Coursework 2 - Exploring Autism Spectrum Disorders

October 2022

Contents

1 Introduction 3

2 Data and Methods 3

2.1 Part One – Autism Literature 4
2.2 Part Two – Autism Genes 5
2.3 Part Three – Autism Gene Networks 7

3 Results 7

3.1 Part One – Autism Literature 7
3.1.1 Task 1 7
3.1.2 Extension 8
3.1.3 Task 2 8
3.1.4 Task 3 9
3.1.5 Task 4 9
3.1.6 Task 5 10
3.1.7 Extension 11
3.2 Part Two – Autism Genes 14
3.2.1 Task 1 14
3.2.2 Task 2 15
3.2.3 Task 3 15
3.2.4 Task 4 15
3.2.5 Extension 16
3.2.6 Task 5 19
3.2.7 Extension 21
3.3 Part Three – Autism Gene Networks 22
3.3.1 Task 1 22
3.3.2 Tasks 2 & 3 22
3.3.3 Extension 24

4 Discussion 25

A Data Sources and Tools 29

B Additional Figures and Tables **29**

B.1 Part One 29

B.2 Part Two 31

B.3 Part Three 32

1 Introduction

Autism Spectrum Disorders (ASD), also unitedly referred to as autism, are a group of neurological disabilities impacting an affected person's early development. While the specific symptoms vary between individuals, they all exhibit 'deficits in social communication and social interaction across multiple contexts' as well as 'restricted, repetitive patterns of behaviour' such as 'inflexible adherence to routines, or ritualised patterns of verbal or nonverbal behaviour'. [1] Given the lack of medical test, such as blood tests, and the wide variety of symptoms, diagnosing ASD based solely on an individual's behaviour, interactions and interests can be difficult.

While autism can occur alone, it often does so together with other neurological disorders. This is called syndromic ASD. Examples of such co-occurrences include epilepsy, Fragile X Mental Retardation or Rett Syndrome. [2, 3]. This differentiation extends onto the genes that are linked to autism. While a syndromic gene has only been associated with autism where it co-occurs with another disorder, a non-syndromic (idiopathic) gene has been linked to autism directly and independently of other syndromes.

ASD currently affects about 1 in 100 children world-wide [4], yet to date no cure has been developed and treatments focus on reducing symptoms that interfere with everyday life [5]. Further research into the causes and potential treatments is therefore urgently needed.

In this report I will thus investigate the genes that have been linked to autism in three parts: a literature analysis, an ontological study, and an inspection of the genes' networks. These will shine a light on how research into autism-linked genes has been developing over the years and what the common properties of these genes are.

My investigation will be based on the gene list published by the Simons Foundation Autism Research Initiative (SFARI). Mutations in these genes have been shown to different extents to be linked to either idiopathic or syndromic autism and are accordingly split into the scores 1 (clearly implicated in ASD), 2 (implicated with sufficient evidence by a genome-wide association), 3 (significant evidence but unreplicated) for idiopathic autism, and S for a link to syndromic autism. A gene may be labelled with a number and S if it has been implicated in syndromic and non-syndromic ASD. [6]

2 Data and Methods

The SFARI gene list that I will be using throughout was published on the 11-07-2022 release and retrieved on 11-14-2022. Further details on the data sources and tools can be found in Appendix A.

To aid reproducibility, the code used in this report is available at `https://gitfront.io/r/p/hfh1ee6sPzAQ/bioinf-cw2-autism/`.

Where applicable, automated data collection and scraping were done in line with the UK Office for National Statistics guidelines [7]. This includes providing a means of contact in the user agent for each request and minimising the burden on websites by using their

API to batch multiple queries into a single web request, instead of sending an individual request for each query.

The steps undertaken for each part are enumerated below such that the step numbers match the part's subsection numbers in the Results section. The reasoning for each step is detailed in the Results.

2.1 Part One – Autism Literature

To gain an overview of the field, I first study the literature around autism using the SFARI gene list as a starting point.

1. Retrieve the SFARI gene list from `https://gene.sfari.org/database/human-gene/`. For a direct link to the list see Appendix A. Each row contains a gene. The list has a `gene-score` column that indicates the strength of evidence implicating the row's gene with idiopathic ASD (integer from 1 to 3). Count the genes that are in each `gene-score` class and plot them in a bar chart.

2. The SFARI gene list also has a `syndromic` column containing is a boolean for each gene, indicating whether it has been linked to syndromic ASD. Split the gene list into two by the `syndromic` values (True or False). For each split list, count the genes in each `gene-score` class and plot the two lists in a stacked bar chart to show the ratio of each gene score that is syndromic or not syndromic.

3. Split the SFARI gene list such that it only contains genes of score 1. Sort each of the new lists by the `number-of-reports` column in descending order and select the five genes that have the highest number of reports. Let us define this list of five as `top5Genes`.

4. For each gene in the `top5Genes`, search PubMed to find the number of reports related to the gene and autism. For this, query PubMed through its biopython API for '`{gene} AND (autism OR ASD OR "Autism Spectrum Disorder")`', where `{gene}` is replaced with the value in each gene's `gene-symbol` column. Store the returned PubMedIds in `reportIdsGeneX` (`X` $\in$ `top5Genes.gene-symbol`). For each gene, count the number of papers returned by the request and list them in a table. The result should be comparable to Table 2.
I name this search method *Basic Search.*

5. For each gene `X` $\in$ `top5Genes.gene-symbol`, use the biopython PubMed API to retrieve a summary per PubMedId in `reportIdsGeneX`. Count the occurrences of each year per gene.
Create a table of the number of papers found in each year for each gene `X`. Columns are years and rows are genes. The result should be comparable to Table B.1.

6. Plot a stacked bar chart from the table created in the previous step. Years are on the x-axis and the number of papers is on the y-axis. The result should be comparable to Fig. 3.

7. Repeat steps 4 and 5 for the following search queries and filters.
   Queries:
   Topic-independent Search: ‘`{gene}`’
   Hit Filtering: ‘`{gene} AND (autism OR ASD OR "Autism Spectrum Disorder")`’
   Major Topic Search: ‘`{gene} AND (autism[Majr] OR "Autism Spectrum Disorder"[Majr])`’
   Major Topic Hit Filtering: ‘`{gene} AND (autism[Majr] OR "Autism Spectrum Disorder"[Majr])`’

   Filters:
   Hit Filtering: Only keep reports whose title or keywords contain the queried gene and ‘autism’ or ‘asd’. The title containing one and keywords containing the other is permitted.
   Major Topic Hit Filtering: Only keep reports whose title or keywords contain the queried gene.

   The total number of reports obtained per gene by each of the above methods should approximately match those in Table 5.

   Plot a bar chart as in step 6 for Topic-independent Search, Major Topic Search and Hit Filtering. They should be similar to Fig. 4.

2.2 Part Two – Autism Genes

1. Load the SFARI gene list used in Part One. Retrieve the `Homo_sapiens.gene_info` table from NCBI at `https://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz`. Perform an inner merge of the two lists on the SFARI list’s `gene-symbol` column and the gene info’s `Symbol_from_nomenclature_authority` column. Use the nomenclature authority’s symbol instead of `Symbol`, as the latter may also include unofficial duplicates of gene symbols, such as ‘MEMO1’. Call this new table `sfariGeneIds`.

2. Gene Ontologies (GO) are a tree-representation of knowledge in a biological domain, consisting of GO terms or classes (nodes) with which individual genes may be annotated. The most general terms are at the root and the most specific at the leaves. Each node’s parent is its generalisation. NCBI keeps a table mapping their gene IDs to the GO terms that each gene has been annotated with.
   Retrieve the `gene2go` table from NCBI at `https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz`. Perform an inner merge of this table with `sfariGeneIds`. Both are to be merged on the `GeneID` column. The resulting table should match the preview in Table B.3. Call this table `sfariGOs`.

3. Split the `sfariGOs` table into three by the `gene-score` column. Call these `gosScore1`, `gosScore2` and `gosScore3`.

4. For each `gosScoreS` table, where `S` $\in$ [1,3], count the number of occurrences of each GO term in the table and select the ten terms with the highest counts. Let this list

of ten terms with their counts be called `topGosScoreS`. The resulting tables should be similar to Tables 6a-c.

5. For each Gene Ontology term in each `topGosScoreS` list, determine the p-value and odds-ratio of the term's count compared to the other scores.
   For score S and GO term G, let
   a = Genes in score S with GO term G
   b = Genes not in score S with GO term G
   c = Genes in score S without GO term G
   d = Genes not in score S without GO term G

   $$\text{p-value} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ b!\ c!\ d!\ (a+b+c+d)!}$$

   $$\text{odds-ratio} = \frac{a/c}{b/d} = \frac{a \cdot d}{b \cdot c}$$

   Computing the p-value and odds-ratio for each row in all three `topGosScoreS` lists should give results similar to Tables 7a-c.
   Similarly, the p-value and odds-ratio of the most common GO terms in all SFARI genes compared to all annotated human genes in `gene2go` can be computed. Use the same equations as above for
   a = Genes in SFARI with GO term G
   b = Genes not in SFARI with GO term G
   c = Genes in SFARI without GO term G
   d = Genes not SFARI without GO term G
   This should give results similar to Table 8.

   This report considers p-values of < 0.05 as significant.

6. Split the `sfariGeneIds` table into three by the `gene-score` column. Call these `geneIdsScore1`, `geneIdsScore2` and `geneIdsScore3`. For each of the new tables, extract the list of gene IDs as "GeneID:{id1} GeneID:{id2} GeneID:{id3} ..." where {idX} is replaced with an ID from the list. Insert this string into the IDs field on `http://pantherdb.org`. The 'GeneID:' notation is required so that PantherDB does not confuse the number for an ID from another source. On the website, select the Homo Sapiens organism, choose the 'Functional classification viewed in graphic charts → Bar chart' analysis option and submit. Change the drop-down menu to 'Biological Process' and export (small blue link in the top left corner). Load the text file in TSV format and plot it as a bar chart with horizontal bars. The third column is on the x-axis (% of Genes with the process) and the first column is on the y-axis (Gene Ontology term and ID). I chose this format to make the long GO terms more easily readable. The % of Genes is used instead of the absolute number of genes (second column) to make the plots more comparable between the scores, as they have different total numbers of genes.

