DATABASE PROJECT

Author: Ridhiwan Mseya

HEPATOCELLULAR CARCINOMA

DESCRIPTION

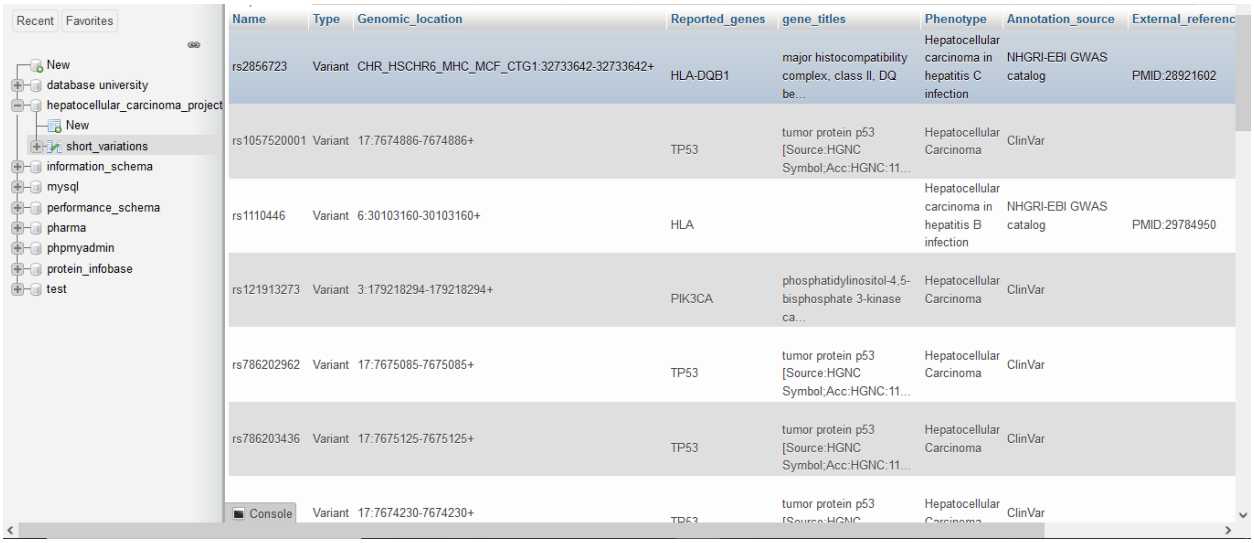
Hepatocellular carcinoma is a cancer malignancy of the liver that is the most common in the world and is first in the leader board of cancer related deaths. The main risk factors that lead to this malignancy are excessive alcohol consumption and chronic viral hepatitis. Suppression of the hepatitis virus can result in the delayed occurrence of malignancy for up to 5 years. Other risk factors are; sex whereby males are more at risk than females at a ratio of 2:1, metabolic and genetic diseases like hemochromatosis, exposure to aflatoxins and smoking cigarettes.

Cirrhosis in the liver causes the formation of regenerative nodules which are the result of increased proliferation of hepatocytes. Detection of these lesions depends on their size, the smaller lesions are harder to detect and lead to misdiagnosis. The use of AFP which regulates the proliferation of the adult liver cells as a biomarker together with DCP which is produced by malignant hepatocytes leads to better sensitivity and specificity in the detection of the disease. Treatment of the disease is possible only when it is diagnosed early with retained liver function. A lot of interventions are possible some surgical and others nonsurgical but for the full recovery a patients will always need a multidisciplinary team that will monitor all developments.

SHORT VARIATIONS, UCSC DATA AND GEO DATA FOR HCC

In order to retrieve a table of SNPs and indels for the phetotype, the ensemble database was visited and the specie of choice was human and a query was launched for hepatocellular carcinoma. The search results were filtered of the feature type column to include only variants, the variants are all sourced from the dbSNP database. The table was then exported in terms of a csv file. The table was then edited in excel to remove duplicate column names and imported into my database. While in the database it was further edited to remove unnecessary columns and the first row which contained column names.

The table of “short_variations” is shown below:



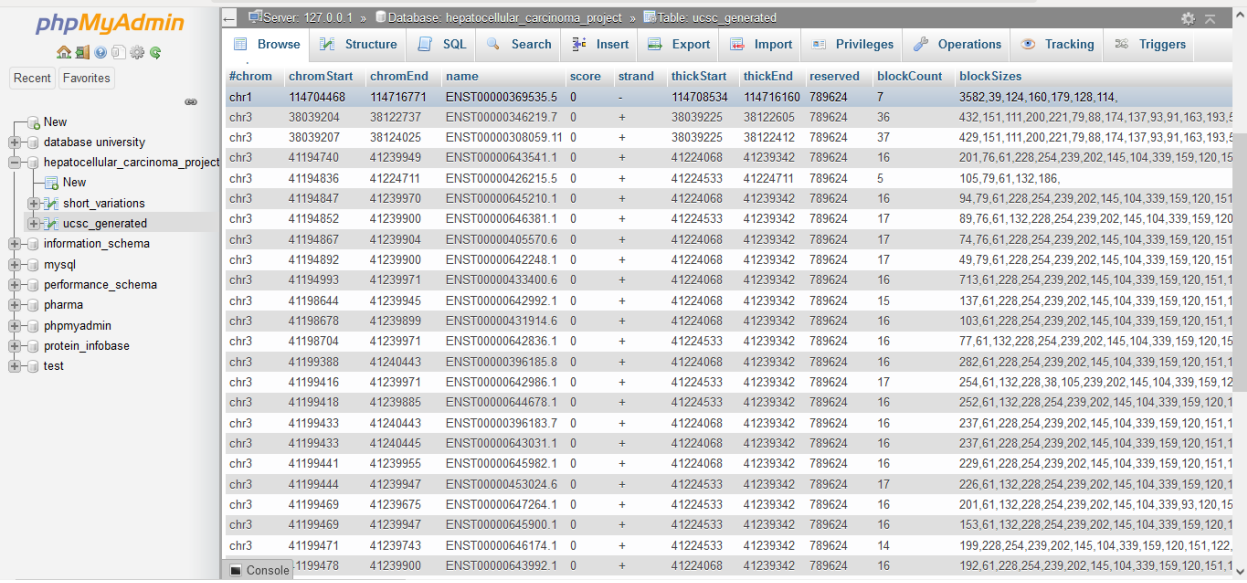
Name	Type	Genomic_location	Reported_genes	gene_titles	Phenotype	Annotation_source	External_referenc
rs2856723	Variant	CHR_HSCHR6_MHC_MCF_CTG1:32733642-32733642+	HLA-DQB1	major histocompatibility complex, class II, DQ be...	Hepatocellular carcinoma in hepatitis C infection	NHGRI-EBI GWAS catalog	PMID:28921602
rs1057520001	Variant	17:7674886-7674886+	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11...	Hepatocellular Carcinoma	ClinVar	
rs1110446	Variant	6:30103160-30103160+	HLA		Hepatocellular carcinoma in hepatitis B infection	NHGRI-EBI GWAS catalog	PMID:29784950
rs121913273	Variant	3:179218294-179218294+	PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase ca...	Hepatocellular Carcinoma	ClinVar	
rs786202962	Variant	17:7675085-7675085+	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11...	Hepatocellular Carcinoma	ClinVar	
rs786203436	Variant	17:7675125-7675125+	TP53	tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11...	Hepatocellular Carcinoma	ClinVar	
	Variant	17:7674230-7674230+	TP53	tumor protein p53 [Source:HGNC	Hepatocellular Carcinoma	ClinVar	

