Universiteit Leiden
The Netherlands

Opleiding Bioinformatica

Network analysis of polyQ proteins in the human brain

Dylan Kossen

Supervisors:
Katy Wolstencroft

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

July 11, 2021

Abstract

PolyQ diseases are diseases caused by the elongation of regions with glutamine repeats.These disease cause degenerative neurological disorders, which are not curable at this date. There is not much known about the native functions or structure of polyQ proteins. As polyQ proteins cause neurodegenerative diseases, the expression and interaction data from brain tissue is needed. Until recently it was difficult to gather protein expression and interaction data from human brain tissue. This is why most research focuses on the expression data gathered from the mouse. However, it is not well known how those results translate to the functions of polyQ proteins in humans. This is why this thesis created a protein-protein interaction network for proteins expressed in the human brain. The protein-protein interaction network was used to be able to analyse functional enrichment and general network and node properties. The functional enrichment has shown that there are only minor differences in the functions of human polyQ proteins and the orthologs of mouse polyQ proteins. The general node properties of degree and closeness centrality showed that the human polyQ proteins are more connected and more important for the speed of information in the network than the orthologs of mouse polyQ proteins and the non-polyQ proteins. This, in combination with the functional enrichment results, indicates that polyQ proteins play an important role in binding. As there are some major differences in properties of the human polyQ proteins and the orhtologs of mouse polyQ proteins, the experiments done in mice cannot be directly translated to humans, without looking at them more carefully.

Contents

1 Introduction

This introduction will give some background information to explain the subjects of this thesis. The following subjects will be talked about in this section: polyQ proteins, Huntington's disease and protein-protein interaction networks.

1.1 PolyQ proteins

PolyQ proteins are proteins which have a region of glutamine (Q) repeats. PolyQ repeats are one of the more common homopeptide repeats. The elongation of a polyQ region can have disastrous consequences and as of now nine different diseases caused by this have been identified. These diseases are: spinocerebellar ataxias types 1, 2, 6, 7, and 17; dentaorubral pallidoluysian atrophy; spinal and bulbar muscular atrophy; Machado-Joseph disease; and Huntington's disease. [FHC+14] There is not much known about the native functions and structures of polyQ proteins. This is because most studies focus on the proteins which cause the aforementioned diseases. As the polyQ disease cause these neurodegenerative diseases, the expression and interaction data from brain tissue is needed, however, this data was until recently difficult to gather. Even though studying these types of proteins has been extremely difficult, there have been studies which suggest that polyQ proteins are involved in binding other proteins. [TANM17] [HSW08] This is a reason why this thesis will look at the protein expression data from the human brain to create a network and analyse the polyQ proteins in order to see if the polyQ proteins are indeed involved in protein-protein interactions.

1.2 Huntington's disease

Huntington's disease is a progressive neurodegenerative disease caused by the elongation of the polyglutamine region in *huntingtin*. The disease causes symptoms like cognitive decline and chorea. A pathological characteristic of the disease are nuclear and cytoplasmic inclusions of polyglutamine and mutant *huntingtin*. [Wal07] A normal *huntingtin* protein has less than 35 glutamine repeats. Signs of Huntington's start showing between 36-40 glutamine repeats. The disease fully develops when *huntingtin* has 41 or more glutamine repeats. [Fin11] As of now there is no cure for this disease or any of the other diseases caused by the elongation of the polyQ region in proteins.

1.3 Protein-protein interaction networks

Protein-protein interactions can be visualised in networks. These networks can then be analysed to gain more information about the proteins in the networks. This is possible because it has been proven that protein-protein interaction networks behave just like other complex networks like social network or the internet. [VFMV03] It is thus possible to look at the mathematical properties of the nodes and come to conclusions about the biological functions of the proteins.

The network properties that are most important for protein-protein interaction networks are degree, closeness centrality, betweenness centrality, and the assumption that the network is scale-free. A network is scale-free when a large number of nodes have a small number of edges, and a small number of nodes have a large number of edges, so the degree distribution of the network follows

a power law. It has been proven that biological networks are scale-free. [Alb05] When a network is scale-free it has the small world property. The small world property means that you can go from any random node in the network to any other random node within at most 6 steps. This can be seen by looking at the characteristic path length of the network. The characteristic path length is the average of all the shortest paths in the network, the characteristic path length will certainly be below 6 for a network that is scale-free. The fact that biological networks are scale-free brings another important property with it. It makes it possible to identify hub nodes. Hub nodes in protein-protein interaction networks are of great interest, because a mutation in one of these proteins could lead to enormous consequences for the organism, as the protein is involved in so many interactions with other proteins. In the strict definition a hub is a node with a higher degree than average. For the protein-protein interaction networks in this thesis this would that there would be hundreds of hubs. This not feasible, this is why the definition of hub for this thesis has been put on around 2-3 times as high degree than average.
The closeness centrality property of nodes is another important property used in the analysis of protein-protein interaction networks. Closeness centrality can be seen as the speed of information flow through a node. The higher the closeness centrality is, the faster information can go from one side of the network to the other side of the network while going through the node. So a node with a high closeness centrality makes it important for the network.
Betweeness centrality is another type of centrality that is important in protein-protein interaction networks. The betweenness centrality measures the amount of shortest paths that come through a node. So a high betweenness centrality means that the node is important due to the amount of information that flows through it. Proteins that have a high betweenness centrality can be seen as bottlenecks in the network. These proteins are most likely key proteins in important pathways. [YKS+07]

1.4 Related work

This thesis is based on a previous thesis in which the research was done on proteins expressed in the mouse brain instead of the human brain. [Jan20] That research focused on the polyQ proteins expressed in the mouse brain, and used the orthologs of human polyQ proteins to come to conclusions. The way this thesis is setup is to follow the methods of the mouse thesis so that it is possible to compare the results of this thesis and the thesis on the mouse polyQ proteins. The mouse thesis found that there are not many differences in GO functions or GO processes between the orthologs of human polyQ proteins and the mouse polyQ proteins. The disease proteins, however, did show differences in GO functions and GO processes. Another conclusion, based on the node properties degree, betweenness centrality, and closeness centrality, was that polyQ proteins are more important in the network than the average protein. If the results of this thesis and the mouse thesis turn out to be similar, then that is really important to the field. As more experimental protein expression and interaction data is available on the mouse than on humans.

1.5 Research question

As most studies focus on polyQ proteins when they are in a disease state, more research into the native functions of polyQ proteins is needed. This thesis aims to gain information on the native functions of polyQ proteins in the human brain. This thesis only focuses on proteins expressed in

the brain, because the known polyQ disease proteins cause neurological illnesses. Previous research on this topic has used expression data from the mouse [Jan20] to study the native functions of polyQ proteins. As this thesis uses human expression data, the results will be compared with the research done in mouse in order to understand whether how well mouse studies translate to humans for polyQ proteins. So the goal for this thesis is to investigate the differences and similariteies between polyQ and non-polyQ proteins in the human brain and to compare these results with results from the previous work done in mouse. This goal makes it possible to create two main research questions:

RQ1: What are the common functions of polyQ proteins in the human brain?

RQ2: Is there a difference in properties for the disease polyQ proteins compared to regular polyQ proteins?

To include the comparison between these results and the results acquired from research in mouse, the first research question can be extended:

RQ1.1: Do these functions differ form the functions of polyQ proteins in the mouse brain?

In order to answer these questions protein-protein interaction networks will be used to create a network of proteins expressed in the human brain. These networks will be analysed and the results compared to those of the research done in mouse.

1.6 Thesis overview

Section 1 gives background information on important subjects for this thesis. Section 2 explains what methods and tools were used for the data gathering, network creation, and analysis. Section 3 shows all the results retrieved by this research. Section 4 discusses these results and section 5 gives the conclusion for this research.
This bachelor thesis is written for the bachelor bioinformatics at LIACS and was supervised by Katy Wolstencroft.
The data for this thesis can be found at: *https://git.liacs.nl/s2351765/thesisdata*

2 Methods

This section will give detailed information on how the research was done. The processes of data gathering, network creation, clustering and functional enrichment will be fully explained in their own subsections.

2.1 Data gathering

The Humanmine database was used to collect the data on proteins expressed in the human brain. Humanmine is an integrated database which contains human genomic data. [SAB+12] [KLB+14] The Humanmine database uses an interface which makes it possible to create very specific queries. To retrieve the protein data for the human brain the query was started with the following filter:

$$\text{Tissue group} = \text{Central nervous system (Brain)}.$$

This filter only returns proteins which are expressed in the different kinds of human brain tissue, like the hippocampus or the cerebral cortex. Next there needs to be filters on the expression level and the reliability of the data. It is best to gather only reliable data, but as discussed in Section 1, there is not much known about PolyQ proteins. This is why two datasets were gathered. One dataset (dataset lenient) has a more lenient filter on the reliability and the expression level. The more lenient filter is as follows: the expression level can be either low, medium or high and the reliability can be either approved, enhanced, supported or uncertain. The stricter dataset (dataset strict) has the filter as follows: expression level can be either medium or high and the reliability can be approved, enhanced or supported. The query was made so UniProt accession numbers were returned. The results were downloaded as a .csv file. This file contains many duplicate proteins, as a lot of proteins are expressed in the multiple different brain tissues. For instance, if a protein is expressed in the hippocampus and cerebral cortex the protein will be returned twice, once for the hippocampus and once for the cerebral cortex. These duplicates were filtered out of the .csv file by using a small script which first sorted the file alphanumerically and then removed the duplicates.

At this point there is still is no information on which proteins are PolyQ. To get this information the ScanProsite tool was used. This tool makes it possible to scan a protein sequence database for certain motifs.[DCSG+06] At this point it needed to be established when a protein is qualified as a PolyQ protein. The rule is if eight out of ten amino acids in sequence at some region in the protein are glutamine then the protein is qualified as a PolyQ protein. The different motifs that fit the rule were used to scan the UniProtKB protein sequence database. A filter was applied to only return proteins from Homo sapiens and the output format was set to matchlist. This returned a list of proteins which fit the rule and the list was saved in .txt file.

To be able to compare the results with the results from [Jan20], the ortholgs of PolyQ proteins from the mouse needs to be gathered. To get the the PolyQ proteins from the mouse, the ScanProsite tool was used again. The database used is the same as for human, UniProtKB. The same rule for when a protein is classified as a PolyQ protein was used and thus the same motifs were used. However, this time the filter was set to return only proteins from Mus musculus. The output was set to matchlist again. This returned a list of proteins which are PolyQ in mice. This dataset needs to be converted to get orthologs of these proteins in humans. To do this the UniProt accession

numbers needed to be converted to Ensembl Gene ID. This was done by using the retrieve/id mapping tool on the UniProt site. [Uni21] The tool lets the user upload a list of proteins and select to what the Accession numbers need to be converted. The Biomart tool was used to get the human ortholog Ensemble Protein IDs. [KKH+11] After this the UniProt retrieve/id mapping tool was used to convert the Ensembl Protein IDs to UniProt accession numbers.

2.2 Network creation

The networks created for this thesis are all protein-protein interaction networks. The nodes in the network are proteins and the edges are the interactions between proteins. The interactions are not just physical interactions between proteins in protein complexes, other associations between proteins are also taken into account. The network is undirected, as an interaction is always going both ways. To create the network two main data sets are needed:

- A dataset of proteins, in this case proteins expressed in the human brain
- A dataset of interactions between these proteins

The networks were created by using the open source software Cytoscape. Cytoscape makes it possible to create and visualise molecular interaction networks. [SMO+03] To gather the interaction data for the proteins, an app called stringApp had to be downloaded from the app store within Cytoscape. StringApp makes it possible to gather the interaction data out of the STRING database. [SGL+19] The network was created by going to $File \rightarrow Import \rightarrow Network\ From\ Public\ Databases$. Here the STRING: protein query database is selected as the data source. The list of proteins is then entered into the according field. The lenient dataset and the strict dataset were created using different confidence cutoffs. For the lenient network the confidence cutoff is 0.40 and for the stict dataset the confidence cutoff is 0.70. After this a disambiguation dialog appeared and showed the proteins which could not be matched with a unique STRING protein. All the proteins in this list were manually reviewed and the right STRING protein was selected to be in the network.

At this point there is a network but without information on what proteins are PolyQ proteins in human and what orthologs are PolyQ in mouse. To get this information into the network table, a table was created by using the datasets about the PolyQ proteins. The table was import into the network table by using $File \rightarrow Import \rightarrow Table\ from\ file$. The key column used to match the two tables is stringdb::canonical name.

To be able to analyse and compare the PolyQ proteins with regular proteins, multiple smaller networks were created. The following subnetworks were created: a first neighbour network of human PolyQ proteins, a first neighbour network of mouse ortholog PolyQ proteins, a first neighbour network of proteins of which are both PolyQ in human and mouse, a first neighbour network of all the PolyQ proteins, so they are PolyQ in human or their ortholog is PolyQ in mouse or both, a network of proteins PolyQ in human, a network of mouse ortholog PolyQ proteins, a network of proteins which are both PolyQ in human and in mouse, and a network of all the PolyQ proteins. The first neighbour networks were created by firstly selecting all the corresponding PolyQ proteins, then the first neighbours were added to the selection by going to $Select \rightarrow Nodes \rightarrow First\ Neighbours\ Of\ Selected\ Nodes \rightarrow Undirected$.

The network was created by going to *File* → *New Network* → *From Selected Nodes*, *All Edges*. The other subnetworks were created in the same manner but without the selection of the first neighbours.

2.3 Clustering

Clusters in a network are sets of nodes which have a higher amount of edges between them than a certain predetermined cut-off score. Clusters are interesting parts of the network as the higher connectivity between the nodes implies the involvement in a common biological process. By selecting the clusters which have the most amount of polyQ proteins in them, the biological processes in which these polyQ proteins are involved could be identified. The clustering was done by using an app within the Cytoscape [SMO+03] programme: MCODE [BH03]. MCODE is a clustering algorithm designed for finding protein complexes in protein networks. The lenient network was the network used to do clustering on. The parameter settings for MCODE were: k-score = 3, node score cut-off = 0.3, and degree cut-off = 3. The max depth parameter was left to its default value of 100.

2.4 Network analysis

A network on itself without any knowledge about the values of its properties does not give much information, this is why network analysis is needed. The in Cytoscape [SMO+03] integrated "Analyze Network" function returns a set of important network properties with its values. The properties for which the "Analyze Network" function returns the corresponding values are:

- Number of nodes in the network
- Number of edges in the network
- Average number of neighbours a node has
- Network diameter
- Network radius
- Characteristic path length
- Clustering coefficient
- Network density
- Network heterogeneity
- Network centralisation
- Connected components

There are also node specific properties like degree, closeness centrality, and betweenness centrality added to the node table. So the "Analyze Network" function also makes it possible to look into more detail for specific nodes, in this case the polyQ proteins. The "Analyze Network" function in Cytoscape [SMO+03] can be found under *Tools* → *Analyze Network*.

2.5 Functional enrichment

Functional enrichment of a network makes it possible to understand what the functions of the proteins are and in what biological processes they are involved. The Gene Ontology consists of data of molecular function, cellular component, and biological process [ABB+00] [gen21]. GO terms are used to identify the specific process, function, or component.

The proteins in the network can have annotation terms. These annotation terms are gathered from the Gene Ontology database. The frequencies of the annotation terms in the network and the frequencies of these annotation terms happening by chance are used to calculate the p-values for the annotation terms. The p-value displays the certainty that a term is not just present in the network due to chance. The p-value can also be used to visualise the results of the functional enrichment.

The functional enrichment is done by using the StringApp. The StringApp has a function, under *Apps* in Cytoscape, called *STRING Enrichment*. The functional enrichment is done by then selecting *Retrieve functional enrichment*. The functional enrichment returns a table with information on the different GO functions, GO components, GO processes the proteins are involved in. The results are ordered based on the p-value for the GO term based on the proteins and their interactions in the network. The lower the p-value the more certain the algorithm is that that GO term is in the network for the associated proteins. To better understand the results in this table the GO terms for each category, GO component, GO function, and GO process, are selected and exported to a table. This table is then filtered to retrieve only the GO terms in the first column and the associated p-value in the second column. This table is then entered into REVIGO [SBŠŠ11]. REVIGO allows to retrieve a tree map table for the entered GO terms and their p-values. This tree map table is then entered into CirGO [KLS+19]. CirGO visualises the entered tree map in a circular graph with the names of the GO terms instead of the identifier to make interpretation easier.

