Published in final edited form as:

Mol Psychiatry. 2017 January ; 22(1): 82–88. doi:10.1038/mp.2016.20.

Pathways to smoking behaviors: biological insights from the Tobacco and Genetics Consortium meta-analysis

Camelia C. Minică[@], Hamdi Mbarek, René Pool, Conor V. Dolan, Dorret I. Boomsma^{*}, and Jacqueline M. Vink^{*}

Department of Biological Psychology, Vrije Universiteit, Amsterdam, The Netherlands

Abstract

By running gene and pathway analyses for several smoking behaviors in the Tobacco and Genetics Consortium (TAG) sample of 74,053 individuals, twenty-one genes and several biological pathways were implicated. Analyses were carried out using the HYbrid Set-based Test (HYST) as implemented in the Knowledge-based mining system for Genome-wide Genetic studies software. Fifteen genes are novel and were not detected with the SNP-based approach in the original TAG analysis. For quantity smoked, fourteen genes passed the false discovery rate of 0.05 (corrected for multiple testing), with the top association signal located at the *IREB2* gene ($P = 1.57E-37$). Three genomic loci were significantly associated with ever smoked. The top signal is located at the noncoding antisense RNA transcript *BDNF-AS* ($P = 6.25E-07$) on 11p14. The *SLC25A21* gene ($P = 2.09E-08$) yielded the top association signal in the analysis of smoking cessation. The 19q13 noncoding RNA locus exceeded the genome-wide significance in the analyses of age at initiation ($P = 1.33E-06$). Pathways belonging to the Neuronal system pathways, harbouring the nicotinic acetylcholine receptor genes expressing the α (*CHRNA* 1-9), β (*CHRNA* 1-4), γ , δ and ϵ subunits, yielded the smallest p-values in the pathway analysis of quantity smoked (lowest $P = 4.90E-42$). Additionally, pathways belonging to 'a subway map of cancer pathways' regulating the cell-cycle, mitotic DNA replication, axon growth and synaptic plasticity, were found significantly enriched for genetic variants in ever smokers relative to never smokers (lowest $P = 1.61E-07$). In addition, these pathways were also significantly associated with quantity smoked (lowest $P = 4.28E-17$). Our results shed light on one of the world's leading causes of preventable death and open a path to potential therapeutic targets. These results are informative in decoding the biological bases of other disease traits, such as depression and cancers, with which smoking shares genetic vulnerabilities.

Keywords

Set-based genome-wide analyses; HYST; TAG; smoking; initiation; quantity; cessation

[@]To whom correspondence should be addressed: Camelia C. Minică, Department of Biological Psychology, Vrije Universiteit Amsterdam, Van der Boechehorststraat 1, 1081 BT, Amsterdam, The Netherlands. Telephone number: +31205983035; camelia.minica@gmail.com.

^{*}These authors contributed equally to this work.

Supplementary Information is available in the online version of the paper at <http://www.nature.com/mp/>.

Conflict of Interest: No conflict of interest has been declared.

Introduction

Tobacco smoking kills almost 6 million people each year (World Health Organization, 2015)¹. Despite smoking prevalence decreasing in the past 30 years, there has been a steady increase in the absolute number of smokers, i.e., from 721 million in 1980 to 967 million in 2012 due to accelerated population growth². Aside from environmental influences such as e.g., tobacco availability and social pressure³, genetic factors are implicated in all stages of smoking⁴, from experimentation⁵ to dependence^{5, 6} and cessation⁷. Although genome-wide association studies (GWASs; see⁸⁻¹⁰ for a recent overview) and the meta-analyses conducted by large consortia¹¹⁻¹³ have provided important insights into the biological bases of smoking, progress has been slow due to the small effects of the single nucleotide polymorphisms (SNPs) and to the multiple testing burden. In genetic studies, the power to detect small effects is linked primarily to sample size (given a fixed alpha level). However, power may also depend on the nature of the genotype-phenotype relation^{14, 15}. Most consortia focus initially on individual SNPs showing the strongest evidence for association. The TAG consortium with the largest sample yet of 74,053 individuals identified 130 SNPs (tagging the 15q25 locus) that passed the genome-wide threshold of 5×10^{-8} in the quantity smoked analysis. Subsequently, 1025 SNPs that passed the significance threshold of 10^{-4} were included in the follow-up SNP-based analyses in the combined TAG, ENGAGE¹² and the Oxford-GlaxoSmithKline¹¹ sample. We note that the remaining data (summary statistics for up to ~2.5 million SNPs in each analysis) have remained largely unexploited. Set-based tests (with a gene or a biological pathway as the unit of analysis) are more powerful than SNP based tests^{14, 16} as they consider jointly the weak effects of multiple SNPs within a gene or biological pathway. Furthermore, by targeting genomic regions rather than individual SNPs, the number of tests drops from millions to thousands, thus alleviating the multiple testing burden.

Here we take this next step and mine the publicly available TAG meta-analysis results by conducting pathway analyses and gene-based tests. Our aim is to identify biological pathways implicated in smoking behaviors. Our analyses, which are not limited to the most significant loci identified by the TAG meta-analysis, serve to provide additional insights into the biological bases of smoking behaviors.

