INSIGHT INTO NICOTINE AND ALCOHOL USE THROUGH GENETIC ASSOCIATION META-ANALYSES

A DISSERTATION SUBMITTED TO THE FACULTY OF THE UNIVERISTY OF MINNESOTA BY

MENGZHEN LIU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

ADVISOR: SCOTT VRIEZE

DECEMBER 2019

Acknowledgements

I want to acknowledge the acknowledgements in the supplementary material of the two meta-analyses. Without the participants who have contributed the data and the study-level analysts and researchers who have provided and analyzed data, none of this would have been possible. I want to thank my collaborators at Pennsylvania State University who developed the methods and responded to my numerous questions. I also want to thank everyone at the Institute for Behavioral Genetics at the University of Colorado – Boulder where I spent the first 3 years. My cohort, the administration staff and professors there are invaluable to help me start on my journey as a graduate student in behavioral genetics. Similarly, I also want to thank the professors in Personality, Individual Differences and Behavioral Genetics at University of Minnesota where I spent the latter half of my PhD journey. Lastly, I want to thank the Vrieze lab where I've spent way too much time. I want to thank all my lab mates (past and present) for the times when we were stuck in the same office and building for over 12 hours a day trying to meet and finish unreasonable requests. I want to thank Scott Vrieze for accepting me as a graduate student, the support given to me over 5.5 years and the everything bagels.

I also want to thank both Krishna.human and Krishna.cat for their love and support.

Table of Contents

List of Tables……..iii

List of Figures……..iv

Notes…….v

Introduction……..1

Chapter 1 - GWAS Meta-Analysis of Alcohol and Nicotine Use in 1.2M…….…..…5

Chapter 2 – Exome Meta-analysis of Alcohol and Nicotine Use in 622,409 individuals……..….16

Chapter 3 – Predicting Endophenotypes with Alcohol and Nicotine use PRS ……..39

Discussion.…….…….51

Conclusion……..…….53

List of Tables

Table 1 Nonsynonymous sentinel variants……...…..14

Table 2 Association results for SNVs identified in single variant association meta-analyses and taken forward to replication are provided…….……. 23

Table 3 Association results for novel SNVs identified in the combined meta-analysis of the discovery and replication cohorts ……..27

Table 4 Results from conditional analyses at previously reported smoking behavior loci ……33

Table 5 Summary of endophenotypes…….49

Table 6 Heritability calculated using GCTA based on the whole sample……..50

List of Figures

Figure 1 Genetic correlations between substance use phenotypes and phenotypes from other large GWAS……...6

Figure 2 Multivariate analysis of pleiotropy…….7

Figure 3 Heritability and polygenic prediction…….8

Figure 4 Correlations among exemplary DEPICT gene sets…...……...11

Figure 5 Study design including the discovery and replication stages……...17

Figure 6 A concentric Circos plot of the association results ……..25

Figure 7 Correlation between PRS…….45

Figure 8 Correlation between PRS and endophenotype…….47

Notes

Chapter 1 of the dissertation has already been published. Slight edits have been made to fit the format of the dissertation [*asterisk denotes equal contribution, i.e. joint first-authorship]:

Liu, M.*, Jiang, Y.*, Wedow, R.*, Li, Y.*, Brazel, D., Chen, F., … & Vrieze, S. (2018) Association studies of up to 1.2 million individuals yield new insights in the genetic etiology of tobacco and alcohol use. Accepted at *Nature Genetics*.

Contributions

G.A., D.J.L., and S.V. designed the study. D.J.L. and S.V. led and oversaw the study. M.L. was the study's lead analyst. She was assisted by Y.J., D.J.L., S.V., R.W., D.M.B., and G.D. Bonferroni thresholds were calculated by D.M. Phenotype definitions were developed by L.J.B., M.C.C., D.A.H., J.K., E.J., D.J.L., M.M., M.R.M., S.V., and L.Z. Software development was carried out by Y.J., D.J.L., and X.Z. Conditional analyses were performed by Y.J. and M.L.. Heritability, genetic correlation, and polygenic scoring analyses were performed by R.W. Multivariate analyses were performed by Y.J., M.L., and D.J.L. Bioinformatics analyses were performed and interpreted by F. Chen, J.D., J.J.L., Y. Li, M.L., J. A. Stitzel, S.V., and R.W. The LocusZoom website was designed by G.D. Figures were created by M.L., R.W., Y. Li, and S.V. M.A.E. and M.C.K. helped with data access. R.W. coordinated authorship and acknowledgement details. M.C.C., S.P.D., E.J., J.K., and J. A. Stitzel provided helpful advice and feedback on study design and the manuscript. All authors contributed to and critically reviewed the manuscript. Y. Li, D.J.L., M.L., S.V., and R.W. made major contributions to the writing and editing.

Chapter 2 of the dissertation has already been published. Slight edits have been made to fit the format of the dissertation [*asterisk denotes equal contribution, i.e. joint first-authorship]:

Erzurumluoglu, M.*, Liu, M.*, Jackson, V.*, Barnes, D., Datta, G., Melbourne, C., … & Liu, D. (2018) Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behavior associated genetic loci. Accepted at *Molecular Psychiatry*.

Contributions

These authors contributed equally and share the first author position: A. Mesut Erzurumluoglu, Mengzhen Liu, Victoria E. Jackson

These authors contributed equally and share the last author position: Martin D. Tobin, Scott Vrieze, Dajiang J. Liu, Joanna M. M. Howson.

Introduction

Since the Surgeon General's report in 1964, cigarette use has been declining in the US and in most high-income countries across the rest of the world (Antman, Arnett, Jessup, & Sherwin, 2014; Spanagel, 2017) (M. Ng et al., 2014) (Jha & Peto, 2014). There have been multiple studies of cigarette smoking as the primary risk factor for many diseases including lung cancer and heart disease, contributing to 5 million deaths globally (Jha & Peto, 2014). The risk of cigarette smoking is not limited to regular smokers alone, a higher risk of cancer and cardiovascular disease also exists for intermittent smokers (Schane, Ling, & Glantz, 2010). Moreover, the effects of secondhand smoke are also a public health concern since mere exposure has been linked to an increased risk in lung cancer and stroke (Kim, Ko, Kwon, & Lee, 2018). Children are particularly vulnerable because they breathe at a faster rate than adults and secondhand smoke has been associated with increased rates of bacterial infections, acute respiratory illness and rates of hospitalization of asthma attacks (Cao, Yang, Gan, & Lu, 2015; Z. Wang et al., 2015).

The only other substance to approximate the public health burden of tobacco use is alcohol use. There are an estimated 2.4 billion people across the globe who use alcohol (Gowing et al., 2015; Griswold et al., 2018; Jha & Peto, 2014). Alcohol use has been clearly linked to risk for a wide variety of diseases (e.g., liver cirrhosis), but also to unintentional injuries such as traffic accidents and falls (Griswold et al., 2018; Rehm, 2011).

Twin studies routinely find that nicotine and alcohol use are heritable (Polderman et al., 2015) with ~50% of the phenotypic variance being accounted for by additive genetic effects. Despite this substantial heritability, prior to 2019 only a handful of specific genetic variants or genomic regions have been reliably found to be associated with substance use or dependence. The decreasing cost of array genotyping and genome sequencing since the mid-2000s (Mardis, 2011) has led to a marked increase in the number of studies that use these technologies to study genetic associations for complex traits and disease.

The standard analytical approach for gene-disease mapping has become the genome-wide association study (GWAS). Very simply, GWAS is a series of correlations between individuals' genotypes, most commonly single-nucleotide polymorphisms (SNPs), with the phenotype of interest. GWAS and GWAS meta-analyses have successfully found several functional, and potentially causal variants associated with substance dependence (Laura J. Bierut et al., 2010; Hancock et al., 2018; Walters et al., 2018). However, these variants account for a tiny fraction of the phenotypic variation, prompting the conclusion that behavioral phenotypes are highly polygenic; many variants, each of small effect, work in conjunction to influence the phenotype. In fact, the effect of any single variant is so small that very large study samples are necessary to detect them (Visscher et al., 2017). Alcohol and nicotine dependence are clinically relevant phenotypes and tend to have higher heritability estimates than measures of consumption

(Verhulst, Neale, & Kendler, 2015; Vink, Willemsen, & Boomsma, 2005) which would make them more ideal GWAS candidate phenotypes. However, there is substantial difficulty in achieving the desired sample size in substance abuse as cases would require clinical diagnosis and further work would be needed to find a suitable control (Dick, Meyers, Rose, Kaprio, & Kendler, 2011). It is more practical to work with simple substance use phenotypes that are regularly collected in biomedical studies using short survey questions ("Do you smoke regularly?" or "How many drinks per week do you typically consume?") and are common in medical records as part of regular health check-ups or hospital intakes. There have been GWAS meta-analyses of alcohol and nicotine use that have found several significant loci (Schumann et al., 2016; Tobacco and Genetics Consortium, 2010) and we aim to increase the sample size further in order to capture more substance use associated variants.

In the first chapter, we used GWAS meta-analysis to discover common variants (variants with allele frequency > 0.1%) that are associated with alcohol and nicotine use. It is the largest GWAS meta-analysis of alcohol and nicotine use to date combining summary statistics from over 30 GWASs and reached over 1.2 million participants of European descent.

For nicotine use, we examined cigarette smoking from initiation to cessation. The four phenotypes are

- Smoking Initiation: Binary phenotype on whether the participants have ever been a regular smoker (also commonly defined as having smoked more than 100 cigarettes). 2 coded as regular smoker and 1 as never a regular smoker.
- Age of initiation: Quantitative phenotype on when the participants started regularly smoking. Individuals who are not regular smokers were set to missing.
- Cigarettes per Day: Binned phenotype (1-5) on how many cigarettes smoked per day. Individuals who are not regular smokers were set to missing.
- Smoking cessation: Binary phenotype on whether the participants is a current or former regular smoker. 2 coded as current and 1 as former. Individuals who are not regular smokers were set to missing.

We had one alcohol use phenotype which measures heaviness of use.

- Drinks per week: Quantitative phenotype on how many alcoholic drinks per week they consume. Studies were asked to left-anchor and log transform this phenotype.

We discovered 566 conditionally independent variants in 406 loci associated with nicotine and alcohol use. Using these results, we performed cell, tissue, gene-set, and pathway enrichment analyses on each set of meta-analysis results to understand the specific biological mechanisms of those traits. An advantage to including both alcohol and nicotine use phenotypes is that we can

jointly explore the results for any common variants that may contribute to a more general substance use factor. Alcohol and nicotine use are highly comorbid behaviors (Meyerhoff et al., 2006) so there may be common variants that are affecting both substances pleiotropically. We examined the genetic correlations between the phenotypes and did a pleiotropy analysis to see if any genes overlap across the five traits. Lastly, to see the utility of these results, we calculated polygenic risk scores (PRS) with the meta-analyses results and it significantly predicted the same phenotypes in two other independent samples.

Previous large-scale GWAS meta-analysis of alcohol and nicotine use have found several significant loci (Schumann et al., 2016; Tobacco and Genetics Consortium, 2010) but the heritability derived from these SNPs are far from the heritability estimated from twin studies. SNP-based heritability from published GWAS meta-analysis results are generally under 10% (Zheng et al., 2017), much less than the 30-60% (Grant et al., 2009; Polderman et al., 2015; Verhulst et al., 2015; Vink et al., 2005) typically found in twin studies. The discrepancy between heritabilities estimated from twin studies and genotyped variants has been termed the “missing heritability” (Eichler et al., 2010; Gibson, 2012; Maher, 2008). One hypothesis is the effect of each individual variant is much smaller than previous expectations and we may need hundreds of thousands of individuals to detect them. Another common hypothesis concerns the genetic architecture underlying the trait where rare variants with large effects are what’s contributing to the missing heritability. There are examples of highly penetrant mendelian diseases that are due to low frequency variants such as cystic fibrosis, therefore, it stands to reason that the same may be true for behavioral traits as well. From an evolutionary theory perspective, if the variant has a large deleterious effect then it is expected to be selected against in a population and thus exists at a lower frequency (Gibson, 2012).

In the second chapter, we performed exome meta-analysis in parallel to the first chapter in order to find rare variants that may be associated specifically with nicotine use. We examined 4 nicotine use phenotypes:

- Smoking Initiation: Binary phenotype on whether the participants have ever been a regular smoker (also commonly defined as having smoked more than 100 cigarettes). 2 coded as regular smoker and 1 as never a regular smoker.
- Cigarettes per day (CPD; quantitative trait) average number of cigarettes smoked per day by ever smokers.
- 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 initiation to current age for current smokers or age at cessation for former smokers.

- Smoking cessation: Binary phenotype on whether the participants is a current or former regular smoker. 2 coded as current and 1 as former. Individuals who are not regular smokers were set to missing.

The exome-metanalysis was done simultaneously as the GWAS meta-analysis and found 40 common loci (also implicated in the GWAS meta-analysis) associated with nicotine use but no conclusive rare variant associations. We also checked for conditionally independent rare variants within previously associated loci and found one low-frequency variant (allele frequency~1%). In order to characterize these loci, we queried the GWAS catalogue, QTL in GTEx V7, Brain xQT, and BRAINEAC and also performed pathway enrichment analysis. Lastly, we used mendelian randomization with our results and some key phenotypes associated with smoking. We found causal associations between smoking initiation and educational attainment.

In order to understand the mechanisms of these addictions, there have been many animal studies, most commonly mice, that model drug addiction from use to relapse (Lynch, Nicholson, Dance, Morgan, & Foley, 2010). The biology and chemistry of alcohol and nicotine have been studied extensively (Benowitz, Hukkanen, & Jacob, 2009; Cederbaum, 2012; Edenberg, 2007), yet there are still gaps in the knowledge of how and why there are individual differences in the metabolism of these substances. The underlying biology of substance metabolism may be common amongst mammalian species, but human-specific traits and behaviors are much harder to model and replicate in mice.

A common method to measure these underlying mechanisms in humans is to examine endophenotypes that are associated with the complex phenotype of interest. Endophenotypes are stable, simple, and heritable traits within individuals that are useful as measures associated with a more complex phenotype; some examples of endophenotypes are biomarkers such as cotinine and brain-based measures like electroencephalography. These endophenotypes are viewed as measures that are closer to acute underlying biological pathways or cognitive processes which may be expressed as part of the heterogeneity of a complex phenotype.

There have been studies linking alcohol use disorder and various brain-based endophenotype in the literature (Carlson, Iacono, & McGue, 2002; Malone, Iacono, & McGUE, 2001). In the third chapter, we associated the results from the imputed GWAS meta-analyses to these endophenotypes in order to understand its connection to substance use. None of the associations were significant after correcting for multiple tests.

Chapter 1

Tobacco and alcohol use are leading causes of mortality that influence risk for many complex diseases and disorders (Ezzati, Lopez, Rodgers, Vander Hoorn, & Murray, 2002). They are heritable (Hicks, Schalet, Malone, Iacono, & McGue, 2011; Polderman et al., 2015) and etiologically related (Kenneth S. Kendler, Prescott, Myers, & Neale, 2003; Kenneth S. Kendler, Schmitt, Aggen, & Prescott, 2008) behaviors that have been resistant to gene discovery efforts (Bierut et al., 2012; Jorgenson et al., 2017; Schumann et al., 2016; T. E. Thorgeirsson et al., 2016; Thorgeir E. Thorgeirsson et al., 2010; Tobacco and Genetics Consortium, 2010).

An analysis overview is provided in Supplementary Fig. 1; all independent associated variants are in Supplementary Tables 1–5; and quantile-quantile, Manhattan, and LocusZoom plots are shown in Supplementary Figs. 2–12. Smoking initiation phenotypes included age of initiation of regular smoking (AgeSmk; n=341,427; 10 associated variants) and a binary phenotype indicating whether an individual had ever smoked regularly (SmkInit; n=1,232,091; 378 associated variants). Heaviness of smoking was measured with cigarettes per day (CigDay; n=337,334; 55 associated variants). Smoking cessation (SmkCes; n=547,219; 24 associated variants) was a binary variable contrasting current versus former smokers. Available measures of alcohol use were simpler, with drinks per week (DrnkWk; n=941,280; 99 associated variants) widely available and similarly measured across studies. See the Supplementary Note and Supplementary Tables 6 and 7 for phenotype definition details. An analysis overview is provided in Supplementary Fig. 1; all independent associated variants are in Supplementary Tables 1–5; and quantile-quantile, Manhattan, and LocusZoom plots are shown in Supplementary Figs. 2–12.

The four smoking phenotypes were genetically correlated with one another (Fig. 1 and Supplementary Table 8). DrnkWk was not highly genetically correlated with the smoking phenotypes (rg~0.10) except for SmkInit (rg~0.34, $p=6.7\times10-63$), suggesting that sequence variations affecting alcohol use and those affecting initiation of smoking overlap substantially. The phenotypes were highly genetically correlated across constituent studies (Supplementary Table 9), suggesting a minor effect of phenotypic heterogeneity in the present results, even across Western Europe and the United States. Smoking phenotypes were genetically correlated in expected directions with many behavioral, psychiatric, and medical phenotypes (Fig. 1 and Supplementary Table 10). Genetic variation associated with increased alcohol use was associated with greater levels of risky behavior (rg=0.20, $p=1.8\times10-7$) and cannabis use(rg=0.36, $p=6.2\times10-10$), but with less risk of disease for almost all diseases (Fig. 1 and Supplementary Table 10).

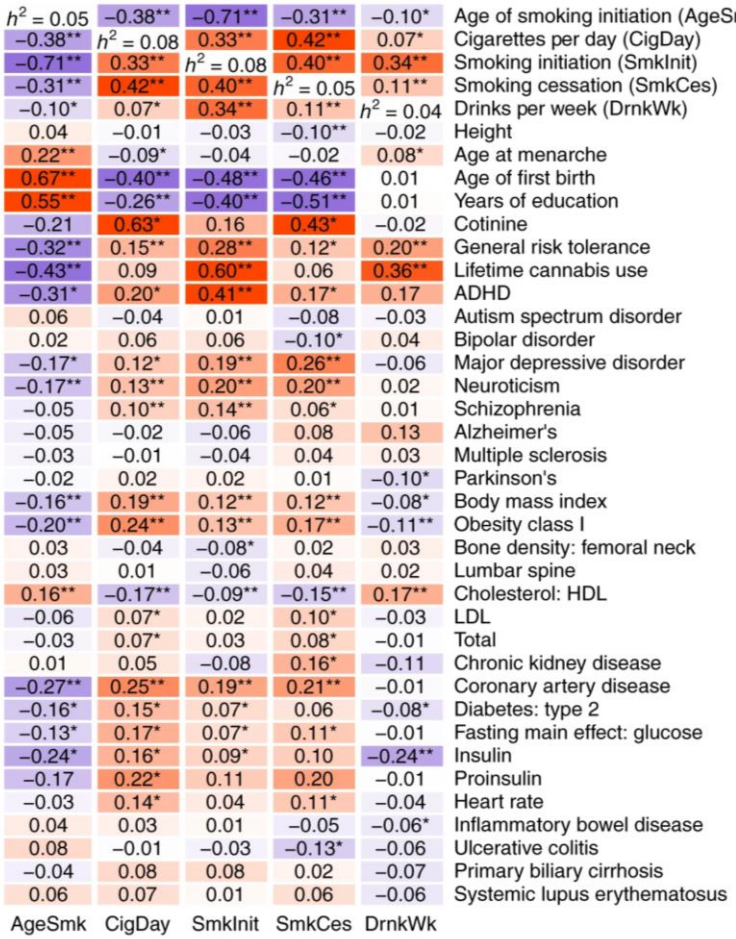

Fig. 1 Genetic correlations between substance use phenotypes and phenotypes from other large GWAS. Genetic correlations between each of the phenotypes are shown in the first five rows, with heritability estimates displayed down the diagonal. All genetic correlations and heritability estimates were calculated using LD score regression. Purple shading represents negative genetic correlations, and red shading represents positive correlations, with increasing color intensity reflecting increasing correlation strength. A single asterisk reflects a significant genetic correlation at the $P < 0.05$ level. Double asterisks reflect a significant genetic correlation at the Bonferroni-correction $P < 0.000278$ level (corrected for 180 independent tests). Note that SmkCes was oriented such that higher scores reflected current smoking, and for AgeSmk, lower scores reflect earlier ages of initiation, both of which are typically associated with negative outcomes

Using a novel method to evaluate multivariate genetic correlation at the locus (versus global) level, we observed 150 loci that affected multiple substance use phenotypes (Fig. 2 and Supplementary Table 11). Patterns of pleiotropy across phenotypes were highly diverse, with only three loci significantly associated with all five phenotypes. These three loci included associations implicating phosphodiesterase 4B (*PDE4B*) and cullin 3 (*CUL3*). *PDE4B* regulates cyclic AMP second messenger availability and thereby affects signal transduction, and it is downregulated by chronic nicotine administration in rats (Polesskaya, Smith, & Fryxell, 2007). *CUL3* has wide-

ranging effects, including on ubiquination and protein degradation, and de novo mutations in *CUL3* are associated with rare diseases affecting response to the mineralocorticoid aldosterone (Boyden et al., 2012), which itself is affected by smoking (W. Wang et al., 2016) and is associated with alcohol use (Aoun et al., 2018). In addition to testing for pleiotropy, we also used MTAG (Turley et al., 2018) to leverage the observed genetic correlations to increase power for locus discovery. Using this method, we discovered 1,193 independent, genome-wide significantly associated common variants (minor allele frequency (MAF), >1%; AgeSmk, 173; CigDay, 89; SmkCes, 83; SmkInit, 692; DrnkWk, 156) listed in Supplementary Table 12 and described further in the supplementary information.

