Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci

A Mesut Erzurumluoglu
*University of Leicester*

Mengzhen Liu
*University of Minnesota*

Victoria E. Jackson
*University of Leicester*

Arpana Agrawal
*Washington University School of Medicine in St. Louis*

Laura Bierut
*Washington University School of Medicine in St. Louis*

*See next page for additional authors*

Follow this and additional works at: https://digitalcommons.wustl.edu/open_access_pubs

Please let us know how this document benefits you.

**Recommended Citation**

Erzurumluoglu, A Mesut; Liu, Mengzhen; Jackson, Victoria E.; Agrawal, Arpana; Bierut, Laura; Chou, Yi-Ling; Heath, Andrew; Madden, Pamela A F; Rice, John; and et al, "Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci." Molecular Psychiatry. 25, 10. 2392 - 2409. (2019).
https://digitalcommons.wustl.edu/open_access_pubs/11756

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact vanam@wustl.edu.
Authors
A Mesut Erzurumluoglu, Mengzhen Liu, Victoria E. Jackson, Arpana Agrawal, Laura Bierut, Yi-Ling Chou, Andrew Heath, Pamela A F Madden, John Rice, and et al
Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci

A. Mesut Erzurumluoglu1 et al.

Received: 2 May 2018 / Revised: 30 September 2018 / Accepted: 14 November 2018 / Published online: 7 January 2019
© The Author(s) 2019. This article is published with open access

Abstract
Smoking is a major heritable and modifiable risk factor for many diseases, including cancer, common respiratory disorders and cardiovascular diseases. Fourteen genetic loci have previously been associated with smoking behaviour-related traits. We tested up to 235,116 single nucleotide variants (SNVs) on the exome-array for association with smoking initiation, cigarettes per day, pack-years, and smoking cessation in a fixed effects meta-analysis of up to 61 studies (up to 346,813 participants). In a subset of 112,811 participants, a further one million SNVs were also genotyped and tested for association with the four smoking behaviour traits. SNV-trait associations with \( P < 5 \times 10^{-8} \) in either analysis were taken forward for replication in up to 275,596 independent participants from UK Biobank. Lastly, a meta-analysis of the discovery and replication studies was performed. Sixteen SNVs were associated with at least one of the smoking behaviour traits \( (P < 5 \times 10^{-8}) \) in the discovery samples. Ten novel SNVs, including rs12616219 near TMEM182, were followed-up and five of them (rs462779 in REV3L, rs12780116 in CNNM2, rs1190736 in GPR101, rs11539157 in PJA1, and rs12616219 near TMEM182) replicated at a Bonferroni significance threshold \( (P < 4.5 \times 10^{-3}) \) with consistent direction of effect. A further 35 SNVs were associated with smoking behaviour traits in the discovery plus replication meta-analysis (up to 622,409 participants) including a rare SNV, rs150493199, in CCDC141 and two low-frequency SNVs in CEP350 and HDGFRP2. Functional follow-up implied that decreased expression of REV3L may lower the probability of smoking initiation. The novel loci will facilitate understanding the genetic aetiology of smoking behaviour and may lead to the identification of potential drug targets for smoking prevention and/or cessation.

Introduction
Smoking is a major risk factor for many diseases, including common respiratory disorders such as chronic obstructive pulmonary disease (COPD) [1, 2], cancer [3] and cardiovascular diseases [4], and is reported to cause 1 in 10 premature deaths worldwide [5]. A greater understanding of the genetic aetiology of smoking behaviour has the potential to lead to new therapeutic interventions to aid smoking prevention and cessation, and thereby reduce the global burden of such diseases.

Previous genome-wide association studies (GWASs) identified 14 common SNVs [1, 6–12] (with minor allele frequency, MAF >0.01) robustly associated with smoking behaviour-related traits \( (P < 5 \times 10^{-8}) \). The 15q25 (CHRNA3/5-CHRNB4) region has the largest effect, explaining ~1% and 4–5% of the phenotypic variance of smoking quantity [13] and cotinine, a biomarker of nicotine intake [14], respectively. Overall, genetic loci identified to date explain ~2% of the estimated genetic heritability of smoking behaviour [6], which is reported to be between 40–60% [15–17]. A recent study suggested that an important proportion (~3.3%) of the phenotypic variance of smoking behaviour-related traits was explained by rare nonsynonymous variants (MAF <0.01)
Hence, well-powered studies of rare variants are needed.

To investigate the effect of rare coding variants on smoking behaviour, we studied 346,813 participants (of which 324,851 were of European ancestry) from 61 cohorts (Supp. Tables 1 and 2) at up to 235,116 SNVs from the exome array. As we had access to UK Biobank, we also interrogated SNVs present on the UK Biobank and UK BiLEVE Axiom arrays to identify additional associations across the genome beyond the exome array. To our knowledge, these datasets are an order of magnitude larger than the previous studies [6], and constitute the most powerful exome-array study of smoking behaviour to date.

Materials and methods

Participants

Our study combined study-level summary association data from up to 59 studies of European ancestry and two studies of South Asian ancestry from three consortia (Consortium for Genetics of Smoking Behaviour (CGSB), GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN) and the Coronary Heart Disease (CHD) Exome+ consortium), INTERVAL and UK Biobank. In total, up to 324,851 individuals of European ancestry and 21,962 South Asian individuals were analysed in the discovery stage (Fig. 1). Further information about the participating cohorts and consortia is given in Supp. Table 1 and the Supp. Material. All participants provided written informed consent and studies were approved by local Research Ethics Committees and/or Institutional Review boards.

Phenotypes

We chose to analyse the following four smoking behaviour-related traits because of their broad availability in existing epidemiological and medical studies, as well as their biological relevance for addiction behaviours:

i. Smoking initiation (binary trait: ever vs never smokers). Ever smokers were defined as individuals who have smoked >99 cigarettes in their lifetime, which is consistent with the definition by the Centre for Disease Control [19];

ii. Cigarettes per day (CPD; quantitative trait: average number of cigarettes smoked per day by ever smokers);
iii. Pack-years (quantitative trait; Packs per day x Years smoked, with a pack defined as 20 cigarettes); years smoked is typically formed from age at smoking commencement to current age for current smokers or age at cessation for former smokers.

iv. Smoking cessation (binary trait: former vs current smokers).

In UK Biobank, phenotypes were defined using phenotype codes 1239, 1249, and 2644 for smoking initiation and smoking cessation, and 1239, 3436, 3456 for CPD and pack-years. CPD was inverse normal transformed in the CHD Exome+, INTERVAL and CGSB studies and categorised (1–10, 11–20, 21–30, and 31+ CPD) by the GSCAN studies and UK Biobank (Supp. Table 2). All studies performed an inverse normal transformation of pack-years. Summary statistics of study level phenotype distributions are provided in Supp. Table 1.

Genotyping and quality control

Fifty-nine cohorts were genotyped using exome arrays (up to 235,116 SNVs) and two (UK Biobank and INTERVAL) were genotyped using Axiom Biobank Arrays (up to 820,000 SNVs; Supp. Table 2). In total, ~1.06M SNVs were analysed including ~64,000 SNVs on both the Axiom and Exome Arrays. Furthermore, two studies (NAGOZALC and GFG) genotyped their participants using arrays with custom content, increasing the total number of variants analysed to 1,207,583 SNVs. Individual studies performed quality control (QC; Supp. Material, Supp. Table 2) and additional QC was conducted centrally (i) to ensure alleles were consistently aligned, (ii) that there were no major sample overlaps between contributing studies, and (iii) variants conformed to Hardy–Weinberg equilibrium and call rate thresholds. We also examined the distribution of the effect sizes and test statistics across cohorts to ensure the test statistics were well-calibrated.

