Data supplement for Sanchez-Roige et al., Item-Level Genome-Wide Association Study of the Alcohol Use Disorders Identification Test in Three Population-Based Cohorts. Am J Psychiatry (doi: 10.1176/appi.ajp.2020.20091390)

CONTENTS

1 Cohorts

The present study used six large, genotyped cohorts: three for discovery and three for validation. To be eligible for inclusion, individuals were required to: (i) be primarily of non-Hispanic European ancestry, (ii) have data on all relevant outcomes and covariates, and (iii) have high-quality, imputed genotyping data. All cohorts were also required to apply appropriate individual- and variant-level quality control thresholds. The cohorts that met these criteria for both analytic stages (discovery and validation) are briefly described below in alphabetical order within each stage.

1.1 Discovery cohorts

Discovery analyses were performed in the three cohorts described below. As genotyping and data collection procedures for each cohort have been extensively described elsewhere, we refer the reader to those studies for details. Quality control thresholds for these cohorts are described in **Supplementary Section 4**.

1.1.1 Avon Longitudinal Study of Parents and Children

The Avon Longitudinal Study of Parents and Children (ALSPAC) is a longitudinal study of 14,541 pregnant mothers living in the former county of Avon, United Kingdom (UK). Details on data collection and genotyping have been described in previous publications (1–3). The ALSPAC website also contains details on all available data through a fully searchable data dictionary (http://www.bristol.ac.uk/alspac/researchers/our-data/). In the present study, only the offspring with all relevant data were used in the analyses. Accordingly, the maximum sample size used in discovery analyses was 3,582. Using REDCap (4, 5), participants completed questionnaires that included the AUDIT; for the current analyses, we used responses from the age 22 assessment where available, or responses from age 20 if the individual did not participate at age 22.

All participants provided informed consent at the time of recruitment following the recommendations of the ALSPAC Ethics and Law Committee. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee, and the Local Research Ethics

Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and S.S-R. will serve as guarantors for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). This research was specifically funded by NIH AA018333.

1.1.2 Netherlands Twin Register

The Netherlands Twin Registry (NTR) is a national register of twins and multiples, as well as their family members. The NTR follows adolescent and adult twins and their relatives over time, collecting data on numerous complex traits. NTR is a representative sample of the Dutch population (6). Details on data collection and genotyping have been described in previous publications (7, 8). In the present study, the maximum sample size for discovery analyses was 9,975.

All participants provided informed consent at the time of recruitment. Study procedures were performed in accordance with the ethical standards of institutional and/or national research committee, as well as the 1964 Helsinki declaration and its later amendments or comparable ethical standards. The NTR studies were approved by the Central Ethics Committee on Research Involving Human Subjects of the VU University Medical Centre, Amsterdam, an Institutional Review Board certified by the U.S. Office of Human Research Protections (IRB number IRB00002991 under Federal-wide Assurance- FWA00017598; IRB/institute codes, NTR 03-180).

1.1.3 UK Biobank - Discovery

The UK Biobank (UKB) is a population-based biomedical study of genetic and environmental influences on human health and wellbeing. Genotyped participants completed questionnaires related to a myriad of complex traits. A subset of participants also completed an online follow-up protocol, where they answered additional surveys related to mental health and substance use. Details on data collection have been described in previous publications (9), and can also be viewed via the Data Showcase (https://biobank.ctsu.ox.ac.uk/crystal/). The maximum sample size used for discovery analyses was 147,267.

All participants provided informed consent at the time of recruitment. Participants also provided consent for follow-up via linkage to their electronic health records. As described by Sudlow and colleagues (10) the UKB investigators developed a robust Ethical and Governance Framework to guide their study procedures, and performed all procedures in accordance with the guidelines established by the relevant ethical, institutional, and regulatory bodies.

1.2 Validation cohorts

Validation analyses were performed in the three cohorts described below. Similar to the discovery cohorts, genotyping and data collection procedures for each validation cohort have been extensively described elsewhere, and we refer the reader to those studies for details. However, as there is more variability in the quality control procedures applied across validation cohorts (*i.e.*, there was no single study-wide pipeline), quality control filters are briefly described below.

1.2.1 Collaborative Studies on Genetics of Alcoholism Study

The Collaborative Study on the Genetics of Alcoholism (COGA) was established to examine familial underpinnings of alcohol use disorders and related behaviors (11–13). Alcohol dependent probands were ascertained from inpatient or outpatient treatment facilities across 7 sites in the United States. Family members of the probands were also invited to participate. Community individuals and their family members were selected from a variety of sources and invited to participate. A proportion of the families in COGA are large and have a high density of alcoholism. In the present study, we excluded participants who (i) were not of European ancestry, (ii) elected not to drink due to personal (*e.g.*, religious, cultural, health) reasons, or (iii) did not report being “regular drinkers” (having at least one drink per month for 6 consecutive months). After these exclusions, our maximum analytic sample consisted of 6,390 individuals.

