**BIN503 FINAL PROJECT**

**Sevda Rafatova**

**Fatma Rabia Fidan**

**Firuza Rahimova**

**pgAdmin account that we jointly used for the project: e212642**

1. Introduction - Sevda and Firuza
2. a - Rabia

b - Sevda

c - Firuza

All constraints are overcome by Sevda and Rabia

3. mostly Firuza but everyone helped

4. 3 by Firuza, 1 by Rabia

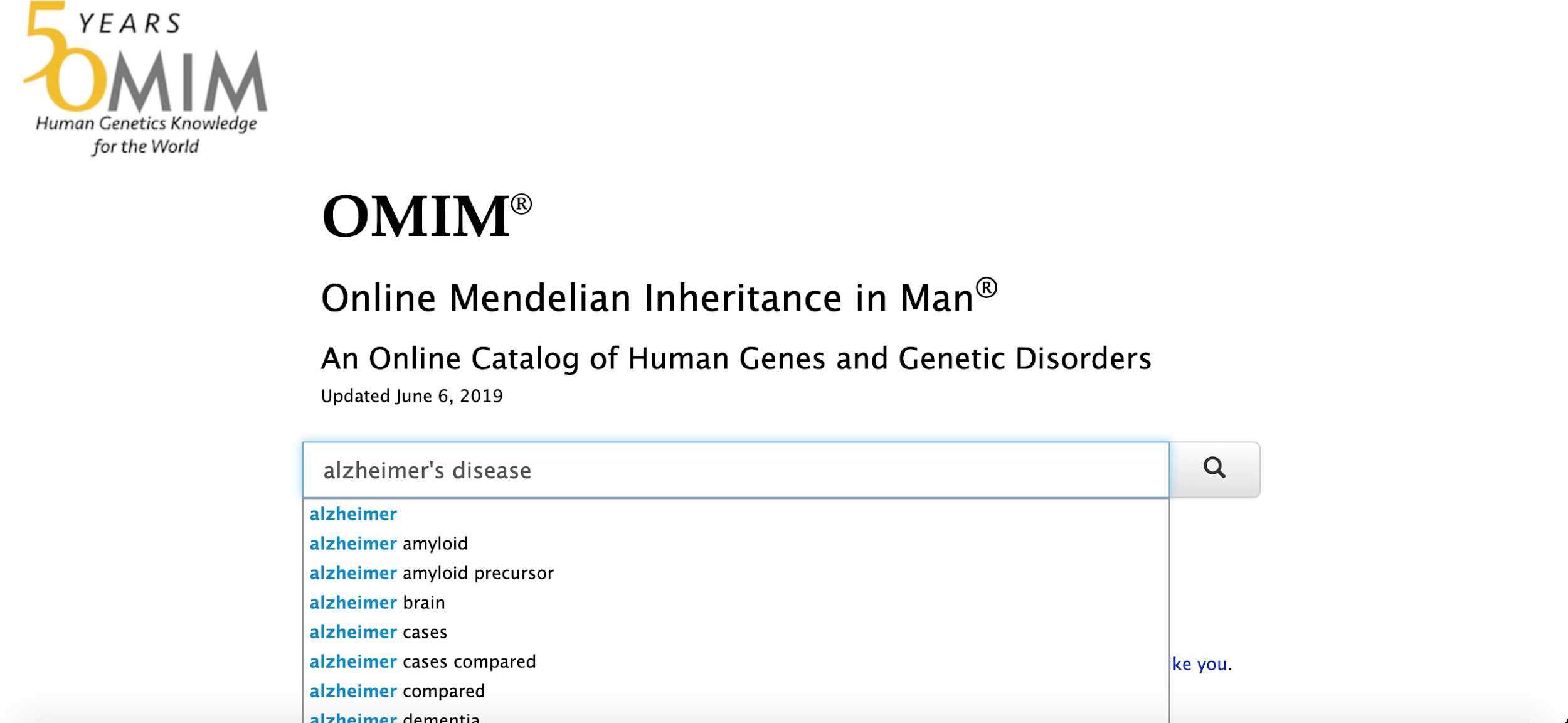
5. Sevda and Rabia

Fortunately, we were a good team and nobody avoided working; on the contrary, everyone tried to do their best. We believe all of us have equal total contribution to this project.

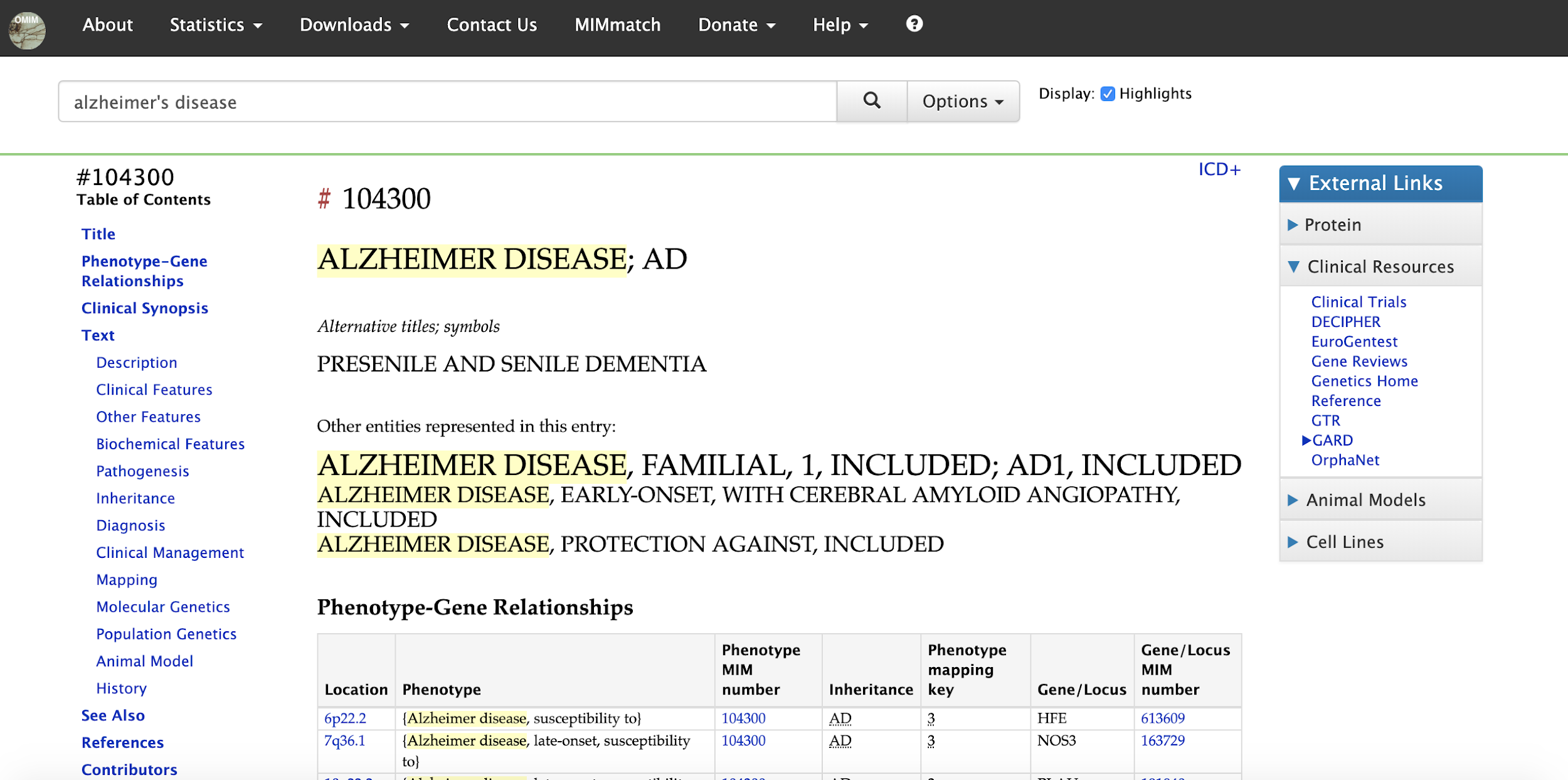
**1.** Alzheimer’s disease is associated with loss of function and death of nerve cells in several areas of the brain which leads to a disease which leads to loss of cognitive function such as memory and language, it is a progressive and a neurodegenerative disorder.

In old age, people are suffering from progressive dementia and AD is the most common version of this neurodegenerative problem. Neurofibrillary tangles (intracellular) and amyloid beta plaques (extracellular) are main pathological phenotype of this disorder. AD is associated with APOE ε4 allele, mutations in APP, PSEN1 and PSEN2 genes. Moreover, mutations in A2M gene is increasing susceptibility. In most occurrences, AD is familial but it can also be sporadic. Disease starts by memory loss and deficits in cognitive performance, and onset depends on the type of AD which can even be seen in younger people.

We used OMIM - Online Mendelian Inheritance in Man Database to find the information above. OMIM contains data of human gene and genetic phenotypes, primarily focusing on the link between them.

