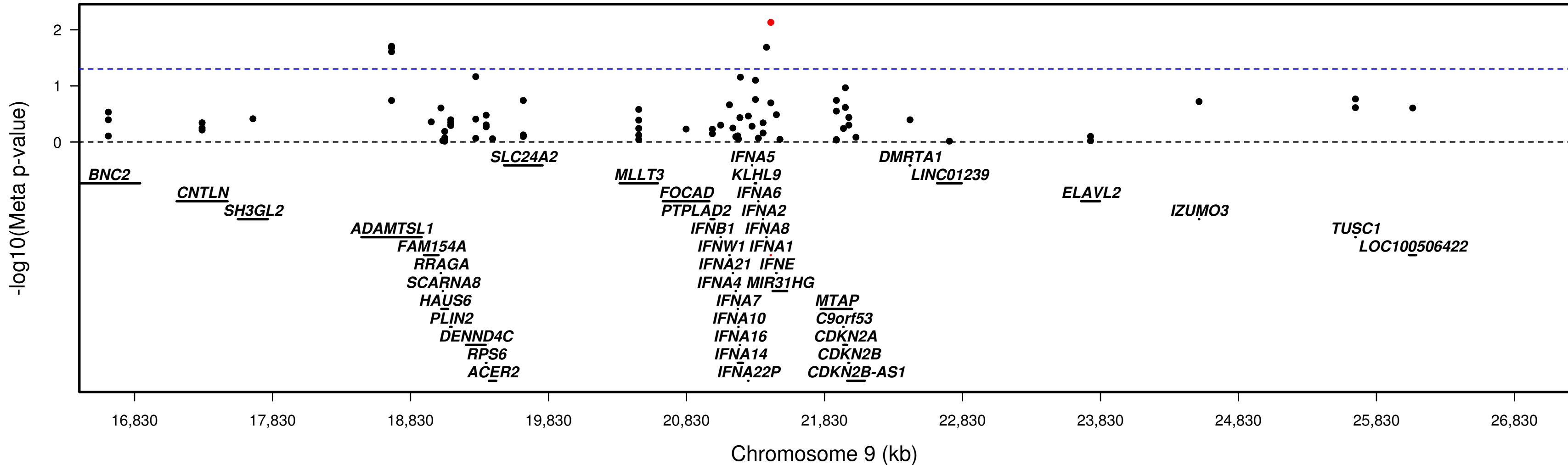
10q24.3 - rs11591710
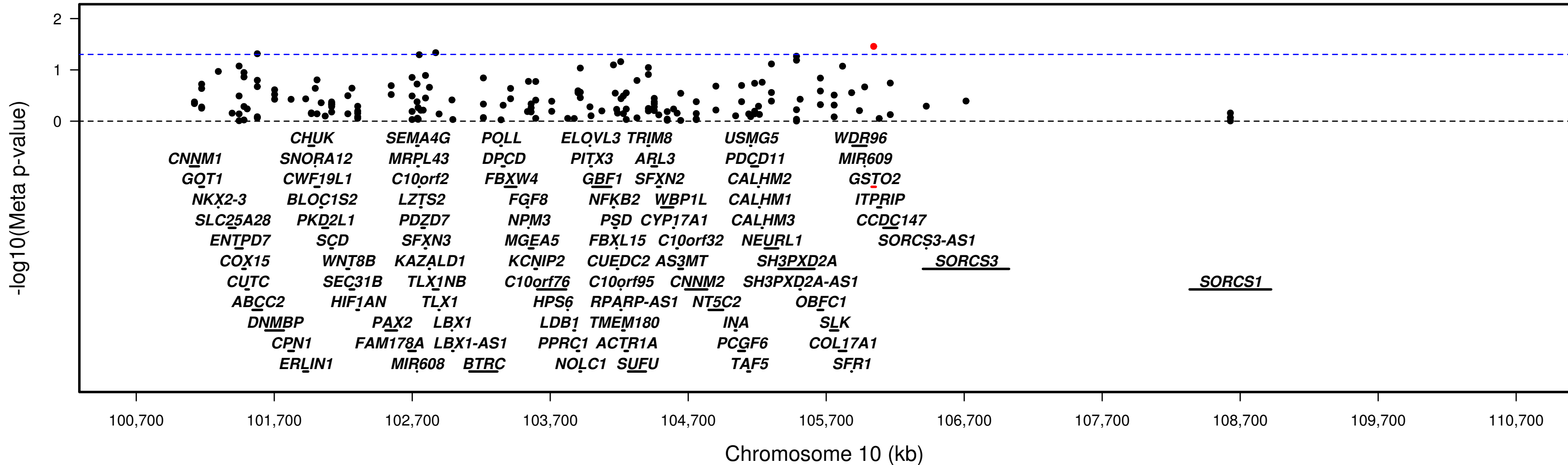
11q23.3 - rs1056562
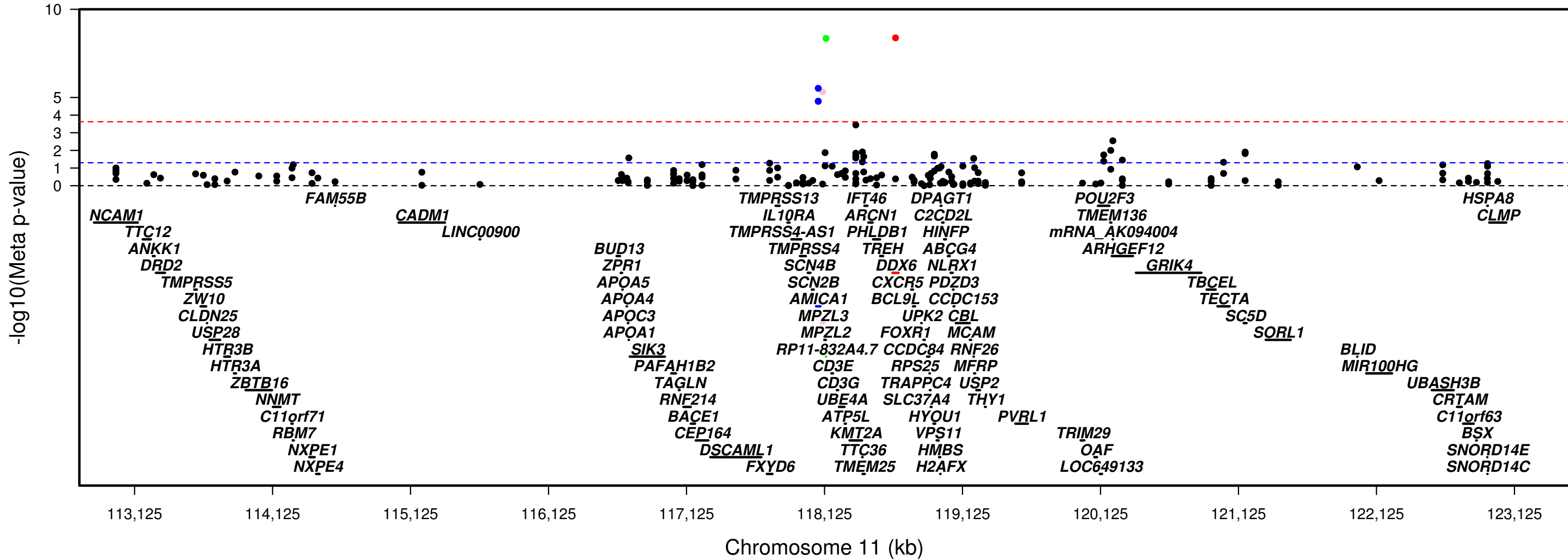
12p13.33 - rs7953330
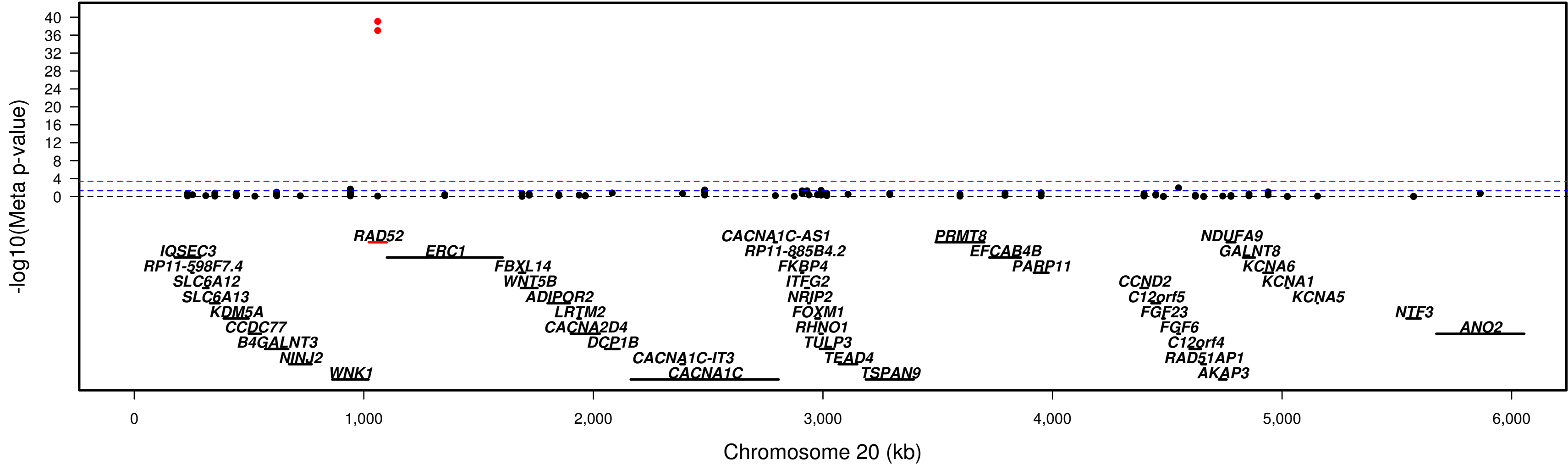
13q13.1 - rs11571833
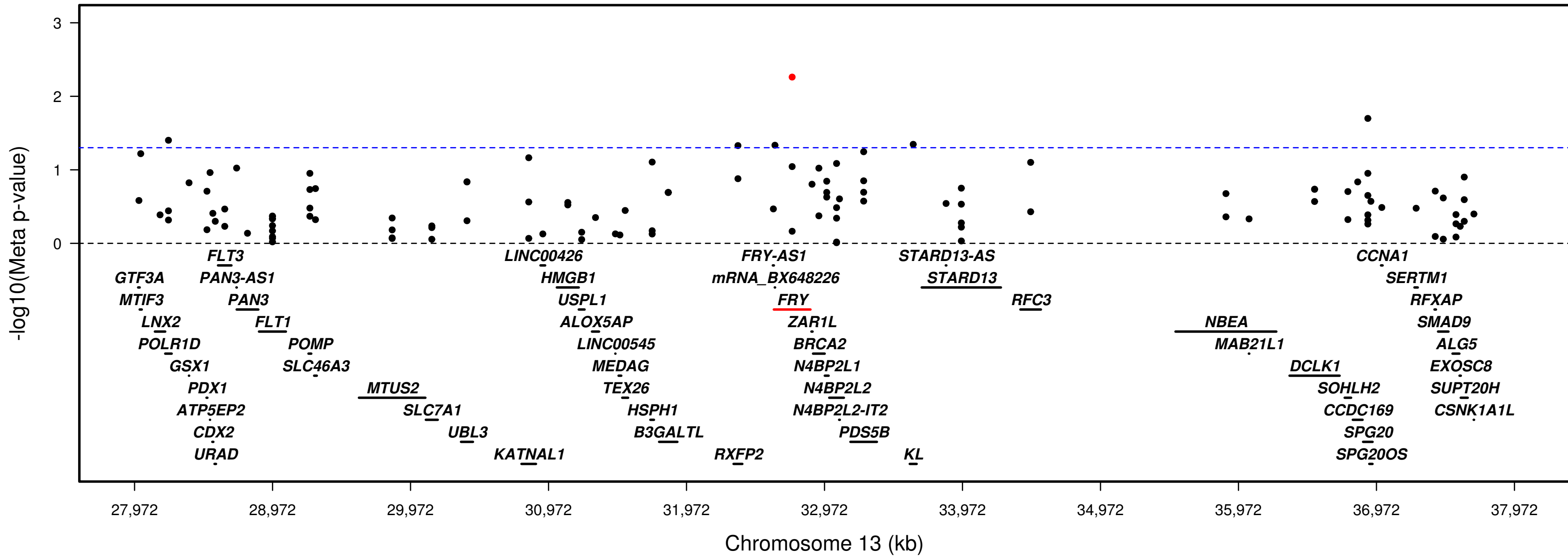
15q21.1 - rs66759488
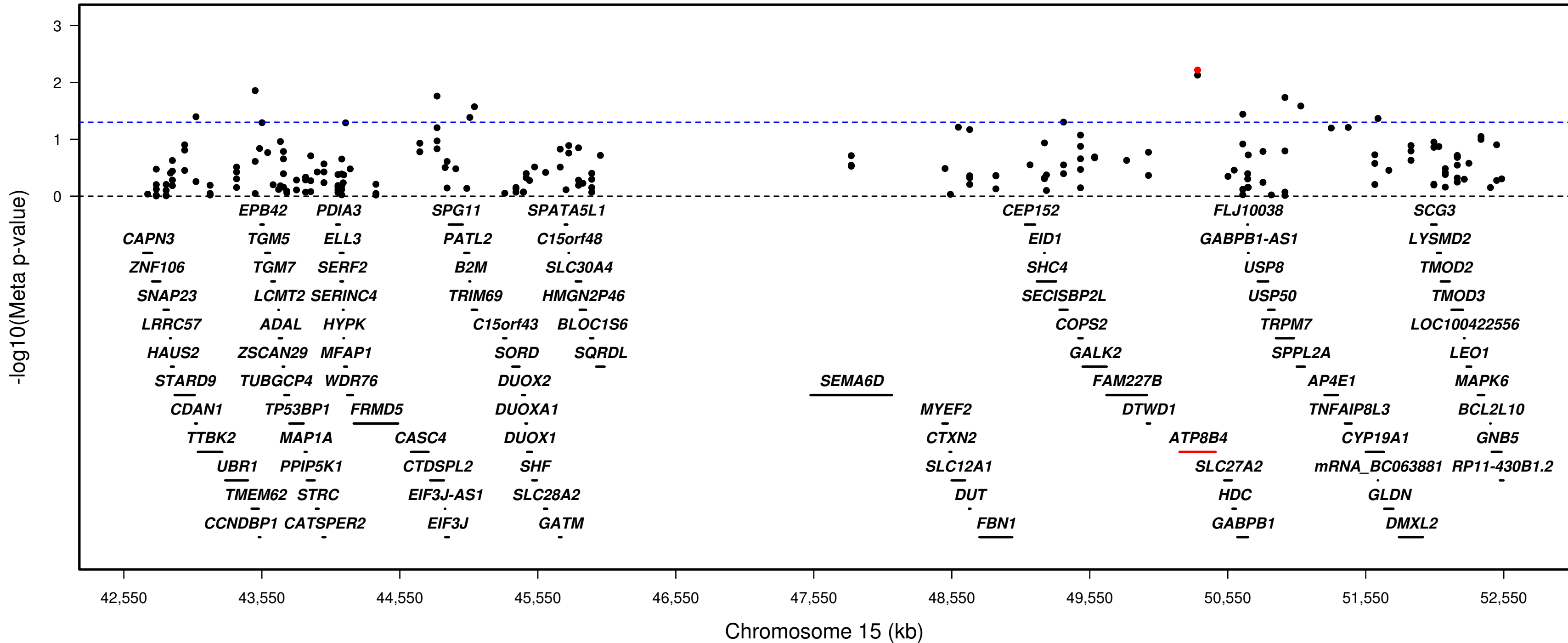