Retrieved a table of genes from the database using;

SELECT DISTINCT `Reported_genes` FROM `short_variations` WHERE Reported_genes != ""

Exported the table as csv file to be used in UCSC browser. The csv file had to also be edited to remove the title and make it comma delimited. When uploading to the UCSC table browser an error occurred where some of the genes couldn’t be recognized because of the format of the csv so they were pasted as a list manually and edited accordingly.

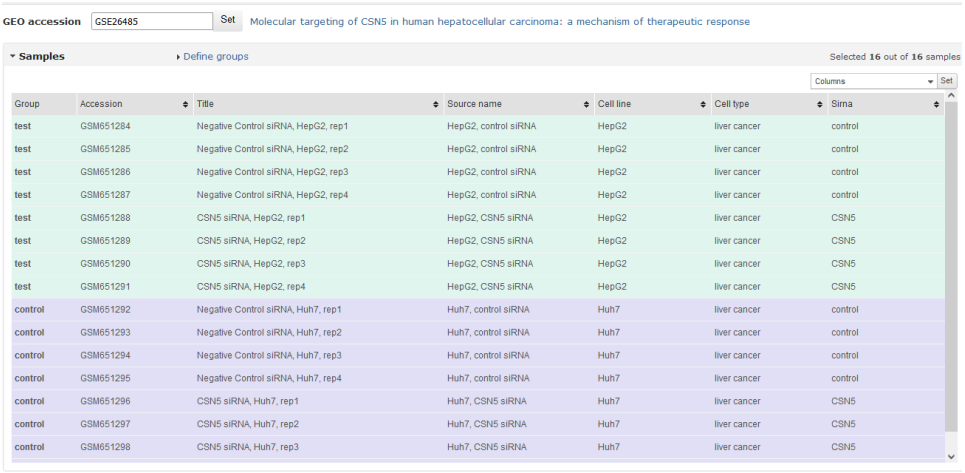
In the UCSC table browser I selected my track to be GENECODE V38 as it has the most genes recorded and selected region as genome and my table was knownGene. I also selected NCBI RefSeq track in order to get the refseqIDs and exon counts. These configurations had information on all of the genes except seven genes; HLA, HLA-DQ, HLA-DR, HLA-DQ, HLA_DRB1, HLA_DQA1, LOC110806263. The obtained tables were in the form of text and had to be converted to a csv files for importation into the database. Shown below is the table for first configuration;



To get a GEO dataset starting from the first day of 2010 I used advanced search option in the browser and selected the organism to be human, the published date to be 1/1/2010 to present, the filter to be Datasets and searched for HCC as phenotype. Below are the results of the query;

I chose to work with the sixth dataset: “CSN5 depletion effect on hepatocellular carcinoma cell lines”. This dataset involves analysis of Huh7 and HepG2 hepatocellular carcinoma (HCC) cells depleted for CSN5, the fifth subunit of the COP9 signalosome. CSN5 expression is elevated in early HCC. Results provide insight into the role of CSN5 in the pathogenesis of HCC. The organism is Homo Sapien, the platform is “GPL6883: Illumina HumanRef-8 v3.0 expression beadchip”, GSE26485 and published on 2011/12/27.

To analyze the GEO dataset with GEO2R I used its reference number to search for the dataset and selected half of my data to be the test set and the other half as the control.