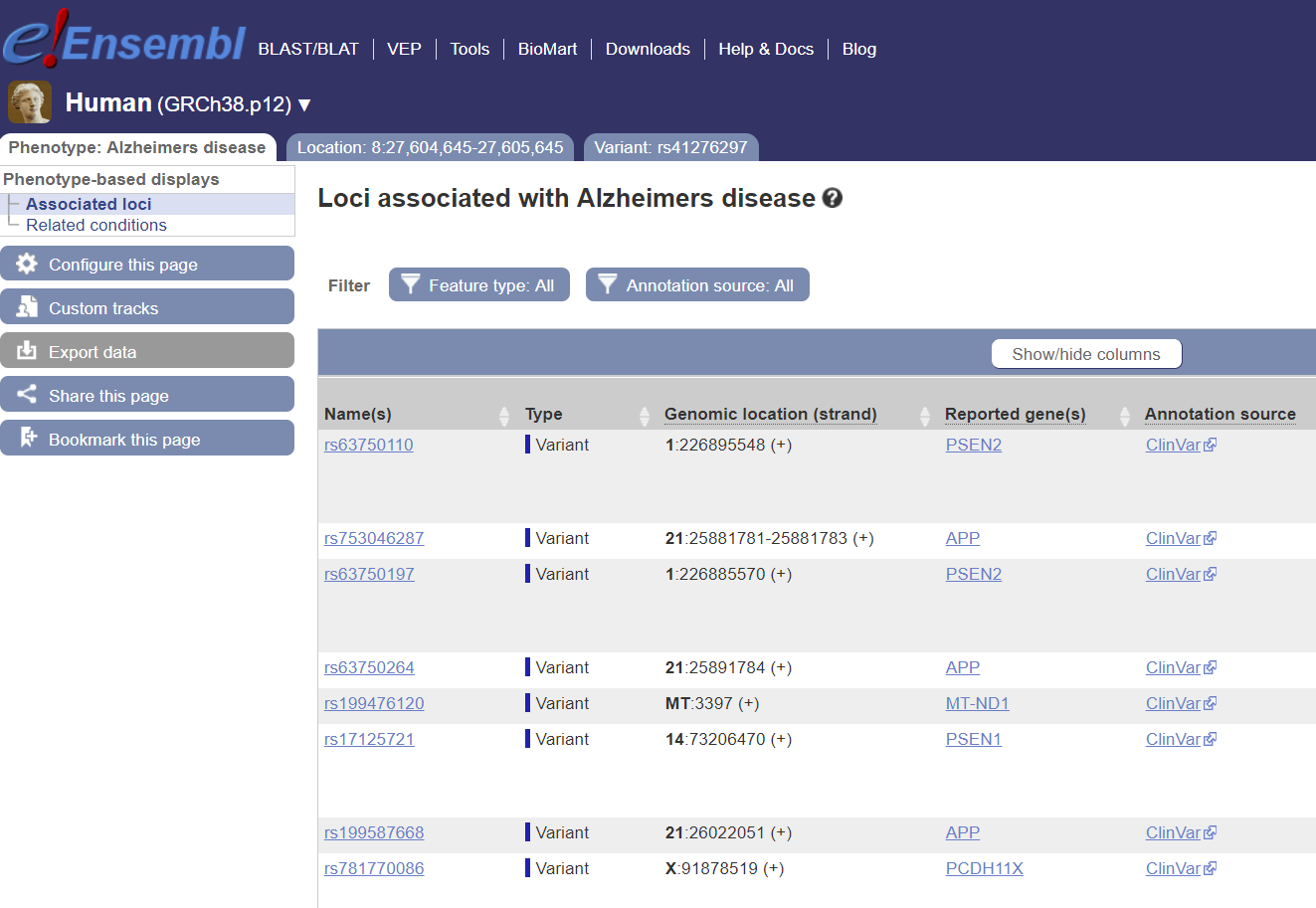


**Figure 1: Searching for Alzheimer’s Disease in OMIM**



**Figure 2: Search results in OMIM**

The other database that used to get information about disease phenotype is Ensembl.



**Figure 3: Variations related with Alzheimer’s Disease**

**2.a)**

Ensembl is a genome database constructed by European Bioinformatics Institute and the Wellcome Trust Sanger Institute in 1999. It contains information about genes, transcripts, snps and everything about our and many species’ genomes. It has many tools one of which is Biomart. Biomart allows us retrieve genomic information from many databases with a user-friendly interface. It has many filters and attributes to choose from. And indeed we used Biomart for the first 3 table. We retrieved information from HGNC, UniProt, PDB, and dbSNP other than Ensembl by using Biomart.

HGNC is a committee which gives unique names and symbols to human genes. The online database is curated. HGNC tries to unite the scientific society in terms of gene names usage to avoid confusion.

UniProt is a protein database which is explained in more detail in 2b.

PDB is a protein 3-D Structure Database as well as nucleic acids, and complex assemblies. The structures that are experimentally resolved are submitted as entries. There may be multiple entries per Protein.

dbSNP is a database which contains information about single nucleotide polymorphisms. It was constructed by National Center for Biotechnology Information (NCBI) in collaboration with the National Human Genome Research Institute (NHGRI).

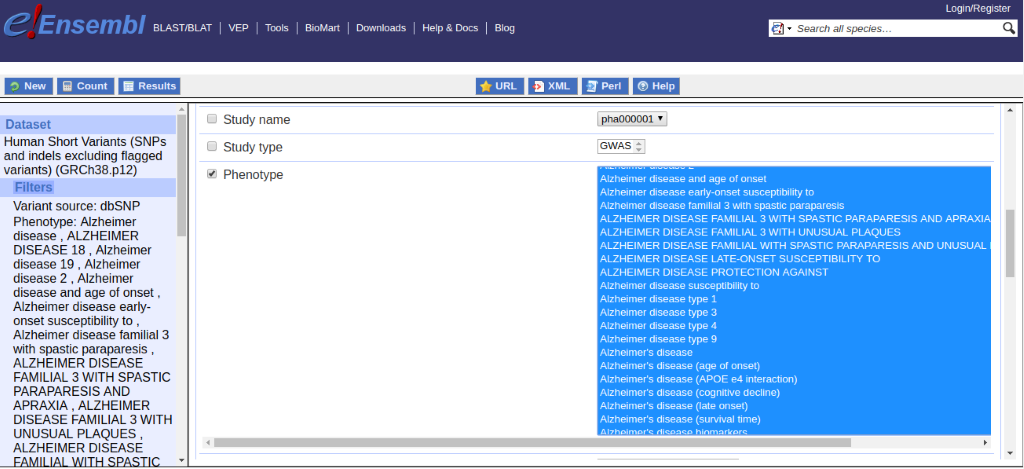
**GENES** table contains information about the genes on which a variant about Alzheimer’s disease lies. Table 1 has 3 columns: Ensembl Gene ID of all the genes containing an Alzheimer’s variant, HGNC symbols of them and start position of the genes in base pairs.

**VARIATIONS** table contains information about the variants directly. This table has 3 columns, too. Ensembl Gene ID of the gene containing the variant -if the variant is on a coding gene-, Variation Name of the variant, variation start in translation if it is on a translated region in the genome. Not all variants are on coding genes, so the first column may be NA and similarly, not all variants are translated, they can be intergenic or they can be on introns, so, the third column may be NA too.

**ENSEMBLUNIPROT** and **UNIPROTPDB** tables contain information about the proteins which are coded by Alzheimer’s associated variants’ genes. Both have 2 columns. "Ensembl Gene ID, UniProt/SwissProt ID of the protein and UniProt/SwissProt ID, PDB ID. PDB column may be NA since 3-D structure may or may not be determined.

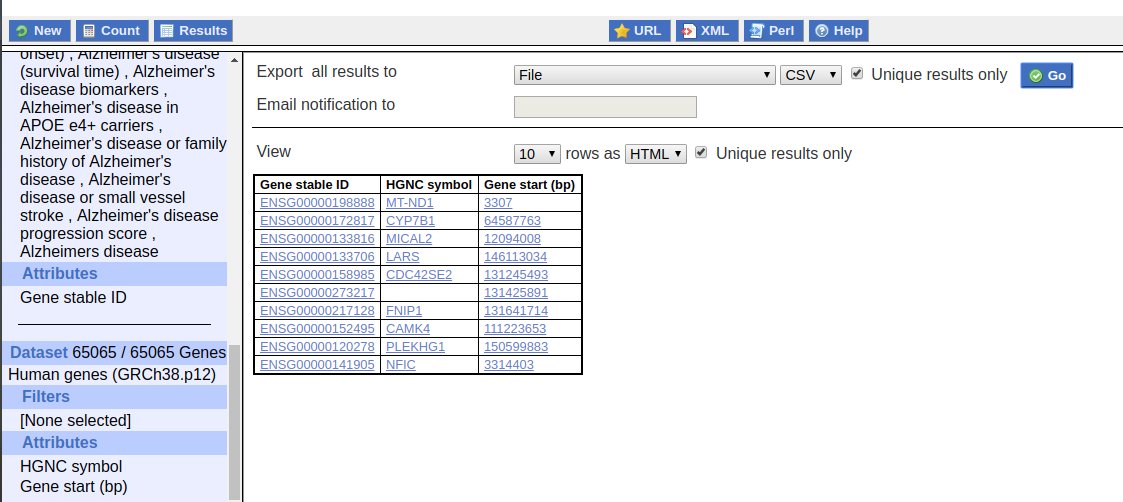
We have used Biomart to retrieve the information for TABLE1: GENES as follows:

We chose Human Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p12) as Dataset. Then for filters we added Variant source: dbSNP and Phenotype: everything about the Alzheimer’s.



