Autism spectrum disorder

BB226659

Introduction

Autism spectrum disorder (ASD) is a neurodevelopmental condition which even though is known to manifest itself differently and at varying degrees, is commonly characterized in people by their difficulty interacting with other people, forming friendships, and having uncommon responses to stimuli (Bonati Maurizio, 2022). There are a lot of environmental risk factors that have been described by literature at both prenatal and perinatal stages of development. Advanced parental age and maternal metabolic conditions are just two examples that have been shown to be significant ASD risk factors (Catherine Lord, 2020). In the late 20th century, more scientists started to focus on genetic autism research and that led to an influx of new discoveries (Anita Thapar, 2021). Some relatively recent papers report that 75-93% of ASDs are the result of heritable, genetic factors.

In this research project, we will be looking at the genes that have been linked to autism and curated into a list by the Simons Foundation Autism Research Initiative (SFARI). We will analyse how ASD research has evolved in the past 30 years and which genes have had the most spotlight. We will then analyse the genes that are highly linked to autism and attempt to learn a little bit more about them and try to identify various clusters of said genes and see how their function might be of interest to us.

Data and Methods

All of the analyses were carried out on the SFARI gene list which was downloaded from the following website:
https://gene.sfari.org/database/human-gene/ (Release: 11-07-2022, Downloaded: 08-11-2022).

Python coding language (ver. 3.8.10) was used for most of the data manipulation and output creation (Van Rossum, 2009). This paper will include sections of the code, which were used during the research. This should improve the reproducibility of the results, although the same results can be achieved in a variety of approaches.

Part 1 – Autism Literature

The SFARI dataset had to be divided into sub-tables, based on the gene-score category. And the number of genes in each category is counted with the following lines of python code.

```
df = pd.read_csv('SFARI.csv')
gene_count = df.groupby(['gene-score']).size().reset_index(name='number-of-occurrences')
```

To then get the top 5 SFARI genes of the gene-score 1 category, we ranked the table based on the contents of the number-of-reports column.

```
df_sorted = df.sort_values("number-of-reports", ascending=False)
df_1 = df_sorted.loc[df_sorted['gene-score'] == 1]
```

The symbols of the top 5 genes were then used to create a query term which was then sent to the PubMed (Eric W. Sayers, 2022) database through Biopython's Entrez module and the number of hits recorded (Count). The query which was sent was *"GeneSymbol AND Autism"*

To get a table with the number of papers publishes in a specific year (year range used 2022-1990) we run the following code:

```
for name in task3_pnames: #task3_pnames is a list of the 5 gene symbols
    new_row=[name]

    for year in reversed(range(1990, 2023)):
        query_year=name+" AND Autism AND ("+str(year)+":"+str(year)+"[pdat])"
        task4_search = Entrez.esearch(db="pubmed", term=query_year)
        task4_read = Entrez.read(task4_search)

        new_row.append(task4_read['Count'])


    #print(new_row)

    dates.loc[len(dates.index)] = new_row

print(dates)
```

The term created at this stage was: *"GeneSymbol AND Autism AND (year1:year2[pdat])"*

The same steps were carried out for gene-score 2 and 3 to add an additional piece of analysis.

Part 2 – Autism Genes

To map the gene-symbols of the SFARI gene list to an NCBI UID we created a new query term: "GeneSymbol[sym] AND Homo Sapiens" and sent it through Biopython Entrez to the 'gene' database and only collected the first hit.

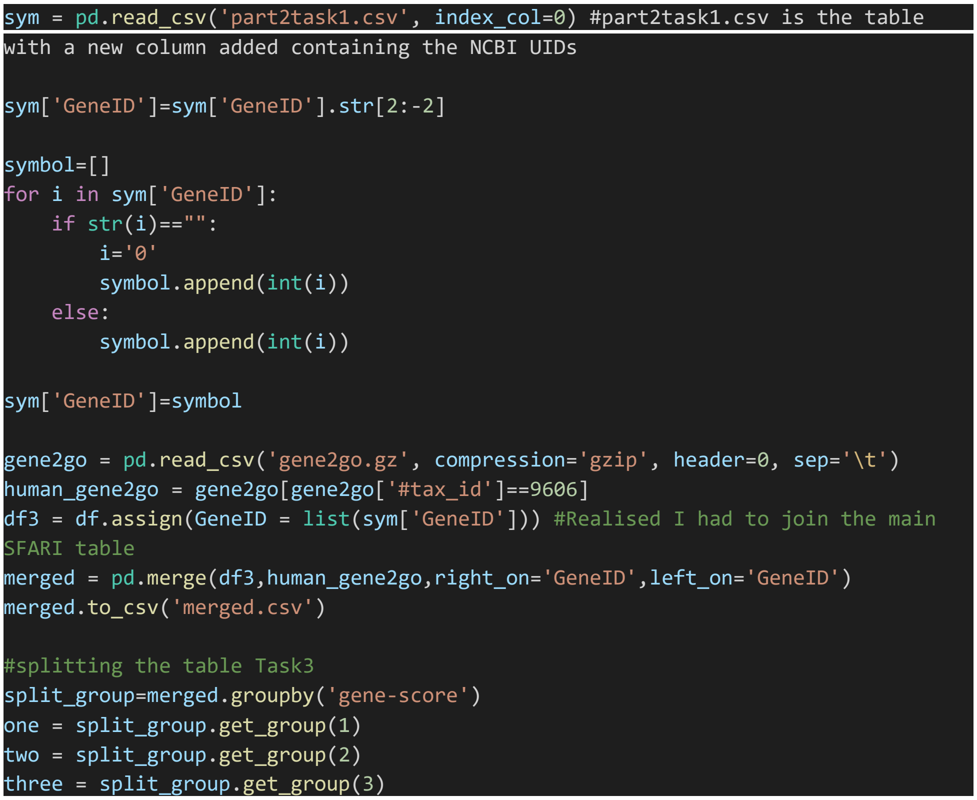
```
df = pd.read_csv('SFARI.csv')

task1_pnames=list(df["gene-symbol"]) #List of all the gene symbols from SFARI
from Bio import Entrez
Entrez.email = ''


df2 = pd.DataFrame(columns=['symbol', 'GeneID'])

for s in task1_pnames:
    task1_search = Entrez.esearch(db="gene", term=s+"[sym] AND Homo Sapiens",
retmax=1)
    task1_read = Entrez.read(task1_search)
    #print(s)
    #print(task1_read['IdList'])
    task1[s]=task1_read['IdList']
    df2.loc[len(df2.index)] = [s, task1_read['IdList']]
    df2.to_csv('part2task1.csv')
print(df2)
```

