PROJECT

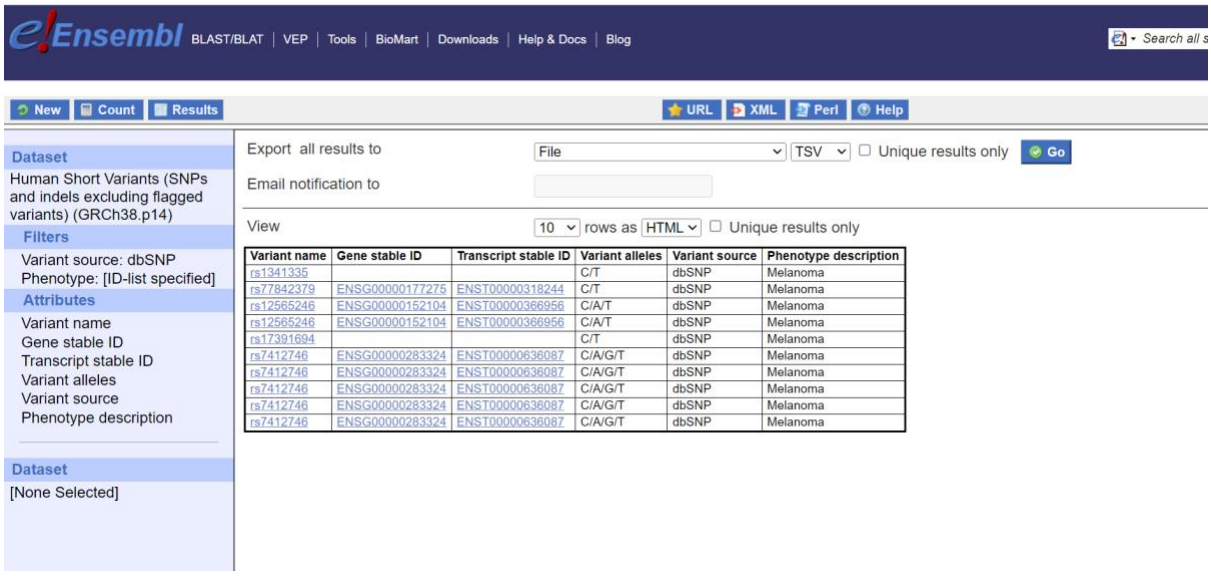
1. Introduction

Melanoma is a malignant melanocytic tumor arising from pigment-producing cells, melanocytes, primarily induced by prolonged ultraviolet exposure to the skin. The characteristics of melanoma are asymmetry, irregularities in shape and uneven colourization. (National Cancer Institute, 2022) It can appear anywhere on the skin, where melanoma is often spotted as a mole that stands out from the rest of the moles. Melanoma is known to be more common among people with pale skin, blue eyes and red or fair colored hair. The depth of melanoma is the most important prognostic factor. Two staging systems can be used to assess the depth of the mole: Breslow and Clark levels, whereas Breslow is used as the standard approach today (National Cancer Institute, 2022) A skin exam will be made if there is a suspicion of melanoma and a biopsy may be required. A biopsy removes the abnormal tissues and the tissue is examined under a microscope to look for cancer cells (National Cancer Institute, 2022)

Melanoma, compared to other types of cancer, spreads very rapidly when not treated in its early stages and there are various types of melanoma. (Skin Cancer Foundation, 2021) There are 5 stages of melanoma and the stage depends on the thickness of the tumor and whether the cancer has spread to other parts of the body(3). There are different treatment options including surgery chemotherapy, radiation therapy and immunotherapy (National Cancer Institute, 2023) Preventive measures include the use of sun protection and regular skin checks at the doctor (National Cancer Institute, 2022)

2. Retrieve short variations table

The dbSNP database retrieves 0 results when searching for “Melanoma”. To retrieve the needed data we used the ENSEMBL Biomart database. Choosing ‘Human Short Variants (SNPs and indels excluding flagged variants’ followed by filtering for Phenotype: ‘Melanoma’ and choosing dbSNP as the variant source. We chose the attribute columns ‘Variant name’, Gene stable ID’, ‘Transcript stable ID, ‘Variant alleles’, ‘Variant source’, ‘Phenotype description’ as shown in figure 1.



Variant name	Gene stable ID	Transcript stable ID	Variant alleles	Variant source	Phenotype description
rs1341335			C/T	dbSNP	Melanoma
rs77842379	ENSG00000177275	ENST00000318244	C/T	dbSNP	Melanoma
rs12565246	ENSG00000152104	ENST00000366956	C/A/T	dbSNP	Melanoma
rs12565246	ENSG00000152104	ENST00000366956	C/A/T	dbSNP	Melanoma
rs17391694			C/T	dbSNP	Melanoma
rs7412746	ENSG00000283324	ENST00000636087	C/A/G/T	dbSNP	Melanoma
rs7412746	ENSG00000283324	ENST00000636087	C/A/G/T	dbSNP	Melanoma
rs7412746	ENSG00000283324	ENST00000636087	C/A/G/T	dbSNP	Melanoma
rs7412746	ENSG00000283324	ENST00000636087	C/A/G/T	dbSNP	Melanoma
rs7412746	ENSG00000283324	ENST00000636087	C/A/G/T	dbSNP	Melanoma

Figure 1 - Ensembl / Biomart interface

3. **Retrieve table of genomic coordinates and related information of the RefSeq genes.**

To retrieve genomic coordinates and other related information we used the Ensembl gene ID's we retrieved in part 2 (Figure 1).

Data retrieval steps:

1. Enter the UCSC table browser website and under 'Select dataset' category we chose:
    a. 'Human' in the genome option
    b. 'Dec.2013 (GRCh38/hg38)' in assembly
    c. 'Genes and Gene Predictions' under group
    d. 'ALL GENCODE V44' under track
    e. 'Basic (qgEncodeGencodeBasicV44) under table

2. In the 'Define Region of interest' category we wanted to upload a list of Ensembl Gene ID's retrieved from our short_variants table. But the UCSC data format required the identifiers to be in the Ensembl gene Stable ID version's format. We therefore went back to Ensembl Biomart to retrieve the Ensembl stable ID version by uploading a list of Ensembl Gene ID's from our short_variants table from Biomart (figure 1) We could then upload this new list of Ensembl stable ID version to USCS table browser in the 'Define region of interest' category (figure 2).