7. Use the PantherDB API through the python library 'bioservices' to retrieve the expanded biological processes for each of `geneIdsScore1-3`. See `https://gitfront.io/r/p/hfh1ee6sPzAQ/bioinf-cw2-autism/blob/code/part2.ipynb` (Figures → Task5-Extension) for details. The resulting table of processes should be comparable to Tables 9, B.5 and B.6.

2.3 Part Three – Autism Gene Networks

1. Load the `geneIdsScore1` list from part two. Extract its gene ids line-by-line. Copy the approximately 214 line string into the 'List of Names' field on `https://string-db.org` ('Multiple proteins' function in the left pane). Select the Homo Sapiens organism and submit the search. Select the correct genes and press 'Continue ->'. The 'Network Stats' of the 'Σ Analysis' section should be similar to those listed in Table B.7. This is a protein-protein interaction network.

2. On the network page from the previous step, go to the 'Clusters' section, select MCL clustering with 'Inflation parameter' set to 3 (See Appendix A for details on the algorithm and the inflation parameter) and apply. The resulting network should look similar to the one in Fig. B.1, but may be in a different orientation. At the bottom of the 'Clusters' section, download the TSV file.
   Load the TSV file and extract the rows with `cluster number` = 1 into one list and `cluster number` = 2 into another. These are the proteins corresponding to the two largest clusters found in the network. For each of the two tables, use `geneIdsScore1` to find the gene id corresponding to the `protein name` column. Note that here gene and protein are considered interchangeable. For each of the two clusters, use their gene ids on PantherDB as outlined in part two, step 6, but instead of selecting 'Biological Process', use 'Pathway'. Plot them as in part two, step 6. The results should be similar to Fig. 8 and 9.
   PantherDB's pathways, in contrast to their biological processes, are not based on Gene Ontologies but instead the pathways from the Reactome database (see Appendix A).

3. Use the gene ids of each cluster that were extracted in the previous step to obtain the expanded Pathways from PantherDB through the bioservices library. See `https://gitfront.io/r/p/hfh1ee6sPzAQ/bioinf-cw2-autism/blob/code/part3.ipynb` (Figures → Task2-Extension) for details. Plot them as before. The results should look similar to Fig. 10 and 11.

3 Results

3.1 Part One – Autism Literature

3.1.1 Task 1

To obtain an overview of the ratio of genes in each SFARI score category, I have shown their distribution in Figures 1 and 2. While Fig. 1 focuses only on the idiopathic genes, Fig. 2 also highlights the syndromic genes. Of the total 1095 genes listed by SFARI, 214 have substantial evidence for their link to ASD (score 1), 695 have been implicated with sufficient evidence by a genome-wide association study (score 2) and 91 have been linked without replication (score 3). Genome-wide studies can identify mutations in many genes simultaneously. Depending on the genome-wide significance achieved and other evidence gathered, the genes are sorted into each score, with score 1 having the highest threshold and score 3 the lowest. The genes that stand out in such studies tend to at least have some evidence. If, however, they don't achieve the highest threshold, this puts the genes in category 2.

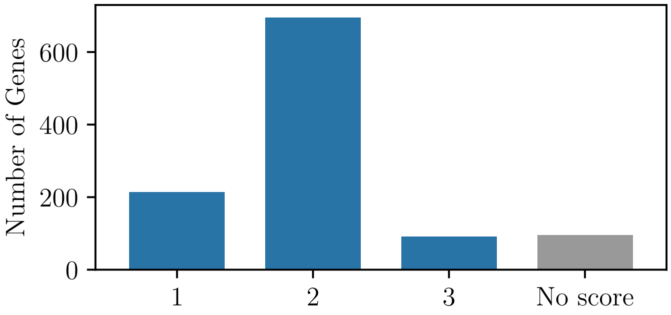


Figure 1: Number of genes in each idiopathic SFARI gene score category.

3.1.2 Extension

Looking at Figure 2, it is interesting to note that while most idiopathic genes are in score 2, most syndromic genes are either in score 1 or have not been linked to idiopathic ASD at all. This can be explained by 'the greatest progress in autism genetics being made in the gene discovery of monogenic disorders, which cause syndromic disease' [8]. Since more research is done on genes linked to syndromic autism, more is known about their mechanism and more evidence has been gathered regarding its involvement in idiopathic ASD. Thus, for most syndromic genes it is known whether they also cause non-syndromic ASD (score 1) or only syndromic ASD (no score assigned).

3.1.3 Task 2

In the following analysis, I use a set of five representative genes to investigate the historical progress of ASD genetics. To ensure my selection is relevant to autism, I limit myself to score 1 genes and select the five genes with the most reports quoted by SFARI as a proxy for the evidence gathered on them. The resulting five most-reported-on genes are found in Table 1. As the institute manually curates its report selection, I attempt to reproduce their results using automated search-and-filter methods. The yearly trend in publications, relating each gene to autism, is obtained from the results.

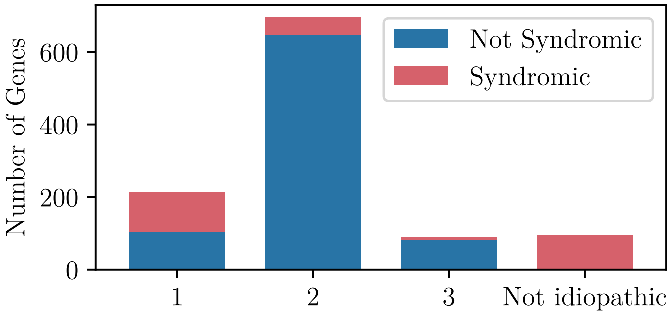


Figure 2: Number of genes in each idiopathic SFARI gene score category. Genes that have only been linked to idiopathic ASD are shown in blue, while genes that have been linked only to syndromic ASD or to both are in red.

Gene Symbol	Gene Name	Reports
SHANK3	SH3 and multiple ankyrin repeat domains 3	120
MECP2	Methyl CpG binding protein 2	107
NRXN1	Neurexin 1	100
SCN2A	Sodium channel, voltage-gated, type II, alpha subunit	96
SCN1A	Sodium channel, voltage-gated, type I, alpha subunit	84

Table 1: Top five genes in SFARI category 1, sorted in descending order by the number of reports cited by SFARI.

3.1.4 Task 3

Retrieving publications from PubMed that relate each gene with autism is non-trivial, as papers may mention either the gene or autism without it being the paper's subject. An example of this is Gerosa et al.'s review [9]. Even though the publication's topic is PCDH19, they mention SCN1A as another relevant gene in their abstract. This could produce a false match.

One method for searching would be to query PubMed with '`(`*gene*`) AND (`*autism query*`)`' where *gene* and *autism query* is to be replaced with the relevant gene symbol and *autism query* is replaced with a list of autism-related search terms. This query expands to, for example, '`(SHANK3) AND (autism OR ASD OR "Autism Spectrum Disorder")`'. The number of autism-related articles returned for each gene using this search technique is shown in Table 2. A yearly breakdown by gene is in Table 3, also found in large in Table B.1. To differentiate this search method from others that I introduce later, I will henceforth refer to it as the *basic* method.

3.1.5 Task 4

Worth noting from the yearly breakdown (Table 3) is that for three of the five genes, their first mention with autism is in 1993, any subsequent mentions are not until much later.

Gene Symbol	Number of Reports
SHANK3	498
MECP2	546
NRXN1	195
SCN2A	105
SCN1A	82

Table 2: Number of publications returned by the basic search for ASD-related reports on each gene.

Responsible for this is GeneReviews [10], first published in 1993 and continuously updated with information on a wide variety of genes. Since PubMed lists the same publication date for each chapter, all relevant chapters on the genes and autism are listed as 1993, even if published later. For SHANK3 and SCN1A GeneReviews has only a single chapter where ASD is mentioned [11, 12] while it has three on MECP2 [12, 13, 14].

Another interesting year is 2023. Even though I am writing this report in 2022, there is already a publication from 2023 on PubMed. While this is surprising, it is merely caused by an article that has already been uploaded to PubMed but will be published in January next year.

While the above are only small artefacts of the PubMed database, what is most interesting is that the number of PubMed reports retrieved (Tables 2 and 3) is significantly higher for SHANK3, MECP2, and NRXN1, than that quoted by SFARI (Table 1). While they may have skipped some reports in their manual curation, a difference of 378 articles must be caused by something else. A likely explanation is that my automated retrieval unintentionally also included reports that merely mention the search terms without them being the main topic of the study. The previously mentioned review by Gerosa et al. [9] is one such example that was retrieved by the basic search method for SCN1A. An approach to eliminate these false hits is explored further down in the extension.

Gene	1993	'99	2000	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12	'13	'14	'15	'16	'17	'18	'19	'20	'21	'22	'23
SHANK3	1	-	-	-	-	-	-	1	-	4	7	9	5	18	17	17	32	27	40	37	49	44	53	63	73	1
MECP2	3	1	7	4	9	8	9	18	12	19	17	23	23	32	26	33	39	45	46	35	22	34	30	22	29	-
NRXN1	-	-	-	-	-	-	1	-	-	1	6	15	8	18	18	17	16	18	6	11	9	18	10	10	13	-
SCN2A	-	-	-	-	-	1	2	-	-	-	2	-	1	1	1	2	5	4	7	8	11	16	9	19	16	-
SCN1A	1	-	-	-	-	1	2	1	-	-	1	1	-	2	6	4	2	5	3	5	6	8	11	13	10	-

Table 3: (See Table B.1 for a larger version) Number of publications returned by the basic search for ASD-related reports on each gene.