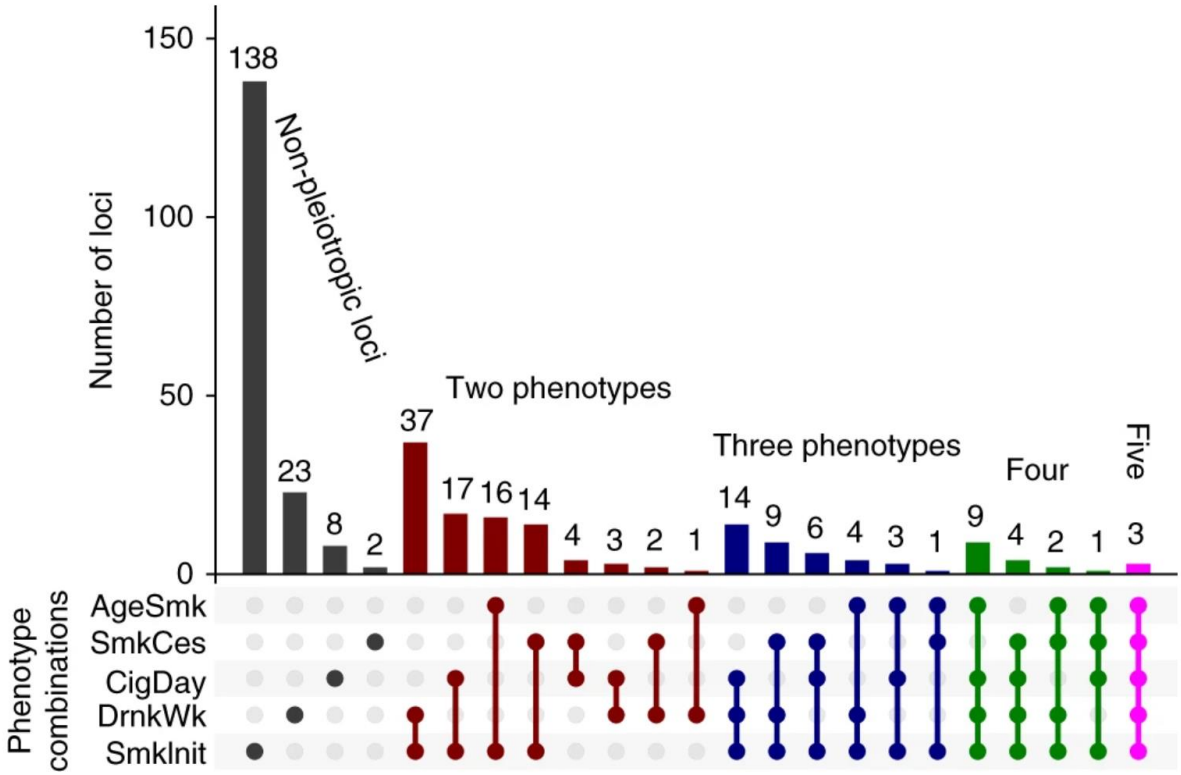

Fig. 2 Pleiotropy. Depicted here are results from the multivariate analysis of pleiotropy. For each locus, the method returns the best-fitting solution of which phenotypes were associated with that locus. All loci with one or more associated phenotypes are shown here. For example, every locus associated with AgeSmk was found to be pleiotropic for other phenotypes (green, blue, red, purple, and fuchsia bars), and no locus showed association with only AgeSmk (no dark gray bar for AgeSmk). When sample sizes are unequal across phenotypes, the method also improves power for those phenotypes with smaller samples. The total numbers of loci associated with each trait (whether pleiotropic or not) from these analyses were 40 (AgeSmk), 48 (SmkCes), 72 (CigDay), 111 (DrnkWk), and 278 (SmkInit). Full information is in Supplementary Table 11.

Phenotypic variation accounted for by our initial 566 conditionally independent genome-wide significant variants from the initial genome-wide association study (GWAS) ranged from 0.1%

(SmkCes) to 2.3% (SmkInit; see Fig. 3). SNP heritability calculated using linkage disequilibrium (LD) score regression (Bulik-Sullivan et al., 2015) ranged from 4.2% for DrnkWk to 8.0% for CigDay (Fig. 3 and Supplementary Table 13), consistent with estimates made using individual-level data (Yang, Lee, Goddard, & Visscher, 2011), SNP heritabilities calculated from the largest individual contributing studies (Supplementary Table 13), and prior work (Zheng et al., 2017). The results suggest that these phenotypes are highly polygenic, and that the majority of the heritability is accounted for by variants below standard GWAS thresholds.

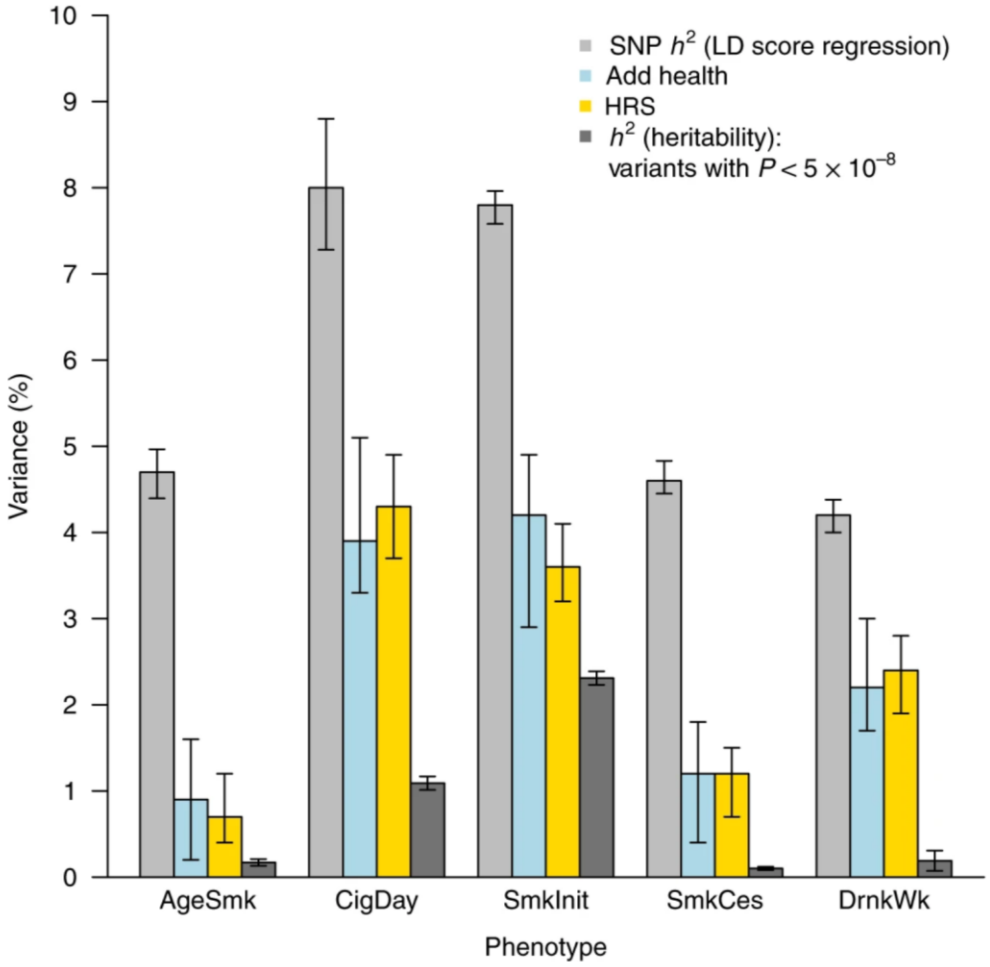

Fig. 3 Heritability and polygenic prediction. The light gray bars reflect SNP heritability, estimated with LD score regression. The light blue and gold bars reflect the predictive power of a PRS in Add Health and the HRS, respectively. Despite the 41 year generational gap between participants from these two studies, and major tobacco-related policy changes during that time, the polygenic scores are similarly predictive in both samples. Error bars are 95% confidence intervals estimated

with 1,000 bootstrapped repetitions. Dark gray bars represent the total phenotypic variance explained by only genome-wide significant SNPs.

To further investigate the polygenicity, polygenic risk scores (PRS; Supplementary Table 14) were computed on the National Longitudinal Study of Adolescent to Adult Health (Add Health) (Harris, Halpern, Haberstick, & Smolen, 2013) and the Health and Retirement Study (HRS) (Sonnega et al., 2014) datasets, which are representative of their birth cohorts in the United States and represent exposures to different tobacco policy environments. Add Health participants were born, on average, in 1979; average birth year in the HRS was 1938. Despite these generational differences, the polygenic score performed similarly in both samples. It accounted for approximately 1%, 4%, 1%, 4%, and 2.5% of variance in AgeSmk, CigDay, SmkCes, SmkInit, and DrnkWk, respectively, about half of the estimated SNP heritability of these traits (Fig. 3). More concretely, in Add Health and the HRS, respectively, a 1s.d. increase in the CigDay risk score resulted in two and three additional daily cigarettes; a 1s.d. increase on the SmkInit risk score resulted in a 12% and 10% increased risk of regularly smoking; and a 1s.d. increase on the DrnkWk risk score reflected one additional drink per week in both datasets.

Cell and tissue enrichment (Finucane et al., 2015) was observed across all five phenotypes within core histone marks from multiple central nervous system tissues (Supplementary Figs. 13–15 and Supplementary Tables 15 and 16). Enrichment was observed in tissues from cortical and sub-cortical regions in the central nervous system. Structure and function of these regions have been robustly associated with individual differences in frequencies, magnitudes, and clinical characteristics of alcohol use, and substance use/misuse generally, in human imaging research. Our results include significant enrichment across phenotypes and histone marks in the hippocampus (Wilson, Bair, Thomas, & Iacono, 2017), inferior temporal pathways (Feldstein Ewing, Sakhardande, & Blakemore, 2014), dorsolateral and medial prefrontal cortex (Goldstein & Volkow, 2011), caudate, and striatum (Volkow & Morales, 2015). Consistent with gene and pathway findings described below, alcohol and nicotine use affect dopaminergic and glutamatergic neurotransmission among these brain regions, compromising reward-based learning and facilitating drug-seeking behavior (Volkow & Morales, 2015). Enrichment within other cell or tissue groups and specific cell or tissue types included immune and liver cells, but was less consistent across analytical approaches.

We manually reviewed all of the genes implicated by the GWAS or gene-based tests (see Supplementary Tables 1–5 for the full catalog of implicated genes and Supplementary Tables 17–21 for gene and gene set test results). We replicated known associations between multiple variants in the nicotine metabolism gene *CYP2A6* with CigDay ($P=4.0\times10-99$) and SmkCes ($P=1.6\times10-48$). We replicated an association signal in the alcohol metabolism gene *ADH1B*

associated with DrnkWk, identifying in that locus 11 conditionally independently associated variants (lowest P <2.2×10−303).

All drugs of abuse activate the mesolimbic dopamine system reward pathway (Koob & Volkow, 2010), and dopamine-related genes have long been popular candidate genes. We found that variants near the widely studied dopamine receptor D2 (*DRD2*) (Koob & Volkow, 2016) were associated across phenotypes, including CigDay, SmkCes, and DrnkWk (P=6.5×10−12, 1.1×10−10, and 4.9×10−11, respectively), but not with AgeSmk or SmkInit, suggesting that these variants are less relevant in early stages of nicotine use. Other specific dopamine-related genes only showed associations with smoking phenotypes, including multiple associations between CigDay and SmkCes with dopamine β-hydroxylase (*DBH*; P=9.8×10−24 and 1.2×10−35, respectively) (Tobacco and Genetics Consortium, 2010), an enzyme necessary to convert dopamine to norepinephrine. SmkInit was associated with variation near protein phosphatase 1 regulatory subunit 1B (*PPP1R1B*; P=3.9×10−8), a signal transduction gene that affects synaptic plasticity and reward-based learning in the striatum (Fernandez, Schiappa, Girault, & Le Novère, 2006; Yagishita et al., 2014) and contributes to the behavioral effects of nicotine in mice (Zhu et al., 2005). In pathway analyses, dopamine gene sets were enriched only in SmkInit, where the exemplar ‘reactome dopamine neurotransmitter release cycle’ pathway was enriched (P=9.2×10−5; Fig. 4 and Supplementary Table 18).

Fig. 4 Correlations among exemplary DEPICT gene sets. (a,b) There were 68 clusters available for SmkInit (a) and 10 for DrnkWk (b) (CigDay, AgeSmk, and SmkCes did not have >1 exemplary set). Purple shading represents negative correlations, and red shading represents positive correlations, with increasing color intensity reflecting increasing correlation strength. Cluster names are truncated for space, with a full list of all names in Supplementary Table 18. The number after each name is the number of gene sets in each cluster. The matrix naturally falls into three red superclusters along the diagonal. The largest supercluster contains primarily gene sets related to neurotransmitter receptors, ion channels (sodium, potassium, calcium), learning/memory, and other aspects of central nervous system function. The middle supercluster includes gene sets defined by regulation of transcription and translation, including RNA binding and transcription factor activity. The final supercluster is composed primarily of gene sets related to development of the nervous system.

Neuronal acetylcholine nicotinic receptors are the initial site of nicotine action in the brain and have long been implicated in nicotine use and dependence (Stoker & Markou, 2013). With the exception of *CHRNA7*, all central-nervous-system-expressed nicotinic receptor genes were significantly associated with one or more smoking phenotypes, many reported here for the first

time. Enrichment was also noted for nicotinic-receptor-related pathways and genes in smoking phenotypes (Supplementary Tables 17–21). There was no evidence of association between nicotinic receptor genes or pathways with DrnkWk, despite the use of nicotinic receptor partial agonists (for example, varenicline) in the treatment of alcohol dependence (Litten et al., 2013).

Associations with SmkInit highlighted structures and functions related to long-term potentiation and reward-related learning and memory, systems that affect reward processing and addiction (Hyman, Malenka, & Nestler, 2006; Kalivas, 2009; Koob & Volkow, 2016). Glutamate is an important neurotransmitter mediating these processes, and exemplar pathways related to glutamate were significantly enriched in SmkInit (for example, 'extracellular-glutamate-gated ion channel', P=9.9×10−7 ; 'post-NMDA receptor activation events', P=5.5×10−5 ; and 'DLG4 PPI subnetwork', P=4.5×10−12; Supplementary Table 18). *DLG4* affects NMDA receptors and potassium channel clusters and has a central role in glutamatergic models of reward-related learning (Kalivas, 2009). Individual associated genes related to these pathways included glutamate ionotropic receptor NMDA type subunit 2 (*GRIN2A*; P=3.4×10−11) and homer scaffolding protein 2 (*HOMER2*; P=3.1×10−14), which affects addictive behavior in mice (Kalivas, 2009; Szumlinski et al., 2017) and regulates glutamate metabotropic receptor 1 (*GRM1*). Pathways enriched in SmkInit also included sodium-, potassium-, and calcium voltage-gated channels (Fig. 4 and Supplementary Table 18), essential to neuronal excitability and signaling.

Alcohol is known to affect glutamatergic signaling pathways (Gass & Olive, 2008), and more than half of the enriched pathways for DrnkWk clustered within the exemplar 'glutamate ionotropic receptor kainate type subunit 2 (*GRIK2*) PPI subnetwork' (Fig. 4 and Supplementary Table 18). However, not all DrnkWk-enriched pathways involved the brain as glucose and carbohydrate processing pathways were associated with DrnkWk but no smoking phenotype, perhaps suggesting that alcohol consumption is influenced by individual differences in one's ability to process calorie-rich alcoholic beverages. Finally, we discovered variation in and around gene-rich regions, including corticotropin-releasing hormone receptor 1 (*CRHR1*; P=1.6×10−17) and urocortin (UCN; P=8.1×10−45), associated with DrnkWk, but not smoking. *UCN* encodes an endogenous ligand for *CRHR1* and *CRHR2* (ref. (Vaughan et al., 1995)). *CRH* affects hormones involved in the stress response, including cortisol, and has been associated with the stress response and relapse to drug taking in animals (Logrip, Koob, & Zorrilla, 2011; Volkow, Koob, & McLellan, 2016).

Specific mechanisms by which implicated genes influence substance use in humans are largely unknown, even for those genes reported above involving systems, such as neurotransmission, reward-related learning and memory, and the stress response. To prioritize genes for functional experimentation, we tabulated conditionally independent genome-wide significant non-synonymous variants (Table 1). In the 406 GWAS loci, 4% of sentinel variants were

nonsynonymous, representing a significant enrichment (P=2.5×10−10; 0.4% of variants with MAF>0.1% in the imputation panel(McCarthy et al., 2016) were non-synonymous). Several genes in Table 1 have been previously associated with substance use/addiction (see Supplementary Table 22 for a list of previous associations), and two variants have been functionally validated (rs1229984 and rs16969968) (Edenberg, 2007; Lassi et al., 2016). The others have not, but in some cases their genes interact with established molecular targets of addiction and may themselves be suitable targets for further investigation. For example, rs1024323 in G-protein-coupled receptor kinase 4 (*GRK4*) was associated with CigDay (P=8.7×10−9) and lies within a locus associated with AgeSmk. GRK4 is involved in the regulation of G-protein-coupled receptors, including metabotropic glutamate receptor 1 (*GRM1*) (Sallese et al., 2000), *GABAB* receptors (Perroy, Adam, Qanbar, Chénier, & Bouvier, 2003), and dopamine receptors D1 (*DRD1*) and D3 (*DRD3*) in the kidneys and cerebellum, and is involved in essential hypertension (Yang, Villar, Armando, Jose, & Zeng, 2016). *GRK4* is also expressed in the midbrain and forebrain (GTEx Consortium et al., 2017; Yang et al., 2016), but no research has evaluated its impact on substance use behavior. To take one more example, the non-synonymous variant in *SLC39A8* affects zinc and manganese transport, is highly pleiotropic for complex phenotypes, and may impair inflammation, glutamatergic neurotransmission, and regulation of various metals in the body (Costas, 2018).

