nature research

Corresponding author(s): Joel Gelernter, Howard Edenberg, and Arpana Agrawal

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Data collecting (phenotyping and genotyping) was performed previously by each cohort. Details on data collection in each cohort have been described previously, see the Supplementary Information for a full list of references. Broadly, phenotyping was not specifically dependent on specialized software (though may have, for example, been stored in a software system such as REDCap, https://projectredcap.org/), and genotyping was performed using standard genotype calling pipelines outside of the scope of the current study.
Data analysis	Quality control, imputation, and GWAS of case/control cohorts was performed using ricopili (https://github.com/Nealelab/ricopili), which includes wrappers around PLINK, Eigenstrat, LifeOver, SHAPEIT, IMPUTE2, and METAL. Quality control, imputation, and GWAS of family-based cohorts was performed with picopili (https://github.com/Nealelab/picopili), which additionally includes wrappers of PRIMUS, ADMIXTURE, REAP, GMMAT, and geepack. Trans-ancestral meta-analyses were performed with METASOFT and MANTRA. PRSice was used for polygenic risk score analyses. LD score regression analyses were performed with ldsc (https://github.com/bulik/ldsc) and the LD Hub web tool (ldsc.broadinstitute.org). Gene-based analyses were performed with the FUMA web tool (http://fuma.ctglab.nl/), which uses MAGMA. Analyses related to local ancestry calling included scripts from Alicia Martin (https://github.com/armartin/ancestry_pipeline) using HAPI-UR and RFMix. Power analysis was performed using CaTS. Plots were generated using R, LocusZoom (http://locuszoom.org/), LDlink (https://analysistools.nci.nih.gov/LDlink/), and HUGIn (https://yunliweb.its.unc.edu/hugin/).

All of the above are publicly available, with the exception of MANTRA which is available from the method's developer (Andrew Morris). Relevant links and citations are all provided in the manuscript.

Remaining calculations, most notably the effective sample size calculations used for meta-analyses across study designs/association models, were performed using ad hoc scripts. Example code is available from the first author by request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Summary statistics from the genome-wide meta-analyses are available on the Psychiatric Genomics Consortium's downloads page (http://www.med.unc.edu/pgc/results-and-downloads), including the source data for Figures 1 and 2. Individual-level data from the genotyped cohorts and cohort-level summary statistics will be made available to researchers following an approved analysis proposal through the PGC Substance Use Disorder group with agreement of the cohort PIs; contact the corresponding authors for details. Cohort data are also available from dbGaP except where prohibited by IRB or European Union data restrictions. Expression data used to evaluate variants in ADH1B is available from GTEx (https://gtexportal.org/home/). Hi-C data used to evaluate the chromosome 3 variant can be queried with HUGIn (https://yunliweb.its.unc.edu/hugin/). Publicly available genome-wide summary statistics used for testing genetic correlations are accessible through LD Hub (http://ldsc.broadinstitute.org/), or from the Psychiatric Genomics Consortium (http://www.med.unc.edu/pgc/results-and-downloads), the Social Science Genetic Association Consortium (SSGAC; https://www.thessgac.org/data), Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA; http://enigma.ini.usc.edu/research/download-enigma-gwas-results/), and the Neale Lab (http://www.nealelab.is/uk-biobank); for availability of summary statistics from other studies contact the respective authors. The source data for Figure 3 is included in Supplementary Table S6.