3 Results

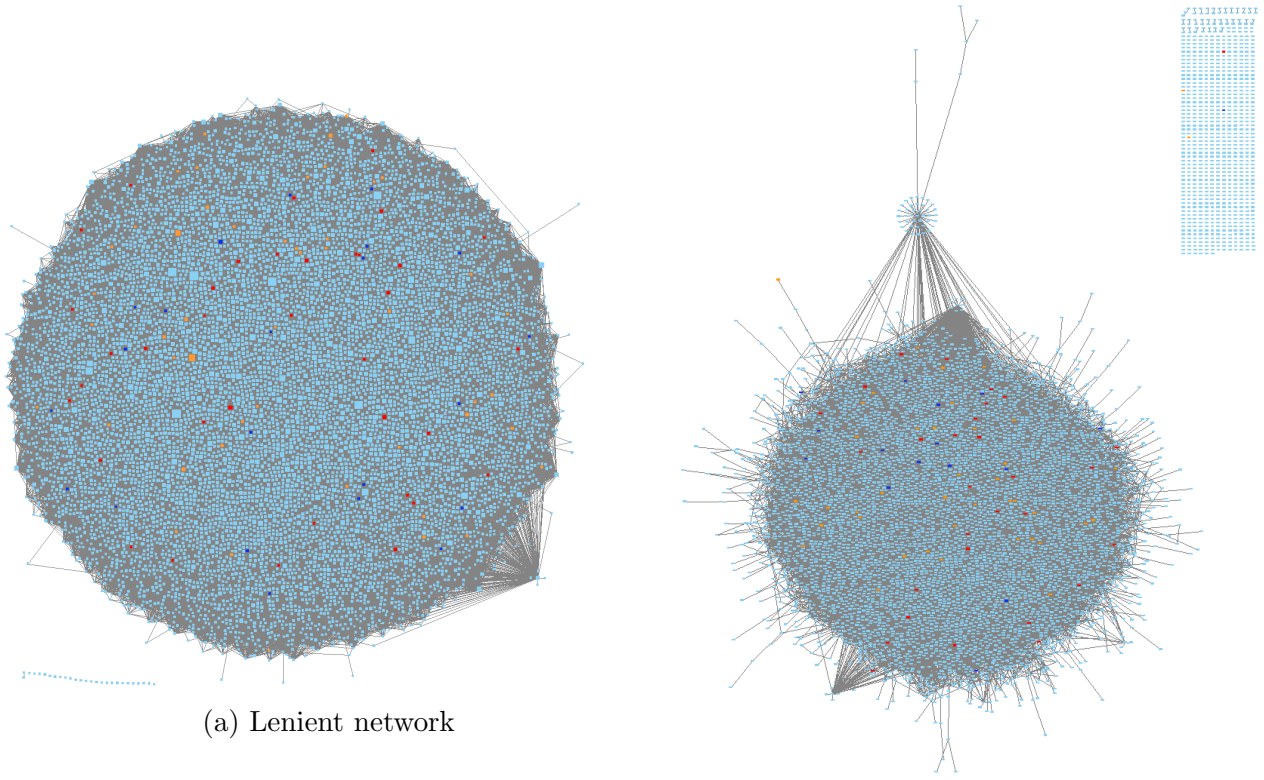
This section will show the results acquired from the different kinds of network analyses. In the first subsection, Network, the network properties of the two main networks, the network created from the lenient dataset and the network created from the strict dataset, will be shown. In the second section, Network clustering, the results of the clustering of the network will be shown. In the section Network analysis the results of the network analysis from all the networks will be shown. In the final section, Functional enrichment, the results from the functional enrichment will be shown.

3.1 Network

The different properties of the lenient and the strict network are compared in table 1. This table shows how the strictness of the data collection and network creation strongly influences the different network properties. It can be seen that the lenient network has 3176 more proteins in the network than the strict network. The lenient network also has a lot more interactions between the proteins as there are 304875 more edges in the lenient network than in the strict network. The giant components (GC) are also quite different. In the lenient network there are only 27 proteins which are not in the giant component, while in the strict network there are 807 proteins not in the giant component, this can also be seen in figure 1. The amount of edges outside of the giant component is small for both of the networks. So most proteins which are not in the giant component have no edges at all. The lenient network has 13 more human polyQ proteins and 12 more orthologs of mouse polyQ proteins in the network than the strict network. These results are the main reason why the network analysis and most of the functional enrichment was done on the lenient network and not on the strict network.

	Lenient Network	Strict Network
# nodes	10042	6866
GC # nodes	10015	6059
# edges	419821	114946
GC # edges	419820	114914
Avg. # neighbours	83.838	37.932
Diameter	6	11
Radius	4	6
CPL	2.740	3.562
Clustering coef.	0.232	0.428
Density	0.008	0.006
Heterogeneity	1.096	1.397
Centralisation	0.106	0.125
# Human PQ	84	71
# Mouse PQ	58	46

Table 1: Comparison of the properties of the lenient and the strict network, GC = giant component, CPL = characteristic path length



(a) Lenient network

(b) Strict network

Figure 1: The lenient and strict networks

3.2 Clustering

The clustering algorithm MCODE [BH03] returned 29 clusters. Most of these clusters contained none or just one polyQ protein. The clusters with 4 or more human polyQ proteins were selected. These clusters can be seen in figures 2, 3, and 4. Cluster 11 contains 4 human polyQ proteins. Clusters 6, 14, and 17 contain 5 human polyQ proteins each. Cluster 7 contains 8 human polyQ proteins. And lastly, cluster 9 contains 10 human polyQ proteins.

3.3 Network analysis

The results of the network analysis are shown in this section. First, the properties of the polyQ proteins and non-polyQ proteins from the lenient network will be shown, after that the results of the polyQ first neighbour networks, then the results of the polyQ networks, after that the analysis of the disease proteins, and finally the results of the clusters.
Table 2 shows that the human polyQ proteins are much more connected on average than the orthologs of mouse polyQ proteins and the non polyQ proteins, as the degree of the human polyQ proteins is on average higher than that of the other proteins. The human polyQ proteins also have higher average closeness and betweenness than orthologs of mouse polyQ proteins and the non-polyQ proteins.

	Only PQ Human	PQ Human	Only PQ Mouse	PQ Mouse	non-PQ
Avg. degree	105.979	97.167	64.545	77.5	83.541
Avg. betweenness	0.000298384	0.000239870	0.000111667	0.000142815	0.000172841
Avg. closeness	0.381809827	0.378570226	0.365302595	0.370856627	0.367129494

Table 2: Comparison of the PolyQ proteins in human and the orthologs of PolyQ proteins in mice.

Table 3 shows the comparison of the properties of the disease proteins, and the human polyQ first neighbour networks and orthologs of mouse polyQ first neighbour networks of the lenient and the strict networks, and figures 6, 5 show the networks. The average number of neighbours for the human polyQ first neighbour network is larger than that of the orthologs of mouse polyQ proteins in the lenient network, but in the strict network this is the other way around: the human polyQ first neighbour network has less average neighbours than the orthologs of mouse polyQ proteins first neighbour network. The characteristic path length of the human polyQ first neighbour network is smaller than that of the orthologs of mouse polyQ proteins of the lenient networks, while again in the strict network this is the other way around. The disease first neighbour network has a lower average amount of neighbours and characteristic path length than the lenient human polyQ first neighbour network, but a larger centralisation. There are some orthologs of mouse polyQ proteins which do not have an interaction with a human polyQ protein and vice versa.

	Disease FN	Lenient HPQ FN	Lenient MPQ FN	Strict HPQ FN	Strict MPQ FN
# nodes	799	3376	2741	1135	914
# edges	22868	141988	103472	31191	27386
Avg. # neighbours	57.242	84.116	75.499	55.204	60.444
Diameter	3	5	5	9	8
Radius	2	3	3	5	4
CPL	2.229	2.471	2.511	2.691	2.659
Clustering coef.	0.459	0.322	0.331	0.585	0.631
Density	0.072	0.025	0.028	0.049	0.067
Heterogeneity	0.751	0.965	1.009	0.993	0.937
Centralisation	0.353	0.176	0.173	0.224	0.265
# human polyQ	27	84	69	71	44
# mouse polyQ	7	46	58	37	46

Table 3: Comparison between the first neighbour (FN) networks of the disease proteins, and human polyQ proteins (HPQ) and orthologs of mouse polyQ proteins (MPQ) of the lenient and strict networks.

Table 4 shows the comparison between the human polyQ proteins network, the orthologs of mouse polyQ proteins network, and the combination of human and the orthologs of mouse polyQ proteins. The networks are shown in figures 8, 9. It can be seen that the orthologs of mouse polyQ proteins are not very well connected to each other as there are only 33 edges but 58 nodes in the network. The human polyQ proteins are much more connected with each other as there are 193 edges and 84 nodes in the network.

	Human PolyQ	Mouse PolyQ	Human & Mouse PolyQ
# nodes	84	58	106
# edges	193	33	224
Avg. # neighbours	6.508	3.059	6.147
Diameter	5	6	8
Radius	3	3	4
CPL	2.610	2.551	2.913
Clustering coef.	0.460	0.138	0.401
Density	0.112	0.191	0.087
Heterogeneity	0.904	0.649	0.986
Centralisation	0.419	0.279	0.403

Table 4: Comparison of the polyQ networks

The properties of the different polyQ disease proteins are shown in table 5. All but one of the disease proteins in the network have higher degree than the average degree of the network, lenient network 1. The disease proteins also have a higher closeness centrality and all but one have a higher betweenness centrality than the non-polyQ proteins 2.

	In network	Degree	Closeness centrality	Betweenness centrality	In cluster
Ar	No	-	-	-	-
Atn1	Yes	101	0.3948738	2.0098796E-4	-
Atxn1	Yes	75	0.4009609	9.0041892E-5	-
Atxn2	No	-	-	-	-
Atxn3	Yes	122	0.4117598	1.7214686E-4	14
Atxn7	No	-	-	-	-
Cacna1a	Yes	153	0.3932765	3.1041061E-4	-
Htt	Yes	210	0.4330940	5.8263933E-4	9
Tbp	Yes	338	0.4415538	0.0010416	-

Table 5: Comparison of the properties of the disease proteins.

Table 6 shows the different properties for the clusters. The density of the clusters shows that they are much more interconnected than the network as a whole, lenient network 1. The polyQ proteins in clusters 6, 7, and 9 have on average more interactions than the cluster as a whole. For clusters 11, 14, and 17 this is vice versa, the polyQ proteins have on average less interactions than the cluster as a whole.

	Cluster 6	Cluster 7	Cluster 9	Cluster 11	Cluster 14	Cluster 17
# nodes	631	448	577	423	327	260
# edges	8865	5582	3667	1910	843	547
Avg # neighbours	28.098	24.920	12.711	9.031	5.156	4.208
Diameter	5	6	8	7	10	13
Radius	3	4	5	4	6	7
CPL	2.710	2.802	3.270	3.424	4.618	5.548
Clustering coef.	0.554	0.656	0.445	0.463	0.407	0.374
Density	0.045	0.056	0.022	0.021	0.016	0.016
Heterogeneity	0.687	0.497	0.637	0.769	0.582	0.527
Centralisation	0.156	0.119	0.142	0.198	0.043	0.034
# human polyQ	5	8	10	8	5	5
# mouse polyQ	3	3	4	4	3	3
Avg. degree polyQ	33.833	25.375	18.5	8	4.429	4.167

Table 6: Comparison of the properties of the clusters

3.4 Functional enrichment

In this section the results of the functional enrichment done on the networks will be shown. These results consist of the GO components, GO functions, and GO processes. First the results of the first neighbour networks will be shown, then the results for the polyQ only networks, and lastly the results for the clusters.

The results of the GO components for the first neighbour network of human polyQ proteins show that the top component category is nuclear chromosome with 40.4%. In that category the nucleoplasm, nuclear lumen, and intracellular organelle lumen are the three largest components (figure 10). For the human orthologs of mouse polyQ proteins this category, nuclear chromosome, is also the largest with 72.2%. In this category the nucleoplasm, nuclear lumen, and intracellular organelle lumen are also the three largest components (figure 13). Table 7 shows two components which are not present in the mouse polyQ first neighbour network. These components are part of the transferase complex category. This category is third largest in the human polyQ first neighbour network, but it is not present in the mouse polyQ first neighbour network.

	FN human polyQ	FN mouse polyQ
Nucleoplasm	x	x
Nuclear lumen	x	x
Nucleus	x	x
Protein-containing complex	x	x
Catalytic complex	x	
Transcription regulator complex	x	

Table 7: Comparison of some of the larger catergories of GO components between the human polyQ first neighbour network and the human orthologs of mouse polyQ proteins first neighbour network.

The GO component categories for the first neighbour network of the disease proteins are (figure 16):

- Transferase complex, 29.5%
- Nuclear chromosome, 23.7%
- Nucleus, 6.5%
- Presynapse, 4.9%
- Dendrite, 4.6%

The results of the GO functions for the first neighbour network of human polyQ proteins show that most of the functions in the network are some sort of binding (figure 11). This is the same for the mouse polyQ proteins first neighbour network (figure 14). Table 8 shows that there are some differences between the GO function categories of the human polyQ first neighbour network and the mouse polyQ first neighbour network. The human polyQ first neighbour network contains the GO function category mRNA binding, while the mouse polyQ first neighbour does not have this category. The mouse polyQ first neighbour network contains the category catalytic activity, while the human polyQ first neighbour network does not have this category.

	FN human polyQ	FN mouse polyQ
mRNA binding	x	
Transcription coregulator activity	x	x
DNA binding transcription factor binding	x	x
Catalytic activity		x

Table 8: Comparison of some of the categories of GO functions between the human polyQ first neighbour network and the human orthologs of mouse polyQ proteins first neighbour network.

The GO function categories for the first neighbour network of the disease proteins are (figure 17):

- Core promoter sequence specific DNA binding, 13.3%
- DNA-directed 5'-3' RNA polymerase activity, 9.2%
- Ubiquitin-like protein ligase binding, 7.3%
- Transcription coregulator activity, 4.6%
- DNA-binding transcription factor binding, 2.9%

The results of the GO processes show that for both the human polyQ first neighbour network and the mouse polyQ first neighbour network the category transcription is the largest with 36.3% and 39.8% respectively (figures 12, 15). Table 9 shows that there are some differences again between the human polyQ first neighbour network and the mouse polyQ first neighbour network. The human polyQ first neighbour network contains the category regulation of metabolic process, while the mouse polyQ first neighbour network does not contain this category. The mouse polyQ first neighbour network contains the categories cellular response to DNA damage stimulus, and chromatin organisation which are not present in the human polyQ first neighbour network.

	FN human polyQ	FN mouse polyQ
Transcription	x	x
Negative regulation of transcription by RNA polymerase II	x	x
Regulation of metabolic process	x	
Cellular response to DNA damage stimulus		x
Chromatin organisation		x

Table 9: Comparison of the GO processes categories of the human polyQ first neighbour network and the human orthologs of mouse polyQ proteins first neighbour network.

The GO processes for the first neighbour network of the disease proteins are (figure 18):

- Negative regulation of transcription by RNA polymerase II, 24.5%
- Transcription 23.8%
- Response to organic cyclic compound, 17.9%
- Regulation of biological process, 6.5%

For the polyQ only network the top GO component category of the human polyQ network is nuclear chromosome with 38.6% (figure 19). In this category the nucleoplasm, nuclear lumen, and chromatin are the top components. Some of the other categories for the human polyQ network are transcription regulator complex (21.4%) and nucleus (12.8%). The functional enrichment for the human orthologs of mouse polyQ proteins did not return any GO components. So a comparison between the two is in this case not possible.

