Alcohol and Alcoholism, 2021, 1–9
https://doi.org/10.1093/alcalc/agab031
Article

Article

Investigation of Association of Rare, Functional Genetic Variants With Heavy Drinking and Problem Drinking in Exome Sequenced UK Biobank Participants

David Curtis[1,2,*]

[1]UCL Genetics Institute, UCL, Darwin Building, Gower Street, London WC1E 6BT, UK, and [2]Centre for Psychiatry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK

*Corresponding author: Tel.: 00 44 7973 906 143; E-mail: d.curtis@ucl.ac.uk

Received 4 February 2021; Revised 22 March 2021; Editorial Decision 23 March 2021; Accepted 23 March 2021

Abstract

Aims: The study aimed to identify specific genes and functional genetic variants affecting susceptibility to two alcohol-related phenotypes: heavy drinking and problem drinking.
Methods: Phenotypic and exome sequence data were downloaded from the UK Biobank. Reported drinks in the last 24 hours were used to define heavy drinking, while responses to a mental health questionnaire defined problem drinking. Gene-wise weighted burden analysis was applied, with genetic variants which were rarer and/or had a more severe functional effect being weighted more highly. Additionally, previously reported variants of interest were analysed inidividually.
Results: Of exome sequenced subjects, for heavy drinking, there were 8166 cases and 84,461 controls, while for problem drinking, there were 7811 cases and 59,606 controls. No gene was formally significant after correction for multiple testing, but three genes possibly related to autism were significant at $P < 0.001$, *FOXP1*, *ARHGAP33* and *CDH9*, along with *VGF* which may also be of psychiatric interest. Well established associations with rs1229984 in *ADH1B* and rs671 in *ALDH2* were confirmed, but previously reported variants in *ALDH1B1* and *GRM3* were not associated with either phenotype.
Conclusions: This large study fails to conclusively implicate any novel genes or variants. It is possible that more definitive results will be obtained when sequence data for the remaining UK Biobank participants become available and/or if data can be obtained for a more extreme phenotype such as alcohol dependence disorder. This research has been conducted using the UK Biobank Resource.

INTRODUCTION

There is a heritable contribution to the related phenotypes of heavy drinking, problematic drinking and alcohol dependency disorder, and this has been explored in a number of ways. As reviewed recently, genome wide association studies (GWAS) using common variants produce signals at a number of genes including some involved in alcohol metabolism, *ADH1B*, *ADH1C* and *ADH4*, along with others with less obvious roles, *DRD2*, *KLB*, *CADM2*, *CRHR1*, *FGF14*, *GCKR*, *SIX3*, *SLC39A8* and *FTO* (Johnson *et al.*, 2020). A subsequent GWAS of heavy alcohol consumption replicated some of these and additionally implicated the locus for a pseudogene, *BTF3P13* (Thompson *et al.*, 2020).

GWAS results typically implicate genomic regions where common variation is associated with moderate differences in susceptibility to a trait but it is often difficult to elucidate which individual genes and variants are responsible and the biological mechanisms involved. In this respect, alcohol intake is somewhat exceptional among complex traits in that there are individual functional variants which have been

well characterized. These are in genes involved in the metabolism of alcohol and consist of a non-synonymous gain of function variant in *ADH1B*, rs1229984, and a damaging variant in *ALDH2*, rs671, which both reduce the risk development of alcohol-related disorders by increasing the levels of circulating acetaldehyde following alcohol ingestion (Edenberg and McClintick, 2018). As they discuss, other variants in *ADH* genes have also been reported to be associated in some studies, but linkage disequilibrium (LD) relationships in this region make interpretation difficult, and likewise, it is not clear if other variants in aldehyde dehydrogenase genes, including rs2228093 and rs2073478 in *ALDH1B1*, have effects (Edenberg and McClintick, 2018). We have reported that a Kozak sequence variant in *GRM3*, rs148754219, is associated with alcohol dependence as well as schizophrenia and bipolar disorder but this finding has to date not been replicated (O'Brien *et al.*, 2014).

