Investigation of existing data on ASD demonstrates variability in research methods.

B217754-2022

Contents

Introduction

Autism Spectrum Disorder, more commonly known as ASD, is a disorder presenting in a range of severities. This complex disease's etiology is poorly elucidated but progressively expanding year on year .

SFARI Gene is a curated database concerning genes with a relationship to ASD. It scores these genes based on evidence in the literature on likely causative mutations [14] and categorises them into levels of 'gene-score' with a score of 3 having the weakest (but potentially present) chance of being involved in ASD development and score of 1 having the strongest. There is also another metric to rank the genes, known as the EAGLE score, which evaluates whether the gene is associated with ASD specifically over other neurodevelopmental phenotypes (it has been shown that SFARI genes are more tightly regulated than other genes with neuronal function. [14]

This paper is a demonstration of data exploration, comparing different information mining methods when used to pursue similar information.

Data & Methods

Part 1: ASD literature mining

The SFARI gene list was downloaded from SFARI's webpage. Microsoft Excel was used to find the counts of gene per gene-score category and plot this information on a bar chart. The genes with a 'gene-score of 1' and '5 highest number of reports' according to the SFARI database were then identified and summarised using Microsoft Excel. A Python3 script was written to query PubMed iteratively with the following query: *"Autism Spectrum Disorder OR Autism [MH] AND {gene}'|efilter -mindate {year} -maxdate {year}"*, varying the fields *{gene}* and *{year}* in each search to parse through list of genes and years of publication of interest. The list of query genes was dictated by the previously

identified genes of interest when filtering the SFARI gene list by report-count. Searches were filtered with a code of command-line pipe including *efilter* with year, in range 1995-2023. The starting point of 1995 was chosen to predate the completion of the human genome project and to capture any changes in trends of genomics publications at major milestones of Pyosequencing[1] and Ion-Torrent sequencing [2] as well as others since [3].

The data was collected for each year & gene and used to generate a stacked histogram in Microsoft Excel. Subsequently, an estimation of literature concerning genes which are more specific to ASD was performed by using the genes with an EAGLE [4] score > 0 as the query list of genes. The result was a chart of publications over years for EAGLE-ranked genes.

Part 2: ASD-related gene functions

Python3 script used to search for an NCBI UID for each SFARI gene and use this for future searching. This process included downloading the human gene data subset (Size: 4.0M) and the gene2go Accession ID translation table (Size: 26M) from the NCBI FTP [5] . Some Gene symbols gave results which were incongruous with the rest of the result dataset, so these were investigated directly by consulting the SFARI webpage. Once UIDs were obtained, the Gene Ontology (GO) terms annotated to the SFARI genes were retrieved programmatically. To identify the functions and other associations recorded though Gene Ontology Resource (GO), the SFARI genes were paired with GO ID numbers by using previously mapped UIDs. The most commonly annotated GO terms for each level of gene-score were then programmatically synthesised.

The list of the three levels of gene UIDs were passed to the PantherDB [6-8] and ran "Functional classification viewed in graphic charts" by selecting the "Bar chart" display option, with other default settings and no species selected. The resulting page provided information on gene function in a downloadable format once ontology was changed to "Biological Process". The data was converted to csv, and presented using Microsoft Excel. Enrichment analysis was carried out on the most commonly annotated GO term in our list, assessing the probability of its prevalence in our data by chance against what was observed.

Part 3: ASD-related gene networks and interactions

The protein-protein interaction database STRING [9] was used to assess the known interactions between the SFARI genes and highlight any other genes of interest. The lists of NCBI UIDs divided by gene-score were provided separately to STRING with default settings and species set as 'homo sapiens'. Select statistics of the resulting STRING networks were recorded and are presented in the Results section.

Each gene-score-separated list of genes was further stratified with STRING MCL clustering, the corresponding data was downloaded and sorted by size of cluster. The two largest clusters of genes were re-exported for analysis in PANTHERDB to act as a comparison to the functional classification of genes related to ASD in Part 2.

In the case that reference is not made to the version number of a particular software, assume that versions were latest publicly available releases as of December 1st 2022.

Software same	Version or Release
Python3	3.8.10
Microsoft Excel	version 2210
PANTHERDB	Release 17.0
SFARI GENE	'Latest release 2018 Q4'
STRING	Version 11.5

Results

[Part 1] The genes linked the most strongly to ASD are not the most frequently studied

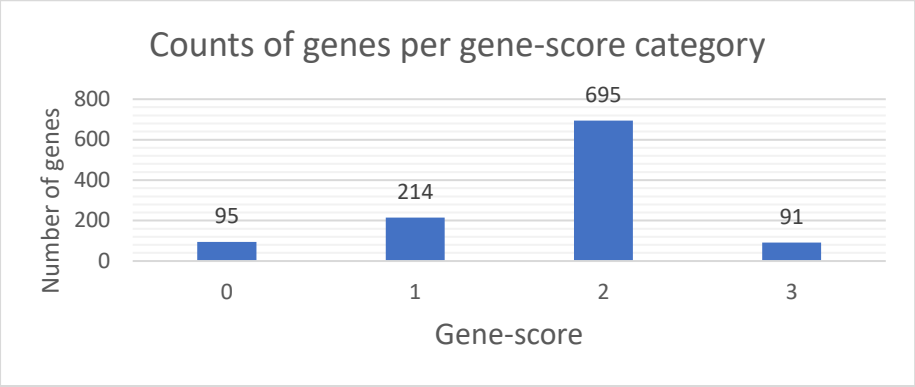


Figure 1 Bar chart of the number of genes in each SFARI gene-score category (Task1.1)

The count of genes by gene-score shows that 695 genes in the SFARI gene list were 'strong candidates' to have an association with autism risk. Meanwhile, there were fewer genes classed as having the strongest body of evidence of a link to ASD (gene score 1). This immediately highlights the where further evidence could be gathered, as gene-score allocation is reliant on not only the frequency of times a gene appears in literature, but also the type of literature it appears in.

gene-symbol	gene-name	chromosome	genetic-category	gene-score	eagle	number-of-reports
SHANK3	SH3 and multiple ankyrin repeat domains 3	22	Rare Single Gene Mutation, Syndromic, Genetic Association, Functional	1	74.85	120
MECP2	Methyl CpG binding protein 2	X	Rare Single Gene Mutation, Syndromic, Functional	1		107
NRXN1	neurexin 1	2	Rare Single Gene Mutation, Syndromic, Genetic Association, Functional	1	143.8	100
SCN2A	sodium channel, voltage-gated, type II, alpha subunit	2	Rare Single Gene Mutation, Syndromic, Functional	1	109.3	96
SCN1A	sodium channel, voltage-gated, type I, alpha subunit	2	Rare Single Gene Mutation, Syndromic, Genetic Association, Functional	1		84

Figure 2 From the SFARI genes with a gene-score of 1, these are the top 5 genes with the most publications according to the SFARI database. (Task 1.2)

This table shows the count of literary evidence for the most-reported genes contained in the SFARI gene list according to SFARI. The functions of these genes are related to foundational cell-cell signalling processes, and logically would cause systemic changes if mutated into dysfunction.

