Genetic insights into smoking behaviours in 10,558 men of African ancestry from continental Africa and the UK

Noemi-Nicole Piga[1], Palwende Romuald Boua[2,3,4], Chisom Soremekun[5,6,7], Nick Shrine[1], Kayesha Coley[1], Jean-Tristan Brandenburg[3], Martin D. Tobin[1,8], Michèle Ramsay[3,4], Segun Fatumo[6,7,9], Ananyo Choudhury[3*], Chiara Batini[1*T]
*These authors contributed equally

[1]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Leicester, United Kingdom
[2]Clinical Research Unit of Nanoro, Institut de Recherche en Sciences de la Santé, Nanoro, Burkina Faso
[3]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[4]Division of Human Genetics, National Health Laboratory Service and School of Pathology, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[5]Department of Immunology and Molecular Biology, College of Health Science, Makerere University, Kampala, Uganda.
[6]H3Africa Bioinformatics Network (H3ABioNet) Node, Center for Genomics Research and Innovation (CGRI)/National Biotechnology Development Agency CGRI/NABDA, Abuja, Nigeria
[7]The African Computational Genomics (TACG) Research Group, MRC/UVRI LSHTM Uganda Research Unit, Entebbe, Uganda
[8]National Institute for Health Research Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester, UK
[9]Department of Non-communicable Disease Epidemiology (NCDE), London School of Hygiene and Tropical Medicine, London, United Kingdom

[T]corresponding author: cb334@leicester.ac.uk

keywords: smoking cessation, smoking initiation, GWAS, Africa, UK

Abstract

Smoking is a leading risk factor for many of the top ten causes of death worldwide. Of the 1 billion smokers globally, 80% live in low- and middle-income countries, where the number of deaths due to tobacco use is expected to double in the next decade according to the World Health Organization. Genetic studies have helped to identify biological pathways for smoking behaviours, but have mostly focussed on individuals of European ancestry or living in either North America or Europe.

Here we present a genome-wide association study of two smoking behaviour traits in 10,558 men of African ancestry living in five African countries and the UK. Eight independent variants were associated with either smoking initiation or cessation at p-value $< 5 \times 10^{-6}$. Of these, four were monomorphic or rare in European populations. Gene prioritization strategy highlighted five genes, including *SEMA6D,* previously described as associated with several smoking behaviour traits. These results confirm the importance of genetic epidemiological studies in underrepresented populations.

Introduction

Smoking is a leading risk factor for many of the top ten causes of death worldwide, including heart and lung diseases[1]. Each year, tobacco use is directly responsible for 7 million deaths and 25% of all cancer fatalities globally[2]. However, smoking prevalence varies among world regions, and of the 1.3 billion tobacco users worldwide, 80% live in low- and middle-income countries (LMICs)[3].

Reassuringly, prediction models by the World Health Organization (WHO) show a reduction in smoking prevalence in most areas from 2010 to 2025 thanks to tobacco control strategies[4,5]. However, in parallel to the American and European decline, the tobacco industry and market has seen a recent expansion in the African continent, due to the fast population growth and the improvement in buying power[6]. In line with this, sub-Saharan Africa has experienced a 52% increase in tobacco use from 1980 to 2016[7] and a further 9 million people are expected to take up smoking in the African region by 2025[8]. In addition, the support for smokers wishing to quit and for healthcare professionals training in smoking cessation is still very limited across the continent[9]. To help counteract these changes, as of November 2018 more than forty African countries supported the WHO Framework Convention on Tobacco Control and twenty countries were involved in the Protocol to Eliminate Illicit Trade in Tobacco Products[6].

Currently, there are 94 million male and 13 million female tobacco users in Africa, with one in five adolescents using tobacco, and an increasing smoking prevalence among young women[6]. The type of tobacco used varies among countries and sexes, and its consumption is overall associated with alcohol drinking, lower income status, education levels, and professional activity[10,11]. Using Demographic Health Surveys data for 30 sub-Saharan African countries, Sreeramareddy et al. compared smoking with the use of smokeless tobacco and showed that while the first is more prevalent among men, the majority of women preferentially uses the second, and specifically chewing tobacco[10]. Similar patterns were confirmed by a recent study focusing on tobacco and alcohol use in rural and urban settings in four African countries[11].

The patterns described above change when we focus on communities of African descent in the UK. In 2019, 14% of the adult UK general population smoked regularly, with slightly different figures between sexes: 15.9% of men and 12.5% of women smoked, resulting in a ratio between male and female smoking proportions of 1.27. This ratio was 1.87 among Black ethnic minorities in the same year[12]; however, contrary to what is observed in continental Africa, between 2014 and 2019 the smoking prevalence among Black adults decreased from 13.3% to 9.6%[12].

In the US, African-American individuals have been shown, on average, to initiate smoking later and smoke fewer cigarettes per day than European-American individuals[13]. However,

they show comparable levels of nicotine equivalents, are less likely to successfully quit smoking, and have a higher risk of smoking-related lung cancer[13,14].

Genetics has been shown to play a role in smoking behaviour traits. Genome-wide association studies (GWASs) identify genetic variants associated with the trait of interest, which inform biological understanding and highlight functional pathways and potential drug targets for precision medicine approaches to treatment[15]. The strongest associations with smoking behaviours have been consistently shown at locus 15q25.1 for amount smoked and nicotine dependence, containing the cluster of *CHRNA5-A3-B4* genes which encode subunits of the nicotinic acetylcholine receptors (nAChRs)[16]. In the brain, nicotine binds to nAChRs stimulating the release of several neurotransmitters, impacting the reward pathway, learning and memory[16]. To date, the largest GWAS of smoking behaviour traits includes 1.2 million individuals and highlights genes coding for proteins involved in neurotransmission of nicotine, dopamine and glutamate[17]. However, as in most GWASs, it only includes individuals of European origin failing on the representation of global human diversity[18,19].

Previous genetic epidemiology studies of smoking behaviour in individuals of African descent have only included African-American (AA) participants and have identified four genetic loci associated with different traits. Located ~6kb from the 5' of *CHRNA5*, rs2036527 on chromosome (chr) 15, was the only variant associated with cigarettes per day in 32,389 AA individuals in the Study of Tobacco in Minority Populations[13]. In a study including ~1000 AA participants, Chenoweth et al. found three independent variants on chr 19 (rs12459249, rs111645190, rs185430475) associated with nicotine metabolite ratio (NMR)[14]. These variants showed low linkage disequilibrium with four signals previously identified in a Finnish cohort and located in the genomic region of the *CYP2A6* gene which encodes for the key enzyme of nicotine metabolism[14,16]. Hancock et al. performed a trans-ethnic analysis including 28,677 European and 9,925 AA smokers, and identified variant rs910083 on chr 20 as associated with nicotine dependence[20]. Finally, Xu et al. studied smoking trajectories in almost 300,000 individuals from the Million Veteran Program (MVP), including >54,000 AAs, which aimed to capture the variation of smoking status over time[21]. They found an association on chr 1 with variant rs4478781 in AAs only, and 14 associated loci in a trans-ethnic meta-analysis including European-Americans and Hispanic-Americans, mainly driven by the results from the European ancestry group[21].

