S1803764 December 2021

Bioinformatics 1 – coursework 2
Report

Introduction

Autism spectrum disorder (ASD) is a broad term used to describe a group of neurodevelopmental conditions: autistic disorder, pervasive developmental disorder not otherwise specified (PDD-NOS), and Asperger syndrome.[1] These conditions are all characterized by developmental issues such as differences in communication and social interaction, as well as, restricted and repetitive interests or patterns of behaviour.[2]

The exact cause of ASD is still unknown. Given the complexity of the disorder, and the fact that symptoms and severity vary so greatly from person to person it is highly likely there are multiple causes. The main suspected causes being genetics, and environmental.[3] However, we must note that these environmental factors likely all link back to genetics through epigenetic regulation.[4]

In the 2000s, the advent of high throughput sequencing revolutionized genetic research and allowed researchers to study ASD on a genome-wide level. Sequencing technology quickly identified that the aetiology of ASD was multigenic and highly heterogeneous, with very few of the same pathogenic variants present in a significant percentage of afflicted individuals. Dozens of large-scale genetic studies have been conducted on ASD patients and their families, leading to hundreds of risk genes being identified.[4]

In this project I will be investigating the relevance of different SFARI genes in relation to autism spectrum disorder. I will first look at how the literature is evolving for these genes over the years, and which genes the majority of literature covers. Next I will use methods such as gene ontology enrichment analysis to try learn a bit more about the biological processes, cellular locations, and molecular functions of these SFARI genes. Lastly, I will perform some basic network analysis and extract the largest gene clusters to gain a better insight as to what these groups of genes do and how they interact with each other.

Data & Methods

Data & Resources Used		
Data	**Release date**	**Download date**
SFARI gene list	02-09-2021	02-11-2021
gene2go	23:13, 30-11-2021	01-12-2021
Online Resource	**Access date**	
PubMed	02-12-2021	
PantherDB	29-11-2021	
StringDB	30-11-2021	

DATA:

The SFARI human gene module was created to consolidate the extensive amount of information embedded in peer-reviewed journals into a more readily accessible collection of genetic data pertaining to ASD. This data includes genetic variants across the entire ASD risk spectrum, from rare monogenic causes to common variants of weaker effect. Utilising such a representative dataset of ASD will allow us to identify the most significant risk genes for varying gene-scores, how they function, and how they interact with other genes.

The gene2go module was created to add annotations to genes describing their cellular components, molecular functions, and biological processes. This will allow us to add meaningful information to our analyses that directly tells us how these genes work, and thus allow us to better understand the functions of varying groups of genes.

METHODS:

I used Python throughout this entire project as a means to collect data, automate tasks, and visualise data effectively to maximise interpretability. All of these files will be attached separately for anyone who wants to replicate/understand my data collection and manipulation processes.

Part 1

In part 1, I used Entrez from Biopython to collect all my PubMed data. This enabled me to easily create complex queries, work with this data immediately, and create useful visualisations for analysis. However, the usefulness of my PubMed results is entirely dependant on the quality of my queries. To ensure my queries were valid in the way they filtered the PubMed database I made sure to make use of search field tags so I could customise how each term was to be interpreted. I only made use of the MeSH [MH] and Text Words [TW] tags in my queries. I decided to use MeSH tags on any words I wanted synonyms for (such as “autism”) and Text Words tags for any phrases or acronyms that I did not want to be split up into compound queries. My query structure was as follows:

(<*gene-symbol*>[TW] OR <*gene-ensembl-id*>[TW] OR <*gene-name*>[TW])
AND
(autism[MH] OR autistic[MH] OR ASD[TW] OR "autism spectrum disorder"[TW] OR "pervasive developmental disorder"[TW] OR PDD-NOS[TW] OR PDD[TW] OR asperger[MH])

The first term in this conjunction was used to ensure the paper is related to the given gene, it does this by ensuring that the gene’s name, symbol, or ensembl-ID appears in the text of the paper. The next term in this conjunction ensures that this paper is related to autism, it does this by checking whether any words/terms related to “autism” or ASD are present in the text of the paper. It checks words related to “autism” using the MeSH tags, and it checks for words/terms related to ASD by simply checking if any of the subset of ASD disorders/syndromes are present in the text.