Analyzed variants were imputed using the 1000 Genomes Phase 3 (14) reference panel. Imputed SNPs with information scores < 0.30 or individual genotype probability scores < 0.90 were excluded, as were SNPs that did not pass Hardy-Weinberg equilibrium (HWE p< 1E-6), and SNPs with a minor allele frequency less than 0.05%. Additional details on genotyping procedures and related quality control can be found in previously published studies (15).

The Institutional Review Boards at all sites approved this study and all participants provided informed consent at every assessment.

1.2.2 UK Biobank - Validation

As the UK Biobank was described in **Supplementary Section 1.1.3**, it is not described again here. For the validation analyses, we selected a subset of UKB, consisting of unrelated White British participants who were not included in the discovery stage analyses ($n_{\max}$ = 239,719).

Analyzed variants were selected from the third release of the UKB imputed genotype data. In the validation cohort, imputed SNPs were subjected to the quality thresholds reported by UKB investigators, but they were also filtered on several additional thresholds: imputation quality score > .90, minor allele frequency > 1E-5, missingness < .05; unique identifiers only, bi-allelic variants only. SNPs were converted to hard-call format (certainty > .90) prior to analysis.

We included a lower number of PCs (10) in the validation analyses than GWAS (40) to avoid overfitting, given the fact that the analyses are performed on a smaller sample and use aggregate genetic predictors instead of individual variants.

1.2.3 Vanderbilt University Medical Center Biobank

The Vanderbilt University Medical Center (BioVU) biobank cohort is been extensively described previously (16, 17). BioVU is one of the largest biobanks in the United States, consisting of electronic health record data from the Vanderbilt University Medical Center on ~250,000 patients spanning 1990 to 2017. A subset of BioVU patients of European descent ($n_{\max}$ = 91,602) have been genotyped as part of various institutional and investigator-initiated projects on the Illumina MEGAEX platform, which contains more than 2 million markers. Genotyping and Quality control of this sample have been described elsewhere (16, 17).

Analyzed variants were imputed using SHAPEIT(18)/IMPUTE4(9) with the 1000 genomes phase I reference panel, and variants with INFO < .3 were excluded. A subset of SNPs in linkage disequilibrium was used to calculate relatedness and principal components of ancestry using multidimensional scaling in PLINK v1.9 (19). We randomly removed one individual from pairs of highly related individuals (pihat > .1) to avoid spurious results driven by cryptic relatedness, and restricted to a homogenous population of European descent defined by principal components of ancestry to avoid population stratification effects. Variants were removed if allele frequencies differed significantly ($P < 5x10^{-5}$) between any batch. Finally, we filtered multi-allelic and structural variants, converted dosage data to hard genotype calls, and excluded variants with uncertainty > .1 or INFO < .95

2 Phenotype construction

The AUDIT consists of 10 self-report items that are scored via 3- or 5-point ordinal scales. Items 1 through 8 are scored on an ordinal scale ranging from 0 to 4, while items 9 and 10 are scored as 0-2-4. The scoring scheme has been described in our previous GWAS of AUDIT composite scores (20). For consistency, the same phenotyping strategy was applied across all three discovery cohorts. However, in UKB, the AUDIT is administered with ‘gating’ or ‘skip’ logic. This is described in the **Supplementary Section 2.1**, where we also describe the steps taken to minimize the influence of this survey structure.

The distribution of AUDIT scores is shown in **Supplementary Table 2**. Of note, samples broadly differed in their general characteristics: UKB participants were generally older than NTR and ALSPAC, and UKB showed higher socioeconomic backgrounds than the general population (9, 10, 21). The ratio of females was similar across the cohorts (61%, 62%, 57% in ALSPAC, NTR, UKB, respectively).

2.1 Multiple imputation via chained equations in UK Biobank

As noted above, the AUDIT was administered with skip logic in UKB, such that only items 1, 9, and 10 were administered to all participants. Items 2 and 3 were administered based on the response to item 1, and items 4 through 8 were subsequently administered based on the responses to items 2 and 3. Unfortunately, this created patterns of missingness across the AUDIT that affected the sample size, power, and covariance between the items. We have previously assumed that missing responses in the AUDIT represented ‘structural zeroes’ and were re-coded as zeroes (*i.e.*, zero imputation) (20). This approach relies heavily on the assumption that the range of valid responses on item B are entirely contingent on item A (*e.g.*, it is logically impossible for someone to binge drink if they never consume alcohol). While straightforward, this *is* still an imputation model, albeit a restrictive and suboptimal model.

