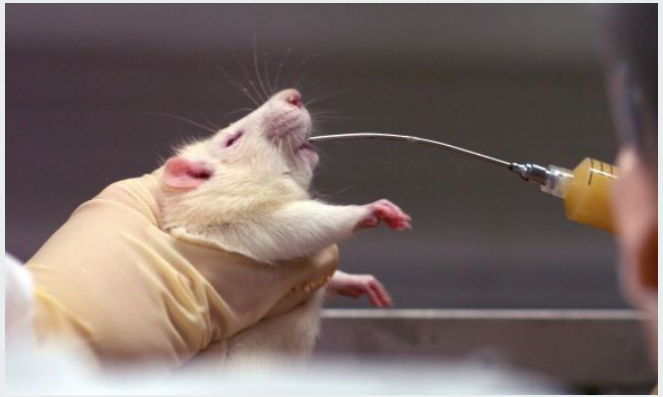
Binge Drinking Mice: Identifying Candidate Genes For a Classifier Model

Jack Saigusa, Arshia Verma, Xu Zhang



Introduction:

1. Excessive alcohol use, caused more than 140,000 deaths annually in the United States between 2015 and 2019. (CDC, 2022)

2. We wanted to know how binge drinking might impact brain in different brain regions.

3. Investigated 4 brain regions related to reward pathways and alcohol consumption. (Ferguson et al., 2019)

4. Mice models are commonly used to identify genes that contribute to the genetic predisposition to alcoholism. (Ferguson et al., 2019)

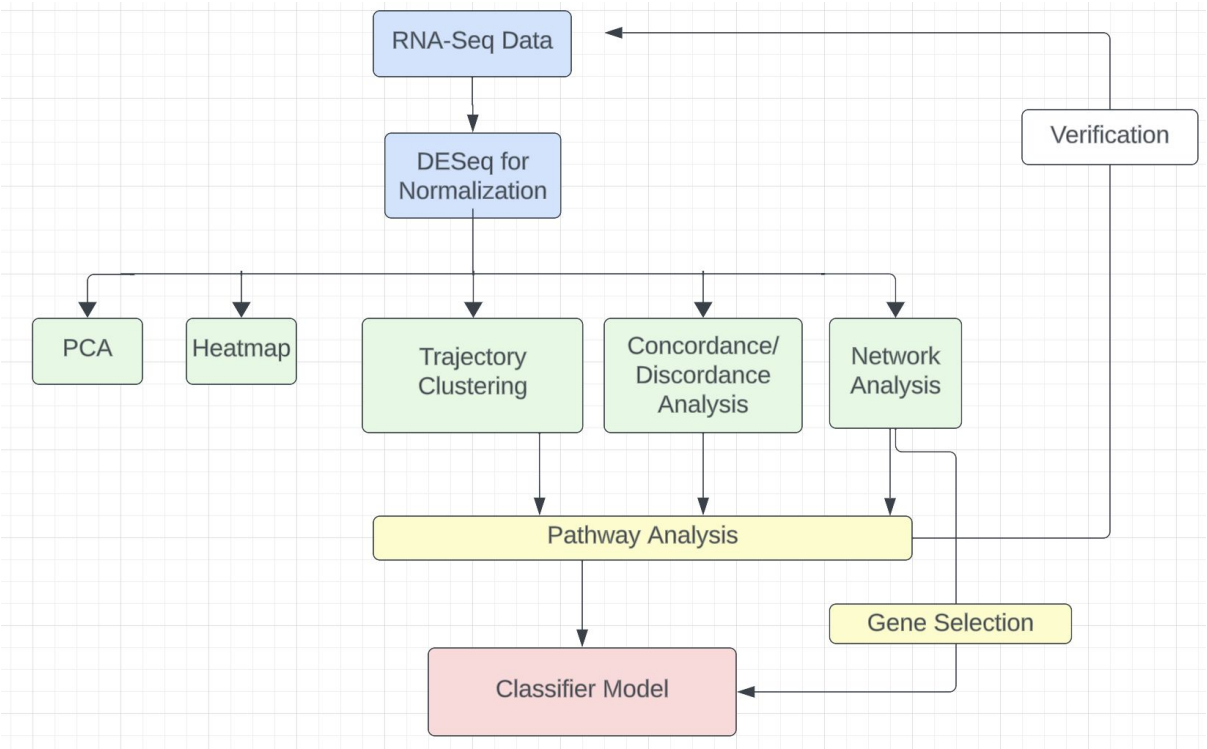
Our Samples:

3 genetic lines

6 brain regions

	HSNpt	iHDID1	iHDID2	Total
BLA (pooled)	10	10	10	30
BNST	40	39	40	119
CEA (pooled)	10	9	9	28
NAC	40	39	40	119
PFC	40	39	39	118
VTA	40	38	38	116

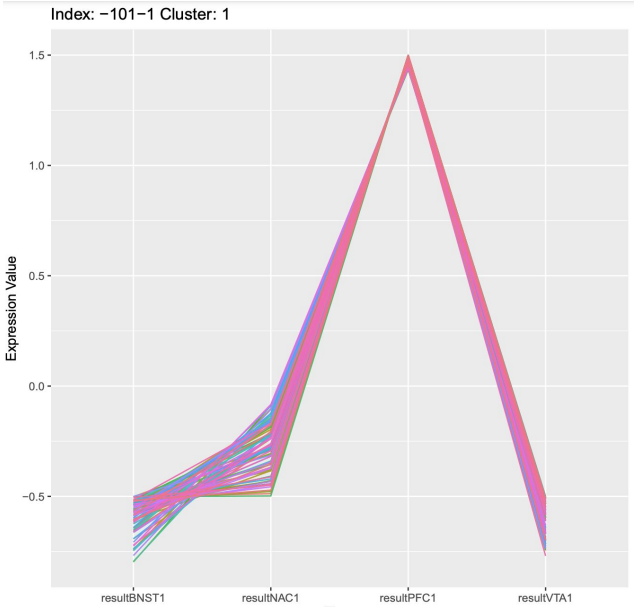
Goal: To identify candidate genes related to binge drinking in order to classify samples into mouse lines

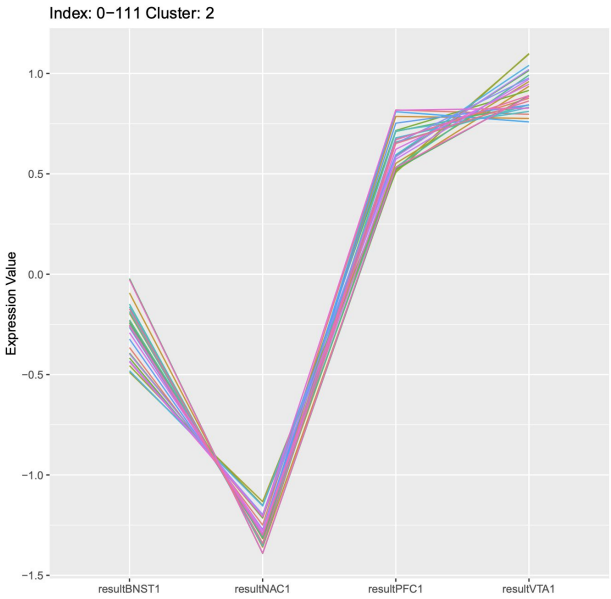


Trajectory Clustering: ctsGE

Goal: group genes using expression indexes and K-means clustering

ctsGE





DID1: 42 Clusters
DID2: 42 Clusters

Significantly enriched with DEGS

14 Clusters
15 Clusters

Top 8 Genes by Log2FC

68 genes
71 genes

Concordance and Discordance Analysis

Goal: Select DEGs with a consistent level of expressions across brain regions.

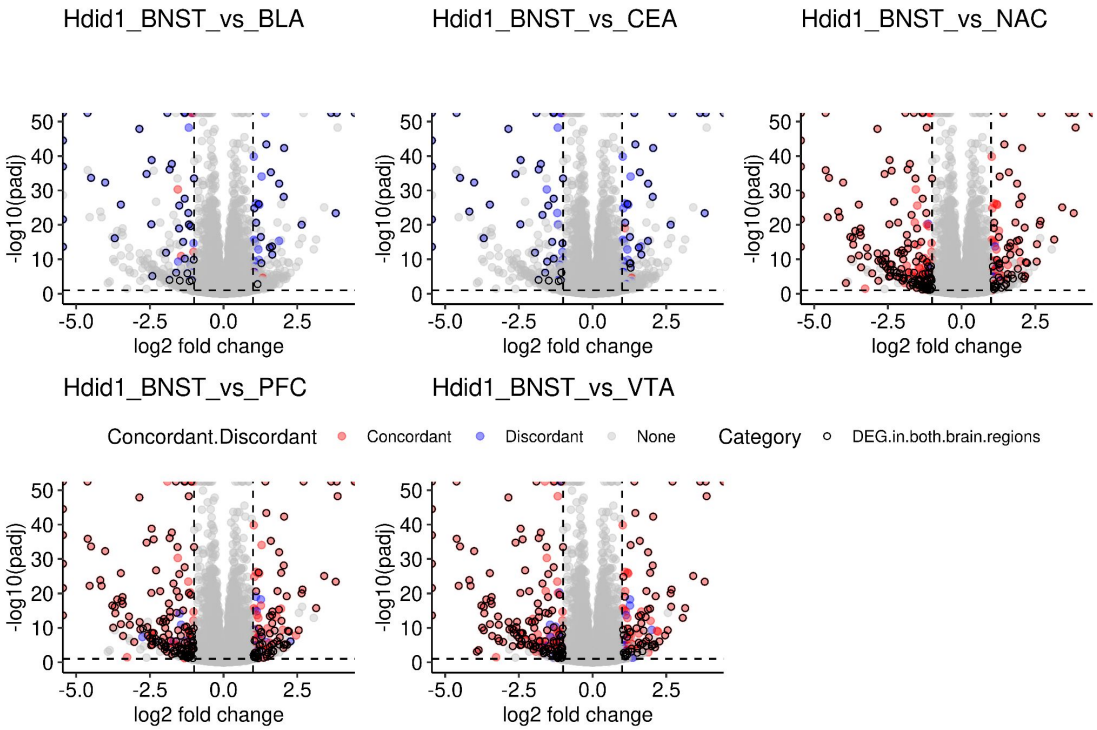
Method: Gene concordance and Measure their quality

