Meta-analysis and imputation refines the association of 15q25 with smoking quantity


Author Contributions

IZL carried out most of the analysis for this study. JM and CF conceived and directed this study and wrote the manuscript. FT, DMW and VM were involved in study design and helped to coordinate the inclusion of many of the GSK cohorts. SGP, PM, LM, WB, CWK, XY, GW, PV, MP, NJW, HZH, RJFL, IB, KK, SG, PB, RM, AK, RM, JBV, JS, JLK, AF, PM, RD, KM, PB, AG, SL, MI, TB, SH, HEW, RR, ND, CL, OP, LZ, JH, SC, JK, JCC, MSB, JMD, ADP, KMK, LS, JML, RW, SE, JFW, SHW, HC, VV, MPR, ML, LQ, RW, WM, HHH, DJF, AF, MW, AS, MU, AT, XX, FB, PS, DS, DStC, DR, GRA, HIG, AT, HV, AP, UJ, IR, CH, AFW, IK, BJW, JRT, ABJ, ASH, NJS, CAA, TA, CGM, MF, JS, MC, PBM, MF, AD, JW, WT, SE, AB, & WTCCC prepared and shared datasets and, in some cases, cohort-specific results from their own primary analysis.

Competing Interests Statement

FT, CF, DMW, VM, PM, SGP, CWK are/were full time employees of the company GlaxoSmithKline (GSK). GSK also funded several aspects of the study as detailed in ACKNOWLEDGEMENTS. There were no competing interests arising from GSK’s involvement in this study.

URLs

ProbABEL software: http://mga.bionet.nsc.ru/yurii/ABEL/
SNPMETA software: http://www.stats.ox.ac.uk/~marchini/software/gwas/snpmeta.html
1000 Genomes Project: http://www.1000genomes.org/
UCSC Genome Browser: http://genome.ucsc.edu/
1Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK.
2Genetics Division, GlaxoSmithKline, Verona, Italy 3Genetics Division, GlaxoSmithKline, Upper Merion, PA, USA. 4Division of Neurosciences and Mental Health, Imperial College London, UK. 5Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, PA, USA. 6Genetics Division, GlaxoSmithKline, Research Triangle Park, NC, USA. 7University Hospital Center, University of Lausanne, Lausanne, Switzerland. 8Department of Internal Medicine, University of Lausanne, Lausanne, Switzerland. 9Department of Psychiatry, University of Lausanne, Lausanne, Switzerland. 10MRC Epidemiology Unit, Institute of Metabolic Science, Cambridge, UK. 11Wellcome Trust Sanger Institute, Hinxton, UK. 12Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. 13Center for Human Nutrition, University of Texas Southwestern Medical Center, Dallas, Texas, USA. 14The Heart Research Institute, Sydney, New South Wales, Australia. 15Gladstone Institute of Cardiovascular Disease, University of California, San Francisco, California, USA. 16American Hospital, Istanbul, Turkey. 17Department of Internal Medicine, University of Oulu, Oulu, Finland. 18Biocenter Oulu, University of Oulu, Oulu, Finland. 19Division of Cardiology, University of Ottawa Heart Institute, Ottawa, Ontario, Canada. 20Centre for Addiction and Mental Health, University of Toronto, ON, Canada. 21Medical Research Council Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King’s College London, UK. 22Center for Neuroscience, Division of Medical Sciences, University of Dundee, UK. 23Institute of Medicine, University of Bergen, Bergen, Norway. 24Max Planck Institute of Psychiatry, Munich, Germany. 25Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. 26Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany. 27Klinikum Grosshadern, Munich, Germany. 28Psychiatrie Klinik und Poliklinik, University of Mainz, Germany. 29Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria. 30School of Public Health, School of Medicine, University of Zagreb, Croatia. 31Centre for Population Health Sciences, University of Edinburgh, UK. 32Institute of Genetics and Molecular Medicine, MRC Human Genetics Unit, Edinburgh, UK. 33National Heart and Lung Institute, Imperial College London, UK. 34Division of Epidemiology, Imperial College London, UK. 35Cardiovascular Research Institute, MedStar Research Institute, Washington Hospital Center, Washington, DC, USA. 36The Cardiovascular Institute, University of Pennsylvania, Philadelphia, PA, USA. 37The Institute for Translational Medicine and Therapeutics, School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. 38Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, PA, USA. 39The Center for Applied Genomics, Children’s Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. 40Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. 41Istituto di Neurogenetica e Neurofarmacologia, CNR, Monserrato, Cagliari, Italy. 42National Institute on Aging, Baltimore, Maryland 21224, USA. 43Department of Epidemiology, University of Texas M. D. Anderson Cancer Center, Houston, Texas, USA. 44Department of Mental Health, University of Aberdeen, Aberdeen, United Kingdom. 45Division of Molecular and Clinical Neurobiology, Department of Psychiatry, Ludwig-Maximilians-University, Munich, Germany. 46Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. 47Department of Psychiatry and Psychotherapy, University of Greifswald, Germany. 48Interfaculty Institute for Genetics and Functional Genomics, University of Greifswald, Germany. 49Institute for Community Medicine, University of Greifswald, Germany. 50Institute of Clinical Chemistry and Laboratory Medicine, University of Greifswald, Germany. 51Department of Social Medicine and Epidemiology, University of Greifswald, Germany. 52Croatian Centre for Global Health, University of Split, Croatia. 53Department of Health Sciences, University of Leicester, Leicester, UK. 54Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds, UK. 55Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK. 56Peninsula College of
Abstract