In the present paper, we used multiple imputation by chained equations to minimize the impact of missing data on our item-level analyses. Specifically, we used the MICE package in R to employ a more advanced imputation model based on participants’ income, frequency of any alcohol consumption, frequency of alcohol consumption with meals, current alcohol intake versus previous intake 10 years ago, alcohol use disorder diagnosis status (derived from electronic health records), propensity for risk taking, and depression. We imputed item-level responses that were missing due to skip logic, but did not alter participants’ responses that were intentionally left blank. We imputed the dataset 50 times and used the mean imputed value for each item in subsequent item-level association analyses; however, the composite scores (AUDIT-C and AUDIT-P) were constructed and analyzed using zero imputation (to maintain consistency with previous GWAS). For phenotypic analyses, we used a simpler zero imputation approach (i.e., recoding missing values as zeroes), as the lavaan v0.6.5 package in R cannot readily integrate multiple imputation when using the weighted least squares estimator for categorical variables.

3 Genome-wide association analyses

We conducted genome-wide association sample-size weighted meta-analyses in three population-based cohorts using METAL. The selection of the sample size weighted approach was two-fold. First, we sought to maintain consistency between the meta-analyses in the present manuscript and those previously reported (20). Second, we believe this approach is slightly more careful than applying weights based on standard errors, which require additional assumptions to be valid. For example, if standard errors are not in the same units for any reason (*e.g.*, transformations were applied inconsistently across cohorts), weights will be inaccurate. While no such errors are present in the current analyses, we nevertheless believe that weighting by sample size is the more prudent approach. Each cohort is briefly described below.

3.1 Avon Longitudinal Study of Parents and Children

In ALSPAC, we used PLINK v2 to conduct univariate association analyses for each of the ten AUDIT items, AUDIT-C composite score, and AUDIT-P composite score. As noted in **Supplementary Section 1.1.1**, only offspring genotypes were used in the discovery analyses, as the AUDIT was not administered to mothers. Analyses were further restricted to participants of non-Hispanic European ancestries. The association model included covariate for sex, age (20 or 22), and the first ten principal components of ancestry.

Although we applied the same quality control pipeline to all GWAS results in the present study, some quality control was applied centrally by ALSPAC investigators before we obtained the data. Specifically, markers with > 2% missingness were excluded, as were those with MAF < 0.005, INFO < 0.3, and/or HWE $p < 5E-7$. Individuals missing data on > 5% of markers were excluded. We note this for transparency, but also note that we generally applied more stringent filters ourselves, as outlined in **Supplementary Section 4**.

3.2 Netherlands Twin Register

In NTR, we used the fastGWA function of the GCTA software (22) to perform univariate association analyses for each of the ten AUDIT items, AUDIT-C composite score, and AUDIT-P composite score. The mixed model employed by fastGWA included a random effect with a sparse genetic relatedness matrix in order to correct for kinship present in the sample. We included sex, year of birth, genotyping platform, and the first five principal components of ancestry as covariates in the models.

3.3 UK Biobank - Discovery

In UKB, we used BOLT-LMM v2.3.2 (23) to perform univariate association analyses for each of the ten AUDIT items, as well as the AUDIT-C and AUDIT-P composite scores. We selected BOLT-LMM as our preferred software because it employed a linear mixed model that included a genetic variance component, which (i) permitted the analysis of related individuals and (ii) improved control of population stratification. The genetic variance component was estimated using a set of 483,680 directly genotyped, autosomal SNPs that passed the genotype quality-control, had minor allele frequency greater than 0.005, Hardy-Weinberg-Equilibrium (HWE) *p*-value greater than E–16, and light LD-pruning (window size = 50 kb; variant step-size = 5; r^2 = 0.9) (24). Further details for this pipeline have been previously reported elsewhere (9).

We excluded participants (i) that did not self-report their ethnic background as “White”, “White British”, “White Irish”, or “Any other white background”; (ii) whose self-reported sex and genetic sex were incongruent; (iii) that were identified as purported sex chromosome aneuploidy cases; (iv) that did not pass the sample-level quality control thresholds (9); or (v) that were missing data for the required outcome(s) and covariates. Never drinkers and non-current drinkers were included in the analyses.

The BOLT-LMM association model included covariates for sex, birth year, sex-by-birth year effects, genotyping batch (that also effectively indexed genotyping array), as well as the first 40 principal components of ancestry, which we estimated ourselves (see (25) for details). Analyzed variants were selected from the third release of the UKB imputed genotype data.

4 Quality control of association results

Similar to our previous work (25), we used EasyQC (26) to apply a series of thorough quality control filters to our association results. For each cohort, we applied the following thresholds to each set of GWAS summary statistics. Filters were applied in the order that they are listed below.

1. We excluded SNPs if either allele corresponded to a value other than "A", "C", "G", or "T".
2. We excluded SNPs if any of the following statistics were missing: beta, standard error, *p*-value, effect allele frequency, sample size, and imputation quality score (for imputed SNPs).
3. We excluded SNPs if they had nonsensical or impossible values (*e.g.*, negative standard errors, allele frequencies greater than 1 or below 0).
4. We filtered SNPs with minor allele frequencies less than .005.
5. We removed SNPs with an imputation quality score < .90.
6. In the case of duplicated base pair positions (based on GRCh37), we retained the SNP with the largest sample size.