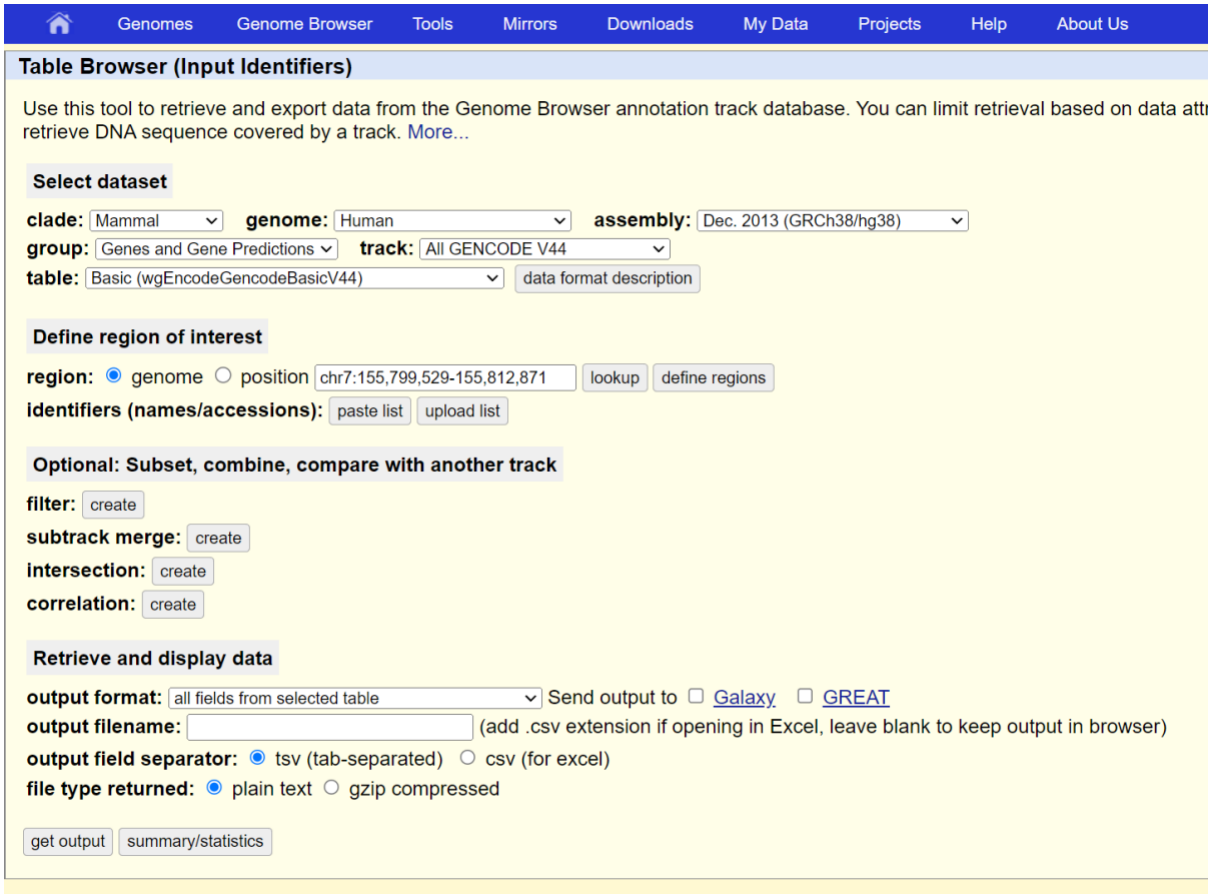


Figure 2 - UCSC Table Browser interface

3. Table from analysis with GEO2R.

Data retrieval steps:

1. Enter NCBI website and choose option 'GEO DataSets'
2. Search for "Melanoma" and add the following filters:
    a. Choose "Homo Sapiens" under 'Organisms'
    b. publication data from January 1, 2010
3. Choosing the dataset: 'Characterization of macrophages influence in human melanoma' (National Center for Biotechnology Information, 2023)
4. Description of dataset:

In tumors, certain cells (macrophages) interact with cancer cells like melanoma. These interactions affect the behavior of these immune cells by turning them into tumor-associated macrophages (TAMs). TAMs often help tumors grow and hide from the immune system. This experiment aims to understand how these interactions change the behavior of macrophages to find new treatment ways for skin cancer by targeting these changes. The experiment used

the method type: “Expression profiling by high throughput sequencing”. Overall design includes growing melanoma cells in a lav alongside macrophages created from a type of immune cell called CD14+ monocytes from human blood cells. These macrophages were made in the lab using the following substances: GM-CSF or M-CSF for seven days in total.

5. To analyze it with GEO2R, we defined 2 groups: Control group (3 samples) and Coculture GM-CSF and M-CSF and retrieved a table, which could then be imported to our database in phpMyAdmin.
6. Analyzing with GEO2R enables us to compare the sample groups to identify genes that are differentially expressed across the two experimental conditions. And the result is presented as a table of genes expression levels between the groups ordered by significance (P-value).

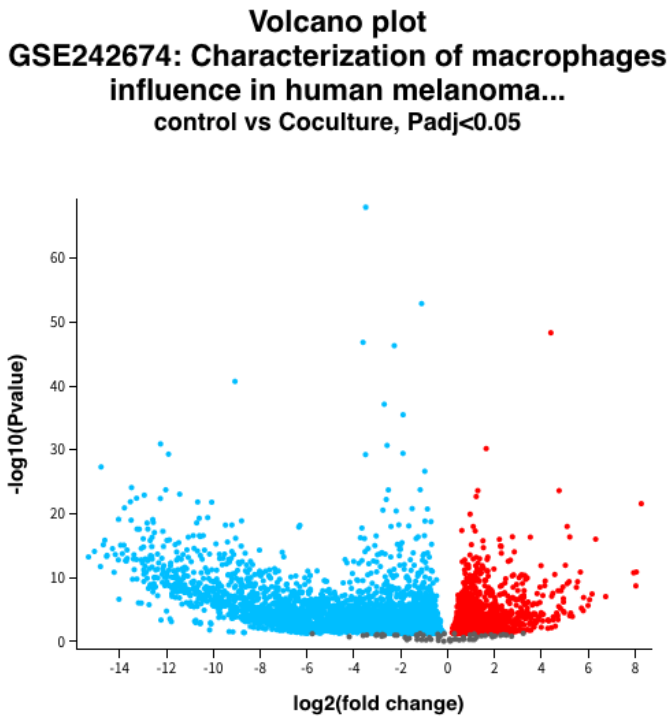


Figure 3 - Vulcano Plot from Geo2R

The vulcano plot (figure 3) gives insight into the experiment ‘Characterization of macrophages influence in human melanoma’ (National Center for Biotechnology Information, 2023) by representing the statistical significance and fold changes of gene expression between the control group vs. Coculture group. The red points in the graph represent genes that are significantly upregulated, and the blue points represent genes that are significantly downregulated.

5. Table with Melanoma related genes mapped to human proteins

Data retrieval steps:

1. Enter UniProt and click on 'Id Mapping'
2. Choosing From database → Genome Annotation Databases → Ensembl and to database → UniProt → UniProtKB/Swiss-Prot
3. Loading a text file containing all the Ensembl Gene Id's retrieved from Biomart table