Accession Codes
Comorbidity and Trauma Study (CATS): dbGAP accession phs000277.v1.p1
Center for Education and Drug Abuse Research (CEDAR): dbGAP accession phs001649.v1.p1
Christchurch Health and Development Study (CHDS): dbGAP submission in process
The Collaborative Study on the Genetics of Alcoholism (COGA): dbGaP accession numbers phs000125.v1.p1, phs000763.v1.p1, and phs000976.v1.p1
Study of Addiction: Genetics and Environment (SAGE): dbGAP accession phs000092.v1.p1
Collaborative Genetic Study of Nicotine Dependence (COGEND): dbGAP accession phs000404.v1.p1
Gene-Environment-Development Initiative (GEDI) – Duke University (GSMS): dbGAP accession phs000852.v1.p1
Center on Antisocial Drug Dependence (CADD): dbGAP submission in process
Spit for Science: dbGAP submission in process
NIAAA: available via https://btris.nih.gov/
Gene-Environment-Development Initiative (GEDI) –Virginia Commonwealth University (VTSABD): dbGAP submission in process
Minnesota Center for Twin and Family Research (MCTFR): dbGAP accession phs000620.v1.p1
Yale-Penn: dbGAP accession phs000425.v1.p1 and phs000952.v1.p1

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Sample size was not predetermined, but instead reflects our best effort to aggregate all possible studies with genome-wide genotype data and robust phenotyping of alcohol dependence according to the DSM-IV criteria used in the current study. This open, international collaboration supported by the Psychiatric Genomics Consortium includes contributions from 28 studies (plus 3 replication cohorts and 2), and to our knowledge represents the largest genome-wide study of alcohol dependence to date.

Based on the available data, we have made efforts to maximize the use of the genotyped samples. This includes developing the infrastructure and appropriate statistical modeling to include both family-based and case/control cohorts in the same genome-wide analysis, and including trans-ancestral analysis of both European and African ancestry individuals.

We have also performed power analysis for the current genome-wide study, as detailed in the manuscript. For instance, we estimate that the full discovery meta-analysis has >80% power to detect variants associated with alcohol dependence with true odds ratios >= 1.15 and minor allele frequency > 0.15. We have also provided comparisons of power with other GWAS (e.g., schizophrenia, depression, obesity). This power and sample size are consistent with successful GWAS of many other complex traits.

Data exclusions	Data exclusions were performed based on (a) failure of pre-determined data quality control criteria and (b) planned phenotype exclusions to avoid confounds in defining alcohol dependence cases and controls. For quality control, individuals were excluded if they were observed to have low genotyping quality (i.e. high missingness rates), excess heterozygosity (an indicator of possible sample contamination or other technical artifacts), or if they deviated from reported family pedigree structures (i.e. excessive mendelian error rates, discordance between genetically-inferred and reported sex, cryptic genetic relatedness to unrelated individuals, or lack of expected genetic relatedness to members of the same pedigree). Observed outliers of genetic ancestry, as determined by principal components analysis, were also excluded in order to avoid the known risk of population stratification in genome-wide studies including such individuals. Ancestries other than African or European were excluded due to insufficient sample size for a meaningful analysis in the currently available data. For phenotype-based exclusions, we omit individuals lacking phenotype information for alcohol dependence, individuals who report never being exposed to alcohol, and individuals meeting criteria for alcohol abuse (i.e. qualifying neither as alcohol dependence cases or healthy controls). Cohorts with other exclusion criteria as part of their original study recruitment are detailed in the Supplementary Information. The metrics used as exclusion criteria were established prior to the analyses, but some thresholds used for exclusion (e.g. threshold from principal components analysis to define ancestry strata) were evaluated during the QC process. All of the above exclusions were made in accordance with the planned study protocol, and are detailed in the manuscript.
Replication	The primary genome-wide significant locus identified in the current study (i.e. the ADH1B locus) is itself a replication of previous studies of alcohol dependence (see manuscript for references). For the novel genome-wide significant locus on chromosome 3, we present more targeted replication analysis from 3 additional cohorts of African and Finnish ancestries relevant to the putative signal. As described in the manuscript, replication was not found. We rely on this lack of replication to conclude that there is not sufficient evidence for an effect of the chr. 3 locus, with the result observed in the discovery sample potentially reflecting confounding from ancestry or an increased multiple testing burden. We also evaluate the consistency of effects in this locus between European and African ancestry cohorts and across study designs as a form of internal replication. These tests find very little evidence of any heterogeneity, indicating that the reported results have generally consistent evidence across ancestry and study design. Polygenic risk score analyses also provide generalizability of the overall results in both European and African ancestry cohorts. In all instances, polygenic risk scores derived from effect sizes in this study successfully predicted alcohol-related phenotypes in other studies as expected. The only instance of poor prediction was that effect sizes from the EA discovery GWAS in this study only weakly predicted alcohol dependence in an independent AA sample (COGA AAfGWAS), which is consistent with prior observations about cross-population polygenic prediction. The strong sample size requirements of the secondary analyses (most notably LD score regression analyses of heritability and genetic correlation to other traits) and dependence on LD reference panels limits options for direct replication of those findings. We instead focus on comparisons to existing GWAS of other alcohol-related phenotypes to get potential insight into how genome-wide results appear to generalize between these phenotypes in different study populations. The compelling findings from those comparisons are a key result for the current analysis and are discussed at length in the manuscript.
Randomization	Randomization of experimental groups was not applicable to this study. The experimental conditions are determined by each individual's genetics, which are fixed at conception. Conceptually this reflects a randomization of the alleles inherited from each individual's parents (i.e. mendelian randomization), but it does not involve randomization of experimental conditions by the researchers in a classical sense. Our study assess the observed association between that natural randomization of genotype and the ascertained phenotype of alcohol dependence.
Blinding	Blinding is not relevant to the current study. Samples were not allocated to different conditions by the researches, and the phenotype ascertainment process is fully separate from the genotyping process.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants]