Study level analyses

Each study (including the case-cohort studies [20]) undertook analyses of up to four smoking traits using RAREMETALWORKER [21] or RVTESTS [22] (Supp. Table 2), which generated single variant score statistics and their covariance matrices within sliding windows of 1Mb. CPD and pack-years were analysed using linear models or linear mixed models. Smoking initiation and smoking cessation were analysed using logistic models or linear mixed models. All studies adjusted each trait for age, sex, at least three genetic principal components and any study-specific covariates (Supp. Table 2). Chromosome X variants were analysed using the above-described approach, but coding males as 0/2. This coding scheme ensures that on average females and males have equal dosages and so is optimal for genes that are inactivated (due to X chromosome inactivation) and is valid for genes that do not undergo X chromosome activation. Males and females were analysed together adjusting for sex as a covariate.

Single variant meta-analyses

Fixed effects meta-analyses across the individual contributing studies of single variant associations were undertaken using the Cochran-Mantel-Haenszel method in RAREMETAL. Z-score statistics were used in the meta-analysis to ensure that the association results are robust against potentially different units of measurement in the phenotype definitions across studies [23]. We performed genomic control correction on the meta-analysis results. Variants with \( P < 1 \times 10^{-6} \) in tests of heterogeneity were excluded. Variants with \( P \leq 5 \times 10^{-8} \) were taken forward for replication. In addition, rs12616219 was also taken forward for replication as its \( P \)-value was very close to this threshold (smoking initiation, \( P = 5.49 \times 10^{-8} \)). None of the rare SNVs were genome-wide significant, therefore we also took forward the rare variant with the smallest association \( P \)-value, rs141611945 (\( P = 2.95 \times 10^{-7} \); MAF < 0.0001).

Replication and combined meta-analysis of discovery and replication data

As UK biobank genetic data were released in two phases, we took the opportunity to replicate findings from the discovery stage in a further 275,596 individuals made available in the phase two release of UK Biobank genetic data. To avoid potential relatedness between discovery and replication samples, the replication samples were screened and individuals with relatedness closer than second degree with the discovery sample in the UK Biobank were removed [24]. Phenotypes were defined in the same way as the discovery samples (described above). Since the exome array and the UK Biobank Axiom arrays do not fully overlap, we used both genotyped exome variants (approx. 64,000) as well as the additional ~90,000 well-imputed exome array variants from UK Biobank (imputation quality score >0.3) for replication of single variant and gene-based tests. The rare \( \text{ATF6} \) variant was absent from the UK Biobank array and is more prevalent in Africans (MAF = 0.01) than Europeans (MAF = 0.0007). Therefore, replication was sought in 1,437 individuals of African American-ancestry from the HRS and COGA studies. Analysis methods for replication cohorts were the same as for discovery cohorts, including methods to analyse chromosome X (Supp. Table 2). The criteria set for the replication were (i) the same direction of
effect as the discovery analysis and (ii) \( P \leq 0.0045 \) in the replication studies (Bonferroni-adjusted for eleven SNVs at \( \alpha = 0.05 \)).

Finally, in order to fully utilise all available data, we carried out a combined meta-analysis of the discovery and replication samples across the exome array content using the same protocols mentioned above.

**Conditional analyses**

To identify conditionally independent variants within previously reported and novel loci a sequential forward stepwise selection was performed [25]. A 1 MB region was defined around the reported or novel sentinel variant (500 kb either side) and conditional analyses performed with all variants within the region. If a conditionally independent variant was identified, \( (P < 5 \times 10^{-6}; \) Bonferroni-adjusted for \( \sim 10,000 \) independent variants in the test region) the analysis was repeated conditioning on both the most significant conditionally independent variant and the sentinel variant. This stepwise approach was repeated (conditioning on the variants identified in current and earlier iterations) until there were no variants remaining in the region that were conditionally independent. The same protocol was followed for the novel SNVs identified in this study.

**Gene-based analyses**

For discovery gene-based meta-analyses, we utilised three statistical methods as part of the RAREMETAL package: the Weighted Sum Test (WST) [26], the burden test [27] and the Sequence Kernel Association test (SKAT) [28]. EPACTS (v.3.3.0) [29] was used to annotate variants (for use in gene-based meta-analyses), as recommended by RAREMETAL. Two MAF cut-offs were used, one used low-frequency (MAF < 0.05) and rare variants, the second only used rare variants (MAF < 0.01). Nonsynonymous, stop gain, splice site, start gain, start loss, stop loss, and synonymous variants were selected for inclusion. A sensitivity analysis to exclusion of synonymous variants was also performed. Gene-level associations with \( P < 8 \times 10^{-7} \) were deemed statistically significant (Bonferroni-adjusted for \( \sim 20,000 \) genes and three tests at \( \alpha = 0.05 \)). To examine if the gene associations were driven by a single variant, the gene tests were conducted conditional on the SNV with the smallest \( P \)-value in the gene, using the shared single variant association statistic and covariance matrices [21, 25].

**Mendelian randomization analyses**

To evaluate the causal effect of SI and CPD on BMI, schizophrenia and educational attainment (EA), we conducted Mendelian randomization (MR) analyses using three complementary approaches available in MR-Base [30]: inverse variance weighted regression [31], MR-Egger [32, 33], and weighted median [34]. We used both the previously reported smoking-associated SNVs and the SNVs from the current report (as provided in Tables 1–3 and Supp. Table 3) as instrumental variables. The BMI [35], schizophrenia [36] and educational attainment [37] data came from previously published publicly available data. To assess possible reverse causation, we also used outcome associated SNVs as instrumental variables and conducted MR analyses using SI and CPD as outcome. We considered \( P < 0.05 / 3 = 0.017 \) as statistically significant (Bonferroni-adjusted for three traits).

**In silico functional follow up of associated SNVs**

To identify whether the (replicated) SNVs identified here affected other traits, we queried the GWAS Catalog [38] (version: e91/28/02/2018, downloaded on 01/03/18) for genome-wide significant \( (P < 5 \times 10^{-8}) \) associations using all proxy SNVs \( (r^2 \geq 0.8) \) within 2 Mb of the top variant in our study.

eQTL lookups were carried out in the 13 brain tissues available in GTEx V7 [39], Brain xQTL (dorsolateral prefrontal cortex) [40] and BRAINEAC [41] databases, all of which had undergone QC by the individual studies. We did not perform additional QC on these data. In brief, GTEx used Storey’s q-value method to correct the FDR for testing multiple transcripts based upon the empirical \( P \)-values for the most significant SNV for each transcript [43, 42]. BRAINEAC calculated the number of tests per transcript and used Benjamini–Hochberg procedure to calculate FDR per transcript using a FDR < 1% as significant. BRAINxQTL used \( P < 8 \times 10^{-8} \) as a cut-off for significance for any given transcript. SNVs that met the study specific significance and FDR thresholds, which were in LD \( (r^2 > 0.8 \text{ in 1000 Genomes Europeans}) \) with the top eQTL or the sentinel eQTL for a given tissue/transcript combination were considered significant. The genes implicated by these eQTL databases and/or coding changes (e.g., missense and nonsense SNVs) were put into ConsensusPathDB [44] to identify whether these genes were over-represented in any known biological pathways. Replicated missense SNVs were also put into PolyPhen-2 [45] and FATHMM (unweighted) [46] to obtain variant effect prediction.

