Bioinformatics Coursework 2 – Exploring Autism Genes

November 2020

1 Introduction

Autism affects approximately 1 in 160 children worldwide [2], with the onset of symptoms appearing in the early developmental stages of a child. According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), there are two primary sets of symptoms for the diagnosis of Autism: 1) persistent deficits in social communication and interactions; 2) restricted and repetitive patterns of behaviour, interests or activities [1]. Autism is highly heritable, and as such gene analyses are being performed in an attempt to characterise the specific genes that are associated with Autism. This has been exemplified in a recent study by Grove et al (2019) [3], who identified a greater number of genetic risk variants in Autism than had been performed in previous studies. They were able to perform their study due to the increasing availability of genetic data. As the pool of genetic data continues to grow, the need to analyse that data effectively is of ever greater importance.

This report details the results of some basic data analysis techniques performed on a set of genes associated with Autism. It is split into three parts: part one explores the literature on genes and Autism, by identifying the most commonly studied genes associated with Autism. Part two takes a look at the ontologies of the genes, discovering which ontological terms are most frequently found in the gene set, and particularly those within the Biological Process ontology. Part three then analyses the interactivity of the genes, by discovering the protein-protein network of interactions within the gene set.

2 Data and Methods

The gene data set that has been used to undertake this study is taken from the Simons Foundation Autism Research Initiative (SFARI) [8], which comprises a list of genes associated with Autism. The data set used in this study is the 29th October 2020 release. Each gene in the data set has the following attributes: ‘Gene Symbol’, ‘Gene Name’, ‘Ensemble ID’, ‘Chromosome’, ‘Genetic Category’, ‘Gene Score’, ‘Syndromic’ and ‘Number of Reports’. Without detailing all of these, what is important to note, as it is used throughout this study for dividing the data up, is the *Gene Score*. The genes in this data set are split into three gene-score categories that indicate the confidence levels of a gene and its implication in Autism. Those with a score of 1 have the highest confidence, indicating they have clearly been shown in the literature to be implicated in Autism. Score 2 genes are those that could be strong candidates, whilst those with scores of 3 have only suggestive evidence of their links to Autism.

In part one, the number of genes within each of the three *Gene Score* categories are first measured. Genes that fall into the *Gene Score* category of 1 are then ranked based on the highest *Number of Reports* for each of those genes. The five genes with the highest number of reports are then used to query PubMed [7] for the number of papers published, per year, that include both the gene name and the term “Autism”. Such a search request can be made on PubMed by inputting “*Gene Symbol* AND Autism” into the PubMed search bar.

In part two, all the genes in the data set are mapped to an NCBI Unique Identifier, which can be found by searching the Gene database at NCBI for a specific gene term and filtering for “Human”. With each gene mapped to a UID, the Gene Ontology terms for each gene can be found using NCBI’s Gene2Go Correspondence File (download date was 17th November 2020). From this, the most commonly annotated terms across each of the three gene scores are summarised, with a further analysis of these genes performed using the PantherDB tool [6]. Here, the Biological Process ontology for each of the genes are searched for and compared to the annotated terms as gathered from the Gene2Go file.

Part three then uses STRING [9] in order to discover the protein-protein interaction network across the genes in *Gene Score* category 1. The genes that form the two largest MCL clusters in this network are used in the PantherDB, this time searching for Pathway ontology.

Most of the analyses performed in this study have been conducted using Jupyter Notebooks with Python 3.7. These notebooks have been included as supporting material in the assignment submission.

3 Results

3.1 Part One – Autism Literature

For each of the three *Gene Score* categories, the number of genes that fell into each are shown in Figure 1. This plot shows that the number of genes in gene score 3 is the highest, with 507 total genes. This is followed by gene score 2, with a total of 207 genes, which is closely followed by gene score 1 with a total of 194.

The five genes with the highest *Number of Reports* from *Gene Score* category 1 are displayed in Table 1. These five genes were each used as search queries in PubMed, along with the keyword "Autism". The total number of results for each of these five genes are given in Table 2, with a stacked bar chart summarising this table shown in Figure 2.