15q21.1 - rs77468143
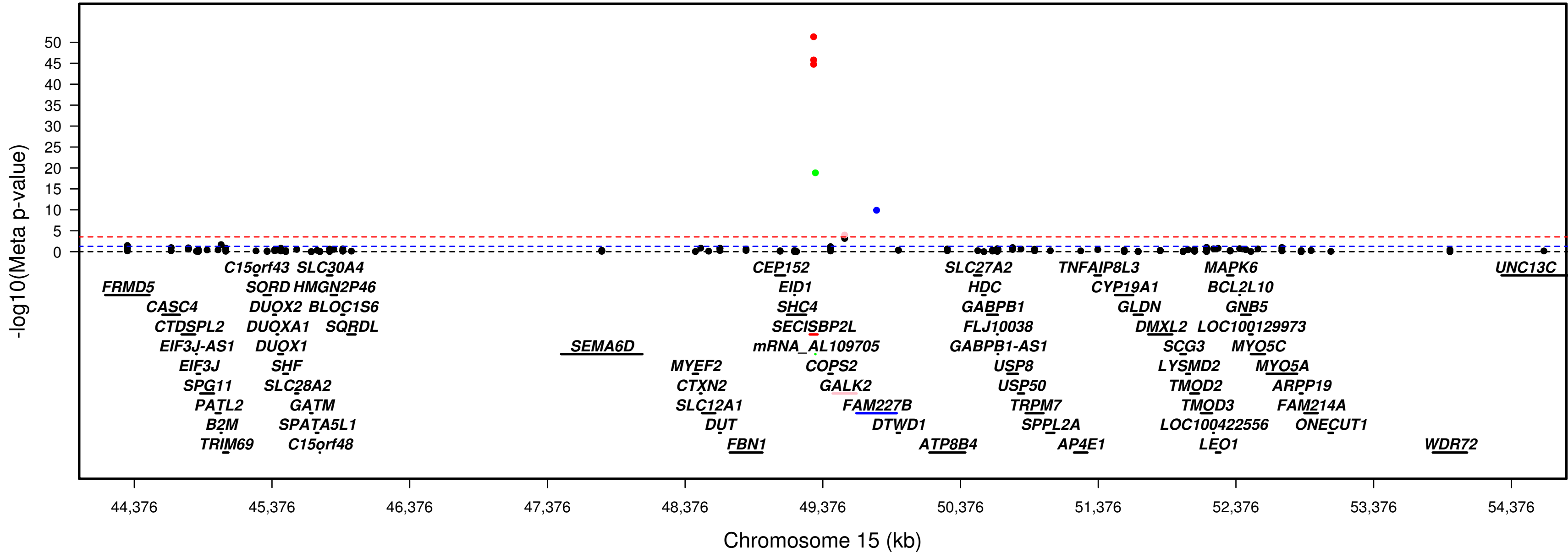
15q25.1 - rs55781567
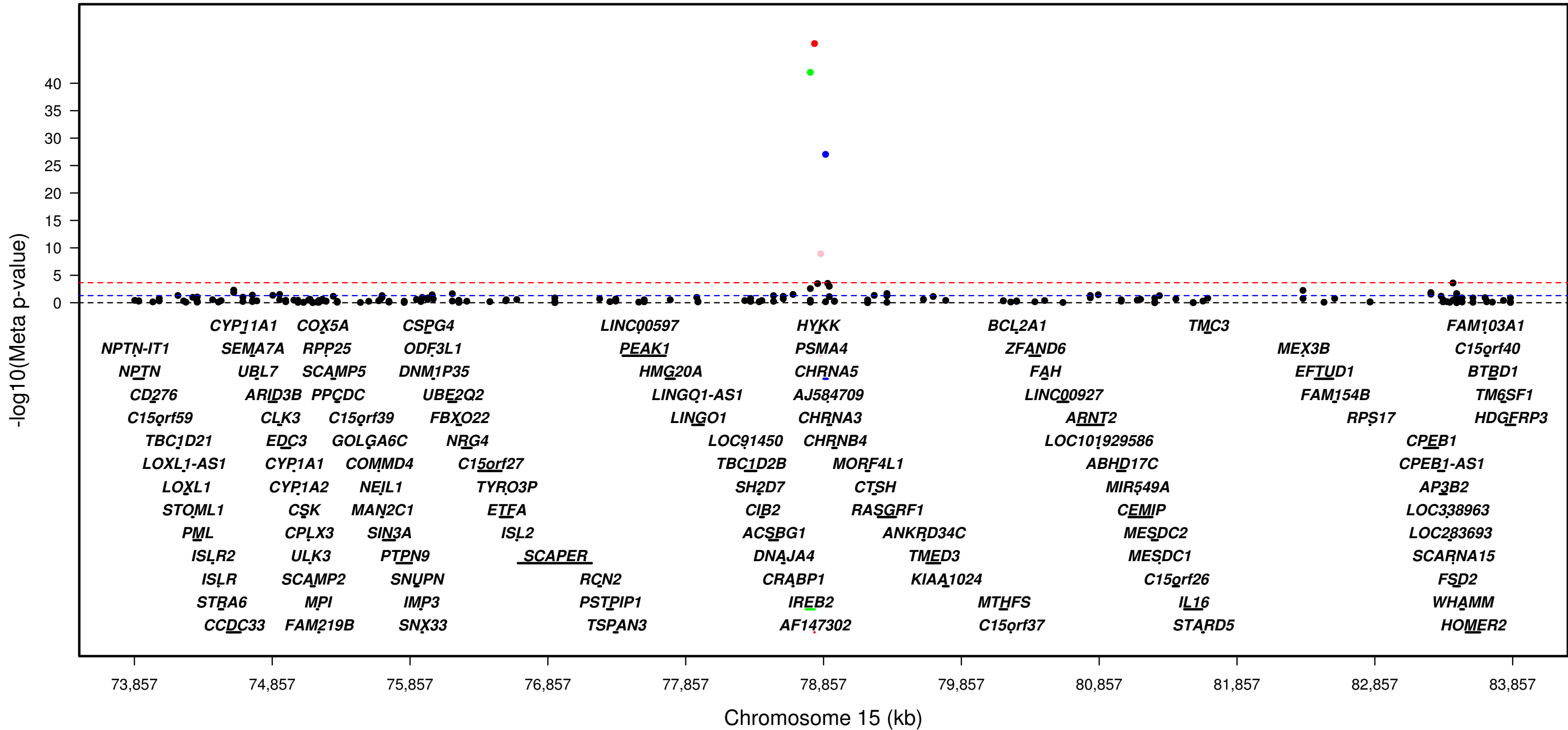
19q13.2 - rs56113850
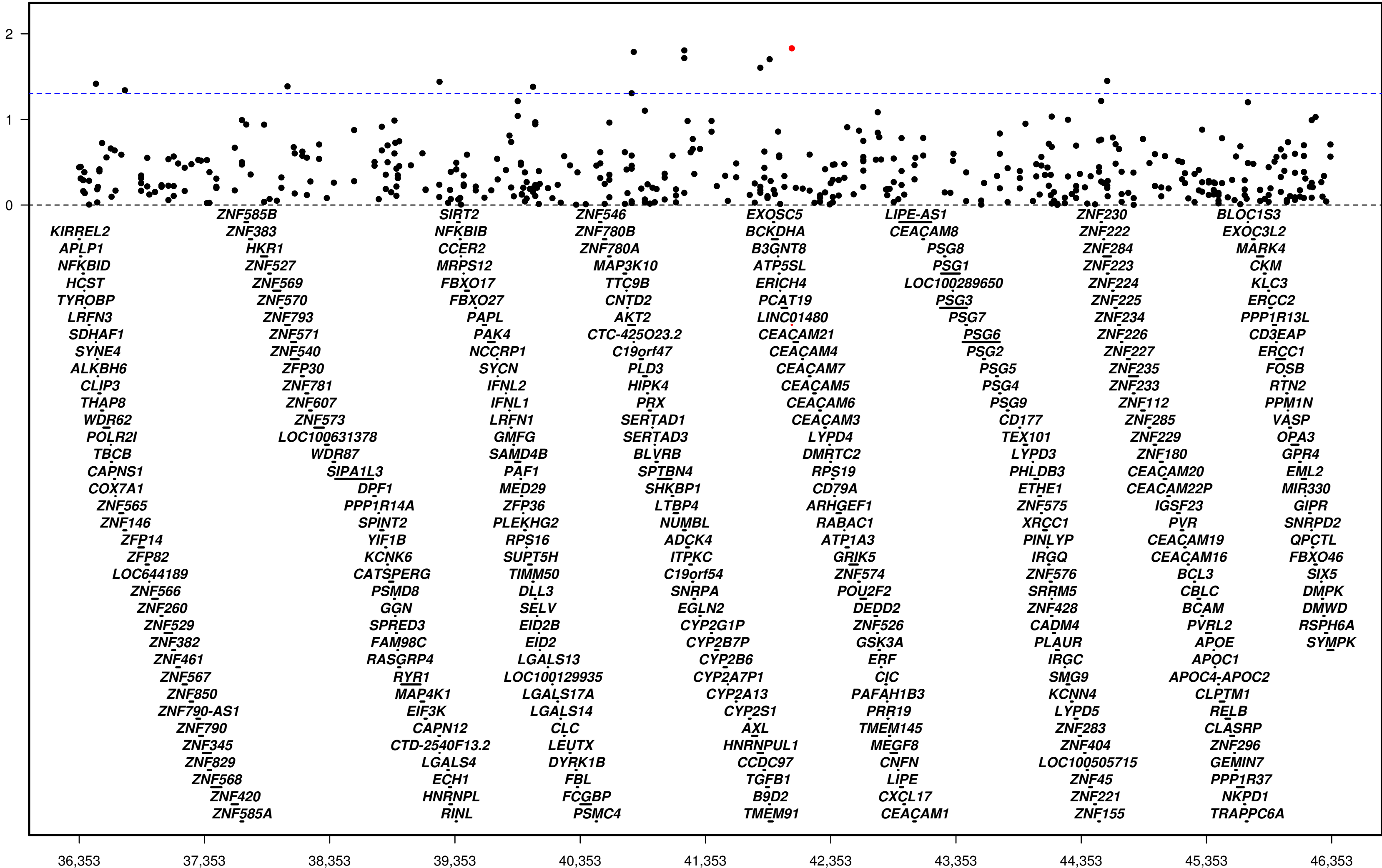
20q13.33 - rs41309931
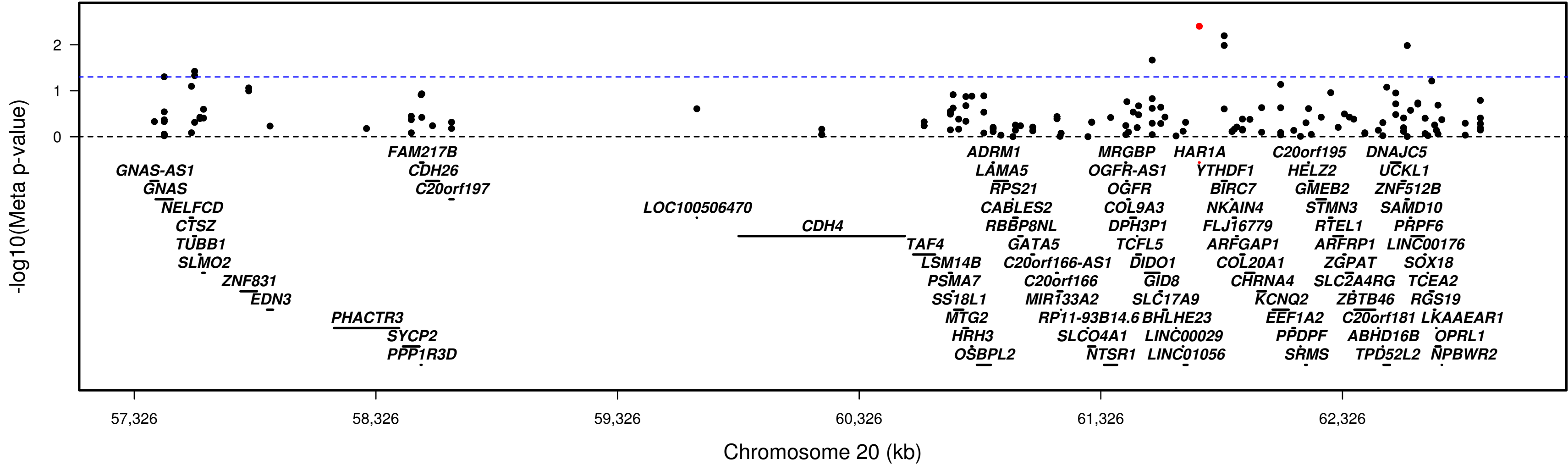
22q12.1 - rs17879961
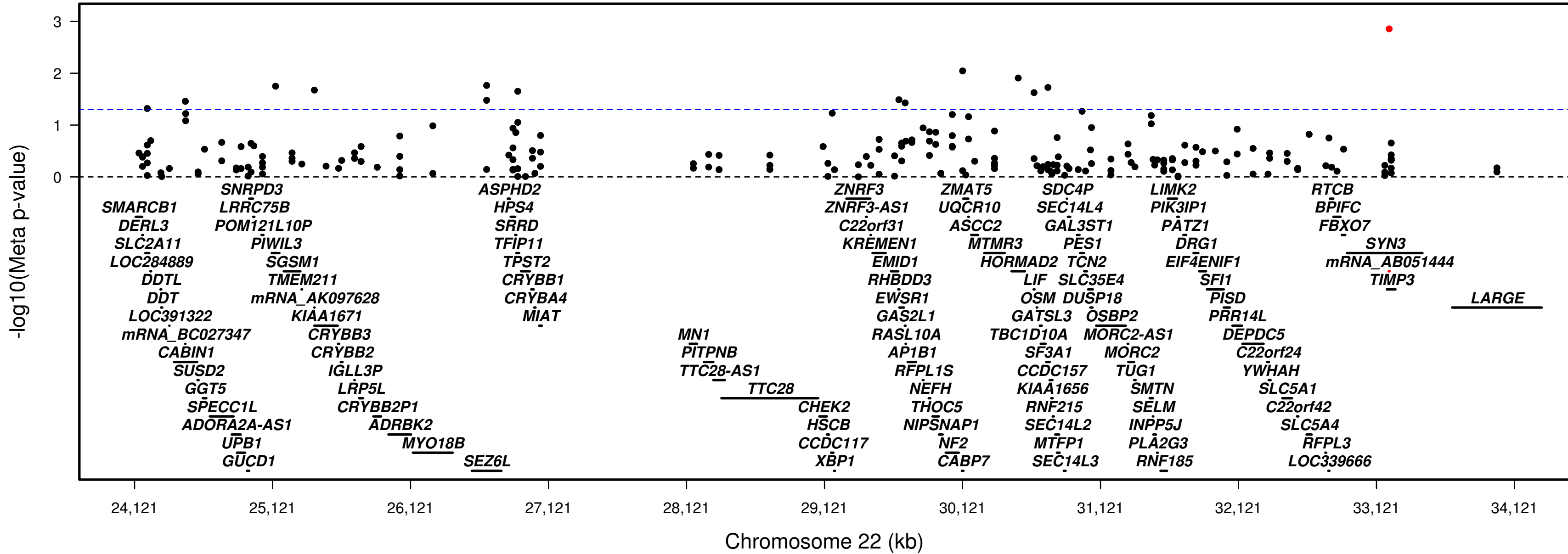
Supplementary Figure 4