**Results**

**Single variant associations**

In the discovery meta-analyses, we identified 15 common SNVs that were genome-wide significant \( (P < 5 \times 10^{-8}) \) for one or more of the smoking behaviour traits, of which 9
<table>
<thead>
<tr>
<th>dbSNP ID (Exome ID)</th>
<th>Chr:Pos</th>
<th>EA/OA</th>
<th>Gene</th>
<th>Consequence</th>
<th>Trait</th>
<th>Discovery stage</th>
<th>Replication stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>EAF</td>
<td>DoE</td>
<td>P-value</td>
<td>Beta (SE)</td>
<td>P-value</td>
<td></td>
</tr>
<tr>
<td>rs141611945 (exm118559)</td>
<td>1:161771868</td>
<td>G/A</td>
<td>ATF6</td>
<td>Missense CPD</td>
<td>128,746</td>
<td>0.0065%</td>
<td>+</td>
</tr>
<tr>
<td>rs1190736 ** (exm1659559)</td>
<td>X:136113464</td>
<td>A/C</td>
<td>GPR101</td>
<td>Missense CPD (PY)</td>
<td>99,037</td>
<td>46.6%</td>
<td>-</td>
</tr>
<tr>
<td>rs462779 (exm572256)</td>
<td>6:111695887</td>
<td>A/G</td>
<td>REV3L</td>
<td>Missense SI</td>
<td>346,682</td>
<td>80.1%</td>
<td>-</td>
</tr>
<tr>
<td>rs216195 (exm1276230)</td>
<td>17:2203167</td>
<td>G/T</td>
<td>SMG6</td>
<td>Missense SI</td>
<td>335,406</td>
<td>27.3%</td>
<td>-</td>
</tr>
<tr>
<td>rs11539157 (exm1643833)</td>
<td>X:68381264</td>
<td>A/C</td>
<td>PJA1</td>
<td>Missense SI</td>
<td>289,917</td>
<td>16.5%</td>
<td>+</td>
</tr>
<tr>
<td>rs12616219</td>
<td>2:104352495</td>
<td>A/C</td>
<td>TMEM182</td>
<td>Intergenic SI</td>
<td>112,811</td>
<td>46.4%</td>
<td>-</td>
</tr>
<tr>
<td>rs1505691</td>
<td>6:28168033</td>
<td>G/A</td>
<td>ZSCAN9</td>
<td>Missense SI</td>
<td>112,811</td>
<td>34.8%</td>
<td>-</td>
</tr>
<tr>
<td>rs2841334</td>
<td>9:128122320</td>
<td>A/G</td>
<td>GAPVD1</td>
<td>Intronic SI</td>
<td>112,811</td>
<td>20.9%</td>
<td>-</td>
</tr>
<tr>
<td>rs202664</td>
<td>22:41813886</td>
<td>C/T</td>
<td>TOB2</td>
<td>Intergenic SC</td>
<td>51,043</td>
<td>19.9%</td>
<td>-</td>
</tr>
<tr>
<td>rs11895381</td>
<td>2:60053727</td>
<td>A/G</td>
<td>BCL11A</td>
<td>Intergenic SI</td>
<td>112,811</td>
<td>34.2%</td>
<td>-</td>
</tr>
<tr>
<td>rs12780116</td>
<td>10:104821946</td>
<td>A/G</td>
<td>CNNM2</td>
<td>Intergenic SI</td>
<td>112,811</td>
<td>13.9%</td>
<td>+</td>
</tr>
</tbody>
</table>

