TITLE

Case-only exome variation analysis of severe alcohol dependence using a multivariate hierarchical gene clustering approach.

SHORT TITLE

Case-only exome variation analysis of severe alcohol dependence

AUTHORS

Amanda E. Gentry[1,2], Jeffry C. Alexander[1*], Mohammad Ahangari[1,3], Roseann E. Peterson[1,2,4,5], VCU Alcohol Research Center working group, Michael F. Miles[5,6], Jill C. Bettinger[5,6], Andrew G. Davies[5,6], Mike Groteweil[5,7], Silviu A. Bacanu[1,2,5], Kenneth S. Kendler[1,2,5], Brien P. Riley[1,2,5,7], Bradley T. Webb[5,8]

1. Virginia Institute for Psychiatric and Behavioral Genetics, Virginia Commonwealth University, Richmond, VA, 23219, USA.
2. Department of Psychiatry, Virginia Commonwealth University, Richmond, VA, 23219, USA.
3. Integrative Life Sciences Ph.D. Program, Virginia Commonwealth University, Richmond, VA 23219, USA.
4. Department of Psychiatry and Behavioral Sciences, Institute for Genomics in Health, SUNY Downstate Health Sciences University, Brooklyn, NY, 11203, USA.
5. VCU Alcohol Research Center, Virginia Commonwealth University, Richmond, VA, 23219, USA.
6. Department of Pharmacology and Toxicology, Virginia Commonwealth University, Richmond, VA, 23219, USA.
7. Department of Human and Molecular Genetics, Virginia Commonwealth University, Richmond, VA, 23219, USA.
8. GenOmics, Bioinformatics, and Translational Research Center, Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, 27709, USA.
*Currently at Veterinary Information Network

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: Leverage a novel case only design and Whole Exome Sequencing (WES) of severe AD cases from the island of Ireland to quantify differences in functional variation between genes associated with ethanol metabolism and/or response and their matched control genes.

Methods: First, three sets of ethanol related genes were identified including those a) involved in alcohol metabolism in humans b) showing altered expression in mouse brain after alcohol exposure, and altering ethanol behavioral responses in invertebrate models. These genes of interest (GOI) sets were matched to control gene sets using multivariate hierarchical clustering of gene-level summary features from gnomAD. Using WES data from 190 individuals with severe AD, GOI were compared to matched control genes using logistic regression to detect aggregate differences in abundance of loss of function, missense, and synonymous variants, respectively.

Results: Three non-independent sets of 10, 117, and 359 genes were queried against control gene sets of 139, 1522, and 3360 matched genes, respectively. Significant differences were not detected in the number of functional variants in the primary set of ethanol-metabolizing genes. In both the mouse expression and invertebrate sets, we observed an increased number of synonymous variants in GOI over matched control genes. Post-hoc simulations showed the estimated effects sizes observed are unlikely to be under-estimated.

Conclusion: The proposed method demonstrates a computationally viable and statistically appropriate approach for genetic analysis of case-only data for hypothesized gene sets supported by empirical evidence.

Introduction

Alcohol use disorder (AUD) is a common, moderately heritable disorder with significant social and economic impact. Twin(1–11), family(12), and adoption studies(13–16) consistently show that genetic influences have a large impact on the risk for AUD and alcohol-related phenotypes, with twin-based heritability estimates of ~0.50(17) and SNP-based heritability estimates of ~0.056(18). In recent years, genome-wide association studies (GWAS) have successfully identified common single nucleotide variants (cSNV) robustly associated with alcohol consumption(19), the Alcohol Use Disorders Identification Test (AUDIT)(20), and problematic alcohol use(18,21). Many of these identified cSNVs impact genes encoding the alcohol metabolizing enzymes such as the cluster of alcohol dehydrogenase (*ADH*) genes on chromosome 4. Additionally, there is accumulated evidence that variation in *CYP*, *CAT,* and *ALDH* genes are also involved in alcohol metabolism and AUD risk. *CYP2E1* expression is induced by chronic alcohol consumption and is thought to contribute to ethanol metabolism in the brain where *ADH* activity is limited(22). *CAT* is also widely expressed in the brain and a number of studies suggest that polymorphisms in the *CAT* gene are involved in the level of response to ethanol and alcohol dependence (AD) and abuse(23,24). *CYP2E1* and *CAT* are considered part of the canonical set of genes contributing to oxidative metabolism of ethanol(25).

While the effect of variants on *ADH*, *CYP2E1*, and *CAT* genes are more subtle, the well documented *ALDH2*2* (rs671) loss of function (LOF) allele shows only 20-40% of wild type enzymatic activity in heterozygote carriers due to the homo-tetrameric structure of mature *ALDH2*. This variant is common in individuals of East Asian ancestry, but largely absent in other populations(26) and is associated with lower rates of alcohol abuse/dependence(27) because the

reduced enzymatic activity leads to accumulation of acetaldehyde and unpleasant symptoms such as excessive flushing in carriers.

Although significant progress has been made in cSNV identification in AUD and alcohol-related phenotypes, rare or intermediate frequency single nucleotide variant (rSNV) investigations are largely limited by sample power. Findings from the 1000 Genomes Project suggest that rare functional variation is frequent in the genome with approximately 400 premature stop, splice-site disrupting and frame-shift alleles affecting 250-300 genes per individual genome(28) and display strong population specificity(29). Sequencing studies in psychiatric disorders suggest that rare functional variation is an important element of risk for intellectual disability(30), autism spectrum disorders(31), and schizophrenia(32). While rare functional variants have not yet been widely studied through exome sequencing in AUD, early evidence points to their significant contribution to the genetic architecture of alcohol use phenotypes(18,20,33,34). Furthermore, studies across psychiatric phenotypes such as autism spectrum disorders(31), and schizophrenia(32) show an excess rate of rSNV in genes identified from common variant GWAS signals in cases compared to controls, suggesting that there is a convergence between cSNV and rSNV signals and disease risk is likely influenced by multiple alleles of varying frequencies in the same loci(35).

As a complement to genetic studies in humans, mice and invertebrate model organisms can also facilitate the identification of genetic mechanisms or orthologous genes that influence AUD in humans. Studies in mice identified genes showing altered expression in prefrontal cortex (PFC) after intraperitoneal injection of 1.8 g/kg of ethanol versus saline control, and identified a subset of these as hub genes defined by both high connectivity and high centrality in co-expression networks(36). Introduction of mutations or knockdown strategies in invertebrate

models such as *D. melanogaster* or *C. elegans* can also identify genes involved in behavioral response to ethanol(37–39).

In this study, we sought to investigate exome variation in a sample of 190 severely affected alcohol dependence (AD) cases from the island of Ireland using a novel case-only analysis framework. Because of previous evidence of the protective effects of LOF alleles in ethanol metabolizing genes, we hypothesized that individuals with severe AD would show less functional coding variation in these genes compared to control genes with similar attributes. Furthermore, we extended this work to test sets of genes identified in model organisms and also hypothesized that genes robustly shown to impact alcohol-related outcomes in mice and invertebrates would show divergent patterns of exome variation in comparison to control genes with similar attributes. For each hypothesis, we sought to compare the numbers of LOF, missense (MIS), and synonymous (SYN) variants between genes of interest and a matched set of control genes. Given the absence of implemented methods to test our hypotheses in a case-only framework with related subjects, we sought to develop a novel framework to address these challenges. As an alternative to comparing aggregate exome variation between cases and controls, we matched genes of interest to control genes with similar attributes using gnomAD(40) as an external source of independent information, and unbiased comparisons across sets were made in a case-only analysis framework. Since comparison sets would be derived from within an individual's genome, they would not be subject to any potential inflation from stratification due to population structure or other sources of bias. We present this framework as a complementary method to case-control association designs that can be performed in samples without matched controls and provides a rigorous framework for hypothesis testing where robust prior sources of evidence are available.

Materials and Methods

Sample description: The Irish Affected Sib-Pair Study of Alcohol Dependence (IASPSAD) sample(8) was collected from 1998-2002 in treatment facilities and hospitals in the Republic of Ireland and Northern Ireland. Probands with all four grandparents born in Ireland or Britain were ascertained for a diagnosis of DSM-IV AD(41) with one or more affected siblings. Lifetime history of AD was assessed using a modification of the Semi-Structured Assessment for the Genetics of Alcoholism (SSAGA) version 11(42) which permits evaluation of International Classification of Disease (ICD)-10, Feighner(43), RDC(44), DSM-III-R(45) 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 Structured Clinical Interview for DSM (SCID)(46), the CAGE Assessment (Cutting Down, Annoyance by Criticism, Guilty Feeling, and Eye Openers)(47), and Fast Alcohol Screening Test (FAST) items developed to screen for drinking problems(48).

Exome capture and sequencing: Exome capture was performed using the Agilent SureSelect V5 71Mb exome + untranslated regions 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, with the first (n=57) and second (n=76) batches via 2x90 and the third batch (n=57) via 2x100 paired end sequencing.

Variant calling: Sequence data was processed and called according to GATK3 best practices and summarized for quality control using FastQC (v 0.11.4). Sequence read alignment was performed using BWA-MEM (v 0.7.12) to hs37d5 reference genome, followed by

reordering, duplicate marking and insertion deletion realignment with Picard (v 2.0.1). Variant calling was done using HaplotypeCaller, and variant quality score recalibration was carried out using Variant Quality Score Recalibration (VQSR) in GATK (v 3.5).