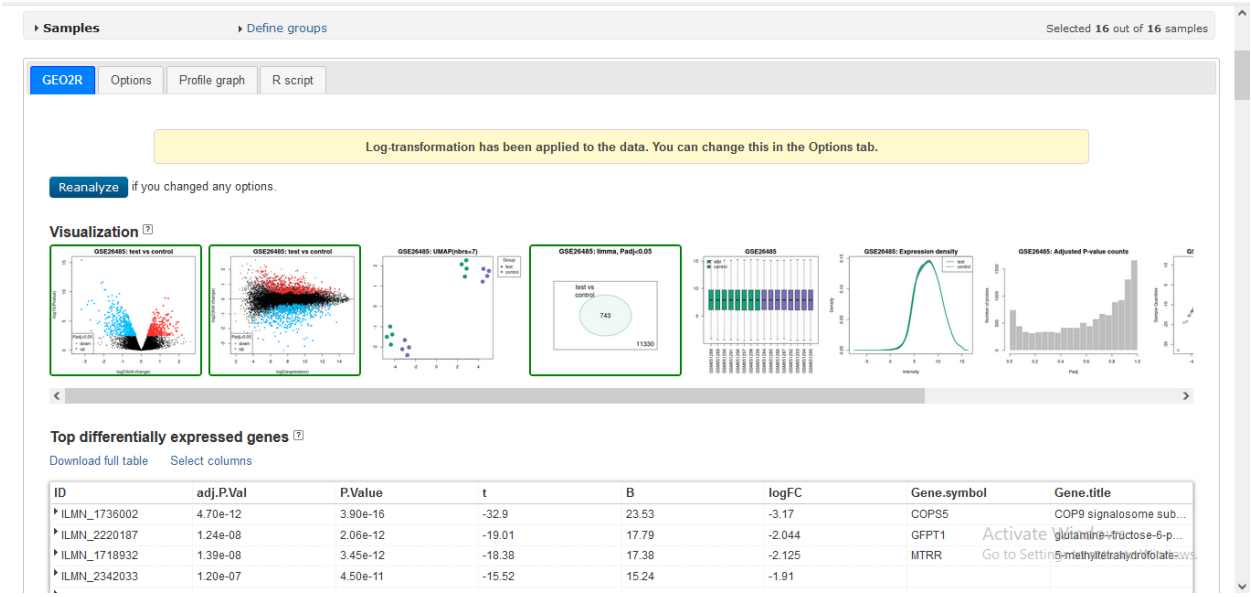
GEO accession: GSE26485 Set — Molecular targeting of CSN5 in human hepatocellular carcinoma: a mechanism of therapeutic response

Samples — Define groups — Selected 16 out of 16 samples

Columns Set

Group	Accession	Title	Source name	Cell line	Cell type	Sirna
test	GSM651284	Negative Control siRNA, HepG2, rep1	HepG2, control siRNA	HepG2	liver cancer	control
test	GSM651285	Negative Control siRNA, HepG2, rep2	HepG2, control siRNA	HepG2	liver cancer	control
test	GSM651286	Negative Control siRNA, HepG2, rep3	HepG2, control siRNA	HepG2	liver cancer	control
test	GSM651287	Negative Control siRNA, HepG2, rep4	HepG2, control siRNA	HepG2	liver cancer	control
test	GSM651288	CSN5 siRNA, HepG2, rep1	HepG2, CSN5 siRNA	HepG2	liver cancer	CSN5
test	GSM651289	CSN5 siRNA, HepG2, rep2	HepG2, CSN5 siRNA	HepG2	liver cancer	CSN5
test	GSM651290	CSN5 siRNA, HepG2, rep3	HepG2, CSN5 siRNA	HepG2	liver cancer	CSN5
test	GSM651291	CSN5 siRNA, HepG2, rep4	HepG2, CSN5 siRNA	HepG2	liver cancer	CSN5
control	GSM651292	Negative Control siRNA, Huh7, rep1	Huh7, control siRNA	Huh7	liver cancer	control
control	GSM651293	Negative Control siRNA, Huh7, rep2	Huh7, control siRNA	Huh7	liver cancer	control
control	GSM651294	Negative Control siRNA, Huh7, rep3	Huh7, control siRNA	Huh7	liver cancer	control
control	GSM651295	Negative Control siRNA, Huh7, rep4	Huh7, control siRNA	Huh7	liver cancer	control
control	GSM651296	CSN5 siRNA, Huh7, rep1	Huh7, CSN5 siRNA	Huh7	liver cancer	CSN5
control	GSM651297	CSN5 siRNA, Huh7, rep2	Huh7, CSN5 siRNA	Huh7	liver cancer	CSN5
control	GSM651298	CSN5 siRNA, Huh7, rep3	Huh7, CSN5 siRNA	Huh7	liver cancer	CSN5

The analysis is shown on the screenshot below:



Samples — Define groups — Selected 16 out of 16 samples

GEO2R | Options | Profile graph | R script

Log-transformation has been applied to the data. You can change this in the Options tab.

Reanalyze if you changed any options.