Materials and Methods

The Tobacco and Genetics Consortium Sample

The published TAG meta-analysis results (downloaded from the Psychiatric Genomics Consortium website <https://www.med.unc.edu/pgc/downloads>) combined the SNP-based summary statistics of 16 participating cohorts including individuals of European ancestry from Europe and United States. The TAG sample has been described in detail elsewhere (see Table 1 and the Supplementary Information in¹³). In short, the samples were collected mainly between 1970 and 2006. Across the participating studies, the mean age ranged from 39.6 to 72.3 years. In the total combined sample, 64.37% of the participants were females. Data on smoking status (ever/never smoked regularly) were available for N=69,409 individuals. Of these, 39,022 individuals declared to have smoked regularly (i.e., a regular smoker - according to the Center for Disease Control's definition - declared to have smoked

at least 100 cigarettes during the lifetime). Data on quantity smoked were available in N=38,181 ever regular smokers, measured either as the average or the maximum number of cigarettes per day (two measures highly correlated over time (>0.71) in longitudinally followed samples¹⁷). The mean quantity smoked per day varied between 13.1 and 23.4 cigarettes across the participating studies. Data on smoking cessation were available in N=35,845 individuals. Of these, 20,619 declared to have quit regular smoking (see above for the ‘ever regular smoker’ definition) at least 1 year prior to the time of the assessment, while the remaining participants were current smokers. Age at initiation, defined as when first smoked or age when started to smoke regularly (the genetic correlation between these two definitions exceeds .97, see ¹⁷) was observed in N=22,438 individuals. The average value as observed in the participating samples ranged from 16.4 to 32.2 years of age. Multidimensional scaling ¹⁸, STRUCTURE ¹⁹ or EigenSoft principal component analysis ²⁰ were used to address population stratification prior to the study specific GWAS. Genomic control was further used to correct for potential effects of residual inflation ¹³. All the subjects from all studies participating in the TAG consortium provided informed consent.

Statistical analysis

We lifted the SNP positions from HapMap 2 NCBI build 36, Human Genome references 18 to the Human Genome references 19 positions based on dbSNP build 138. Following conversion, there were meta-analysis results available for quantity smoked (2,342,540 SNPs), ever smoking (2,339,474 SNPs), smoking cessation (2,340,171 SNPs) and age at initiation (2,341,140 SNPs). The SNPs passed the TAG consortium's quality criteria (see Supplementary Table 3 in ¹³).

Analyses were carried out using the HYbrid Set-based Test (HYST) as implemented in the Knowledge-based mining system for Genome-wide Genetic studies software ^{14, 16}. In the HYST approach, the SNPs within the tested set (i.e., where the set is a gene or a pathway, consisting of SNPs in linkage disequilibrium (LD)) are firstly partitioned into n weakly correlated blocks based on the LD information. Next, for each of these n LD blocks, a p-value (P_{block}) is obtained with the Gene-based Association Test that uses the Extended Simes procedure (GATES)¹⁴; that is, for each i th LD-block ($i=1..n$), GATES (a) orders ascendingly the p-values ($p_{(1)}..p_{(m)}$) of the m SNPs assigned to that block, (b) estimates the effective number of independent SNPs m_e and $m_{\alpha(j)}$, among the m SNPs and among the top j SNPs (i.e., key SNPs), respectively, by taking account of the LD structure, and (c) computes

the block-based p-value as $P_{block} = \min(\frac{m_e p_{(j)}}{m_{\alpha(j)}})$. Lastly, HYST uses the scaled chi-square test to combine the n LD block-based p-values as to obtain the overall p-value for the gene/pathway ¹⁶. The LD blocks were inferred ¹⁴ based on the 1000 Genomes Haplotypes Phase 3 reference panel ALL (downloaded from <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>). We opted for the 1000 Genomes as this is the highest resolution genetic map to date, derived based on a powerful design and utilizing the latent genomic information²¹. Furthermore, the use of this map will facilitate any follow-up analyses in the current GWAS samples as these often include imputed data based on the 1000 Genomes reference panel. Both the gene-based tests and the pathway analyses were conducted genome-wide and utilized the unweighted form of HYST. Variants were assigned to at least one of the ~24,100

genes (or within a region extended 5 kb at the 5' and at the 3' end of the gene) based on the UCSC RefGene (hg19) coordinates. Variants missing LD information were discarded. For the pathway analysis, we considered 1320 canonical pathways collected from the Molecular Signature Database ²² (manually curated pathways originating from Reactome ²³, KEGG ²⁴, and BioCarta ²⁵ databases), and restricted the analysis to pathways comprising 10 to 300 genes. We chose a genome-wide significance level of 0.05, and used the Benjamini and Hochberg (BH) method ²⁶ to correct the gene- and pathway based p-values for multiple testing. The BH method is expected to show better power relative to a Bonferroni correction, which is conservative when the number of tests is large and the tests are correlated ^{26, 27}.

Results

The gene-based analyses revealed twenty-one genes significantly associated with smoking behaviours, including eight genes in the 15q25 region. Of the twenty-one genes, fifteen genes were not previously reported as significantly associated with any smoking or addiction phenotype in the GWAS catalogue (as of 2015-10-23, see www.genome.gov/gwastudies ²⁸), and were not detected with the SNP-based approach in the original TAG analysis. One gene was associated with both quantity smoked and smoking cessation. Results are displayed in Table 1.

Fourteen genes passed the corrected for multiple testing False Discovery Rate (FDR) of 0.05 ²⁹ in the gene-based analysis of quantity smoked (see Table 1a), with the *IREB2* gene on 15q25 being the top association ($P=1.57E-37$).

Three genomic loci exceeded genome-wide significance in the ever smoking analysis (Table 1b). The top signal was yielded by the noncoding antisense RNA transcript *BDNF-AS* ($P=6.25E-07$) on 11p14. Note also that the *CDC27* gene, harbouring the rs16941640 SNP with the lowest p-value in the initial TAG analysis ($P=2.2E-07$), achieved significance in our gene-based analysis ($P=3.04E-06$). Two genes were significantly associated with smoking cessation (see Table 1c), with the *SLC25A21* gene on 14q13 yielding the lowest p-value ($P=2.09E-08$). The gene-based tests revealed the non-coding RNA *MIR1323-MIR512-1-MIR512-2* locus on 19q13.42 as significantly associated with age at initiation ($P=1.33E-06$; Table 1d). Note that no individual SNP reached significance in the original TAG analysis of age at initiation. We refer to the Supplementary Table S1 for details on gene functions and to the Tables S2-S5 for details on SNPs assigned to genes.