3.1.6 Task 5

Plotting the yearly distribution of the autism-related report counts for each gene from Table 3 gives Fig. 3. The plot shows that there has been a significant increase in papers mentioning both, the specified gene and autism. Using these five genes as a proxy for the entire autism genetics space, it has seen a steady increase in research over the past years.

Contrasted with the approximated global rise in publication output by 50% (4% annually) between 2010 and 2020, this field has experienced a 205% (12% annually) rise in those years, if the generalisation from these five genes is to be trusted.

Notably, the main driver of this increase is SHANK3 with an average annual rate of $\sqrt[2022-2005]{73/1} - 1 = 29\%$ since its first mention with autism. The other genes are at 18% (MECP2), 19% (NRXN1), 16% (SCN2A), and 12% (SCN1A) annually since their first connection to autism. NRXN1 and MECP2 had their highest number of reports published around 2012 and 2016 respectively and have not attained a rate as high since then.

The representativeness of these five genes and the retrieved papers, however, is uncertain. Given that these genes are in score 1 and much evidence has been gathered for them, they are likely a good representation of genes that are implicated with ASD. Further experiments are needed to be certain about their role as a proxy for the field. The following extension addresses the search methods used to retrieve the papers and with what certainty these are related to both, the gene in question and autism.

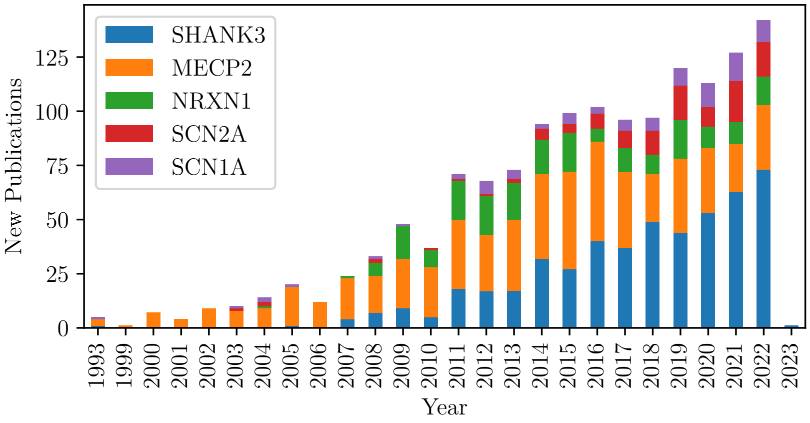


Figure 3: Distribution of each gene's autism-related articles by publication year. Uses values from Table 3

3.1.7 Extension

To address the problem that the basic search method also returns papers which may include the search terms, but do not study the gene in regards to ASD, I propose three alternative methods: *Major Topic Search*, *Hit Filtering* and *Major Topic Hit Filtering*. A summary of all four methods, including *Basic*, is found in Table 4.

Major Topic Search makes use of PubMed's Medical Subject Heading (MeSH) functionality. MeSH is a thesaurus containing synonyms of special terms. For instance, the entry for 'Autistic Disorder' also includes 'Autism', 'Infantile Autism', 'Early Infantile Autism', and 'Kanner's Syndrome' [15]. If a PubMed query contains a registered MeSH term with the suffix [mesh] or [mh], the search also matches its synonyms. Further, if a mesh term is specified to be a major topic using the suffix [Majr], only publications that explicitly list one of its synonyms as such a topic are matched.

Hit Filtering starts off with a wide selection of papers returned by the basic method and filters them to only include publications where the gene and autism are specially featured. For a term to produce a hit, it must be included in either the paper's title or its keywords. For the paper to be a match, the gene must produce a hit and autism must produce a hit. The title and keywords are chosen, as only terms that are relevant to the study are included there while the abstract may also include background information or further mentions.

Major Topic Hit Filtering is a combination of the above, where the selection of autism-related papers is done using the [Majr] suffix and relatedness to the given gene is ensured by its inclusion in the title or keywords.

Topic-independent gene reports include all reports on a given gene independent of what genetic function they study. I have included them here to put the number of autism-related reports into perspective.

SFARI Autism reports are manually curated ASD-related reports cited by SFARI. Notably, not all of their as relevant classified reports are related to autism, but may also study other important characteristics.

Method Name	Autism query	Filter
Basic	`autism OR ASD OR "Autism Spectrum Disorder"`	None
Major Topic	`autism[Majr] OR "Autism Spectrum Disorder"[Majr]`	None
Hit Filtering	Same as Basic	({gene} & (autism|asd)) in title or keywords
Major Topic Hit Filtering	Same as Major Topic	{gene} in title or keywords

Table 4: Search-and-filter methods used to find publications that relate a gene with ASD. The filter for 'Hit filtering' can be read as 'Both, the gene symbol and a reference to ASD must be found in the title or the keywords. Finding one in the title and one in the keywords is acceptable.'

The number of reports found by each search-and-filter method as well as the number of topic-independent reports and SFARI autism reports are presented in Table 5. Evaluating the methods against each other, Basic Search produces the most matches but also includes false positives such as [16]. Major Topic Search improves upon this by requiring autism to be explicitly specified as a major topic by the authors. However, this may still produce false positives for the gene such as in [9], which was returned by Major Topic Search. Major Topic Hit Filtering removes the irrelevant gene mentions by ensuring the gene is in the title or the keywords. While this method's restrictiveness now prevents almost all false positives from making it into the reports list, it may be a little too restrictive, given that it only finds 5 autism-related reports on SCN1A, while SFARI registers 29. Hit Filtering (without Major Topic) may thus achieve the best balance between avoiding papers that mention the specific gene or autism merely as background information, and not being too restrictive so that as many relevant publications as possible are found.

Gene	Topic-Independent Gene Reports	Basic Search	Major Topic	Hit Filtering	Major Topic Hit Filtering	SFARI Autism
SHANK3	719	498	231	152	70	62
MECP2	3612	547	127	60	19	32
NRXN1	499	195	69	21	12	53
SCN2A	747	105	51	19	13	51
SCN1A	1534	82	21	11	5	29

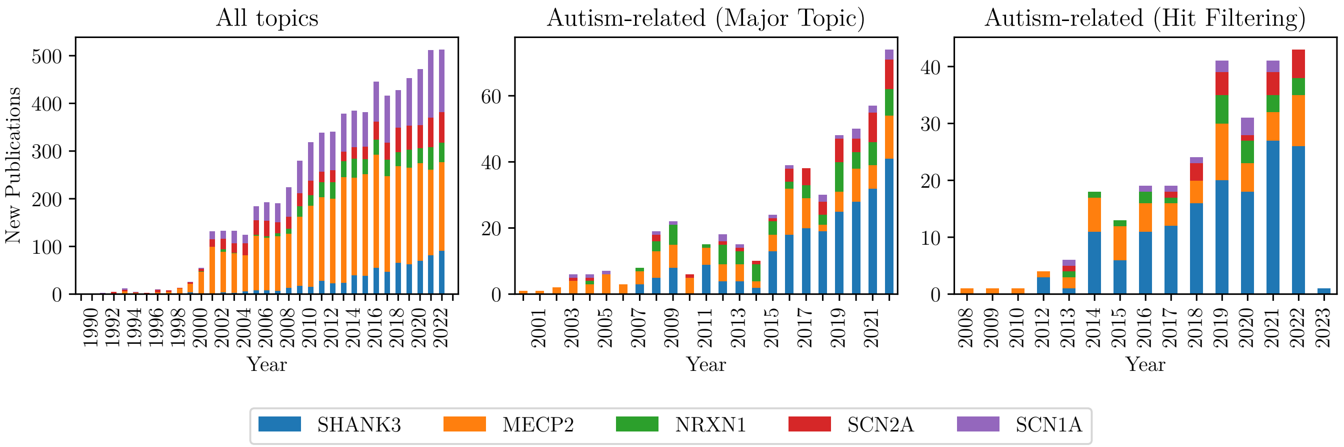
Table 5: Extension of Table 2. Number of articles found by each of the four search-and-filter methods. The total number of (topic-independent) reports published and the number of autism-related reports, as quoted by SFARI, are also shown for each gene.

The Major Topic Search results, Hit Filtering results, as well as the total number of reports on each gene per year, are presented in Figure 4. While they show the same overall trend as the Basic search, there are some key differences. Major Topic and Hit Filtering show a recent surge in MECP2 reports compared to the 2016 peak presented by the basic search. Major Topic's results still hint towards this peak in 2016, whereas Hit Filtering does not show this at all. Not enough data is available from the two more-restrictive methods to make a statement about NRXN1, SCN2A or SCN1A.

Analysing the plot showing the number of publications across all topics for each gene, we can observe that the link to autism resulted in a large surge in research for SHANK3, MECP2, NRXN1, and SCN1A. The first mention of SHANK3 in 2005, in relation to syndromic autism [11], and in 2006 linking it to idiopathic ASD [17] was the starting point of the significant rise in its research. Similarly, MECP2 was not much researched when it was first discovered in 1992 [18] but surged when it was linked to syndromic ASD in 2000 [19] and to idiopathic autism in 2002 [20]. NRXN1, just like the other two, was only researched very little until it was implicated with ASD in 2007 [21]. SCN1A and SCN2A were both connected to ASD in 2003 [22], though SCN1A was previously linked to syndromic autism [23]. Research into it increased rapidly while SCN2A's rise remains muted.