To assign the gene ontology terms to the SFARI genes, we used the newly generated NCBI UIDs and matched them with the IDs from the gene2go file (provided by the School of Informatics from the University of Edinburgh). This created a merged file, which was subsequently split into 3 tables based on the gene-score categories and the top 10, most commonly annotated oncology terms were collected.



```
sym = pd.read_csv('part2task1.csv', index_col=0) #part2task1.csv is the table
with a new column added containing the NCBI UIDs

sym['GeneID']=sym['GeneID'].str[2:-2]

symbol=[]
for i in sym['GeneID']:
    if str(i)=="":
        i='0'
        symbol.append(int(i))
    else:
        symbol.append(int(i))

sym['GeneID']=symbol

gene2go = pd.read_csv('gene2go.gz', compression='gzip', header=0, sep='\t')
human_gene2go = gene2go[gene2go['#tax_id']==9606]
df3 = df.assign(GeneID = list(sym['GeneID'])) #Realised I had to join the main
SFARI table
merged = pd.merge(df3,human_gene2go,right_on='GeneID',left_on='GeneID')
merged.to_csv('merged.csv')

#splitting the table Task3
split_group=merged.groupby('gene-score')
one = split_group.get_group(1)
two = split_group.get_group(2)
three = split_group.get_group(3)
```

The three lists of UIDs were then exported to text and imported into the pantherdb tool (Paul D. Thomas, 2003) on the website: ‘http://pantherdb.org/’ (Release: PANTHER 17.0). Functional classification viewed in graphic charts analysis was carried out on ‘Biological Databases’ ontology. Consequently, bar chart information was exported for all the gene-score categories and later used to create the bar charts which can be seen in the results.

As an additional piece of analysis, the same UID lists were imported into Reactome Pathway Database analysis tool (Version: Reactome v82) and the option “projected to human” was chosen. Pathway analysis results were downloaded into a .tsv table and the 5 most statistically significant pathways were chosen to display. A genome-wide overview image was also downloaded which highlights the Reactome pathways found in the genes submitted.

Part Three - Autism Networks

To create a protein-protein interaction matrix, we first had to export a plain text file of the NCBI UIDs for the SFARI genes in gene-score 1. Then using the STRING website (https://string-db.org/) (STRING Version 11.5) upload the file and select ‘Homo Sapiens’ as the organism of interest. The key statistics were then recorded from the ‘Analysis’ tab and the whole matrix was saved as a ‘bitmap image’ file. MCL clustering technique was chosen, and the output was saved in a TSV format. From the output table we picked the two largest clusters and used the protein UIDs to search their ontology using the Panther database again, but with a changed ontology to "Pathway".

Results

Part One – Autism Literature

To analyse the literature related to ASD, we used the downloaded data from the SFARI database and divided it into three groups based on how likely the genes are associated with autism. The first thing we wanted to assess was the number of genes in each gene-score category (Fig. 1). This gave us an overview of how many false positives were expected (mostly genes with score 3), that number was relatively small (91), which mean that only a small portion of the database had only suggestive evidence. It is also worth noting that there are some genes in the database that do not have a score assigned, they are in a syndromic category, which includes mutations that are associated with an increased risk, but not specifically required for an ASD diagnosis.

As we can see from figure 1, most of the genes have a score of 2, which means they are strong candidates that have reported de novo likely-gene-disrupting mutations, but the number of high-confidence genes is relatively low (214).