Despite these interesting findings, individuals of African descent, and especially those living in continental Africa, remain heavily unrepresented in large-scale studies[18,19]. Here we present the first GWAS of smoking behaviour traits in 10,558 men of African ancestry living in five African countries and the UK. In agreement with the sex biases previously described[10,11], for the African datasets included here, smoking prevalence in females ranged between 0.5% and 7%, which prompted their exclusion from all discovery analyses.

Materials and methods

Participants

This study used genetic data and smoking behaviour information from three different cohorts: the Africa Wits-INDEPTH Partnership for Genomic Studies (AWI-Gen)[22], the Uganda Genome Resource[23](UGR) and the individuals of African ancestry in UK Biobank (UKB-AFR)[24]. AWI-Gen is a cross-sectional population study including ~12,000 individuals from Ghana, Burkina Faso, Kenya and South Africa, aged 40-60+ years, aimed at understanding the genomic and environmental factors that contribute to body composition and cardio-metabolic diseases. For the purpose of this study, with the aim of accounting for the local population structure, the AWI-Gen cohort was divided into three datasets representing the main geographical areas: AWI-East (Kenya), AWI-South (South Africa) and AWI-West (Ghana and Burkina Faso).

The UGR includes ~6,400 individuals, and it is a subset of the General Population Cohort (GPC)[25], a population-based open cohort aimed at understanding HIV infections in Uganda. The UGR was built with the aim of improving the local resources for public health and to allow large genetic epidemiology studies[23].

UKB is a large longitudinal study in the UK which includes samples from over 500,000 volunteers aged 40-69 years at baseline[24]. Since 2006, it has collected extensive phenotypic and biological data in order to allow approved researchers to investigate the genetic and/or environmental determinants of a wide range of diseases and health-related phenotypes. Using genetic data for population stratification, we identified ~7,800 individuals of African ancestry in UKB[26].

Description of phenotypes

We have analysed two binary smoking behaviour traits: Smoking Initiation (SI) and Smoking Cessation (SC), defined using relevant answers in questionnaire data available in each cohort. Overall, SI compares 'never' smokers (controls) versus 'ever' smokers (cases), in which the former represents individuals who have never, or only very rarely, smoked in their life and the latter are participants who smoked or currently smoke every day or occasionally. SC only includes 'ever' smokers and it compares 'current' smokers (cases), who were smoking every day or occasionally at the time of answering the questionnaire, to 'previous' smokers (controls), who were not. The detailed description of the phenotype definitions and the specific questions used for each cohort are available in Supplementary Note 1.

The number of cases and controls for each phenotype in each cohort is presented in Table 1.

Genotyping and Imputation Quality Control

The AWI-Gen individuals were genotyped at ~2.4M SNPs using the Illumina Infinium H3Africa SNP array, which is designed to be specific and sensitive to the genomic diversity of African populations[22]. Imputation at ~39M autosomal variants was performed using the

Haplotype Reference Consortium (HRC) panel[27] on the Michigan Imputation Server. Details of quality control (QC) and imputation settings are presented in Choudhury et al al. 2020[28].
The UGR individuals were genotyped at ~2.2M autosomal markers using the HumanOmni2.5-8 chip array. Imputation at ~98M variants was performed using a combined reference panel with sequence data from three different studies (African Genome Variation Project[29], Uganda 2000 Genomes and 1000 Genomes Project Phase 3 [1000GP][30]). Details of QC and imputation settings are presented in Gurdasani et al.[23].
UK Biobank individuals were genotyped at ~800,000 variants using the UK Biobank Axiom Array. Imputation at ~93M autosomal variants was performed using the HRC[27], the UK10K[31] and the 1000GP[30] reference panels combined. Details of QC and imputation settings are presented in Bycroft et al.[24].
Details of genotyping QC are reported in Supplementary Table 1. Additional QC was performed across all datasets to ensure that effect alleles were consistently aligned between cohorts. Imputed autosomal variants with a minor allele count (MAC) >= 20 or minor allele frequency (MAF) > 0.01, and an imputation info score >= 0.3 were included in all further analyses (Supplementary Figure 1).

Study level genome-wide association analyses

Genome-wide association analyses were performed using a univariate linear mixed model (LMM) and significance was evaluated using a likelihood ratio test as implemented in GEMMA v0.98.1[32]. Covariates included age, age^2 and as many principal components (PCs) as required in each dataset (east 5PCs; south 14PCs;west 11PCs; UGR 10PCs; UKB 9PCs). PCs were calculated for each dataset from a PC analysis using independent genotyped SNPs in PLINK v1.90[33]. For AWI-Gen and UKB-AFR datasets, we determined the number of PCs to include by using the eigenvalues to assess when adding further components would not contribute additional information. In an iterative process, we stopped at the first PC for which the contribution of the three previous PCs was not greater than the contribution of the following three. For UGR, 10 PCs were used as in Gurdasani et al.[23]. The genetic relatedness matrix (GRM) included in the LMM was calculated for each full cohort with GEMMA v0.98.1[32] using independent SNPs. Details of the QC used for genotyped variants and the number of variants used to perform the PCA and to calculate the GRM for each cohort are reported in Supplementary Table 1. Manhattan and quantile-quantile (qq-) plots were visualized using the qqman package in R[34]. The LD score regression intercept was calculated to assess the presence of genomic inflation using ldsc v1.0.1[35]. LD scores were calculated including the African superpopulation from 1000GP[30]. When the LD score intercept was above 1.05, the association p-value was recalculated as follows:
$corrected.se = se * \sqrt{(LDScoreregressionintercept)}$; $z = -|\frac{estimate}{corrected.se}|$;
$corrected.pvalue = P(Z < -z) + P(Z > z)$. Where $corrected.se$and se are the corrected and original standard error of the estimate effect size respectively; z is the z-score used to calculate the corrected p-value using two-sided test statistics.