Smoking is a leading global cause of disease and mortality. We performed a genomewide meta-analytic association study of smoking-related behavioral traits in a total sample of 41,150 individuals drawn from 20 disease, population, and control cohorts. Our analysis confirmed an effect on smoking quantity (SQ) at a locus on 15q25 (P=9.45e-19) that includes three genes encoding neuronal nicotinic acetylcholine receptor subunits (CHRNA5, CHRNA3, CHRNB4). We used data from the 1000 Genomes project to investigate the region using imputation, which allowed analysis of virtually all common variants in the region and offered a five-fold increase in coverage over the HapMap. This increased the spectrum of potentially causal single nucleotide polymorphisms (SNPs), which included a novel SNP that showed the highest significance, rs55853698, located within the promoter region of CHRNA5. Conditional analysis also identified a secondary locus (rs6495308) in CHRNA3.

Smoking behavior and Nicotine Dependence (ND) are multifactorial traits with substantial genetic influences. There is an urgent need to better understand the molecular neurobiology of ND, in order to design targeted, more effective therapies. Recently, genome-wide association scans (GWAS) have established one locus in ND and Smoking Quantity (SQ), which implicates a cluster of three genes encoding neuronal nicotinic acetylcholine receptor subunits, CHRNA5, CHRNA3, and CHRNB4, on chromosome 15q25. The locus is also associated with lung cancer, peripheral arterial disease, and chronic obstructive pulmonary disease and lung function.

We initially performed a GWAS meta-analytic study of smoking-related traits in a total sample of 41,150 individuals of white European descent, sourced from multiple disease, population and control cohorts (Table 1, Supplementary Table 1, Online Methods). As the cohorts were genotyped on a variety of different genome-wide SNP arrays (Table 1, Supplementary Table 1), we first imputed genotypes for all datasets, for all SNPs in the HapMap version release 22. The main focus of our analysis was on SQ within current or past smokers, as a semi-quantitative trait based on the self-reported variable of Cigarettes-per-Day (CPD). We performed association analysis separately within each cohort under an additive model, using covariate effects for age and sex, disease case/control status where applicable, and other cohort-specific covariates (Supplementary Table 1). The meta-analysis was then carried out by combining study-specific β-estimates using a fixed effects model. In total, 15,574 subjects reported CPD values >0 and were used for meta-analysis of SQ (Table 1, Supplementary Table 1). We followed up our most promising association findings by
comparing them with results from two concurrent GWAS meta-analyses of smoking; the ENGAGE study of 46,481 subjects\textsuperscript{15}, and the TAG study of 74,035 subjects\textsuperscript{16}. We also made our meta-analysis results available to the authors of those studies to check their top findings for replication.

Our meta-analysis of SQ identified the \textit{CHRNA5/CHRNA3} locus on 15q25 as the single outstandingly significant locus in the genome (Figure 1, Table 2, Supplementary Table 2), with a minimum $P=9.45\times 10^{-19}$ for rs1051730, which has been a SNP commonly reported\textsuperscript{4–8}, and very low $P$ values for many other SNPs in the region (Supplementary Figure 1, Supplementary Table 2). All cohorts in the analysis contributed at least somewhat to the 15q25 association (Supplementary Figure 1). Each copy of the ‘high-smoking’ A allele (34\% frequency) had a quantitative effect size on SQ of $0.079 \pm 0.088$ which is inline with previous estimates\textsuperscript{7}. Joint analysis of our total dataset together with TAG and ENGAGE, for rs1051730, yielded $P=1.71\times 10^{-66}$ (Table 2).