Annotation: SnpEff (v 4.3, database GRCh37.75) was used to obtain gene annotations and included only the transcripts found in the gnomAD release 2.1.1 gene constraints.

Description of candidate genes of interest: Selection of genes of interest (GOI) was carried out in collaboration with investigators from the Virginia Commonwealth University Alcohol Research Center (VCU-ARC), focused on cross-species discovery and functional interpretation of genes involved in AUD and related phenotypes. Three sets of alcohol related GOI were constructed for this analysis which crossed three taxonomic categories including human, mouse, and invertebrates. The first GOI set contains 11 ethanol metabolizing genes whose products are known to be involved with ethanol metabolism in humans. These included the *ADH* genes (n=7), *ALDH1A1, ALDH2*, *CAT*, and *CYP2E1* (Table 1). The second GOI set contains 109 hub genes with both high connectivity and high centrality in co-expression networks that show altered expression in mouse PFC 4 hours after intraperitoneal injection of 1.8 g/kg of ethanol versus saline control(36) (Supplemental Table S1). The third GOI set contains 358 genes for which manipulation in invertebrate model organisms results in altered ethanol response phenotypes(49) (Supplemental Table S2).

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

Gene	Transcript	Present in gnomAD	Present in the invertebrate GOI set	Present in the mouse brain expression GOI 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

Annotations for gene clustering: In order to create matched sets of genes for our case-only analytic approach, we utilized gene-level annotation information from the gnomAD database (v 2.11)(40). The seven gnomAD annotations used to cluster all genes included: (1-3) the ratios of observed to expected (O/E) counts for each variant class (LOF, MIS, and SYN), (4-6) a z-score for each O/E ratio, and (7) the probability of loss of function intolerance (pLI) score. In addition to these metrics from gnomAD, genes were annotated for clustering with metrics for (8) genomic length, (9) transcript length, and (10) 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 to warrant dropping from clustering.

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

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

Gene sets for analysis: Given that gnomAD metrics were needed to annotate the GOI, we had to drop 1 gene from the human ethanol metabolizing set because it did not appear in gnomAD (Table 1), 1 gene from the mouse hub gene set (Supplemental Table S1), and 5 genes from the ethanol behavioral response in invertebrates set (Supplemental Table S2). The final GOI sets contained 10, 108, and 353 genes in the human, mouse, and invertebrate sets, respectively, for inclusion in clustering and subsequent analyses. For hypothesis testing, we considered 3 testing sets: Set 1 contained only the 10 GOI for human ethanol metabolization, Set 2 contained Set 1, plus the addition of the 108 mouse hub genes. One gene appeared in both the human and mouse sets, therefore Set 2 contained a total of 117 genes. Set 3 contained Set 2, plus the addition of the ethanol behavior response in invertebrates set. One hundred and eleven of the invertebrate set genes appeared in Set 2, therefore Set 3 contained a total of 359 genes. Figure 1 shows the flow chart of filtering and sample sizes (panel a), as well as the Venn diagram (panel b) illustrating overlap between the three GOI sets, after removing genes absent from gnomAD.

Caption, Figure 1: (a) Flow diagram with gene counts for the human GOI set, the mouse expression set, and the invertebrate set, through filtering and combination for hypothesis testing sets, and (b) Venn diagram illustrating the overlap of genes in the sets of interest, after filtering for presence in gnomAD. (*Figure created with BioRender.com*)

Clustering: Multivariable single linkage agglomerative clustering of all canonical gnomAD genes was carried out using the *hclust* package (method = "single") in R (v 3.6.0). The clustering algorithm was used to identify genes similar to each GOI based on the gene-level variables described above to create a control set of genes against which the GOI could be compared in a regression framework. This procedure utilizes information from gnomAD only, separate from the IASPSAD sample exomes. Clustering was performed on 10 gene metrics using a dissimilarity metric of $1 - |cor(X)|$, where X is the $q \times p$ matrix of the $q = 10$ normalized metrics of each gene ($p = 19{,}108$). Figure 2 provides an example of gene clusters plotted according to three of these gnomAD gene metrics (SYN z-score, transcript length, and pLI). The data shown in this figure represent 4 gene clusters (shown in different colors) and a set of genes chosen at random (in gray), with ellipses highlighting the shape of the cluster. Matrix $cor(X)$ represents the full set of pairwise correlations for all 19,108 genes in the set.

Caption, Figure 2: Example gene clusters plotted to show three of the 10 clustering gene 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.

The hierarchical clustering was carried out 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 were joined together. We then pruned the final trees for all three GOI 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 were termed Clusters of Interest (COI). After a cut height is chosen, all other clusters not containing a GOI were pruned from the tree.

Association testing: We used logistic regression to compare observed counts of LOF, MIS, and SYN variants aggregated from the IASPSAD sample exome data between GOI and the matched control genes. In this framework, instead of human subjects, the dependent variable is

the gene which is either a GOI for a given hypothesis or not and coded 1 or 0, respectively. Therefore, 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 aggregated across the subjects in the IASPSAD sample. In other words, this framework tests whether or not the number of LOF, MIS, and/or SYN variants contributes to the probability that a given gene contributes to a hypothesized gene set such as alcohol metabolism. For each of the three GOI sets, the model is constructed as follows:

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

where GOI represents case/control status of the genes as described above for each set, and LOF_{Alc}, $MIS_{Alc}$and SYN_{Alc} represent the observed counts of LOF, MIS, and SYN variants in that GOI, 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.

Simulations: To empirically demonstrate the utility of this approach, we applied the method to a series of simulated datasets and assessed their performance. These simulations were conducted post-hoc and designed to follow distributional patterns and estimated effects observed in the real data model for the overarching purpose of increasing confidence in the real data results. Full details of the simulation parameters are given in the Supplemental Section S1 and Supplemental Table S3, but in brief, the simulations replace the observed LOF, MIS, and SYN variant counts for all 19,108 genes from the sample data with randomly generated data according to the following steps:

1. Set the distributions of the simulated data mimicking the observed distributions of LOF, MIS, and SYN variant counts in the alcohol sample data according to negative binomial distributions
2. Simulate the random LOF, MIS, and SYN data using the *SimCorrMix* package, according to the correlation structure in the observed alcohol sample data, such that:

$$\begin{bmatrix} 1 & 0.166 & 0.257 \\ 0.166 & 1 & 0.688 \\ 0.257 & 0.688 & 1 \end{bmatrix}$$

3. Generate the GOI probabilities using a logit model with beta effects for the intercept, LOF, MIS, and SYN values set at (-3, 0.75, 0.05, 0.1), (-5, 1.25, 0.1, 0.2), or (-7, 1.75, 0.2, 0.3). These values were chosen to broadly reflect the estimated effects observed in the data.
4. Generate random error according to a normal distribution with mean = 0 and standard deviation ranging from 0.55-1.45
5. Utilizing the clusters generated from gnomAD metrics, assign COIs
6. Apply the logistic regression models to each scenario, using the genes assigned to COIs as control genes

Each simulation scenario was iterated across 1000 random datasets. The performance of the approach was assessed by summarizing the mean LOF, MIS, and SYN and the standard error of those estimates in each simulation scenario.

Results

<u>WES variant call and filtering</u>

WES data for all 190 IASPSAD subjects passed quality control measurements, with mean sequencing depth across all samples at 60.6x (standard deviation 12.02), with 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 insertion/deletions (indels). For quality control, SNPs and indels were excluded if they fell into GATK VQSR tranches 99.0 or greater, indicating that 99% of the true variants present in the sample will be retained in the filtered set (38,503 variants removed), to avoid the rising rates of false positive variant calls at this and more inclusive thresholds. Variants with MAF ≥ 0.05 in gnomAD European (non-Finnish), non-cancer samples were excluded which resulted in a final set of 652,428 variants (91,780 removed) with 2,328 LOF, 31,015 MIS, and 46,046 SYN variants left for exome analysis.

Description of GOIs

Figure 3 shows the overview of the analysis framework. The correlation between the LOF, MIS, and SYN counts for all genes from gnomAD and the observed counts from the IASPSAD subjects were 0.26, 0.80, and 0.76, respectively. This indicates that at least for MIS and SYN variants, there is sufficiently strong evidence that the gnomAD database information can be utilized to group genes for a case-only within-sample analysis. Table 2 shows the correlation between the 10 gene metrics used for identifying control genes from the gnomAD database, indicating that these annotations are measuring disparate genomic features, which supports the inclusion of all ten metrics in the multivariate clustering. Table 3 describes the resulting trees cut at various heights in the hierarchical clustering analysis. We chose to cut all three GOI set 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 all genes in the largest COIs. Figure 4 provides a visual representation of one of the branches, representing one cluster of the pruned tree with the GOI (*ADH4*) 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 all genes included in the COIs (for tight clusters with high similarity across metrics). Figure 5 illustrates the relationship between GOI set size, median cluster size, and proportion of the exome included in the COI.

