Molecular Psychiatry

www.nature.com/mp

ARTICLE

Fine-mapping the *CYP2A6* regional association with nicotine metabolism among African American smokers

Jennie G. Pouget[1,2], Haidy Giratallah[1,3], Alec W. R. Langlois[1,3], Ahmed El-Boraie[1,3], Caryn Lerman[4], Jo Knight[5], Lisa Sanderson Cox[6], Nikki L. Nollen[6], Jasjit S. Ahluwalia[7], Christian Benner[8], Meghan J. Chenoweth[1,2,3] and Rachel F. Tyndale[1,2,3]✉

The nicotine metabolite ratio (NMR; 3’hydroxycotinine/cotinine) is a stable biomarker for CYP2A6 enzyme activity and nicotine clearance, with demonstrated clinical utility in personalizing smoking cessation treatment. Common genetic variation in the *CYP2A6* region is strongly associated with NMR in smokers. Here, we investigated this regional association in more detail. We evaluated the association of *CYP2A6* single-nucleotide polymorphisms (SNPs) and * alleles with NMR among African American smokers (N = 953) from two clinical trials of smoking cessation. Stepwise conditional analysis and Bayesian fine-mapping were undertaken. Putative causal variants were incorporated into an existing African ancestry-specific genetic risk score (GRS) for NMR, and the performance of the updated GRS was evaluated in both African American (n = 953) and European ancestry smokers (n = 933) from these clinical trials. Five independent associations with NMR in the *CYP2A6* region were identified using stepwise conditional analysis, including the deletion variant *CYP2A6*4* (beta = −0.90, p = 1.55×10^{-11}). Six putative causal variants were identified using Bayesian fine-mapping (posterior probability, PP = 0.67), with the top causal configuration including *CYP2A6*4*, rs116670633, *CYP2A6*9*, rs28399451, rs8192720, and rs10853742 (PP = 0.09). Incorporating these putative causal variants into an existing ancestry-specific GRS resulted in comparable prediction of NMR within African American smokers, and improved trans-ancestry portability of the GRS to European smokers. Our findings suggest that both * alleles and SNPs underlie the association of the *CYP2A6* region with NMR among African American smokers, identify a shortlist of variants that may causally influence nicotine clearance, and suggest that portability of GRSs across populations can be improved through inclusion of putative causal variants.

Molecular Psychiatry; https://doi.org/10.1038/s41380-024-02703-5

INTRODUCTION

Tobacco use remains the leading cause of preventable death and disease in North America [1]. Nicotine (the primary addictive agent in tobacco) [2] is metabolized to cotinine primarily by the liver enzyme CYP2A6, and then to 3’hydroxycotinine exclusively by hepatic CYP2A6 [3, 4]. The nicotine metabolite ratio (NMR; 3’hydroxycotinine/cotinine) is a stable biomarker for nicotine metabolism by CYP2A6 in smokers [5, 6]. Individual differences in NMR predict total nicotine clearance, and thus smoking behaviours (including cessation) as well as health outcomes. In particular, higher NMR (i.e. faster nicotine inactivation and CYP2A6 activity) is associated with greater nicotine dependence, cigarette consumption, and lung cancer risk along with lower cessation [7, 8]. Furthermore, NMR has translational potential in personalizing cessation treatment given that smokers with higher NMR show greater benefit from treatment with varenicline (compared to nicotine replacement therapy) [9, 10].

The NMR can only be reliably measured in current, regular smokers. This limits its use as a biomarker in longitudinal studies of smoking initiation or smoking-related disease risk in occasional/non-smokers, and limits the potential clinical utility of using NMR to guide personalized counselling on smoking-related risks to promote prevention efforts and behavioural change. However, because NMR is highly heritable (h^2 = 60–80% [11, 12]), an individual’s NMR could potentially be estimated using their genetic information regardless of their current smoking status (i.e. using a genetic risk score that predicts NMR). To achieve this, large-scale genetic studies of NMR are required to robustly identify the underlying genetic risk variants.

To date, most genetic studies of NMR have been undertaken in European ancestry smokers, and the genetic architecture of NMR in non-European smokers remains only partially understood, contributing to potential health disparities [13]. In European smokers, the largest GWAS of NMR conducted (n = 5,185) identified a strong genome-wide association near *CYP2A6* on chromosome 19, and a second association near *TMPRSS11E* on chromosome 4 [14]. The *CYP2A6* association pattern in European smokers was complex, with six independent variants identified in conditional analysis and a top causal configuration including 13 variants identified in Bayesian fine-mapping [14]. To our

[1]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health, Toronto, ON, Canada. [2]Department of Psychiatry, University of Toronto, Toronto, ON, Canada. [3]Department of Pharmacology & Toxicology, University of Toronto, Toronto, ON, Canada. [4]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA. [5]Data Science Institute and Medical School, Lancaster University, Lancaster, UK. [6]Department of Population Health, University of Kansas School of Medicine, Kansas City, KS, USA. [7]Departments of Behavioral and Social Sciences and Medicine, Brown University, Providence, RI, USA. [8]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. ✉email: r.tyndale@utoronto.ca

Received: 2 August 2023 Revised: 14 August 2024 Accepted: 16 August 2024
Published online: 31 August 2024

knowledge we have conducted the largest GWAS of NMR in African American smokers to date (n = 954), finding a single genome-wide association near *CYP2A6*. The association pattern in African American smokers was unique compared to that observed in Europeans [15], with 58 of the 96 genome-wide significant hits not reaching genome-wide significant in Europeans and a different lead variant (rs12459249) that was not in high linkage disequilibrium (LD) with the top variant in Europeans ($r^2 < 0.6$) [16].

While GWAS provide comprehensive coverage of single nucleotide polymorphisms (SNPs), there are several well characterized *CYP2A6* * alleles with known functional effects on CYP2A6 activity that are not well captured using standard GWAS approaches [17]. Incorporating both *CYP2A6* * alleles and common genetic variants identified by GWAS, we previously developed ancestry-specific genetic risk scores (GRSs) to estimate an individual's NMR from their genetic information [18, 19]. These GRSs explained 33.8% and 32.4% of variance in NMR in European [18] and African [19] ancestry populations, respectively, and showed reasonable prediction of slow vs. normal nicotine metabolizer status in these populations (AUC = 0.78 and 0.73, respectively) [18, 19]. As has been previously described for GRS more broadly [13], given differences in LD structure across ancestral populations these ancestry-specific GRSs showed poor portability across populations, with the European and African ancestry GRSs explaining only 18–20% of variance in NMR in the alternate population [19]. Additionally, Bloom *et al*. developed an ancestry-specific GRS for a different nicotine metabolism measure (D_2-cotinine:[D_2-nicotine+D_2-cotinine]) in Europeans using * alleles and other variants from the literature [20]. Development of a universal GRS using multi-ancestry cohorts is another promising approach, with Baurley et al. reporting similar predictive performance across African, Asian, and European ancestry smokers using machine learning algorithms to predict NMR based on age, sex, ancestry, BMI, and a set of 263 SNPs prioritized from GWAS (of which 198 were located in the *CYP2A6* region) [21].

In summary, previous large-scale efforts have been undertaken to fine-map the *CYP2A6* regional association with NMR in European ancestry smokers [14]. However, to our knowledge there has been no previous study fine-mapping the genome-wide *CYP2A6* association in African ancestry smokers. Given growing interest in developing genetic tools to assist with smoking counseling and cessation, in the current study we address this knowledge gap and the potential health disparities it creates. Building on our previous studies in a group of African Americans participating in two large smoking cessation trials (Fig. S1), here we investigated the *CYP2A6* association with NMR in more detail using an updated conditional analysis and new Bayesian fine-mapping approach to analyze both SNPs and * alleles (including structural variants) in the region. We also evaluated whether incorporating the putative causal variants identified by fine-mapping improved an existing ancestry-specific GRS to genetically predict NMR in African American populations, and the portability of this GRS to predict NMR in those of European ancestry.

MATERIALS AND METHODS

Participants

Our study sample comprised African and European ancestry smokers from two clinical trials of cessation: Pharmacogenetics of Nicotine Addiction Treatment 2 (PNAT-2; NCT01314001) [10] and Kick-it-at-Swope 3 (KIS-3; NCT00666978) [22]. The clinical trial protocols were approved by institutional review boards at all participating sites and the University of Toronto.

Study design of both PNAT-2 and KIS-3 have been described in detail elsewhere [10, 22]. Briefly, PNAT-2 randomized eligible adult smokers (aged 18-65 years, smoking ≥10 cigarettes/day) by NMR group (normal metabolizers vs. slow metabolizers) to treatment with placebo, nicotine patch, or varenicline for smoking cessation; all three treatment arms received behavioural counselling [10]. Approximately 37% of the total PNAT-2 sample were African ancestry (genetically determined based on comparison of genome-wide data to population reference panels as previously described [19], see Quality Control below for further details), and were included in the primary analyses here (n = 506, Table 1). We conducted additional analyses evaluating the portability of GRSs developed to predict NMR in African populations to the subset of PNAT-2 participants that were European ancestry (genetically determined as previously described [18], n = 933).