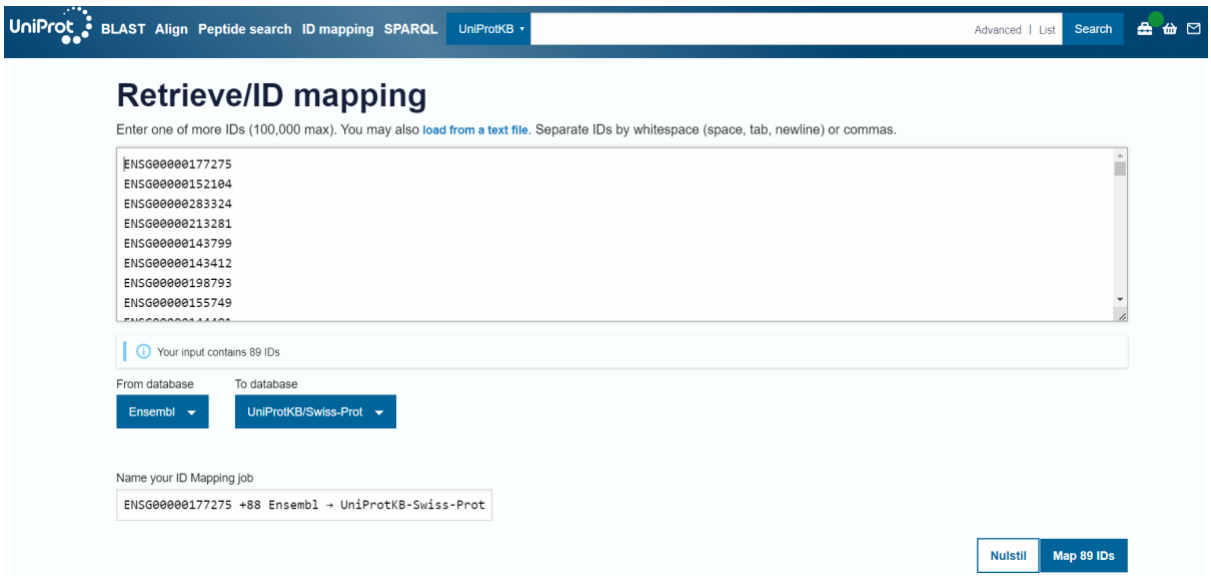


Figure 4 - UniProt interface showing ID mapping

4. Click on 'Map 89 IDs'
5. Download the Id-mapping result with 90 entries (figure 5).

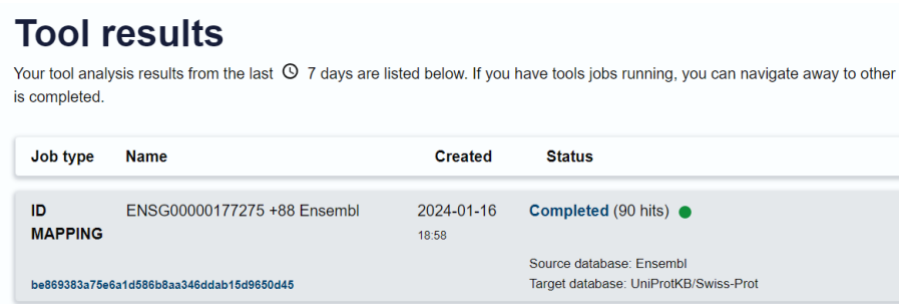


Figure 5 - The result from UniProt ID-mapping

6. Customize columns to retrieve the relevant attributes (figure 6).

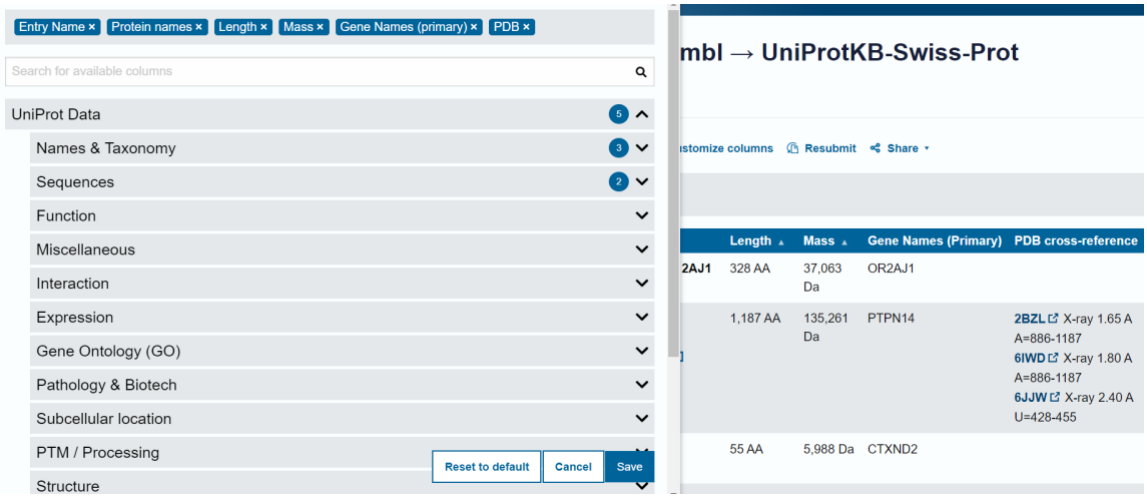


Figure 6 - filtering on attributes in UniProt ID-mapping

7. Download it as a TSV file and convert it to a CSV file.
8. Upload CSV file to our database.

6. Construct the database with the collected data:

The following tables are retrieved from the Ensembl, USCS, GEO2r and UniProt databases. The database consists of the following tables:

<u>geo2r:</u>

This table contains data related to the gene expressions. It includes two identifiers GeneID and EnsemblGeneID. Further, it contains statistical measures such as pajd, pvalue, log2FoldChange, baseMean, stat, ifcSE. The table provides insights into the differential expression of genes.

GeneID	padj	pvalue	lfcSE	stat	log2FoldChange	baseMean	Symbol	Description	EnsemblGeneID
107986948	1.00	1.00	0.4081	-0.000014	-0.000005	15.08	LOC107986948		
100288203	1.00	1.00	0.1926	0.000242	0.000047	180.14	WHAMMP4	WHAMM pseudogene 4	
83939	1.00	1.00	0.1366	-0.000004	-0.000001	3087.28	EIF2A	eukaryotic translation initiation factor 2A	ENSG00000144895
4899	1.00	1.00	0.3055	0.000051	0.000016	354.86	NRF1	nuclear respiratory factor 1	ENSG00000106459
114932	6.70e-0	1.00e-0	0.1962	3.890000	0.763000	735.82	MRFAP1L1	Morf4 family associated protein 1 like 1	ENSG00000178988
3030	1.84e-0	1.00e-0	0.0815	1.640000	0.134000	5832.74	HADHA	hydroxyacyl-CoA dehydrogenase trifunctional multie...	ENSG00000084754
389119	6.70e-0	1.00e-0	0.3048	-3.890000	-1.190000	52.19	INKA1	inka box actin regulator 1	ENSG00000185614
51307	1.84e-0	1.00e-0	0.2422	1.640000	0.398000	1539.29	FAM53C	family with sequence similarity 53 member C	ENSG00000120709
5471	3.05e-0	1.00e-0	0.5588	2.580000	1.440000	624.46	PPAT	phosphoribosyl pyrophosphate amidotransferase	ENSG00000128059
54954	1.85e-0	1.00e-0	0.4555	-1.640000	-0.748000	202.09	FAM120C	family with sequence similarity 120C	ENSG00000184083

Figure 7 - Screenshot of geo2r table from phpMyAdmin

proteins:

This table stores information about proteins, including identifiers from Ensembl and UniProt, protein names, length, mass, gene names, and PDB identifiers. The attributes cover various aspects of protein information, enabling a comprehensive representation of protein data in the database. UniProtID is the primary key for this table as it is a unique identifier for the proteins. Further, EnsemblGeneID is the foreign key in this table linking it to the short_variants table.

EnsemblGeneID	UniProtID	Entry_Name	Protein_names	Length	Mass	Gene_Names_primary	PDB
ENSG00000177275	Q8NGZ0	O2AJ1_HUMAN	Olfactory receptor 2AJ1	328	37063	OR2AJ1	
ENSG00000152104	Q15678	PTN14_HUMAN	Tyrosine-protein phosphatase non-receptor type 14 ...	1187	135261	PTPN14	2BZL;6IWD;6JJW;
ENSG00000283324	A0A1B0GV90	CTXD2_HUMAN	Cortexin domain containing 2	55	5988	CTXND2	
ENSG00000213281	P01111	RASN_HUMAN	GTPase NRas (EC 3.6.5.2) (Transforming protein N-R...	189	21229	NRAS	2N9C;3CON;5UHV;6E6H;6MPP;6ULI;6ULK;6ULN;6ULR;6UON;...

Figure 8 - Screenshot of proteins table from phpMyAdmin

refseq:

This table stores genomic information of genes, including their location on chromosomes, directionality, transcript variations, and details about exon organization. This table helps us to get an understanding of the genomic structure of genes in the context of the entire genome. The primary key of this table is the Transcript_stable_ID_version. This one is linked to the cross-reference table 'cross_ref' mentioned later in the section on Conceptual design.