Table 3: For each of the three GOI test sets, median cluster size (Med. Size), number of total genes, combined across all clusters of interest (No.Genes), and proportion of the exome retained in the clusters of interest (Prop. Gen) for hierarchical clustering trees cut at various heights.

	Primary human ethanol metabolizing GOI set			**Including mouse brain expression set GOI set**			**Including invertebrate GOI set**		
Height	**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	359	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

Lengend:
Med. Size: Median cluster size
No. Genes: Number of genes across all clusters of interest
Prop. Gen.: Proportion of the exome included

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

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

Caption, Figure 5: Relationship between median number of genes per cluster of interest and proportion of the exome included in the cluster of interest.

Simulation results

Full simulation results appear in Supplemental Table S4. In summary, these simulations demonstrated the validity of the logistic regression framework to answer the directed hypotheses regarding counts of LOF, MIS, and SYN variants in GOIs, as compared to matched control genes. The approach was able to accurately identify significant LOF, MIS, and SYN effects where they were simulated to exist in the data. Point estimates were somewhat overestimated in most cases, but for MIS and SYN variants, the true effect fell within 2 standard deviations of the estimated values. Figure 6 shows the estimates, plus 2 standard deviations, as well as the true simulated effect for the LOF (panel a), MIS (panel b), and SYN (panel c) variants across 9 simulation scenarios.

Caption, Figure 6: Simulation estimates of parameter values with standard deviations and true simulated values. One standard deviation from the point estimate is shown in the same color as the plotted point with one additional standard deviation shown in gray. The starred point shows the true, simulated point value. Estimates shown for (a) LOF, (b) MIS, and (c) SYN variants.

Association analysis

For each of the three GOI sets tested, logistic regression was performed where all GOIs served as "cases" and all other genes in the associated COIs served as "controls". The total set size for the three GOI sets were 149 for the primary human alcohol metabolizing GOI, 1,639 for the mouse PFC brain expression hub GOI, and 3,719 for altered ethanol response invertebrate GOI, respectively. Supplemental Tables S5-S7 list the control genes for each model. Estimates for the logistic regression models are shown in Table 4. Using an alpha cutoff of 0.05, the human

ethanol metabolizing GOI did not show a statistically significant difference between the number of LOF, MIS, or SYN variants in GOI compared to control genes. For the mouse PFC brain expression GOI, a significant difference in counts of SYN variants with odds ratio (OR) of 1.21 (*p*-value = 0.0006) was observed, indicating that the addition of a single synonymous variant conferred odds of a gene being a GOI 1.21 times higher than the odds of being a control gene. Finally, for the altered ethanol response invertebrate GOI, a significant difference in the number of SYN variants with an OR of 1.06 was observed (*p*-value 0.0169).

Table 4: Logistic regression results.

Primary Human Ethanol Metabolizing GOI set 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*

Mouse Brain Expression GOI 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 GOI Set N=3,719 (359 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*

Discussion

In this study, we sought to perform a hypothesis driven case-only exome analysis in severe AD cases using three GOI sets with prior evidence for ethanol metabolism in humans, or implicated in alcohol phenotypes in model organism studies in mouse and invertebrates. While we observed differences in the number of SYN variants for mouse and invertebrate GOIs, the results do not support the hypothesis that ethanol metabolizing genes, in particular those directly involved in humans, are largely depleted for LOF variants in severe AD cases. While MIS or LOF variants have more direct effect on gene products, SYN mutations can impact the speed of messenger RNA (mRNA) translation processes by changing the codon to one with different transfer RNA (tRNA) abundance or by altering folding and stability of mRNA, producing secondary structures that are less efficiently recognized for mRNA processing(50). In particular, recent oncological findings suggest that SYN mutations might play a role in the development of cancer by altering codon optimization and translational velocity. However, SYN mutations are still largely considered to be functionally silent.

Genetic studies of cardiac disease(51–53), obesity(54), Alzheimer disease(55), and non-alcoholic fatty liver disease(56) consistently show that MIS and LOF alleles are common in the human genome, alter disease risk, and in some cases are protective(57,58). Such protective effects of functional variants are well documented for alcohol-related phenotypes. Most recently, rs75967634 in *ADH1B* was also found to be associated with problematic alcohol use in individuals of European ancestry(21). Although of low frequency outside of east Asia,

*ADH1B*2* (rs1229984) is associated with both AUD diagnosis(34,59) and the problem drinking component of the AUDIT(20) in European populations. Together, these results suggest that although effects of variants in ethanol metabolizing enzymes are well documented across different populations, our findings based on our modestly sized sample of severe AD cases does not provide evidence in support of a significant depletion of LOF or MIS variation in these genes.

In this study, we presented a framework for analyzing case-only exome variation data in the absence of appropriate control subjects. It is not uncommon in genetic studies to obtain molecular data on a set of subjects who are all positive for a dichotomous phenotype. Therefore, methods have been developed for intentional case-only study designs, in particular for gene-by-environment interaction studies(60) or polygenic risk scores(61). However, current case-only analysis frameworks generally consider an environmental exposure as the dichotomous outcome being tested for association with genetic variation within a sample of all cases, and offer no information regarding phenotypic variation due to the primary disease state. Furthermore, when controls have not been genotyped or sequenced alongside cases in the same study, publicly available datasets may also provide appropriate population controls with careful ancestral matching. Additionally, methods such as the Robust Variance Score Statistic (RVS) method(62), or the burden test implemented in the TASER software(63) which is an extension of RVS method that offers improved adjustment for sample differences in case/control analyses that can be used in such cases. Other methods such as iECAT(64) does not require individual-level genotype data from population controls, but rather conducts an adjusted association testing using only allele counts. However, due to differences in sampling techniques or technology, filtering criteria, and variant calling, using these methods may not always be feasible. Importantly, none

of these methods offers correction or adjustment for related samples such as the IASPSAD sample analyzed in this study. We therefore sought to model case/control status of individual genes within a case only sample of severely affected subjects with AD cases using a novel design. In contrast to studies that use external control subjects, our proposed framework used individual cases' exome data to assign case status to GOI and control status to matched genes as described in the methods section. This strategy removes the need for careful correction or adjustment for subtle population structure in exome studies, and further leverages external information from large sequencing studies such as gnomAD to ensure gene matching is robust by using a multivariate agnostic hierarchical clustering approach. Therefore, this methodological framework represents an interpretable, straight-forward, and computationally affordable approach that is easily implemented using existing software and tools. We additionally note that this framework can also be extended to model quantitative measures within cases, such as maximum number of drinks per day with minimal adjustment.

Finally, our simulation studies demonstrated that while the approach has sufficient power to detect real effects, the estimates of these effects may show some inflation under certain distributional conditions. It is recommended that future applications of this approach utilize some simulations to determine empirically expectations of power and identify potential sources of bias *a priori*. Conservatively, the simulation results indicate that the significant estimate of the effect of synonymous variants in the mouse expression set is unlikely attributable to positive bias alone. Furthermore, while such exploration was beyond the scope of the current study, different experimental designs may warrant further simulations to empirically guide the choice of tree cutpoints to balance cluster size and proportion of the exome included in the COIs.

The findings presented in this study should be interpreted in the context of three important limitations. First, our modest sample size of 190 affected subjects with exome data is limited, and the tests conducted on each GOI set may not be sufficiently powered to detect significant differences. Second, while we empirically modeled the median number of matched genes in each cluster against the proportion of the genome in the final clustered set to choose an appropriate pruning parameter, simulations to evaluate the impact of various parameter choices such as tree pruning height, or proportion of the genome included in gene clusters could provide better benchmarking for selecting these parameters in future studies. Third, as more exome data from ancestrally diverse populations become available, future analyses could attempt to replicate these findings in larger, more diverse samples to improve the generalizability of these findings and the utility of our case-only exome analysis framework in other populations and phenotypes.

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

References:

1. Heath AC, Bucholz KK, Madden PA, Dinwiddie SH, Slutske WS, Bierut LJ, et al. Genetic and environmental contributions to alcohol dependence risk in a national twin sample: consistency of findings in women and men. Psychol Med. 1997 Nov;27(6):1381–96.

2. Hrubec Z, Omenn GS. Evidence of Genetic Predisposition to Alcoholic Cirrhosis and Psychosis: Twin Concordances for Alcoholism and Its Biological End Points by Zygosity among Male Veterans. Alcohol Clin Exp Res. 03/1981;5(2):207–15.

3. Kendler KS, Heath AC, Neale MC, Kessler RC, Eaves LJ. A population-based twin study of alcoholism in women. JAMA. 1992 Oct 14;268(14):1877–82.

4. Kendler KS, Prescott CA, Neale MC, Pedersen NL. Temperance board registration for alcohol abuse in a national sample of Swedish male twins, born 1902 to 1949. Arch Gen Psychiatry. 1997 Feb;54(2):178–84.

5. McGue M, Pickens RW, Svikis DS. Sex and age effects on the inheritance of alcohol problems: a twin study. J Abnorm Psychol. 1992 Feb;101(1):3–17.

6. Pickens RW. Heterogeneity in the Inheritance of Alcoholism: A Study of Male and Female Twins. Arch Gen Psychiatry. 1991 Jan 1;48(1):19.

