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

Anna Saboe, Ryan Mower, Grace Walker, Luke Bessant, Muhammad Abuzar, Greg Shobert

Introduction

Biological Question: Are certain genes associated with addiction expressed higher or differentially across areas of the brain impacted by addiction related disorders?

Motivation:
➔ Large gap in knowledge regarding genetic basis of addiction
➔ Personal interest in addiction
➔ Initially, wanted to understand relationship between age and addiction phenotypes
   ◆ Lack of public data availability
➔ *Could different tissues display different gene expression patterns?*
➔ *Overall, can genetics predict risk of addiction?*

Approach - Obtaining Data

1. SNPs associated with drug dependence
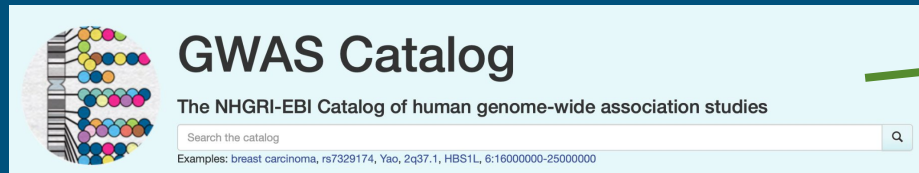
Variant_Risk_Allele	Location	Mapped gene	Trait(s)
rs2133896	12:99455122	ANKS1B	alcohol dependence, heroin dependence, methamp...
rs147247472	1:49441901	AGBL4, AGBL4-IT1	alcohol dependence, heroin dependence, methamp...
rs10196867	2:79751234	CTNNA2	alcohol dependence, heroin dependence, methamp...
rs11731003	4:7388545	SORCS2	alcohol withdrawal
rs528301	2:44927769	LINC01833	alcohol and nicotine codependence
...	...	...	...
rs642899	13:33106717	STARD13	nicotine use, generational effect measurement,...
rs9782914	1:241728933	WDR64	nicotine use, generational effect measurement,...
rs10037670	5:154431930	SAP30L-AS1	nicotine use, generational effect measurement,...
rs2071214	17:78223510	BIRC5	nicotine use, generational effect measurement,...
rs2380220	6:95519536	CYCSP17, MANEA-DT	nicotine use, generational effect measurement,...

2. TPM values from RNA-Seq data
   a. Basal Ganglia
   b. Frontal Cortex
   c. Amygdala
   d. Skeletal Tissue

Description	GTEX.1192X.0011.R7b.SM.DNZZC	GTEX.11DXW.0011.R7b.SM.DO119	GTEX.11DXY.0011.R7a.SM.DO11L	GTEX.11DYG.0011.R7b.SM.EXUS6
DDX11L1	0.0000	0.0000	0.0127	0.000
WASH7P	1.8930	0.8327	1.9290	2.145
MIR6859-1	0.0000	0.0000	0.0000	0.000
MIR1302-2HG	0.1037	0.1256	0.0000	0.000
FAM138A	0.0000	0.0000	0.0359	0.000
...	...	...	...	...
MT-ND6	8320.0000	20060.0000	8295.0000	6948.000
MT-TE	63.7700	68.0400	648.2000	19.520
MT-CYB	34410.0000	51280.0000	42030.0000	41970.000
MT-TT	5.4420	7.9650	24.8800	1.962
MT-TP	1.9810	3.4660	3.5420	3.047