Multiple variants at the 15q25 locus have been suggested to underlie its effect, including a non-synonymous SNP in \textit{CHRNA5}, together with variants that affect mRNA expression levels\textsuperscript{17–19}. We decided to use our very large sample, in combination with data from the 1000 Genomes Project (see URL below), to perform fine mapping and modeling of the 15q25 locus in relation to SQ. We reasoned that, with the near complete information on common variants derived from 1000 Genomes, it might be possible to pinpoint a variant, or combination of variants, that can explain all the signal of association at 15q25. We used data from 108 estimated CEU haplotypes from the April 2009 release of the 1000 Genomes Pilot 1 data. This contained 2189 SNPs in our region of interest (See Online Methods), approximately a five-fold increase in coverage compared to 437 SNPs in release 22 of the HapMap. By imputing genotypes for all SNPs across this locus from 1000 Genomes, and repeating the meta-analysis, we found that the most significant association was with a novel and previously untested SNP, not in the HapMap, located within the 5′ untranslated region (UTR) of \textit{CHRNA5}, which makes it a candidate for affecting mRNA transcription (rs55853698, $P = 1.31\times 10^{-16}$; Figure 2). The p-value for the commonly reported SNP rs1051730 in this analysis was similar but a little higher, $P=1.47\times 10^{-15}$. (P values for our 1000 Genomes analysis are generally higher than our HapMap-based analysis because not all cohorts were included in the 1000 Genomes imputation - see Online Methods.) SNP rs55853698 is a G/T substitution where the G allele has a frequency ranging from 0.313 to 0.378 across the various cohorts.

To investigate whether the association at 15q25 can be explained completely by rs55853698, we carried out tests of association for all SNPs spanning the \textit{CHRNA5/CHRNA3} locus conditional upon this SNP (Figure 2). Residual association was still detected at many SNPs in the region, with the most significant signal occurring at rs6495308 ($P=3.96\times 10^{-5}$), located within an intron of \textit{CHRNA3} (Figure 2). In unconditioned analysis rs6495308 has a marginal association in the meta-analysis of $P=3.30\times 10^{-10}$. Further conditioning on rs6495308, after already conditioning on rs55853698, leaves no obvious signal of association in the region (Supplementary Figure 2), suggesting that these two SNPs together could be sufficient to explain this genetic effect.

Wang \textit{et al.}\textsuperscript{18} suggested that a non-synonymous SNP rs16969968, in \textit{CHRNA5}, is functional for ND risk (and lung cancer risk), but also that variants that cause high expression of \textit{CHRNA5} mRNA, tagged by SNP rs588765, increase the risk for ND independently. The marginal $p$-values of rs16969968 and rs588765 in our meta-analysis were $P=1.64\times 10^{-18}$ and $P=1.74\times 10^{-3}$. Conditional analysis on rs16969968 within our cohorts still left residual association within the region (Supplementary Figure 2), with the most significant signal again occurring at rs6495308 ($P=1.54\times 10^{-5}$). Conditioning on both...
rs16969968 and rs588765, i.e. the proposed combination of Wang et al.\textsuperscript{18}, leaves no obvious signal of association (Supplementary Figure 2). To further investigate which pair of SNPs best explains the signal of association we used the Bayesian Information Criteria (BIC) measure of model fit\textsuperscript{20}. For the model of Wang et al.\textsuperscript{18}, i.e. conditioning on both rs16969968 and rs588765, we obtained $\text{BIC} = 22719.87$, posterior probability 0.15. For the model conditioning on the novel promoter SNP rs55853698, and rs6495308, we obtained $\text{BIC} = 22716.49$, posterior probability 0.85, which indicates a better model fit.

Examination of the LD structure between the SNPs that we have considered shows that rs1051730, rs16969968, and rs55853698 are all close tagging proxies of each other (all pairwise $R^2 > 0.96$). These variants tag, or cause, the principal risk for high SQ attributable to the 15q25 locus, but the high LD makes it difficult to assign causality. The ‘residual association’ SNPs rs588765 and rs6495308 are in low LD with each other ($R^2 = 0.21$), and are both only in modest LD with the principal SNPs (maximum $R^2 = 0.47$). It is not therefore clear that this locus can be completely understood in the way proposed by Wang et al.\textsuperscript{18}. While the non-synonymous SNP in CHRNA5, rs16969968, may be important, we have identified a novel and potentially functional SNP in the 5' UTR of this gene that is a close proxy to the non-synonymous SNP in terms of LD, but shows a slightly more significant association in our meta-analysis. Then, while rs588765 can explain much of the secondary or residual association at this locus, we find that a largely independent variant within CHRNA3, rs6495308, is the best tagger of the residually associated variation, while also contributing to a better fitting 2-SNP model, and having a much stronger marginal significance in unconditioned analysis ($P=3.30\times 10^{-10}$ for rs6495308 compared to $P=1.74\times 10^{-03}$ for rs588765).