7. Prescott CA, Aggen SH, Kendler KS. Sex differences in the sources of genetic liability to alcohol abuse and dependence in a population-based sample of U.S. twins. Alcohol Clin Exp Res. 1999 Jul;23(7):1136–44.

8. Prescott CA, Caldwell CB, Carey G, Vogler GP, Trumbetta SL, Gottesman II. The Washington University Twin Study of alcoholism. Am J Med Genet. 2005 Apr 5;134B(1):48–55.

9. Reed T, Page WF, Viken RJ, Christian JC. Genetic predisposition to organ-specific endpoints of alcoholism. Alcohol Clin Exp Res. 1996 Dec;20(9):1528–33.

10. Romanov K, Kaprio J, Rose RJ, Koskenvuo M. Genetics of alcoholism: effects of migration on concordance rates among male twins. Alcohol Alcohol Suppl. 1991;1:137–40.

11. True WR, Heath AC, Bucholz K, Slutske W, Romeis JC, Scherrer JF, et al. Models of treatment seeking for alcoholism: the role of genes and environment. Alcohol Clin Exp Res. 1996 Dec;20(9):1577–81.

12. Cotton NS. The familial incidence of alcoholism: a review. J Stud Alcohol. 01/1979;40(1):89–116.

13. Cadoret RJ, Troughton E, O'Gorman TW. Genetic and environmental factors in alcohol abuse and antisocial personality. J Stud Alcohol. 01/1987;48(1):1–8.

14. Cloninger CR. Inheritance of Alcohol Abuse: Cross-Fostering Analysis of Adopted Men. Arch Gen Psychiatry. 1981 Aug 1;38(8):861.

15. Goodwin DW. Alcohol Problems in Adoptees Raised Apart From Alcoholic Biological Parents. Arch Gen Psychiatry. 1973 Feb 1;28(2):238.

16. Sigvardsson S. Replication of the Stockholm Adoption Study of Alcoholism: Confirmatory Cross-Fostering Analysis. Arch Gen Psychiatry. 1996 Aug 1;53(8):681.

17. Verhulst B, Neale MC, Kendler KS. The heritability of alcohol use disorders: a meta-analysis of twin and adoption studies. Psychol Med. 2015 Apr;45(5):1061–72.

18. Kranzler HR, Zhou H, Kember RL, Vickers Smith R, Justice AC, Damrauer S, et al. Genome-wide association study of alcohol consumption and use disorder in 274,424 individuals from multiple populations. Nat Commun. 2019 Apr 2;10(1):1499.

19. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. Nat Genet. 2019 Feb;51(2):237–44.

20. Sanchez-Roige S, Fontanillas P, Elson SL, The 23andMe Research Team, Gray JC, de Wit H, et al. Genome-wide association study of alcohol use disorder identification test (AUDIT) scores in 20 328 research participants of European ancestry: GWAS of AUDIT. Addict Biol. 01/2019;24(1):121–31.

21. Zhou H, Sealock JM, Sanchez-Roige S, Clarke T-K, Levey DF, Cheng Z, et al. Genome-wide meta-analysis of problematic alcohol use in 435,563 individuals yields insights into biology and relationships with other traits. Nat Neurosci. 2020 Jul;23(7):809–18.

22. Zimatkin SM, Deitrich RA. Ethanol metabolism in the brain. Addict Biol. 1997 Oct;2(4):387–400.

23. Hu X, Oroszi G, Chun J, Smith TL, Goldman D, Schuckit MA. An expanded evaluation of the relationship of four alleles to the level of response to alcohol and the alcoholism risk. Alcohol Clin Exp Res. 2005 Jan;29(1):8–16.

24. Plemenitas A, Kastelic M, Porcelli S, Serretti A, Rus Makovec M, Kores Plesnicar B, et al. Genetic variability in CYP2E1 and catalase gene among currently and formerly alcohol-dependent male subjects. Alcohol Alcohol. 2015 Mar;50(2):140–5.

25. Zakhari S. Overview: how is alcohol metabolized by the body? Alcohol Res Health. 2006;29(4):245–54.

26. Goedde HW, Agarwal DP, Fritze G, Meier-Tackmann D, Singh S, Beckmann G, et al. Distribution of ADH2 and ALDH2 genotypes in different populations. Hum Genet. 1992 Jan;88(3):344–6.

27. Thomasson HR, Crabb DW, Edenberg HJ, Li TK, Hwu HG, Chen CC, et al. Low frequency of the ADH2*2 allele among Atayal natives of Taiwan with alcohol use disorders. Alcohol Clin Exp Res. 1994 Jun;18(3):640–3.

28. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

29. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science. 2012 Feb 17;335(6070):823–8.

30. Lelieveld SH, Reijnders MRF, Pfundt R, Yntema HG, Kamsteeg E-J, de Vries P, et al. Meta-analysis of 2,104 trios provides support for 10 new genes for intellectual disability. Nat Neurosci. 2016 Sep;19(9):1194–6.

31. Antaki D, Guevara J, Maihofer AX, Klein M, Gujral M, Grove J, et al. A phenotypic spectrum of autism is attributable to the combined effects of rare variants, polygenic risk and sex. Nat Genet. 2022 Sep;54(9):1284–92.

32. Singh T, Poterba T, Curtis D, Akil H, Al Eissa M, Barchas JD, et al. Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature. 2022 Apr;604(7906):509–16.

33. Brazel DM, Jiang Y, Hughey JM, Turcot V, Zhan X, Gong J, et al. Exome Chip Meta-analysis Fine Maps Causal Variants and Elucidates the Genetic Architecture of Rare Coding Variants in Smoking and Alcohol Use. Biol Psychiatry. 2019 Jun 1;85(11):946–55.

34. Walters RK, Polimanti R, Johnson EC, McClintick JN, Adams MJ, Adkins AE, et al. Transancestral GWAS of alcohol dependence reveals common genetic underpinnings with psychiatric disorders. Nat Neurosci. 2018 Dec;21(12):1656–69.

35. Panagiotou OA, Evangelou E, Ioannidis JPA. Genome-wide Significant Associations for Variants With Minor Allele Frequency of 5% or Less—An Overview: A HuGE Review. Am J Epidemiol. 2010 Oct 15;172(8):869–89.

36. Wolen AR, Phillips CA, Langston MA, Putman AH, Vorster PJ, Bruce NA, et al. Genetic Dissection of Acute Ethanol Responsive Gene Networks in Prefrontal Cortex: Functional and Mechanistic Implications. Palmer AA, editor. PLoS One. 2012 Apr 12;7(4):e33575.

37. Davies AG, Pierce-Shimomura JT, Kim H, VanHoven MK, Thiele TR, Bonci A, et al. A central role of the BK potassium channel in behavioral responses to ethanol in C. elegans. Cell. 2003 Dec 12;115(6):655–66.

38. Davies AG, Bettinger JC, Thiele TR, Judy ME, McIntire SL. Natural variation in the npr-1 gene modifies ethanol responses of wild strains of C. elegans. Neuron. 2004 Jun 10;42(5):731–43.

39. Bhandari P, Hill JS, Farris SP, Costin B, Martin I, Chan C-L, et al. Chloride intracellular channels modulate acute ethanol behaviors in Drosophila, Caenorhabditis elegans and mice. Genes Brain Behav. 2012 Jun;11(4):387–97.

40. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020 May;581(7809):434–43.

41. American Psychiatric Association, Task Force on Nomenclature and Statistics, American Psychiatric Association, Committee on Nomenclature and Statistics. Diagnostic and statistical manual of mental disorders. Washington, D.C.: American Psychiatric Association; 2000.

42. Bucholz KK, Cadoret R, Cloninger CR, Dinwiddie SH, Hesselbrock VM, Nurnberger JI, et al. A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA. J Stud Alcohol. 1994 Mar;55(2):149–58.

43. Feighner JP, Robins E, Guze SB, Woodruff RA, Winokur G, Munoz R. Diagnostic criteria for use in psychiatric research. Arch Gen Psychiatry. 1972 Jan;26(1):57–63.

44. Spitzer RL, Endicott J, Robins E. Research diagnostic criteria: rationale and reliability. Arch Gen Psychiatry. 1978 Jun;35(6):773–82.

45. American Psychiatric Association, American Psychiatric Association, editors. Diagnostic and statistical manual of mental disorders: DSM-III-R. 3rd ed., rev. Washington, DC: American Psychiatric Association; 1987. 567 p.

46. Spitzer RL, Williams JB, Gibbon M, First MB. The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description. Arch Gen Psychiatry. 1992 Aug;49(8):624–9.

47. Ewing JA. Detecting alcoholism. The CAGE questionnaire. JAMA. 1984 Oct 12;252(14):1905–7.

48. Hodgson R, Alwyn T, John B, Thom B, Smith A. The FAST Alcohol Screening Test. Alcohol Alcohol. 2002 Jan;37(1):61–6.

49. Grotewiel M, Bettinger JC. *Drosophila* and *Caenorhabditis elegans* as Discovery Platforms for Genes Involved in Human Alcohol Use Disorder. Alcohol Clin Exp Res. 08/2015;39(8):1292–311.

50. Chu D, Wei L. Parsing the synonymous mutations in the maize genome: isoaccepting mutations are more advantageous in regions with codon co-occurrence bias. BMC Plant Biol. 2019 Oct 14;19(1):422.

51. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004 Aug 6;305(5685):869–72.

52. Cohen JC, Boerwinkle E, Mosley TH Jr, Hobbs HH. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. N Engl J Med. 2006 Mar 23;354(12):1264–72.

53. Kotowski IK, Pertsemlidis A, Luke A, Cooper RS, Vega GL, Cohen JC, et al. A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. Am J Hum Genet. 2006 Mar;78(3):410–22.

54. Benzinou M, Creemers JWM, Choquet H, Lobbens S, Dina C, Durand E, et al. Common nonsynonymous variants in PCSK1 confer risk of obesity. Nat Genet. 2008 Aug;40(8):943–5.

55. Guerreiro R, Wojtas A, Bras J, Carrasquillo M, Rogaeva E, Majounie E, et al. TREM2 variants in Alzheimer's disease. N Engl J Med. 2013 Jan 10;368(2):117–27.

56. Romeo S, Kozlitina J, Xing C, Pertsemlidis A, Cox D, Pennacchio LA, et al. Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. Nat Genet. 2008 Dec;40(12):1461–5.

57. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science. 2009 Apr 17;324(5925):387–9.

58. Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, et al. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nat Genet. 2007 Apr;39(4):513–6.

59. Gelernter J, Kranzler HR, Sherva R, Almasy L, Koesterer R, Smith AH, et al. Genome-wide association study of alcohol dependence:significant findings in African- and European-Americans including novel risk loci. Mol Psychiatry. 2014 Jan;19(1):41–9.

60. Sahebi L, Dastgiri S, Ansarin K, Sahebi R, Mohammadi SA. Study Designs in Genetic Epidemiology. ISRN Genetics. 2013 May 12;2013:1–8.

61. Meisner A, Kundu P, Chatterjee N. Case-only analysis of gene-environment interactions using polygenic risk scores [Internet]. Genetics; 2019 Feb [cited 2019 May 23]. Available from: http://biorxiv.org/lookup/doi/10.1101/555300

62. Derkach A, Chiang T, Gong J, Addis L, Dobbins S, Tomlinson I, et al. Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. Bioinformatics. 2014 Aug 1;30(15):2179–88.

63. Hu Y-J, Liao P, Johnston HR, Allen AS, Satten GA. Testing Rare-Variant Association without Calling Genotypes Allows for Systematic Differences in Sequencing between Cases and Controls. Hoffmann TJ, editor. PLoS Genet. 2016 May 6;12(5):e1006040.

64. Lee S, Kim S, Fuchsberger C. Improving power for rare-variant tests by integrating external controls. Genet Epidemiol. 11/2017;41(7):610–9.

Supplemental Information Captions:

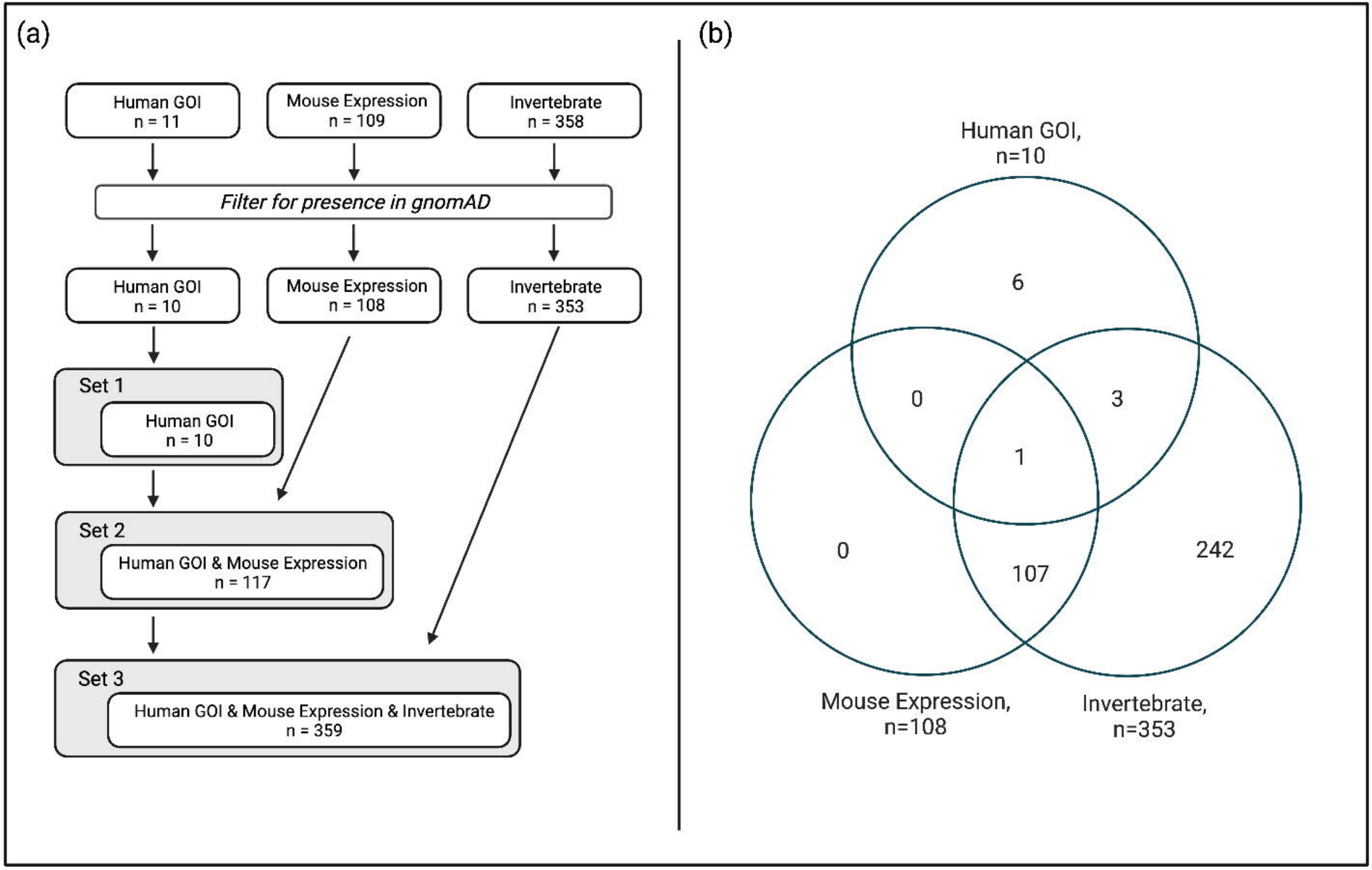
Supplemental Table S1: The genes in the expression set (n=109 genes), with annotation indicating which were present in gnomAD (108), which were present in the invertebrate set (108), and which were present in the human GOI set (1).

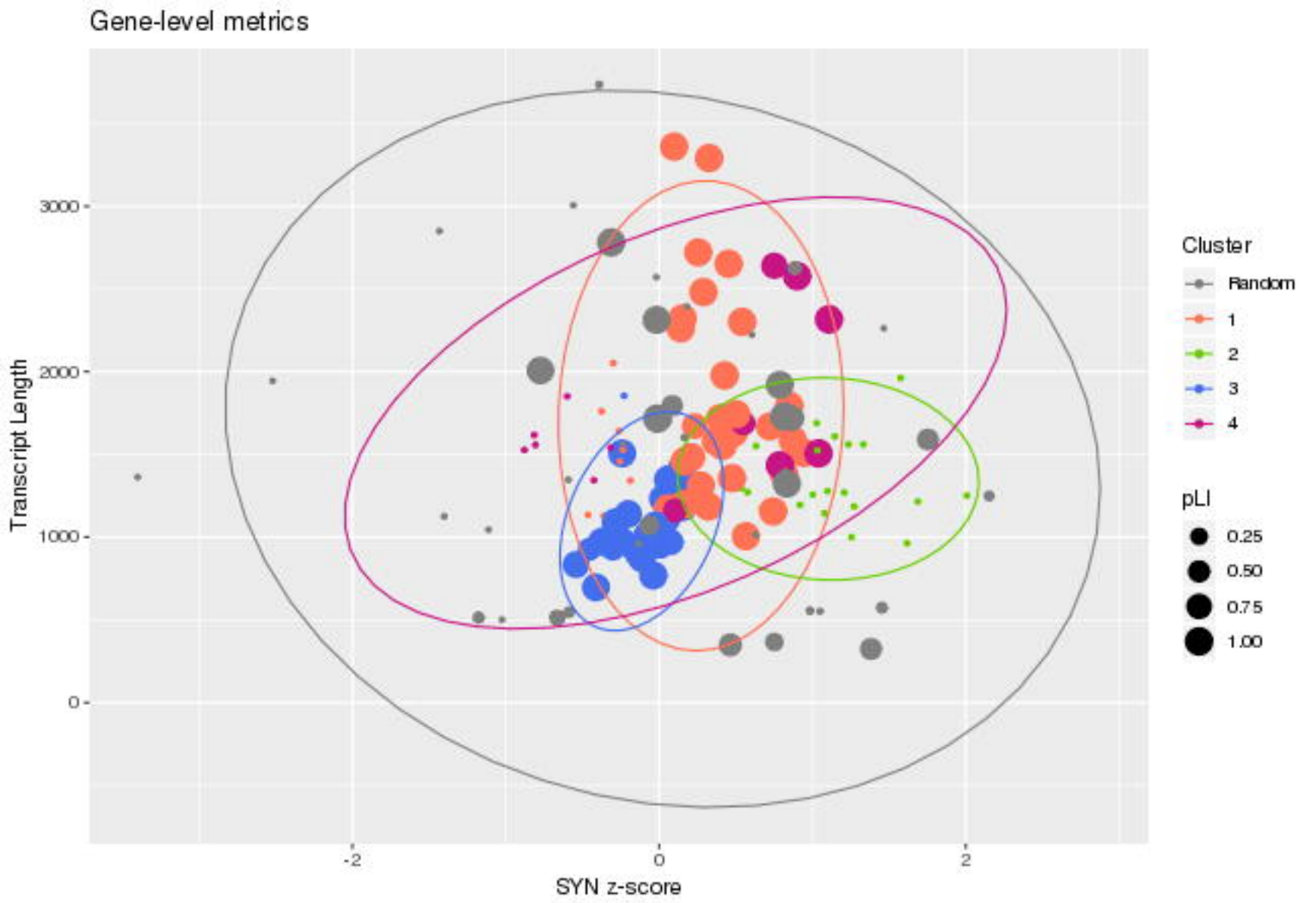
Supplemental Table S2: The list of invertebrate genes (n = 358 genes) with annotation indicating which were present in gnomAD (353), which were present in the expression set (108), and which were present in the human GOI set (11).