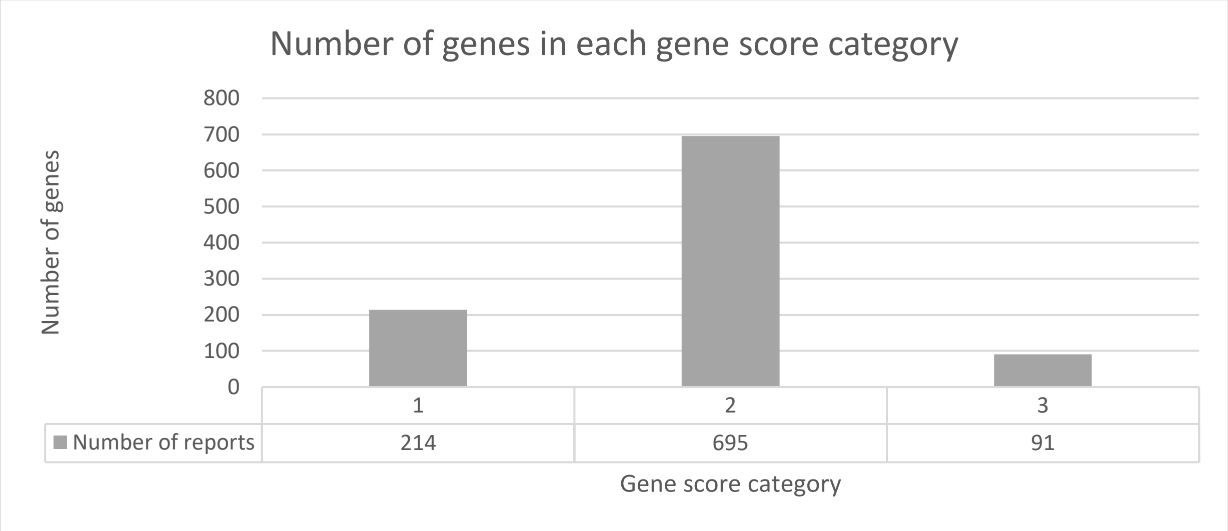


Figure 1 (Task 1). Graph representing the number of genes in each gene score category. The gene score categories represent the confidence we have in that gene being associated with ASD.

- gene score 1 – highest confidence
- gene score 3 – lowest confidence

Since the genes with a score of one, are known to be clearly implicated in autism, we decided to focus on those genes (with a score of 1) and see which ones were most frequently reported (Kyle Satterstrom, 2020).

Top Referenced Genes	Ensembl ID	Gene Score	Number of reports	Papers in PubMed
SHANK3	ENSG00000251322	1	**120**	**500**
MECP2	ENSG00000169057	1	**107**	**532**
NRXN1	ENSG00000179915	1	**100**	**184**
SCN2A	ENSG00000136531	1	**96**	**105**
SCN1A	ENSG00000144285	1	**84**	**80**

Table 1 (Task 2 and 3). This table represents the most reported genes from a SFARI gene score 1 category. The table includes the symbols of the genes, their ensembl ID, gene score, the number of reports and the number of papers related to autism that reference the gene in the PubMed

As we can see in table 1, the highest reported gene was SHANK3 with 120 reports, followed by MECP2 and NRXN1 with all of them scoring more than 100 reports each. We thought it would be interesting to see how the number of reports compares with genes from other score categories. Therefore, we repeated the ranking process for the genes from scores 2 and 3 and displayed the results in table 2.

To no surprise the number of reports and the number of references in the literature for the top reported genes decreased in each subsequent category, suggesting that the higher number of reports correlates with the confidence that our gene is implicated in autism risk. However, it is important to point out that a high number of reports is not required for a gene to qualify in score 1 and have high confidence. When we looked at the lowest numbers of reports for the score 1 category, there were a lot of genes with just one report to their name.

Top Referenced Genes	Ensembl ID	Gene Score	Number of reports	Papers in PubMed
CNTNAP2	ENSG00000174469	2	**82**	**209**
OXTR	ENSG00000180914	2	**49**	**175**
KCNQ2	ENSG00000075043	2	**42**	**24**
RBFOX1	ENSG00000078328	2	**41**	**43**
SATB2	ENSG00000119042	2	**40**	**20**
YWHAZ	ENSG00000164924	3	**14**	**5**
CDK5RAP2	ENSG00000136861	3	**13**	**1**
GRB10	ENSG00000106070	3	**11**	**5**
KLF7	ENSG00000118263	3	**11**	**6**
MSRA	ENSG00000175806	3	**11**	**1**

Table 2. This table represents the most reported genes from a SFARI gene score 2 and 3 categories for comparison with table 1. The table includes the symbols of the genes, their ensembl ID, gene score, the number of reports and the number of papers related to autism that reference the gene in the PubMed database

After we have identified the top reported genes, we decided to see the number of times these top genes were referenced in the scientific literature related to autism (Tables 1 and 2). We then split this data and ended up with the number of publications per year for each year since 1990. This data is displayed in table 3 and to help visualise the data, a histogram was included in Figure 2. It shows the exponential increase in research in the last decade for the top genes associated with autism.

	2022	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010	2009	2008	2007	2006
SHANK3	76	67	57	56	57	44	48	32	35	32	21	21	9	9	9	4	1
MECP2	30	22	32	39	26	41	48	51	43	37	30	38	23	26	20	20	13
NRXN1	12	13	12	22	13	15	8	17	18	20	17	21	10	14	7	4	0
SCN2A	16	22	12	18	11	9	8	7	6	3	1	1	1	0	2	0	0
SCN1A	9	12	13	9	6	6	5	5	3	4	7	3	0	1	1	0	0

	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990	1990
SHANK3	2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
MECP2	17	11	8	9	4	7	1	0	0	0	0	0	3	0	0	0	0
NRXN1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SCN2A	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SCN1A	1	2	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0

Table 3 (Task 4). This table represents the 5 most reported genes for gene score 1 category of SFARI database and the number of publications that were autism related each year for that gene. The table covers the timeline up to the beginning of the 90s as there were no significant publications before that time concerning these 5 genes.