The results of the GO functions show that most of the functions of the human polyQ network are some sort of binding (figure 20). This is the same for the mouse polyQ network (figure 22). Table 10 shows that are some differences between the mouse polyQ network and the human polyQ network. The human polyQ network contains the category E-box binding which the mouse polyQ network does not have. The mouse polyQ network has the category sequence-specifice DNA binding which the human polyQ network does not have.

	Human polyQ	Mouse polyQ
E-box binding	x	
Transcription regulator activity	x	x
DNA-binding transcription factor binding	x	
Sequence-specific DNA binding		x

Table 10: Comparison of the GO functions of the human polyQ network and the human orthologs of mouse polyQ proteins network.

The top GO process category for the human polyQ network is positive regulation of Notch signaling pathway with 50.0% (figure 21). For the mouse polyQ network the top category is positive regulation of muscle contraction with 59.7%. Table 11 shows that there are some differences between the GO processes of the human polyQ network and the mouse polyQ network. The transcription category is the only category which both the human polyQ network and the mouse polyQ network have. The rest of the categories are all different from each other.

	Human polyQ	Mouse polyQ
Positive regulation of Notch signaling pathway	x	
Postive regulation of muscle contraction		x
Transcription	x	x
Neural nucleus development	x	
Chromatin organisation	x	
Cardiac muscle cell differentiation		x

Table 11: Comparison of the GO processes of the human polyQ network and the human orthologs of mouse polyQ proteins network.

Table 12 shows some of the GO components in the clusters. The GO components of the clusters show that the clusters have many different components, but that there is also a lot of similarities between certain clusters. Clusters 9 and 11 contain multiple of the same components while they share less of the same components with clusters 6 and 7 (figures 30, 33, 24, 27). Clusters 14 and 17 also have similar components (figures 36, 39).

	Cl. 6	Cl. 7	Cl. 9	Cl. 11	Cl. 14	Cl. 17
Protein DNA-complex	x		x			
Nuclear chromosome	x	x				
Postsynapse			x	x		
Cytoplasm	x		x	x	x	x
Membrane		x	x	x	x	x
Axon			x	x		

Table 12: Comparison of the different GO components in the clusters.

Table 13 shows the some of the GO functions in the clusters. Unfortunately, the functional enrichment of cluster 17 did not return any GO functions, so this clusters cannot be compared to the other clusters. The table shows that, just like the other networks, the clusters almost all contain binding. Cluster 14 is the only cluster which does not contain binding (figure 37). Clusters 9 and 11 show similarity the GO functions (figures 31, 34), just like with the GO components. In this case, however, the are also similar to clusters 6 and 7 (figures 25, 28), which was not the case with the GO components.

	Cl. 6	Cl. 7	Cl. 9	Cl. 11	Cl. 14
Binding	x	x	x	x	
Protein binding	x	x	x	x	
DNA binding	x	x			
Transferase activity	x		x	x	x
Catalytic activity	x	x			x
Kinase activity	x		x	x	x
Hydrolase activity	x	x		x	
Phosphatase activity			x	x	

Table 13: Comparison of the different GO functions in the clusters.

The GO processes are quite different for the different clusters, they only have the more general processes in common. This is why there is no comparison table for the clusters here, instead a list with the larger percentage processes will be given for the different clusters.

Cluster 6 (figure 26):

- Innate immune response 24.4%
- Autophagy 21.1%
- Intraciliary transport involved in cilium assembly 17.1%

Cluster 7 (figure 29):

- Histone deacetylation 34.7%
- Antigen processing and presentation 11.2%
- Oxidative phosphorylation 10.1%

Cluster 9 (figure 32):

- Chromatin organisation 18.9%
- Regulation of TOR signaling 10.8%
- Negative chemotaxis 9.8%

Cluster 11 (figure 35):

- Cellular component morphogenesis 25.7%
- Cell surface receptor signaling pathway involved in cell-cell signaling 19.0%
- Positive regulation of synaptic transmission 16.1%

Cluster 14 (figure 38):

- Acylglycerol metabolic process 51.2%
- Transition metal ion transport 8.8%
- Cellular transition metal ion homeostasis 6.5%

Cluster 17 only has one category which is cellular lipid catabolic process 57.4% (figure 40).

4 Discussion

The two (the strict and the lenient network) have shown that human brain proteins engage in many interactions. However, what the networks have also shown is that the strictness of the rules for data collection and network creation can have a massive impact on the amount of proteins shown in the network and the amount of interactions shown in the network. This shows that research on protein-protein interaction networks is difficult, as a small change in the search parameters or network creation parameters, can dramatically increase or decrease the size of the network. This shows that balancing the parameters in the network creation process is difficult as one does not want to miss out on proteins, but one also does not want to have many false positive interactions. Getting this balance wrong may lead to bias, as has already been stated by [HPRL08].

For the strict network almost all proteins in the network are contained in the giant component. This is different from the strict network where more than 10% of the proteins in the network are not contained in the giant component. For both networks the degree distribution indicates that they are scale-free. This is important as some conclusion and analyses done in this research rely on this fact. The lenient network was chosen as the main focus for the most part of the research, this was due to the fact that the lenient network contained more human polyQ proteins, more proteins and more interactions.

The clustering has shown that most clusters did not contain many human polyQ proteins. The ones with four or more human polyQ proteins were used for further analysis. Of those clusters only 2 contained disease polyQ proteins. The fact that there are only a few polyQ proteins per clusters makes it more difficult to come to any conclusion based on the results of the clustering. It would have been better if more polyQ proteins were found in a single network. That this did not happen may indicate that polyQ proteins are involved in many different processes and pathways.

The network analysis of the polyQ proteins in the lenient network has shown that polyQ proteins are on average more connected than non-polyQ proteins. This shown by the higher average degree, higher average betweenness centrality, and higher average closeness centrality of the polyQ proteins. This indicates that polyQ proteins play an important role in certain biological processes.

The analysis of the first neighbour networks has shown that the inter connectivity of the proteins connected to the polyQ proteins is higher than that of the average of the other proteins in the human brain. This is shown by the higher network density of the first neighbour networks in comparison to the lenient and strict network. This higher density in the first neighbour network may again indicate that polyQ proteins play an important role in protein-protein interactions in the human brain.

The analysis of the polyQ only networks has shown that the human polyQ proteins are more interconnected than the orthologs of mouse polyQ proteins. This is shown by the higher amount of edges in the human polyQ network compared to the mouse polyQ network. However, there are polyQ proteins in both the human polyQ network as the orthologs of mouse polyQ network which have no edges to any other polyQ protein.

Almost all the disease proteins have shown a higher degree, higher betweenness centrality, and

higher closeness centrality than the average of all the polyQ proteins. This indicates that the disease proteins play an even more important role in the network than the regular polyQ proteins. However, as stated before, protein-protein interaction network contain biases. For the case of the disease proteins this bias is quite clear: research focuses more on these disease proteins than on other polyQ proteins. This results in more knowledge and more certainty about the knowledge than for the regular polyQ proteins. In the future the differences in properties in the network of the disease proteins and the regular polyQ proteins may disappear or strengthen due to more knowledge being available.

The analysis of the disease problem raised an issue. One of the proteins, *Atxn2*, is available in the HumanMine database and the StringApp database. It also satisfies all the rules set in the data gathering process and the network creation process. This means that *Atxn2* should be in the network, even if it does not have interactions with other proteins as a node without edges, but it is not. This raises the question as to how many other proteins, that fit the set rules, get lost in the process of gathering the data and creating the network.

The results of the functional enrichment have shown that there are not many differences in GO components, GO functions, or GO processes between the human polyQ proteins and the orthologs of mouse polyQ proteins. This is in line with what was expected as the results of the thesis on mouse polyQ proteins showed the same behaviour. The functional enrichment of the disease proteins also shows these similarities, indicating that the cellular components, molecular functions and biological processes for regular polyQ proteins and disease polyQ proteins in largely the same. However, the polyQ proteins results are influenced by the disease polyQ proteins as they are in the network too. This again leads back to the bias issue raised before. The disease proteins are more studied and thus more is know about these polyQ proteins than other polyQ proteins.

The functional enrichment of the clusters has shown, as expected, many differences between the clusters. Clusters 9 and 14 are of more interest due to them containing a disease protein. The results of cluster 9, which contains *huntingtin* have shown involvement in response to several different kinds of stimuli. With components including chloride ion channels, axons, and presynapses. This could indicate why certain symptoms of Huntington's disease, like chorea, happen. As the pathways associated with these symptoms might not function properly when *huntingtin* is mutated.

5 Conclusion

To research the goal of this thesis, which was to investigate the differences and similarities between polyQ and non-polyQ proteins in the human brain and to compare these results with results from the previous work done in mouse, three research questions were asked. Research question 1 was answered by the functional enrichment. Which has shown that polyQ proteins are involved in the binding of different types of proteins and other molecules, with the main category being different types of DNA binding. It has been shown that the polyQ proteins are involved in many different processes, but are mainly found in the nucleoplasm and cytoplasm. To answer research question 1.1, the functions of the polyQ proteins in humans do not differ much in function from that of the orthologs of human polyQ proteins in mouse. However there are differences between the mouse polyQ proteins, as they were found to be on average less connected to other proteins [Jan20] which is in contrast with the results for humans. This indicates that research, when done in mouse, should focus on looking at the orthologs of human polyQ proteins and not at the mouse polyQ proteins. As the mouse polyQ proteins are involved in different biological pathways and have different properties in the protein-protein interaction network. To answer the final research question, research question 2, there are no signs of any major differences between regular polyQ proteins in function or involvement in biological process. However, there are signs that the disease polyQ proteins are more connected and more important in the protein-protein interaction network than regular polyQ proteins.

The amount of available data on proteins expressed in the human brain and on protein-protein interaction is unfortunately still limited. The need for more research into protein-protein interactions in general is high. However, research into this is difficult, as live humans need to be studied. This means that it will most likely take some time before research like this thesis is more reliable with less bias in the data and the results. For now the best recommendation is to do this research on polyQ proteins for more different organisms, especially organisms on which more protein-protein interaction data is available, to compare the results with the results of this thesis and with those of the mouse thesis. [Jan20] This could give more insight into the general workings of polyQ regions in proteins and why the elongation can be so devastating.

References

[ABB+00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[Alb05] Reka Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957, 2005.

[BH03] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC bioinformatics*, 4(1):1–27, 2003.

[DCSG+06] Edouard De Castro, Christian JA Sigrist, Alexandre Gattiker, Virginie Bulliard, Petra S Langendijk-Genevaux, Elisabeth Gasteiger, Amos Bairoch, and Nicolas Hulo. Scanprosite: detection of prosite signature matches and prorule-associated functional and structural residues in proteins. *Nucleic acids research*, 34(suppl_2):W362–W365, 2006.

[FHC+14] Hueng-Chuen Fan, Li-Ing Ho, Ching-Shiang Chi, Shyi-Jou Chen, Giia-Sheun Peng, Tzu-Min Chan, Shinn-Zong Lin, and Horng-Jyh Harn. Polyglutamine (polyq) diseases: genetics to treatments. *Cell transplantation*, 23(4-5):441–458, 2014.

[Fin11] Steven Finkbeiner. Huntington's disease. *Cold Spring Harbor perspectives in biology*, 3(6):a007476, 2011.

[gen21] The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.

[HPRL08] Luke Hakes, John W Pinney, David L Robertson, and Simon C Lovell. Protein-protein interaction networks and biology—what's the connection? *Nature biotechnology*, 26(1):69–72, 2008.

[HSW08] Sarah Hands, Christopher Sinadinos, and Andreas Wyttenbach. Polyglutamine gene function and dysfunction in the ageing brain. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1779(8):507–521, 2008.

[Jan20] L.M.I. Janssens. A network analysis approach to studying polyq protein interactions. *LIACS*, 2020.

[KKH+11] Rhoda J Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, et al. Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database*, 2011, 2011.

[KLB+14] Alex Kalderimis, Rachel Lyne, Daniela Butano, Sergio Contrino, Mike Lyne, Joshua Heimbach, Fengyuan Hu, Richard Smith, Radek Štěpán, Julie Sullivan, et al. Intermine: extensive web services for modern biology. *Nucleic acids research*, 42(W1):W468–W472, 2014.

[KLS+19] Irina Kuznetsova, Artur Lugmayr, Stefan J Siira, Oliver Rackham, and Aleksandra Filipovska. Cirgo: an alternative circular way of visualising gene ontology terms. *BMC bioinformatics*, 20(1):1–7, 2019.

[SAB+12] Richard N Smith, Jelena Aleksic, Daniela Butano, Adrian Carr, Sergio Contrino, Fengyuan Hu, Mike Lyne, Rachel Lyne, Alex Kalderimis, Kim Rutherford, et al. Intermine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinformatics*, 28(23):3163–3165, 2012.

[SBŠŠ11] Fran Supek, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PloS one*, 6(7):e21800, 2011.

[SGL+19] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

[SMO+03] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[TANM17] Franziska Totzeck, Miguel A Andrade-Navarro, and Pablo Mier. The protein structure context of polyq regions. *PLoS One*, 12(1):e0170801, 2017.

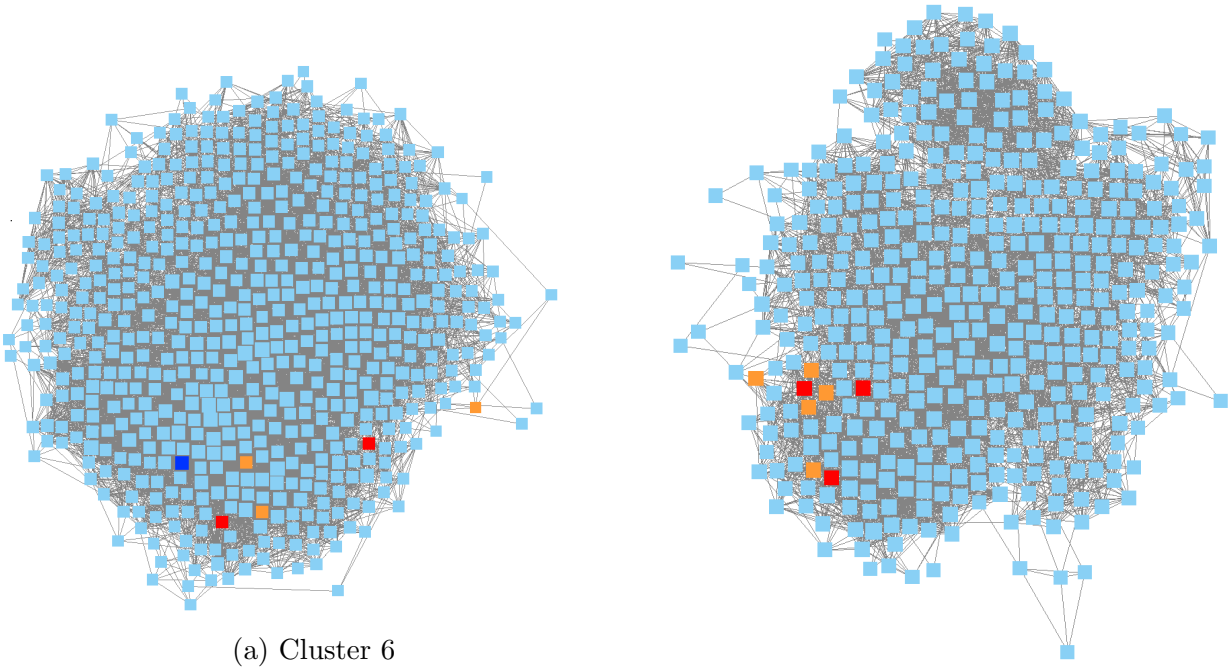
[Uni21] UniProt. Uniprot: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2021.

[VFMV03] Alexei Vázquez, Alessandro Flammini, Amos Maritan, and Alessandro Vespignani. Modeling of protein interaction networks. *Complexus*, 1(1):38–44, 2003.

[Wal07] Francis O Walker. Huntington's disease. *The Lancet*, 369(9557):218–228, 2007.