Interesting to note is that the three syndromic genes are those with the most research done. This matches with the statement by Ziats et al. that syndromic ASD is where most of the progress is made. [8]

Concluding this extension into different search techniques for finding autism-related


All topics

Autism-related (Major Topic)

Autism-related (Hit Filtering)

Figure 4: Extension of Fig. 3. Left is the total number of publications on SHANK3, MECP2, NRXN1, SNC2A and SCN1A per year, independent of their investigation topic. In the middle is a plot of the autism-related articles found by Major Topic Search per year. Right is the yearly trend in autism-related work done on each gene that is found using the Hit Filtering method.

reports on genes, it has been shown in particular by the similarity of the plots in Fig. 4 to that in Fig. 3 that the retrieved reports are indeed representative of the research space. The syndromic-to-idiopathic ratio in Fig. 4 (left) matching the hypothesis from [8] is supporting evidence that the five selected genes are representative of the research space. However, for this claim to be made compelling, more evidence is needed.

3.2 Part Two – Autism Genes

With a background on the gene scores and the progress of ASD genetics research, this section will study the properties of ASD-linked genes using their Gene Ontology annotations, as defined in the Data and Methods section.

3.2.1 Task 1

To retrieve the Gene Ontology annotations for each gene symbol in the SFARI list, I first convert the symbols to NCBI gene IDs to then retrieve the onotologies from the NCBI `gene2go` list (see Appendix A). For a detailed procedure, see the Data and Methods section. A sample of the converted gene IDs is in Table B.2. During the conversion process MSNP1AS, RP11-1407O15.2 and RPS10P2-AS1 were left without a Gene ID. The ID for MSNP1AS could be found with a manual search of the NCBI database, but not for the other two. As these are only two out of 695 genes in score 2, their effect on the outcome of the following analyses is negligible and I can leave them out without compromising my results.

3.2.2 Task 2

A preview of the GO terms associated with each SFARI gene is shown in Table B.3. Five genes have no corresponding GO term for Homo Sapiens in `gene2go`. A manual search of the Gene Ontology database (reference in Appendix A) provided the same results.

3.2.3 Task 3

As the final preparatory step, I split the genes with their annotations into three lists according to their score, only using the idiopathic genes (including syndromic genes implicated with ASD). The three resulting lists are shown as Tables B.4a-c.

3.2.4 Task 4

To identify common characteristics of genes with likely ASD-causing mutations, I count the occurrences of each ontology term and list the ten most common annotations, as shown in Tables 6a-c. While the terms' order differs among the scores, all genes are mostly related to the same GO terms, such as nucleus, protein binding, membrane, plasma membrane, etc.

These appear to be the areas where mutations in genes could cause ASD. Noticably, some of these areas are more specific than others. For instance, the nucleoplasm is the substance inside the nucleus, the cytosol is the area of the cytoplasm that has no organelles, and the plasma membrane is a specific type of membrane. As all plasma membranes are also membranes, it would be natural to assume that all genes annotated as relating to the former are also related to the latter, but not vice versa. Thus, more genes should be annotated as membrane than plasma membrane. However, this is not the case. Since `gene2go` only contains the explicit annotations of each gene without adding their generalisations, it appears that for ASD-linked genes more genes were annotated as plasma membrane than membrane. A possible explanation for this may be that the more general term is only used if the protein that is produced from the gene either interacts with a wide variety of membranes or the researcher making the annotation is unsure which membrane it interacts with.

In the following extension work I will provide a more thorough analysis of the annotation differences between score groups and of the SFARI genes compared to all annotated genes in humans.

GO term ID	GO term description	GO term count
GO:0005634	nucleus	177
GO:0005515	protein binding	170
GO:0005654	nucleoplasm	140
GO:0005886	plasma membrane	114
GO:0005829	cytosol	106
GO:0005737	cytoplasm	84
GO:0045944	positive regulation of transcription by RNA polymerase II	74
GO:0006357	regulation of transcription by RNA polymerase II	55
GO:0016020	membrane	53
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	53

(a) Score 1

GO term ID	GO term description	GO term count
GO:0005886	plasma membrane	495
GO:0005515	protein binding	470
GO:0005829	cytosol	308
GO:0005634	nucleus	302
GO:0005654	nucleoplasm	243
GO:0005737	cytoplasm	243
GO:0016020	membrane	167
GO:0070062	extracellular exosome	89
GO:0046872	metal ion binding	89
GO:0003723	RNA binding	85

(b) Score 2

GO term ID	GO term description	GO term count
GO:0005515	protein binding	75
GO:0005829	cytosol	61
GO:0005737	cytoplasm	48
GO:0005634	nucleus	47
GO:0005886	plasma membrane	47
GO:0005654	nucleoplasm	32
GO:0016020	membrane	25
GO:0046872	metal ion binding	17
GO:0003723	RNA binding	17
GO:0005524	ATP binding	17

(c) Score 3

Table 6: The ten most commonly annotated Gene Ontology terms for the genes in each SFARI score group.

3.2.5 Extension

To better compare the GO term counts among the score groups, Tables 7a-c have additional columns showing the '% of Genes' in a given score group which have the annotation, a 'p-value', as determined by Fischer's exact text, and the 'odds-ratio'. Adding the percentage of genes that have a Gene Ontology term helps compare counts, as the scores have different numbers of genes. The p-value indicates how likely it is to obtain a value as extreme or more extreme than by chance. The closer the p-value is to 0, the more likely it is that the group's term count is significantly higher or lower than the average. If the odds-ratio > 1, the group has a higher term count than average, and if it is < 1, the group's term count is lower. I also included the category to which each GO term belongs for a later comparison.

We can see that in score 1 (Table 7a) most of the listed GO terms that have a significantly different % of Genes than the other scores (p-value < 0.05) also have a higher-than-average representation (odds-ratio> 1). In score 2 (Table 7b) most listed terms with a p-value < 0.05 are underrepresented (odds-ratio < 1). The only exception to this is the plasma membrane. It is more represented in score 2 and less represented in score 1.

To explain the general overrepresentation in score 1 we must first analyse Table 8. This table shows the most commonly annotated genes in the full SFARI gene set and the p-value is calculated in comparison to all annotations of human genes in `gene2go`. Noticably, nine of the ten most common terms are significantly overrepresented in genes linked to idiopathic ASD. These are likely the areas which have been either studied more excessively in regards to autism, or where a gene mutation is more likely to cause autism than a mutation in other areas, or a link between the mutation and autism is more easily made. If these are indeed studied more, or links to autism are more easily found, it would also explain why the most common terms in score 1 have an odds-ratio > 1. As the areas are easier to study in regards to autism, studies of them are more likely to produce strong evidence, placing the gene into score 1.

Further, as activity in the plasma membrane effects gene expression more indirectly than activity in the nucleus, studying proteins that act in the plasma membrane is less likely to produce strong evidence while studying proteins acting in the nucleus or nucleoplasm is more likely to, placing 'plasma membrane' into score 2 or 3 and nucleus and nucleoplasm into score 1. This hypothesis is supported by genes that are directly related to the regulation of transcription, and thus gene expression, have the highest odds-ratios in score 1 while not on the lists for scores 2 and 3.

In score 3 (Table 6c) none of the ten most commonly annotated terms have a p-value below the common significance threshold of 0.05, even though some of the terms have odds-ratios comparable to those in score 1 or 2. This is because a more extreme odds-ratio is needed in smaller groups. For the same ratios to be significant, they need to be obtained over a larger set of genes than are in score 3.

GO term ID	GO term description	Term count	Category	% of Genes	p-value	odds-ratio
GO:0005634	**nucleus**	177	Component	82.7	**0.000**	2.41
GO:0005515	**protein binding**	170	Function	79.4	**0.008**	1.62
GO:0005654	**nucleoplasm**	140	Component	65.4	**0.000**	2.21
GO:0005886	**plasma membrane**	114	Component	53.3	**0.022**	0.68
GO:0005829	cytosol	106	Component	49.5	0.380	1.15
GO:0005737	cytoplasm	84	Component	39.3	0.867	1.03
GO:0045944	**pos. regulation of transcription by RNA polymerase II**	74	Process	34.6	**0.000**	2.61
GO:0006357	**regulation of transcription by RNA polymerase II**	55	Process	25.7	**0.000**	2.79
GO:0016020	membrane	53	Component	24.8	0.643	0.90
GO:0000981	**DNA-binding transcription factor activity, RNA polymerase II-specific**	53	Function	24.8	**0.001**	2.34

(a) Score 1

GO term ID	GO term description	Term count	Category	% of Genes	p-value	odds-ratio
GO:0005886	**plasma membrane**	495	Component	71.9	**0.000**	1.64
GO:0005515	**protein binding**	470	Function	68.3	**0.000**	0.57
GO:0005829	**cytosol**	308	Component	44.8	**0.004**	0.68
GO:0005634	**nucleus**	302	Component	43.9	**0.000**	0.46
GO:0005654	**nucleoplasm**	243	Component	35.3	**0.000**	0.58
GO:0005737	**cytoplasm**	243	Component	35.3	**0.098**	0.79
GO:0016020	membrane	167	Component	24.3	0.760	0.96
GO:0070062	extracellular exosome	89	Component	12.9	0.924	1.02
GO:0046872	**metal ion binding**	89	Function	12.9	**0.017**	0.65
GO:0003723	**RNA binding**	85	Function	12.4	**0.041**	0.66

(b) Score 2

GO term ID	GO term description	Term count	Category	% of Genes	p-value	odds-ratio
GO:0005515	protein binding	75	Function	82.4	0.087	1.62
GO:0005829	cytosol	61	Component	67.0	0.051	1.57
GO:0005737	cytoplasm	48	Component	52.7	0.054	1.56
GO:0005634	nucleus	47	Component	51.6	0.910	0.95
GO:0005886	plasma membrane	47	Component	51.6	0.364	0.79
GO:0005654	nucleoplasm	32	Component	35.2	0.233	0.72
GO:0016020	membrane	25	Component	27.5	0.690	1.10
GO:0046872	metal ion binding	17	Function	18.7	0.286	1.34
GO:0003723	RNA binding	17	Function	18.7	0.372	1.33
GO:0005524	ATP binding	17	Function	18.7	0.104	1.60

(c) Score 3

Table 7: Extension of Table 6. The ten most commonly annotated Gene Ontology terms for the genes in each SFARI score group. The p-value and odds-ratio are computed in relation to the other idiopathic genes in the SFARI list. The GO term description is highlighted where the p-value < 0.05.

