Case-only rare variant analysis of severe alcohol dependence (AD) using a multivariate hierarchical gene clustering approach.

AUTHORS
Amanda E. Gentry[1], Jeffrey Alexander[2], Roseann E. Peterson[1], Silviu A. Bacanu[1], Brien P. Riley[1], Bradley T. Webb[3]
1. Virginia Institute for Psychiatric and Behavioral Genetics, Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, 23219, USA.
2. Unaffiliated
3. GenOmics, Bioinformatics, and Translational Research Center, Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, 27709, USA.

ABSTRACT
Background: Variation in genes involved in ethanol metabolism has been shown to influence risk for alcohol dependence (AD) including protective loss of function alleles in ethanol metabolizing genes. We therefore hypothesized that people with severe AD would exhibit different patterns of rare functional variation in genes with strong prior evidence for influencing ethanol metabolism and response when compared to genes not meeting these criteria.
Objective: Apply a novel analysis method to leverage a case only design and Whole Exome Sequencing (WES) of severe AD cases from Ireland in order to quantify the difference in functional variation between ethanol-metabolizing genes and their matched controls.
Methods: First, three sets of genes known to influence ethanol metabolism or response were generated. Set 1 comprised genes involved in alcohol metabolism in human including alcohol dehydrogenase (ADH), aldehyde dehydrogenase (ALDH), the Cytochrome P450 Family 2 Subfamily E Member 1 (CYP2E1), and catalase (CAT) genes (n=11). Set 2 comprised set 1, in addition to genes showing altered expression in mouse brain after alcohol exposure (n=117). Set 3 comprised sets 1 and 2, in addition to genes shown to alter ethanol behavioral response in invertebrate models (n=358). These sets of genes of interest (GOI) were then matched to sets of control genes using multivariate hierarchical clustering of gene-level summary features obtained from the gnomAD database. Using WES data from 195 individuals with severe alcohol dependence, GOI were then compared to their matched control gene sets using multivariate logistic regression to detect aggregate differences in abundance of synonymous, missense, and loss of function (LOF) variants, respectively.
Results: A significant difference was not found in the number of functional variants in the primary set of 10 ethanol-metabolizing genes, compared to a matched set of control genes. In both the mouse expression and invertebrate sets, we observed an increased number of synonymous variants in GOI over matched control genes.
Conclusion: The proposed method demonstrates a computationally viable and statistically appropriate approach for genetic analysis of case-only data.

Introduction

Genetics of Alcohol Use Disorder: A large literature supports the existence of genetic influences on risk for alcohol use disorder (AUD) and alcohol-related phenotypes (ARP) in family (Cotton, 01/1979), adoption (Cadoret et al., 01/1987; Cloninger, 1981; Goodwin, 1973; Sigvardsson, 1996) and twin(Heath et al., 1997; Hrubec and Omenn, 03/1981; Kendler et al., 1992, 1997; McGue et al., 1992; Pickens, 1991; Prescott et al., 1999, 2005; Reed et al., 1996; Romanov et al., 1991; True et al., 1996) studies with twin study estimates of heritability reaching 50% (Verhulst et al., 2015) and SNP-based heritability estimates in Europeans of around 5.6% (Kranzler et al., 2019). Recent genome-wide association studies (GWAS) have successfully identified common simple nucleotide variants (cSNV) robustly associated with alcohol consumption (Liu et al., 2019), the Alcohol Use Disorders Identification Test (AUDIT) (Sanchez-Roige et al., 02/2019), and problematic alcohol use (Kranzler et al., 2019). Studies of other complex traits suggest that many additional genetic loci remain to be identified for alcohol-related measures including alcohol use disorder (AUD). In parallel to the success of GWAS, evidence from a variety of phenotypes is consistent in suggesting that rare simple nucleotide variants (rSNV) also influence complex traits. However, the relative contribution of cSNV and rSNV to the overall genetic architecture of risk for specific alcohol-related traits is not yet precisely known.

Rare variants, the human genome, common disorders: The 1000 Genomes Project estimated there are ~340-400 premature stop, splice-site disrupting and frame-shift alleles affecting 250-300 genes per individual genome (1000 Genomes Project Consortium et al., 2010) . A more stringent analysis of a subset of 185 individuals from this study estimated this total to be approximately 100 true loss of function (LOF) variants with a mean of 20 genes completely inactivated per genome (MacArthur et al., 2012). Importantly, a large survey of 202 drug-target loci also observed that rare variants are population-specific, supporting the necessity of studying rare functional variation impacting AUD risk in an ethnically homogeneous sample (Nelson et al., 2012). Excess rare variation in cases compared to controls is also observed in genes targeted by common variant GWAS signals in studies of blood lipids (Johansen et al., 2010, 2011; Razzaghi et al., 2012; Sanna et al., 2011), height (Lanktree et al., 01/2011), celiac disease (Trynka et al., 2011) inflammatory bowel disease (Rivas et al., 2011) and asthma (Torgerson et al., 2012), suggesting that disease risk is influenced by multiple alleles of varying frequencies in the same genes (Panagiotou et al., 2010).

Rare variants in psychiatric disorders: Rare functional variation is an important element of risk for intellectual disability (Lelieveld et al., 2016), autism spectrum disorders (Iossifov et al., 2014) and schizophrenia (Purcell et al., 2014). Rare functional variants have not yet been widely studied through exome sequencing in AUD, but early evidence points to their significant contribution to the genetic architecture of alcohol use phenotypes (Brazel et al., 2019; Kranzler et al., 2019; Sanchez-Roige et al., 02/2019; Walters et al., 2018). A population-specific variant in the serotonin receptor 2B (*HTR2B*) gene is strongly associated with impulsivity and violent behavior in Finnish AUD cases (Bevilacqua et al., 2010). Functional variants in the nicotinic cholinergic receptor subunit alpha 4 (*CHRNA4*) gene are protective against nicotine dependence (Xie et al., 2011).

Ethanol-metabolizing genes: Four gene families exhibit accumulated evidence of alcohol metabolism*, ALDH, ADH, CYP*, and *CAT*. *ADH1A*, *ADH1B*, and *ADH1C* are expressed in liver and encode subunits of ADH1. ADH1 isoforms metabolize ethanol at a range of low concentrations from 0.05 mM (ADH1B homo-dimer) to 1 mM (ADH1C homo-dimer) to 4 mM

(ADH1A homo-dimer) (Cederbaum, 2012) and account for ~70% of the ethanol metabolizing capacity of the liver (Lee et al., 2006). *ADH4* is expressed in liver and encodes ADH4. The ADH4 homo-dimer has activity for ethanol metabolism at high concentration (30 mM) (Bosron et al., 1993). *ADH5* is expressed in stomach and liver, and while the homo-dimer has activity for ethanol metabolism only at concentrations much higher than found in liver (>1M), it may play an important role in ethanol metabolism in the stomach because gastric alcohol concentrations can reach molar range during alcohol consumption (Lee et al., 2003). *ADH6* is expressed in liver, stomach and duodenum. While less is known about the function of ADH6 in vivo than the other ADH classes, ADH6 expressed in vitro in a cell-based system showed activity for ethanol metabolism similar to that of ADH7 (Chen and Yoshida, 1991). *ADH7* is preferentially expressed in the oral and nasal cavities, esophagus and stomach. While ADH7 is thought to be primarily involved in metabolism of retinol and related compounds, it does have activity for ethanol metabolism at high concentration (28 mM) (Xie et al., 1997) and the presence of ethanol inhibits both all-trans-retinol and 9-cis-retinol oxidation (Allali-Hassani et al., 1998). *CYP2E1* expression is induced by chronic alcohol consumption and CYP2E1 in hepatic microsomes is important in metabolizing ethanol at moderate concentrations (8-10 mM) in liver (Lieber and DeCarli, 1970) and also contributes to ethanol metabolism in the brain where ADH activity is limited (Zimatkin and Deitrich, 1997). And finally, *CAT* shows wide expression in brain, and brain peroxisomal catalase provides slow but detectable ethanol metabolism (Gill et al., 1992). Both CYP2E1 and catalase are considered part of canonical EtOH metabolism (Zakhari, 2006).