Using SFARI's rankings above as a guide in the search, PubMed was programmatically queried for the highly-reported genes by year (full results are in the appendix). Over time, there was an increase in papers published in line with the uptake of NGS use in research. but the total counts for each gene are below.

Gene Symbol	Total Count of papers on Pubmed
SHANK3	427
MECP2	317
NRXN1	157
SCN2A	88
SCN1A	50

Figure 3 Number of papers in PubMed that include the gene symbol and are related to Autism (Task 1.3). Totals are taken from a table of papers per year as seen in Appendix (Task 1.4).

These values show that the highlighted genes appear in the literature in with the same rankings as suggested by SFARI, however the number of reports including these genes does vary from the SFARI Gene database. However, as discussed later, the query methods for PubMed may have had room for adjustment to make results more refined, which could have led to a count of reports closer to that presented by SFARI for the genes.

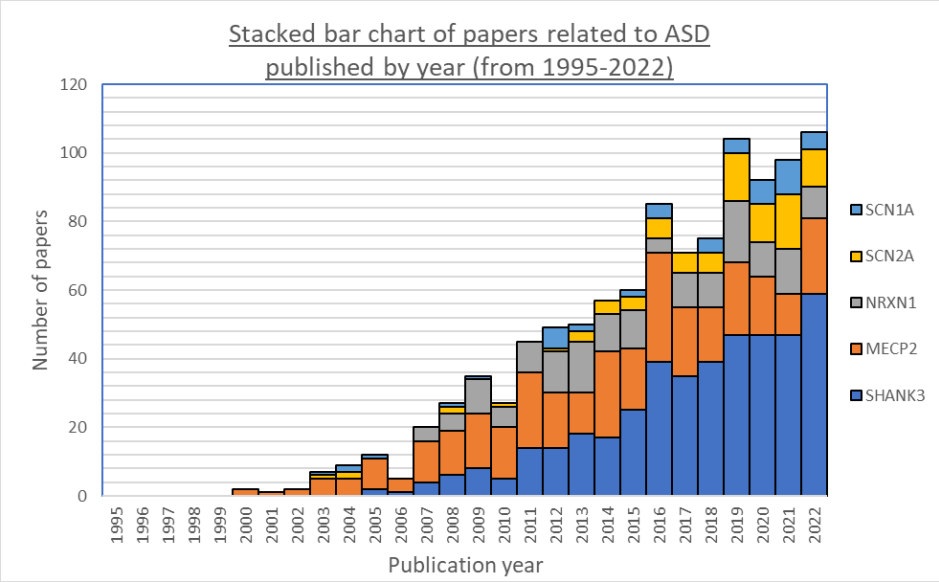


Figure 4 Stacked bar chart of papers on genes related to ASD, published by year (representing the data in appendix). A stacked column chart was plotted instead of a histogram to preserve the representation of papers by year as well as by gene.

As an extension of investigation, a graph of papers published on all SFARI genes with an eagle score above 0 since 2013 was plotted to show trends in the research of the genes which are classed as related to ASD with a different metric. This was done because the top 5 highlighted genes by SFARI may not be representative of the truly most studied genes related to ASD (we selected these based off of SFARI's own report counts). Most genes were not frequently cited, however the gene 'SON' had 423 results in 2022. Due to limitations in data presentation, this record was removed from the graph.

Investigation of existing data on ASD demonstrates variability in research methods.

B217754-2022

[Part 2] ASD-related gene functions

The table of Gene symbols mapped to UIDs was unsuitable for presentation in this document, however the data was used in the generation of GO term analysis later. Whilst mapping gene symbols to NCBI UIDs, the symbols below returned multiple or no UIDs when mapping programmatically. Manual inspection found that the dimensions for the dataframe to contain the results were not flexible enough to handle exceptions. Therefore, this list of symbols was addressed in a new segment of code, and their values appended to the full list of mapped UIDs in the file 'NCBI_UID_list.csv'. If running the code afresh, they are also saved as 'NCBI_tricky_list.csv'.

Gene Symbol	GeneID	Explanation for mapping issues
DUPLICATE_MEMO1_1	7795	MEMO1 is used as a gene symbol twice in the SFARI gene table. This GeneID is for the MEMO1 *Methylation modifier for class I HLA* and appears unrelated to ASD or the SAFRI gene list so was struck from further investigation
DUPLICATE_MEMO1_2	51072	MEMO1 *mediator of cell motility 1* was updated on late October 2022 [10] and appears to the correct entry for the SAFRI gene list.
MSNP1AS	NO VALUE FOUND	Although this gene has a gene-score of 2 (Strong evidence of relationship to ASD), it does not have a UID because it is a non-coding RNA encoded by the antisense strand. Literature cited on SFARI suggests that the transcription product's high sequence identity with that of MSN which plays a role in neuronal architecture control [11] .
RP11-1407O15.2	NO VALUE FOUND	This entry in SAFRI represents a set of observed single point mutations on the nucleotide level from a study in 2017 [12]
RPS10P2-AS1	NO VALUE FOUND	Another non-coding RNA transcript existing in an intronic region near an SNP that has a known association with ASD [13]

The 10 most commonly annotated GO terms for the SFARI geneset, grouped by gene-score, is presented in the Appendices.

Task 2.5 Biggest two Clusters from STRING, ran in PANTHER for biological process/function identification. There were many genes with no identifiable function in PANTHER:

Summary statistics for largest cluster		Summary statistics for second largest cluster	
Count of Biological function categories	Genes in Unclassified PANTHER category	Count of Biological function categories	Genes in Unclassified PANTHER category
17	62.10%	38	70.50%

Enrichment analysis showed that the most commonly annotated GO term in the list of SFARI genes was GO:0005515: 'Protein binding', appearing 1.18 more frequently than was expected by chance. This suggests that by its association to the genes related to ASD as categorised by the SFARI list, some modulation of the protein binding is an important aspect of ASD.

"The core of PANTHERDB is the library of phylogenetic trees." [6-8], meaning that the biological function of a query protein is assigned by association with other proteins (relationships including protein family/subfamily, orthologs, paralogs and pathways dictated by Reactome). PANTHERDB also uses information from gene-associated GO terms to return information about the Biological function of a gene, so there is influence between the two sources of information in Part 2 of this project.