Part 2

For task 1, I mapped all the SFARI genes to their respective NCBI UIDs by iterating through each of the gene-symbols in the SFARI gene list, using Entrez.esearch to query the NCBI gene database, and using Entrez.read to read the results of the query. To be able to perform this query properly I made use of NCBI’s advanced search syntax which allowed me to filter my search by directly setting the gene symbol and the organism:

(<*gene-symbol*>[sym]) AND homo sapiens[Organism]

When performing these queries a few of them returned more than 1 possible ID for a given gene-symbol, however, after further inspection I could see these IDs were typically for closely related genes or deprecated aliases, thus I always used the ID from the first/best hit found in my search.

For task 2, I filtered the gene2go file by taxonomic ID (where tax_id=9606) to ensure that we were only looking through human genes. Once this was done I iterated through every symbol in the SFARI gene list, retrieved it’s NCBI UID using the gene-mappings I found in task 1, and retrieved the Gene Ontology terms for the given gene-symbol.

For task 3, I created text files with all the SFARI gene-symbols for each possible gene-score by filtering the SFARI genes Dataframe.

For task 4, I created tables of the 10 most commonly annotated GO terms for each gene-score. I did this by iterating through the gene-symbols for each gene-score (using results from task 3) and mapped these to GO terms (using the mappings found in task 2). I then put all these GO mappings into a value count algorithm to return the counts of each unique GO_term and then ranked these in order to retrieve the top 10 counts for each gene-score.

For task 5, I used the gene-score files from task 3 and input these into PantherDB's gene list analyser one by one and retrieved the relevant biological process ontology data for all of these gene-scores. I plotted the data for all the gene-scores on the same horizontal bar plot, and normalized their widths so the results for varying gene-scores were more comparable.

For the extension task, I put each of the gene lists found in task 5 into the Reactome pathway analysis tool (using homo sapiens as the organism), and downloaded the analysis report for each of these gene-scores. This allowed me to get a visualisation of genome-scale pathway analysis results and find the most significant pathways for each gene-score.

Part 3

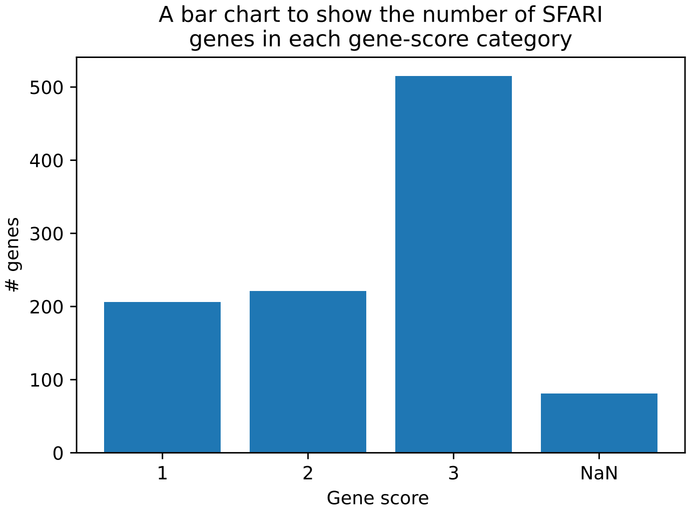
For task 1, I input the text file of gene-score 1 SFARI gene-symbols created in task 3 into StringDB's multiple protein search (using homo sapiens as the organism).

For task 2, I set the clustering for this network to be MCL and downloaded the clusters. I then passed the genes in the top 2 clusters of this network into PantherDB's gene list analyser separately to retrieve their pathway ontology data. I then plotted all this data onto a single horizontal bar plot, and normalized the cluster widths so the results for the different clusters were more comparable.

Results

Part One – Autism Literature

1) Plot a bar chart of the number of genes in each SFARI gene-score category



2) Rank the genes by 'number-of-reports' and find the top 5 SFARI genes that are in gene-score category 1

Top 5 reported SFARI genes with gene-score 1		
Rank	**Gene symbol**	**Number of reports**
1	NRXN1	94
2	SHANK3	92
3	MECP2	90

