Investigation into Autism Spectrum Disorder
(B164439) Bioinformatics, University of Edinburgh

Introduction:

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder which causes adverse effects on social communication and presence of restricted interests[1][2]. ASD is a complicated disorder where statistically it affects more than 1 in 100 people in the UK[2]; there are around 700,000 autistic adults in the UK[2]. This is an interesting prevalent field of research where researchers now are attempting to understand the exact genetic mechanism behind ASD due to the increase in accessibility in next generation sequencing (NGS)[3] and popularity of advancing scientific technologies such as high throughput genotyping[3][4]. Through these advanced research methodologies, scientists have shown that mutations occurring in specific genes on particular chromosomes of humans[5], would lead to ASD. Using a bioinformatics and genetic approach to overcome and understand the complexity of this disease, the workflow includes literature mining search to research on gene-ASD associations through Simons Foundation Autism Research Initiative (SFARI) database[6], autism gene analysis by utilizing gene ontology (GO)[7] for attempting to functionally characterize these genes and autism network analysis to decipher which genes interact with each other in certain pathways in autism. Development in this area of research could potentially uncover more accurate molecular mechanisms behind why these mutations occur and how this leads to ASD.

Data and Methods:

Autism Literature:

Investigation into literature to discover gene-ASD relationships was conducted through using the SFARI database. Number of genes in gene score category 1, 2 and 3 were plotted to imply which genes were implicated in ASD. Gene score from SFARI database was heavily curated to present which genes were most confident in their contribution to secondary ASD. To refine this search, gene score Syndromic (S) was not taken into consideration in this study as it covers idiopathic autism-autism of unknown origin[8]. The top 5 genes in gene score 1 based on number of reports published were extracted; these genes were consecutively investigated on NCBI PUBMED to find the number of reports published related to autism and the respective genes. Search field tags (MeSH) was employed to refine the literature search. Publications over a span of 12 years in relation to these genes were plotted. Further investigation into citation data such as Eagle Score[8][9] was utilized for ensuring confidence of the relevance of the gene to ASD.

Autism Genes:

In order to functionally classify the genes associated with ASD, the functional terms annotated to these SFARI genes were identified. Gseapy package's BioMart tool[10][11] was utilized to retrieve NCBI Entrez Gene IDs (UIDs) and merge these IDs with the SFARI gene symbols. Duplicates were removed in this list and were further analyzed. Then, by retrieving the gene2go file from NCBI (through urllib package), gene ontology terms annotated with SFARI genes could be extracted. GO IDs which did not present evidence (NAS and ND terms) were remove to improve reliability of our results. The merged results which contain the UIDs from gene2go NCBI file and SFARI gene list were grouped by their gene score from SFARI. Top 10 commonly annotated GO terms were recorded for each gene score list in descending order. Subsequently, the 3 different text files were exported from code to carry out inferences on the functional classification of these genes through PantherDB. Protein Analysis Through Evolutionary Relationships (PantherDB)[12] provides functional classification of genes through the Gene Ontology (GO) project. These 3 different text files (GeneScore1, GeneScore2, GeneScore3) were uploaded to PantherDB to present functional classification viewed in graphic charts. The ontology domain "Biological Processes" was used. To further investigate the GO annotations to functionally classify the genes, specifically for the top 5 genes in their relation to ASD, the HumanCyc[13] pathway database was investigated to compare the

GO terms associated with these 5 genes. This was conducted to obtain further insights into molecular pathways and gene ontology (GO) that overlap. We considered 4 pathway databases[14][15]: KEGG[15], PantherDB[12], HumanCyc pathway[13] and GO Domain: Biological Processes (BP)[7][14][15].

Autism Network:

As pathways are central to human's responses to stimuli, pathway-based analysis was approached to understand how complex diseases such as ASD may be related to each other through their underlying molecular mechanisms. Network connectivity analysis (NCA) was conducted to test for direct functional association between top identified genes in gene score 1. This was implemented to assess the both the functional profile and significance of these genes through STRING and PantherDB software. STRING (Search Tool for Retrieval of Interacting Genes/ Proteins) database[16] was implemented as the classification systems are based on high-throughput text-mining as well as hierarchical clustering on the association network itself. Network statistics were analyzed in terms of number of nodes, number of edges and number of degrees. Through MCL clustering, the top two clusters were extracted. PantherDB Functional Classification through Bar Charts was implemented with Pathway ontology to annotating genes to pathways. From Markov (MCL) clustering[17], the top cluster in the network would be used to create a new, separate network in Cytoscape[15]. In order to functionally characterize the cluster, stringApp[15] was implemented to perform functional enrichment analysis with a false discovery rate (FDR) threshold of 5%. Then, the filter functionality was used to eliminate redundant terms (using default redundancy cutoff of 0.5).