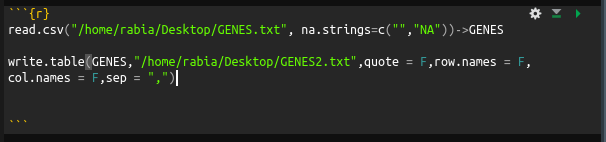
**Figure 4: Selecting filters in Biomart**

Then as attributes, we chose Gene stable ID. For HGNC symbol, we had to add the second dataset: Human genes (GRCh38.p12). No additional filters were added for this. As attributes, we chose HGNC symbol and Gene Start (bp).



**Figure 5: Biomart results overview**

Then we edited the table replacing empty cells with NA in order to make our database consistent and our data compatible with programming languages when people download it.



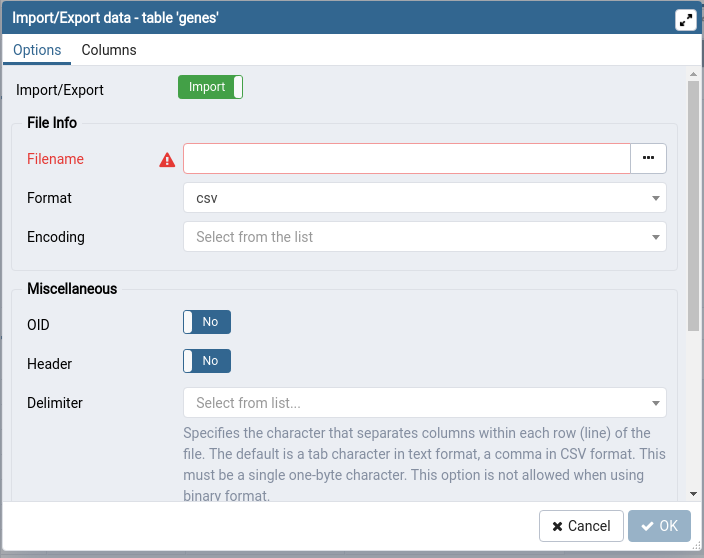
**Figure 6: R codes for adding NAs and exporting new table**

Now we use SQL to create the table:



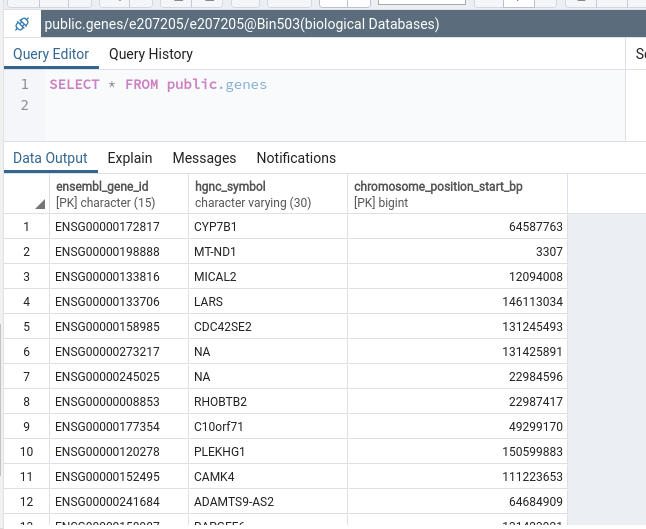
**Figure 7: SQL code for creating tables**

Then imported the data from the file by using pgAdmin interface



**Figure 8: importing Data with interface**

viewing the table:



**Figure 9: GENES table**

similarly, VARIATIONS, ENSEMBLUNIPROT and UNIPROTPDB tables are created.

**b)** In this part UniProt Knowledge base (UniProtKB) database is used. UniProt is a database with aim to provide freely accessible resource for protein sequence and functional information. It consists of Swiss Prot and TrEMBL parts. Swiss Prot is manually annotated and reviewed. It consists of records with information extracted from literature and curator-evaluated computational analysis, it is more reliable than TrEMBL.TrEMBL is automatically annotated and not reviewed. It consists of records that waiting to get full manually annotated.

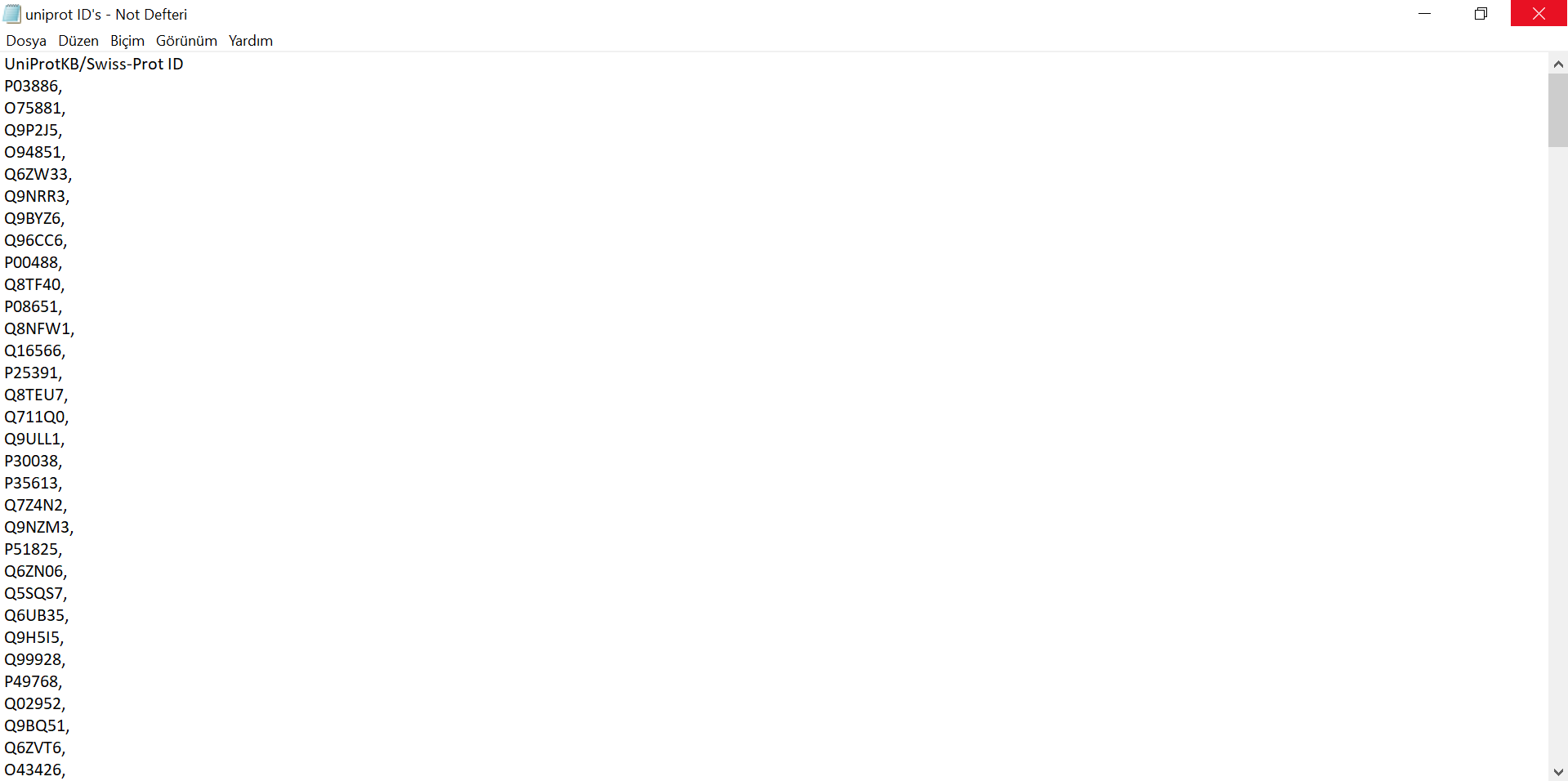
Retrieve/ ID mapping is a tool that takes as input a list of identifiers to do one from the below;

-Get corresponding UniProt ID’s and use them on UniProt Database

- Convert identifiers which are of a different type to UniProt ID’s or vice versa and use them.

The UniProt ID’s that are used in this part are obtained from previous part. From Proteins table (ensmbluniprot+uniprotpdb).