#bin	Transcript_stable_ID_version	chrom	strand	txStart	txEnd	cdsStart	cdsEnd	exonCount	exonStarts
2476	ENST00000318244.4	chr1	+	247924888	247935339	247933768	247934755	2	247924888,247933746,
83	ENST00000361445.9	chr1	-	11106534	11262551	11107484	11259409	58	11106534,11108180,11109289,11109648,11112851,11114...
288	ENST00000366794.10	chr1	-	226360690	226408093	226361459	226407929	23	226360690,226361968,226363098,226363942,226365001,...
277	ENST00000366956.10	chr1	-	214348699	214551602	214357921	214464803	19	214348699,214364511,214369456,214372710,214376218,...
3	ENST00000368947.9	chr1	+	150982291	150995634	150983105	150995322	14	150982291,150982486,150983089,150983337,150983974,...
1460	ENST00000369535.5	chr1	-	114704468	114716771	114708534	114716160	7	114704468,114708153,114708530,114709568,114713799,...
277	ENST00000543945.5	chr1	-	214357750	214551223	214397946	214464803	18	214357750,214364511,214369456,214372710,214376218,...
1736	ENST00000636087.1	chr1	+	150887135	150913292	150912314	150912482	2	150887135,150912241,
288	ENST00000677203.1	chr1	-	226360696	226408100	226361459	226407929	22	226360696,226361968,226363098,226363942,226365001,...
83	ENST00000703140.1	chr1	-	11106534	11262530	11107484	11259409	56	11106534,11108180,11109289,11109648,11112851,11114...

exonEnds	score	symbol	cdsStartStat	cdsEndStat	exonFrames
247925168,247935339,	0	OR2AJ1	cmpl	cmpl	-1,0,
11107500,11108286,11109370,11109729,11112917,11114...	0	MTOR	cmpl	cmpl	2,1,1,1,1,0,0,2,0,0,2,1,0,0,0,0,0,2,0,0,2,0,0,0,0,...
226361541,226362083,226363160,226364070,226365154,...	0	PARP1	cmpl	cmpl	2,1,2,0,0,0,0,0,0,0,2,1,1,1,1,0,0,0,2,0,1,0,0,
214358050,214364675,214369691,214372839,214376437,...	0	PTPN14	cmpl	cmpl	0,1,0,0,0,0,1,0,2,0,2,0,2,0,1,2,0,0,-1,
150982379,150982583,150983180,150983434,150984069,...	0	ANXA9	cmpl	cmpl	-1,-1,0,0,1,0,0,1,0,0,1,1,0,0,
114708050,114708192,114708654,114709728,114713978,...	0	NRAS	cmpl	cmpl	-1,-1,0,2,0,0,-1,
214358050,214364675,214369691,214372839,214376437,...	0	PTPN14	cmpl	cmpl	-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1,2,0,1,2,0,0,-1,
150887313,150913292,	0	CTXND2	cmpl	cmpl	-1,0,
226361541,226362083,226363160,226364070,226365154,...	0	PARP1	cmpl	cmpl	2,1,2,0,0,0,0,0,0,2,1,1,1,1,0,0,0,2,0,1,0,0,

Figure 9 - - Screenshot of refseq table from phpMyAdmin

short_variants:

This table contains information about genetic variations, the SNPs and indels. The Variant_name is the primary key as it is a unique identifier. Further is the Transcript_stable_ID and Gene_stable_ID foreign keys associated with the genomic region, in this case the region affected by the variation. From these keys it is possible to link the table with another table.

Variant_name	Variant_source	Phenotype_description	Gene_stable_ID	Transcript_stable_ID	Variant_alleles
rs1341335	dbSNP	Melanoma			C/T
rs77842379	dbSNP	Melanoma	ENSG00000177275	ENST00000318244	C/T
rs12565246	dbSNP	Melanoma	ENSG00000152104	ENST00000366956	C/A/T
rs17391694	dbSNP	Melanoma			C/T
rs7412746	dbSNP	Melanoma	ENSG00000283324	ENST00000636087	C/A/G/T
rs11554290	dbSNP	Melanoma	ENSG00000213281	ENST00000369535	T/A/C/G
rs3219090	dbSNP	Melanoma	ENSG00000143799	ENST00000366794	T/C
rs61815526	dbSNP	Melanoma			G/A
rs1722784	dbSNP	Melanoma	ENSG00000143412	ENST00000368947	A/G

Figure 10 - Screenshot of short_variants table from phpMyAdmin

Additional tables from other databases?

Example 1: Pathway info

It would be interesting to look at e.g. Reactome database or another pathway database. Here we can gain further information on the pathways in which each gene is involved and thereby understand more about the biological processes and the pathways involved when macrophages influence melanoma cells. We can more contextually gain information about each gene's function in this context. Thus, by making use of pathway data, we can gain a better understanding of the complex biological system of melanoma and identify the key genes involved and are especially relevant to Melanoma and the macrophage interaction.

Example 2: clinical studies

The Genomic Data Commons (GDC)[1] database would be interesting to use for the melanoma phenotype, since it stores genomic and clinical data related to cancer research. It contains information about clinical studies, disease types, survival rates, mutations etc. for cancer

[1] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

studies. We can dive into specific case studies and gain insight into lifestyle exposure factors such, gender-specific associations and filter on specific intervals for diagnosis age and stages.

Conceptual design:

We have added a fifth table called cross_ref, as the refseq table is not linked to any of the other three tables it is necessary to create a cross reference table that can link the refseq table with short_variants. We have created the cross_ref table with the following four attributes: Gene_stable_ID,Gene_stable_ID_version,Transcript_stable_ID,and Transcript_stable_ID_version. The Transcript_stable_ID is the primary key, the unique attribute, and the Gene_stable_ID is the foreign key from the short_variants, and Transcript_stable_ID_version is the foreign key from the refseq table.

Gene_stable_ID	Gene_stable_ID_version	Transcript_stable_ID	Transcript_stable_ID_version
ENSG00000134871	ENSG00000134871.19	ENST00000360467	ENST00000360467.7
ENSG00000198646	ENSG00000198646.14	ENST00000359003	ENST00000359003.7
ENSG00000183036	ENSG00000183036.11	ENST00000328619	ENST00000328619.10
ENSG00000125970	ENSG00000125970.12	ENST00000246194	ENST00000246194.8
ENSG00000104044	ENSG00000104044.16	ENST00000354638	ENST00000354638.8
ENSG00000101079	ENSG00000101079.21	ENST00000349004	ENST00000349004.6
ENSG00000183486	ENSG00000183486.14	ENST00000330714	ENST00000330714.8
ENSG00000156052	ENSG00000156052.11	ENST00000286548	ENST00000286548.9

Figure 11- Screenshot of cross_ref table from phpMyAdmin

Following (figure 12) is an overview of the five tables, their attributes, foreign keys, primary keys and how they are linked with each other:

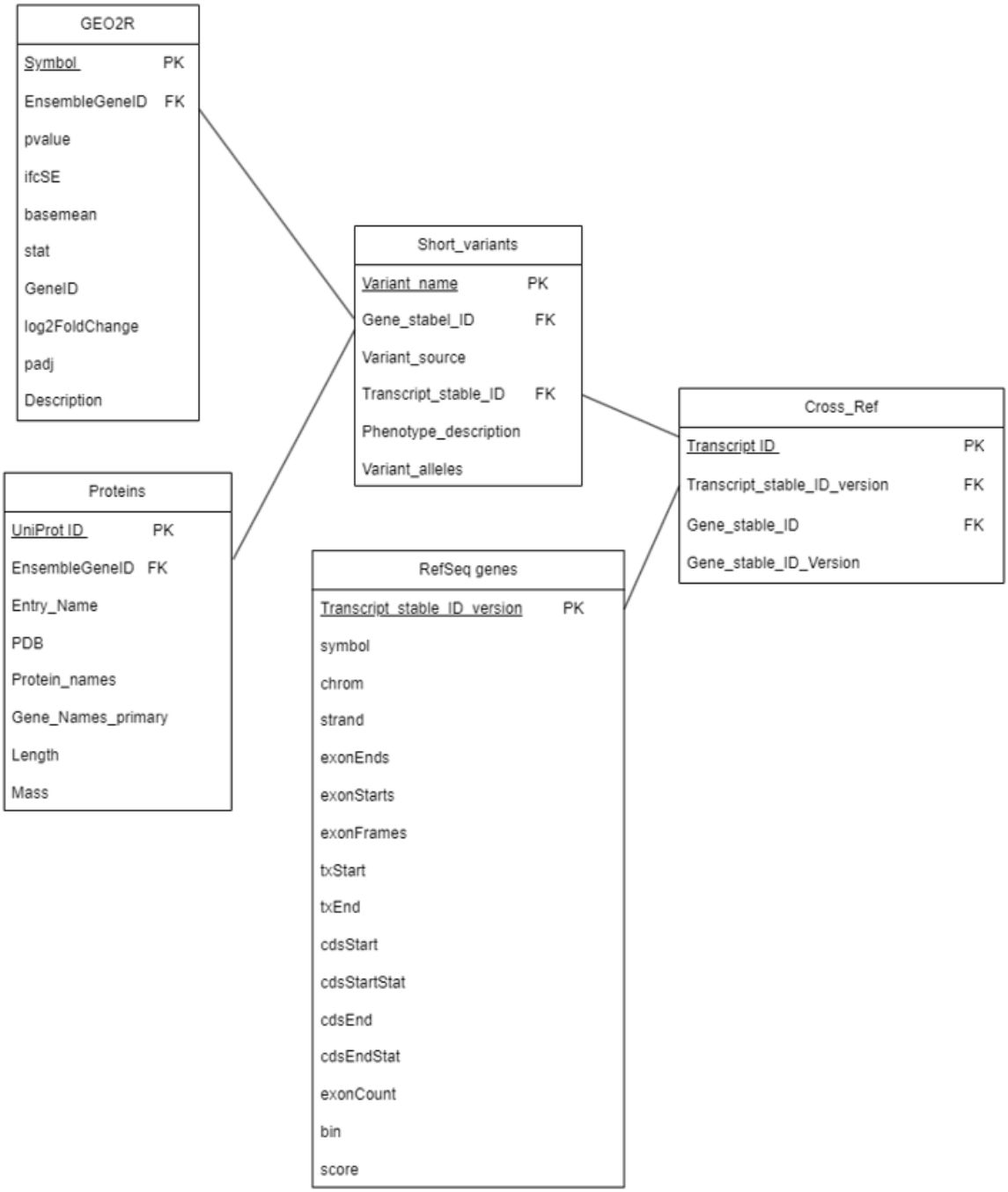


Figure 12 - all the final tables uploaded to phpMyAdmin

In the following ER-diagram (figure 13) the database is mapped out, with the five tables as entries. All the entries have relevant attributes retrieved from the biological databases in the earlier exercises. Further is the relationships between the entries mapped out, with the according cardinalities.

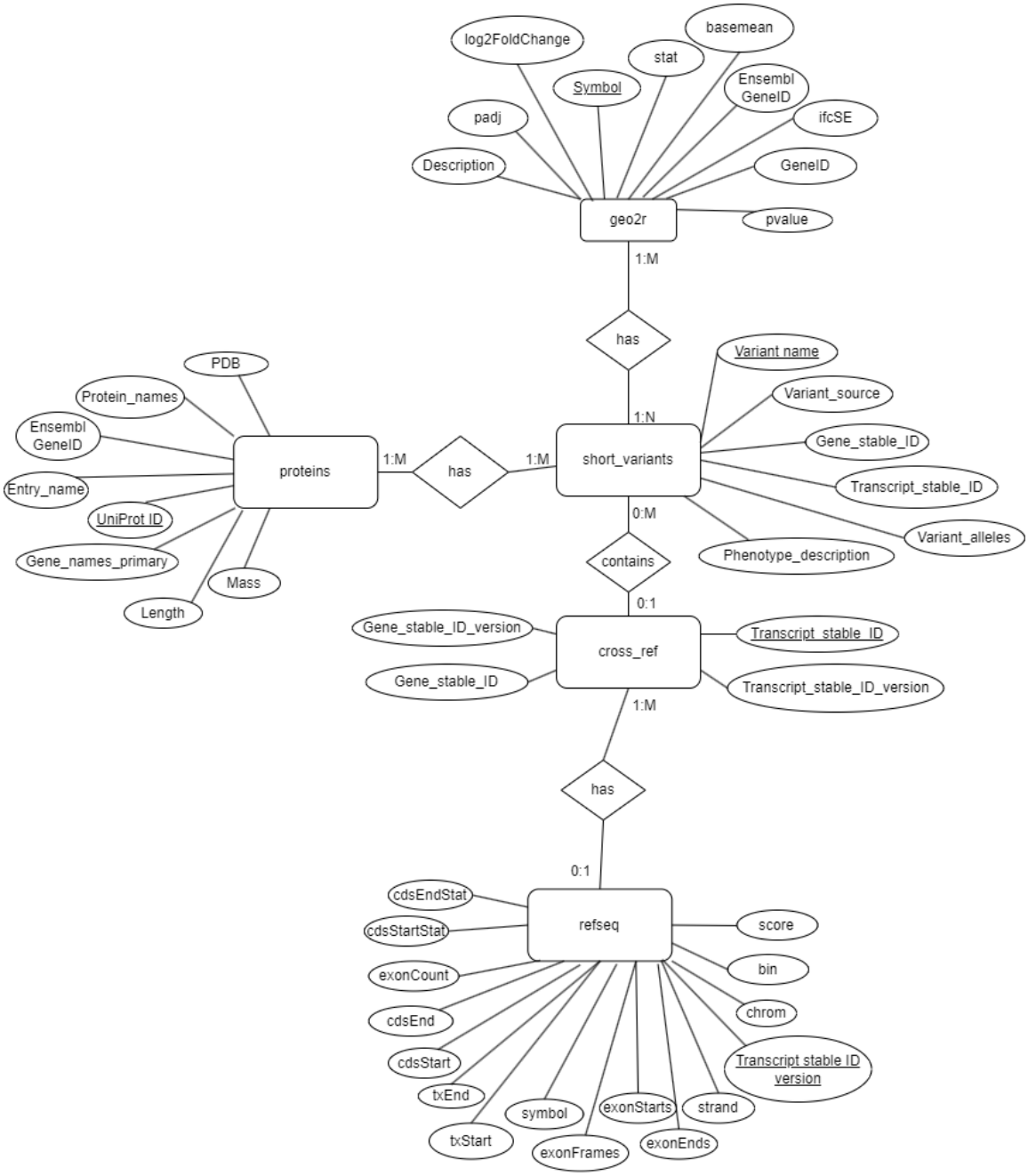


Figure 13 - ER diagram

7. Four queries and SQL commands:

<u>Query 7.1:</u> List all the variants of genes with a related protein length bigger than 1000.

SELECT short_variants.Variant_names.

FROM short_variants

JOIN proteins ON proteins.EnsemblGeneID = short_variants.Gene_stable_ID WHERE proteins.Length > 1000;

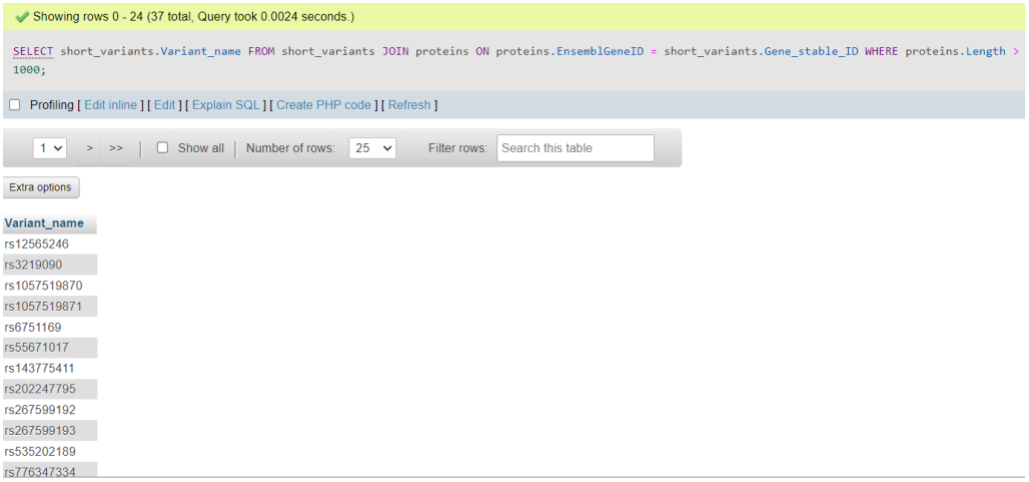


Figure 14 - query 7.1 result from phpMyAdmin

<u>Query 7.2:</u> Find the genes and their transcript Id that has less than 5 exons and are located in chromosome 2