Novel smoking trait associated SNVs that replicated with \( P < 0.005 \) and had consistent direction of effect in discovery and replication are highlighted in bold. The replication sample size for smoking initiation (SI), cigarettes per day (CPD), pack-years (PY), and smoking cessation (SC) were 275,596, 80,015, 78,897, and 123,851 respectively. Chromosome (Chr) and position (Pos) for hg19 build 37. EA effect allele, OA other allele, Gene closest gene, \( N \) number of individuals, EAF effect allele frequency in the pooled samples, MAC minor allele count, DoE direction of effect, SE standard error. All SNVs had heterogeneity \( P > 0.02 \) in the discovery stage. *Replication was sought in 1,437 individuals of African American-ancestry from the HRS and COGA studies; **The beta(005) for the association of rs1190736 with PY in the replication stage was \( -0.026 (0.0039) \)
<table>
<thead>
<tr>
<th>dbSNP ID (Exome-chip ID)</th>
<th>Chr:Pos</th>
<th>EA/OA</th>
<th>Gene</th>
<th>Consequence</th>
<th>Trait</th>
<th>EAF</th>
<th>Beta (se) in replication stage</th>
<th>P-value in combined meta-analysis</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs1514175</td>
<td>1:74991644</td>
<td>G/A</td>
<td>TNNT3K</td>
<td>Intronic</td>
<td>SI</td>
<td>0.57</td>
<td>−0.011 (0.003)</td>
<td>(5.42 \times 10^{-8}) (9.03 \times 10^{-7})</td>
<td>1.0 \times 10^{-5}</td>
</tr>
<tr>
<td>rs7096169</td>
<td>10:104618695</td>
<td>G/A</td>
<td>BORCS7 (CNNM2# in Table 1)</td>
<td>Intronic</td>
<td>SI</td>
<td>0.31</td>
<td>0.016 (0.003)</td>
<td>(2.17 \times 10^{-12}) (3.38 \times 10^{-12})</td>
<td>7.3 \times 10^{-12}</td>
</tr>
<tr>
<td>rs2292259</td>
<td>12:56482180</td>
<td>G/F</td>
<td>ERBB3</td>
<td>Intronic</td>
<td>SI</td>
<td>0.66</td>
<td>0.0121 (0.003)</td>
<td>(2.78 \times 10^{-8}) (7.56 \times 10^{-8})</td>
<td>1.5 \times 10^{-7}</td>
</tr>
<tr>
<td>rs216195</td>
<td>17:2203167</td>
<td>G/T</td>
<td>SMG6#</td>
<td>Missense</td>
<td>SI</td>
<td>0.29</td>
<td>−0.0076 (0.003)</td>
<td>(2.41 \times 10^{-8}) (2.80 \times 10^{-8})</td>
<td>5.5 \times 10^{-8}</td>
</tr>
<tr>
<td>rs2900306 (exm383568)</td>
<td>4:2990499</td>
<td>T/G</td>
<td>GRK4</td>
<td>Missense</td>
<td>CPD</td>
<td>0.34</td>
<td>−0.024 (0.005)</td>
<td>(1.66 \times 10^{-10}) (3.99 \times 10^{-10})</td>
<td>3.8 \times 10^{-10}</td>
</tr>
<tr>
<td>rs4908760</td>
<td>1:44037685</td>
<td>A/G</td>
<td>PTPRF</td>
<td>Intronic</td>
<td>SI</td>
<td>0.35</td>
<td>0.0078 (0.003)</td>
<td>(1.76 \times 10^{-8}) (3.36 \times 10^{-8})</td>
<td>4.7 \times 10^{-8}</td>
</tr>
<tr>
<td>rs6692219 (exm127721)</td>
<td>1:179989584</td>
<td>C/G</td>
<td>CEP350</td>
<td>Missense</td>
<td>SI</td>
<td>0.028</td>
<td>−0.0257 (0.003)</td>
<td>(4.69 \times 10^{-8}) (1.08 \times 10^{-7})</td>
<td>1.3 \times 10^{-6}</td>
</tr>
<tr>
<td>rs11971186</td>
<td>7:12637897</td>
<td>G/A</td>
<td>GRM8</td>
<td>Intronic</td>
<td>SI</td>
<td>0.20</td>
<td>−0.0080 (0.003)</td>
<td>(1.45 \times 10^{-9}) (3.39 \times 10^{-9})</td>
<td>3.9 \times 10^{-9}</td>
</tr>
<tr>
<td>rs150493199 (exm249655)</td>
<td>2:179721072</td>
<td>A/T</td>
<td>CCDC141</td>
<td>Missense</td>
<td>SC</td>
<td>0.0098</td>
<td>0.048 (0.134)</td>
<td>(1.28 \times 10^{-8}) (6.45 \times 10^{-8})</td>
<td>6.0 \times 10^{-8}</td>
</tr>
<tr>
<td>rs3001723</td>
<td>4:14037685</td>
<td>A/G</td>
<td>PTPRF</td>
<td>Intronic</td>
<td>SI</td>
<td>0.21</td>
<td>0.0159 (0.003)</td>
<td>(6.64 \times 10^{-11}) (0.00015)</td>
<td>4.1 \times 10^{-11}</td>
</tr>
<tr>
<td>rs1937455</td>
<td>1:6616939</td>
<td>G/A</td>
<td>PDE4B</td>
<td>Intronic</td>
<td>SI</td>
<td>0.30</td>
<td>−0.0146 (0.0027)</td>
<td>(1.23 \times 10^{-9}) (0.00073)</td>
<td>5.6 \times 10^{-9}</td>
</tr>
<tr>
<td>rs2947411</td>
<td>1:45154908</td>
<td>G/A</td>
<td>SIX3</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.38</td>
<td>0.0136 (0.002)</td>
<td>(4.12 \times 10^{-11}) (1.77 \times 10^{-11})</td>
<td>3.8 \times 10^{-11}</td>
</tr>
<tr>
<td>rs6738833</td>
<td>2:104150891</td>
<td>T/C</td>
<td>TMEM182#</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.33</td>
<td>−0.018 (0.003)</td>
<td>(8.66 \times 10^{-14}) (1.63 \times 10^{-13})</td>
<td>4.4 \times 10^{-14}</td>
</tr>
<tr>
<td>rs13026471</td>
<td>2:137564022</td>
<td>T/C</td>
<td>TMEM182#</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.18</td>
<td>0.0217 (0.003)</td>
<td>(2.45 \times 10^{-8}) (0.00028)</td>
<td>3.0 \times 10^{-8}</td>
</tr>
<tr>
<td>rs6724928</td>
<td>2:15600591</td>
<td>C/T</td>
<td>KCN13</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.22</td>
<td>−0.011 (0.003)</td>
<td>(4.47 \times 10^{-8}) (0.00194)</td>
<td>3.8 \times 10^{-7}</td>
</tr>
<tr>
<td>rs13022438</td>
<td>2:162800372</td>
<td>G/A</td>
<td>SLCA4A10</td>
<td>Intronic</td>
<td>SI</td>
<td>0.27</td>
<td>0.0146 (0.003)</td>
<td>(1.41 \times 10^{-6}) (0.0005)</td>
<td>8.1 \times 10^{-6}</td>
</tr>
<tr>
<td>dbSNP ID (Exome-chip ID)</td>
<td>Chr:Pos</td>
<td>EA/OA</td>
<td>Gene</td>
<td>Consequence</td>
<td>Trait</td>
<td>EAF</td>
<td>Beta (se) in replication stage</td>
<td>P-value in combined meta-analysis</td>
<td>Notes</td>
</tr>
<tr>
<td>--------------------------</td>
<td>---------</td>
<td>-------</td>
<td>------</td>
<td>-------------</td>
<td>-------</td>
<td>-----</td>
<td>------------------------------</td>
<td>---------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>rs1869244</td>
<td>3:5724531</td>
<td>A/G</td>
<td>LOC105376939</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.32</td>
<td>0.0123 (0.003)</td>
<td>2.76 × 10^{-9} (0.00040)</td>
<td>rs11776293 is a eQTL for CHRNA2 in cerebellum in GTEx</td>
</tr>
<tr>
<td>rs35438712</td>
<td>3:8588205</td>
<td>T/C</td>
<td>CADM2</td>
<td>Intronic</td>
<td>SI</td>
<td>0.25</td>
<td>0.017 (0.003)</td>
<td>1.99 × 10^{-10} (1.15 × 10^{-5}/3.2 × 10^{-6})</td>
<td></td>
</tr>
<tr>
<td>rs6883351</td>
<td>5:22193967</td>
<td>T/C</td>
<td>CDH12</td>
<td>Intronic</td>
<td>SI</td>
<td>0.34</td>
<td>0.0129 (0.003)</td>
<td>4.69 × 10^{-9} (0.0010/1.4 × 10^{-6})</td>
<td></td>
</tr>
<tr>
<td>rs6414946</td>
<td>5:87729711</td>
<td>C/A</td>
<td>TMEM161B</td>
<td>Intronic</td>
<td>SI</td>
<td>0.32</td>
<td>−0.0137 (0.003)</td>
<td>5.2 × 10^{-9}</td>
<td></td>
</tr>
<tr>
<td>rs11747772</td>
<td>5:16692708</td>
<td>C/T</td>
<td>TENM2</td>
<td>Intronic</td>
<td>SI</td>
<td>0.25</td>
<td>0.0144 (0.003)</td>
<td>6.20 × 10^{-9} (0.011/2.2 × 10^{-6})</td>
<td></td>
</tr>
<tr>
<td>rs9320995</td>
<td>6:98726381</td>
<td>G/A</td>
<td>POU3F2</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.18</td>
<td>0.0150 (0.003)</td>
<td>1.70 × 10^{-9} (0.00079)</td>
<td></td>
</tr>
<tr>
<td>rs10255516</td>
<td>7:1675621</td>
<td>G/A</td>
<td>ELF1</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.33</td>
<td>−0.0139 (0.003)</td>
<td>2.86 × 10^{-9} (0.0021/1.8 × 10^{-6})</td>
<td></td>
</tr>
<tr>
<td>rs10807839</td>
<td>7:10807839</td>
<td>G/A</td>
<td>SDK1</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.32</td>
<td>0.0139 (0.003)</td>
<td>4.69 × 10^{-9}</td>
<td></td>
</tr>
<tr>
<td>rs11776293</td>
<td>7:11776293</td>
<td>T/C</td>
<td>EPHX2</td>
<td>Intronic</td>
<td>SI</td>
<td>0.12</td>
<td>−0.0200 (0.003)</td>
<td>2.23 × 10^{-12} (0.00013/8.9 × 10^{-9})</td>
<td></td>
</tr>
<tr>
<td>rs1562612</td>
<td>8:9817068</td>
<td>G/A</td>
<td>TOX</td>
<td>Intronic</td>
<td>SI</td>
<td>0.35</td>
<td>−0.0112 (0.003)</td>
<td>1.15 × 10^{-9} (1.42 × 10^{-5}/2.9 × 10^{-5})</td>
<td></td>
</tr>
<tr>
<td>rs3857914</td>
<td>8:93184065</td>
<td>C/T</td>
<td>RUNX1T1</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.19</td>
<td>0.0157 (0.003)</td>
<td>1.54 × 10^{-9} (0.00571)</td>
<td></td>
</tr>
<tr>
<td>rs2799849</td>
<td>9:86725641</td>
<td>C/T</td>
<td>RM10</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.22</td>
<td>−0.0156 (0.003)</td>
<td>1.94 × 10^{-9} (0.002648)</td>
<td></td>
</tr>
<tr>
<td>rs6482190</td>
<td>10:22037809</td>
<td>A/G</td>
<td>LOC105376939</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.17</td>
<td>0.0146 (0.003)</td>
<td>8.85 × 10^{-9} (0.00219.5 × 10^{-5})</td>
<td></td>
</tr>
<tr>
<td>rs4523689</td>
<td>11:7950797</td>
<td>G/A</td>
<td>OR10A6</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.27</td>
<td>−0.012 (0.003)</td>
<td>7.77 × 10^{-9} (0.00306/2.2 × 10^{-5})</td>
<td></td>
</tr>
<tr>
<td>rs933006</td>
<td>13:79305193</td>
<td>A/G</td>
<td>TRPC4</td>
<td>Intronic</td>
<td>SI</td>
<td>0.30</td>
<td>−0.0413 (0.003)</td>
<td>3.50 × 10^{-9} (0.002296)</td>
<td></td>
</tr>
<tr>
<td>rs557899</td>
<td>15:87643795</td>
<td>A/C</td>
<td>SEMA6D</td>
<td>Intergenic</td>
<td>SI</td>
<td>0.26</td>
<td>0.0157 (0.003)</td>
<td>2.99 × 10^{-10} (4.46 × 10^{-5}/1.0 × 10^{-5})</td>
<td></td>
</tr>
<tr>
<td>rs76085882</td>
<td>19:4474725</td>
<td>A/C</td>
<td>HDGF2P2</td>
<td>Intronic</td>
<td>SI</td>
<td>0.029</td>
<td>−0.0360 (0.007)</td>
<td>8.50 × 10^{-9} (0.001243)</td>
<td></td>
</tr>
</tbody>
</table>