Concordance/Discordance

- Concordance: find genes whose expression changes in the same direction across brain regions.

  Overexpression in one region implies overexpression in other regions.

- Excluded ‘low-quality’ concordant genes



Concordance and Discordance Analysis:

How did we find ‘ Low-quality ’ DEGs:

We want

Gene	Gm38405			
Brain Region	BNST	NAC	PFC	VTA
Log2FC	**-1.95**	1.129	**0.14**	-0.68

DEGs that have close Log2FC values in all brain regions

Conclusion

- With these criteria found:
  107 genes from the DID1 group .
  126 genes from the DID2 group.

- Ran pathway enrichment and left with most significant genes

- Conclusion:
  24 genes found in concordance analysis.

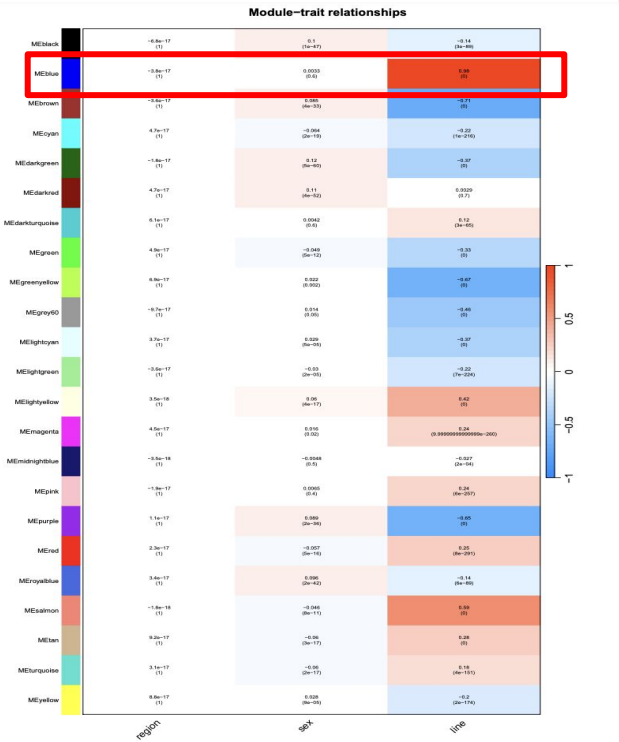
Co-Expression Network Analysis:

Goal: Identify groups of coexpressed genes that are expressed differently in HDID and control samples.

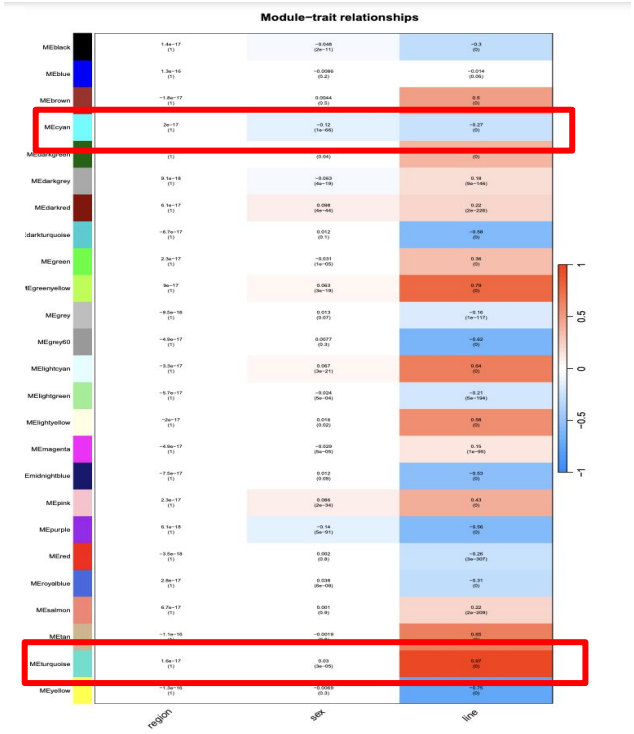
Method: Create WGCNA consensus modules.