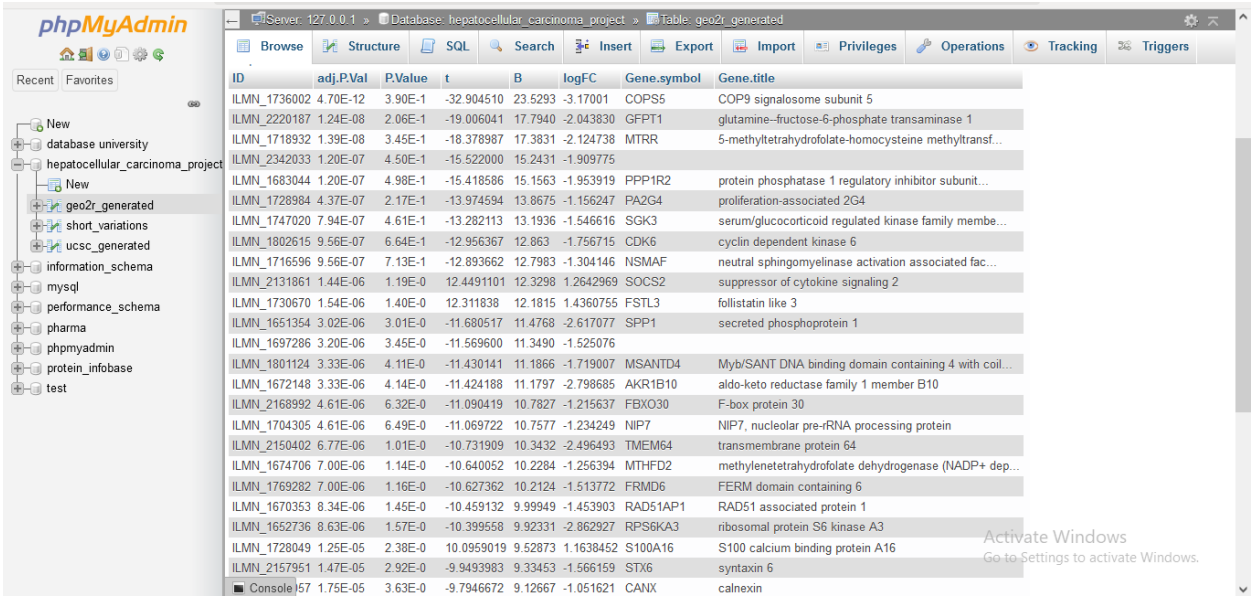
Visualization

Top differentially expressed genes

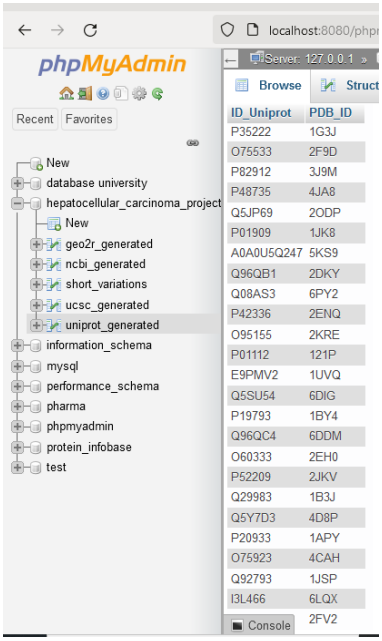
Download full table Select columns

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
ILMN_1736002	4.70e-12	3.90e-16	-32.9	23.53	-3.17	COPS5	COP9 signalosome sub...
ILMN_2220187	1.24e-08	2.06e-12	-19.01	17.79	-2.044	GFPT1	glutamine--fructose-6-p...
ILMN_1718932	1.39e-08	3.45e-12	-18.38	17.38	-2.125	MTRR	5-methyltetrahydrofolate-...
ILMN_2342033	1.20e-07	4.50e-11	-15.52	15.24	-1.91		

The whole table contained more than 10,000 rows which reached the data limit for importation in the database, thus I used only the first 250 rows of the whole dataset. Shown below is the table generated in the database:



With all the 4 tables generated until this point, I have most of the data that is needed to perform different queries but I am missing PDB IDs. These can be generated from Uniprot IDs in the Retrieve/ID web application. I made a query to select only the UniprotID column from the table generated from UCSC data and exported it. All the exported IDs where mapped to their respective PDB ids in the web application. The mapping table was edited to and added to the database. Shown below is the table with PDB ids:

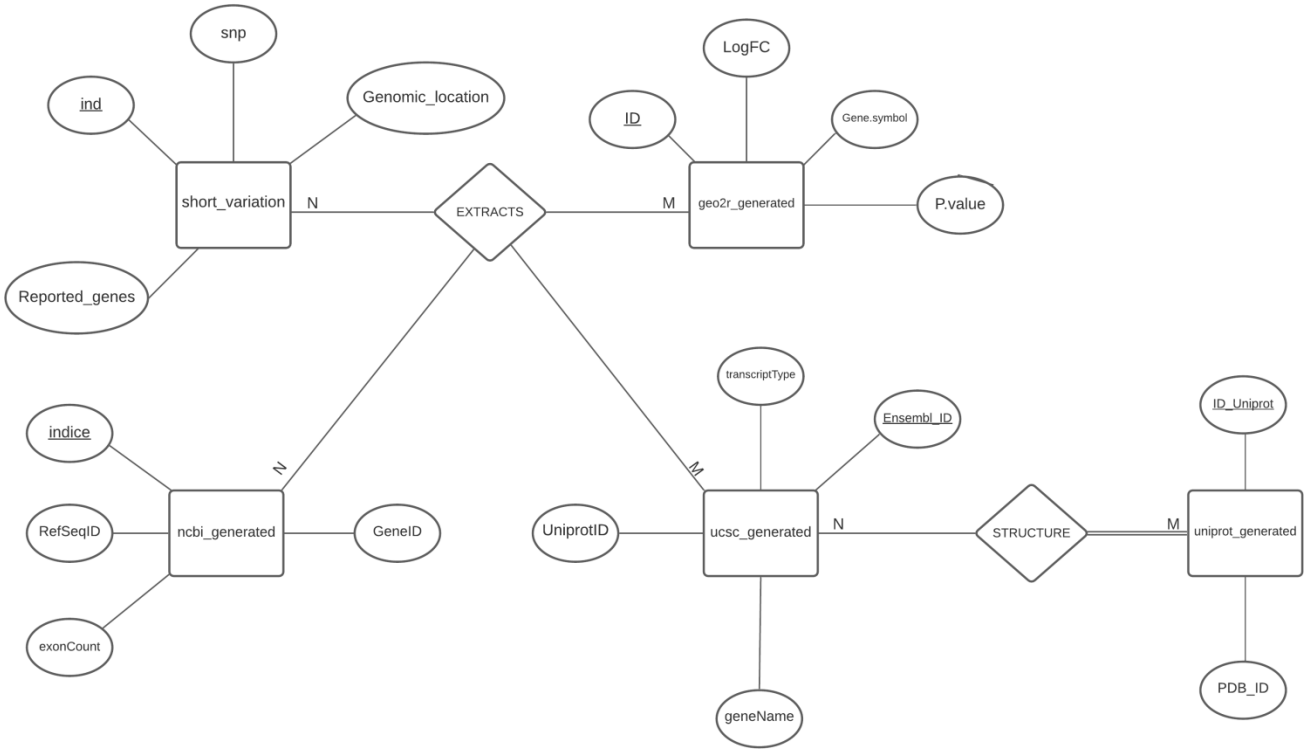


THE HCC DATABASE

The database consists of five relations namely; geo2r_generated, ncbi_generated, short_variations, ucsc_generated and uniprot_generated. These entities are related to each other by gene ids and Uniprot ids. Not all entries will have data for given columns which is common for biological datasets. geo2r_generated, ncbi_generated, short_variations and ucsc_generated entities are related to each other by gene ids while the uniprot_generated entity is related to ucsc_generated by Uniprot ids. Therefore, the HCC database can provide important information of a gene and its relation to the phenotype, the protein structure associated with that gene, how the gene is regulated and its structure and position in the chromosome.