Results for the statistically significant pathways are summarized in Table 2. Details on genes assigned to the pathways are given in Supplementary Tables S6 and S7. The reported p-values passed the (corrected for multiple testing) FDR of 0.05.

As indicated in Table 2, the pathways reaching statistical significance form chains of pathways, rather than being isolated pathways spread across the genome. We identified several biological pathways significantly associated with quantity smoked (see Table 2a). Highly enriched are the neuronal system pathways harbouring the nicotinic acetylcholine receptor genes expressing the α (*CHRNA* 1-9), β (*CHRNA* 1-4), γ , δ and ϵ subunits; the three most strongly associated pathways were the highly calcium permeable postsynaptic

nicotinic acetylcholine receptors pathway ($P=4.90E-42$), the acetylcholine binding and downstream events ($P=3.11E-41$) and the presynaptic nicotinic acetylcholine receptors ($P=8.42E-32$) pathway. Note that some of the genes are assigned to more than one pathway, indicating that they play roles in multiple biological processes underlying the smoking behaviours. For instance, the *CHRNA3/5* gene cluster on 15q25 (see Supplemental Table S6), was assigned to all three top pathways in the quantity smoked analysis. Excluding this gene cluster from the analysis showed that while the highly calcium permeable postsynaptic nicotinic acetylcholine receptors and the acetylcholine binding and downstream events pathways remained highly significant (i.e., $P=1.27E-22$ and $P=1.18E-21$, respectively), the p-value for the presynaptic nicotinic acetylcholine receptors dropped to $P=1.10E-04$, suggesting that the association with this pathway was driven in part by the inclusion of these well-established candidate loci. Furthermore, quantity smoked was statistically associated with pathways regulating the immune system (with the cross presentation of soluble exogenous antigens endosomes pathway yielding the strongest association; $P=4.38E-17$), metabolism (where the regulation of ornithine decarboxylase ODC pathway yielded the strongest signal; $P=6.71E-16$), signal transduction (with the signaling by WNT pathway exceeding the significance threshold; $P=1.18E-17$) and asthma pathways ($P=3.10E-04$). Additionally, particularly enriched are the cell-cycle checkpoints pathways governing the transition of the new cell from one stage to another and the programmed cell-death, with strong association signals coming from the P53-dependent G1 DNA damage response pathway ($P=7.24E-20$) and from the pathway regulating apoptosis ($P=1.84E-09$).

Note that the cell-cycle pathways (displayed in italics in Table 2) are implicated also in ever smoking. These pathways relate to the Anaphase-Promoting Complex/Cyclosome-mediated degradation of Cdc20 and other APC/C^{Cdh1} targeted proteins. Although none of these pathways is associated with both smoking phenotypes, they are inter-connected, forming chains of pathways governing different stages of cell division. Especially enriched in ever smokers is the conversion from APC^{Cdc20} to APC^{Cdh1} in late anaphase pathway ($P=1.61E-07$), while in the quantity smoked analysis, enrichment signals come from the autodegradation of Cdh1 by Cdh1:APC/C pathway ($P=4.28E-17$) and from the APC/C^{Cdh1}-mediated degradation of Cdc20 and other APC/C^{Cdh1}-targeted proteins in late mitosis/early G1 pathway ($P=1.99E-16$). We identified no significant pathways for age at initiation or smoking cessation.

Discussion

We have extended the original TAG analysis and the previous work that has implicated SNPs and genes in smoking behaviours. Our results illustrate the utility of secondary analyses of published, publicly-available datasets, and demonstrate that following-up the SNP-based tests by gene- and pathway analyses can lead to considerable gains in statistical power and can yield important insights into the biological mechanisms underlying the trait of interest. The gene-based analysis revealed twenty-one genes implicated in smoking behaviours. Of these, fifteen are novel and were missed with the SNP-based approach in the original TAG analysis. Aside from the known cluster of genes (*IREB2-CHRNA3-CHRNA5-CHRNA4-HYKK-PSMA4*), we identified a cluster of three loci on the same chromosome 15 (*ADAMTS7, MORF4L1 and LOC646938*, $P=4.48E-17$, $1.18E-06$ and $1.66E-05$,

respectively). The *ADAMTS7-MORF4L1* locus has been previously associated with e.g., coronary artery disease³⁰ for which smoking is a known risk factor. Based on a joint-analysis, SNPs within *ADAMTS7* and *MORF4L1* were recently listed as candidate signals for smoking behaviour independent of those yielded by the known loci rs16969968 or rs588765 in *CHRNA5*³¹. Regarding the two significant associations on chromosome 1 (i.e., *SHISA4* and *IPO9* genes, $P=1.52E-06$ and $5.45E-06$, respectively), we note that neither of these have been previously associated with any addiction behaviour.

We also located statistically significant signal coming from the known HLA locus on chromosome 6, result suggesting a link between smoking and the immune system. Loci in the HLA region have been previously implicated in e.g., schizophrenia³², with which smoking dependence shares genetic vulnerabilities. Three genomic loci were significantly associated with ever smoking. The top signal came from the noncoding antisense RNA transcript *BDNF-AS* locus ($P=6.25E-07$) on chromosome 11. *BDNF-AS* downregulates the *BDNF* gene³³ which in turn has a key role in regulating neuronal growth and synaptic plasticity^{33,34}. The *BDNF* gene did not reach genome-wide significance in our gene-based analysis ($P=1.3E-04$), result indicating that the eight significant SNPs reported in the combined TAG analysis were tagging better the overlapping non-coding RNA *BDNF-AS* locus (see Supplementary Figures³⁵ for LD information).