WGCNA Consensus Modules

- WGCNA consensus modules - conserved gene co-expression relationships.



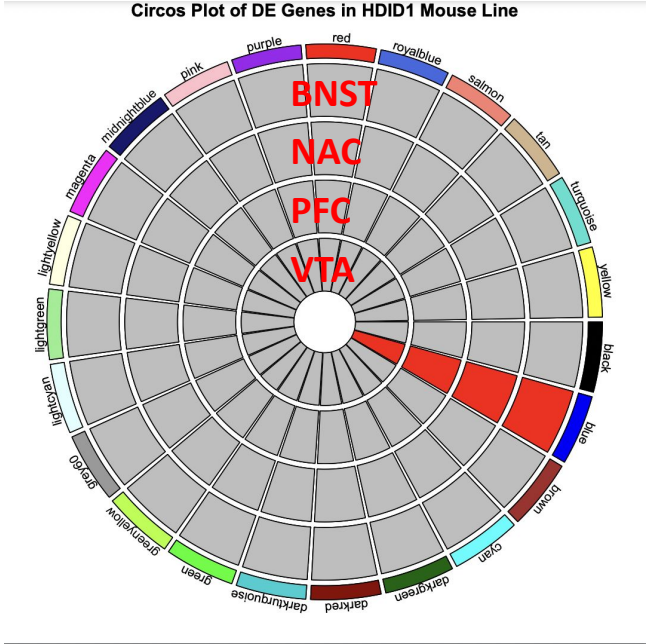
HDID1

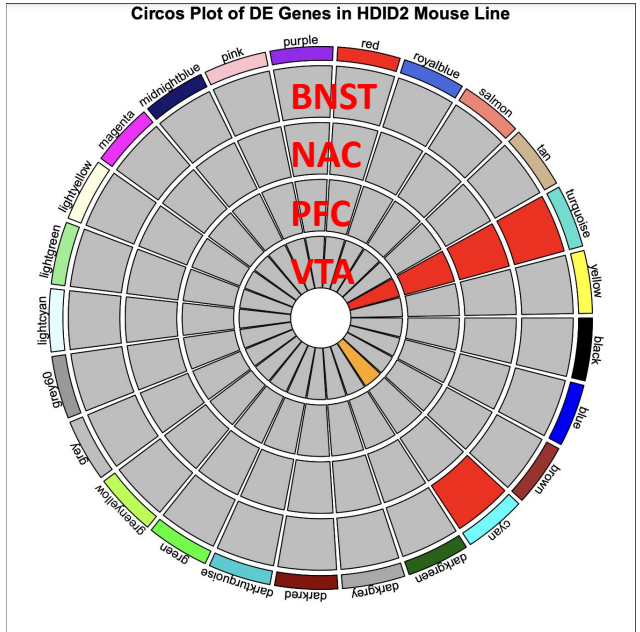


HDID2

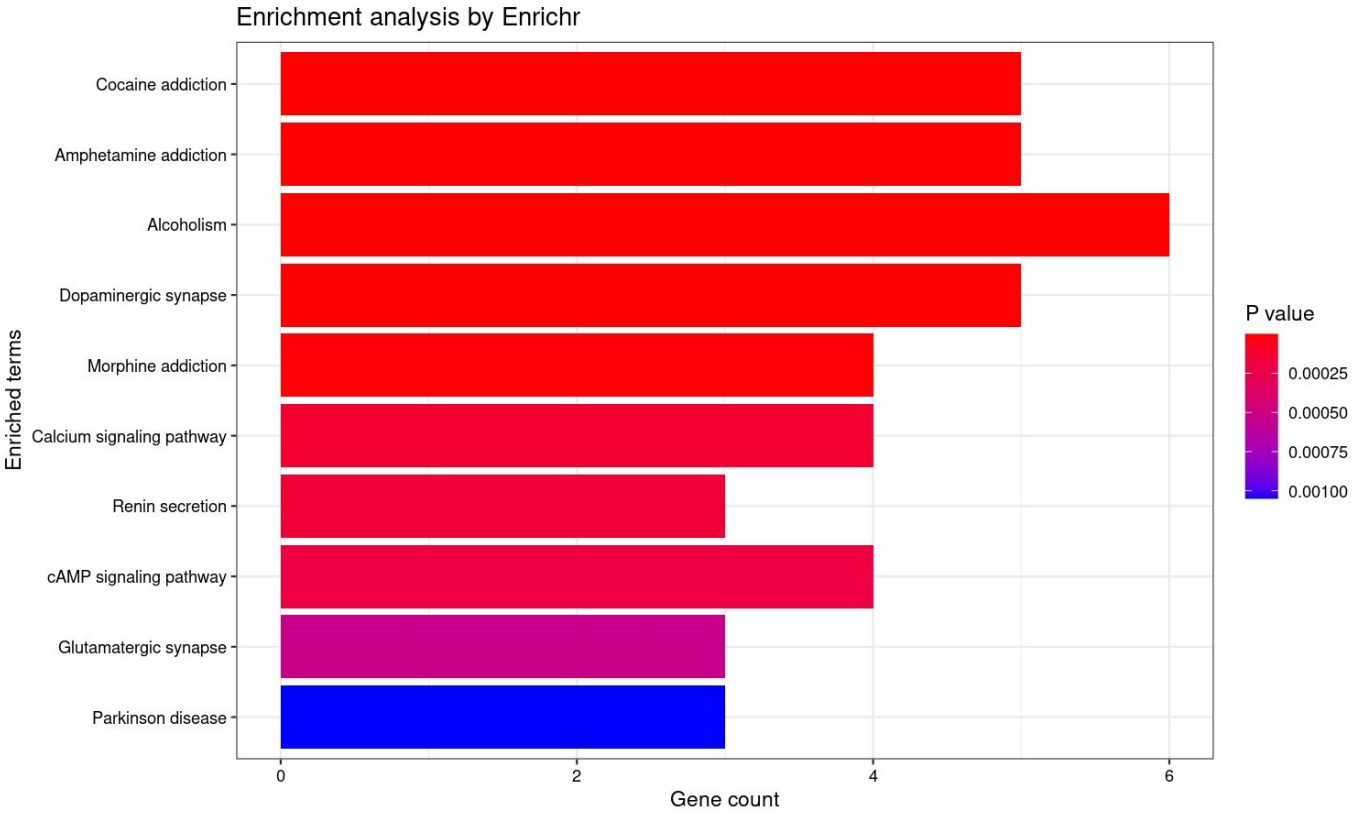
Identification of Significant Modules

