

SBV887 Assignment

Era Sarda
2020MT10801

May 2023

1. Introduction

Neurodegenerative disorder : Parkinson's disease

Details of the experiment:

Geo Accession ID:GSE206308

Title: Genes critical for development and differentiation of dopaminergic neurons are downregulated in Parkinson's disease

Organisms:Homo sapiens; Mus musculus

Experiment type: Expression profiling by high throughput sequencing

Summary:performed transcriptome analysis using RNA sequencing on substantia nigra pars compacta (SNpc) from mice after acute and chronic 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) treatment and Parkinson's disease (PD) patients.

Overall design: Comparative gene expression profiling analysis of RNA-seq data for substantia nigra pars compacta from MPTP mice treated for 14 days and 24 hours (treated vs. control) and for human PD (PD patients vs. Control).

In this report, only analysis of Homo sapiens data is done of Control v/s Diseased condition.

2. RNA Sequence Analysis

- There are 3 samples of control condition (Sample 1, 2 3), and 4 samples of diseased condition (Sample 4, 5, 6 7)
- Total no. of genes: 63677
- Gene wise normalization of data is done.
- Then for comparing 2 conditions of control and diseased, t-test is applied, with significance level 5 percent.
- According to p-values we obtain the list of differentially expressed genes in the 2 conditions.
- Dataset is obtained of 3 columns: Gene ID, logfc value, p-value.
- The genes in the dataset is sorted according to logfc value, and the ones with value greater than 0 are upregulated genes, and with value less than 0 are downregulated genes.
- No. of differentially expressed genes:714
- No. of upregulated differentially expressed genes: 239
- No. of downregulated differentially expressed genes: 475
- The most upregulated gene came out to be:
ENSG000000002016:
GeneCards Symbol: RAD52
RAD52 Homolog, DNA Repair Protein
log2fc value: 36.6453501738774
pvalue: 0.000990736516687783
- The most downregulated gene came out to be: ENSG00000273373
GeneCards Symbol: SLC16A4-AS1
SLC16A4 Antisense RNA
lof2fc value:-0.976074887866455
pvalue: 0.0046289050161359

Top 8 genes sorted according in decreasing order w.r.t. log2fc values:

Gene Accession Conversion Tool							Help
Gene Accession Conversion Statistics							Download File
Conversion Summary			Submit Converted List to DAVID as a Gene List		Submit Converted List to DAVID as a Background		
ID Count	In DAVID DB	Conversion	From	To	Species	David Gene Name	
8	Yes	Successful	ENSG00000003989	6542	Homo sapiens	solute carrier family 7 member 2(SLC7A2)	
0	Yes	None	ENSG00000003509	55471	Homo sapiens	NADH:ubiquinone oxidoreductase complex assembly factor 7(NDUFAF7)	
0	No	None	ENSG00000007129	90273	Homo sapiens	CEA cell adhesion molecule 21(CEACAM21)	
0	Ambiguous	Pending	ENSG00000004961	3052	Homo sapiens	holocytochrome c synthase(HCCS)	
Total Unique User IDs: 8			ENSG00000006757	8228	Homo sapiens	patatin like phospholipase domain containing 4(PNPLA4)	
Summary of Ambiguous Gene IDs			ENSG00000006071	6833	Homo sapiens	ATP binding cassette subfamily C member 8(ABCC8)	
All Possible Sources For Ambiguous IDs			ENSG00000002016	5893	Homo sapiens	RAD52 homolog, DNA repair protein(RAD52)	
Ambiguous ID			ENSG00000004866	93655	Homo sapiens	ST7 overlapping transcript 3(ST7-OT3)	
			ENSG00000004866	7982	Homo sapiens	suppression of tumorigenicity 7(ST7)	

Figure 1: source: DAVID ontology software

- Volcano plot is used display the genes, with the red ones as differentially expressed genes, and blue the other ones.
- Heat map also clearly shows the upregulation and downregulation of genes between the 2 conditions of control and diseased. The red portion shows high logfc values i.e upregulated genes (> 0) and blue as low logfc values i.e. downregulated genes (< 0).