The UK Biobank consists of a sample of 500,000 participants who have provided DNA samples and who have undergone phenotyping on a wide variety of measures on which GWAS have been performed, including the study of heavy drinking referred to above (Sudlow *et al.*, 2015; Thompson *et al.*, 2020); 94.6% of participants are of white ethnicity and as we have discussed regarding a previous analysis of BMI, it has become standard practice for investigators to simply discard data from participants with other ancestries and we regard this as regrettable (Curtis, 2021a). In that study, we carried out a weighted burden analysis of rare variants in 50,000 exome sequenced participants and demonstrated that if population principal components are included as covariates then it is possible to include all participants, regardless of ancestry, without any inflation of the test statistic. In this respect, the kind of weighted burden analysis we apply here is far less sensitive to population stratification than conventional GWAS because one is seeking to detect a difference in the overall frequency of rare variants within a gene rather than focussing on differences in frequencies of individual variants. Exome sequence data have now been made available for 200,000 participants (Szustakowski *et al.*, 2020). This makes it possible to investigate whether there may be rare sequence variants which have major effects on gene function and whether such variants within a gene are collectively associated with a trait of interest. Potentially, the identification of rare variants with functional effects could facilitate biological insights. Previous weighted burden analyses successfully implicated functional genes in the case of hyperlipidaemia and type 2 diabetes but did not produce any positive results for a phenotype related to common mental illness (Curtis 2021b, 2021c).

Here, we use this approach to carry out weighted burden analysis of rare, functional variants on two alcohol-related phenotypes, heavy drinking and problem drinking.

MATERIALS AND METHODS

The UK Biobank dataset was downloaded along with the variant call files for 200,632 subjects who had undergone exome-sequencing and genotyping by the UK Biobank Exome Sequencing Consortium using the GRCh38 assembly with coverage 20X at 95.6% of sites on average (Szustakowski et al. 2020). UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (approval number: 11/NW/0382) and had obtained informed consent from all participants. The UK Biobank approved an application for use of the data (ID 51119) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11527/001). All variants were annotated using the standard software packages VEP, PolyPhen and SIFT (Kumar et al. 2009; Adzhubei et al. 2013; McLaren et al. 2016). To obtain population principal components reflecting ancestry, version 2.0 of *plink* (https://www.cog-genomics.org/plink/2.0/) was run with the options *—maf 0.1 —pca 20 approx* (Chang et al. 2015; Galinsky et al. 2016).

Two alcohol-related phenotypes were assigned as follows. The phenotype for heavy drinking was defined in a similar way to that used in the previous GWAS using 125,249 UK Biobank participants, except that was based on reported average drinks in a week, whereas instead, we used reported drinks in the previous 24 hours (Thompson et al. 2020). On a number of occasions, a subset of participants were asked about their dietary intake in the previous 24 hours, and if they had consumed alcohol, they were asked to provide the number of drinks for different types of beverage. From these reports, the total daily intake was then estimated using the same number of units per drink as in the previous study: beer/cider = 2.6; white wine = 1.5; red wine = 1.5; fortified wine = 1.1; spirits = 1 and other = 1.5. The same sex-specific criteria were used as in the previous study, meaning that if 7 times this daily intake exceeded 50 for males or 35 for females, the participant was defined as a case in respect of heavy drinking, whereas controls were taken as participants who had completed the questionnaire on at least one occasion but did not exceed these limits (Thompson et al. 2020).

The phenotype of problem drinking was based on the 157,379 participants who completed a mental health questionnaire. Cases were defined as participants answering Yes to any of these three questions:

> Have you been physically dependent on alcohol?
> Has a relative or friend or a doctor or another health worker been concerned about your drinking or suggested you cut down?
> Have you been addicted to alcohol?

Participants selecting a response of Monthly or more frequently to any of these questions were also classified as cases:

> How often during the last year have you failed to do what was normally expected from you because of drinking?
> How often during the last year have you been unable to remember what happened the night before because you had been drinking?
> How often during the last year have you had a feeling of guilt or remorse after drinking?
> How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?
> How often during the last year have you found that you were not able to stop drinking once you had started?
> All other participants who had completed the mental health questionnaire were counted as controls.

The SCOREASSOC program was used to carry out a weighted burden analysis to test whether, in each gene, sequence variants which were rarer and/or predicted to have more severe functional effects occurred more commonly in cases than controls. Attention was restricted to rare variants with minor allele frequency (MAF) ≤ 0.01 in both cases and controls. As previously described, variants were weighted by overall MAF so that variants with MAF = 0.01 were given a weight of 1, while very rare variants with MAF close to zero were given a weight of 10 (Curtis 2021a). Variants were also weighted according to their functional annotation using the GENEVARASSOC program, which was used to generate input files for weighted burden analysis by SCOREASSOC (Curtis 2016, 2012). The weights were partially informed from the analysis of the effects of different

Table 1. The weight which was assigned to each type of variant as annotated by VEP, Polyphen and SIFT as well as the broad categories which were used for multivariate analyses of variant effects (Kumar et al. 2009; Adzhubei et al. 2013; McLaren et al. 2016)