- Circos plot used to identify modules significantly enriched with DEGs across multiple brain regions.
- Hub genes from modules of interest were identified.





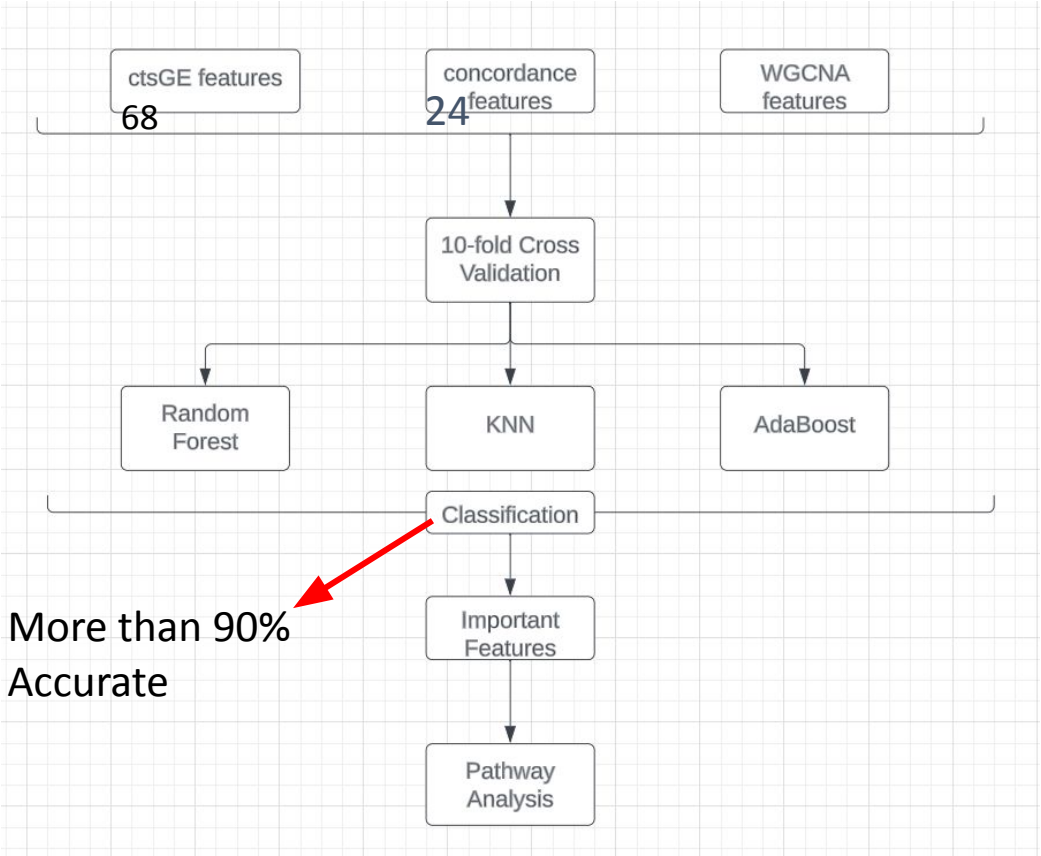
EnrichR Pathways for Cyan Module Hub Genes