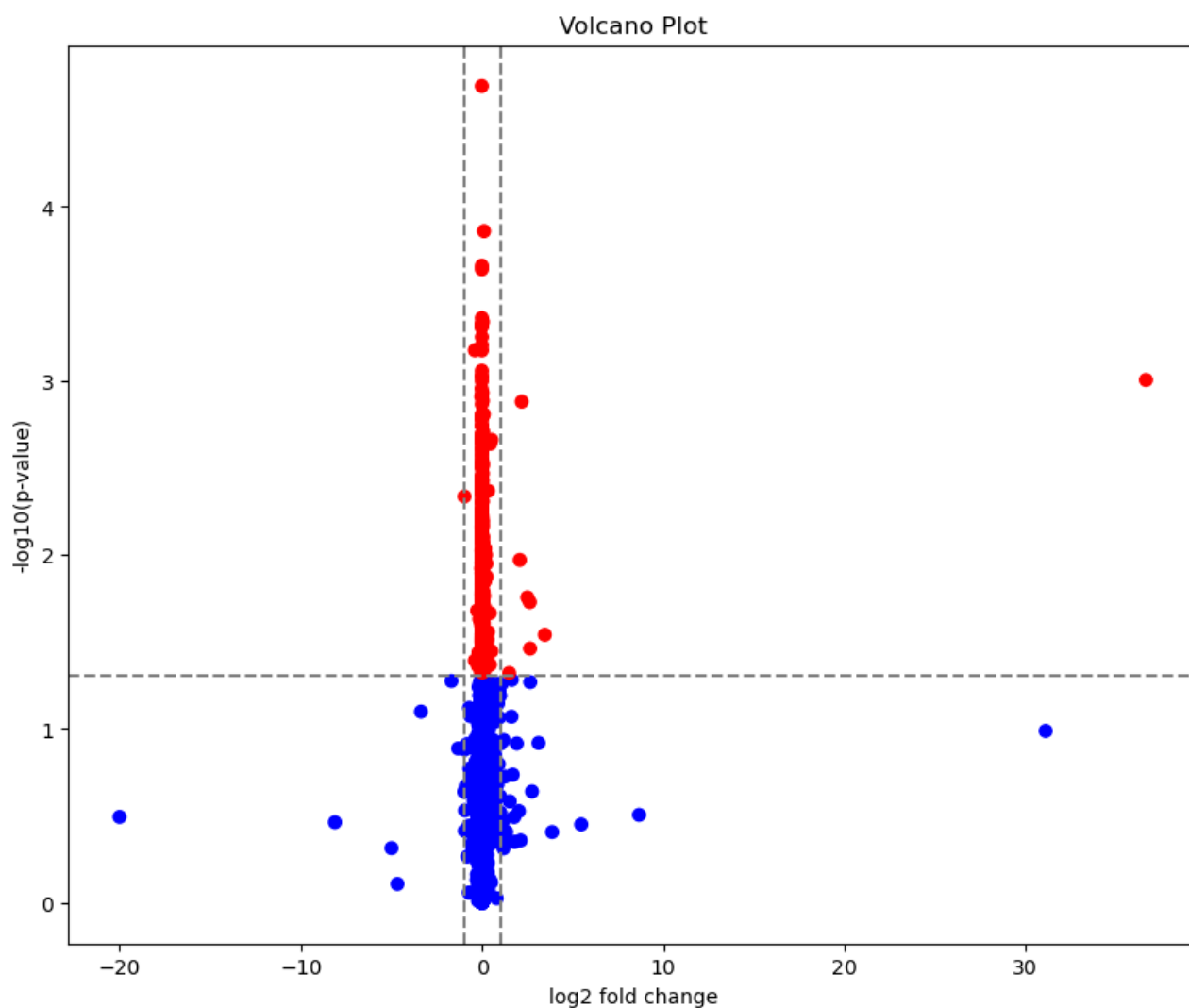


Figure 2: log2 fold change v/s -log10(p-value)