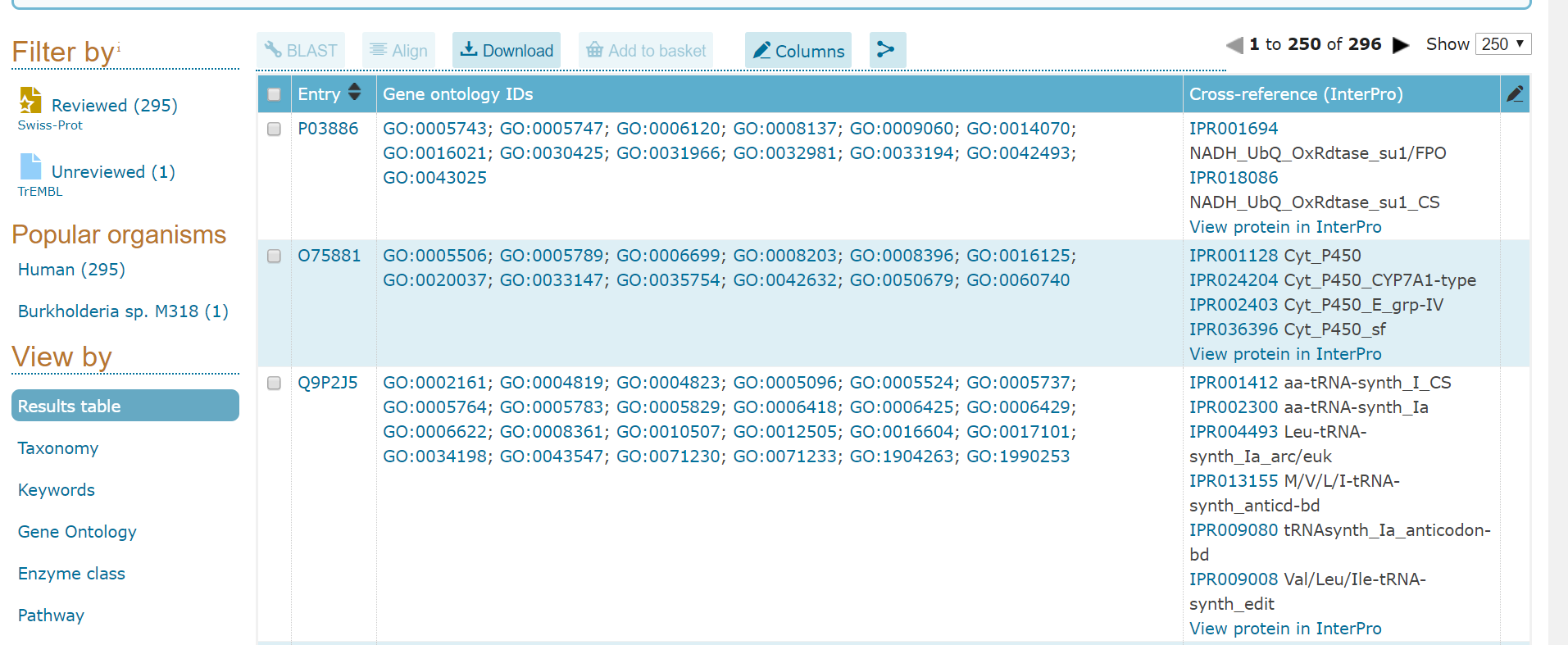
The UniProt ID list that we’ve used in UniProt Retrieve/ID mapping is;



**Figure 10: UniProt ID list**

Then, by using UniProt Retrieve/ID mapping we’ve obtained:

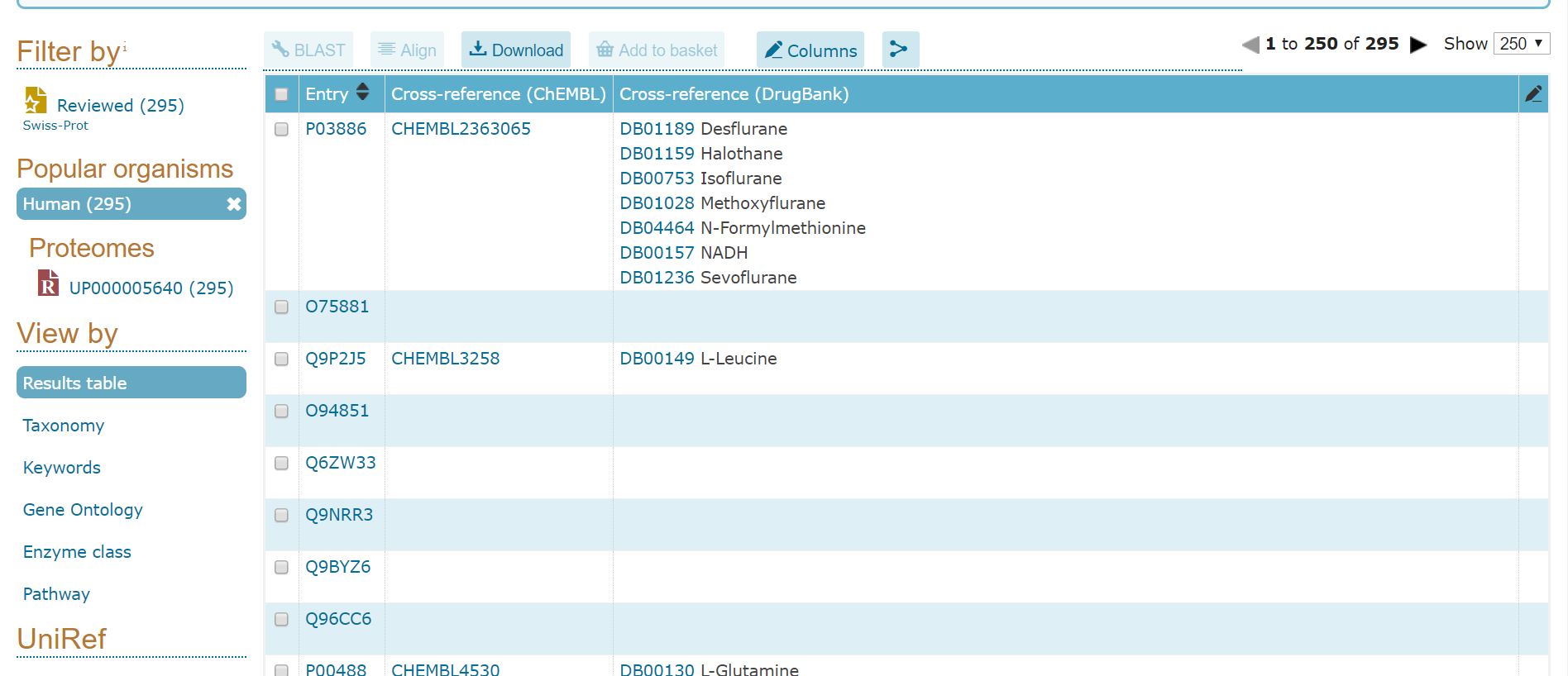
Results for table 4: Function;



**Figure 11: Results of UniProt Retrieve/ID mapping**

Thisdata downloaded in tab separated format. And saved in txt format in Notepad.

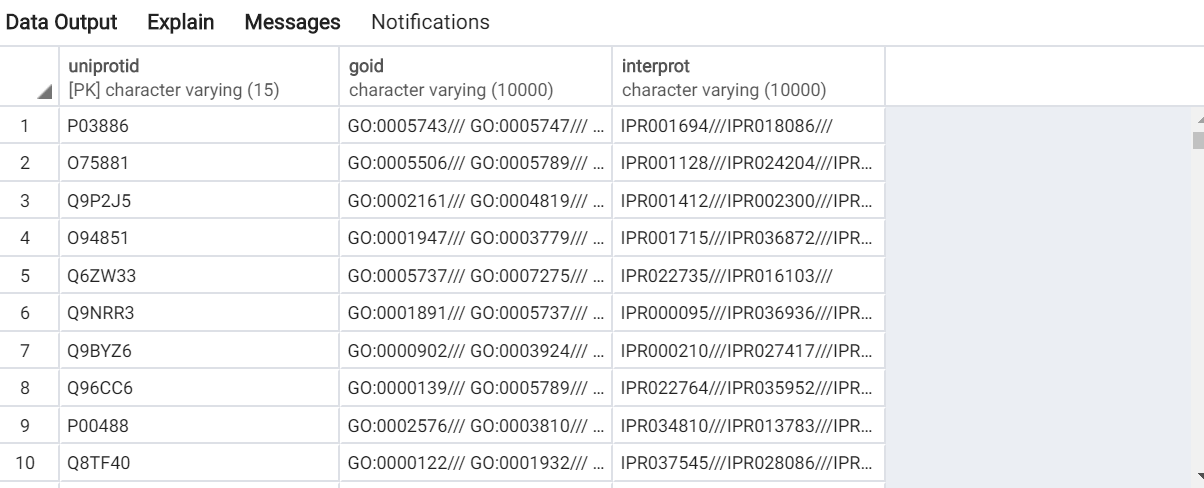
Results for table 5: Bioactivity;



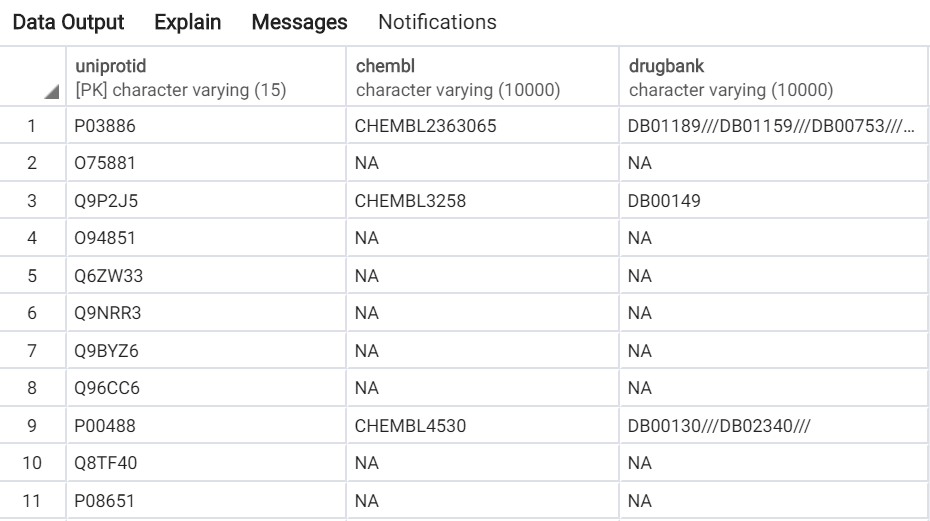
**Figure 12: Results of UniProt Retrieve/ID mapping**

We have downloaded data from UniProtKB in tab separated format. Then by using R programming we have added ‘NA’ to the empty entries as shown in 2a. We have created tables in pgAdmin interface.(<https://bin503.kansil.org/pgadmin/browser/>).