They are a lot of duplicates in the entities and it is hard to use the existing columns as primary keys so we will add row number column as a primary key to such entities. Four of the entities are connected by one relationship and the remaining one is related to only one of the entities and is dependent on it. All primary keys are underlined and there is only one foreign key found in the ucsc_generated entity. Some of the attributes are not show on the ER Diagram below, for the sake of simplicity and better visualization.

ER DIAGRAM

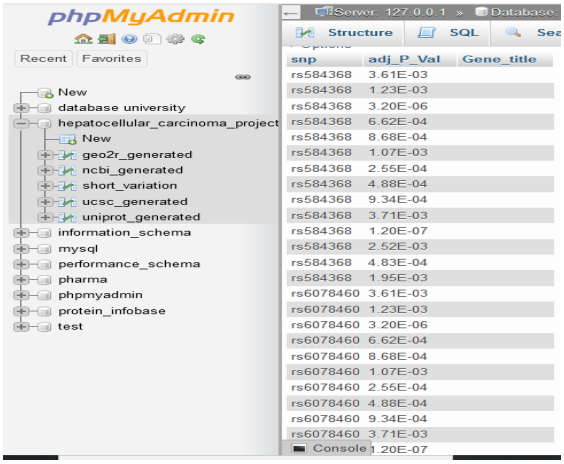


1. First query

   SELECT snp, adj_P_Val, Gene_title

   FROM short_variation as s, geo2r_generated as g

   WHERE s.Reported_genes = g.Gene_symbol

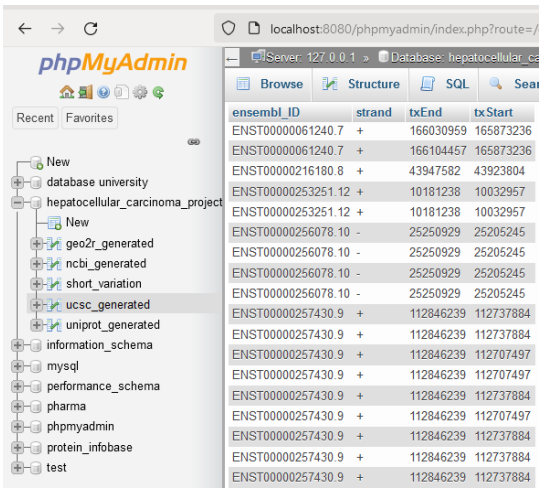


2. Second query

   SELECT ensembl_ID, u.strand, txEnd, txStart

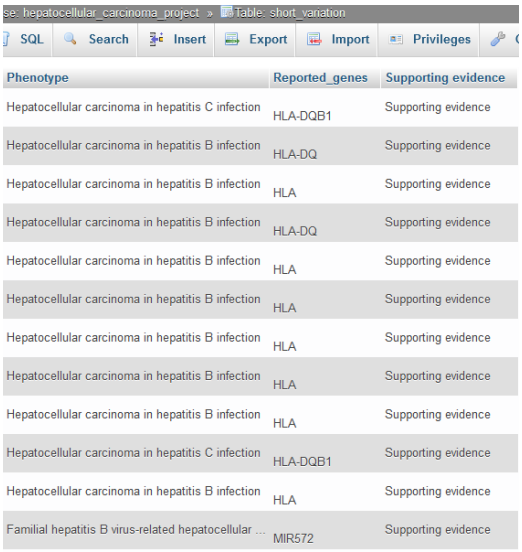
   FROM ncbi_generated as n, ucsc_generated as u

WHERE n.GeneID = u.geneName



3. Third query

   SELECT Phenotype, Reported_genes, "Supporting evidence"
   FROM short_variation as s
   WHERE s.Phenotype LIKE '%hepatitis%'



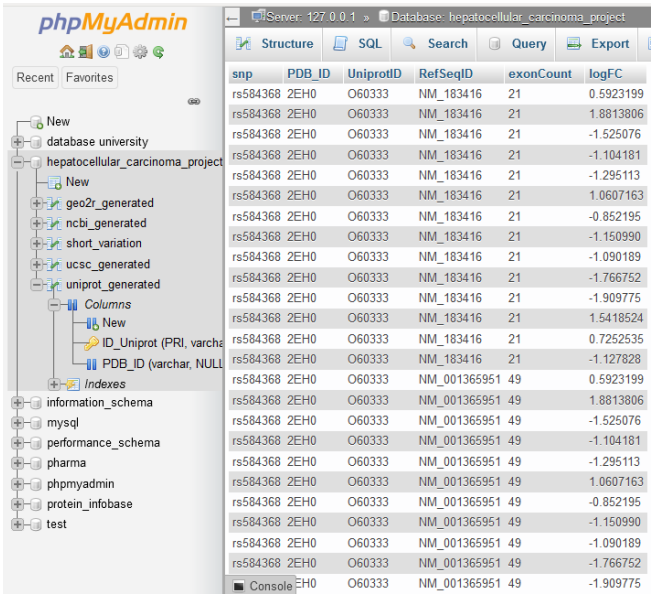
Phenotype	Reported_genes	Supporting evidence
Hepatocellular carcinoma in hepatitis C infection	HLA-DQB1	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA-DQ	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA-DQ	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA	Supporting evidence
Hepatocellular carcinoma in hepatitis C infection	HLA-DQB1	Supporting evidence
Hepatocellular carcinoma in hepatitis B infection	HLA	Supporting evidence
Familial hepatitis B virus-related hepatocellular ...	MIR572	Supporting evidence