KIS-3 randomized eligible adult light smokers (aged ≥18 years, smoking ≤10 cigarettes/day) who self-identified as African American to treatment with bupropion or placebo for smoking cessation; both treatment arms received health education counselling [22]. Recruitment for KIS-3 was from a community-based clinic in Kansas, MO [22]. Participants who were African ancestry (genetically determined, as previously described [19], n = 458) were included in the primary analyses (Table 1).

Outcome measure

Nicotine metabolite ratio (NMR, 3'hydroxycotinine/cotinine ratio). We measured NMR as a continuous variable by determining the ratio of 3'hydroxycotinine/cotinine concentrations in blood samples collected at the time of clinical trial enrollment, when participants were smoking regularly. Cotinine and 3'hydroxycotinine concentrations were determined using liquid chromatography-tandem mass spectrometry, as previously described [23].

Genetic data collection

Genotyping. To capture common SNPs, we conducted genome-wide genotyping using the Illumina HumanOmniExpressExome-8 v1.2 array (Illumina, San Diego, CA, USA) at the Centre for Applied Genomics, Hospital for Sick Children (Toronto, ON, Canada). We also included a previously described custom iSelect® add-on, capturing an additional 2,688 variants associated with nicotine metabolism and/or smoking behaviours for richer coverage of regions of interest including *CYP2ABFGST* (chromosome 19), *CHRNA5-A3-B4* (chromosome 15), *OCT2* (chromosome 6), and *UGT2B* (chromosome 4) [15].

We directly genotyped the following 12 *CYP2A6* * alleles: *CYP2A6*46* (formerly *CYP2A6*1B*), *CYP2A6*1×2*, *CYP2A6*4*, *CYP2A6*9*, *CYP2A6*12*, *CYP2A6*17*, *CYP2A6*20*, *CYP2A6*23*, *CYP2A6*25/*26/*27* (all tagged by

Table 1. Sociodemographic and clinical characteristics of the final study sample.

	Total Sample (n = 953)	PNAT-2 (n = 504)	KIS-3 (n = 449)	Standardized Difference[a]
% Female (n)	57.9 (552)	50.4 (254)	66.4 (298)	0.33
Age ± SD (range)	47.1 ± 10.7 (19–80)	47.3 ± 9.8 (20–65)	46.8 ± 11.6 (19–80)	0.04
BMI ± SD (range)	30.8 ± 7.5 (15–68)	30.5 ± 7.1 (18–58)	31.2 ± 7.8 (15–68)	0.10
Cigarettes/day ± SD (range)	12.3 ± 6.4 (1–40)	16.3 ± 6.3 (5–40)	7.8 ± 2.6 (1–17)	1.76
Cotinine (ng/mL) ± SD (range)	260 ± 128 (14–837)	274 ± 130 (32–837)	244 ± 123 (14–681)	0.24
NMR ± SD (range)	0.35 ± 0.23 (0.01–1.79)	0.33 ± 0.20 (0.01–1.17)	0.38 ± 0.26 (0.02–1.79)	0.23

[a]Standardized differences (SD) were used to evaluate differences in study covariates between the two clinical trial samples included in the current study, with SD < 0.1 generally accepted as indicating a minimal difference between groups [58]. SD compare differences in mean/prevalence in units of the pooled standard deviation, which allows for comparison of the relative balance of variables in different units, and are not influenced by sample size [58].

rs28399440), *CYP2A6*28*, *CYP2A6*31*, *CYP2A6*35* as previously described [18, 19]. These *CYP2A6* * alleles have demonstrated functional effects on CYP2A6 activity, and include structural variants (*CYP2A6* gene deletions and duplications) as well as amino acid changes (see Table S2 for details). Individuals with structural variants (*CYP2A6*1×2*, *CYP2A6*4*, *CYP2A6*12*, *CYP2A6*34*, and *CYP2A6*53*) were re-genotyped using an approach with improved accuracy, as previously described [24].

Quality control. We performed quality control for samples and raw genotype data using PLINK [25], following standard protocols as previously described [15]. Individuals with discrepant sex, genotype call rate < 0.98, heterozygosity rate > 3 SDs from sample mean, substantial cryptic relatedness (PI_HAT > 0.185), or substantial non-African admixture (determined by visual inspection of multidimensional scaling (MDS) plots) were excluded. Self-reported African American ancestry was highly concordant with genetically determined ancestry in our sample (>95% concordance rate) [15]. Variants with call rate < 0.98, minor allele frequency (MAF) < 0.01, or Hardy-Weinberg equilibrium (HWE) p-value $< 1 \times 10^{-6}$ were excluded.

Imputation. We imputed chromosome 19 using the Michigan Imputation Server, which utilizes Minimac4 [26]. Accurately sequencing the *CYP2A6* region is challenging due to extensive variability, regions of high homology (i.e. including the pseudogene *CYP2A7*), and complex structural variation [17]; poor sequencing quality in this region reduces the quality of imputed genotype calls made using standard reference panels. Therefore, we compared the results of imputation using two different cosmopolitan reference panels: the TOPMED Version R2 reference panel (N = 97,256 with ~30% African ancestry from African, African Caribbean, or African American populations) [27], and the 1000 Genomes Phase 3 reference panel (N = 2504 with ~25% African ancestry from the following populations: Esan in Nigeria (ESN), Gambian in Western Division, Mandinka (GWD), Luhya in Webuye, Kenya (LWK), Mende in Sierra Leone (MSL), Yoruba in Ibadan, Nigeria (YRI), African Caribbean in Barbados (ACB), people with African ancestry in Southwest USA (ASW)) [28]. The TOPMED imputation was performed with prephasing of haplotypes using Eagle v2.4 and human genome build hg38 [29]. The 1000 Genomes Phase 3 imputation was performed with prephasing of haplotypes using ShapeIT v2.r79034 [30] and human genome build hg37, as previously described [31].

Post-imputation quality control was performed using PLINK [25] to exclude duplicate and multi-allelic variants, as well as variants with poor imputation quality (INFO < 0.6) or HWE p-value $< 1 \times 10^{-6}$. We then compared the density of coverage and imputation quality across the two imputation methods.

Statistical analyses

Association testing. All statistical analyses were done using R Statistical Software unless otherwise specified [32]. We used a mega-analytic approach, pooling data from both clinical trials (PNAT-2 and KIS-3) for all analyses unless otherwise specified.

Based on LD patterns in our sample, and in keeping with prior *CYP2A6* fine-mapping efforts in European ancestry smokers [14], we included variants within 5 Mb of *CYP2A6* in our analyses (chromosome 19:38,000,000–43,000,000 bp; Genome Reference Consortium Human Build 38, hg38). We evaluated the association of these variants in the *CYP2A6* region with NMR. Given the non-normal distribution of NMR in our sample, we applied rank-based inverse normal transformation using the R package RNOmni [33] and used these transformed NMR values for all analyses unless otherwise specified (Fig. S2).

Association testing was done in SNPTEST v2.5.2 [34] using linear regression to test the association of imputed genotype dosages with normalized NMR using an additive genotypic model with adjustment for age, sex, body mass index (BMI), and two ancestry-informative dimensions to account for population substructure as covariates.

Stepwise conditional analysis. To identify the number of independent associations in the *CYP2A6* region, we completed stepwise conditional analysis in SNPTEST v2.5.2 [34] by including genotype dosages for the top variant as an additional covariate in the base model described above (effectively conditioning on additive effects of the top variant), and repeating this procedure until no further association signals reached genome-wide significance ($p < 5 \times 10^{-8}$). Regional association plots were constructed using LocusZoom, with LD information from the 1000 Genomes Phase 3 African populations reference panel [35].

Bayesian fine-mapping. To identify potentially causal variants in the *CYP2A6* region, we used FINEMAP v1.4 specifying a maximum of 20 potential causal variants [36]. FINEMAP performs Bayesian fine-mapping using a shotgun stochastic search method to identify the most likely causal configuration of variants, given association summary statistics and local LD patterns [36]. We also performed exploratory functionally informed fine-mapping in FINEMAP [36] by assigning a higher prior probability to *CYP2A6* * alleles (prior probability = 0.70 for these variants being causal) compared to non-* allele variants (prior probability = 0.50). Input summary statistics for FINEMAP were obtained as described above using SNPTEST v2.5.2 [34], and the input SNP correlation matrix was computed from genotype dosages in our sample using LDstore v2.0 [37]. Regional association plots were constructed using R [32].