Meta-analysis

We performed a two-step meta-analysis. In step1, we included the three AWI-Gen datasets (AWI-East, AWI-South, AWI-West) to obtain AWI-Gen cohort-level summary statistics. In step2, we included AWI-Gen, UGR and UKB-AFR. In each step, we used the modified random effect model (RE2) as implemented in METASOFT[36]. Variants were included in the analysis if they were present in at least two out of three studies. Supplementary Figure 1 shows the number of variants for each meta-analysis and phenotype. Results were visualized using Manhattan and qq-plots, and LD score regression intercept was calculated to evaluate genomic inflation, as done for the study-level association analyses.

Definition of associated and sentinel variants

For each trait, variants were divided into three tiers defined using different significance thresholds which take into account the results of the meta-analysis and the cohort-level summary statistics (AWI-Gen, UGR and UKB-AFR). Tier1 included variants with meta-analysis p-value $< 5 \times 10^{-8}$ and p-value < 0.05 in each cohort; tier2, variants with a meta-analysis p-value $< 5 \times 10^{-6}$ and p-value < 0.05 in each cohort; tier3, variants with a p-value $< 5 \times 10^{-8}$ in at least one of the cohort genome-wide association analyses.

For each tier, sentinel variants were defined in an iterative process as the variants with the lowest p-value in a region of 200kb centered on the variant.

Conditional analysis

To assess the presence of any additional independent signals, we utilized the GCTA v1.93.2[37] stepwise model selection for the conditional analysis (option --vcojo-slct) in a region of +/- 100kb from each sentinel variant and using the populations of African ancestry in 1000GP[30] as a reference for LD patterns[30].

Fine-mapping analysis

Regions of +/- 100kb from each sentinel variant were analysed to retrieve the 99% credible set variants using FINEMAP v1.4 software[38] with a shotgun stochastic algorithm assuming one causal variant.

We then created the refined credible sets by filtering for a Posterior Inclusion Probability (PIP) greater than or equal to 1%. LD proxy variants of these SNPs were identified in the African populations in 1000GP[30] within the regions defined by the 99% credible sets extended by +/-100kb using PLINK v1.90[33]. Variants with both D' >= 0.9 and r^2 >= 0.6 were retained. Supplementary Figure 1 and Supplementary Table 2 show the 99% credible sets, the refined credible sets and the proxy variants for each locus. The refined credible set variants and their proxies were used for all follow up analyses.

Replication and lookup analyses

The replication of our meta-analysis step2 results was performed using two publicly available datasets: (a) the genome-wide summary statistics for smoking trajectories in individuals of African ancestry included in the MVP[21]; (b) and the genome-wide summary statistics from the SI and SC meta-analyses in individuals of European ancestry released by the GWAS & Sequencing Consortium of Alcohol and Nicotine use (GSCAN)[17]. The number of independent sentinel variants was used to calculate the Bonferroni corrected p-value threshold for the replication analyses (0.007 SI; 0.05 SC).

We performed two lookup analyses aiming to understand if our discovered loci were previously described as associated with any smoking phenotypes or any other trait.

For the first analysis, we compiled a list of variants described in 14 studies[13,17,20,21,39,40,41,42,43,44,45,46,47,48] as associated with 9 smoking behaviour traits (Smoking Initiation; Smoking Cessation; Age of Initiation; Cigarettes per Day; Fagerström Test for Nicotine Dependence; Pack Years; Trajectory Contrast I; Trajectory Contrast II; Time to the First Cigarette).

For the second analysis, we queried GWAS Catalog v1.0.2[49] (see URLs) to identify variants associated with any other phenotype.

Follow up analyses

Gene prioritisation

We combined results from four analyses in order to identify the genes influenced by the SNPs in the refined credible sets and their proxies. First, we assessed the predicted pathogenicity of the variants using the Combined Annotation Dependent Depletion (CADD) score[50] as implemented in the Ensembl Variant Effect Predictor (VEP)[51]. We defined as pathogenic those variants with a CADD score greater or equal than 15.

We then investigated if the variants influenced the expression of a gene or the protein level using eQTL and pQTL data respectively. We used the eQTL Mapping option in the Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) v1.3.6[52] which collects eQTL data from 14 data sources (see URLs). Significant eQTLs are defined on either p-value or FDR thresholds based on the specific data source (see URLs) and we retrieved results from blood, brain or lung tissues. For the pQTL analysis, we used data from 90 cardiovascular genes of the SCALLOP Consortium[53] and followed their definition of significant pQTLs. They defined signals more than 1Mb away from the protein gene as *trans*-pQTLs, and signals that were closer than 1Mb as *cis*-pQTL[53].

Finally, we used the Chromatin Interaction Mapping tool as implemented in FUMA v1.3.6[52] to gain insights into possible epigenetic properties of the interrogated variants using the blood, brain and lung tissues (see URLs).

We selected those variants that were identified by at least one of the analyses described above and retrieved the list of prioritised genes for SI and SC.

Pathway analysis

The prioritised genes were queried to investigate any biological function and implication in smoking behaviour using the webtool Metascape[54] which performs pathway enrichment and protein-protein interaction analyses combining information from different databases and -omics data using hierarchical clustering[54]. Pathway enrichment is based on an over-representation analysis[54], while the protein-protein interaction makes use of the MCODE algorithm[55], which captures densely connected regions in a complex network[54,55]. For both analyses, we used the default option of Metascape and the Entrez Gene ID as gene name. We considered a pathway being enriched if represented by prioritised genes linked to distinct meta-analysis step2 associated loci.

PheWAS analysis

Variants in the refined credible set for the locus of chr 15 and passing at least one of the criteria for our gene prioritisation strategy were included in a PheWAS using the PheWAS R package[56] in three datasets available from the Integrative Epidemiology Unit (IEU) OpenGWAS project (see URLs)[57,58]: (i) 'IEU analysis of UK Biobank phenotypes'[59] (ukb-b, see URLs) and (ii) 'Neale lab analysis of UK Biobank phenotypes, round 2' (ukb-d, see URLs) for European individuals, and (iii) 'Pan-ancestry genetic analysis of the UK Biobank performed at the Broad Institute'[60] (ukb-e, see URLs) for African individuals only. PheWAS results from ukb-b and ukb-d were combined together since they include distinct phenotypes and refer to the same ancestry group. For each ancestry group we filtered for significant associations after applying a Bonferroni correction for each variant based on the number of tested phenotypes.

