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

Jason Z Liu[1*], Federica Tozzi[2], Dawn M Waterworth[3], Sreekumar G Pillai[3], Pierandrea Muglia[2], Lefkos Middleton[4], Wade Berrettini[5], Christopher W Knouff[6], Xin Yuan[3], Gérard Waeber[7,8], Peter Vollenweider[7,8], Martin Preisig[7,9], Nicholas J Wareham[10], Jing Hua Zhao[10], Ruth J F Loos[10], Inês Barroso[11], Kay-Tee Khaw[12], Scott Grundy[13], Philip Barter[14], Robert Mahley[15,16], Antero Kesaniemi[17,18], Ruth McPherson[19], John B Vincent[20], John Strauss[20], James L Kennedy[20], Anne Farmer[21], Peter McGuffin[21], Richard Day[22], Keith Matthews[22], Per Bakke[23], Amund Gulsvik[23], Susanne Lucae[24], Marcus Ising[24], Tanja Brueckl[24], Sonja Horstmann[24], H-Erich Wichmann[25–27], Rajesh Rawal[25], Norbert Dahmen[28], Claudia Lamina[25,29], Ozren Polasek[30], Lina Zgaga[31], Jennifer Huffman[32], Susan Campbell[32], Jaspal Kooner[33], John C Chambers[34], Mary Susan Burnett[35], Joseph M Devaney[35], Augusto D Pichard[35], Kenneth M Kent[35], Lowell Satler[35], Joseph M Lindsay[35], Ron Waksman[35], Stephen Epstein[35], James F Wilson[31], Sarah H Wild[31], Harry Campbell[31], Veronique Vitart[32], Muredach P Reilly[36,37], Mingyao Li[38], Liming Qu[38], Robert Wilensky[36], William Matthai[36], Hakon H Hakonarson[39], Daniel J Rader[36,37], Andre Franke[40], Michael Wittig[40], Arne Schäfer[40], Manuela Uda[41], Antonio Terracciano[42], Xiangjun Xiao[43], Fabio Busonero[41], Paul Scheet[43], David Schlessinger[42], David St Clair[44], Dan Rujescu[45], Gonçalo R Abecasis[46], Hans Jörgen Grabe[47], Alexander Teumer[48], Henry Völzke[49], Astrid Petersmann[50], Ulrich John[51], Igor Rudan[52,31], Caroline Hayward[32], Alan F Wright[32], Ivana Kolcic[30], Benjamin J Wright[53], John R Thompson[53], Anthony J Balmforth[54], Alistair S Hall[54], Nilesh J Samani[55], Carl A Anderson[11], Tariq Ahmad[56], Christopher G Mathew[57], Miles Parkes[58], Jack Satsangi[59], Mark Caulfield[60], Patricia B Munroe[60], Martin Farrall[61], Anna Dominiczak[62], Jane Worthington[63], Wendy Thomson[63], Steve Eyre[63], Anne Barton[63], The Wellcome Trust Case Control Consortium[65], Vincent Mooser[3], Clyde Francks[2,64] & Jonathan Marchini[1]

Smoking is a leading global cause of disease and mortality[1]. We established the Oxford-GlaxoSmithKline study (Ox-GSK) to perform a genome-wide meta-analysis of SNP association with smoking-related behavioral traits. Our final data set included 41,150 individuals drawn from 20 disease, population and control cohorts. Our analysis confirmed an effect on smoking quantity at a locus on 15q25 ($P = 9.45 \times 10^{-19}$) that includes *CHRNA5, CHRNA3* and *CHRNB4*, three genes encoding neuronal nicotinic acetylcholine receptor subunits. We used data from the 1000 Genomes project to investigate the region using imputation, which allowed for analysis of virtually all common SNPs in the region and offered a fivefold increase in marker density over HapMap2 (ref. 2) as an imputation reference panel. Our fine-mapping approach identified a SNP showing 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 are multifactorial traits with substantial genetic influences[3]. There is an urgent need to better understand the molecular neurobiology of nicotine dependence in order to design targeted, more effective therapies[4]. Recently, genome-wide association studies (GWAS) have established one locus associated with nicotine dependence and smoking quantity, which implicates a cluster of three genes, *CHRNA5*, *CHRNA3* and *CHRNB4* on chromosome 15q25, which encode neuronal nicotinic acetylcholine receptor subunits[5–9]. This locus is also associated with lung cancer[8,10,11], peripheral arterial disease[8] and chronic obstructive pulmonary disease and lung function[12].

We initially performed a GWAS meta-analytic study of smoking-related traits in a total sample of 41,150 individuals of European descent, sourced from several disease, population and control cohorts (**Table 1**, **Supplementary Table 1** and Online Methods). As the cohorts were genotyped on a variety of different genome-wide SNP arrays (**Table 1** and **Supplementary Table 1**), we first imputed genotypes for all data sets[13] for all SNPs in the HapMap version release 22 (ref. 2).