Conclusion

Ultimately, substance use is embedded in a complex web of causal relations (Kong et al., 2018) (for example, see Fig. 1), and caution must be exercised in drawing strong causal conclusions. However, our findings represent a major step forward in understanding the etiology of these complex, disease-relevant behaviors. In particular, statistical and interpretive power were both enabled by simultaneously studying multiple related substance use behaviors representing different stages of use and different substances. More precise measurements, including evaluating age and environment as moderators for these dynamic phenotypes (Vrieze, Hicks, Iacono, & McGue, 2012), functional research, and complementary gene mapping approaches (for example, sequencing) will aid in the discovery of mechanisms by which implicated genes may affect substance use and related disease risk.

Phenotype	Gene	rsID	Chr	Position	REF	ALT	AF	Beta	*P*	*N*	*Q*
CigDay (SmkCes)	*CHRNA5*	rs16969968[a]	15	78,882,925	G	A	0.34	0.075	1.2×10^{-278}	330,721	0.34
CigDay	*HIST1H2BE*	rs7766641	6	26,184,102	G	A	0.27	−0.014	2.9×10^{-10}	335,553	0.78
CigDay (AgeSmk)	*GRK4*	rs1024323	4	3,006,043	C	T	0.38	−0.012	8.7×10^{-9}	337,334	0.17
SmkInit	*REV3L*	rs462779[a]	6	111,695,887	G	A	0.81	−0.019	4.5×10^{-29}	1,232,091	0.67
SmkInit (DrnkWk)	*BDNF*	rs6265	11	27,679,916	C	T	0.20	−0.016	2.8×10^{-19}	1,232,091	0.13
SmkInit	*RHOT2*	rs1139897	16	720,986	G	A	0.23	−0.012	1.8×10^{-15}	1,232,091	0.61
SmkInit (DrnkWk)	*ZNF789*	rs6962772[a]	7	99,081,730	A	G	0.15	−0.015	2.1×10^{-14}	1,232,091	0.92
SmkInit	*BRWD1*	rs4818005[a]	21	40,574,305	A	G	0.58	−0.010	3.9×10^{-14}	1,232,091	0.75
SmkInit	*ENTPD6*	rs6050446	20	25,195,509	A	G	0.97	0.035	8.8×10^{-13}	1,225,969	0.33
SmkInit	*RPS6KA4*	rs17857342[a]	11	64,138,905	T	G	0.38	−0.010	9.8×10^{-12}	1,232,091	0.16
SmkInit	*FAM163A*	rs147052174	1	179,783,167	G	T	0.02	0.037	2.3×10^{-10}	1,232,091	0.59
SmkInit	*PRRC2B*	rs34553878	9	134,907,263	A	G	0.11	0.016	1.2×10^{-9}	1,232,091	0.28

SmkInit	*ADAM15*	rs45444697[a]	1	155033918	C	T	0.21	0.010	5.3×10^{-9}	1,232,091	0.46
SmkInit	*MMS22L*	rs9481410[a]	6	97,677,118	G	A	0.76	0.010	1.1×10^{-8}	1,232,091	0.04
SmkInit	*QSER1*	rs62618693	11	32,956,492	C	T	0.04	−0.020	2.1×10^{-8}	1,232,091	1.00
DrnkWk	*ADH1B*	rs1229984	4	100,239,319	T	C	0.96	0.060	2.2×10^{-308}	941,280	0.05
DrnkWk	*GCKR*	rs1260326	2	27,730,940	T	C	0.60	0.008	8.1×10^{-45}	941,280	0.10
DrnkWk	*SLC39A8*	rs13107325	4	103,188,709	C	T	0.07	−0.009	1.5×10^{-22}	941,280	0.33
DrnkWk	*SERPINA1*	rs28929474	14	94,844,947	C	T	0.02	−0.012	1.3×10^{-11}	941,280	0.50
DrnkWk (SmkInit)	*ACTR1B*	rs11692465	2	98,275,354	G	A	0.09	0.008	2.5×10^{-11}	937,516	0.40
DrnkWk	*TNFSF12-13*	rs3803800	17	7,462,969	A	G	0.79	0.004	1.5×10^{-10}	941,280	0.67
DrnkWk	*HGFAC*	rs3748034	4	3,446,091	G	T	0.14	−0.005	1.7×10^{-8}	941,280	0.65

Table 1 Non-synonymous sentinel variants. The sentinel variant in approximately 4% of loci was non-synonymous. Shown here are all non-synonymous sentinel variants, and all non-synonymous variants in near-perfect LD with a sentinel variant. If the listed gene was also associated (through single variant or gene-based test) with another phenotype, that phenotype is listed in parentheses. Several genes have been implicated in previous studies of substance use/addiction, including *CHRNA5*, *BDNF*, *GCKR*, and *ADH1B*. Phenotype abbreviations are defined in Fig. 1. Chr, chromosome; REF, reference allele; ALT, alternate allele; AF, allele frequency of ALT; Q, Cochrane's Q statistic P value. [a]These variants were not themselves sentinel, but were in near-perfect LD with a sentinel variant ($r2 > 0.99$, from the 1000 Genomes European population). The scale of Beta is on the unit of the standard deviation of the phenotype. For binary phenotypes the standard deviation was calculated from the weighted average prevalence across all studies included in the meta-analysis (available in Supplementary Table 7

Chapter 2

Introduction

Smoking is a major risk factor for many diseases, including common respiratory disorders such as chronic obstructive pulmonary disease (COPD) (Wain et al., 2017, 2015), cancer (McKay et al., 2017) and cardiovascular diseases (O'Donnell & Nabel, 2011), and is reported to cause 1 in 10 premature deaths worldwide (Reitsma et al., 2017). A greater understanding of the genetic etiology of smoking behavior 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 (Bloom et al., 2014; Hancock et al., 2018; Siedlinski et al., 2011; Thakur, Sengupta, Grizenko, Choudhry, & Joober, 2012; Thorgeir E. Thorgeirsson et al., 2010; Timofeeva et al., 2011; Tobacco and Genetics Consortium, 2010; Wain et al., 2015)(with minor allele frequency, MAF >0.01) robustly associated with smoking behavior-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 (Munafo, Tilling, Taylor, Evans, & Davey Smith, 2018) and cotinine, a biomarker of nicotine intake (Keskitalo et al., 2009), respectively. Overall, genetic loci identified to date explain ~2% of the estimated genetic heritability of smoking behavior (Tobacco and Genetics Consortium, 2010), which is reported to be between 40–60% (Carmelli, Swan, Robinette, & Fabsitz, 1992; Kaprio, Koskenvuo, & Sarna, 1981; Vink, Willemsen, & Boomsma, 2005). A recent study suggested that an important proportion (~3.3%) of the phenotypic variance of smoking behavior-related traits was explained by rare nonsynonymous variants (MAF <0.01) (Brazel et al., 2019). Hence, well-powered studies of rare variants are needed.

To investigate the effect of rare coding variants on smoking behavior, we studied 346,813 participants (of which 324,851 were of European ancestry) from 61 cohorts (Supp. Tables 40 and 41) 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 (Tobacco and Genetics Consortium, 2010), and constitute the most powerful exome-array study of smoking behavior to date.

Method

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 analyzed in the discovery stage (Fig. 5). Further information about the participating cohorts and consortia is given in Supp. Table 40 and the Suppl Material. All participants provided written informed consent and studies were approved by local Research Ethics Committees and/or Institutional Review boards.

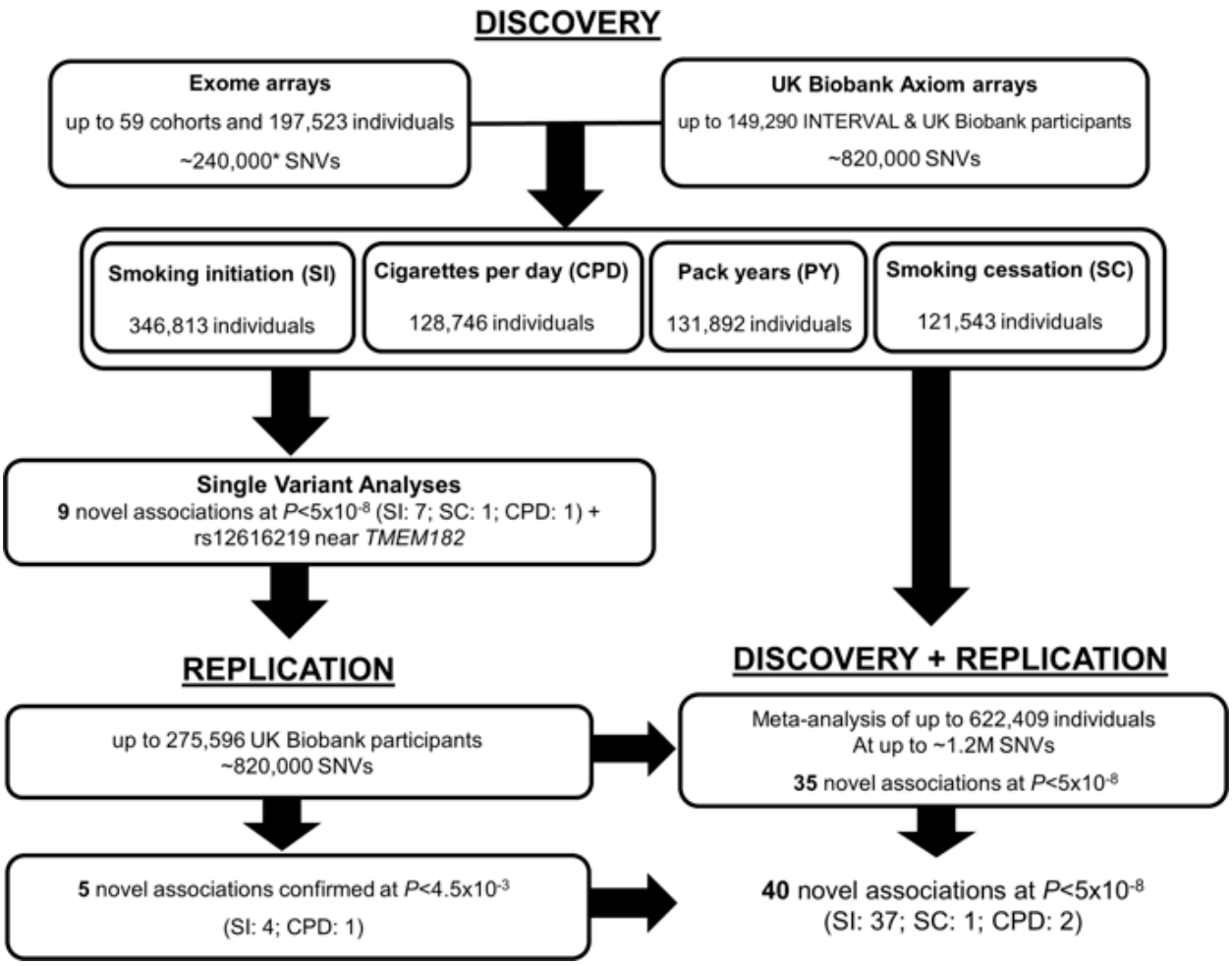

Fig. 5 Study design including the discovery and replication stages. NB: Gene-based studies, conditional analyses, and replication in African American ancestry samples not shown here for clarity. *GFG and NAGOZALC studies contributed additional custom content

Phenotypes

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

1. 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 (Centers for Disease Control and Prevention (CDC), 2008);
2. Cigarettes per day (CigDay; quantitative trait: average number of cigarettes smoked per day by ever smokers);
3. 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.
4. 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 CigDay and pack-years. CigDay was inverse normal transformed in the CHD Exome+, INTERVAL and CGSB studies and categorized (1–10, 11–20, 21–30, and 31+ CigDay) by the GSCAN studies and UK Biobank (Supp. Table 41). All studies performed an inverse normal transformation of pack-years. Summary statistics of study level phenotype distributions are provided in Supp. Table 40.

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 41). In total, ~1.06M SNVs were analyzed 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 analyzed to 1,207,583 SNVs. Individual studies performed quality control (QC; Supp. Material, Supp. Table 41) 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 (Staley et al., 2017)) undertook analyses of up to four smoking traits using RAREMETALWORKER (Feng, Liu, Zhan, Wing, & Abecasis, 2014) or RVTESTS (Zhan, Hu, Li, Abecasis, & Liu, 2016) (Supp. Table 41), which generated single variant score statistics and their covariance matrices within sliding windows of 1Mb. CigDay and pack-years were analyzed using linear models or linear mixed models. Smoking initiation and smoking cessation were analyzed 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 41). Chromosome X variants were analyzed 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 analyzed 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 (Willer, Li, & Abecasis, 2010). 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 (Bycroft et al., 2018). 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 *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 analyze chromosome X (Supp. Table 41). 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 utilize 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 (Jiang et al., 2018). 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 ~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 utilized three statistical methods as part of the RAREMETAL package: the Weighted Sum Test (WST) (Madsen & Browning, 2009), the burden test (Morris & Zeggini, 2010) and the Sequence Kernel Association test (SKAT) (Wu et al., 2011). EPACTS (v.3.3.0) (Zhan & Liu, 2015) 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 ~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 (Feng et al., 2014; Jiang et al., 2018).

Mendelian randomization analyses

To evaluate the causal effect of SmkInit and CigDay on BMI, schizophrenia and educational attainment (EA), we conducted Mendelian randomization (MR) analyses using three complementary approaches available in MR-Base (Hemani et al., 2018): inverse variance weighted regression (Pierce & Burgess, 2013), MR-Egger (Bowden, Davey Smith, & Burgess, 2015; Rees, Wood, & Burgess, 2017), and weighted median (Bowden, Smith, Haycock, & Burgess, 2016). We used both the previously reported smoking-associated SNVs and the SNVs from the current report (as provided in Tables 2-5 and Supp. Table 42) as instrumental variables. The BMI (Locke et al., 2015), schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014) and educational attainment (Okbay et al., 2016) 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 SmkInit and CigDay 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 (MacArthur et al., 2017) (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 (Battle, Brown, Engelhardt, & Montgomery, 2017), Brain xQTL (dorsolateral prefrontal cortex) (Ng et al., 2017) and BRAINEAC (Trabzuni et al., 2011) 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 (Ongen, Buil, Brown, Dermitzakis, & Delaneau, 2016; Storey & Tibshirani, 2003). 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 ($r^2 > 0.8$ 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 (Kamburov, Wierling, Lehrach, & Herwig, 2009) to identify whether these genes were over-represented in any known biological pathways. Replicated missense SNVs were also put into PolyPhen-2 (Adzhubei et al., 2010) and FATHMM (unweighted) (Shihab et al., 2013) 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 behavior traits, of which 9 were novel (Table 2, Supp. Table 42). Seven novel loci were identified for smoking initiation, one for both CigDay and pack-years and one for smoking cessation (Figs. 5, 6, Table 2 and Supp. Figure 19). Results for the significant loci were consistent across participating cohorts and there was at least nominal evidence of association ($P < 0.05$) at the novel loci within each of the contributing consortia (Supp. Table 43). Full association results for all novel SNVs across the four traits are provided in Supp. Table 44. No rare variants were genome-wide significant; the rare variant with the smallest *P*-value was a missense variant in *ATF6*, rs141611945 (MAF < 0.0001, CigDay $P = 2.95 \times 10^{-7}$).

dbSNP ID (Exome ID)	Chr:Pos	EA/OA	Gene	Consequence	Trait	Discovery stage				Replication stage	
						N	EAF	DoE	*P*-value	Beta (SE)	*P*-value
rs141611945 (exm118559)	1:161771868	G/A	*ATF6*	Missense	CigDay	128,746	0.0065% MAC = 9	+	2.95×10^{-7}	0.184 (0.169)	*P = 0.276 in African American samples
rs1190736 ** (exm1659559)	X:136113464	A/C	*GPR101*	Missense	CigDay (PY)	99,037 (96,824)	46.6% (47.0%)	-	**1.40×10^{-11}** (**4.98×10^{-9}**)	−0.028 (0.0041) −0.027 (0.0049) −0.028 (0.0073)	All samples: **8.20×10^{-12}** (**2.70×10^{-11}**) Males only: **1.90×10^{-8}** (**6.0×10^{-8}**) Females only: **1.10×10^{-4}** (**7.1×10^{-4}**)
rs462779 (exm572256)	6:111695887	A/G	*REV3L*	Missense	SmkInit	346,682	80.1%	-	**4.52×10^{-8}**	−0.023 (0.0034)	**9.7×10^{-12}**
rs216195 (exm1276230)	17:2203167	G/T	*SMG6*	Missense	SmkInit	335,406	27.3%	-	2.80×10^{-8}	−0.008 (0.0029)	8.5×10^{-3}
rs11539157 (exm1643833)	X:68381264	A/C	*PJA1*	Missense	SmkInit	289,917	16.5%	+	**1.39×10^{-11}**	0.022 (0.0026) 0.0158 (0.0033) 0.0185 (0.0039)	All samples: **5.40×10^{-17}** Males only: **1.30×10^{-6}** Females

											only: **2.20×10^{-6}**
Non - Exome - chip SNVs											
rs12616219	2:104352495	A/C	*TMEM182*	Intergenic	SmkInit	112,811	46.4%	-	**5.49×10^{-8}**	−0.015 (0.0027)	**5.5×10^{-8}**
rs1150691	6:28168033	G/A	*ZSCAN9*	Missense	SmkInit	112,811	34.8%	-	4.95×10^{-8}	−0.007 (0.0028)	8.0×10^{-3}
rs2841334	9:128122320	A/G	*GAPVD1*	Intronic	SmkInit	112,811	20.9%	-	2.28×10^{-8}	−0.009 (0.0033)	7.5×10^{-3}
rs202664	22:41813886	C/T	*TOB2*	Intergenic	SmkCes	51,043	19.9%	-	1.02×10^{-8}	−0.011 (0.0050)	2.1×10^{-2}
rs11895381	2:60053727	A/G	*BCL11A*	Intergenic	SmkInit	112,811	34.2%	-	5.61×10^{-9}	−0.007 (0.0028)	1.2×10^{-2}
rs12780116	10:104821946	A/G	*CNNM2*	Intronic	SmkInit	112,811	13.9%	+	**9.19×10^{-10}**	0.017 (0.0039)	**1.1×10^{-5}**

Table 2. 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 (SmkInit), cigarettes per day (CigDay), pack-years (PY), and smoking cessation (SmkCes) 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(se) for the association of rs1190736 with PY in the replication stage was −0.026 (0.0039)