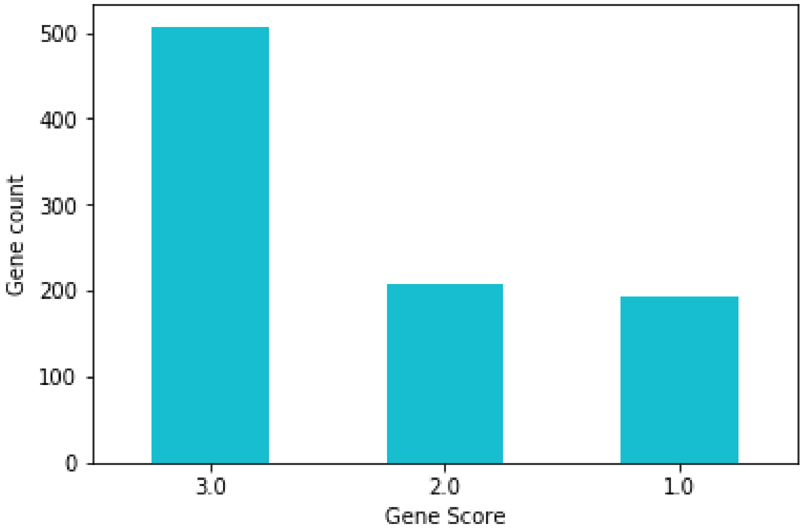

Figure 1: Bar chart of the number of genes in each of the three *Gene Score* categories.

Gene Symbol	Number of Reports
SHANK3	88
NRXN1	88
MECP2	87
SCN2A	71
SCN1A	67

Table 1: The five genes and their Gene Symbols from Gene Score category 1 with the highest Number of Reports.

	2020	**2019**	**2018**	**2017**	**2016**	**2015**	**2014**	**2013**	**2012**	**2011**	**2010**	**2009**	**2008**	**2007**	**2006**	**2005**	**2004**	**2003**	**2002**	**2001**	**2000**	**1999**	**1993**	**TOTAL**
SHANK3	46	56	57	44	47	32	35	32	20	21	9	9	9	4	1	2	0	0	0	0	0	0	1	425
NRXN1	10	21	12	15	7	17	18	20	17	21	10	14	7	4	0	0	0	0	0	0	0	0	0	193
MECP2	31	39	26	42	49	51	43	38	30	38	24	28	22	21	14	19	11	8	9	5	7	1	3	559
SCN2A	11	18	11	9	8	7	6	3	1	1	1	0	2	0	0	0	2	1	0	0	0	0	0	81
SCN1A	11	9	6	6	5	5	3	4	7	3	0	1	1	0	0	1	2	1	0	0	0	0	1	66

Table 2: The number of PubMed results per year for the five genes from Table 1.

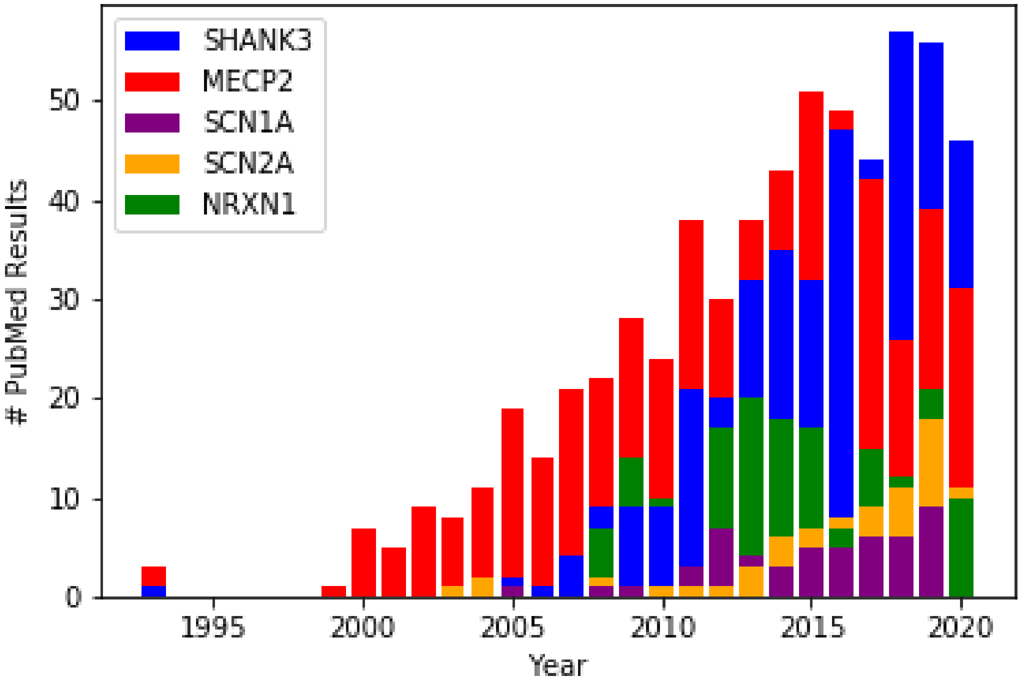

Figure 2: Stacked bar chart of the number of PubMed results as per Table 2.