The main focus of our analysis was on smoking quantity within current and past smokers, defined as a semiquantitative trait based on the self-reported variable of cigarettes smoked per day (CPD)[8]. We performed association analyses separately within each cohort under

[*] Complete lists of author affiliations appear at the end of this paper.

Received 19 October 2009; accepted 18 March 2010; published online 25 April 2010; doi:10.1038/ng.572

Table 1 Summary information for the cohorts used in meta-analysis

			Sample Sizes					
Label	Description	Genotyping	All	CPD > 0	Ever	Never	Current	Non-current
WTCCC-RA	Rheumatoid arthritis cases	Affymetrix 500K	1,860	NA	NA	NA	262	558
EPIC	Obesity case-control	Affymetrix 500K	3,516	NA	1,927	1,589	353	1,574
WTCCC-HT	Hypertension cases	Affymetrix 500K	1,952	830	NA	NA	1,274	672
GEMS	Dyslipidemia case-control	Affymetrix 500K	1,847	862	910	793	268	642
GSK-COPD	COPD case-control	Illumina 550	1,633	1,632	NA	NA	725	905
GSK-Bipolar	Bipolar depression case-control	Illumina 550	1,805	944	1,008	790	498	510
GSK-UPD	Unipolar depression case-control	Illumina 550	1,792	899	935	856	503	432
WTCCC-IBD	Crohn's disease cases	Affymetrix 500K	1,748	NA	713	540	713	420
KORA	Population-based	Affymetrix 500K	1,644	253	811	831	217	1,425
KORCULA	Population-based	Illumina 300	827	NA	376	451	179	654
LOLIPOP	Population-based	Affymetrix 500K	1,288	650	653	635	258	395
MedStar	Coronary artery disease case-control	Affymetrix 6.0	1,322	820	853	469	300	553
ORCADES	Population-based	Illumina 300	692	NA	288	404	60	632
PENNCATH	Coronary artery disease case-control	Affymetrix 6.0	1,401	NA	NA	NA	464	612
POPGEN	Population-based	Affymetrix 6.0	1,107	573	495	608	NA	NA
CoLaus	Population-based	Affymetrix 500K	5,636	3,132	3,357	2,275	1,485	1,872
SardiNIA	Population-based	Affymetrix 500+10K	4,305	1,731	1,743	2,562	873	3,432
SHIP	Population-based	Affymetrix 6.0	4,080	2,011	2,631	1,449	1,240	2,840
VIS	Population-based	Illumina 300	769	NA	441	328	212	557
WTCCC-CAD	Coronary artery disease cases	Affymetrix 500K	1,926	1,237	1,457	461	239	1,218
TOTALS			41,150	15,574	18,598	15,041	10,123	19,903

Further details are given in Online Methods and **Supplementary Table 1**; NA, not applicable.

an additive model using covariate effects for age, sex, disease case or control status where applicable, and other cohort-specific covariates (**Supplementary Table 1**). A meta-analysis was then carried out by combining study-specific β (regression coefficient) estimates using a fixed effects model[14]. In total, 15,574 subjects reported CPD values over zero and were used for the meta-analysis of smoking quantity (**Table 1** and **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[15] and the TAG study of 74,035 subjects[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 smoking quantity identified the *CHRNA5–CHRNA3* locus on 15q25 as the single significant locus of note in the genome (**Fig. 1**, **Table 2** and **Supplementary Table 2**), with a minimum $P = 9.45 \times 10^{-19}$ for rs1051730, a SNP which has been previously reported to be associated with traits related to smoking[5–9]; we also found highly significant P values for many other SNPs in the region (**Supplementary Fig. 1** and **Supplementary Table 2**). All cohorts in the analysis contributed at least somewhat to the 15q25 association (**Supplementary Fig. 1**). Each copy of the A allele (34% frequency) had a quantitative effect size on smoking quantity of 0.079 (95% confidence interval 0.070–0.088), which is in line with previous estimates[8]. A joint analysis of our total data set, together with the TAG and ENGAGE data sets, for rs1051730 yielded $P = 1.71 \times 10^{-66}$ (**Table 2**).

Multiple variants at the 15q25 locus have been suggested to underlie its effect on smoking quantity, including a nonsynonymous SNP in *CHRNA5* and variants that affect mRNA expression levels[17–19]. We utilized our very large sample, in combination with data from the 1000 Genomes Project (see URLs), to perform fine mapping and modeling of the 15q25 locus in relation to smoking quantity. We reasoned that with the near complete information on common SNPs derived from the 1000 Genomes data set, it might be possible to pinpoint a variant or combination of variants that can explain the entirety of the signal of association at 15q25. We used data from 108 estimated CEU European-ancestry haplotypes from the April 2009 release of the 1000 Genomes Pilot 1 data. This data set contained 2,189 SNPs in our region of interest (Online Methods), which was approximately a fivefold increase in density compared to the 437 SNPs in release 22 of HapMap. By imputing genotypes for all SNPs across this locus from 1000 Genomes and by repeating the meta-analysis, we found that the most significant association was with a new and previously untested SNP which is not in the HapMap and is located within the 5′ untranslated region of *CHRNA5*; this location makes it a candidate for affecting mRNA transcription (rs55853698, $P = 1.31 \times 10^{-16}$; **Fig. 2**). The P value for the commonly reported SNP rs1051730 in this