Functional variants and alcohol: Variation in these genes involved in ethanol metabolism has been shown to influence risk for alcohol dependence (AD). The well documented *ALDH2*2* (rs671) LOF allele shows only 20-40% of wild type enzymatic activity in heterozygote carriers due to the homotetrameric structure of mature ALDH2. When ethanol is consumed, the reduced enzymatic activity leads to the accumulation of acetaldehyde, the first product of ethanol catabolism, and unpleasant symptoms including flushing. This variant is common in East Asian populations, associated with lower rates of alcohol abuse/dependence (Thomasson et al., 1994), but practically absent in others (Goedde et al., 1992). Evidence for effects of *ALDH2* in European populations is inconsistent (Dickson et al., 2006; Luo et al., 2006), but the protective impact of these alleles has been clearly demonstrated.

Beyond ALDH2*2: Functional variants also exist in the alcohol dehydrogenase (ADH) genes with allele frequencies that differ considerably across populations. Of 7 ADH genes, located in a cluster on chromosome 4q, three code for Class 1 enzymes involved with alcohol metabolism (Guengerich, 1999). Variants in two of these, *ADH1B* and *ADH1C*, show protective effects against AUD by causing an aversive reaction to ethanol (Osier et al., 1999; Shen et al., 1997). The *ADH1B*2* (rs1229984) and *ADH1C*1* (rs698) alleles are common and are in strong linkage disequilibrium (LD) in East Asian populations (Goedde et al., 1992; Shen et al., 1997) and in moderate LD in other populations (Meyers et al., 2013), while *ADH1B*3* (rs2066702) is restricted to populations of African descent (Bosron and Li, 1987; Edenberg, 2007). Recently, rs75967634 also in *ADH1B* was found to be associated with problematic alcohol use in European populations (Zhou et al., 2020). Although of low frequency outside of east Asia, *ADH1B*2* (rs1229984) is associated with both AUD diagnosis (Gelernter et al., 2014; Walters et al., 2018) and the problem drinking component of the AUDIT (Sanchez-Roige et al., 02/2019) in European populations and may have an epistatic interaction with ADH7 variation in protection against AUD (Osier et al., 2004).

CYP2E1: In addition to primary metabolism by ADH/ALDH, ethanol is catalyzed by CYP2E1 which is inducible, more widely expressed, and has significant activity at higher blood alcohol concentrations (BACs). The *CYP2E1*1D* allele has been reported to have greater enzyme inducibility and is a target of candidate gene studies for alcohol and nicotine dependence (Gemma et al., 2006). Common polymorphisms in CYP2E1 have been studied extensively for their role influencing consumption and alcoholic liver disease. However, there are no common alleles with effect sizes similar to those observed in ALDH or ADH.

CAT: Few studies of genetic variation in the CAT gene and alcohol-related problems have been reported. One study of the association of the CAT -262C>T (rs1001179) promoter polymorphism with level of response to ethanol and diagnosis of alcohol abuse or dependence in 85 male subjects found no evidence of association (Hu et al., 2005). A second study of this promoter polymorphism in 101 current alcohol-dependent cases, 100 former alcohol-dependence cases, and 97 controls detected a higher frequency of CAT -262T in all alcohol-dependent subjects ($p<0.007$), while current case CAT -262T allele carriers had higher AUDIT scores ($P = 0.023$) (Plemenitas et al., 2015).

Protective LOF: More broadly, studies of cardiac risk factors (Cohen et al., 2004, 2006; Kotowski et al., 2006), obesity (Benzinou et al., 2008), Alzheimer disease (Guerreiro et al., 2013) and non-alcoholic fatty liver disease (Romeo et al., 2008) consistently show that nonsynonymous and LOF alleles are common in the human genome, alter disease risk, and in some cases are protective (Nejentsev et al., 2009; Romeo et al., 2007) as in *ADH* and *ALDH*.

Cases only designs: It is not uncommon in genetic research to obtain molecular data on a sample of subjects who are all positive for a dichotomous phenotype, such as case only studies. Methods have been developed for intentional case-only design studies, including some to investigate gene-by-environment interactions (Sahebi et al., 2013) and extended to include polygenic risk scores (Meisner et al., 2019). These case-only methods however, generally consider an environmental exposure as the dichotomous outcome being tested for association with genetic variation within a sample of all cases with a particular disease and offer no information regarding phenotypic variation due to the primary disease state.

When controls have not been sequenced alongside cases in the same study, publicly available datasets may provide population controls when appropriately matched. When population controls are available, it is still necessary to account for differences in sampling techniques, sequencing technology, filtering criteria, calling methods, and sequencing depth. Failing to adjust for these by application of naive, traditional case/control methods may result in spurious findings (Persyn et al., 2018).

Existing case only rare variant methods: Some novel methods for case/control studies of rare variants with population controls have been developed. The Robust Variance Score Statistic (RVS) method (Derkach et al., 2014) was developed to adjust for systematic differences between sequenced cases and population controls which might otherwise confound true case/control differences. The burden test implemented in the TASER software (Hu et al., 2016) which is as an extension of the Derkatch method utilizes sequencing read data directly, rather than called genotypes. This method offers improved adjustment for sample differences. Unlike RVS and TASER, the iECAT method (Lee et al., 11/2017) does not require individual-level genotype data from population controls, but rather conducts an adjusted rare variant association test using only allele counts. While powerful, the iECAT requires at least some number of internal controls to be sequenced before external controls may be incorporated. Finally, ProxECAT (Hendricks et al., 2018) does not require any internal controls, utilizes only allele counts, However, none of these

methods (RVS, TASER, iECAT, or ProxECAT) offers correction or adjustment for related samples as currently implemented.

Hypothesis testing in case only designs: While some of the aforementioned methods allow for the use of population controls, none are appropriate for case only studies of closely related individuals since they do not allow for any relatedness corrections. Given the existence of protective LOF alleles in ethanol metabolizing genes, we hypothesized that people with severe AUD would carry less rare functional variation than expected. Second, we hypothesized that genes robustly involved with alcohol related traits in model organisms would show different patterns of rare variation in comparison to genes with similar attributes. In the second hypothesis, the direction of effect is not specified.

Given the absence of implemented methods to test our hypotheses in a case-only framework with related subjects, we sought to develop a novel method to address these challenges. As an alternative to comparing aggregate rare variants on a gene-by-gene basis between individual case and control subjects, sets of genes derived from a specific hypothesis can be compared to a matched set of control genes. If genes of interest (GOI) can be matched to similar genes not in the hypothesized set using external sources of independent information, then unbiased comparisons across sets can be made in a case-only study. Since comparison sets are derived from within an individual's genome, they would not be subject to any potential inflation from stratification due to rare variant population structure. Further, this provides a rigorous framework for hypothesis testing where robust prior sources of evidence are available. This method is complementary to case-control association testing and can be performed in samples without matched controls which can be non-trivial for studies of rare variants.

In the current study, we developed and implemented such an approach and applied it to a sample of 195 Irish individuals with severe alcoholism with exome-wide sequencing data. For each hypothesis, we sought to compare the numbers of synonymous (SYN, base changes that do not alter the amino acid product), missense (MIS, base changes that alter the amino acid product), and loss-of-function (LOF, base changes resulting in loss of gene function) variants between genes of interest (GOI) to a matched set of control genes.