4	SCN2A	75
5	SCN1A	68

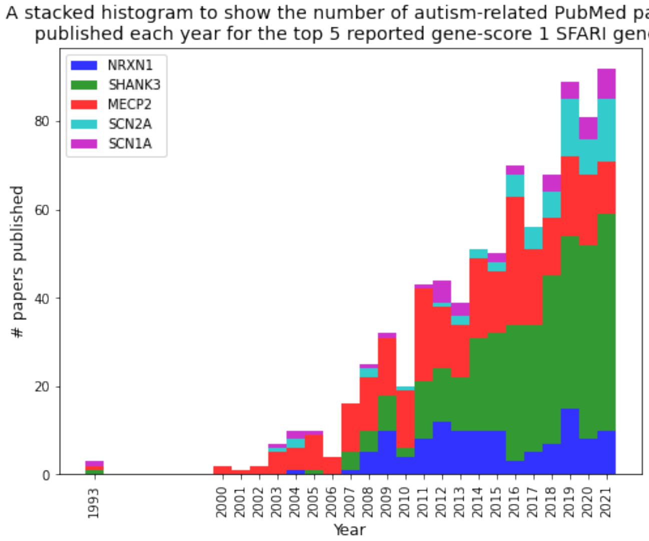
3) For each of these genes find the number of papers in PubMed that include the gene AND are related to Autism

# ASD-related PubMed articles for the top 5 reported SFARI genes with gene-score 1	
Gene symbol	**# articles published**
NRXN1	119
SHANK3	331
MECP2	261
SCN2A	64
SCN1A	40

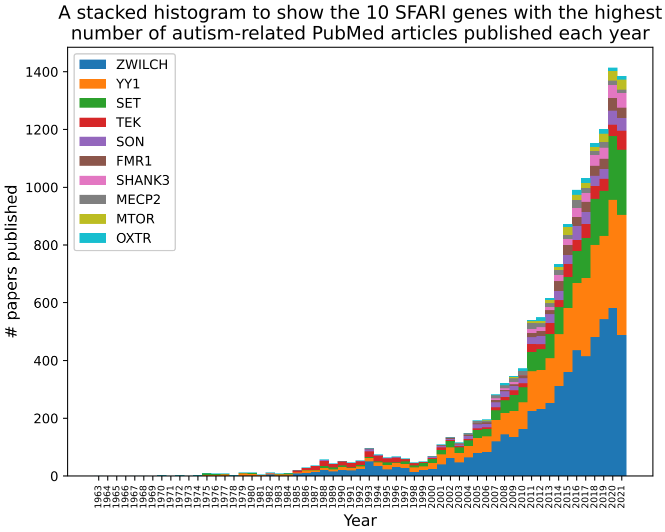
4) From this data fill a table with genes as rows and paper count by year as column

# autism-related PubMed articles published each year for the top 5 reported gene-score 1 SFARI genes																						
gene symbol	**# articles published**																					
	1993	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**
NRXN1	-	-	-	-	1	-	-	1	5	10	4	8	12	10	10	10	3	5	7	15	8	10
SHANK3	1	-	-	-	-	1	-	4	5	8	2	13	12	12	21	22	31	29	38	39	44	49
MECP2	1	1	2	5	5	8	4	11	12	13	13	21	14	12	18	14	29	17	13	18	16	12
SCN2A	-	-	-	1	2	-	-	-	2	-	1	-	1	2	2	2	5	5	6	13	8	14
SCN1A	1	-	-	1	2	1	-	-	1	1	-	1	5	3	-	2	2	-	4	4	5	7

5) Plot a single stacked histogram displaying the data from the table



Extension 1) Extend this analysis from part 5 to all the SFARI genes



Part Two – Autism Genes

1) Map the gene-symbol for every gene in the SFARI gene list to an NCBI UID

Results summary statistics	
# genes	**# gene mappings found**
1023	1020

There were 3 gene symbols where the query did not return any hits at all: MSNP1AS, RP11-1407O15.2, RPS10P2-AS1. After further inspection of these genes’ SFARI profiles (added as hyperlinks on each of these genes), when clicking “Entrez Gene” in the “External Links” section it reroutes us to the main NCBI search page rather than the relevant NCBI gene profile thus it is safe to assume that missing these mappings was not an error on my part. I will now conduct some research to try and understand why these genes do not exist in NCBI’s gene database, and whether any close relatives are present.