After importation of data our tables are ready.



**Figure 13: Function table obtained in pgAdmin**

****

**Figure 14: Bioactivity table obtained in pgAdmin**

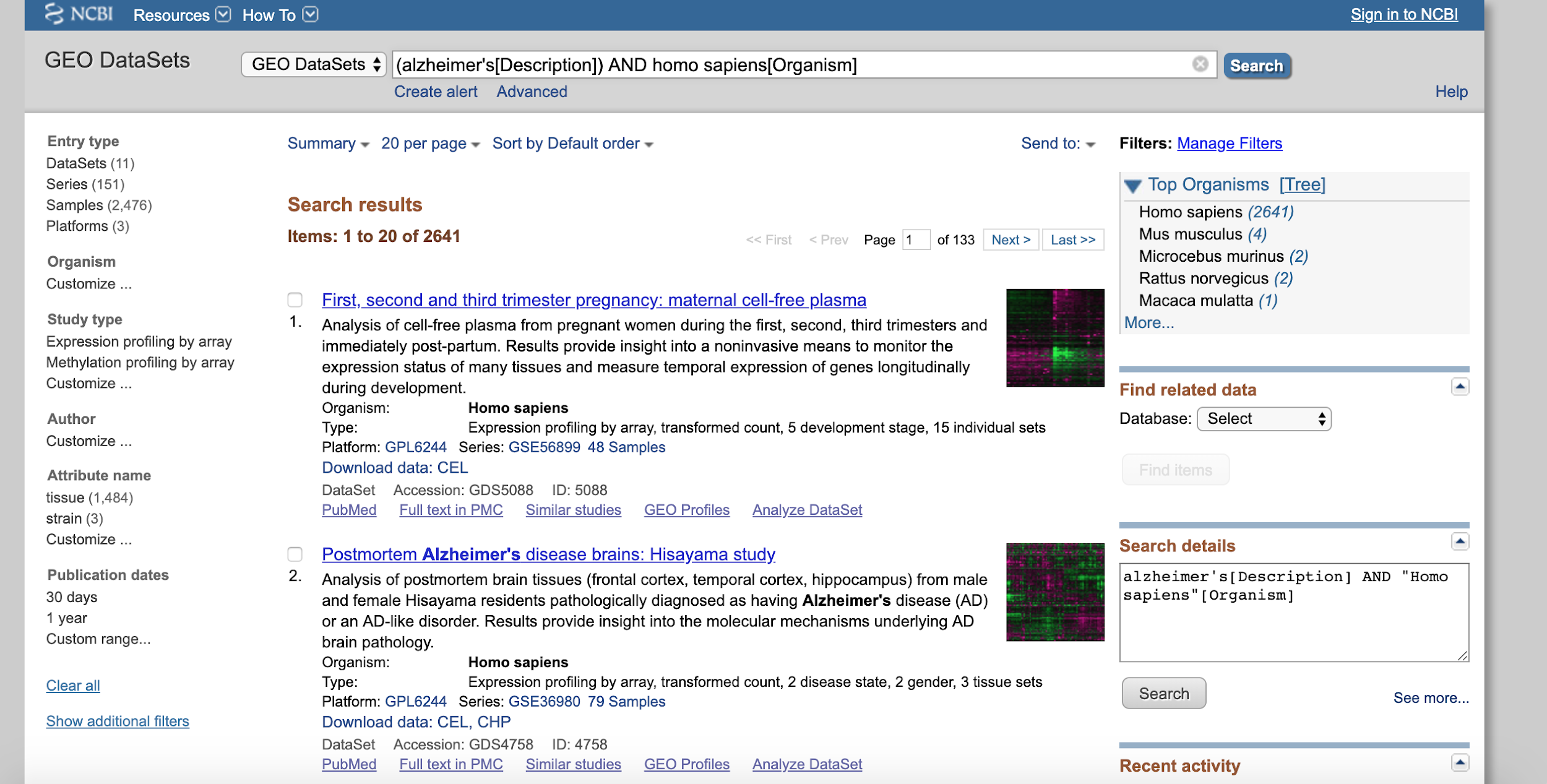
**Function Table:** By using this table we can get information about the functions of the proteins associated with the Alzheimer’s Disease. This table consists of 3 columns. Which are uniprotid, goid and interport. UniProt ID is a primary key for this table and foreign key for database. The symbol “///” shown in the table represents that there are multi values in the same cell. NA represents that there is no information about that ID.

**Bioactivity Table:** By using this table we can get information about the bioactivity of the proteins associated with the Alzheimer’s Disease.This table consists of 3 columns. Which are uniprotid, chembl and drugbank. UniProt ID is a primary key for this table and foreign key for database. The symbol “///” shown in the table represents that there are multi values in the same cell. NA represents that there is no information about that ID.

**c)**

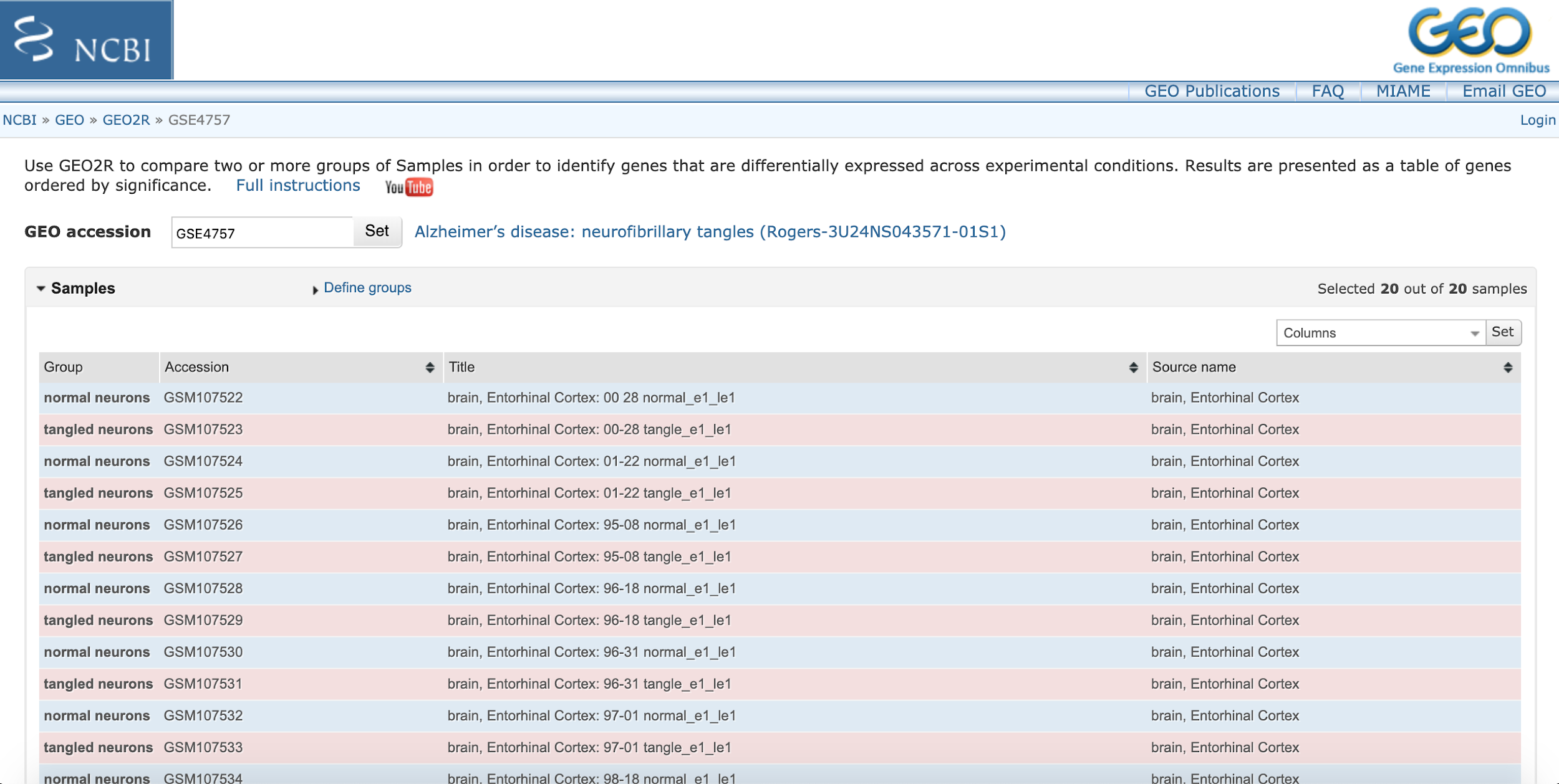
We have found a dataset which contains tissues from 10 mid-stage Alzheimer’s Disease patients. Neurofibrillary tangles(NFT) are a main pathological phenotype of AD. Dunckley T. et al compared checked gene expression in neurons showing NFT phenotype and normal neurons.

Data is retrieved from GEO Datasets. Its GEO accession code is GSE4757. GEO is vital in genomics, as it freely shares all the high throughput sequencing, microarray and other types of genomic data. Data is submitted by researchers easily, and downloading is very simple.