Materials and Methods

Sample description: The Irish Affected Sib-Pair Study of Alcohol Dependence (ISP) sample was collected from 1998-2002 in treatment facilities and hospitals in Ireland and Northern Ireland. Probands with all four grandparents born in Ireland or Britain were ascertained for a diagnosis of DSM-IV (American Psychiatric Association and American Psychiatric Association, 2000) alcohol dependence (AD) and one or more affected siblings. Lifetime history of AD was assessed using a modification of the SSAGA version 11 (Bucholz et al., 1994) which permits evaluation of ICD-10, Feighner (Feighner et al., 1972), RDC (Spitzer et al., 1978), DSM-III-R (American Psychiatric Association and American Psychiatric Association, 1987) and DSM-IV diagnostic criteria. The sample is severely affected, with ~87% of probands and ~78% of siblings endorsing ≥6 of the 7 DSM-IV AD criteria and 92% reported withdrawal symptoms. Parents were evaluated for lifetime history of alcohol abuse and AD based on the SCID (Spitzer et al., 1992), the CAGE (Ewing, 1984), and FAST (Hodgson et al., 2002) items developed to screen for drinking problems.

Exome capture and sequencing: Exome capture was performed using the Agilent SureSelect V5 71Mb exome + UTR target kit, followed by library preparation and sequencing on the Illumina HiSeq X Ten system at BGI. The 190 samples were sequenced in 3 batches, batches 1 (n=57) and 2 (n=76) via 2x90 and batch 3 (n=57) via 2x100 paired end sequencing.

Variant calling: We processed and called the sequenced data according to GATK3 best practices and summarized the raw data for quality control using FastQC (v 0.11.4). We performed sequence read alignment with BWA-MEM (v 0.7.12) to hs37d5 reference, followed by reordering, duplicate marking and in/del realignment with Picard (v 2.0.1), variant calling with HaplotypeCaller, and variant quality score recalibration with VQSR in GATK (v 3.5).
Annotation: We obtained gene annotations using SnpEff (v 4.3, database GRCh37.75) and included only the transcripts found in the gnomAD (release 2.1.1) gene constraints
Description of candidate genes: Construction of GOIs was performed in collaboration with investigators from the Alcohol Research Center (ARC) at Virginia Commonwealth University which exists to investigate cross-species gene discovery and functional interpretation for the genetics of AUD. Three sets of alcohol related GOI were constructed for this analysis which crossed three taxonomic groups including human, mouse, and invertebrates. The first metabolizing set contains 11 genes whose products are known to be involved with ethanol metabolism in humans including the ADH (n=7), ALDH (n=2), CAT (n=1), and CYP (n=1) families (Table 1). The second mouse expression set includes 109 hub genes with both high connectivity and high centrality in coexpression networks showing altered expression in mouse prefrontal cortex (PFC) 4 hours after intraperitoneal injection of saline or 1.8 g/kg of ethanol (Wolen et al., 2012) (Supplemental Table 1). The final invertebrate set contains 358 genes for which manipulation in invertebrate models results in altered ethanol response phenotypes (Grotewiel and Bettinger, 08/2015) (Supplemental Table 2). Figure 1 illustrates the overlap between the gene sets.

Annotations for gene clustering: In order to create matched sets of genes for our novel case-only analytical approach, we utilized independent external gene-level annotations from the gnomAD database (version 2.1.1) (Karczewski et al., 2020) as the basis for gene matching. The seven gnomAD annotations used to cluster all genes included the ratios of observed-to-expected (O/E) counts for each variant class (LOF, MIS, and SYN), a z-score for each O/E ratio, and the probability of being loss-of-function intolerant (*pLI*) score. In addition to these metrics from gnomAD, genes were annotated for clustering with metrics for genomic length, transcript length, and number of exons. In total, ten annotations were utilized for clustering. While correlations between some variables were high (see Table 2), none were considered close enough to one to warrant dropping from clustering. Integration of the hypothesized gene lists with gnomAD yielded sets sizes of 10, 108, and 353 for analysis in the metabolizing, expression, and invertebrate GOIs, respectively.

Clustering: Multivariable single linkage agglomerative clustering of all canonical gnomAD genes was accomplished in R (v 3.6.0) using the *hclust* package (*method="single"*). The clustering algorithm groups gene-level information from the ten previously described annotations to identify genes similar to the GOIs in size and function to create a within-subject control set of genes against which the GOIs could be compared in a regression framework. Importantly, this clustering into gene sets procedure utilizes information from gnomAD only, separate from the sample exomes. Clustering was performed using a dissimilarity metric of (1-|cor(X)|), where X is the qxp matrix of the q=10 normalized features of each gene (p=19,108). The 10 variables included in the clustering were (1-3) the ratio of observed to expected LOF, MIS, and SYN, and (4-6) a z-score for the deviation of observed from expected for each, (7) pLI (which describes the probability that a gene is loss-of-function intolerant,) (8) transcript length, (9) genomic length, and (10) number of exons. Figure 2 shows example gene clusters plotted according to three gnomAD metrics, SYN z-score, transcript length, and pLI. The data shown

represents 4 gene clusters in color and a set of genes chosen at random in gray, with ellipses highlighting the shape of the cluster.

Matrix cor(X) therefore represents the full set of pairwise correlations for all 19,108 genes in the set. The hierarchical clustering algorithm proceeds in a stepwise fashion, beginning with all observations (genes) in a separate cluster of their own. At each successive step, the two least dissimilar clusters are joined together. We prune the final trees for all three gene sets, choosing a height with enough genes clustered with each GOI for useful comparison while not exceeding 20% of the exome included in the final branches. Branches in the tree which contain at least one GOI are termed Clusters of Interest (COI). After a cut height is chosen, all other non-COIs are pruned from the tree.

Association testing:

After clustering and pruning, we retained all COIs containing at least one GOI and dropped all other genes. We then used logistic regression to compare observed counts of SYN, MIS, and LOF variants from the ISP exomes, in aggregate, between GOI and matched control genes. In this framework, the dependent variable is the gene, not human subject, so that the logistic regression quantifies the probability that a given gene is a GOI as a function of the observed counts of LOF, MIS, and SYN variants as aggregated across the subjects.

For each of the three, successively larger gene sets, the model is constructed as follows:

$$GOI \sim LOF_{Alc} + MIS_{Alc} + SYN_{Alc},$$

where GOI represents case/control status of the gene, and LOF_{Alc}, MIS_{Alc} and SYN_{Alc} represent the observed counts of LOF, MIS, and SYN variants in that gene in the alcohol study sample, respectively. This model tests a null hypothesis of no association between the LOF, MIS, and SYN variant counts and probability of being a GOI versus a non-GOI. Evidence against the null hypothesis indicates that the alcohol-related genes of interest contain differing amounts of variation, as measured by counts of LOF, MIS, and SYN variants, from their matched control genes within the alcohol sample.

Results

WES data for all 195 subjects passed quality control standards, with mean on-target sequencing depth across all samples of 60.6x (standard deviation 12.02) and 96.7% of the target covered at ≥10x. A total of 782,711 variants were detected with 677,758 SNPs and an additional 109,526 indels. For quality control, SNPs and indels were excluded if they fell into GATK Variant Quality Score Recalibration (VQSR) tranches 99.0 or greater, indicating that in the filtered set, 99% of the true variants present in the sample will be retained (38,503 variants removed). Variants were excluded if common as defined as by having a MAF≥5% in gnomAD non-Finnish European non-cancer samples which resulted in a final set of 652,428 variants (91,780 removed) with 2,328, 31,015, and 46,046 LOF, MIS, and SYN, respectively.

The correlation between the LOF, MIS, and SYN observed counts for all genes from gnomAD and the observed exome-wide counts from the alcohol subject data were 0.26, 0.80, and 0.76, respectively. This indicates that at least for MIS and SYN variants, there is good evidence that the gnomAD database information may be reasonably utilized to group genes for a within-sample analysis. Table 2 shows the correlation between the 10 gene measures from the gnomAD database and indicates these annotations are measuring disparate genomic features. This difference supports including all ten in the multivariate clustering.