GO term ID	GO term description	Term count	Category	% of Genes	p-value	odds
GO:0005515	**protein binding**	788	Function	72.4	**0.0**	17.38
GO:0005886	**plasma membrane**	700	Component	64.3	**0.0**	3.95
GO:0005634	**nucleus**	597	Component	54.9	**0.0**	2.00
GO:0005829	**cytosol**	525	Component	48.3	**0.0**	4.01
GO:0005654	**nucleoplasm**	462	Component	42.5	**0.0**	7.92
GO:0005737	**cytoplasm**	416	Component	38.2	**0.0**	1.76
GO:0016020	membrane	270	Component	24.8	0.8768	0.98
GO:0045944	**positive regulation of transcription by RNA polymerase II**	176	Process	16.2	**0.0**	5.47
GO:0046872	**metal ion binding**	163	Function	15.0	**0.0**	1.85
GO:0003723	**RNA binding**	160	Function	14.7	**0.0**	3.25
...	...	...	...	...	...	...
GO:0015293	symporter activity	1	Function	0.1	0.7373	0.42
GO:0050891	multicellular organismal water homeostasis	1	Process	0.1	0.2156	4.19
GO:2001271	negative regulation of cysteine-type endopeptidase activity involved in execution phase of apoptosis	1	Process	0.1	0.067	15.36

Table 8: The ten most commonly annotated Gene Ontology terms for all idiopathic genes in the SFARI list. The p-value and odds-ratio are computed in relation to all annotated human genes in `gene2go`. The GO term description is highlighted where the p-value < 0.05.

3.2.6 Task 5

Next I use PantherDB to get the main biological processes in each of the score groups, as described in the methods. The results are plotted in Figures 5, 6 and 7. PantherDB found no results for four score 1 genes, seven score 2 genes and one score 3 gene, but missing twelve out of 998 genes with NCBI IDs will unlikely have a significant effect on the results. The plots use the percentage of genes that are part of the biological process as their x-axis, as that makes them more comparable than stating the absolute number of genes for each process.

From the figures we can see that most genes in all three scores (between 60% and 70% for each score) are involved cellular processes. Many are also involved in biological regulation as well as metabolic processes. Many of the processes that mutations in the SFARI genes may interrupt are linked to known issues in individuals with ASD, such as metabolism [24], development [25], rhythm [26], response to stimulus [27], or the immune system [28]. The prevalence with which ASD-linked genes are found in each process is very similar across all scores.

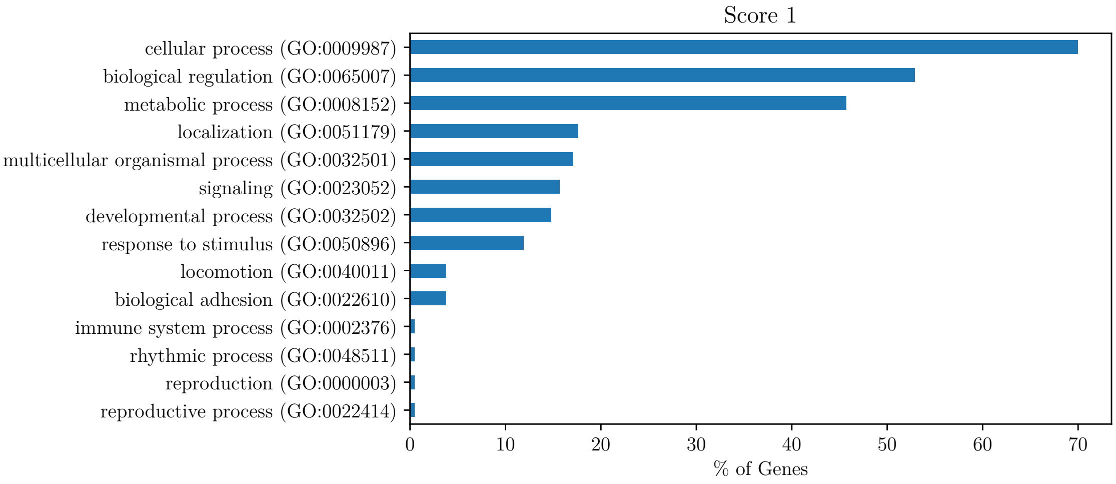


Figure 5: Biological processes of genes with SFARI score 1 as generated by PantherDB.

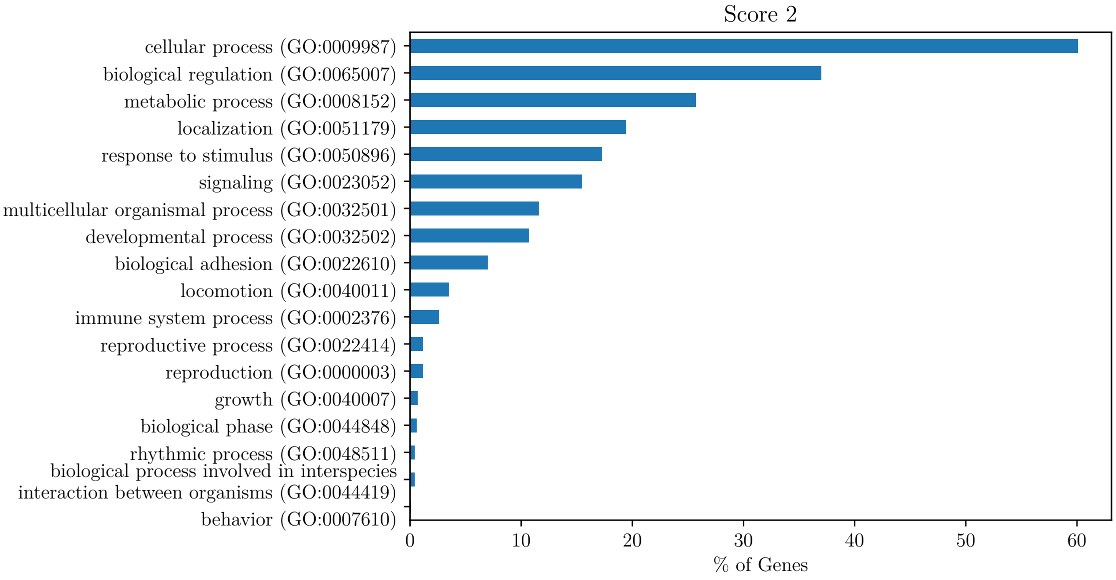


Figure 6: Biological processes of genes with SFARI score 2 as generated by PantherDB.

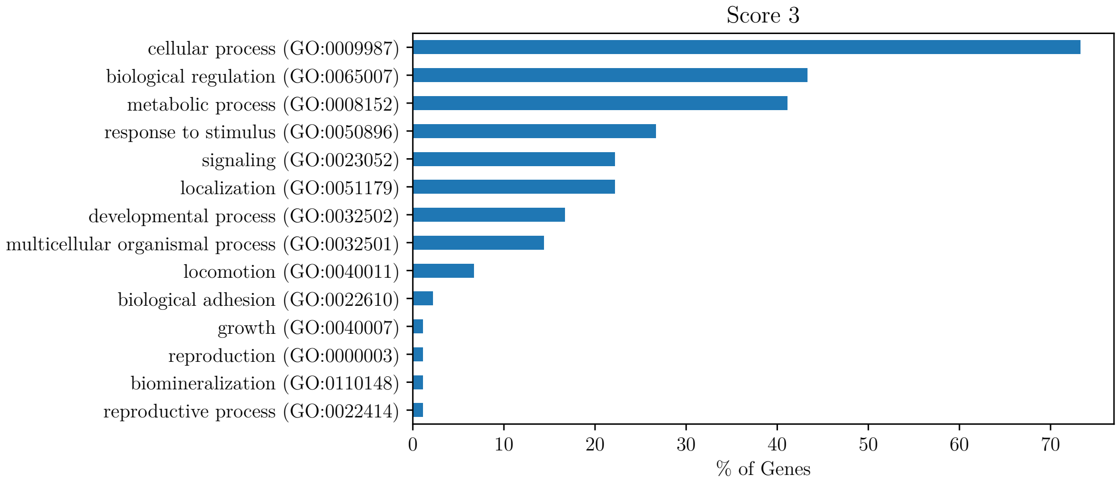


Figure 7: Biological processes of genes with SFARI score 3 as generated by PantherDB.

3.2.7 Extension

Given that the plots in Figures 5 - 7 are derived from the genes' GO terms just like Table 8, we might expect 'regulation of transcription by RNA polymerase II' to be present in one or more of the figures, as this is the most common process in SFARI genes according to Table 8. While PantherDB's use of phylogenic trees to augment each gene's annotations using its orthologs, paralogs and xenologs [29] adds further terms to each gene, that should not cause other GO terms, such as the above mentioned, to get removed from genes and fall to a 0% gene cover.

There appears to be no overlap between the gene2go results (Table 8) and pantherdb results (Figures 5 - 7). This is because the Gene Ontology classes shown in the pantherdb plot are kept very general. Most terms are direct children or grandchildren of the top-level 'Biological Process' term (see GO:0008150 in [30]). However, these terms can be expanded to reveal the more specific terms for each gene. Representatively, the most common expanded biological processes for score 1 genes are shown in Table 9. The tables for the other scores can be found in Tables B.5 and B.6. We can see that these do include the 'regulation of transcription by RNA polymerase II'.