After applying the filters described above, we inspected diagnostic plots to further ensure that results were not prone to systematic error or bias.

7. We inspected a plot of the allele frequencies in our GWAS sample against those in a non-Hispanic European reference sample in order to check for discrepancies in allele frequencies and errors in strand orientation.
8. To identify discrepancies in reported *p*-values, we examined their relationship the reported coefficient estimates and their SEs.
9. We examined a quantile-quantile plot to evaluate whether population stratification had been sufficiently accounted for in the GWAS.

5 Genomic structural equation modeling

Genomic Structural Equation Modelling (Genomic SEM) (27) is a novel statistical method for applying SEM to GWAS summary statistics in order to model the joint genetic architecture of complex traits. Genomic SEM is a flexible framework that allows for empirical modeling of multivariate genetic covariance matrices, such as those estimated via linkage disequilibrium (LD) score regression (28). Here, we used Genomic SEM to conduct a series of analyses in order to study the multivariate genetic architecture of AUDIT items. We highlight three specific aims: (i) identify the latent genetic factor(s) structure that best represent the genetic covariance of AUDIT items, (ii) evaluate genetic relationships between the latent genetic factors and exogenous phenotypes, and (iii) perform multivariate GWAS of the latent genetic factors (pending good evidence of an underlying factor structure).

We have extensively described the general background of Genomic SEM in previous publications (25, 27). We suggest that readers interested in the methodology of Genomic SEM review the original study (27), recent applications to genetic discovery (25, 29), and the online tutorial (https://github.com/MichelNivard/GenomicSEM/wiki). The interested reader can find example code and additional resources in the extensive online tutorial, which can be used to reproduce the analyses carried out in the present manuscript.

5.1 Confirmatory factor analysis

Genomic SEM can be used to conduct confirmatory factor analysis (CFA), where theoretical models are used to parsimoniously explain the genetic covariances among a set of phenotypes. Here, we use CFA to test a series of competing genetic factor models in order to identify the model that best fit the data. As in traditional CFA, good fit means that the specified latent genetic factor structure adequately explains the observed genetic covariances among the phenotypes.

The AUDIT has been the subject of many factor analytic studies at the phenotypic level (**Supplementary Table 1)**. As the phenotypic null hypothesis suggests that the genetic factor structure will largely be similar to the phenotypic factor structure, we based our models on the existing literature and our own phenotypic factor analysis in UKB (the largest sample), as described in the main text. We used unit variance identification was to set the scale of the latent factors.

We assessed model fit using conventional indices in SEM: the model χ^2 statistic, the Akaike information criterion (AIC), the comparative fit index (CFI), and the standardized root mean square residual (SRMR). As previously described (27), all of these fit indices retain their standard interpretation within Genomic SEM except for the model χ^2 statistic, which is used as a comparative measure of fit to evaluate competing models (akin to AIC) rather than a measure of statistical significance. CFI values greater than .90 and SRMR values less than .08 were considered reflective of good model fit (30). While models with good fit also traditionally have non-significant model χ^2 statistics, the test is extremely sensitive in large samples like the ones used in this study. As such, the model χ^2 statistic is interpreted as a comparative measure of fit to evaluate competing models rather than a measure of statistical significance.

5.1.1 Factor extension

Factor extension is a method for estimating factor loadings for variables that were not included in the original analysis. In the present study, item 6 was excluded from the final analysis on the basis of its non-significant SNP heritability. Nonetheless, we were still interested in its relationship to the *Problems* factor to which it should theoretically related.

To estimate the putative loading for item 6, we employed a straightforward two-step procedure. In the first step, we freely estimated the correlated factors model again excluding item 6. In the second step, we included item 6 and re-estimated the model; however, we fixed all of the factor loadings and residual variances for items 1-5 and 7-9 to be the same across models. Only the parameters for item 6 were freely estimated in this second step. This yielded an unattenuated estimate of the relationship between item 6 and *Problems* without affecting the overall factor structure.

5.1.2 Genetic correlation

As described in our previous work (25), Genomic SEM can also be used to estimate the genetic correlation between a latent factor and an exogenous phenotype that is not included in the factor model. We note that this method is preferable over bivariate LD score regression for reasons that have been previously explained (25). In the present study, we used this method genetic correlations between 100 exogenous phenotypes and the latent genetic factors, *Consumption* and *Problems*, as well as *Frequency Residual* (*i.e.*, the residual genetic variance in item 1).