Chromosome (Chr) and position (Pos) for each SNV is given for hg19 build 37. Only SNVs reaching genome-wide significance (P < 5 × 10^{-8}, in bold) in the combined meta-analysis are shown. Magnitude of the effect size estimates are not presented as traits were transformed in differently by the three consortia analysed. SNVs identified in the discovery stage of this study (see Table 1) are denoted #. The discovery sample size for smoking initiation (SI), CPD, pack-years (PY), and smoking cessation (SC) were 346,813, 128,746, 131,892, and 121,543, respectively; and the replication sample size for SI, CPD, PY, and SC were 275,596, 80,015, 78,897, and 123,851, respectively. NB: rs6673752 (intronic to UBAP2L) was not available in the discovery cohorts. EA effect allele, OA other allele, Beta (se) beta and standard error for association in the replication stage. All SNVs had heterogeneity P > 0.0001

Bold font highlights the genome-wide significant P-values from the meta-analysis of discovery plus replication studies.
Table 3 Results from conditional analyses at previously reported smoking behaviour loci

<table>
<thead>
<tr>
<th>Gene region</th>
<th>dbSNP ID</th>
<th>Chr:Pos</th>
<th>EA/OA</th>
<th>Consequence</th>
<th>Trait</th>
<th>EAF</th>
<th>( P ) (unconditional)</th>
<th>SNVs conditioned on</th>
<th>Discovery Conditional ( P )</th>
<th>Conditional ( P ) in replication [DoE]</th>
</tr>
</thead>
<tbody>
<tr>
<td>19q13 (RAB4B)</td>
<td>rs8102683</td>
<td>19:41363765 C/T</td>
<td>Intergenic</td>
<td>CPD</td>
<td>74.8%</td>
<td>( 4.53 \times 10^{-16} )</td>
<td>rs7937</td>
<td>1.44 ( \times 10^{-12} ) [+ -]</td>
<td>3.5 ( \times 10^{-4} ) [+ -]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs28399442</td>
<td>19:41354458 A/C</td>
<td>Intronic (CYP2A6)</td>
<td>CPD</td>
<td>1.3%</td>
<td>( 2.27 \times 10^{-12} )</td>
<td>rs7937, rs8102683</td>
<td>2.63 ( \times 10^{-12} ) [+ -]</td>
<td>8.1 ( \times 10^{-14} ) [+ -]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs3865453</td>
<td>19:41338556 T/C</td>
<td>Intergenic</td>
<td>CPD</td>
<td>6.54%</td>
<td>( 2.96 \times 10^{-12} )</td>
<td>rs7937, rs8102683, rs28399442</td>
<td>4.96 ( \times 10^{-16} ) [- -]</td>
<td>2.3 ( \times 10^{-11} ) [- -]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs11694518</td>
<td>2:416125523 A/C</td>
<td>Intronic (CYP2A6)</td>
<td>CPD</td>
<td>29.5%</td>
<td>( 2.63 \times 10^{-10} )</td>
<td>rs10193706</td>
<td>3.43 ( \times 10^{-7} ) [- -]</td>
<td>4.0 ( \times 10^{-11} ) [- -]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs3865453</td>
<td>19:41438556 T/C</td>
<td>Intergenic</td>
<td>CPD</td>
<td>76.4%</td>
<td>( 1.83 \times 10^{-6} )</td>
<td>rs1051730</td>
<td>7.77 ( \times 10^{-21} ) [+ -]</td>
<td>1.0 ( \times 10^{-13} ) [+ -]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rs938682</td>
<td>15:78882925 A/G</td>
<td>Intergenic (CYP2A6)</td>
<td>CPD</td>
<td>6.4%</td>
<td>( 2.63 \times 10^{-10} )</td>
<td>rs10193706</td>
<td>3.43 ( \times 10^{-7} ) [- -]</td>
<td>4.0 ( \times 10^{-11} ) [- -]</td>
<td></td>
</tr>
</tbody>
</table>

SNVs with \( P < 5 \times 10^{-8} \) are highlighted in bold. The discovery sample size for smoking initiation (SI) and CPD was 346,813 and 128,746, respectively. The replication sample size for SI and CPD were 275,596 and 80,015, respectively. Chr Chromosome, Pos position for hg19 build 37, EA effect allele, OA other allele, EAF effect allele frequency in the pooled samples, DoE Direction of effect.
We assessed evidence for associations at the 14 SNVs previously reported for smoking behaviour-related traits. Seven were genotyped on the exome array and proxies ($r^2 > 0.3; \pm 2\text{ Mb}$) were identified for the remaining seven (Supp. Table 3). All showed nominal evidence of association at $P < 0.05$ and six of these were genome-wide significant in the meta-analysis of the trait for which it was previously reported (Supp. Tables 3 and 5).

Conditional analyses identified five independent associations within three previously reported loci and all five replicated (Table 3). At the 19q13 (RAB4B) locus, there were three variants in or near CYP2A6 associated with CPD.

**Fig. 2** A concentric Circos plot of the association results for smoking initiation (SI; outer ring), cigarettes per day (CPD) and smoking cessation (SC; inner ring) for chromosomes 1–22 (Pack-years results, which can be found in Supp. Figure 1, are omitted for clarity). Each dot represents a SNV, with the X and Y axes corresponding to genomic location in Mb and $-\log_{10} P$-values, respectively. Labels show the nearest gene to the novel sentinel variants identified in the discovery stage and taken forward to replication. The top signals were truncated at $10^{-10}$ for clarity. Novel and previously reported signals are highlighted in red and dark blue, respectively. Grey rings on the y-axis increase by increments of 2 (initial ring corresponding to $P = 0.001$, then 0.000001 etc.); and the outer and inner red rings correspond to the genome-wide significance level ($P = 5 \times 10^{-8}$) and $P = 5 \times 10^{-7}$, respectively. Image was created using Circos (v0.65).
independently of the established variant (rs7937) and each other: rs8102683 (conditional \( P = 4.53 \times 10^{-16} \)), rs28399442 (conditional \( P = 2.63 \times 10^{-12} \)) and rs3865453 (conditional \( P = 4.96 \times 10^{-10} \)) and rs28399442 was a low-frequency variant. The same SNVs also showed evidence of independent effects with pack-years, albeit with larger \( P \)-values (\( P < 5 \times 10^{-6} \); Supp. Table 5). At the TEX41/PABPC1P2 locus, rs11694518 (conditional \( P = 3.43 \times 10^{-7} \)) was associated with smoking initiation independently of the established variant (rs10427255). At 15q25, rs938682 (\( P = 7.78 \times 10^{-21} \)) was associated with CPD independently of the established variant (rs1051730) and (in agreement with a previous report [49]) is an eQTL for CHRNA5 in brain putamen basal ganglia tissues in GTEx.