With PantherDB we have augmented the ontologies with those of each gene's orthologs, paralogs and xenologs and filtered them to contain only biological processes. While the broad topics are too imprecise to give us much insight into what processes are effected, the expanded list clearly shows that mutations disturbing the correct function of transcription by RNA polymerase II, gene expression as well as neural and synaptic processes (chemical synaptic transmission, regulation of membrane potential) are most often correlated to ASD.

GO term ID	GO term description	GO term count
GO:0006366	transcription by RNA polymerase II	53
GO:0006357	regulation of transcription by RNA polymerase II	33
GO:0045944	positive regulation of transcription by RNA polymerase II	13
GO:0007268	chemical synaptic transmission	11
GO:0000122	negative regulation of transcription by RNA polymerase II	9
GO:0006351	DNA-templated transcription	7
GO:0006338	chromatin remodeling	6
GO:0097553	calcium ion transmembrane import into cytosol	6
GO:0070509	calcium ion import	6
GO:0050804	modulation of chemical synaptic transmission	6

Table 9: The ten most common biological processes among genes with SFARI score 1. Extracted from PantherDB using its API through the python `bioservices` library (see Appendix A).

3.3 Part Three – Autism Gene Networks

After looking at the genes' GO terms, and in particular their biological processes, for SFARI score in the previous section, this section will investigate the genes' functions by looking at their protein-protein interaction network and analysing their pathways. I will focus only on score 1 genes to guarantee a higher level of confidence that the obtained results are truly relevant to ASD-linked genes. Including score 2 and 3 genes would not guarantee this, as the evidence implicating them with ASD is not without doubt.

3.3.1 Task 1

Using StringDB's Multiple Proteins function (see Appendix A) I first create a protein-protein interaction network from the NCBI gene IDs in score 1. Details of the resulting network are in Table B.7.

3.3.2 Tasks 2 & 3

Next, I cluster the interactions using the MCL clustering algorithm with the inflation parameter set to 3 (details on the algorithm and parameter choice in Appendix A). The clustered network is visualised in Figure B.1. From the network I select the two largest clusters as representatives for the protein-protein interactions of score 1 and use PantherDB to identify their pathways, treating the proteins from StringDB as equivalent to their corresponding genes.

Given that the proteins are predicted to interact with one another, they share surprisingly few pathways. Of the 33 proteins in the largest cluster, only 4 (12%) share the 'Huntington disease' pathway or the 'Ionotropic glutamate receptor pathway' (Fig. 8) and of the 25 proteins in the second-largest cluster only 7 (28%) share the 'Wnt signaling pathway' (Fig. 9). The other listed pathways are shared by still fewer proteins.

The clusters share only a few pathways between them. This is to be expected, as the clustering process aims to group proteins that interact with one another and thus are more likely to share pathways as well as separate proteins that don't interact with each other or with the same third party and are therefore less likely to share pathways.

At first impression, it may seem surprising to see the 'Huntington disease' (HD) pathway in both clusters and 'Parkinson disease' (PD) pathway in cluster 1, as other disorders that commonly co-occur with syndromic ASD seem more likely to appear. However, PD has been shown to be more prevalent in adults with ASD [31]. While HD is not a disease that typically co-occurs with ASD, they share similar symptoms to the extent that Juvenile Huntington's disease can be misdiagnosed as ASD [32], which may be partially explained by mutations of genes in the same pathways causing either HD or ASD. This, though, is a hypothesis that would need further investigation.

The 'Wnt signalling pathway', which is the most prevalent in the second-largest cluster (Table 9), ' regulates important aspects of early development, such as the formation of organs (organogenesis) or formation of neural networks (neural patterning) [33]. Mutations in this pathway's genes potentially causing ASD is in accordance with 'signalling', 'developmental process' and 'growth' being significant biological processes of genes in the SFARI list (Fig. 5 - 7).

Similarly, the 'TGF-β signalling pathway' is responsible for cell growth, cell differentiation (change of cell type, typically from a stem cell to a more specialised type), cell migration and apoptosis (cell death), which also aligns with the 'signalling', 'development' and 'growth' processes.

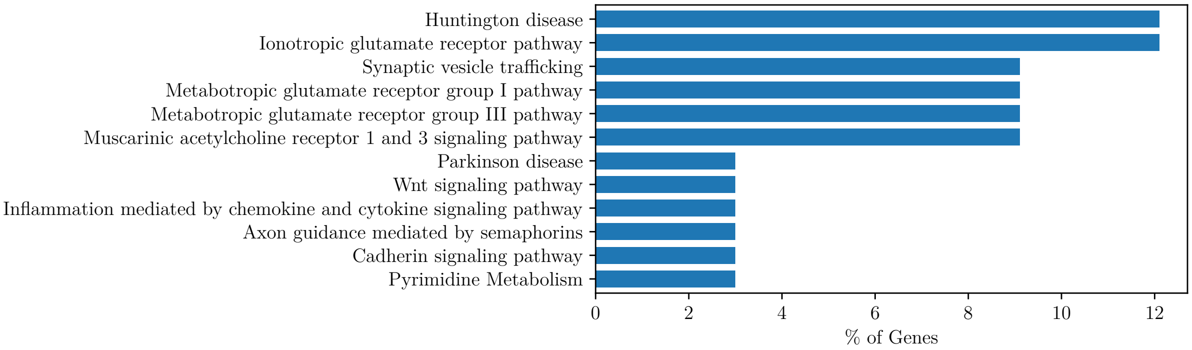


Figure 8: High-level pathways of genes in the largest protein-protein interaction cluster with SFARI score 1.

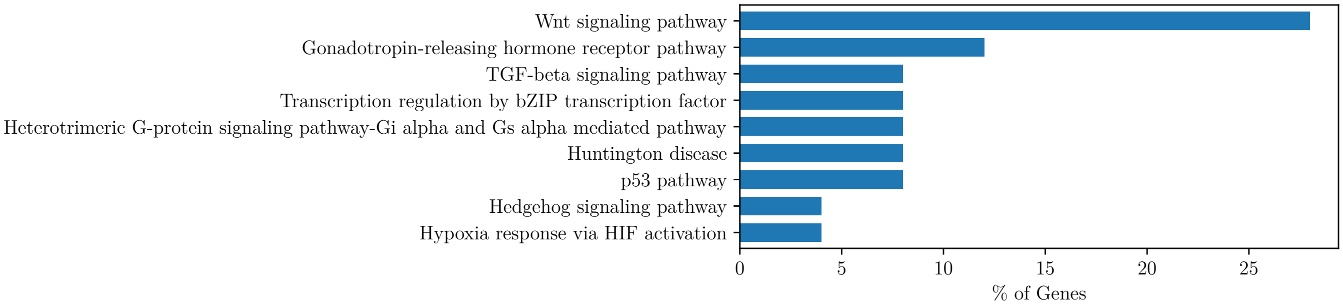


Figure 9: High-level pathways of genes in the second-largest protein-protein interaction cluster with SFARI score 1.

3.3.3 Extension

As with the biological processes derived from PantherDB, I will expand the high-level pathways from Fig. 8 and 9 to reveal the more specific pathways that each of the clusters' genes is involved in. The resulting pathways are found in Fig. 10 and 11. Not only do the pathways in the resulting plots have a much higher gene correspondence, but they also give a much clearer view of what the proteins in each cluster do.

Most pathways in the largest cluster are related to neural processes. The most common pathway 'Neuronal System', with over 50% of the cluster's genes relating to this, makes this rather clear. The other pathways, apart from 'Signal Transduction' and 'Developmental Biology' are also related to synapses or axons and therefore neurons.

The second-largest cluster, on the other hand, is more related to gene expression. Again, this is made clear by over 50% of the cluster's genes sharing the 'Gene expression (Transcription)' pathway. Other than 'Developmental Biology' and 'Disease' they are all directly related to this topic.

Referring back to the common biological processes in Table 9, it contains mostly processes related to either the neuronal system (largest cluster) or gene expression (second-largest cluster).

Based on the analyses of the prevalent Gene Ontology terms and common biological processes of ASD-linked genes in part two and the study of common pathways in the two largest protein-protein interaction clusters it can be deduced with confidence that the two main areas in which genes whose mutations may cause ASD act are regulating gene expression and modulation of neural activity. For more details on the specific biological processes, refer to Figures 5 - 7 and Tables 9, B.5 and B.6. For details on the specific pathways, refer to Figures 8 - 11.

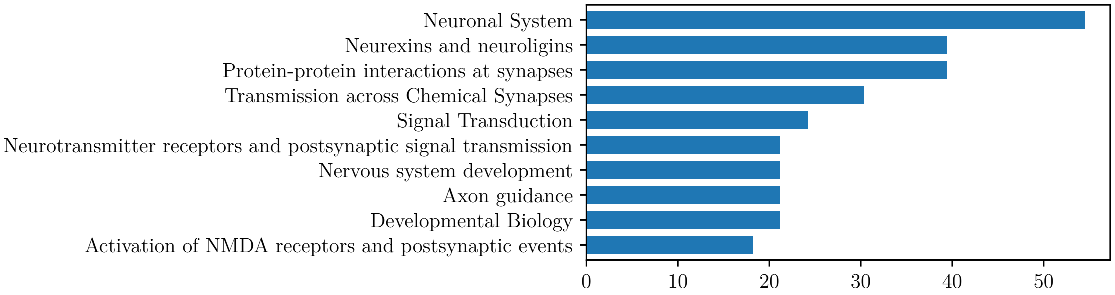


Figure 10: Expanded pathways of genes in the largest protein-protein interaction cluster with SFARI score 1.