[YKS+07] Haiyuan Yu, Philip M Kim, Emmett Sprecher, Valery Trifonov, and Mark Gerstein. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol*, 3(4):e59, 2007.

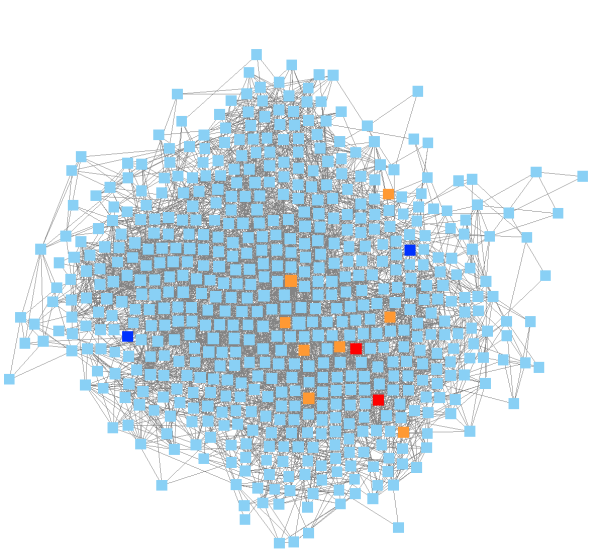
6 Appendix



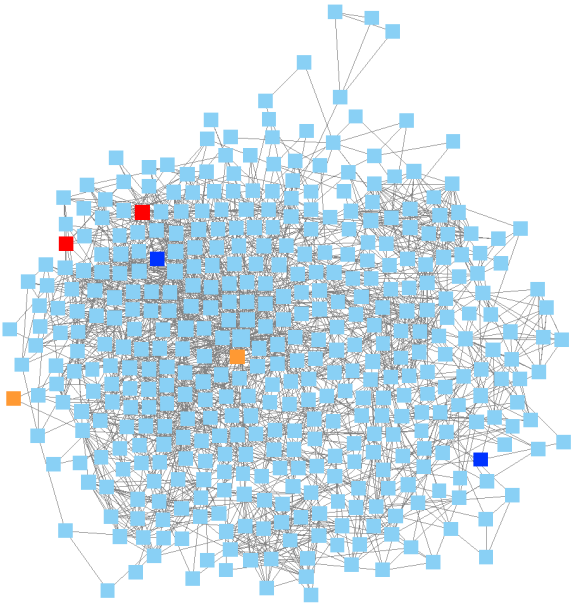
(a) Cluster 6

(b) Cluster 7

Figure 2: Clusters 6 & 7

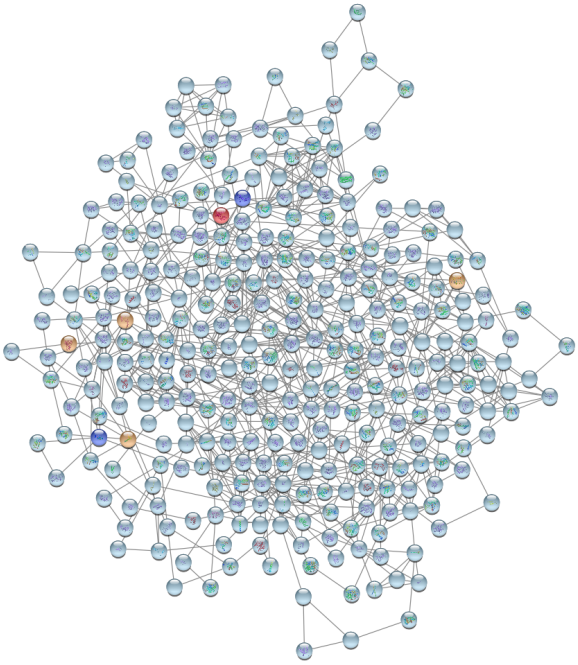


(a) Cluster 9

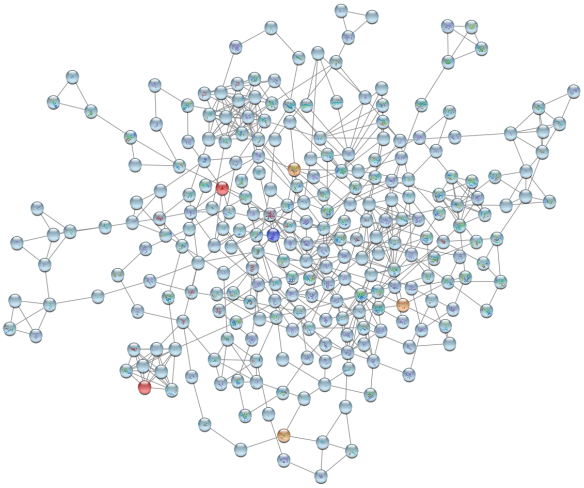


(b) Cluster 11

Figure 3: Clusters 9 & 11

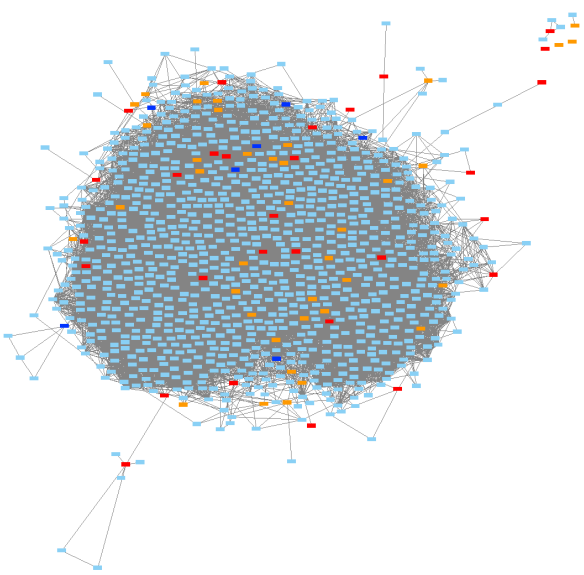


(a) Cluster 14

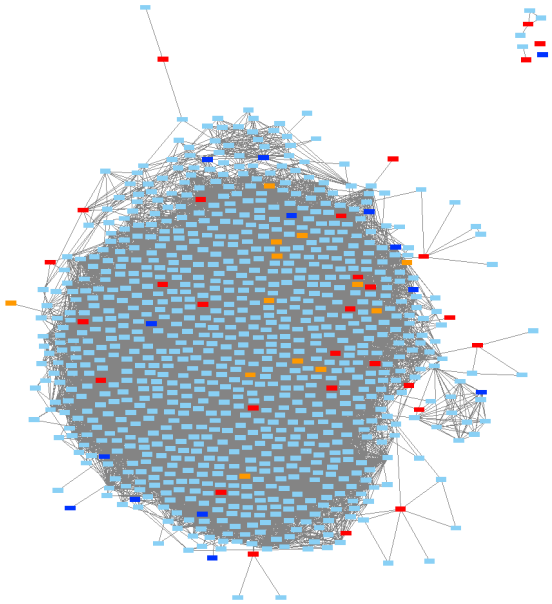


(b) Cluster 17

Figure 4: Clusters 14 & 17

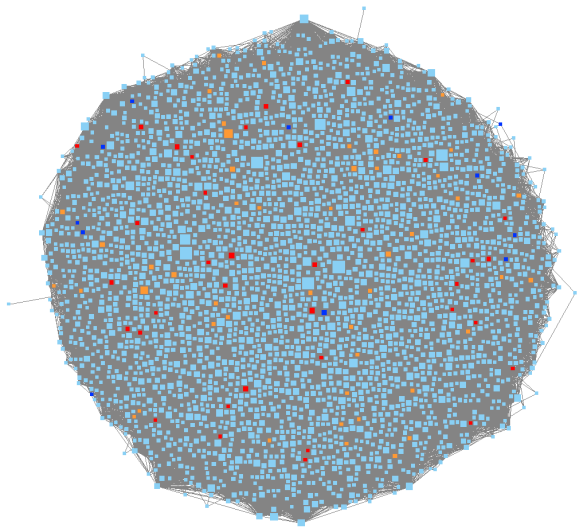


(a) The first neighbour network for proteins polyQ in human.

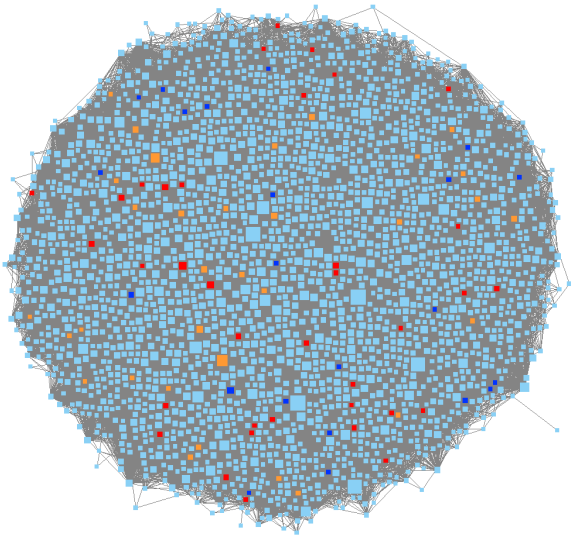


(b) The first neighbour network for human orthologs of proteins polyQ in mouse

Figure 5: First neighbour networks for polyQ proteins (created from the strict network).



(a) The first neighbour network for proteins polyQ in human.



(b) The first neighbour network for human orthologs of proteins polyQ in mouse

Figure 6: First neighbour networks for polyQ proteins (created from the lenient network).

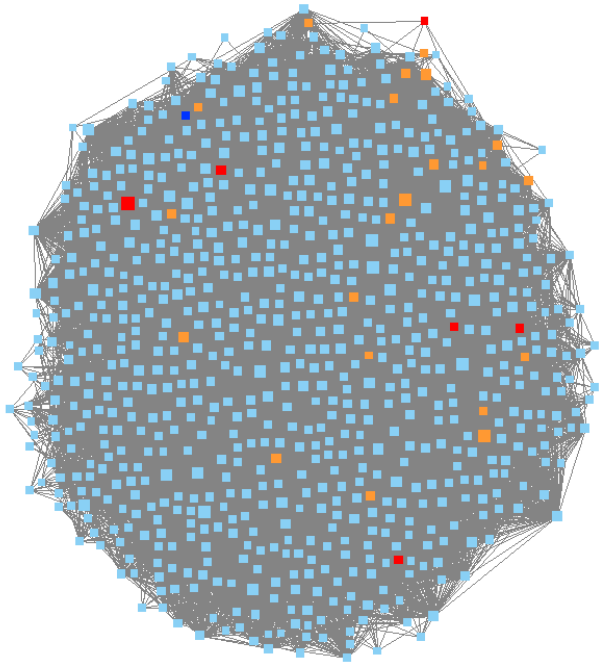
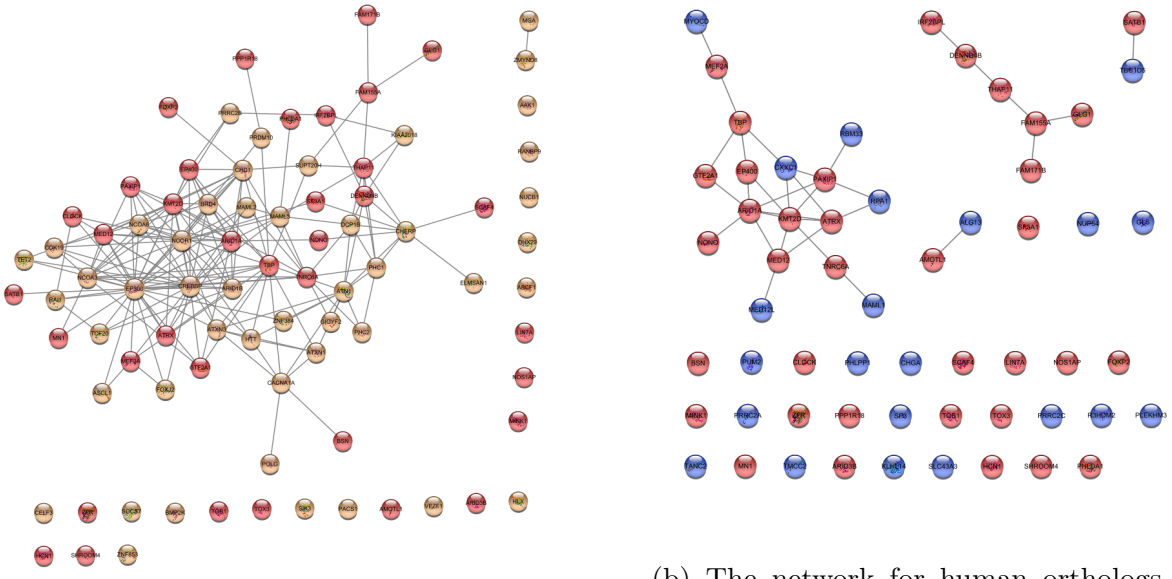


Figure 7: The first neighbour network for the disease proteins (created from the lenient network).



(a) The network for proteins polyQ in human.

(b) The network for human orthologs of proteins polyQ in mouse

Figure 8: PolyQ only networks (created from the lenient network).

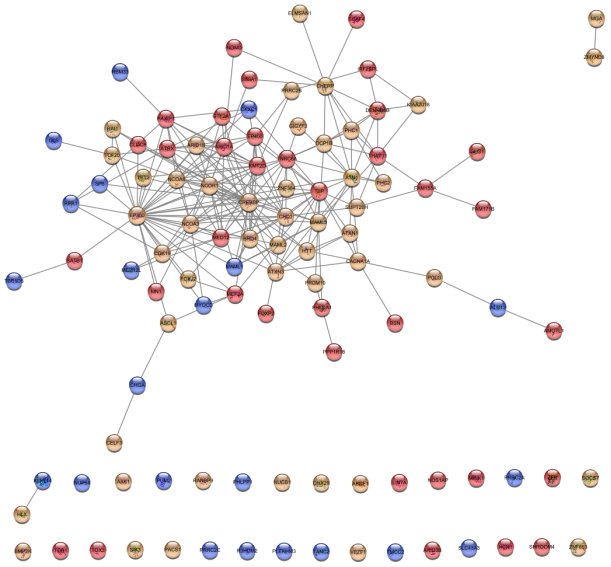


Figure 9: PolyQ only network for human polyQ proteins and human orthologs of polyQ proteins in mouse (created from the lenient network).

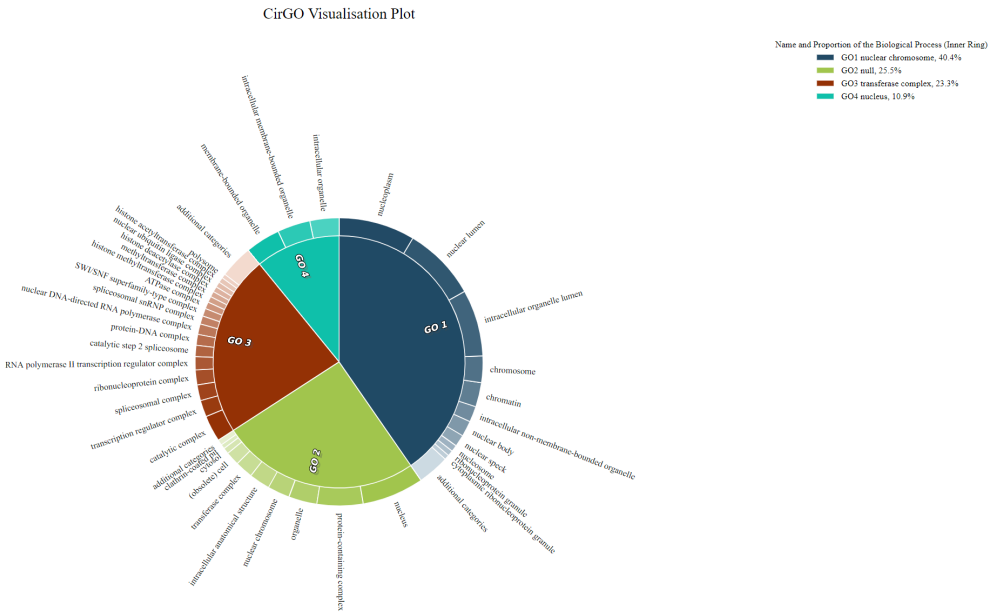


Figure 10: The GO components of the first neighbour network of human polyQ proteins (created from the strict network).

