

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 phenotype, 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_reference
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	
rs786203436	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;

phpMyAdmin

Server: 127.0.0.1 Database: hepatocellular_carcinoma_project Table: ucsc_generated

#chrom	chromStart	chromEnd	name	score	strand	thickStart	thickEnd	reserved	blockCount	blockSizes
chr1	114704468	114716771	ENST00000369535.5	0	-	114708534	114716160	789624	7	3582,39,124,160,179,128,114,
chr3	38039204	38122737	ENST00000346219.7	0	+	38039225	38122605	789624	36	432,151,111,200,221,79,88,174,137,93,91,163,193,5
chr3	38039207	38124025	ENST00000308059.11	0	+	38039225	38122412	789624	37	429,151,111,200,221,79,88,174,137,93,91,163,193,5
chr3	41194740	41239949	ENST00000643541.1	0	+	41224068	41239342	789624	16	201,76,61,132,228,254,239,202,145,104,339,159,120,15
chr3	41194836	41224711	ENST00000426215.5	0	+	41224533	41224711	789624	5	105,79,61,132,186,
chr3	41194847	41239970	ENST00000645210.1	0	+	41224068	41239342	789624	16	94,79,61,132,228,254,239,202,145,104,339,159,120,151
chr3	41194852	41239900	ENST00000646381.1	0	+	41224533	41239342	789624	17	89,76,61,132,228,254,239,202,145,104,339,159,120
chr3	41194867	41239904	ENST00000405570.6	0	+	41224068	41239342	789624	17	74,76,61,132,228,254,239,202,145,104,339,159,120,151
chr3	41194892	41239900	ENST00000642248.1	0	+	41224068	41239342	789624	17	49,79,61,132,228,254,239,202,145,104,339,159,120,151
chr3	41194993	41239971	ENST00000433400.6	0	+	41224068	41239342	789624	16	713,61,228,254,239,202,145,104,339,159,120,151,1
chr3	41198644	41239945	ENST00000642992.1	0	+	41224068	41239342	789624	15	137,61,228,254,239,202,145,104,339,159,120,151,1
chr3	41198678	41239899	ENST00000431914.6	0	+	41224068	41239342	789624	16	103,61,228,254,239,202,145,104,339,159,120,151,1
chr3	41198704	41239971	ENST00000642836.1	0	+	41224533	41239342	789624	16	77,61,132,228,254,239,202,145,104,339,159,120,15
chr3	41199388	41240443	ENST00000396185.8	0	+	41224068	41239342	789624	16	282,61,228,254,239,202,145,104,339,159,120,151,1
chr3	41199416	41239971	ENST00000642986.1	0	+	41224533	41239342	789624	17	254,61,132,228,38,105,239,202,145,104,339,159,12
chr3	41199418	41239885	ENST00000644678.1	0	+	41224533	41239342	789624	16	252,61,132,228,254,239,202,145,104,339,159,120,1
chr3	41199433	41240443	ENST00000396183.7	0	+	41224068	41239342	789624	16	237,61,228,254,239,202,145,104,339,159,120,151,1
chr3	41199433	41240445	ENST00000643031.1	0	+	41224068	41239342	789624	16	237,61,228,254,239,202,145,104,339,159,120,151,1
chr3	41199441	41239955	ENST00000645982.1	0	+	41224068	41239342	789624	16	229,61,228,254,239,202,145,104,339,159,120,151,1
chr3	41199444	41239947	ENST00000453024.6	0	+	41224533	41239342	789624	17	226,61,132,228,254,239,202,145,104,339,159,120,1
chr3	41199469	41239675	ENST00000647264.1	0	+	41224533	41239342	789624	16	201,61,132,228,254,239,202,145,104,339,93,120,15
chr3	41199469	41239947	ENST00000645900.1	0	+	41224533	41239342	789624	16	153,61,132,228,254,239,202,145,104,339,159,120,1
chr3	41199471	41239743	ENST00000646174.1	0	+	41224533	41239342	789624	14	199,228,254,239,202,145,104,339,159,120,151,122,
chr3	1199478	41239900	ENST00000643992.1	0	+	41224068	41239342	789624	16	192,61,228,254,239,202,145,104,339,159,120,151,1