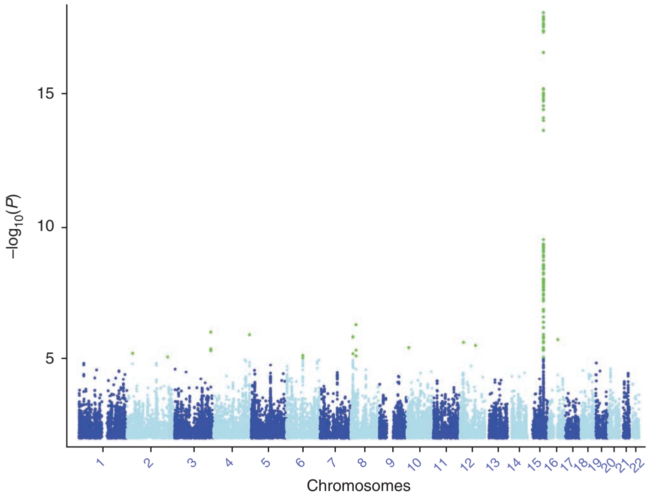

Figure 1 Plot showing the significance of association of all SNPs in the genome-wide smoking quantity meta-analysis. SNPs are plotted on the x axis according to their positions on each chromosome against association with smoking quantity on the y axis ($-\log_{10}$ P value). SNPs with P values $< 1.0 \times 10^{-5}$ are highlighted in green.

Table 2 Summary information for selected SNPs at 15q25 from meta-analysis of association with the Smoking Quantity (SQ) phenotype

					Ox-GSK		TAG	ENGAGE	Combined		
SNP	Chr.	Position	Coded allele	Coded allele freq.	P	P_{het}	P	P	P	β	s.e.m.
rs588765	15	76,652,480	T	0.43	1.74×10^{-3}	0.50	NA	NA	NA	NA	NA
rs16969968	15	76,669,980	G	0.65	1.64×10^{-18}	0.86	1.85×10^{-27}	1.53×10^{-23}	4.29×10^{-65}	–0.078	0.0046
rs1051730	15	76,681,394	G	0.66	9.45×10^{-19}	0.68	3.62×10^{-27}	9.98×10^{-25}	1.71×10^{-66}	–0.079	0.0046
rs6495308	15	76,694,711	T	0.77	3.30×10^{-10}	0.10	7.99×10^{-24}	1.60×10^{-13}	5.82×10^{-44}	0.073	0.0052

Our study is referred to as Ox-GSK. Information for all SNPs spanning the 15q25 locus in our genome-wide analysis is given in **Supplementary Table 2**. Chr., chromosome; Freq., frequency; P_{het}, heterozygosity P value; NA, not applicable.

analysis was similar but slightly less significant ($P = 1.47 \times 10^{-15}$). The P values for our 1000 Genomes analysis were generally higher than those from our HapMap-based analysis because not all of our study cohorts were included in the 1000 Genomes imputation (see Online Methods). rs55853698 is a G/T substitution, where the G allele has a frequency ranging from 0.313–0.378 across the various cohorts.

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

It has previously been suggested[18] that a nonsynonymous SNP, rs16969968, in *CHRNA5* is associated with nicotine dependence risk and lung cancer risk, but also that variants that cause high expression of *CHRNA5* mRNA, tagged by rs588765, increase the risk for nicotine dependence 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 Fig. 2**), with the most significant signal again occurring at rs6495308 ($P = 1.54 \times 10^{-5}$). Conditioning on both rs16969968 and rs588765, that is, the combination previously proposed[18], leaves no obvious signal of association (**Supplementary Fig. 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, in which smaller values indicate a better fit[20]. For the previous model[18], that is, conditioning on both rs16969968 and rs588765, we obtained BIC = 22,719.87 and a posterior probability 0.15. For the model conditioning on the new promoter SNP rs55853698 and rs6495308, we obtained BIC = 22,716.49 and a posterior probability 0.85, which indicates a better model fit.

Examination of the linkage disequilibrium (LD) structure between the SNPs considered here shows that rs1051730, rs16969968 and rs55853698 are all close-tagging proxies of each other (all pairwise $r^2 > 0.96$). These variants either tag or potentially cause the principal risk for high smoking quantity attributable to the 15q25 locus, but the high LD makes it difficult to assign specific causality. The SNPs that show residual association, rs588765 and rs6495308, are in low LD with each other ($r^2 = 0.21$) and are both in only 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 previously proposed[18]. Although the nonsynonymous SNP in *CHRNA5*, rs16969968, may be important, we have identified a new and potentially functional SNP in the 5′ untranslated region of this gene that is a close proxy for the nonsynonymous SNP in terms of LD, but which shows a slightly more significant association in our meta-analysis. Furthermore, although 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; this variant also contributes to a better-fitting two-SNP model and has a much stronger marginal significance in our unconditioned analysis ($P = 3.30 \times 10^{-10}$ for rs6495308 as compared to $P = 1.74 \times 10^{-3}$ for rs588765).