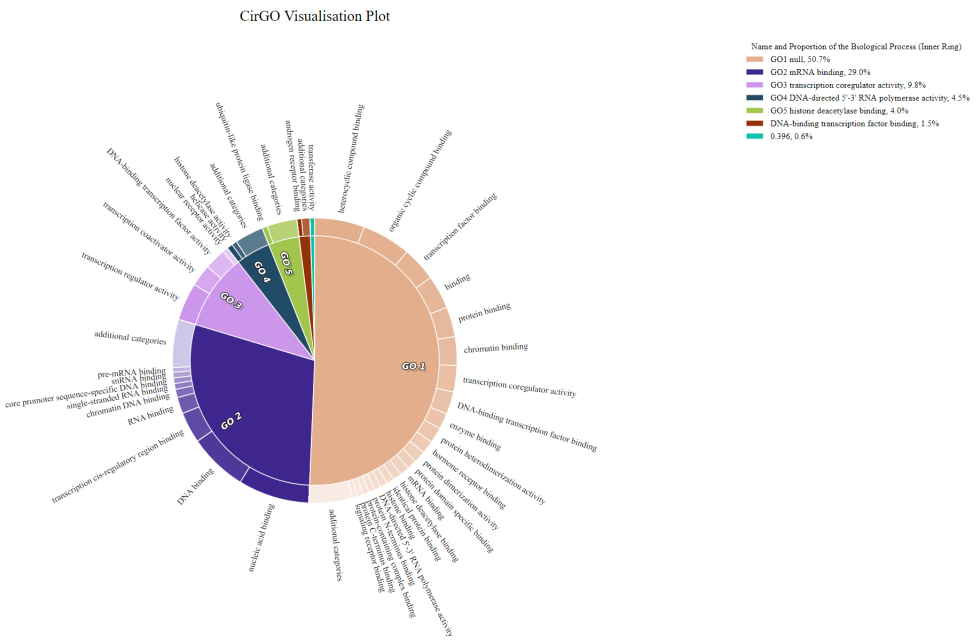


Figure 11: The GO functions of the first neighbour network of human polyQ proteins (created from the strict network).

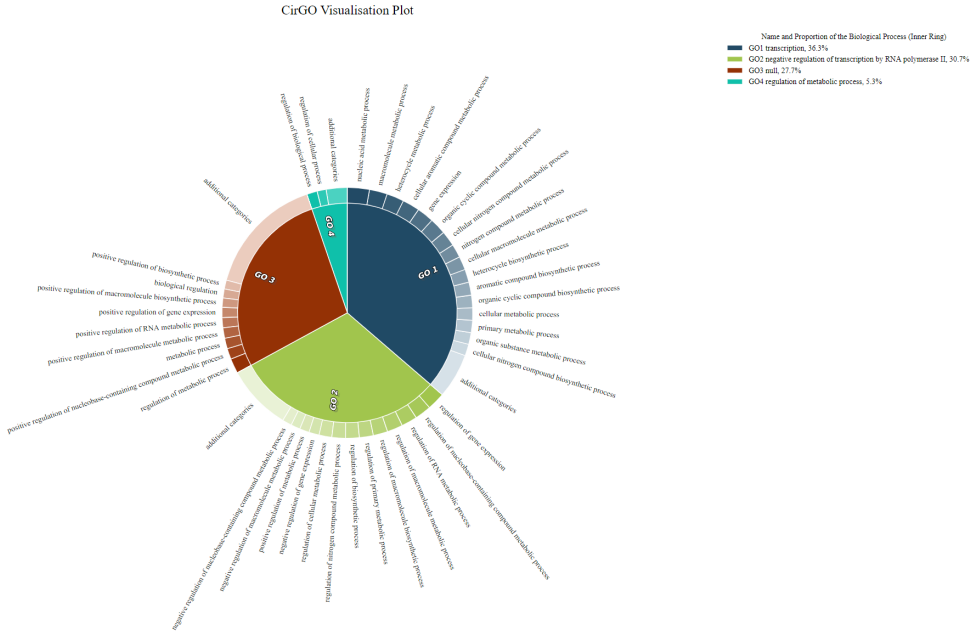


Figure 12: The GO processes of the first neighbour network of human polyQ proteins (created from the strict network).

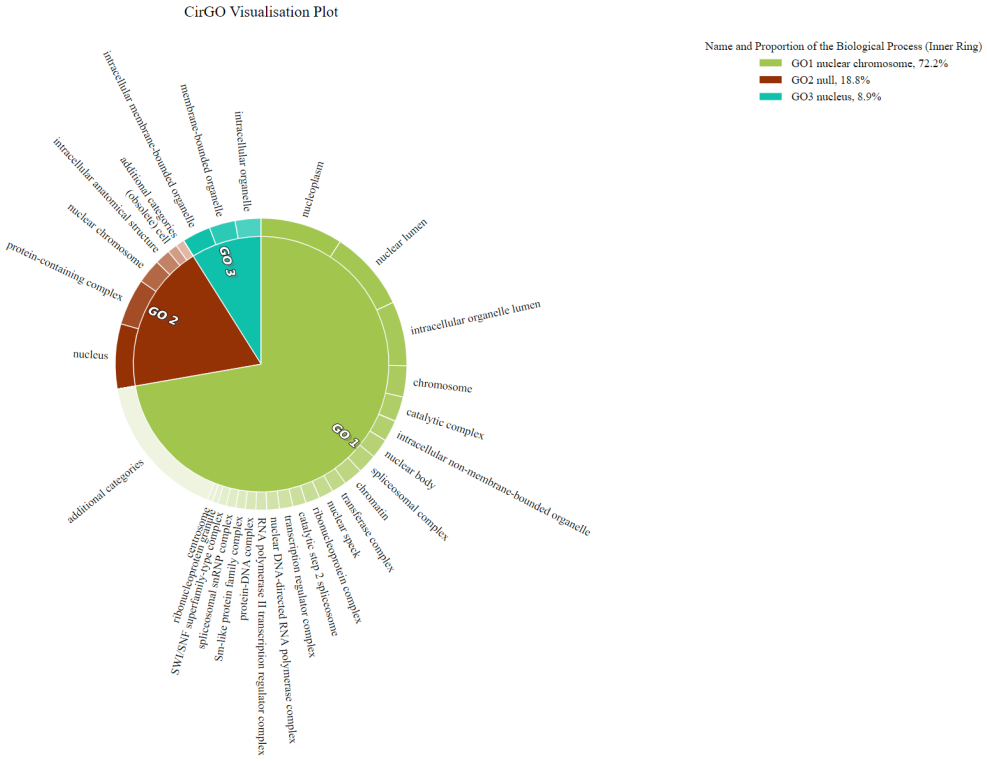


Figure 13: The GO components of the first neighbour network of human orthologs of mouse polyQ proteins (created from the strict network).

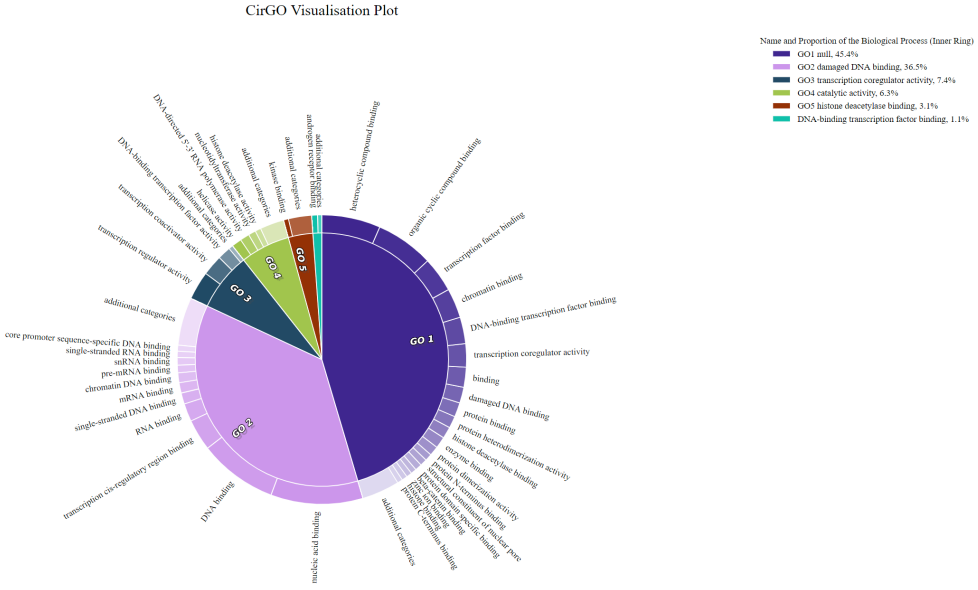


Figure 14: The GO functions of the first neighbour network of human orthologs of mouse polyQ proteins (created from the strict network).

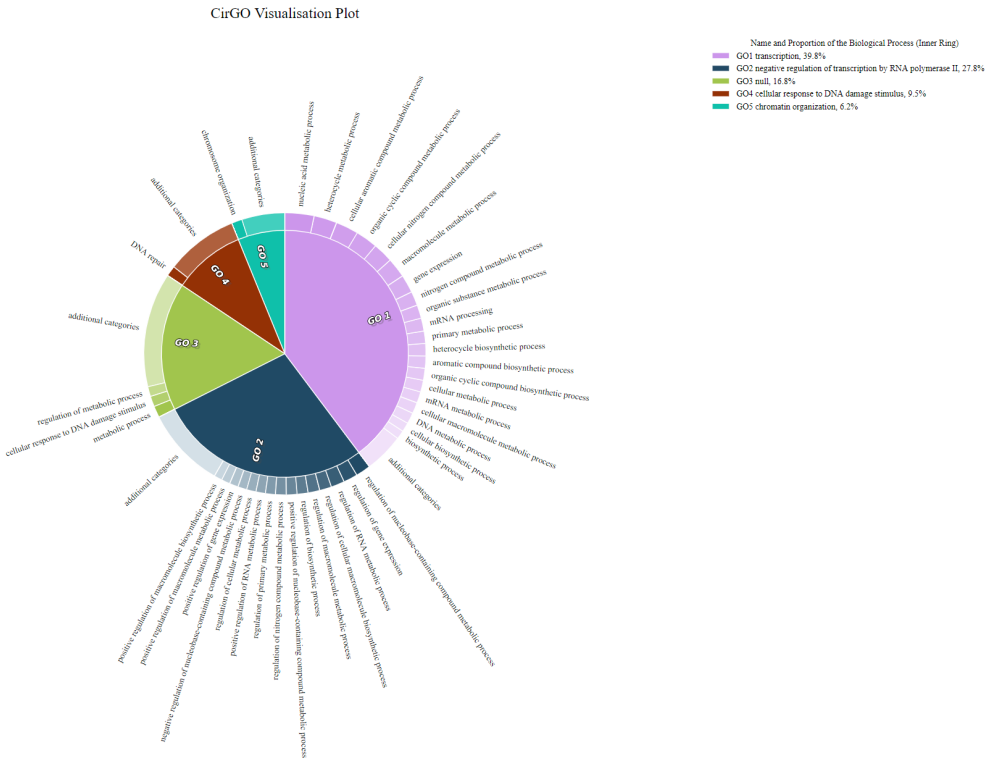


Figure 15: The GO processes of the first neighbour network of human orthologs of mouse polyQ proteins (created from the strict network).

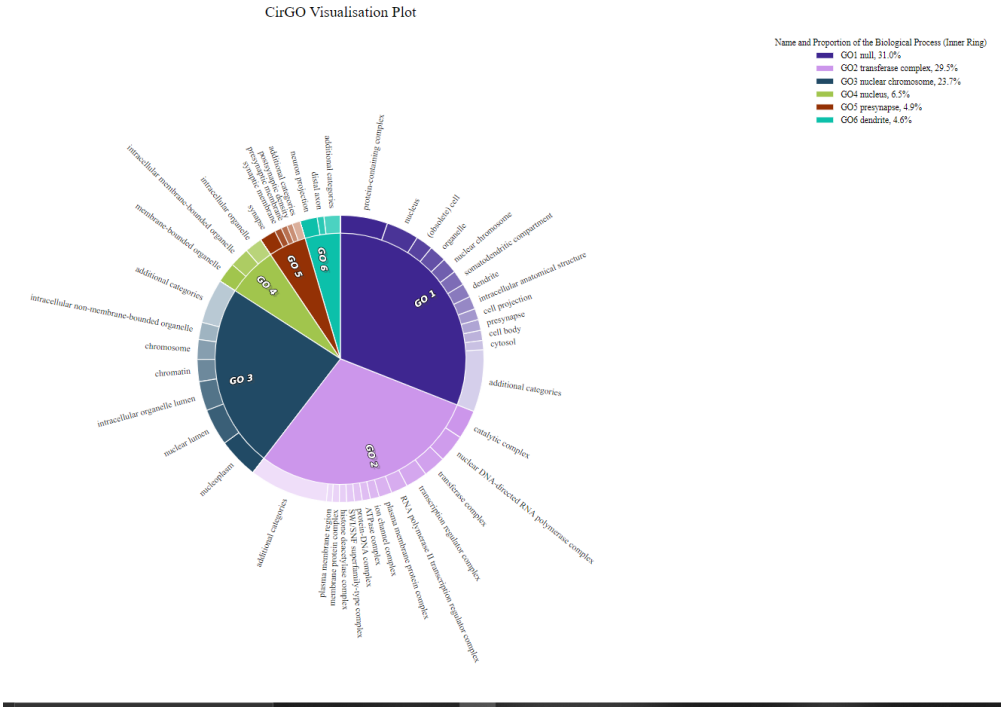


Figure 16: The GO components of the first neighbour network of disease proteins (created from the lenient network)

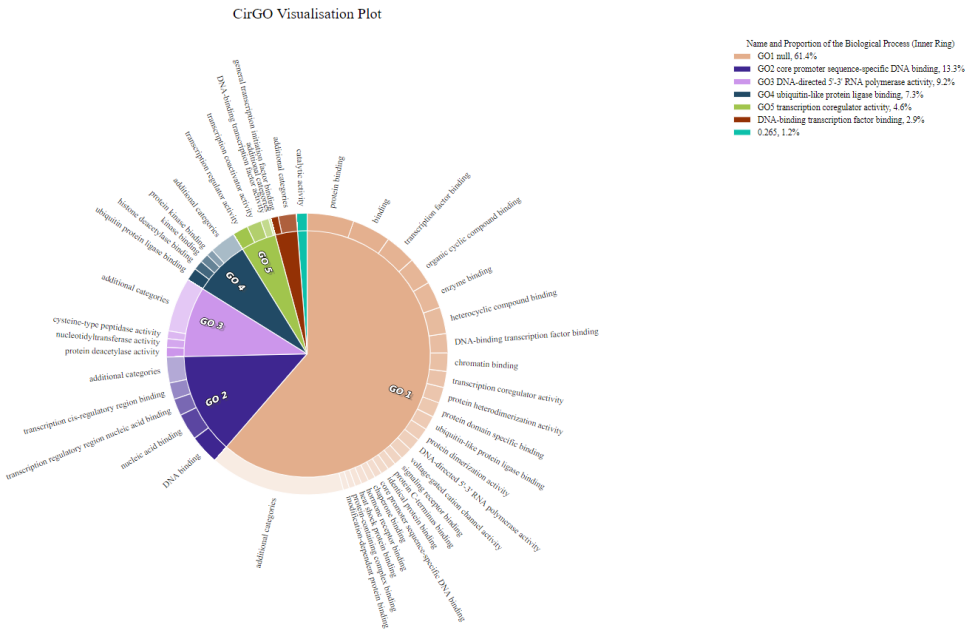


Figure 17: The GO funtions of the first neighbour network of disease proteins (created from the lenient network).