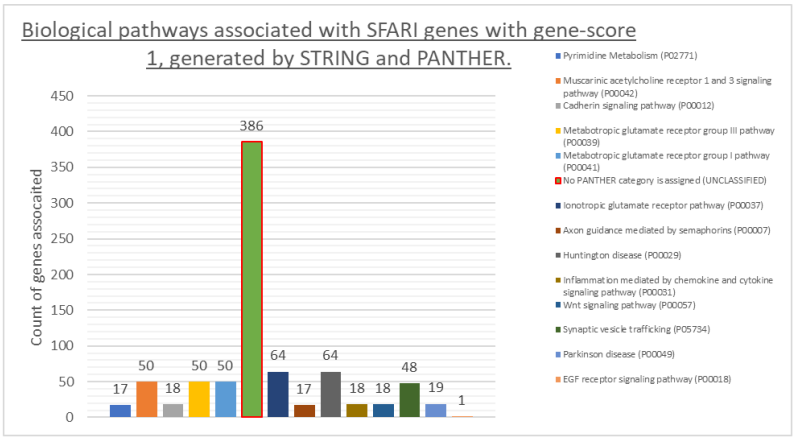
[Part 3] ASD-related gene networks ad interactions

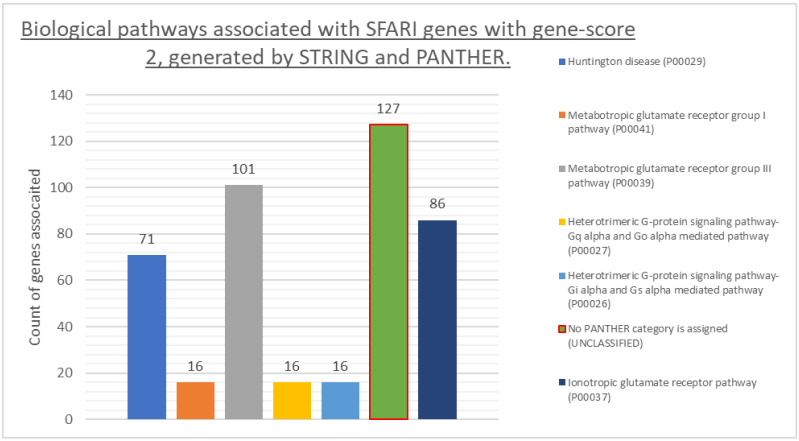
(Task3.1) STRING analysis

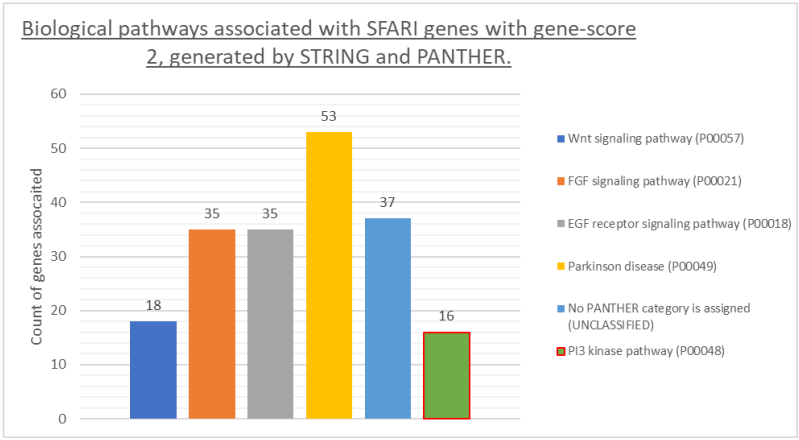
Gene-score group	number of nodes	number of edges	average node degree
1	213	1555	14.6
2	682	3380	9.91
3	91	63	1.38

For the geneset with gene-score of 2, STRING provided a warning that the network was large, and reduced visual presentation of the network. Furthermore, "the network edges show interaction confidence only".

(Task3.2) STRING-derived pathway analysis employing PANTHERDB







(Task 3.2) STRING clustering images are shown in the appendix

Largest clusters were in list of gene-score 2 genes.

Improvements to the method

The search using PantherDB could have included “GeneID:x” for each GeneID fed into the searching page to ensure that PantherDB recognises the query gene names as the Gene-Symbols. This was not done and may have returned more accurate results. The PubMed query might have returned more accurate results if the query Gene Symbols were translated to the official HGNC gene symbols before searching. Similar to the PantherDB, this would have allowed the gene identifier to be correctly recognised in the indexing system of PubMed. The query term could have been further improved by using the search field identifier [majr] instead of [MH]. Although the use of [MH] was a result of inexperience with PubMed querying, the resulting dataset when searching with [MH] was large and

the use of [majr] may have returned broader results which were less relevant. Manual inspection of the some of the returned papers by using the web interface showed that many papers did not make mention of the gene in the query.

Discussion

The search query was sensitive to the order of terms and the fields entered: using ASD and Autism vs just ASD gave far more results, but little variation in counts between gene:years. This suggests that 'autism' as mesh term is used broadly across ranges of literature subjects, which may not truly include information relevant to the gene being queried. This questions the subjectivity of manual publication labelling by authors, but also reflects the importance of a good search query to ensure that the papers identified are truly to Autism. Using MeSH terms was a way to capture papers related to the topic of Autism, but did not restrict the results to specific types of paper such as genetic study.

The purpose of re-analysing data thrgouh different methods was to provide an analysis of the ways we can retrieve information about these genes based on different data management strategies. Agreement of gene significance in disease or function is important across annotation sources, but by pulling information on a dataset from multiple sources it is possible to widen the search and collect more data than would have been found otherwise. Much of the information held in older literature which could enhance our analysis were it to be included in functional enrichment such as that seen above is inaccessible. This is due to the limited capacity for ontology-parsing platforms to annotate the infeasible diverse nature of published material…the most prolific annotation services employ a degree of manual annotation to ensure that data is being continually added and correctly allocated to the ontologies in existence.

Investigation of existing data on ASD demonstrates variability in research methods.

B217754-2022

References

1. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*. 1996;242(1):84-89. Accessed Dec 2, 2022. doi: 10.1006/abio.1996.0432.

2. Rothberg JM, Hinz W, Rearick TM, et al. An integrated semiconductor device enabling non-optical genome sequencing. *Nature*. 2011;475(7356):348-352. Accessed Dec 2, 2022. doi: 10.1038/nature10242.

3. Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation sequencing technologies. *Current Protocols in Molecular Biology*. 2018;122(1):e59-n/a. https://onlinelibrary.wiley.com/doi/abs/10.1002/cpmb.59. doi: 10.1002/cpmb.59.

4. SFARI. EAGLE score (evaluation of autism gene link evidence). SFARI Web site. https://gene.sfari.org/about-gene-scoring/eagle-score/. Accessed Dec 2, 2022.

5. NCBI. NCBI FTP gene DATA. NCBI FTP Web site. https://ftp.ncbi.nlm.nih.gov/gene/DATA/. Updated 2022. Accessed Dec 02, 2022.

6. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou L, Mi H. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein science*. 2022;31(1):8-22. https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4218. doi: 10.1002/pro.4218.

7. Mi H, Muruganujan A, Huang X, et al. Protocol update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature protocols*. 2019;14(3):703-721. https://www.ncbi.nlm.nih.gov/pubmed/30804569. doi: 10.1038/s41596-019-0128-8.