5.1.3 Multivariate genome-wide association analyses

We performed multivariate GWASs analyses of the audit latent genetic factors from the best-fitting model. First, we used the *sumstats* function to standardize the univariate GWAS summary statistics for each of the AUDIT items. Next, we used the *userGWAS* function to run the model in an iterative manner for each SNP, regressing the latent genetic factors on the SNP. As the latent genetic factors were the dependent variables in these models, unit loading identification was specified for scaling purposes (item 3 for *Consumption*, item 4 for Problems). We applied the field standard genome-wide significance threshold of $p < 5\text{e}{-8}$.

6 Biological annotation

To link our genome-wide association results to putative risk genes and associated biology, we performed a series of biological annotation analyses described below.

6.1 Functional mapping and annotation

We used FUMA v1.3.5e (31) to study the functional consequences of the lead SNPs for Consumption and Problems, which included ANNOVAR categories (*i.e.*, the functional consequence of SNPs on genes), Combined Annotation Dependent Depletion (CADD) scores (*i.e.*, scores > 12.37 are the suggested threshold to classify a SNP as deleterious), RegulomeDB scores (*i.e.*, biological evidence that the SNP is a regulatory element, scores with 1a representing the strongest evidence).

We used FUMA (31) to identify whether the loci associated with *Consumption* and *Problems* (and SNPs in LD, r2 > 0.1) were previously associated (p < 1E-5) with other traits from the GWAS Catalog (version e96 2019-05-03). This information is shown in **Supplementary Tables 13-14**).

6.1.1 Supplemental results for AUDIT-C and AUDIT-P

We identified 8 independent loci for AUDIT-C, showing convergence with *Consumption* with a few exceptions: *GCKR* [previously associated with forms of alcohol behavior [*e.g.* (20, 32–35)], physical activity, sleep patterns and metabolic traits [*e.g.* (36–39)], *ORC5* [implicated in acute myeloid leukemia (40), brain volume (41), alcohol consumption (33)] and *GPR139* [previously associated with smoking behavior (42), anthropometric traits (42, 43), age at menarche (38), insomnia (44)] were only associated with AUDIT-C, whereas others, such as *CPS1* [associated with metabolic and anthropometric phenotypes (44–46) and chronic kidney disease (47)], *FUT2* [pleiotropic gene associated with *e.g.,* alcohol behaviors (20, 33), blood protein levels (48), diarrhea (49), serum cancer antigen levels (50), tumor biomarkers (51), vitamin B12 levels (52)] and *BTF3P13* [associated with alcohol consumption (51, 52) and behavioral disinhibition (53)] were only associated with *Consumption*. The regions including the genes *ADH1B*, *KLB*, *MAPT*, *SLC39A8* and *RP11-89K21.1* were common to both *Consumption* and AUDIT-C.

Compared to the three independent loci that we identified for *Problems*, we identified 6 independent loci for AUDIT-P, including a region containing the well-known dopamine receptor gene, *DRD2*. Several of the loci significant in AUDIT-P [*KLB*, replicating previous GWAS of alcohol phenotypes (20, 33–35)], and two novel candidate genes for alcohol, *RP11-478B11.2*, *AC064875.2*) were not associated with *Problems*, with the exception of the main signal on chromosome 4, including the genes *ADH1B* and *SLC39A8*.

6.2 Gene-based, gene-set, and gene-property analyses

We performed competitive gene-based association analyses using the multivariate GWAS for *Consumption* and *Problems* by applying the “Multi-marker Analysis of GenoMic Annotation” (MAGMA v1.08) (31) in FUMA. SNPs were mapped to 18,857 protein-coding genes from Ensembl build 85 based on physical position. This method accounts for LD patterns, using the 1000 Genomes European sample as a reference. We used a Bonferroni-corrected significance threshold (p < 2.65E-6, adjusted for testing 18,857 genes).

We also performed a MAGMA (54) gene-set analysis in FUMA. We used 18,116 curated gene sets and Gene Ontology (GO) terms obtained from the Molecular Signatures Database (MSigDB v7.0, https://www.gsea-msigdb.org/gsea/msigdb/index.jsp) (55), to study the relationships between *Consumption* and *Problems* and the biological processes, molecular function and cellular component of the individual gene products. We used a Bonferroni-corrected significance threshold (p < 3.23E-6), adjusted for testing 15,485 gene sets).

6.3 Gene-based analysis of cortical chromatin interaction data

We used H-MAGMA (56), an extension of MAGMA v1.08 that uses Hi-C data to assign non-coding (intergenic and intronic) SNPs, to identify trait-associated genes based on their chromatin interactions. Exonic and promoter SNPs were assigned to genes based on physical position. We used four Hi-C datasets derived from adult brain (57), fetal brain (58), and iPSC derived neurons and astrocytes (59) (all available for download: https://github.com/thewonlab/H-MAGMA). The results are reported in **Supplementary Tables 21-28**. We used a Bonferroni-corrected significance threshold (p < 9.40E-7, p < 9.41E-7, p < 9.41E-7, p < 9.41E-7, for each analysis, respectively).