VEP/SIFT/Polyphen annotation	Weight	Category
intergenic_variant	0	Unused
feature_truncation	0	Intronic, etc.
regulatory_region_variant	0	Intronic, etc.
feature_elongation	0	Intronic, etc.
regulatory_region_amplification	1	Intronic, etc.
regulatory_region_ablation	1	Intronic, etc.
TF_binding_site_variant	1	Intronic, etc.
TFBS_amplification	1	Intronic, etc.
TFBS_ablation	1	Intronic, etc.
downstream_gene_variant	0	Intronic, etc.
upstream_gene_variant	0	Intronic, etc.
non_coding_transcript_variant	0	Intronic, etc.
NMD_transcript_variant	0	Intronic, etc.
intron_variant	0	Intronic, etc.
non_coding_transcript_exon_variant	0	Intronic, etc.
3_prime_UTR_variant	1	3 prime UTR
5_prime_UTR_variant	1	5 prime UTR
mature_miRNA_variant	5	Unused
coding_sequence_variant	0	Unused
synonymous_variant	0	Synonymous
stop_retained_variant	5	Unused
incomplete_terminal_codon_variant	5	Unused
splice_region_variant	1	Splice region
protein_altering_variant	5	Protein altering
missense_variant	5	Protein altering
inframe_deletion	10	InDel, etc
inframe_insertion	10	InDel, etc
transcript_amplification	10	InDel, etc
start_lost	10	Unused
stop_lost	10	Unused
frameshift_variant	100	Disruptive
stop_gained	100	Disruptive
splice_donor_variant	100	Splice site variant
splice_acceptor_variant	100	Splice site variant
transcript_ablation	100	Disruptive
SIFT deleterious	20	Deleterious
PolyPhen possibly damaging	5	Possibly damaging
PolyPhen probably damaging	10	Probably damaging

categories of variant in *LDLR* on hyperlipidaemia risk, which showed for example that variants predicted to cause complete loss of function (LOF) of the gene had large effect sizes, while non-synonymous variants annotated as deleterious by SIFT had, on average, smaller effects (Curtis 2021b). Variants predicted to cause LOF were assigned a weight of 100. Non-synonymous variants were assigned a weight of 5, but if PolyPhen annotated them as possibly or probably damaging, then 5 or 10 was added to this, and if SIFT annotated them as deleterious, then 20 was added. The choice of weights is to some extent arbitrary but serves the purpose of applying higher weights to variants expected to have larger effects. In order to allow more systematic exploration of the effects of different types of variant on risk, the variants were also grouped into broader categories to be used in multivariate analyses as described below. The full set of weights and categories is displayed in Table 1.

As described previously, the weight due to MAF and the weight due to functional annotation were multiplied together to provide an overall weight for each variant. Variants were excluded if there were more than 10% of genotypes missing in the controls or cases or if the heterozygote count was smaller than both homozygote counts in the controls or cases. If a subject was not genotyped for a variant, then they were assigned the subject-wise average score for that variant. For each subject, a gene-wise weighted burden score was derived as the sum of the variant-wise weights, each multiplied by the number of alleles of the variant which the given subject possessed. For variants on the X chromosome, hemizygous males were treated as homozygotes.

For each gene, logistic regression analysis was carried out including the first 20 population principal components and sex as covariates and a likelihood ratio test was performed comparing the likelihoods of the models with and without the gene-wise burden score. The statistical significance was summarized as a signed log *P* value (SLP), which is the log base 10 of the *P* value given a positive sign if the score is higher in cases and negative if it is higher in controls. This provides a concise way to describe small *P* values as well as capturing the direction of effect.