Results

Discovery analyses

We performed a genome-wide association analysis for each dataset and phenotype combination in a total of 10,558 men for SI and 4,257 for SC. A modified random effect model was implemented for both steps of the meta-analysis on variants present in at least two of the individual datasets. Step1 included the three AWI-Gen datasets and step2 meta-analysed the results of step1 with UGR and UKB-AFR (Supplementary Figure 1). Results for the individual studies and meta-analysis step1 are presented in Supplementary Note 2.

For SI, step2 analysed 14,459,454 SNPs: no genome-wide significant variant was observed, while 99 variants passed the suggestive significance threshold (Figure 1a). The qq-plot showed no residual population structure (Supplementary Figure 2a) and the LD score regression intercept was 0.94. For SC, step2 analysed a total of 14,057,868 variants: no SNPs passed the genome-wide significant threshold and 45 SNPs were below the suggestive significance threshold (Figure 1b). The qq-plot showed no residual population structure (Supplementary Figure 2b), confirmed by an LD-score regression intercept value of 0.88. Following our tier criteria and our definition of sentinel variants, we identify (i) no variant in tier1 for either trait; (ii) 7 sentinel variants for SI and one for SC in tier2; and (iii) one variant in tier3 for SI. In the meta-analysis, the 8 sentinel variants from tier2 show low heterogeneity (I^2) of effect sizes, as well as having a consistent direction of effect among studies (Table 2 and Supplementary Figure 3). We focused our follow up analyses on the sentinel variants in tier2.

Conditional analyses did not identify any additional independent signal in the 200kb loci we defined around our 8 sentinel variants in tier2. For each locus, we first identified the 99% credible set with a Bayesian approach, which included a total of 2,243 potentially causal variants for SI and 120 for SC (Supplementary Table 2). These 99% credible sets spanned regions of 47.6 to 200kb (see Supplementary Figure 4), with the only exception of the 99% credible set on chr 19 which included only one variant. We then identified the refined credible sets to include only those variants with a PIP > 0.01 and their LD proxies (D' >= 0.9 and r^2 >= 0.6) reducing the number of SNPs to 136 for SI and 36 for SC (Supplementary Figure 1 and Supplementary Table 2).

Replication and literature lookup analyses

We performed replication analyses using the refined credible sets and their proxy variants in two different datasets: (a) an analysis of smoking trajectory contrasts performed in individuals of African ancestry[21]; and (b) two meta-analyses of SI and SC in European individuals[17]. The smoking trajectory contrasts represent a comparison of either (I) current vs never or (II) current vs mixed smokers, defined using electronic health records data, and capturing SI and SC respectively[21].

When using dataset (a)[21], 9 out of the 69 variants in the refined credible set on chr 15 passed the Bonferroni corrected threshold for SI (Supplementary Table 3a). Four out of 5 variants on chr 1 passed the nominal significance threshold, and no variants on chr 10 showed evidence of replication. No dataset (a) variants were found for the loci on chr 4, 7, 16 and 19. When using (b)[17], all the variants in the refined credible set on chr 15 passed the Bonferroni corrected threshold for SI, with four of these being genome-wide significant and having the same direction of effect as in our analysis (Supplementary Table 3b). No variants on chr 7, 16 and 19 showed evidence of replication for SI or SC and no dataset (b) variants were found for the loci on chr 1, 4 and 10.

We investigated if any of the variants included in the credible sets had been previously described as associated with any smoking behaviour trait or any other phenotype. From both analyses, we only obtained results for the SI trait. One variant on chr 15 (rs9646181) was previously described as associated with smoking trajectory contrast I (p-value 4.8×10^{-10}) (Supplementary Table 3c), in a trans-ethnic meta-analysis including individuals of African, European and Hispanic ancestries and it was mapped as an intronic variant in the gene *SEMA6D*[21]. Looking beyond smoking behaviour traits, we interrogated the GWAS Catalog and found that the variant rs4624724 on chromosome 4 had been previously described as associated with adolescent idiopathic scoliosis (p-value 4×10^{-8}; Supplementary Table 3d)[61].

<u>Functional follow up analyses</u>

Gene prioritisation analyses

We combined the results of four different analyses: variant annotation with VEP and its CADD score estimate, eQTL and pQTL mapping, and chromatin interaction mapping.
For both traits, all queried variants were annotated as non-coding, or as part of upstream or downstream regulatory regions (VEP[51]) and no variants were identified as a pQTL using the 90 cardiovascular genes of the SCALLOP Consortium[53].

For the SI trait, we retrieved a total of 95 genes. Specifically, the CADD scores highlighted a total of six possible deleterious variants affecting two genes, *RP11-552E10.1* and *SEMA6D* (Supplementary Table 4a). The eQTL mapping showed that 6 variants affect the expression of 4 genes (*AC073133.1*, *FBN1*, *MAN2B2* and *SLC1A6*) in different brain tissue datasets (BRAINEAC (see URLs), Common Mind Consortium[62], eQTLGen[63], PsychENCODE[64]; Supplementary Table 4a). Finally, the chromatin interaction analysis highlighted 42 unique variants, at least one for each associated locus, having an effect on a total of 92 genes in either brain tissues or lung fibroblast cells[64,65,66] (Supplementary Table 4a).
Three of the 95 genes identified were mapped by two of the prioritisation analyses (*AC073133.1*, *MAN2B2*, and *SEMA6D*; Supplementary Table 4a). *AC073133.1* showed significant results for eQTLs for brain tissue with one variant (rs6969023) and for chromatin interaction in lung fibroblasts mediated by three variants (rs116530211,rs76374118,rs79338905; Supplementary Table 4a). *MAN2B2* was identified by

variant rs73207830 as being an eQTL in blood, and variant rs116755844 indicated chromatin interaction in the Promoter anchored Hi-C loops data from PsychENCODE[64] (Supplementary Table 4a). *SEMA6D* showed evidence of both pathogenicity based on a high CADD score and chromatin interaction in lung fibroblast cells (Supplementary Table 4a).

For the SC trait, only one variant had a CADD score higher than or equal to 15 but it was annotated as intergenic (data not shown). The combination of the eQTL and the chromatin interaction mapping defined a total of 30 genes associated with 18 distinct variants (Supplementary Table 4b). The eQTL mapping identified two genes, *CES1* and *LPCAT2*. While *LPCAT2* was highlighted by only one dataset in blood (eQTLGen[63]), *CES1* was retrieved by six datasets (BIOSQTL[67], DICE[68], eQLTCatalogue[69], eQTLGen[63], GTExv8[70], PsychENCODE[64]) including several blood cell types (B cells, monocytes, and T cells), and lung and brain tissues (Supplementary Table 4b). Both genes showed significant chromatin interaction values discovered in IMR90, a lung fibroblast cell line[65] (Supplementary Table 4b). The remaining 28 genes showed SNP-mediated chromatin interaction both in the IMR90 cell line[65] and in the Promoter anchored Hi-C loops data from PsychENCODE[64] (Supplementary Table 4b).