To our knowledge, our analysis has, for the first time, surveyed virtually all of the common SNPs in the 15q25 region and provides

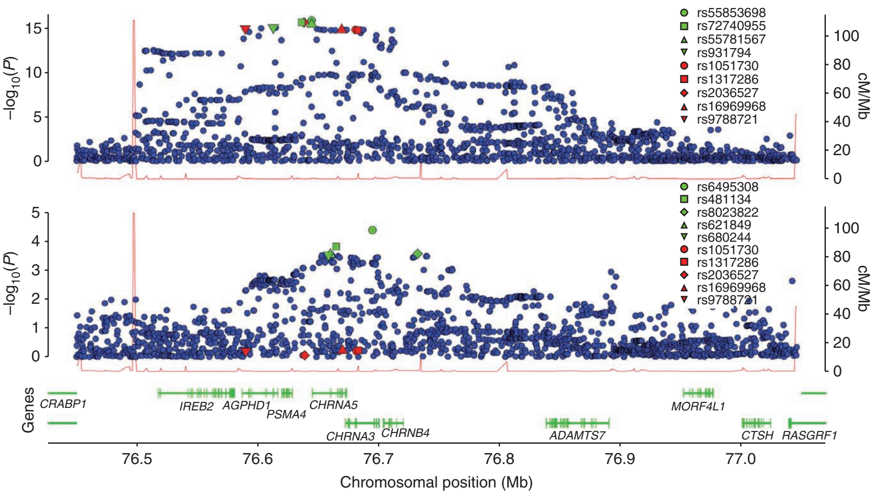

Figure 2 Chromosome 15q25 signal plots. Signal plot based on the 1000 Genomes imputation and meta-analysis of smoking quantity association (top). SNPs are plotted by their positions on the chromosome against association with smoking quantity ($-\log_{10}$ P value) on the left y axis. The five SNPs with 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 colored 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. Chromosome 15q25 signal plot based on the 1000 Genomes imputation and meta-analysis of smoking quantity association, conditional on rs55853698 (middle). 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. Genes and the positions of exons using data from the UCSC genome browser (bottom; see URLs).

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 nonsynonymous SNP rs16969968. It is notoriously difficult to distinguish functional variation when there is high LD across a region[21]. There are many ways in which variants can be functional, including expression regulatory changes that affect either close or distant genes, epigenetic changes, splicing effects, alterations to microRNA binding sites, or noncoding RNAs[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. In addition, common insertions or deletions can have functional effects, and the 1000 Genomes data will allow for analysis of this class of variant via an imputation framework.

The second-strongest association with smoking quantity within the genome in our meta-analysis was at a locus on 8p21 that received modest support from the TAG and ENGAGE studies (**Supplementary Table 2** and **Supplementary Fig. 3**; $P = 5.26 \times 10^{-7}$ for rs11782673). This locus would not remain significant after correcting for genome-wide multiple testing; however, it is noteworthy that the locus spans *CHRNA2*, another gene that encodes a neuronal nicotinic acetylcholine receptor subunit.

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

Our study identified association at some loci that, although not reaching genome-wide significance in our own meta-analysis, supported findings from the concurrent TAG and ENGAGE studies[15,16]. These include new loci on chromosomes 8 and 19 for smoking quantity, on chromosome 11 for ever/never and on chromosome 9 for current versus non-current smokers[15,16]. These findings have provided further new insights into the biology of smoking behavior.

URLs. ProbABEL software, http://mga.bionet.nsc.ru/~yurii/ABEL/; SNPTEST, IMPUTE and SNPMETA software, http://www.stats.ox.ac.uk/~marchini/software/gwas/gwas.html; 1000 Genomes Project: http://www.1000genomes.org/; April 2009 release of the 1000 Genomes Pilot 1 data, ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/release/2009_04/; UCSC Genome Browser, http://genome.ucsc.edu/; MERLIN, http://www.sph.umich.edu/csg/abecasis/merlin/; R, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

GlaxoSmithKline (GSK), a pharmaceuticals company that is interested in developing new cessation therapies for smoking, funded a postdoctoral fellowship for J.Z.L. 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. A. Roses and P. Matthews played crucial roles in establishing and funding the Medical Genetics activities at GSK. Acknowledgments that are specific to individual cohorts are given in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