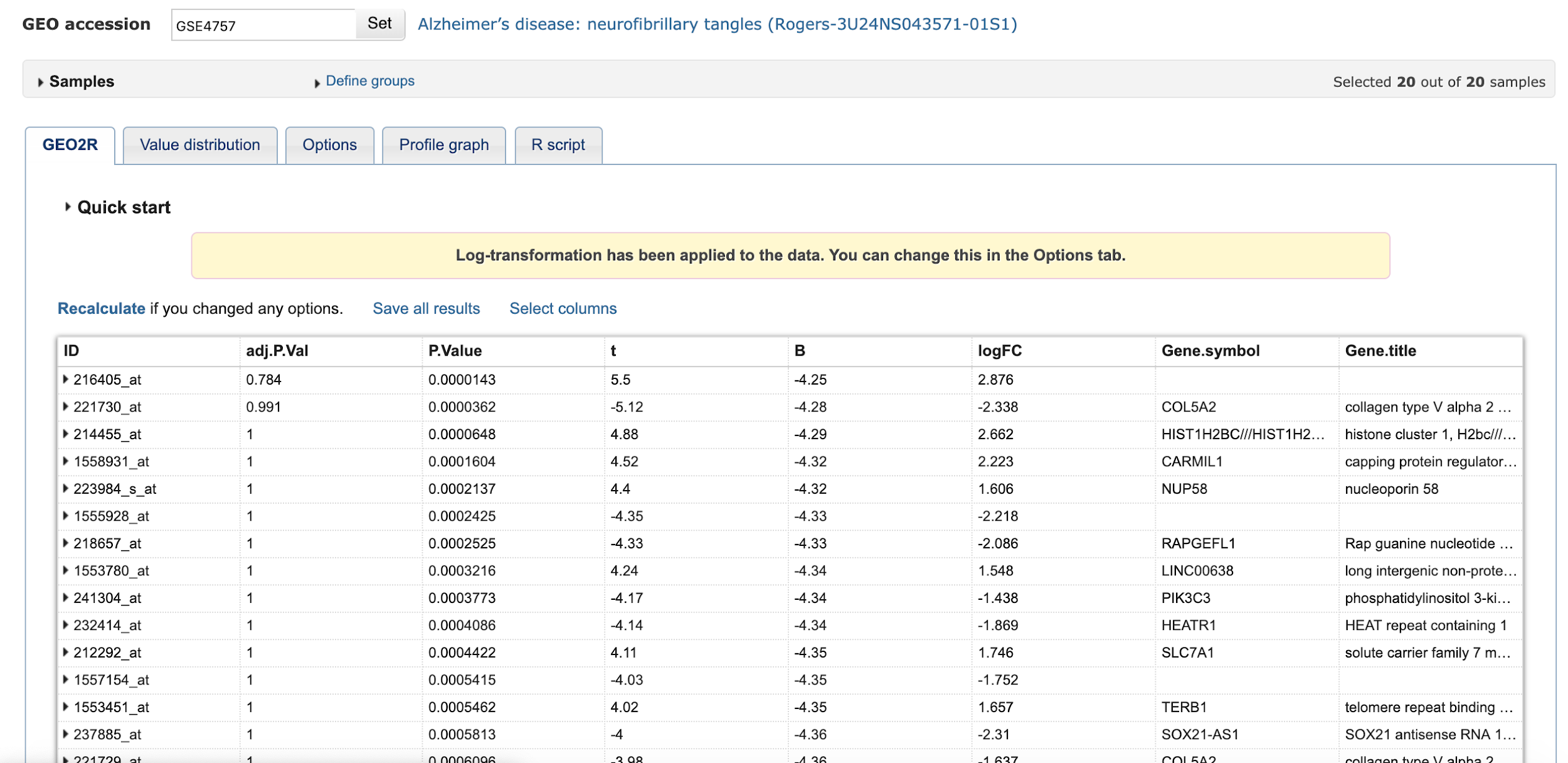
**Figure 15: Searching for AD datasets in GEO.**

Grouping is performed as shown below: normal and tangled neurons.



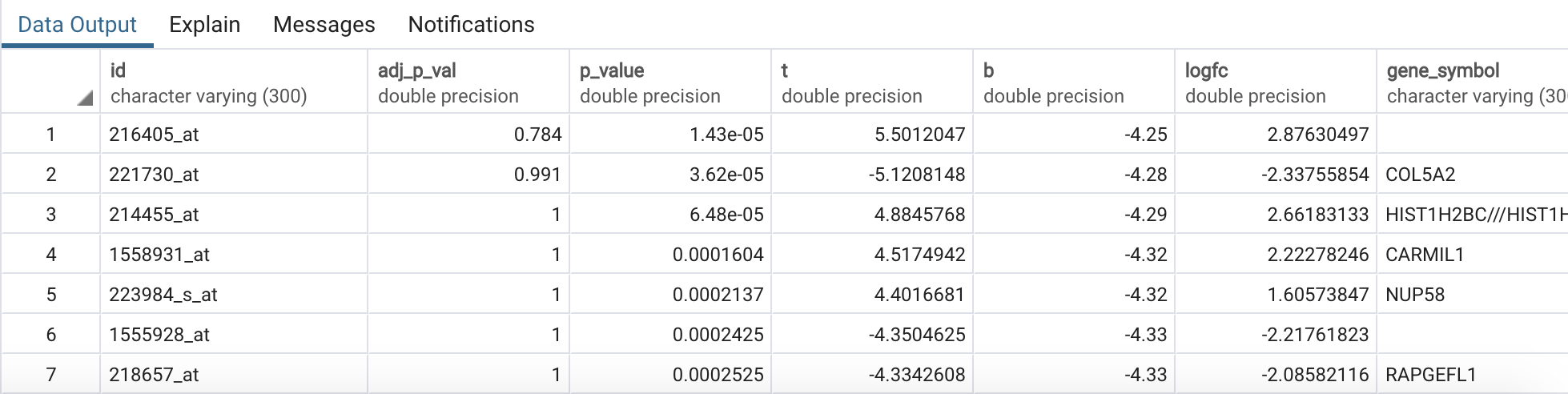
**Figure 16: Grouping for GEO2R analysis.**

Analysis of this data is done by GEO2R with default settings. GEO2R is a web tool with which one can compare several groups of samples to check differential gene expression easily.



**Figure 17: GEO2R analysis results.**

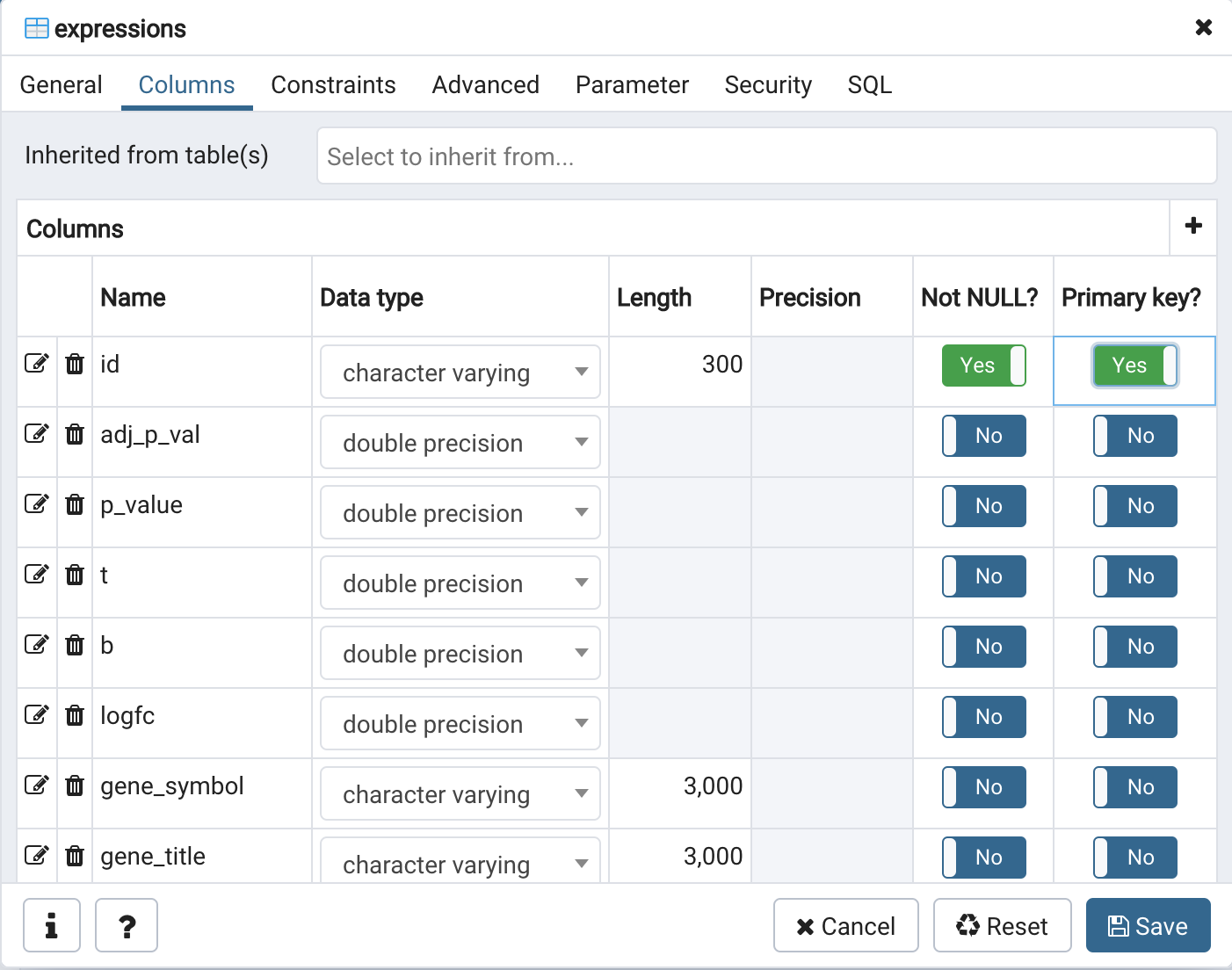
We imported GEO2R data into PostgreSQL by creating expressions table and using Import function. Data pre-processing in R changed empty column values into NA. All of these are as shown in 2a. Table looks as below:



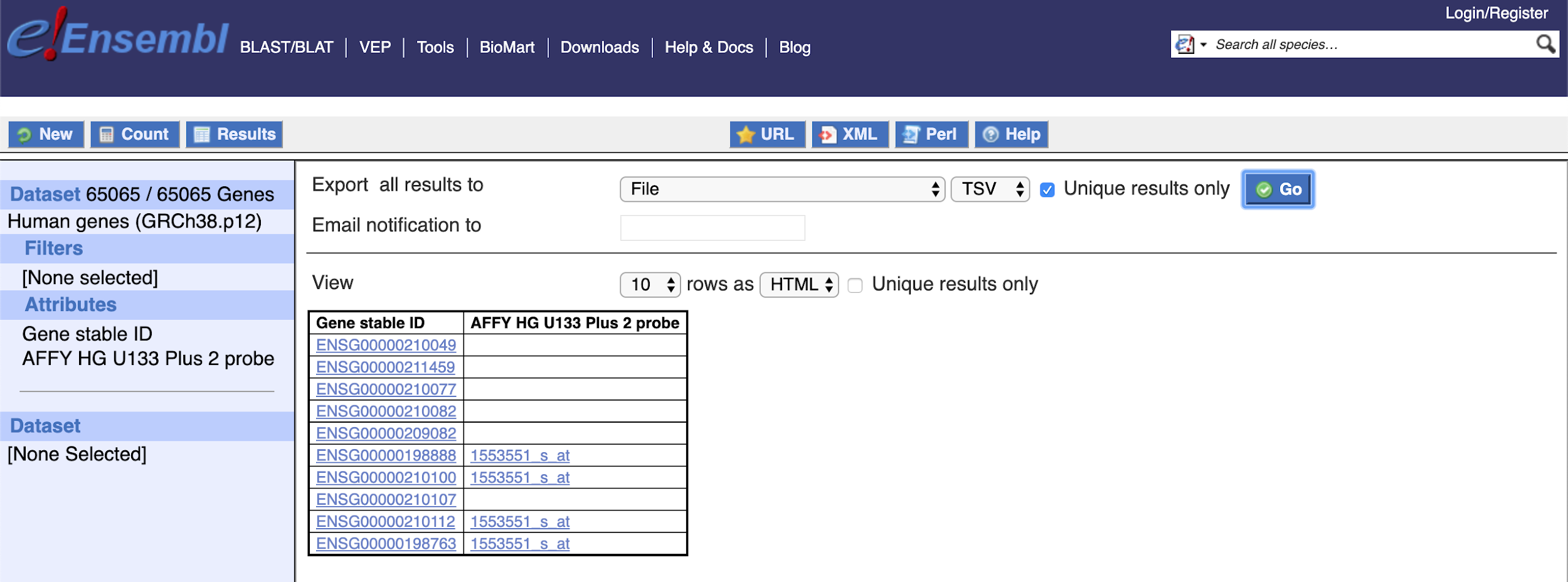
**Figure 18: Expressions table.**

In ID column, IDs of sequencing probes are listed. While comparing healthy and NFT neurons, p\_values smaller than 0.05 gives us significantly different genes. t is t-statistic result of t-test, B and logfc are statistical values that we are not familiar with but comes with the default GEO2R analysis. Gene symbol corresponds to HGNC symbol to identify genes, and gene titles are their long names.

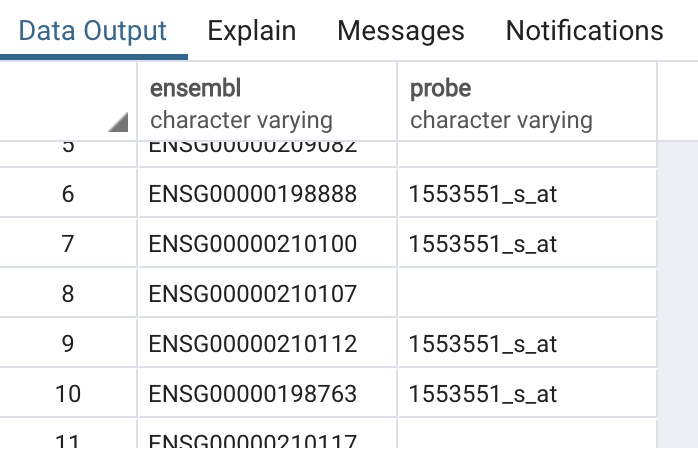
Primary key is defined as:

**Figure 19: Definition of Primary key for Expressions table.** 

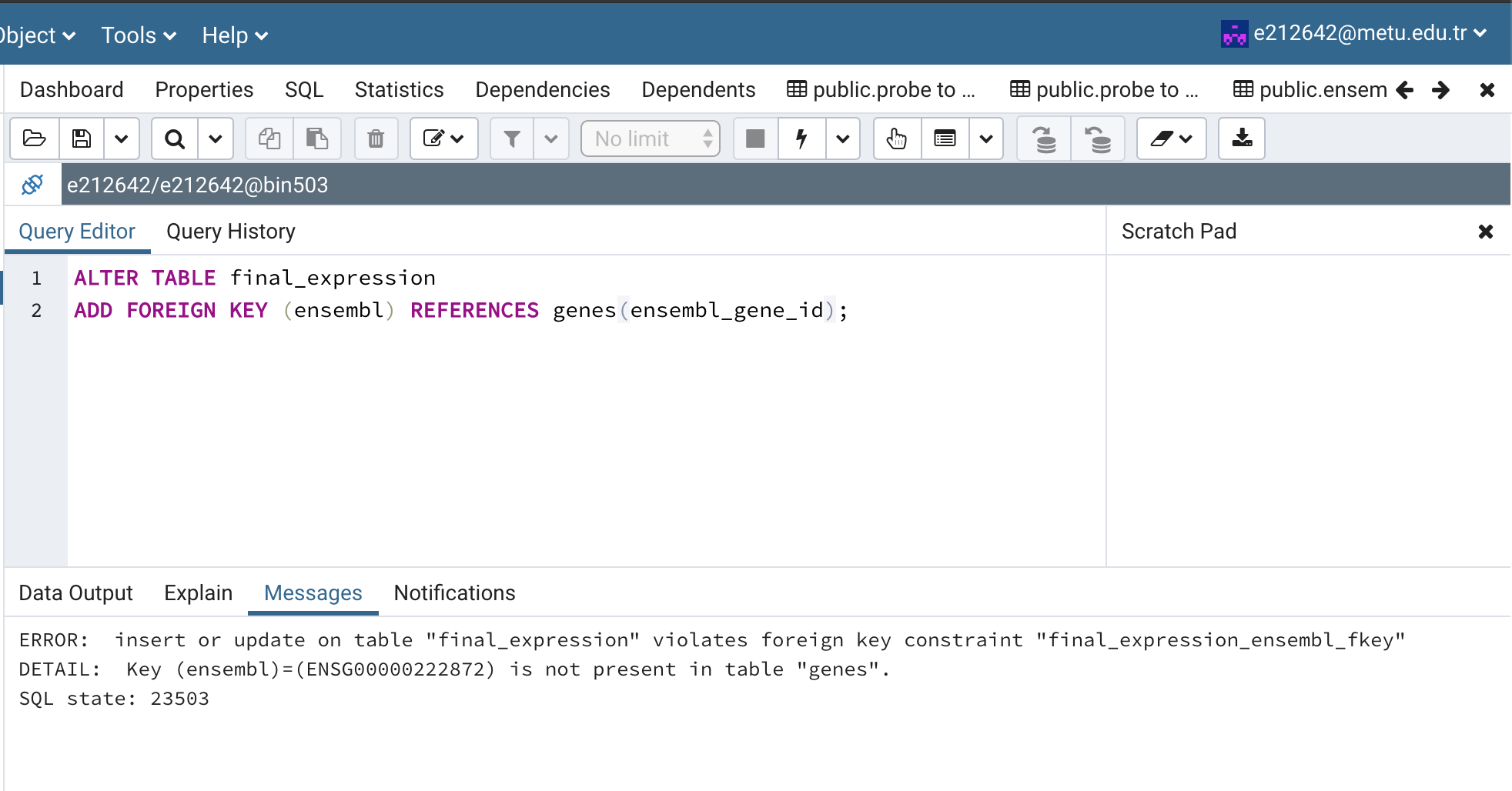
We had to download probe ID corresponding to ensembl gene id values from Biomart.

****

**Figure 20: Probe ID - Ensembl ID correspondence download from Biomart**

Another table called **ENSEMBL** was created in order to use Ensembl gene ID as foreign key. 