8. Mi H, Thomas P. PANTHER pathway: An ontology-based pathway database coupled with data analysis tools. In: *Methods in molecular biology (clifton, N.J.).* Vol 563. Totowa, NJ: Humana Press; 2009:123-140. http://link.springer.com/10.1007/978-1-60761-175-2_7. 10.1007/978-1-60761-175-2_7.

9. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(D1):D605-D612. Accessed Dec 2, 2022. doi: 10.1093/nar/gkaa1074.

10. NCBI. MEMO1 mediator of cell motility 1. NCBI Gene Web site. https://www.ncbi.nlm.nih.gov/gene/?term=51072%5Buid%5D. Updated 2022. Accessed Dec 1, 2022.

11. Kerin T, Ramanathan A, Rivas K, Grepo N, Coetzee GA, Campbell DB. A noncoding RNA antisense to moesin at 5p14.1 in autism. *Sci Transl Med*. 2012;4(128):128ra40. Accessed Dec 2, 2022. doi: 10.1126/scitranslmed.3003479.

12. Lim ET, Uddin M, De Rubeis S, et al. Rates, distribution, and implications of post-zygotic mosaic mutations in autism spectrum disorder. *Nat Neurosci*. 2017;20(9):1217-1224. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5672813/. Accessed Dec 2, 2022. doi: 10.1038/nn.4598.

13. Bilinovich SM, Lewis K, Grepo N, Campbell DB. The long noncoding RNA RPS10P2-AS1 is implicated in autism spectrum disorder risk and modulates gene expression in human neuronal progenitor cells. *Front Genet*. 2019;10:970. Accessed Dec 2, 2022. doi: 10.3389/fgene.2019.00970.

14. Arpi MNT, Simpson TI. SFARI genes and where to find them; modelling autism spectrum disorder specific gene expression dysregulation with RNA-seq data. *Sci Rep*. 2022;12:10158.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9203566/. Accessed Dec 2, 2022. doi:

10.1038/s41598-022-14077-1.

Appendices

Task 1.4 Paper count by year per gene

Gene_Name	1995	1996	1997	1998	1999	2000	2001
SHANK3	0	0	0	0	0	0	0
MECP2	0	0	0	0	0	2	1
NRXN1	0	0	0	0	0	0	0
SCN2A	0	0	0	0	0	0	0
SCN1A	0	0	0	0	0	0	0
	2002	2003	2004	2005	2006	2007	2008
SHANK3	0	0	0	2	1	4	6
MECP2	2	5	5	9	4	12	13
NRXN1	0	0	0	0	0	4	5
SCN2A	0	1	2	0	0	0	2
SCN1A	0	1	2	1	0	0	1
	2009	2010	2011	2012	2013	2014	2015
SHANK3	8	5	14	14	18	17	25
MECP2	16	15	22	16	12	25	18
NRXN1	10	6	9	12	15	11	11
SCN2A	0	1	0	1	3	4	4
SCN1A	1	0	0	6	2	0	2
	2016	2017	2018	2019	2020	2021	2022
SHANK3	39	35	39	47	47	47	59
MECP2	32	20	16	21	17	12	22
NRXN1	4	10	10	18	10	13	9
SCN2A	6	6	6	14	11	16	11
SCN1A	4	0	4	4	7	10	5

Task1 Extension: Quantification of literature on more specific ASD-disease genes.

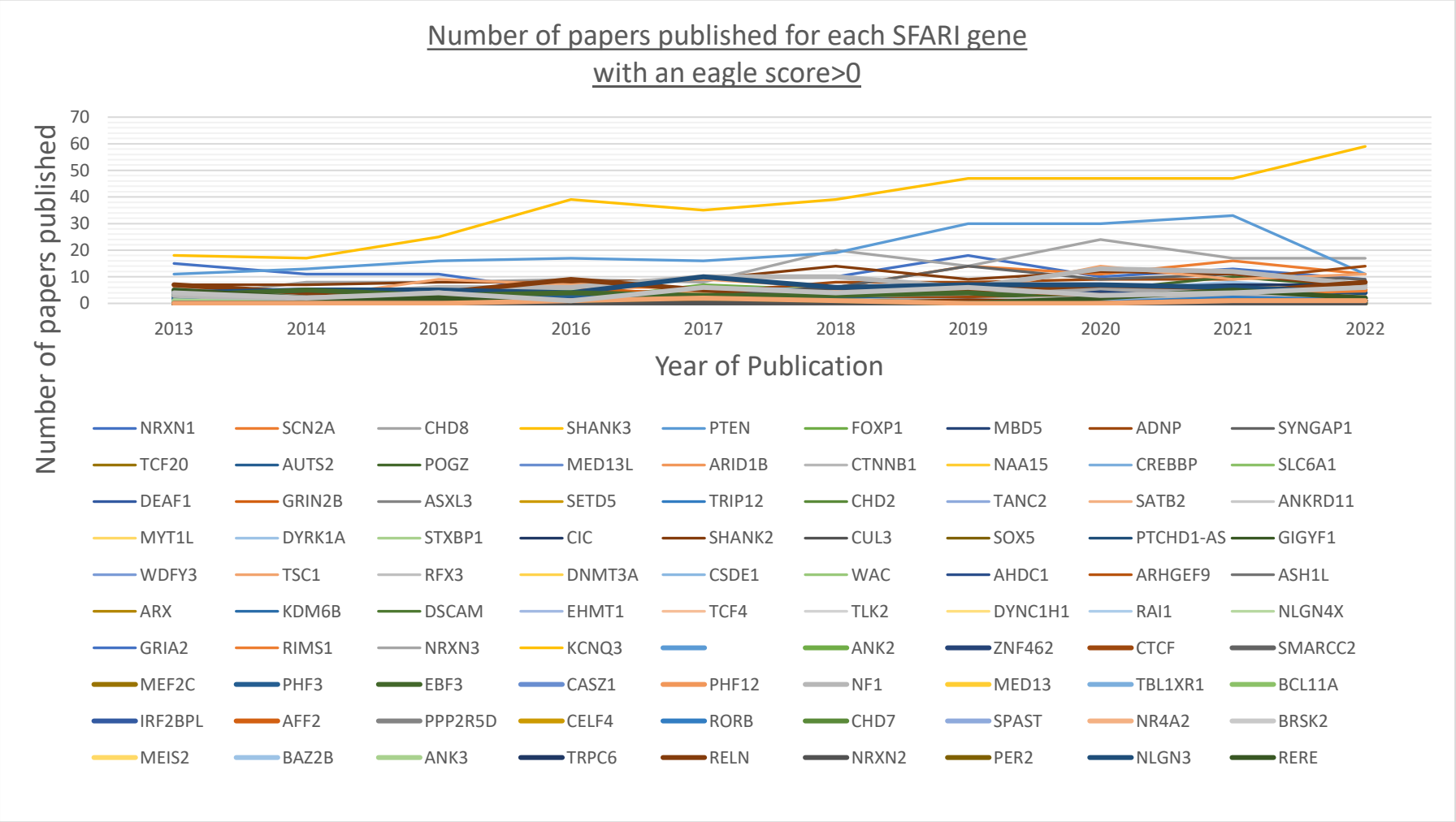
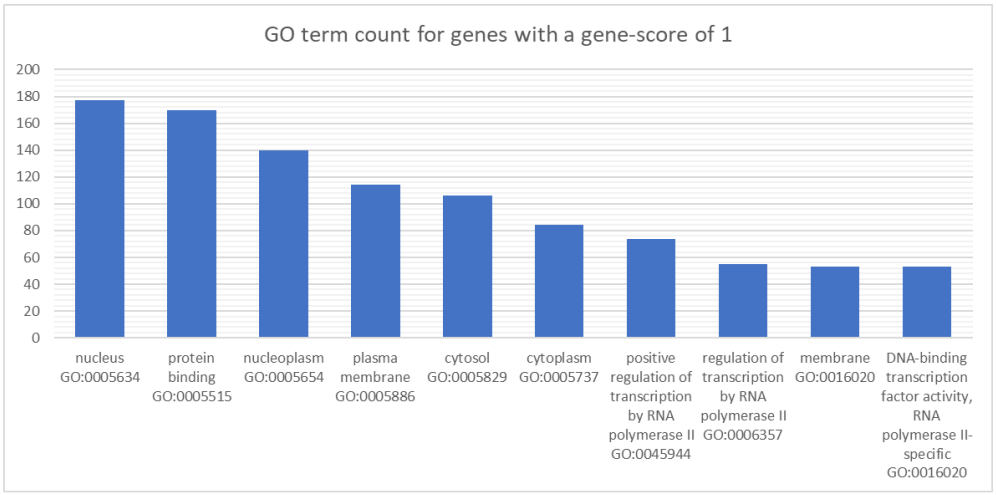
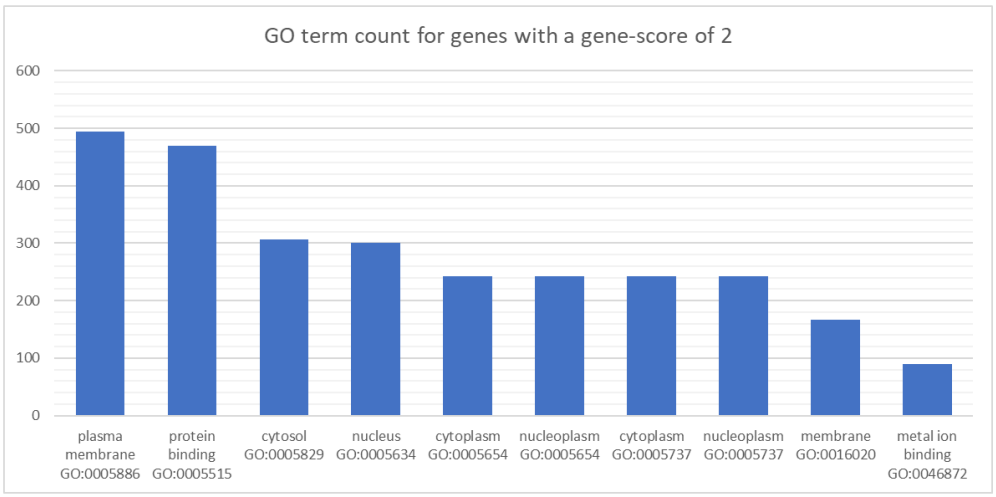