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;

Entry type: clear Summary - 20 per page - Sort by Default order -

Send to: Filters: Manage Filters

Top Organisms [Tree]

Homo sapiens (11)

Find related data

Database: Select

Find items

Search details

("Homo sapiens"[Organism] AND "2010/01/01"[PDAT] : "3000"[PDAT]) AND ("carcinoma, hepatocellular"[MeSH Terms] OR Hepatocellular Carcinoma[All Fields]) AND "gds"[Filter]

Search

See more...

Important Links

GEO Home

GEO Documentation

About GEO DataSets

Construct a Query

Download Options

Search results

Items: 11

Filters activated: DataSets. Clear all to show 38003 items.

1. [Quinazolinone RVX-208 effect on liver carcinoma cell line](#)

Analysis of HepG2 liver carcinoma cells treated with the quinazolinone RVX-208 or the triazolothienodiazepine JQ1, a BET bromodomain (BD) inhibitor. BET protein dysfunction is linked to cancer. Results provide insight into the difference in affinity of RVX-208 for the 1st and 2nd BDs in BET proteins.

Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 3 agent sets
Platform: GPL6244 Series: GSE51143 8 Samples
Download data: CEL
DataSet Accession: GDS5340 ID: 5340
[PubMed](#) [Full text in PMC](#) [Similar studies](#) [GEO Profiles](#) [Analyze DataSet](#)

2. [Hepatocellular carcinoma peripheral blood mononuclear cells](#)

Analysis of peripheral blood mononuclear cells from hepatocellular carcinoma (HCC), pancreatic carcinoma, and gastric carcinoma patients. Results provide insight into a blood-based gene signature for detection of early-stage HCC.

Organism: Homo sapiens
Type: Expression profiling by array, transformed count, 4 disease state sets
Platform: GPL570 Series: GSE49515 26 Samples
Download data: CEL
DataSet Accession: GDS4882 ID: 4882
[PubMed](#) [Similar studies](#) [GEO Profiles](#) [Analyze DataSet](#)

3. [IL-28B polymorphism effect on hepatitis C virus-related hepatocellular carcinoma](#)

resected liver

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 Set Molecular targeting of CSN5 in human hepatocellular carcinoma: a mechanism of therapeutic response

Selected 16 out of 16 samples

Group	Accession	Title	Source name	Cell line	Cell type	Slma
test	GSM951284	Negative Control siRNA, HepG2, rep1	HepG2, control siRNA	HepG2	liver cancer	control
test	GSM951285	Negative Control siRNA, HepG2, rep2	HepG2, control siRNA	HepG2	liver cancer	control
test	GSM951286	Negative Control siRNA, HepG2, rep3	HepG2, control siRNA	HepG2	liver cancer	control
test	GSM951287	Negative Control siRNA, HepG2, rep4	HepG2, control siRNA	HepG2	liver cancer	control
test	GSM951288	CSN5 siRNA, HepG2, rep1	HepG2, CSN5 siRNA	HepG2	liver cancer	CSN5
test	GSM951289	CSN5 siRNA, HepG2, rep2	HepG2, CSN5 siRNA	HepG2	liver cancer	CSN5
test	GSM951290	CSN5 siRNA, HepG2, rep3	HepG2, CSN5 siRNA	HepG2	liver cancer	CSN5
test	GSM951291	CSN5 siRNA, HepG2, rep4	HepG2, CSN5 siRNA	HepG2	liver cancer	CSN5
control	GSM951292	Negative Control siRNA, Huh7, rep1	Huh7, control siRNA	Huh7	liver cancer	control
control	GSM951293	Negative Control siRNA, Huh7, rep2	Huh7, control siRNA	Huh7	liver cancer	control
control	GSM951294	Negative Control siRNA, Huh7, rep3	Huh7, control siRNA	Huh7	liver cancer	control
control	GSM951295	Negative Control siRNA, Huh7, rep4	Huh7, control siRNA	Huh7	liver cancer	control
control	GSM951296	CSN5 siRNA, Huh7, rep1	Huh7, CSN5 siRNA	Huh7	liver cancer	CSN5
control	GSM951297	CSN5 siRNA, Huh7, rep2	Huh7, CSN5 siRNA	Huh7	liver cancer	CSN5
control	GSM951298	CSN5 siRNA, Huh7, rep3	Huh7, CSN5 siRNA	Huh7	liver cancer	CSN5