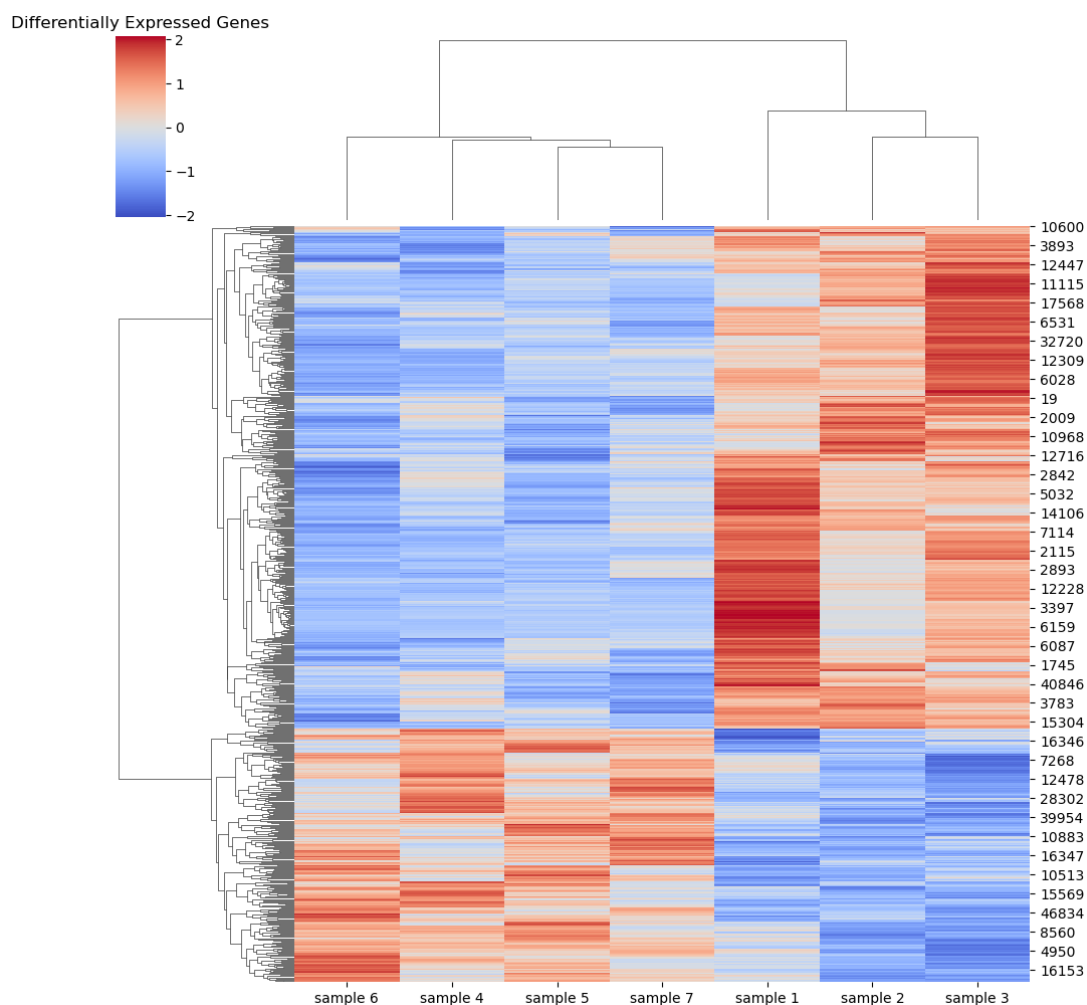


Figure 3: Differentially Expressed genes

3. Protein class

	Gene ID	Mapped IDs	Gene Name Gene Symbol Persistent id Orthologs	PANTHER Family/Subfamily	<input checked="" type="checkbox"/> PANTHER Protein Class	<input checked="" type="checkbox"/> Species <input checked="" type="checkbox"/>
<input type="checkbox"/>	1. HUMAN HGNC=4837 UniProtKB=P53701	ENSG00000004961	Holocytochrome c-type synthase HCCS PTN002489554 orthologs	HOLOCYTOCHROME C-TYPE SYNTHASE (PTHR12743:SF0)	lyase	Homo sapiens
<input type="checkbox"/>	2. HUMAN HGNC=24887 UniProtKB=P41247	ENSG00000006757	Patatin-like phospholipase domain-containing protein 4 PNPLA4 PTN002487927 orthologs	PATATIN-LIKE PHOSPHOLIPASE DOMAIN-CONTAINING PROTEIN 4 (PTHR12406:SF7)	phospholipase	Homo sapiens
<input type="checkbox"/>	3. HUMAN HGNC=11060 UniProtKB=P52569	ENSG00000003989	Cationic amino acid transporter 2 SLC7A2 PTN002534894 orthologs	CATIONIC AMINO ACID TRANSPORTER 2 (PTHR43243:SF35)	amino acid transporter	Homo sapiens
<input type="checkbox"/>	4. HUMAN HGNC=28834 UniProtKB=Q3KPI0	ENSG00000007129	Carcinoembryonic antigen-related cell adhesion molecule 21 CEACAM21 PTN002486103 orthologs	CARCINOEMBRYONIC ANTIGEN-RELATED CELL ADHESION MOLECULE 21 (PTHR12080:SF82)	immunoglobulin receptor superfamily	Homo sapiens