Approach- Hierarchical clustering

1. **Complete-linkage hierarchical clustering**
   a. Gave most uniform distribution of genes into clusters
      i. 20-100 genes to capture biological entities
      ii. n = 200 clusters
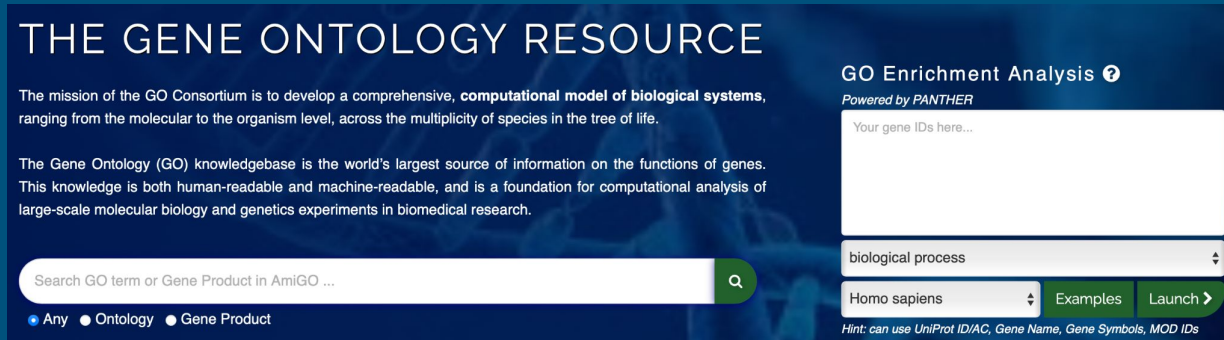   b. Has proven effective for gene expression data (D'haeseleer)

Furthest Neighbor (Complete Linkage Clustering)
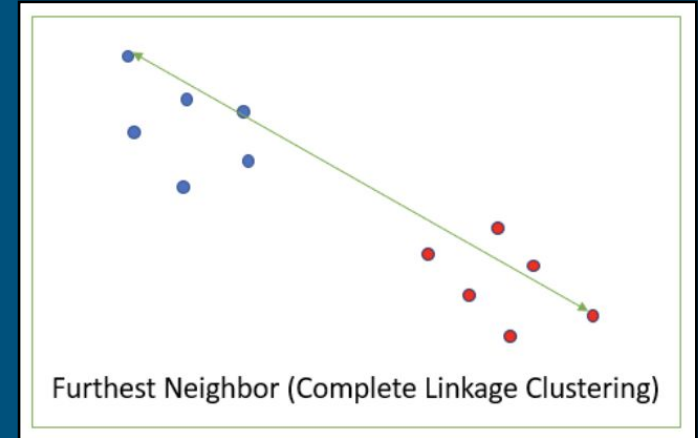
2. **Pearson correlation coefficient**
   a. Works well with complete-linkage clustering for gene expression data (Jaskowiak et al.)

$$d_{fg} = 1 - r_{fg}, \quad \text{with} \quad r_{fg} = \frac{\sum_c (e_{fc} - \bar{e}_f)(e_{gc} - \bar{e}_g)}{\sqrt{\sum_c (e_{fc} - \bar{e}_f)^2 \sum_c (e_{gc} - \bar{e}_g)^2}}$$

Sources: https://www.statisticshowto.com/complete-linkage-clustering/
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
D'haeseleer, P. How does gene expression clustering work?. *Nat Biotechnol* 23, 1499–1501 (2005). https://doi.org/10.1038/nbt1205-1499

Approach- Gene Ontology Enrichment

1. Determine most interesting cluster based on how many genes within the cluster match with the list of genes from our SNP data
2. Perform Gene Ontology (GO) using PANTHER to determine if most interesting clusters are enriched for particular functions
    a. Biological process
    b. Molecular function
    c. Cellular component