Unmapped SFARI gene		Closest matching NCBI gene	
Gene symbol	**Gene name**	**Gene symbol**	**Gene name**
MSNP1AS	Moesinpseudogene 1, antisense	MSNP1	Moesin pseudogene 1
RP11-1407O15.2	-	-	-
RPS10P2-AS1	Ribosomal protein S10 pseudogene 2 anti-sense 1	RPS10	Ribosomal protein S10

To map each of these genes to their closest match I first searched their gene symbols in NCBI’s gene database if no outputs were given I then tried to search their names. This method was successful for MSNP1AS and RPS10P2-AS1, but not RP11-1407O15.2. After looking at the SFARI profile for RP11-1407O15.2 I could see that it was categorised as a rare single gene mutation, and had no molecular function description indicating this gene is most likely not present on NCBI as it is a rare mutation.

Excluding RP11-1407O15.2, it seems the main difference between the unmapped SFARI genes and the closest NCBI gene matches is that these SFARI genes are both antisense pseuodgenes, and are both in the “Genetic Association, Functional” SFARI genetic category.

2) Using the gene2go file from NCBI find the Gene Ontology terms that have been annotated to all of the SFARI genes

After inspection of these results given some genes matched with multiple GO terms there were some repeated terms associated with some genes. Thus I thought it would be useful to represent the statistics of our results with and without GO term repetition.

Results summary statistics			
# of genes	# of genes with GO terms	Average # GO terms per gene (with repetitions)	Average # GO terms per gene (with no repetitions)
1023	1015	27.15	23.55

As seen from the table above we can see that 8 of our SFARI genes could not be mapped to GO terms. We know that 3 of these genes were not mapped to a GO term because they could not be mapped to NCBI UIDs in task 1 (MSNP1AS, RP11-1407O15.2, RPS10P2-AS1) and thus could not be queried against the gene2go file. However, the 5 remaining unmapped genes (CCSER1, FAM47A, METTL26, MSANTD2, PTCHD1-AS) did all have valid UIDs so we should try understand why these genes were not present in the gene2go dataset.

One common feature of all these 5 unmapped genes is that they are all rare single genetic mutations, so it is possible these genes were not included in the GO database due to their rarity.

3) Split the genes up into three lists by their SFARI gene-score

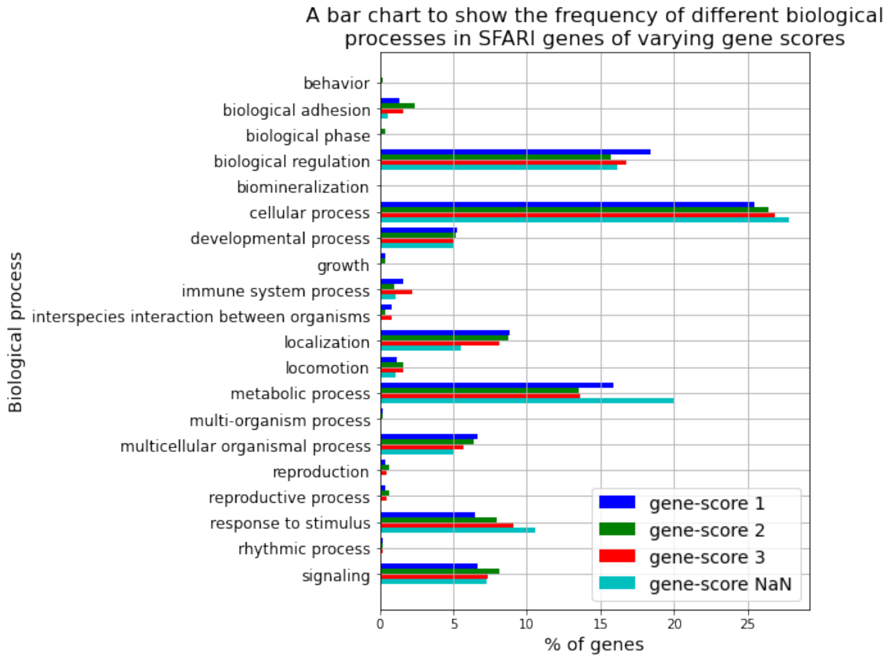
Number of SFARI genes for each unique gene score	
Gene score	Number of genes
1	206
2	221
3	515
NaN	81

4) Create tables of the 10 most commonly annotated terms for each gene list. The tables should have the following columns: GO term ID, GO term Description, GO term count