<input type="checkbox"/>	5. HUMAN HGNC=28816 UniProtKB=Q7L592	ENSG00000003509	Protein arginine methyltransferase NDUF7, mitochondrial NDUF7 PTN002486034 orthologs	PROTEIN ARGININE METHYLTRANSFERASE NDUF7, MITOCHONDRIAL (PTHR12049:SF7)	-	Homo sapiens
<input type="checkbox"/>	6. HUMAN HGNC=9824 UniProtKB=P43351	ENSG00000002016	DNA repair protein RAD52 homolog RAD52 PTN002486322 orthologs	DNA REPAIR PROTEIN RAD52 HOMOLOG (PTHR12132:SF1)	DNA metabolism protein	Homo sapiens
<input type="checkbox"/>	7. HUMAN HGNC=59 UniProtKB=Q09428	ENSG00000006071	ATP-binding cassette sub-family C member 8 ABCC8 PTN002517148 orthologs	ATP-BINDING CASSETTE SUB-FAMILY C MEMBER 8 (PTHR24223:SF187)	ATP-binding cassette (ABC) transporter	Homo sapiens
<input type="checkbox"/>	8. HUMAN HGNC=11351 UniProtKB=Q9NRC1	ENSG00000004866	Suppressor of tumorigenicity 7 protein ST7 PTN002489557 orthologs	SUPPRESSOR OF TUMORIGENICITY 7 PROTEIN (PTHR12745:SF10)	-	Homo sapiens

Figure 4: Top 8 genes (source:Panther)

- (a) defense/immunity protein (PC00090): 1 gene : CEACAM21
- (b) transporter (PC00227): 2 genes: SLC7A2, ABCC8
- (c) metabolite interconversion enzyme (PC00262): 2 genes : HCCS, PNPLA4
- (d) DNA metabolism protein (PC00009) 1 gene : RAD52
- (e) No PANTHER category is assigned (UNCLASSIFIED): 2 genes : NDUFAF7, ST7