We had to consider the possibility that not all of the papers identified were actually referencing Autism, but carefully built query terms suggest that the papers should have at least mentioned the two terms.

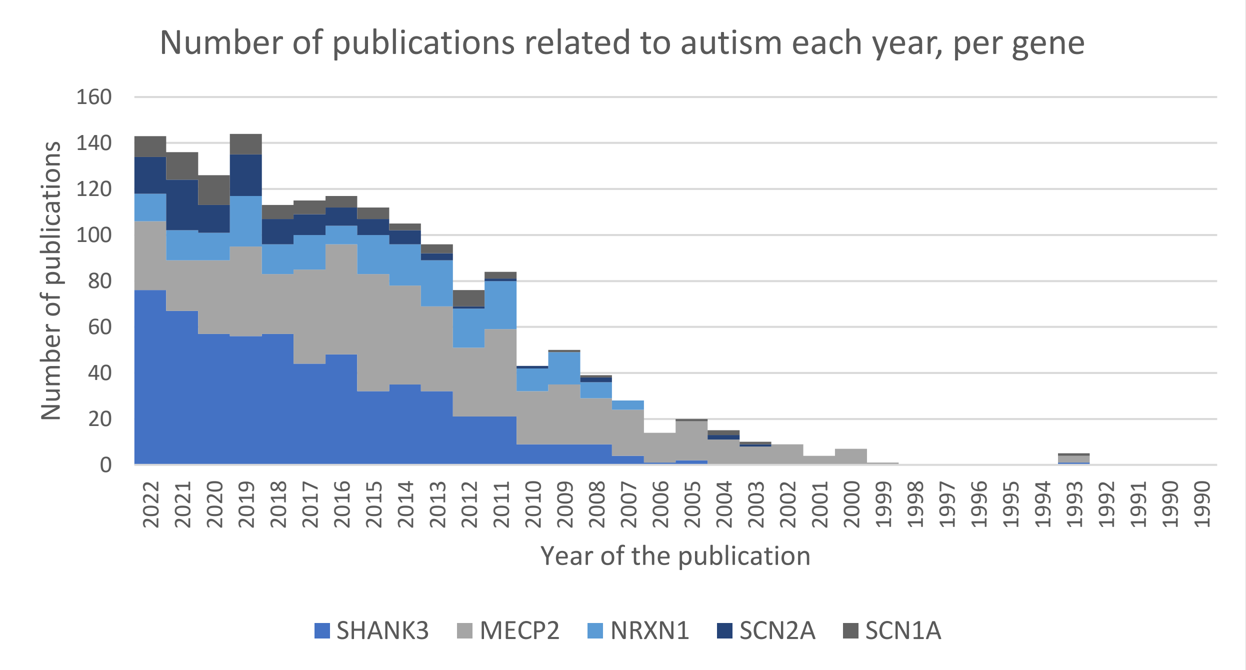


Figure 2 (Task 5). Stacked histogram that visualises the number of publications related to autism, that reference the gene name, per year. The 5 genes represented (SHANK3, MECP2, NRX1, SCN2A and SCN1A) are the top 5 reported genes in ASD from SFARI database, gene score 1.

Part Two – Autism Genes

In this part, we wanted to look at some of the functional terms that have been annotated to SFARI genes and compare the annotations between different gene-scores. After applying the gene2go ontology table, we identified the top ten gene ontology terms for all the gene-scores, the data is presented in table 4.

10 Most commonly annotated terms for genes with a gene-score 1

GO term ID	GO term Description	GO term count
GO:0005886	plasma membrane	498
GO:0005515	protein binding	471
GO:0005829	cytosol	306
GO:0005634	nucleus	303
GO:0005737	cytoplasm	243
GO:0005654	nucleoplasm	243
GO:0016020	membrane	169
GO:0070062	extracellular exosome	90
GO:0046872	metal ion binding	88
GO:0003723	RNA binding	87

