Identifying Addiction-Associated Genes: Analyzing RNA-Seq Data from Brain Tissues to Reveal Genetic Drivers of Addiction Phenotypes

CSCI 5461 Final Project

May 5, 2023

Authors: Anna Saboe (*saboe009*), Ryan Mower (*mower023*), Grace Walker (*walk1195*),

Muhammad Abuzar (*abuza006*), Greg Shobert (*shobe051*), Luke Bessant (*bessa028*)

Abstract

Addiction-related disorders are known to have a strong impact on the function of various areas of the brain, but the genetic implications of addiction are largely understudied. Further understanding the relationship between addiction and gene expression in the brain serves to enrich knowledge of how these diseases can be treated from a genomic lens. In this project we aimed to illustrate trends in gene expression within three types of brain tissues with high genetic association to addiction phenotypes: the basal ganglia, amygdala, and prefrontal cortex. Skeletal tissue was used as a control tissue not typically predicted to be altered by addiction as brain tissue would. We performed hierarchical clustering of RNA-Seq TPM data from each tissue to identify modules of similar gene expression within each tissue type. Resulting clusters for each tissue were analyzed for enrichment of genes on which alcohol dependence-related SNPs have been identified. Gene Ontology (GO) enrichment for each tissue was performed on the genes belonging to the clusters most highly enriched for addiction-related genes to search for patterns in biological pathways and functions across clustered genes to thereby determine genetic drivers of addiction. We found that clusters containing the highest number of addiction-related genes showed enrichment for biological terms pertaining largely to neuron signaling, such as synapse maintenance, dendrite structure and synaptic regulation. This indicates the kinds of biological processes most affected by addiction phenotypes, potentially as a result of an individual containing an addiction-associated SNP in their genome. We further conclude that our approach profiles genes from significant clusters as modules, revealing groups of genes with correlated implications for brain addiction pathways that could be investigated further.

Introduction

Addiction is a chronic brain disorder characterized by functional changes to brain circuits involved in reward, stress, and self-control, resulting in compulsive drug seeking behavior and use beyond the influence of consequences[1]. 46.3 million people in the past year alone meet the DSM-5 criteria for a substance abuse disorder, 29.5 million of them being alcohol-related[2]. While extremely prevalent, addiction and addiction related disorders are largely underrepresented in current genomics research. At least half of addiction susceptibility is associated with genetic factors[3], but many of those factors are unknown or understudied in the context of expression relating to addiction disorders.

One current research avenue is identifying novel genes expressed more highly in addiction-associated tissues than other tissues. Since addiction is a brain disorder, investigating gene expression in brain tissues may be especially informative of the relationship between genetic expression and addiction phenotype. The brain plays a critical role in addiction development for individuals across a variety of regions. Well-supported evidence suggests that the amygdala, basal ganglia, and prefrontal cortex are three regions most significant to the onset, development, and maintenance of substance use disorders[4]. The basal ganglia controls pleasurable effects of substance use, contributing to habit formation, while the amygdala contributes to stress, anxiety, and irritability associated with substance withdrawal. Most notably, the prefrontal cortex facilitates executive function, closely associated with the lack of control characteristic of substance abuse[4]. These three brain tissues jointly form a three-stage 'addiction cycle,' a biological addiction mechanism which intensifies over time, inducing an increasingly powerful reinforcement of addiction-related neural pathways.

Little is currently known about differential gene expression across the amygdala, basal ganglia, and prefrontal cortex, or how this differential expression might relate to addiction disorders across individuals. Identifying highly and differentially expressed genes from these tissues of individuals with addiction disorders could illuminate biological drivers of addiction, including individual genes, gene clusters, and related physiological pathways. Treatment for addiction currently falls behind those for other brain disorders, due in part to uncertainty over genetic factors involved. Identifying key genes and tissues that cluster together in association with addiction could reduce this uncertainty, constituting substantial therapeutic potential.

[1] NIDA. 2021, August 3. Introduction. Retrieved from https://nida.nih.gov/publications/drugs-brains-behavior-science-addiction/introduction on 2023, April 26.