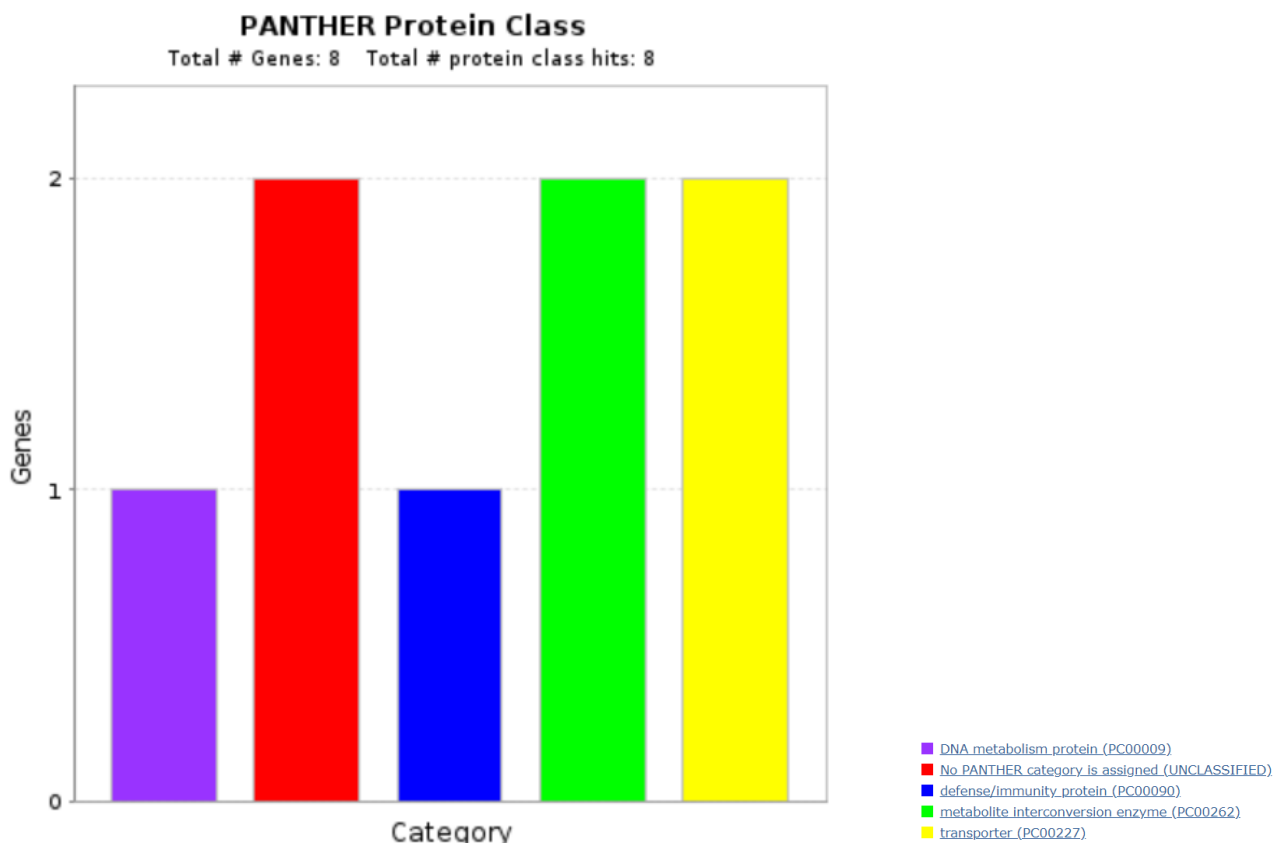


Figure 5: Protein classes

Figure 6: Labels

Although there are higher number of genes belonging to the protein classes Metabolite inter-conversion enzyme (PC00262) and transporter (PC00227), but they caused similar fold change. But DNA metabolism protein (RAD52) caused a very high fold change (\log_2 fc values 36.64535017). Therefore DNA Metabolism protein might be important in the disorder.

DNA Metabolism Protein: A protein that acts on DNA to modify it either topologically (e.g. bending), its base-pairing, or covalently.

Class ID: PC00009

Child: DNA glycosylase

DNA helicase

DNA ligase

DNA methyltransferase

DNA photolyase

DNA polymerase processivity factor

DNA strand-pairing protein

DNA topoisomerase etc.

4. Structural Analysis of DNA Helicase Chosen DNA Helicase: RECQL4

ATP-dependent DNA helicase Q1; DNA helicase that may play a role in the repair of DNA that is damaged by ultraviolet light or other mutagens. Exhibits a magnesium-dependent ATP-dependent DNA-helicase activity that unwinds single- and double-stranded DNA in a 3'-5' direction; Belongs to the helicase family. RecQ subfamily (649 aa)