QUERIES

1. Gene of choice is KIF1B

   SELECT snp, PDB_ID, UniprotID, RefSeqID, exonCount, logFC

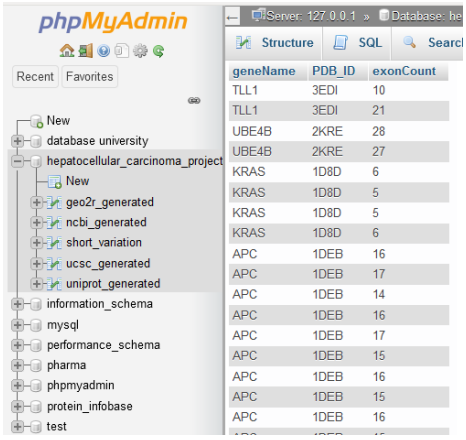
FROM ncbi_generated as n, ucsc_generated as u, short_variation as s, geo2r_generated as g, uniprot_generated as p
WHERE s.Reported_genes = g.Gene_symbol and u.UniprotID = p.ID_Uniprot and n.GeneID = u.geneName and n.GeneID = 'KIF1B'



snp	PDB_ID	UniprotID	RefSeqID	exonCount	logFC
rs584368	2EH0	O60333	NM_183416	21	0.5923199
rs584368	2EH0	O60333	NM_183416	21	1.8813806
rs584368	2EH0	O60333	NM_183416	21	-1.525076
rs584368	2EH0	O60333	NM_183416	21	-1.104181
rs584368	2EH0	O60333	NM_183416	21	-1.295113
rs584368	2EH0	O60333	NM_183416	21	1.0607163
rs584368	2EH0	O60333	NM_183416	21	-0.852195
rs584368	2EH0	O60333	NM_183416	21	-1.150990
rs584368	2EH0	O60333	NM_183416	21	-1.090189
rs584368	2EH0	O60333	NM_183416	21	-1.766752
rs584368	2EH0	O60333	NM_183416	21	-1.909775
rs584368	2EH0	O60333	NM_183416	21	1.5418524
rs584368	2EH0	O60333	NM_183416	21	0.7252535
rs584368	2EH0	O60333	NM_183416	21	-1.127828
rs584368	2EH0	O60333	NM_001365951	49	0.5923199
rs584368	2EH0	O60333	NM_001365951	49	1.8813806
rs584368	2EH0	O60333	NM_001365951	49	-1.525076
rs584368	2EH0	O60333	NM_001365951	49	-1.104181
rs584368	2EH0	O60333	NM_001365951	49	-1.295113
rs584368	2EH0	O60333	NM_001365951	49	1.0607163
rs584368	2EH0	O60333	NM_001365951	49	-0.852195
rs584368	2EH0	O60333	NM_001365951	49	-1.150990
rs584368	2EH0	O60333	NM_001365951	49	-1.090189
rs584368	2EH0	O60333	NM_001365951	49	-1.766752
rs584368	2EH0	O60333	NM_001365951	49	-1.909775

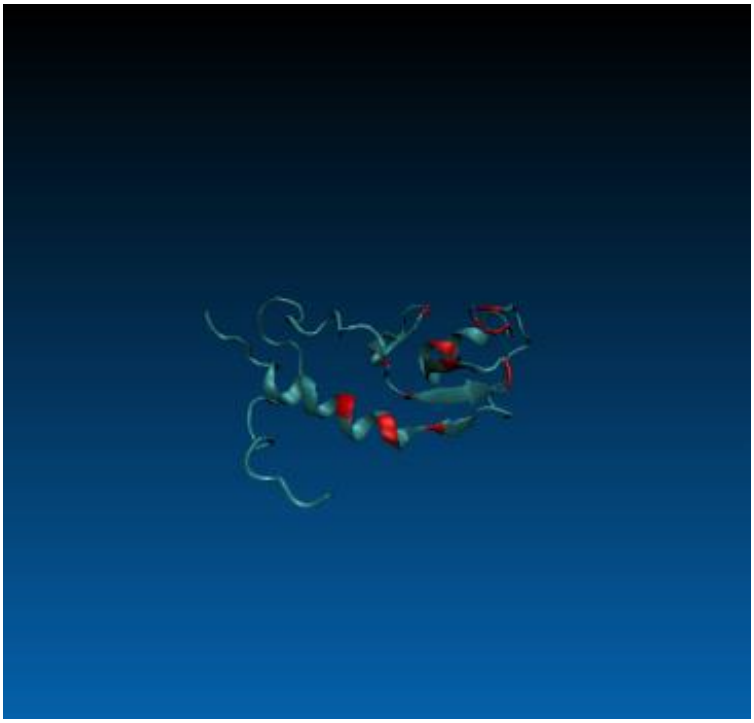
Other genes are UBE4B, CASP8, DLC1

2. Query for PDB_ID structures with exon count greater or equal to 4
SELECT geneName, PDB_ID, exonCount
FROM ncbi_generated as n, ucsc_generated as u, uniprot_generated as p
WHERE n.GeneID = u.geneName and u.UniprotID = p.ID_Uniprot and n.exonCount >= 4



geneName	PDB_ID	exonCount
TLL1	3EDI	10
TLL1	3EDI	21
UBE4B	2KRE	28
UBE4B	2KRE	27
KRAS	1D8D	6
KRAS	1D8D	5
KRAS	1D8D	5
KRAS	1D8D	6
APC	1DEB	16
APC	1DEB	17
APC	1DEB	14
APC	1DEB	16
APC	1DEB	17
APC	1DEB	15
APC	1DEB	16
APC	1DEB	15
APC	1DEB	16