6.4 Gene-based analysis of cortical transcriptomic data

We used S-PrediXcan v0.6.2 (41) to analyze predicted gene expression levels in multiple brain tissues, and to test whether the gene expression correlated with the genetic liability of Consumption and Problems. As input data, we used the summary statistics for Consumption and Problems, pre-computed tissue weights from the Genotype-Tissue Expression (GTEx, v8) project database (https://www.gtexportal.org/) as the reference transcriptome dataset (60), and covariance matrices of the SNPs within each gene model (based on HapMap SNP set; available to download at the PredictDB Data Repository, http://predictdb.org) from 13 brain tissues: anterior cingulate cortex, amygdala, caudate basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex, hippocampus, hypothalamus, nucleus accumbens basal ganglia, putamen basal ganglia, spinal cord and substantia nigra. It should be noted that, although GTEx is one of the most comprehensive genetic expression databases available to date, brain tissue sample sizes are still modest and thus the statistical power for eQTL discovery may be limited. We used a Bonferroni transcriptome-wide significance threshold of $p < 3.52\text{E-}6$ (184,691 gene-tissue pairs).

7 Polygenic risk score analyses

To validate our genome-wide association analyses, we performed polygenic risk score analyses in three independent cohorts. We calculated either R^2 or pseudo R^2 to quantify the proportion of variance explained by each PRS for each outcome. Specifically, we calculated R^2 for the continuous outcomes in UK Biobank, but pseudo R^2 for all other outcomes. For the dichotomous outcomes in UK Biobank and COGA, we calculated pseudo R^2 using the r.squaredLR() function of the MuMIn package in R. However, for the continuous outcomes in COGA, we used the r.squaredGLMM() function of the MuMIn package to calculate pseudo R^2, as that is standard protocol for PRS analyses conducted in the cohort. Please note that the R^2 for each PRS was calculated using a covariates-only base model, which means that the R^2 values are not independent of each other (due to shared variance between the PRSs). Finally, please note that the COGA models are not directly comparable to the UK Biobank models, as mixed-effect models were used in COGA to account for family structure. The phenotyping procedures in each cohort are briefly described below.

7.1 UK Biobank - Validation

The phenotypes included in the UKB analyses were: (1) drinking quantity (*i.e.*, typical grams of ethanol consumed per month based on reports of monthly or weekly quantities of various types of alcoholic beverages), (2) drinking frequency (*i.e.*, pseudo-continuous number of typical days drinking per month), and (3) lifetime alcohol use disorder (AUD) diagnosis (*i.e.*, ICD10 codes of alcohol use disorders (F10.1, F10.2, F10.3) or alcoholic liver disease (K70) in hospitalization records, death records, or in-person medical interviews). Because of the small number of AUD cases (n = 4,463), a random subset of 20,000 control participants was used in the AUD prediction models to improve the imbalanced ratio. We excluded current (n = 10,093) and lifetime non-drinkers (n = 9,323) from polygenic analyses of current consumption measures (quantity/frequency) and lifetime non-drinkers from analyses of AUD. Townsend Deprivation

Index scores account for a large amount of variance in alcohol phenotypes; for that reason, we included this variable as an additional covariate.

7.2 Collaborative Studies on Genetics of Alcoholism Study

In COGA, the phenotypes included in the analyses were: (1) number of drinks per week (*i.e.*, the sum of 4 separate measures: average number of beers/wine/liquor/other alcohol consumed per week); (2) maximum drinks in 24 hours; (3) DSM-5 AUD. The measure of maximum drinks in 24 hours was winsorized [anyone with a value ≥ 66 (mean + 3SD's) was set to 65]. The number of drinks per week and maximum drinks in 24 hours were log-transformed before the analysis to account for skew. We included 6,390 individuals who reported being regular drinkers (having at least one drink per month for 6 consecutive months).

7.3 Vanderbilt University Medical Center Biobank

In BioVU, 1,335 phecodes with at least 100 cases were analyzed. We required the presence of at least two International Classification of Disease codes that mapped to a PheWAS disease category (Phecode Map 1.2 (https://phewascatalog.org/phecodes) to assign case status (61).

REFERENCES

1. Boyd A, Golding J, Macleod J, et al.: Cohort Profile: The 'Children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol 2013; 42:111–127

2. Fraser A, Macdonald-Wallis C, Tilling K, et al.: Cohort Profile: The Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol 2013; 42:97–110

3. Northstone K, Lewcock M, Groom A, et al.: The Avon Longitudinal Study of Parents and Children (ALSPAC): an update on the enrolled sample of index children in 2019. Wellcome Open Res 2019; 4:51

4. Harris PA, Taylor R, Minor BL, et al.: The REDCap consortium: Building an international community of software platform partners. J Biomed Inform 2019; 95:103208

5. Harris PA, Taylor R, Thielke R, et al.: Research Electronic Data Capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform 2009; 42:377–381