Table 3 describes the resulting trees cut at various heights. We chose to cut all three trees at height=0.09 (representing 9% of the tree), a value which achieves a median cluster size of 8 genes while still only utilizing just under 20% of the exome in the clusters for the largest gene set

of interest. Figure 3 shows one branch, representing one cluster, of the pruned tree, with the GOI labeled in red. From a methodological standpoint, the size of GOI sets will depend on the hypothesis and application. Therefore, the decision on cut height will need to be made on an experiment-wise basis in order to balance the need for large clusters (for maximum statistical power for comparisons) and smaller overall proportion of the exome included in the COIs (for tight clusters with high similarity across metrics). For each of the three hypotheses, Figure 4 illustrates the relationship between GOI set size, median cluster size and proportion of the exome included in the COIs.

For each of the three hypotheses tested, the logistic regression model was constructed using all genes within COIs. Total sample sizes for the three models tested were 149, 1,639, and 3,719 for the primary GOI, expression, and invertebrate sets, respectively. Supplemental tables 3-5 list the control genes for each model.

Using an alpha cutoff of 0.05, the first model with only 10 alcohol-related GOI did not show a statistically significant difference between number of LOF, MIS, or SYN variants in GOIs and control genes. For the expression set, the second model found a significant difference in counts of synonymous variants with odds ratios of exp(0.18)=1.21, p-value 0.0006, indicating that the additional of a single synonymous variant conferred a 1.21 increase in the odds of that gene being an ethanol-metabolizing gene of interest. Finally, for the invertebrate set, the third model found a significant difference in the number of synonymous variants with an odds ratio of exp(0.06) = OR 1.06, p-value 0.0169. Model estimates are shown in Table 4.

Discussion

Rationale: We sought to answer the hypothesis that persons with severe AD will have different amounts of variation within genes that are 1) involved in ethanol metabolism or 2) implicated in alcohol phenotypes via model organism studies.

Results: While some differences in SYN variants were found, the results do not support the hypothesis that ethanol metabolizing genes are largely depleted for loss of function variants in severe AD cases. The current results offer clues towards answers to an important question in the field of AD genetics.

Limitations: Several limitations inherent in this work bear recognition. First, our sample size of 195 affected subjects is limited and these tests may not be sufficiently powered. Additionally, we empirically modeled the median number of matched genes in each cluster against the proportion of the genome included in the final clustered set in order to choose an appropriate pruning parameter. Simulations to evaluate the impact of various parameters choices, such as tree pruning height or proportion of the genome included in gene clusters, were not performed.

Contributions: For this analysis, we modeled case/control status of individual genes within a case only sample of severely affected alcohol-dependent subjects. For situations where data is available only for cases, in the absence of an appropriate control set, we have presented a reasonable modeling approach. In contrast to studies using external controls subjects, the controls were matched genes from the cases' own genomes. This strategy removes the need for careful correction or adjustment for subtle population structure in rare variant studies which can be challenging. This novel approach leverages external information from large sequencing studies such as gnomAD to ensure gene matching is robust by using multivariate agnostic hierarchical clustering. Furthermore, this methodological contribution represents an interpretable, straight-forward, and computationally inexpensive approach that is easily implemented in existing software. Alternatives, such as modeling a quantitative measure within

cases, such as max number of drinks per day, would also be appropriate for a case-only analysis, but address a fundamentally different hypothesis than the one addressed here.

Acknowledgements: This work was supported in part by the Virginia Commonwealth University Alcohol Research Center and NIAAA P50AA022537.

References:

1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467(7319): 1061–1073.

Allali-Hassani A, Peralba JM, Martras S, et al. (1998) Retinoids, omega-hydroxyfatty acids and cytotoxic aldehydes as physiological substrates, and H2-receptor antagonists as pharmacological inhibitors, of human class IV alcohol dehydrogenase. *FEBS letters* 426(3): 362–366.

American Psychiatric Association and American Psychiatric Association (eds.) (1987) *Diagnostic and Statistical Manual of Mental Disorders: DSM-III-R*. 3rd ed., rev. Washington, DC: American Psychiatric Association.

American Psychiatric Association and American Psychiatric Association (eds.) (2000) *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR*. 4th ed., text revision. Washington, DC: American Psychiatric Association.

Benzinou M, Creemers JWM, Choquet H, et al. (2008) Common nonsynonymous variants in PCSK1 confer risk of obesity. *Nature genetics* 40(8): 943–945.

Bevilacqua L, Doly S, Kaprio J, et al. (2010) A population-specific HTR2B stop codon predisposes to severe impulsivity. *Nature* 468(7327): 1061–1066.

Bosron WF and Li TK (1987) Catalytic properties of human liver alcohol dehydrogenase isoenzymes. *Enzyme* 37(1–2): 19–28.

Bosron WF, Ehrig T and Li TK (1993) Genetic factors in alcohol metabolism and alcoholism. *Seminars in liver disease* 13(2): 126–135.

Brazel DM, Jiang Y, Hughey JM, et al. (2019) Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use. *Biological psychiatry* 85(11): 946–955.

Bucholz KK, Cadoret R, Cloninger CR, et al. (1994) A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. *Journal of studies on alcohol* 55(2): 149–158.

Cadoret RJ, Troughton E and O'Gorman TW (01/1987) Genetic and environmental factors in alcohol abuse and antisocial personality. *Journal of studies on alcohol* 48(1): 1–8.

Cederbaum AI (2012) Alcohol metabolism. *Clinics in liver disease* 16(4): 667–685.

Chen CS and Yoshida A (1991) Enzymatic properties of the protein encoded by newly cloned human alcohol dehydrogenase ADH6 gene. *Biochemical and biophysical research communications* 181(2): 743–747.

Cloninger CR (1981) Inheritance of Alcohol Abuse: Cross-Fostering Analysis of Adopted Men. *Archives of general psychiatry* 38(8): 861.

Cohen JC, Kiss RS, Pertsemlidis A, et al. (2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305(5685): 869–872.

Cohen JC, Boerwinkle E, Mosley TH Jr, et al. (2006) Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *The New England journal of medicine* 354(12): 1264–1272.

Cotton NS (01/1979) The familial incidence of alcoholism: a review. *Journal of studies on alcohol* 40(1): 89–116.

Derkach A, Chiang T, Gong J, et al. (2014) Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics* 30(15): 2179–2188.

Dickson PA, James MR, Heath AC, et al. (2006) Effects of variation at the ALDH2 locus on alcohol metabolism, sensitivity, consumption, and dependence in Europeans. *Alcoholism, clinical and experimental research* 30(7): 1093–1100.

Edenberg HJ (2007) The genetics of alcohol metabolism: role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol research & health: the journal of the National Institute on Alcohol Abuse and Alcoholism* 30(1): 5–13.

Ewing JA (1984) Detecting alcoholism. The CAGE questionnaire. *JAMA: the journal of the American Medical Association* 252(14): 1905–1907.

Feighner JP, Robins E, Guze SB, et al. (1972) Diagnostic criteria for use in psychiatric research. *Archives of general psychiatry* 26(1): 57–63.

Gelernter J, Kranzler HR, Sherva R, et al. (2014) Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. *Molecular psychiatry* 19(1): 41–49.

Gemma S, Vichi S and Testai E (2006) Individual susceptibility and alcohol effects:biochemical and genetic aspects. *Annali dell'Istituto superiore di sanita* 42(1): 8–16.

Gill K, Menez JF, Lucas D, et al. (1992) Enzymatic production of acetaldehyde from ethanol in rat brain tissue. *Alcoholism, clinical and experimental research* 16(5): 910–915.

Goedde HW, Agarwal DP, Fritze G, et al. (1992) Distribution of ADH2 and ALDH2 genotypes in different populations. *Human genetics* 88(3): 344–346.

Goodwin DW (1973) Alcohol Problems in Adoptees Raised Apart From Alcoholic Biological Parents. *Archives of general psychiatry* 28(2): 238.

Grotewiel M and Bettinger JC (08/2015) *Drosophila* and *Caenorhabditis elegans* as Discovery Platforms for Genes Involved in Human Alcohol Use Disorder. *Alcoholism, clinical and experimental research* 39(8): 1292–1311.

Guengerich FP (1999) *Comprehensive Toxicology, Volume 3: Biotransformation*. Elsevier Science.