3.2 Part Two – Autism Genes

All of the genes in the SFARI data set have been have been mapped to their NCBI Unique IDs (UID), and then their gene ontology terms have been retrieved for each gene using NCBI's Gene2Go Correspondance File. These steps are not shown here, but can be found in the supporting Jupyter notebooks. Tables 3-5 show the 10 most commonly annotated terms for each of the three Gene Score categories.

The Gene UIDs for each of the three Gene Scores were then input into the PantherDB Gene Analysis Tool. The gene count of Biological Processes for the three Gene Scores are shown in Figures 3-5.

GO ID	GO Description	Count
GO:0005515	Protein binding	150
GO:0005634	Nucleus	144
GO:0005654	Nucleoplasm	123
GO:0005886	Plasma membrane	91
GO:0005829	Cytosol	83
GO:0005737	Cytoplasm	68
GO:0045944	Positive regulation of transcription by RNA polymerase II	65
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	55
GO:0000122	Negative regulation of transcription by RNA polymerase II	49
GO:0046872	Metal ion binding	41

Table 3: The 10 most annotated Gene Ontology terms for genes in Gene Score 1.

GO ID	GO Description	Count
GO:0005515	Protein binding	136
GO:0005886	Plasma membrane	118
GO:0005634	Nucleus	100
GO:0005829	Cytosol	91
GO:0005737	Cytoplasm	79
GO:0005654	Nucleoplasm	75
GO:0005887	Integral component of plasma membrane	37
GO:0016020	Membrane	36
GO:0016021	Integral component of membrane	31
GO:0098978	Glutamatergic synapse	30

Table 4: The 10 most annotated Gene Ontology terms for genes in Gene Score 2.

GO ID	GO Description	Count
GO:0005515	Protein binding	326
GO:0005886	Plasma membrane	253
GO:0005829	Cytosol	191
GO:0005634	Nucleus	165
GO:0005737	Cytoplasm	151
GO:0005654	Nucleoplasm	136
GO:0005887	Integral component of plasma membrane	96
GO:0016021	Integral component of membrane	88
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	80
GO:0016020	Membrane	76

Table 5: The 10 most annotated Gene Ontology terms for genes in Gene Score 3.