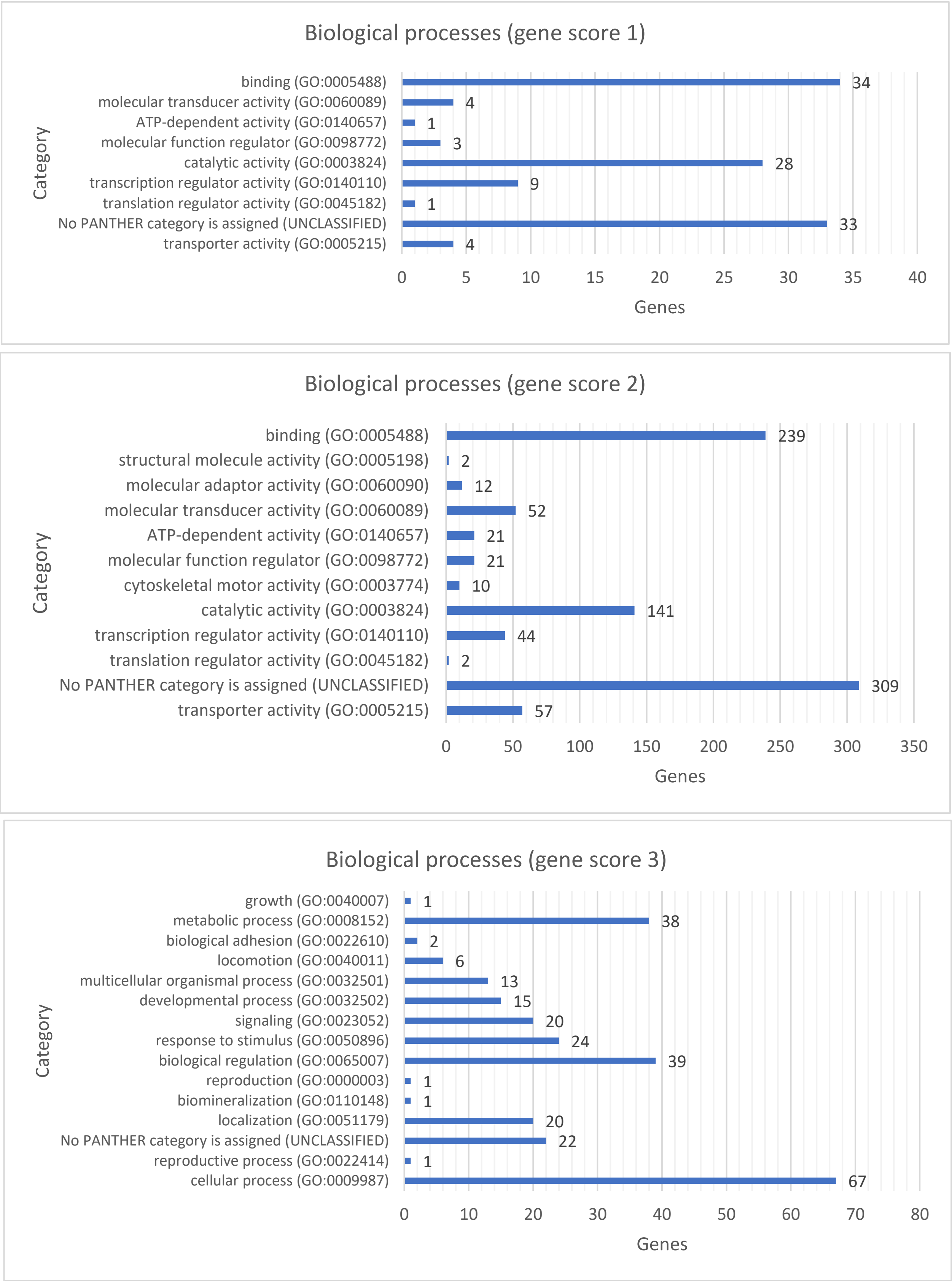
10 Most commonly annotated terms for genes with a gene-score 2

GO term ID	GO term Description	GO term count
GO:0005634	nucleus	176
GO:0005515	protein binding	170
GO:0005654	nucleoplasm	141
GO:0005886	plasma membrane	116
GO:0005829	cytosol	105
GO:0005737	cytoplasm	84
GO:0045944	positive regulation of transcription by RNA polymerase II	73
GO:0016020	membrane	55
GO:0006357	regulation of transcription by RNA polymerase II	55
GO:0000122	negative regulation of transcription by RNA polymerase II	53

10 Most commonly annotated terms for genes with a gene-score 3

GO term ID	GO term Description	GO term count
GO:0005515	protein binding	73
GO:0005829	cytosol	57
GO:0005886	plasma membrane	50
GO:0005634	nucleus	46
GO:0005737	cytoplasm	45
GO:0005654	nucleoplasm	32
GO:0016020	membrane	26
GO:0003723	RNA binding	17
GO:0005576	extracellular region	17
GO:0046872	extracellular exosome	16

Table 4 (Task 1, 2, 3 and 4). This table represents ten most commonly annotated gene ontology terms for all the 3 gene-score genes of the SFARI database. These ontology terms were derived by comparing the autism-related genes with the gene2go database.



Figures 3, 4, 5 (Task 5). These bar charts show the biological process gene ontology terms describing the function of the gene products for three different gene-score genes of the SFARI database. The terms were identified using panther-db database.

As an additional piece of analysis, the same UID lists were imported into the Reactome Pathway Database analysis tool. Figure 6 represents the genome-wide overview of the gene-score 1 SFARI results pathway analysis. It highlights the functions that are associated with our dataset. The centre of each of the 'bursts' is the top-level pathway and each step away from the middle represents a lower level of the hierarchy pathway.