Results:

Autism Literature:

From the 1095 genes found from SFARI database, the number of genes in gene score 1, 2, and 3 are 214, 695 and 91 genes respectively. The top 5 genes found in gene score 1 category are SHANK3, MECP2, NRXN1, SCN2A and SCN1A with SHANK3 having the highest number of reports at 120 in the curated SFARI database, as shown in Figure 2 below. When investigated with PUBMED, 495 reports were published which is lower than the number of reports published for the 2nd most reported gene, MECP2 on SFARI. MECP2 has 524 papers published on PUBMED. This is presented by Table 2, where the number of PUBMED papers for the top 5 genes from SFARI are shown. From supplementary table I (Appendix) regarding a breakdown of published PUBMED papers on the 5 genes and the bar chart indicating the number of papers published, SHANK3 has shown an increasing number of publications over the last 12 years. For MECP2 gene, the number of published papers have declined over the past 12 years with a steady increase from 2010 and a moderate decline until 2022. The other 3 genes, NRXN1, SCN2A and SCN1A display a similar decline to MECP2 but with lower publications over the last 12 years.

Whilst investigating the other citation data, it is also quite surprising that the eagle score for the 2nd and 5th most reported gene on SFARI, had NaN values for their eagle score. The NaN values shown in the Eagle score category for MECP2 and SCN1A indicates there is no current evidence to support a causal role in for this gene in ASD[9]. Although this may be the case contradicting results showing gene-ASD association on PUBMED shown in Table 2, this could potentially be due to not-up-date, efficient curation of the eagle scores from the SFARI research group. Therefore, the role of gene in ASD does not have convincing evidence was demonstrated but rather potentially a range of neurodevelopmental phenotypes that are related to autism. From Table 1, SHANK3, NRXN1 and SCN2A have shown significantly high values for the eagle score above 70, which indicates the functional role of the gene in ASD has been demonstrated repeatedly in research and clinical settings. This is quite indicative of the roles of these genes in ASD. Collating data from bar chart shown in Figure 2, and Table 2, MECP2 is quite indicative of its gene-ASD association, which the highest number of reports on PUBMED, relatively high publications shown through the years despite having a NaN eagle score. It is quite important to note as well that this gene was also reported initially as a novel gene in 1993 for its association with ASD, and the number of publications since has increased from then, where improved biomedical research quality can be assumed throughout the years to further infer this genetic association with ASD.

Figure 1:

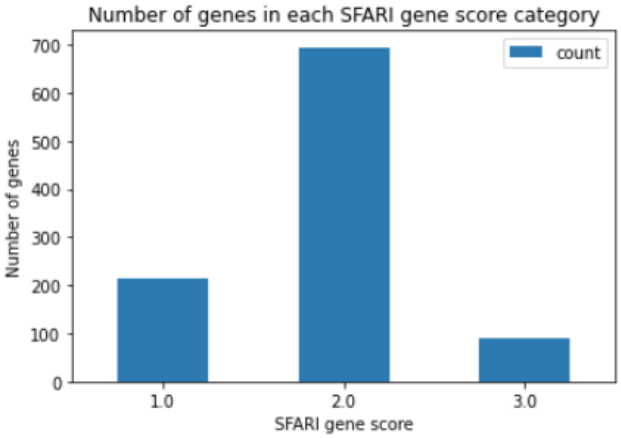


Table 1:

gene-symbol	gene-name	ensembl-id	chromosome	genetic-category	gene-score	syndromic	eagle	number-of-reports
SHANK3	SH3 and multiple ankyrin repeat domains 3	ENSG00000251322	22	Rare Single Gene Mutation, Syndromic, Genetic ...	1.0	1	74.85	120
MECP2	Methyl CpG binding protein 2	ENSG00000169057	X	Rare Single Gene Mutation, Syndromic, Functional	1.0	1	NaN	107
NRXN1	neurexin 1	ENSG00000179915	2	Rare Single Gene Mutation, Syndromic, Genetic ...	1.0	0	143.75	100
SCN2A	sodium channel, voltage-gated, type II, alpha ...	ENSG00000136531	2	Rare Single Gene Mutation, Syndromic, Functional	1.0	0	109.30	96
SCN1A	sodium channel, voltage-gated, type I, alpha s...	ENSG00000144285	2	Rare Single Gene Mutation, Syndromic, Genetic ...	1.0	1	NaN	84

Table 2:

Genes	Number of PUBMED papers
SHANK3	495
MECP2	524
NRXN1	184
SCN2A	104
SCN1A	80

Figure 2:

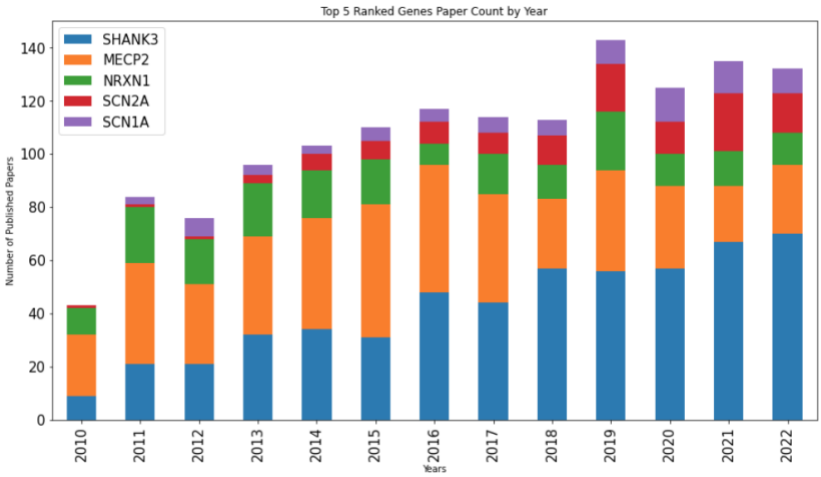


Figure 1: Bar chart presents number of genes in SFARI Gene Score 1, 2 and 3
Table 1: Top 5 ranked genes based on number of Reports on SFARI, citation data includes Eagle score, and number of publications heavily curated by SFARI
Table 2: Top 5 ranked genes (SHANK3, MECP2, NRXN1, SCN2A & SCN1A) and the number of PUBMED reports
Figure 2: Number of PUBMED publications from 2010-2022 for top 5 Ranked genes on SFARI: SHANK3, MECP2, NRXN1, SCN2A and SCN1A; summarized breakdown of number of PUBMED publications of each gene through 2010-2022 shown in Supplementary Table 1 (Appendix)

Autism Genes:

From Table 3 below, through a biological standpoint, most of the GO terms annotated with genes in gene score 1, 2, 3 are quite generic. Terms such as nucleus (GO: 0005634) and protein binding (GO: 0005515) which are found in the 3 gene score categories are generic- these terms are commonly annotated in most cellular and molecular biological/ biomedical literature[18]. In this case, when text-mining for GO annotations in publications and databases such as KEGG and REACTOME, molecular and genetic mechanisms behind autism research will be more likely have GO annotations regarding these cellular components which make up more than half of the 10 annotated terms in these gene score categories. These would most likely be parent terms when investigating more specific GO annotations relevant to the functional classification of autism genes. This is further proven by Guzzi et. al who has mentioned that many annotations are shallow in a directed acyclic graph (DAG). GO utilizes 3 DAGs to define functions of a gene product: molecular function ontology (MFO), biological process ontology (BPO) and cellular component ontology (CCO)[18][19]. Therefore, this may limit the accuracy of the inferences made about the functionality of these autism genes. Furthermore, as calculation of information-based methods (IC) in Gene Ontology website depends on annotation

corpus which links the large number of gene ontology terms given (7319 GO terms in Gene Score 1; 17603 GO terms in Gene Score 2, and 2601 GO terms in Gene Score 3), this has an issue in terms of the same GO term having different IC values when different corpora are used[19].

With investigating the top most commonly annotated GO terms, the results obtained from "Functional Classification" with Biological Processes ontology, these GO terms indicate that these genes are responsible in the cellular and metabolic processes in multicellular organisms such as *Homo sapiens*. For example, GO terms such as plasma membrane (GO: 0005886) was frequently found in all gene score categories as this is a crucial organelle within higher eukaryotes. Therefore, this does not provide any functional implications surrounding autism genes in general despite the frequency of these GO terms being high.

Table 3: Top 10 Commonly annotated terms with respective SFARI Gene Score Category 1, 2, and 3; GO term IDs, term description and GO term count (frequency of GO terms in relevant gene-ASD publications) *i) Gene Score 1; ii) Gene Score 2; iii) Gene Score 3*

i) Gene Score 1:

GO term ID	GO term Description	GO term count
GO:0005634	Nucleus	173
GO:0005515	Protein Binding	170
GO:0005654	Nucleoplasm	140
GO:0005886	Plasma membrane	113
GO:0005829	Cytosol	106
GO:0005737	Cytoplasm	83
GO:0045944	Positive regulation of transcription by RNA polymerase II	72
GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	53
GO:0000122	Negative regulation of transcription by RNA polymerase II	52
GO:0006357	Regulation of transcription by RNA polymerase II	52

ii) Gene Score 2:

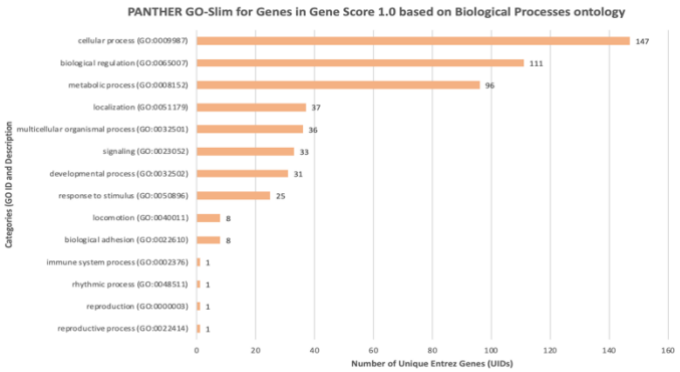
GO term ID	GO term Description	GO term count
GO:0005886	Plasma membrane	484
GO:0005515	Protein Binding	470
GO:0005829	Cytosol	307
GO:0005634	Nucleus	294
GO:0005654	Nucleoplasm	243
GO:0005737	Cytoplasm	241
GO:0016020	Membrane	143
GO:0046872	Metal ion binding	89
GO:0070062	Extracellular exosome	89
GO:0003723	RNA binding	81

iii) Gene Score 3:

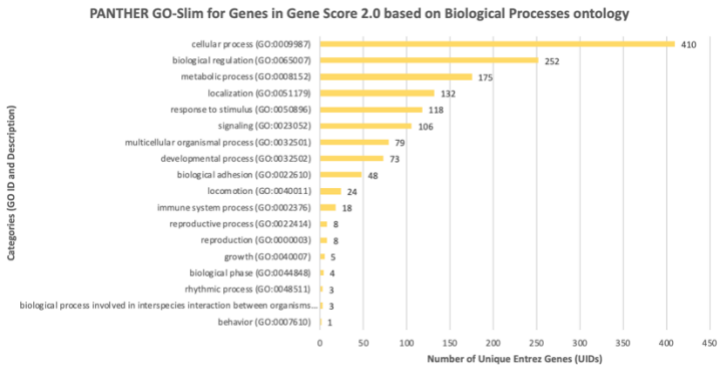
GO term ID	GO term Description	GO term count
GO:0005515	Protein Binding	75
GO:0005829	Cytosol	61
GO:0005737	Cytoplasm	48
GO:0005886	Plasma membrane	47
GO:0005634	Nucleus	46
GO:0005654	Nucleoplasm	32
GO:0016020	Membrane	24
GO:0005524	ATP binding	17
GO:0046872	Metal ion binding	17
GO:0003723	RNA binding	17

Figure 3: Number of Unique Entrez ID genes annotated to GO annotated terms in biological processes ontology domain through PantherDB GO-SLIM analysis; three bar charts represent number of genes in gene score 1, 2, 3 with annotated GO terms in Biological processes ontology.
i) Gene Score 1, ii) Gene Score 2; iii) Gene Score 3 (colour coded from above).

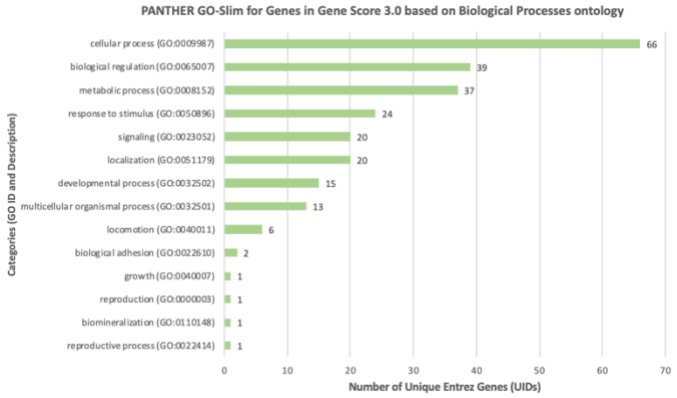
i)



ii)



iii)



From Figure 3 above, PantherDB GO-Slim has provided relatively less specific results regarding GO terms annotated with these autism genes[20]. For example, in Figure 3i, GO terms in the biological processes ontology, such as "biological regulation" (GO: 0065007) and "locomotion" (GO: 0040011) are more relevant to autism compared to the GO terms when finding the top 10 commonly annotated terms. However, there is higher consistency with the PantherDB results provided. There is more agreement with the top GO annotations in PantherDB than the top 10 commonly GO terms shown in Table 3. This is quite indicative of the limitations of implementing GO-Slims[20][21]. With a list of GO terms provided with respect to the gene2go file, enriched with overexpressed genes, if a child term has highly statistically significant enrichment, the parent term potentially would appear significantly enriched purely for including all the genes from the child term. GO-slims, are limited to general GO terms which are less interesting than more specific terms- which GO-Slim has removed[21].

For the investigating of GO annotations with respect to the top 5 genes found in autism literature, HumanCyc database was used[13]. SHANK3 has 56 GO terms associated with the biological processes' ontology. From Table 3i above regarding top 10 most commonly annotated terms in gene score 1 category, it is known that SHANK3 colocalizes in the plasma membrane[22]. This is also presented in other databases such HumanCyc and KEGG which were utilized to crosscheck this study. From the table above, GO terms in biological processes ontology was not informative as compared to GO terms associated with SHANK3 gene in the HumanCyc database. In the biological processes ontology domain, GO terms associated with SHANK3 included relatively specific GO terms such as dendritic spine morphogenesis (GO: 0060997)[23] and positive regulation of synapse structural plasticity (GO: 0051835)[24] which are more specific to the molecular mechanism to ASD. This was also demonstrated in the other 4 genes where the GO terms annotated with respective genes in the HumanCyc database were more relevant to autism. This provides a higher confidence measure for gene-ASD association.

When investigating duplicates found when merging the SFARI dataset to gene2go file for obtaining Unique Entrez IDs, there were 8 duplicate IDs found for 8 gene symbols from SFARI dataset which are shown in the Supplementary Table 2. Further investigation into these genes on NCBI, it was shown that these gene IDs were equivalent to each other. For example, for gene symbol ST7, when queried on NCBI, ST7 (Gene ID: 7982) was also known as ST7OT3 (Gene ID: 93655). Therefore, in order to decrease redundancy in the results in this section, the duplicates were removed.