J.Z.L. carried out most of the analysis for this study. J.M. and C.F. conceived and directed this study and wrote the manuscript. F.T., D.M.W. and V.M. were involved in study design and helped to coordinate the inclusion of many of the GSK cohorts. S.G.P., P. Muglia, L.M., W.B., C.W.K., X.Y., G.W., P.V., M. Preisig, N.J.W., J.H.Z., R.J.F.L., I.B., K.-T.K., S.G., P. Barter, R. Mahley, A.K., R. McPherson, J.B.V., J. Strauss, J.L.K., A. Farmer, P. McGuffin, R.D., K.M., P. Bakke, A.G., S.L., M.I., T.B., S.H., H.-E.W., R.R., N.D., C.L., O.P., L.Z., J.H., S.C., J.K., J.C.C., M.S.B., J.M.D., A.D.P., K.M.K.. L.S., J.M.L., R. Waksman, S. Epstein, J.F.W., S.H.W., H.C., V.V., M.P.R., M.L., L.Q., R. Wilensky, W.M., H.H.H., D.J.R., A. Franke, M.W., A.S., M.U., A. Terracciano, X.X., F.B., P.S., D.S., D.St.C., D.R., G.R.A., H.J.G., A. Teumer, H.V., A.P., U.J., I.R., C.H., A.F.W., I.K., B.J.W., J.R.T., A.J.B., A.S.H., N.J.S., C.A.A., T.A., C.G.M., M. Parkes, J. Satsangi, M.C., P.B.M., M.F., A.D., J.W., W.T., S. Eyre, A.B. and W.T.C.C.C. prepared and shared data sets and, in some cases, cohort-specific results from their own primary analysis.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturegenetics/.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Ezzati, M., Lopez, A.D., Rodgers, A., Vander Hoorn, S. & Murray, C.J. Selected major risk factors and global and regional burden of disease. *Lancet* **360**, 1347–1360 (2002).
2. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
3. Li, M.D. The genetics of nicotine dependence. *Curr. Psychiatry Rep.* **8**, 158–164 (2006).
4. Benowitz, N.L. Neurobiology of nicotine addiction: implications for smoking cessation treatment. *Am. J. Med.* **121**, S3–S10 (2008).
5. Berrettini, W. *et al.* α-5/α-3 nicotinic receptor subunit alleles increase risk for heavy smoking. *Mol. Psychiatry* **13**, 368–373 (2008).
6. Bierut, L.J. *et al.* Novel genes identified in a high-density genome wide association study for nicotine dependence. *Hum. Mol. Genet.* **16**, 24–35 (2007).
7. Li, M.D. Identifying susceptibility loci for nicotine dependence: 2008 update based on recent genome-wide linkage analyses. *Hum. Genet.* **123**, 119–131 (2008).
8. Thorgeirsson, T.E. *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638–642 (2008).
9. Caporaso, N. *et al.* Genome-wide and candidate gene association study of cigarette smoking behaviors. *PLoS One* **4**, e4653 (2009).
10. Amos, C.I. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* **40**, 616–622 (2008).
11. Hung, R.J. *et al.* A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* **452**, 633–637 (2008).
12. Pillai, S.G. *et al.* A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.* **5**, e1000421 (2009).
13. 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**, 906–913 (2007).
14. Normand, S.L. Meta-analysis: formulating, evaluating, combining, and reporting. *Stat. Med.* **18**, 321–359 (1999).
15. Thorgeirsson, T. *et al.* Sequence variants at *CHRNB3–CHRNA6* and *CYP2A6* affect smoking behavior. *Nat. Genet.* **42**, 448–453 (2010).
16. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat. Genet.* **42**, 441–447 (2010).
17. Falvella, F.S. *et al.* Transcription deregulation at the 15q25 locus in association with lung adenocarcinoma risk. *Clin. Cancer Res.* **15**, 1837–1842 (2009).
18. Wang, J.C. *et al.* Risk for nicotine dependence and lung cancer is conferred by mRNA expression levels and amino acid change in CHRNA5. *Hum. Mol. Genet.* **18**, 3125–3135 (2009).
19. Wang, J.C. *et al.* Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence. *Mol. Psychiatry* **14**, 501–510 (2008).
20. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
21. Ioannidis, J.P., Thomas, G. & Daly, M.J. Validating, augmenting and refining genome-wide association signals. *Nat. Rev. Genet.* **10**, 318–329 (2009).