Machine Learning Model:

Goal: Predict mouse line of a given sample

Training and Testing Data: gene counts data set



Our 3 Models

RF

KNN

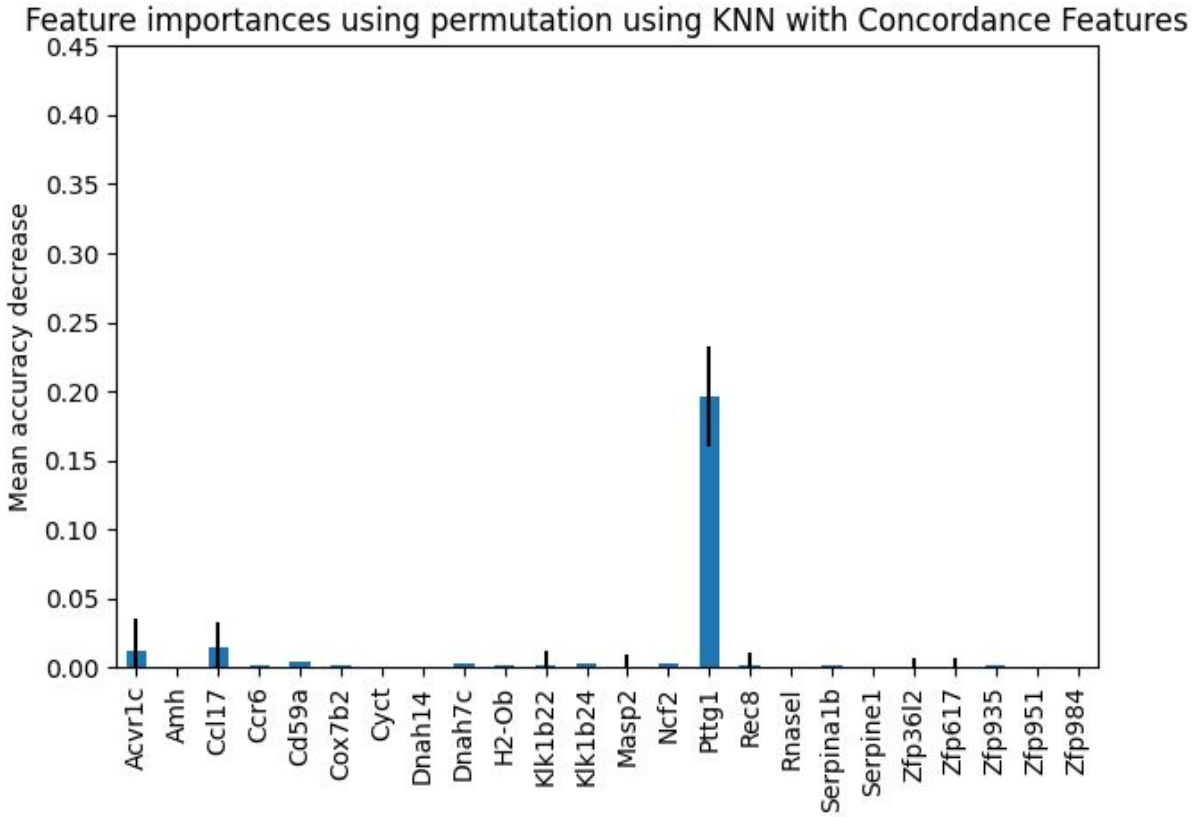
Popular in studies that classify genetic lines