Pathway analysis

We performed a pathway analysis to investigate biological interactions between the prioritised genes using the web-based tool Metascape developed for overrepresentation analysis of genes in biological pathways and protein-protein interaction[54]. We decided to focus only on pathways enriched with genes implicated by different loci, and so we performed this analysis only on the genes prioritised for SI. Fifty-five out of the ninety-five genes had an Entrez Gene ID and were analysed by Metascape resulting in two enriched pathways from Gene Ontology (GO) Resource. 'Metanephros development', the process to form the definitive kidney (GO:0001656), was enriched for *FBN1*, *SHH* and *WFS1* genes (Log(p-value): -3.50; Supplementary Table 5); the 'developmental growth involved in morphogenesis' (GO:0060560), a large GeneOntology category including several classes of morphogenesis activities was enriched for *PDPN*, *SALL1*, *SEMA6D* and *SHH* genes (Log(p-value): -2.25; Supplementary Table 5). The protein-protein interaction network analysis identified two interactions: *CYP4F3* with *CYP4F8*, and *FBN1* with *WFS1*.

PheWAS analysis

We selected variants in the locus on chr 15 for a PheWAS analysis, as this was the only locus to replicate and it harboured *SEMA6D*, which was previously identified by other studies on smoking behaviour traits (see Discussion). We limited the PheWASs to variants supported by at least one of the four criteria of our gene prioritisation analysis obtaining a total of 4 variants: rs11634974, rs11636198, rs12905212, and rs7273389. The number of tested phenotypes for European individuals (ukb-d and ukb-d) varied among variants: from 2443 (rs7273389), to 3338 (rs11634974), to 3342 (rs11636198 and rs12905212). For PheWASs in African ancestry individuals, all variants were tested for 1152 distinct phenotypes.

We found Bonferroni-corrected significant results for fifteen traits only in European individuals: rs11634974, rs11636198 and rs12905212 showed association with the same 6 traits, including 'Current tobacco smoking' which had the same direction of effect as in this study, and 'Qualifications: College or University degree' with opposite direction of effect (Supplementary Table 6). The fourth variant, rs7273389, showed nine significant associations, seven of which were with body fat measures (Supplementary Table 6).

Discussion

Smoking is a preventable risk factor for several diseases worldwide[1] with 80% of smokers living in low- and middle-income countries and a rising prevalence in Africa[3]. GWASs have shown that genetics plays a role in smoking behaviours[16,17], but similarly to other traits, most studies have been performed in individuals of European ancestry, thus underestimating the role of genetic diversity for these traits globally[18,19]. Disentangling the genetics of smoking in sub-Saharan Africa is essential to shed light onto its biology in this region and globally, and to help elucidate its role as a risk factor for non-communicable diseases, either directly or through interaction[71].

In this study we focussed on understanding the genetics of two smoking behaviour traits, smoking initiation and cessation, in 10,558 men of African ancestry living in five countries in the African continent and the UK, including participants from three cohorts: AWI-Gen[22], divided into three geographical areas (East, South and West), UGR[23], and UKB-AFR[24]. After a two-step meta-analysis, we identified 7 loci associated with SI and one with SC, all in tier2 (variants with a meta-analysis p-value $< 5 \times 10^{-6}$ and p-value < 0.05 in each cohort). We selected variants for *in silico* functional follow up analyses based on their posterior inclusion probability of being causal, obtaining 136 variants for SI and 36 for SC. We compared the allele frequencies at these variants between the African (AFR) and European (EUR) superpopulations from 1000GP[30] obtained via VEP[51]. All but one variant of five associated loci with SI (on chr 1, 4, 7, 10 and 16) were monomorphic or had a MAF < 2% in EUR, while they were common (MAF ranging 2-25%) in AFR (Supplementary Table 7). Despite being common in both AFR and EUR, most variants on chr 15 showed allele frequencies 2-5 times higher in AFR, and the only variant for chr 19 was common for both groups (Supplementary Table 7). Variants in the chr 16 locus associated with SC showed a general higher frequency in EUR (Supplementary Table 7).
The variants identified by the few studies including African-American individuals described in the introduction[13, 14, 20] did not replicate in our study, with the caveat that they focused on smoking phenotypes different from our traits.

Our gene prioritisation strategy highlighted *AC073133.1, MAN2B2,* and *SEMA6D* for SI and *CES1* and *LPCAT2* for SC, as genes supported by two out of the four analyses included (CADD score, eQTLs, pQTLs and chromatin interaction). A detailed description of their function, and of the genes highlighted by the pathway and protein-protein interaction analyses is included in Supplementary Note 3. *SEMA6D* on chr 15, however, will be described in detail here as this locus shows strong evidence of replication, is involved in one of the two pathways identified, and includes eQTLs for a gene involved in a protein-protein interaction (*FBN1*). This gene is a member of the semaphorin family that encodes both secreted and membrane proteins involved in axon guiding, that may have a role in maintaining and remodelling neuronal connections (see URLs). It was already identified as associated with smoking initiation, cessation and amount by five studies[17,21,44,45,46] including the largest study to date

by GSCAN[17] and the recent trans-ethnic GWAS meta-analysis of smoking trajectories in the MVP cohort[21]. Querying GWAS Catalog for *SEMA6D* (as of July 2021), we found reported associations for 63 traits (see URLs), including smoking and drinking behaviour phenotypes, depression and cognitive ability (see URLs). In agreement with our PheWAS, educational attainment and body mass index phenotypes were among the top five trait classes associated with *SEMA6D* (Supplementary Table 6).

We are aware this study has its limitations. The underrepresentation of individuals of African ancestry in biobank-scale cohorts affects several aspects of this work: from the limited sample sizes to the availability of additional datasets for larger meta-analyses, replication and follow up analyses. Not only does this influence the number of datasets available for genomics studies, but also the number of variants that can be tested for association, as shown by half of the tier2 loci being monomorphic or rare in EUR from 1000GP[30] (Supplementary Table 7). The sex bias observed in our datasets restricted the analyses to men only, impacting further on sample sizes and calling for specific attention to the cultural habits that may generate it. While it is true that smoking prevalence is very low among women in many African countries[10,11], it is growing among girls[6]. The widespread use of chewing tobacco in some areas suggests that a new way of collecting data on tobacco use should be considered to aim to study nicotine dependence at a population level.