Two genes exceeded genome-wide significance with smoking cessation: the *SLC25A21* gene on chromosome 14 and with the *SEMA6D* gene on chromosome 15 ($P=1.64E-06$). The role these genes play in smoking cessation and their potential as therapeutic targets has yet to be elucidated. *SEMA6D* expression levels were recently shown to significantly predict survival in patients with breast cancer³⁶; also the gene was listed among potential targets in cancer therapy owing to its role in regulation of the immune response³⁷. The *SLC25A21* gene was previously proposed as a plausible candidate for smoking cessation³⁸ and current smoking³⁹. Notice that the *SLC25A21* gene was the only one significantly associated with two smoking behaviours, namely with quantity smoked ($P=1.41E-05$) and smoking cessation ($P=2.09E-08$), result consistent with the large genetic correlation between the two smoking behaviours ($r_g = -.62$; $P<.001$)⁴⁰ observed in this sample.

The 19q13 *MIR1323-MIR512-1-MIR512-2* noncoding RNA locus exceeded genome-wide significance in the age at initiation analysis ($P=1.33E-06$). Importantly, loci in the 19q13 region were previously implicated in quantity smoked by the TAG Consortium in their combined analysis and also by the ENGAGE¹¹ analysis and, recently, by Loukola et al.⁴¹; the region has been proposed as a plausible candidate for further investigation in relation to smoking quantity given its role in nicotine metabolism (see e.g.,¹³). We now add age at initiation to the list of smoking phenotypes to be further investigated in relation to this biologically plausible region.

By extending the tested genomic region from genes to pathways - where a pathway is a 'meta-gene' comprising multiple genes having the same biological function - we identified chains of pathways statistically associated with ever smoking and quantity smoked (see Supplementary Figures⁴²). The strongest association signal in the quantity smoked analysis was yielded by the highly calcium permeable postsynaptic nicotinic acetylcholine receptors

pathway ($P=4.90E-42$) harboring the nicotinic acetylcholine receptors. The results are consistent with the hypothesis that mechanisms underlying smoking dependence involve the mesocorticolimbic dopamine system^{10, 43}. In short, as Benowitz described⁴⁴, the biological mechanisms underlying nicotine addiction involve nicotine binding to the nicotinic acetylcholine receptors which, in turn, release several neurotransmitters (dopamine, glutamate and γ -aminobutyric acid) in regions of the brain known to be involved in the perception of pleasure and reward (i.e., in the ventral tegmental area and the shell of the nucleus accumbens). Following repeated exposure, the nicotinic acetylcholine receptors adapt to nicotine and become unresponsive. It is hypothesized that the reactivation of these closed receptors following abstinence/cessation gives rise to symptoms of craving and withdrawal which, in turn, reinforce continuing smoking/relapse. We implicated in quantity smoked and ever smoking pathways relating to the Anaphase-Promoting Complex/Cyclosome mediated degradation of Cdc20 and other APC/C^{Cdh1}-targeted proteins. Although none of these pathways was associated with both smoking phenotypes, they are inter-connected, forming chains of pathways governing different stages of cell division. These pathways control not only the mitotic regulators of DNA replication (i.e., APC^{Cdc20}) but also axon growth and synaptic plasticity (i.e., APC^{Cdh1})⁴⁵; hence our results lend support to the idea that neuronal plasticity and learning play a paramount role in the development of nicotine addiction⁴³. Furthermore, for the cell-cycle pathways are known to belong to ‘a subway map of cancer pathways’⁴⁶ given their role in cancer development (a disease of unregulated cell proliferation), our results suggest that some of the same biological mechanisms underlie both smoking and cancer, as first conjectured by Fisher⁴⁷: “[...] *cigarette smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype*”. While the mediation study by VanderWeele et al.⁴⁸ demonstrated that variants at the 15q25 locus have a direct effect on both smoking and lung cancer, it will be of interest to determine whether their conclusion generalizes at the pathway level as suggested by the present results. If that is the case, the novel cancer therapies targeting the cell-cycle regulators⁴⁹ might yield clues of particular importance for the development of efficient smoking cessation therapy.

As detailed in the Supplementary Tables S6 and S7, our results provide clues on many other genes sharing the same pathways, genes that act in concert to give rise to individual differences in smoking behaviours. The genes and pathways reported herein worth further investigation in relation to other addiction phenotypes and disease traits such as depression⁵⁰, schizophrenia⁵¹ or cancers^{48, 52} with which smoking shares genetic factors⁵³.

Of the existing algorithms⁵⁴⁻⁵⁷ we chose HYST to conduct the pathway analyses because: (a) it has adequate control of the type I error rate, (b) good power (i.e., it is at least as powerful as - and in some circumstances - more powerful than alternative set-based tests^{14, 16}), (c) it is computationally efficient (i.e., it does not rely on permutations/simulations to account for LD between the SNPs when computing the p-values), (d) it exploits all the available information (i.e., the SNP p-values and their LD structure) to compute the pathway p-value, and thus does not require arbitrarily chosen p-value thresholds for selecting the list of input genes, and (e) importantly, it uses up to date annotation information on variants, genes and pathways from bioinformatics databases. Yet, despite the differences in terms of assumptions and the algorithm's power, we note that in large samples as the one we used,

results obtained based on alternative pathway-based approaches converge⁵⁶ (see also Supplementary Tables S8 and S9 for results obtained using an alternative overrepresentation test as implemented in Reactome²³). In view of this and to avoid the loss in power associated with increasing the number of tests, we limited our analysis strategy to HYST.