Figure 5 trends in the research of the genes which are classed as related to ASD according to the EAGLE scoring method. Most genes were not frequently cited The gene 'SON' had 423 results in 2022. Due to limitations in data presentation, this record was removed from the graph.

Genes with a gene-score of 1		
GO term count	GO_ID	GO_term
177	GO:0005634	nucleus
170	GO:0005515	protein binding
140	GO:0005654	nucleoplasm
114	GO:0005886	plasma membrane
106	GO:0005829	cytosol
84	GO:0005737	cytoplasm
74	GO:0045944	positive regulation of transcription by RNA polymerase II
55	GO:0006357	regulation of transcription by RNA polymerase II
53	GO:0016020	membrane
53	GO:0016020	DNA-binding transcription factor activity, RNA polymerase II-specific
Genes with a gene-score of 2		
GO term count	GO_ID	GO_term
495	GO:0005886	plasma membrane
469	GO:0005515	protein binding
306	GO:0005829	cytosol
301	GO:0005634	nucleus
243	GO:0005654	cytoplasm
243	GO:0005654	nucleoplasm
243	GO:0005737	cytoplasm
243	GO:0005737	nucleoplasm
167	GO:0016020	membrane
89	GO:0046872	metal ion binding
Genes with a gene-score of 3		
GO term count	GO_ID	GO_term
75	GO:0005515	protein binding
61	GO:0005829	cytosol
48	GO:0005737	cytoplasm
47	GO:0005634	plasma membrane
47	GO:0005634	nucleus
47	GO:0005886	plasma membrane
47	GO:0005886	nucleus
32	GO:0005654	nucleoplasm
25	GO:0016020	membrane
17	GO:0046872	metal ion binding

Task 2.4 The 10 most commonly annotated terms for the SFARI geneset, grouped by gene-score