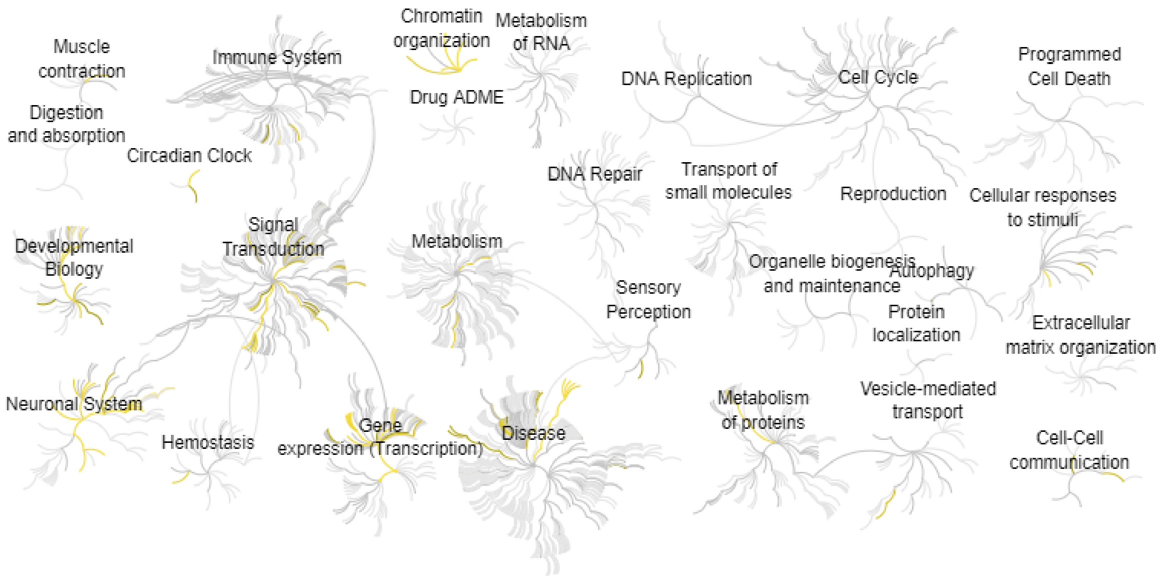


Figure 6. This figure represents a genome-wide, hierarchical visualisation of Reactome pathways. It summarises the pathway analysis results.

In the bottom left corner, we can see a part of the Neuronal System pathway being highlighted. It is the Neurexins and neuroligins pathway, which was the most significant one for gene-score 1 dataset. There were 17 out of 60 entities found and this resulted in an extremely small p-value. The Neuroligins and Neurexins are involved in synaptic signalling (Südhof, 2008), which supports the findings from table 3 which had transducer activity function represented in its oncology

Part Three – Autism Networks

In this stage, we wanted to look for evidence and examples of how these proteins work together. The first step involved checking if the proteins physically interact in a coordinated way. This was done by visualising the protein-protein interaction network with help of software from a STRING website. We noted down statistics like 'Terms' - Nodes (genes), edges (their relations/connections) and the average node degree. These values are presented in table 5.

number of nodes	212
number of edges	1523
average node degree	14.4

Table 5 (Task 1). This table represents key statistics from a derived protein-protein interaction network of the gene-score 1 genes of SFARI database

From that MCL clustered interaction matrix, we chose the two biggest clusters (sizes of 34 and 27) and tried analysing the functions of the genes within them using the ‘Pathway’ ontology (Figures 6 and 7)

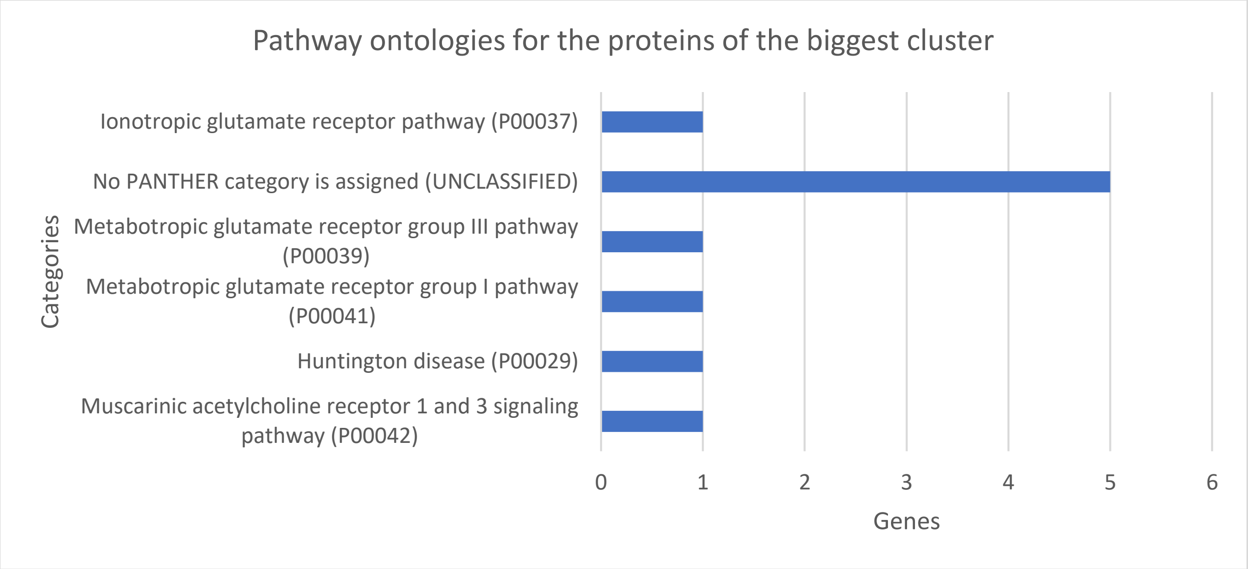


Figure 7 (Task 2). The figure represents the pathway ontologies of the proteins from the biggest MCL cluster of the protein-protein interaction matrix