This study adds support to a previously identified locus from large European and trans-ethnic studies, but importantly it highlights the need for additional African cohorts to be developed and maintained. This is essential if we aim to overcome the limitations described above and be in a position to perform statistically powerful large-scale association studies across smoking behaviour phenotypes, as well as many other traits currently understudied in African populations.

Data access statement

Complete summary statistics from the meta-analysis step2 for SI and SC are being deposited to NHGRI-EBI GWAS catalog [https://www.ebi.ac.uk/gwas/] (Study accession numbers: SI, GCST90091238; SC, GCST90091239; under embargo until publication is accepted). Individual-level genetic and phenotypic data from the AWI-Gen, Uganda Genome Resource and UK Biobank are available to approved researchers upon application or data access request.

File handling and individual analyses were performed using a combination of bash and R scripts, available upon request from the authors.

Ethics approvals

The AWI-Gen study was approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (Wits) (protocol numbers M121029 and M170880). In addition, each research site obtained approval from their local ethics review board prior to commencing any participant-related activities. Uganda Genome Resource was approved by the Science and Ethics Committee of the UVRI, the Ugandan National Council for Science and Technology, and the East of England-Cambridge South NHS Research Ethics Committee United Kingdom. This research has been conducted using the UK Biobank Resource under approved Application 4892. Informed consent was obtained from all participants and all research was performed in accordance with relevant guidelines and regulations.

Author contributions

Study design: MDT, MR, SF, AC, CB; Data collection: PRB, MR, SF, AC; Data preparation and initial processing: N-NP, CS, NS, J-TB, AC, CB; Data analyses: N-NP, PRB, CS, NS, KC, J-TB, CB; Manuscript writing: N-NP, CB with contributions from all authors. All authors critically evaluated and approved the manuscript.

MDT, MR, SF, AC and CB designed the study. PRB, MR, SF and AC collected the data. N-NP, CS, NS, J-TB, AC and CB prepared the data for analyses and performed initial quality controls. N-NP, PRB, CS, NS, KC, J-TB and CB performed data analyses. N-NP and CB wrote the manuscripts and prepared figures and tables, with contributions from all authors. All authors critically evaluated and approved the manuscript.

Acknowledgements

We thank all volunteers from the AWI-Gen, UGR and UKB studies who have made this project possible. The AWI-Gen Collaborative Centre is funded by the National Human Genome Research Institute (NHGRI), Office of the Director (OD), Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), the National Institute of Environmental Health Sciences (NIEHS), the Office of AIDS Research (OAR) and the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), of the National Institutes of Health under award number U54HG006938 and its supplements, as part of the H3Africa

Consortium. This research has been conducted using the UK Biobank Resource under Application 4892. This work was supported by the University of Leicester and Health Data Research UK, an initiative funded by UK Research and Innovation, Department of Health and Social Care (England) and the devolved administrations, and leading medical research charities. This study used the ALICE and SPECTRE High Performance Computing Facilities at the University of Leicester and the ZA-Wits-Core Cluster at University of the Witwatersrand. As this research was funded in whole, or in part, by the Wellcome Trust, for the purpose of open access, the author will apply a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.
C.B. was supported by an internal fellowship at the University of Leicester from the Wellcome Trust Institutional Strategic Support Fund (204801/Z/16/Z) and a UKRI Innovation Fellowship at Health Data Research UK (MR/S003762/1). M.D.T. is supported by a Wellcome Trust Investigator Award (WT202849/Z/16/Z). M.R. is a South African Research Chair in Genomics and Bioinformatics of African populations hosted by the University of the Witwatersrand, funded by the Department of Science and Technology, and administered by the National Research Foundation. C.S. acknowledges H3Africa Bioinformatics Network (H3ABioNet) Node, National Biotechnology Development Agency (NABDA), and the Center for Genomics Research and Innovation (CGRI) Abuja, Nigeria. S.F. is funded by the Wellcome International Intermediate fellowship (220740/Z/20/Z) at the MRC/UVRI and LSHTM.
The views expressed here do not necessarily reflect the views of the funders.

Tables

Table 1. Sample size (and mean age) for each phenotype in each dataset for cases (1) and controls (0).

	SI - Smoking initiation ever (1) vs never (0)		SC - Smoking cessation current (1) vs previous (0)	
Cohort	ever	never	current	previous
AWI-east	425 (50yrs)	374 (49yrs)	185 (49yrs)	239 (50yrs)
AWI-south	1343 (53yrs)	811 (56yrs)	876 (51yrs)	466 (56yrs)
AWI-west	782 (50yrs)	1045 (50yrs)	481 (50yrs)	301 (51yrs)
UGR	544 (50yrs)	2100 (27yrs)	420 (49yrs)	124 (57yrs)
UKB-AFR	1167 (52yrs)	1967 (51yrs)	509 (50yrs)	656 (54yrs)