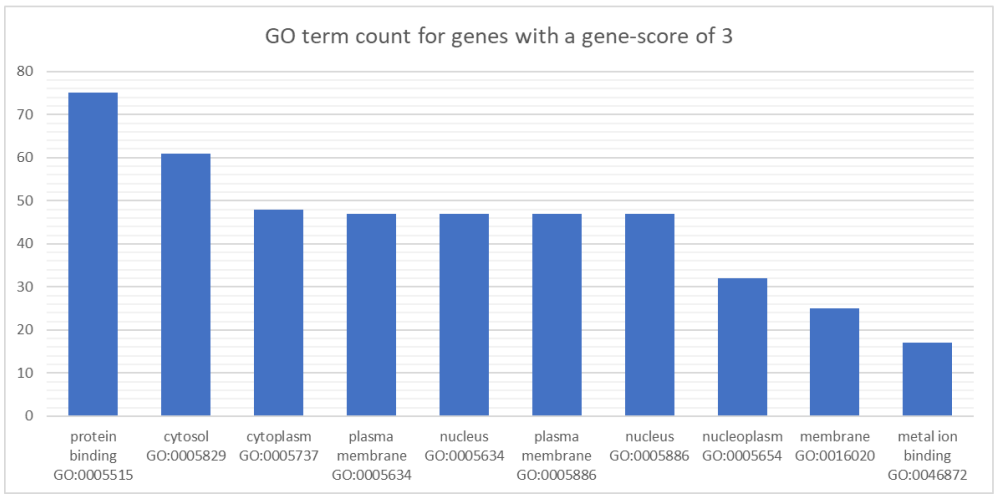
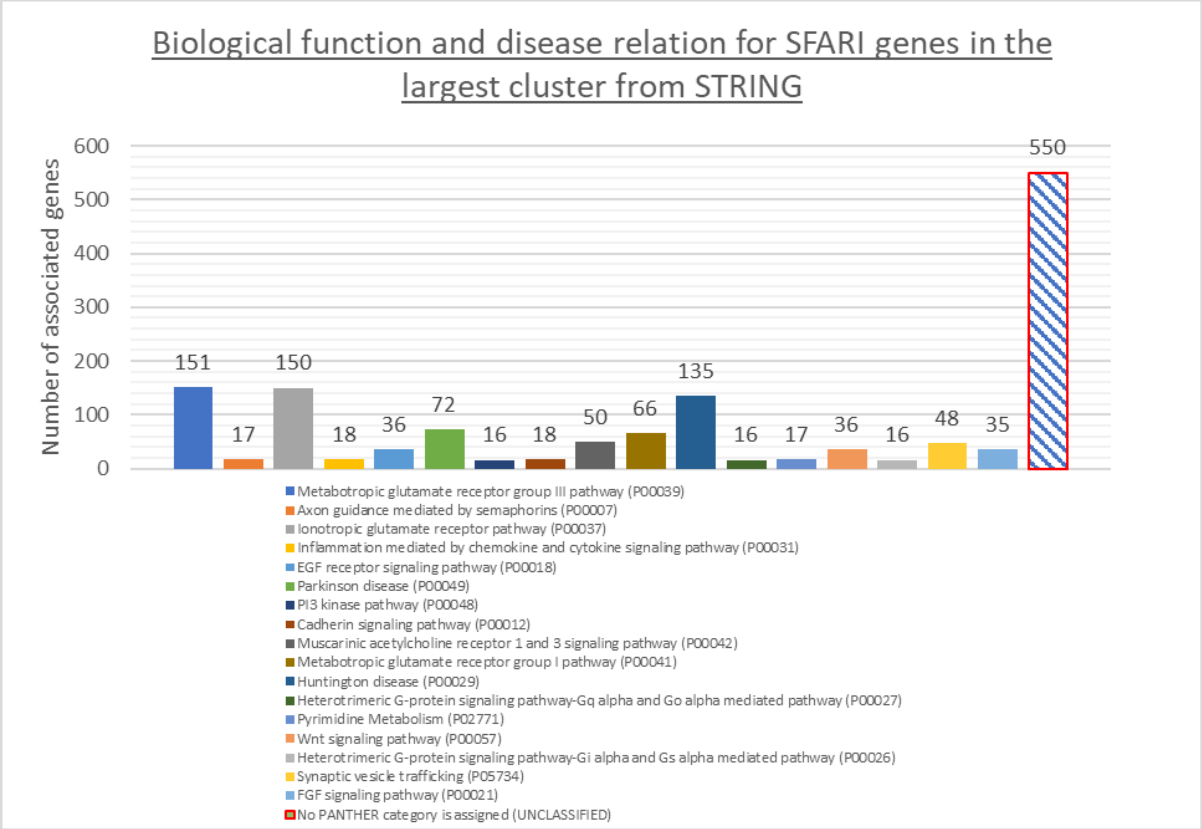
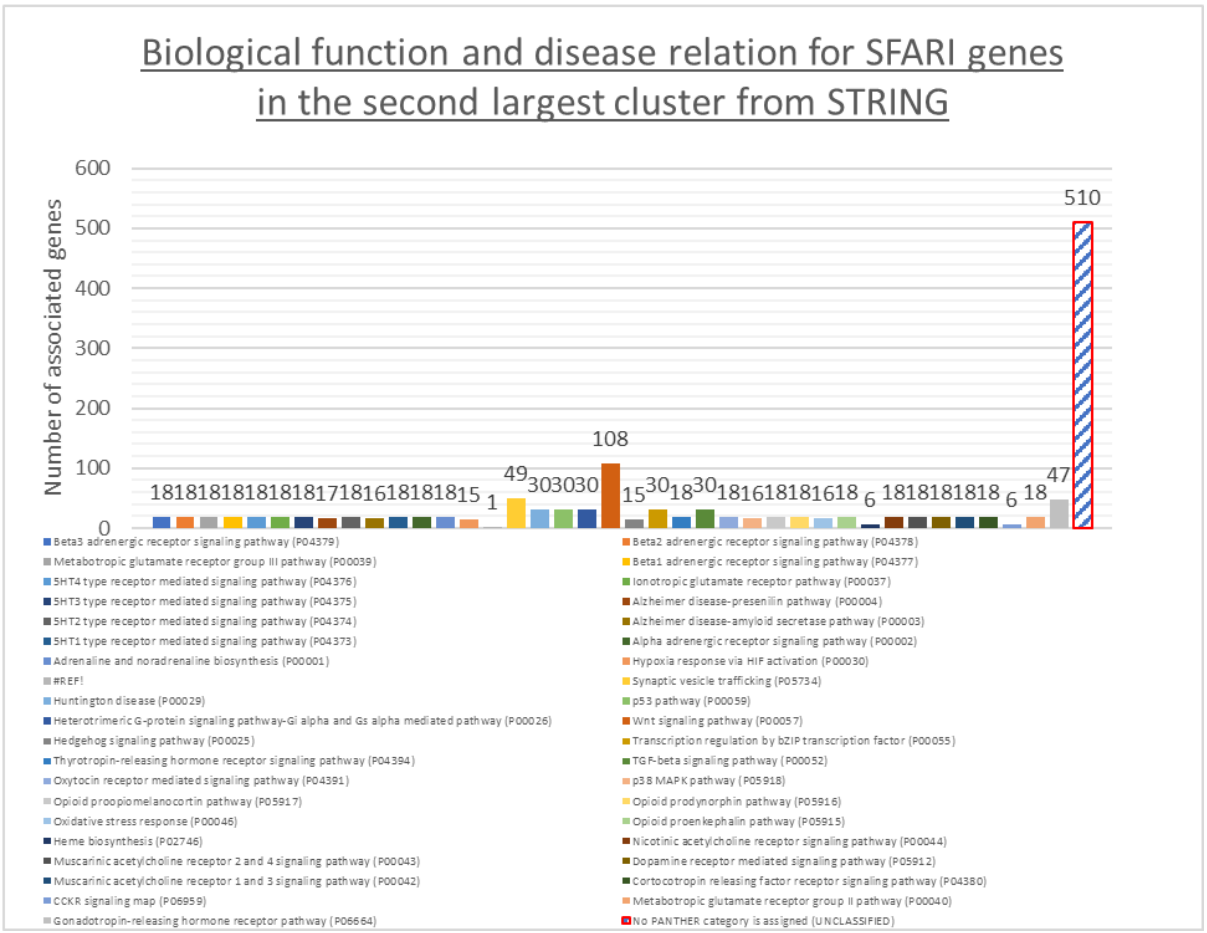


Figure 6 The 10 most commonly annotated terms for the SFARI geneset, grouped by gene-score