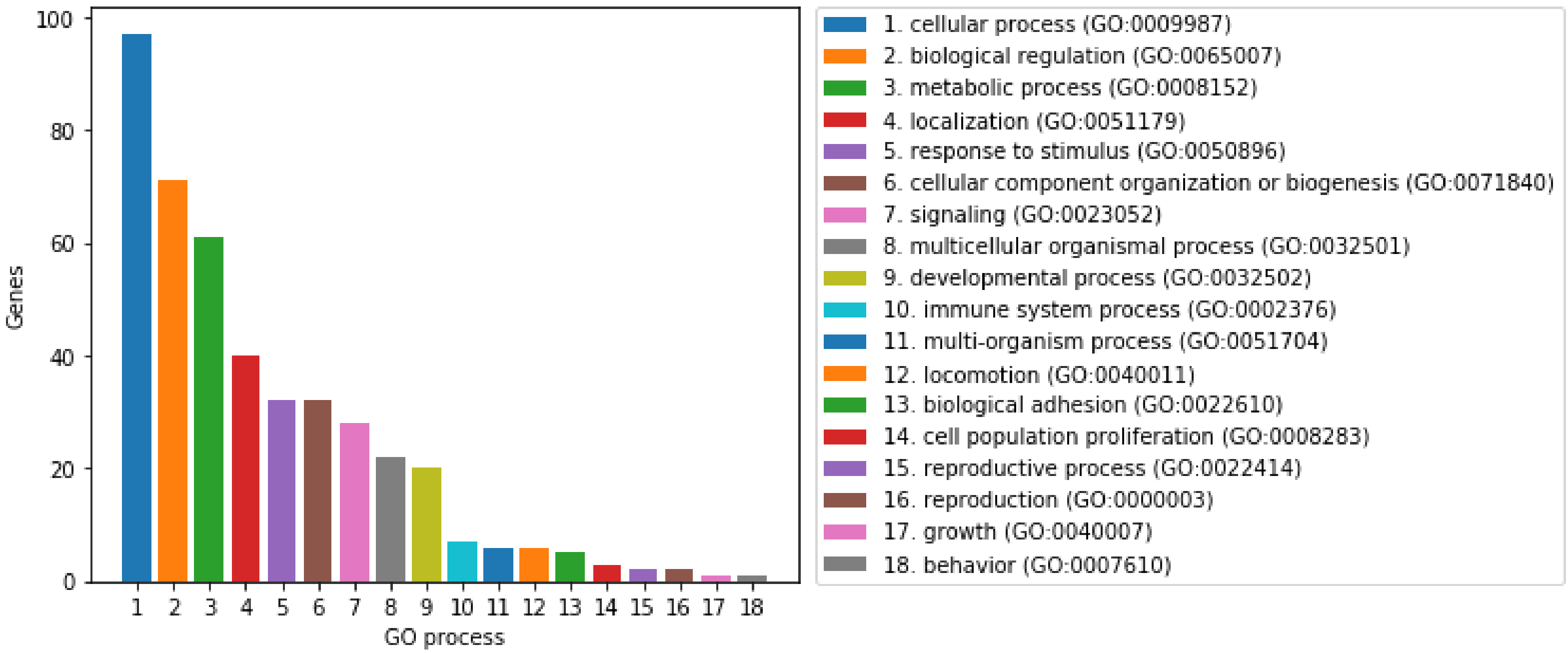

Figure 3: Bar chart of the Biological Processes as gathered from the PantherDB for genes with Gene Scores 1. Each bar gives the number of genes having that Biological Process.

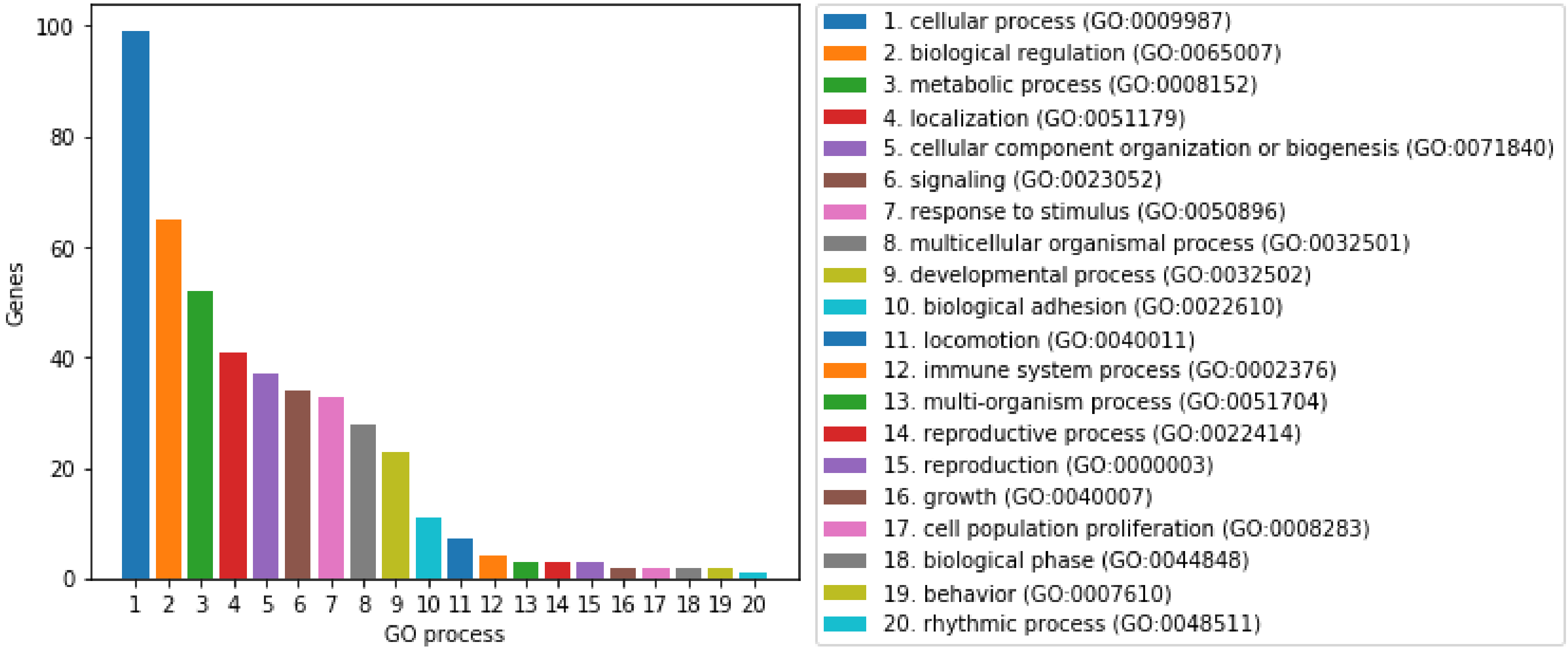

Figure 4: Bar chart of the Biological Processes as gathered from the PantherDB for genes with Gene Scores 2. Each bar gives the number of genes having that Biological Process.