Select cross_ref.Gene_stable_ID, refseq.Transcript_stable_ID_version
FROM cross_ref
JOIN refseq on refseq.Transcript_stable_ID_version = cross_ref.Transcript_stable_ID_version
Where refseq.exonCount > 5 and refseq.chrom = "chr2"

Figure 15 - query 7.2 result from phpMyAdmin

Query 7.3: Find the count of the Ensembl gene ID's that geo2r and proteins database have in common:

```
SELECT COUNT(proteins.EnsemblGeneID)
FROM proteins
JOIN geo2r ON geo2r.EnsemblGeneID = proteins.EnsemblGeneID
WHERE geo2r.EnsemblGeneID = proteins.EnsemblGeneID;
```

COUNT(proteins.EnsemblGeneID)
75

Figure 16 - query 7.3 result from phpMyAdmin

Query 7.4: How many variants has each protein:

```
SELECT proteins.UniProtID, COUNT(short_variants.Variant_name)
FROM proteins
JOIN short_variants ON short_variants.Gene_stable_ID = proteins.EnsemblGeneID
GROUP BY proteins.UniProtID;
```

Showing rows 0 - 24 (90 total, Query took 0.0156 seconds.)

```
SELECT proteins.UniProtID, COUNT(short_variants.Variant_name) FROM proteins JOIN short_variants ON short_variants.Gene_stable_ID =
proteins.EnsemblGeneID GROUP BY proteins.UniProtID;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> | Show all | Number of rows: 25 Filter rows: Search this table

Extra options

UniProtID	COUNT(short_variants.Variant_name)
A0A1B0GV90	1
A7E2Y1	1
C9J7I0	1
O14746	1
O15360	1
O60733	2
O75030	2
O76027	1
O94887	1
O95049	1
O95995	1
O96017	1

Figure 17 - query 7.4 result from phpMyAdmin

9. **SQL command** output: variation name, PDB IDs, Uniprot/SwissProt IDs, RefSeq IDs, exon counts, logFC value of a given gene symbol.

Refseq IDs = **protein ID, Gene Id, Genomic ID, Transcript ID**

Query 8.1:

SELECT short_variants.Variant_name, proteins.PDB, proteins.UniProtID, short_variants.Gene_stable_ID, refseq.chrom, refseq.txStart, refseq.txEnd, short_variants.Transcript_stable_ID, refseq.exonCount, geo2r.log2FoldChange
FROM short_variants
JOIN geo2r ON geo2r.EnsemblGeneID = short_variants.Gene_stable_ID
JOIN proteins ON proteins.EnsemblGeneID = short_variants.Gene_stable_ID
JOIN cross_ref ON cross_ref.Gene_stable_ID = short_variants.Gene_stable_ID
JOIN refseq ON cross_ref.Transcript_stable_ID_Version = refseq.Transcript_stable_ID_Version
WHERE refseq.symbol = "MTOR";


Showing rows 0 - 1 (2 total, Query took 0.0628 seconds.)

```
SELECT short_variants.Variant_name, proteins.PDB, proteins.UniProtID, short_variants.Gene_stable_ID, refseq.chrom, refseq.txStart, refseq.txEnd, short_variants.Transcript_stable_ID, refseq.exonCount,
geo2r.log2FoldChange FROM short_variants JOIN geo2r ON geo2r.EnsemblGeneID = short_variants.Gene_stable_ID JOIN proteins ON proteins.EnsemblGeneID = short_variants.Gene_stable_ID JOIN cross_ref ON
cross_ref.Gene_stable_ID = short_variants.Gene_stable_ID JOIN refseq ON cross_ref.Transcript_stable_ID_Version = refseq.Transcript_stable_ID_Version WHERE refseq.symbol = "MTOR";
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

Variant_name	PDB	UniProtID	Gene_stable_ID	chrom	txStart	txEnd	Transcript_stable_ID	exonCount	log2FoldChange
rs1057519870	1AUE;1FAP;1NSG;2FAP;2GAQ;2NPU;2RSE;3FAP;3JBZ;4DRH;...	P42345	ENSG00000198793	chr1	11106534	11262551	ENST00000361445	58	0.551000
rs1057519871	1AUE;1FAP;1NSG;2FAP;2GAQ;2NPU;2RSE;3FAP;3JBZ;4DRH;...	P42345	ENSG00000198793	chr1	11106534	11262551	ENST00000361445	58	0.551000

Figure 18 - query 8.1 result from phpMyAdmin

Query 8.2:

SELECT short_variants.Variant_name, proteins.PDB, proteins.UniProtID, short_variants.Gene_stable_ID, refseq.chrom, refseq.txStart, refseq.txEnd, short_variants.Transcript_stable_ID, refseq.exonCount, geo2r.log2FoldChange
FROM short_variants
JOIN geo2r ON geo2r.EnsemblGeneID = short_variants.Gene_stable_ID
JOIN proteins ON proteins.EnsemblGeneID = short_variants.Gene_stable_ID
JOIN cross_ref ON cross_ref.Gene_stable_ID = short_variants.Gene_stable_ID
JOIN refseq ON cross_ref.Transcript_stable_ID_Version = refseq.Transcript_stable_ID_Version

WHERE refseq.symbol = "FARP2";

Showing rows 0 - 0 (1 total, Query took 0.0661 seconds.)

```
SELECT short_variants.Variant_name, proteins.PDB, proteins.UniProtID, short_variants.Gene_stable_ID, refseq.chrom, refseq.txStart, refseq.txEnd,
short_variants.Transcript_stable_ID, refseq.exonCount, geo2r.log2FoldChange FROM short_variants JOIN geo2r ON geo2r.EnsemblGeneID =
short_variants.Gene_stable_ID JOIN proteins ON proteins.EnsemblGeneID = short_variants.Gene_stable_ID JOIN cross_ref ON cross_ref.Gene_stable_ID =
short_variants.Gene_stable_ID JOIN refseq ON cross_ref.Transcript_stable_ID_Version = refseq.Transcript_stable_ID_Version WHERE refseq.symbol =
"FARP2";
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

Variant_name	PDB	UniProtID	Gene_stable_ID	chrom	txStart	txEnd	Transcript_stable_ID	exonCount	log2FoldChange
rs143775411		O94887	ENSG00000006607	chr2	241356284	241494841	ENST00000264042	27	0.277000

Figure 19 query 8.2 result from phpMyAdmin

Query 8.3:

SELECT short_variants.Variant_name, proteins.PDB, proteins.UniProtID, short_variants.Gene_stable_ID, refseq.chrom, refseq.txStart, refseq.txEnd, short_variants.Transcript_stable_ID, refseq.exonCount, geo2r.log2FoldChange
FROM short_variants
JOIN geo2r ON geo2r.EnsemblGeneID = short_variants.Gene_stable_ID
JOIN proteins ON proteins.EnsemblGeneID = short_variants.Gene_stable_ID
JOIN cross_ref ON cross_ref.Gene_stable_ID = short_variants.Gene_stable_ID
JOIN refseq ON cross_ref.Transcript_stable_ID_Version = refseq.Transcript_stable_ID_Version
WHERE refseq.symbol = "FLACC1";

Showing rows 0 - 0 (1 total, Query took 0.0601 seconds.)