Our analysis has, for the first time, surveyed virtually all of the common variants in the 15q25 region, and provides one of the first examples of how data from the 1000 Genomes Project can contribute new information to mapping and characterizing loci for complex traits. We recommend that further analysis of this locus should not be limited in focus to CHRNA5, nor particularly to the common, non-synonymous SNP rs16969968. It is notoriously difficult to distinguish functional variation in the context of high LD across a region\textsuperscript{21}. There are numerous ways in which variants can be functional, including expression regulatory changes that affect close or distant genes, epigenetic changes, splicing effects, alterations to microRNA binding sites, or non-coding RNAs\textsuperscript{21}. It is also conceivable that association with common variants can arise through the effects of multiple rarer variants that happen to be relatively restricted to specific haplotype backgrounds.

The second strongest association within the genome in our meta-analysis, for SQ, was at a locus on 8p21 that received modest support from the TAG and ENGAGE studies (Supplementary Table 2, Supplementary Figure 3; $P=5.26\times 10^{-07}$ for rs11782673). This locus would not survive correcting for genome-wide multiple testing, although it is noteworthy that the locus spans another neuronal nicotinic acetylcholine receptor subunit gene, CHRNA2.

In addition to our analysis of SQ, we also tested genome-wide for allelic differences between those who reported currently smoking, or smoking in the past, versus those who said they had never been smokers (the EVER/NEVER phenotype; sample sizes in Table 1, Supplementary Table 1). This was in order to identify genetic effects on the establishment of a smoking habit. No locus achieved genome-wide significance, and none of the top 15 loci showed evidence for replication (Supplementary Table 2, Supplementary Figure 4). Likewise, no consistent results emerged when we tested for allelic differences between those who reported currently smoking versus those who had smoked in the past but had stopped at
the time of interview (Supplementary Table 2, Supplementary Figure 4). When age-
adjusted, this is a rough measure of smoking cessation.

Our study identified association at some loci which, while not reaching genomewide
significance in our own meta-analysis, supported findings from the concurrent TAG and
ENGAGE studies\textsuperscript{15,16}. These include novel loci on chromosomes 8 and 19 for SQ, 11 for
EVER/NEVER, and 9 for Current/Non-Current\textsuperscript{15,16}. These findings have provided further
novel insights into the biology of smoking behavior.

**ONLINE METHODS**

**Study samples**

Study collections and their basic characteristics are listed in Table 1 and Supplementary
Table 1. Subjects used in our analysis were adults of white European descent. Summary
descriptions of the collections are given below, together with primary citations that describe
the collections fully. Data were used in accordance with the ethical permissions and
consents relating to each collection.

**GEMS\textsuperscript{22}:** The Genetic Epidemiology of Metabolic Syndrome (GEMS) study consists of
dyslipidaemic cases (age 20–65 years) matched with normolipidaemic controls by sex and
recruitment site, drawn from non-Mediterranean subjects of the Genetic Epidemiology of
Metabolic Syndrome study (Finland, Switzerland, Canada, Australia, USA).

**CoLaus\textsuperscript{23}:** The *Cohorte Lausannoise* (CoLaus) is a single-center, cross-sectional population-
based study, including individuals aged 35 to 75 years randomly selected from the list of
residents of the city of Lausanne (Switzerland).

**GSK COPD\textsuperscript{11}:** This collection includes cases with chronic obstructive pulmonary disease
diagnosed according to Global Initiative of Chronic Obstructive Lung Disease (GOLD)
criteria, and unaffected controls recruited from Bergen, Norway.

**GSK UPD\textsuperscript{24}:** This collection includes cases with recurrent major depression according to
DSM-IV criteria and age- and gender-matched non-affected controls, recruited at the Max-
Planck Institute of Psychiatry in Munich, Germany; patients were also recruited at two
satellite recruiting hospitals (BKH Augsburg and Klinikum Ingolstadt) in the Munich area.

**GSK Bipolar\textsuperscript{25}:** The Bipolar collection includes DSM-IV Bipolar cases and controls from
subjects recruited at 3 study sites: the Institute of Psychiatry (IOP) in London, U.K.; the
Centre for Addiction and Mental Health in Toronto, Canada; and the University of Dundee,
U.K.