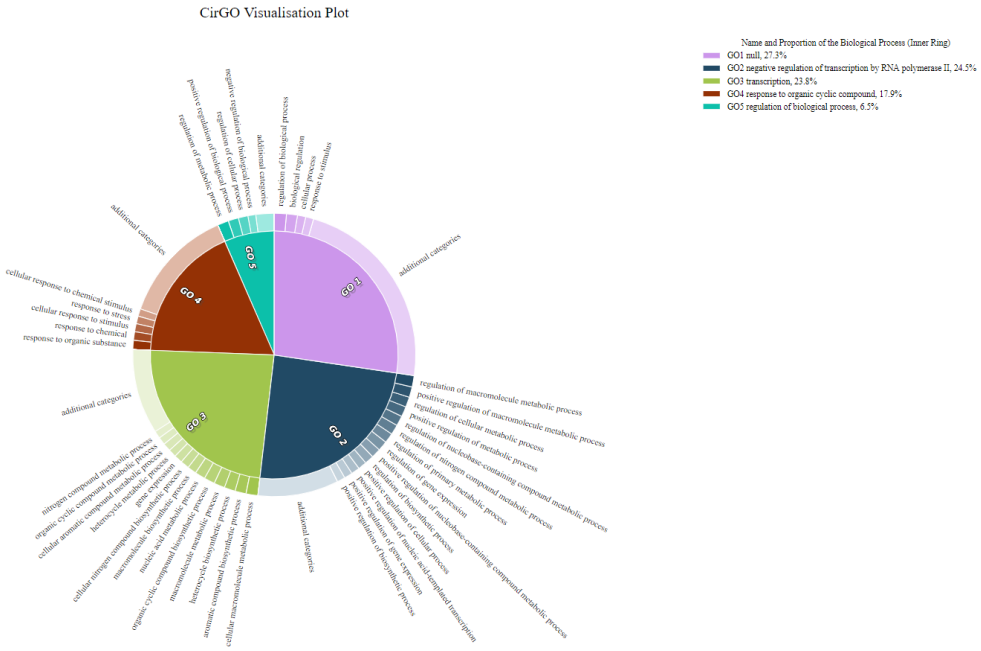


Figure 18: The GO processes of the first neighbour network of disease proteins (created form the lenient network).

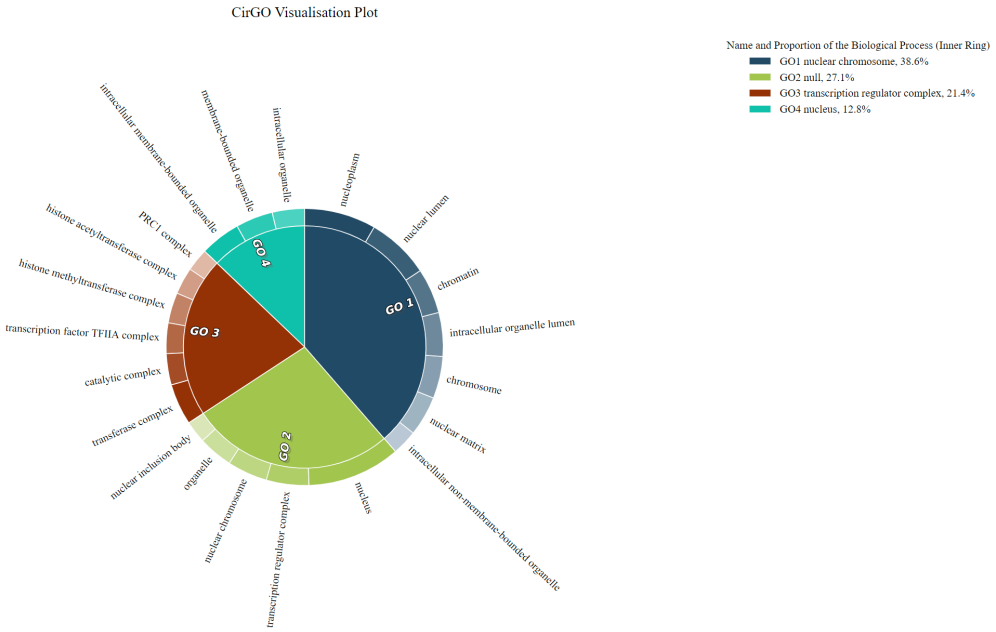


Figure 19: The GO components of the human polyQ network (created from the lenient network).

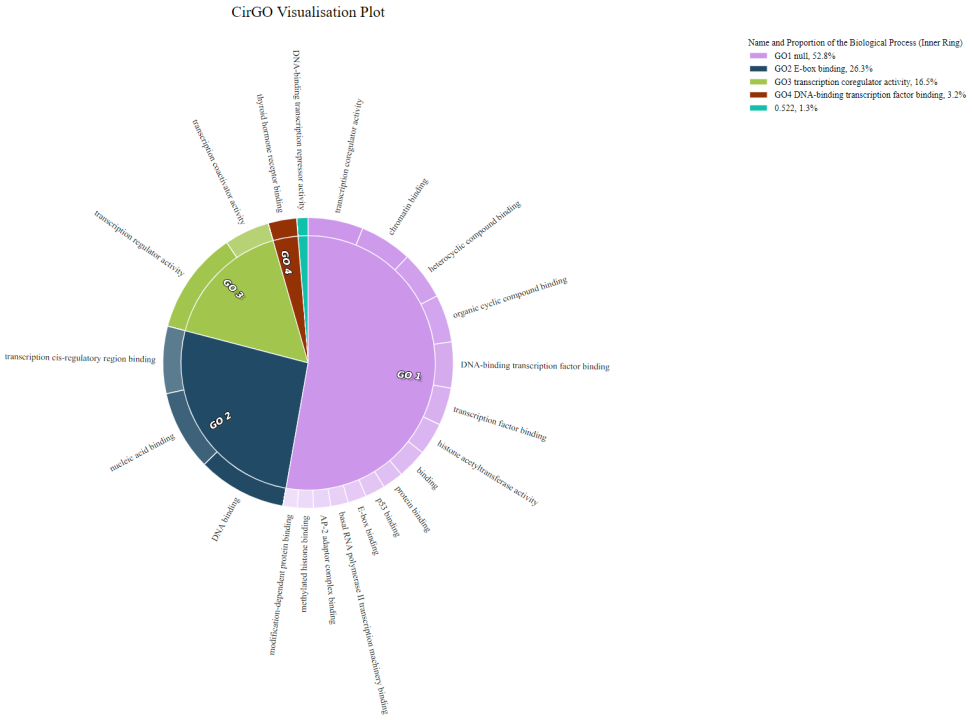


Figure 20: The GO functions of the human polyQ network (created from the lenient network).

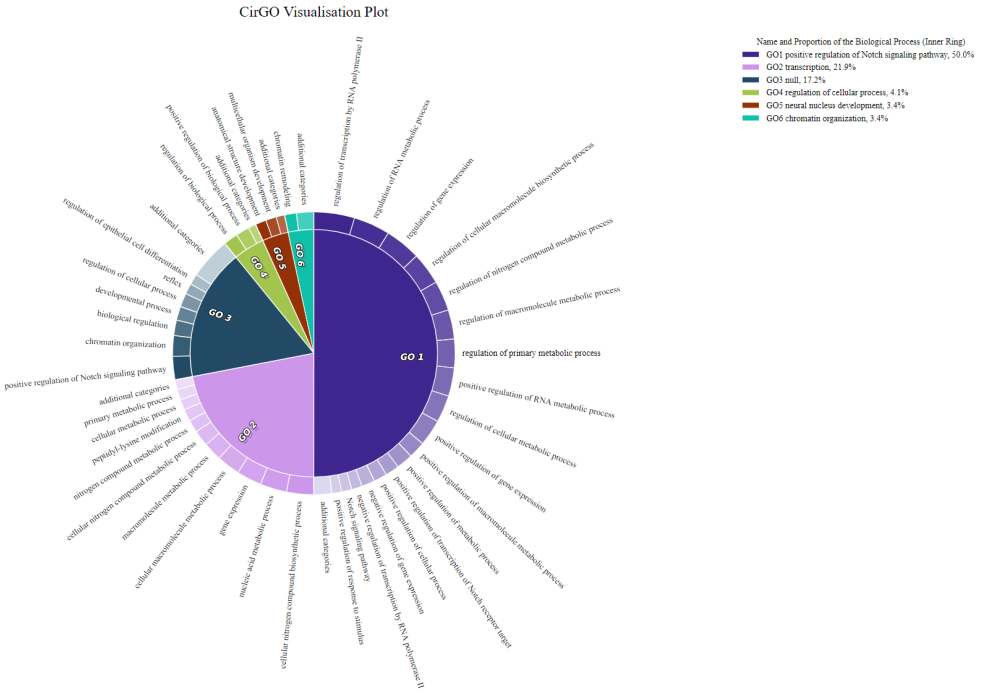


Figure 21: The GO processes of the human polyQ network (created from the lenient network).

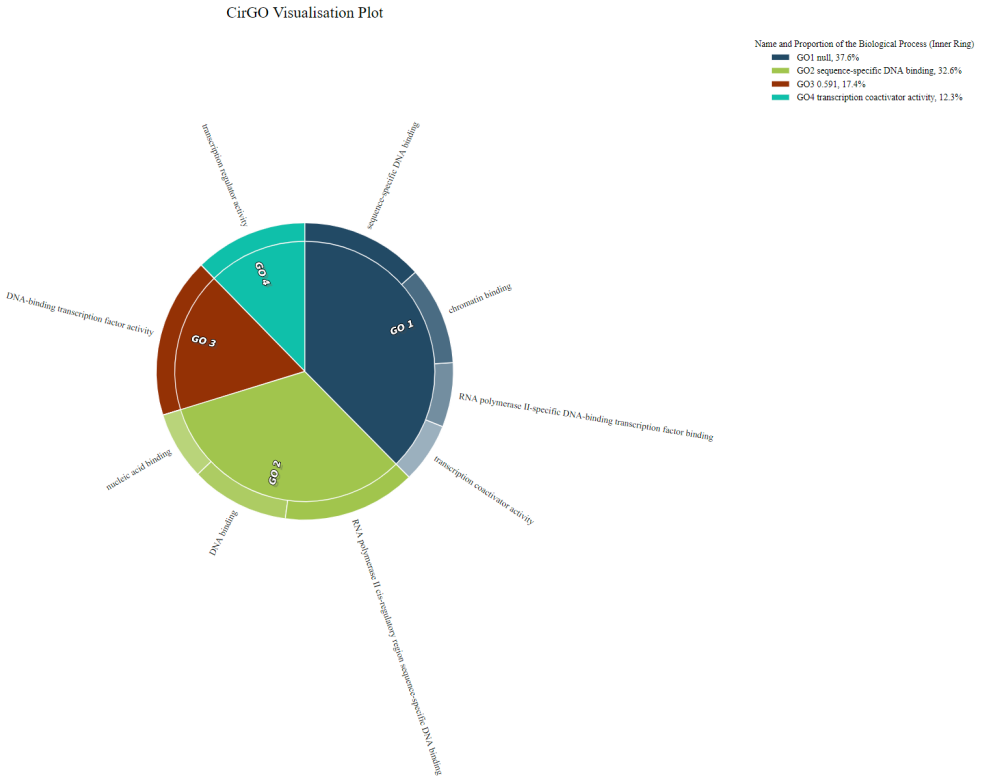


Figure 22: The GO functions of the human orthologs of mouse polyQ proteins network (created from the lenient network).

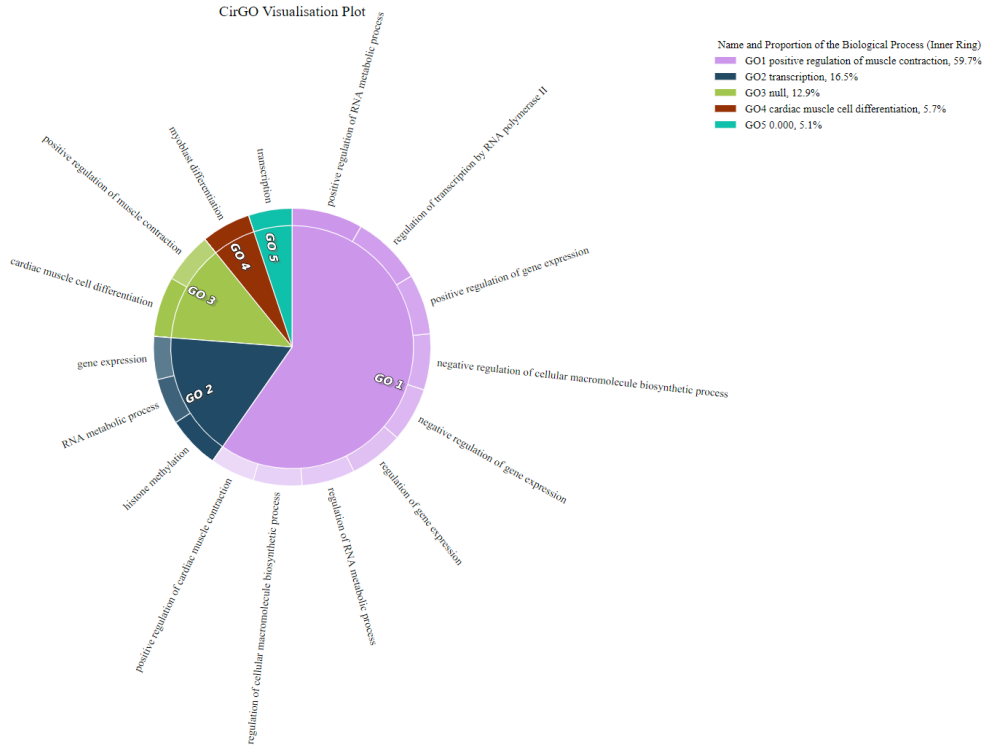


Figure 23: The GO processes of the human orthologs of mouse polyQ proteins network (created from the lenient network).

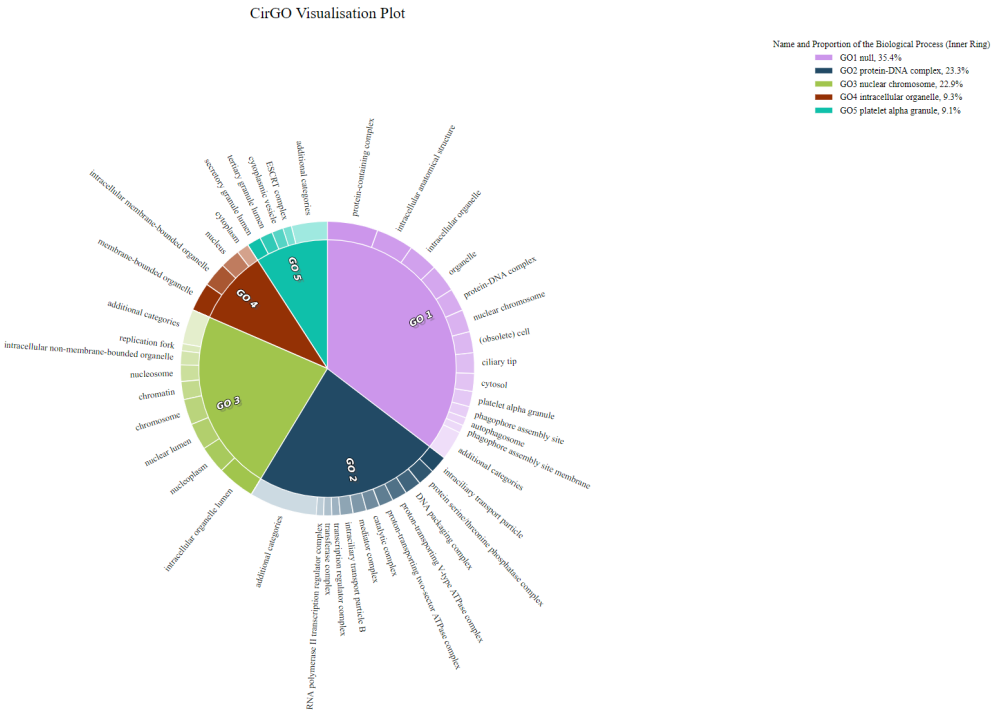


Figure 24: The GO componennts of cluster 6.