**Gene-based association studies**

Gene-based collapsing tests using MAF < 0.01 variants, did not identify any associated genes at the pre-specified \( P < 8 \times 10^{-7} \) threshold. Of the top four gene associations, three were novel (CHRNA2, MMP17, and CRCP) and one was known (CHRNA5), and had \( P < 7 \times 10^{-4} \), with CPD and/or pack-years (Supp. Table 6). Analyses conditional on the variant with the smallest \( P \)-value in the gene, revealed the associations at CHRNA2, MMP17 and CRCP were due to more than one rare variant (conditional \( P < 0.05 \); Supp. Table 6). In contrast, the CHRNA5 gene association was attributable to a single variant (rs2229961).

**Mendelian randomization analyses**

We conducted MR analyses to elucidate the potential causal impact of SI and CPD on BMI, schizophrenia and EA using the MR-Egger, median weighted and inverse variance weighted methods. We found a causal association between SI and EA using both the median weighted and inverse variance weighted methods (\( P < 0.0001 \); Supp. Table 7) but not with MR-Egger (\( P = 0.2 \)). There was an association of SI with BMI using MR-Egger only (\( P = 0.01 \); Supp. Table 7), but there was evidence of horizontal pleiotropy (\( P = 0.001 \)) and no support from the other methods. Similarly, increased CPD was only associated with reduced BMI using the weighted median approach (\( P = 0.009 \)) and not the other methods (\( P > 0.017 \)). We also tested if schizophrenia, EA or BMI causally influence CPD or SI using SNVs associated with schizophrenia, EA and BMI, respectively, as instrumental variables. No evidence of such reverse causation was found (Supp. Table 7). These results were consistent with previous analyses [50]. There was no evidence of a causal effect of SI on schizophrenia, or CPD on educational attainment (Supp. Table 7).

**Functional characterization of novel loci**

Using proxies with \( r^2 \geq 0.8 \) in 1000 Genomes Europeans, we queried the GWAS catalogue [38] (\( P \leq 5 \times 10^{-8} \)) for pleiotropic effects of our novel sentinel SNVs. Two, rs11539157 and rs3001723 were previously associated with schizophrenia [36], suggesting shared biological pathways between schizophrenia and smoking behaviours (Table 2). This fits with the known association of smoking with schizophrenia [51]. Two, rs1514175 and rs2947411 have previously been associated with BMI [52], and extreme obesity [53].

eQTL lookups in GTEx V7 (13 Brain tissues with ≥80 samples) [39], Brain xQTL [40] and BRAINEAC [41] databases revealed that the A allele at rs462779, which decreases risk of smoking initiation, also decreased expression of REV3L in cerebellum in GTEx (A allele \( P = 4.8 \times 10^{-8} \); \( \beta = -0.40 \)) and was in strong LD with the top eQTL for REV3L in cerebellum (\( r^2 = 0.86 \) with rs9487668 in 1000 Genomes Europeans). The smoking initiation-associated SNV, rs12780116, was an eQTL for BORCS7 in four brain tissues, and NT5C2 in the cerebellar hemisphere (A allele \( P = 4.5 \times 10^{-7} \); \( \beta = -0.32 \)) and the cerebellum (\( P = 5.6 \times 10^{-6} \); \( \beta = -0.415 \)) in strong LD with the top eQTL, \( r^2 = 0.97 \) with rs11191546). The G allele of a second variant in the region, rs7096169 (intronic to BORCS7 and only in weak LD with rs12780116, \( r^2 = 0.18 \) in 1000G Europeans) increases smoking initiation and reduces expression of BORCS7 and AS3MT in eight brain tissues (including dorsolateral prefrontal cortex in the Brain xQTL and was the top BORCS7 eSNP in GTEx in the Cerebellar Hemisphere, Cerebellum, and Spinal cord cervical-C1). The same variant also reduced expression of ARL3 in cerebellum in GTEx (Table 2).

Biological pathway enrichment analyses carried out in ConsensusPathDB [44] using the genes implicated by the eQTL databases (Table 2) and/or a coding SNVs (i.e., PJA1, GPR101) showed that the (i) pyrimidine metabolism and (ii) activation of nicotinic acetylcholine receptors pathways are enriched for these smoking behaviour associated genes (false discovery rate < 0.01; \( P < 0.0001 \)).

**Discussion**

Smoking is the most important preventable lifestyle risk factor for many diseases, including cancers [3, 54], heart disease [4, 55] and many respiratory diseases such as COPD [1, 2]. Not initiating is the best way to prevent smoking-related diseases and genetics can play a considerable part in smoking behaviours including initiation. We have performed the largest exome-wide genetic association study of smoking behaviour-related traits to date involving up to
622,409 individuals, and identified and replicated five associations, including two on the X-chromosome (Table 1). We identified a further 35 novel associations in a meta-analysis of discovery and replication cohorts (Table 2). We validated 14 previously reported SNV-smoking trait associations (Supp. Table 3) and identified secondary independent associations at three loci, including three in the 19q13 region (rs8102683, rs28399442, and rs3865453; Table 3).

Gene-based tests improve power by aggregating effects of rare variants. While no genes reached our Bonferroni-adjusted P-value threshold, we identified three candidate genes with multiple rare variant associations for future replication: calcitonin gene-related peptide-receptor component (CRCP) with CPD and CHRNA2 and MMP17 with pack-years (Supp. Table 6; also see ‘Genes of Interest’ section in Supp. Material). CRCP’s protein product is expressed in brain tissues amongst others and functions as part of a receptor complex for a neuropptide that increases intracellular cyclic adenosine monophosphate levels [56]. MMP17 encodes a matrix metalloproteinase that is also expressed in the brain and is a member of the peptidase M10 family, and proteins in this family are involved in the degradation of extracellular matrix in normal physiological processes [57]. Given, we were not able conclusively to identify rare variant associations, even larger studies, are required to identify rare variants associated with smoking behaviours. In addition, phenotypes such as cotinine levels [58] and nicotine metabolism speed [59] could be interrogated using methods such as MTAG [60] to improve power.

As recommended by UK Biobank, we analysed UK Biobank samples by adjusting for genotyping array because a subset of (extreme smokers in) UK Biobank were genotyped on a different array (UK BiLEVE). However, this adjustment could potentially introduce collider bias in analyses of smoking traits. Given that the UK BiLEVE study is relatively small compared to the full study, and the genetic effect sizes for smoking-associated variants are small, we expect the influence of collider bias to be small [61]. Nevertheless, we performed sensitivity analyses to assess the impact of collider bias. Firstly, we performed a meta-analysis excluding the UK BiLEVE samples, and secondly, we re-analysed UK Biobank without adjusting for genotype array. As expected, the estimated genetic effects from these additional analyses were very similar to our reported results suggesting collider bias is not a concern (Suppl. Table 8).