Polymer Visualization



Figure 7: DNA Helicase

Ligand Interaction with polymer

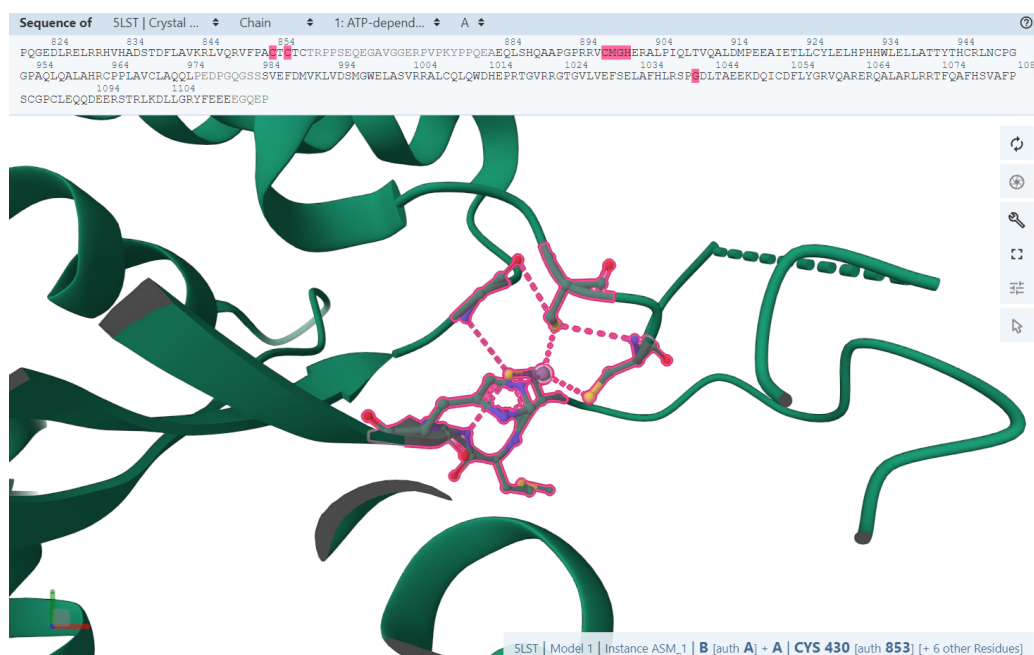


Figure 8: source: Protein Data Bank

Ligand view

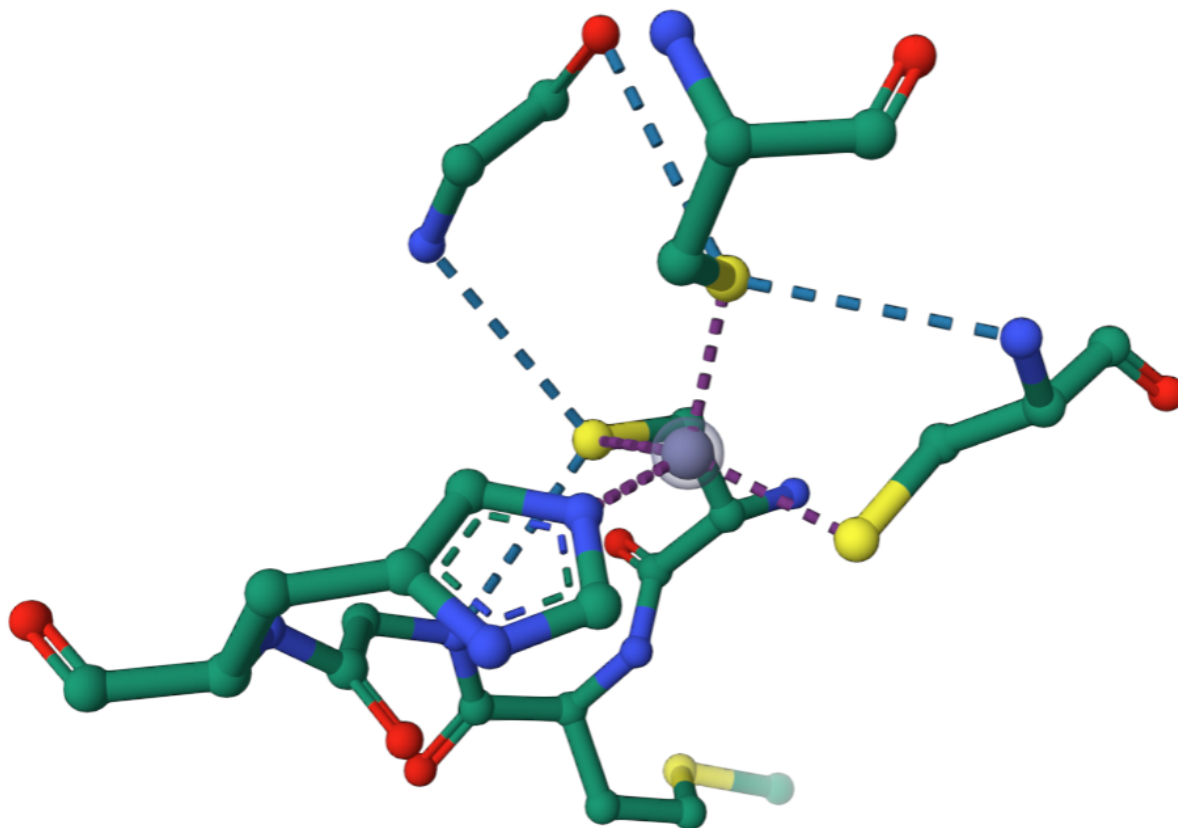


Figure 9:

ION: Zinc Ion ZN 1201
Interaction Fetaure: Types Transition Metal
Valence Model: charge 0
Idea Geometry: Unknown
Implicit H and Total H: 0

Ligands: CYS (Cysteine), GLY (Glycine)

5. Network Analysis of DNA helicase Chosen DNA Helicase: RECQL4

Note: STRING uses a spring model to generate the network images. Nodes are modeled as masses and edges as springs; the final position of the nodes in the image is computed by minimizing the 'energy' of the system. It gives high confidence edges a higher 'spring strength' so that they will reach an optimal position before lower confidence edges.