Autism Network:

Table 5: STRINGDB cluster analysis results

Number of Nodes	213
Number of Edges	1555
Average Node Degree	14.6

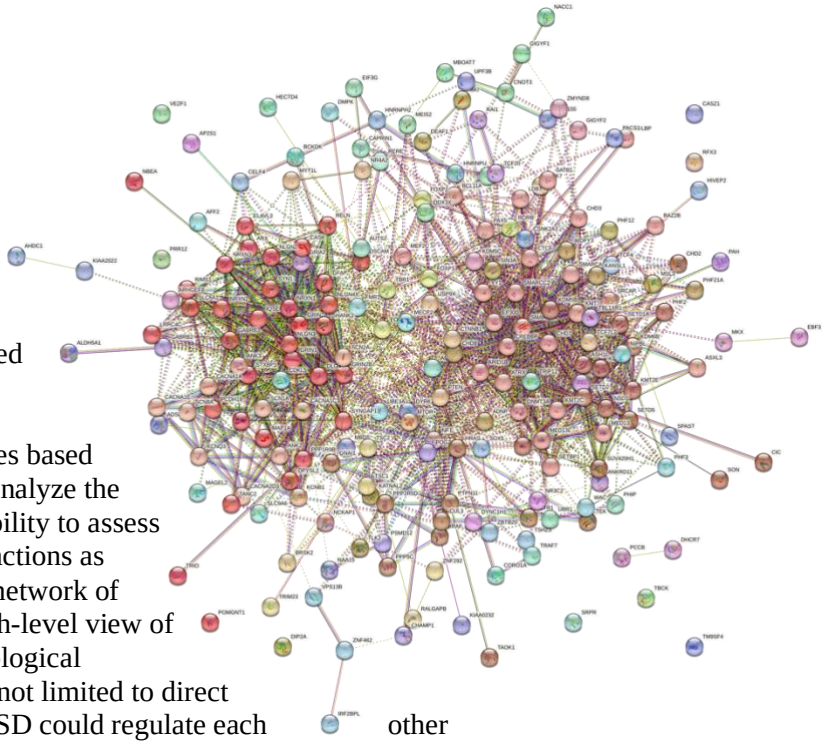
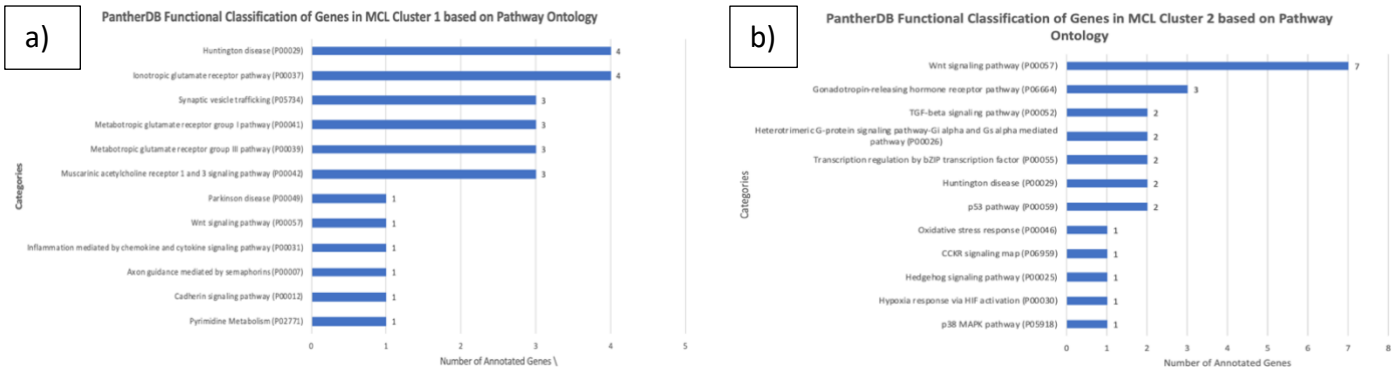


Figure 4: Biological network representing protein-protein interactions between genes in gene score 1. Spheres represent nodes where different colours indicate different clusters, i.e. red (Cluster 1) and edges are represented by lines.

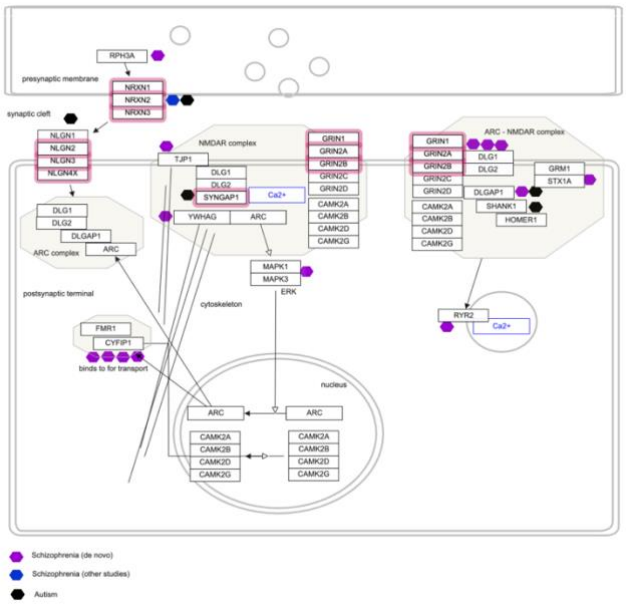
The concept of functional coupling between genes based on conservation of gene clusters can be used to analyze the gene-ASD associations[25]. With STRING, the ability to assess and compare the significance of individual interactions as STRING provides a graphical representation of network of inferred protein interactions which provide a high-level view of functional linkage, facilitating the analysis in biological processes[26]. As protein-protein interactions are not limited to direct physical interactions, proteins associated with ASD could regulate each other transcriptionally, or participate in a multiprotein complex assembly[26]. From Table 5, which provides statistics about this network accumulated from genes in Gene Score 1 (214 genes), demonstrated 213 nodes and 1555 edges which indicate relatively high number of connections either in terms of regulation or co-expression within the network.