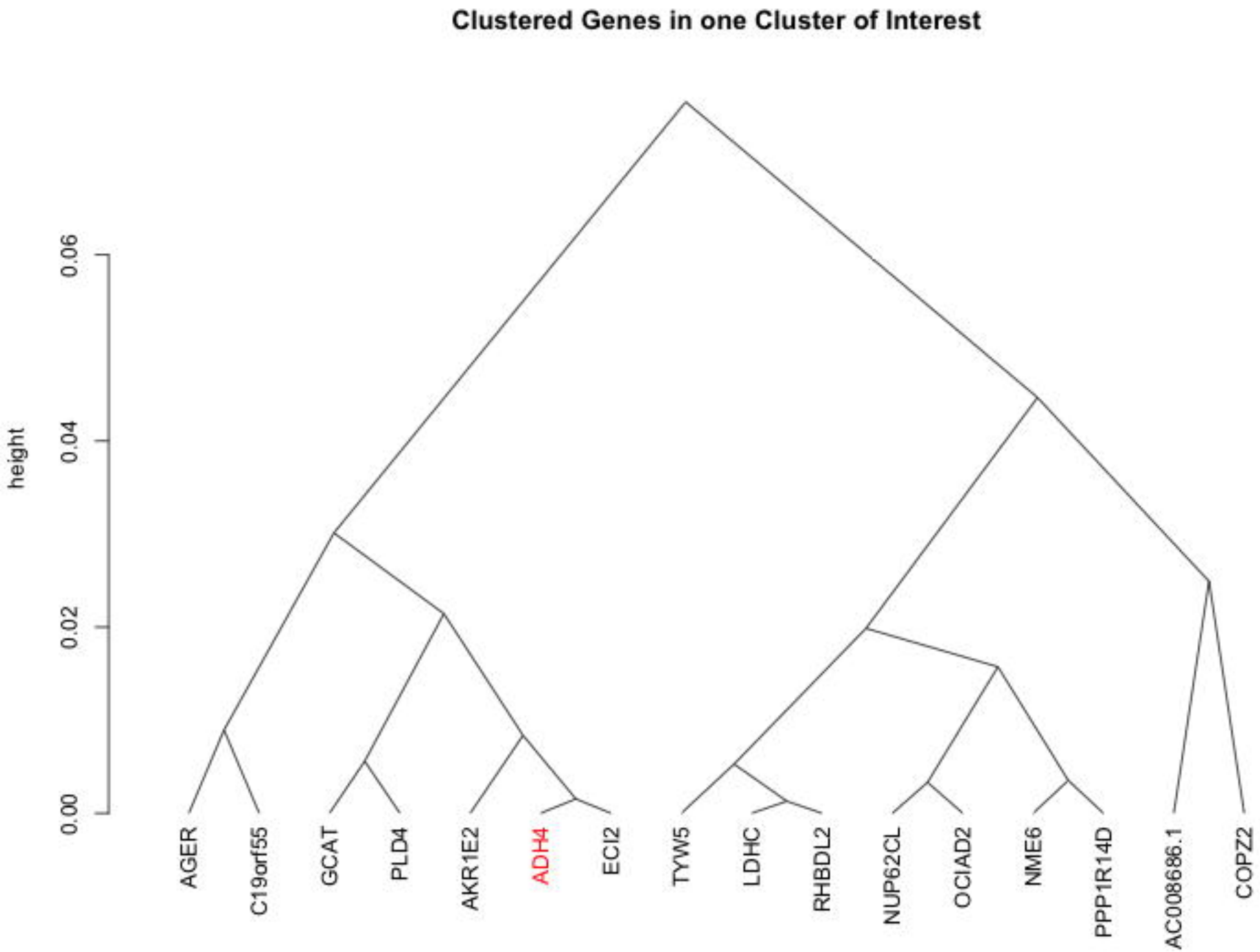
Supplemental Section S1: Simulation Details

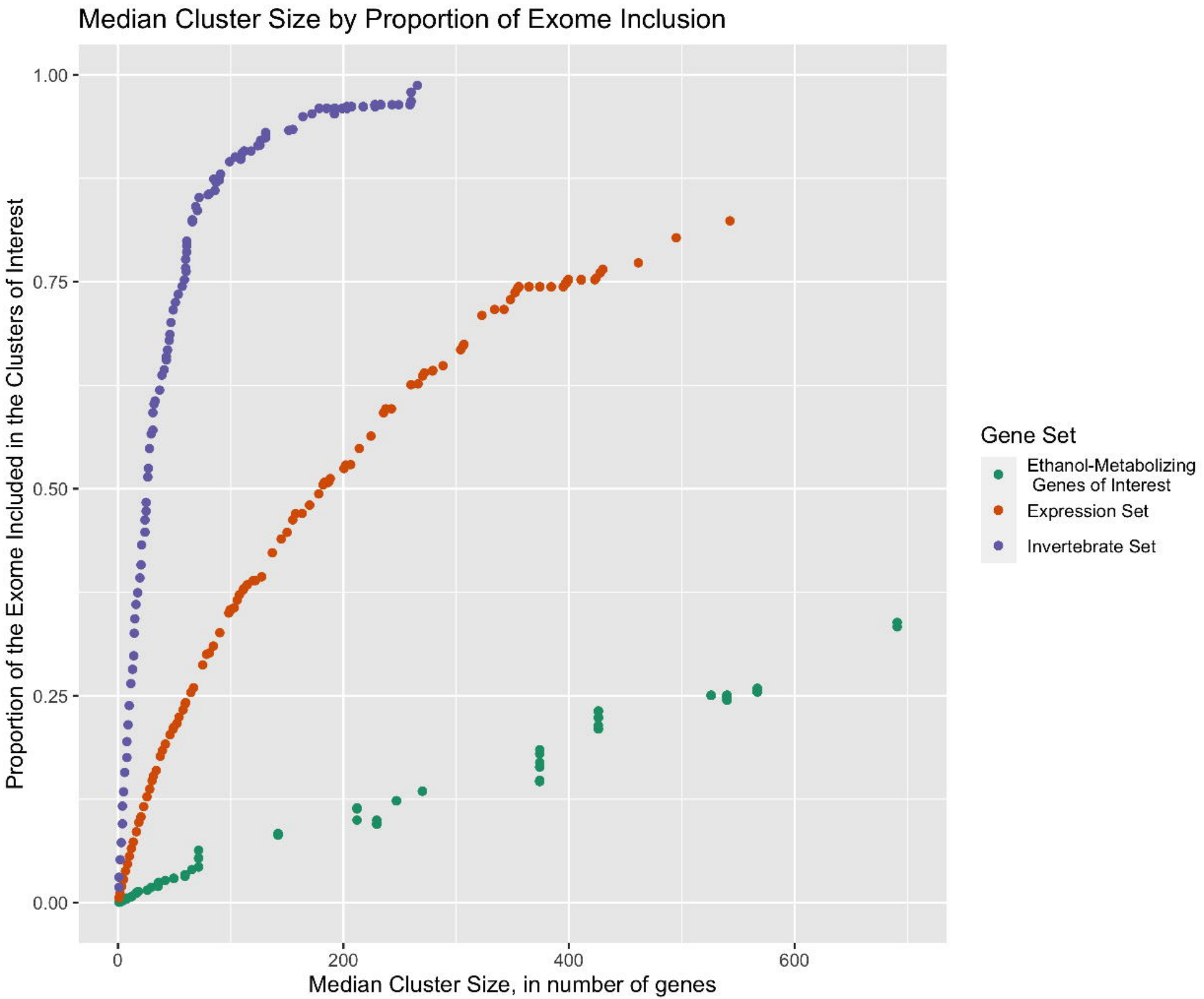
Supplemental Table S3: Parameters used to generate random data for the simulations.