Line color indicates the type of interaction evidence

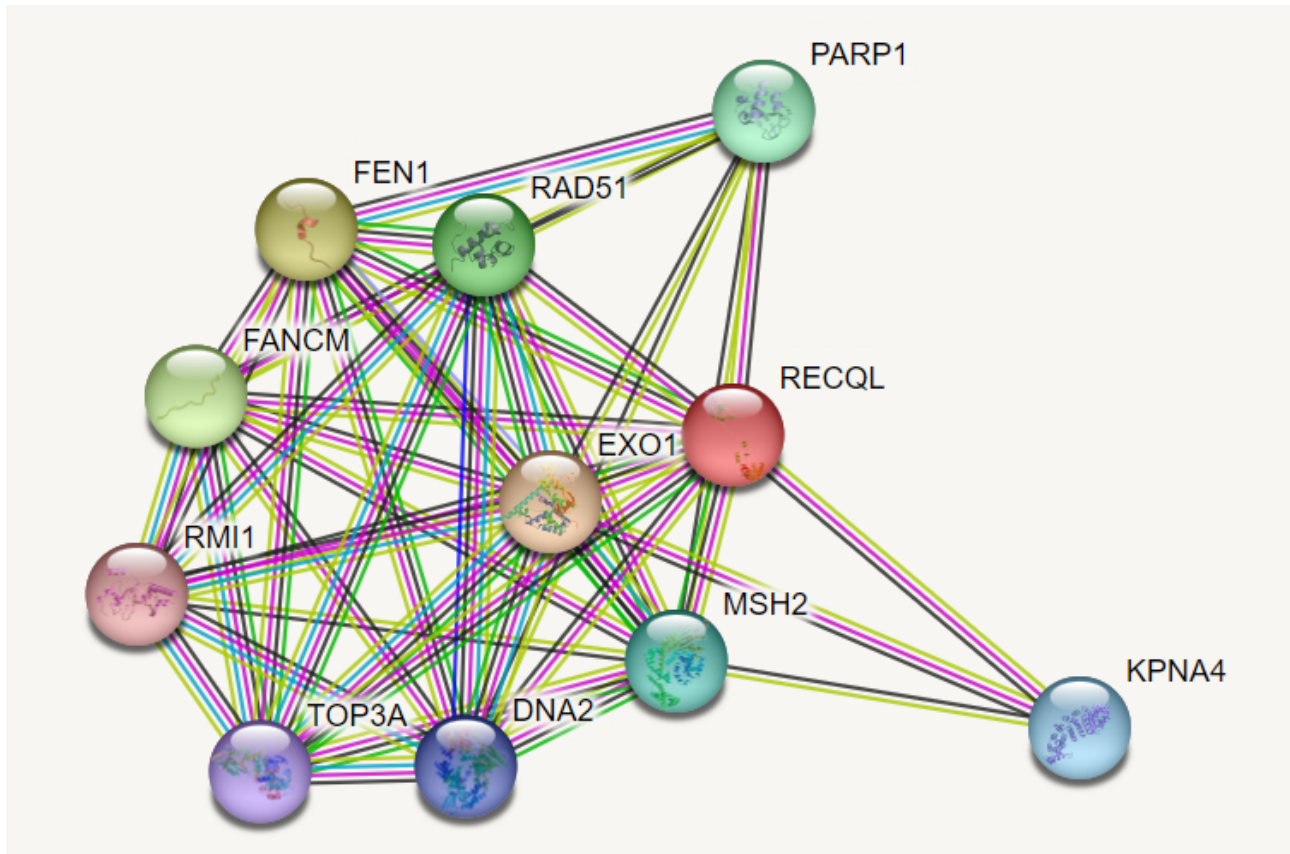


Figure 10: full STRING Network (source:STRING)

Proteins which may be interacting with RECQL:

Protein	Evidence score
EXO1	0.977
FEN1	0.939
FANCM	0.931
RAD51	0.902
PARP1	0.894
MSH2	0.883
KPNA4	0.879
DNA2	0.850
TOP3A	0.840
RMI1	0.842

Network Stats:

number of nodes:11

number of edges:46

average node degree:8.36

avg. local clustering coefficient: 0.925

expected number of edges:13

PPI enrichment p-value: 1.31e-12

The network has significantly more interactions than expected.