As it is based on gene-level statistics obtained in a meta-analytic sample, our study overcomes some limitations of previous exploratory pathway studies^{10, 53, 58} such as e.g., the literature selection bias, the bias arising from the use of incongruent analysis protocols or the heterogeneity problem^{56, 59, 60} arising from incongruent definition, assessment, and/or distributional properties of the phenotype in question across the selected studies. More importantly, we (and, previously, Vink et al. in³⁹) use a hypothesis-free/unbiased genome-wide approach in deriving the list of input genes to be included in the pathway enrichment analyses. In so doing, our study surmounts the ‘circularity’ bias of previous pathway studies^{10, 58} built on lists of input genes mostly derived from published confirmatory/candidate gene studies. Furthermore, by being based on HYST¹⁴ – which employs the scaled chi-square test to combine the LD-block-based p-values calculated with the GATES procedure – the study presented herein overcomes the limitations of other pathway-based tests such as e.g., the overrepresentation tests which do not take into account the correlation structure among the genes within pathways and among the SNPs within genes, or require arbitrary p-value thresholds to select the list of input genes.

We note the following limitations of the current study. First, the analyses were limited to the main effects of the genetic variants on smoking behaviors. It is possible that the aggregate effect of genetic variants within the significant pathways on smoking behaviors depends on exposure to environmental risk factors. Such gene-environment interaction remains to be studied, but will become feasible in the coming years, as the requisite large samples with raw genotype data and harmonized (across cohorts) measures of environmental exposures become available⁶¹. Second, HYST identified several genes on chromosome 15q25 as significantly associated with quantity smoked. Aside from the known cluster of genes (*IREB2-CHRNA3-CHRNA5-CHRNA4-HYKK-PSMA4*), we identified a cluster of three genes on the same chromosome 15 (*ADAMTS7*, *MORF4L1* and *LOC646938*, $P=4.48E-17$, $1.18E-06$ and $1.66E-05$, respectively). We note that based on a joint-analysis, SNPs within *ADAMTS7* and *MORF4L1* were recently listed as candidate signals for smoking behaviour independent of those yielded by the known loci rs16969968 or rs588765 in *CHRNA5*³¹. Further zooming into this genomic region through SNP conditional analysis and molecular experiments is desirable to completely disentangle which of these signals are independent, as well as for gene function validation¹⁶. Third, this study is based on summary statistics obtained by the 16 TAG participating groups in GWASs, which controlled for the effects of age and sex. It is a limitation of the present study that we could not consider other covariates, which may have an effect on the phenotypes of interests or on the genetic effects on the present phenotypes. Possible covariates include cohort, the availability of tobacco, and social pressure to start/quit smoking. Finally, we emphasize that the present results are based on a sample limited to individuals of European ancestry. Whether the biological pathway and genes identified here will replicate in populations of other ancestries remains to be established.

In conclusion, by running gene and pathway analyses in the Tobacco and Genetics Consortium sample, we found twenty-one genes and several biological pathways implicated in several smoking behaviors. Using the largest sample amassed yet in a GWAS of smoking we found that aside from the nicotinic acetylcholine receptors - known for the rewarding role they play in nicotine dependence, the cell-cycle regulators are possible targets in smoking cessation therapy. Our results shed light on one of the world's leading causes of preventable death and open a path to potential targets for therapeutics. These results are informative in decoding the biological bases of the addiction phenotypes and disease traits such as schizophrenia, depression and cancers with which smoking shares risk loci and biological pathways.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank to the reviewers for their valuable comments on the manuscript. We thank to the Tobacco and Genetics Consortium participants whose data we analyzed in this study. We also thank the Tobacco and Genetics Consortium, to all the researchers who contributed and lead this consortium. C.C.M., H.M. and J.M.V. are supported by the ERC starting grant 284167.

References

1. Organisation WH. 2015. http://www.who.int/tobacco/health_priority/en/
2. Ng M, Freeman MK, Fleming TD, Robinson M, Dwyer-Lindgren L, Thomson B, et al. Smoking prevalence and cigarette consumption in 187 countries, 1980-2012. *Jama*. 2014; 311(2):183-192. [PubMed: 24399557]
3. Giovino GA, Mirza SA, Samet JM, Gupta PC, Jarvis MJ, Bhala N, et al. Tobacco use in 3 billion individuals from 16 countries: an analysis of nationally representative cross-sectional household surveys. *The Lancet*. 2012; 380(9842):668-679.
4. Amos CI, Spitz MR, Cinciripini P. Chipping away at the genetics of smoking behavior. *Nat Genet*. 2010; 42(5):366-368. [PubMed: 20428092]
5. Vink JM, Willemsen G, Boomsma DI. Heritability of smoking initiation and nicotine dependence. *Behavior Genetics*. 2005; 35(4):397-406. [PubMed: 15971021]
6. Lubke GH, Hottenga JJ, Walters R, Laurin C, de Geus EJ, Willemsen G, et al. Estimating the genetic variance of major depressive disorder due to all single nucleotide polymorphisms. *Biol Psychiatry*. 2012; 72(8):707-709. [PubMed: 22520966]
7. Xian H, Scherrer JF, Madden PA, Lyons MJ, Tsuang M, True WR, et al. The heritability of failed smoking cessation and nicotine withdrawal in twins who smoked and attempted to quit. *Nicotine & Tobacco Research*. 2003; 5(2):245-254. [PubMed: 12745498]
8. Bühler KM, Giné E, Echeverry-Alzate V, Calleja-Conde J, de Fonseca FR, López-Moreno JA. Common single nucleotide variants underlying drug addiction: more than a decade of research. *Addiction Biology*. 2015
9. Loukola A, Wedenoja J, Keskitalo-Vuokko K, Broms U, Korhonen T, Ripatti S, et al. Genome-wide association study on detailed profiles of smoking behavior and nicotine dependence in a twin sample. *Mol Psychiatry*. 2014; 19(5):615-624. [PubMed: 23752247]
10. Wang J, Li MD. Common and unique biological pathways associated with smoking initiation/progression, nicotine dependence, and smoking cessation. *Neuropsychopharmacology*. 2010; 35(3):702-719. [PubMed: 19890259]

11. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature genetics*. 2010; 42(5): 436–440. [PubMed: 20418889]
12. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. Sequence variants at CHRNA3-CHRNA6 and CYP2A6 affect smoking behavior. *Nature genetics*. 2010; 42(5):448–453. [PubMed: 20418888]
13. Tobacco, Consortium G. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*. 2010; 42(5):441–447. [PubMed: 20418890]
14. Li MX, Gui HS, Kwan Johnny S, Sham Pak C. GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. *American Journal of Human Genetics*. 2011; 88(3):283–293. [PubMed: 21397060]
15. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*. 2014; 15(5):335–346.
16. Li MX, Kwan JS, Sham PC. HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. *The American Journal of Human Genetics*. 2012; 91(3):478–488. [PubMed: 22958900]
17. Consortium TAG. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics*. 2010; 42(5):441–447. [PubMed: 20418890]
18. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449(7164):851–861. [PubMed: 17943122]
19. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics*. 2000; 155(2):945–959. [PubMed: 10835412]
20. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*. 2006; 38(8):904–909. [PubMed: 16862161]
21. Consortium GP. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012; 491(7422):56–65. [PubMed: 23128226]
22. Database MS. http://www.broadinstitute.org/gsea/msigdb/collection_details.jsp#CP
23. Reactome. <http://www.reactome.org/>
24. KEGG. <http://www.genome.jp/kegg/>
25. BioCarta. http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
26. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995; 57(1):289–300.
27. Laird, NM., Lange, C. *The fundamentals of modern statistical genetics*. Springer Science & Business Media; 2010.
28. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*. 2014; 42(D1):D1001–D1006. [PubMed: 24316577]
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995:289–300.
30. Consortium CADG. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nature genetics*. 2011; 43(4):339–344. [PubMed: 21378988]
31. Schwantes-An TH, Culverhouse R, Duan W, Ramnarine S, Rice JP, Saccone NL. Interpreting joint SNP analysis results: when are two distinct signals really two distinct signals? *Genet Epidemiol*. 2013; 37(3):301–309. [PubMed: 23404318]
32. Consortium SWGotPG. Biological insights from 108 schizophrenia-associated genetic loci. *Nature*. 2014; 511(7510):421–427. [PubMed: 25056061]

33. Lipovich L, Dachet F, Cai J, Bagla S, Balan K, Jia H, et al. Activity-dependent human brain coding/noncoding gene regulatory networks. *Genetics*. 2012; 192(3):1133–1148. [PubMed: 22960213]
34. Modarresi F, Faghihi MA, Lopez-Toledano MA, Fatemi RP, Magistri M, Brothers SP, et al. Inhibition of natural antisense transcripts in vivo results in gene-specific transcriptional upregulation. *Nature biotechnology*. 2012; 30(5):453–459.
35. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005; 21(2):263–265. [PubMed: 15297300]
36. Chen D, Li Y, Wang L, Jiao K. SEMA6D Expression and Patient Survival in Breast Invasive Carcinoma. *International Journal of Breast Cancer*. 2015; 2015:10.
37. Tamagnone L. Emerging Role of Semaphorins as Major Regulatory Signals and Potential Therapeutic Targets in Cancer. *Cancer Cell*. 22(2):145–152.
38. Uhl GR, Drgon T, Johnson C, Ramoni MF, Behm FM, Rose JE. Genome-wide association for smoking cessation success in a trial of precessation nicotine replacement. *Molecular medicine*. 2010; 16(11-12):513. [PubMed: 20811658]
39. Vink JM, Smit AB, de Geus EJ, Sullivan P, Willemsen G, Hottenga JJ, et al. Genome-wide association study of smoking initiation and current smoking. *The American Journal of Human Genetics*. 2009; 84(3):367–379. [PubMed: 19268276]
40. Nivard M, Verweij K, Minica C, Treur J, Consortium IC, Vink J, et al. Connecting the dots, genome wide association studies in substance use. *Molecular Psychiatry*. 2016 in press.
41. Loukola A, Buchwald J, Gupta R, Palviainen T, Hällfors J, Tikkanen E, et al. A Genome-Wide Association Study of a Biomarker of Nicotine Metabolism. *PLoS Genet*. 2015; 11(9):e1005498. [PubMed: 26407342]
42. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*. 2003; 13(11):2498–2504. [PubMed: 14597658]
43. Benowitz NL. Nicotine addiction. *N Engl J Med*. 2010; 362(24):2295. [PubMed: 20554984]
44. Benowitz NL. Nicotine addiction. *The New England journal of medicine*. 2010; 362(24):2295. [PubMed: 20554984]
45. Li M, Zhang P. The function of APC/CCdh1 in cell cycle and beyond. *Cell division*. 2009; 4(1):2. [PubMed: 19152694]
46. Hahn, WC., Weinberg, RA. A subway map of cancer pathways. *Nature Publishing Group*; 2002.
47. Fisher R. Cigarettes, cancer, and statistics. *Centennial Review of Arts & Science*. 1958:151–166.
48. VanderWeele TJ, Asomaning K, Tchetgen EJT, Han Y, Spitz MR, Shete S, et al. Genetic variants on 15q25. 1, smoking, and lung cancer: an assessment of mediation and interaction. *American journal of epidemiology*. 2012; 175(10):1013–1020. [PubMed: 22306564]
49. Diaz-Moralli S, Tarrado-Castellarnau M, Miranda A, Cascante M. Targeting cell cycle regulation in cancer therapy. *Pharmacology & therapeutics*. 2013; 138(2):255–271. [PubMed: 23356980]
50. Edwards AC, Kendler KS. A twin study of depression and nicotine dependence: shared liability or causal relationship? *Journal of affective disorders*. 2012; 142(1-3):90–97. [PubMed: 22901332]
51. Dani JA, Harris RA. Nicotine addiction and comorbidity with alcohol abuse and mental illness. *Nature neuroscience*. 2005; 8(11):1465–1470. [PubMed: 16251989]
52. Saccone NL, Culverhouse RC, Schwantes-An TH, Cannon DS, Chen X, Cichon S, et al. Multiple independent loci at chromosome 15q25. 1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet*. 2010; 6(8):e1001053. [PubMed: 20700436]
53. de Viron S, Morré SA, Van Oyen H, Brand A, Ouburg S. Genetic similarities between tobacco use disorder and related comorbidities: an exploratory study. *BMC medical genetics*. 2014; 15(1):85. [PubMed: 25060307]
54. Pers TH, Karjalainen JM, Chan Y, Westra HJ, Wood AR, Yang J, et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature communications*. 2015:6.
55. Segrè AV, Groop L, Mootha VK, Daly MJ, Altshuler D, Consortium D, et al. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet*. 2010; 6(8):e1001058. [PubMed: 20714348]