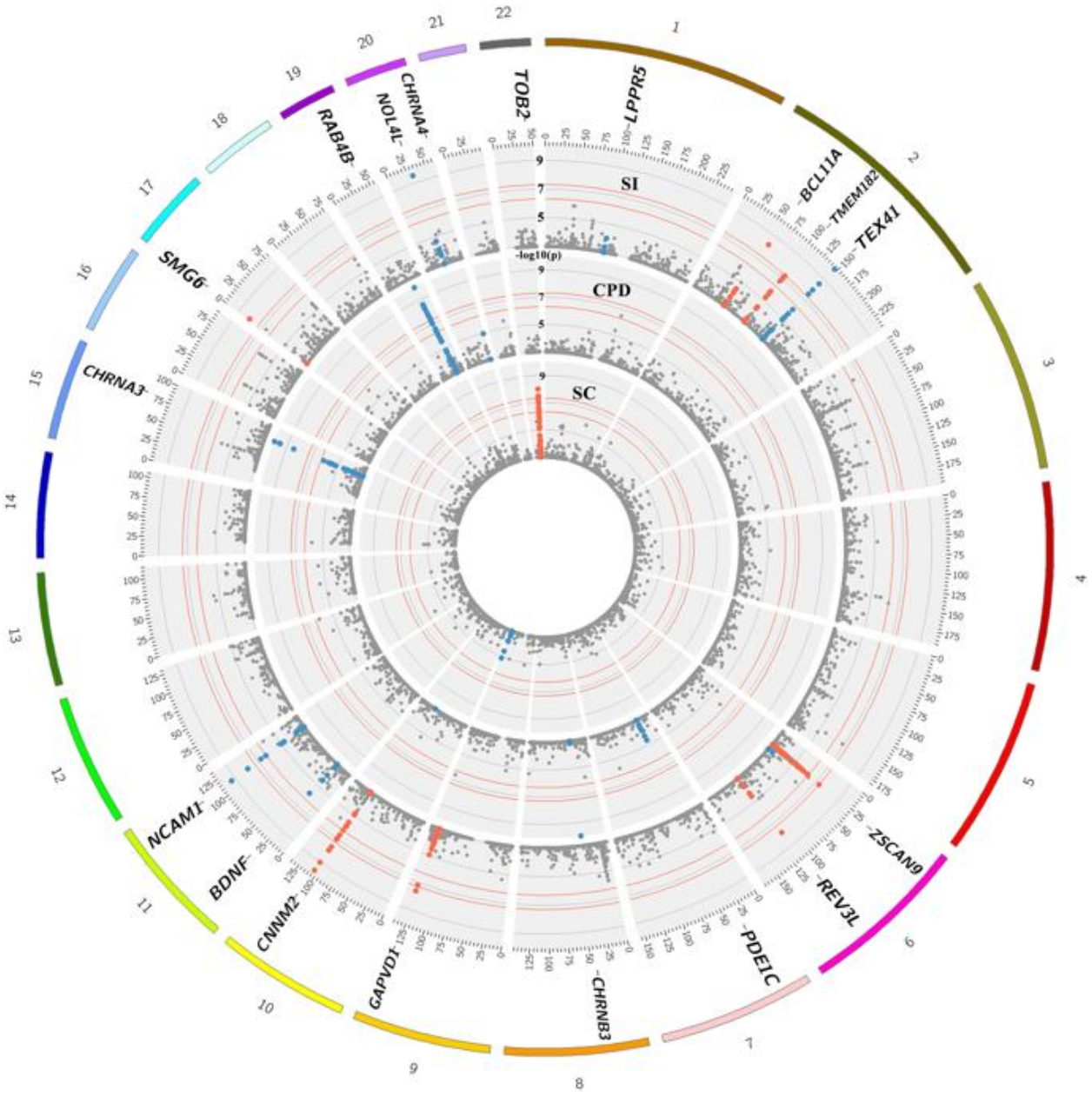

Fig. 6 A concentric Circos plot of the association results for SmkInit (SI; outer ring), CigDay (CPD) and SmkCes (SC ; inner ring) for chromosomes 1–22 (Pack-years results, which can be found in Supp. Figure 19, are omitted for clarity). Each dot represents a SNV, with the X and Y axes corresponding 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 0.00001 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)

Eleven SNVs (including rs12616219 near *TMEM182* with $P = 5.49 \times 10^{-8}$, and the rare variant, rs141611945) were taken forward for replication in independent samples (Table 2). The latest release of European UK Biobank individuals not included in the discovery stage (smoking initiation, $n = 275{,}596$; smoking cessation $n = 123{,}851$; CigDay $n = 80{,}015$; pack-years $n = 78{,}897$), was used for replication of the common variants (Fig. 6). Five of the common variants replicated (four for smoking initiation and one with CigDay and pack-years) at $P < 0.0045$. Two coding variants (rs11539157, rs1190736) were predicted to be 'probably damaging' by PolyPhen-2 and FATHMM. The remaining five SNVs were at least nominally associated ($P < 0.01$) in the replication samples and had consistent direction of effect across discovery and replication. Replication for the rare variant rs141611945 could not be carried out in UK Biobank as the SNV nor its proxies ($r^2 > 0.3$) were available. Thus we initiated replication in African American samples of the COGA ($n = 476$) and HRS ($n = 961$) cohorts (overall MAF≈0.01). The direction of effect was consistent in the two replication cohorts and consistent with the discovery meta-analysis but a meta-analysis of the two replication cohorts yielded a $P = 0.28$. Further data are required to replicate this association.

We also performed a meta-analysis combining the discovery and replication samples (up to 622,409 individuals). LD score regression showed that the λ (intercept) for all traits was ~1.00, which indicated that confounding factors inflating the results was not an issue (Bulik-Sullivan et al., 2015; Zheng et al., 2017). The combined analysis identified 35 additional novel SNV-smoking trait associations, 33 with smoking initiation, one with CigDay and one with smoking cessation at $P < 5 \times 10^{-8}$ (Table 3). We note that among our four SNVs that did not replicate, rs216195 (in *SMG6*) was genome-wide significant in the combined meta-analysis of discovery and replication studies ($P = 2.41 \times 10^{-9}$; Table 3).

We also calculated the phenotypic variance explained for novel and known variants. Results can be found in the 'Calculation of Phenotypic Variance Explained' section in the Supplementary Material.

dbSNP ID (Exome-chip ID)	Chr:Pos	EA/OA	Gene	Consequence	Trait	EAF	Beta (se) in replication stage	*P*-value in combined meta-analysis (*P*-value in Discovery/Replication stage)	Notes
Combining only genotyped Exome - chip content on the Axiom array									
rs1514175	1:74991644	G/A	*TNNI3K*	Intronic	SI	0.57	−0.011 (0.003)	**5.42×10^{-9}** ($9.03 \times 10^{-5}/1.0 \times 10^{-5}$)	Previously associated with BMI
rs7096169	10:104618695	G/A	*BORCS7* (*CNNM2*[#] in Table 1)	Intronic	SI	0.31	0.016 (0.003)	**2.17×10^{-13}** ($3.38 \times 10^{-7}/7.3 \times 10^{-9}$)	$r^2 = 0.28$ between rs7096169 and rs12780116 (Table 2) in 1000 Genomes EUR. Previously associated with Schizophrenia. rs7096169 an eQTL for *ARL3*, BORCS7, and *AS3MT* in ≥1 of the brain tissues in GTEx
rs2292239	12:56482180	G/T	*ERBB3*	Intronic	SI	0.66	0.0121 (0.003)	**2.78×10^{-8}** ($7.56 \times 10^{-5}/1.5 \times 10^{-5}$)	Previously associated with type-1 diabetes and years of

									educational attainment. rs2292239 is an eQTL for *RPS26* and *SUOX* in ≥4 of the brain tissues in GTEx
rs216195	17:2203167	G/T	*SMG6* [#]	Missense	SI	0.29	−0.0076 (0.003)	**2.41×10^{-9}** ($2.80 \times 10^{-8}/8.5 \times 10^{-3}$)	Same SNV as in Table 2
Combining well - imputed Exome - chip content on the Axiom array									
rs2960306 (exm383568)	4:2990499	T/G	*GRK4*	Missense	CPD	0.34	−0.024 (0.005)	**1.06×10^{-9}** ($3.99 \times 10^{-5}/3.8 \times 10^{-6}$)	rs2960306 is an eQTL for *GRK4* in four of the brain tissues in GTEx
rs4908760	1:8526142	A/G	*RERE*	Intronic	SI	0.35	0.0078 (0.003)	**1.76×10^{-8}** ($3.36 \times 10^{-6}/4.7 \times 10^{-3}$)	Previously associated with Vitiligo
rs6692219 (exm127721)	1:179989584	C/G	*CEP350*	Missense	SI	0.028	−0.0257 (0.008)	**4.69×10^{-9}** ($1.08 \times 10^{-6}/1.3 \times 10^{-3}$)	
rs11971186	7:126437897	G/A	*GRM8*	Intronic	SI	0.20	−0.0080 (0.003)	**1.45×10^{-8}** ($1.38 \times 10^{-6}/3.9 \times 10^{-3}$)	
rs150493199 (exm249655)	2:179721072	A/T	*CCDC141*	Missense	SC	0.0098	0.048 (0.134)	**1.28×10^{-8}** ($6.45 \times 10^{-8}/0.72$)	
Non - Exome - chip SNVs									
rs3001723	1:44037685	A/G	*PTPRF*	Intronic	SI	0.21	0.0159 (0.003)	**6.64×10^{-11}** ($0.00015/4.1 \times 10^{-8}$)	Previously associated with Schizophrenia and

									Years of educational attainment
rs1937455	1:66416939	G/A	*PDE4B*	Intronic	SI	0.30	−0.0146 (0.0027)	**1.23×10^{-9}** ($0.00073/5.6 \times 10^{-8}$)	
rs72720396	1:91191582	G/A	*BARHL2*	Intergenic	SI	0.16	−0.0150 (0.003)	**9.86×10^{-9}** ($5.63 \times 10^{-5}/1.9 \times 10^{-6}$)	
rs6673752	1:154219177	C/G	*UBAP2L*	Intronic	SI	0.055	−0.027 (0.004)	**1.1×10^{-11}** (NA/1.1×10^{-11})	
rs2947411	2:614168	G/A	*TMEM18*	Intergenic	SI	0.83	0.0189 (0.004)	**4.97×10^{-10}** ($0.00017/7.1 \times 10^{-8}$)	Previously associated with BMI
rs528301	2:45154908	A/G	*SIX3*	Intergenic	SI	0.38	0.0136 (0.002)	**4.12×10^{-11}** ($1.77 \times 10^{-6}/3.8 \times 10^{-7}$)	
rs6738833	2:104150891	T/C	*TMEM182* [#]	Intergenic	SI	0.33	−0.018 (0.003)	**8.66×10^{-14}** ($1.63 \times 10^{-6}/4.4 \times 10^{-11}$)	$r^2 = 0.69$ between rs6738833 and rs12616219 (Table 2) in European samples of the 1000 Genomes Project
rs13026471	2:137564022	T/C	*THSD7B*	Intronic	SI	0.18	0.0127 (0.003)	**2.45×10^{-8}** ($0.00028/3.0 \times 10^{-5}$)	
rs6724928	2:156005991	C/T	*KCNJ3*	Intergenic	SI	0.32	−0.011 (0.003)	**4.47×10^{-8}** ($0.0019/4.8 \times 10^{-5}$)	
rs13022438	2:162800372	G/A	*SLC4A10*	Intronic	SI	0.27	0.0146 (0.003)	**1.41×10^{-11}** ($0.0005/8.1 \times 10^{-8}$)	

rs1869244	3:5724531	A/G	*LOC105376939*	Intergenic	SI	0.32	0.0123 (0.003)	**2.76×10^{-9}** ($0.00040/4.1 \times 10^{-6}$)	
rs35438712	3:85588205	T/C	*CADM2*	Intronic	SI	0.25	0.017 (0.003)	**1.99×10^{-13}** ($1.15 \times 10^{-5}/3.2 \times 10^{-10}$)	
rs6883351	5:22193967	T/C	*CDH12*	Intronic	SI	0.34	0.0129 (0.003)	**4.69×10^{-8}** ($0.0010/1.4 \times 10^{-6}$)	
rs6414946	5:87729711	C/A	*TMEM161B*	Intronic	SI	0.32	−0.0137 (0.003)	**5.27×10^{-10}** ($3.63 \times 10^{-5}/2.8 \times 10^{-7}$)	
rs11747772	5:166992708	C/T	*TENM2*	Intronic	SI	0.25	0.0144 (0.003)	**6.20×10^{-9}** ($0.011/2.2 \times 10^{-7}$)	
rs9320995	6:98726381	G/A	*POU3F2*	Intergenic	SI	0.18	0.0150 (0.003)	**1.70×10^{-8}** ($0.00079/6.1 \times 10^{-7}$)	
rs10255516	7:1675621	G/A	*ELFN1*	Intergenic	SI	0.33	−0.0139 (0.003)	**2.86×10^{-10}** ($0.0021/1.8 \times 10^{-7}$)	
rs10807839	7:3344629	G/A	*SDK1*	Intronic	SI	0.19	0.0162 (0.003)	**8.93×10^{-11}** ($0.0026/4.4 \times 10^{-8}$)	
rs6965740	7:117514840	T/G	*CTTNBP2*	Intergenic	SI	0.31	−0.0126 (0.003)	**9.66×10^{-9}** ($5.56 \times 10^{-6}/2.8 \times 10^{-6}$)	
rs11776293	8:27418429	T/C	*EPHX2*	Intronic	SI	0.12	−0.0200 (0.003)	**2.23×10^{-12}** ($0.00011/8.9 \times 10^{-9}$)	rs11776293 is an eQTL for *CHRNA2* in cerebellum in GTEx
rs1562612	8:59817068	G/A	*TOX*	Intronic	SI	0.35	−0.0112 (0.003)	**1.15×10^{-9}** ($1.42 \times 10^{-5}/2.9 \times 10^{-5}$)	
rs3857914	8:93184065	C/T	*RUNX1T1*	Intergenic	SI	0.19	0.0157 (0.003)	**1.54×10^{-9}** ($0.065/7.1 \times 10^{-8}$)	

rs2799849	9:86752641	C/T	*RMI1*	Intergenic	SI	0.22	−0.0156 (0.003)	**1.94×10^{-8}** ($0.026/4.8 \times 10^{-8}$)	
rs6482190	10:22037809	A/G	*LOC107984214*	Intronic	SI	0.17	0.0146 (0.003)	**8.85×10^{-9}** ($0.0021/9.5 \times 10^{-7}$)	
rs4523689	11:7950797	G/A	*OR10A6*	Intergenic	SI	0.27	−0.012 (0.003)	**7.77×10^{-9}** ($0.00030/2.2 \times 10^{-5}$)	
rs933006	13:38350193	A/G	*TRPC4*	Intronic	SI	0.32	−0.0143 (0.003)	**3.50×10^{-8}** ($0.022/9.6 \times 10^{-8}$)	
rs557899	15:47643795	A/C	*SEMA6D*	Intronic	SI	0.26	0.0157 (0.003)	**2.99×10^{-13}** ($4.46 \times 10^{-5}/1.0 \times 10^{-8}$)	
rs76608582	19:4474725	A/C	*HDGFRP2*	Intronic	SI	0.029	−0.0360 (0.007)	**8.50×10^{-9}** ($0.012/4.3 \times 10^{-8}$)	

Table 3. Chromosome (Chr) and position (Pos) for each SNV is given for hg19 build 37. Only SNVs reaching genome-wide significance ($P < 5 \times 10\text{-}8$, 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 analyzed. 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

Associations at known smoking behavior loci

We assessed evidence for associations at the 14 SNVs previously reported for smoking behavior-related traits. Seven were genotyped on the exome array and proxies ($r^2 > 0.3$; ±2 Mb) were identified for the remaining seven (Supp. Table 42). 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 42 and 44).

Conditional analyses identified five independent associations within three previously reported loci and all five replicated (Table 4). At the 19q13 (*RAB4B*) locus, there were three variants in or near *CYP2A6* associated with CigDay 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 6). 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 CigDay independently of the established variant (rs1051730) and (in agreement with a previous report (J. C. Wang et al., 2009)) is an eQTL for *CHRNA5* in brain putamen basal ganglia tissues in GTEx.

Gene region	dbSNP ID	Chr:Pos	EA/OA	Consequence	Trait	EAF	*P* (unconditional)	SNV(s) conditioned on	Discovery Conditional *P* [DoE]	Conditional *P* in replication [DoE]
19q13 (*RAB4B*)	rs8102683	19:41363765	C/T	Intergenic	CPD	74.8%	**4.53×10^{-16}**	rs7937	**1.44×10^{-13}** [+]	3.5×10^{-4} [+]
	rs28399442	19:41354458	A/C	Intronic (*CYP2A6*)	CPD	1.3%	**2.27×10^{-12}**	rs7937, rs8102683	**2.63×10^{-12}** [+]	**8.1×10^{-14}** [+]
	rs3865453	19:41338556	T/C	Intergenic	CPD	6.54%	**2.96×10^{-12}**	rs7937, rs8102683, rs28399442	**4.96×10^{-10}** [−]	**2.3×10^{-13}** [−]
TEX41 - PABPC1P2	rs11694518	2:146125523	T/C	Intergenic	SI	29.5%	**2.90×10^{-9}**	rs10193706	3.43×10^{-7} [−]	**4.0×10^{-31}** [−]
15q25 (*CHRNA3*)	rs938682	15:78882925	A/G	Intronic (*CHRNA3*)	CPD	76.4%	**1.83×10^{-69}**	rs1051730	**7.77×10^{-21}** [+]	**1.0×10^{-13}** [+]

Table 4. 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

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 CigDay and/or pack-years (Supp. Table 45**)**. 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 45). 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 SmkInit and CigDay on BMI, schizophrenia and EA using the MR-Egger, median weighted and inverse variance weighted methods. We found a causal association between SmkInit and EA using both the median weighted and inverse variance weighted methods ($P < 0.0001$; Supp. Table 46) but not with MR-Egger ($P = 0.2$). There was an association of SmkInit with BMI using MR-Egger only ($P = 0.01$; Supp. Table 46), but there was evidence of horizontal pleiotropy ($P = 0.001$) and no support from the other methods. Similarly, increased CigDay 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 CigDay or SmkInit using SNVs associated with schizophrenia, EA and BMI, respectively, as instrumental variables. No evidence of such reverse causation was found (Supp. Table 46). These results were consistent with previous analyses (Gage et al., 2017). There was no evidence of a causal effect of SmkInit on schizophrenia, or CigDay on educational attainment (Supp. Table 46).

Functional characterization of novel loci

Using proxies with $r^2 \geq 0.8$ in 1000 Genomes Europeans, we queried the GWAS catalogue (MacArthur et al., 2017) ($P \leq 5 \times 10^{-8}$) for pleiotropic effects of our novel sentinel SNVs. Two, rs11539157 and rs3001723 were previously associated with schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014), suggesting shared biological pathways between schizophrenia and smoking behaviors (Table 3). This fits with the known association of smoking with schizophrenia (Kelly & McCreadie, 2000). Two, rs1514175 and rs2947411 have previously been associated with BMI (Speliotes et al., 2010), and extreme obesity (Wheeler et al., 2013).

eQTL lookups in GTEx V7 (13 Brain tissues with ≥80 samples) (Battle et al., 2017), Brain xQTL (Ng et al., 2017) and BRAINEAC (Trabzuni et al., 2011) 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 $P = 4.8x10^{-8}$; $\beta = -0.40$) 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 3).

Biological pathway enrichment analyses carried out in ConsensusPathDB (Kamburov et al., 2009) using the genes implicated by the eQTL databases (Table 3) 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 behavior 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 (Hecht, 1999; McKay et al., 2017), heart disease (Ockene & Miller, 1997; O'Donnell & Nabel, 2011)and many respiratory diseases such as COPD (Wain et al., 2017, 2015). Not initiating is the best way to prevent smoking-related diseases and genetics can play a considerable part in smoking behaviors including initiation. We have performed the largest exome-wide genetic association study of smoking behavior-related traits to date involving up to 622,409 individuals, and identified and replicated five associations, including two on the X-chromosome (Table 2). We identified a further 35 novel associations in a meta-analysis of discovery and replication cohorts (Table 3). We validated 14 previously reported SNV-smoking trait associations (Supp. Table 42) and identified secondary independent associations at three loci, including three in the 19q13 region (rs8102683, rs28399442, and rs3865453; Table 42).

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 CigDay and *CHRNA2* and *MMP17* with pack-years (Supp. Table 45; 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 neuropeptide that increases intracellular cyclic adenosine monophosphate levels (Uhlen et al., 2015). *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 breakdown of extracellular matrix in normal physiological processes (O'Leary et al., 2016). Given, we were not able conclusively to identify rare variant associations, even larger studies, are required to identify rare variants associated with smoking behaviors. In addition, phenotypes such as cotinine levels (Ware et al., 2016)and nicotine metabolism speed (Loukola et al., 2015) could be interrogated using methods such as MTAG (Turley et al., 2018) to improve power.

As recommended by UK Biobank, we analyzed 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 (Munafo et al., 2018). 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-analyzed 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 47).