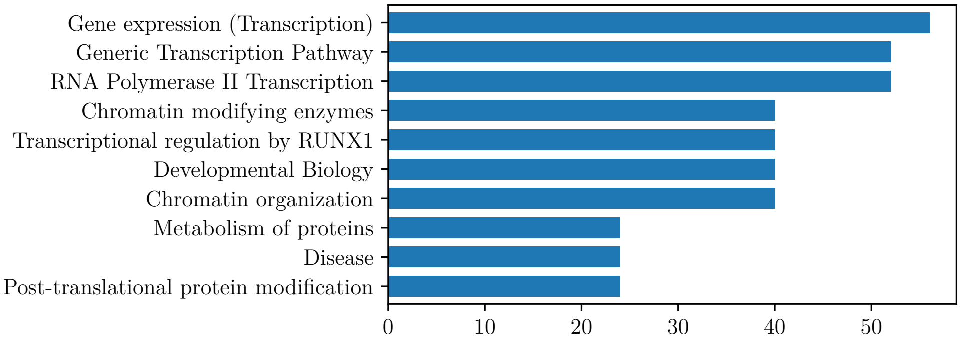


Figure 11: Expanded pathways of genes in the second-largest protein-protein interaction cluster with SFARI score 1.

4 Discussion

In this report, I have analysed the state of research in autism genetics in three parts entailing a quantitative interpretation of the literature, a study of the gene ontology and a network-aided pathway analysis.

Literature As discussed in the Results section, the publication volume among the five most-reported-on has been steadily increasing over the past years. I went into detail on their initial link to ASD and their research trend over time since then. I also analysed how this link affected their total research volume across all topics. With finding all and exclusively autism-related papers for each gene being non-trivial, I explore a variety of search-and-filter methods and their effectiveness in searching for reports. As these are only five of the 1095 genes in the SFARI database, their representativeness of the whole field has to be justified. Evidence for this is that the amount of research done on the three syndromic genes is significantly higher than that done on the two that are not syndromic, as this closely matches the hypothesis stated in the literature (see section 3.1.7). However, this syndromic to non-syndromic research ratio could be merely incidental given my sample consists of only 5 genes. Further investigation into how representative these genes are is needed, for instance by selecting a larger, possibly random sample of genes to compare the stated results to.

Gene Analysis In Part Two of the Results I investigate the components with which the SFARI genes interact and the biological processes in which they are involved through a study of their Gene Ontology. While the most common Gene Ontology classes that the genes are annotated with, according to NCBI's `gene2go` mapping, are primarily biological components, the results obtained from PantherDB are filtered for biological processes. We can see from this that the ASD-linked genes mostly act in and around the nucleus and membrane, engaging in processes related to gene expression. I have thoroughly discussed the difference in term prevalence between different SFARI scores and of the entire SFARI gene group compared to all annotated human genes in `gene2go`. However, in this report's analysis, I have disregarded the evidence supporting each GO annotation as well as the qualifier specifying how the gene relates to the annotation (e.g. 'located_in' or 'enables'). The evidence may shine further light on the state of research while the qualifier could provide more information on how the genes interact with each component or process.

I have extensively contrasted the results extracted from `gene2go` and PantherDB, expanding the latter's process terms to be more comparable to those of the former. This concluded in a comparison between mostly components from `gene2go`, high-level biological processes from the PantherDB website and the expanded processes from the PantherDB API.

Gene Networks The third part uses StringDB to form a protein-protein interaction network of the genes with score 1, cluster them with the MCL algorithm and analyse the pathways of the two largest clusters. I used the pathways returned directly from PantherDB's website as well as the broader and more detailed set returned by their API. In the Results section, I specify how the clusters form two distinct groups of genes involved in the two main processes where mutations in genes are likely to be related to ASD. These are the modulation of neural activity and gene expression. While the website's results are more specific, the two fields can more easily be distinguished in the API's returned pathways. This is in line with the results found in Part Two, as detailed in the Part Three Results section.

An analysis of more clusters or other score groups may also prove insightful in determining further biological areas related to ASD or gathering additional evidence on the hypotheses made on the state of research.

The conclusions and hypotheses made in this report are by no means certain and may need further reworking or may be shown to be entirely incorrect, provided the investigations' breadth but therefore lack of depth. Depending on the claim, a more detailed bioinformatics study or an experimental approach would be appropriate to gather further evidence.

References

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders.* Fifth Edition. American Psychiatric Association, May 22, 2013.
2. Canitano R. Epilepsy in autism spectrum disorders. *European Child & Adolescent Psychiatry.* 2007; 16(1):61–66.
3. Singh SK, Eroglu C. Neuroligins Provide Molecular Links Between Syndromic and Nonsyndromic Autism. *Science Signaling.* 2013; 6(283).
4. Zeidan J, Fombonne E, Scorah J, et al. Global prevalence of autism: A systematic review update. *Autism Research.* 2022; 15(5):778–790.
5. Cdc. *Basics About Autism Spectrum Disorder (ASD) — NCBDDD — CDC.* Centers for Disease Control and Prevention. Mar. 31, 2022. URL: `https://www.cdc.gov/ncbddd/autism/facts.html` [visited on 11/25/2022].
6. Simons Foundation Autism Research Initiative. *Human Gene Module.* SFARI Gene. URL: `https://gene.sfari.org/database/human-gene/` [visited on 11/27/2022].
7. Office For National Statistics (ons). *Web Scraping Policy.* Apr. 27, 2020.
8. Ziats CA, Patterson WG, Friez M. Syndromic Autism Revisited: Review of the Literature and Lessons Learned. *Pediatric Neurology.* 2021; 114():21–25.
9. Gerosa L, Francolini M, Bassani S, Passafaro M. The Role of Protocadherin 19 (PCDH19) in Neurodevelopment and in the Pathophysiology of Early Infantile Epileptic Encephalopathy-9 (EIEE9): PCDH19 and Neurons. *Developmental Neurobiology.* 2019; 79(1):75–84.
10. Adam MP, Everman DB, Mirzaa GM, et al., eds. *GeneReviews®.* Seattle (WA): University of Washington, Seattle, 1993.
11. Phelan K, Rogers RC, Boccuto L. Phelan-McDermid Syndrome. *GeneReviews®.* Ed. by Adam MP, Everman DB, Mirzaa GM, et al. Seattle (WA): University of Washington, Seattle, 1993.
12. Khaikin Y, Mercimek-andrews S. STXBP1 Encephalopathy with Epilepsy. *GeneReviews®.* Ed. by Adam MP, Everman DB, Mirzaa GM, et al. Seattle (WA): University of Washington, Seattle, 1993.
13. Kaur S, Christodoulou J. MECP2 Disorders. *GeneReviews®.* Ed. by Adam MP, Everman DB, Mirzaa GM, et al. Seattle (WA): University of Washington, Seattle, 1993.
14. Van esch H. MECP2 Duplication Syndrome. *GeneReviews®.* Ed. by Adam MP, Everman DB, Mirzaa GM, et al. Seattle (WA): University of Washington, Seattle, 1993.
15. *MeSH Browser.* URL: `https://meshb.nlm.nih.gov/record/ui?ui=D001321` [visited on 11/29/2022].
16. O'roak BJ, Vives L, Fu W, et al. Multiplex Targeted Sequencing Identifies Recurrently Mutated Genes in Autism Spectrum Disorders. *Science.* 2012; 338(6114):1619–1622.

17. Durand CM, Betancur C, Boeckers TM, et al. Mutations in the gene encoding the synaptic scaffolding protein SHANK3 are associated with autism spectrum disorders. *Nature Genetics*. 2007; 39(1):25–27.

18. Lewis JD, Meehan RR, Henzel WJ, et al. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*. 1992; 69(6):905–914.

19. Orrico A, Lam CW, Galli L, et al. *MECP2* mutation in male patients with non-specific X-linked mental retardation. *FEBS Letters*. 2000; 481(3):285–288.

20. Beyer KS, Blasi F, Bacchelli E, et al. Mutation analysis of the coding sequence of the MECP2 gene in infantile autism. *Human Genetics*. 2002; 111(4):305–309.

21. Bourgeron T. The possible interplay of synaptic and clock genes in autism spectrum disorders. *Cold Spring Harbor Symposia on Quantitative Biology*. 2007; 72():645–654.

22. Weiss LA, Escayg A, Kearney JA, et al. Sodium channels SCN1A, SCN2A and SCN3A in familial autism. *Molecular Psychiatry*. 2003; 8(2):186–194.

23. Baulac S, Gourfinkel-an I, Picard F, et al. A second locus for familial generalized epilepsy with febrile seizures plus maps to chromosome 2q21-q33. *American Journal of Human Genetics*. 1999; 65(4):1078–1085.

24. Manzi B, Loizzo AL, Giana G, Curatolo P. Autism and Metabolic Diseases. *Journal of Child Neurology*. 2008; 23(3):307–314.

25. Lord C, Cook EH, Leventhal BL, Amaral DG. Autism Spectrum Disorders. *Neuron*. 2000; 28(2):355–363.

26. Amos P. Rhythm and timing in autism: learning to dance. *Frontiers in Integrative Neuroscience*. 2013; 7().

27. Lovaas OI, Koegel RL, Schreibman L. Stimulus overselectivity in autism: A review of research. *Psychological Bulletin*. 1979; 86(6):1236–1254.

28. Warren RP, Margaretten NC, Pace NC, Foster A. Immune abnormalities in patients with autism. *Journal of Autism and Developmental Disorders*. 1986; 16(2):189–197.

29. Mi H, Huang X, Muruganujan A, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*. 2017; 45():D183–D189.

30. Carbon S, Mungall C. *Gene Ontology Data Archive*. Version Number: 2022-11-03 Type: dataset. July 2, 2018.

31. Starkstein S, Gellar S, Parlier M, Payne L, Piven J. High rates of parkinsonism in adults with autism. *Journal of Neurodevelopmental Disorders*. 2015; 7(1):29.