10 most commonly annotated terms for each SFARI gene-score list				
Gene score	Rank	GO term ID	GO term description	GO term count
1	1	GO:0005739	mitochondrion	14
	2	GO:0005634	nucleus	13
	3	GO:0005515	protein binding	10
	4	GO:0009507	chloroplast	5
	5	GO:0005886	plasma membrane	5
	6	GO:0003674	molecular_function	5
	7	GO:0006355	regulation of transcription, DNA-templated	4
	8	GO:0005737	cytoplasm	4
	9	GO:0008150	biological_process	4
	10	GO:0005576	extracellular region	3
2	1	GO:0005634	nucleus	19
	2	GO:0005739	mitochondrion	13
	3	GO:0003674	molecular_function	10

	4	GO:0009507	chloroplast	9
	5	GO:0003700	DNA-binding transcription factor activity	5
	6	GO:0008150	biological_process	5
	7	GO:0005737	cytoplasm	5
	8	GO:0005886	plasma membrane	4
	9	GO:0005794	golgi apparatus	4
	10	GO:0005515	protein binding	4
3	1	GO:0005634	nucleus	51
	2	GO:0009507	chloroplast	23
	3	GO:0005886	plasma membrane	18
	4	GO:0008150	biological_process	15
	5	GO:0005737	cytoplasm	14
	6	GO:0003700	DNA-binding transcription factor activity	13
	7	GO:0005515	protein binding	13
	8	GO:0003674	molecular_function	12
	9	GO:0005739	mitochondrion	11
	10	GO:0005576	extracellular region	8
NaN	1	GO:0005634	nucleus	7
	2	GO:0005739	mitochondrion	4
	3	GO:0003700	DNA-binding transcription factor activity	4
	4	GO:0005737	cytoplasm	3
	5	GO:0006355	regulation of transcription, DNA-templated	3
	6	GO:0008150	biological_process	3
	7	GO:0009773	photosynthetic electron transport in photosystem I	2
	8	GO:0005515	protein binding	2
	9	GO:0000976	transcription cis-regulatory region binding	2
	10	GO:0005886	plasma membrane	2
Combined	1	GO:0005634	nucleus	90
	2	GO:0005739	mitochondrion	42
	3	GO:0009507	chloroplast	38
	4	GO:0005515	protein binding	29
	5	GO:0005886	plasma membrane	29
	6	GO:0005515	molecular_function	28
	7	GO:0008150	biological_process	27
	8	GO:0005737	cytoplasm	26
	9	GO:0003700	DNA-binding transcription factor activity	24
	10	GO:0005576	extracellular region	14

5) Take the three lists of UIDs created above and use the PantherDB tool to retrieve data to create a bar chart that displays the biological processes for each SFARI gene for varying gene-scores



Extension 1) Explore other pathway analysis tools and websites such as Reactome

Top 3 most significant pathways for each SFARI gene score from Reactome							
Gene score	**Pathway name**	**Entities**				**Reactions**	
		found	**ratio**	**p-value**	**FDR***	**found**	**ratio**
1	Chromatin organization	27 / 256	0.018	2.81e-13	1.54e-10	32 / 85	0.006
	Chromatin modifying enzymes	27 / 256	0.018	2.81e-13	1.54e-10	32 / 85	0.006
	Neuronal system	36 / 489	0.034	9.52e-13	3.47e-10	88 / 216	0.016
2	Neuronal system	26 / 489	0.034	3.9e-6	0.002	92 / 216	0.016
	Axon guidance	29 / 585	0.041	4.07e-6	0.002	100 / 298	0.022
	Post-transcriptional silencing by small RNAs	4 / 7	4.91e-4	1.21e-5	0.003	3 / 3	2.21e-4
3	Neuronal system	57 / 489	0.034	4.53e-11	6.62e-8	135 / 216	0.016
	Transmission across chemical synapses	40 / 343	0.024	3.88e-8	2.84e-5	107 / 163	0.012
	Protein-protein interactions at synapses	16 / 93	0.007	5.37e-6	0.002	24 / 33	0.002
NaN	DSCAM interactions	3 / 11	9.79e-4	1.41e-4	0.099	3 / 6	4.42e-4
	Toxicity of botulinum toxin type G (botG)	2 / 4	3.56e-4	6.13e-4	0.133	4 / 5	3.68e-4
	Ephrin signalling	3 / 19	0.002	6.91e-4	0.133	11 / 11	8.1e-4
Combined	Neuronal system	120 / 489	0.034	1.11e-16	2.1e-13	174 / 216	0.016
	Protein-protein interaction at synapses	39 / 93	0.007	2.55e-15	2.41e-12	29 / 33	0.002
	Neurexins and neuroligins	31 / 60	0.004	8.66e-15	5.46e-12	19 / 19	0.001