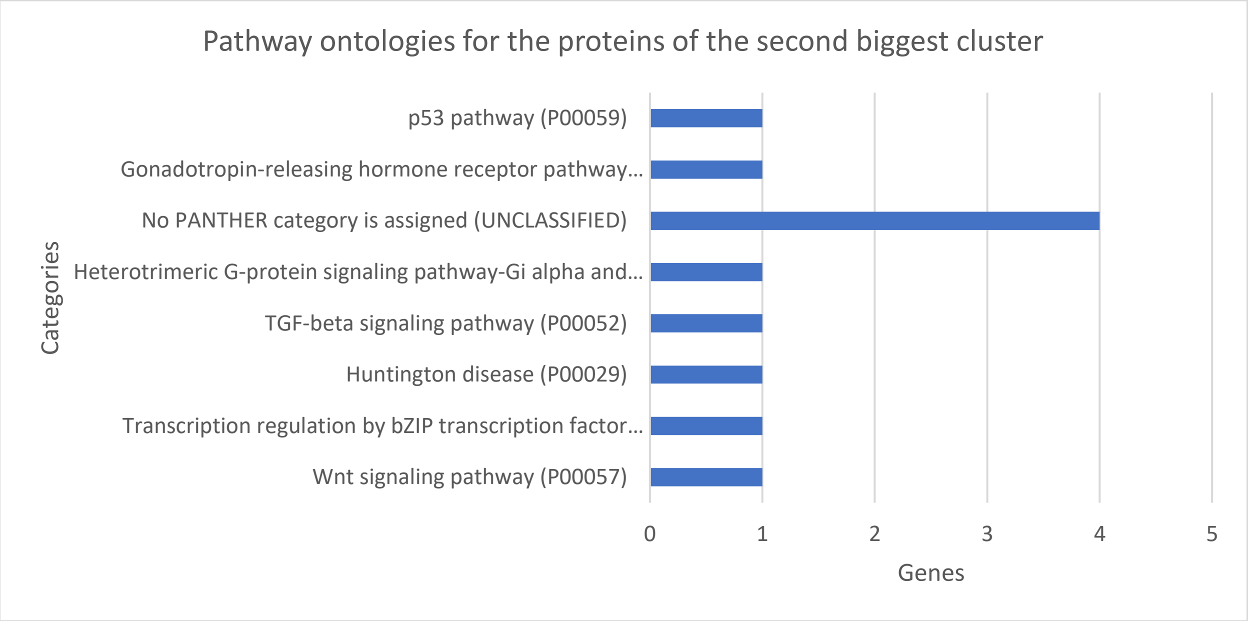


Figure 8 (Task 2). The figure represents the pathway ontologies of the proteins from the second largest MCL cluster of the protein-protein interaction matrix

When we compare these two tables, we can see that the function of most of the genes from the largest cluster (Fig. 7) is related to the glutamate receptor pathway (Koochekpour, 2013) (signalling pathways), while the second largest cluster (Fig. 8) has proteins involved in the cellular homeostatic pathway

(Sandra L Harris, 2005). This shows that the proteins from each cluster are interacting and therefore are related, and this can help ensure that the predicted functions are correct. As we can also see the functions of the proteins from these two clusters are different, which explains why the proteins do not interact as strongly with the members of the other cluster

When we compare these figures to the results from part 2, gene-score 1, we can see that the functions overlap as well. The oncology terms like transcription and translation regulation are all referring to the homeostatic pathways, while molecular transducer/transporter activity is all associated with the signalling pathways. This supports the results and strongly supports the results of the part 2 function predictions.

We also decided to perform the same analysis for gene-scores 2 and 3 (appendix) and see if the clustering of proteins and predicting their function will also be supporting the previous results. The protein-protein interaction matrices had small clusters with

As the final results figure we decided to visualise the protein-protein interaction matrix to highlight the vast number of protein interactions for gene-score 1 members of the SFARI dataset (Figure 9).