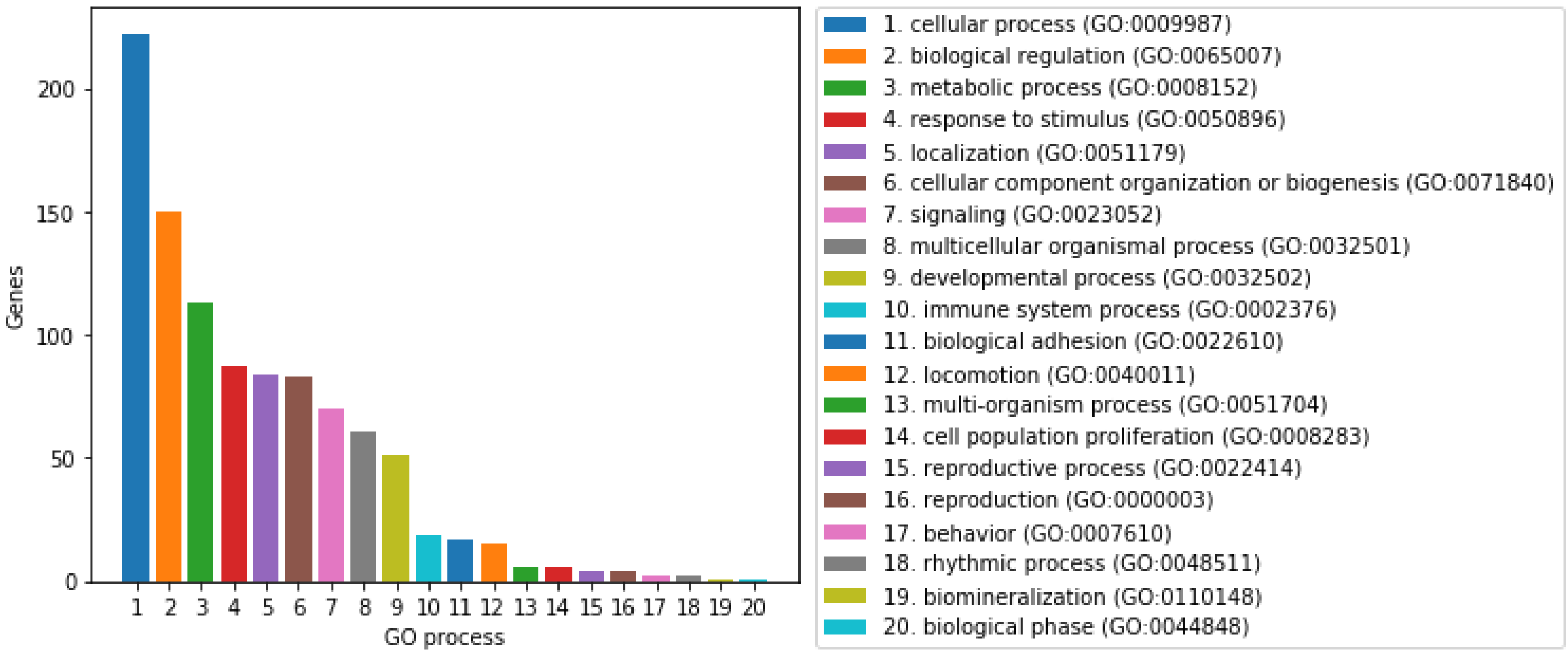

Figure 5: Bar chart of the Biological Processes as gathered from the PantherDB for genes with Gene Scores 3. Each bar gives the number of genes having that Biological Process.

3.3 Part Three – Autism Gene Networks

All the genes in Gene Score 1 were uploaded into STRING, giving a visualisation of the protein-protein network of interactions. The total number of nodes in this network came to 192, the number of edges at 955, and the average node degree was 9.95. A visualisation of the network is shown in Figure 8. However, the network in this image can be difficult to see clearly, due to the size of the network. The full sized version of the image is therefore included in the supporting documents, which makes viewing the network easier.

Results of using the PantherDB tool with the two largest MCL clusters from the STRING network are shown in Figures 6 and 7. In this instance, the Pathway ontology is displayed for the genes. Table 6 summarises the Pathway descriptions for the Pathways displayed in the two Figures.