56. Network T, Consortium PASotPG. Psychiatric genome-wide association study analyses implicate neuronal, immune and histone pathways. *Nature neuroscience*. 2015; 18(2):199–209. [PubMed: 25599223]
57. Jin L, Zuo XY, Su WY, Zhao XL, Yuan MQ, Han LZ, et al. Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & bioinformatics*. 2014; 12(5):210–220.
58. Liu M, Fan R, Liu X, Cheng F, Wang J. Pathways and Networks-Based Analysis of Candidate Genes Associated with Nicotine Addiction. 2015
59. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human molecular genetics*. 2009; 18(11):2078–2090. [PubMed: 19286671]
60. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, Russell RK, et al. Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *The American Journal of Human Genetics*. 2009; 84(3):399–405. [PubMed: 19249008]
61. Vinkhuyzen A, Wray N. Novel directions for G×E analysis in psychiatry. *Epidemiology and psychiatric sciences*. 2015; 24(01):12–19. [PubMed: 25295644]
62. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nature genetics*. 2010; 42(5):436–440. [PubMed: 20418889]
63. Bloom AJ, Hartz SM, Baker TB, Chen LS, Piper ME, Fox L, et al. Beyond Cigarettes Per Day. A Genome-Wide Association Study of the Biomarker Carbon Monoxide. *Annals of the American Thoracic Society*. 2014; 11(7):1003–1010. [PubMed: 25072098]
64. David S, Hamidovic A, Chen G, Bergen A, Wessel J, Kasberger J, et al. Genome-wide meta-analyses of smoking behaviors in African Americans. *Translational psychiatry*. 2012; 2(5):e119.

Table 1

Results gene-based analyses for smoking behaviours. Variants were assigned to genes (or within a region extended 5 kb at the 5' and at the 3' end of the gene) based on the UCSC RefGene (hg19) coordinates.

a. genes implicated in quantity smoked in a sample of 38181 ever regular smokers							
Symbol	Nominal P	Corrected P	CHR	Start position	End position	Group	Reported associations*
<i>IREB2</i>	1.57E-37	3.79E-33	15q25	78730517	78793798	protein-coding gene	none
<i>CHRNA3</i>	6.45E-34	7.78E-30	15q25	78887646	78913637	protein-coding gene	Smoking behaviour ^{12,13,62} Nicotine dependence ¹² Exhaled carbon monoxide levels ⁶³
<i>CHRNA5</i>	3.17E-32	2.55E-28	15q25	78857861	78887611	protein-coding gene	Smoking behaviour ^{62,64} Nicotine dependence ¹²
<i>HYKK</i>	2.14E-30	1.29E-26	15q25	78799905	78829715	protein-coding gene	none
<i>CHRNA4</i>	3.87E-28	1.87E-24	15q25	78916635	78933587	protein-coding gene	Nicotine dependence ¹² Exhaled carbon monoxide levels ⁶³
<i>PSMA4</i>	6.47E-23	2.60E-19	15q25	78832746	78841563	protein-coding gene	Smoking behaviour ⁶⁴
<i>ADAMTS7</i>	4.48E-17	1.55E-13	15q25	79051544	79103773	protein-coding gene	none
<i>MORF4L1</i>	1.18E-06	3.55E-03	15q25	79165330	79190081	protein-coding gene	none
<i>SHISA4</i>	1.52E-06	4.07E-03	1q32	201857804	201861715	protein-coding gene	none
<i>IP09</i>	5.45E-06	1.32E-02	1q32	201798287	201853422	protein-coding gene	none
<i>HLA-DPB1</i>	7.65E-06	1.68E-02	6p21	33043702	33057473	protein-coding gene	none
<i>HLA-DPA1</i>	9.52E-06	1.92E-02	6p21	33032345	33048555	protein-coding gene	none
<i>SLC25A21</i>	1.41E-05	2.61E-02	14q13	37147125	37641865	protein-coding gene	none
<i>LOC646938</i>	1.66E-05	2.86E-02	15q25	79044378	79045734	Unknown	none
b. genes implicated in ever smoking in a sample of 69,409 individuals, of whom 39,022 were ever regular smokers, and the remaining 30,387 declared to have never smoked regularly							
Symbol	Nominal P	Corrected P	CHR	Start position	End position	Group	Reported associations*
<i>BDNF-A5</i>	6.25E-07	1.51E-02	11p14	27528398	27719718	non-coding RNA	Smoking behaviour ¹³
<i>APBB2</i>	2.43E-06	2.44E-02	4p13	40812043	41216635	protein-coding	none
<i>CDC27</i>	3.04E-06	2.44E-02	17q21	45195062	45266678	protein-coding gene	none
c. genes implicated in smoking cessation in a sample of 35,845 individuals of whom 20,619 declared to have quit smoking at least one year prior to the data collection and 15,226 were current smokers.							