Follow-up of the replicated SNVs in the literature and eQTL databases implicated some potentially interesting genes: *NT5C2* is known to hydrolyze purine nucleotides and be involved in maintaining cellular nucleotide balance, and was previously associated with schizophrenia (Aberg et al., 2013). *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 (Zhang et al., 2013). 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 SmkInit in the combined meta-analysis, and is in LD with rs56372821 ($r^2 = 0.83$), which is associated with reduced cannabis use disorder (Demontis et al., 2019). 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 (Steenaard et al., 2015). The combined meta-analysis also identified a rare missense variant in *CCDC141*, rs150493199 (MAF < 0.01; Table 3). Coding variants in *CCDC141* were previously

associated with heart rate (van den Berg et al., 2017) and blood pressure (Hoffmann et al., 2017; Warren et al., 2017).

Smoking behaviors 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 (Kong et al., 2018). They also show that there is an effect of educational attainment and genetic nurture on smoking behavior. Four of our sentinel SNVs (or a strong proxy; $r^2 > 0.8$) were associated with years of educational attainment (Okbay et al., 2016) (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 CigDay that was previously identified by Kumasaka et al. in a Japanese population (Kumasaka et al., 2012) 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. (Loukola et al., 2015) 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.

CigDay and pack-years are two correlated measures of smoking. In the ~40,000 individuals from UK Biobank with CigDay and pack-years calculated, correlation between CigDay 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, CigDay 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 (Tobacco and Genetics Consortium, 2010)) was associated with increased pack-years ($P = 1.29 \times 10^{-14}$) and increased CigDay ($P = 2.93 \times 10^{-9}$) 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 CigDay ($n = 128,746$). These findings may help elucidate the genetic basis of these correlated addiction phenotypes.

We performed the largest exome-wide genetic association study of smoking behavior-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 etiology of smoking behavior 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.

Chapter 3

Introduction

Endophenotypes have been highly regarded as a measurable and close proxy between genetics and a range of psychiatric disorders and related phenotypes. They are viewed as a manifesting and measurable intermediate phenotype between genetics and disorders that signal a more acute underlying biological process. Popularly defined, endophenotypes are traits or behaviors that are heritable, manifest in the individual regardless of disorder onset, reproducibly associate with the behavior and have a higher prevalence in cases and probands than the general population (Gottesman & Gould, 2003). However, there have also been numerous other definitions that attempt to identify and pinpoint to a more functional and exact definition (Cannon & Keller, 2006; Iacono & Malone, 2011; K. S. Kendler & Neale, 2010). They can manifest in many forms, but commonly accepted ones include physiological measures like electroencephalogram (EEG) measures or other simple tasks. The antisaccade task is a classic example of an endophenotype associated with schizophrenia. In most antisaccade tasks, participants are shown a target (like a light source) either from the left or right of their vision. The participants are then tasked to inhibit their response to look towards the target and look in the opposite direction (Hutton & Ettinger, 2006). Patients and their relatives on average make more errors than the general population and this phenomenon has been documented and replicated in many independent experiments (Calkins, Curtis, Iacono, & Grove, 2004; Radant et al., 2010). Endophenotypes like antisaccade are viewed as a smaller and simpler component of larger complex traits like schizophrenia as it measures a more concise aspect, in this case, an impaired inhibitory function (Radant et al., 2010), which is commonly seen in schizophrenia patients. Moreover, these smaller measurable qualities are not only state-independent, but also thought to be potentially useful to identify susceptible individuals within a population without the need of a full diagnosis for the disorder. While there may be heterogeneity in a lot of complex trait disorders, endophenotypes have been consistently reproduced in association with the traits, similar to many biomarkers, such as cholesterol levels for cardiovascular diseases, that are used in the medical fields. Since these endophenotype are seen regardless of disease onset and are also highly heritable, they are hypothesized to be a simpler trait, wherein the effect size of each gene (or variant in gene) on the endophenotype is presumed to be much higher than the effect size for the complex trait of interest.

If genetic influence is hypothesized to be larger on the endophenotype then the sample size required to detect any effect of the genes decreases, and this assumption makes endophenotypes a very good candidate for genetic association studies. However, to date, endophenotypes have yet to find convincing biological mechanisms of complex trait (Iacono, Malone, & Vrieze, 2017). Previous genome-wide association studies (GWAS) surveying different

candidate endophenotypes in a community sample (with sample size over 4,000) (Iacono, Malone, Vaidyanathan, & Vrieze, 2014; Malone, Burwell, et al., 2014; Malone, McGue, & Iacono, 2017; Malone, Vaidyanathan, et al., 2014; Vaidyanathan, Isen, et al., 2014; Vrieze, Malone, et al., 2014) have had close to null results. Moreover, meta-analysis of different endophenotypes across studies have also not had any successes in variant discovery (Flint & Munafò, 2007). Contrary to our belief, the genetic effect sizes of loci associated with endophenotypes seem to be on the same scale as the effect sizes of the complex traits that they are associated with, many variants with small effect sizes. Endophenotypes consequently do not seem to have a simpler genetic architecture as previously hypothesized.

While endophenotype may not be helpful in genetic discovery, they can still shed light on the underlying biology that are not yet fully understood in these complex disorders. Our previous paper (M. Liu et al., 2017) focused on predicting endophenotypes by using polygenic risk scores (PRS) based on a large schizophrenia GWAS meta-analysis by the psychiatric genomics consortium. We hypothesized that by aggregating genetic variants associated with the disorder (as opposed to looking at a single variant or gene), we can construct a genetic liability index that is associated with the endophenotype. We found no significant correlation after correcting for multiple tests but the strength of PRS is highly dependent on a strong and well-powered discovery GWAS to provide the weights (Dudbridge, 2013). As components and proxies for more highly complex trait, more well-powered summary statistics may be needed to explain and understand these heritable and persisting endophenotype traits. Moreover, while we have previously focused on schizophrenia, there are still several other disorders linked to endophenotypes that have been studied over the years. In our current study, we have gathered summary statistics from 3 major GWAS meta-analysis efforts looking at 3 different set of complex traits including disorder, substance use and regular cognitive function.

Schizophrenia is most commonly associated with classic endophenotypes like antisaccade (patients and probands tend to have difficulty suppressing the saccade response) and P300 ERPs (patients and probands have lower amplitude). While our previous schizophrenia PRS did not predict any of our endophenotypes, we aim to expand upon our earlier paper by updating our previous schizophrenia PRS with a recent meta-analysis done by Pardiñas et. al (Pardiñas et al., 2018) which has a larger number of cases and improved methods that found more significant loci than the previous one from by the Psychiatric Genomics Consortium (Schizophrenia Working Group of the Psychiatric Genomics Consortium, 2014).

Earlier research have shown that candidate endophenotypes like P300 to be related to substance use disorders (Begleiter, Porjesz, Bihari, & Kissin, 1984; Iacono, 1998; Iacono & Malone, 2011; Malone, Iacono, & McGUE, 2001) where alcohol-dependent participants seem to have a reduced P300 amplitude. While there have not been large scale substance abuse GWAS, the GWAS

Sequencing Consortium of Alcohol and Nicotine (GSCAN) have recently conducted a large-scale GWAS meta-analysis of over 1.2 million people in phenotypes of alcohol and nicotine use (Mengzhen Liu et al., 2019). The phenotypes include commonly asked survey questions that measure heaviness of use (cigarettes per day and drinks per week), initiation of use (initiation of regular smoking and age of initiation of regular smoking) and cessation (smoking cessation). While most work has looked at substance abuse and endophenotypes, we hypothesize that given the larger sample size of the discovery cohort, we can predict the endophenotypes based on our substance use PRS.

Similarly, the Social Science Genetic Association Consortium also recently published a large 1.1 million people GWAS meta-analysis on years of educational attainment (EA) (Lee et al., 2018), which has been regarded as a proxy phenotype for cognitive ability (Lynn & Vanhanen, 2012). In their study, they have shown that not only is EA genetically correlated with cognitive ability but EA PRS also significantly predicts cognitive performance. There has been a number of studies that show relationship between endophenotypes and cognitive ability and IQ (Dauwels, Vialatte, & Cichocki, 2011; Thatcher, North, & Biver, 2005); moreover, EEG measures have also been shown to have high discriminative ability between participants with low IQ and high IQ (Thatcher et al., 2005). Given these relationships, we hypothesize that PRS of EA will also be significantly associated with our endophenotypes.

Given the vast amount of endophenotype and well-powered GWAS summary statistics data available, our main goal in this paper is to do a comprehensive analysis predicting endophenotypes with PRS based on schizophrenia, substance use and cognition.

Method

Sample and Endophenotypes

Participants were assessed as part of the Minnesota Center for Twin and Family Research (MCTFR), a community-based study of twins and their parents. Participants came in for in-person assessments, questionnaires and laboratory-based tests and then they were followed up in waves (Iacono et al., 2017, 2014; Iacono & McGue, 2002; Wilson et al., n.d.). Participants were genotyped on the Illumina 660W-Quad as described previously (Miller et al., 2012; Vrieze, Feng, et al., 2014, p. 20) and then imputed to the Haplotype Reference Consortium (McCarthy et al., 2016) panel using the Michigan imputation server (Das et al., 2016). We used a subset of the data that was primarily of European descent. We calculated four principal components (PC) on the European population in the 1000G (The 1000 Genomes Project Consortium, 2015) using PLINK (Chang et al., 2015) and projected the MCTFR genotypes based on the PC weights. We then selected only for participants that fall within each of the four 1000G European PCs for analysis.

The endophenotypes have been described previously (Iacono et al., 2014; Vrieze, Malone, et al., 2014). Out of the previously reported 17 endophenotypes, we chose not to include electro-dermal activity and acoustic startle and affect startle modulation to focus on the brain-based measures as brain-related genes and pathways have been implicated in enrichment analysis from the seven GWAS meta-analyses. All the endophenotypes were corrected for sex, age, age cohort and 10 principal components (PCs) and task-specific factors as described in previous papers. Here is a brief overview of the endophenotypes presented in this paper:

Antisaccade response. Participants are asked to fixate on a point in the center of their field of view. A light is flashed to either side, and participants are instructed to look away from the light. The antisaccade endophenotype is a proportion of trials in which they fail to inhibit their prepotent response to saccade toward the light. As mentioned earlier, there has been a number of studies that have linked antisaccade response to schizophrenia (Calkins et al., 2004; Levy, Mendell, & Holzman, 2004; McDowell et al., 2002; Radant et al., 2010) where patients and relatives have a higher proportion of errors as compared to controls.

Resting EEG. Participants are asked to relax with eyes closed for 5 minutes while listening to soft white noise. We obtained power in the alpha, beta, theta and delta frequency bands from a fast Fourier transformation of EEG at the Cz electrode. Schizophrenia has been associated with low-frequency power (Narayanan et al., 2014) and while not used directly as separate resting EEG powers, resting EEG have shown associations IQ (Langer et al., 2012; Thatcher et al., 2005).

P300 event-related potential (ERP). Participants are asked to complete a rotated heads visual oddball task (Begleiter et al., 1984). The P300 ERP is derived from the average ERP waveform of midline parietal electrode across all target trials. P300 has been associated with both schizophrenia and alcohol-dependency where in both, patients had reduced amplitude in the decision making process as compared to controls. (Bramon, Rabe-Hesketh, Sham, Murray, & Frangou, 2004; Malone et al., 2001).

Total energy and inter-trial phase coherence of Theta and Delta within the P3 window We derived inter-trial phase coherence (ITPC) and average time-frequency energy using a reduced interference distribution (RID) applied to the Rihaczek distribution (Aviyente, Bernat, Evans, & Sponheim, 2011) from the rotated heads visual oddball task. The P300 may be too downstream as an endophenotype, as it is not a unitary phenomenon and can be composed of many brain regions during recording (Malone et al., 2017). We chose to include these endophenotypes as components of P3 as there have been studies linking these endophenotype to alcohol dependence(Chen et al., 2009; Jones et al., 2004; Zlojutro et al., 2011).

SNP- and family-based heritability of endophenotypes

Heritability for each endophenotype was calculated using a method by Zaitlen et al. (Zaitlen et al., 2013), implemented in GCTA (Yang et al., 2010, 2011), where both SNP-based (phenotypic variance explained due to genotyped SNPs) heritability and narrow-sense (phenotypic variance explained due to additive genetics) heritability can be estimated from the same sample using both related and unrelated individuals. In their method, two matrices are constructed, an identity by state (IBS) matrix with a certain relatedness threshold, which is used as an approximation of identity by descent (IBD) matrix, and a full genetic relatedness matrix. Using these two covariance matrices, we can jointly estimate additive genetic heritability and SNP-based heritability. We used the genotyped data filtered by basic quality control metrics (removing variants that have a minor allele count less than 10, have a Hardy-Weinberg equilibrium less than 1e-6 and a call rate less than 0.9) to build the genetic relationship matrix (GRM) used by GCTA. As suggested in Zaitlen et al.'s paper, we used a relatedness threshold of 0.05 as a cutoff for unrelated individuals and set the GRM off-diagonal elements that are below the threshold to 0 to create the IBS matrix. We can also get, what is commonly referred to as, missing (or unexplained) heritability by subtracting the SNP-based heritability from the narrow-sense heritability.

Creation of polygenic risk scores

Summary statistics for the educational attainment GWAS was from the largest to date GWAS meta-analysis done on educational attainment (Lee et al., 2018). The present participants were included in this GWAS meta-analysis and, as such, we used a version of the publicly available summary statistics that had excluded MCTFR. Association summary statistics for schizophrenia Center for Neuropsychiatric Genetics and Genomics (https://walters.psycm.cf.ac.uk/) (Pardiñas et al., 2018). Polygenic scores for substance use were generated from summary statistics reported in (Liu et al., 2019). Once again, MCTFR was one of the discovery cohorts in this GWAS meta-analysis and, as such, was not included in the set of summary statistics used to create polygenic scores in the present sample Substance use phenotypes included age of initiation of regular smoking, cigarettes per day among smokers, smoking cessation (a binary phenotype of former v. current smoker) and smoking initiation (a binary phenotype of ever versus never regular smoker). There was also a measure of alcohol use, scaled as drinks per week.

We calculated the polygenic risk scores (PRS) using the software LDPred (Vilhjálmsson et al., 2015), a Bayesian method of PRS calculation that estimates posterior mean causal effect sizes from GWAS summary statistics conditioning on a point-normal mixture distribution for the genetic architecture of effects and a reference sample for LD patterns. We first pruned the MCTFR genotypes to only those with imputation quality score RSQ greater than 0.7. We then further limited the variants to those with a MAF > 0.01 and are present in HapMap3 since these tend to be variants that have stable and well-known properties. We assumed proportion of causal

variants to be 1 and the final PRS for Smoking Initiation (N = 1,225,910) contains 1,093,640 variants, Age of Smoking Initiation (N = 341,427) contains 1,093,797 variants, Cigarettes per Day (N = 337,334) contains 1,093,797 variants, Smoking Cessation (N = 547,219) contains 1,097,755 variants, Drinks per Week (N = 937,381) and 1,093636 variants, Educational Attainment (N = 762,526) contains 1,093,298 variants and Schizophrenia (N = 105,318) contains 1,073,315 variants.

We calculated correlations among PRSs and between the endophenotypes and the PRS using Rapid Feasible Generalized Least Squares, a R package running generalized least-squares regression method in families accounting for parents, monozygotic (MZ) twins and dizygotic (DZ) twins (Li, Basu, Miller, Iacono, & McGue, 2011). The endophenotypes and the PRSs were scaled to be have mean zero and variance of 1 prior to the regression analysis, such that the resulting slope is interpretable as a correlation coefficient with appropriate standard errors.

Results

Endophenotype descriptions

Table 2 shows a summary of the endophenotypes including their sample size, mean age, gender distribution and within family correlation. The parents are roughly uncorrelated, except for Delta at Cz Power. The MZ correlation is roughly twice the DZ correlation showing that the endophenotypes presented here are heritable, which agrees with the results from the heritability analysis in Table 3.

Heritability

Heritability results are shown in Table 3. Our narrow-sense heritability is very similar to the biometric twin-based heritabilities calculated in the previous papers (Malone, Burwell, et al., 2014; Malone et al., 2017; Malone, Vaidyanathan, et al., 2014; Vaidyanathan, Isen, et al., 2014; Vaidyanathan, Malone, et al., 2014). The heritability due to additive genetic variance is different since previous estimates all had relatively large standard errors (around 0.2 for most estimates). Almost all of the SNP-based heritability overlaps with previous estimates. The only exception is our P300 phenotype where while the narrow-sense heritability is very close to the previously reported twin-based heritability but the variance explained by genotyped SNPs increased drastically.

Polygenic risk scores

Figure 7 show the correlation matrix of the seven PRS with each other. The results are all in the same direction as the genetic correlation previously reported in the Liu et, al. 2019 paper for the significant correlations but with a lower magnitude. This is to be expected as this is a correlation of PRSs but it is supportive that the PRSs are constructed as expected. We see strong negative

correlations between educational attainment and our smoking behavior PRSs where the less one smokes (either starting later, smoking less cigarettes per day, quitting or just never initiating), the higher their predicted level of educational attainment, a common pattern seen in many studies (Wedow et al., 2018). Figure 8 shows the correlation between the endophenotypes on the Y-axis and the PRS on the X-axis. We do see some expected trends such as the negative correlation between the P300 endophenotypes and both, the drinks per week PRS and the schizophrenia PRS. None of the correlations are significant after correcting for multiple testing.

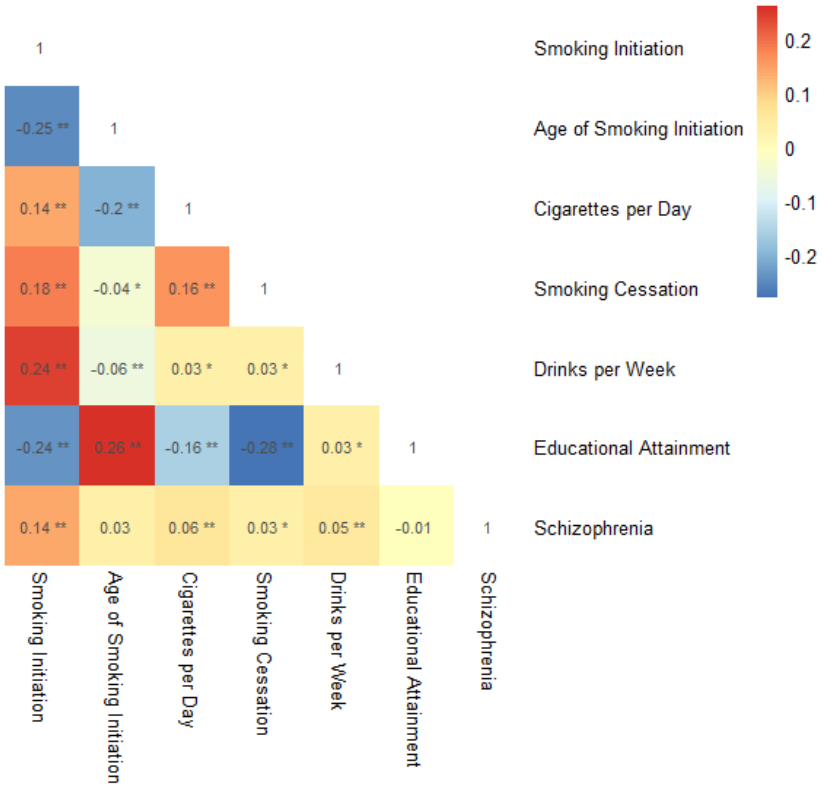

Fig. 7. Correlation between the PRSs. Each PRS was scaled with mean 0 and analysis was done using the RFGLS R package accounting for family structure. * denotes significance at $p < 0.05$ and ** denotes significance after Bonferoni correction $p < 0.003$.