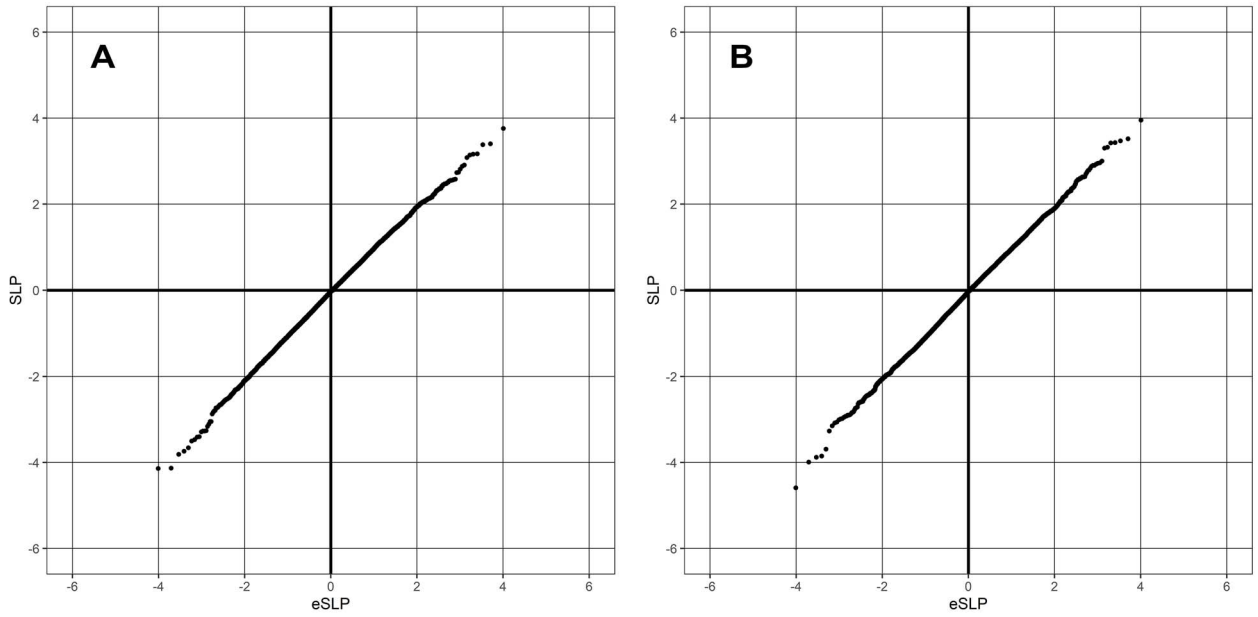

Fig. 1. QQ plot of SLPs obtained for weighted burden analysis of association with alcohol phenotypes showing observed against expected SLP for each gene. (A) Heavy drinking phenotype. (B) Problem drinking phenotype.

Gene set analyses were carried out as before using the 1454 'all GO gene sets, gene symbols' pathways as listed in the file *c5.all.v5.0.symbols.gmt* downloaded from the Molecular Signatures Database at http://www.broadinstitute.org/gsea/msigdb/collections.jsp (Subramanian et al. 2005). For each set of genes, the natural logs of the gene-wise *P* values were summed according to Fisher's method to produce a chi-squared statistic with degrees of freedom equal to twice the number of genes in the set and this was then used as a test of association of the set with the alcohol phenotypes.

Individual coding variants, which had previously been reported to be associated with alcohol-related phenotypes, were entered into logistic regression analyses with principal components and sex as covariates. The variants in the *ADH* gene cluster were analysed jointly in order to avoid spurious results due to LD relationships.

For selected genes, additional analyses were carried out to clarify the contribution of different categories of variant. As described previously, logistic regression analyses were performed on the counts of the separate categories of variant as listed in Table 1, again including principal components and sex as covariates, to estimate the effect size for each category (Curtis 2021b). The odds ratios associated with each category were estimated along with their standard errors and the Wald statistic was used to obtain a *P* value, except for categories in which variants occurred fewer than 50 times in which case Fisher's exact test was applied to the variant counts. In the case of *ADH1C*, the individual variants referred to above within the *ADH* region were included as additional covariates.

RESULTS

Exome sequence data were available for only a subset of the subjects who had completed the relevant questionnaires. For the heavy drinking phenotype, there were 8166 cases and 84,461 controls. For the problem drinking phenotype, there were 7811 cases and 59,606 controls. There were 20,384 genes for which there were qualifying variants. Given this and the fact that two phenotypes were tested, the critical threshold for the absolute value of the SLP to declare a result as formally statistically significant is $-\log10(0.05/(20384*2)) = 5.91$ and no gene achieved this for either phenotype. Figure 1A shows the QQ plot for the heavy drinking phenotype, which shows that the SLPs conform well with the null hypothesis expectation. If the genes with the 100 highest and lowest SLPs, which might be capturing a biological effect, are disregarded, then the gradient for the positive SLPs is 0.99 with intercept at −0.038 and the negative SLPs have gradient 1.03 with intercept at −0.042. Figure 1B shows the QQ plot for problem drinking. With the 100 highest and lowest SLPs removed, the positive SLPs have intercept −0.032 and gradient 0.99, while the negative SLPs have intercept −0.046 and gradient 1.04. Given that the null hypothesis expectation is that the gradient should be 1 with intercept at 0, these results suggest that the approach used is sound and that including the principal components and sex as covariates means that there is little inflation of the test statistics due to population stratification or other artefacts.

By chance, one would expect about 20 genes to have SLP with absolute value exceeding 3 (equivalent to $P < 0.001$) for each phenotype, whereas the actual numbers are 24 for heaving drinking and 18 for problem drinking. These genes are listed in Table 2 and, although not formally significant after correction for multiple testing, a few seem to be of potential interest.