Guerreiro R, Wojtas A, Bras J, et al. (2013) TREM2 variants in Alzheimer's disease. *The New England journal of medicine* 368(2): 117–127.

Heath AC, Bucholz KK, Madden PA, et al. (1997) Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. *Psychological medicine* 27(6): 1381–1396.

Hendricks AE, Billups S, Pike HNC, et al. (2018) *ProxECAT: Proxy External Controls Association Test. A new case-control gene region association test using allele frequencies from public controls*. 24 February. Genetics. Available at: http://biorxiv.org/lookup/doi/10.1101/271007 (accessed 23 May 2019).

Hodgson R, Alwyn T, John B, et al. (2002) The FAST Alcohol Screening Test. *Alcohol and alcoholism* 37(1): 61–66.

Hrubec Z and Omenn GS (03/1981) Evidence of Genetic Predisposition to Alcoholic Cirrhosis and Psychosis: Twin Concordances for Alcoholism and Its Biological End Points by Zygosity among Male Veterans. *Alcoholism, clinical and experimental research* 5(2): 207–215.

Hu X, Oroszi G, Chun J, et al. (2005) An expanded evaluation of the relationship of four alleles to the level of response to alcohol and the alcoholism risk. *Alcoholism, clinical and experimental research* 29(1): 8–16.

Hu Y-J, Liao P, Johnston HR, et al. (2016) Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls. *PLoS genetics* Hoffmann TJ (ed.) 12(5): e1006040.

Iossifov I, O'Roak BJ, Sanders SJ, et al. (2014) The contribution of de novo coding mutations to autism spectrum disorder. *Nature* 515(7526): 216–221.

Johansen CT, Wang J, Lanktree MB, et al. (2010) Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature genetics* 42(8): 684–687.

Johansen CT, Wang J, Lanktree MB, et al. (2011) An increased burden of common and rare lipid-associated risk alleles contributes to the phenotypic spectrum of hypertriglyceridemia. *Arteriosclerosis, thrombosis, and vascular biology* 31(8): 1916–1926.

Karczewski KJ, Francioli LC, Tiao G, et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809): 434–443.

Kendler KS, Heath AC, Neale MC, et al. (1992) A population-based twin study of alcoholism in women. *JAMA: the journal of the American Medical Association* 268(14): 1877–1882.

Kendler KS, Prescott CA, Neale MC, et al. (1997) Temperance board registration for alcohol abuse in a national sample of Swedish male twins, born 1902 to 1949. *Archives of general psychiatry* 54(2): 178–184.

Kotowski IK, Pertsemlidis A, Luke A, et al. (2006) A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *American journal of human genetics* 78(3): 410–422.

Kranzler HR, Zhou H, Kember RL, et al. (2019) Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. *Nature communications* 10(1): 1499.

Lanktree MB, Guo Y, Murtaza M, et al. (01/2011) Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height. *American journal of human genetics* 88(1): 6–18.

Lee S, Kim S and Fuchsberger C (11/2017) Improving power for rare-variant tests by integrating external controls. *Genetic epidemiology* 41(7): 610–619.

Lee S-L, Wang M-F, Lee A-I, et al. (2003) The metabolic role of human ADH3 functioning as ethanol dehydrogenase. *FEBS letters* 544(1–3): 143–147.

Lee S-L, Chau G-Y, Yao C-T, et al. (2006) Functional assessment of human alcohol dehydrogenase family in ethanol metabolism: significance of first-pass metabolism. *Alcoholism, clinical and experimental research* 30(7): 1132–1142.

Lelieveld SH, Reijnders MRF, Pfundt R, et al. (2016) Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. *Nature neuroscience* 19(9): 1194–1196.

Lieber CS and DeCarli LM (1970) Hepatic microsomal ethanol-oxidizing system. In vitro characteristics and adaptive properties in vivo. *The Journal of biological chemistry* 245(10): 2505–2512.

Liu M, Jiang Y, Wedow R, et al. (2019) Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature genetics* 51(2): 237–244.

Luo X, Kranzler HR, Zuo L, et al. (2006) Diplotype trend regression analysis of the ADH gene cluster and the ALDH2 gene: multiple significant associations with alcohol dependence. *American journal of human genetics* 78(6): 973–987.

MacArthur DG, Balasubramanian S, Frankish A, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335(6070): 823–828.

McGue M, Pickens RW and Svikis DS (1992) Sex and age effects on the inheritance of alcohol problems: a twin study. *Journal of abnormal psychology* 101(1): 3–17.

Meisner A, Kundu P and Chatterjee N (2019) *Case-only analysis of gene-environment interactions using polygenic risk scores*. 19 February. Genetics. Available at: http://biorxiv.org/lookup/doi/10.1101/555300 (accessed 23 May 2019).

Meyers JL, Shmulewitz D, Aharonovich E, et al. (2013) Alcohol-metabolizing genes and alcohol phenotypes in an Israeli household sample. *Alcoholism, clinical and experimental research* 37(11): 1872–1881.

Nejentsev S, Walker N, Riches D, et al. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* 324(5925): 387–389.

Nelson MR, Wegmann D, Ehm MG, et al. (2012) An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337(6090): 100–104.

Osier M, Pakstis AJ, Kidd JR, et al. (1999) Linkage disequilibrium at the ADH2 and ADH3 loci and risk of alcoholism. *American journal of human genetics* 64(4): 1147–1157.

Osier MV, Lu R-B, Pakstis AJ, et al. (2004) Possible epistatic role of ADH7 in the protection against alcoholism. *American journal of medical genetics. Part B, Neuropsychiatric genetics: the official publication of the International Society of Psychiatric Genetics* 126B(1): 19–22.

Panagiotou OA, Evangelou E and Ioannidis JPA (2010) Genome-wide Significant Associations for Variants With Minor Allele Frequency of 5% or Less—An Overview: A HuGE Review. *American journal of epidemiology* 172(8): 869–889.

Persyn E, Redon R, Bellanger L, et al. (2018) The impact of a fine-scale population stratification on rare variant association test results. *PloS one* 13(12): e0207677.

Pickens RW (1991) Heterogeneity in the Inheritance of Alcoholism: A Study of Male and Female Twins. *Archives of general psychiatry* 48(1): 19.

Plemenitas A, Kastelic M, Porcelli S, et al. (2015) Genetic variability in CYP2E1 and catalase gene among currently and formerly alcohol-dependent male subjects. *Alcohol and alcoholism* 50(2): 140–145.

Prescott CA, Aggen SH and Kendler KS (1999) Sex differences in the sources of genetic liability to alcohol abuse and dependence in a population-based sample of U.S. twins. *Alcoholism, clinical and experimental research* 23(7): 1136–1144.

Prescott CA, Caldwell CB, Carey G, et al. (2005) The Washington University Twin Study of alcoholism. *American journal of medical genetics* 134B(1): 48–55.

Purcell SM, Moran JL, Fromer M, et al. (2014) A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 506(7487): 185–190.

Razzaghi H, Santorico SA and Kamboh MI (2012) Population-Based Resequencing of LIPG and ZNF202 Genes in Subjects with Extreme HDL Levels. *Frontiers in genetics* 3. DOI: 10.3389/fgene.2012.00089.

Reed T, Page WF, Viken RJ, et al. (1996) Genetic predisposition to organ-specific endpoints of alcoholism. *Alcoholism, clinical and experimental research* 20(9): 1528–1533.

Rivas MA, Beaudoin M, Gardet A, et al. (2011) Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics* 43(11): 1066–1073.

Romanov K, Kaprio J, Rose RJ, et al. (1991) Genetics of alcoholism: effects of migration on concordance rates among male twins. *Alcohol and alcoholism . Supplement* 1: 137–140.

Romeo S, Pennacchio LA, Fu Y, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nature genetics* 39(4): 513–516.

Romeo S, Kozlitina J, Xing C, et al. (2008) Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nature genetics* 40(12): 1461–1465.