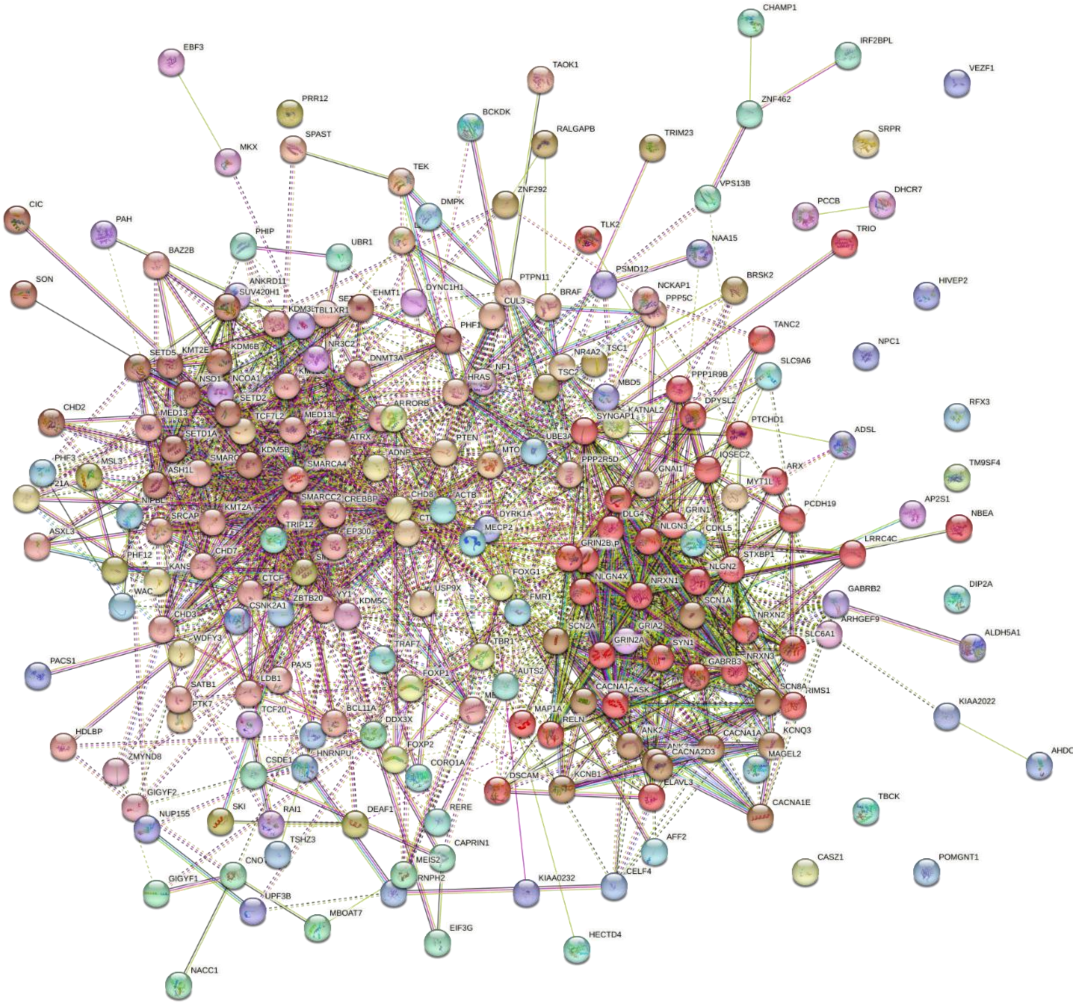


Figure 9 (Task 3). The figure represents the protein-protein interaction matrix for gene-score 1 members of the SFARI database.

Discussion

References

Anita Thapar, M. R. (2021). Genetic Advances in Autism. *Journal of Autism and Developmental Disorders* , 4321–4332.

Bonati Maurizio, M. C. (2022). Still too much delay in recognition of autism spectrum disorder. *Epidemiol Psychiatr Sci*.

Catherine Lord, T. S.-V. (2020). Autism spectrum disorder. *Nat Rev Dis Primers*.

Eric W. Sayers, E. E.-B. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Res*. doi:10.1093/nar/gkab1112

Huaiyu Mi, A. M. (2019). Large-scale gene function analysis with PANTHER Classification System. *Nat Protoc*, 1551-1566.

Koochekpour, S. S. (2013). Glutamate, Glutamate Receptors, and Downstream Signaling Pathways. *Int J Biol Sci*, 948-959.

Kyle Satterstrom, K. J. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, 568-584.

Paul D. Thomas, M. J. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res., 13:* , 2129-2141 .

Sandra L Harris, A. J. (2005). The p53 pathway: positive and negative feedback loops. *Nature. Oncogene*, 2899–2908.

Südhof, T. C. (2008). Disease, Neuroligins and Neurexins Link Synaptic Function to Cognitive. *Nature*, 903-911.

Van Rossum, G. a. (2009). Python 3 Reference Manual. *CreateSpace*.

Appendix

number of nodes	686
number of edges	3375
average node degree	9.84

Table A1. This table represents key statistics from a derived protein-protein interaction network of the **gene-score 2** genes of SFARI database

number of nodes	91
number of edges	61
average node degree	1.34

Table A2. This table represents key statistics from a derived protein-protein interaction network of the **gene-score 3** genes of SFARI database

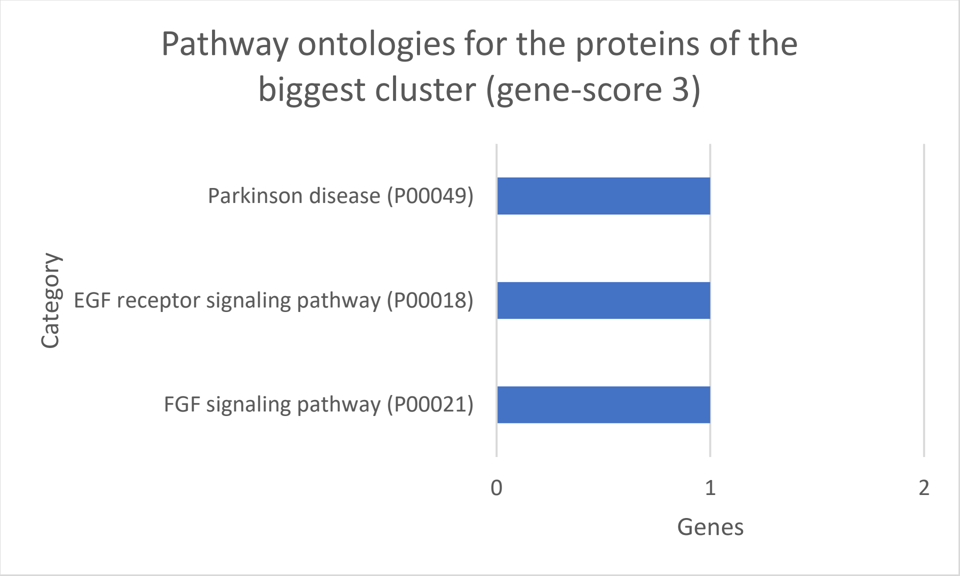


Figure A1. The figure represents the pathway ontologies of the proteins from the biggest MCL cluster of the protein-protein interaction matrix of the gene-score 3 dataset.

Included the link to the google drive containing my code used in this assignment: https://drive.google.com/drive/folders/1UznEksgtqFYgUS8SMMHgOvwY35mmohsx?usp=sharing