Using heavy drinking as the phenotype, *VGF* produces SLP = −3.05, suggesting that functional variants in it might be associated with lower alcohol intake. As reviewed recently, *VGF* is expressed extensively in central nervous system, and there are reports that its expression is reduced in patients with depression but increased in schizophrenia, while both heterozygous *Vgf* knockout mice and *Vgf* over-expressing mice display some depressive features (Mizoguchi et al. 2019). *CFAP206* produces SLP = −3.11, but although expression of *Cfap206* is up-regulated in the hippocampus by alcohol administration in rats, there seems little else to suggest that it is likely to affect alcohol intake (Choi et al. 2020). Disruption of *FOXP1* (SLP = −3.4) has been reported to be associated with cognitive dysfunction including intellectual disability, autism spectrum disorder and language impairment as well as psychiatric symptoms (Bacon and

Table 2. Genes with absolute value of SLP exceeding 3 or more (equivalent to $P < 0.001$) for test of association of weighted burden score with alcohol phenotypes

A Results for heavy drinking phenotype

Symbol	SLP	Name
PAEP	3.76	Progestagen Associated Endometrial Protein
MYDGF	3.40	Myeloid Derived Growth Factor
PLA2G12B	3.38	Phospholipase A2 Group XIIB
ZNRF4	3.17	Zinc And Ring Finger 4
SYT16	3.16	Synaptotagmin 16
RAB29	3.14	RAB29, Member RAS Oncogene Family
TIAF1	3.08	TGFB1-Induced Anti-Apoptotic Factor 1
SEC24D	−3.05	SEC24 Homolog D, COPII Coat Complex Component
VGF	−3.05	VGF Nerve Growth Factor Inducible
CFAP206	−3.11	Cilia And Flagella Associated Protein 206
SMIM18	−3.16	Small Integral Membrane Protein 18
VASH2	−3.26	Vasohibin 2
TBXAS1	−3.27	Thromboxane A Synthase 1
TFEC	−3.27	Transcription Factor EC
CASQ2	−3.29	Calsequestrin 2
FOXP1	−3.40	Forkhead Box P1
WNT8B	−3.41	Wnt Family Member 8B
ATP13A2	−3.47	ATPase Cation Transporting 13A2
DNM2	−3.50	Dynamin 2
SALL4	−3.66	Spalt Like Transcription Factor 4
PRR7	−3.74	Proline Rich 7, Synaptic
MIR6505	−3.81	MicroRNA 6505
FBXL12	−4.13	F-Box And Leucine Rich Repeat Protein 12
KRT38	−4.14	Keratin 38
B Results for problem drinking phenotype		
Symbol	**SLP**	**Name**
SLC2A13	3.95	Solute Carrier Family 2 Member 13
IL19	3.52	nterleukin 19
ZNF714	3.47	Zinc Finger Protein 714
HCFC1R1	3.43	Host Cell Factor C1 Regulator 1
ZSWIM1	3.42	Zinc Finger SWIM-Type Containing 1
MYO5C	3.32	Myosin VC
GPR61	3.30	G Protein-Coupled Receptor 61
PHKA2	3.00	Phosphorylase Kinase Regulatory Subunit Alpha 2
PTGDR	−3.01	Prostaglandin D2 Receptor
ASAH2B	−3.06	N-Acylsphingosine Amidohydrolase 2B
HOXB4	−3.08	Homeobox B4
IRX5	−3.15	roquois Homeobox 5
HNRNPC	−3.27	Heterogeneous Nuclear Ribonucleoprotein C
ARHGAP33	−3.69	Rho GTPase Activating Protein 33
KLHDC8A	−3.85	Kelch Domain Containing 8A
HIST1H1D	−3.88	H1.3 Linker Histone, Cluster Member
NKIRAS2	−3.99	NFKB Inhibitor Interacting Ras Like 2
CDH9	−4.59	Cadherin 9

Rappold 2012; Siper et al. 2017). The two genome wide significant hits in a study of self-reported childhood maltreatment in 124,711 UK Biobank participants were at *FOXP1* and *FOXP2*, leading the authors to suggest that these genes might be involved in externalizing symptoms (Dalvie et al. 2020).

Two of genes with SLP < −3 for the problem drinking phenotype are also possibly implicated in autism risk, *ARHGAP33* (SLP = −3.69) and *CDH9* (SLP = −4.59). *ARHGAP33* is involved in dendrite and synapse development and its loss is associated with autism-like behaviours in mice (Rosário et al. 2007; Schuster et al. 2015; Nakazawa et al. 2016). A homozygous non-synonymous variant in *ARHGAP33* has been reported as a possible cause of a case of generalized developmental delay with seizures, microcephaly and dysmorphic features (Anazi et al. 2017). Common variants between *CDH9* and *CDH10* were shown to be associated with autism, and it is now recognized that several cadherins are involved in neuronal development and are designated as autism risk genes (Wang et al. 2009; Redies et al. 2012; Lin et al. 2016).