[2] Administration (SAMHSA), S. A. and M. H. S. (2023, January 4). *SAMHSA Announces National Survey on Drug Use and Health (NSDUH) Results Detailing Mental Illness and Substance Use Levels in 2021*. HHS.gov. https://www.hhs.gov/about/news/2023/01/04/samhsa-announces-national-survey-drug-use-health-results-detailing-mental-illness-substance-use-levels-2021.html#:~:text=Drug%20Use%20and%20Substance%20Use

[3] Price, M. (2008, June 1). Genes matter in addiction. *Monitor on Psychology*, *39*(6). https://www.apa.org/monitor/2008/06/genes-addict

[4] Administration (US), S. A. and M. H. S., & General (US), O. of the S. (2016). The neurobiology of substance use, misuse, and addiction. In *Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health [Internet]*. US Department of Health and Human Services. https://www.ncbi.nlm.nih.gov/books/NBK424849/.

Genome-wide association studies (GWAS) are commonly employed to identify individual genes whose expression is associated with a given trait or disease by analyzing the genomes of many individuals for single nucleotide polymorphisms (SNPs). GWAS have identified 731 SNPs related to the search term "drug dependence", which constitutes dependence and addiction to substances such as alcohol, nicotine, heroine, cannabis, and stimulants, therefore illuminating vast genetic variation associated with addiction[5]. Additionally, the Genotype-Tissue Expression (GTEx) project reports RNA-sequencing (RNA-seq) data and calculated TPM values for ~56,200 genes from a variety of body tissues[6]. Analyzing RNA-seq data by tissue and observing the expression trends for the genes associated with addiction-related SNPs could potentially reveal brain tissues more closely associated with addiction.

In this project, we analyzed RNA-Seq TPM tissue data from the three brain tissues of interest: the amygdala, prefrontal cortex, and basal ganglia, as well as a control tissue for which we chose skeletal muscle. After performing hierarchical clustering analysis, we then looked within each cluster of a given tissue and determined if certain clusters were enriched with genes from the list of SNPs associated with alcoholism. We sought to determine if particular tissues would display greater cluster enrichment than the others, which may inform future studies as to whether gene expression in particular tissues would be more informative of alcohol dependency.

Methods

Obtaining Datasets

GWAS Addiction-Related SNPs

The GWAS Catalog from the National Human Genome Research Institute's (NHGRI) database of human genome-wide association studies[7] was used to find SNPs and their associated genes that have been identified as relating to addiction. We used the search term 'drug dependency' to gather a list of studies associating SNPs with various conditions such as alcohol, heroin and cannabis dependency. 731 SNPs associated with these conditions were found from 125 studies, with most having been mapped to a particular gene in the human genome. This list of SNPs was then filtered to remove those that had not been mapped to either a chromosome location or specific gene, which resulted in a list of 635 SNPs. Once these SNPs had been filtered, we then further filtered the list by traits (i.e. what type of addiction phenotype each SNP was correlated with) and extracted only the SNPs associated with alcohol dependence. This resulted in a list of 277 genes of interest with which downstream analysis was performed.

Tissue Expression Data

[5] Baurley, J.W., Edlund, C.K., Pardamean, C.I. et al. Smokescreen: a targeted genotyping array for addiction research. BMC Genomics 17, 145 (2016). https://doi.org/10.1186/s12864-016-2495-7

[6] The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this report were obtained from the GTEx Portal on 04/01/23.

[7] Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical Genomic Database. Proc Natl Acad Sci U S A. 2013 May 21. [Epub ahead of print] [PubMed]

The regions of the brain whose functions have been found to be most impacted by addiction related disorders include the basal ganglia, amygdala, and the prefrontal cortex. We downloaded gene expression data from the Genotype-Tissue Expression (GTEx) project database for each of these three tissues, as well as skeletal muscle tissue to be used as a control. Data was provided in the form of Transcripts Per Million, representing a normalized form of RNA-seq read count data scaled to the number of RNA transcripts for every 1,000,000 within the given sample[8]. In the context of our research, TPM values are beneficial to use because they account for the length of genes, therefore acting as a more accurate metric for measuring gene expression across samples. TPM values for each tissue were provided for a number of individuals across various age groups. Total samples are shown in Table 1 alongside a breakdown of gender and age group distribution. All four tissues had higher numbers of male samples than females, and the greatest number of samples from the 60-70 age range, with 50-59 being the second highest. Younger age groups were far less represented in this dataset.