Furthermore, protein-protein interaction enrichment p-value is less than 1.0e-16 which indicates proteins have more interactions among themselves when compared to a random combination of proteins[26][17]. This provides substantial evidence that the gene products in gene score 1 are biologically and functional relevant together. As shown in Figure 5, when the top 2 clusters were extracted in tsv format for PantherDB functional classification analysis, using Pathway ontology, MCL Cluster 1 (which has 33 genes) showed 4 genes each interacting in Huntingtin disease (P00029) and Ionotropic glutamate receptor pathway (P00037) whereas MCL Cluster 2 (26 genes) showed 7 genes interacting with each other in Wnt signaling (P00057). Biomedical literature evidence has suggested that the Wnt signaling pathway[27][28] is crucial during nervous system development- mutations in genes responsible for this pathway (7 genes found in Gene Score 1) would lead to adverse effects in neurodevelopment[28]. For MCL Cluster 1, there is no direct biomedical literature which indicates a direct relationship between Huntington's disease[29] and ASD but there are some overlaps in terms of both these diseases being a disorder of the nervous system. However, in Ionotropic glutamate receptor pathway, there is direct evidence of the genes responsible in this pathway, where if mutated, will cause alterations in glutamate levels[30], which leads to disrupt neuronal function in autism and induce autism phenotypes. To further analyze the results from PantherDB, KEGG pathways obtained from STRING was used. The KEGG pathway annotated which showed the lowest false discovery rate (FDR) was cell adhesion molecules pathway (hsa04514) which showed a moderately good strength value at 1.41 and 4.30e-05 for FDR value. As strength value measures how large enrichment effect is, it can be implicated that there are a high number of proteins annotated with a term in MCL Cluster 1. FDR values, which describe how significant the enrichment is, shows that the results here are statistically significant. The other KEGG pathways were also relevant to neuronal development such as synaptic vesicle cycle(hsa04721) and glutamatergic synapses (hsa04724).

Figure 5 (PantherDB Functional Classification based on Pathway Ontology): a) Barchart demonstrating Number of annotated Genes (in Cluster 1) within different GO pathways b) Barchart demonstrating Number of annotated Genes (in Cluster 2) within different GO pathways



When further investigating the top cluster (MCL Cluster 1 genes) on CytoScape[15], this provided more accurate results in terms of functionally characterizing these cluster genes. The most reliable pathway was shown in Disruption of postsynaptic signaling by Copy Number Variations (CNV)- Homo sapiens (WP4875) where 10 of the 33 cluster genes were found to interact in a pathway. A pathway visualization is shown in Figure 6 below, where genes such as NRXN1 and GRIN2A interact with each other in this pathway and where certain genes here are localized as a whole. This pathway was extracted as it provided the lowest p-value at 1.98e-16 which indicates the effect is large and result is of major theoretical or clinical importance. Similarity score as compared to the other pathways was also the highest at 0.39 where a major proportion of the genes (10/33) in this cluster has shown evidence to interact with each other (indirectly or directly). The other pathways shown are similar to ones shown in Figure 5a where glutamatergic synapse pathway is related to at least 3 of the pathways from PantherChart 5a above (P00037, P00041, P00039). Utilizing stringApp to perform a functional enrichment analysis with an FDR threshold of 5%, this was used to obtain the most significant term in KEGG pathways which was Ionotopic glutamate receptor pathway (P00037) which covers a substantial proportion of pathways shown in Figure 5a where alterations in glutamate levels would lead to autistic phenotypes. This is quite indicative of genes in cluster are functionally responsible for the glutamate receptor pathway.



MCL Cluster 2 genes, on the other hand, yielded insignificant and uninformative results as compared to Cluster 1 genes. Through CytoScape, this yielded the Prion disease pathway- Homo sapiens with 5 proteins interacting with each other, which is known to be indirectly related to autism, however, this is mostly in terms of an overlap in neurodevelopmental phenotypes rather than ASD itself. KEGG pathways provided on StringDB, did not yield significant results as compared to MCL Cluster 1. Overall, autism networks showed more promising GO annotations in terms of functional characterization of these genes compared to autism genes.

Figure 6: CytoSCAPE-illustrated diagram for evidence of genes in gene score 1 interacting with each other. NRXN1/2/3, GRIN1, GRIN2A/B, NLGN2/3/4X and SYNGAP1 interact with each other directly/ indirectly as shown.