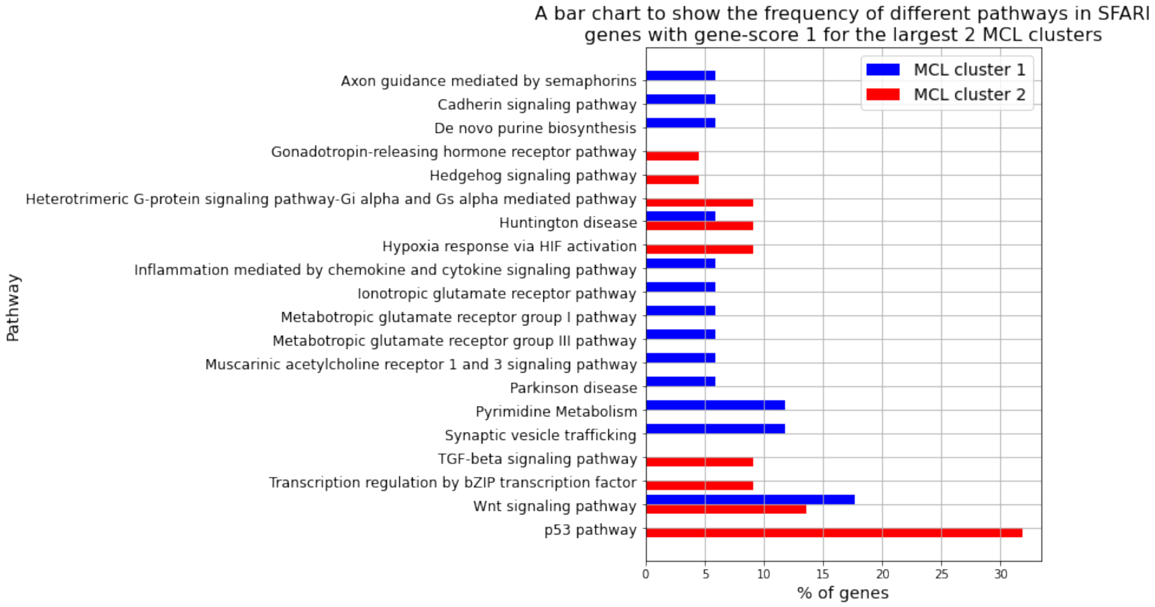
Part Three – Autism Gene Networks

1) Visualise the protein-protein interaction network for gene-score 1 SFARI genes using the STRING website (https://string-db.org/)

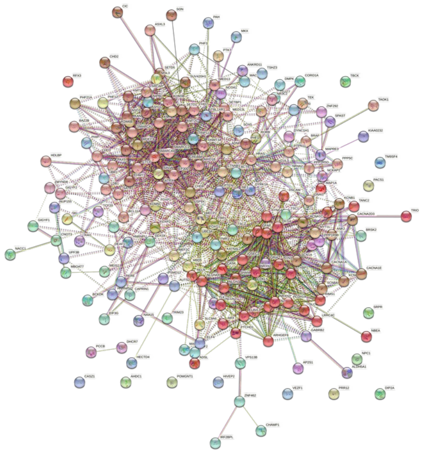
Protein-protein interaction network statistics for gene-score 1 SFARI genes		
# of nodes	# of edges	Average node degree
204	1376	13.5

***Note:** when performing a multiple protein search on https://string-db.org/ some reason some of my genes do not get matched. I input 206 genes, however, the clustered output reports there only being 204 nodes (where each node represents a gene). After reviewing my search’s gene matches it seemed to only not be able to match a single gene (85358) where it gave me the following error: "*Sorry, STRING found no proteins by this name in Homo sapiens*". However, I know this ID is valid as shown here: https://www.ncbi.nlm.nih.gov/gene/?term=85358%5Buid%5D.