Table 1. Donor data for samples collected from four tissue types, including total sample counts, percent male:female, and distribution of samples across age groups.

	Sample Total	**Male:Female (%)**	**18-29 years**	**30-39 years**	**40-49 years**	**50-59 years**	**60-70 years**
Basal ganglia	205	76	5	5	19	67	109
Amygdala	152	70	5	4	13	45	85
Prefrontal cortex	209	73	5	4	17	63	120
Skeletal muscle *(control)*	803	66	67	65	124	255	292

Hierarchical Clustering Algorithm

We performed hierarchical clustering of the TPM data to identify clusters of similar gene expression patterns within each tissue. Variance was first determined across all samples, and the dataset was reduced to only the top 10,000 genes exhibiting the highest variance due to both computational technology limitations and biological considerations. This step ensured removal of genes with little to no expression measurements. Additionally, this step removed genes with uniform expression across all tissues, likely to be housekeeping genes or those with well-conserved expression relating to essential biological functions that are outside the context of our research interests in addiction.

Hierarchical clustering was then performed on TPM expression data for each tissue independently using the scipy.cluster.hierarchy library. We performed complete-linkage clustering, a form of agglomerative clustering, to return the maximum of all pairwise distances between nodes. The Pearson correlation coefficient was used as a metric to calculate distances

[8] Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA (New York, N.Y.)*, *26*(8), 903–909. https://doi.org/10.1261/rna.074922.120

for our clustering. Initially, a handful of other hierarchical clustering strategies and distance metrics were tested with the goal of obtaining clusters containing approximately 20-100 genes. This was done because it was hypothesized that clusters of this size would be more likely to capture interesting biological information upon analysis of function enrichment. However, other methods were not able to provide an even distribution of samples into the clusters, and thus, complete-linkage clustering was chosen. In addition to the favorable cluster distribution results obtained from using this method, complete linkage clustering has also been demonstrated to be an effective way to reduce the impact of noise and outliers on clustering[9]. Additionally, this method is useful for identifying homology in order to help understand cellular processes[10]. The Pearson correlation coefficient has been cited in the literature as one of the best-performing metrics for complete-linkage clustering compared to 14 other distance measures[11]. Using these metrics to perform hierarchical clustering, the number of clusters in each tissue was chosen to obtain a mean of 50 genes contained throughout all clusters, resulting in 200 total clusters for each tissue.

Following clustering, we then sorted our clusters by those that contained the greatest number of intersecting genes between our list of alcohol-associated genes and the total genes in each cluster. The cluster containing the greatest degree of overlap was deemed the “most interesting” cluster, and the list of all of its members was used to test for GO enrichment.

Statistical Analysis for Enrichment

To assess the biological relevance and significance of the results from our clustering algorithm, we performed a Gene Ontology (GO) enrichment analysis of the most interesting clusters from each tissue using a statistical overrepresentation test. This was done to check for biological terms associated with the genes found in the most interesting clusters. Such information may indicate what types of biological processes are most affected by addiction phenotypes, potentially as a result of the content of addiction-associated SNPs in an individual’s genome. Using the PANTHER classification system available through the Gene Ontology Resource website[12], we performed the enrichment analysis using the Fisher's Exact statistic to determine if particular terms were overrepresented that relate to biological processes, molecular functioning, and cellular components. The process works by inputting the list of genes that belonged to the most interesting cluster for each tissue separately. Then the system searches through each gene in the list and determines all the biological terms the gene is linked to. This is concurrently done with the *Homo sapiens* reference gene list (20589 total genes). The total

[9] Jayanthi Ranjan and Saani Khalil, 2007. Clustering Methods for Statistical Analysis of Genome Databases. Information Technology Journal, 6: 1217-1223.