Discussion:

When mining autism literature, it was shown that SHANK3 and MECP2 produced the most number of reports on PUBMED in gene score 1. This potentially illustrates a clearer understanding over the years in terms of the research development for deciphering the molecular and genetic mechanism behind ASD. This also would provide a higher confidence metric in terms of gene-ASD association through literature mining. For example, it can be implied that SHANK3 and MECP2 has a higher confidence that they are associated with autism due to the number of reports published on PUBMED. SFARI's gene ranking also shows that there is a substantial proportion of genes in gene score category 2 at 695 genes out of the 1095 provided. This should be further evaluated as SFARI's database was manually curated- was it due to limitations of the research foundation in terms of curating the scores or was there not sufficient publications/ evidence? Furthermore, literature could be evaluated through a comprehensive literature data-mining analysis. Several research papers have shown that through literature metric analysis[31], has shown a broad genetic network functionality associated with diseases. This was conducted through considering factors such as number of citations, quality weight, and novelty of gene in terms of how recent was research conducted for the gene associated with ASD[31]. This literature metric analysis can be conducted using Elsevier's proprietary MedScan natural language processing (NLP) system)[32] where data extracted included genes and their association types including binding and regulation where network analysis and KEGG pathways can be explored on further.

For autism gene analysis, BioMart, a data mining tool for ENSEMBL, allowed fast and efficient querying when mapping the SFARI gene symbols to Entrez Gene IDs (UIDs)[10][11]. However, there are a few limitations to using this module in the gseapy package as there were duplicates risen. BioMart could not differentiate the unique gene IDs on NCBI for the same gene symbols on SFARI on Supplementary Table 2. This increases noise and redundancy in our results. Furthermore, in some cases, there were missing GeneID entries where manual deduction on NCBI website had to be retrieved such as for MAPT-AS1 and PTCHD1-AS. These were antisense strands where unique gene IDs had to be manually curated into the text file which are RPS10P2 and MSNP1. In this case, the disadvantage of using BioMart is that this module in gseapy only considers genes which are already annotated under ENSEMBL. In the top 10 most commonly annotated terms for genetic analysis, these terms were also mostly generic GO terms which could potentially be parental terms, where this did not aid in functionally characterizing the genes more specifically. The GO terms annotated here has helped in identifying that the genes supposedly related to ASD are responsible in the molecular and cellular processes in humans- in terms of protein localization within the cell but not how the genes function within the cell. PantherDB Go-Slim, in these results, also did not show much promise in functionally characterizing these GO terms to the genes. However, they were more specific than the aforementioned method. PantherDB extracted GO terms such as locomotion, and biological adhesion which would be more specific to the mechanism and effects of ASD but not necessarily in specific way. Therefore, analyzing the KEGG pathways of genes recognized from autism literature would provide a better standpoint in understanding the functional relevance of these genes and the mutations caused in these genes into the genetic and molecular mechanisms behind ASD, such as shown in the example of SHANK3.

In relevance to both network analysis and gene analysis, measuring semantic similarity between GO terms would also be useful in functional bioinformatics research[33][19]. This would allow a more quantitative approach in understanding GO annotations in both result sections. For autism network analysis, results produced provided a better insight into how these genes interact with each other and a better understanding regarding the mechanism behind mutations in genes responsible for autism. This also provided us with higher confidence of some genes such as NRXN1 are crucial and when mutated, will disrupt a neuronal signaling pathway which causes adverse effects in locomotion. Further network analysis in terms of such as TOMAS which a novel Topology-aware Meta-analysis approach for pathway analysis was known to overcome noise and bias to identify pathways implicated in diseases[34]. Overall, further research should be conducted in order to overcome challenges and limitations in the methods used for biological network analysis. This would be useful in terms of understanding complex diseases such as autism.

References:

1. Hodges H, Fealko C, Soares N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. *Translational Pediatrics*. 2020;9(Suppl 1):S55-S65. doi:10.21037/tp.2019.09.09

2. Campisi L, Imran N, Nazeer A, Skokauskas N, Azeem MW. Autism spectrum disorder. *British Medical Bulletin*. 2018;127(1):91-100. doi:10.1093/bmb/ldy026

3. Qin D. Next-generation sequencing and its clinical application. *Cancer biology & medicine*. 2019;16(1):4-10. doi:10.20892/j.issn.2095-3941.2018.0055

4. AlAyadhi LY, Hashmi JA, Iqbal M, et al. High-resolution SNP genotyping platform identified recurrent and novel CNVs in autism multiplex families. *Neuroscience*. 2016;339:561-570. doi:10.1016/j.neuroscience.2016.10.030

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

6. Banerjee-Basu S, Packer A. SFARI Gene: an evolving database for the autism research community. *Disease Models & Mechanisms*. 2010;3(3-4):133-135. doi:10.1242/dmm.005439

7. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*. 2000;25(1):25-29. doi:10.1038/75556

8. Casanova MF, Casanova EL, Frye RE, et al. Editorial: Secondary vs. Idiopathic Autism. *Frontiers in Psychiatry*. 2020;11. doi:10.3389/fpsyt.2020.00297

9. Abrahams BS, Arking DE, Campbell DB, et al. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular Autism*. 2013;4(1):36. doi:10.1186/2040-2392-4-36