Stratified results and regional plots. Left panel, forrest plots for the 18 sentinel variants and association with lung cancer (Overall) as well as the association stratified by major lung cancer histological subtypes (Adeno, lung adenocarcinoma,Squam, squamous cell lung cancer, Small, small cell lung cancer) and smoking status (ever vs never smokers). Phet; P-value for heterogeneity. Right panel. Regional plots of each susceptibility locus. Y-axis association with lung cancer (or lung cancer histological subtype as appropriate). X-axis genomic position (hg19). Red diamond indicates the sentinel variant (squares other variants at the loci), coloring indicates the degree of linkage disequilibrium (r^2) relative to the sentinel variant (calculated using 1000G CEU individuals).

Lung cancer 1p31.1 rs71658797

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.14	1.09–1.18
By histology (Phet=0.47)				
Adeno	11273	55483	1.18	1.12–1.25
Squam	7426	55627	1.13	1.06–1.20
Small	2664	21444	1.13	1.03–1.24
By Smoking status (Phet=0.59)				
Ever	23223	16964	1.15	1.09–1.20
Never	2355	7504	1.19	1.05–1.34
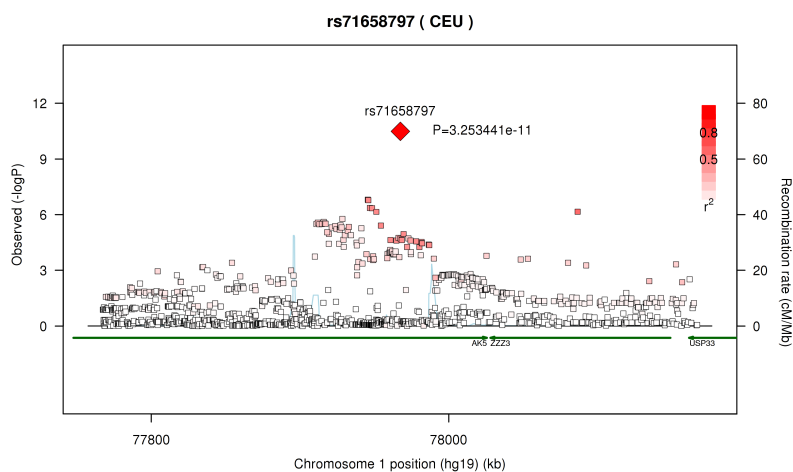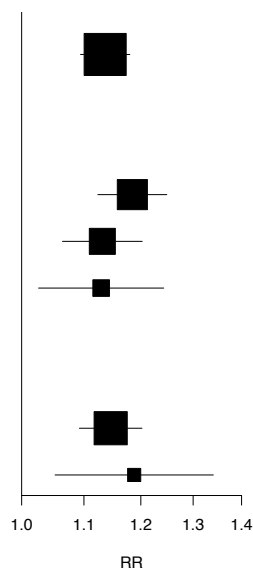
Lung cancer 6q27 rs6920364

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.07	1.05–1.10
By histology (Phet=0.84)				
Adeno	11273	55483	1.07	1.04–1.11
Squam	7426	55627	1.07	1.03–1.11
Small	2664	21444	1.09	1.03–1.15
By Smoking status (Phet=0.34)				
Ever	23223	16964	1.06	1.03–1.09
Never	2355	7504	1.10	1.03–1.18
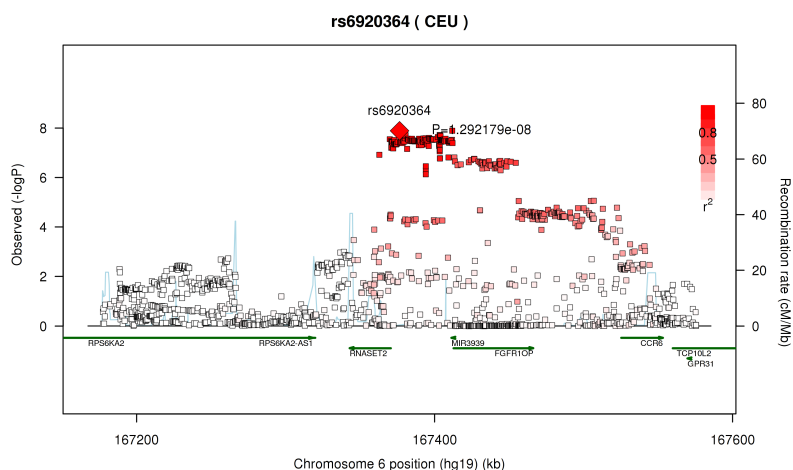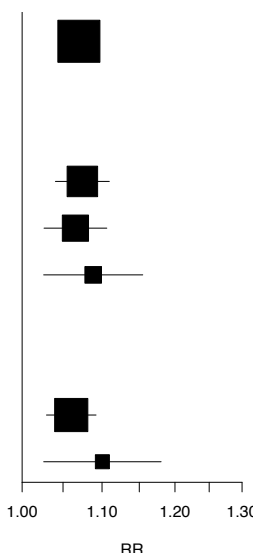
Lung cancer 8p21.1 rs11780471

	Cases	Controls	OR	95%CI
Overall	29266	56450	0.87	0.83–0.91
By histology (Phet=0.99)				
Adeno	11273	55483	0.87	0.81–0.93
Squam	7426	55627	0.87	0.81–0.94
Small	2664	21444	0.88	0.78–0.99
By Smoking status (Phet=0.34)				
Ever	23223	16964	0.86	0.81–0.92
Never	2355	7504	0.93	0.81–1.07
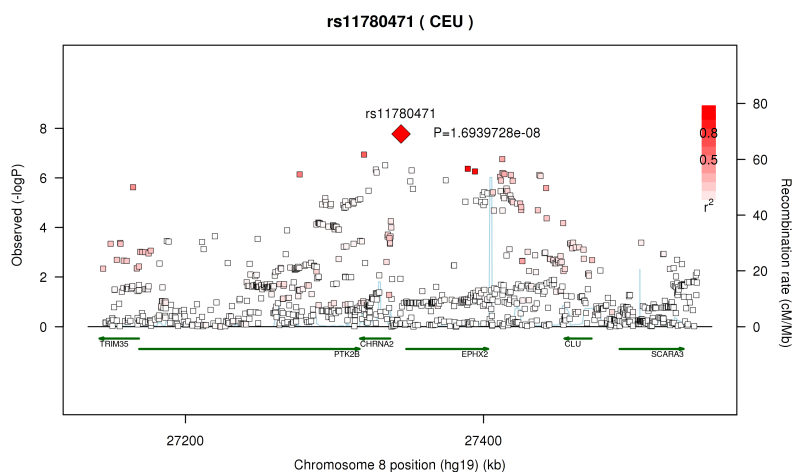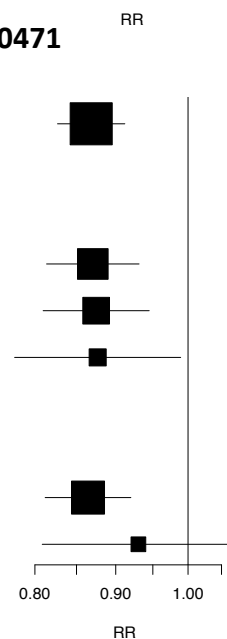
Lung cancer 13q13.1 rs11571833

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.60	1.43–1.80
By histology (Phet=0.0001)				
Adeno	11273	55483	1.30	1.10–1.54
Squam	7426	55627	2.12	1.77–2.55
Small	2664	21444	2.19	1.66–2.90
By Smoking status (Phet=0.58)				
Ever	23223	16964	1.49	1.29–1.73
Never	2263	7343	1.68	1.14–2.47
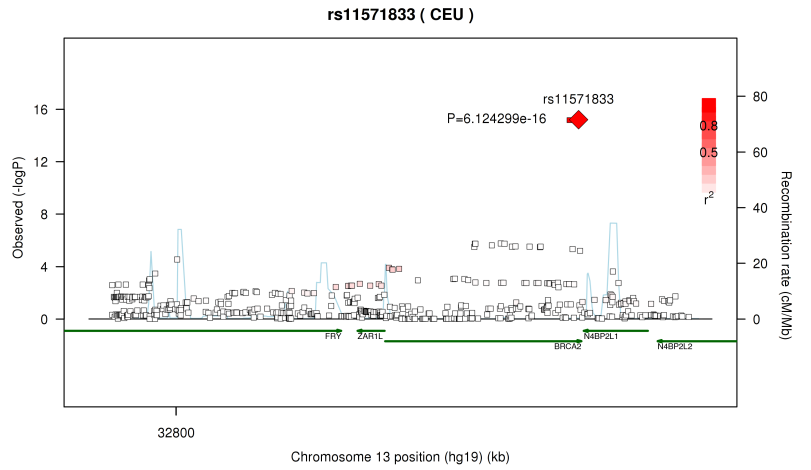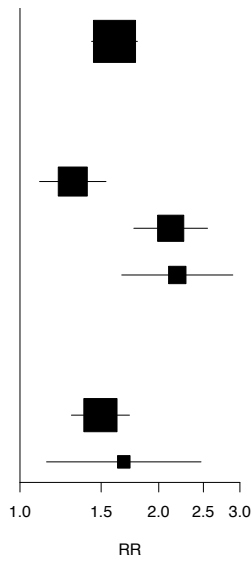
Lung cancer 15q21.1 rs66759488

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.07	1.04–1.10
By histology (Phet=0.29)				
Adeno	11273	55483	1.05	1.02–1.09
Squam	7426	55627	1.06	1.02–1.11
Small	2664	21444	1.11	1.05–1.18
By Smoking status (Phet=0.19)				
Ever	23223	16964	1.07	1.04–1.10
Never	2355	7504	1.01	0.94–1.09
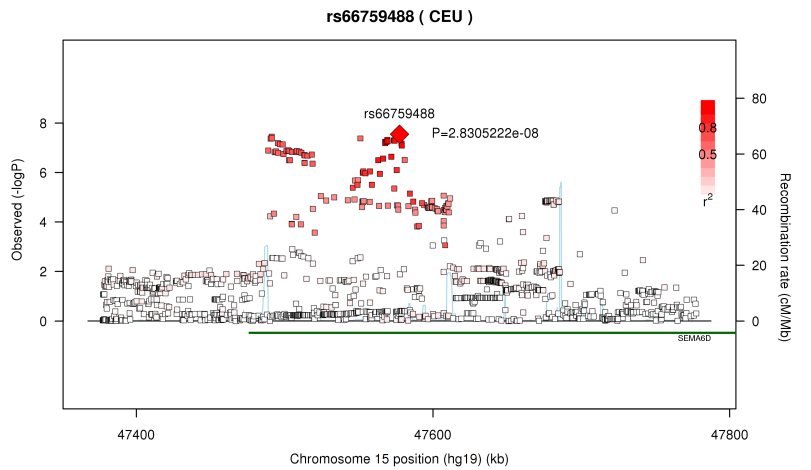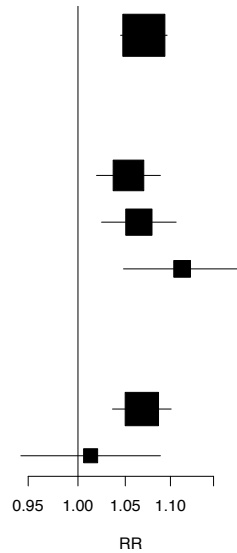
Lung cancer 15q25.1 rs55781567