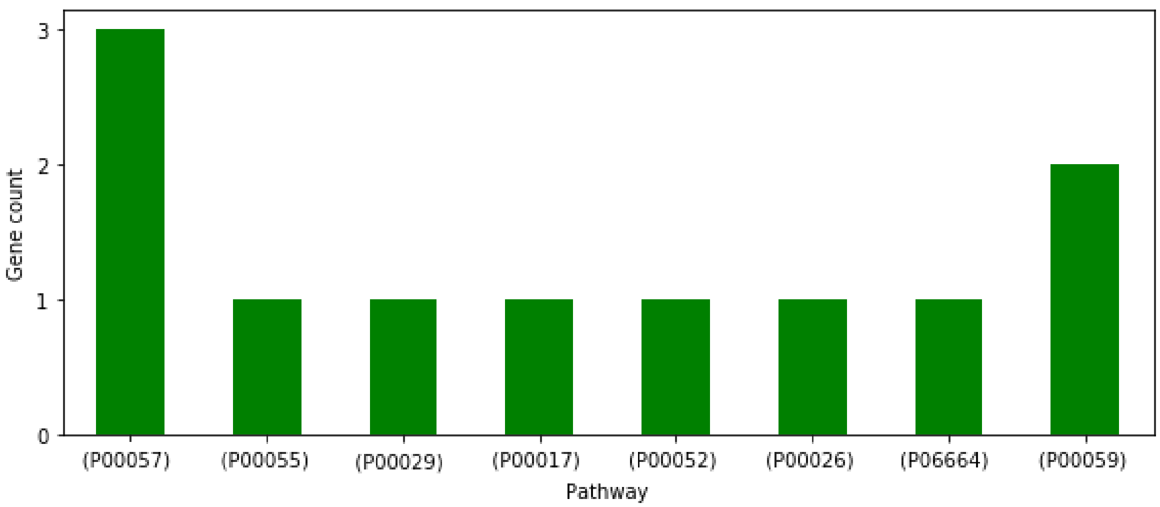

Figure 6: A gene count of the types of Pathways for the genes in the largest MCL cluster from the protein-protein network. The Pathway descriptions are given in Table 6

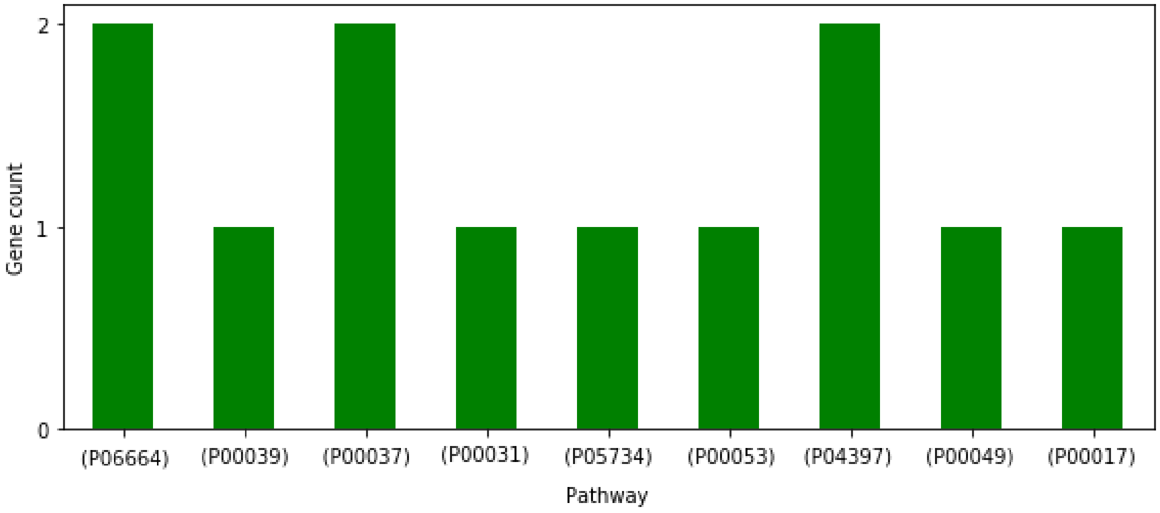

Figure 7: A gene count of the types of Pathways for the genes in the second largest MCL cluster from the protein-protein network. The Pathway descriptions are given in Table 6

Pathway ID	Description
P00017	DNA replication
P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway
P00029	Huntington disease
P00031	Inflammation mediated by chemokine and cytokine signaling pathway
P00037	Ionotropic glutamate receptor pathway
P00039	Metabotropic glutamate receptor group III pathway
P00049	Parkinson disease
P00052	TGF-beta signaling pathway
P00053	T cell activation
P00055	Transcription regulation by bZIP transcription factor
P00057	Wnt signaling pathway
P00059	p53 pathway
P04397	p53 pathway by glucose deprivation
P05734	Synaptic vesicle trafficking
P06664	Gonadotropin-releasing hormone receptor pathway

Table 6: Summary of the Pathway descriptions for the genes in Figures 6 and 7.