6. Boomsma DI, Vink JM, Beijsterveldt TCEM van, et al.: Netherlands Twin Register: A Focus on Longitudinal Research. Twin Res Hum Genet 2002; 5:401–406

7. Mbarek H, Milaneschi Y, Fedko IO, et al.: The Genetics of Alcohol Dependence: Twin and SNP-Based Heritability, and Genome-Wide Association Study Based on AUDIT Scores. Am J Med Genet B Neuropsychiatr Genet 2015; 168:739–748

8. Ligthart L, et al.: The Netherlands Twin Register: Longitudinal Research Based on Twin and Twin-Family Designs. Twin Research and Human Genetics 2019; 22:623-636

9. Bycroft C, Freeman C, Petkova D, et al.: The UK Biobank resource with deep phenotyping and genomic data. Nature 2018; 562:203–209

10. Sudlow C, Gallacher J, Allen N, et al.: UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med 2015; 12:e1001779

11. Begleiter: The Collaborative Study on the Genetics of Alcoholism. Alcohol Health Res World 1995; 19:228–236

12. Nurnberger JI, Wiegand R, Bucholz K, et al.: A family study of alcohol dependence: coaggregation of multiple disorders in relatives of alcohol-dependent probands. Arch Gen Psychiatry 2004; 61:1246–1256

13. Schuckit MA, Smith TL, Danko G, et al.: A 22-Year Follow-Up (Range 16 to 23) of Original Subjects with Baseline Alcohol Use Disorders from the Collaborative Study on Genetics of Alcoholism. Alcohol Clin Exp Res 2018; 42:1704–1714

14. Auton A, Abecasis GR, Altshuler DM, et al.: A global reference for human genetic variation. Nature 2015; 526:68–74

15. Johnson EC, Sanchez-Roige S, Acion L, et al.: Polygenic contributions to alcohol use and alcohol use disorders across population-based and clinically ascertained samples. Psychol Med 2020; 1–10

16. Dennis J, Sealock J, Levinson RT, et al.: Genetic risk for major depressive disorder and loneliness in sex-specific associations with coronary artery disease. Mol Psychiatry 2019;

17. Roden DM, Pulley JM, Basford MA, et al.: Development of a large-scale de-identified DNA biobank to enable personalized medicine. Clin Pharmacol Ther 2008; 84:362–369

18. Delaneau O, Coulonges C, Zagury J-F: Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics 2008; 9:540

19. Purcell S, Neale B, Todd-Brown K, et al.: PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet 2007; 81:559–575

20. Sanchez-Roige S, Palmer AA, Fontanillas P, 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; 176:107–118

21. Fry A, Littlejohns TJ, Sudlow C, et al.: Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. Am J Epidemiol 2017; 186:1026–1034

22. Jiang L, Zheng Z, Qi T, et al.: A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet 2019; 51:1749–1755

23. Loh P-R, Tucker G, Bulik-Sullivan BK, et al.: Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 2015; 47:284–290

24. Chang CC, Chow CC, Tellier LC, et al.: Second-generation PLINK: rising to the challenge of larger and richer datasets [Internet]. GigaScience 2015; 4[cited 2020 Jul 23] Available from: https://academic.oup.com/gigascience/article/4/1/s13742-015-0047-8/2707533

25. Mallard TT, Karlsson Linnér R, Grotzinger AD, et al.: Multivariate GWAS of psychiatric disorders and their cardinal symptoms reveal two dimensions of cross-cutting genetic liabilities. bioRxiv 2020; 603134

26. Rangamaran VR, Uppili B, Gopal D, et al.: EasyQC: Tool with Interactive User Interface for Efficient Next-Generation Sequencing Data Quality Control. J Comput Biol 2018; 25:1301–1311

27. Grotzinger AD, Rhemtulla M, de Vlaming R, et al.: Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nat Hum Behav 2019; 3:513–525

28. Bulik-Sullivan BK, Loh P-R, Finucane HK, et al.: LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 2015; 47:291–295

29. Linnér RK, Mallard TT, Barr PB, et al.: Multivariate genomic analysis of 1.5 million people identifies genes related to addiction, antisocial behavior, and health. bioRxiv, 2020 doi/10.1101/2020.10.16.342501

30. Hu L, Bentler PM: Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Model Multidiscip J 1999; 6:1–55

31. Watanabe K, Taskesen E, van Bochoven A, et al.: Functional mapping and annotation of genetic associations with FUMA. Nat Commun 2017; 8:1826

32. Evangelou E, Gao H, Chu C, et al.: New alcohol-related genes suggest shared genetic mechanisms with neuropsychiatric disorders. Nat Hum Behav 2019; 3:950–961

33. Liu M, Jiang Y, Wedow R, et al.: Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet 2019; 51:237–244

34. Clarke T-K, Adams MJ, Davies G, et al.: Genome-wide association study of alcohol consumption and genetic overlap with other health-related traits in UK Biobank (N=112 117). Mol Psychiatry 2017; 22:1376–1384