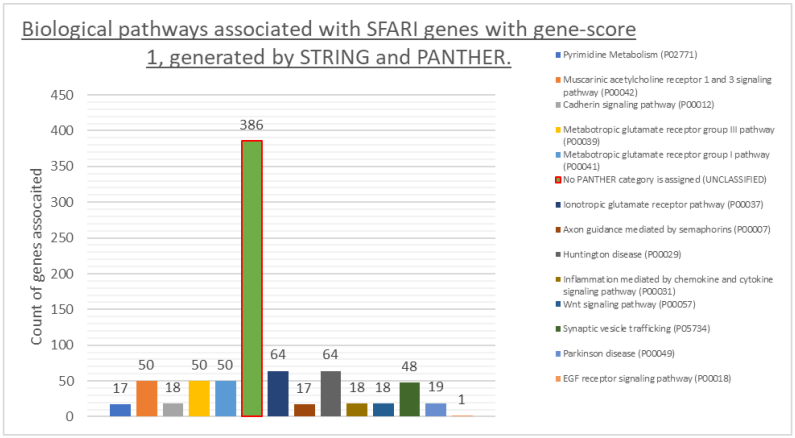
Task 2.5

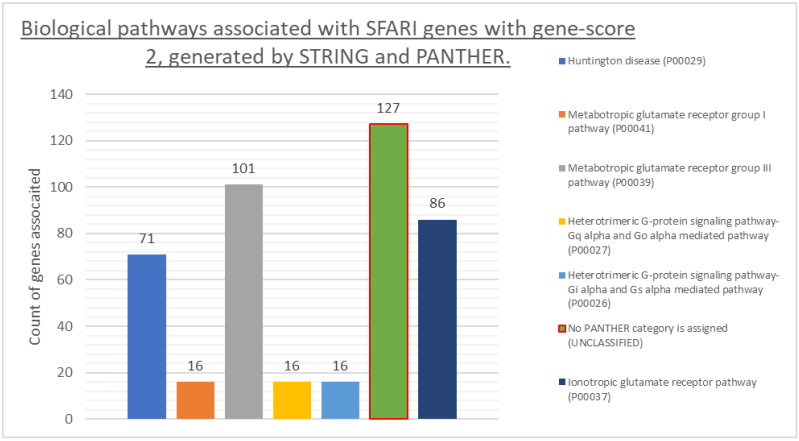

Biological function and disease relation for SFARI genes in the largest cluster from STRING
Number of associated genes
600
500
400
300
200
100
0
151
17
150
18
36
72
16
18
50
66
135
16
17
36
16
48
35
550
Metabotropic glutamate receptor group III pathway (P00039)
Axon guidance mediated by semaphorins (P00007)
Ionotropic glutamate receptor pathway (P00037)
Inflammation mediated by chemokine and cytokine signaling pathway (P00031)
EGF receptor signaling pathway (P00018)
Parkinson disease (P00049)
PI3 kinase pathway (P00048)
Cadherin signaling pathway (P00012)
Muscarinic acetylcholine receptor 1 and 3 signaling pathway (P00042)
Metabotropic glutamate receptor group I pathway (P00041)
Huntington disease (P00029)
Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (P00027)
Pyrimidine Metabolism (P02771)
Wnt signaling pathway (P00057)
Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026)
Synaptic vesicle trafficking (P05734)
FGF signaling pathway (P00021)
No PANTHER category is assigned (UNCLASSIFIED)

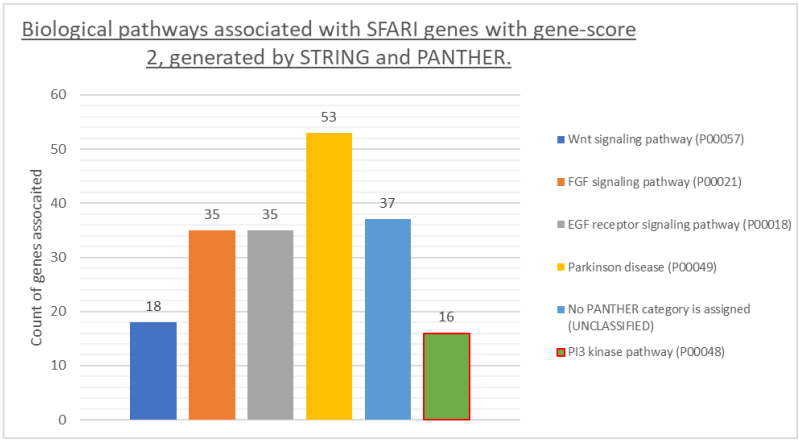

Biological function and disease relation for SFARI genes in the second largest cluster from STRING
Number of associated genes
600
500
400
300
200
100
0
510
Beta3 adrenergic receptor signaling pathway (P04379)
Beta2 adrenergic receptor signaling pathway (P04378)
Metabotropic glutamate receptor group III pathway (P00039)
Beta1 adrenergic receptor signaling pathway (P04377)
5HT4 type receptor mediated signaling pathway (P04376)
Ionotropic glutamate receptor pathway (P00037)
5HT3 type receptor mediated signaling pathway (P04375)
Alzheimer disease-presenilin pathway (P00004)
5HT2 type receptor mediated signaling pathway (P04374)
Alzheimer disease-amyloid secretase pathway (P00003)
5HT1 type receptor mediated signaling pathway (P04373)
Alpha adrenergic receptor signaling pathway (P00002)
Adrenaline and noradrenaline biosynthesis (P00001)
Hypoxia response via HIF activation (P00030)
#REF!
Synaptic vesicle trafficking (P05734)
Huntington disease (P00029)
p53 pathway (P00059)
Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026)
Wnt signaling pathway (P00057)
Hedgehog signaling pathway (P00025)
Transcription regulation by bZIP transcription factor (P00055)
Thyrotropin-releasing hormone receptor signaling pathway (P04394)
TGF-beta signaling pathway (P00052)
Oxytocin receptor mediated signaling pathway (P04391)
p38 MAPK pathway (P05918)
Opioid proopiomelanocortin pathway (P05917)
Opioid prodynorphin pathway (P05916)
Oxidative stress response (P00046)
Opioid proenkephalin pathway (P05915)
Heme biosynthesis (P02746)
Nicotinic acetylcholine receptor signaling pathway (P00044)
Muscarinic acetylcholine receptor 2 and 4 signaling pathway (P00043)
Dopamine receptor mediated signaling pathway (P05912)
Muscarinic acetylcholine receptor 1 and 3 signaling pathway (P00042)
Corticotropin releasing factor receptor signaling pathway (P04380)
CCKR signaling map (P06959)
Metabotropic glutamate receptor group II pathway (P00040)
Gonadotropin-releasing hormone receptor pathway (P06664)
No PANTHER category is assigned (UNCLASSIFIED)

Task 3.2

(Task3.2) STRING-derived pathway analysis employing PANTHERDB







Task 3.3 Bitmap images of the STRING networks of the SFARI genes stratified by gene-score and clustered using STRING MCL clustering.