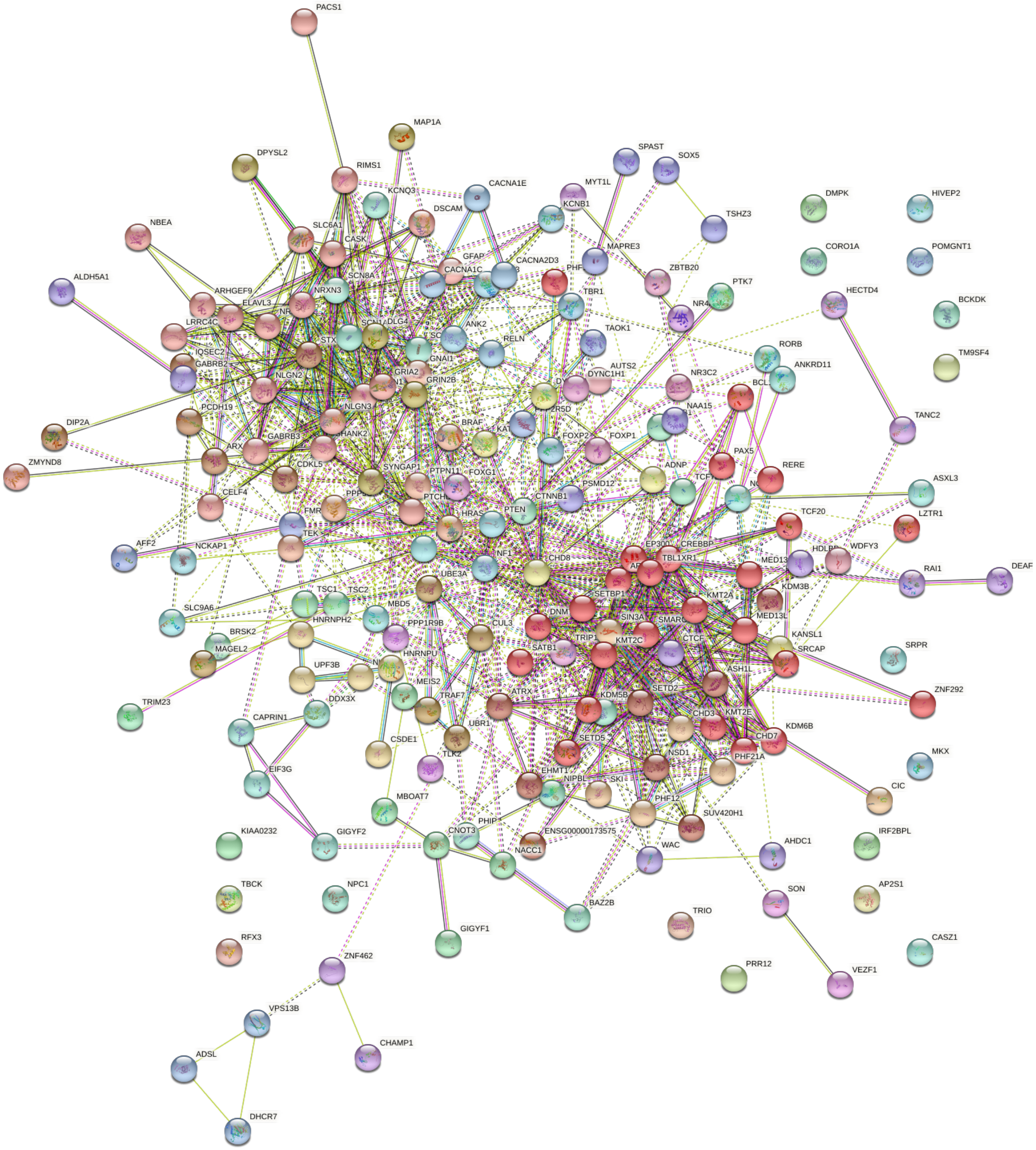

Figure 8: Image of the STRING protein-protein network interaction. Whilst the gene symbols are not clearly distinguishable in this image, it does provide a general overview of the degree of interactivity across the proteins. The full size image for the network can be found in the supporting material.