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.30	1.27–1.33
By histology (Phet=0.47)				
Adeno	11273	55483	1.28	1.24–1.32
Squam	7426	55627	1.30	1.25–1.35
Small	2664	21444	1.33	1.25–1.41
By Smoking status (Phet=3X10−12)				
Ever	23223	16964	1.33	1.29–1.37
Never	2355	7504	1.01	0.94–1.08
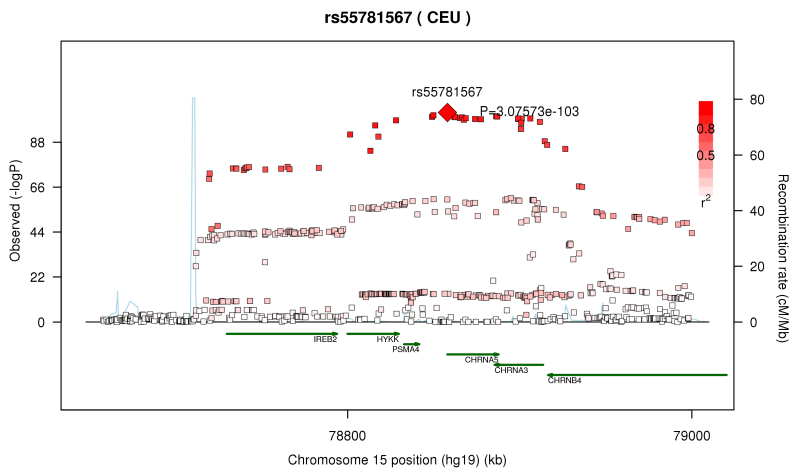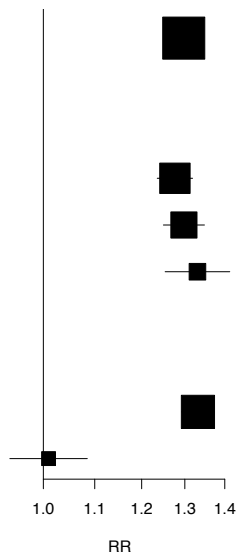
Lung cancer 19q13.2 rs56113850

	Cases	Controls	OR	95%CI
Overall	25583	51525	0.88	0.86–0.91
By histology (Phet=0.54)				
Adeno	10137	51525	0.89	0.86–0.93
Squam	6209	51525	0.87	0.83–0.90
Small	2276	18476	0.89	0.83–0.96
By Smoking status (Phet=0.002)				
Ever	19706	13322	0.88	0.85–0.91
Never	2196	6251	1.00	0.93–1.09
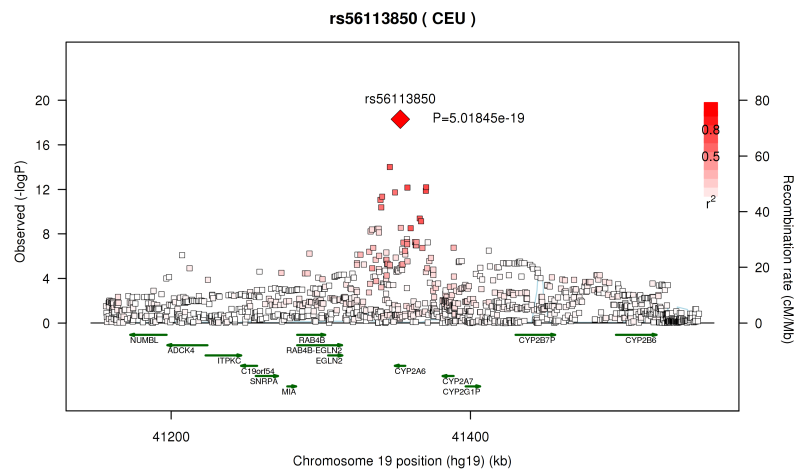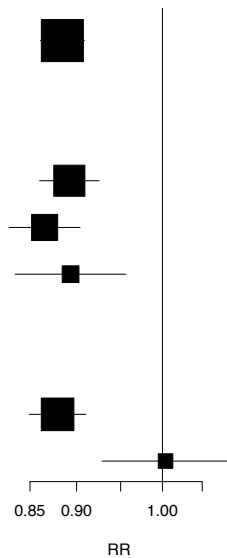
Adenocarcinoma 3q28 rs13080835

	Cases	Controls	OR	95%CI
Overall	29266	56450	0.94	0.92–0.97
By histology (Phet=1X10−6)				
Adeno	11273	55483	0.90	0.87–0.92
Squam	7426	55627	1.01	0.97–1.05
Small	2664	21444	0.99	0.94–1.05
By Smoking status (Phet=0.02)				
Ever	23223	16964	0.96	0.93–0.99
Never	2355	7504	0.88	0.82–0.94
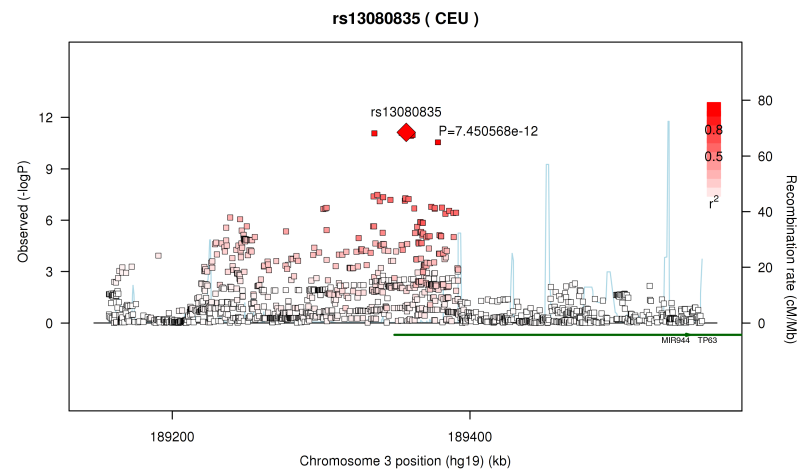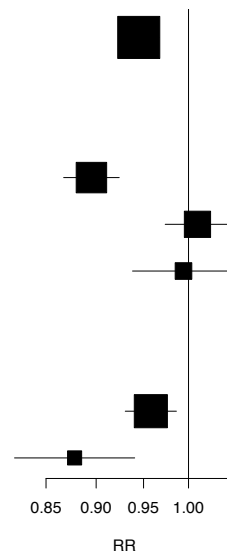
Adenocarcinoma 5p15.33 rs7705526

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.12	1.10–1.15
By histology (Phet=1X10−14)				
Adeno	11273	55483	1.25	1.21–1.29
Squam	7426	55627	1.01	0.97–1.05
Small	2664	21444	1.06	0.99–1.13
By Smoking status (Phet=0.02)				
Ever	23223	16964	1.11	1.07–1.15
Never	2355	7504	1.23	1.14–1.32
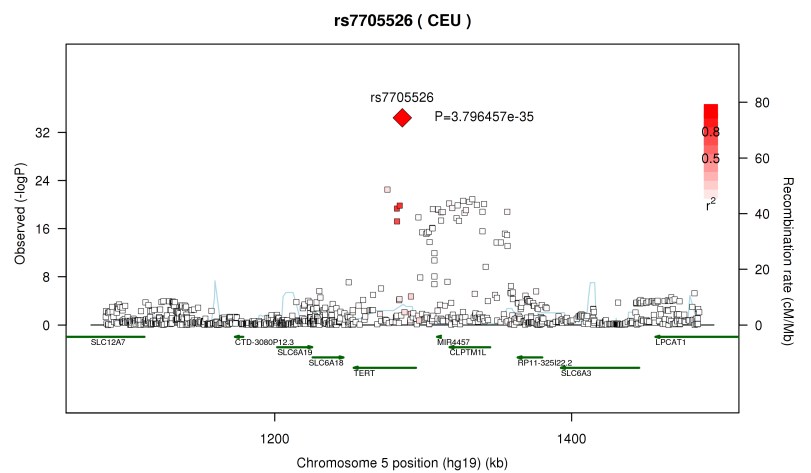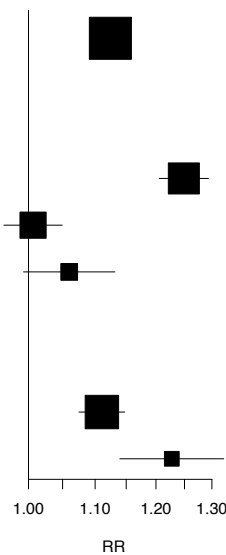
Adenocarcinoma 8p12 rs4326709

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.07	1.04–1.10
By histology (Phet=0.0003)				
Adeno	11273	55483	1.13	1.09–1.18
Squam	7426	55627	1.03	0.99–1.08
Small	2664	21444	0.98	0.92–1.06
By Smoking status (Phet=0.34)				
Ever	23223	16964	1.06	1.03–1.10
Never	2355	7504	1.11	1.02–1.21

RR: 0.95 1.05 1.15

rs4236709 (CEU)

rs4236709 P=1.277147e-10

Chromosome 8 position (hg19) (kb)

32200 32400 32600

NRG1-IT3 NRG1
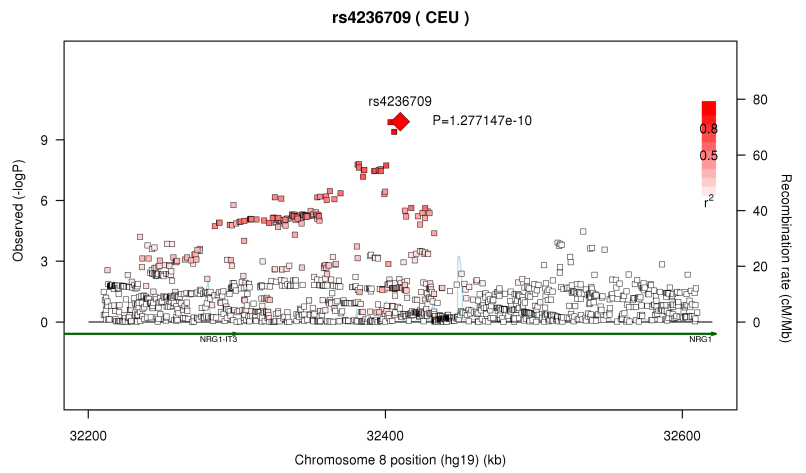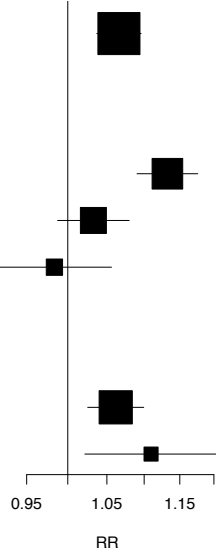
Adenocarcinoma 9p21.3 rs885518

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.09	1.05–1.13
By histology (Phet=0.002)				
Adeno	11273	55483	1.17	1.11–1.23
Squam	7426	55627	1.04	0.98–1.10
Small	2664	21444	1.01	0.92–1.11
By Smoking status (Phet=0.008)				
Ever	23223	16964	1.07	1.02–1.12
Never	2355	7504	1.25	1.12–1.39

RR: 1.0 1.1 1.2 1.3

rs885518 (CEU)

rs885518 P=9.9587e-10

Chromosome 9 position (hg19) (kb)

21800 22000

MTAP CDKN2A CDKN2B-AS1 CDKN2B
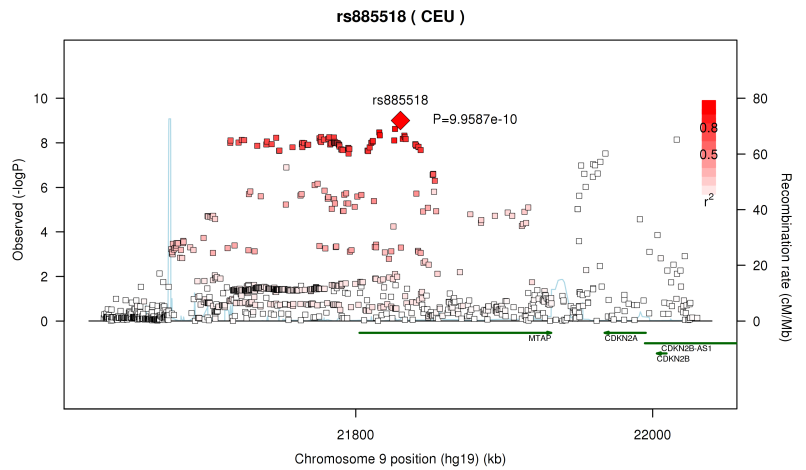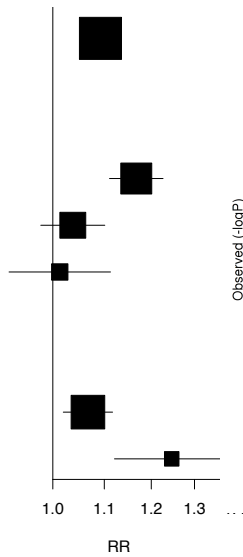
Adenocarcinoma 10q24.33 rs11591710

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.07	1.04–1.11
By histology (Phet=1X10−5)				
Adeno	11273	55483	1.16	1.11–1.22
Squam	7426	55627	0.99	0.93–1.04
Small	2664	21444	1.02	0.94–1.11
By Smoking status (Phet=0.23)				
Ever	23223	16964	1.06	1.02–1.10
Never	2355	7504	1.13	1.03–1.25

RR: 0.95 1.05 1.15 1

rs11591710 (CEU)

rs11591710 P=6.29897e-11

Chromosome 10 position (hg19) (kb)