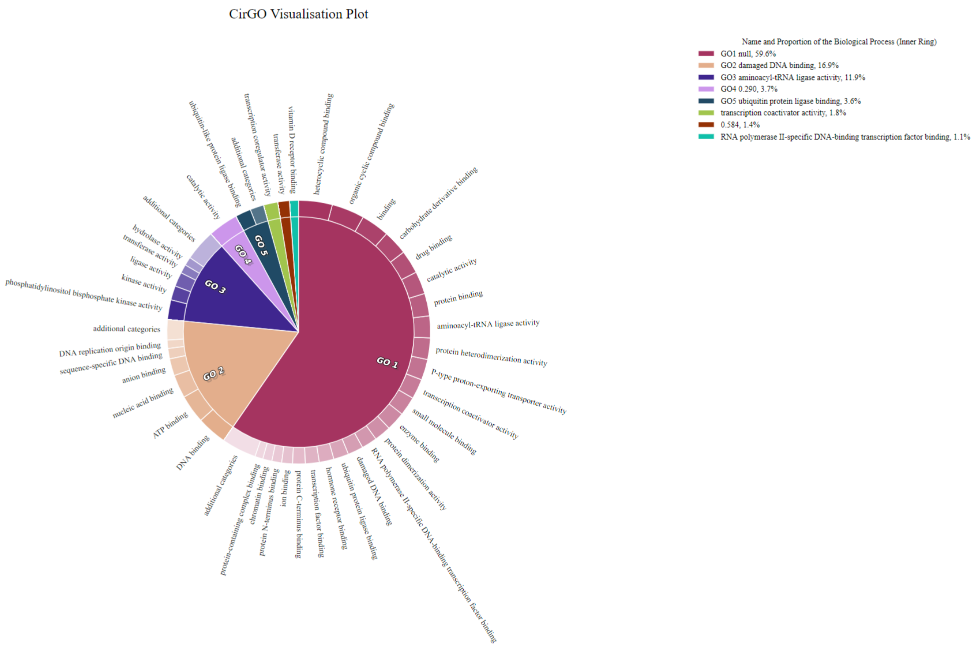


Figure 25: The GO functions of cluster 6.

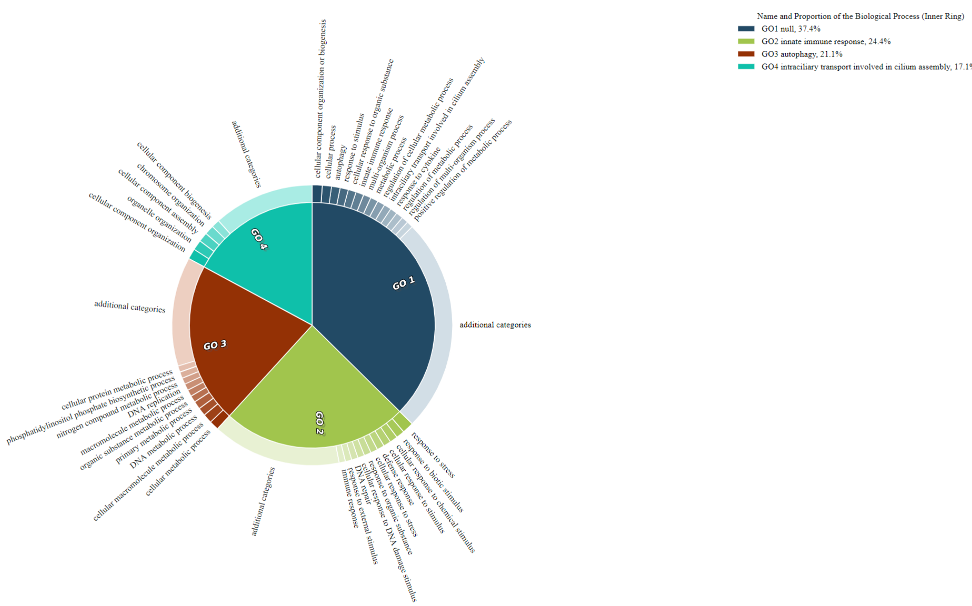


Figure 26: The GO processes of cluster 6.

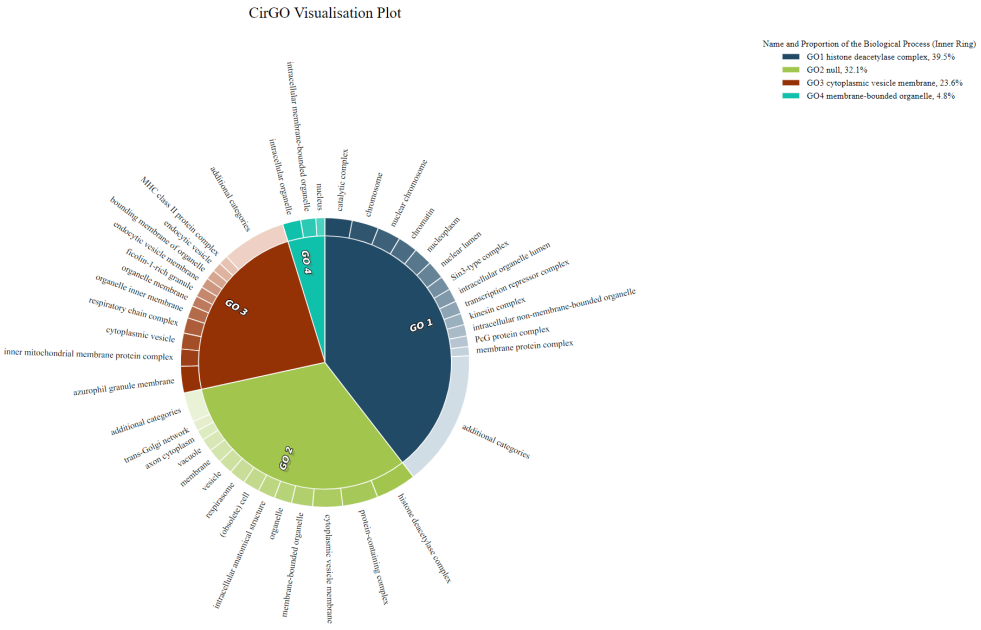


Figure 27: The GO components of cluster 7.

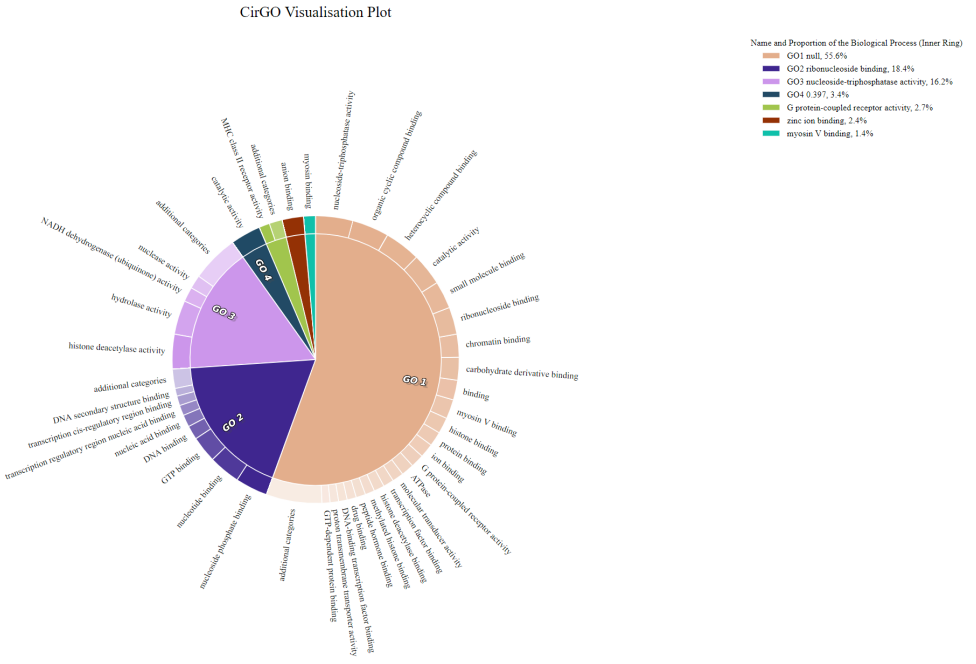


Figure 28: The GO functions of cluster 7.

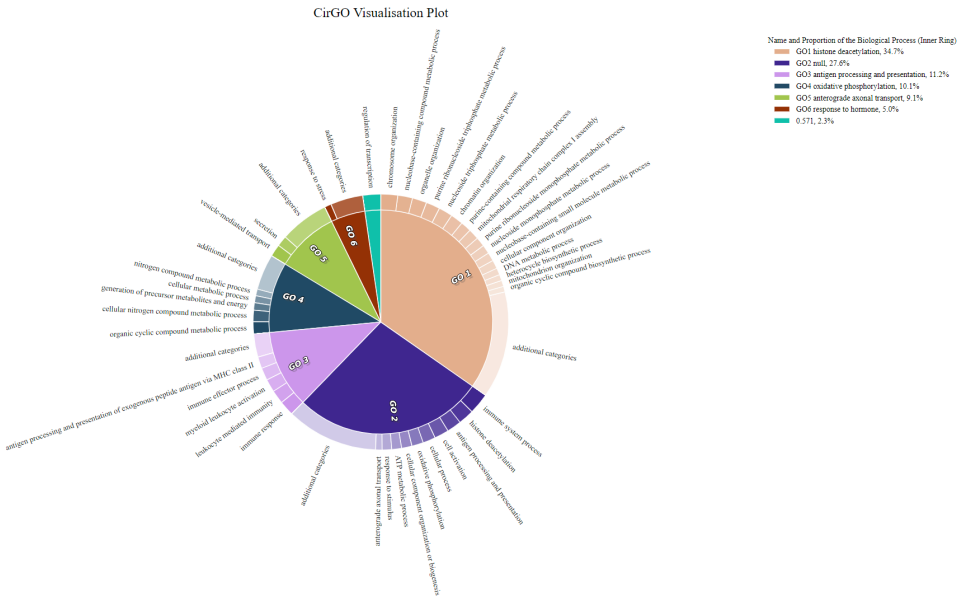


Figure 29: The GO processes of cluster 7.

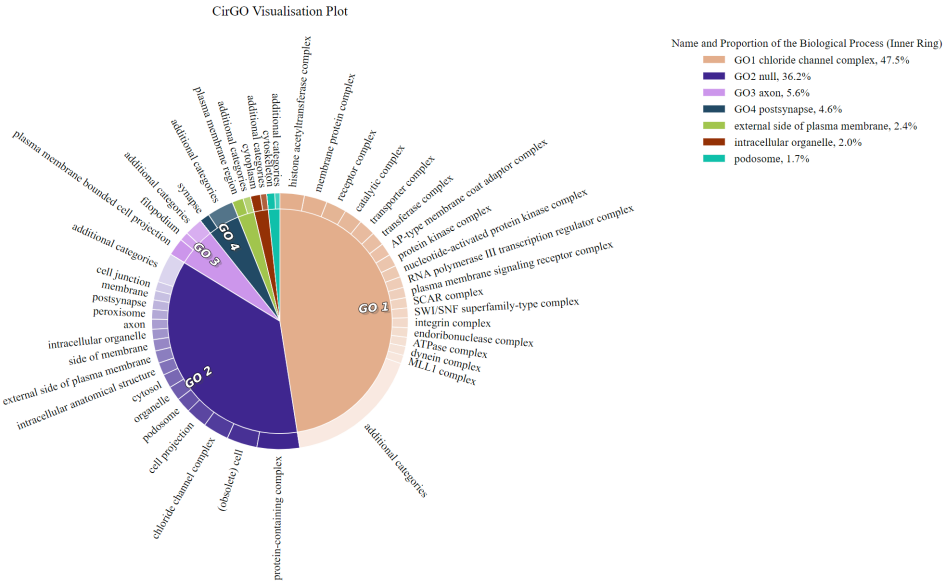


Figure 30: The GO components of cluster 9.

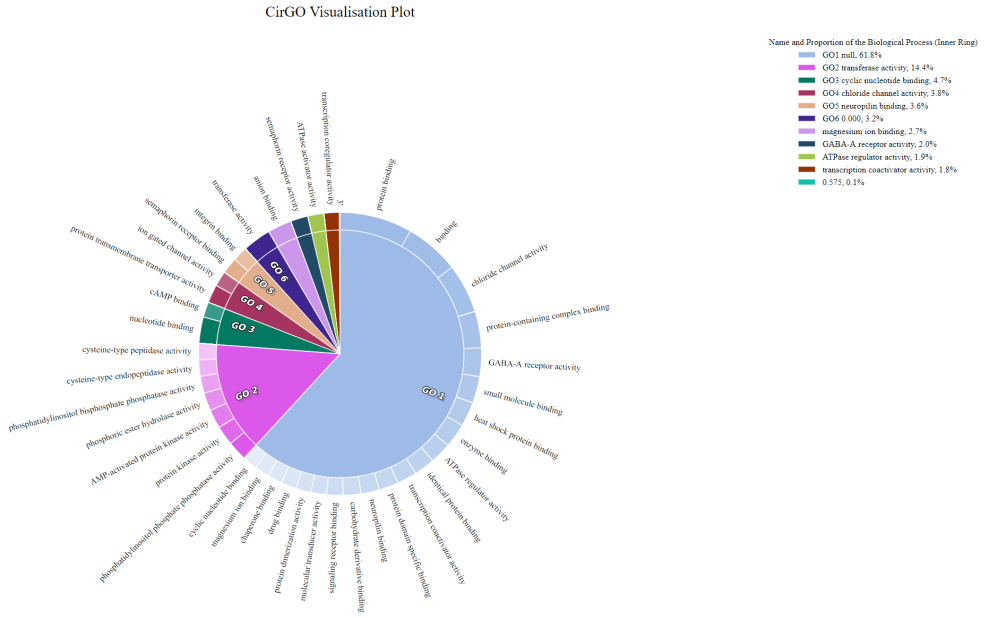


Figure 31: The GO functions of cluster 9.

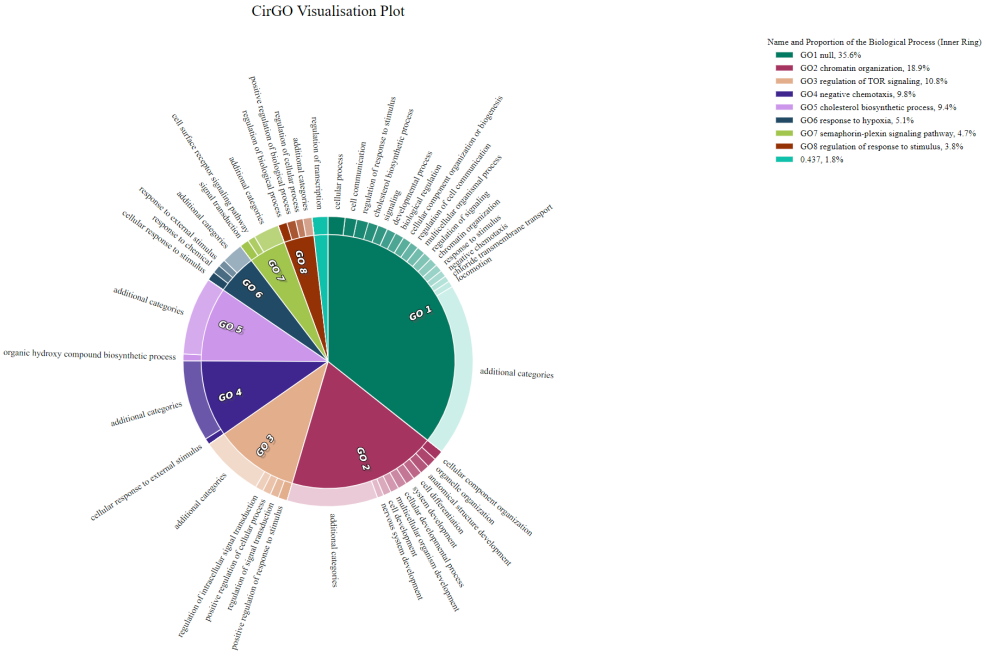


Figure 32: The GO processes of cluster 9.