Variant annotation. To annotate variants identified in our analyses we used RegulomeDB [38], a publicly available database that estimates a variant's likelihood of having a regulatory function using a probability score that ranges from 0 to 1 (with 1 being most likely to be a regulatory variant). The probability score is constructed based on a machine learning model integrating functional genomic data including ChIP-seq signal, DNase-seq signal, information content change, and DeepSEA scores [38].

We also evaluated whether variants were known to influence expression of genes encoding functional proteins using publicly available expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression (GTEx) Project [39]. The GTEx Project eQTL analysis was based on whole genome sequencing and RNA-seq data collected from 838 donors (~ 13% African ancestry) across 49 tissues. Given the potential misidentification of *CYP2A6* transcripts as pseudogene *CYP2A7* due to high sequence homology, we considered eQTL data for pseudogene *CYP2A7* along with all other protein-coding genes. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 12/04/2024.

Incorporation of putative causal variants into an existing genetic risk score (GRS) for NMR. To investigate whether Bayesian fine-mapping improved the predictive power of genetically determined NMR in African American smokers, we compared our previously described GRS for this ancestral population [19] (referred to here as the **original GRS**) to GRSs including putative causal variants identified by fine-mapping in the current study. The original GRS included eight *CYP2A6* * alleles *(*1×2*, **4*, **9*, **12*, **17*, **20*, **25/*26/*27*, **35*) and three LD-independent genome-wide significant SNPs (rs12459249, rs111645190, rs185430475) identified in an earlier conditional analysis of the *CYP2A6* region [15]. The initial GRS estimation was constructed using mentholated cigarette use as an additional covariate, and explained 32.4% of the variance in log-NMR [19]. We elected to not adjust for menthol in the current study in order to maximize sample size (10% of participants were missing menthol data) and because menthol adjustment did not appreciably alter SNP effects on NMR [31]. For harmonization with data used in the current study, we therefore recalculated the weights for all variants in the original GRS using the analytic approach described below (without adjustment for mentholated cigarette use), and with *CYP2A6* * allele genotypes obtained using a more recent genotyping approach with improved accuracy [24].

The **updated GRS** included all eight *CYP2A6* * alleles from the original GRS and the six LD-independent putative causal variants identified by FINEMAP as the lead variant in their respective credible set. We did not include the three GWAS conditional hits in the *CYP2A6* region from the original GRS [19] in our updated GRS given that two of these SNPs (rs12459249 and rs111645190) were in high LD ($r^2 > 0.80$) with putative causal variants identified by fine-mapping (rs10853742 and rs28399451, respectively) and the remaining SNP (rs185430475) did not show robust association with NMR in our updated analysis ($p > 1 \times 10^{-4}$). To construct the updated GRS, the effect size of each putative causal variant was estimated separately in KIS-3 and PNAT-2 by association testing in SNPTEST v2.5.2 [34] using linear regression to test the association of imputed genotype dosages with square-root transformed NMR as the outcome variable using an additive genotypic model with adjustment for age, sex, BMI, and two ancestry-informative dimensions to account for population substructure as covariates. Given that the overall variance in log-NMR explained was comparable for GRSs with variant weights derived from linear regression against square-root or rank-transformed NMR, square-root transformed NMR was used for comparability of weights with the original GRS [20]. The overall effect size for each variant was then estimated in the total sample (KIS-3 and PNAT-2) by fixed-effects meta-analysis using the meta v1.7 R package [40], followed by multiplication of the resultant β coefficient by the standard deviation of the sqrt-NMR to

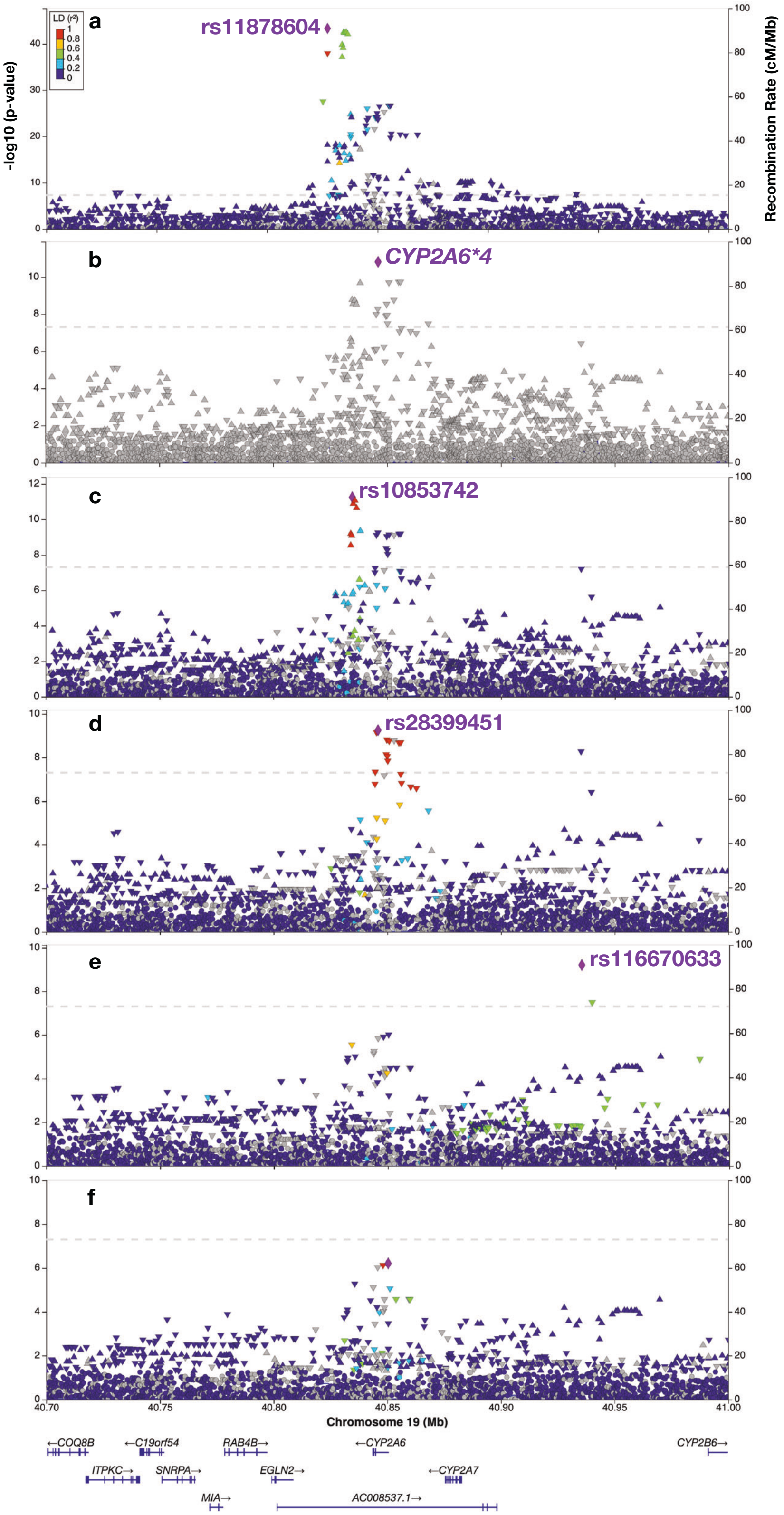

Fig. 1 Conditional analysis of the *CYP2A6* regional association with NMR in African ancestry smokers. Five independent associations were identified by conditional analysis (**a–e**), including CYP2A6 deletion variant *CYP2A6*4* (**b**); after conditioning on these five variants (**a–e**), there were no genome-wide significant associations remaining in the region (**f**). Genomic positions based on Genome Reference Consortium build 38, hg38.

unstandardize the scores [19]. The GRS was then computed for each n individual in the total sample as follows, where d refers to the number of risk alleles and β refers to the effect size for each i variant included in the GRS:

$$wGRS = \sum_{i=1}^{n} \beta_i * d_i$$

To evaluate the performance of the updated and original GRSs [19], we first calculated the variance in log-transformed NMR (log-NMR, which best represents the nicotine clearance rate [41]) explained by each GRS in linear regression models of log-NMR ~ GRS using the R function lm [32]. We also evaluated the variance in log-NMR explained by a GRS that included only the five variants identified by conditional analysis, and the six putative causal variants identified by FINEMAP.

Next, we compared the transferability of the updated and original GRSs [19] from African to European populations by calculating the variance explained in log-NMR by each GRS in the European ancestry subset of PNAT-2 (N = 933).

RESULTS

Clinical characteristics of the final discovery sample are presented in Table 1. From PNAT-2, two samples were excluded due to missing or outlying normalized NMR values. From KIS-3, eight samples were excluded due to cotinine concentrations <10 ng/mL (which suggest non-daily smoking [42]), and one sample was excluded due to missing BMI. After quality control, our final sample therefore comprised 953 African American smokers (n = 504 from PNAT-2, and n = 449 from KIS-3).

Following imputation using the TOPMED reference panel, 104,131 variants in the *CYP2A6* region (chromosome 19:38,000,000-43,000,000 bp; Genome Reference Consortium Human Build 38, hg38) were available for analysis. The median INFO score for variants in the *CYP2A6* region was 0.97 (mean = 0.92, SD = 0.096), suggesting high imputation quality. After imputation using the 1000 Genomes reference panel, 46,154 variants in the *CYP2A6* region were available for analysis with median INFO score 0.91 (mean = 0.88, SD = 0.110). Given the denser coverage and higher quality genotypes obtained from imputation using the TOPMED reference panel (Fig. S3), we used imputed genotype dosages from these data for our analyses along with 12 directly genotyped *CYP2A6* * alleles.

Within the *CYP2A6* region a total of 113 variants showed robust association ($p < 5 \times 10^{-8}$) with NMR, including four of the 12 * alleles genotyped in our sample (*CYP2A6*17*, *CYP2A6*9*, *CYP2A6*4*, and *CYP2A6*25/*26/*27*, Table S2). Overall, these *CYP2A6* * alleles were less strongly associated with NMR than other variants in the region (p-values ranging from $p = 2.06 \times 10^{-26}$ for *CYP2A6*17* to $p = 4.40 \times 10^{-8}$ for *CYP2A6*25/*26/*27*, Table S2). The strongest association was observed for rs11878604 (beta = −0.689, $p = 4.75 \times 10^{-44}$), a SNP located ~16 kb 3′ of *CYP2A6* (Fig. 1). This lead variant had a RegulomeDB probability score of 0.69 (scores range from 0 to 1, with 1 most likely to represent a variant with regulatory function) [38]; rs11878604 was also identified as an adrenal eQTL for *CYP2A6* in the GTEx Project, with the allele associated with lower NMR (i.e. reduced CYP2A6 activity) showing association with decreased *CYP2A6* expression in adrenal gland tissue (Table S1, Fig. S4).

Stepwise conditional analysis with SNPTEST [34] identified five independent associations with NMR in the *CYP2A6* region (Fig. 1, Table S1). Only the lead variant (rs11878604) was identified as an eQTL for *CYP2A6* in GTEx. After conditioning on imputed rs11878604 genotype dosage, a second independent association was identified with the directly genotyped *CYP2A6*4* allele (beta = −1.033, $p = 8.54 \times 10^{-13}$). The *CYP2A6*4* allele confers a whole gene deletion of *CYP2A6*, and individuals with this allele have correspondingly decreased CYP2A6 activity [43, 44]. Notably, in our sample *CYP2A6*4* was not in LD with any other individual variant in the region (all $r^2 < 0.15$), consistent with previous literature indicating that *CYP2A6*4* cannot be tagged by nearby SNPs [45]. *CYP2A6*4* was not genotyped in the 1000 Genomes Phase 3 African populations used as an LD reference for construction of regional association plots by LocusZoom, and as such there is no LD information displayed on the *CYP2A6*4* regional association plot (Fig. 1b). Conditioning on rs11878604 and *CYP2A6*4* revealed a third independent association with rs10853742 located ~9 kb 3′ of *CYP2A6* (beta = 0.405, $p = 5.65 \times 10^{-12}$), a SNP with a RegulomeDB probability score of 0.61 that was identified as a skin eQTL for *CYP2A7* in the GTEx Project (Table S2, Fig. S4). Conditioning on rs11878604, *CYP2A6*4*, and rs10853742 identified a fourth independent association with rs28399451 (beta = −0.340, $p = 5.59 \times 10^{-10}$). Located within intron 6 of *CYP2A6*, rs28399451 had a RegulomeDB probability score of 0.135 and was identified as a skin and peripheral nerve eQTL for *CYP2A7* in the GTEx Project (Table S1, Fig. S4). Conditioning on genotype dosages of these four variants (rs11878604, *CYP2A6*4*, rs10853742, rs28399451) identified a fifth independent association with rs116670633 (beta = −0.676, $p = 6.27 \times 10^{-10}$); this SNP was located ~85 kb 5′ of *CYP2A6*, had a RegulomeDB probability score of 0.135, and was not identified as an eQTL in the GTEx Project. After conditioning on these five variants, there were no remaining genome-wide associations with NMR (Fig. 1). These findings were consistent when association testing was run independently in PNAT-2 and KIS-3 and then meta-analyzed using an inverse-variance weighting approach (Table S1).

Bayesian fine-mapping with FINEMAP [36] identified six causal variants contributing to the *CYP2A6* region association with NMR (posterior probability of six causal variants in the region, PP = 0.67). The top causal configuration included *CYP2A6*4*, rs116670633, *CYP2A6*9*, rs28399451, rs8192720, and rs10853742; the posterior probability of these six variants representing the true causal configuration was 0.090, and together they explained 31% of the heritability of NMR **(**Fig. 2**)**. In addition to the top causal configuration, Bayesian fine-mapping identified six "credible sets" **(**Fig. 2, Table 2); each credible set can be interpreted as containing a causal variant with 95% coverage probability. The lead variants in credible sets 1–5 were highly likely to be causal (*CYP2A6*4*, rs116670633, *CYP2A6*9*, rs28399451, rs8192720; PIP for these variants being truly causal >0.50). Four of the putative causal variants identified by FINEMAP were also identified by conditional analysis (*CYP2A6*4*, rs116670633, rs28399451, rs10853742).

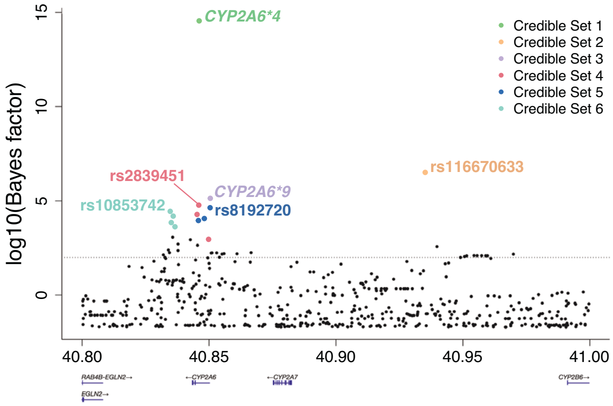

Fig. 2 Bayesian fine-mapping of *CYP2A6* association with NMR. Top causal configuration included *CYP2A6*4*, rs116670633, *CYP2A6*9*, rs28399451, rs8192720, and rs10853742; posterior probability of this top configuration being truly causal = 0.090; NMR heritability explained by top configuration (h^2) = 0.31.

Table 2. Association with NMR and functional annotations for *CYP2A6* region variants identified by fine-mapping.

95% Credible Set	Variant	Chromosome 19 Position (bp)[a]	Location Relative to *CYP2A6*	Ref Allele	Effect Allele	MAF[b]	INFO[c]	Beta[d]	SE[d]	PIP[e]	$\log_{10}$BF[f]	GTEx Project eQTLs[g]	RegulomeDB Probability Score[h]
1	*CYP2A6*4*	40843541–40850447	Whole gene deletion	—	Deletion	0.02	Typed	−1.033	0.143	1	14.57	Not available[h]	Not available[i]
2	rs116670633	40935245	84.8 kb 5′	T	G	0.03	0.99	−0.409	0.129	0.989	6.53	None	0.135
3[j]	*CYP2A6*9* (rs28399433)	40843541–40850447	Promoter (TATA box)	A	C	0.08	Typed	−0.493	0.077	0.788	5.14	*CYP2A6* (adrenal): NES = −0.51; $p = 6.0 \times 10^{-6}$ *CYP2A7* (lung): NES = 0.45; $p = 6.4 \times 10^{-6}$ *EGLN2* (artery): NES = −0.25; $p = 1.2 \times 10^{-5}$	0.554
4	rs28399451	40845938	Intron 6	G	A	0.14	0.93	−0.689	0.065	0.616	4.77	*CYP2A7* (skin): NES = 0.73; $p = 6.6 \times 10^{-9}$ *CYP2A7* (peripheral nerve): NES = 0.52; $p = 7.8 \times 10^{-5}$	0.135
	rs4803380	40845264	Intron 7	C	T	0.13	0.95	−0.691	0.066	0.339	4.27	*CYP2A7* (skin): NES = 0.73; $p = 5.3 \times 10^{-9}$ *CYP2A7* (peripheral nerve): NES = 0.52; $p = 7.8 \times 10^{-5}$	0.778
	rs28399439	40849808/12	Intron 2	AC	A	0.13	0.98	−0.700	0.065	0.022	2.92	*CYP2A6* (adipose): NES = 0.69; $p = 5.9 \times 10^{-5}$	0.983
5	rs8192720	40850405	Exon 1, synonymous	G	A	0.04	0.99[k]	−0.792	0.113	0.546	4.65	None	0.609
	rs72549439	40848131	Intron 4	G	A	0.04	0.96[k]	−0.754	0.106	0.228	4.04	None	0.244
	rs72549445	40845791	Intron 6	T	G	0.04	0.93	−0.775	0.110	0.195	3.95	None	0.981
6	rs10853742	40834668	8.9 kb 3′	G	C	0.33	0.99[k]	0.623	0.043	0.433	4.45	*CYP2A7* (skin): NES = 0.23; $p = 1.7 \times 10^{-5}$	0.609
	rs7251570	40835845	7.7 kb 3′	A	G	0.34	0.95	0.636	0.044	0.300	4.20	*CYP2A7* (skin): NES = 0.22; $p = 2.7 \times 10^{-5}$	0.590

Table 2. continued

95% Credible Set	Variant	Chromosome 19 Position (bp)[a]	Location Relative to *CYP2A6*	Ref Allele	Effect Allele	MAF[b]	INFO[c]	Beta[d]	SE[d]	PIP[e]	$\log_{10}$BF[f]	GTEx Project eQTLs[g]	RegulomeDB Probability Score[h]
	rs11667314	40835078	8.5 kb 3′	T	C	0.34	0.95	0.634	0.044	0.160	3.85	*CYP2A7* (skin): NES = 0.22; $p = 2.7 \times 10^{-5}$	0.507
	rs3865454	40836554	7.0 kb 3′	T	G	0.34	0.95	0.635	0.044	0.087	3.55	*CYP2A7* (skin): NES = 0.22; $p = 2.7 \times 10^{-5}$	0.729

[a] Human genome reference hg38.

[b] Minor allele frequency (MAF) observed in our sample.

[c] Imputation quality INFO scores were using R^2 values representing the estimated true correlation between imputed and real genotypes based on sample allele frequencies, as implemented in Minimac4 [26].

[d] beta and standard error (SE) reported are from association testing using linear regression in SNPTEST of genotype dosage ~ NMR with adjustment for age, sex, BMI, and two ancestry-informative dimensions.

[e] FINEMAP output, marginal Posterior Inclusion Probabilities (PIP) for each SNP represent the posterior probability that this SNP is causal.

[f] FINEMAP output, the Bayes factor quantifies the evidence that a particular SNP is causal, with log10 Bayes factors greater than 2 suggesting considerable evidence for causality.

[g] Publicly available expression quantitative trait loci (eQTL) data from the Genotype-Tissue Expression (GTEx) Project [39] was used to evaluate whether variants were known to influence gene expression of protein coding genes. eQTL effect alleles correspond to the effect alleles for NMR in our study, allowing for direct comparison of the directions of effect on NMR (beta) and gene expression (normalized effect size, NES).

[h] RegulomeDB is a publicly available database that estimates a variant's likelihood of regulatory function using a probability score ranging from 0 to 1 (with 1 being most likely to be a regulatory variant). The score is constructed based on a machine learning model integrating functional genomic data including ChIP-seq signal, DNase-seq signal, information content change, and DeepSEA scores [38].

[i] Because *CYP2A6*4* is a structural variant (whole gene deletion), CYP2A6*4 genotypes are not available in existing eQTL datasets which use array-based technology for genotyping.

[j] Credible set 3 also included 22 SNPs with low PIPs (mean PIP = 0.003, range=0.001 – 0.02) which tagged *CYP2A6*9* to varying degrees (mean D′ = 0.91, range = 0.41 – 1) and were therefore not included in the main table above but are detailed in Table S3.

[k] These variants were directly genotyped in our sample, but imputed genotype dosages were used for association testing (mean correlation between direct genotyping and imputed genotype dosages = 0.88, range = 0.62–0.97).

Exploratory functionally-informed FINEMAP analyses specifying a maximum of six causal variants and upweighting the 12 *CYP2A6* * alleles, which have well characterized functional effects on CYP2A6 activity (summarized in Table S2), provided consistent results and did not identify any alternative putative causal variants.

The six credible sets were made up of differing numbers of putatively causal variants, typically in high LD with each other (Fig. S5**)**. Credible set 1 included only *CYP2A6*4* (PIP = 1), which was not in significant LD with any other variant in the region. As described above, *CYP2A6*4* is a whole-gene deletion variant conferring absent CYP2A6 activity [44]; because it is a structural variant, *CYP2A6*4* eQTL data is not available in existing eQTL datasets which use array-based technology for genotyping. Credible set 2 included only rs116670633, which as described above, is a SNP located ~85 kb upstream of *CYP2A6* with limited evidence of regulatory function (PIP = 0.985); this variant was not in LD with any of the variants in other credible sets, but was in low LD with *CYP2A6*35* ($r^2 = 0.46$). Credible set 3 included *CYP2A6*9* (PIP = 0.890), a functional promoter region variant that decreases CYP2A6 activity, along with 22 other SNPs in LD with *CYP2A6*9* that each had very low PIPs (PIP range = 0.001–0.02, Table S3). Credible set 4 included three variants in high LD with each other (Fig. S5), with lead variant rs28399451 (PIP = 0.603). The variants in credible set 4 were also in moderate LD with *CYP2A6*17* ($r^2 = 0.67$–0.70). One variant in credible set 4 (rs28399439) was an adipose eQTL for *CYP2A6* in GTEx, although unexpectedly the allele associated with lower NMR (i.e. slower CYP2A6 activity) was associated with increased *CYP2A6* expression (Table 2, Fig. S4). The remaining two variants in credible set 4 (lead variant rs28399451 and rs4803380) were skin and peripheral nerve eQTLs for *CYP2A7*. Credible set 5 included three variants in high LD with each other (Fig. S5), with the top variant being rs8192720 (PIP = 0.574). The variants in credible set 5 were in moderate LD with *CYP2A6*25/*26/*27* ($r^2 = 0.50$–0.53) and low LD with *CYP2A6*20* ($r^2 = 0.37$–0.39); these three variants were not identified as eQTLs in GTEx (Table 2). Credible set 6 included four variants, with lead variant rs10853742 (PIP = 0.448). The variants in credible set 6 were in low LD with the lead variant from conditional analysis (rs11878604, $r^2 = 0.46$). All four variants in credible set 6 were skin eQTLs for *CYP2A7* in GTEx (Table 2, Fig. S4).

Incorporating the putative causal variants identified through fine-mapping into our existing ancestry-specific GRS [19] resulted in a new "updated GRS." As a benchmark, the "original GRS" comprising eight *CYP2A6* * alleles and three SNPs (rs12459249, rs111645190, rs185430475) identified in an earlier conditional analysis [15] explained 33.2% of the variance in log-NMR in our sample of African American smokers (Fig. 3a, Table 3). The updated GRS included the same eight CYP2A6 * alleles, excluded rs185430475, and included four new SNPs identified by fine-mapping (rs11667603, rs8192720, rs10853742, rs28399451). Two of these new putative causal variants (rs10853742, rs28399451) were represented by tag SNPs in the original GRS in the African ancestry sample (Fig. S5), while in the European ancestry sample only rs10853742 was represented by a proxy variant in the original GRS ($r^2 = 0.95$ with rs12459249). The updated GRS showed similar prediction of NMR as the original GRS within the African ancestry training sample (variance in log-NMR $R^2 = 0.345$ vs. 0.332 for the original GRS; Fig. 3a, c, Table 3), and improved prediction of NMR in an independent European ancestry sample ($R^2 = 0.282$ vs. 0.228 for the original GRS; Fig. 3b, d). In comparison, a GRS including the six FINEMAP putative causal variants alone improved prediction of NMR to a lesser degree ($R^2 = 0.334$ vs. 0.332 for the original GRS in African and $R^2 = 0.251$ vs. 0.228 for the original GRS in European ancestry; Table 3), suggesting the SNPs identified by fine-mapping provide independent predictive information from *CYP2A6* * alleles.

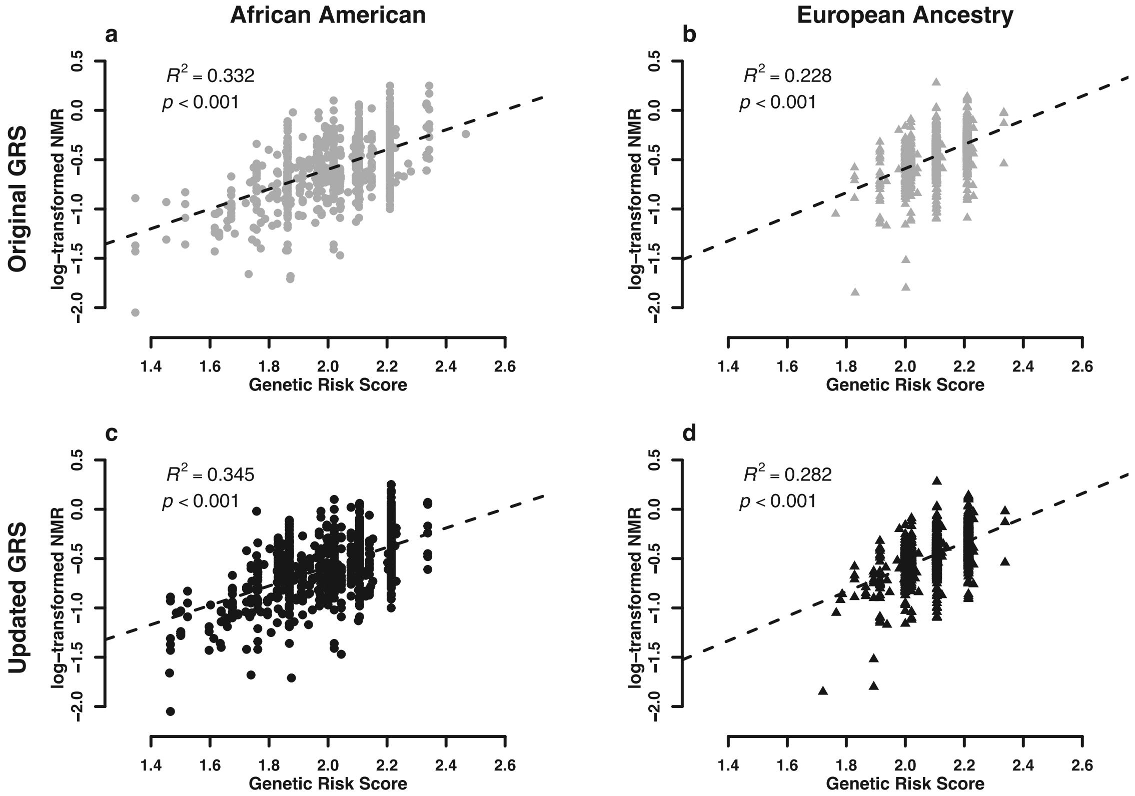

Fig. 3 Comparison of an existing African ancestry-specific genetic risk score ("Original GRS") for NMR with a genetic risk score incorporating newly identified putative causal variants ("Updated GRS"). Variance in log-NMR explained by the original GRS in African American smokers (**a**) and its portability to European ancestry smokers (**b**), as well as the updated GRS in African American smokers (**c**) and its portability to European ancestry smokers (**d**). The original GRS comprised * alleles and SNPs identified in a previous conditional analysis, whereas the updated GRS replaced these SNPs with putative causal SNPs identified by fine-mapping (for details of the variants included in the original and updated GRS, see Table 3). R^2 represents the variance in log-NMR explained.

DISCUSSION

In this study we evaluated the strong regional association of *CYP2A6* with NMR among African Americans participating in two large clinical trials of smoking cessation, performing an updated conditional analysis and novel fine-mapping analyses which improved an existing tool to genetically predict NMR. Importantly, our analyses focused on treatment-seeking individuals participating in clinical trials of smoking cessation, which excluded individuals with serious medical or psychiatric comorbidities (including comorbid substance use) and those who were pregnant or breastfeeding. As such, an important future direction will be to expand these analyses in community samples of smokers to evaluate external validity in the general population.

Previous conditional analysis of the *CYP2A6* regional association in this sample described by Chenoweth *et al* identified three independent associations (rs12459249, rs111645190, rs185430475) [15]; this earlier work did not include *CYP2A6* * alleles, and used an older reference panel for genotype imputation resulting in low-density SNP coverage. The conditional analyses and fine-mapping presented here included denser SNP genotyping coverage and 12 directly genotyped *CYP2A6* * alleles (several of which are structural variants with robust functional effects on CYP2A6 activity) [46–55], providing a more comprehensive view of variation in the *CYP2A6* region than any previous study in this population. In addition to confirming two previously reported *CYP2A6* associations with NMR in African American smokers, our conditional analysis identified three novel associations: rs11878604, *CYP2A6*4* (full *CYP2A6* gene deletion), and rs116670633.

In this first fine-mapping effort of the *CYP2A6* regional association with NMR in African populations to date, we identified six causal variants in the region (posterior probability, PP = 0.67). Prior fine-mapping using a similar analytic approach in European populations identified 13 causal variants in the region. The variants comprising the top causal configuration in our African ancestry sample were distinct from those in Europeans (*CYP2A6*4*, rs116670633, *CYP2A6*9*, rs28399451, rs8192720, rs1085374; PP = 0.090), and explained 31% of the heritability of NMR. Interestingly, *CYP2A6*9* is a known functional allele conferring reduced CYP2A6 activity [49], while the remaining four lead SNPs identified by FINEMAP were not associated with altered *CYP2A6* expression in GTEx (recognizing that regulatory information in publicly available databases is limited by methodological challenges inherent in measuring *CYP2A6* gene expression levels due to structural and copy number variation in this region, as well as high sequence homology with pseudogene *CYP2A7*). Importantly, the top putative causal variant identified was *CYP2A6*4* (PIP = 1), a loss-of-function mutation conferring whole gene deletion of *CYP2A6*. *CYP2A6*4* is not included in the vast majority of genomic studies because it cannot by genotyped accurately using array-based technologies, and is not tagged by any individual SNP in the region [45]. The strong evidence we observed for a causal association between *CYP2A6*4* and NMR highlights the importance of including *CYP2A6* structural variants in future genetic studies of tobacco-related phenotypes. To help facilitate their inclusion we recently developed a method to impute *CYP2A6* structural variants from SNP haplotypes obtained using standard genotyping array data (sensitivity >60%, false positive rate <1% in both African and European ancestry populations) [24].

Finally, we demonstrated that an updated GRS including the putative causal variants identified in African American smokers (versus those identified by conditional analysis in an earlier GRS)

Table 3. Effects of incorporating top putative causal variants identified by fine-mapping into an existing genetic risk score (“Original GRS”) to predict NMR in African American smokers.

Model	Variants Included	Ref Allele	Effect Allele	Beta[a]	GRS Weight[b]	African American		European	
						Effect Allele Freq[c]	R^{2}, [d]	Effect Allele Freq[c]	R^{2}, [d]
1 - Original GRS	CYP2A6*4[e,f]	—	Deletion	−0.935	−0.169	0.023	0.332	0.003	0.228
	CYP2A6*1×2[e]	—	Duplication	0.686	0.124	0.013		0.008	
	CYP2A6*9 (rs28399433)[e,f]	A	C	−0.473	−0.086	0.083		0.066	
	CYP2A6*12[e]	—	CYP2A6/2A7 hybrid	−0.570	−0.103	0.006		0.023	
	CYP2A6*17 (rs28399454)[e]	C	T	−0.699	−0.127	0.107		0.001	
	CYP2A6*20 (rs568811809)[e]	TT	—	−0.704	−0.127	0.015		0.000	
	CYP2A6*25/*26/*27 (rs28399440)[e]	A	G	−0.782	−0.142	0.022		0.000	
	CYP2A6*35 (rs143731390)[e]	T	A	−0.345	−0.062	0.020		0.000	
	rs12459249[e,g]	T	C	0.578	0.105	0.674		0.670	
	rs111645190[e,g]	G	A	−0.633	−0.115	0.139		0.000	
	rs185430475[e,g]	C	G	0.735	0.133	0.013		0.000	
2 - Conditional analysis variants	rs11878604	T	C	−0.651	−0.118	0.232	0.295	0.077	0.224
	CYP2A6*4[f]	—	Deletion	−0.935	−0.169	0.023		0.003	
	rs10853742[f]	G	C	0.591	0.107	0.669		0.664	
	rs28399451[f]	G	A	−0.611	−0.111	0.139		0.024	
	rs116670633[f]	T	G	−0.407	−0.074	0.031		0.002	
3 - FINEMAP top causal variants	CYP2A6*4[e,f]	—	Deletion	−0.935	−0.169	0.023	0.334	0.003	0.251
	CYP2A6*9 (rs28399433)[e,f]	A	C	−0.473	−0.086	0.083		0.066	
	rs10853742[f]	G	C	0.591	0.107	0.669		0.664	
	rs28399451[f]	G	A	−0.611	−0.111	0.139		0.024	
	rs8192720[f]	G	A	−0.743	−0.134	0.039		0.003	
	rs116670633[f]	T	G	−0.407	−0.074	0.031		0.002	
4 - Updated GRS *Original GRS * alleles + FINEMAP top causal variants*	CYP2A6*4[e,f]	—	Deletion	−0.935	−0.169	0.023	0.345	0.003	0.282
	CYP2A6*1×2[e]	—	Duplication	0.686	0.124	0.013		0.008	
	CYP2A6*9 (rs28399433)[e,f]	A	C	−0.473	−0.086	0.083		0.066	
	CYP2A6*12[e]	—	CYP2A6/2A7 hybrid	−0.570	−0.103	0.006		0.023	
	CYP2A6*17 (rs28399454)[e]	C	T	−0.699	−0.127	0.107		0.001	
	CYP2A6*20 (rs568811809)[e]	TT	—	−0.704	−0.127	0.015		0.000	
	CYP2A6*25/*26/*27 (rs28399440)[e]	A	G	−0.782	−0.142	0.022		0.000	
	CYP2A6*35 (rs143731390)[e]	‘	A	−0.345	−0.062	0.020		0.000	
	rs10853742[f]	G	C	0.591	0.107	0.669		0.664	
	rs28399451[f]	G	A	−0.611	−0.111	0.139		0.024	
	rs8192720[f]	G	A	−0.743	−0.134	0.039		0.003	
	rs116670633[f]	T	G	−0.407	−0.074	0.031		0.002	

Bold font indicates novel putative causal variants identified in the present study that were not in linkage disequilibrium with variants identified in previous non-Bayesian analyses.

[a] Beta reported is from fixed-effects meta-analysis of association testing results in PNAT-2 and KIS-3 samples using linear regression in SNPTEST of genotype dosage ~ sqrt-NMR with adjustment for age, sex, BMI, and two ancestry-informative dimensions.

[b] GRS weights were calculated as β * SD(sqrt-NMR) to unstandardize the scores.

[c] Effect allele frequency observed in our sample.

[d] Variance in log-NMR explained (R^2) by the GRS, estimated using linear regression of log-NMR ~ GRS.

[e] These variants were included in the original GRS for NMR in African American smokers described by El-Boraie *et al.* [19], with beta and GRS weights updated in the current study as described in Methods.

[f] These variants were identified as top putative causal variants by fine-mapping in the current study.

[g] These variants were identified by earlier conditional analysis of the *CYP2A6* regional association with NMR conducted in the current study sample, described by Chenoweth et al. [15].

captured similar amounts of variation in log-NMR in African ancestry individuals, and improved the portability of the GRS to European ancestry individuals. Future work evaluating the performance of our updated GRS in independent validation samples including diverse ancestry smokers is needed to evaluate whether this improved portability extends across other ancestries. One potential explanation for the improved performance of our African ancestry-specific updated GRS within European smokers is that fine-mapping identified novel variants influencing NMR that were not represented in the original GRS (i.e. rs11670633, rs8192720). Additionally, prior work has demonstrated that including putative causal variants identified by fine-mapping improves the transferability of GRS across diverse populations because of differences in LD structure which result in tag SNPs from one ancestral population no longer being good proxies for the underlying true causal variants in other ancestral populations [56, 57]. Consistent with this, the LD patterns between tag SNPs included in our original GRS and the four putatively causal SNPs included in the updated GRS were different in our African and European samples.

Overall, our results further elucidate the genetic architecture of the *CYP2A6* regional association with NMR among African American smokers and provide a shortlist of variants that may causally influence nicotine clearance in this population, which could be prioritized for investigation in future functional studies of CYP2A6 activity. In particular, the strong evidence for a causal association observed between *CYP2A6*4* and NMR highlights the importance of including *CYP2A6* structural variants in future genetic studies of tobacco-related phenotypes. Finally, the potential utility of genomic data - including genetic risk scores (GRS) - in medical decision making is growing and complements the utility of other biomarkers such as NMR, particularly in situations where NMR measurements are not available or feasible (i.e. non-smokers). Given that incorporating putative causal variants improved trans-ancestry portability of an existing GRS for NMR in this study, our results demonstrate the broader value of fine-mapping efforts as a tool to refine and improve the potential clinical utility of GRS across diverse populations which may ultimately help address potential health disparities exacerbated by existing Euro-centric GWAS data [13].

DATA AVAILABILITY

The data generated and analyzed during the current study are not publicly available due to the lack of research ethics approval for sharing full genotype or summary-level data. Researchers interested in accessing the data should contact the corresponding author, who will facilitate the review of such requests in accordance with ethical guidelines and data use agreements.

REFERENCES

1. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health, Atlanta (GA): Centers for Disease Control and Prevention (US). The health consequences of smoking—50 years of progress: A report of the Surgeon General. 2014.
2. Benowitz NL. Nicotine addiction. N. Engl J Med. 2010;362:2295–303.
3. Nakajima M, Yamamoto T, Nunoya K, Yokoi T, Nagashima K, Inoue K, et al. Role of human cytochrome P4502A6 in C-oxidation of nicotine. Drug Metab Disposition. 1996;24:1212–7.
4. Nakajima M, Yamamoto T, Nunoya K, Yokoi T, Nagashima K, Inoue K, et al. Characterization of CYP2A6 involved in 3′-hydroxylation of cotinine in human liver microsomes. J Pharmacol Exp Therapeutics. 1996;277:1010–5.
5. Benowitz NL, St Helen G, Dempsey DA, Jacob P III, Tyndale RF. Disposition kinetics and metabolism of nicotine and cotinine in African American smokers: Impact of CYP2A6 genetic variation and enzymatic activity. Pharmacogenet Genomics. 2016;26:340–50.
6. Dempsey D, Tutka P, Jacob P, Allen F, Schoedel K, Tyndale RF, et al. Nicotine metabolite ratio as an index of cytochrome P450 2A6 metabolic activity. Clin Pharm Ther. 2004;76:64–72.
7. Chenoweth MJ, Tyndale RF. Pharmacogenetic optimization of smoking cessation treatment. Trends Pharm Sci. 2017;38:66.
8. Murphy SE. Biochemistry of nicotine metabolism and its relevance to lung cancer. J Biol Chem. 2021;296:1–16.
9. Lerman C, Tyndale R, Patterson F, Wileyto EP, Shields PG, Pinto A, et al. Nicotine metabolite ratio predicts efficacy of transdermal nicotine for smoking cessation. Clin Pharm Ther. 2006;79:600–8.
10. Lerman C, Schnoll RA, Hawk LW, Cinciripini P, George TP, Wileyto EP, et al. Use of the nicotine metabolite ratio as a genetically informed biomarker of response to nicotine patch or varenicline for smoking cessation: A randomised, double-blind placebo-controlled trial. Lancet Respir Med. 2015;3:131–8.
11. Loukola A, Buchwald J, Gupta R, Palviainen T, Hällfors J, Tikkanen E, et al. A genome-wide association study of a biomarker of nicotine metabolism. PLoS Genet. 2015;11:e1005498.
12. Swan GE, Lessov-Schlaggar CN, Bergen AW, He Y, Tyndale RF, Benowitz NL. Genetic and environmental influences on the ratio of 3′hydroxycotinine to cotinine in plasma and urine. Pharmacogenet Genomics. 2009;19:398.
13. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Current clinical use of polygenic scores will risk exacerbating health disparities. Nat Genet. 2019;51:584–91.
14. Buchwald J, Chenoweth MJ, Palviainen T, Zhu G, Benner C, Gordon S, et al. Genome-wide association meta-analysis of nicotine metabolism and cigarette consumption measures in smokers of European descent. Mol Psychiatry. 2021;26:2223.
15. Chenoweth MJ, Ware JJ, Zhu AZX, Cole CB, Cox LS, Nollen N, et al. Genome-wide association study of a nicotine metabolism biomarker in African American smokers: impact of chromosome 19 genetic influences. Addiction 2018;113:523.
16. Alexander TA, Machiela MJ. LDpop: an interactive online tool to calculate and visualize geographic LD patterns. BMC Bioinforma. 2020;21:14.
17. Wassenaar CA, Zhou Q, Tyndale RF. CYP2A6 genotyping methods and strategies using real-time and end point PCR platforms. Pharmacogenomics 2015;17:147–62.
18. El-Boraie A, Taghavi T, Chenoweth MJ, Fukunaga K, Mushiroda T, Kubo M, et al. Evaluation of a weighted genetic risk score for the prediction of biomarkers of CYP2A6 activity. Addiction Biol. 2020;25:e12741.
19. El-Boraie A, Chenoweth MJ, Pouget JG, Benowitz NL, Fukunaga K, Mushiroda T, et al. Transferability of ancestry-specific and cross-ancestry CYP2A6 activity genetic risk scores in African and European populations. Clin Pharm Ther. 2021;110:975–85.
20. Bloom J, Hinrichs AL, Wang JC, Von Weymarn LB, Kharasch ED, Bierut LJ, et al. The contribution of common CYP2A6 alleles to variation in nicotine metabolism among European-Americans. Pharmacogenet Genomics. 2011;21:403–16.
21. Baurley JW, Bergen AW, Ervin CM, Park SL, Murphy SE, McMahan CS. Predicting nicotine metabolism across ancestries using genotypes. BMC Genomics. 2022;23:663.
22. Cox LS, Nollen NL, Mayo MS, Choi WS, Faseru B, Benowitz NL, et al. Bupropion for smoking cessation in African American light Smokers: A randomized controlled trial. J Natl Cancer Inst. 2012;104:290–8.
23. Jacob P, Yu L, Duan M, Ramos L, Yturralde O, Benowitz NL. Determination of the nicotine metabolites cotinine and trans-3′-hydroxycotinine in biologic fluids of smokers and non-smokers using liquid chromatography-tandem mass spectrometry: biomarkers for tobacco smoke exposure and for phenotyping cytochrome P450 2A6 activity. J Chromatogr B Anal Technol Biomed Life Sci. 2011;879:267–76.
24. Langlois AWR, El-Boraie A, Pouget JG, Cox LS, Ahluwalia JS, Fukunaga K, et al. Genotyping, characterization, and imputation of known and novel CYP2A6 structural variants using SNP array data. J Hum Genet. 2023;68:533–41. https://doi.org/10.1038/S10038-023-01148-Y.
25. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
26. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016;48:1284–7.
27. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 2021;590:290–9.
28. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, et al. A global reference for human genetic variation. Nature 2015;526:68–74.
29. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet. 2016;48:1443–8.
30. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 2014;10:e1004234.

31. Chenoweth MJ, Cox LS, Nollen NL, Ahluwalia JS, Benowitz NL, Lerman C, et al. Analyses of nicotine metabolism biomarker genetics stratified by sex in African and European Americans. Sci Rep. 2021;11:1–12.
32. R Core Team. R: A language and environment for statistical computing. 2021.
33. McCaw Z. RNOmni: Rank Normal Transformation Omnibus Test. R package version 1.0.0. https://CRAN.R-project.org/package=RNOmni. 2020.
34. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007;39:906–13.
35. Boughton AP, Welch RP, Flickinger M, Vandehaar P, Taliun D, Abecasis GR, et al. LocusZoom.js: interactive and embeddable visualization of genetic association study results. Bioinformatics 2021;37:3017–8.
36. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. Bioinformatics 2016;32:1493–501.
37. Benner C, Havulinna AS, Järvelin MR, Salomaa V, Ripatti S, Pirinen M. Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies. Am J Hum Genet. 2017;101:539–51.
38. Dong S, Boyle AP. Predicting functional variants in enhancer and promoter elements using RegulomeDB. Hum Mutat. 2019;40:1298.
39. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.
40. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: A practical tutorial. Evid Based Ment Health. 2019;22:153–60.
41. Tanner JA, Novalen M, Jatlow P, Huestis MA, Murphy SE, Kaprio J, et al. Nicotine metabolite ratio (3-hydroxycotinine/cotinine) in plasma and urine by different analytical methods and laboratories: implications for clinical implementation. Cancer Epidemiol Biomark Prev. 2015;24:1239–46.
42. Vartiainen E, Seppälä T, Lillsunde P, Puska P. Validation of self reported smoking by serum cotinine measurement in a community-based study. J Epidemiol Community Health (1978). 2002;56:167–70.
43. Gaedigk A, Ingelman-Sundberg M, Miller NA, Leeder JS, Whirl-Carrillo M, Klein TE. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the human cytochrome P450 (CYP) allele nomenclature database. Clin Pharm Ther. 2018;103:401.
44. Mwenifumbo JC, Zhou Q, Benowitz NL, Sellers EM, Tyndale RF. New CYP2A6 gene deletion and conversion variants in a population of Black African descent. Pharmacogenomics 2010;11:189.
45. Kumasaka N, Aoki M, Okada Y, Takahashi A, Ozaki K, Mushiroda T, et al. Haplotypes with copy number and single nucleotide polymorphisms in CYP2A6 locus are associated with smoking quantity in a Japanese population. PLoS One. 2012;7:e44507.
46. Mwenifumbo JC, Lessov-Schlaggar CN, Zhou Q, Krasnow RE, Swan GE, Benowitz NL, et al. Identification of novel CYP2A6*1B variants: The CYP2A6*1B allele is associated with faster in vivo nicotine metabolism. Clin Pharm Ther. 2008;83:121.
47. Rao Y, Hoffmann E, Zia M, Bodin L, Zeman M, Sellers EM, et al. Duplications and defects in the CYP2A6 gene: Identification, genotyping, and in vivo effects on smoking. Mol Pharm. 2000;58:747–55.
48. Nunoya K, Yokoi T, Kimura K, Inoue K, Kodama T, Funayama M, et al. A new deleted allele in the human cytochrome P450 2A6 (CYP2A6) gene found in individuals showing poor metabolic capacity to coumarin and (+)-cis-3,5-dimethyl-2-(3-pyridyl)thiazolidin-4-one hydrochloride. Pharmacogenetics 1998;8:239–49.
49. Kiyotani K, Yamazaki H, Fujieda M, Iwano S, Matsumura K, Satarug S, et al. Decreased coumarin 7-hydroxylase activities and CYP2A6 expression levels in humans caused by genetic polymorphism in CYP2A6 promoter region (CYP2A6*9). Pharmacogenetics 2003;13:689–95.
50. Oscarson M, McLellan RA, Asp V, Ledesma MC, Bernal Ruiz ML, Sinues B, et al. Characterization of a novel CYP2A7/CYP2A6 hybrid allele (CYP2A6*12) that causes reduced CYP2A6 activity. Hum Mutat. 2002;20:275–83.
51. Fukami T, Nakajima M, Yoshida R, Tsuchiya Y, Fujiki Y, Katoh M, et al. A novel polymorphism of human CYP2A6 gene CYP2A6*17 has an amino acid substitution (V365M) that decreases enzymatic activity in vitro and in vivo. Clin Pharm Ther. 2004;76:519–27.
52. Fukami T, Nakajima M, Higashi E, Yamanaka H, McLeod HL, Yokoi T. A novel CYP2A6*20 allele found in African-American population produces a truncated protein lacking enzymatic activity. Biochem Pharm. 2005;70:801–8.
53. Ho MK, Mwenifumbo JC, Zhao B, Gillam EMJ, Tyndale RF. A novel CYP2A6 allele, CYP2A6*23, impairs enzyme function in vitro and in vivo and decreases smoking in a population of Black-African descent. Pharmacogenet Genomics. 2008;18:67–75.
54. Mwenifumbo JC, Al Koudsi N, Man KH, Zhou Q, Hoffmann EB, Sellers EM, et al. Novel and established CYP2A6 alleles impair in vivo nicotine metabolism in a population of Black African descent. Hum Mutat. 2008;29:679–88.
55. Al Koudsi N, Ahluwalia JS, Lin SK, Sellers EM, Tyndale RF. A novel CYP2A6 allele (CYP2A6*35) resulting in an amino-acid substitution (Asn438Tyr) is associated with lower CYP2A6 activity in vivo. Pharmacogenomics J. 2009;9:274–82.
56. Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot WJ, Khera AV, et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat Genet. 2022;54:450–8.
57. Amariuta T, Ishigaki K, Sugishita H, Ohta T, Koido M, Dey KK, et al. Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. Nat Genet. 2020;52:1346–54.
58. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivar Behav Res. 2011;46:399–424.

ACKNOWLEDGEMENTS

Computations were performed on the CAMH Specialized Computing Cluster, funded by the Canada Foundation for Innovation Research Hospital Fund and the computational resources and staff expertise provided by the Digital Research Alliance of Canada (alliancecan.ca). The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. This work was funded by a Canadian Institutes of Health Research (CIHR) Project grant (PJY-159710), National Institutes of Health (NIH) Grants PGRN DA020830, CA091912, and R35 CA197461, and a Canada Research Chair in Pharmacogenomics (Tyndale). Ahluwalia funded in part by P20GM130414, a NIH funded Center of Biomedical Research Excellence (COBRE).

AUTHOR CONTRIBUTIONS

JP and RT conceived and designed the research, with all other authors providing input regarding study design and interpretation of results. AL, MC, and AEB generated the genotype data analyzed. JP, HG, AL, and CB analyzed the data. JP, HG, AL, AEB, CL, JK, CL, LC, NN, JA, CB, MC, and RT wrote the manuscript. All authors approved the submitted draft of this manuscript.

COMPETING INTERESTS

The other authors declare no conflicts of interest. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Dr. Ahluwalia received sponsored funds for travel expenses as a speaker for the 2021 and 2022 annual GTNF conference. Dr. Ahluwalia serves as a consultant and has equity in Qnovia, a start-up company developing a prescription nicotine replacement product for FDA approval. At the time of manuscript submission, Dr. El-Boraie was employed by AbbVie Corporation on unrelated topics. Other authors declare that they have no competing interests.

ETHICS APPROVAL

The clinical trial protocols for KIS-3 and PNAT-2 were approved by institutional review boards at all participating sites and the University of Toronto. Informed consent was obtained from all participants prior to data collection.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41380-024-02703-5.

Correspondence and requests for materials should be addressed to Rachel F. Tyndale.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.