**GSK Lolipop\textsuperscript{26}:** The *London Life Sciences Prospective Population* (LOLIPOP) is a
population based study including Indian Asian and European white men and women
recruited from the lists of 58 General Practitioners in West London.

**GSK Medstar\textsuperscript{27}:** The MedStar cohort includes cases with acute coronary syndrome or
chronic coronary artery disease from Washington DC, and unaffected controls.

**PennCath\textsuperscript{27}:** The Penn-CATH cohort is a University of Pennsylvania Medical Center based
angiographic study, from which cases with coronary artery disease (CAD) and controls with
no evidence of CAD at the coronary angiography were derived.
EPIC\textsuperscript{28}: The EPIC-Obesity cohort is a case-control cohort for obesity drawn from the EPIC-Norfolk cohort, which includes white European men and women aged 39–79 years recruited in Norfolk, UK.

KORA\textsuperscript{29}: The Co-operative Health Research in the Region of Augsburg (KORA) study is an epidemiological survey of the general population living in the city of Augsburg, Southern Germany, and two adjacent counties.

WTCCC HT\textsuperscript{30}: The WTCCC-HT collection comprises severely hypertensive probands ascertained from families with multiple affected members in the UK as part of the BRIGHT study.

WTCCC CAD, WTCCC CD, WTCCC RA\textsuperscript{30}: include patients with Coronary Artery Disease, Crohn’s disease and Rheumatoid Arthritis from the Wellcome Trust Case Control Consortium Study.

POPGEN study\textsuperscript{31}: The Population Genetic Cohort (POPGEN) is a cross sectional epidemiological surveys of regional German populations from Schleswig-Holstein, northern Germany.

SHIP Study\textsuperscript{32}: The Study of Health in Pomerania (SHIP) is a longitudinal, population-based survey from West Pomerania, Germany. Data from the baseline cohort were used for this study.

VIS Study\textsuperscript{33}: This study includes unselected Croatians, aged 18–93 years, recruited from the villages of Vis and Komiza on the Dalmatian island of Vis.

ORCADES Study\textsuperscript{34}: The Orkney Complex Disease Study is a family-based, cross-sectional study that seeks to identify genetic factors influencing cardiovascular and other disease risk in the population isolate of the Orkney Isles in northern Scotland.

KORCULA Study\textsuperscript{35}: The KORCULA study includes healthy volunteers aged 18 and over from the villages of Lumbarda, Žrnovo, and Račišće on the Island of Korcula, Croatia.

SardiNIA Study\textsuperscript{36}: The SardiNIA is a population-based longitudinal cohort study that includes male and female related individuals, aged 14 years and above, from a cluster of four towns in the Ogliastra province of Sardinia, Italy.

**Genotyping, quality control and imputation**

Supplementary Table 1 lists the various genotype platforms used for each cohort, genotype calling algorithms, SNP and sample quality control, and details of the imputation and association analysis software used. The quality control measures from previous analyses of each cohort were adopted for this study and are detailed in the table. We used NCBI Build 36 co-ordinates for SNP base-pair positions so that all the cohorts could be combined seamlessly.

We imputed all SNPs reported in the CEU sample in HapMap Phase II using various imputation algorithms\textsuperscript{12,37} (see the URL section for a link to the software ProbABEL). Imputations were performed after excluding samples and SNPs that did not meet the study-specific quality control criteria. Genotypes were imputed for SNPs not present in the genome-wide arrays or for those where genotyping had failed to meet the QC criteria.

Only imputed SNPs with good imputation quality were included in the meta-analysis. This was defined as proper_info $\geq 0.5$ (for studies analysed with IMPUTE/SNPTEST\textsuperscript{12}) or rsq-
hat ≥ 0.5 (for studies analysed using MACH)\(^3\) and Imp_info ≥ 0.5 (for studies analysed using ProbABEL).

**Derivation of smoking phenotypes**

We used the categorical SQ levels defined by Thorgeirsson et al.\(^7\). The SQ levels were 0 (1–10 cigarettes per day), 1, (11–20), 2 (21–30) and 3 (31 or more). Each increment represents an increase in SQ of 10 cigarettes per day. Most of the cohorts in our study have maximal CPD recorded on each sample but a few have collected average CPD (Supplementary Table 1). We examined the distributions of CPD across cohorts and found no large differences between those cohorts with average or maximal CPD. The mean and standard deviation of the CPD measurements in each cohort are given in Supplementary Table 1. The Ever/Non and Current/Non-current phenotypes used were those collected by the individual cohorts. Not all cohorts had all three phenotypes collected. Precise details of the phenotypes collected in each cohort are given in Supplementary Table 1. An assessment would typically be questionnaire-based, following a structure such as:

Tick the option that best describes you:

- I smoke now
- I don’t smoke now. I have stopped for … years.
- I have never smoked

About how many cigarettes do you or did you smoke per day?