[10] Oyelade et al. clustering algorithms: their application to gene Expression data. Bioinformatics and Biology Insights 2016:10 237–253 doi: 10.4137/BBi.s38316.

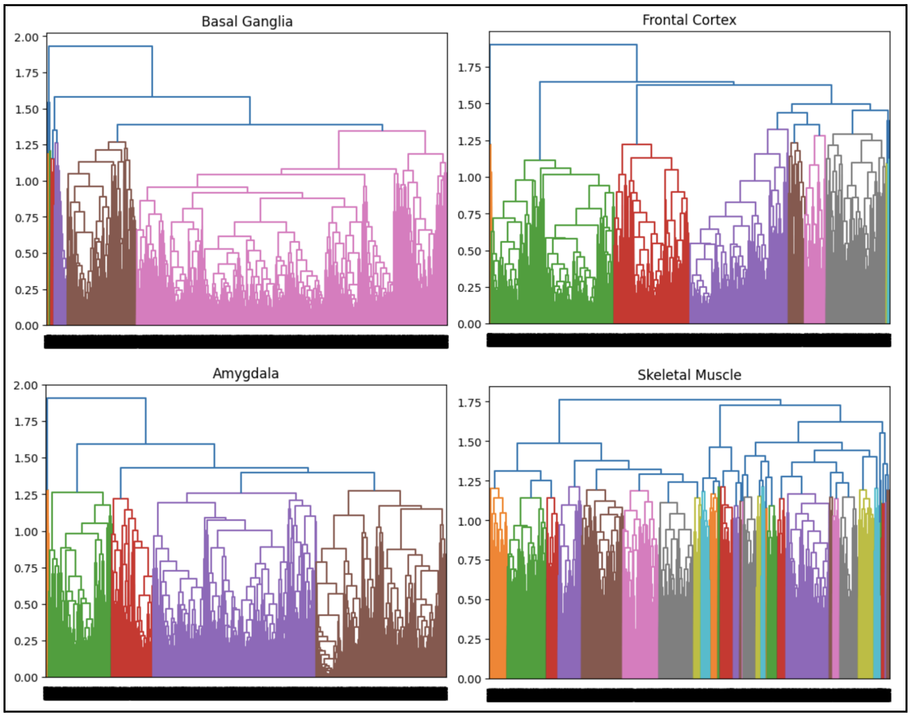
[11] Jaskowiak, P.A., Campello, R.J. & Costa, I.G. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics 15 (Suppl 2), S2 (2014). https://doi.org/10.1186/1471-2105-15-S2-S2

[12] Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. Jan 2019;47(D1):D419-D426

number of genes in the genome pertaining to each biological term that was also a hit for the input list are counted. Then, the system determines the expected degree of enrichment based on how many genes out of the entire genome map to each particular term. This is then used to calculate the fold enrichment within the input gene list, where the number of genes that match with each biological term is divided by the expected enrichment score to then determine if the input list is overrepresented or underrepresented by a particular biological term[13]. The most enriched biological components were chosen based on the highest enrichment scores for particular biological terms that also reported a significant p-value and False Discovery Rate (FDR) of less than 0.05. We also performed enrichment analyses for the genes that showed overlap across the most significant clusters from each tissue to see if there were particular biological terms shared across tissues.

Results

Our clustering algorithm produced dendrograms for each tissue, shown in Figure 1. Figure 2 contains histograms for each tissue type that display the distribution of genes across all resulting clusters. Mapping the set of alcohol addiction-related genes to all clusters and counting the number of these genes that appear in each cluster produced a single cluster from each tissue containing the greatest amount of addiction-related genes. The amount of addiction-related SNPs for each of these four clusters is shown in Table 2, as well as the relative cluster size and ratio of addiction-related genes to the total genes in the cluster.



[13] Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, *8*(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092.

Figure 1: Dendrograms obtained from complete-linkage clustering algorithm performed using Pearson correlation coefficient for three brain tissues associated with addiction. Clustering was also performed on skeletal muscle tissue expression data as a control.