Sahebi L, Dastgiri S, Ansarin K, et al. (2013) Study Designs in Genetic Epidemiology. *ISRN Genetics* 2013: 1–8.

Sanchez-Roige S, Palmer AA, Fontanillas P, et al. (02/2019) Genome-Wide Association Study Meta-Analysis of the Alcohol Use Disorders Identification Test (AUDIT) in Two Population-Based Cohorts. *American Journal of Psychiatry* 176(2): 107–118.

Sanna S, Li B, Mulas A, et al. (2011) Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLoS genetics* Gibson G (ed.) 7(7): e1002198.

Shen YC, Fan JH, Edenberg HJ, et al. (1997) Polymorphism of ADH and ALDH genes among four ethnic groups in China and effects upon the risk for alcoholism. *Alcoholism, clinical and experimental research* 21(7): 1272–1277.

Sigvardsson S (1996) Replication of the Stockholm Adoption Study of Alcoholism: Confirmatory Cross-Fostering Analysis. *Archives of general psychiatry* 53(8): 681.

Spitzer RL, Endicott J and Robins E (1978) Research diagnostic criteria: rationale and reliability. *Archives of general psychiatry* 35(6): 773–782.

Spitzer RL, Williams JB, Gibbon M, et al. (1992) The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description. *Archives of general psychiatry* 49(8): 624–629.

Thomasson HR, Crabb DW, Edenberg HJ, et al. (1994) Low frequency of the ADH2*2 allele among Atayal natives of Taiwan with alcohol use disorders. *Alcoholism, clinical and experimental research* 18(3): 640–643.

Torgerson DG, Capurso D, Mathias RA, et al. (2012) Resequencing candidate genes implicates rare variants in asthma susceptibility. *American journal of human genetics* 90(2): 273–281.

True WR, Heath AC, Bucholz K, et al. (1996) Models of treatment seeking for alcoholism: the role of genes and environment. *Alcoholism, clinical and experimental research* 20(9): 1577–1581.

Trynka G, Hunt KA, Bockett NA, et al. (2011) Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nature genetics* 43(12): 1193–1201.

Verhulst B, Neale MC and Kendler KS (2015) The heritability of alcohol use disorders: a meta-analysis of twin and adoption studies. *Psychological medicine* 45(5): 1061–1072.

Walters RK, Polimanti R, Johnson EC, et al. (2018) Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. *Nature neuroscience* 21(12): 1656–1669.

Wolen AR, Phillips CA, Langston MA, et al. (2012) Genetic Dissection of Acute Ethanol Responsive Gene Networks in Prefrontal Cortex: Functional and Mechanistic Implications. *PloS one* Palmer AA (ed.) 7(4): e33575.

Xie P, Parsons SH, Speckhard DC, et al. (1997) X-ray structure of human class IV sigmasigma alcohol dehydrogenase. Structural basis for substrate specificity. *The Journal of biological chemistry* 272(30): 18558–18563.

Xie P, Kranzler HR, Krauthammer M, et al. (2011) Rare nonsynonymous variants in alpha-4 nicotinic acetylcholine receptor gene protect against nicotine dependence. *Biological psychiatry* 70(6): 528–536.

Zakhari S (2006) Overview: how is alcohol metabolized by the body? *Alcohol research & health: the journal of the National Institute on Alcohol Abuse and Alcoholism* 29(4): 245–254.

Zhou H, Sealock JM, Sanchez-Roige S, et al. (2020) Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. *Nature neuroscience* 23(7): 809–818.

Zimatkin SM and Deitrich RA (1997) Ethanol metabolism in the brain. *Addiction biology* 2(4): 387–400.

Tables

Table 1: Alcohol-metabolizing genes of interest (n=11), with annotation indicating which were present in the gnomAD gene constraints file (10), which were present in the mouse set (1), and which were present in the invertebrate set (5).

Gene	Transcript	Present in gnomAD	Present in the invertebrate set	Present in the expression set
ADH1A	ENST00000209668	TRUE	TRUE	FALSE
ADH1B	ENST00000305046	TRUE	TRUE	FALSE
ADH1C		FALSE	TRUE	FALSE
ADH4	ENST00000265512	TRUE	FALSE	FALSE
ADH5	ENST00000296412	TRUE	FALSE	FALSE
ADH6	ENST00000394899	TRUE	FALSE	FALSE
ADH7	ENST00000476959	TRUE	FALSE	FALSE
ALDH1A1	ENST00000297785	TRUE	TRUE	FALSE
ALDH2	ENST00000261733	TRUE	TRUE	TRUE
CAT	ENST00000241052	TRUE	FALSE	FALSE
CYP2E1	ENST00000463117	TRUE	FALSE	FALSE

Table 2: Correlations between the gene metrics from the gnomAD database.

	O/E LOF	O/E MIS	O/E SYN	LOF z-score	MIS z-score	SYN z-score	pLI	Gen. Length	Tran. Length	No. of Exons
O/E LOF	1	0.551	0.207	-0.728	-0.548	-0.137	-0.619	-0.141	-0.156	-0.161
O/E MIS	0.551	1	0.416	-0.472	-0.874	-0.395	-0.495	-0.062	0.002	-0.071
O/E SYN	0.207	0.416	1	-0.113	-0.326	-0.848	-0.031	-0.007	0.008	-0.028
LOF z-score	-0.728	-0.472	-0.113	1	0.648	0.069	0.654	0.335	0.572	0.59
MIS z-score	-0.548	-0.874	-0.326	0.648	1	0.403	0.564	0.135	0.149	0.245
SYN z-score	-0.137	-0.395	-0.848	0.069	0.403	1	0.016	-0.032	-0.101	-0.03
pLI	-0.619	-0.495	-0.031	0.654	0.564	0.016	1	0.148	0.158	0.141
Gen. Length	-0.141	-0.062	-0.007	0.335	0.135	-0.032	0.148	1	0.315	0.378
Tran. Length	-0.156	0.002	0.008	0.572	0.149	-0.101	0.158	0.315	1	0.802
No. of Exons	-0.161	-0.071	-0.028	0.59	0.245	-0.03	0.141	0.378	0.802	1

O/E: Observed/Expected
Gen. Length: Genomic Length
Tran. Length: Transcript Length
No. of Exons: Number of Exons

Table 3: Median cluster size (Med. Size), number of genes per cluster (No.Genes) and proportion of the exome retained in the clusters of interest (Prop. Gen) for hierarchical clustering trees cut at various heights.

	Primary alcohol-related GOI			**Including expression set GOI**			**Including invertebrate GOI**		
eight	**Med. Size**	**No. Genes**	**Prop.**	**Med. Size**	**No. Genes**	**Prop.**	**Med. Size**	**No. Genes**	**Prop.**
0	1	10	0.001	1	117	0.006	1	358	0.019
0.01	2.5	22	0.001	2	214	0.011	1	589	0.031
0.02	3.5	35	0.002	3	376	0.02	2	993	0.052
0.03	4	42	0.002	4	538	0.028	3	1388	0.073
0.04	6.5	75	0.004	5	731	0.038	4	1822	0.095
0.05	7.5	83	0.004	5	891	0.047	4	2230	0.117
0.06	8	97	0.005	8	1072	0.056	5	2561	0.134
0.07	8	108	0.006	8	1252	0.066	6	3007	0.157
0.08	11.5	130	0.007	10	1403	0.073	8	3351	0.175
.09	12.5	149	0.008	11	1639	0.086	8	3719	0.195
0.15	26	290	0.015	22	2818	0.147	14.5	6220	0.326
0.2	36	470	0.025	26	3662	0.192	20.5	7798	0.408
0.25	59.5	608	0.032	34	4285	0.224	25	9238	0.483
0.5	247	2353	0.123	94	8549	0.447	60	14849	0.777
0.75	426	4272	0.224	152	12227	0.64	118	17349	0.908
1.00	19,108	19,108	1.00	19,108	19,108	1.00	19,108	19,108	1.00

Med. Size: Median cluster size
No. Genes: Number of genes
Prop. Gen.: Proportion of the exome included

Table 4: Logistic regression results.