The analysis is shown on the screenshot below:

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	COP55	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:

ID	adj.P.Val	P.Value	t	B	logFC	Gene.symbol	Gene.title
ILMN_1736002	4.70E-12	3.90E-1	-32.904510	23.5293	-3.17001	COPS5	COP9 signalosome subunit 5
ILMN_2220187	1.24E-08	2.06E-1	-19.006041	17.7940	-2.043830	GFPT1	glutamine-fructose-6-phosphate transaminase 1
ILMN_1718932	1.39E-08	3.45E-1	-18.378987	17.3831	-2.124738	MTRR	5-methyltetrahydrofolate-homocysteine methyltransf...
ILMN_2342033	1.20E-07	4.50E-1	-15.522000	15.2431	-1.909775		
ILMN_1683044	1.20E-07	4.98E-1	-15.418586	15.1563	-1.953919	PPP1R2	protein phosphatase 1 regulatory inhibitor subunit...
ILMN_1728984	4.37E-07	2.17E-1	-13.974594	13.8675	-1.156247	PA2G4	proliferation-associated 2G4
ILMN_1747020	7.94E-07	4.61E-1	-13.282113	13.1936	-1.546616	SGK3	serum/glucocorticoid regulated kinase family membe...
ILMN_1802615	9.56E-07	6.64E-1	-12.956367	12.863	-1.756715	CDK6	cyclin dependent kinase 6
ILMN_1716596	9.56E-07	7.13E-1	-12.893662	12.7983	-1.304146	NSMAF	neutral sphingomyelinase activation associated fac...
ILMN_2131861	1.44E-06	1.19E-0	12.4491101	12.3298	1.2642969	SOCS2	suppressor of cytokine signaling 2
ILMN_1730670	1.54E-06	1.40E-0	12.311838	12.1815	1.4360755	FSTL3	folliculin like 3
ILMN_1651354	3.02E-06	3.01E-0	-11.680517	11.4768	-2.617077	SPP1	secreted phosphoprotein 1
ILMN_1697286	3.20E-06	3.45E-0	-11.569600	11.3490	-1.525076		
ILMN_1801124	3.33E-06	4.11E-0	-11.430141	11.1866	-1.719007	MSANTD4	Myb/SANT DNA binding domain containing 4 with coil...
ILMN_1672148	3.33E-06	4.14E-0	-11.424188	11.1797	-2.798685	AKR1B10	aldo-keto reductase family 1 member B10
ILMN_2168992	4.61E-06	6.32E-0	-11.090419	10.7827	-1.215637	FBXO30	F-box protein 30
ILMN_1704305	4.61E-06	6.49E-0	-11.069722	10.7577	-1.234249	NIP7	NIP7, nucleolar pre-rRNA processing protein
ILMN_2150402	6.77E-06	1.01E-0	-10.731909	10.3432	-2.496493	TMEM64	transmembrane protein 64
ILMN_1674706	7.00E-06	1.14E-0	-10.640052	10.2284	-1.256394	MTHFD2	methylenetetrahydrofolate dehydrogenase (NADP+ dep...
ILMN_1769282	7.00E-06	1.16E-0	-10.627362	10.2124	-1.513772	FRMD6	FERM domain containing 6
ILMN_1670353	8.34E-06	1.45E-0	-10.459132	9.9949	-1.453903	RAD51AP1	RAD51 associated protein 1
ILMN_1652736	8.63E-06	1.57E-0	-10.399558	9.92331	-2.862927	RPS6KA3	ribosomal protein S6 kinase A3
ILMN_1728049	1.25E-05	2.38E-0	10.0959019	9.52873	1.1638452	S100A16	S100 calcium binding protein A16
ILMN_2157951	1.47E-05	2.92E-0	-9.9493983	9.33453	-1.566159	STX6	syntaxin 6
ILMN_157	1.75E-05	3.63E-0	-9.7946672	9.12667	-1.051621	CANX	calnexin