105600 105800

SH3PXD2A OBFC1 SLK COL17A1 MIR936 SFR1
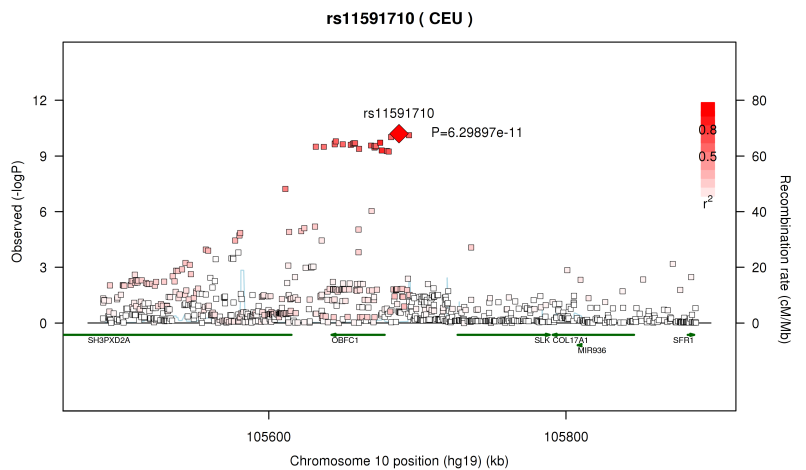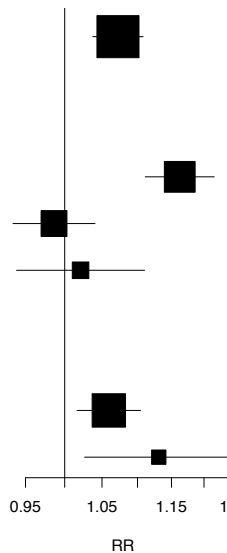
Adenocarcinoma 11q23.3 rs1056562

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.07	1.04–1.09
By histology (Phet=0.004)				
Adeno	11273	55483	1.11	1.07–1.14
Squam	7426	55627	1.02	0.98–1.06
Small	2664	21444	1.05	0.99–1.11
By Smoking status (Phet=0.64)				
Ever	23223	16964	1.07	1.04–1.10
Never	2355	7504	1.05	0.98–1.12

rs1056562 (CEU)

rs1056562 P=2.756228e-10
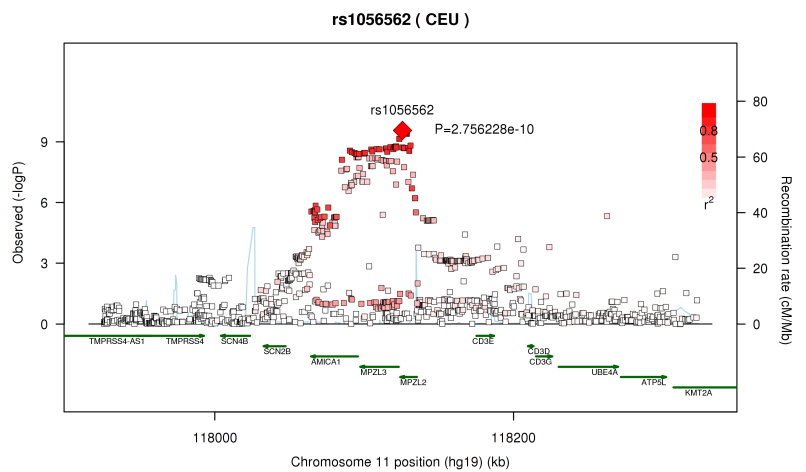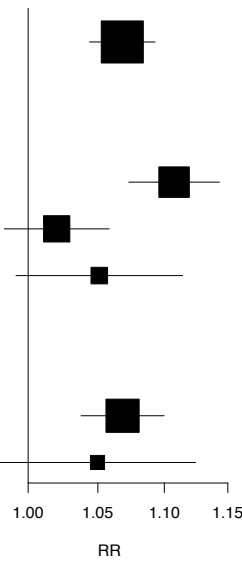
Adenocarcinoma 15q21.1 rs77468143

	Cases	Controls	OR	95%CI
Overall	29266	56450	0.92	0.90–0.95
By histology (Phet=2X10−8)				
Adeno	11273	55483	0.86	0.83–0.89
Squam	7426	55627	1.01	0.97–1.05
Small	2664	21444	0.97	0.91–1.04
By Smoking status (Phet=0.99)				
Ever	23223	16964	0.92	0.89–0.95
Never	2355	7504	0.92	0.85–0.99

rs77468143 (CEU)

rs77468143 P=1.686382e-16
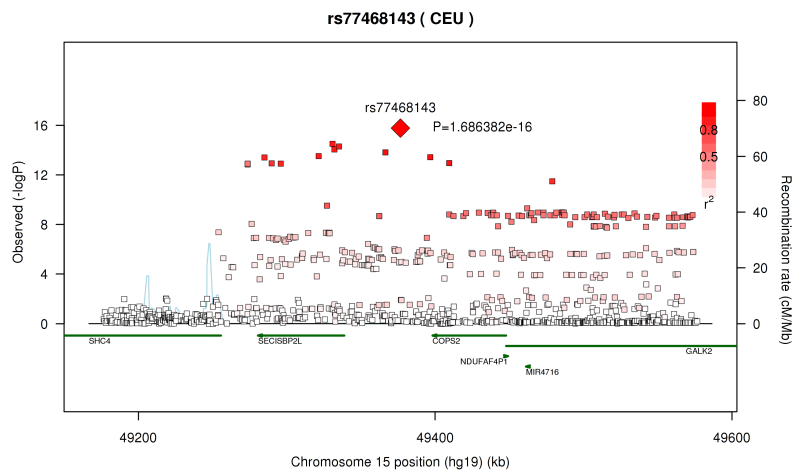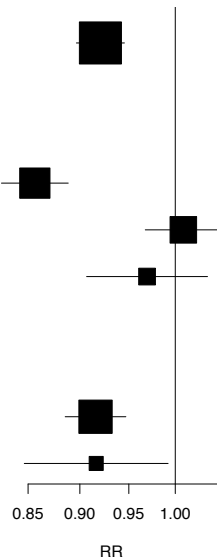
Adenocarcinoma 20q13.33 rs41309931

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.08	1.04–1.12
By histology (Phet=0.0002)				
Adeno	11273	55483	1.17	1.11–1.23
Squam	7426	55627	1.00	0.94–1.06
Small	2664	21444	1.05	0.95–1.15
By Smoking status (Phet=0.69)				
Ever	23223	16964	1.07	1.02–1.13
Never	2355	7504	1.10	0.98–1.23

rs41309931 (CEU)

rs41309931 P=1.3052529e-09
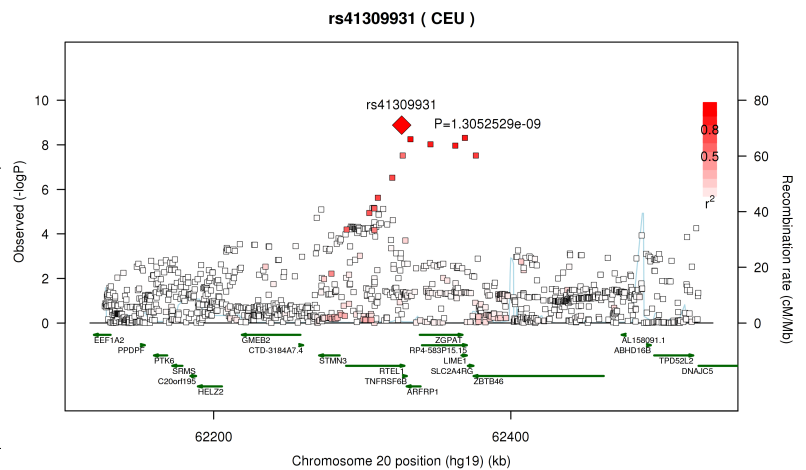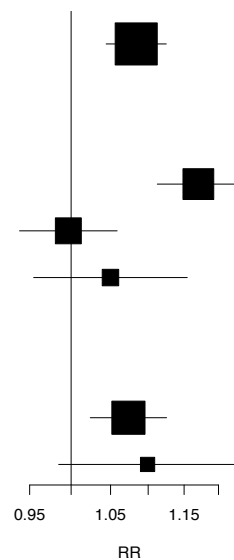
Squamous 6p21.33 rs116822326

	Cases	Controls	OR	95%CI
Overall	29266	56450	1.15	1.12–1.19
By histology (Phet=0.0001)				
Adeno	11273	55483	1.08	1.03–1.13
Squam	7426	55627	1.25	1.19–1.31
Small	2664	21444	1.18	1.09–1.27
By Smoking status (Phet=5X10−5)				
Ever	20856	14456	1.17	1.12–1.22
Never	2196	6251	0.94	0.86–1.04

RR

rs116822326 (CEU)

rs116822326
P=3.826103e-19
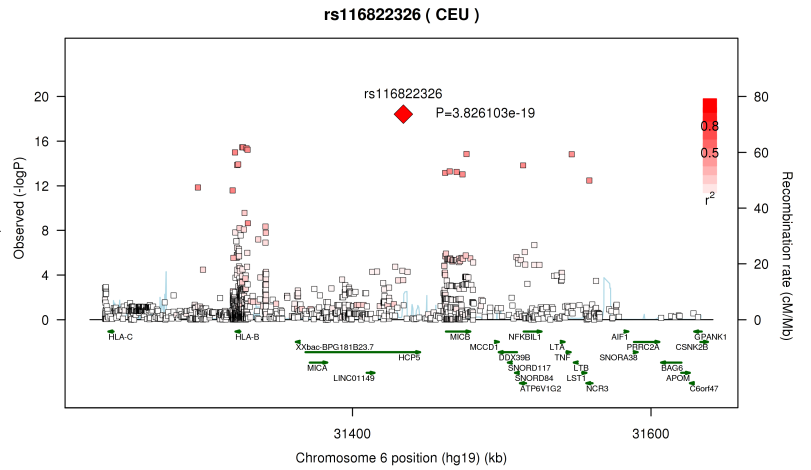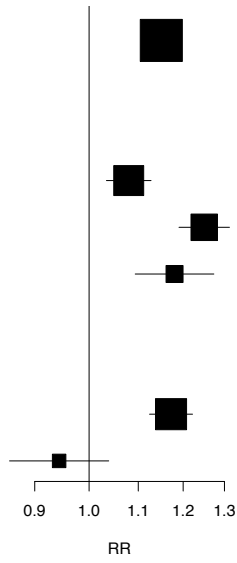
Squamous 12p13.33 rs7953330

	Cases	Controls	OR	95%CI
Overall	29266	56450	0.92	0.89–0.94
By histology (Phet=0.0001)				
Adeno	11273	55483	0.96	0.93–1.00
Squam	7426	55627	0.86	0.83–0.90
Small	2664	21444	0.87	0.81–0.92
By Smoking status (Phet=0.096)				
Ever	23223	16964	0.91	0.88–0.94
Never	2355	7504	0.97	0.90–1.05

RR

rs7953330 (CEU)

rs7953330
P=7.256218e-13
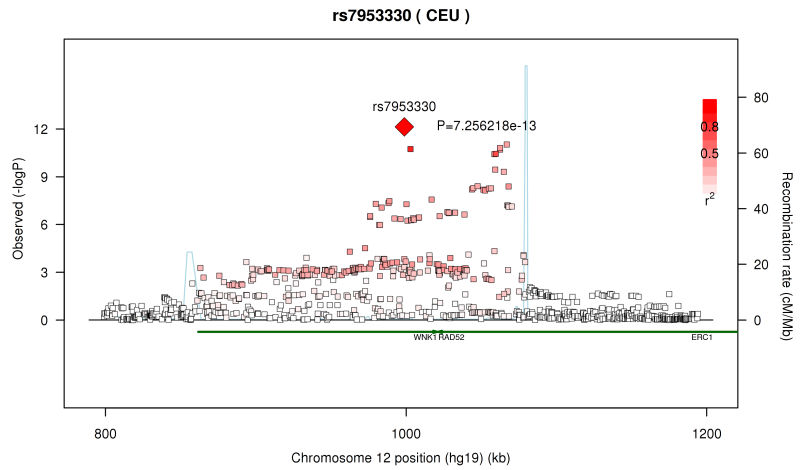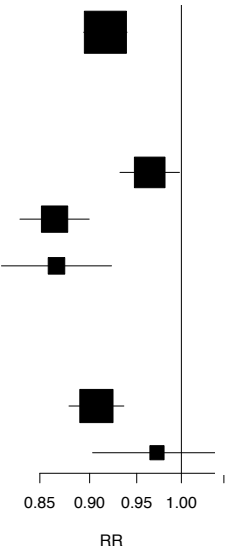
Squamous 22q12.1 rs17879961

	Cases	Controls	OR	95%CI
Overall	28282	55480	0.60	0.52–0.70
By histology (Phet=7X10−6)				
Adeno	10676	54513	0.92	0.74–1.14
Squam	7210	54657	0.41	0.32–0.52
Small	2664	21444	0.59	0.41–0.84
By Smoking status (Phet=4X10−5)				
Ever	22331	16155	0.53	0.44–0.64
Never	2263	7343	1.42	0.93–2.19

RR

rs17879961 (CEU)

rs17879961
P=5.700129e-13
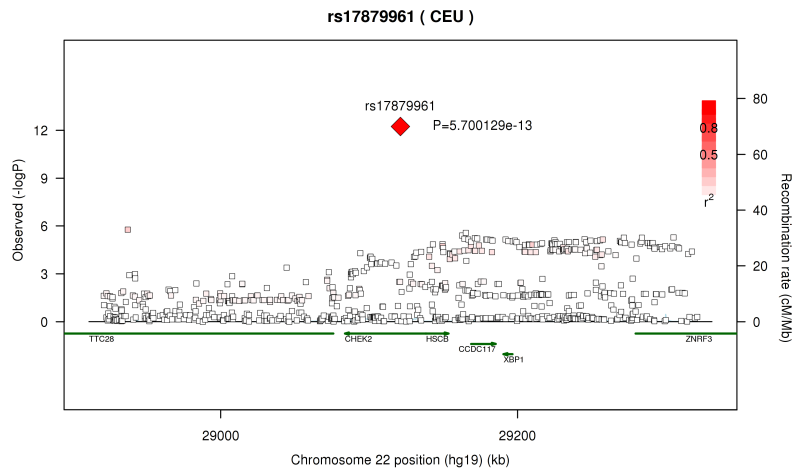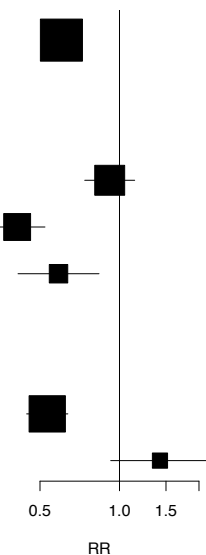
Supplementary Figure 5. Upper panels. Positions of susceptibility variants relative to genes at these loci and relevant genomic annotations. Lower Panel. eQTL box plots for association between expression levels in lung epithelium and susceptibility variants in the Laval, UBC and Groningen cohorts.