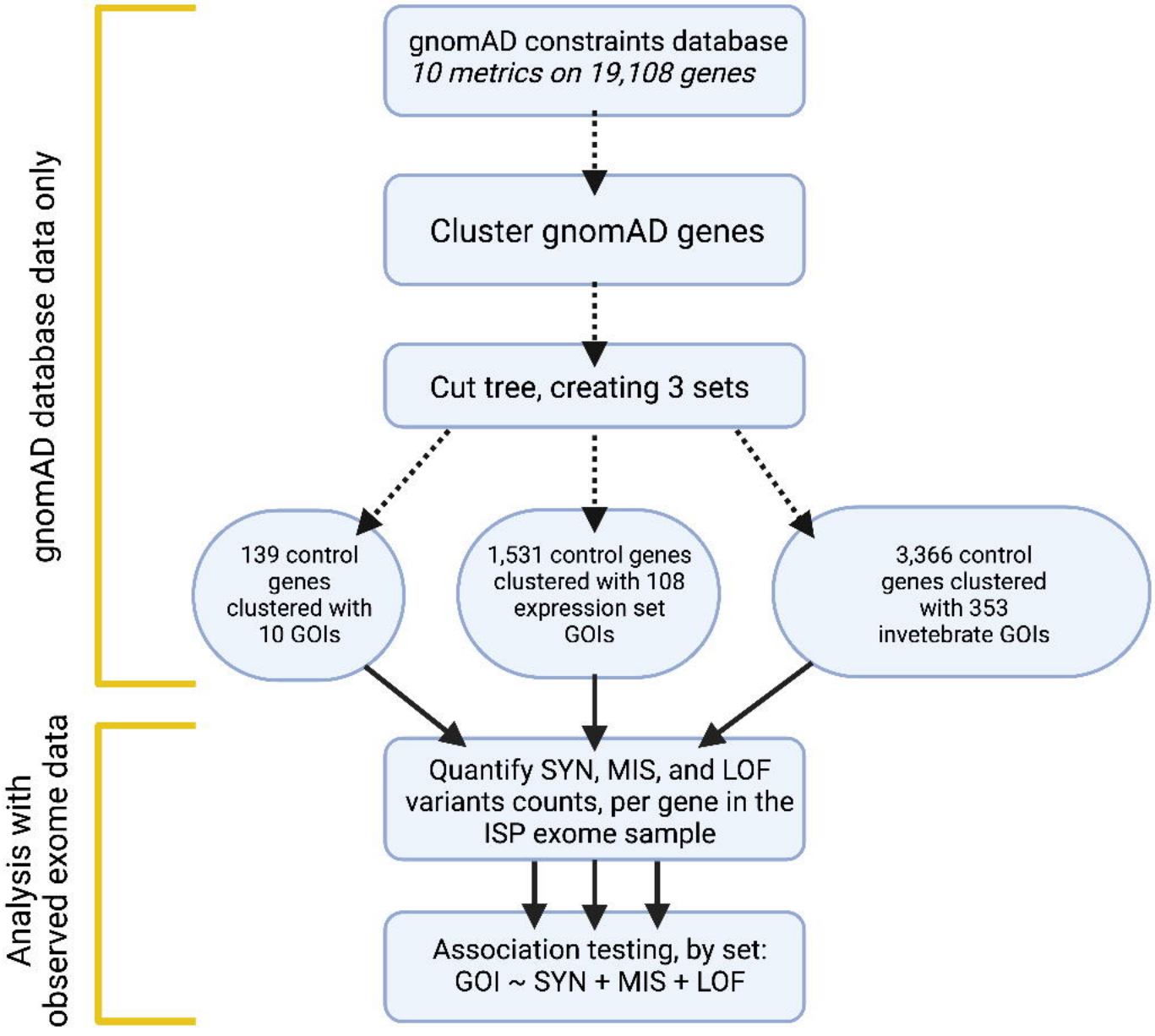
Supplemental Table S4: Parameter estimates from simulations.

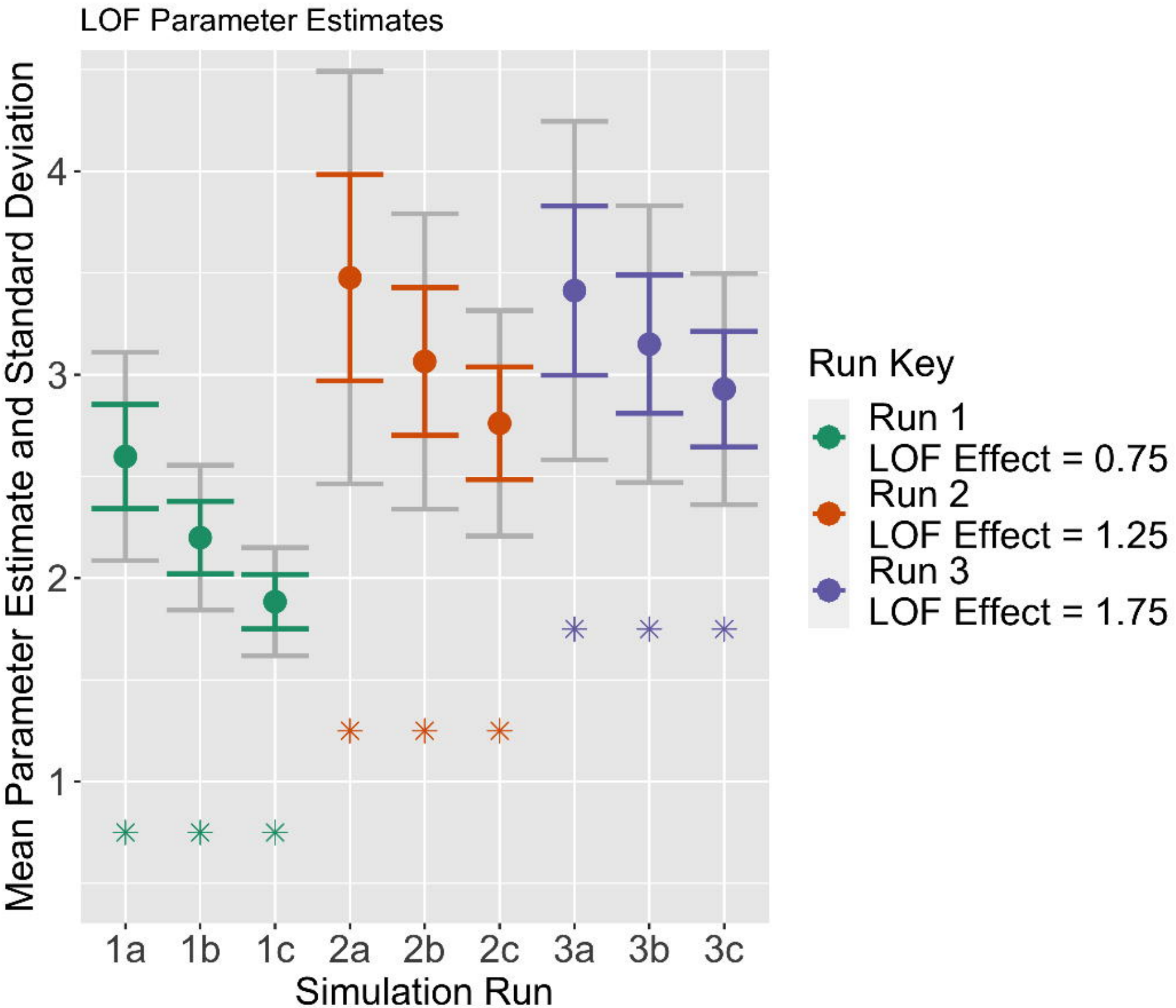

(a)
Human GOI
n = 11
Mouse Expression
n = 109
Invertebrate
n = 358
Filter for presence in gnomAD
Human GOI
n = 10
Mouse Expression
n = 108
Invertebrate
n = 353
Set 1
Human GOI
n = 10
Set 2
Human GOI & Mouse Expression
n = 117
Set 3
Human GOI & Mouse Expression & Invertebrate
n = 359
(b)
Human GOI,
n=10
6
0
3
1
0
107
242
Mouse Expression,
n=108
Invertebrate,
n=353