Line thickness indicates the strength of data support

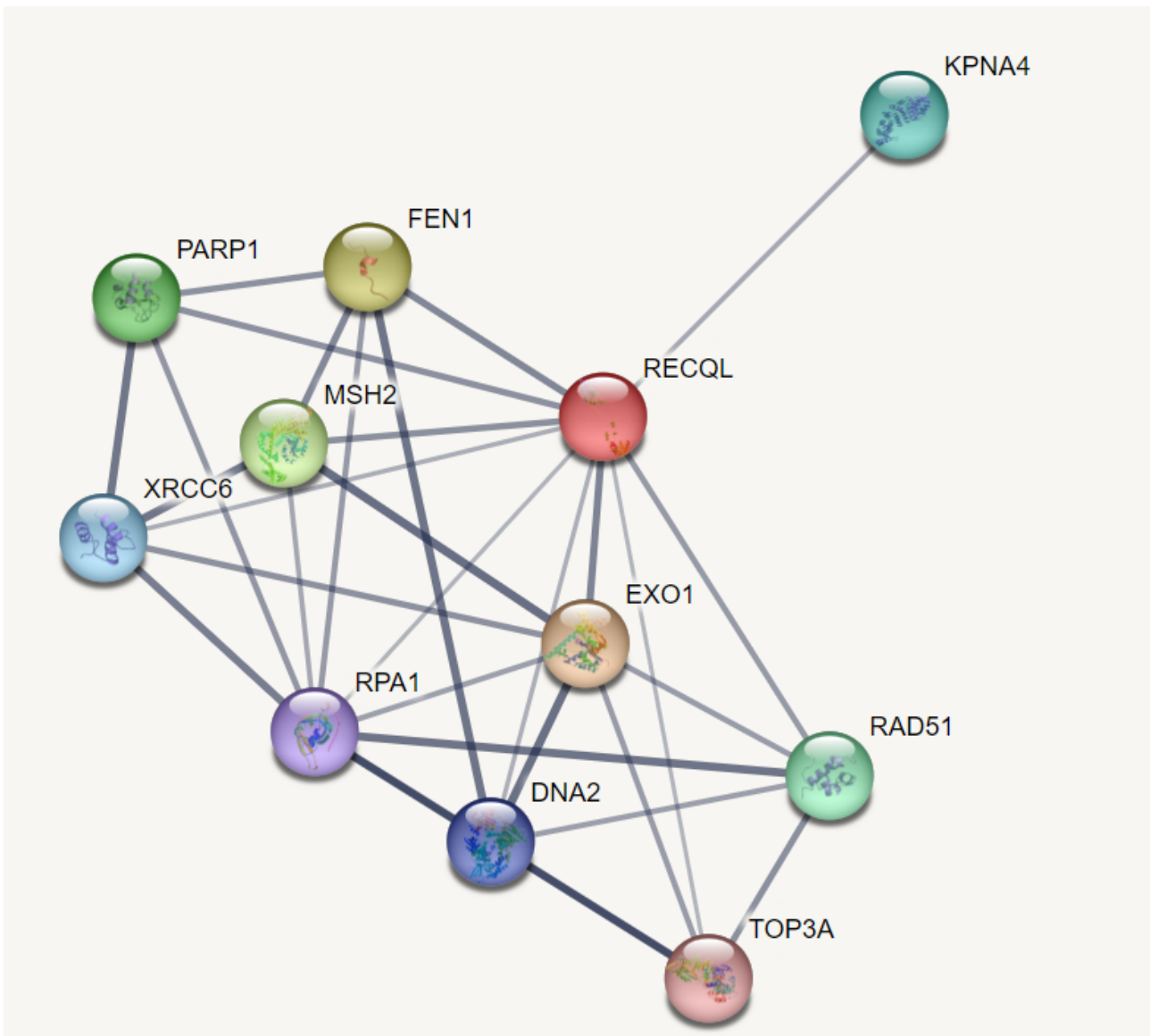


Figure 11: physical subnetwork

Predicted Physical Partners:

Protein	Confidence score
EXO1	0.834
FEN1	0.811
MSH2	0.753
PARP1	0.740
RAD51	0.676
KPNA4	0.570
XRCC6	0.534
DNA2	0.526
RPA1	0.492
TOP3A	0.482

Network Stats:

number of nodes: 11

number of edges: 31

average node degree: 5.64

average local clustering coefficient: 0.791

expected number of edges: 11

PPI enrichment p-value: 1.02e-06

The network has significantly more interactions than expected