15q21.1

Chr15 position (Mb)

GENCODE Version 19
Basic Gene Annotation

Top SNPs (phenotype)

Correlated (r²>0.7) variants

IM-PET enhancers overlapping SNP

Hnisz enhancers overlapping SNP

PreSTIGE enhancers overlapping SNP

TF-chip peaks overlapping candidate SNP

Histone modification ChIP-seq peaks
overlapping candidate SNP

DNase HS peaks overlapping candidate SNP

Lung ChromHMM-25state:
ENHANCERS
PROMOTERS
TRANSCRIBED

All Hi-C interactions in IMR90

A **Laval** — P = 7.51e-35

B **UBC** — P = 1.93e-14

C **Groningen** — P = 4.08e-08

8p12

Chr8 position (Mb)

GENCODE Version 19
Basic Gene Annotation

Top SNPs (phenotype)

Correlated (r^2>0.7) variants

IM-PET enhancers overlapping SNP

Hi-C loop hub overlapping SNP

TF-chip peaks overlapping candidate SNP

Histone modification ChIP-seq peaks
overlapping candidate SNP

DNase HS and FAIRE-seq peaks
overlapping candidate SNP

Lung ChromHMM-25state
ENHANCERS
PROMOTERS
TRANSCRIBED

All Hi-C interactions in IMR90 (Rao et al., 2014)

NHGRI GWAS catalog

rs4236709 (Adeno)

NRG1 (IMR90, NHLF)

RNU6-528P (IMR90), *NRG1* (IMR90), *NRG1-IT2* (IMR90)

A   Laval
P = 0.001233

B   UBC
P = 0.1252

C   Groningen
P = 0.004915

11q23.3

Chr11 position (Mb)

117.90 | 118.00 | 118.10 | 118.20 | 118.30 | 118.40 |

GENCODE Version 19
Basic Gene Annotation

Top SNPs (phenotype)
rs1056562 (Adeno) | rs61900932 (Adeno) |
rs112753655 (Ever) |

Correlated (r²>0.7) variants

IM-PET enhancers overlapping SNP

Hnisz enhancers overlapping SNP

TF-chip peaks overlapping candidate SNP

Histone modification ChIP-seq peaks
overlapping candidate SNP

DNase HS and FAIRE-seq peaks
overlapping candidate SNP

Lung ChromHMM-25state
ENHANCERS
PROMOTERS
TRANSCRIBED

All Hi-C interactions in IMR90

(a)

(b)

SNP	Chr.	Gene	Effect allele		Telomere (CODD)			Lung adenocarcinoma				Lung squam cell carcinoma				Small cell lung cancer				Phet^
			TRICL	CODD et al	β	SE	P-value	OR	L95	U95	P-value	OR	L95	U95	P-value	OR	L95	U95	P-value	
rs10936599	3	*TERC*	T	T	−0.097	0.008	3×10^{-31}	0.924	0.890	0.959	3×10^{-5}	1.023	0.981	1.068	0.280	NA	NA	NA	NA	0.002
rs2736100	5	*TERT*	A	A	−0.078	0.009	4×10^{-19}	0.834	0.808	0.861	6×10^{-29}	1.006	0.969	1.043	0.770	0.938	0.885	0.994	0.030	2×10^{-13}
rs7675998	4	*NAF1*	A	A	−0.074	0.009	4×10^{-16}	0.915	0.880	0.950	4×10^{-6}	0.980	0.938	1.024	0.370	0.982	0.917	1.053	0.620	0.040
rs9420907^^^	10	*OBFC1*	*C***	*A*	−0.069	0.01	7×10^{-11}	1.149	1.100	1.201	5×10^{-10}	0.981	0.932	1.034	0.480	1.014	0.935	1.100	0.730	0.000
rs8105767	19	*ZNF208*	*G***	*A*	−0.048	0.008	1×10^{-9}	1.042	1.006	1.079	0.020	0.982	0.943	1.023	0.380	0.997	0.934	1.063	0.920	0.080
rs755017^^	20	*RTEL1*	*G***	*A*	−0.062	0.011	7×10^{-9}	1.042	0.992	1.094	0.100	0.995	0.939	1.053	0.860	1.022	0.934	1.119	0.630	0.480
rs11125529	2	*ACYP2*	*A***	*C*	−0.056	0.01	5×10^{-8}	1.028	0.982	1.076	0.240	1.004	0.952	1.058	0.890	1.038	0.954	1.129	0.380	0.720
rs2967374*	16	*MPHOSPH6*	*A***	*G*	−0.045	0.009	3×10^{-7}	1.085	1.045	1.127	2×10^{-5}	1.038	0.993	1.085	0.100	1.011	0.942	1.085	0.770	0.13

*Variant was not consistently replicated as being associated with telomere length in Codd et al Nat Genetics 2014

**Variant effect alleles flipped between TRICL and Codd et al

^heterogeneity test between estimates across lung cancer histologies

^^r^2 between the lung adenocarcinoma sentinel variant rs41309931 and rs755017 is 0.002

^^^r^2 between the lung adenocarcinoma sentinel variant rs11591710 and rs9420907 is 0.96

Supplementary Figure 6. Panel (a). Scatter plots variants at 10q23 (left) and 20q13.33 (right) comparing the association with lung adenocarcinoma (Y-axis) and association with Telomere length (TL) (X-axis). At 10q23 the variants associated with lung adenocarincoma appear to be those associated with TL. At 20q13.33 however, the allele associated with lung adenocarcinoma is not clearly associated with TL. **Panel (b).** Association between variants associated with telomere length and lung cancer histological subtypes (alleles normalized to the TRICL reference allele).
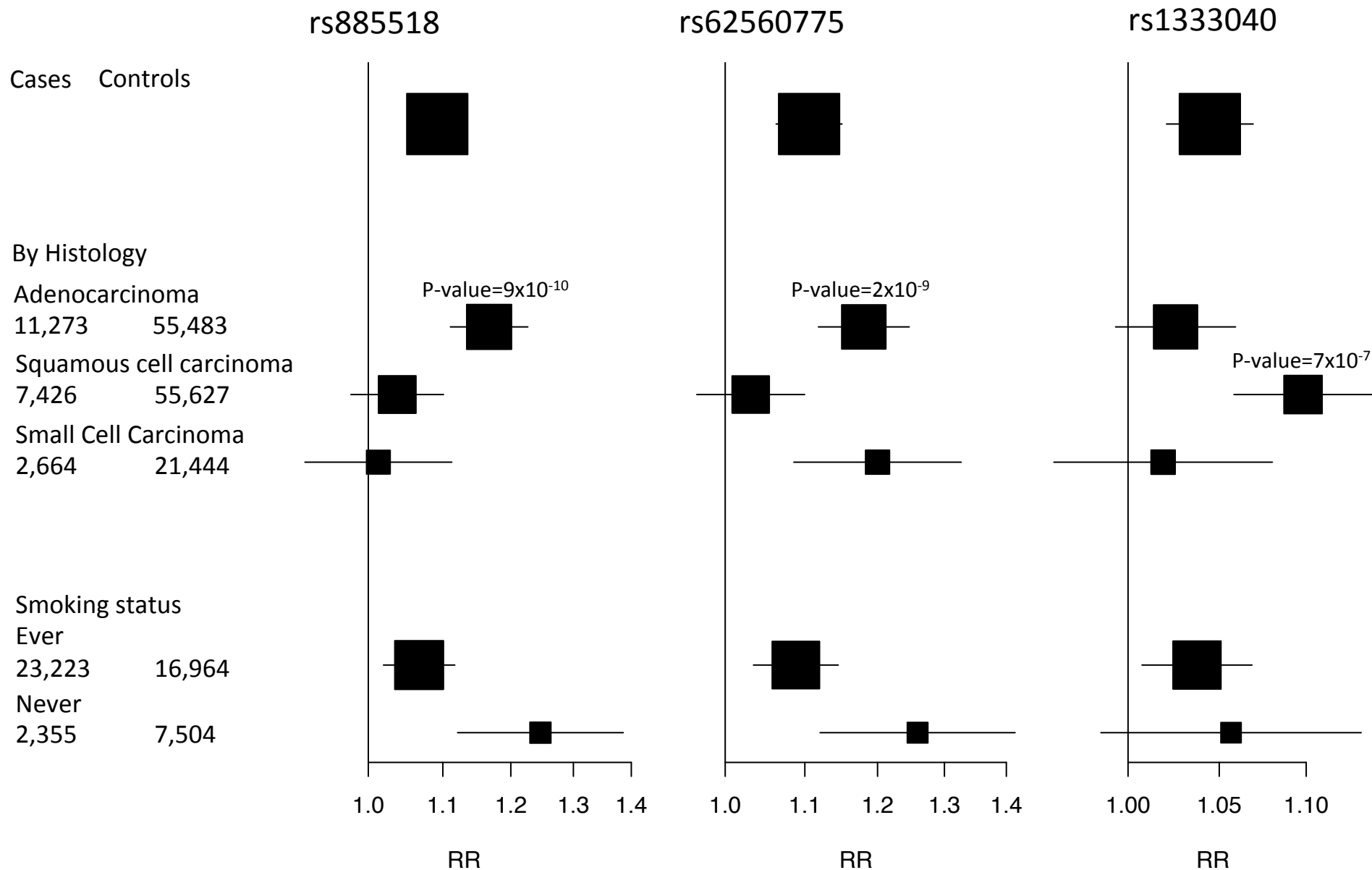
Supplementary Figure 7. Forest plots for three uncorrelated variants at 9p21.3. These three variants appear relevant to different lung cancer histological subtypes at this locus; rs885518 associated with lung adenocarcinoma uniquely, rs62560775 associated with lung adenocarcinoma/small cell carcinoma and rs1333040 associated with squamous cell carcinoma.

Supplementary Figure 8. CPD/Genetic correlations.

a. Conditional Quantile-Quantile plots for smoking intensity (expressed as cigarettes per day, CPD) and lung cancer risk based on TRICL-ILCCO OncoArray data. The X- and Y-axis represent the -log10(p-value) for lung cancer. The color coding represent different p-value thresholds based on the associations with smoking intensity (expressed as cigarettes per day, CPD) (supplementary methods). Increasing deviations from the expected line with increasing threshold would indicate excessive number of variants associated with both CPD and lung cancer. After excluding 15q25 and *CYP2A6*, We observed some evidence of increasing level of deviations from expected distribution along with increasing CPD for lung squamous cell carcinoma, but not for adenocarcinoma. This is compatible with the etiology of SQC and adenocarcinoma.

b. Genetic correlations between association with lung cancer smoking intensity (expressed as cigarettes per day, CPD) from the TRICL-ILCCO Oncoarray data. Excluded regions include the 15q25 region surrounding the *CHRNA5* gene and 19q13.2 *CYP2A6* both loci well described to be linked with both lung cancer and smoking behavior (supplementary methods). Both Panels 1 and 2 imply correlation between the alleles associated with propensity to smoke and those that are related with lung cancer and beyond those described by the 15q25 and 19q13.2 loci. Both observations were most prominent in lung cancer, but less marked in lung adenocarcinoma.

c. Genetic correlations between association with lung cancer (TRICL-ILCCO Oncoarray data) and traits investigated by the TAG consortium. The phenotypes include A. CPD, B. Ever vs never, and C. smoking cessation that contrasted former versus current smokers (supplementary methods). The genetic correlation observed with the TRICL lung cancer and CPD was also present in this second dataset. We additionally noted that the genetic alleles making it more difficult to quit smoking are correlated with lung cancer risk.

(a)

(b)

	N SNPs	Genetic correlation (rG) with Cigarettes per day (CPD)		Lower rG	Upper rG	P-value
None	1,008,696		0.56	0.32 : 0.80		4.1×10^{-6}
Excluding regions	1,004,089		0.50	0.27 : 0.74		1.9×10^{-5}
Adenocarcinoma						
None	1,008,188		0.37	0.04 : 0.70		0.03
Excluding regions	1,003,582		0.30	-0.03 : 0.64		0.08
Squamous Cell						
None	1,008,011		0.45	0.15 : 0.74		2.9×10^{-3}
Excluding regions	1,003,405		0.37	0.05 : 0.70		0.02
Small Cell						
None	1,008,093		0.54	0.24 : 0.84		4.9×10^{-4}
Excluding regions	1,003,485		0.50	0.17 : 0.83		3.2×10^{-3}
Ever smokers						
None	1,008,244		0.53	0.22 : 0.83		6.9×10^{-4}
Excluding regions	1,003,634		0.47	0.15 : 0.78		4.1×10^{-3}
Never smokers						
None	1,007,892		0.41	-1.22 : 2.04		0.62
Excluding regions	1,003,285		0.38	-1.49 : 2.25		0.69

-1 -0.5 0 0.5 1

(c)

Lung Cancer	Smok stat	N SNPs	Genetic correlation (rG)	Lower rG : Upper rG	P-value
Overall	CPD	924,716	0.78	0.48 : 1.08	3.9×10^{-7}
	Smoking	924,708	0.50	0.32 : 0.67	2.4×10^{-8}
	Cessation	924,634	-0.56	-0.77 : -0.35	1.9×10^{-7}
Adenocarcinoma	CPD	924,385	0.55	0.22 : 0.88	1.1×10^{-3}
	Smoking	924,374	0.40	0.21 : 0.59	2.9×10^{-5}
	Cessation	924,302	-0.40	-0.65 : -0.16	1.3×10^{-3}
Squamous cell	CPD	924,248	0.84	0.44 : 1.23	3.2×10^{-5}
	Smoking	924,237	0.45	0.24 : 0.65	2.6×10^{-5}
	Cessation	924,171	-0.59	-0.86 : -0.31	3.7×10^{-5}
Small cell	CPD	923,798	0.63	0.25 : 1.01	1.3×10^{-3}
	Smoking	923,789	0.47	0.24 : 0.70	4.8×10^{-5}
	Cessation	923,717	-0.57	-0.87 : -0.27	1.8×10^{-4}
Ever smokers	CPD	923,883	0.73	0.40 : 1.06	1.6×10^{-5}
	Smoking	923,881	0.29	0.10 : 0.47	2.2×10^{-3}
	Cessation	923,802	-0.49	-0.74 : -0.25	9.1×10^{-5}
Never smokers	CPD	923,769	0.16	-1.19 : 1.52	0.81
	Smoking	923,756	0.85	-2.35 : 4.05	0.60
	Cessation	923,686	-0.02	-0.90 : 0.87	0.97

Axis: -1, -0.5, 0, 0.5, 1, 1.5

Supplementary Table 1 Characteristics of OncoArray Studies sorted by geographical regions, study design and recruitment period.