35. Schumann G, Liu C, O'Reilly P, et al.: KLB is associated with alcohol drinking, and its gene product β-Klotho is necessary for FGF21 regulation of alcohol preference. Proc Natl Acad Sci U S A 2016; 113:14372–14377

36. Klarin D, Damrauer SM, Cho K, et al.: Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nat Genet 2018; 50:1514–1523

37. Noordam R, Bos MM, Wang H, et al.: Multi-ancestry sleep-by-SNP interaction analysis in 126,926 individuals reveals lipid loci stratified by sleep duration. Nat Commun 2019; 10:5121

38. Kichaev G, Bhatia G, Loh P-R, et al.: Leveraging Polygenic Functional Enrichment to Improve GWAS Power. Am J Hum Genet 2019; 104:65–75

39. Hoffmann TJ, Theusch E, Haldar T, et al.: A large electronic-health-record-based genome-wide study of serum lipids. Nat Genet 2018; 50:401–413

40. Lv H, Zhang M, Shang Z, et al.: Genome-wide haplotype association study identify the FGFR2 gene as a risk gene for acute myeloid leukemia. Oncotarget 2017; 8:7891–7899

41. GTEx Consortium, Barbeira AN, Dickinson SP, et al.: Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun 2018; 9:1825

42. Justice AE, Winkler TW, Feitosa MF, et al.: Genome-wide meta-analysis of 241,258 adults accounting for smoking behaviour identifies novel loci for obesity traits. Nat Commun 2017; 8:14977

43. Zhao B, Luo T, Li T, et al.: Genome-wide association analysis of 19,629 individuals identifies variants influencing regional brain volumes and refines their genetic co-architecture with cognitive and mental health traits. Nat Genet 2019; 51:1637–1644

44. Jansen PR, Watanabe K, Stringer S, et al.: Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. Nat Genet 2019; 51:394–403

45. Pulit SL, Stoneman C, Morris AP, et al.: Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Hum Mol Genet 2019; 28:166–174

46. Kettunen J, Demirkan A, Würtz P, et al.: Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. Nat Commun 2016; 7:11122

47. Schlosser P, Li Y, Sekula P, et al.: Genetic studies of urinary metabolites illuminate mechanisms of detoxification and excretion in humans. Nat Genet 2020; 52:167–176

48. Sun BB, Maranville JC, Peters JE, et al.: Genomic atlas of the human plasma proteome. Nature 2018; 558:73–79

49. Bustamante M, Standl M, Bassat Q, et al.: A genome-wide association meta-analysis of diarrhoeal disease in young children identifies FUT2 locus and provides plausible biological pathways. Hum Mol Genet 2016; 25:4127–4142

50. Olafsson S, Alexandersson KF, Gizurarson JGK, et al.: Common and Rare Sequence Variants Influencing Tumor Biomarkers in Blood. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol 2020; 29:225–235

51. He M, Wu C, Xu J, et al.: A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19-9, carcinoembryonic antigen and α fetoprotein and their associations with cancer risk. Gut 2014; 63:143–151

52. Lin X, Lu D, Gao Y, et al.: Genome-wide association study identifies novel loci associated with serum level of vitamin B12 in Chinese men. Hum Mol Genet 2012; 21:2610–2617

53. Sanchez-Roige S, Fontanillas P, Elson SL, et al.: Genome-Wide Association Studies of Impulsive Personality Traits (BIS-11 and UPPS-P) and Drug Experimentation in up to 22,861 Adult Research Participants Identify Loci in the CACNA1I and CADM2 genes. J Neurosci Off J Soc Neurosci 2019; 39:2562–2572

54. de Leeuw CA, Mooij JM, Heskes T, et al.: MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 2015; 11:e1004219

55. Liberzon A, Subramanian A, Pinchback R, et al.: Molecular signatures database (MSigDB) 3.0. Bioinforma Oxf Engl 2011; 27:1739–1740

56. Sey NYA, Hu B, Mah W, et al.: A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. Nat Neurosci 2020; 23:583–593

57. Wang D, Liu S, Warrell J, et al.: Comprehensive functional genomic resource and integrative model for the human brain [Internet]. Science 2018; 362[cited 2020 Jul 3] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6413328/

58. Won H, de la Torre-Ubieta L, Stein JL, et al.: Chromosome conformation elucidates regulatory relationships in developing human brain. Nature 2016; 538:523–527

59. Rajarajan P, Borrman T, Liao W, et al.: Neuron-specific signatures in the chromosomal connectome associated with schizophrenia risk. Science 2018; 362:eaat4311

60. GTEx Consortium: The Genotype-Tissue Expression (GTEx) project. Nat Genet 2013; 45:580–585

61. Wei W-Q, Teixeira PL, Mo H, et al.: Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016; 23:e20–e27