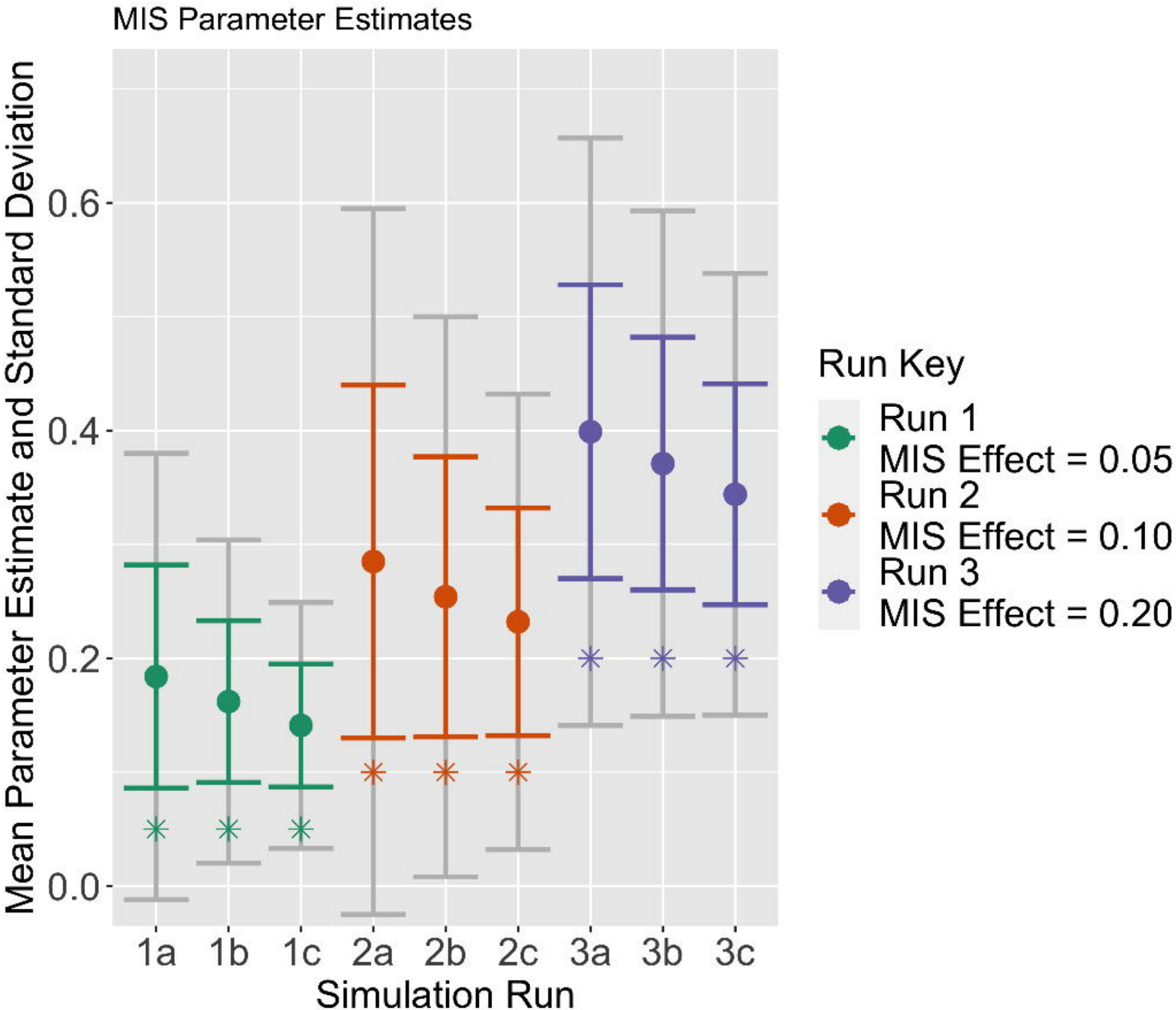

Gene-level metrics
Transcript Length
3000
2000
1000
0
-2
0
2
SYN z-score
Cluster
Random
1
2
3
4
pLI
0.25
0.50
0.75
1.00

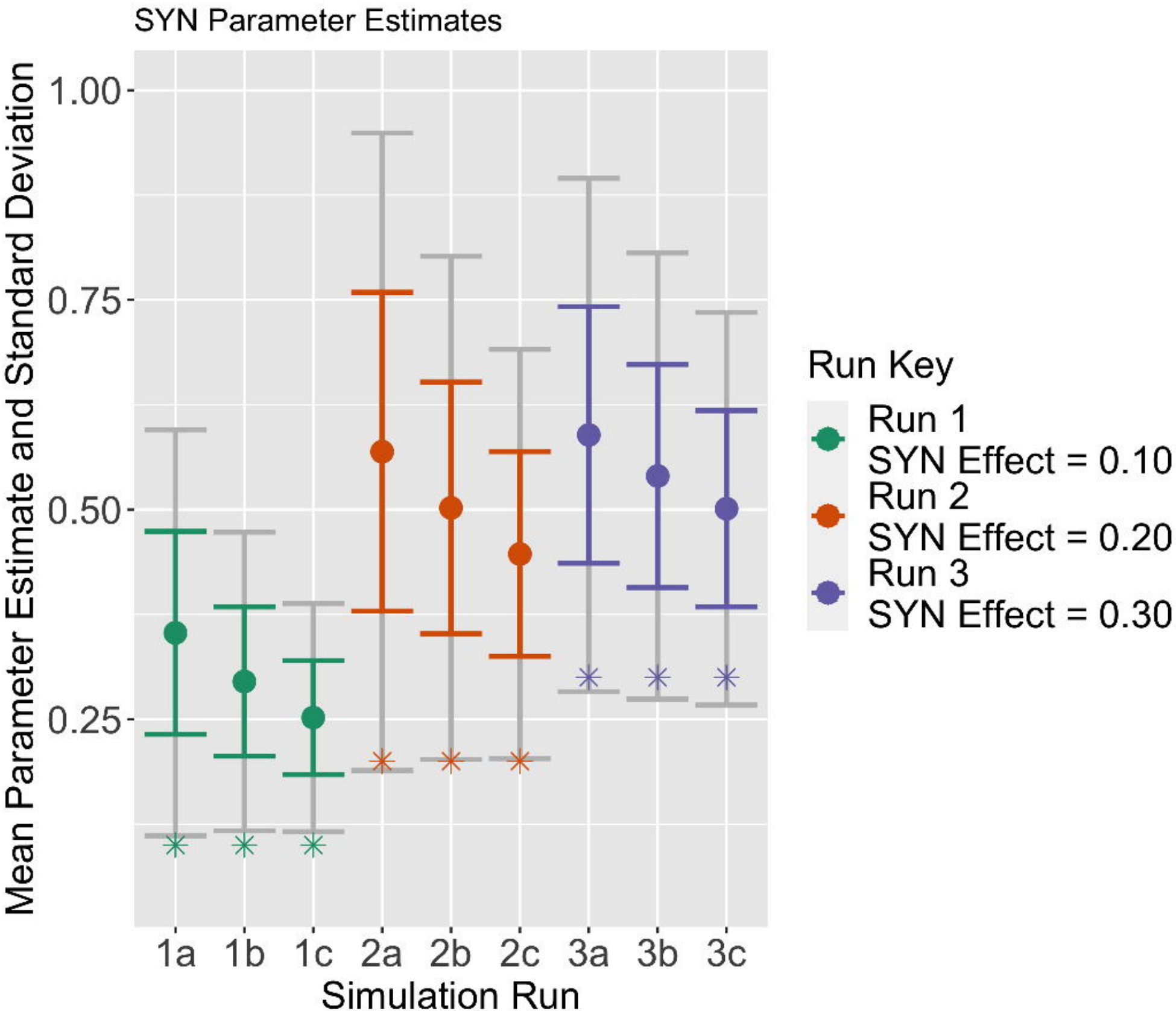

Clustered Genes in one Cluster of Interest
height
0.00
0.02
0.04
0.06
AGER
C19orf55
GCAT
PLD4
AKR1E2
ADH4
ECI2
TYW5
LDHC
RHBDL2
NUP62CL
OCIAD2
NME6
PPP1R14D
AC008686.1
COPZ2


Median Cluster Size by Proportion of Exome Inclusion
Proportion of the Exome Included in the Clusters of Interest
1.00
0.75
0.50
0.25
0.00
0
200
400
600
Median Cluster Size, in number of genes
Gene Set
Ethanol-Metabolizing Genes of Interest
Expression Set
Invertebrate Set


gnomAD constraints database
10 metrics on 19,108 genes
Cluster gnomAD genes
Cut tree, creating 3 sets
139 control genes clustered with 10 GOIs
1,531 control genes clustered with 108 expression set GOIs
3,366 control genes clustered with 353 invetebrate GOIs
Quantify SYN, MIS, and LOF variants counts, per gene in the ISP exome sample
Association testing, by set: GOI ~ SYN + MIS + LOF
gnomAD database data only
Analysis with observed exome data


LOF Parameter Estimates
Mean Parameter Estimate and Standard Deviation
4
3
2
1
1a
1b
1c
2a
2b
2c
3a
3b
3c
Simulation Run
Run Key
Run 1
LOF Effect = 0.75
Run 2
LOF Effect = 1.25
Run 3
LOF Effect = 1.75


MIS Parameter Estimates
Mean Parameter Estimate and Standard Deviation
0.0
0.2
0.4
0.6
1a
1b
1c
2a
2b
2c
3a
3b
3c
Simulation Run
Run Key
Run 1
MIS Effect = 0.05
Run 2
MIS Effect = 0.10
Run 3
MIS Effect = 0.20


SYN Parameter Estimates
Mean Parameter Estimate and Standard Deviation
1.00
0.75
0.50
0.25
1a
1b
1c
2a
2b
2c
3a
3b
3c
Simulation Run
Run Key
Run 1
SYN Effect = 0.10
Run 2
SYN Effect = 0.20
Run 3
SYN Effect = 0.30