Population characteristics	The current study encompasses 14,904 cases and 37,944 controls from 28 cohorts in the primary analysis (after quality control), with an additional 2,997 cases and 25,318 controls from 3 replication cohorts and 9,629 individuals in 2 other cohorts used for

polygenic risk score analysis. Details on each cohort are provided in the manuscript, with summary descriptives in Table 1 and full descriptions in the Supplementary Information.

Briefly, included participants represent a mix of ascertainment schemes across cohorts, including both population-based collections and ascertained research cohorts. These include studies of genetically unrelated cases and controls, as well as family-based studies ranging from sibling pairs to extended pedigrees ascertained for enrichment of substance abuse. Overall, the participants include roughly equal numbers of males and females, with ages fully distributed across the lifespan for adults. Participants are from North America, Europe, and Australia and are of European or African ancestry (confirmed in genetic data), with African ancestry individuals predominantly reflecting African-American admixture.

Genome-wide genotype data has been collected for all participants. Most individuals in the primary analysis were analyzed using the individual level genotype data, but a subset (N=9,929 from 5 cohorts) are only represented in summary statistics from their respective cohorts. The 3 replication cohorts are also only analyzed through contributed summary statistics.

Phenotyping criteria vary by cohort (full descriptions in the manuscript supplement), but for most cohorts a standardized measure such as the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA) or the Composite International Diagnostic Interview (CIDI) has been administered to ascertain lifetime alcohol dependence status in accordance with DSM-IV diagnostic guidelines. Current treatment data has not been collected for all participants, but is not critical to the current research question of genetic associations with lifetime dependence diagnosis.

Recruitment

Participants were recruited separately for each cohort according to their respective study design. Descriptions of the design for each cohort can be found in the Supplementary Information, along with references to previous publications containing complete details.

Overall, the cohorts represent a mix of population-based cohorts without targeted ascertainment (e.g. birth cohorts from a specified region), cohorts recruited for studies of alcohol dependence (e.g. families of probands from inpatient or outpatient treatment facilities), or cohorts originally recruited for studies of other substance dependence (e.g. cocaine or nicotine) or other phenotypes where measures of alcohol dependence were included in phenotyping (e.g. schizophrenia, high-risk populations involved in the criminal justice system, or pharmacogenetics studies).

These recruitment strategies could yield biases in the results for a given cohort, but the mix of recruitment strategies used across the cohorts is unlikely to produces consistent biases across the current analysis. Instead, any different biases resulting from the variety of recruitment strategies and study designs would be more likely to manifest as heterogeneity or noise in results across the cohorts, potentially reducing power.