Acronym	Study Name	Institute	PI	Country	study design	study period	Case no*.	Control no*.	Total	Ref (PMID)
CARET	The Carotene and Retinol Efficacy Trial	FHCRC	J. Doherty, C. Chen	USA	Cohort	Recruitment 1985-1996.	612	613	1225	15572756
PLCO	The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial	NCI	N. Caporaso	USA	Cohort	1992-2001	1700	1437	3137	22031728
MEC	Multiethnic Cohort Study	U of Hawaii, USC	L. Le Marchand	USA	Cohort	Recruitment 1993-1996	918	962	1880	15229477
NELCS	New England Lung Cancer Study	Dartmouth College	A. Andrew	USA	Pop. CC	2005-2007	189	184	373	20049123
HLCS	Harvard Lung Cancer Study	Harvard/Mass General Hospital	D. Christiani	USA	Hosp. CC	1992-2004	3170	827	3997	16214908
MDACC	MD Anderson Cancer Center Study	MDACC	X. Wu	USA	Hosp. CC	1994-current	999	1000	1999	17470739
Tampa	Tampa Lung Cancer Study	WSU	P. Lazarus	USA	Hosp. CC	1999-2003	403	400	803	
BioVU	Vanderbilt 2	Vanderbilt	M. Aldrich	USA	hospital CC	2007-present	798	799	1597	
LCRI-DOD	LCRI-DOD	Markey Cancer Center	S. Arnold	USA	Pop. CC	2012-2014	99	137	236	
TLC	Total Lung Cancer: Molecular Epidemiology of Lung Cancer Survival	Moffitt Cancer Center, Tampa, FL	M.B. Schabath	USA	case only	2012--present	454		454	23839018
Canada	Canadian screening study	UHN, BCCA	Lam/Tsao/Liu	Canada	screening cohort	2004-2011, 2008-2013	267	463	730	24004118
MSH-PMH	Mount Sinai Hospital-Princess Margaret Hospital Study	MSH, PMH	R.J. Hung, G. Liu	Canada	Hosp. CC	2008-2012	1436	998	2456	24880342 24947688
CAPUA	Cancer de Pulmon en Asturias	U. Oviedo	A. Tardon	Spain	Hosp.CC	2002-2012	796	772	1568	23013535
ATBC	The Alpha-Tocopherol, Beta-Carotene Cancer Prevention	NCI	D. Albanes	Finland	Cohort	1985-1993	1138	670	1808	8205268
NSHDS	Northern Sweden Health and Disease Study	Umeå University	M. Johansson	Sweden	Cohort	1985-present	244	243	487	14660243
MDCS	The Malmö Diet and Cancer Study	Lund University	J. Manjer	Sweden	Cohort	1991-1996	169	176	345	11916347
EPIC	European Prospective Investigation into Cancer and Nutrition	IARC	M. Johansson	Europe	Cohort	1992-1999	1221	1235	2456	12639222
LLP	Liverpool Lung Cancer Project	U. Liverpool	J.K. Field	UK	Cohort	1999-2007, 1999-2011	457	513	970	25873368
Norway	Norway Lung Cancer Study	NIOH	A. Haugen	Norway	Pop. CC	1986-2005	335	422	757	18258609
EAGLE	Environment and Genetics in Lung Cancer Study Etiology	NCI	M.T. Landi	Italy	Pop.CC	2002-2005	1845	1789	3634	18538025
Nijmegen	The Nijmegen Lung Cancer Study	Radboud University Medical Centre	L. A. Kiemeney	The Netherlands	Pop. CC	2002-2008	453	457	910	Control: 17568781; Case: 20418888

NICCC-LCA	Clalit National Israeli Cancer Control Center- lung cancer study	Carmel Medical Center & Technion	G. Rennert	Israel/ Europe	Pop.CC	2005-ongoing	714	554	1268	25924736
L2	The IARC L2 Study	IARC	P. Brennan	CE Europe	Pop./Hosp. CC	2005-2013	1134	1134	2268	
Copenhagen	Copenhagen lung cancer study	Copenhagen	S. E. Bojesen	Denmark	Hosp.CC	2004 (controls) 2012-2013 (cases)	657	1460	2117	22441734
Germany	German lung cancer study	HMGU, DKFZ	H. Bickeböller, A. Risch	Germany	Mixed CC	1998-2013	1075	602	1677	18483334
ReSoLucent	ReSoLucent	U. Sheffield,	Woll	UK	Mixed CC	2005-2014	689	443	1132	
						TOTAL	**23878**	**19520**	**43398**	

Hosp. Hospital-based; Pop., Population-based; CC, case-control study; CE, Central and Eastern

[*] numbers of unique cases and controls submitted for genotyping (without expected duplicates and pre-quality control filters)

Supplementary Table 3a. Observed heritability explained by SNPs for overall lung cancer as well as for subtypes as defined by histology and smoking status. Analyses are based on HapMap 3 SNPs with MAF > 0.01.

	Observed h^2_g (se)	(Maximum) Number of cases	(Maximum) Number of controls
Overall lung cancer	0.089 (0.0131)	29,266	56,450
Adenocarcinoma	0.067 (0.0094)	11,273	55,483
Squamous Cell	0.062 (0.013)	7,426	55,627
Small Cell	0.116 (0.023)	2,664	21,444
Ever smokers	0.110 (0.024)	23,223	16,964
Never smokers	0.027 (0.048)	2,355	7,504

Supplementary Table 3b. Genetic correlations between subtypes of lung cancer as defined by histology and smoking status. Analyses are based on HapMap 3 SNPs with MAF > 0.01.

	Adenocarcinoma	Squamous Cell	Small Cell	Ever Smokers	Never Smokers
Adenocarcinoma	1	0.584 (0.104) p=1.71e-08	0.407 (0.109) p=1.92e-04	0.892 (0.048) p=5.08e-75	0.920 (1.159) p=0.427
Squamous Cell		1	0.693 (0.118) p=3.97e-09	0.934 (0.052) p=1.55e-71	0.059 (0.341) p=0.862
Small Cell			1	0.693 (0.118) p=3.97e-09	-0.118 (0.474) p=0.803
Ever smokers				1	0.878 (0.963) p=0.360
Never smokers					1

Supplementary Table 4. Association for sentinel variants separated by Oncoarray and previous GWAS studies.

	Locus	rs number	Position	Gene	EAF	Previous GWAS OR(95%CI)	P-value	Allele	EAF	Oncoarray OR(95%CI)	P-value
Lung	1p31.1	rs71658797	77967507	*AK5*	0.10	1.12 (1.07-1.18)	9.64E-06	T/A	0.10	1.15 (1.08-1.21)	4.37E-06
Lung	6q27	rs6920364	167376466	*RNASET2*	0.45	1.08 (1.05-1.11)	2.42E-06	G/C	0.48	1.06 (1.03-1.10)	6.29E-04
Lung	8p21.2	rs11780471	27344719	*EPHX2,CHRNA2*	0.06	0.86 (0.80-0.92)	5.32E-06	G/A	0.06	0.88 (0.81-0.95)	8.50E-04
Lung	13q13.1	rs11571833	32972626	*BRCA2*	0.01	1.75 (1.50-2.05)	2.95E-12	A/T	0.01	1.43 (1.20-1.70)	5.24E-05
Lung	15q21.1	rs66759488	47577451	*SEMA6D*	0.35	1.08 (1.04-1.11)	1.45E-05	G/A	0.38	1.07 (1.03-1.11)	4.35E-04
Lung	15q21.1	rs77468143	49376624	*SECISBP2L,COPS2*	0.26	0.92 (0.89-0.95)	4.66E-06	T/G	0.25	0.92 (0.88-0.96)	8.54E-05
Lung	15q25.1	rs55781567	78857986	*CHRNA5*	0.36	1.34 (1.29-1.38)	2.29E-70	C/G	0.38	1.25 (1.21-1.30)	1.16E-34
Lung	19q13.2	rs56113850	41353107	*CYP2A6*	0.45	0.88 (0.84-0.91)	4.64E-11	C/T	0.42	0.89 (0.86-0.93)	3.62E-09
Adeno	3q28	rs13080835	189357199	*TP63*	0.50	0.89 (0.85-0.93)	1.09E-06	G/T	0.48	0.90 (0.86-0.94)	1.79E-06
Adeno	5p15.33	rs7705526	1285974	*TERT*	0.34	1.29 (1.22-1.36)	2.69E-18	C/A	0.36	1.23 (1.17-1.28)	7.71E-18
Adeno	8p12	rs4236709	32410110	*NRG1*	0.22	1.11 (1.05-1.17)	2.31E-04	A/G	0.22	1.15 (1.09-1.21)	6.04E-07
Adeno	9p21.3	rs885518	21830157	*MTAP*	0.10	1.11 (1.04-1.20)	3.77E-03	A/G	0.11	1.22 (1.14-1.31)	2.97E-08
Adeno	10q24.3	rs11591710	105687632	*OBFC1/SLK*	0.13	1.16 (1.08-1.24)	1.34E-05	A/C	0.14	1.17 (1.09-1.24)	2.65E-06
Adeno	11q23.3	rs1056562	118125625	*MPZL3/AMICA1*	0.47	1.12 (1.07-1.17)	3.80E-06	C/T	0.48	1.10 (1.06-1.15)	1.59E-05
Adeno	15q21.1	rs77468143	49376624	*SECISBP2L*	0.26	0.85 (0.80-0.89)	1.57E-09	T/G	0.25	0.86 (0.82-0.91)	1.89E-08
Adeno	20q13.33	rs41309931	62326579	*RTEL1*	0.12	1.19 (1.10-1.28)	9.80E-06	G/T	0.12	1.17 (1.09-1.26)	8.54E-06
SQC	6p21.33	rs116822326	31434111	*MHC*	0.15	1.24 (1.16-1.33)	7.54E-10	A/G	0.17	1.25 (1.16-1.34)	8.74E-10
SQC	12p13.33	rs7953330	998819	*RAD52*	0.32	0.85 (0.81-0.90)	8.54E-09	G/C	0.30	0.89 (0.84-0.94)	7.42E-05
SQC	22q12.1	rs17879961	29121087	*CHEK2*	0.005	0.38 (0.28-0.52)	5.53E-10	A/G	0.004	0.45 (0.30-0.67)	9.73E-05

Supplemental Table 5. Statistical strength of evidence for each locus according to alternative significance criteria.

Strata	Locus*	Rs number	Position	Gene	Allele[a]	EAF	OR (95%CI)	P-value	GWAS Correction	Bonferroni Correcton	Bonferroni Correction Adjusting for Strata	ABF	BFDP	
													$Pa=10^{-6}$	$Pa=10^{-8}$
Lung	1p31.1*	rs71658797	77967507	*FUBP*	T/A	0.1	1.14 (1.09-1.18)	3.25E-11	3.25E-05	3.84E-05	1.15E-04	5.42E-09	5.39E-03	3.52E-01
Lung	6q27*	rs6920364	167376466	*RNASET2*	G/C	0.46	1.07 (1.05-1.10)	1.29E-08	1.29E-02	1.53E-02	4.58E-02	2.87E-06	7.41E-01	9.97E-01
Lung	8p21.1*	rs11780471	27344719	*EPHX2,CHRNA2*	G/A	0.06	0.87 (0.83-0.91)	1.69E-08	1.69E-02	2.00E-02	5.99E-02	1.89E-06	6.54E-01	9.95E-01
Lung	13q13.1	rs11571833	32972626	*BRCA2*	A/T	0.01	1.60 (1.43-1.80)	6.12E-16	6.12E-10	7.24E-10	2.17E-09	9.20E-14	9.20E-08	9.20E-06
Lung	15q21.1*	rs77468143	49376624	*SEMA6D*	T/G	0.25	1.07 (1.04-1.09)	1.00E-08	1.00E-02	1.18E-02	3.55E-02	5.98E-06	8.57E-01	9.98E-01
Lung	15q25.1	rs55781567	78857986	*CHRNA5*	C/G	0.37	1.29 (1.26-1.32)	3.08E-103	3.08E-97	3.64E-97	1.09E-96	3.18E-100	3.18E-94	3.18E-92
Lung	19q13.2	rs56113850	41353107	*CYP2A6*	C/T	0.44	0.88 (0.86-0.91)	5.02E-19	5.02E-13	5.94E-13	1.78E-12	1.55E-16	1.55E-10	1.55E-08
Adeno	3q28	rs13080835	189357199	*TP63*	G/T	0.49	0.89 (0.87-0.92)	7.45E-12	7.45E-06	8.81E-06	2.64E-05	1.50E-09	1.50E-03	1.30E-01
Adeno	5p15.33	rs7705526	1285974	*TERT*	C/A	0.34	1.25 (1.21-1.29)	3.80E-35	3.80E-29	4.49E-29	1.35E-28	1.42E-32	1.42E-26	1.42E-24
Adeno	8p12*	rs4236709	32410110	*NRG1*	A/G	0.22	1.13 (1.09-1.18)	1.28E-10	1.28E-04	1.51E-04	4.54E-04	2.05E-08	2.01E-02	6.72E-01
Adeno	9p21.3*	rs885518	21830157	*MTAP*	A/G	0.1	1.17 (1.11-1.23)	9.96E-10	9.96E-04	1.18E-03	3.53E-03	1.20E-07	1.08E-01	9.23E-01
Adeno	10q24.3*	rs11591710	105687632	*OBFC1*	A/C	0.14	1.16 (1.11-1.22)	6.30E-11	6.30E-05	7.45E-05	2.23E-04	8.88E-09	8.80E-03	4.70E-01
Adeno	11q23.3*	rs1793151	118126920	*MPZL3/AMICA1*	T/C	0.48	1.11 (1.07-1.14)	3.25E-10	3.25E-04	3.84E-04	1.15E-03	5.09E-08	4.84E-02	8.36E-01
Adeno	15q21.1*	rs77468143	49376624	*SECISBP2L*	T/G	0.25	0.86 (0.83-0.87)	1.69E-16	1.69E-10	2.00E-10	5.99E-10	3.66E-14	3.66E-08	3.66E-06
Adeno	20q13.33*	rs41309931	62326579	*RTEL1*	G/T	0.12	1.17 (1.11-1.23)	1.31E-09	1.31E-03	1.55E-03	4.65E-03	1.53E-07	1.33E-01	9.39E-01
SQC	6p21.33	rs116822326	31434111	*MHC*	A/G	0.16	1.25 (1.19-1.31)	3.83E-19	3.83E-13	4.53E-13	1.36E-12	7.51E-17	7.51E-11	7.51E-09
SQC	12p13.33	rs7953330	998819	*RAD52*	G/C	0.32	0.86 (0.83-0.90)	7.26E-13	7.26E-07	8.58E-07	2.58E-06	1.26E-10	1.26E-04	1.25E-02
SQC	22q12.1	rs17879961	29121087	*CHEK2*	A/G	0	0.41 (0.32-0.52)	5.70E-13	5.70E-07	6.74E-07	2.02E-06	2.70E-10	2.70E-04	2.63E-02

* denote novel locus identified to GWAS significance by this study; [a], reference/effect. EAF, effective allele frequency; OR, odds ratio; 95%CI, 95% confidence interval. P-value, based on fixed-effect meta-analysis adjusted for age, sex and genetically derived ancestry; ^ variant excluded due to discordance between imputed and validation genotyping.