AdaBoost

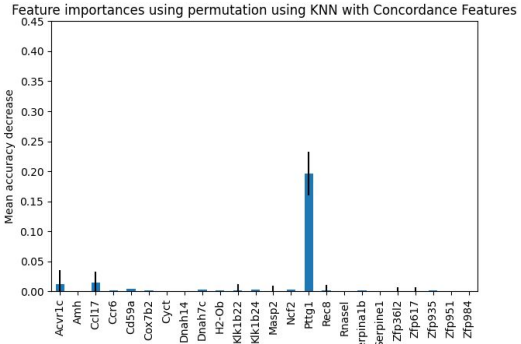
Often used for predicting locations in genomes
or cancer prediction

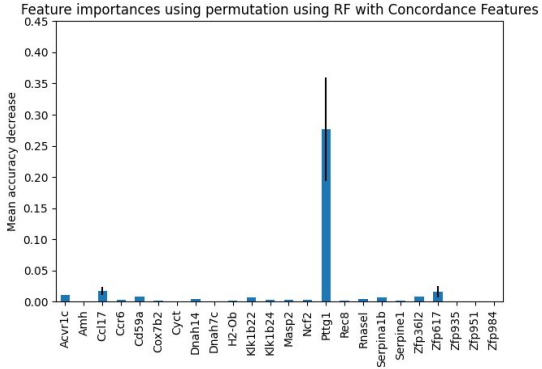
Permutation Importance

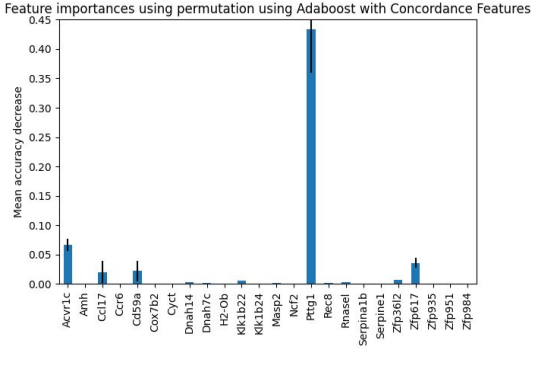
- Remove a feature
- see how the predictions get worse.
- Pttg1, Ccl17, and Acvrlc influenced the classifier performance most (Concordance features)



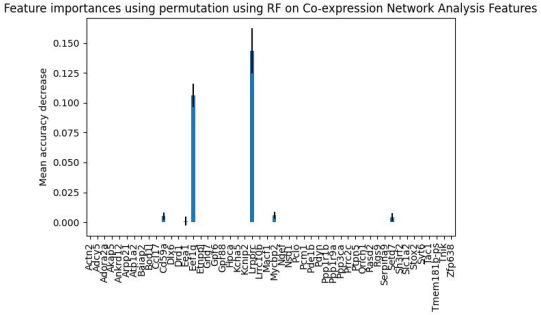
Concordance Feature Importances:

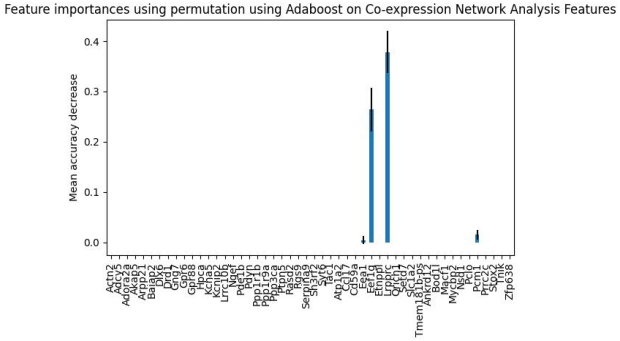


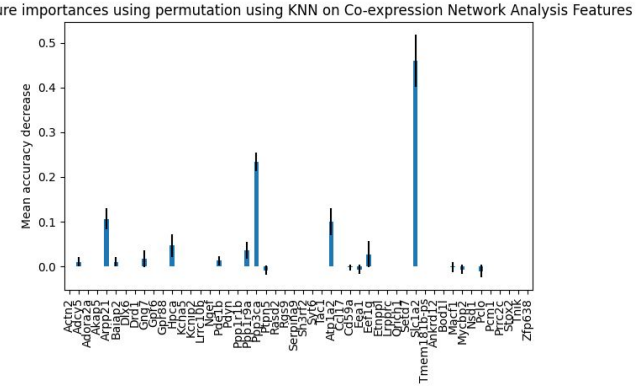




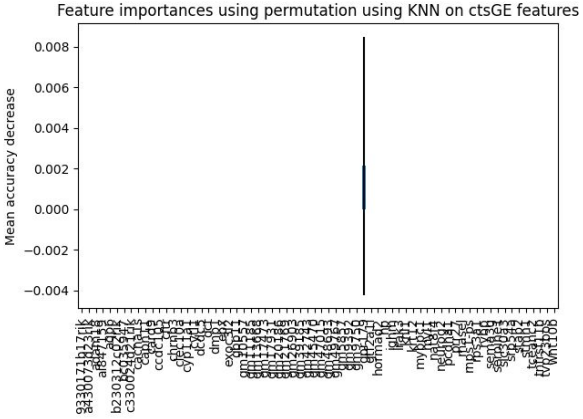
Co-expression Network Feature Importance