Discussion

The predictive power of PRS is highly dependent on the size of the discovery sample. Previous attempt at predicting 8 of these 12 endophenotypes with PRS constructed using weights from a large schizophrenia GWAS meta-analysis was unsuccessful (M. Liu et al., 2017). In this paper, we have used more well-powered GWAS meta-analysis such as the GWAS meta-analysis of smoking initiation where the discovery sample size has exceeded 1 million participants. However,

despite the increase in discovery sample size, none of our PRS significantly predicts any of the endophenotypes post p-value correction. While the previous schizophrenia GWAS result did not predict any of the endophenotypes (M. Liu et al., 2017), the updated Schizophrenia GWAS results is negatively correlated with P300, which is consistent with previous literature (Ford, 1999; Jeon & Polich, 2003), and trending towards significance. This is encouraging news for continued studying of endophenotypes as by increasing the power of the discovery sample, we may be getting closer to understanding how endophenotypes and these complex traits are related.

Another aim of this paper is to potentially understand the genetics and biology underlying these endophenotypes through correlation with PRSs constructed from a related complex trait. In the case for P300, the underlying biology is not fully understood, but there has been data suggesting that it is caused by the effect of glutamatergic neurotransmission (Chen et al., 2009; Frodl-Bauch, Bottlender, & Hegerl, 1999; Zlojutro et al., 2011) and specific study have shown association between hippocampal glutamate and frontal theta activity in participants while doing simple tasks (Gallinat et al., 2006). Smoking initiation, the PRS with the largest sample size, show a negative correlation approaching significance with ITPC Theta at FZ and the discovery GWAS meta-analysis results for show an enrichment for glutamate-related pathways. These pathways have been implicated and studied extensively in substance abuse (Kalivas, 2009; Koob & Volkow, 2016) and here we present converging evidences from genes implicated in substance use to endophenotypes that have been implicated in substance use. Together, these converging evidences point towards an imbalance in glutamate that may be affecting the reduced amplitude of the P300.

In this paper we have also included total energy and inter-trial phase coherence of Theta and Delta within the P3 window since multiple brain regions are active during the visual oddball task and contribute to the P3 amplitude. These endophenotype measures are not only associated with the P3 amplitude but have also been previously studied to be associated with alcohol-dependence (Chen et al., 2009; Jones et al., 2004; Kang et al., 2010; Zlojutro et al., 2011). In our analysis, while drinks per week is not significantly associated with this endophenotype, it is still in the expected negative direction. Drinks per week in the Liu et al (2019) paper does not correlate with any psychiatric disorder which implies that it more generally captures regular alcohol use and not dependent or problematic use. This could indicate a difference in the underlying biology between becoming alcohol dependent and regular alcohol use. Another assumption of our study assumes that the same variants that significantly affect the complex trait directly influences the endophenotypes. There could have been an environmental effect wherein the environment actually mediates heaviness of drinking and our endophenotypes (K. S. Kendler & Neale, 2010).

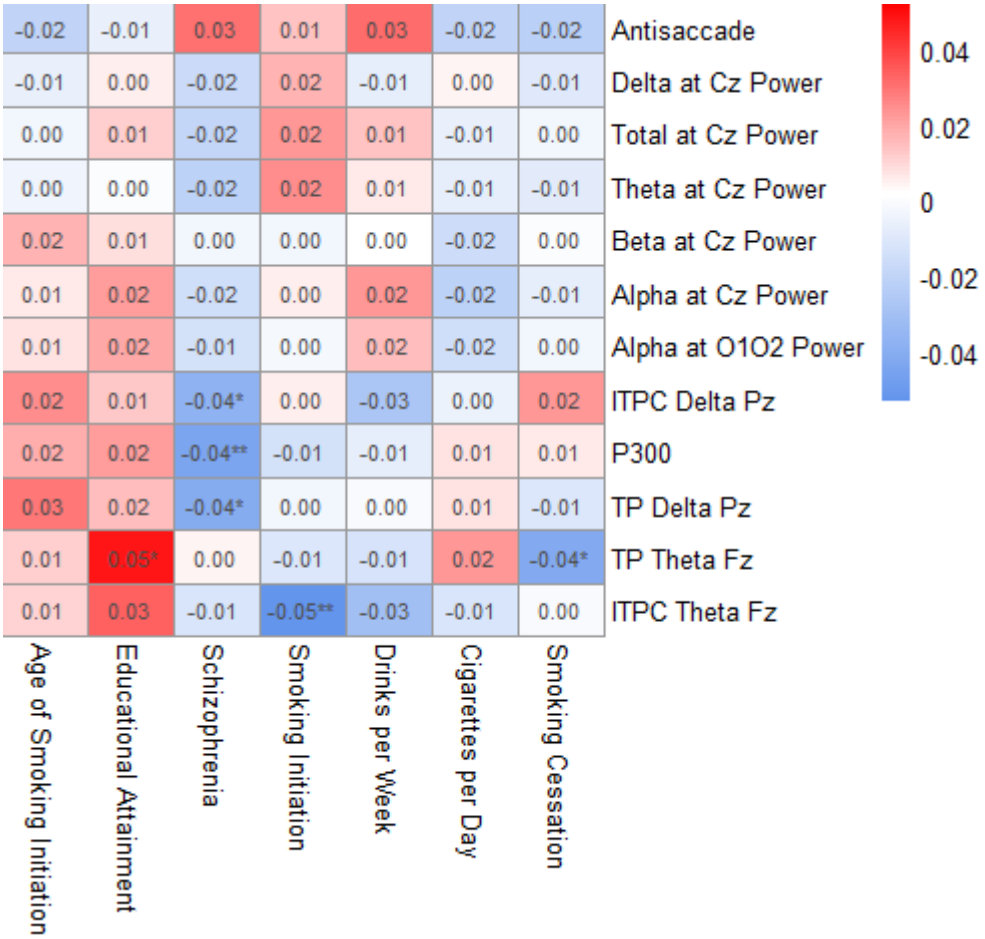

Fig. 8 Correlation between each PRS on the horizontal axis and endophenotype on the vertical axis. Both endophenotype and PRS was scaled with a mean 0 and analysis was done using the RFGLS R package accounting for family structure. * denotes significance at $p < 0.05$ and ** denotes significance at $p < 0.01$. No correlations remain significant after Bonferoni correction.

Limitations

One major limitation of our current study is that it is based on a community sample and many previous studies of endophenotypes are done on clinical probands and their relatives. Our study may have been limited in the power to detect any PRS-endophenotype correlation due to the healthy population, and these endophenotypes may manifest exclusively in populations with a heavier genetic burden for diseases. Since we assume that endophenotypes are supposedly more genetically simple, it could be argued that a PRS, which aggregates many variants across the genome, may not be a good fit to predict these simpler measures. By aggregating many other

variants that are not associated with the endophenotype but rather only with the phenotype, it may be adding noise and reducing the correlation. Another issue that deals more with the current state of GWAS is the lack of large non-European GWAS meta-analysis that exists. The GWAS meta-analysis we used here are all based on individuals with European descent thus limiting the target population to only Europeans as well. More research needs to be done in non-European populations in order to generalize the effectiveness of both the PRS and endophenotype to understanding the biological underpinnings of these complex traits.

Conclusion and Future Directions

While we did not find a significant prediction of endophenotypes by any of the PRS, the well-powered smoking initiation PRS and ITPC Theta Fz endophenotype supports existing theory of glutamate-related pathways influencing both P300 and substance abuse. As endophenotypes are measured and studied right now, it may not be feasible to use them to understand the biological processes of the complex traits that they are associated with. However, these measures could be used as another line of evidence for underlying cognitive processes in a similar method to how GWAS has helped understand how certain tissues or genes may be associated with the trait. Future studies may need to include a large set of endophenotypes to understand one complex trait. The UK Biobank, for example, has over 3,000 functional and structural brain imaging phenotypes (Elliott et al., 2018) that could be used to more finely pin down the cognitive processes of these complex traits

	N	*Mean Age*	*% Female*	*Mother-Father Corr.*	*Offspring-Mother Corr.*	*Offspring-Father Corr.*	*MZ Corr.*	*DZ Corr.*
Antisaccade	4457	28.98	44	0.04	0.24	0.17	0.53	0.18
P300	4155	29.01	44	0.00	0.26	0.19	0.64	0.39
Total Power Theta Fz	3420	29.27	50	-0.03	0.17	0.20	0.64	0.17
Total Power Delta Pz	4138	29.02	43	-0.03	0.18	0.16	0.61	0.34
ITPC Theta Fz	3427	29.28	50	-0.05	0.08	0.06	0.41	0.15
ITPC Delta Pz	4153	29.01	43	-0.04	0.11	0.04	0.46	0.20
Total at Cz Power	3938	28.75	44	-0.04	0.28	0.20	0.78	0.37
Alpha at Cz Power	3938	28.75	44	0.07	0.27	0.30	0.85	0.45
Alpha at O1O2 Power	3956	28.77	44	0.05	0.30	0.28	0.80	0.41
Beta at Cz Power	3938	28.75	44	0.00	0.36	0.23	0.85	0.38
Delta at Cz Power	3938	28.75	44	-0.12	0.21	0.08	0.56	0.24
Theta at Cz Power	3938	28.75	44	-0.07	0.22	0.15	0.73	0.36

Table 5. Summary of the endophenotypes included in the analysis. These are from the European subset of the MCTFR sample and we obtained the within family correlation from RFGLS output.

	SNP h^2	SE	unexplained h^2	SE	total h^2	SE
Antisaccade	0.22	0.10	0.29	0.10	0.51	0.02
P300	0.55	0.11	0.02	0.11	0.57	0.02
Total Power Theta Fz	0.23	0.13	0.38	0.13	0.62	0.02
Total Power Delta Pz	0.38	0.11	0.20	0.11	0.58	0.02
ITPC Theta Fz	0.11	0.12	0.24	0.12	0.35	0.03
ITPC Delta Pz	0.16	0.10	0.22	0.10	0.38	0.03
Total at Cz Power	0.26	0.11	0.54	0.11	0.80	0.01
Alpha at Cz Power	0.39	0.11	0.46	0.11	0.85	0.01
Alpha at O1O2 Power	0.39	0.11	0.39	0.11	0.78	0.01
Beta at Cz Power	0.39	0.11	0.48	0.11	0.87	0.01
Delta at Cz Power	0.22	0.11	0.33	0.11	0.55	0.03
Theta at Cz Power	0.23	0.11	0.50	0.11	0.73	0.02

Table 6. Heritability calculated using GCTA based on the whole sample.

Discussion

In this thesis, we examined the genetic architecture of alcohol and nicotine use by simultaneously doing an extensive meta-analysis to look for both common (GWAS meta-analysis) and rare (exome meta-analysis) variants that contribute to phenotypic variance. Based on the results from the two meta-analyses, the underlying genetic architecture of nicotine and alcohol is influenced by many common variants but no individual rare variants with large effect. The results from the GWAS meta-analysis were then used to construct a polygenic risk score (PRS) in a community sample to test for associations with substance use related endophenotypes.

We completed a GWAS meta-analysis at the same time as the exome meta-analysis on four nicotine use related phenotypes - age of regular smoking initiation (AgeSmk), smoking initiation (SmkInit), cigarettes smoked per day (CigDay) and smoking cessation (SmkCes) and one alcohol related phenotype, drinks per week (DrnkWk).

SmkInit is our most well-powered substance use meta-analysis. It is a common medical intake question and is comorbid with a wide number of psychiatric and medical diseases (Rojewski et al., 2016). Smoking initiation presents itself as a highly polygenic trait with many variants of small effects in contrast to CigDay or DrnkWk where we have a few loci with very large effects. In our first chapter, gene-based pathway analysis has shown us a complicated and intricate set of genes and systems implicated in smoking initiation. For instance, the dopamine reward system, which has been studied extensively in substance use as a key system that regulates addiction, is enriched in our smoking initiation results (Koob & Volkow, 2010). Another widely studied system is the glutamatergic system, which has also been implicated in not only human addiction research but also mice models as well (Kalivas, 2009). These results serve as an important bridge in translating how genotypes may affect these systems and thus affect behavior.

CigDay is an indication of heaviness of use of nicotine. We have replicated many well-studied genes such as the *CHRNA5-CHRNA3-CHRNB4* region and the *CYP2A6* region in both GWAS and exome meta-analysis. In the exome meta-analysis, we have also included pack-years as an indication of heaviness of use as it is commonly used in epidemiological and clinical studies. However, it does not perform as well as CigDay in variant discovery. In the GWAS meta-analysis, we have also found that all central nervous system nicotinic receptor genes (except *CHRNA7)* were significantly associated with CigDay.

The discovered loci in SmkCes is very similar to the results from CigDay. In this aspect, it is interesting to note that quitting smoking may have some underlying biology that is associated with the metabolism of nicotine. We have replicated previously reported loci for DrnkWk like *ADH1B, GCKR,* and *KLB.* An estimated 30% of alcohol users would show symptoms of alcohol use

disorder in their lifetime (Grant et al., 2015) and previous research has shown that alcohol dependence and consumption have a very high genetic correlation estimated at 0.97 (Grant et al., 2009). In the GWAS meta-analysis, DrnkWk measures alcohol consumption across the population and is not limited to those with problematic use. Despite consumption being correlated with dependence, we do not see genetic correlation with any of the medical diseases or any behavioral traits (except risk tolerance and lifetime cannabis use). Moreover, DrnkWk PRS does not predict any of the endophenotypes associated with alcohol dependence. There may be a more nuanced relationship between alcohol consumption and dependence that needs to be studied more in depth.

We also performed an exome meta-analysis on four nicotine related phenotypes – smoking initiation (SmkInit), cigarettes per day (CigDay), pack-years (PY) and smoking cessation (SmkCes) in 622,409 individuals and found 40 novel loci but no rare variants that replicated. The replicated results are all common variants which have also been implicated in the common variant GWAS meta-analysis. The one rare novel variant that was found in CigDay discovery is not significant when replicated in another replication sample, but we did discover a conditionally independent rare variant near a known gene, *CYP2A6*. The lack of rare variant discovery could be due to our sample size (which is half the sample size of the common variant GWAS meta-analysis) or that rare functional variants may not actually be in the exome but acts on a gene some distance away from the gene. In order to find these non-exomic rare variants, a next step could be to inspect GWAS with whole-genome sequences. The Trans-Omics for Precision Medicine (TOPMed) has whole-genome sequenced 53,581 participants and found that 97% of its >400 million variants have frequencies < 1% (Taliun et al., 2019) indicating just how many more rare and undiscovered variations are in the genome beyond the exome.

As the size of the discovery sample increases, the results become more robust and thus the use of these results in subsequent analysis becomes feasible. Cell and tissue enrichment analysis results showed a significant enrichment in the central nervous system across all five phenotypes, particularly in tissues from the cortical and sub-cortical regions. One way to understand these underlying biological mechanisms is to look at endophenotypes that act as a simpler measurable trait that is hypothesized to be affected by the same biology as the complex trait it is associated with. Endophenotypes associated with substance use, such as the P300, are assumed to be less complex thus more directly influenced by genes than the complex trait that they tag. In Chapter 3, we tested if the GWAS meta-analysis would predict any of these endophenotypes in a community-representative sample. None of the associations were significant after correcting for multiple testing, but there was converging evidence between one endophenotype, P300, and SmkInit that implicate glutamatergic neurotransmission system as significant in substance use.

Future Directions and Conclusion

The meta-analyses presented here have primarily been conducted on samples of individuals restricted to European ancestry. However, there are many more variants that are private to other populations and continents which are missed by excluding them from major studies (The 1000 Genomes Project Consortium, 2015). For example, variations in *ALDH2,* a well-studied gene involved in alcohol metabolism, is almost entirely private to East Asian populations and thus missing from our DrnkWk meta-analysis despite the sample size being close to a million individuals. Similarly, PRSs do not work well when the discovery population and the target population are different (Duncan et al., 2019). More diverse populations are needed in order to fully understand the individual differences between individuals.

Overall, the nicotine use phenotypes are significantly correlated with each other and a lot of nicotinic receptors are pleiotropic across the four smoking traits and particularly between CigDay, SmkCes and SmkInit. DrnkWk is most significantly genetically correlated with SmkInit but not so much with the other smoking phenotypes. This is consistent with current research where alcohol consumption is comorbid with smoking initiation (Meyerhoff et al., 2006) however, the biology underlying the two traits are dissimilar (Dani & Harris, 2005). In conclusion, based on the distribution of effect sizes, the genetic underpinning of substance use is split where heaviness of use phenotypes seems to be influenced by a few loci of large genetic effects (eg. *CHRNA3* for CigDay and *ADH1B* for DrnkWk) and initiation and cessation of use are more polygenic traits influenced by many variants with small effects.

Aberg, K. A., Liu, Y., Bukszár, J., McClay, J. L., Khachane, A. N., Andreassen, O. A., … Oord, E. J. van den. (2013). A Comprehensive Family-Based Replication Study of Schizophrenia Genes. JAMA Psychiatry, 70(6), 573–581. https://doi.org/10.1001/jamapsychiatry.2013.288

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature Methods, 7(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Aoun, E. G., Jimenez, V. A., Vendruscolo, L. F., Walter, N. a. R., Barbier, E., Ferrulli, A., … Leggio, L. (2018). A relationship between the aldosterone-mineralocorticoid receptor pathway and alcohol drinking: Preliminary translational findings across rats, monkeys and humans. Molecular Psychiatry, 23(6), 1466–1473. https://doi.org/10.1038/mp.2017.97

Aviyente, S., Bernat, E. M., Evans, W. S., & Sponheim, S. R. (2011). A phase synchrony measure for quantifying dynamic functional integration in the brain. Human Brain Mapping, 32(1), 80–93. https://doi.org/10.1002/hbm.21000

Battle, A., Brown, C. D., Engelhardt, B. E., & Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. Nature, 550. https://doi.org/10.1038/nature24277

Begleiter, H., Porjesz, B., Bihari, B., & Kissin, B. (1984). Event-related brain potentials in boys at risk for alcoholism. Science, 225(4669), 1493–1496. https://doi.org/10.1126/science.6474187

Bierut, L. J., Goate, A. M., Breslau, N., Johnson, E. O., Bertelsen, S., Fox, L., … Edenberg, H. J. (2012). ADH1B is associated with alcohol dependence and alcohol consumption in populations of European and African ancestry. Molecular Psychiatry, 17(4), 445–450. https://doi.org/10.1038/mp.2011.124

Bloom, A. J., Baker, T. B., Chen, L.-S., Breslau, N., Hatsukami, D., Bierut, L. J., & Goate, A. (2014). Variants in two adjacent genes, EGLN2 and CYP2A6, influence smoking behavior related to disease risk via different mechanisms. Human Molecular Genetics, 23(2), 555–561. https://doi.org/10.1093/hmg/ddt432

Bowden, J., Davey Smith, G., & Burgess, S. (2015). Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression. International Journal of Epidemiology, 44(2), 512–525. https://doi.org/10.1093/ije/dyv080

Bowden, J., Smith, G. D., Haycock, P. C., & Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. Genetic Epidemiology, 40(4), 304–314. https://doi.org/10.1002/gepi.21965

Boyden, L. M., Choi, M., Choate, K. A., Nelson-Williams, C. J., Farhi, A., Toka, H. R., … Lifton, R. P. (2012). Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. Nature, 482(7383), 98–102. https://doi.org/10.1038/nature10814

Bramon, E., Rabe-Hesketh, S., Sham, P., Murray, R. M., & Frangou, S. (2004). Meta-analysis of the P300 and P50 waveforms in schizophrenia. Schizophrenia Research, 70(2–3), 315–329. https://doi.org/10.1016/j.schres.2004.01.004

Brazel, D. M., Jiang, Y., Hughey, J. M., Turcot, V., Zhan, X., Gong, J., … Vrieze, S. (2019). Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use. Biological Psychiatry, 85(11), 946–955. https://doi.org/10.1016/j.biopsych.2018.11.024