10. Smedley D, Haider S, Ballester B, et al. BioMart – biological queries made easy. *BMC Genomics*. 2009;10(1):22. doi:10.1186/1471-2164-10-22

11. Smedley D, Haider S, Durinck S, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*. 2015;43(W1):W589-W598. doi:10.1093/nar/gkv350

12. Mi H, Ebert D, Muruganujan A, et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*. 2020;49(D1):D394-D403. doi:10.1093/nar/gkaa1106

13. Trupp M, Altman T, Fulcher CA, et al. Beyond the genome (BTG) is a (PGDB) pathway genome database: HumanCyc. *Genome Biology*. 2010;11(Suppl 1):O12. doi:10.1186/gb-2010-11-s1-o12

14. Stobbe MD, Houten SM, Jansen GA, van Kampen AH, Moerland PD. Critical assessment of human metabolic pathway databases: a stepping stone for future integration. *BMC Systems Biology*. 2011;5(1). doi:10.1186/1752-0509-5-165

15. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape stringApp: Network analysis and visualization of proteomics data. Published online October 11, 2018. doi:10.1101/438192

16. Mering C v. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Research*. 2003;31(1):258-261. doi:10.1093/nar/gkg034

17. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*. 2019;47(Database issue):D607-D613. doi:10.1093/nar/gky1131

18. Guzzi PH, Mina M, Guerra C, Cannataro M. Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in Bioinformatics*. 2011;13(5):569-585. doi:10.1093/bib/bbr066

19. Zhao C, Wang Z. GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms. *Scientific Reports*. 2018;8(1). doi:10.1038/s41598-018-33219-y

20. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*. 2019;47(D1):D419-D426. doi:10.1093/nar/gky1038

21. 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(D1):D183-D189. doi:10.1093/nar/gkw1138

22. Durand CM, Perroy J, Loll F, et al. SHANK3 mutations identified in autism lead to modification of dendritic spine morphology via an actin-dependent mechanism. *Molecular Psychiatry*. 2011;17(1):71-84. doi:10.1038/mp.2011.57

23. Lo LHY, Lai KO. Dysregulation of protein synthesis and dendritic spine morphogenesis in ASD: studies in human pluripotent stem cells. *Molecular Autism*. 2020;11(1). doi:10.1186/s13229-020-00349-y

24. Rubio MD, Johnson R, Miller CA, Huganir RL, Rumbaugh G. Regulation of Synapse Structure and Function by Distinct Myosin II Motors. *Journal of Neuroscience*. 2011;31(4):1448-1460. doi:10.1523/jneurosci.3294-10.2011

25. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences*. 1999;96(6):2896-2901. doi:10.1073/pnas.96.6.2896

26. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*. 2020;49(D1):D605-D612. doi:10.1093/nar/gkaa1074

27. Zhang Y, Yuan X, Wang Z, Li R. The Canonical Wnt Signaling Pathway in Autism. *CNS & Neurological Disorders - Drug Targets*. 2014;13(5):765-770. doi:10.2174/18715273126661312231141 49

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

29. Roos RA. Huntington's disease: a clinical review. *Orphanet Journal of Rare Diseases*. 2010;5(1). doi:10.1186/1750-1172-5-40

30. Nisar S, Bhat AA, Masoodi T, et al. Genetics of glutamate and its receptors in autism spectrum disorder. *Molecular Psychiatry*. 2022;27(5):2380-2392. doi:10.1038/s41380-022-01506-w

31. Xu C, Cao H, Zhang F, Cheadle C. Comprehensive literature data-mining analysis reveals a broad genetic network functionally associated with autism spectrum disorder. *International Journal of Molecular Medicine*. Published online August 28, 2018. doi:10.3892/ijmm.2018.3845

32. Novichkova S, Egorov S, Daraselia N. MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics*. 2003;19(13):1699-1706. doi:10.1093/bioinformatics/btg207

33. Kamran AB, Naveed H. GOntoSim: a semantic similarity measure based on LCA and common descendants. *Scientific Reports*. 2022;12(1). doi:10.1038/s41598-022-07624-3

34. Nguyen T, Diaz D, Draghici S. TOMAS. *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. Published online October 2, 2016. doi:10.1145/2975167.2975168

Appendix/ Supplementary material:

Table 1: Number of publications for top 5 ranked genes in SFARI from 2010 to 2022

Genes	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
SHANK3	9	21	21	32	35	32	48	44	57	56	57	67	73
MECP2	23	38	30	37	43	51	48	41	26	39	32	22	28
NRXN1	10	21	17	20	18	17	8	15	13	22	12	13	12
SCN2A	1	1	1	3	6	7	8	9	11	18	12	22	16
SCN1A	0	3	7	4	3	5	5	6	6	9	13	12	9

Table 2: Duplicate Entrez Gene IDs of SFARI Gene Symbols

Gene Symbol	ID_1	ID_2
ST7	93655	7982
TECTA	116804918	7007
H4C5	554313	8367
RHOXF1	104797536	158800
CACNA1C	100874369	775
USP9Y	64595	8287
H4C3	554313	8364