Put the number of years you have smoked.

**Statistical Analysis and Meta-analysis**

Each cohort was analyzed separately for each of the 3 phenotypes considered. The majority of the analysis was carried out on the raw genotype data in Oxford but some cohorts (SardiNIA, VIS, KORCULA, ORCADES, SHIP) carried out their own analysis and submitted results for the meta-analysis. For the binary traits (Ever/Non, Current/Non-Current) tests for additive genetic effects on the logodds scale were carried out using logistic regression. For the categorical SQ phenotype, tests for additive genetic effects were carried out on a linear scale using linear regression. The programs SNPTEST, probABEL and MERLIN were used on the various cohorts to fit these models taking account of the genotype uncertainty at imputed SNPs. All tests conditioned on Sex and Age and for some cohorts other covariates of self-reported ancestry, country of origin or PCA-derived covariates were included (a complete list is given in Supplementary Table 1). A Genomic Control (GC) lambda estimate was calculated for each phenotype and each cohort (Supplementary Table 3).

The meta-analysis was carried out by combining study-specific \( \beta \)-estimates using a fixed effects model\(^1\)\(^4\) using the inverse of the variance of the study-specific \( \beta \)-estimates to weight the contribution of each study. The variance of each cohort’s \( \beta \)-estimate was multiplied by the GC lambda estimate to correct for observed inflation\(^3\). Specifically,

\[
\beta_{\text{META}} = \frac{\sum_i \beta_i / (\lambda_i \sigma_i^2)}{\sum_i 1 / (\lambda_i \sigma_i^2)}, \quad \sigma_{\text{META}} = \sqrt{\sum_i 1 / (\lambda_i \sigma_i^2)}, \quad Z_{\text{META}} = \frac{\beta_{\text{META}}}{\sigma_{\text{META}}},
\]

\( L \) Liu et al. Page 8
\( N \) Nat Genet. Author manuscript; available in PMC 2013 April 01.
where $\beta_i$, $\beta_i^*$, $\sigma^2_i$ and $\lambda_i$ are the $\beta$-estimate, $\beta$-estimate variance and GC lambda estimate for the $i$th cohort. This method is appropriate when the same phenotype and measurement scale are used in each cohort and has the advantage that measures of effect size ($e^\beta$ is an estimate of the Odds Ratio of the risk allele) and its standard error can be calculated. We also repeated the analysis of SQ by combining $Z$-scores from each cohort weighted by their sample size and obtained almost identical results. All meta-analysis was carried out using the SNPMETA program (see URL list). After performing each meta-analysis the overall lambda estimate for each phenotype was: SQ 1.0145, Ever/Never 1.002, Current/Non-Current 0.998. For each SNP we also calculated a p-value for the heterogeneity across the studies.

**SNP selection for replication**

In collaboration with two other groups carrying out similar meta-analysis of smoking related traits (ENGAGE\textsuperscript{15} and TAG\textsuperscript{16}) we agreed to an *in-silico* replication strategy in which for each phenotype (SQ, EVER/NEVER, CURRENT/NON-CURRENT) each group would select 15 regions of the genome showing evidence for association, and summary data (p-values, $\beta$-estimate, $\beta$-estimate variances, sample sizes, GC-lambda estimates and sample sizes) would be shared across groups to facilitate replication. We selected the top 15 regions for each phenotype based on the p-values we obtained in our own meta-analysis. We excluded regions in which only a small number of cohorts contributed to the study because the information measure at the SNPs in the excluded cohorts were below our thresholds, or where the heterogeneity between the studies was high. Each selected region consisted of several SNPs showing evidence of association in our meta-analysis with p-values below $1\times10^{-5}$. For each of the three phenotypes the results from all the cohorts in all three concurrent studies were combined together using the same GC-corrected inverse-variance meta-analysis method described above. A full list of the selected regions and the summary information from all 3 phenotypes is given in Supplementary Table 2.