[1]Department of Statistics, University of Oxford, Oxford, UK. [2]Genetics Division, GlaxoSmithKline, Verona, Italy. [3]Genetics Division, GlaxoSmithKline, Upper Merion, Pennsylvania, USA. [4]Division of Neurosciences and Mental Health, Imperial College London, UK. [5]Department of Psychiatry, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA. [6]Genetics Division, GlaxoSmithKline, Research Triangle Park, North Carolina, USA. [7]University Hospital Center, University of Lausanne, Lausanne, Switzerland. [8]Department of Internal Medicine, University of Lausanne, Lausanne, Switzerland. [9]Department of Psychiatry, University of Lausanne, Lausanne, Switzerland. [10]Medical Research Council Epidemiology Unit, Institute of Metabolic Science, Cambridge, UK. [11]Wellcome Trust Sanger Institute, Hinxton, UK. [12]Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [13]Center for Human Nutrition, University of Texas Southwestern Medical Center, Dallas, Texas, USA. [14]The Heart Research Institute, Sydney, New South Wales, Australia. [15]Gladstone Institute of Cardiovascular Disease, University of California, San Francisco, California, USA. [16]American Hospital, Istanbul, Turkey. [17]Department of Internal Medicine, University of Oulu, Oulu, Finland. [18]Biocenter Oulu, University of Oulu, Oulu, Finland. [19]Division of Cardiology, University of Ottawa Heart Institute, Ottawa, Ontario, Canada. [20]Centre for Addiction and Mental Health, University of Toronto, Toronto, Ontario, Canada. [21]Medical Research Council Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, UK. [22]Center for Neuroscience, Division of Medical Sciences, University of Dundee, Dundee, UK. [23]Institute of Medicine, University of Bergen, Bergen, Norway. [24]Max Planck Institute of Psychiatry, Munich, Germany. [25]Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [26]Institute of Medical Informatics, Biometry and Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany. [27]Klinikum Grosshadern, Munich, Germany. [28]Psychiatrische Klinik und Poliklinik University of Mainz, Mainz, Germany. [29]Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Innsbruck Medical University, Innsbruck, Austria. [30]School of Public Health, School of Medicine, University of Zagreb, Croatia. [31]Centre for Population Health Sciences, University of Edinburgh, Edinburgh, UK. [32]Institute of Genetics and Molecular Medicine, MRC Human Genetics Unit, Edinburgh, UK. [33]National Heart and Lung Institute, Imperial College London, London, UK. [34]Division of Epidemiology, Imperial College London, London, UK. [35]Cardiovascular Research Institute, MedStar Research Institute, Washington Hospital Center, Washington DC, USA. [36]The Cardiovascular Institute, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [37]The Institute for Translational Medicine and Therapeutics, School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [38]Biostatistics and Epidemiology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [39]The Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [40]Institute of Clinical Molecular Biology, Christian-Albrechts-University, Kiel, Germany. [41]Istituto di Neurogenetica e Neurofarmacologia, Consiglio Nazionale delle Ricerche, Monserrato, Cagliari, Italy. [42]National Institute on Aging, Baltimore, Maryland, USA. [43]Department of Epidemiology, University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [44]Department of Mental Health, University of Aberdeen, Aberdeen, UK. [45]Division of Molecular and Clinical Neurobiology, Department of Psychiatry, Ludwig-Maximilians-University, Munich, Germany. [46]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA. [47]Department of Psychiatry and Psychotherapy, University of Greifswald, Greifswald, Germany. [48]Interfacultary Institute for Genetics and Functional Genomics, University of Greifswald, Greifswald, Germany. [49]Institute for Community Medicine, University of Greifswald, Greifswald, Germany. [50]Institute of Clinical Chemistry and Laboratory Medicine, University of Greifswald, Greifswald, Germany. [51]Department of Social Medicine and Epidemiology, University of Greifswald, Greifswald, Germany. [52]Croatian Centre for Global Health, University of Split, Split, Croatia. [53]Department of Health Sciences, University of Leicester, Leicester, UK. [54]Mulitdisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds, UK. [55]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK. [56]Peninsula College of Medicine and Dentistry, Exeter, UK. [57]Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London, UK. [58]Gastroenterology Research Unit, Addenbrooke's Hospital, Cambridge, UK. [59]Gastrointestinal Unit, Molecular Medicine Centre, University of Edinburgh, Western General Hospital, Edinburgh, UK. [60]Clinical Pharmacology and Barts and the London Genome Centre, William Harvey Research Institute, Barts and the London School of Medicine, Queen Mary University of London, London, UK. [61]Department of Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, UK. [62]British Heart Foundation Glasgow Cardiovascular Research Centre, Division of Cardiovascular and Medical Sciences, University of Glasgow, Western Infirmary, Glasgow, UK. [63]arc Epidemiology Research Unit, School of Translational Medicine, Faculty of Medical and Human Sciences, University of Manchester, UK. [64]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [65]A full list of members is provided in the **Supplementary Note**. Correspondence should be addressed to J.M. (marchini@stats.ox.ac.uk) or C.F. (clyde.francks@well.ox.ac.uk).

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 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[22]: The Genetic Epidemiology of Metabolic Syndrome (GEMS) study consists of dyslipidemic case individuals (age 20–65 years) matched with normolipidemic controls by sex and recruitment site, drawn from non-Mediterranean subjects of the GEMS study (from Finland, Switzerland, Canada, Australia and the United States).

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

GSK COPD[12]: This collection includes case individuals with chronic obstructive pulmonary disease, diagnosed according to Global Initiative for Chronic Obstructive Lung Disease (GOLD) criteria, and unaffected controls recruited from Bergen, Norway.

GSK UPD[24]: This collection includes case individuals with recurrent major depression according to DSM-IV criteria and age- and gender-matched unaffected controls, recruited at the Max-Planck Institute of Psychiatry in Munich, Germany. Subjects were also recruited at two satellite recruiting hospitals (Bezirkskrankenhaus Augsburg and Klinikum Ingolstadt) in the Munich area.