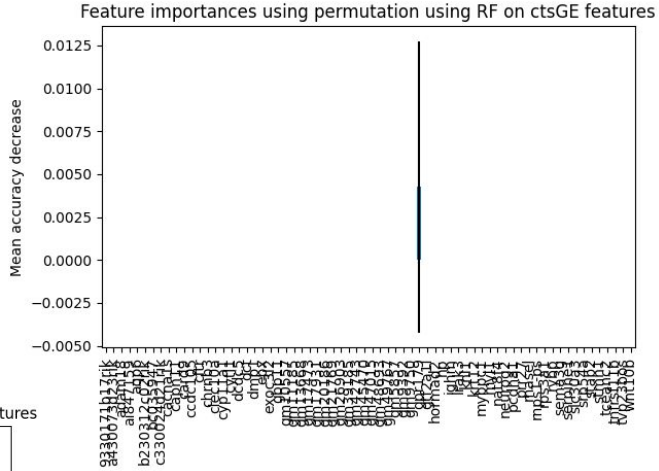


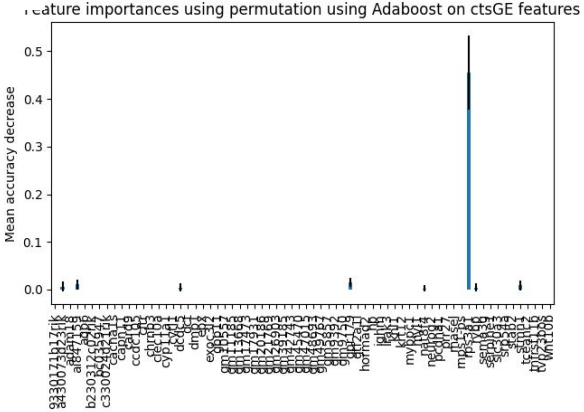




Trajectory Clustering(ctsGE) Model Results







Summary of Important Features

Features	Random Forest Most Signficant Features	Adaboost Most Signficant Features	KNN Most Signficant Features
Concordance Analysis	Pttg1, Acvrlc, Ccl17	Pttg1, Acvrlc, Ccl17	Pttg1, Acvrlc, Ccl17
Co-Expression Network Analysis	Lrpprc, Eef1g, Mycbp2	Lrpprc, Eef1g, Pcm1	Slc1a2, Ppp3ca, Arpp21
ctsGE	GPR179	RPS4A1	GPR179

Conclusion:

1. Resulting genes have biological implications: e.g. Pttg1.

2. Computational methods did great job in classifying samples regardless of brain regions.

   >>2.1 These computational methods can be applied to other tasks that study gene expressions across different sample categories.

3. Caveat: classifier features selected from the same dataset as the model testing.

Future Research

- Extract features with current data, but test the models on a different dataset (to avoid bias).

- Pathway analysis of significant genes from feature permutation

- Incorporate eigengene network as part of the computational methods

Thank you!

Dhivya Arasappan
Michael Keist
Dr. Dayne Mayfield

References:

CDC. (2022, July 6). Deaths from excessive alcohol use in the United States. Centers for Disease Control and Prevention. https://www.cdc.gov/alcohol/features/excessive-alcohol-deaths.html

Ferguson, L. B., Zhang, L., Kircher, D., Wang, S., Mayfield, R. D., Crabbe, J. C., Morrisett, R. A., Harris, R. A., & Ponomarev, I. (2019). Dissecting Brain Networks Underlying Alcohol Binge Drinking Using a Systems Genomics Approach. Molecular neurobiology, 56(4), 2791–2810. https://doi.org/10.1007/s12035-018-1252-0

Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008). https://doi.org/10.1186/1471-2105-9-559

Sharabi-Schwager, M., Or, E., & Ophir, R. (2017). ctsGE-clustering subgroups of expression data. *Bioinformatics (Oxford, England)*, *33*(13), 2053–2055. https://doi.org/10.1093/bioinformatics/btx116