**1000 Genomes imputation analysis of the 15q22 associated region for SQ**

We used 108 estimated CEU haplotypes from the April 2009 release of the 1000 Genomes Pilot 1 data to carry out our fine-mapping experiments at the 15q25 locus (see the URL list for a link to the data source). We used these haplotypes to carry out imputation in the interval 76.4–77.0Mb on chr15 in 12 of the cohorts (GSK-Bipolar, GSK-Unipolar, GSK-COPD, KORA, POPGEN, Lausanne, GSLOlipop, GSK-GEMS, Medstar, SHIP, WTCCC-CAD and WTCCC-HT) using the program IMPUTE\textsuperscript{12}. This release contains 2189 SNPs in this interval compared to 437 SNPs in release 22 of the HapMap data. Meta-analysis of the imputed data was then carried out in the same way as described above. An important technical detail when carrying out imputation using the 1000 Genomes haplotype data is how to align it with the genotype data from genome-wide studies. The program IMPUTE aligns SNPs between the haplotype and genotype data based on base-pair position (and not using SNP identifiers such as rs IDs) so as long as the same co-ordinate system is used for both the haplotype and genotype data the alignment is automatic.

**Conditional analysis and modeling**

The analysis conditional upon SNPs was carried out using all of the centrally analyzed cohorts (Bipolar, Unipolar, COPD, KORA, POPGEN, Lausanne, LOLIPOP, GEMS, MEDSTAR, SHIP, WTCCC-CAD and WTCCC-HT). At the SNP being conditioned upon we used expected genotype counts as this allowed us to combine data from cohorts which had imputed the SNP and cohorts which had genotyped the SNP. These expected counts where included into the baseline null model as an additional covariate along with the other covariates such as Age, Sex and covariates coding for population structure. The same

\textit{Nat Genet. Author manuscript; available in PMC 2013 April 01.}
method was used when conditioning upon two SNPs. The model selection analysis of the two pairs of SNPs in the 15q25 region was carried out using the expected genotype counts. Analysis was carried out using the R statistical package.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

GlaxoSmithKline plc (GSK), a pharmaceuticals company that is interested to develop novel cessation therapies for smoking, funded a postdoctoral fellowship for JL at Oxford University. GSK also funded the collection, characterization, and in some cases the genotyping and genotype data preparation for several of the cohorts used in this study. Allen Roses and Paul Matthews played crucial roles in establishing and funding the Medical Genetics activities at GSK. Acknowledgements that are specific to individual cohorts are given in the Supplementary Note.

References

15. ENGAGE Smoking Consortium. Sequence variants at CHRNA3-CHRNB4 and CYP2A6 influence smoking behavior. (Submitted).

Nat Genet. Author manuscript; available in PMC 2013 April 01.


Figure 1.
Manhattan plot showing the significance of association of all SNPs in genome-wide SQ meta-analysis. SNPs are plotted on the x-axis according to their positions on each chromosome, against association with SQ on the y-axis (−log10 P-value). SNPs with p-values less than 1.0E-05 are highlighted in green.
Figure 2.
Chromosome 15q25 signal plots. **Top:** Signal plot based on 1000 Genomes imputation and meta-analysis of SQ association. SNPs are plotted by their positions on the chromosome, against association with SQ (−log10 p-value) on the left Y-axis. The five SNPs with the lowest p-values from the HapMap imputation are highlighted in red. The five SNPs with the lowest p-values from the 1000 Genomes imputation are highlighted in green (unless already coloured red). The rs identities of highlighted SNPs are given in the box. Recombination rates across the region are shown by the red line plotted against the right y-axis. **Middle:** Chromosome 15q25 signal plot based on 1000 Genomes imputation and meta-analysis of SQ association, conditional on the SNP rs55853698. The five SNPs with the lowest p-values from the conditional analysis are highlighted in green. The five SNPs with the lowest p-values from the unconditioned HapMap-imputation analysis are highlighted in red. **Bottom:** Genes and the positions of exons (using data from the UCSC genome browser; URL is given below).
Table 1

Summary information for the cohorts used in meta-analysis. Further details are given in Online Methods and Supplementary Table 1.