GSK Bipolar[25]: The Bipolar collection included DSM-IV-diagnosed bipolar case individuals and controls from subjects recruited at three study sites: the Institute of Psychiatry (IOP) in London, UK; the Centre for Addiction and Mental Health in Toronto, Canada; and the University of Dundee, UK.

GSK LOLIPOP[26]: The London Life Sciences Prospective Population (LOLIPOP) was 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[27]: The MedStar cohort included case individuals with acute coronary syndrome or chronic coronary artery disease (CAD) from Washington DC, together with unaffected controls.

Penn-CATH[27]: The Penn-CATH cohort was a University of Pennsylvania Medical Center-based angiographic study from which case individuals with CAD and controls with no evidence of CAD at the coronary angiography were derived.

EPIC[28]: The EPIC-Obesity cohort was a case-control cohort for obesity drawn from the EPIC-Norfolk cohort which included men and women of European ancestry aged 39–79 years recruited in Norfolk, UK.

KORA[29]: The Cooperative Health Research in the Region of Augsburg (KORA) study was an epidemiological survey of the general population living in the city of Augsburg, southern Germany, and two adjacent counties.

WTCCC HT[30]: The WTCCC-HT collection comprised severely hypertensive probands ascertained from families with multiple affected members in the UK as part of the BRIGHT study.

WTCCC CAD, WTCCC CD and WTCCC RA[30]: These studies included individuals with CAD, Crohn's disease and rheumatoid arthritis from the Wellcome Trust Case Control Consortium Study.

POPGEN study[31]: The Population Genetic Cohort (POPGEN) was a cross sectional epidemiological survey of regional German populations from Schleswig-Holstein, northern Germany.

SHIP study[32]: The Study of Health in Pomerania (SHIP) was a longitudinal, population-based survey from West Pomerania, Germany. Data from the baseline cohort were used for this study.

VIS study[33]: This population cohort comprised Croatians aged 18–93 years recruited from the villages of Vis and Komiza on the Dalmatian island of Vis.

ORCADES study[34]: The Orkney Complex Disease Study (ORCADES) was a family-based, cross-sectional study that sought to identify genetic factors influencing cardiovascular and other disease risk in the population isolate of the Orkney Isles in northern Scotland.

KORCULA study[35]: The KORCULA study included healthy volunteers aged 18 and over from the villages of Lumbarda, Žrnovo, and Račišće on the Island of Korcula, Croatia.

SardiNIA study[36]: The SardiNIA was a population-based longitudinal cohort study that included 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, the genotype calling algorithms, SNP and sample quality control measures 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 coordinates for SNP base-pair positions so that all the cohorts could be successfully combined.

We imputed all SNPs reported in the CEU sample in HapMap Phase II using various imputation algorithms[13,37] (see URLs for a link to 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 quality control criteria.

Only imputed SNPs with good imputation quality were included in the meta-analysis. This was defined as proper_info ≥ 0.5 (a software-specific statistic for the studies analyzed with IMPUTE/SNPTEST[13]) or rsq-hat ≥ 0.5 (a statistic used for studies analyzed using MACH[37]) and Imp_info ≥ 0.5 (a statistic used for studies analyzed using ProbABEL).

Derivation of smoking phenotypes. We used the categorical smoking quantity levels previously defined[8]. The smoking quantity levels were 0 (defined as 1–10 CPD), 1 (11–20 CPD), 2 (21–30 CPD) and 3 (31 or more CPD). Each increment represents an increase in smoking quantity of 10 cigarettes per day. Most of the cohorts in our study have maximal CPD recorded on each sample, but a few collected average CPD (**Supplementary Table 1**). We examined the distributions of CPD across cohorts and found no large differences between those cohorts using average CPD and those using maximal CPD. The mean and standard deviation of the CPD measurements in each cohort are given in **Supplementary Table 1**. The ever/never and current/non-current phenotypes used were those collected by the individual cohorts. Not all cohorts had all three phenotypes (smoking quantity, ever/never and current/non-current) 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 the following:

> 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?
> List the number of years you have smoked.

Statistical analysis and meta-analysis. Each cohort was analyzed separately for each of the three phenotypes considered. The majority of the analysis was carried out on the raw genotype data at the Department of Statistics, University of Oxford, but some cohorts (SardiNIA, VIS, KORCULA, ORCADES and SHIP) carried out their own analysis and submitted results for the meta-analysis. For the binary traits (ever/never and current/non-current) tests for additive genetic effects on the log-odds scale were carried out using logistic regression. For the categorical smoking quantity 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 principal components analysis-derived covariates were included (a complete list of covariates is given in **Supplementary Table 1**). A genomic control inflation factor (λ) estimate was calculated for each phenotype and each cohort (**Supplementary Table 3**).