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 were mapped to their respective PDB ids in the web application. The mapping table was edited and added to the database. Shown below is the table with PDB ids:

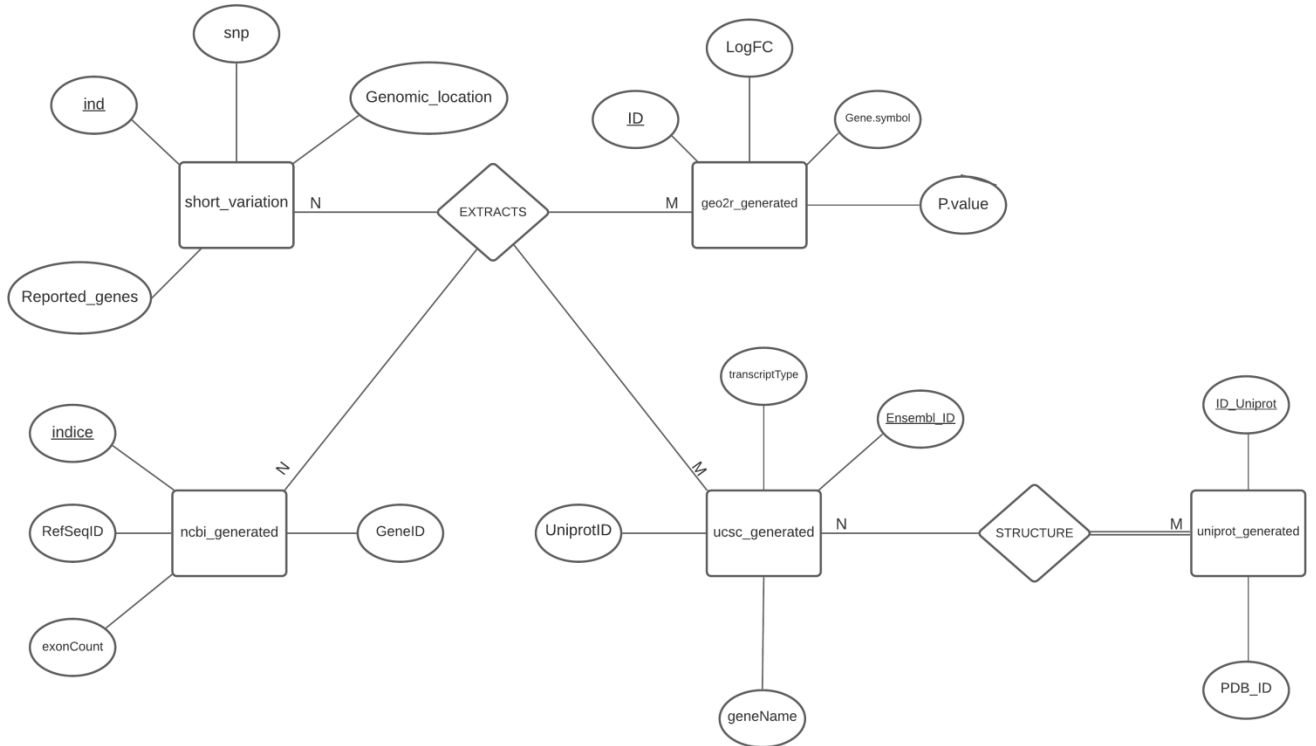
ID_Uniprot	PDB_ID
P35222	1G3J
O75533	2F9D
P82912	3J9M
P48735	4JA8
Q5JF69	2ODP
P01909	1JK8
A0A0U5Q247	5K59
Q96QB1	2DKY
Q08AS3	6PY2
P42336	2ENQ
O95155	2KRE
P01112	121P
E9PMV2	1UVQ
Q5SU54	6DIG
P19793	1BY4
Q96QC4	6DDM
O60333	2EH0
P52209	2JKV
Q29983	1B3J
Q5Y7D3	4D8P
P20933	1APY
O75923	4CAH
Q92793	1JSP
I3L466	6LQX
	2FV2