<table>
<thead>
<tr>
<th>Label</th>
<th>Description</th>
<th>Genotyping</th>
<th>Sample Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>All</td>
<td>CPD&gt;0</td>
</tr>
<tr>
<td>WTCCC-RA</td>
<td>Rheumatoid Arthritis cases</td>
<td>Affymetrix 500K</td>
<td>1860</td>
</tr>
<tr>
<td>EPIC</td>
<td>Obesity case-control</td>
<td>Affymetrix 500K</td>
<td>3516</td>
</tr>
<tr>
<td>WTCCC-HT</td>
<td>Hypertension cases</td>
<td>Affymetrix 500K</td>
<td>1952</td>
</tr>
<tr>
<td>GEMS</td>
<td>Dyslipidemia case-control</td>
<td>Affymetrix 500K</td>
<td>1847</td>
</tr>
<tr>
<td>GSK-COPD</td>
<td>COPD case-control</td>
<td>Illumina 550</td>
<td>1633</td>
</tr>
<tr>
<td>GSK-BIPOLAR</td>
<td>Bipolar depression case-control</td>
<td>Illumina 550</td>
<td>1805</td>
</tr>
<tr>
<td>GSK-UPD</td>
<td>Unipolar depression case-control</td>
<td>Illumina 550</td>
<td>1792</td>
</tr>
<tr>
<td>WTCCC-IBD</td>
<td>Crohn's disease cases</td>
<td>Affymetrix 500K</td>
<td>1748</td>
</tr>
<tr>
<td>KORA</td>
<td>Population-based</td>
<td>Affymetrix 500K</td>
<td>1644</td>
</tr>
<tr>
<td>KORCULA</td>
<td>Population-based</td>
<td>Illumina 300</td>
<td>827</td>
</tr>
<tr>
<td>LOLIPOL</td>
<td>Population-based</td>
<td>Affymetrix 500K</td>
<td>1288</td>
</tr>
<tr>
<td>MEDSTAR</td>
<td>Coronary Artery Disease case-control</td>
<td>Affymetrix 6.0</td>
<td>1322</td>
</tr>
<tr>
<td>ORCADES</td>
<td>Population-based</td>
<td>Illumina 300</td>
<td>692</td>
</tr>
<tr>
<td>PENNCATH</td>
<td>Coronary Artery Disease case-control</td>
<td>Affymetrix 6.0</td>
<td>1401</td>
</tr>
<tr>
<td>POPGEN</td>
<td>Population-based</td>
<td>Affymetrix 6.0</td>
<td>1107</td>
</tr>
<tr>
<td>CoLaus</td>
<td>Population-based</td>
<td>Affymetrix 500K</td>
<td>5636</td>
</tr>
<tr>
<td>SardiNIA</td>
<td>Population-based</td>
<td>Affymetrix 500+10K</td>
<td>4,305</td>
</tr>
<tr>
<td>SHIP</td>
<td>Population-based</td>
<td>Affymetrix 6.0</td>
<td>4080</td>
</tr>
<tr>
<td>VIS</td>
<td>Population-based</td>
<td>Illumina 300</td>
<td>769</td>
</tr>
<tr>
<td>WTCCC-CAD</td>
<td>Coronary Artery Disease cases</td>
<td>Affymetrix 500K</td>
<td>1926</td>
</tr>
<tr>
<td>TOTALS</td>
<td></td>
<td></td>
<td>41150</td>
</tr>
</tbody>
</table>
Table 2

Summary information for selected SNPs at 15q25 from meta-analysis of association with the Smoking Quantity (SQ) phenotype. Our study is referred to as OX-GSK. Information for all SNPs spanning the 15q25 locus in our genomewide analysis is given in Supplementary Table 2.

<table>
<thead>
<tr>
<th>SNP (rs ID)</th>
<th>Chr</th>
<th>Position</th>
<th>Coded Allele</th>
<th>Coded Allele MAF</th>
<th>OX-GSK</th>
<th>TAG</th>
<th>ENGAGE</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs588765</td>
<td>15</td>
<td>76652480</td>
<td>T</td>
<td>0.43</td>
<td>1.74×10⁻³</td>
<td>0.50</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>rs16969968</td>
<td>15</td>
<td>76669980</td>
<td>G</td>
<td>0.65</td>
<td>1.64×10⁻¹⁸</td>
<td>0.86</td>
<td>1.85×10⁻²⁷</td>
<td>1.53×10⁻²³</td>
</tr>
<tr>
<td>rs1051730</td>
<td>15</td>
<td>76681394</td>
<td>G</td>
<td>0.66</td>
<td>9.45×10⁻¹⁹</td>
<td>0.68</td>
<td>3.62×10⁻²⁷</td>
<td>9.98×10⁻²⁵</td>
</tr>
<tr>
<td>rs6495308</td>
<td>15</td>
<td>76694711</td>
<td>T</td>
<td>0.77</td>
<td>3.30×10⁻¹⁰</td>
<td>0.10</td>
<td>7.99×10⁻²⁴</td>
<td>1.60×10⁻¹³</td>
</tr>
</tbody>
</table>