<u>Legend</u>



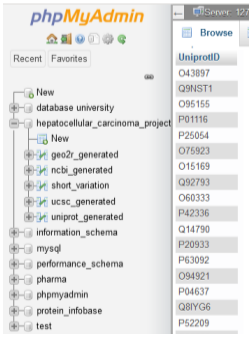
Conserved Region

The PDB_ID is *'2KRE'*

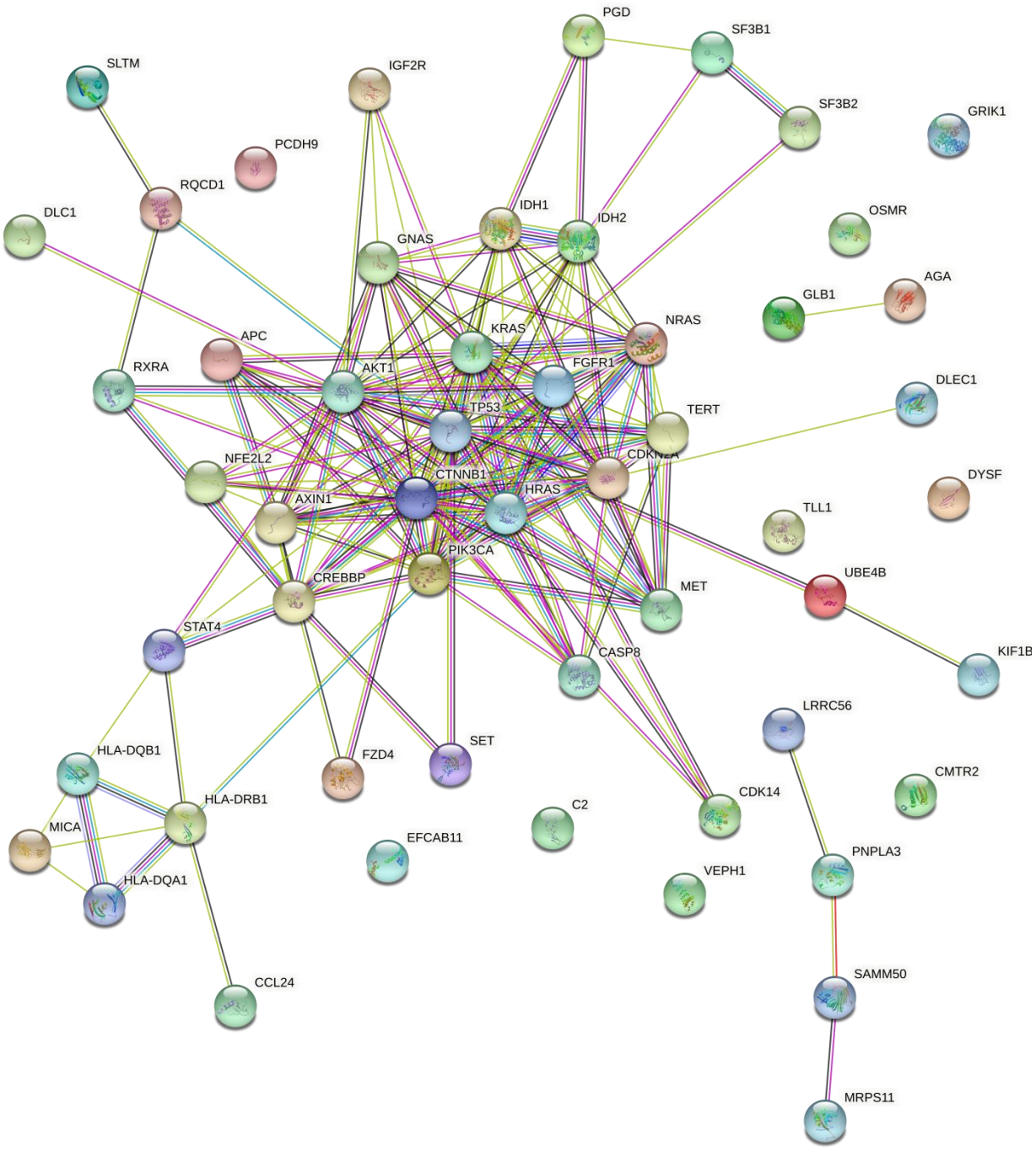
3. Count genes with adjusted P-value less than 0.05
   SELECT COUNT(DISTINCT Gene_symbol)
   FROM geo2r_generated as g
   WHERE g.adj_P_Val < 0.05 and g.Gene_symbol != ""

   They are 221 genes.

4. Retrieve Uniprot IDs
   SELECT UniprotID
   FROM geo2r_generated as g, ucsc_generated as u
   WHERE g.adj_P_Val < 0.05 and g.Gene_symbol != ""