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
    
```

snp	adj_P_Val	Gene_title
rs584368	3.61E-03	
rs584368	1.23E-03	
rs584368	3.20E-06	
rs584368	6.62E-04	
rs584368	8.68E-04	
rs584368	1.07E-03	
rs584368	2.55E-04	
rs584368	4.88E-04	
rs584368	9.34E-04	
rs584368	3.71E-03	
rs584368	1.20E-07	
rs584368	2.52E-03	
rs584368	4.83E-04	
rs584368	1.95E-03	
rs6078460	3.61E-03	
rs6078460	1.23E-03	
rs6078460	3.20E-06	
rs6078460	6.62E-04	
rs6078460	8.68E-04	
rs6078460	1.07E-03	
rs6078460	2.55E-04	
rs6078460	4.88E-04	
rs6078460	9.34E-04	
rs6078460	3.71E-03	

2. Second query

```

SELECT ensembl_ID, u.strand, txEnd, txStart
FROM ncbi_generated as n, ucsc_generated as u
    
```

WHERE n.GeneID = u.geneName

The screenshot shows the phpMyAdmin interface. On the left, a tree view shows the database structure. The main panel displays the 'Structure' tab for the 'short_variation' table in the 'hepatocellular_carcinoma_project' database. The table has four columns: 'ensembl_ID', 'strand', 'txEnd', and 'txStart'. The data rows show various Ensembl IDs and their corresponding strand orientations and coordinates.

ensembl_ID	strand	txEnd	txStart
ENST00000061240.7	+	166030959	165873236
ENST00000061240.7	+	166104457	165873236
ENST00000216180.8	+	43947582	43923804
ENST00000253251.12	+	10181238	10032957
ENST00000253251.12	+	10181238	10032957
ENST00000256078.10	-	25250929	25205245
ENST00000256078.10	-	25250929	25205245
ENST00000256078.10	-	25250929	25205245
ENST00000256078.10	-	25250929	25205245
ENST00000256078.10	-	25250929	25205245
ENST00000257430.9	+	112846239	112737884
ENST00000257430.9	+	112846239	112737884
ENST00000257430.9	+	112846239	112707497
ENST00000257430.9	+	112846239	112707497
ENST00000257430.9	+	112846239	112707497
ENST00000257430.9	+	112846239	112737884