In order to see if any additional genes were highlighted by analysing gene sets, gene set analysis was performed as described

Table 3. Results of analyses of individual variants previously reported to be implicated in alcohol-related phenotypes, showing allele frequency in controls and cases; population principal components and sex were included in all analyses in order to calculate the SLP; for the analyses of the variants in *ADH1B* and *ADH1C*, the SLP was derived from a multivariate analysis which included all variants in these genes simultaneously

Variant	Gene	Chr	Position	Heavy drinking			Problem drinking		
				Control allele frequency	Case allele frequency	*P* value	Control allele frequency	Case allele frequency	*P* value
rs1229984	*ADH1B*	4	99318162	0.0345210	0.018185	1.0×10^{-14}	0.034420	0.017219	3.3×10^{-17}
rs2066702	*ADH1B*	4	99307860	0.0044820	0.002265	NS	0.003003	0.001536	NS
rs698	*ADH1C*	4	99339632	0.4008950	0.418677	NS	0.403057	0.420305	NS
rs1693482	*ADH1C*	4	99342808	0.3990130	0.417035	NS	0.401171	0.418408	NS
rs671	*ALDH2*	12	111803962	0.0015270	0.000061	0.0045	0.001300	0.000064	0.0071
rs2228093	*ALDH1B1*	9	38396005	0.1333550	0.128766	NS	0.131952	0.129113	NS
rs2073478	*ALDH1B1*	9	38396068	0.4095980	0.405500	NS	0.407919	0.402344	NS
rs148754219	*GRM3*	7	86644771	0.0074010	0.007396	NS	0.007156	0.008067	NS

above. Given that 1454 sets were tested for two phenotypes, a critical value to achieve to declare results significant after correction for multiple testing would be $P < 0.05/(1454*2) = 0.000017$ and this was not achieved by any set. Two sets achieved $P < 0.001$ with the problem drinking phenotype, NEGATIVE REGULATION OF CYTOSKELETON ORGANIZATION AND BIOGENESIS ($P = 0.000054$) and RESPONSE TO BIOTIC STIMULUS ($P = 0.00048$). Inspection of the detailed results for these sets did not suggest any additional genes as plausible candidates.

None of the genes mentioned in the introduction showed evidence for association with either phenotype using the gene-based burden tests, with the most significant result being SLP = 1.45 for *GCKR* with heavy drinking. The results for all genes are presented in Supplementary Table S1 and for all sets of genes in Supplementary Table S2.

The results for the variants previously reported to be associated with alcohol-related phenotypes are shown in Table 3. This shows that rs1229984, the gain of function variant in *ADH1B*, produces highly significant evidence for a protective effect for both phenotypes but that the other tested variants in *ADH1B* and *ADH1C* are not significantly associated with either phenotype. The damaging variant in *ALDH2*, rs671, is much less frequent in cases than controls for both phenotypes but is so rare in this largely European sample that it only produces modestly significant results. The tested variants in *ALDH1B1* and *GRM3* have equal frequencies in cases and controls for both phenotypes.

For the previously implicated genes listed in the introduction and for *VGF*, *FOXP1*, *ARHGAP33* and *CDH9*, additional analyses were carried out as described in order to see if there were particular categories of variant associated with the phenotypes. With a single exception, no category of variant within any of these genes showed evidence for association with heavy or problem drinking. The exception is that disruptive variants in *ADH1C* initially appeared to be protective for both heavy drinking, $P = 0.00095$, OR = 0.73 (0.60–0.88), and for problem drinking, $P = 0.00016$, OR = 0.68 (0.56–0.83). Although there were seven different disruptive variants, six of them were extremely rare and the effect was driven by a single stop-gained variant with MAF = 0.0099, 4:99347033:C > A. When the other variants within the ADH gene cluster were included as covariates in a multivariate analysis, then this association disappeared and only the effect of rs1229984 itself was significant, suggesting that the overall result was due to LD between the stop variant and rs1229984 rather than any separate effect from impaired functioning of *ADH1C*. The results for the analyses of categories of variant are presented in Supplementary Table S3.

DISCUSSION

No genes attain conventional levels of statistical significance after correction for multiple testing. The two well-established variants in *ADH1B* and *ALDH2*, rs1229984 and rs671, demonstrate association when tested for individually, but there is no evidence to implicate other rare coding variants in alcohol-metabolizing genes. Overall, the conclusion is that this large sample of exome sequenced subjects has failed to identify additional genes or variants influencing susceptibility to heavy drinking or problem drinking.