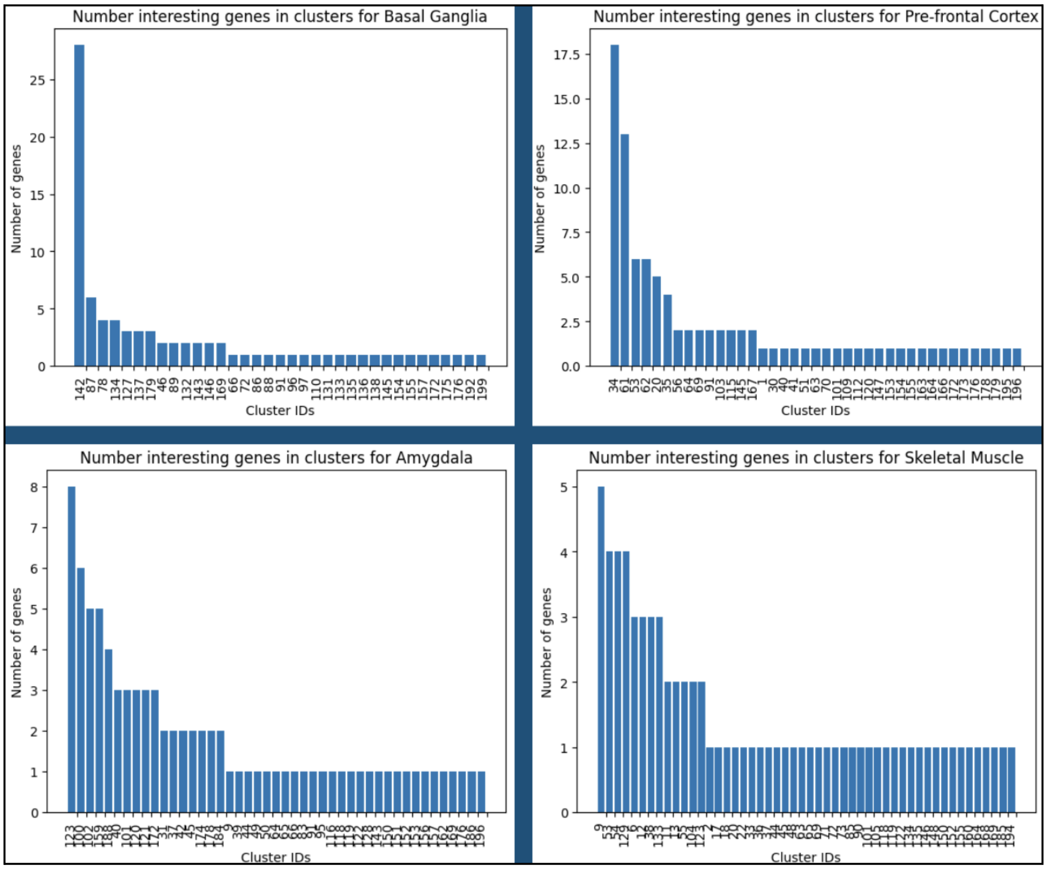


Figure 2: Histograms for basal ganglia, prefrontal cortex, amygdala, and skeletal tissue types outlining the distribution of genes per cluster across all clusters. The mean number of genes per cluster was 50, and the total number of clusters for each tested tissue was 200.

Table 2. Relative cluster size of clusters with the greatest number of addiction-related genes mapped to it for each tissue. Addiction-related SNPs were taken from the NHGRI GWAS database list of addiction-related SNPs and the relative gene they reside on was used for analysis.

Tissue Type	Size of Most Significant Cluster	# Addiction Related Genes	% Addiction Related Genes
Basal ganglia	1,824 genes	28 genes	1.54%
Prefrontal cortex	1,379 genes	18 genes	1.31%
Amygdala	686 genes	8 genes	1.17%
Skeletal muscle	768 genes	5 genes	0.07%

All genes within the most significant cluster for each tissue were enriched to GO terms to identify trends relating to biological functions. Table 3 displays resulting biological terms from GO enrichment mapped to each cluster that were found to be statistically significant by False Discovery Rates.