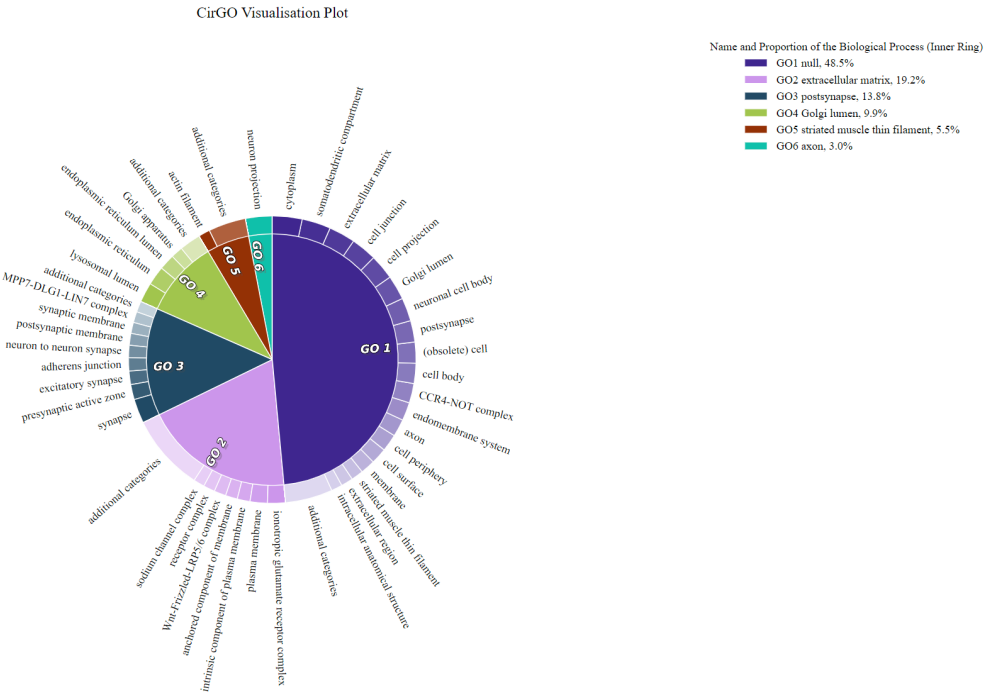


Figure 33: The GO components of cluster 11.

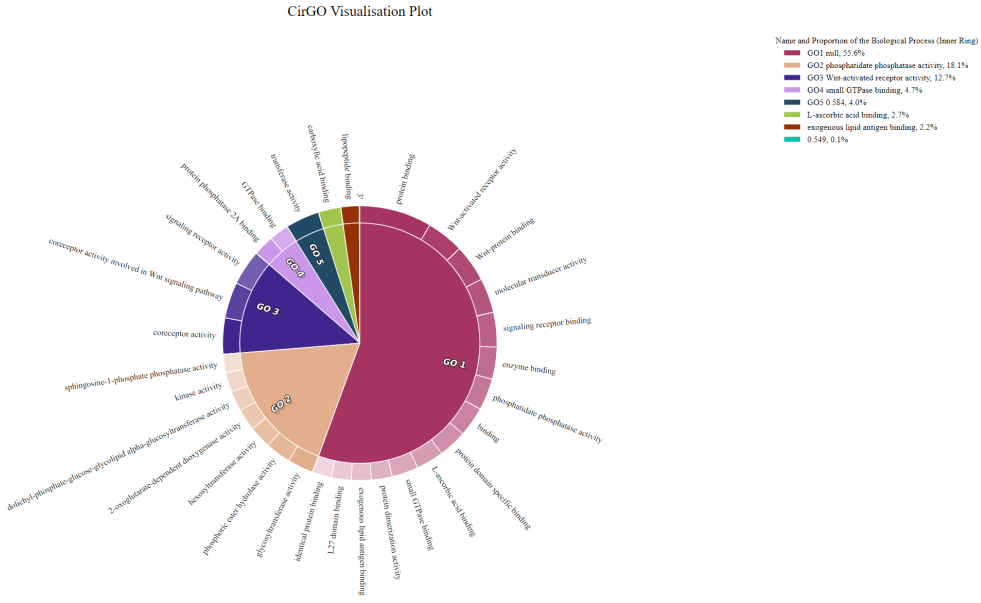


Figure 34: The GO functions of cluster 11.

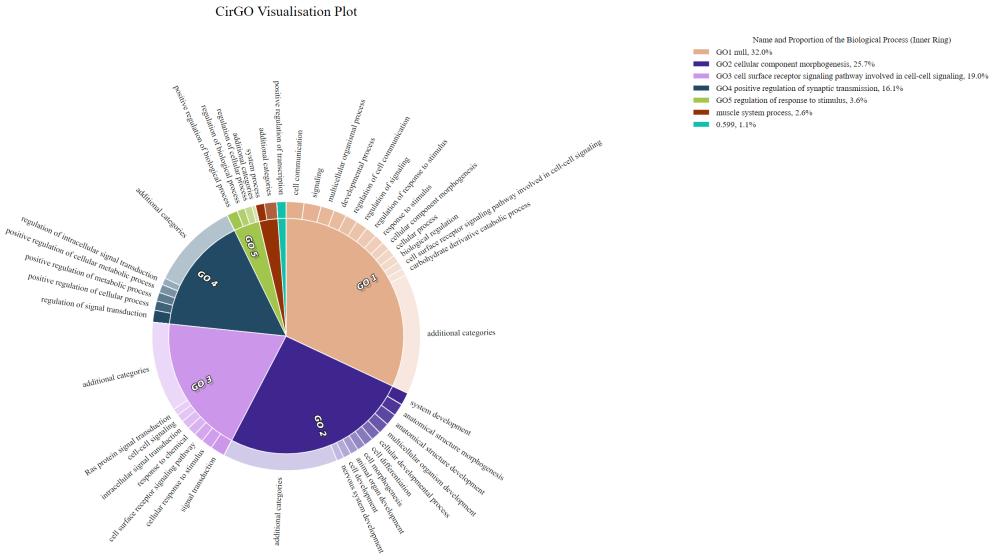


Figure 35: The GO processes of cluster 11.

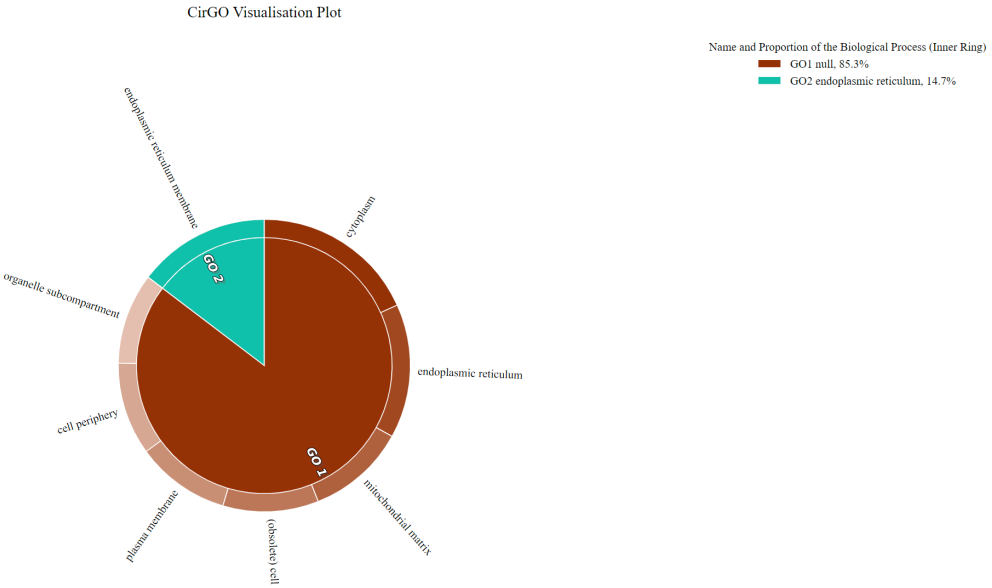


Figure 36: The GO components of cluster 14.

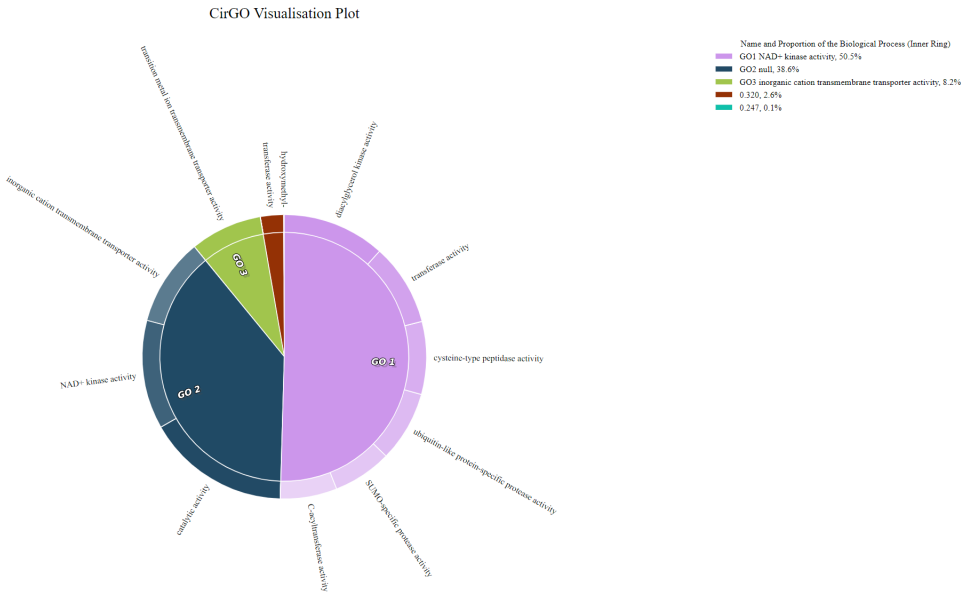


Figure 37: The GO functions of cluster 14.

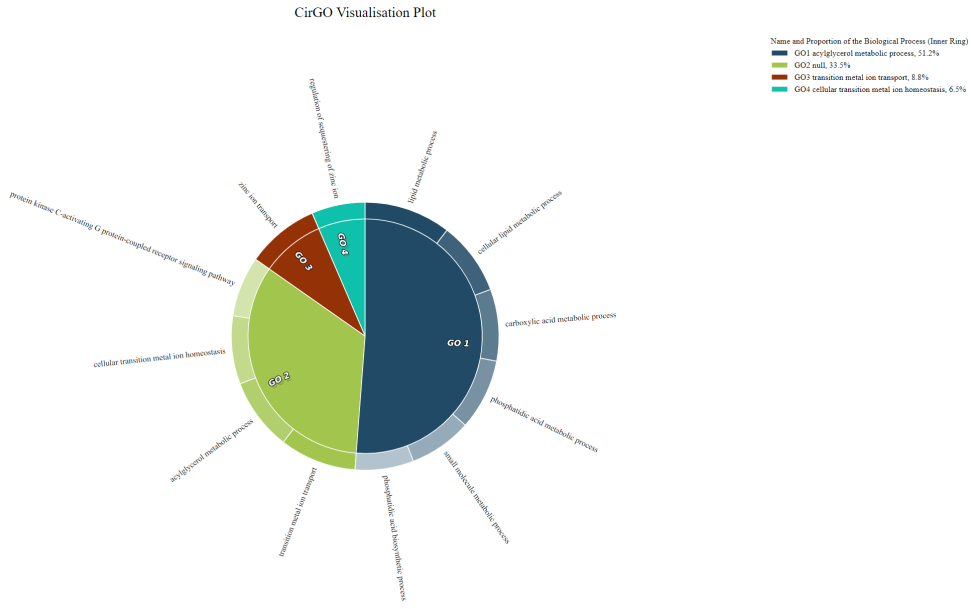


Figure 38: The GO processes of cluster 14.

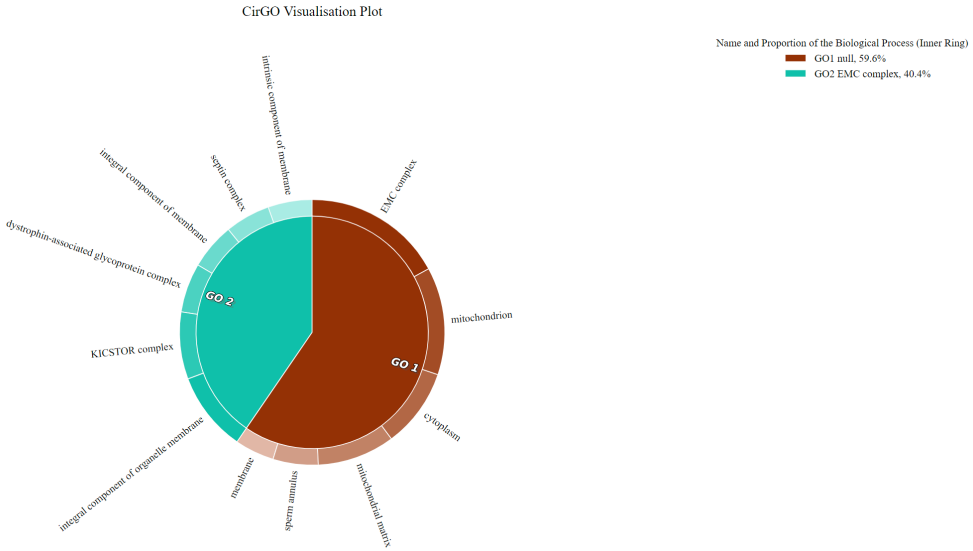


Figure 39: The GO components of cluster 17.

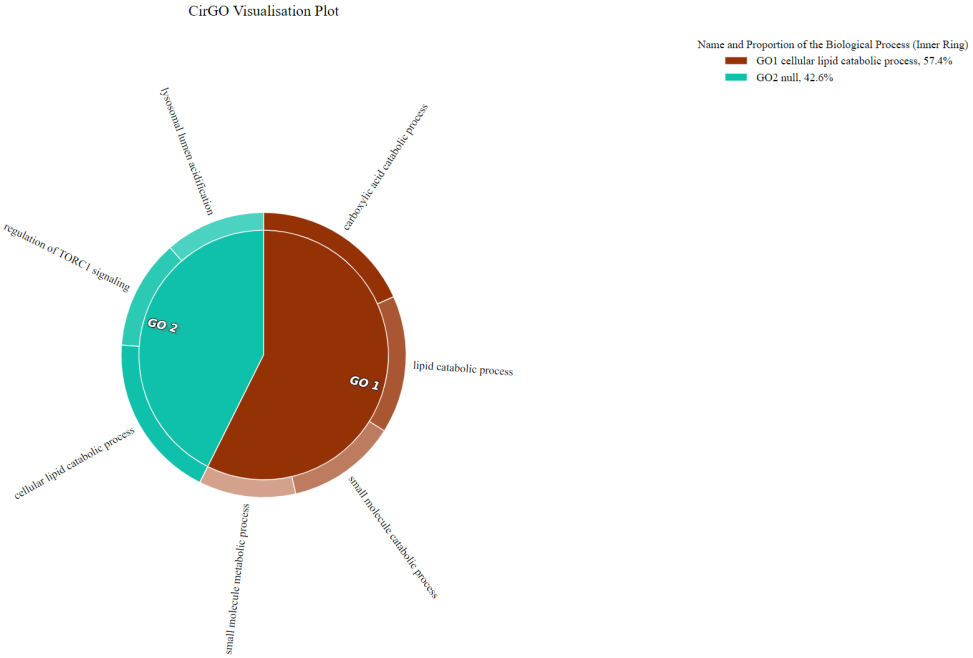


Figure 40: The GO processes of cluster 17.