32. Oosterloo M, Bijlsma EK, Die-smulders C de, Roos RAC. Diagnosing Juvenile Huntington's Disease: An Explorative Study among Caregivers of Affected Children. *Brain Sciences*. 2020; 10(3):155.

33. Komiya Y, Habas R. Wnt signal transduction pathways. *Organogenesis*. 2008; 4(2):68–75.

34. Dongen Sv. Graph clustering by flow simulation. 2000.

A Data Sources and Tools

- Python (version 3.9.7) with libraries:
  - Pandas (version 1.3.4) for data manipulation
  - Matplotlib (version 3.4.3) for data visualisations
  - Biopython (version 1.79) for access to the PubMed API
  - Bioservices (version 1.10.0) for access to the PantherDB API
- SFARI gene list released on 11/07/2022, retrieved on 11/14/2022 at `https://gene.sfari.org//wp-content/themes/sfari-gene/utilities/download-csv.php?api-endpoint=genes`.
- NCBI `Homo_sapiens.gene_info` list, retrieved on 11/14/2022 at `https://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz`.
- NCBI `gene2go` list, retrieved on 11/14/2022 at `https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz`.
- PantherDB version 17.0 on 11/14/2022 at `http://www.pantherdb.org`.
- Reactome pathways index on 11/14/2022 at `https://reactome.org/download/current/ReactomePathways.txt`.
- StringDB website for *Multiple Proteins* at `https://string-db.org`.

MCL clustering The Markov Cluster Algorithm (MCL) is an unsupervised learning algorithm that operates on networks to find clusters of nodes that are tightly interlinked among themselves but sparsely linked to other nodes or clusters. It has a single tunable parameter: inflation. A higher inflation value (5, 6) results in more and smaller clusters whereas a lower value (1.3, 1.4) results in fewer and larger clusters. For this analysis I am choosing a middle value of 3. [34]

B Additional Figures and Tables

B.1 Part One

Gene	1993	’99	2000	’01	’02	’03	’04	’05	’06	’07	’08	’09	’10	’11	’12	’13	’14	’15	’16	’17	’18	’19	’20	’21	’22	’23
SHANK3	1	-	-	-	-	-	-	1	-	4	7	9	5	18	17	17	32	27	40	37	49	44	53	63	73	1
MECP2	3	1	7	4	9	8	9	18	12	19	17	23	23	32	26	33	39	45	46	35	22	34	30	22	29	-
NRXN1	-	-	-	-	-	-	1	-	-	1	6	15	8	18	18	17	16	18	6	11	9	18	10	10	13	-
SCN2A	-	-	-	-	-	1	2	-	-	-	2	-	1	1	1	2	5	4	7	8	11	16	9	19	16	-
SCN1A	1	-	-	-	-	1	2	1	-	-	1	1	-	2	6	4	2	5	3	5	6	8	11	13	10	-

Table B.1: Number of publications returned by the basic search for ASD-related reports on each gene.

Gene symbol	Gene name	NCBI Gene ID
ABAT	4-Aminobutyrate aminotransferase	18
ABCA10	ATP-binding cassette, sub-family A (ABC1), member 10	10349
ABCA13	ATP binding cassette subfamily A member 13	154664
...	...	...
ZSWIM6	Zinc finger SWIM-type containing 6	57688
ZWILCH	Zwilchkinetochore protein	55055
MSNP1AS	Moesinpseudogene 1, antisense	4479
RP11-1407O15.2	-	-
RPS10P2-AS1	Ribosomal protein S10 pseudogene 2 anti-sense 1	-

Table B.2: Preview of the Gene symbol to NCBI Gene ID (Entrez UID) mapping. Two gene symbols remain without a corresponding gene ID.

Gene symbol	Gene ID	GO term ID	GO evidence	GO qualifier	GO term	GO category
ABAT	18	GO:0001666	IEA	involved_in	response to hypoxia	Process
ABAT	18	GO:0003867	IDA	contributes_to	4-aminobutyrate transaminase activity	Function
...	...	...	...	...	...	...
ZWILCH	55055	GO:1990423	IPI	part_of	RZZ complex	Component

Table B.3: Preview of the SFARI gene list with added gene ontologies from `gene2go`. A single gene may have multiple ontologies, thus multiple rows may refer to the same gene with different ontologies. A single ontology can have multiple genes. Gene ID refers to the NCBI Gene ID assigned to each symbol in Table B.2

B.2 Part Two

Gene symbol	Gene ID	GO term ID	GO term description
ACTB	60	GO:0000079	regulation of cyclin-dependent protein serine/threonine kinase activity
ACTB	60	GO:0000776	kinetochore
...	...	...	...
ZNF462	58499	GO:0046872	metal ion binding

(a) Score 1

Gene symbol	Gene ID	GO term ID	GO term description
ABA	18	GO:0001666	response to hypoxia
ABA	18	GO:0003867	4-aminobutyrate transaminase activity
...	...	...	...
ZWILC	55055	GO:1990423	RZZ complex

(b) Score 2

Gene symbol	Gene ID	GO term ID	GO term description
ABL2	27	GO:0000287	magnesium ion binding
ABL2	27	GO:0001784	phosphotyrosine residue binding
...	...	...	...
YWHAZ	7534	GO:0140311	protein sequestering activity

(c) Score 3

Table B.4: Gene Ontologies split by the corresponding gene's SFARI score.

GO term ID	GO term description	GO term count
GO:0006366	transcription by RNA polymerase II	51
GO:0006357	regulation of transcription by RNA polymerase II	36
GO:0007155	cell adhesion	23
GO:0007268	chemical synaptic transmission	22
GO:0099500	vesicle fusion to plasma membrane	15
GO:0016477	cell migration	13
GO:0050804	modulation of chemical synaptic transmission	12
GO:0000122	negative regulation of transcription by RNA polymerase II	12
GO:0007165	signal transduction	11
GO:0035249	synaptic transmission, glutamatergic	11

Table B.5: The ten most common biological processes among genes with SFARI score 2. Extracted from PantherDB using its API through the python `bioservices` library (see Appendix A).

GO term ID	GO term description	GO term count
GO:0006357	regulation of transcription by RNA polymerase II	9
GO:0006366	transcription by RNA polymerase II	9
GO:0016477	cell migration	4
GO:0006468	protein phosphorylation	4
GO:0030154	cell differentiation	3
GO:0007254	JNK cascade	2
GO:0008284	positive regulation of cell population proliferation	2
GO:0008543	fibroblast growth factor receptor signaling pathway	2
GO:0008283	cell population proliferation	2
GO:0009887	animal organ morphogenesis	2

Table B.6: The ten most common biological processes among genes with SFARI score 3. Extracted from PantherDB using its API through the python `bioservices` library (see Appendix A).

B.3 Part Three

Number of nodes	213
Number of edges	1552
Average node degree	14.6

Table B.7: Details of the protein-protein interaction network generated by StringDB (see Appendix A) for score 1 genes.

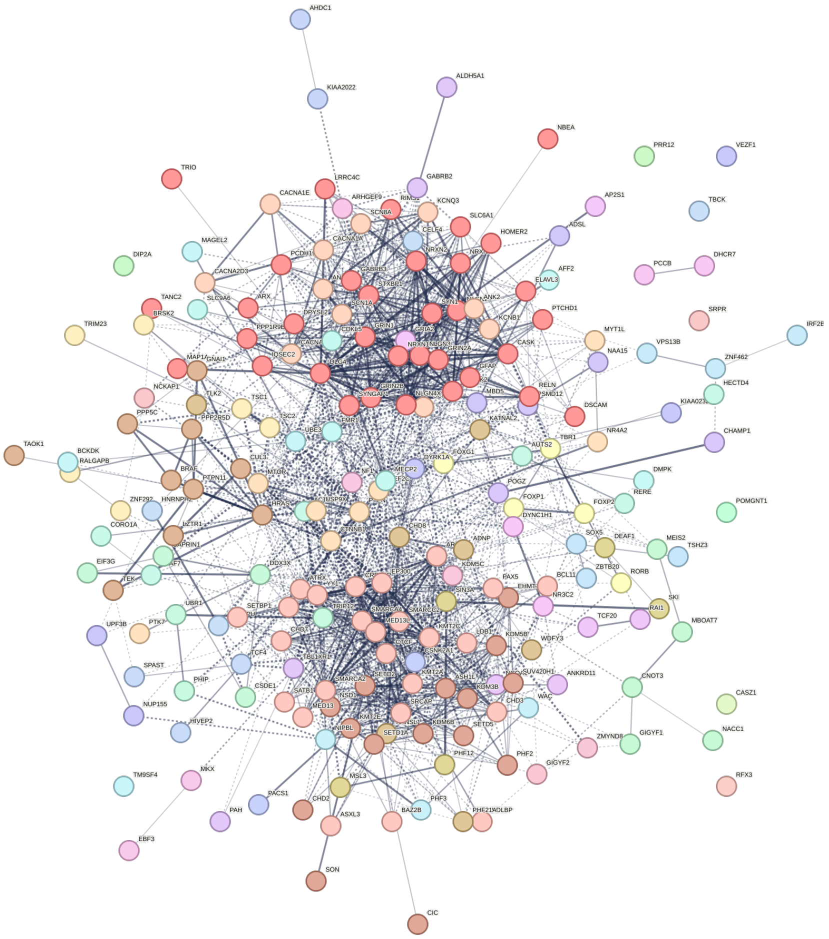


Figure B.1: Visualisation of the clustered protein-protein interaction network generated on SFARI score 1 genes using StringDB. Clustering was done using the MCL algorithm with inflation = 3. Continuous lines are between nodes sharing a cluster. Dotted lines are between nodes in different clusters. The line thickness represents the confidence with which the interaction was predicted. Nodes in the same cluster share the same colouring.