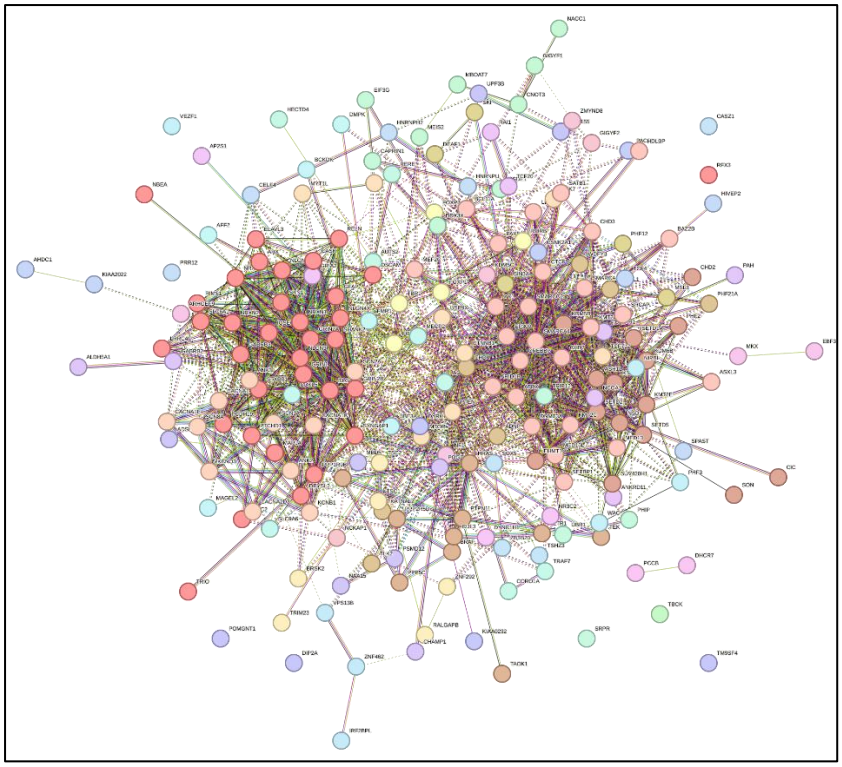


Figure 7 Gene score 2 genes, clustered

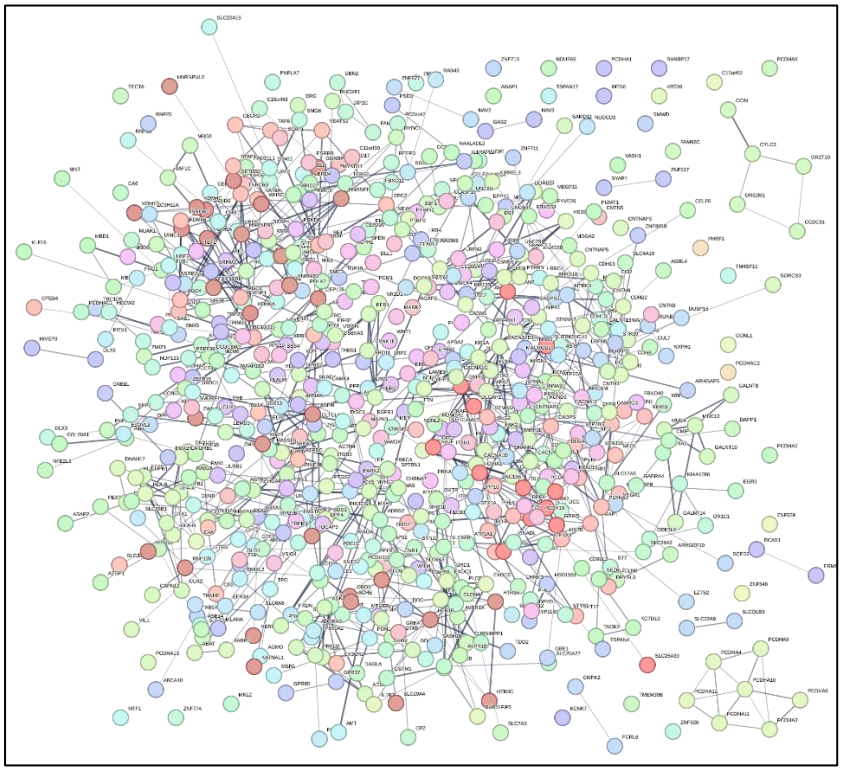


Figure 8 Gene score 1 genes, clustered

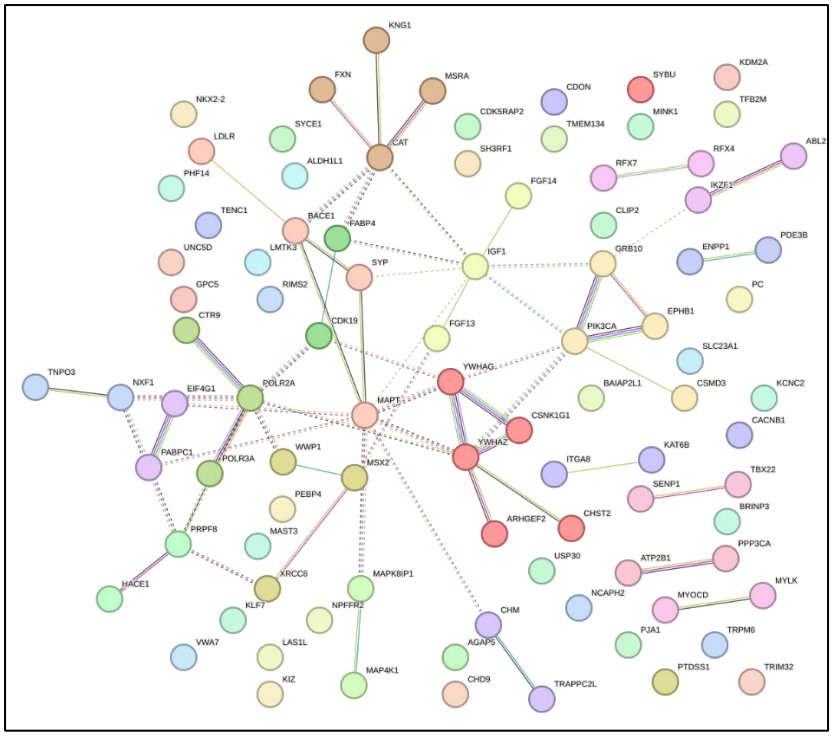


Figure 9 Gene score 3 genes, clustered

Python code

The script used to perform data retrieval and much of the table manipulation can be accessed at https://github.com/B217754-2022/B1_ICA2.git and the ccrypt key is:

Please run the following command:

ccrypt -d B217754--2022.B1.ICA2.tar.gz.cpt

Then enter the key when prompted:

AppletonClass