```
SELECT short_variants.Variant_name, proteins.PDB, proteins.UniProtID, short_variants.Gene_stable_ID, refseq.chrom, refseq.txStart, refseq.txEnd,
short_variants.Transcript_stable_ID, refseq.exonCount, geo2r.log2FoldChange FROM short_variants JOIN geo2r ON geo2r.EnsemblGeneID =
short_variants.Gene_stable_ID JOIN proteins ON proteins.EnsemblGeneID = short_variants.Gene_stable_ID JOIN cross_ref ON cross_ref.Gene_stable_ID =
short_variants.Gene_stable_ID JOIN refseq ON cross_ref.Transcript_stable_ID_Version = refseq.Transcript_stable_ID_Version WHERE refseq.symbol =
"FLACC1";
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

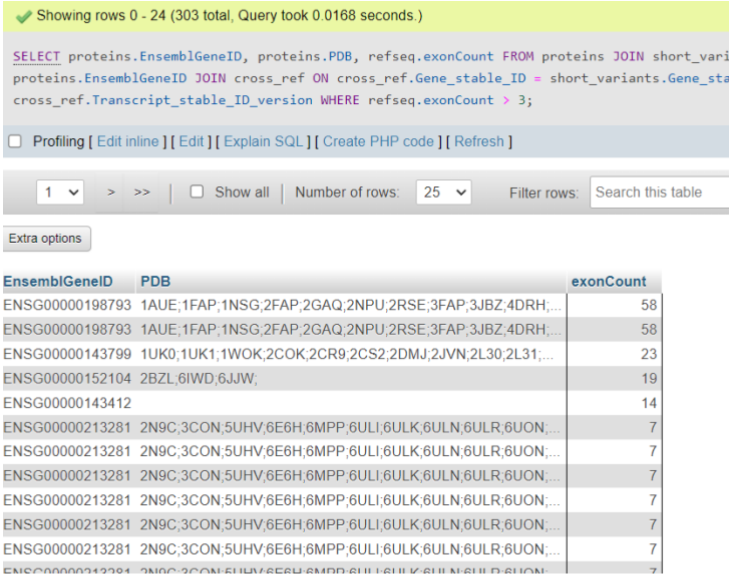
Variant_name	PDB	UniProtID	Gene_stable_ID	chrom	txStart	txEnd	Transcript_stable_ID	exonCount	log2FoldChange
s13016963		Q96Q35	ENSG00000155749	chr2	201288270	201357345	ENST00000392257	15	-0.344000

Figure 20 - query 8.3 result from phpMyAdmin

9.SQL query output: list of genes having PDB IDs and exon count is greater than or equal to 4.

Query 9.1:

```
SELECT proteins.EnsemblGeneID, proteins.PDB, refseq.exonCount
FROM proteins
JOIN short_variants ON short_variants.Gene_stable_ID = proteins.EnsemblGeneID
JOIN cross_ref ON cross_ref.Gene_stable_ID = short_variants.Gene_stable_ID
JOIN refseq on refseq.Transcript_stable_ID_version =
cross_ref.Transcript_stable_ID_version
WHERE refseq.exonCount > 3;
```



Showing rows 0 - 24 (303 total, Query took 0.0168 seconds.)

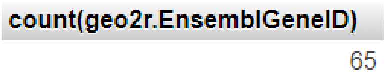
```
SELECT proteins.EnsemblGeneID, proteins.PDB, refseq.exonCount FROM proteins JOIN short_vari
proteins.EnsemblGeneID JOIN cross_ref ON cross_ref.Gene_stable_ID = short_variants.Gene_sta
cross_ref.Transcript_stable_ID_version WHERE refseq.exonCount > 3;
```

EnsemblGeneID	PDB	exonCount
ENSG00000198793	1AUE;1FAP;1NSG;2FAP;2GAQ;2NPU;2RSE;3FAP;3JBZ;4DRH;...	58
ENSG00000198793	1AUE;1FAP;1NSG;2FAP;2GAQ;2NPU;2RSE;3FAP;3JBZ;4DRH;...	58
ENSG00000143799	1UK0;1UK1;1WOK;2COK;2CR9;2CS2;2DMJ;2JVN;2L30;2L31;...	23
ENSG00000152104	2BZL;6IWD;6JJW;	19
ENSG00000143412		14
ENSG00000213281	2N9C;3CON;5UHV;6E6H;6MPP;6ULI;6ULK;6ULN;6ULR;6UON;...	7
ENSG00000213281	2N9C;3CON;5UHV;6E6H;6MPP;6ULI;6ULK;6ULN;6ULR;6UON;...	7
ENSG00000213281	2N9C;3CON;5UHV;6E6H;6MPP;6ULI;6ULK;6ULN;6ULR;6UON;...	7
ENSG00000213281	2N9C;3CON;5UHV;6E6H;6MPP;6ULI;6ULK;6ULN;6ULR;6UON;...	7
ENSG00000213281	2N9C;3CON;5UHV;6E6H;6MPP;6ULI;6ULK;6ULN;6ULR;6UON;...	7
ENSG00000213281	2N9C;3CON;5UHV;6E6H;6MPP;6ULI;6ULK;6ULN;6ULR;6UON;...	7

Figure 21 - query 9.1 result from phpMyAdmin

10. **SQL query** output: count a list of genes whose adjusted p-value in GEO2R is smaller than 0.05.

```
SELECT count(geo2r.EnsemblGeneID)
FROM geo2r
WHERE geo2r.padj < 0.05;
```



count(geo2r.EnsemblGeneID)
65

Figure 22 - query 10.1 result from phpMyAdmin

a. **How many genes are significant?** 65 genes are significant
b. **b. Write down a query to retrieve a list of Uniprot IDs of genes whose adjusted pvalue in GEO2R is smaller than 0.05 and those having a UniProt ID.**

When using 'AS DECIMAL', one result is retrieved by the query below:

```
SELECT proteins.UniProtID
FROM proteins
JOIN geo2r ON geo2r.EnsemblGeneID = proteins.EnsemblGeneID
WHERE CAST( geo2r.padj AS DECIMAL) < 0.05;
```

Figure 23 - query 10.2 result from phpMyAdmin

c+d. Find the protein interactions among the resulting list by referring to STRING and visualize these protein interactions in Cytoscape using the given instructions

Retrieving only one value, we took the liberty to increase to adjusted pvalue to 1.5 (even though it is not what the task entails) in order to work with a few more proteins in Cytoscape

```
SELECT proteins.UniProtID
FROM proteins
JOIN geo2r ON geo2r.EnsemblGeneID = proteins.EnsemblGeneID
WHERE geo2r.padj < 1.5;
```