Table 3. Results from GO enrichment of clusters containing the greatest amount of addiction-related genes for each tissue. Columns reflect each type of biological term enrichment was performed for, and resulting terms are those with a statistically significant FDR calculated using p-values resulting from Fisher's Exact test of significance.

Tissue Type	Biological Process	FDR	Molecular Function	FDR	Cellular Component	FDR
Basal ganglia	Synaptic vesicle budding	*2.21E-02*	Phosphatidylinositol-3-phosphate phosphatase activity	*4.87E-02*	Dendritic spine head	*3.30E-02*
Prefrontal cortex	Neurotransmitter receptor localization to postsynaptic specialization membrane	*1.55E-03*	Voltage-gated monatomic ion channel activity involved in regulation of presynaptic membrane potential	*5.76E-03*	Postsynaptic intermediate filament cytoskeleton	*1.24E-02*
Amygdala	Postsynaptic density structure maintenance	*3.55E-02*	Synaptic receptor adaptor activity	*2.07E-02*	Primary dendrite	*3.53E-02*
Genes shared across all 3 tissues	Synapse organization	*3.43E-02*	*N/A*	*N/A*	Kinesin II complex	*1.94E-02*

Discussion

The results from our clustering algorithm and subsequent enrichment analysis reveals that genes relating to alcohol are significantly enriched to biological functions matching to neural signaling and neuron pathways in the brain. Biological terms enriched to basal ganglia, prefrontal cortex, and amygdala tissues with a statistically significant FDR all support this finding. The skeletal muscle control tissue was not enriched for functions pertaining to neural signaling, demonstrating that this finding is unique to brain tissues. Synaptic vesicle budding, postsynaptic density maintenance, and synaptic receptor adaptor activity all are critical processes enabling proper functioning of neural communication. Synapses are spaces between nerve cells across which numerous neurotransmitters, hormones, and molecules are passed. Mutations affecting their density and structure, as well as the receptors within them, therefore have detrimental effects for the cascade of communication between neurons and in turn for the proper function of various brain areas. Regulation of synaptic communication is another trend of enrichment in significant addiction-related clusters. Genes in the prefrontal cortex revealed enrichment for voltage-gated ion channel activity regarding regulation of presynaptic membrane potential, those in the amygdala mapped to synaptic receptor activity, and the basal ganglia to phosphatidylinositol-3-phosphate phosphatase (PtdIns3P) activity. Synaptic transmission is one of the most closely regulated and sophisticated neural processes. The enrichment of our significant clusters to biological terms relating to this regulation both supports previous findings of a strong relationship between genomic variation of addiction disorders and brain function, while also demonstrating new gene modules that can be the aim of further study.

The results of enrichment for genes shared across significant clusters from all tissue types further supports the relationship between addiction-related genes and neural functioning. Specifically, synapse organization and kinesin II complex were found to be enriched for the complete list of shared genes (Table 3). As previously noted, synapse organization is vital to the cohesion of brain functioning across all regions. Specifically for the amygdala, basal ganglia, and prefrontal cortex, malfunctions in signaling through well-organized synaptic pathways within these brain regions directly impacts processes such as recall, reward processing, and synaptic plasticity. Kinesin II complex is a complex of motor proteins that is predicted to be a transcriptional target for regulation of drug-induced alterations in synaptic plasticity. Both of these terms implicate our clusters as gene modules deeply connected to synaptic functioning in the brain as it relates to addiction.

In a broader sense, our results reveal numerous novel genes sharing expression patterns with those involved in neural functioning pathways that have not previously been related to addiction and addiction disorders. Total gene counts for each most significant cluster (Table 2) range from 686 for amygdala tissue to 1,824 for basal ganglia tissue. Because we enriched the clusters containing the greatest number of genes carrying addiction-related SNPs, these genes are revealed to possibly work as a module impacting the same pathways as are impacted by the previously identified addiction-related genes. This reveals specific genes and groups of genes that are significant to the pathways of addiction and are therefore critical for a holistic understanding of the genetic underpinnings of addiction. As a result of our study, such gene modules are targets for further research into why variation within them and connections between them affect and relate to addiction.