4 Discussion

Looking at the total number of results for the genes in Table 2, it appears that gene MECP2 has been the most studied of these genes with respect to Autism, and could therefore be a gene most commonly associated with Autism in some way. However, searching in PubMed for these genes without the Autism tag gives the following results:

1. SHANK3 - 537 results
2. NRXN1 - 358 results
3. MECP2 - 3,244 results
4. SCN2A - 622 results
5. SCN1A - 1,259 results

These results seem to indicate that despite MECP2 having many results with respect to Autism, there are generally many more results for this gene anyway. As a percentage, the number of results that includes MECP2 and Autism out of the total number of results for MECP2 is 15.3%. Compare this to SHANK3, whose percentage of papers related to Autism is 66.1%, and this would indicate that SHANK3 may be more relevant to Autism than MECP2, since most of the papers on the SHANK3 gene are related to Autism. Furthermore, Figure 2 shows that, whilst papers related to MECP2 in the most recent years has declined, those for SHANK3 have increased, adding further evidence to the possibility that SHANK3 is more closely associated with Autism. However, what is clearer is that the gene SCN1A, despite it having the second highest number of papers on its own (1,259, above), only has 4.7% of its total papers associated with Autism. So despite it being a well studied gene, it appears to be less related with Autism than the other genes might be.

Generally, the biological processes, as gathered from the PantherDB across the three gene score categories, are similar, with all three sharing the top three processes: 'cellular process (GO:0009987)', 'biological regulation (GO:0065007)', 'metabolic process (GO:0008152)' (Figures 3-5). What perhaps differs between those in Gene Score category 1 and those in category 2 is the process 'response to stimulus (GO:0050896)', in which this ranks as 5th in gene score category 1 as opposed to category 2 where it ranks 7th. This could indicate that genes with this biological process are more prominent in gene score category 1, which are those genes with stronger evidence for a link with autism, than those in category 2. But interestingly, this biological process is ranked 4th in gene score category 3. Autism has been linked to sensory-perceptual abnormalities, such as hyper- and hyposensitivity to stimulation [5]. It could be the case then that those genes with the 'response to stimulus' term have some involvement in the sensory abnormalities found in people with autism.

The top biological process across all three gene categories was 'cellular processes'. It is interesting to note that within the top most annotated GO terms (Tables 3-5) a large number of the terms were of cellular components, such as plasma membrane, nucleus, cytosol, cytoplasm, and nucleoplasm. It could be then that these offer a break down of the most common cellular components involved in the cellular processes. Of note is that the cellular component nucleoplasm retains a much larger proportion of the total GO terms in gene category 1 than in the other two. This component might therefore have some increased relevance amongst those genes that are well established in their association with autism.

The pathways in the two MCL gene clusters for genes in category 1 are mostly different across the two biggest clusters (Figures 6 and 7), but they do share two similar pathways: ‘DNA replication (P00017)’ and ’Gonadotropin-releasing hormone receptor pathway (P06664)’. One interesting difference between the two clusters is that the largest cluster has the ‘Huntington disease’ pathway (P00029), whilst the second largest cluster has the ‘Parkinson disease’ pathway (P00049)’. There is evidence to suggest that disordered movement might be a feature of autism, which is a feature that can be found in Huntington’s disease and Parkinson’s disease [4]. These results prove interesting therefore in the study of Autism, and a possible association, through disordered movement, to these two diseases.

References

[1] American Psychiatric Association et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[2] Mayada Elsabbagh, Gauri Divan, Yun-Joo Koh, Young Shin Kim, Shuaib Kauchali, Carlos Marcín, Cecilia Montiel-Nava, Vikram Patel, Cristiane S Paula, Chongying Wang, et al. Global prevalence of autism and other pervasive developmental disorders. *Autism research*, 5(3):160–179, 2012.

[3] Jakob Grove, Stephan Ripke, Thomas D Als, Manuel Mattheisen, Raymond K Walters, Hyejung Won, Jonatan Pallesen, Esben Agerbo, Ole A Andreassen, Richard Anney, et al. Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics*, 51(3):431–444, 2019.

[4] Ashwini Nayate, John L Bradshaw, and Nicole J Rinehart. Autism and asperger’s disorder: are they movement disorders involving the cerebellum and/or basal ganglia? *Brain research bulletin*, 67(4):327–334, 2005.

[5] Meena O’Neill and Robert SP Jones. Sensory-perceptual abnormalities in autism: a case for more research? *Journal of autism and developmental disorders*, 27(3):283–293, 1997.

[6] PantherDB. URL: `http://pantherdb.org/`.

[7] PubMed. URL: `https://pubmed.ncbi.nlm.nih.gov/`.

[8] SFARI. Sfari gene: Human gene. URL: `https://gene.sfari.org/database/human-gene/`.

[9] STRING. URL: `https://string-db.org`.