Showing rows 0 - 8 (9 total, Query took 0.0834 seconds.)

```
SELECT proteins.UniProtID FROM proteins JOIN geo2r ON geo2r.EnsemblGeneID = proteins.EnsemblGeneID WHERE geo2r.padj < 1.5;
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

Show all | Number of rows: 25 Filter rows: Search this table

Extra options

UniProtID
P63000
P20592
O75030
Q9H444
P16234
P52926
P36507
Q9UQ90
Q9UGV2

Figure 24 - query 10.3 result from phpMyAdmin

Based on the first query we retrieved one UniProtID = P63000. This UniProtID was placed in STRING and the default 'max number of interactions to show' setting was set to 'no more than 10 interacters for the first shell. Below (figure 25) is the result of visualizing the protein interaction in Cytoscape using force-directed layout, arranging the size of the protein according to length and finally to color the genes according to the given instructions for downregulation/upregulation genes

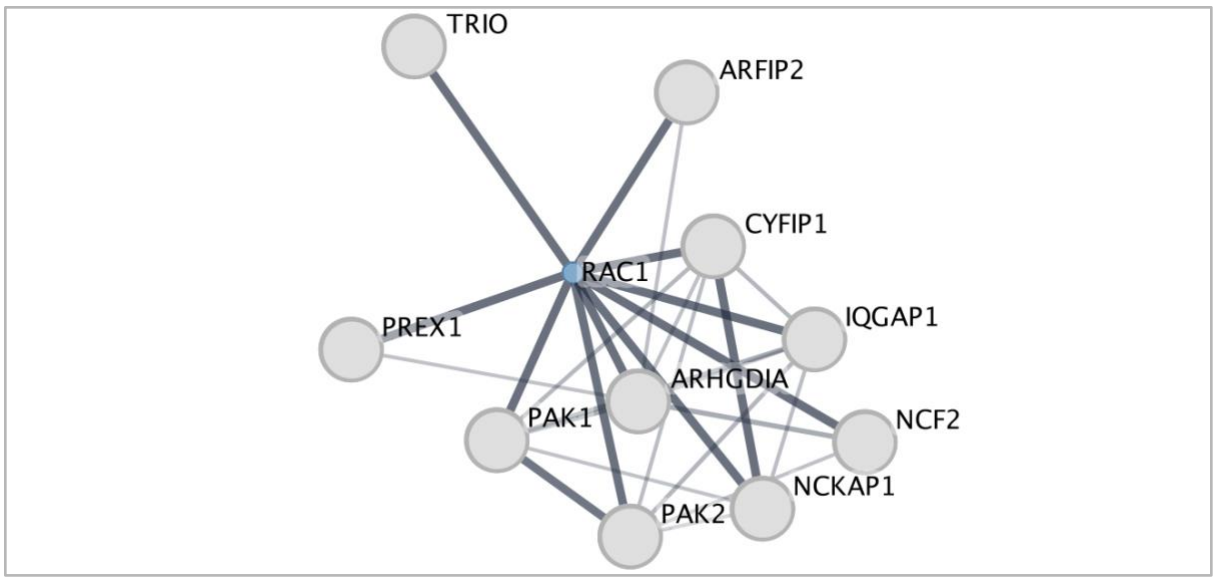


Figure 25 - Cytoscape Visualization of UniProt ID P63000.

Below network figure (figure 26) is based on the result from the second query with the ‘max number of interactions to show’ setting set to ‘no more than 10 interactors for the first shell. This UniProtID was placed in STRING and the default ‘max number of interactions to show’ setting was set to ‘no more than 10 interacters for the first shell. It is easier to see the how the sizes of the proteins varies according to protein length and also the coloring in accordance with negative/positive logFC values

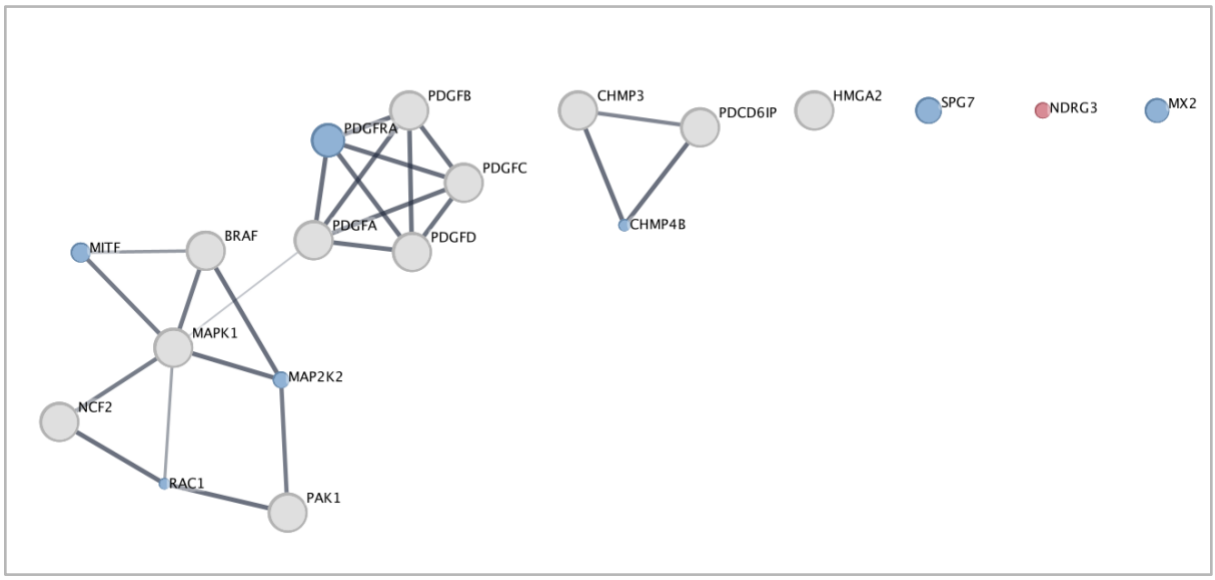


Figure 26 - Cytoscape visualization of UniProt IDs <10 interactions

The last Cytoscape network figure (figure 27) is an extra figure that shows a visualization having the ‘max number of interactions to show’ setting set to ‘no more than 20 interacters’. It follows the same set of instructions in regards to node size, coloring and layout choice.

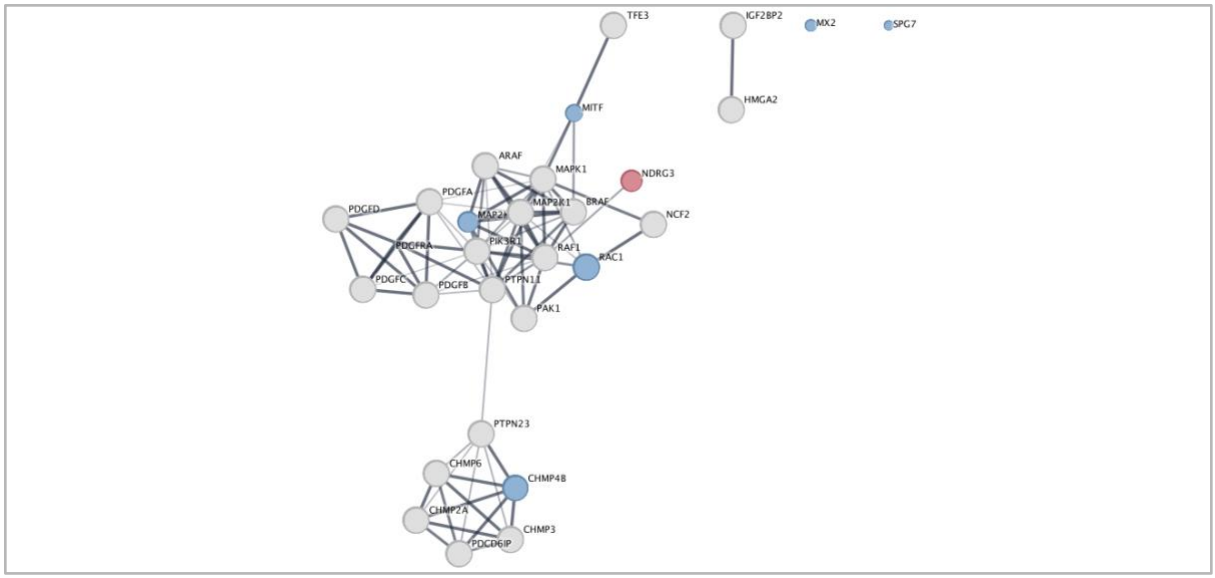


Figure 27 - Cytoscape visualization of UniProt IDs <20 interactions

Literature list:

National Center for Biotechnology Information. (2023). GSE242674. Gene Expression Omnibus (GEO). Retrieved from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE242674

National Cancer Institute. (2022). Melanoma Treatment. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK459367/

National Cancer Institute. (2023). Melanoma Treatment PDQ. Retrieved from https://www.cancer.gov/types/skin/patient/melanoma-treatment-pdq

Skin Cancer Foundation. (2022). Melanoma. Retrieved from https://www.skincancer.org/skin-cancer-information/melanoma/

Skin Cancer Foundation. (2021). Melanoma Warning Signs and Images. Retrieved from https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/#uglyduckling