Although biobanks can provide large samples, they suffer the disadvantage that one must rely on data which are available for a substantial proportion of subjects and one does not have the option to apply standardized operational criteria as one might for a custom-designed case-control study. Both phenotypes used clearly have their limitations. The heavy drinking phenotype is based on a report of drinking over the last 24 hours and so clearly might be more prone to random fluctuations than one based on a self-reported average consumption over a week or month, as was used in the previous study (Thompson et al. 2020). On the other hand, it might have an advantage in terms of being less sensitive to recall and reporting bias. Likewise, the phenotype for problem drinking depended on answering positively to any of a number of questions covering a wide range of severities from being addicted to alcohol to having had somebody suggest cutting down. Nevertheless, the fact that both phenotypes produced highly significant results with rs1229984 supports the thesis that they do have some validity. In fact, it is noteworthy that the effect size for this variant seems very similar for both phenotypes. This appears to be the case also for rs671 although its rarity in this dataset makes it difficult to draw firm conclusions.

Although no gene reached formal significance levels, it is possibly of some interest that three genes potentially related to autism were each individually significant at $P < 0.001$: *FOXP1*, *ARHGAP33* and *CDH9*. There is also a report claiming to show association with alcoholism of clustered SNPs near other cadherin genes, *CDH11* and *CDH13*, though this was not genome-wide significant, and we have previously reported a statistically significant association between alcoholism in bipolar disorder and common variants in

CDH11 (Johnson et al. 2006; Lydall et al. 2011). There is a negative association between alcohol abuse and autism, and adults with autism spectrum disorders have relatively low rates of alcohol and substance use disorders (Vohra et al. 2017). Thus, it would be easy to speculate that if there were genetic variants associated with autistic traits then these might be protective against heavy drinking and problem drinking. However, a study of the polygenic risk for general substance abuse failed to detect a correlation with genetic susceptibility to autism, although this may be a feature of the relatively small sample size used (Carey et al. 2016). The other gene noted to be of potential interest, given its claimed involvement in psychiatric phenotypes, is *VGF*. However, it needs to be stressed that none of these findings is formally significant and they should be regarded merely as hypothesis-generating.

Although the UK Biobank sample consists of 500,000 participants, to date, exome sequence data have only been generally released for 200,000. It is possible that more definitive results may be obtained when the approach described here can be applied to the remaining participants and in other datasets. The results for *CDH9* produced a likelihood ratio chi-squared statistic of 17.7 with this sample, and if we assume that this reflects a real effect size, we can calculate that a sample only 1.3 times larger would have the expectation to produce an exome-wide significant result. It is also possible that studying a more extreme phenotype such as alcohol dependence disorder would have more power but this would involve focused recruitment strategies rather than relying on biobank resources. However, at present, it is not clear whether or not further genetic research will be able to yield actionable insights into factors influencing susceptibility to alcohol-related disorders.

DATA AVAILABILITY

The raw data are available on application to UK Biobank. Scripts and relevant derived variables will be deposited in UK Biobank. Software and scripts used to carry out the analyses are also available at https://github.com/davenomiddlenamecurtis.

ACKNOWLEDGEMENTS

This research has been conducted using the UK Biobank Resource. The author wishes to acknowledge the staff supporting the High-Performance Computing Cluster, Computer Science Department, University College London.

FUNDING

This work was carried out in part using resources provided by BBSRC equipment grant BB/R01356X/1.

CONFLICT OF INTEREST STATEMENT

The author declares that he has no conflict of interest.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Alcohol and Alcoholism* online.

REFERENCES

Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 7:20.

Anazi S, Maddirevula S, Faqeih E, *et al.* (2017) Clinical genomics expands the morbid genome of intellectual disability and offers a high diagnostic yield. *Mol Psychiatry* 22:615–24.

Bacon C, Rappold GA (2012) The distinct and overlapping phenotypic spectra of FOXP1 and FOXP2 in cognitive disorders. *Hum Genet* 131:1687–98.

Carey CE, Agrawal A, Bucholz KK, *et al.* (2016) Associations between polygenic risk for psychiatric disorders and substance involvement. *Front Genet* 7:149.

Chang CC, Chow CC, Tellier LC, *et al.* (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4:7.

Choi MR, Han JS, Chai YG, *et al.* (2020) Gene expression profiling in the hippocampus of adolescent rats after chronic alcohol administration. *Basic Clin Pharmacol Toxicol* 126:389–98.

Curtis D (2012) A rapid method for combined analysis of common and rare variants at the level of a region, gene, or pathway. *Adv Appl Bioinform Chem* 5:1–9.

Curtis D (2016) Pathway analysis of whole exome sequence data provides further support for the involvement of histone modification in the aetiology of schizophrenia. *Psychiatr Genet* 26:223–7.