Bulik-Sullivan, B. K., Loh, P.-R., Finucane, H. K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics Consortium, … Neale, B. M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics, 47(3), 291–295. https://doi.org/10.1038/ng.3211

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., … Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. Nature, 562(7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z

Calkins, M. E., Curtis, C. E., Iacono, W. G., & Grove, W. M. (2004). Antisaccade performance is impaired in medically and psychiatrically healthy biological relatives of schizophrenia patients. Schizophrenia Research, 71(1), 167–178. https://doi.org/10.1016/j.schres.2003.12.005

Cannon, T. D., & Keller, M. C. (2006). Endophenotypes in the Genetic Analyses of Mental Disorders. Annual Review of Clinical Psychology, 2(1), 267–290. https://doi.org/10.1146/annurev.clinpsy.2.022305.095232

Carmelli, D., Swan, G. E., Robinette, D., & Fabsitz, R. (1992). Genetic Influence on Smoking—A Study of Male Twins. New England Journal of Medicine, 327(12), 829–833. https://doi.org/10.1056/NEJM199209173271201

Centers for Disease Control and Prevention (CDC). (2008). Cigarette smoking among adults—United States, 2007. MMWR. Morbidity and Mortality Weekly Report, 57(45), 1221.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. GigaScience, 4, 7. https://doi.org/10.1186/s13742-015-0047-8

Chen, A. C. H., Tang, Y., Rangaswamy, M., Wang, J. C., Almasy, L., Foroud, T., … Porjesz, B. (2009). Association of single nucleotide polymorphisms in a glutamate receptor gene (GRM8) with theta power of event-related oscillations and alcohol dependence. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 150B(3), 359–368. https://doi.org/10.1002/ajmg.b.30818

Costas, J. (2018). The highly pleiotropic gene SLC39A8 as an opportunity to gain insight into the molecular pathogenesis of schizophrenia. American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics, 177(2), 274–283. https://doi.org/10.1002/ajmg.b.32545

Dani, J. A., & Harris, R. A. (2005). Nicotine addiction and comorbidity with alcohol abuse and mental illness. Nature Neuroscience, 8(11), 1465–1470. https://doi.org/10.1038/nn1580

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., … Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. Nature Genetics, 48(10), 1284–1287. https://doi.org/10.1038/ng.3656

Dauwels, J., Vialatte, F.-B., & Cichocki, A. (2011). On the Early Diagnosis of Alzheimer's Disease from EEG Signals: A Mini-Review. In R. Wang & F. Gu (Eds.), Advances in Cognitive Neurodynamics (II) (pp. 709–716). https://doi.org/10.1007/978-90-481-9695-1_106

Demontis, D., Rajagopal, V. M., Thorgeirsson, T. E., Als, T. D., Grove, J., Leppälä, K., … Børglum, A. D. (2019). Genome-wide association study implicates CHRNA2 in cannabis use disorder. Nature Neuroscience, 22(7), 1066–1074. https://doi.org/10.1038/s41593-019-0416-1

Dudbridge, F. (2013). Power and Predictive Accuracy of Polygenic Risk Scores. PLOS Genetics, 9(3), e1003348. https://doi.org/10.1371/journal.pgen.1003348

Duncan, L., Shen, H., Gelaye, B., Meijsen, J., Ressler, K., Feldman, M., … Domingue, B. (2019). Analysis of polygenic risk score usage and performance in diverse human populations. Nature Communications, 10(1), 3328. https://doi.org/10.1038/s41467-019-11112-0

Edenberg, H. J. (2007). The genetics of alcohol metabolism: Role of alcohol dehydrogenase and aldehyde dehydrogenase variants. Alcohol Research & Health: The Journal of the National Institute on Alcohol Abuse and Alcoholism, 30(1), 5–13.

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., … Smith, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature, 562(7726), 210–216. https://doi.org/10.1038/s41586-018-0571-7

Ezzati, M., Lopez, A. D., Rodgers, A., Vander Hoorn, S., & Murray, C. J. (2002). Selected major risk factors and global and regional burden of disease. The Lancet, 360(9343), 1347–1360. https://doi.org/10.1016/S0140-6736(02)11403-6

Feldstein Ewing, S. W., Sakhardande, A., & Blakemore, S.-J. (2014). The effect of alcohol consumption on the adolescent brain: A systematic review of MRI and fMRI studies of alcohol-using youth. NeuroImage: Clinical, 5, 420–437. https://doi.org/10.1016/j.nicl.2014.06.011

Feng, S., Liu, D., Zhan, X., Wing, M. K., & Abecasis, G. R. (2014). RAREMETAL: Fast and powerful meta-analysis for rare variants. Bioinformatics, 30(19), 2828–2829. https://doi.org/10.1093/bioinformatics/btu367

Fernandez, E., Schiappa, R., Girault, J.-A., & Le Novère, N. (2006). DARPP-32 is a robust integrator of dopamine and glutamate signals. PLoS Computational Biology, 2(12), e176. https://doi.org/10.1371/journal.pcbi.0020176

Finucane, H. K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.-R., … Price, A. L. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics, 47(11), 1228–1235. https://doi.org/10.1038/ng.3404

Flint, J., & Munafò, M. R. (2007). The endophenotype concept in psychiatric genetics. Psychological Medicine, 37(2), 163–180. https://doi.org/10.1017/S0033291706008750

Ford, J. M. (1999). Schizophrenia: The broken P300 and beyond. Psychophysiology, 36(6), 667–682. https://doi.org/10.1111/1469-8986.3660667

Frodl-Bauch, T., Bottlender, R., & Hegerl, U. (1999). Neurochemical Substrates and Neuroanatomical Generators of the Event-Related P300. Neuropsychobiology, 40(2), 86–94. https://doi.org/10.1159/000026603

Gage, S. H., Jones, H. J., Taylor, A. E., Burgess, S., Zammit, S., & Munafò, M. R. (2017). Investigating causality in associations between smoking initiation and schizophrenia using Mendelian randomization. Scientific Reports, 7(1), 1–8. https://doi.org/10.1038/srep40653

Gallinat, J., Kunz, D., Senkowski, D., Kienast, T., Seifert, F., Schubert, F., & Heinz, A. (2006). Hippocampal glutamate concentration predicts cerebral theta oscillations during cognitive processing. Psychopharmacology, 187(1), 103–111. https://doi.org/10.1007/s00213-006-0397-0

Gass, J. T., & Olive, M. F. (2008). Glutamatergic substrates of drug addiction and alcoholism. Biochemical Pharmacology, 75(1), 218–265. https://doi.org/10.1016/j.bcp.2007.06.039

Goldstein, R. Z., & Volkow, N. D. (2011). Dysfunction of the prefrontal cortex in addiction: Neuroimaging findings and clinical implications. Nature Reviews. Neuroscience, 12(11), 652–669. https://doi.org/10.1038/nrn3119

Gottesman, I. I., & Gould, T. D. (2003). The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. American Journal of Psychiatry, 160(4), 636–645. https://doi.org/10.1176/appi.ajp.160.4.636

Grant, J. D., Agrawal, A., Bucholz, K. K., Madden, P. A. F., Pergadia, M. L., Nelson, E. C., … Heath, A. C. (2009). Alcohol Consumption Indices of Genetic Risk for Alcohol Dependence. Biological Psychiatry, 66(8), 795–800. https://doi.org/10.1016/j.biopsych.2009.05.018

GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx)

groups, NIH Common Fund, NIH/NCI, … Montgomery, S. B. (2017). Genetic effects on gene expression across human tissues. Nature, 550(7675), 204–213. https://doi.org/10.1038/nature24277

Hancock, D. B., Guo, Y., Reginsson, G. W., Gaddis, N. C., Lutz, S. M., Sherva, R., … Johnson, E. O. (2018). Genome-wide association study across European and African American ancestries identifies a SNP in DNMT3B contributing to nicotine dependence. Molecular Psychiatry, 23(9), 1–9. https://doi.org/10.1038/mp.2017.193

Harris, K. M., Halpern, C. T., Haberstick, B. C., & Smolen, A. (2013). The National Longitudinal Study of Adolescent Health (Add Health) Sibling Pairs Data. Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies, 16(1), 391–398. https://doi.org/10.1017/thg.2012.137

Hecht, S. S. (1999). Tobacco Smoke Carcinogens and Lung Cancer. JNCI: J Natl Cancer Inst, 91. https://doi.org/10.1093/jnci/91.14.1194

Hemani, G., Zheng, J., Elsworth, B., Wade, K. H., Haberland, V., Baird, D., … Haycock, P. C. (2018). The MR-Base platform supports systematic causal inference across the human phenome. ELife, 7, e34408. https://doi.org/10.7554/eLife.34408

Hicks, B. M., Schalet, B. D., Malone, S. M., Iacono, W. G., & McGue, M. (2011). Psychometric and genetic architecture of substance use disorder and behavioral disinhibition measures for gene association studies. Behavior Genetics, 41(4), 459–475. https://doi.org/10.1007/s10519-010-9417-2

Hoffmann, T. J., Ehret, G. B., Nandakumar, P., Ranatunga, D., Schaefer, C., Kwok, P.-Y., … Risch, N. (2017). Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nature Genetics, 49(1), 54–64. https://doi.org/10.1038/ng.3715

Hutton, S. B., & Ettinger, U. (2006). The antisaccade task as a research tool in psychopathology: A critical review. Psychophysiology, 43(3), 302–313. https://doi.org/10.1111/j.1469-8986.2006.00403.x

Hyman, S. E., Malenka, R. C., & Nestler, E. J. (2006). Neural mechanisms of addiction: The role of reward-related learning and memory. Annual Review of Neuroscience, 29, 565–598. https://doi.org/10.1146/annurev.neuro.29.051605.113009

Iacono, W. G. (1998). Identifying psychophysiological risk for psychopathology: Examples from substance abuse and schizophrenia research. Psychophysiology, 35(6), 621–637. https://doi.org/10.1111/1469-8986.3560621

Iacono, W. G., & Malone, S. M. (2011). Developmental Endophenotypes: Indexing Genetic Risk for Substance Abuse With the P300 Brain Event-Related Potential. Child Development Perspectives, 5(4), 239–247. https://doi.org/10.1111/j.1750-8606.2011.00205.x

Iacono, W. G., Malone, S. M., & Vrieze, S. I. (2017). Endophenotype best practices. International Journal of Psychophysiology, 111, 115–144. https://doi.org/10.1016/j.ijpsycho.2016.07.516

Iacono, W. G., & McGue, M. (2002). Minnesota Twin Family Study. Twin Research and Human Genetics, 5, 482–487.

Iacono, William. G., Malone, Stephen. M., Vaidyanathan, U., & Vrieze, S. I. (2014). Genome-wide scans of genetic variants for psychophysiological endophenotypes: A methodological overview. Psychophysiology, 51(12), 1207–1224. https://doi.org/10.1111/psyp.12343

Jeon, Y.-W., & Polich, J. (2003). Meta-analysis of P300 and schizophrenia: Patients, paradigms, and practical implications. Psychophysiology, 40(5), 684–701.

Jiang, Y., Chen, S., McGuire, D., Chen, F., Liu, M., Iacono, W. G., … Liu, D. J. (2018). Proper conditional analysis in the presence of missing data: Application to large scale meta-

analysis of tobacco use phenotypes. PLOS Genetics, 14(7), e1007452. https://doi.org/10.1371/journal.pgen.1007452

Jones, K. A., Porjesz, B., Almasy, L., Bierut, L., Goate, A., Wang, J. C., … Begleiter, H. (2004). Linkage and linkage disequilibrium of evoked EEG oscillations with CHRM2 receptor gene polymorphisms: Implications for human brain dynamics and cognition. International Journal of Psychophysiology, 53(2), 75–90. https://doi.org/10.1016/j.ijpsycho.2004.02.004

Jorgenson, E., Thai, K. K., Hoffmann, T. J., Sakoda, L. C., Kvale, M. N., Banda, Y., … Choquet, H. (2017). Genetic contributors to variation in alcohol consumption vary by race/ethnicity in a large multi-ethnic genome-wide association study. Molecular Psychiatry, 22(9), 1359–1367. https://doi.org/10.1038/mp.2017.101

Kalivas, P. W. (2009). The glutamate homeostasis hypothesis of addiction. Nature Reviews. Neuroscience, 10(8), 561–572. https://doi.org/10.1038/nrn2515

Kamburov, A., Wierling, C., Lehrach, H., & Herwig, R. (2009). ConsensusPathDB—a database for integrating human functional interaction networks. Nucleic Acids Research, 37(suppl_1), D623–D628. https://doi.org/10.1093/nar/gkn698

Kaprio, J., Koskenvuo, M., & Sarna, S. (1981). Cigarette smoking, use of alcohol, and leisure-time physical activity among same-sexed adult male twins. Progress in Clinical and Biological Research, 69 Pt C, 37–46.

Kelly, C., & McCreadie, R. (2000). Cigarette smoking and schizophrenia. Advances in Psychiatric Treatment, 6(5), 327–331. https://doi.org/10.1192/apt.6.5.327

Kendler, K. S., & Neale, M. C. (2010). Endophenotype: A conceptual analysis. Molecular Psychiatry, 15(8), 789–797. https://doi.org/10.1038/mp.2010.8

Kendler, Kenneth S., Prescott, C. A., Myers, J., & Neale, M. C. (2003). The structure of genetic and environmental risk factors for common psychiatric and substance use disorders in men and women. Archives of General Psychiatry, 60(9), 929–937. https://doi.org/10.1001/archpsyc.60.9.929

Kendler, Kenneth S., Schmitt, E., Aggen, S. H., & Prescott, C. A. (2008). Genetic and environmental influences on alcohol, caffeine, cannabis, and nicotine use from early adolescence to middle adulthood. Archives of General Psychiatry, 65(6), 674–682. https://doi.org/10.1001/archpsyc.65.6.674

Keskitalo, K., Broms, U., Heliövaara, M., Ripatti, S., Surakka, I., Perola, M., … Kaprio, J. (2009). Association of serum cotinine level with a cluster of three nicotinic acetylcholine receptor genes (CHRNA3/CHRNA5/CHRNB4) on chromosome 15. Human Molecular Genetics, 18(20), 4007–4012. https://doi.org/10.1093/hmg/ddp322

Kong, A., Thorleifsson, G., Frigge, M. L., Vilhjalmsson, B. J., Young, A. I., Thorgeirsson, T. E., … Stefansson, K. (2018). The nature of nurture: Effects of parental genotypes. Science (New York, N.Y.), 359(6374), 424–428. https://doi.org/10.1126/science.aan6877

Koob, G. F., & Volkow, N. D. (2010). Neurocircuitry of addiction. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 35(1), 217–238. https://doi.org/10.1038/npp.2009.110

Koob, G. F., & Volkow, N. D. (2016). Neurobiology of addiction: A neurocircuitry analysis. The Lancet. Psychiatry, 3(8), 760–773. https://doi.org/10.1016/S2215-0366(16)00104-8

Kumasaka, N., Aoki, M., Okada, Y., Takahashi, A., Ozaki, K., Mushiroda, T., … Kubo, M. (2012). Haplotypes with Copy Number and Single Nucleotide Polymorphisms in CYP2A6 Locus Are Associated with Smoking Quantity in a Japanese Population. PLOS ONE, 7(9), e44507. https://doi.org/10.1371/journal.pone.0044507

Langer, N., Pedroni, A., Gianotti, L. R. R., Hänggi, J., Knoch, D., & Jäncke, L. (2012). Functional brain network efficiency predicts intelligence. Human Brain Mapping, 33(6), 1393–1406. https://doi.org/10.1002/hbm.21297

Lassi, G., Taylor, A. E., Timpson, N. J., Kenny, P. J., Mather, R. J., Eisen, T., & Munafò, M. R. (2016). The CHRNA5–A3–B4 Gene Cluster and Smoking: From Discovery to Therapeutics. Trends in Neurosciences, 39(12), 851–861. https://doi.org/10.1016/j.tins.2016.10.005

Lee, J. J., 23andMe Research Team, COGENT (Cognitive Genomics Consortium), Social Science Genetic Association Consortium, Wedow, R., Okbay, A., … Cesarini, D. (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nature Genetics, 50(8), 1112–1121. https://doi.org/10.1038/s41588-018-0147-3

Levy, D. L., Mendell, N. R., & Holzman, P. S. (2004). The antisaccade task and neuropsychological tests of prefrontal cortical integrity in schizophrenia: Empirical findings and interpretative considerations. World Psychiatry, 3(1), 32–40.

Li, X., Basu, S., Miller, M. B., Iacono, W. G., & McGue, M. (2011). A Rapid Generalized Least Squares Model for a Genome-Wide Quantitative Trait Association Analysis in Families. Human Heredity, 71(1), 67–82. https://doi.org/10.1159/000324839

Litten, R. Z., Ryan, M. L., Fertig, J. B., Falk, D. E., Johnson, B., Dunn, K. E., … Stout, R. (2013). A Double-Blind, Placebo-Controlled Trial Assessing the Efficacy of Varenicline Tartrate for Alcohol Dependence. Journal of Addiction Medicine, 7(4), 277–286. https://doi.org/10.1097/ADM.0b013e31829623f4

Liu, M., Malone, S. M., Vaidyanathan, U., Keller, M. C., Abecasis, G., McGue, M., … Vrieze, S. I. (2017). Psychophysiological endophenotypes to characterize mechanisms of known schizophrenia genetic loci. Psychological Medicine, 47(6), 1116–1125. https://doi.org/10.1017/S0033291716003184

Liu, Mengzhen, Jiang, Y., Wedow, R., Li, Y., Brazel, D. M., Chen, F., … Vrieze, S. (2019). Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nature Genetics, 51(2), 237. https://doi.org/10.1038/s41588-018-0307-5

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., … Speliotes, E. K. (2015). Genetic studies of body mass index yield new insights for obesity biology. Nature, 518(7538), 197–206. https://doi.org/10.1038/nature14177

Logrip, M. L., Koob, G. F., & Zorrilla, E. P. (2011). Role of corticotropin-releasing factor in drug addiction: Potential for pharmacological intervention. CNS Drugs, 25(4), 271–287. https://doi.org/10.2165/11587790-000000000-00000

Loukola, A., Buchwald, J., Gupta, R., Palviainen, T., Hällfors, J., Tikkanen, E., … Kaprio, J. (2015). A Genome-Wide Association Study of a Biomarker of Nicotine Metabolism. PLOS Genetics, 11(9), e1005498. https://doi.org/10.1371/journal.pgen.1005498

Lynn, R., & Vanhanen, T. (2012). National IQs: A review of their educational, cognitive, economic, political, demographic, sociological, epidemiological, geographic and climatic correlates. Intelligence, 40(2), 226–234. https://doi.org/10.1016/j.intell.2011.11.004

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., … Parkinson, H. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Research, 45(D1), D896–D901. https://doi.org/10.1093/nar/gkw1133

Madsen, B. E., & Browning, S. R. (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLOS Genetics, 5(2), e1000384. https://doi.org/10.1371/journal.pgen.1000384

Malone, S. M., Burwell, S. J., Vaidyanathan, U., Miller, M. B., Mcgue, M., & Iacono, W. G. (2014). Heritability and molecular-genetic basis of resting EEG activity: A genome-wide association study. Psychophysiology, 51(12), 1225–1245. https://doi.org/10.1111/psyp.12344

Malone, S. M., Iacono, W. G., & McGUE, M. (2001). Event-related potentials and comorbidity in alcohol-dependent adult males. Psychophysiology, 38(3), 367–376. https://doi.org/10.1111/1469-8986.3830367

Malone, S. M., McGue, M., & Iacono, W. G. (2017). What can time-frequency and phase coherence measures tell us about the genetic basis of P3 amplitude? International Journal of Psychophysiology, 115, 40–56. https://doi.org/10.1016/j.ijpsycho.2016.11.008

Malone, S. M., Vaidyanathan, U., Basu, S., Miller, M. B., Mcgue, M., & Iacono, W. G. (2014). Heritability and molecular-genetic basis of the P3 event-related brain potential: A genome-wide association study. Psychophysiology, 51(12), 1246–1258. https://doi.org/10.1111/psyp.12345

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., … Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics, 48(10), 1279–1283. https://doi.org/10.1038/ng.3643

McDowell, J. E., Brown, G. G., Paulus, M., Martinez, A., Stewart, S. E., Dubowitz, D. J., & Braff, D. L. (2002). Neural correlates of refixation saccades and antisaccades in normal and schizophrenia subjects. Biological Psychiatry, 51(3), 216–223. https://doi.org/10.1016/S0006-3223(01)01204-5

McKay, J. D., Hung, R. J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D. C., … Amos, C. I. (2017). Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. Nature Genetics, 49(7), 1126–1132. https://doi.org/10.1038/ng.3892

Meyerhoff, D. J., Tizabi, Y., Staley, J. K., Durazzo, T. C., Glass, J. M., & Nixon, S. J. (2006). Smoking Comorbidity in Alcoholism: Neurobiological and Neurocognitive Consequences. Alcoholism: Clinical and Experimental Research, 30(2), 253–264. https://doi.org/10.1111/j.1530-0277.2006.00034.x

Miller, M. B., Basu, S., Cunningham, J., Eskin, E., Malone, S. M., Oetting, W. S., … Mcgue, M. (2012). The Minnesota Center for Twin and Family Research Genome-Wide Association Study. Twin Research and Human Genetics : The Official Journal of the International Society for Twin Studies, 15(6), 767–774. https://doi.org/10.1017/thg.2012.62

Morris, A. P., & Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic Epidemiology, 34(2), 188–193. https://doi.org/10.1002/gepi.20450

Munafo, M. R., Tilling, K., Taylor, A. E., Evans, D. M., & Davey Smith, G. (2018). Collider scope: When selection bias can substantially influence observed associations. Int J Epidemiol, 47. https://doi.org/10.1093/ije/dyx206

Ng, B., White, C. C., Klein, H.-U., Sieberts, S. K., McCabe, C., Patrick, E., … Jager, P. L. D. (2017). An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nature Neuroscience, 20(10), 1418–1426. https://doi.org/10.1038/nn.4632

Ockene, I. S., & Miller, N. H. (1997). Cigarette Smoking, Cardiovascular Disease, and Stroke. A Statement Healthc Prof Am Heart Assoc, 96.