GWAS correction, 1,000,000 tests (5×10^{-8}); Bonferroni correction (based on estimated number of effective tests); Bonferroni Adjusting for Strata (number of effective tests x 3), adjusts for analysis of overall lung cancer and ever versus never smoking, which are three tests; ABF, approximate bayes factor; BFDT, bayesian false discovery probability. Pa Prior probability of an association

rsID	RefSeq	Ref_effect	Gene Expression microarray								RNAseq			
			Laval n=409		UBC n=287		Groningen n=342				NCI n=90		GTEx n=297	
			BETA	P-value	BETA	P-value	BETA	P-value	Meta.P		BETA	P-value	BETA	P-value
rs71658797	GIPC2	T_A	-0.11	0.03	-0.23	0.005	-0.04	0.51	0.004		0.4887	0.0649353	0.18096	0.108596
rs6920364	**RNASET2**	**G_C**	**0.05**	**0.001**	**0.11**	**8E-07**	**0.07**	**0.003**	**1.00E-09**		**0.33**	**0.03**	**0.38**	**2.00E-14**
rs6920364	FGFR1OP	G_C	-0.11	0.000002	-0.11	0.0007	-0.07	0.08	4.00E-08		-0.03	0.84	-0.03	0.5
rs11780471	TRIM35	G_A	0.07469	0.007257	0.01437	0.7214	-0.1116	0.05899	0.0128928		0.2568	0.3302795	0.00744	0.93104
rs66759488	AK021664	T_A	-0.071	0.02065	0.01708	0.6849	-0.07906	0.1677	0.0600298		0.03	0.88	NA	
rs55781567	IREB2	C_G	-0.25	1.00E-19	-0.3	5.00E-19	-0.36	3.00E-09	1.00E-42		-0.11	0.45	-0.04	0.24
rs55781567	**CHRNA5**	**C_G**	**-0.14**	**1.00E-13**	**-0.18**	**2.00E-09**	**-0.16**	**2.00E-09**	**9.00E-28**		**-0.2**	**0.18**	**-0.28**	**0.0002**
rs55781567	PSMA4	C_G	0.06	2.00E-06	0.06	0.0001	0.06	0.01	1.00E-09		0.09	0.55	0.06	0.09
rs56113850	NUMBL	C_T	0.06	0.1	0.06	0.08	0.07	0.05	0.02		-0.28	0.12	-0.03	0.48
rs13080835	**TP63**	**G_T**	**-0.3408**	**3.96E-13**	**0.3557**	**7.31E-05**	**-0.4391**	**1.9E-06**	**7.62E-20**		**0.48**	**0.001**	**0.3**	**2.00E-05**
rs7705526	SLC6A19	C_A	0.01494	0.533	-0.05013	0.02837	-0.02712	0.3604	0.1078803		NA		NA	
rs4236709	**NRG1**	**A_G**	**-0.0804**	**0.001233**	**-0.05628**	**0.1252**	**-0.1178**	**0.00492**	**8.678E-05**		**-0.4**	**0.02**	**-0.23**	**0.0001**
rs885518	IFNA1	A_G	0.02	0.12	0.05	0.002	0.006	0.71	0.007		NA		-0.04	0.74
rs11591710	GSTO2		-0.0041	0.9062	-0.1151	0.02414	0.11	0.0517	0.0348346		0.1389	0.4607848	-0.01488	0.779325
rs1056562	DDX6	C_T	0.09	0.00007	0.13	0.00002	0.09	0.009	4.00E-09		0.01	0.95	NA	
rs1056562	**AMICA1**	**C_T**	**-0.09**	**0.0003**	**-0.09**	**0.02**	**-0.14**	**0.002**	**3.00E-06**		**-0.27**	**0.08**	**-0.12**	**0.006**
rs1056562	**MPZL3**	**C_T**	**-0.06**	**0.02**	**-0.12**	**0.00003**	**-0.07**	**0.04**	**4.00E-06**		**-0.06**	**0.7**	**-0.19**	**1.00E-05**
rs1056562	KMT2A	C_T	-0.04	0.02	0.05	0.01	-0.06	0.02	0.0004		NA		-0.01	0.54
rs77468143	**SECISBP2L**	**T_G**	**0.34**	**8.00E-35**	**0.25**	**2E-14**	**0.21**	**4E-08**	**5.00E-52**		**0.87**	**1.00E-06**	**0.31**	**3.00E-16**
rs77468143	**FAM227B**	**T_G**	**0.19**	**0.00005**	**0.34**	**1E-08**	**0.05**	**0.41**	**1.00E-10**		**NA**		**-0.19**	**0.005**
rs41309931	HAR1A	G_T	0.05	0.06	-0.09	0.01	-0.06	0.13	0.004		0.02	0.92	0.14	0.23
rs7953330	**RAD52**	**G_C**	**-0.27**	**6.00E-22**	**-0.18**	**8E-10**	**-0.25**	**3E-13**	**9.00E-40**		**0.24**	**0.14**	**-0.11**	**0.002**
rs11571833	FRY	A_T	0.316	0.006651	0.1674	0.1649	0.27	0.09576	0.0054745		-0.62	0.45	NA	
rs17879961	CABP7	A_G	-0.10	0.7	-0.81	0.002	NA		0.009		-2.15	0.03	NA	

Supplementary Table 6. eQTL analysis of sentinel SNPs at each lung cancer susceptibility locus (variants in complex MHC region excluded). Section A lists the most significant eQT at each locus (ranked by P-value) within the Laval, UBC and Grongingen cohorts. In instances where there were multiple genes targets (P-value< 0.001) at a given locus, are listed. Section B eQTLs for these same variants and gene targets in two independent RNAseq cohorts, NCI and GTEx. Those bolded are those with evidence for association with consistent direction across the five cohorts. The eQTL methods and cohorts studied are described in the supplementary methods.

rs_number	name	Ref_effect	EAF	OR1	P-value1	OR2	P-value2	OR3	P-value3
rs11780471	8:27344719	G_A	0.06	0.87	1×10^{-8}	NA	NA	NA	NA
rs11778371	8:27319905	C_T	0.052	0.87	1×10^{-7}	0.979	0.36	0.98	0.41
rs146729428	8:27412605	A_G	0.116	0.91	2×10^{-7}	0.957	0.02	0.96	0.03
rs67749759	8:27397087	C_T	0.068	1.13	3×10^{-7}	1.121	2×10^{-6}	1.06	0.01
rs71519637	8:27334098	C_T	0.103	1.11	3×10^{-7}	1.102	2×10^{-6}	NA	NA
rs78875955	8:27389456	C_T	0.059	0.88	4×10^{-7}	1.073	0.77	1.07	0.79
rs35505579	8:27337045	A_T	0.103	1.11	4×10^{-7}	1.1	2×10^{-6}	1.19	0.65
rs2565064	8:27327841	G_C	0.289	1.07	5×10^{-7}	1.057	0.00003	1.03	0.03
rs2741348	8:27420782	G_T	0.415	0.94	5×10^{-7}	0.955	0.0001	0.97	0.003
rs2565065	8:27328021	G_A	0.289	1.07	5×10^{-7}	1.056	0.00003	1.03	0.03
rs10091679	8:27351738	G_A	0.103	1.11	5×10^{-7}	1.096	5×10^{-6}	1.02	0.29
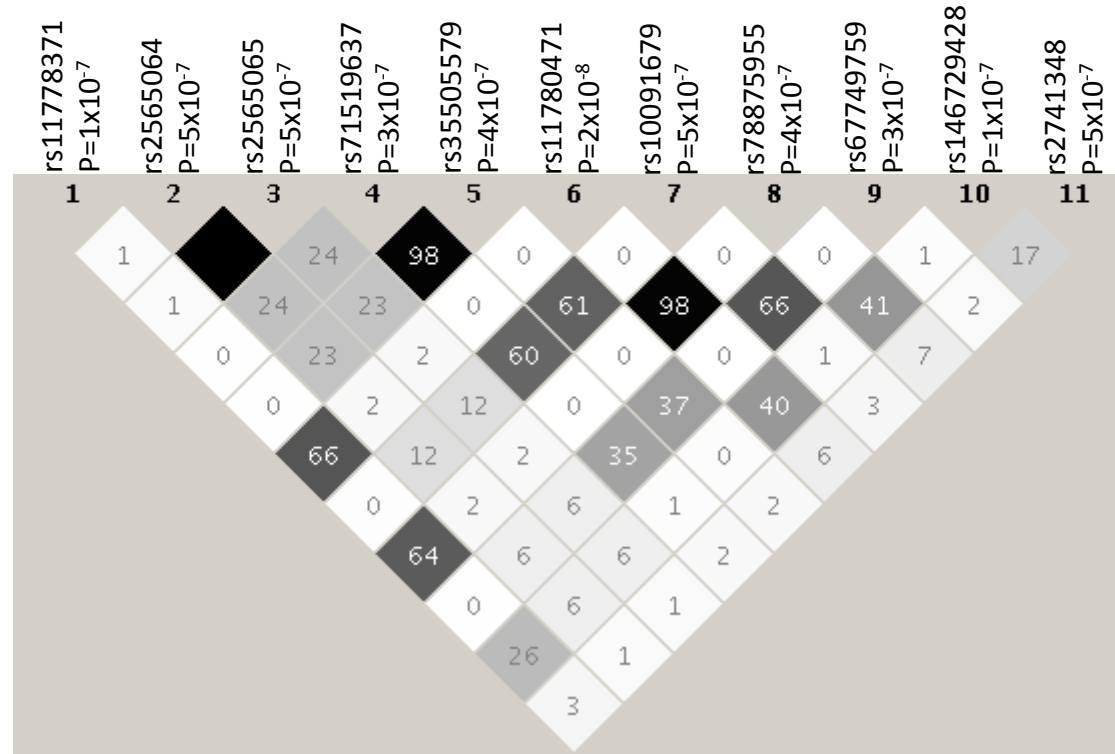
Supplementary Table 8. Upper panel Stepwise multivariate analysis of variants 8p21 associated with lung cancer (29,266 cases and 56,450 controls). Variants included in this table are restricted to the 11 variants associated with lung cancer with a P-value $= 5\times10^{-7}$ or lower. 1: base model. 2: multivariate analysis adjusting for rs11780471. 3: multivariate analysis adjusting for rs11780471 and rs71519637. **Lower panel.** Pairwise r^2 values for these 11 variants associated with lung cancer calculated from individuals of European decent from the 1000genomes project. P-value is the association with lung cancer.

			Section A										Section B			
			Gene Expression microarray						RNAseq				TRICL Oncoarray association			
			Laval n=409		UBC n=287		Groningen n=342		NCI n=90		GTEx n=297					
rsID	RefSeq	Ref_effect	BETA	P	BETA	P	BETA	P	BETA	P	BETA	P	OR	L95	U95	P
rs1056562	*AMICA1*	C_T	-0.09	0.0003	-0.09	0.02	-0.14	0.002	-0.27	0.08	-0.12	0.006	1.108	1.073	1.143	2.76E-10
rs1056562	*MPZL3*	C_T	-0.06	0.02	-0.12	0.00003	-0.07	0.04	-0.06	0.7	-0.19	1.00E-05				
rs3018367	*AMICA1*	T_C	-0.04	0.15	-0.06	0.15	-0.19	0.0004	0.18	0.22	-0.17	0.0002	1.031	0.995	1.069	0.09
rs3018367	*MPZL3*	C_T	-0.08	0.005	-0.14	0.00003	-0.15	0.0002	-0.02	0.91	-0.22	1.10E-06				
rs11216828	*AMICA1*	T_A	-0.09	0.005	-0.07	0.16	0.002	0.98	-0.11	0.58	0.04	0.47	1.121	1.078	1.165	7.91E-09
rs11216828	*MPZL3*	T_A	0.01	0.72	-0.03	0.41	0.05	0.16	-0.17	0.38	0.002	0.98				

Supplementary Table 9. Comparison between variants at 11q23 loci associated with eQTL (Section A) and lung cancer risk (Section B). rs3108367 is associated with expression levels of both *AMICA1* and *MPLZ3,* yet only very modestly associated with lung cancer. By contrast rs1056562 is associated with both, and rs11216828 only with lung adenocarcinoma. The absence of a direct relationship between the eQTL alleles and those associated with lung adenocarcinoma would argue against simple relationship between gene expression levels and lung adenocarcinoma risk.