Primary alcohol-related GOI
N=149 (10 GOI, 139 control genes)

Parameter	Estimate	Standard Error	Marginal P-value
Intercept	*-2.36*	*0.531*	*<0.0001*
LOF Alc	*-0.17*	*0.900*	*0.849*
MIS Alc	*-0.064*	*0.206*	*0.757*
SYN Alc	*-0.11*	*0.276*	*0.678*

Expression Set
N=1,639 (117 GOI, 1,522 control genes)

Parameter	***Estimate***	***Standard Error***	***Marginal P-value***
Intercept	*-2.74*	*0.126*	*<0.001*
LOF Alc	*-0.80*	*0.552*	*0.1462*
MIS Alc	*-0.05*	*0.044*	*0.2296*
SYN Alc	*0.18*	*0.053*	*0.0006*

Invertebrate Set
N=3,719 (358 GOI, 3,361 control genes)

Parameter	***Estimate***	***Standard Error***	***Marginal P-value***
Intercept	*-2.34*	*0.071*	*<0.001*
LOF Alc	*-0.10*	*0.166*	*0.5330*
MIS Alc	*0.004*	*0.022*	*0.8367*
SYN Alc	*0.06*	*0.026*	*0.0169*

Figures

Figure 1: Venn diagram depicting the three GOI sets. (*Figure created with BioRender.com*)

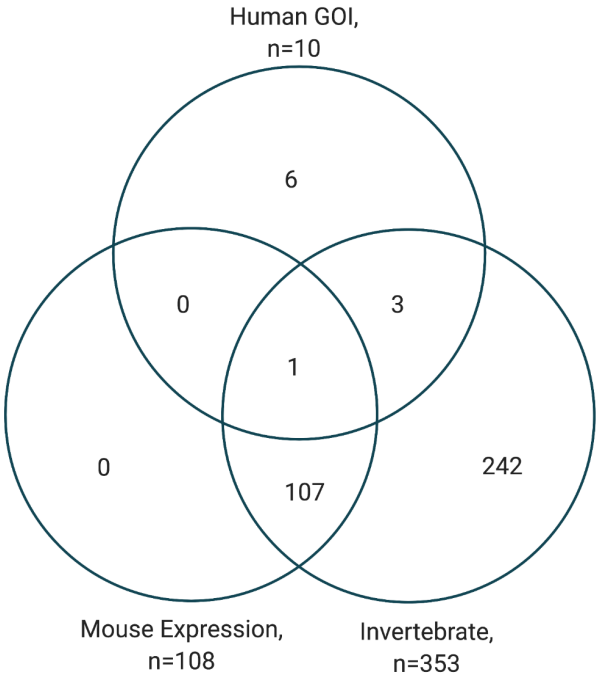

Figure 2: Example gene clusters plotted to show three of the 10 clustering metrics, SYN z-score, transcript length, and pLI. The 4 clusters in orange, green, blue, and pink represent actual gene clusters from the data, while the genes in gray represent a random selection of genes.

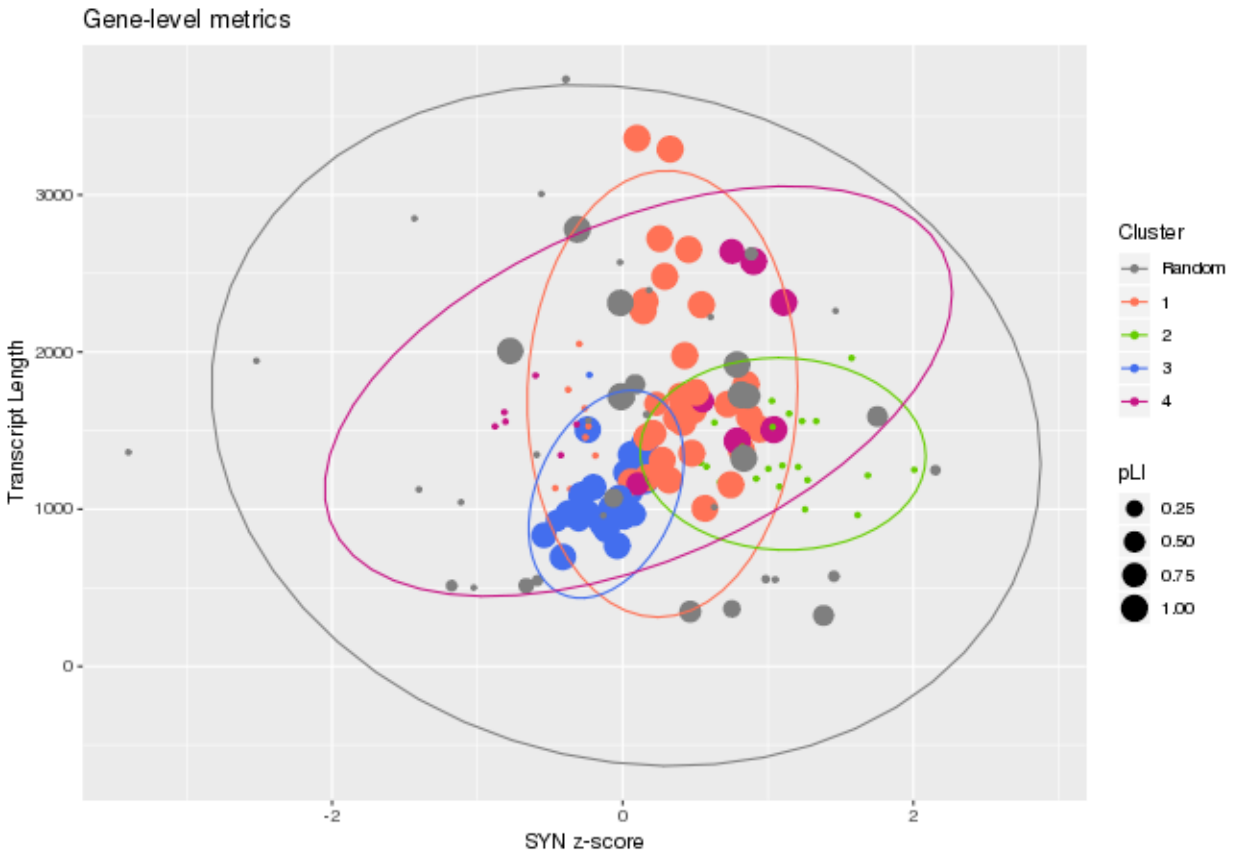

Figure 3: Flow chart demonstrating the analysis method. (*Figure created with BioRender.com*)

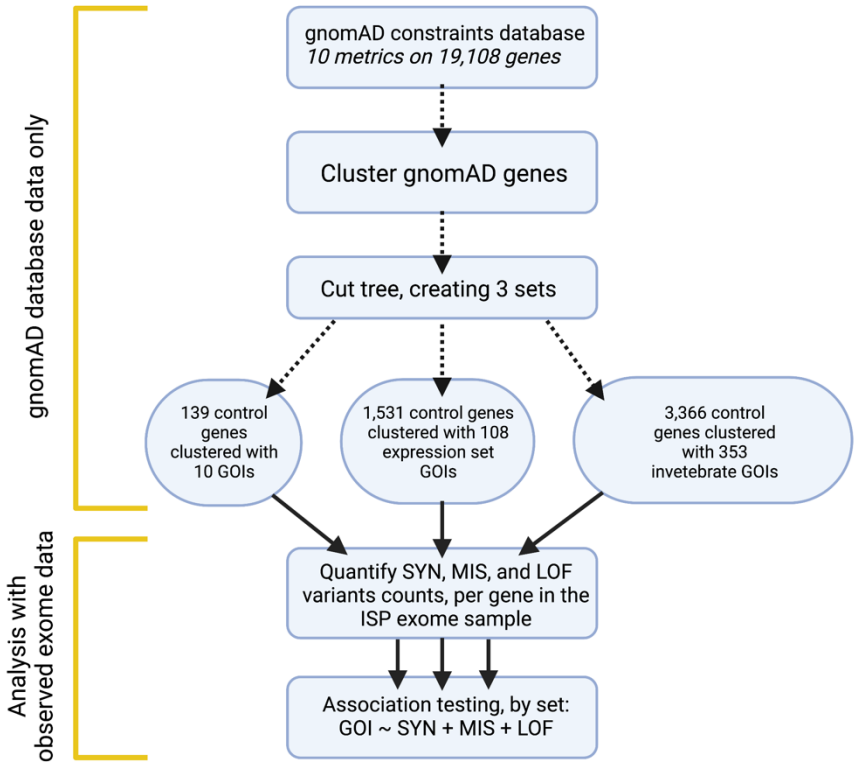

Figure 4: Example of one branch (cluster of interest) from the final, pruned hierarchical clustering tree.