a. genes implicated in quantity smoked in a sample of 38181 ever regular smokers							
Symbol	Nominal P	Corrected P	CHR	Start position	End position	Group	Reported associations *
Symbol	Nominal P	Corrected P	CHR	Start position	End position	Group	Reported associations *
<i>SLC25A21</i>	2.09E-08	5.04E-04	14q13	37147125	37641865	protein-coding gene	none
<i>SEMA6D</i>	1.64E-06	1.99E-02	15q21	48010685	48066420	protein-coding gene	none
d. genes implicated in age at initiation in a sample of 22,438 individuals for whom age when first smoked or age when started to smoke regularly were available.							
Symbol	Nominal P	Corrected P	CHR	Start position	End position	Group	Reported associations *
<i>MIR1323</i>	1.33E-06	1.21E-02	19q13	54175221	54175294	non-coding RNA	none
<i>MIR512-1</i>	1.50E-06	1.21E-02	19q13	54172416	54172500	non-coding RNA	none
<i>MIR512-2</i>	1.50E-06	1.21E-02	19q13	54172410	54172508	non-coding RNA	none

* Reported associations with addiction phenotypes (P < 5E-08) according to GWAS catalogue (release on 2015-10-23).

Table 2
Biological pathways implicated in

a. quantity smoked.		
Pathway	P (HYST)	Total genes
The neuronal system chain		
Highly calcium permeable postsynaptic nicotinic acetylcholine receptors	4.90E-42	11
Acetylcholine binding and downstream events	3.11E-41	14
Presynaptic nicotinic acetylcholine receptors	8.42E-32	12
Neuroactive ligand receptor interaction	2.66E-22	249
Transmission across chemical synapses	2.13E-20	174
Neurotransmitter receptor binding and downstream transmission in the postsynaptic cell	2.87E-20	127
Neuronal system	1.43E-19	263
The cell-cycle chain		
<i>P53-dependent G1 DNA damage response</i>	7.24E-20	52
<i>SCF-Skp2 mediated degradation of p27/p21</i>	7.59E-19	52
<i>Cyclin E associated events during G1/S transition</i>	2.35E-18	61
<i>Autodegradation of the E3 ubiquitin ligase COP1</i>	3.37E-18	46
<i>Proteasome</i>	5.95E-18	44
<i>Autodegradation of Cdh1 by Cdh1:APC/C</i>	4.28E-17	53
<i>CDK-mediated phosphorylation and removal of Cdc6</i>	5.70E-17	45
<i>Assembly of the pre-replicative complex</i>	6.46E-17	61
<i>APC/C^{Cdh1} mediated degradation of Cdc20 and other APC/C^{Cdh1} targeted proteins in late mitosis/early G1</i>	1.99E-16	61
<i>P53-independent G1/S DNA damage checkpoint</i>	2.02E-16	47
<i>CDT1 association with the Cdc6:ORC:origin complex</i>	2.03E-16	53
<i>ORC1 removal from chromatin</i>	3.31E-16	63
<i>SCF (beta-TrCP) mediated degradation of EM11</i>	6.24E-16	48
<i>APC/C^{Cdc20} mediated degradation of mitotic proteins</i>	4.53E-14	62
<i>Regulation of apoptosis</i>	1.84E-09	54
Disease		
Vif-mediated degradation of APOBEC3G	9.55E-18	48
Signal transduction		
Signaling by WNT	1.18E-17	61
The immune system chain		
Cross presentation of soluble exogenous antigens endosomes	4.38E-17	44
Activation of NF-kB in B cells	6.72E-14	60
ER phagosome pathway	7.12E-14	57
Translocation of ZAP-70 to immunological synapse	3.90E-04	13
Allograft rejection	7.50E-04	34
Phosphorylation of CD3 and TCR zeta chains	1.64E-03	15

a. quantity smoked.		
Pathway	P (HYST)	Total genes
Interferon signalling	1.70E-03	148
Metabolism		
Regulation of ornithine decarboxylase ODC	6.71E-16	47
Metabolism of amino acids and derivatives	1.80E-05	187
Gene expression		
Destabilization of mRNA by AUF1 (hnRNP D0)	2.21E-15	49
Asthma	3.10E-04	27
b) ever smoking		
Pathway	P(HYST)	Total genes
Regulation of the mitotic cell-cycle chain		
<i>Conversion from APC^{Cdc20} to APC^{Cdh1} in late anaphase</i>	1.61E-07	14
<i>Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle checkpoint components</i>	2.19E-07	16
<i>Phosphorylation of the APC/C</i>	4.54E-07	15
<i>APC^{Cdc20} mediated degradation of Cyclin B</i>	1.13E-06	17
<i>APC^{Cdc20} mediated degradation of Nek2A</i>	2.30E-06	19