The meta-analysis was carried out by combining study-specific β estimates using a fixed effects model[14,] which used the inverse of the variance of the study-specific β estimates to give weight to the contribution of each study. The variance of each cohort's β estimate was multiplied by the genomic control λ estimate to correct for observed inflation[38]. Specifically,

$$\beta_{\text{META}} = \frac{\sum\limits_i \beta_i \Big/ \left(\lambda_i \sigma_i^2\right)}{\sum\limits_i 1 \Big/ \left(\lambda_i \sigma_i^2\right)}, \quad \sigma_{\text{META}} = \sqrt{\frac{1}{\sum\limits_i 1 \Big/ \left(\lambda_i \sigma_i^2\right)}}, \quad Z_{\text{META}} = \frac{\beta_{\text{META}}}{\sigma_{\text{META}}},$$

where β_i, σ_i^2 and λ_i are the β estimate, β-estimate variance and genomic control λ estimate for the i^{th} cohort. This method is appropriate when the same phenotype and measurement scale are used in each cohort, and it 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 smoking quantity by combining z-scores from each cohort weighted by their sample size[38] and obtained almost identical results. All meta-analysis was carried out using the SNPMETA program (see URLs). After performing each meta-analysis, the overall λ estimate for each phenotype was 1.0145 for smoking quantity, 1.002 for ever/never and 0.998 for current/non-current. For each SNP, we also calculated a P value for the heterogeneity across the studies[38].

SNP selection for replication. In collaboration with two other groups carrying out similar meta-analyses of smoking related traits (ENGAGE[15] and TAG[16]), we agreed to an *in silico* replication strategy in which for each phenotype (smoking quantity, ever/never, current/non-current) each group would select 15 regions of the genome showing evidence for association, and summary data (P values, β estimate, β-estimate variances, sample sizes, genomic control λ estimates and sample sizes) would be shared across groups to facilitate replication. We selected the top 15 regions for each phenotype on the basis of 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. We also excluded regions 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 1×10^{-5}. For each of the three phenotypes, the results from all the cohorts in all three concurrent studies were combined together using the same genomic-control–corrected inverse-variance meta-analysis method described above. A full list of the selected regions and the summary information from all three phenotypes is given in **Supplementary Table 2**.

1000 Genomes imputation analysis. 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 URLs for a link to the data source). We used these haplotypes to carry out imputation in the interval 76.4–77.0 Mb on chromosome 15 in 12 of the cohorts (GSK-Bipolar, GSK-UPD, GSK-COPD, KORA, POPGEN, Lausanne, GSK-LOLIPOP, GSK-GEMS, MedStar, SHIP, WTCCC-CAD and WTCCC-HT) using the program IMPUTE[13]. This release contains 2,189 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 database on the basis of base-pair position (rather than using SNP identifiers such as rs identities) so that as long as the same coordinate system is used for both the haplotype and genotype data, the alignment is automatic.

Conditional analysis and modeling. The analysis conditional upon the SNPs was carried out using all of the centrally analyzed cohorts (Bipolar, UPD, 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 were included in 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 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.

22. Stirnadel, H. *et al.* Genetic and phenotypic architecture of metabolic syndrome-associated components in dyslipidemic and normolipidemic subjects: the GEMS Study. *Atherosclerosis* **197**, 868–876 (2008).
23. Firmann, M. *et al.* The CoLaus study: a population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovasc. Disord.* **8**, 6 (2008).
24. Muglia, P. *et al.* Genome-wide association study of recurrent major depressive disorder in two European case-control cohorts. *Mol. Psychiatry* published online, doi:10.1038/mp.2008.131 (23 December 2008).
25. Scott, L.J. *et al.* Genome-wide association and meta-analysis of bipolar disorder in individuals of European ancestry. *Proc. Natl. Acad. Sci. USA* **106**, 7501–7506 (2009).
26. Chahal, N.S. *et al.* Ethnicity-related differences in left ventricular function, structure and geometry: a population study of UK Indian Asians and European whites. *Heart* **96**, 466–471 (2009).
27. Kathiresan, S. *et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.* **41**, 334–341 (2009).
28. Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br. J. Cancer* **80** (suppl. 1), 95–103 (1999).
29. Wichmann, H.E., Gieger, C. & Illig, T. KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67**, S26–S30 (2005).
30. Wellcome Trust Case-Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
31. Krawczak, M. *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Community Genet.* **9**, 55–61 (2006).
32. John, U. *et al.* Study of Health in Pomerania (SHIP): a health examination survey in an East German region: objectives and design. *Soz. Praventivmed.* **46**, 186–194 (2001).
33. Vitart, V. *et al.* SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat. Genet.* **40**, 437–442 (2008).
34. McQuillan, R. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).
35. Zemunik, T. *et al.* Genome-wide association study of biochemical traits in Korcula Island, Croatia. *Croat. Med. J.* **50**, 23–33 (2009).
36. Pilia, G. *et al.* Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet.* **2**, e132 (2006).
37. Li, Y. & Abecasis, G.R. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am. J. Hum. Genet.* **S79**, 2290 (2006).
38. de Bakker, P.I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).