Future Directions

The work performed in this project opens a variety of avenues for further exploration of the genetics of addiction, both within the scope of specific genes and modules identified and beyond it into the general approach to studying the genotype-phenotype connection. Identification of gene modules correlated to the biological pathways of addiction allows for deeper investigation into the specific genes within them, the proteins they encode, and how they relate to each other transcriptionally. Additionally, the dataset used for our project could be broadened in a variety of ways to include greater representation of younger populations, for example, or to include other measurements of gene expression. Our number of samples could also be increased to further validate the results of our clustering algorithm.

Another direction of further research is to investigate other addiction phenotypes and how they differ in terms of gene expression in the brain tissues we tested. In the context of our study we tested for SNPs specific to alcoholism. However, nicotine, cannabis, and opioids are all drugs affecting similar pathways in the brain. Aiming to identify genes unique to each of these drug dependencies, as well as those sharing expression in a codependent phenotype where more than one substance disorder exists, could reveal key insights to the genetics of drug dependency and addiction across its many varieties.

We also suggest further research that would optimize our clustering methodology, such as implementing an average linkage or k-medoids approach and measuring notable changes in cluster results. Additionally, applying different distance metrics than the one used in our report such as Euclidean distance or Rank Magnitude may impact the results of our clustering method and could be investigated to refine the overall approach we have put forward.

References

1. NIDA. 2021, August 3. Introduction. Retrieved from https://nida.nih.gov/publications/drugs-brains-behavior-science-addiction/introduction on 2023, April 26.
2. Administration (SAMHSA), S. A. and M. H. S. (2023, January 4). *SAMHSA Announces National Survey on Drug Use and Health (NSDUH) Results Detailing Mental Illness and Substance Use Levels in 2021*. HHS.gov. https://www.hhs.gov/about/news/2023/01/04/samhsa-announces-national-survey-drug-use-health-results-detailing-mental-illness-substance-use-levels-2021.html#:~:text=Drug%20Use%20and%20Substance%20Use
3. Price, M. (2008, June 1). Genes matter in addiction. *Monitor on Psychology*, *39*(6). https://www.apa.org/monitor/2008/06/genes-addict
4. Administration (US), S. A. and M. H. S., & General (US), O. of the S. (2016). The neurobiology of substance use, misuse, and addiction. In *Facing Addiction in America: The Surgeon General's Report on Alcohol, Drugs, and Health [Internet]*. US Department of Health and Human Services. https://www.ncbi.nlm.nih.gov/books/NBK424849/.
5. Baurley, J.W., Edlund, C.K., Pardamean, C.I. et al. Smokescreen: a targeted genotyping array for addiction research. BMC Genomics 17, 145 (2016). https://doi.org/10.1186/s12864-016-2495-7
6. Solomon BD, Nguyen AD, Bear KA, Wolfsberg TG. Clinical Genomic Database. Proc Natl Acad Sci U S A. 2013 May 21. [Epub ahead of print] [PubMed]
7. Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA (New York, N.Y.)*, *26*(8), 903–909. https://doi.org/10.1261/rna.074922.120
8. Zhao, S., Ye, Z., & Stanton, R. (2020). Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols. *RNA (New York, N.Y.)*, *26*(8), 903–909. https://doi.org/10.1261/rna.074922.120
9. Jayanthi Ranjan and Saani Khalil, 2007. Clustering Methods for Statistical Analysis of Genome Databases. Information Technology Journal, 6: 1217-1223.
10. Oyelade et al. clustering algorithms: their application to gene Expression data. Bioinformatics and Biology Insights 2016:10 237–253 doi: 10.4137/BBi.s38316.
11. Jaskowiak, P.A., Campello, R.J. & Costa, I.G. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics 15 (Suppl 2), S2 (2014). https://doi.org/10.1186/1471-2105-15-S2-S2
12. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. Jan 2019;47(D1):D419-D426
13. Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature protocols*, *8*(8), 1551–1566. https://doi.org/10.1038/nprot.2013.092.