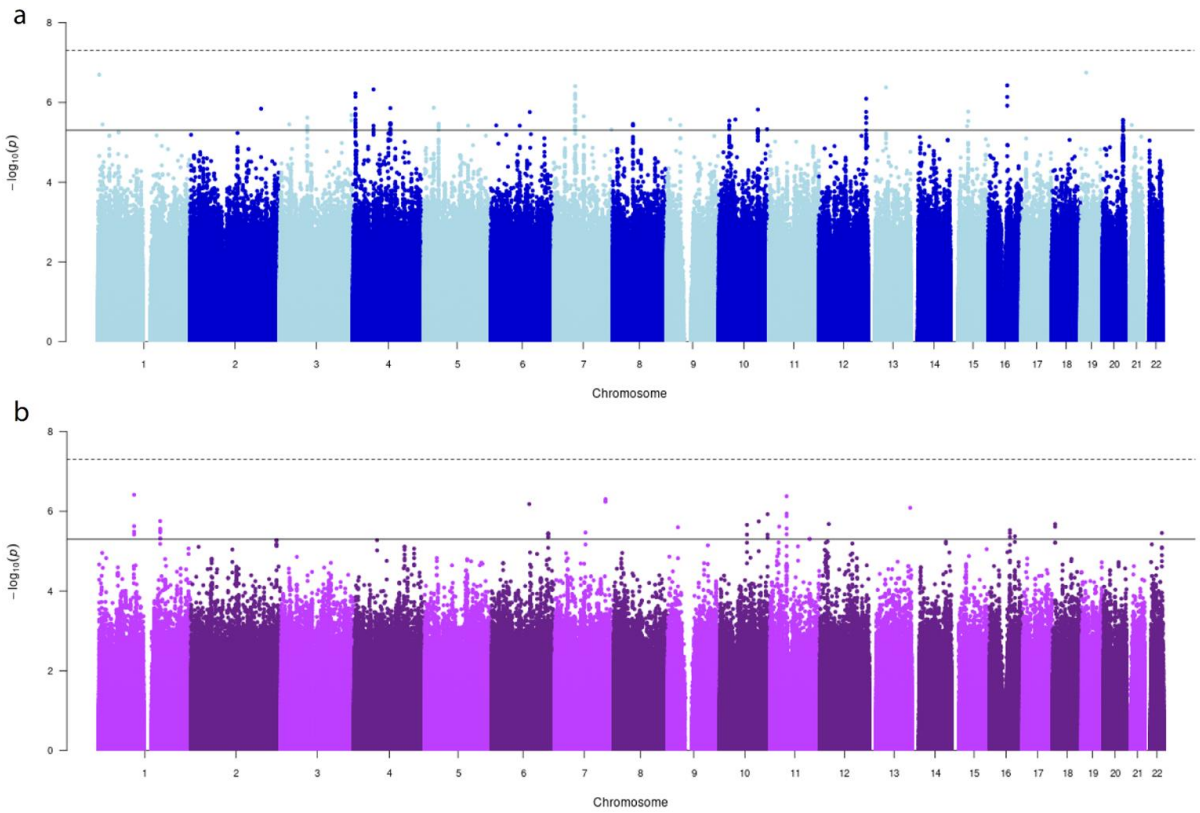
Table 2. Meta-analysis sentinel variants for smoking initiation (SI) and smoking cessation (SC).

Trait	rsID	CHR	POS	EA	NEA	OR	CI	p-value	I2	Direction of effect	AFR_EAF	EUR_EAF
SI	rs116934871	1	12772253	A	G	0.810	[0.742-0.884]	3.57×10^{-6}	0	- / - / -	0.149	0.000
	rs114828540	4	6488077	C	G	1.530	[1.297-1.804]	5.91×10^{-7}	0	+ / + / +	0.028	0.000
	rs75104774	7	156210294	T	C	1.530	[1.273-1.839]	4.77×10^{-6}	4.962	+ / + / +	0.027	0.000
	rs10160111	10	44936177	T	C	1.465	[1.252-1.715]	2.68×10^{-6}	0	+ / + / +	0.033	0.001
	rs11636198	15	47879004	C	A	0.859	[0.808-0.913]	1.70×10^{-6}	0	- / - / -	0.277	0.553
	rs74019935	16	51217700	T	A	0.822	[0.762-0.888]	7.27×10^{-7}	0	- / - / -	0.169	0.013
	rs4808946	19	15193912	C	T	1.175	[1.106-1.249]	1.78×10^{-7}	0	+ / + / +	0.595	0.873
SC	rs12443715	16	55782924	A	G	1.377	[1.205-1.572]	2.97×10^{-6}	0.641	+ / + / +	0.148	0.711

rsID: rs number; CHR: chromosome; POS: b37 position; EA: effect allele; NEA: non-effect allele; OR: odds ratio; CI: 95% confidence interval; I2: percentage of total variation across studies due to heterogeneity; direction of effect: in order, AWI-Gen, UKB-AFR, UGR; AFR_EAF: effect allele frequency in AFR 1000GP[30]; EUR_EAF: effect allele frequency in EUR 1000GP[30].

Figures

Figure 1. Manhattan plots of step2 GWAS meta-analysis: a) smoking initiation, b) smoking cessation; continuous line, suggestive p-value (p) significance threshold (5×10^{-6}); dashed line, genome-wide significance threshold (5×10^{-8}). Number of participants and variants analysed is reported in Supplementary Figure 1.



URLs

GWASCatalog: https://www.ebi.ac.uk/gwas/api/search/downloads/alternative;
eQTLs data sources in FUMA: https://fuma.ctglab.nl/tutorial#eQTLs;
Chromatin Interaction datasets in FUMA: https://fuma.ctglab.nl/tutorial#chromatin-interactions;
BRAINEAC: http://www.braineac.org/;
IEU OpenGWAS project: https://gwas.mrcieu.ac.uk/datasets/;
ukb-b: http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank;
ukb-d: http://www.nealelab.is/uk-biobank;
ukb-e: https://pan.ukbb.broadinstitute.org/;
GeneCards - *SEMA6D*: https://www.genecards.org/cgi-bin/carddisp.pl?gene=SEMA6D;
GWASCatalog- *SEMA6D*: https://www.ebi.ac.uk/gwas/genes/SEMA6D

References

1. WHO. The top 10 causes of death. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (2020).
2. WHO. WHO report on cancer: setting priorities, investing wisely and providing care for all. https://www.who.int/publications/i/item/who-report-on-cancer-setting-priorities-investing-wisely-and-providing-care-for-all (2020).
3. WHO. Fact Sheets Tobacco. https://www.who.int/news-room/fact-sheets/detail/tobacco (2021).
4. WHO. FCTC. https://www.who.int/fctc/text_download/en/ (2003).
5. WHO. The MPOWER package. https://www.who.int/tobacco/mpower/2008/en/ (2008).
6. WHO. Tobacco Control in Africa. https://www.afro.who.int/health-topics/tobacco-control (2021).
7. J D*, et al. The Tobacco Atlas. Atlanta: American Cancer Society and Vital Strategies* (American Cancer Society Inc, Atlanta, 2018).
8. WHO. WHO global report on trends in prevalence of tobacco smoking 2000–2025, third edition. Geneva: World Health Organization. https://www.who.int/publications/i/item/who-global-report-on-trends-in-prevalence-of-tobacco-use-2000-2025-third-edition (2019).
9. Batini C, Ahmed T, Ameer S, Kilonzo G, Ozoh OB, van Zyl-Smit RN. Smoking cessation on the African continent: Challenges and opportunities. *Afr J Thorac Crit Care Med* **25**, 46-48 (2019).
10. Sreeramareddy CT, Pradhan PM, Sin S. Prevalence, distribution, and social determinants of tobacco use in 30 sub-Saharan African countries. *BMC Med* **12**, 243 (2014).
11. Boua PR*, et al.* Prevalence and socio-demographic correlates of tobacco and alcohol use in four sub-Saharan African countries: a cross-sectional study of middle-aged adults. *BMC Public Health* **21**, 1126 (2021).
12. Office for National Statistics. Adult smoking habits in the UK: 2019. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/bulletins/adultsmokinghabitsingreatbritain/2019 (2019).
13. David SP*, et al.* Genome-wide meta-analyses of smoking behaviors in African Americans. *Transl Psychiatry* **2**, e119 (2012).
14. Chenoweth MJ*, et al.* Genome-wide association study of a nicotine metabolism biomarker in African American smokers: impact of chromosome 19 genetic influences. *Addiction* **113**, 509-523 (2018).
15. Visscher PM*, et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* **101**, 5-22 (2017).
16. Loukola A*, et al.* A Genome-Wide Association Study of a Biomarker of Nicotine Metabolism. *PLoS Genet* **11**, e1005498 (2015).
17. Liu M*, et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet* **51**, 237-244 (2019).
18. Gurdasani D, Barroso I, Zeggini E, Sandhu MS. Author Correction: Genomics of disease risk in globally diverse populations. *Nat Rev Genet* **20**, 562 (2019).

19. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet* **51**, 584-591 (2019).
20. Hancock DB*, et al.* Genome-wide association study across European and African American ancestries identifies a SNP in DNMT3B contributing to nicotine dependence. *Mol Psychiatry* **23**, 1911-1919 (2018).
21. Xu K*, et al.* Genome-wide association study of smoking trajectory and meta-analysis of smoking status in 842,000 individuals. *Nat Commun* **11**, 5302 (2020).
22. Ramsay M*, et al.* H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries. *Glob Health Epidemiol Genom* **1**, e20 (2016).
23. Gurdasani D*, et al.* Uganda Genome Resource Enables Insights into Population History and Genomic Discovery in Africa. *Cell* **179**, 984-1002 e1036 (2019).
24. Bycroft C*, et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
25. Asiki G*, et al.* The general population cohort in rural south-western Uganda: a platform for communicable and non-communicable disease studies. *Int J Epidemiol* **42**, 129-141 (2013).
26. Shrine N*, et al.* New genetic signals for lung function highlight pathways and chronic obstructive pulmonary disease associations across multiple ancestries. *Nat Genet* **51**, 481-493 (2019).
27. McCarthy S*, et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* **48**, 1279-1283 (2016).
28. Choudhury A BJ, Chikowore T, et al. Meta-analysis of ~25,000 continental Africans provides insights into the genetic architecture of lipid traits. *Manuscript under review* (2020).
29. Gurdasani D*, et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327-332 (2015).
30. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
31. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
32. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* **44**, 821-824 (2012).
33. Purcell S*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
34. Turner SD. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *The Journal of Open Source Software* **3**, 731 (2018).
35. Bulik-Sullivan BK*, et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015).
36. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet* **88**, 586-598 (2011).
37. Yang J*, et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-375, S361-363 (2012).

38. Benner C, Spencer CC, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493-1501 (2016).
39. Thorgeirsson TE*, et al.* Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet* **42**, 448-453 (2010).
40. Tobacco and Genetics Consortium. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* **42**, 441-447 (2010).
41. Gelernter J*, et al.* Genome-wide association study of nicotine dependence in American populations: identification of novel risk loci in both African-Americans and European-Americans. *Biol Psychiatry* **77**, 493-503 (2015).
42. Hancock DB*, et al.* Genome-wide meta-analysis reveals common splice site acceptor variant in CHRNA4 associated with nicotine dependence. *Transl Psychiatry* **5**, e651 (2015).
43. Wain LV*, et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* **3**, 769-781 (2015).
44. Brazel DM*, et al.* Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use. *Biol Psychiatry* **85**, 946-955 (2019).
45. Karlsson Linner R*, et al.* Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. *Nat Genet* **51**, 245-257 (2019).
46. Erzurumluoglu AM*, et al.* Meta-analysis of up to 622,409 individuals identifies 40 novel smoking behaviour associated genetic loci. *Mol Psychiatry* **25**, 2392-2409 (2019).
47. Chen J*, et al.* Genome-Wide Meta-Analyses of FTND and TTFC Phenotypes. *Nicotine Tob Res* **22**, 900-909 (2020).
48. Quach BC*, et al.* Expanding the genetic architecture of nicotine dependence and its shared genetics with multiple traits. *Nat Commun* **11**, 5562 (2020).
49. Buniello A*, et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
50. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310-315 (2014).
51. McLaren W*, et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
52. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
53. Folkersen L. Genomic evaluation of circulating proteins for drug target characterisation and precision medicine. Preprint at https://www.biorxiv.org/content/10.1101/2020.04.03.023804v1 (2020).
54. Zhou Y*, et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun* **10**, 1523 (2019).
55. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).

56. Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375-2376 (2014).
57. Elsworth B*, et al.* The MRC IEU OpenGWAS data infrastructure. Preprint at https://www.biorxiv.org/content/10.1101/2020.08.10.244293v1 (2020).
58. Hemani G*, et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, e34408 (2018).
59. Mitchell R, Elsworth BL, Mitchell R, Raistrick CA, Paternoster L, Hemani G, Gaunt TR. MRC IEU UK Biobank GWAS pipeline version 2. https://doi.org/10.5523/bris.pnoat8cxo0u52p6ynfaekeigi (2019).
60. Pan-UKB team. https://pan.ukbb.broadinstitute.org (2020).
61. Liu J*, et al.* The coexistence of copy number variations (CNVs) and single nucleotide polymorphisms (SNPs) at a locus can result in distorted calculations of the significance in associating SNPs to disease. *Hum Genet* **137**, 553-567 (2018).
62. Fromer M*, et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442-1453 (2016).
63. Võsa U. Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis. Preprint at https://www.biorxiv.org/content/10.1101/447367v1 (2018)
64. Wang D*, et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
65. Schmitt AD*, et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep* **17**, 2042-2059 (2016).
66. Giusti-Rodríguez P. Using three-dimensional regulatory chromatin interactions from adult and fetal cortex to interpret genetic results for psychiatric disorders and cognitive trait. Preprint at https://www.biorxiv.org/content/10.1101/406330v2 (2019).
67. Zhernakova DV*, et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat Genet* **49**, 139-145 (2017).
68. Schmiedel BJ*, et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715 e1716 (2018).
69. Kerimov* N*, et al.* eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. Preprint at https://www.biorxiv.org/content/10.1101/2020.01.29.924266v2 (2020).
70. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
71. Boua PR*, et al.* Novel and Known Gene-Smoking Interactions With cIMT Identified as Potential Drivers for Atherosclerosis Risk in West-African Populations of the AWI-Gen Study. *Front Genet* **10**, 1354 (2019).