Results- Clustering

Tissue	Size of Most Significant Cluster	Largest # Addiction-Related Genes in Cluster
Basal ganglia	1,824 genes	28 genes
Pre-frontal cortex	1,379 genes	18 genes
Amygdala	686 genes	8 genes
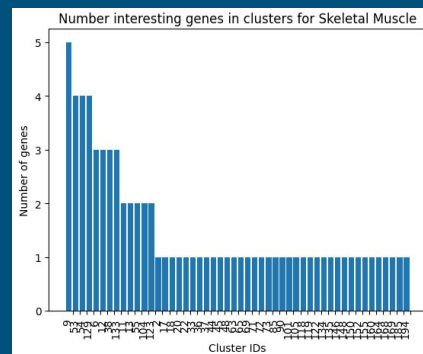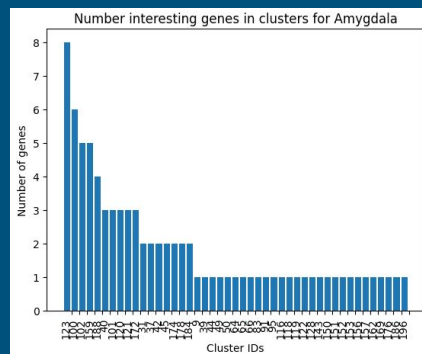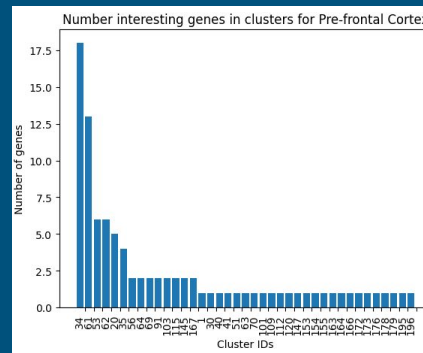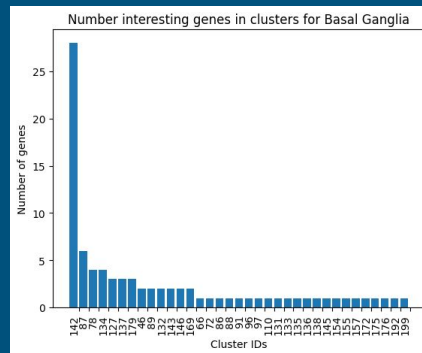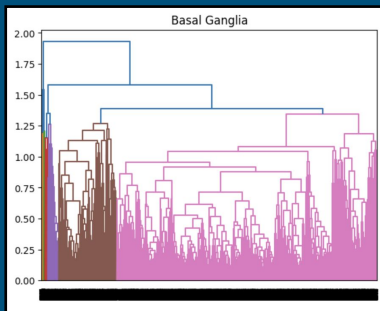
Results- Gene Ontology Enrichment

Tissue	Biological Process	Molecular Function	Cellular Component
Basal ganglia	Synaptic vesicle budding	Phosphatidylinositol-3-phosphate phosphatase activity	Dendritic spine head
	2.21E-02	4.87E-02	3.30E-02
Pre-frontal cortex	Neurotransmitter receptor localization to postsynaptic specialization membrane	Voltage-gated monatomic ion channel activity involved in regulation of presynaptic membrane potential	Postsynaptic intermediate filament cytoskeleton
	1.55E-03	5.76E-03	1.24E-02
Amygdala	Postsynaptic density structure maintenance	Synaptic receptor adaptor activity	Primary dendrite
	3.55E-02	2.07E-02	3.53E-02
Genes shared across all 3 tissues	Synapse organization	*No significant enrichment*	Kinesin II complex
	3.43E-02	*N/A*	1.94E-02

Conclusion and Future Work

Conclusion

➔ Most interesting cluster in all tissues showed enrichment for elements involved in signaling
  ◆ Could indicate a fundamental alteration in neurological communication induced by addiction-related SNPs and the genes they map to
➔ High number of shared genes between each tissue's most enriched cluster
  ◆ Could indicate that expression in all tested areas of the brain are related in an addiction phenotype context

Future Work

➔ Consider looking into other traits
  ◆ Depression, methamphetamine dependence, cannabis dependence, etc…
➔ Try additional clustering methods and metrics
  ◆ Centroid Linkage, average linkage, etc…
➔ Try various clustering algorithms
  ◆ Density or distribution based clustering, varying amount of clusters

Thank you!