ENST00000257430.9	+	112846239	112707497
ENST00000257430.9	+	112846239	112737884
ENST00000257430.9	+	112846239	112737884

3. Third query

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

The screenshot shows the results of a query in phpMyAdmin. The table has three columns: 'Phenotype', 'Reported_genes', and 'Supporting evidence'. The results list various phenotypes related to hepatitis B and C infections and their associated genes.

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 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

- Query for PDB_ID structures with exon count greater or equal to 4

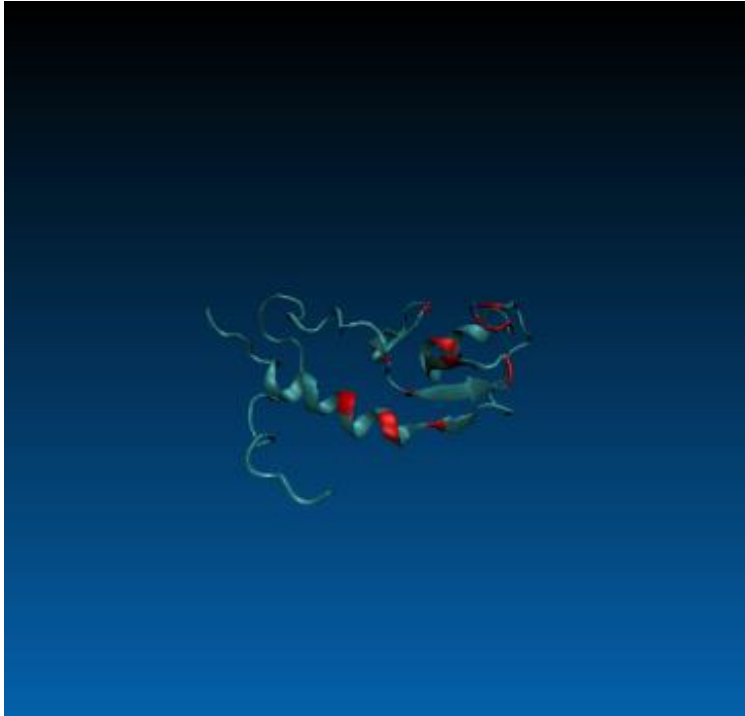
SELECT geneName, PDB_ID, exonCount

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

Legend



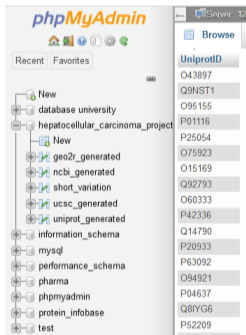
— Conserved Region

The PDB_ID is '2KRE'

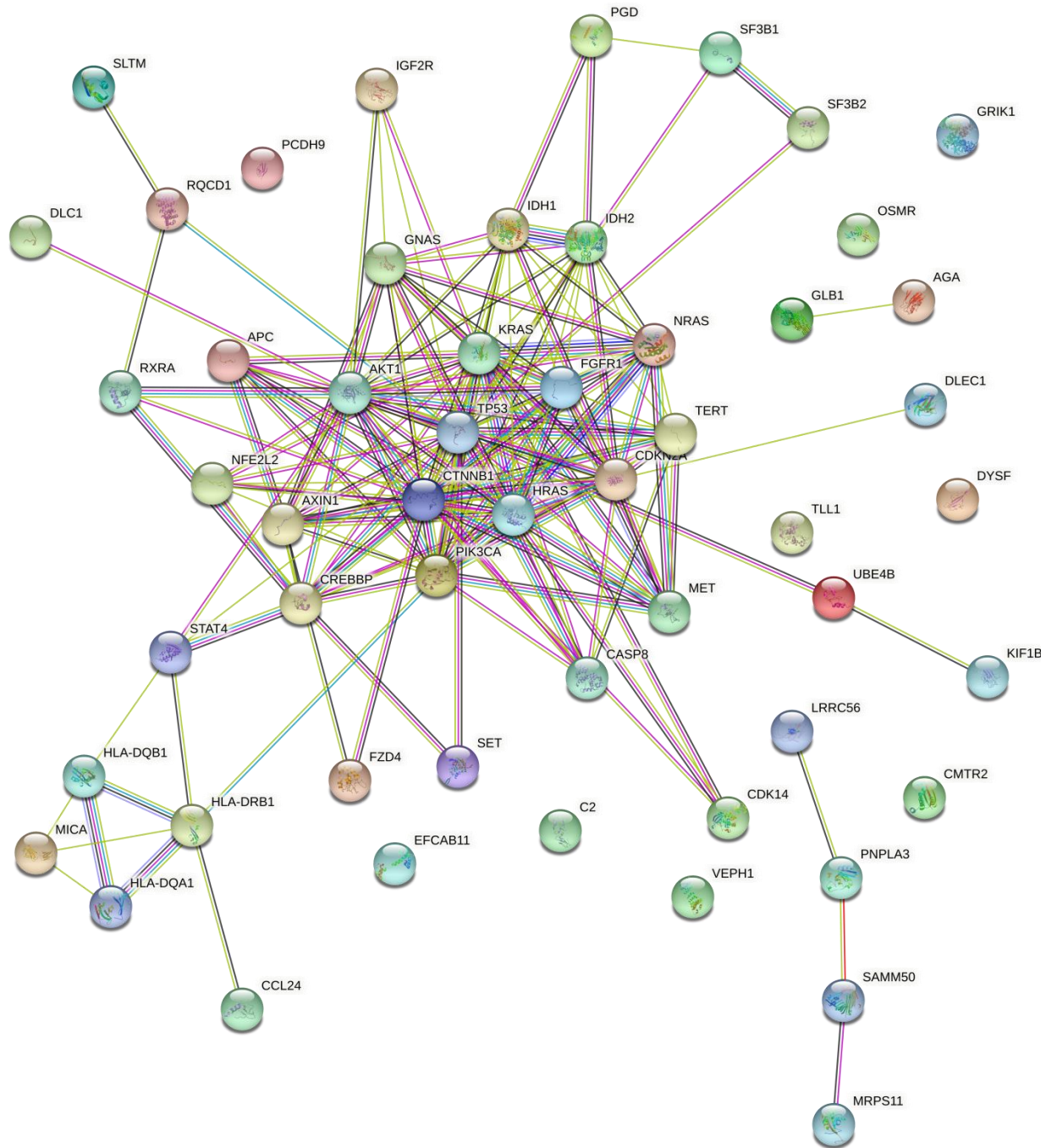