Follow-up of the replicated SNVs in the literature and eQTL databases implicated some potentially interesting genes: NT5C2 is known to hydrolyse purine nucleotides and be involved in maintaining cellular nucleotide balance, and was previously associated with schizophrenia [62]. REV3L, encodes the catalytic subunit of DNA polymerase ζ (zeta) which is involved in translesion DNA synthesis. Previously, polymorphisms in a microRNA target site of REV3L were shown to be associated with lung cancer susceptibility [63]. We showed that decreased expression of REV3L may also lower the probability of smoking initiation. The SNV, rs11776293, intronic in EPHX2, was associated with reduced SI in the combined meta-analysis, and is in LD with rs56372821 ($r^2 = 0.83$), which is associated with reduced cannabis use disorder [64]. rs216195 (in SMG6) was genome-wide significant in the discovery and the combined meta-analysis. SMG6 is a plausible candidate gene as it was previously shown to be less methylated in current smokers compared to never smokers [65]. The combined meta-analysis also identified a rare missense variant in CCDC141, rs150493199 (MAF < 0.01; Table 2). Coding variants in CCDC141 were previously associated with heart rate [66] and blood pressure [67, 68].

Smoking behaviours represent a complex phenotype that are linked to an array of socio-cultural and familial, as well as genetic determinants. Kong et al., recently reported that ‘genetic-nurture’ i.e., effects of non-transmitted parental alleles, affect educational attainment [69]. They also show that there is an effect of educational attainment and genetic nurture on smoking behaviour. Four of our sentinel SNVs (or a strong proxy; $r^2 > 0.8$) were associated with years of educational attainment [37] (rs2292239, rs3001723 ($P < 5 \times 10^{-8}$), rs9320995 ($P = 8.90 \times 10^{-7}$), and rs13022438 ($P = 3.79 \times 10^{-6}$), in agreement with this paradigm and our MR analyses indicated that initiating smoking reduced years in education. Future family studies will be required to disentangle how much of the variance explained in the current analysis is due to direct versus genetic nurturing effects.

Our study primarily focused on European ancestry, but we also included two non-European studies but these non-European studies lacked statistical power on their own to identify ancestry-specific effects. Therefore, we did not perform ancestry-specific meta-analyses. Nevertheless, our results offered cross ancestry replication. One of the associations identified in the conditional analyses, rs8102683 (near CYP2A6), confirmed an association with CPD that was previously identified by Kumaksa et al. in a Japanese population [70] but this is the first time it was associated in Europeans (rs8102683 is also correlated with rs56113850 ($r^2 = 0.43$), a SNV identified previously by Loukola et al. [59] in a genetic association study of nicotine metabolite ratio in Europeans). As more non-European studies become available, it would be of great interest to perform non-European ancestry studies, in order to fine-map causal variants for smoking-related traits.

CPD and pack-years are two correlated measures of smoking. In the ~40,000 individuals from UK Biobank with CPD and pack-years calculated, correlation between CPD and pack-years was 0.640. Interestingly, while pack-years was inversely correlated with smoking cessation (~0.18) i.e., the more years a smoker has been smoking the less
likely they were to cease, CPD was positively correlated with smoking cessation (0.13) i.e., heavier smokers were more likely to stop smoking. In contrast, the DBH SNV, rs3025343, (first identified via its association with increased smoking cessation [6]) was associated with increased pack-years ($P = 1.29 \times 10^{-14}$) and increased CPD ($P = 2.93 \times 10^{-9}$) in our study. The association at DBH also represents the first time that a SNV has a smaller $P$-value for pack-years ($n = 131,892$) compared to CPD ($n = 128,746$). These findings may help elucidate the genetic basis of these correlated addiction phenotypes.

We performed the largest exome-wide genetic association study of smoking behaviour-related traits to date and nearly doubled the number of replicated associations to 24 (including conditional analyses) including associations on the X-chromosome for the first time, which merit further study. We also identified a further 35 novel smoking trait associated SNVs in the combined meta-analysis. The novel loci identified in this study will substantially expand our knowledge of the smoking addiction-related traits, facilitate understanding the genetic aetiology of smoking behaviour and may lead to the identification of drug targets of potential relevance to prevent individuals from initiating smoking and/or aid smokers to stop smoking.

Acknowledgments Author acknowledgements are included in the Supplementary material.

Compliance with ethical standards

Conflict of interest Paul W. Franks has been a paid consultant for Eli Lilly and Sanofi Aventis and has received research support from several pharmaceutical companies as part of European Union Innovative Medicines Initiative (IMI) projects. Neil Poulter has received financial support from several pharmaceutical companies that manufacture either blood pressure lowering or lipid lowering agents or both and consultancy fees. Peter Sever has received research awards from Pfizer. Mark J. Caulfield is Chief Scientist for Genomics England, a UK government company. Adam Butterworth reports grants from Merck and Pfizer and grants outside of this work from AstraZeneca, Biogen, and Novartis, and personal fees from Novartis. John Danesh reports grants, personal fees and non-financial support from Merck Sharp & Dohme (MSD), grants, personal fees and non-financial support from Novartis, grants from British Heart Foundation, grants from European Research Council, grants from NiHR, grants from NHS Blood and Transplant, grants from Pfizer, grants from UK MRC, grants from Wellcome Trust, grants from AstraZeneca, outside the submitted work. Laura J. Bierut is listed as an inventor on Issued U.S. Patent 8,080,371. "Markers for Addiction" covering the use of certain SNPs in determining the diagnosis, prognosis, and treatment of addiction.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

References


Author List

A. Mesut Erzurumluoglu1 · Mengzhen Liu2 · Victoria E. Jackson1,3,4 · Daniel R. Barnes5 · Gargi Datta2,6 · Carl A. Melbourne1 · Robin Young9 · Chiara Batini5 · Praveen Surendran5 · Tao Jiang5 · Sheikh Daud Adnan7 · Saima Afq8 · Arpna Agrawal9 · Elisabeth Altmair20 · Antonis C. Antoniou11 · Folkert W. Asselbergs12,13,14 · Clemens Baumbach10 · Laura Bierut16 · Sarah Bertelsen17 · Michael Boehnke18 · Michiel L. Bots19,20 · David M Brazel6,21 · John C. Chambers8,22,23,24 · Jenny Chang-Claude25,26 · Chu Chen27,28 · Janie Corley29,30 · Yi-Ling Chou9 · Sean P. David31 · Rudolf a. de Boer32 · Christiana A. de Leeuw33 · Joe G. Dennis11 · Anna F. Dominiczak34 · Alison M. Dunning35 · Douglas F. Easton11,35 · Charles Eaton28 · Paul Elliott1 · Evgenios Evangelou20,40 · Jessica D. Fauld41 · Tatiana Foroud41 · Alison Goate42 · Jian Gong43 · Hans J. Grabe44 · Jeff Haessler45 · Christopher Human46 · Göran Hallmans16 · Anke R. Hammerschlag13 · Sarah E. Harris29,47 · Andrew Hattersley48 · Andrew Heath17 · Chris Hsu49 · William G. Iacono2 · Stavroula Kanoni50,51 · Manav Kapoor17 · Jaakko Kaprio52,53 · Sharon L. Kardia54 · Fredrik Karpe55,56 · Jaspal S. Kooner23,24,37,58 · Alina Naheed75 · Aliya Naeheed75 · Matthias Nauck76,77 · Nathan North78 · David R. Weir80 · Xuanxi Zhan81 · Xiaowei Zhan109 · Ian J. Deary79,80 · Panos Deloukas96,116,117 · Emanuela Di Angelantonio5,118 · Marco Ferrario119 ·

Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci
Jean Ferrières120 • Paul W. Franks85,121 • Tim M. Frayling48 • Philippe Frossard84 • Ian P. Hall122 • Caroline Hayward66 • Jan-Håkan Jansson123 • J. Wouter Jukema124,125 • Frank Kee126 • Satu Männisto57 • Andres Metspalu69 • Patricia B. Munroe103,104 • Børge Grønne Nordestgaard78 • Colin N. A. Palmer127 • Veikko Salomaa57 • Naveed Sattar128 • Timothy Spector129 • David Peter Strachan130 • Understanding Society Scientific Group, EPIC-CVD, GSCAN, Consortium for Genetics of Smoking Behaviour, CHD Exome+ consortium • Pim van der Harst32,131 • Eleftheria Zeggini96 • Danish Saleheen5,132,133 • Adam S. Butterworth5,118 • Louise V. Wain1,134 • Goncalo R. Abecasis18 • John Danesh5,96,118 • Martin D. Tobin1,134 • Scott Vrieze2 • Dajiang J. Liu108 • Joanna M. M. Howson5

1 Department of Health Sciences, University of Leicester, Leicester, UK
2 Department of Psychology, University of Minnesota, Minneapolis, MN, USA
3 Population Health and Immunity Division, The Walter and Eliza Hall Institute of Medical Research, 1G Royal Pde, 3052 Parkville, Australia
4 Department of Medical Biology, University of Melbourne, Melbourne 3010 Parkville, Australia
5 MRC/BHF Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK
6 Institute for Behavioral Genetics, University of Colorado Boulder, Boulder, CO, USA
7 National Institute of Cardiovascular Diseases, Sher-e-Bangla Nagar, Dhaka, Bangladesh
8 Department of Epidemiology and Biostatistics, Imperial College London, London W2 1PG, UK
9 Department of Psychiatry, Washington University, St. Louis, MO, USA
10 Research Unit of Molecular Epidemiology, Helmholtz Zentrum München-German Research Center for Environmental Health, Neuherberg, Germany
11 Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK
12 Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, University of Utrecht, Utrecht, The Netherlands
13 Durrer Center for Cardiovascular Research, Netherlands Heart Institute, Utrecht, The Netherlands
14 Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, London, UK
15 Farr Institute of Health Informatics Research and Institute of Health Informatics, University College London, London, UK
16 Department of Psychiatry, Washington University School of Medicine, St. Louis, MO USA
17 Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA
18 Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, MI, USA
19 Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, 3508GA Utrecht, The Netherlands
20 Center for Circulatory Health, University Medical Center Utrecht, 3508GA Utrecht, The Netherlands
21 Department of Molecular, Cellular, and Developmental Biology, University of Colorado Boulder, Boulder, CO, USA
22 Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore
23 Department of Cardiology, Ealing Hospital, Middlesex UB1 3HW, UK
24 Imperial College Healthcare NHS Trust, London W12 0HS, UK
25 Division of Cancer Epidemiology, German Cancer Research Centre (DKFZ), Heidelberg, Germany
26 Cancer Epidemiology Group, University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCCH), Hamburg, Germany
27 Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
28 Department of Epidemiology, University of Washington, Seattle, WA, USA
29 Centre for Cognitive Ageing and Cognitive Epidemiology, University of Edinburgh, Edinburgh EH8 9JZ, UK
30 Psychology, University of Edinburgh, Edinburgh EH8 9JZ, UK
31 Department of Medicine, Stanford University, Stanford, CA, USA
32 Department of Cardiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
33 Department of Complex Trait Genetics, Center for Neurogenomics and Cognitive Research, Amsterdam Neuroscience, VU University Amsterdam, Amsterdam, Netherlands
34 Institute of Cardiovascular and Medical Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK
35 Centre for Cancer Genetic Epidemiology, Department of Oncology, Cambridge Centre, University of Cambridge, Cambridge CB1 8RN, UK
36 Department of Epidemiology and Biostatistics, Imperial College London, London, UK
37 MRC-PHE Centre for Environment and Health, Imperial College London, London W2 1PG, UK
38 National Institute for Health Research Imperial Biomedical
Research Centre, Imperial College Healthcare NHS Trust and Imperial College London, London, UK

39 UK Dementia Research Institute (UK DRI) at Imperial College London, London, UK

40 Department of Hygiene and Epidemiology, University of Ioannina Medical School, Ioannina, Greece

41 Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA

42 Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, NY, USA

43 Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

44 Department of Psychiatry and Psychotherapy, University Medicine Greifswald, 17475 Greifswald, Germany

45 Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

46 Department of Public Health and Clinical Medicine, Nutritional research, Umeå University, Umeå, Sweden

47 Centre for Genomic and Experimental Medicine, University of Edinburgh, Edinburgh EH4 2XU, UK

48 Genetics of Complex Traits, University of Exeter Medical School, Exeter, UK

49 University of Southern California, California, CA, USA

50 William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London EC1M 6BQ, UK

51 Centre for Genomic Health, Queen Mary University of London, London EC1M 6BQ, UK

52 Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland

53 Department of Public Health, University of Helsinki, Helsinki, Finland

54 Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA

55 Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK

56 Oxford National Institute for Health Research, Biomedical Research Centre, Churchill Hospital, Oxford, UK

57 Department of Public Health Solutions, National Institute for Health and Welfare, FI-00271 Helsinki, Finland

58 National Heart and Lung Institute, Imperial College London, London W12 ONN, UK

59 Department of Biostatistics, University of Washington School of Medicine, Seattle, WA, USA

60 University of Eastern Finland, Finland, Finland

61 MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge CB2 0QQ, UK

62 Department of Medical Microbiology, Immunology and Cell Biology, Southern Illinois University School of Medicine, Springfield, IL, USA

63 Montreal Heart Institute, Montreal, Quebec H1T 1C8, Canada

64 Department of Medicine, Faculty of Medicine, Universite de Montreal, Montreal, Quebec H3T 1J4, Canada

65 NIHR Biomedical Research Centre at Guy’s and St Thomas’ Foundation Trust, London SE1 9RT, UK

66 MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

67 Queensland Institute for Medical Research, Brisbane, Australia

68 Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, 1683 Nicosia, Cyprus

69 Estonian Genome Center, University of Tartu, Tartu, Estonia

70 Institute of Health Informatics, University College London, London, UK

71 Department of Epidemiology and Public health, University Hospital of Strasbourg, Strasbourg, France

72 Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

73 Department of Medicine I, Ludwig-Maximilians-University Munich, Munich, Germany

74 DZHK (German Centre for Cardiovascular Research), Partner Site Munich Heart Alliance, Munich, Germany

75 Initiative for Noncommunicable Diseases, Health Systems and Population Studies Division, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh

76 Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, 17475 Greifswald, Germany

77 DZHK (German Centre for Cardiovascular Research), Partner Site Greifswald, University Medicine, Greifswald, Germany

78 Department of Clinical Biochemistry Herlev Hospital, Copenhagen University Hospital, Herlev Ringvej 74, DK-2730 Herlev, Denmark

79 Department of Epidemiology, University of North Carolina, Chapel Hill, NC, USA

80 Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA

81 Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche (CNR), Monserrato, Cagliari, Italy

82 Department of Clinical Genetics, VU University Medical Centre Amsterdam, Amsterdam Neuroscience, Amsterdam, Netherlands

83 International Centre for Circulatory Health, Imperial College London, London, UK

84 Centre for Non-Communicable Diseases, Karachi, Pakistan

85 Genetic and Molecular Epidemiology Unit, Lund University
Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci