

UC San Diego

UC San Diego Previously Published Works

Title

Gene-based polygenic risk scores analysis of alcohol use disorder in African Americans

Permalink

<https://escholarship.org/uc/item/6hh9645g>

Journal

Translational Psychiatry, 12(1)

ISSN

2158-3188

Authors

Lai, Dongbing
Schwantes-An, Tae-Hwi
Abreu, Marco
[et al.](#)

Publication Date

2022

DOI

10.1038/s41398-022-02029-2

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

ARTICLE OPEN



Gene-based polygenic risk scores analysis of alcohol use disorder in African Americans

Dongbing Lai¹✉, Tae-Hwi Schwantes-An¹, Marco Abreu¹, Grace Chan^{2,3}, Victor Hesselbrock², Chella Kamarajan⁴, Yunlong Liu¹, Jacquelyn L. Meyers⁴, John I. Nurnberger Jr^{1,5}, Martin H. Plawecki⁵, Leah Wetherill¹, Marc Schuckit⁶, Pengyue Zhang⁷, Howard J. Edenberg^{1,8}, Bernice Porjesz⁴, Arpana Agrawal⁹ and Tatiana Foroud¹

© The Author(s) 2022

Genome-wide association studies (GWAS) in admixed populations such as African Americans (AA) have limited sample sizes, resulting in poor performance of polygenic risk scores (PRS). Based on the observations that many disease-causing genes are shared between AA and European ancestry (EA) populations, and some disease-causing variants are located within the boundaries of these genes, we proposed a novel gene-based PRS framework (PRS_{gene}) by using variants located within disease-associated genes. Using the AA GWAS of alcohol use disorder (AUD) from the Million Veteran Program and the EA GWAS of problematic alcohol use as the discovery GWAS, we identified 858 variants from 410 genes that were AUD-related in both AA and EA. PRS_{gene} calculated using these variants were significantly associated with AUD in three AA target datasets (*P*-values ranged from 7.61E−05 to 6.27E−03; Betas ranged from 0.15 to 0.21) and outperformed PRS calculated using all variants (*P*-values ranged from 7.28E−03 to 0.16; Betas ranged from 0.06 to 0.18). PRS_{gene} were also associated with AUD in an EA target dataset (*P*-value = 0.02, Beta = 0.11). In AA, individuals in the highest PRS_{gene} decile had an odds ratio of 1.76 (95% CI: 1.32–2.34) to develop AUD compared to those in the lowest decile. The 410 genes were enriched in 54 Gene Ontology biological processes, including ethanol oxidation and processes involving the synaptic system, which are known to be AUD-related. In addition, 26 genes were targets of drugs used to treat AUD or other diseases that might be considered for repurposing to treat AUD. Our study demonstrated that the gene-based PRS had improved performance in evaluating AUD risk in AA and provided new insight into AUD genetics.

Translational Psychiatry (2022)12:266; <https://doi.org/10.1038/s41398-022-02029-2>

INTRODUCTION

Alcohol use disorder (AUD) is one of the most common public health problems [1] and both genetic and environmental factors contribute to risk. Estimates of the heritability of AUD range from 40% to 60% [2–4]. Recently, several large-scale genome-wide association studies (GWAS) of AUD-related phenotypes have reported many genetic variants associated with AUD [5–7]. These GWAS reiterated the highly polygenic underpinnings of AUD and related phenotypes where many variants contribute small effects on AUD. Consequently, polygenic risk scores (PRS) have proven to be a strong approach for assessing AUD genetic liability beyond the genome-wide significant variants [5, 7]. For instance, in our previous study of an European ancestry cohort [8], individuals comprising the top PRS decile were almost twice as likely to meet the criteria for AUD compared to all others, an estimate comparable to those published for the first-degree family history of AUD in national surveys [9, 10]. However, PRS analysis of AUD in admixed populations, such as African Americans (AA), suffer from poor performance due to the much smaller sample sizes of the discovery GWAS [5].

The performance of PRS relies on well-powered discovery GWAS to accurately select the disease-associated variants and estimate their effect sizes, and well-matched target datasets. For admixed populations, the sample sizes of the discovery GWAS comparable to European ancestry (EA) populations (hundreds of thousands to >1 million) require extensive and strategic data collection. Studies have shown that many disease-causing genes are shared among different populations [11–14]. Therefore, large-scale EA GWAS summary statistics can be leveraged to improve the performance of PRS in non-EA populations by increasing the overall discovery GWAS sample size. However, disease-associated variants may have different allele frequencies and effect sizes in different populations, and linkage disequilibrium (LD) patterns are also different [12, 15–18], i.e., the target datasets are not matched to the discovery GWAS. Furthermore, for admixed populations such as AA, the proportions of African ancestry range from close to 0 to almost 100% and are differently distributed across the genome, making AA an extremely heterogeneous population. Therefore, different AA target datasets may also have different LD patterns

¹Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA. ²Department of Psychiatry, University of Connecticut School of Medicine, Farmington, CT, USA. ³Department of Psychiatry, University of Iowa, Carver College of Medicine, Iowa City, IA, USA. ⁴Henri Begleiter Neurodynamics Lab, Department of Psychiatry, State University of New York, Downstate Medical Center, Brooklyn, NY, USA. ⁵Department of Psychiatry, Indiana University School of Medicine, Indianapolis, IN, USA. ⁶Department of Psychiatry, University of California, San Diego Medical School, San Diego, CA, USA. ⁷Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA. ⁸Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN, USA. ⁹Department of Psychiatry, Washington University School of Medicine, St. Louis, MO, USA. ✉email: dlai@iu.edu

Received: 19 January 2022 Revised: 13 June 2022 Accepted: 16 June 2022

Published online: 05 July 2022

and allele frequencies, and PRS results from one study cannot be generalized to other studies. Methods aimed to address these challenges have been proposed, but their performance remains far from ideal [19–23].

The majority of variants in the genome are likely not related to a particular condition and including them in PRS calculations will reduce the performance by introducing noises. Ideally, only variants that act on disease-causing genes should be used in calculating PRS. However, most of these variants and genes remain to be discovered. If a variant located in a gene is nominally associated (e.g., P -values < 0.05) with a disease in both EA and non-EA populations and has the same direction of effect, then it is more likely to be a shared disease-causing variant and that gene is likely to be a shared disease-causing gene across populations. Therefore, using these variants to calculate PRS is expected to improve the performance by excluding many variants in the genome that are unlikely to be related to a disease, thereby increasing the signal-to-noise ratio. Moreover, since these disease-causing variants are shared among different populations, the discovery GWAS and target datasets do not have to be well-matched and the large-scale EA GWAS can be used to increase the overall discovery GWAS sample size. Based on these observations, we propose a novel gene-based PRS framework aimed at enhancing the performance of PRS in admixed populations. We first used an AA GWAS and an EA GWAS to identify genes that were associated with AUD in both populations, then used variants located within these genes' boundaries to calculate PRS (PRS_{gene}). We compared the performance of PRS_{gene} with PRS calculated using variants located in intergenic regions (PRS_{intergenic}) and all available variants (PRS_{all}). Furthermore, for genes included in gene-based PRS calculations, we performed gene enrichment analysis using Gene Ontology (<http://geneontology.org/>) to test whether they were enriched in AUD or other biological processes that could provide novel insight into AUD mechanism. In addition, we tested in which tissues these genes were enriched. We also searched a publically available drug target database [24] to evaluate whether these genes were potential drug targets for AUD treatment, or drug targets for the treatment of other diseases but may be repurposed to treat AUD.

METHODS

Discovery GWAS

The discovery GWAS were from the meta-GWAS of problematic alcohol use in EA cohorts (EA-PAU) ($N = 435,563$) [7] and the AA GWAS of AUD from the Million Veteran Program (AA-AUD) ($N = 56,648$) [5], the largest GWAS of AUD-related phenotypes to date in EA and AA populations, respectively. The EA-PAU was a meta-analysis of problematic alcohol use [7] comprised of the AUD GWAS of the Million Veteran Program Phase I [5] and Phase II data, the alcohol dependence GWAS from the Psychiatric Genomics Consortium [25], and the GWAS of scores from the problem subscale of the Alcohol Use Disorder Identification Test (AUDIT items 4 to 10) in the UK Biobank [6]. Across both discovery GWAS, A/T, or C/G variants were excluded to avoid strand ambiguity. As we were focusing on AUD-associated variants shared between EA and AA, only variants having the same directions of effects in both EA-PAU and AA-AUD were included. Both GWAS summary statistics were downloaded from the database of genotypes and phenotypes (dbGaP: phs001672.v3.p1, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001672.v3.p1).

Target datasets

AA target datasets were drawn from 3 sources: the Collaborative Study on the Genetics of Alcoholism (COGA: $N = 3375$) [26], Study of Addiction: Genetics and Environment (SAGE: $N = 930$) [27], and YalePenn ($N = 2010$) [28]. COGA is a family cohort, in which alcohol-dependent probands and their family members from inpatient and outpatient alcohol dependence treatment facilities in seven sites were invited to participate. COGA also recruited community comparison families from a variety of sources in the same areas [26, 29]. The study was approved by Institutional review boards from all sites. Every participant provided informed consent. The Semi-

Structured Assessment for the Genetics of Alcoholism (SSAGA) was administered to individuals 18 or over and the child version of the SSAGA was used for those younger than 18 [30, 31]. SAGE (phs000092.v1.p1, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000092.v1.p1) and YalePenn (phs000425.v1.p1, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000425.v1.p1) were downloaded from dbGaP. Since COGA had more phenotypic information, if a sample in the COGA dataset was also in SAGE and/or YalePenn, it was only analyzed as part of the COGA data. SAGE and YalePenn were mixes of related and unrelated individuals, although most were unrelated. Only AA samples from COGA, SAGE, and YalePenn were used. Across all three datasets, AUD was defined as meeting lifetime criteria for DSM-IV alcohol dependence [32] or DSM-5 alcohol use disorder [33]. All other individuals were considered as controls.

The gene-based PRS_{gene} were calculated using variants located in AUD genes implicated in both AA and EA (i.e., P -values < 0.05 in both populations), consequently, they should be applicable to both populations. To test this proposition, an EA target dataset was also tested. As some EA samples of COGA and SAGE data were part of EA-PAU, they were not included as the target EA datasets, instead, EA individuals were drawn from the Indiana Biobank (<https://indianabiobank.org/>). The Indiana Biobank is a state-wide collaboration that provides centralized processing and storage of specimens that are linked to participants' electronic medical information via Regenstrief Institute at Indiana University. All Indiana Biobank individuals included in this study were unrelated. AUD in Indiana Biobank was diagnosed based on ICD9 (303 and 305.0) and ICD10 (F10) codes. Individuals not diagnosed as AUD and without AUD-associated conditions such as alcohol-associated pancreatitis were defined as controls.

Genotype data processing and imputation

Detailed information about COGA, SAGE, and YalePenn data processing has been reported previously [34–36]. Briefly, all data were combined and a common set of high quality (minor allele frequency (MAF) $> 10\%$, missing rate $< 2\%$, Hardy-Weinberg Equilibrium (HWE) P -values > 0.001) and independent (defined as $R^2 < 0.5$) variants ($N = 24,135$) was used to identify duplicate samples among different target datasets and confirm the reported family structures using PLINK [37, 38]; family structures were updated as needed. The same set of common variants was also used to estimate the principal components (PCs) of population stratification using Eigenstrat [39] with 1000 Genomes data (Phase 3, version 5, NCBI GRCh37) as the reference panel. These PCs were also used to determine AA samples (first PC between -0.0043 and 0.0115 and second PC between -0.0035 and 0.0059). Due to the different arrays used, each target dataset was imputed separately to 1000 Genomes by using SHAPEIT2 [40] followed by Minimac3 [41]. Before imputation, variants with A/T or C/G alleles, missing rates $> 5\%$, MAF $< 3\%$, and HWE P -values < 0.0001 were excluded. Imputed variants with $R^2 \geq 0.30$ and MAF $\geq 1\%$ were included in all analyses. Indiana Biobank samples were genotyped using Illumina Infinium Global Screening Array (GSA, Illumina, San Diego, CA) by Regeneron (Tarrytown, NY). Variants with missing rate $> 5\%$, MAF $< 1\%$, HWE P -value $< 1E-10$ among cases and $1E-6$ in controls were excluded as reported previously [42]. Population stratification was then estimated using the SNPRelate package [43] from Bioconductor [44]. Indiana Biobank data were also imputed to 1000 Genomes using the Michigan Imputation Server (<https://imputationserver.sph.umich.edu/index.html#pages/home>) [41]. Imputed variants with $R^2 < 0.30$ and MAF $< 1\%$ were excluded.

PRS calculation

We used PRS-CSx, a recently developed method designed for cross-ethnic polygenic prediction that showed better performance than other methods in simulation studies and real data analysis [23]. The posterior effect size of each variant was estimated via a Bayesian regression framework using continuous shrinkage priors. African and European samples from the 1000 Genomes Project were used as the LD reference panels. PRS-CSx can estimate posterior effect sizes of AA only, EA only, and meta-analysis of EA-PAU and AA-AUD. The authors of PRS-CSx recommend using estimated AA- and EA-only posterior effect sizes, then testing different linear combinations of them with different weights in a validating dataset, and choosing the one with the best performance for testing in independent datasets [23]. If the validation dataset and independent datasets are similar, e.g., having similar LD patterns and allele frequencies, this method will have more power. However, if they are different, then the weights estimated from the validating dataset will be biased toward that dataset and different from the independent datasets, resulting in loss of power. As

Table 1. Sample summary.

Population	Target dataset	# case (%male)	# control (%male)	# total	# families
AA	All ^a	2786 (60.80)	3529 (39.27)	6315	3322
	COGA	875 (62.06)	2500 (41.60)	3375	590
	SAGE	387 (59.17)	543 (37.02)	930	869
	YalePenn	1524 (60.50)	486 (29.84)	2010	1863
EA	Indiana Biobank	539 (62.15)	3515 (40.40)	4054	4054

^aCOGA, SAGE, and YalePenn combined.

we noted earlier, AA is a very heterogeneous population. The three AA target datasets in this study were recruited under different ascertainment strategies and in different regions, therefore, meta-analyzed posterior effect sizes were used in this study. In addition, since we only focused on AUD-associated variants implicated in both AA and EA, meta-analysis posterior effect sizes should provide more accurate estimates for those variants. We first selected variants that had P -values < 0.05 in both EA-PAU and AA-AUD (i.e., at least showing marginal associations) and had the same directions of effects (referred to as concordant variants). For our gene-based PRS (PRS_{gene}), only concordant variants located within gene boundaries (defined as within the region containing the gene plus 1 kb upstream of the transcription start site and 1 kb downstream of the transcription end site; annotated using ANNOVAR [45] based on NCBI RefSeq GRCh37) were used. To test whether using any concordant variants regardless of location would do as well, we calculated PRS using concordant variants located outside gene boundaries (referred to as $PRS_{\text{intergenic}}$). We also tested whether further extending gene boundaries used to calculate PRS_{gene} improved results by setting different window sizes: 10 kb, 25 kb, 50 kb, 100 kb, 250 kb, 500 kb, 1 Mb, 50 Mb, and 100 Mb. Lastly, we also used all variants across the entire genome to calculate PRS (PRS_{all}) for comparison purposes. For all AA target datasets, PRS_{gene} , $PRS_{\text{intergenic}}$, and PRS_{all} were calculated using exactly the same sets of variants, respectively, thus they were directly comparable and can be combined. PLINK [37, 38] was used to calculate PRS using the posterior effect sizes estimated by PRS-CSx and imputation dosages. All PRS were standardized as mean = 0 and standard deviation = 1 in AA (all three datasets combined) and EA target datasets separately.

Statistical analysis

As COGA, SAGE, and YalePenn include related individuals, generalized linear mixed models were used with a random effect to adjust for family relationships. For Indiana Biobank, which is a cohort of unrelated individuals, logistic regression models were used. We also stratified individuals based on PRS deciles and compared each to the bottom decile. Since the sample sizes in COGA, SAGE, and YalePenn had insufficient sample sizes in each decile, we combined all three target datasets for the stratified analyses. For all models, sex and the first 10 PCs were included as covariates. For the combined analysis of COGA, SAGE, and YalePenn data, we also included the cohort indicator as an additional covariate. Associations with P -values < 0.05 across all three target datasets were considered statistically significant for PRS_{gene} , $PRS_{\text{intergenic}}$, and PRS_{all} , respectively.

Gene enrichment analyses, searching GWAS catalog and potential drug target genes

For genes included in calculating PRS_{gene} , we performed gene ontology enrichment analysis using PANTHER (released 2021-01-24) [46] implemented in the Gene Ontology (GO) Resource (<http://geneontology.org/>, released 2021-08-18). We focused on GO Biological Processes (GOBPs). We also used Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA) [47] to test whether these genes were enriched in Differentially Expressed Gene (DEG) sets calculated using 54 tissues from The Genotype-Tissue Expression (GTEx V8) project [48]. We searched the GWAS catalog (<https://www.ebi.ac.uk/gwas/>, accessed: 2021-10-11) [49] to check whether these genes had been previously implicated in GWAS of AUD-related phenotypes. Lastly, we checked whether these genes could be potential drug targets by searching the gene list for targets of the FDA-approved drugs as well as those in current clinical trial investigations, compiled by Wang et al. [24] derived from the Informa Pharmaprojects database (<https://pharmaintelligen>

ce.informa.com/products-and-services/data-and-analysis/pharma-projects).

RESULTS

Samples used in this study are summarized in Table 1. In all target datasets, about 60% of cases were males while $< 42\%$ of controls were males.

PRS-CSx estimated posterior meta-analysis effect sizes for 1 126 428 variants and they were used to calculate PRS_{all} . In total, there were 1 533 variants with P -values < 0.05 in both EA-PAU and AA-AUD and having the same directions of effects (i.e., concordant). Among them, 858 (Table S1) and 675 (Table S3) variants were located within (410 genes, Table S2) and outside gene boundaries, respectively. As shown in Table 2, for AA target datasets, both PRS_{gene} and PRS_{all} had P -values < 0.05 in all target datasets except PRS_{all} for COGA. $PRS_{\text{intergenic}}$ had P -values ≥ 0.10 in all target datasets, demonstrating that concordant variants located within genes better stratify risk for AUD than those located in intergenic regions. Effect sizes ranged from 0.15–0.21 for PRS_{gene} , -0.02 to 0.12 for $PRS_{\text{intergenic}}$, and 0.06–0.18 for PRS_{all} , respectively. In EA, 847 of 858 variants (Table S1), and 1 061 130 of 1 126 428 variants were present in Indiana Biobank after QC; both PRS_{gene} and PRS_{all} had P -values < 0.05 (PRS_{gene} Beta = 0.11, SE = 0.02; PRS_{all} Beta = 0.34, SE = 0.05) but not $PRS_{\text{intergenic}}$ (Beta = 0.02, SE = 0.05). Results of using different window sizes to extend gene boundaries are in Table S4 and Fig. S1. The numbers of variants increased slightly with larger window sizes, and windows 50 and 100 Mb had the same number of variants, indicating that most variants are located within or close to genes. Overall, the results were similar, therefore, we kept variants within gene boundaries because it was more straightforward to determine AUD genes, as larger distances often contained multiple genes, and it is challenging to assign intergenic variants to a gene.

The association between PRS and AUD increased from the bottom decile (1st decile) to the top decile (10th decile) (Fig. 1). Using the bottom decile as the reference group, all except the 2nd and 3rd deciles showed statistically significant association with the increased odds for AUD (ORs: 1.37–1.76. Table 3) after adjusting for covariates.

Of the 410 genes, 353 were uniquely mapped to the GO database. The unmapped genes were non-coding RNAs, anti-sense RNAs, pseudo-genes, and read-throughs. Fifty-four GOBPs had false discovery rate (FDR) P -values < 0.05 , including ethanol oxidation, synaptic signaling, synapse organization, synaptic plasticity, startle response, neurogenesis, nervous system development, learning or memory, protein metabolic process, cell adhesion, cell development, cell junction organization, movement of cell or subcellular component, cell-cell signaling, regulation of signaling, etc. (Table S5). Three hundred and seventy-four genes were mapped to GTEx V8 DEG sets and enrichment results are in Fig. S2. The majority of the enrichment sets were found in brain tissues but also included liver, kidney, and other tissues as well (Fig. S2).

Only 47 genes were identified in the previous GWAS of AUD-related phenotypes (Table S2). Twenty-six genes were targets of

Table 2. Associations between AUD and PRS_{gene}, PRS_{intergenic}, and PRS_{all} in AA and EA.

Population	Target dataset	PRS _{gene}				PRS _{intergenic}				PRS _{all}			
		Beta	SE	P-value	# Variants	Beta	SE	P-value	# Variants	Beta	SE	P-value	# Variants
AA	All ^a	0.17	0.03	3.55E-08	858	0.03	0.32	675	0.12	0.03	9.42E-05	1,126,428	
	COGA	0.15	0.04	9.67E-04	858	0.02	0.61	675	0.06	0.04	0.16	1,126,428	
	SAGE	0.18	0.07	6.27E-03	858	0.12	0.10	675	0.18	0.07	0.01	1,126,428	
EA	YalePenn	0.21	0.05	7.61E-05	858	-0.02	0.76	675	0.17	0.06	7.28E-03	1,126,428	
	Indiana Biobank	0.11	0.05	0.02	847	0.02	0.59	666	0.34	0.05	2.35E-21	1,061,130	

Significant P-values are in bold.

PRS_{gene} PRS calculated using concordant variants located in genes associated with AUD in both AA and EA, PRS_{intergenic} PRS calculated using concordant variants located outside genes associated with AUD in both AA and EA, PRS_{all} PRS calculated using all variants.

^aCOGA, SAGE, and YalePenn combined.

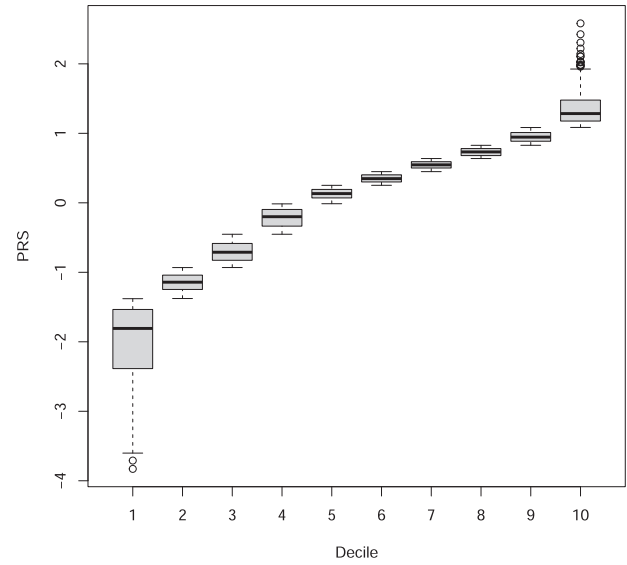


Fig. 1 Distributions of PRS_{gene} in each decile in AA. Box-plots of each PRS_{gene} decile.

Table 3. Odds ratios when comparing each PRS_{gene} decile with the bottom decile in AA (COGA, SAGE, and YalePenn combined).

Decile	OR	OR 95%CI	P-value
10	1.76	1.32–2.34	1.03E-04
9	1.72	1.29–2.29	2.05E-04
8	1.52	1.15–2.00	2.93E-03
7	1.64	1.27–2.12	1.49E-04
6	1.43	1.09–1.87	0.01
5	1.37	1.05–1.79	0.02
4	1.48	1.13–1.93	4.55E-03
3	1.22	0.91–1.62	0.18
2	1.11	0.85–1.45	0.44

Significant P-values are in bold.

drugs approved by the FDA or in clinical trials. Among them, four (*DRD2*, *PDE4B*, *GRM5*, and *SLC6A9*) were drug target genes for AUD treatment (Table S6); for those 22 genes that were targets of drugs to treat diseases other than AUD, 21 were involved in the significant GOBPs identified and five (*EIF4E*, *ESR1*, *MAPT*, *METAP1*, and *TNKS*) were reported by previous GWAS of AUD-related phenotypes (Table S2).

DISCUSSION

In this study, we found that gene-based PRS (PRS_{gene}) calculated using 858 variants from 410 genes were significantly associated with AUD in both AA and EA, and outperformed the PRS calculated using all variants (PRS_{all}) in AA. Compared to the bottom decile, those at the top PRS_{gene} decile were nearly twice as likely to be AUD cases (OR = 1.76) in AA. The 410 genes included in calculating PRS_{gene} were enriched in 54 GOBPs, and many of them are likely to be AUD-related. They were also enriched in brain tissues. In addition, four genes were targets of drugs in Phase II or III clinical trials to treat AUD; 22 genes were targets of drugs approved by the FDA or in clinical trials to treat other diseases but may be repurposed to treat AUD. Together, these findings showed that biologically meaningful polygenic scores can be characterized in non-European ancestry individuals by

leveraging methods that focus on intragenic signals with concordant directions of effects across ancestries. Furthermore, the process identified drugs already under development that could be evaluated for their potential to treat AUD.

To improve the performance of PRS, more disease-associated variants should be included and unrelated variants should be excluded. AUD is caused by many genes with small effects and in GWAS of AUD, due to the large number of variants tested, many variants that are unrelated to AUD show some degree of association (e.g., P -values < 0.05) purely by chance (i.e., false positives). If sample sizes are large (e.g., hundreds of thousands of participants or more), while the majority of AUD-associated variants are still not genome-wide significant, they usually have smaller P -values than those false positives and can still contribute to the calculation of PRS. However, when the discovery GWAS sample sizes are small to moderate, the discrimination between AUD-related and unrelated variants narrow. This leads to a reduction in PRS performance. Using large-scale EA discovery GWAS could mitigate this problem, but the improvement is limited even with sophisticated statistical methods due to the differences between the discovery GWAS and the target datasets [23]. Our gene-based PRS framework leverages the concordant variants across different populations and discriminates variants unrelated to the disease of interest leading to the improved performance of PRS. Using concordant variants also reduces the chance of selecting the wrong independent index variants due to a mismatch of LD patterns among the discovery and target datasets, as well as the external LD reference panels. Moreover, as PRS_{intergenic} were not associated with AUD in our analyses, the performance of PRS_{gene} was further improved by focusing on concordant variants within gene boundaries. PRS_{gene} had superior performance in all our AA target datasets, thus, we conclude that this strategy can be used to improve the performance of PRS when the discovery GWAS sample sizes are not sufficiently large, notable in admixture populations, and other groups that have been underrepresented in GWAS studies to date.

While PRS_{gene} outperformed PRS_{all} in AA, the opposite was observed in EA. This was expected for the following reasons. First, many GWAS findings, such as variants in *KLB* and *GCKR*, which reached genome-wide significance in EA, had P -values > 0.05 in AA (i.e., these genes may not be AUD-related in AA for some unknown mechanisms, or variants acting on these genes in AA have not been identified), therefore, they were not included in calculating PRS_{gene} but were used in calculating PRS_{all} in EA. Second, even within genes that have shown associations with AUD in both AA and EA, different causal variants may have been important in each ancestral group. One example is rs2066702 in the *ADH1B* gene. While relatively common in AA individuals (MAF = 0.18), the variant is rare in EA individuals (MAF = 0.002) (https://www.ncbi.nlm.nih.gov/snp/rs2066702?vertical_tab=true#frequency_tab). This was the only variant selected in *ADH1B* in calculating PRS_{gene}, resulting in no contribution of *ADH1B* when calculating PRS_{gene} in EA individuals from the Indiana Biobank. However, for PRS_{all}, multiple common EA variants in *ADH1B* (e.g., rs2066701, rs1042026, and rs2075633) were included, thus increasing the performance of PRS_{all}. Third, we limited inclusion to variants within gene boundaries. Although PRS_{intergenic} and PRS_{gene} with extended boundaries analyses showed that overall including intergenic concordant variants did not increase the PRS performance, however, some AUD variants are not located within gene boundaries and this may affect AA and EA disproportionately. For example, rs1229978, which is located between *ADH1B* and *ADH1C*, is much more common in EA (MAF = 0.39) than in AA (MAF = 0.15) (https://www.ncbi.nlm.nih.gov/snp/rs1229978?vertical_tab=true); therefore, not including this variant in PRS calculations had a larger impact in EA than in AA. Nevertheless, the significance of

PRS_{gene} in both AA and EA suggested that most of these genes were AUD-related in these two populations.

More than half of the 410 genes (244) were involved in 54 significant GOBPs. As expected, ethanol oxidation was among them and four genes (*ADH1B*, *ADH1C*, *ADH4*, and *ADH5*) were involved. Compromised executive functioning (i.e., neuroadaptation) is one of the major mechanisms contributing to AUD [50] and not surprisingly, several significant GOBPs related to synaptic systems (synaptic signaling, synapse organization, synaptic plasticity, startle response) were identified (46 genes). Although the role of the synaptic system in AUD is well-established [50], however, only nine genes (*CSMD1*, *DCC*, *DRD3*, *EIF4E*, *ERC2*, *LINGO2*, *MAPT*, *NRXN2*, and *TENM2*) were implicated in previous GWAS of AUD-related phenotypes. We also found significant GOBPs related to learning and memory (27 genes), consistent with previous findings that AUD and neurodegenerative diseases share some genetic liability [51]. Nervous system development-related GOBPs were significant (69 genes), and genes involved may predispose to AUD via mechanisms yet to be discovered. GOBPs such as protein metabolic process, cell adhesion, cell development, cell junction organization, movement of cell or subcellular component, cell-cell signaling, and regulation of signaling were also significant. Intuitively, these GOBPs may not seem to be AUD-related, however, among 148 genes only involved in these processes, 20 of them were reported in previous GWAS of AUD-related phenotypes with some of them, e.g., *FTO*, *PDE4B*, and *SLC39A8*, being genome-wide significant in recent large-scale GWAS of AUD [7]. In addition, there were seven genes (*EHBP1*, *EYS*, *FNBP4*, *LOC100507053*, *TNRC6A*, *WDR7*, and *ZNF462*) that were not involved in any significant GOBPs but were reported by previous GWAS of AUD-related phenotypes. Further studies are needed to elucidate the roles of these genes in predisposing to AUD. Tissue-specific enrichment showed that most genes were enriched in brain tissues as expected, however, other tissues such as liver, kidney, and pancreas also showed enrichment. Except liver, how these tissues relate to AUD remain to be discovered. By searching the drug target gene database, we found four genes (*DRD2*, *PDE4B*, *GRM5*, and *SLC6A9*) were already targets of AUD treatment drugs (Table S6). We also found 22 genes that were targets of drugs to treat other diseases (Table S6) and could be examined and/or repurposed to treat AUD. Studies have found that gene-targeted drugs were more likely to get FDA approval [24, 52, 53], therefore, identifying the roles of genes used in calculating PRS_{gene} could facilitate the development of novel treatment methods.

This study has several limitations. First, we limited to concordant variants in both AA and EA, thus, variants that may have discordant but true effects were excluded, reducing the performance of PRS_{gene}. Second, although most intergenic concordant variants did not contribute to the PRS signal as shown in the analyses of PRS_{intergenic} and extended gene boundaries with different window sizes, some of them are truly AUD related and contribute to the risk of AUD, and excluding them leading to a further reduction in the performance of PRS_{gene}. Third, we used posterior effects estimated from the meta-analysis of AA-MVP and EA-PAU. As EA-PAU had a much larger sample size (> 7 times of sample size of AA-MVP), more weight was put on effects estimated from the EA samples. Therefore, for those variants that had different sizes of effects between AA and EA, effects from the meta-analysis were biased toward the EA GWAS. Fourth, studies have shown that using functional annotations can improve the performance of PRS and increase the transferability of PRS between different populations [11, 54–56]. However, most of the available functional annotation databases were generated using European ancestry samples and were not related to AUD, therefore, those functional information were not used in this study, which may also reduce the performance of PRS_{gene}.

In summary, we calculated PRS for evaluating AUD risk that worked cross populations based on our novel gene-based PRS framework. Not only our new framework outperformed the PRS calculated using all variants in AA, but also the genes included in calculating PRS showed enrichment for biological plausible processes and are potential targets for drug development, therefore, this novel framework demonstrates the utilities of PRS beyond disease risk evaluation to the identification of biological processes and drug targets, and shed light on the genetic mechanism of AUD.

REFERENCES

- World Health Organization. Global status report on alcohol and health 2018: executive summary. Geneva: World Health Organization; 2018.
- Heath AC, Martin NG. Genetic influences on alcohol-consumption patterns and problem drinking - results from the Australian Nhand-Mrc Twin Panel Follow-up Survey. *Ann NY Acad Sci.* 1994;708:72–85.
- Prescott CA, Kendler KS. Genetic and environmental contributions to alcohol abuse and dependence in a population-based sample of male twins. *Am J Psychiatry.* 1999;156:34–40.
- Verhulst B, Neale MC, Kendler KS. The heritability of alcohol use disorders: a meta-analysis of twin and adoption studies. *Psychol Med.* 2015;45:1061–72.
- Kranzler HR, Zhou H, Kember RL, Vickers Smith R, Justice AC, Damrauer S, et al. Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nat Commun.* 2019;10:1499.
- Sanchez-Roige S, Palmer AA, Fontanillas P, Elson SL, Adams MJ, Howard DM, et al. Genome-wide association study meta-analysis of the alcohol use disorders identification test (AUDIT) in two population-based cohorts. *Am J Psychiatry.* 2019;176:107–18.
- Zhou H, Sealock JM, Sanchez-Roige S, Clarke T-K, Levey DF, Cheng Z, et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nat Neurosci.* 2020;23:809–18.
- Lai D, Johnson EC, Colbert S, Pandey G, Chan G, Bauer L, et al. Evaluating risk for alcohol use disorder: Polygenic risk scores and family history. *Alcohol Clin Exp Res.* 2022;46:374–83.
- Dawson DA, Harford TC, Grant BF. Family history as a predictor of alcohol dependence. *Alcohol Clin Exp Res.* 1992;16:572–5.
- Karriker-Jaffe KJ, Chartier KG, Bares CB, Kendler KS, Greenfield TK. Intersection of familial risk and environmental social control on high-risk drinking and alcohol dependence in a US national sample of adults. *Addict Behav.* 2021;113:106668.
- 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.
- Bitarello BD, Mathieson I. Polygenic scores for height in admixed populations. *G3.* 2020;10:4027–36.
- Shi H, Burch KS, Johnson R, Freund MK, Kichaev G, Mancuso N, et al. Localizing components of shared transethnic genetic architecture of complex traits from GWAS summary data. *Am J Hum Genet.* 2020;106:805–17.
- Wang Y, Guo J, Ni G, Yang J, Visscher PM, Yengo L. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. *Nat Commun.* 2020;11:3865.
- Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, et al. Human demographic history impacts genetic risk prediction across diverse populations. *Am J Hum Genet.* 2017;100:635–49.
- Cavazos TB, Witte JS. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Hum Genet Genomics Adv.* 2021;2:100017.
- Grinde KE, Qi Q, Thornton TA, Liu S, Shadyab AH, Chan KHK, et al. Generalizing polygenic risk scores from Europeans to Hispanics/Latinos. *Genet Epidemiol.* 2019;43:50–62.
- Schultz LM, Merikangas AK, Ruparel K, Jacquemont S, Glahn DC, Gur RE, et al. Stability of polygenic scores across discovery genome-wide association studies. *HGG Adv.* 2022;3:100091.
- Weissbrod O, Kanai M, Shi H, Gazal S, Peyrot W, Khera A, et al. Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores. *Nat Genet.* 2022;54:450–8.
- Márquez-Luna C, Loh PR, Price AL. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet Epidemiol.* 2017;41:811–23.
- Marnetto D, Pärna K, Läll K, Molinaro L, Montinaro F, Haller T, et al. Ancestry deconvolution and partial polygenic score can improve susceptibility predictions in recently admixed individuals. *Nat Commun.* 2020;11:1628.
- Cai M, Xiao J, Zhang S, Wan X, Zhao H, Chen G, et al. A unified framework for cross-population trait prediction by leveraging the genetic correlation of polygenic traits. *Am J Hum Genet.* 2021;108:632–55.
- Ruan Y, Anne Feng Y-C, Chen C-Y, Lam M, Stanley Global Asia I, Sawa A, et al. Improving polygenic prediction in ancestrally diverse populations. *Nat Genet.* 2022;54:573–80.
- Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature.* 2021;597:527–32.
- Walters RK, Polimanti R, Johnson EC, McClintick JN, Adams MJ, Adkins AE, et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nat Neurosci.* 2018;21:1656–69.
- Reich T, Edenberg HJ, Goate A, Williams JT, Rice JP, Van Eerdewegh P, et al. Genome-wide search for genes affecting the risk for alcohol dependence. *Am J Med Genet.* 1998;81:207–15.
- Bierut LJ, Agrawal A, Bucholz KK, Doheny KF, Laurie C, Pugh E, et al. A genome-wide association study of alcohol dependence. *P Natl Acad Sci USA.* 2010;107:5082–7.
- Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH, et al. Genome-wide association study of alcohol dependence: significant findings in African-and European-Americans including novel risk loci. *Mol Psychiatry.* 2014;19:41–9.
- Numberger JI Jr, Wiegand R, Bucholz K, O'Connor S, Meyer ET, Reich T, et al. A family study of alcohol dependence: coaggregation of multiple disorders in relatives of alcohol-dependent probands. *Arch Gen Psychiatry.* 2004;61:1246–56.
- Bucholz KK, Cadoret R, Cloninger CR, Dinwiddie SH, Hesselbrock VM, Numberger JI Jr, et al. A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. *J Stud Alcohol.* 1994;55:149–58.
- Hesselbrock M, Easton C, Bucholz KK, Schuckit M, Hesselbrock V. A validity study of the SSAGA—a comparison with the SCAN. *Addiction.* 1999;94:1361–70.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th edn. Washington, DC:1994.
- American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5th edn. Washington, DC:2013.
- Lai D, Wetherill L, Kapoor M, Johnson EC, Schwandt M, Ramchandani VA, et al. Genome-wide association studies of the self-rating of effects of ethanol (SRE). *Addict Biol.* 2019:e12800.
- Lai D, Wetherill L, Bertelsen S, Carey CE, Kamarajan C, Kapoor M, et al. Genome-wide association studies of alcohol dependence, DSM-IV criterion count and individual criteria. *Genes Brain Behav.* 2019;18:e12579.
- Lai D, Kapoor M, Wetherill L, Schwandt M, Ramchandani VA, Goldman D, et al. Genome-wide admixture mapping of DSM-IV alcohol dependence, criterion count, and the self-rating of the effects of ethanol in African American populations. *Am J Med Genet B Neuropsychiatr Genet.* 2021;186:151–61.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, 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.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38:904–9.
- Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. *Am J Hum Genet.* 2013;93:687–96.
- Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48:1284–7.
- Schwantes-An TH, Darlay R, Mathurin P, Masson S, Liangpunsakul S, Mueller S, et al. Genome-wide association study and meta-analysis on alcohol-associated liver cirrhosis identifies genetic risk factors. *Hepatology.* 2021;73:1920–31.
- Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics.* 2012;28:3326–8.
- Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods.* 2015;12:115–21.
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 2010;38:e164.
- Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc.* 2013;8:1551–66.
- Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8:1826.
- 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.
- Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies,

- targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 2019;47:D1005–D12.
50. Koob GF, Volkow ND. Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry.* 2016;3:760–73.
 51. Kapoor M, Chao MJ, Johnson EC, Novikova G, Lai D, Meyers JL, et al. Multi-omics integration analysis identifies novel genes for alcoholism with potential overlap with neurodegenerative diseases. *Nat Commun.* 2021;12:5071.
 52. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* 2019;15:e1008489.
 53. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. *Nat Genet.* 2015;47:856–60.
 54. Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput Biol.* 2017;13:e1005589.
 55. Liang Y, Pividori M, Manichaikul A, Palmer AA, Cox NJ, Wheeler H, et al. Polygenic transcriptome risk scores improve portability of polygenic risk scores across ancestries. *bioRxiv.* 2020:2020.11.12.373647.
 56. Weissbrod O, Hormozdiari F, Benner C, Cui R, Ulirsch J, Gazal S, et al. Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat Genet.* 2020;52:1355–63.

ACKNOWLEDGEMENTS

COGA: The Collaborative Study on the Genetics of Alcoholism (COGA), Principal Investigators B. Porjesz, V. Hesselbrock, T. Foroud; Scientific Director, A. Agrawal; Translational Director, D. Dick, includes eleven different centers: University of Connecticut (V. Hesselbrock); Indiana University (H.J. Edenberg, T. Foroud, Y. Liu, M.H. Plawecki); University of Iowa Carver College of Medicine (S. Kuperman, J. Kramer); SUNY Downstate Health Sciences University (B. Porjesz, J. Meyers, C. Kamarajan, A. Pandey); Washington University in St. Louis (L. Bierut, J. Rice, K. Bucholz, A. Agrawal); University of California at San Diego (M. Schuckit); Rutgers University (J. Tischfield, R. Hart, J. Salvatore); The Children's Hospital of Philadelphia, University of Pennsylvania (L. Almasy); Virginia Commonwealth University (D. Dick); Icahn School of Medicine at Mount Sinai (A. Goate, P. Slesinger); and Howard University (D. Scott). Other COGA collaborators include: L. Bauer (University of Connecticut); J. Nurnberger Jr., L. Wetherill, X., Xuei, D. Lai, S. O'Connor, (Indiana University); G. Chan (University of Iowa; University of Connecticut); D.B. Chorlian, J. Zhang, P. Barr, S. Kinreich, G. Pandey (SUNY Downstate); N. Mullins (Icahn School of Medicine at Mount Sinai); A. Anokhin, S. Hartz, E. Johnson, V. McCutcheon, S. Saccone (Washington University); J. Moore, Z. Pang, S. Kuo (Rutgers University); A. Merikangas (The Children's Hospital of Philadelphia and University of Pennsylvania); F. Aliev (Virginia Commonwealth University); H. Chin and A. Parsian are the NIAAA Staff Collaborators. We continue to be inspired by our memories of Henri Begleiter and Theodore Reich, founding PI and Co-PI of COGA, and also owe a debt of gratitude to other past organizers of COGA, including Ting- Kai Li, P. Michael Conneally, Raymond Crowe, and Wendy Reich, for their critical contributions. This national collaborative study is supported by NIH Grant U10AA008401 from the National Institute on Alcohol Abuse and Alcoholism (NIAAA) and the National Institute on Drug Abuse (NIDA). This study was made possible, in part, with support from the Indiana Clinical and Translational Sciences

Institute funded, in part by Award Number UL1TR002529 from the National Institutes of Health, National Center for Advancing Translational Sciences, Clinical and Translational Sciences Award, and the National Center for Research Resources, Construction grant number RR020128 and the Lilly Endowment. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors acknowledge the Indiana University Pervasive Technology Institute for providing [HPC (Big Red II, Karst, Carbonate), visualization, database, storage, or consulting] resources that have contributed to the research results reported within this paper. AA acknowledges K02DA032573, R01DA054869.

AUTHOR CONTRIBUTIONS

DL designed the study, performed the analyses, and wrote the manuscript. DL, TS, and MA processed data. TS, MA, GC, VH, CK, YL, JLM, JIN, MHP, LW, MS, PZ, HJE, BP, AA, and TF contributed to the results interpretation and preparing the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41398-022-02029-2>.

Correspondence and requests for materials should be addressed to Dongbing Lai.

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.



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

© The Author(s) 2022