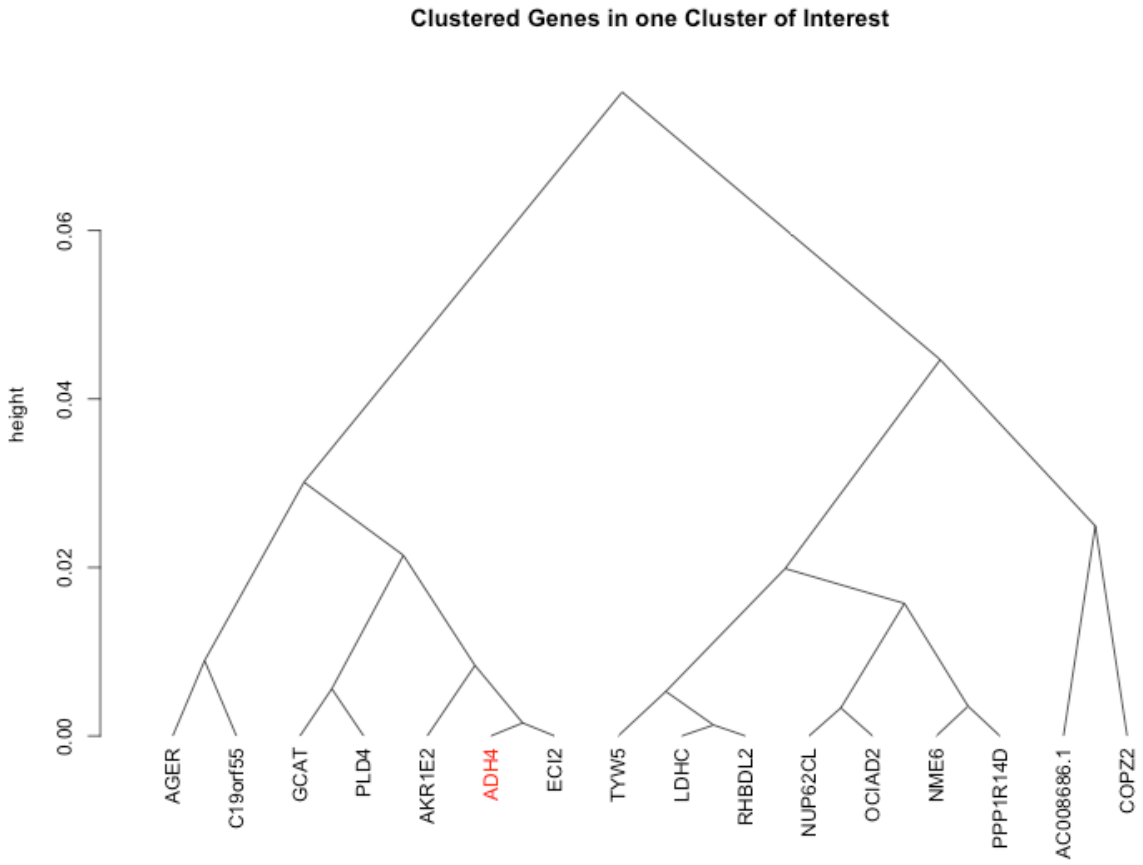
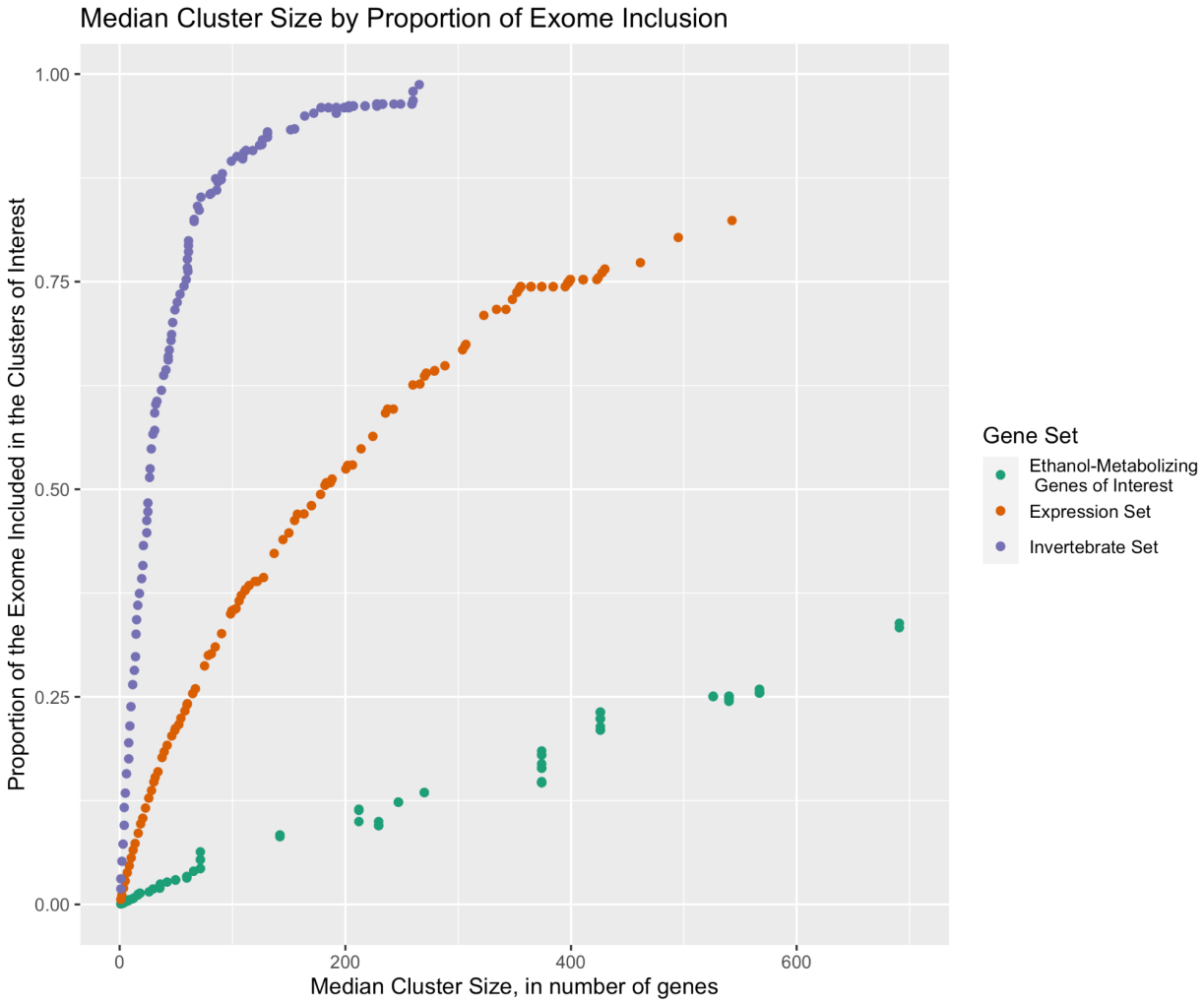
Figure 4: Relationship between median number of genes per COI and proportion of the exome included in the COIs.