O'Donnell, C. J., & Nabel, E. G. (2011). Genomics of Cardiovascular Disease. New England Journal of Medicine, 365(22), 2098–2109. https://doi.org/10.1056/NEJMra1105239

Okbay, A., Beauchamp, J. P., Fontana, M. A., Lee, J. J., Pers, T. H., Rietveld, C. A., … Benjamin, D. J. (2016). Genome-wide association study identifies 74 loci associated with educational attainment. Nature, 533(7604), 539–542. https://doi.org/10.1038/nature17671

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., … Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., & Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. Bioinformatics, 32(10), 1479–1485. https://doi.org/10.1093/bioinformatics/btv722

Pardiñas, A. F., GERAD1 Consortium, CRESTAR Consortium, Holmans, P., Pocklington, A. J., Escott-Price, V., … Walters, J. T. R. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nature Genetics, 50(3), 381–389. https://doi.org/10.1038/s41588-018-0059-2

Perroy, J., Adam, L., Qanbar, R., Chénier, S., & Bouvier, M. (2003). Phosphorylation-independent desensitization of GABAB receptor by GRK4. The EMBO Journal, 22(15), 3816–3824. https://doi.org/10.1093/emboj/cdg383

Polderman, T. J. C., Benyamin, B., de Leeuw, C. A., Sullivan, P. F., van Bochoven, A., Visscher, P. M., & Posthuma, D. (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. Nature Genetics, 47(7), 702–709. https://doi.org/10.1038/ng.3285

Polesskaya, O. O., Smith, R. F., & Fryxell, K. J. (2007). Chronic nicotine doses down-regulate PDE4 isoforms that are targets of antidepressants in adolescent female rats. Biological Psychiatry, 61(1), 56–64. https://doi.org/10.1016/j.biopsych.2006.03.038

Radant, A. D., Dobie, D. J., Calkins, M. E., Olincy, A., Braff, D. L., Cadenhead, K. S., … Tsuang, D. W. (2010). Antisaccade performance in schizophrenia patients, their first-degree biological relatives, and community comparison subjects: Data from the COGS study. Psychophysiology, 47(5), 846–856. https://doi.org/10.1111/j.1469-8986.2010.01004.x

Rees, J. M. B., Wood, A. M., & Burgess, S. (2017). Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. Statistics in Medicine, 36(29), 4705–4718. https://doi.org/10.1002/sim.7492

Reitsma, M. B., Fullman, N., Ng, M., Salama, J. S., Abajobir, A., Abate, K. H., … Gakidou, E. (2017). Smoking prevalence and attributable disease burden in 195 countries and territories, 1990–2015: A systematic analysis from the Global Burden of Disease Study 2015. The Lancet, 389(10082), 1885–1906. https://doi.org/10.1016/S0140-6736(17)30819-X

Sallese, M., Salvatore, L., D'Urbano, E., Sala, G., Storto, M., Launey, T., … De Blasi, A. (2000). The G-protein-coupled receptor kinase GRK4 mediates homologous desensitization of metabotropic glutamate receptor 1. FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology, 14(15), 2569–2580. https://doi.org/10.1096/fj.00-0072com

Schizophrenia Working Group of the Psychiatric Genomics Consortium. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature, 511(7510), 421–427. https://doi.org/10.1038/nature13595

Schumann, G., Liu, C., O'Reilly, P., Gao, H., Song, P., Xu, B., … Elliott, P. (2016). KLB is associated with alcohol drinking, and its gene product β-Klotho is necessary for FGF21 regulation of alcohol preference. Proceedings of the National Academy of Sciences of the United States of America, 113(50), 14372–14377. https://doi.org/10.1073/pnas.1611243113

Shihab, H. A., Gough, J., Cooper, D. N., Stenson, P. D., Barker, G. L. A., Edwards, K. J., … Gaunt, T. R. (2013). Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. Human Mutation, 34(1), 57–65. https://doi.org/10.1002/humu.22225

Siedlinski, M., Cho, M. H., Bakke, P., Gulsvik, A., Lomas, D. A., Anderson, W., … Investigators, the Copdg. I. and E. (2011). Genome-wide association study of smoking behaviours in patients with COPD. Thorax, 66(10), 894–902. https://doi.org/10.1136/thoraxjnl-2011-200154

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W. R., & Weir, D. R. (2014). Cohort Profile: The Health and Retirement Study (HRS). International Journal of Epidemiology, 43(2), 576–585. https://doi.org/10.1093/ije/dyu067

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., … Loos, R. J. F. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics, 42(11), 937–948. https://doi.org/10.1038/ng.686

Staley, J. R., Jones, E., Kaptoge, S., Butterworth, A. S., Sweeting, M. J., Wood, A. M., & Howson, J. M. M. (2017). A comparison of Cox and logistic regression for use in genome-wide association studies of cohort and case-cohort design. European Journal of Human Genetics, 25(7), 854–862. https://doi.org/10.1038/ejhg.2017.78

Steenaard, R. V., Ligthart, S., Stolk, L., Peters, M. J., van Meurs, J. B., Uitterlinden, A. G., … Dehghan, A. (2015). Tobacco smoking is associated with methylation of genes related to coronary artery disease. Clinical Epigenetics, 7(1), 54. https://doi.org/10.1186/s13148-015-0088-y

Stoker, A. K., & Markou, A. (2013). Unraveling the neurobiology of nicotine dependence using genetically engineered mice. Current Opinion in Neurobiology, 23(4), 493–499. https://doi.org/10.1016/j.conb.2013.02.013

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, 100(16), 9440–9445. https://doi.org/10.1073/pnas.1530509100

Szumlinski, K. K., Lominac, K. D., Campbell, R. R., Cohen, M., Fultz, E. K., Brown, C. N., … Kippin, T. E. (2017). Methamphetamine Addiction Vulnerability: The Glutamate, the Bad, and the Ugly. Biological Psychiatry, 81(11), 959–970. https://doi.org/10.1016/j.biopsych.2016.10.005

Taliun, D., Harris, D. N., Kessler, M. D., Carlson, J., Szpiech, Z. A., Torres, R., … Abecasis, G. R. (2019). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. BioRxiv, 563866. https://doi.org/10.1101/563866

Thakur, G. A., Sengupta, S. M., Grizenko, N., Choudhry, Z., & Joober, R. (2012). Family-based association study of ADHD and genes increasing the risk for smoking behaviours. Archives of Disease in Childhood, 97(12), 1027–1033. https://doi.org/10.1136/archdischild-2012-301882

Thatcher, R. W., North, D., & Biver, C. (2005). EEG and intelligence: Relations between EEG coherence, EEG phase delay and power. Clinical Neurophysiology, 116(9), 2129–2141. https://doi.org/10.1016/j.clinph.2005.04.026

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

Thorgeirsson, T. E., Steinberg, S., Reginsson, G. W., Bjornsdottir, G., Rafnar, T., Jonsdottir, I., … Stefansson, K. (2016). A rare missense mutation in CHRNA4 associates with smoking

behavior and its consequences. Molecular Psychiatry, 21(5), 594–600. https://doi.org/10.1038/mp.2016.13

Thorgeirsson, Thorgeir E., Gudbjartsson, D. F., Surakka, I., Vink, J. M., Amin, N., Geller, F., … Stefansson, K. (2010). Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. Nature Genetics, 42(5), 448–453. https://doi.org/10.1038/ng.573

Timofeeva, M. N., McKay, J. D., Davey, S. G., Johansson, M., Byrnes, G. B., Chabrier, A., … Brennan, P. (2011). Genetic Polymorphisms in 15q25 and 19q13 Loci, Cotinine Levels, and Risk of Lung Cancer in EPIC. Cancer Epidemiology and Prevention Biomarkers, 20(10), 2250–2261. https://doi.org/10.1158/1055-9965.EPI-11-0496

Tobacco and Genetics Consortium. (2010). Genome-wide meta-analyses identify multiple loci associated with smoking behavior. Nature Genetics, 42(5), 441–447. https://doi.org/10.1038/ng.571

Trabzuni, D., Ryten, M., Walker, R., Smith, C., Imran, S., Ramasamy, A., … Hardy, J. (2011). Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. Journal of Neurochemistry, 119(2), 275–282. https://doi.org/10.1111/j.1471-4159.2011.07432.x

Turley, P., Walters, R. K., Maghzian, O., Okbay, A., Lee, J. J., Fontana, M. A., … Benjamin, D. J. (2018). Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature Genetics, 50(2), 229–237. https://doi.org/10.1038/s41588-017-0009-4

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., & Mardinoglu, A. (2015). Proteomics. Tissue-based map of the human proteome. Science, 347. https://doi.org/10.1126/science.1260419

Vaidyanathan, U., Isen, J. D., Malone, S. M., Miller, M. B., McGue, M., & Iacono, W. G. (2014). Heritability and molecular genetic basis of electrodermal activity: A genome-wide association study. Psychophysiology, 51(12), 1259–1271. https://doi.org/10.1111/psyp.12346

Vaidyanathan, U., Malone, S. M., Donnelly, J. M., Hammer, M. A., Miller, M. B., McGue, M., & Iacono, W. G. (2014). Heritability and molecular genetic basis of antisaccade eye tracking error rate: A genome-wide association study. Psychophysiology, 51(12), 1272–1284. https://doi.org/10.1111/psyp.12347

van den Berg, M. E., Warren, H. R., Cabrera, C. P., Verweij, N., Mifsud, B., Haessler, J., … Munroe, P. B. (2017). Discovery of novel heart rate-associated loci using the Exome Chip. Human Molecular Genetics, 26(12), 2346–2363. https://doi.org/10.1093/hmg/ddx113

Vaughan, J., Donaldson, C., Bittencourt, J., Perrin, M. H., Lewis, K., Sutton, S., … Rivier, C. (1995). Urocortin, a mammalian neuropeptide related to fish urotensin I and to corticotropin-releasing factor. Nature, 378(6554), 287–292. https://doi.org/10.1038/378287a0

Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., … Price, A. L. (2015). Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. American Journal of Human Genetics, 97(4), 576–592. https://doi.org/10.1016/j.ajhg.2015.09.001

Vink, J. M., Willemsen, G., & Boomsma, D. I. (2005). Heritability of Smoking Initiation and Nicotine Dependence. Behavior Genetics, 35(4), 397–406. https://doi.org/10.1007/s10519-004-1327-8

Volkow, N. D., Koob, G. F., & McLellan, A. T. (2016). Neurobiologic Advances from the Brain Disease Model of Addiction. New England Journal of Medicine, 374(4), 363–371. https://doi.org/10.1056/NEJMra1511480

Volkow, N. D., & Morales, M. (2015). The Brain on Drugs: From Reward to Addiction. Cell, 162(4), 712–725. https://doi.org/10.1016/j.cell.2015.07.046

Vrieze, S. I., Feng, S., Miller, M. B., Hicks, B. M., Pankratz, N., Abecasis, G. R., … McGue, M. (2014). Rare Non-Synonymous Exonic Variants in Addiction and Behavioral Disinhibition. Biological Psychiatry, 75(10), 783–789. https://doi.org/10.1016/j.biopsych.2013.08.027

Vrieze, S. I., Hicks, B. M., Iacono, W. G., & McGue, M. (2012). Decline in Genetic Influence on th Co-Occurrence of Alcohol, Marijuana, and Nicotine Dependence Symptoms from Age 14 to 29? The American Journal of Psychiatry, 169(10), 1073–1081. https://doi.org/10.1176/appi.ajp.2012.11081268

Vrieze, S. I., Malone, S. M., Vaidyanathan, U., Kwong, A., Kang, H. M., Zhan, X., … Iacono, W. G. (2014). In search of rare variants: Preliminary results from whole genome sequencing of 1,325 individuals with psychophysiological endophenotypes. Psychophysiology, 51(12), 1309–1320. https://doi.org/10.1111/psyp.12350

Wain, L. V., Shrine, N., Artigas, M. S., Erzurumluoglu, A. M., Noyvert, B., Bossini-Castillo, L., … Tobin, M. D. (2017). Genome-wide association analyses for lung function and chronic obstructive pulmonary disease identify new loci and potential druggable targets. Nature Genetics, 49(3), 416–425. https://doi.org/10.1038/ng.3787

Wain, L. V., Shrine, N., Miller, S., Jackson, V. E., Ntalla, I., Artigas, M. S., … Hall, I. P. (2015). Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): A genetic association study in UK Biobank. The Lancet Respiratory Medicine, 3(10), 769–781. https://doi.org/10.1016/S2213-2600(15)00283-0

Wang, J. C., Cruchaga, C., Saccone, N. L., Bertelsen, S., Liu, P., Budde, J. P., … Goate, A. M. (2009). Risk for nicotine dependence and lung cancer is conferred by mRNA expression levels and amino acid change in CHRNA5. Human Molecular Genetics, 18(16), 3125–3135. https://doi.org/10.1093/hmg/ddp231

Wang, W., Shen, G., Shahar, E., Bidulescu, A., Kimberly, W. T., Sheth, K. N., … Griswold, M. E. (2016). Forced Expiratory Volume in the First Second and Aldosterone as Mediators of Smoking Effect on Stroke in African Americans: The Jackson Heart Study. Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease, 5(1). https://doi.org/10.1161/JAHA.115.002689

Ware, J. J., Chen, X., Vink, J., Loukola, A., Minica, C., Pool, R., … Munafò, M. R. (2016). Genome-Wide Meta-Analysis of Cotinine Levels in Cigarette Smokers Identifies Locus at 4q13.2. Scientific Reports, 6(1), 1–7. https://doi.org/10.1038/srep20092

Warren, H. R., Evangelou, E., Cabrera, C. P., Gao, H., Ren, M., Mifsud, B., … Wain, L. V. (2017). Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. Nature Genetics, 49(3), 403–415. https://doi.org/10.1038/ng.3768

Wedow, R., Zacher, M., Huibregtse, B. M., Mullan Harris, K., Domingue, B. W., & Boardman, J. D. (2018). Education, Smoking, and Cohort Change: Forwarding a Multidimensional Theory of the Environmental Moderation of Genetic Effects. American Sociological Review, 83(4), 802–832. https://doi.org/10.1177/0003122418785368

Wheeler, E., Huang, N., Bochukova, E. G., Keogh, J. M., Lindsay, S., Garg, S., … Farooqi, I. S. (2013). Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nature Genetics, 45(5), 513–517. https://doi.org/10.1038/ng.2607

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics, 26(17), 2190–2191. https://doi.org/10.1093/bioinformatics/btq340

Wilson, S., Bair, J. L., Thomas, K. M., & Iacono, W. G. (2017). Problematic alcohol use and reduced hippocampal volume: A meta-analytic review. Psychological Medicine, 47(13), 2288–2301. https://doi.org/10.1017/S0033291717000721

Wilson, S., Haroian, K., Iacono, W. G., Krueger, R. F., Lee, J. J., Luciana, M., … Vrieze, S. I. (n.d.). Minnesota Center for Twin and Family Research.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. The American Journal of Human Genetics, 89(1), 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029

Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science (New York, N.Y.), 345(6204), 1616–1620. https://doi.org/10.1126/science.1255514

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. Nature Genetics, 42(7), 565–569. https://doi.org/10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. American Journal of Human Genetics, 88(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yang, J., Villar, V. A. M., Armando, I., Jose, P. A., & Zeng, C. (2016). G Protein-Coupled Receptor Kinases: Crucial Regulators of Blood Pressure. Journal of the American Heart Association, 5(7). https://doi.org/10.1161/JAHA.116.003519

Zaitlen, N., Kraft, P., Patterson, N., Pasaniuc, B., Bhatia, G., Pollack, S., & Price, A. L. (2013). Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. PLoS Genetics, 9(5), e1003520. https://doi.org/10.1371/journal.pgen.1003520

Zhan, X., Hu, Y., Li, B., Abecasis, G. R., & Liu, D. J. (2016). RVTESTS: An efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics, 32(9), 1423–1426. https://doi.org/10.1093/bioinformatics/btw079

Zhan, X., & Liu, D. J. (2015). SEQMINER: An R-Package to Facilitate the Functional Interpretation of Sequence-Based Associations. Genetic Epidemiology, 39(8), 619–623. https://doi.org/10.1002/gepi.21918

Zhang, S., Chen, H., Zhao, X., Cao, J., Tong, J., Lu, J., … Lu, D. (2013). REV3L 3'UTR 460 T>C polymorphism in microRNA target sites contributes to lung cancer susceptibility. Oncogene, 32(2), 242–250. https://doi.org/10.1038/onc.2012.32

Zheng, J., Erzurumluoglu, A. M., Elsworth, B. L., Kemp, J. P., Howe, L., Haycock, P. C., … Neale, B. M. (2017). LD Hub: A centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics, 33(2), 272–279. https://doi.org/10.1093/bioinformatics/btw613

Zhu, H., Lee, M., Guan, F., Agatsuma, S., Scott, D., Fabrizio, K., … Hiroi, N. (2005). DARPP-32 phosphorylation opposes the behavioral effects of nicotine. Biological Psychiatry, 58(12), 981–989. https://doi.org/10.1016/j.biopsych.2005.05.026

Zlojutro, M., Manz, N., Rangaswamy, M., Xuei, X., Flury-Wetherill, L., Koller, D., … Almasy, L. (2011). Genome-wide association study of theta band event-related oscillations identifies serotonin receptor gene HTR7 influencing risk of alcohol dependence. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 156(1), 44–58. https://doi.org/10.1002/ajmg.b.31136