Curtis D (2021a) Multiple linear regression allows weighted burden analysis of rare coding variants in an ethnically heterogeneous population. *Hum Hered* 85:1–10.

Curtis D (2021b) Weighted burden analysis in 200,000 exome-sequenced subjects characterises rare variant effects on risk of type 2 diabetes. *medRxiv* 2021.01.08.21249453.January 09, 2021, doi 10.1101/2021.01.08.21249453, preprint: not peer reviewed.

Curtis D (2021c) Analysis of 200,000 exome-sequenced UK biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia [WWW document]. *medRxiv* . https://www.medrxiv.org/content/10.1101/2021.02.10.21251503v3 January 06, 2021, doi 10.1101/2021.01.05.20249090, preprint: not peer reviewed.

Dalvie S, Maihofer AX, Coleman JRI, *et al.* (2020) Genomic influences on self-reported childhood maltreatment. *Transl Psychiatry* 10:1234567890.

Edenberg HJ, McClintick JN (2018) Alcohol dehydrogenases, aldehyde dehydrogenases, and alcohol use disorders: a critical review. *Alcohol Clin Exp Res* 42:2281–97.

Galinsky KJ, Bhatia G, Loh PR, *et al.* (2016) Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* 98:456–72.

Johnson C, Drgon T, Liu QR, *et al.* (2006) Pooled association genome scanning for alcohol dependence using 104,268 SNPs: validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism. *Am J Med Genet Part B Neuropsychiatr Genet* 141:844–53.

Johnson EC, Chang Y, Agrawal A (2020) An update on the role of common genetic variation underlying substance use disorders. *Curr Genet Med Rep* 8:35–46.

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073–81.

Lin YC, Frei JA, Kilander MBC, *et al.* (2016) A subset of autism-associated genes regulate the structural stability of neurons. *Front Cell Neurosci* doi: https://doi.org/10.3389/fncel.2016.00263.

Lydall GJ, Bass NJ, McQuillin A, *et al.* (2011) Confirmation of prior evidence of genetic susceptibility to alcoholism in a genome-wide association study of comorbid alcoholism and bipolar disorder. *Psychiatr Genet* 21: 294–306.

McLaren W, Gil L, Hunt SE, *et al.* (2016) The ensembl variant effect predictor. *Genome Biol* 17:122.

Mizoguchi T, Hara H, Shimazawa M (2019) VGF has roles in the pathogenesis of major depressive disorder and schizophrenia: evidence from transgenic mouse models. *Cell Mol Neurobiol* 39:721–27.

Nakazawa T, Hashimoto R, Sakoori K, *et al.* (2016) Emerging roles of ARHGAP33 in intracellular trafficking of TrkB and pathophysiology of neuropsychiatric disorders. *Nat Commun* 7:10594.

O'Brien NL, Way MJ, Kandaswamy R, *et al.* (2014) The functional GRM3 Kozak sequence variant rs148754219 affects the risk of schizophrenia

and alcohol dependence as well as bipolar disorder. *Psychiatr Genet* 24: 277–78.

Redies C, Hertel N, Hübner CA (2012) Cadherins and neuropsychiatric disorders. *Brain Res* 1470:130–44.

Rosário M, Franke R, Bednarski C, *et al.* (2007) The neurite outgrowth multiadaptor RhoGAP, NOMA-GAP, regulates neurite extension through SHP2 and Cdc42. *J Cell Biol* 178:503–16.

Schuster S, Rivalan M, Strauss U, *et al.* (2015) NOMA-GAP/ARHGAP33 regulates synapse development and autistic-like behavior in the mouse. *Mol Psychiatry* 20:1120–31.

Siper PM, De Rubeis S, Trelles MDP, *et al.* (2017) Prospective investigation of FOXP1 syndrome. *Mol Autism* 8:57.

Subramanian A, Tamayo P, Mootha VK, *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102:15545–50.

Sudlow C, Gallacher J, Allen N, *et al.* (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 12:e1001779.

Szustakowski JD, Balasubramanian S, Sasson A, *et al.* (2020) Advancing human genetics research and drug discovery through exome sequencing of the UK biobank. *medRxiv* 2020.11.02.20222232.

Thompson A, Cook J, Choquet H, *et al.* (2020) Functional validity, role, and implications of heavy alcohol consumption genetic loci. *Sci Adv* 6:eaay5034.

Vohra R, Madhavan S, Sambamoorthi U (2017) Comorbidity prevalence, healthcare utilization, and expenditures of Medicaid enrolled adults with autism spectrum disorders. *Autism* 21:995–1009.

Wang K, Zhang H, Ma D, *et al.* (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature* 459: 528–33.