- Count genes with adjusted P-value less than 0.05
`SELECT COUNT(DISTINCT Gene_symbol)`
`FROM geo2r_generated as g`
`WHERE g.adj_P_Val < 0.05 and g.Gene_symbol != ""`

They are 221 genes.

- 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 != ""`

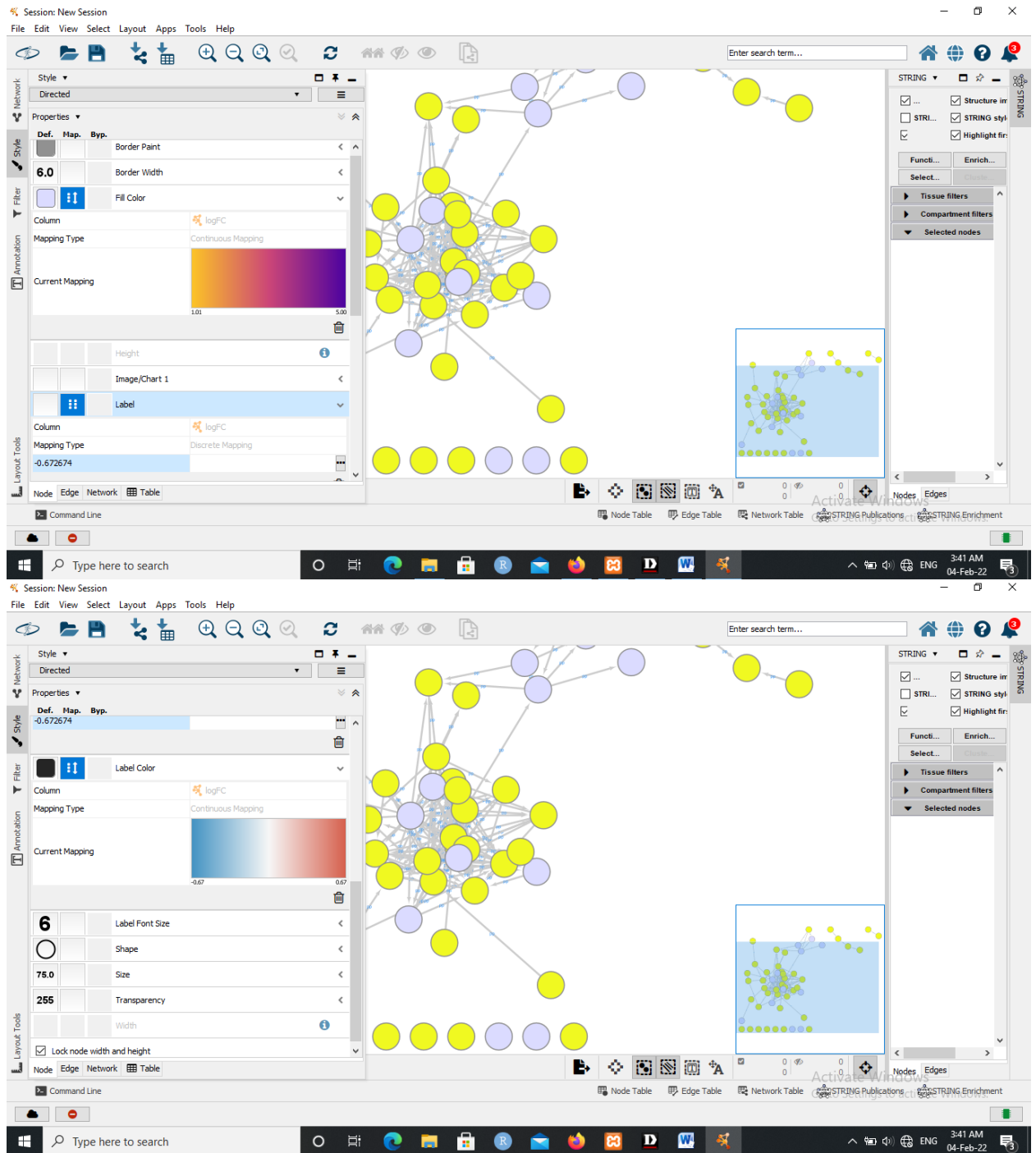


The Image from STRING shown below.

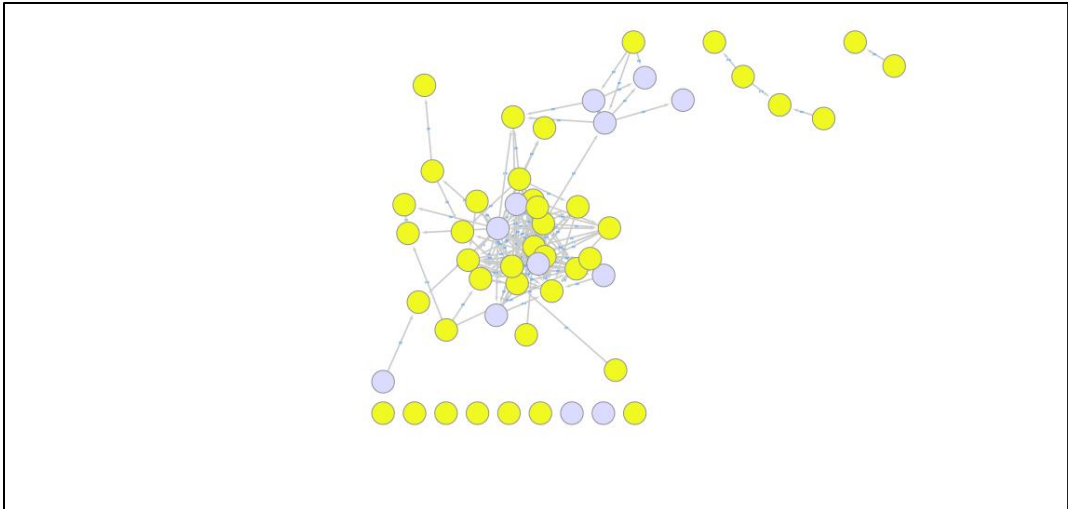


For generation of a network in cytoscape I took the following steps. Firstly, I exported my network from STRING to cytoscape using the stringApp. Secondly, I made a query in my database to get all the UniprotIDs and logFC columns. Then I exported the table into a csv file and imported it into the current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

Below are some snapshots of the cytoscape results;



The string network from cytoscape:



Legend

— Up regulated

— Down regulated