2) Get all the SFARI genes in the 2 largest MCL clusters from our protein-protein interaction network and input these into PantherDB to get their pathway ontology data for analysis

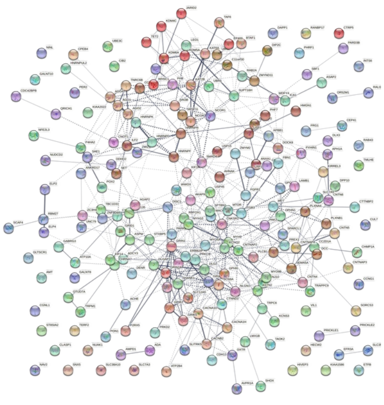


3) Display the MCL clustered protein-protein interaction network

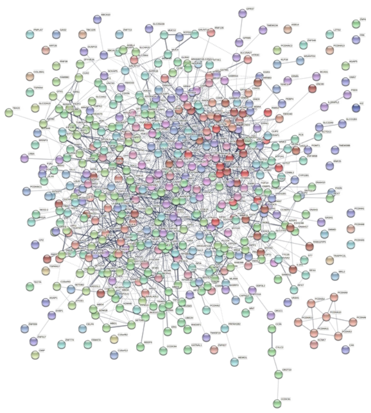


Extension 1) Repeat this analyses with other gene-score restricted lists

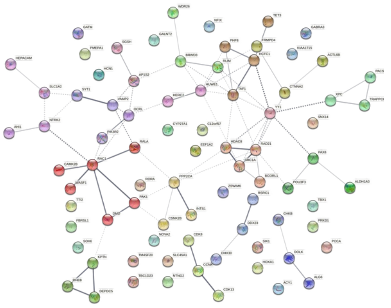
Gene score 2:

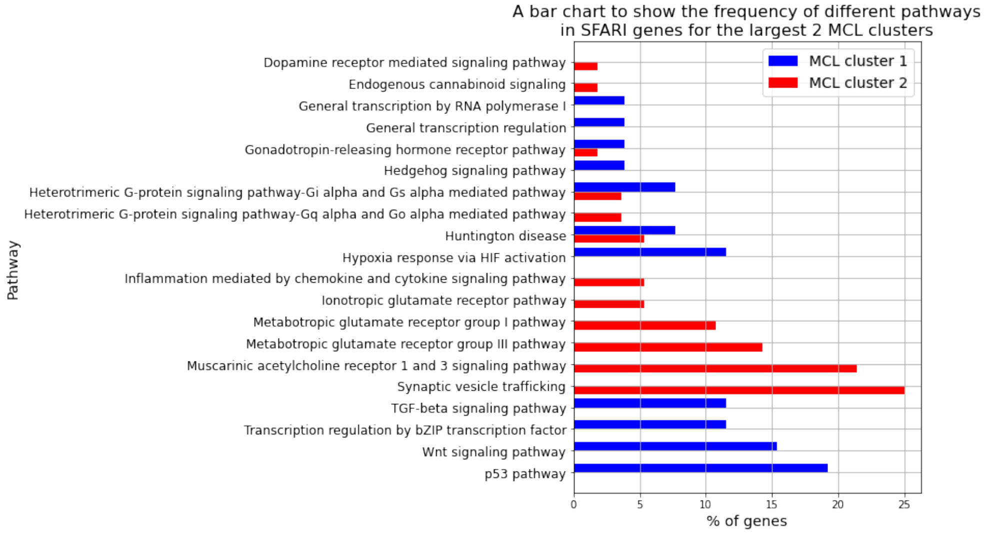


Gene score 3:



Gene score NaN:




A bar chart to show the frequency of different pathways
in SFARI genes for the largest 2 MCL clusters
MCL cluster 1
MCL cluster 2
Pathway
Dopamine receptor mediated signaling pathway
Endogenous cannabinoid signaling
General transcription by RNA polymerase I
General transcription regulation
Gonadotropin-releasing hormone receptor pathway
Hedgehog signaling pathway
Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway
Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway
Huntington disease
Hypoxia response via HIF activation
Inflammation mediated by chemokine and cytokine signaling pathway
Ionotropic glutamate receptor pathway
Metabotropic glutamate receptor group I pathway
Metabotropic glutamate receptor group III pathway
Muscarinic acetylcholine receptor 1 and 3 signaling pathway
Synaptic vesicle trafficking
TGF-beta signaling pathway
Transcription regulation by bZIP transcription factor
Wnt signaling pathway
p53 pathway
0
5
10
15
20
25
% of genes

*Note: this shows the pathways for the 2 largest clusters for the protein-protein interaction network across **all** SFARI genes*

Discussion

Part 1

As shown by my bar graph for task 5 in part 1, it is evident that out of 5 most reported gene-score 1 SFARI genes SHANK3 is the most representative of autism. This is not only due to the fact that this gene has the most autism-related PubMed papers published, but also due to the fact that the number of papers continues to increase each year indicating this gene is still relevant. However, as shown by the extension task I completed ZWILCH has the most autism-related PubMed articles throughout all the SFARI genes, this was very surprising especially considering ZWILCH has gene-score 3.