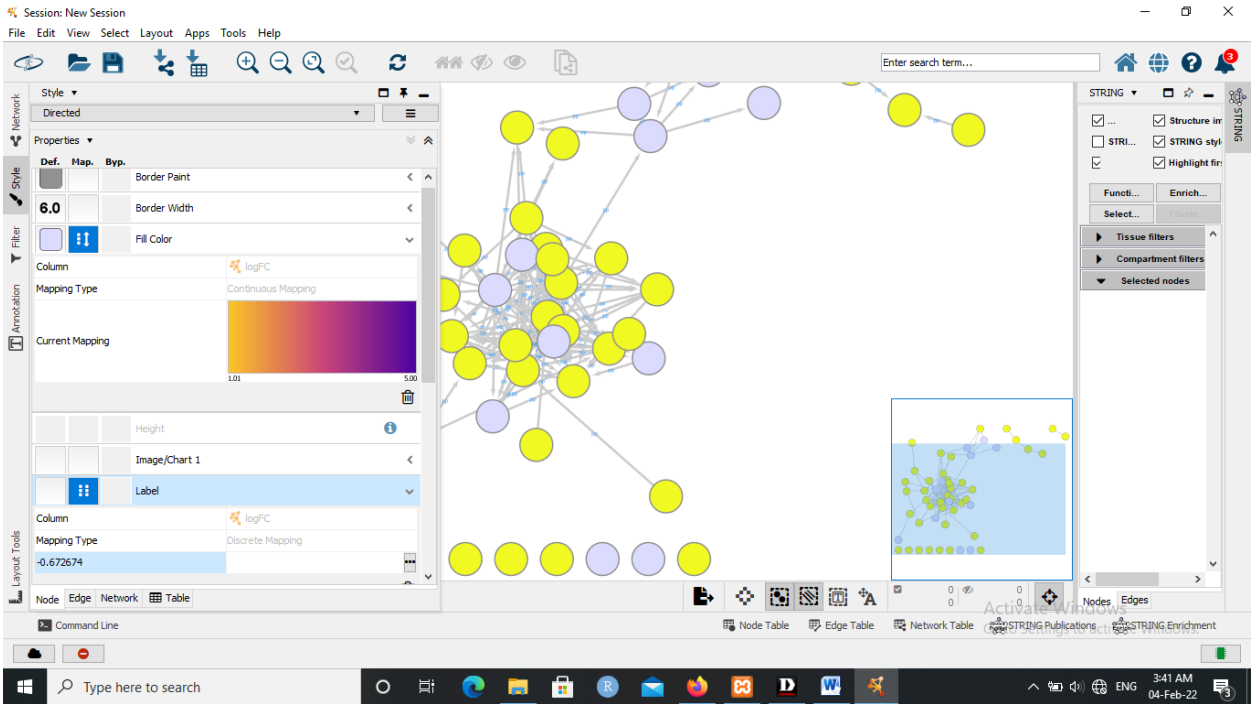


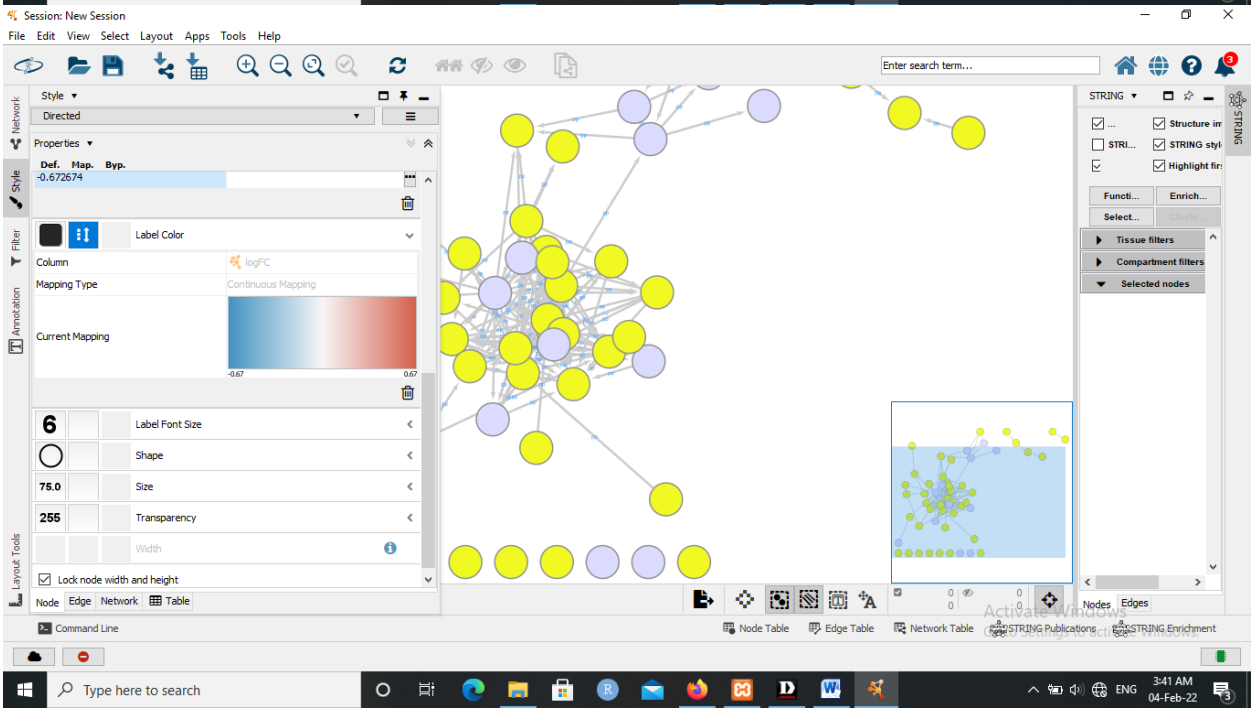
   The Image from STRING shown below.



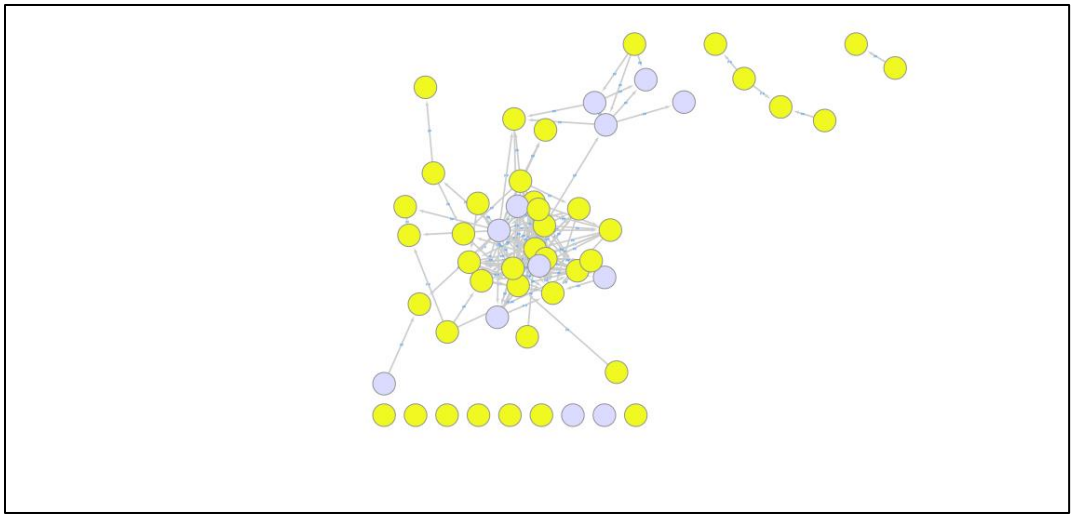
For generation of a network in cytoscape I took the following steps. Firstly, I exported my network from STRING to cytoscape using the stringApp. Secondly, I made a query in my database to get all the UniprotIDs and logFC columns. Then I exported the table into a csv file and imported it into the current node network in cytoscape.

Below are some snapshots of the cytoscape results;





The string network from cytoscape:



Legend

Up regulated

Down regulated