Part 2

From task 4, we can see that the majority of annotated GO terms are cellular components. From task 5 in part 2, we can see that the for the majority of genes across all gene-scores cellular process was the most common biological process annotation. Thus from these tasks we can deduce that the main functions of these autism genes all relate to cellular processes which refers to any process carried out at the cellular level, or cellular communication.

From the extension task it is evident that the neuronal system is the most common pathway found amongst all our SFARI genes across all gene-scores. This neuronal system refers to the chemical and electrical synapses that allow for neuron transmission inside our brain. Thus in conjunction with our findings in task 4 and 5 we can see that it is likely the main cellular process behind all of our GO annotations is synaptic transmission. This in contrast for gene-score 1 SFARI genes, where chromatin organization and chromatin modifying enzymes make up the top 2 pathways, however, we must note that neuronal system was still the 3rd most significant pathway for this gene score. Chromatin remodelling is the rearrangement of chromatin from a condensed state to a transcriptionally accessible state, thus these chromatin-related pathways indicate that the gene-score 1 SFARI genes must have an effect on the gene expression for DNA.[5] The structure of chromatin is regulated by enzymes which add/remove chemical tags on DNA and histone proteins, thus given these gene-score 1 SFARI genes got a significant amount of annotations for the "chromatin modifying enzymes" pathway this is likely due to these SFARI genes causing mutations in these enzymes causing disruption to chromatin gene expression.[5]

Part 3

From tasks 2 and 3, we can see that the majority of genes in cluster 1 are used for the Wnt signalling pathway. Broadly speaking, in the brain Wnt signaling can be split into two main pathways: "canonical" signaling that results in the stabilization of the protein β-catenin which upon stabilization, can exert functions at the plasma membrane or in the nucleus and can act as a transcription factor that modulates the expression of target genes, and "non-canonical" β-catenin-independent signaling.[6] We can see that the function of this first pathway relates to our SFARI genes given our findings from part 2 task 4 as we can see that the nucleus and plasma membrane are the 2nd and 5th most annotated GO terms across all gene-score 1 SFARI genes. Interestingly, many of the proteins in both of these signaling pathways localize to the synapse and play important functions in synaptic growth and maturation.[6] .Thus we can imagine that mutations of the proteins in these pathways could affect a person's synaptic growth and maturation, which may be the root cause of symptoms behind a neurodevelopmental disorder such as ASD.

References

1. CDC. What is Autism Spectrum Disorder? Centers for Disease Control and Prevention. Published March 25, 2020. Accessed November 30, 2021. https://www.cdc.gov/ncbddd/autism/facts.html

2. Cherney K. Everything You Need to Know About Autism Spectrum Disorder (ASD). Healthline. Published November 3, 2021. Accessed November 30, 2021. https://www.healthline.com/health/autism#TOC_TITLE_HDR_1

3. Autism spectrum disorder - Symptoms and causes. Mayo Clinic. Published 2018. Accessed November 30, 2021. https://www.mayoclinic.org/diseases-conditions/autism-spectrum-disorder/symptoms-causes/syc-20352928#:~:text=Genetics.,risk%20of%20autism%20spectrum%20disorder.

4. Rylaarsdam L, Guemez-Gamboa A. Genetic Causes and Modifiers of Autism Spectrum Disorder. *Frontiers in Cellular Neuroscience*. 2019;13. doi:10.3389/fncel.2019.00385

5. Giorgia Guglielmi. Autism's link to chromatin remodeling, explained | Spectrum | Autism Research News. Spectrum | Autism Research News. Published July 6, 2021. Accessed December 3, 2021. https://www.spectrumnews.org/news/autisms-link-to-chromatin-remodeling-explained/

6. Kwan V, Unda BK, Singh KK. Wnt signaling networks in autism spectrum disorder and intellectual disability. *Journal of Neurodevelopmental Disorders*. 2016;8(1). doi:10.1186/s11689-016-9176-3