**Figure 21: Ensembl table**



**Figure 22: Error screenshot**

Ensembl table has duplicate values for ensembl id because one gene can have multiple probes. Also some genes in the final expression table is not present in genes table because they are not associated with Alzheimer’s in dbSNP. So, in order to connect the three tables (genes, final expressions and ensembl) we needed another table which covers three of them. So, we added **ALL\_GENES** table which contains ensembl gene id of all human genes.

**3)ER diagram:** We submitted ER diagram separately.

**4)QUERIES**

**QUERY 1.**

In this query, we count variations of each gene in case that those genes are significantly differentially expressed (between normal and neurofibrillary tangled neurons).

SELECT g.ensembl\_gene\_id, COUNT(variation\_name)

FROM variations v, genes g, final\_expression f, all\_genes a

WHERE v.ensembl\_gene\_id=g.ensembl\_gene\_id

AND a.ensembl\_gene\_id=f.ensembl

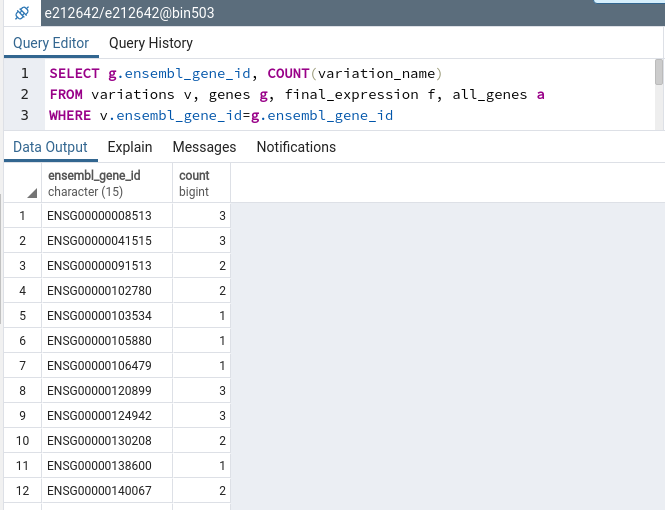
AND a.ensembl\_gene\_id=g.ensembl\_gene\_id

AND a.ensembl\_gene\_id=v.ensembl\_gene\_id

AND f.ensembl=v.ensembl\_gene\_id

AND (f.p\_value < 0.05)

GROUP BY g.ensembl\_gene\_id ;



**Figure 23: Result of Query 1**

**QUERY 2.**

This query gives us Ensembl IDs of genes which are associated with Alzheimer’s Disease. In addition, it informs us about proteins produced by these genes by giving their Chembl ID and Gene Ontology IDs, through which their similars and functions can be learnt.

SELECT g.ensembl\_gene\_id, b.chembl, f.goid

FROM bioactivity b, function f, genes g, ensmbluniprot e

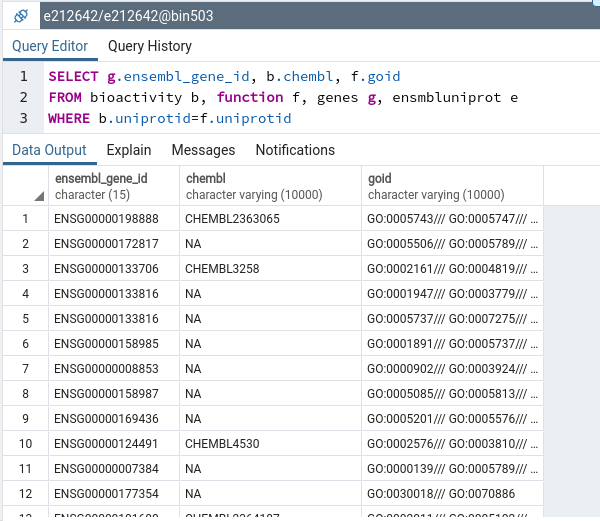
WHERE b.uniprotid=f.uniprotid

AND g.ensembl\_gene\_id = e.ensembl

AND e.uniprot = b.uniprotid

AND e.uniprot = f.uniprotid

AND b.uniprotid = f.uniprotid ;



**Figure 24: Result of Query 2**

**QUERY 3.**

Gene titles containing certain names like transferrin can be searched between the genes significantly differentially expressed and their drugbank IDs can be given as shown in the query below.

SELECT b.drugbank, f.gene\_title

FROM bioactivity b, ensmbluniprot e, genes g, final\_expression f, all\_genes a

WHERE e.uniprot = b.uniprotid

AND f.gene\_title LIKE '%transferrin%'

AND (f.p\_value <= 0.05)

AND a.ensembl\_gene\_id=f.ensembl

AND a.ensembl\_gene\_id=g.ensembl\_gene\_id

AND g.ensembl\_gene\_id=f.ensembl

AND g.ensembl\_gene\_id = e.ensembl



**Figure 25: Result of Query 3**

**DUMP:** Dump can be obtained from pgAdmin interface, student e212642. We have saved it as backup. Dump file name is project\_dump.

**5)DISCUSSION**

We have constructed a database where stands 10 tables in total. By using this database, it is possible to find out genes, proteins, variations, drugs, bioactivity, functions and differential expressions associated with Alzheimer’s Disease. We have found 433 ENSEMBL gene ID’s, 123 unique UniProt ID’s and 596 unique variations related with Alzheimer’s Disease. There are 65 CHEMBL entries for UniProt ID’s. In functions table there are 294 UniProt ID’s and each of them has multiple Gene Ontology ID’s and Interpro data. In differential expression analysis at GEO2R, 1856 genes are significantly different between groups of healthy and tangled neurons, which can be novel targets.

Alzheimer’s is not a simple Mendelian disease. It’s affected by variations throughout genome, which can lead to mutations and inhibit/change resulting proteins’ functions. Not only mutations in proteins but also regulatory functions can be harmful and be important in Alzheimer’s.

By using this database, it is possible to find out if a variation on a gene is associated with Alzheimer’s disease or not. From genes table we can see all genes related with Alzheimer’s Disease. From variations table we can see all variations occurred on these genes. From ensembluniprot table we can find out which genes produce which proteins. From function table we can find out functions of these proteins. By using GO ID and Interpro entries it is possible to further investigate the UniProt ID’s.

Novel path:

The genes, which were not previously associated with Alzheimer’s in dbSNP but actually differently expressed in Alzheimer’s than healthy people/cells, could be a potential novel treatment.

Here we write a query to retrieve all genes that were examined in Geo2R and then we subtract the already associated genes and the genes which were not previously associated+not differentially expressed

(SELECT ensembl\_gene\_id

FROM all\_genes a,final\_expression f

WHERE a.ensembl\_gene\_id=f.ensembl)

EXCEPT(

(SELECT ensembl

FROM final\_expression,genes

WHERE ensembl=ensembl\_gene\_id OR p\_value>0.05 )

UNION

(SELECT ensembl\_gene\_id

FROM genes)

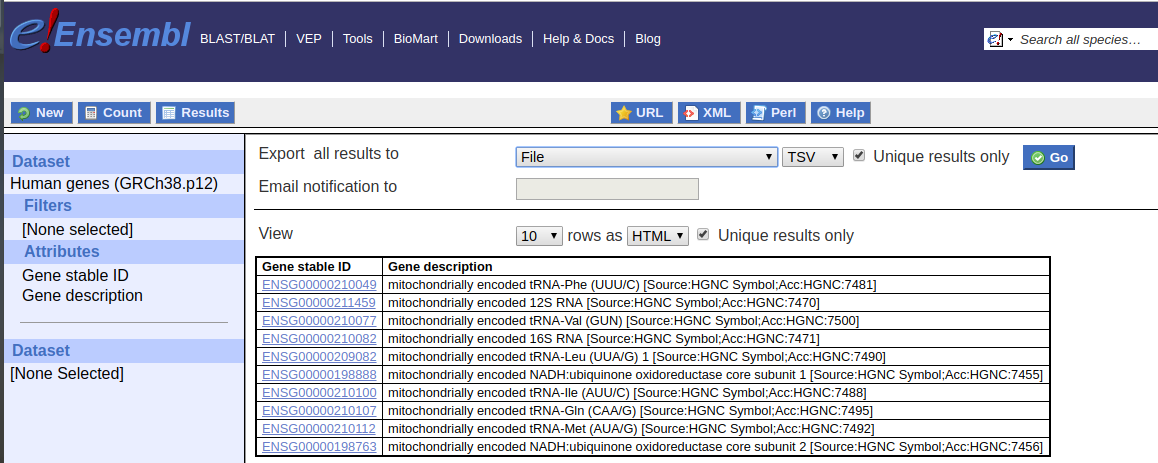
);



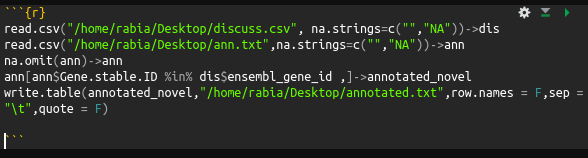
**Figure 26: First 13 rows of the query**

Totally we got 522 such genes. Then we downloaded the result as csv and checked ClinVar, ESP, HGMD-public and PhenCode to see if these genes are associated with Alzheimer’s in other databases than dbSNP. In Biomart, only ClinVar had some variants about Alzheimer’s but they are not on our list. Also we checked DisGeNet, HGV and a couple of more databases. There were no common results. This means, some of our differentially expressed 522 genes may really be a novel treatment path.

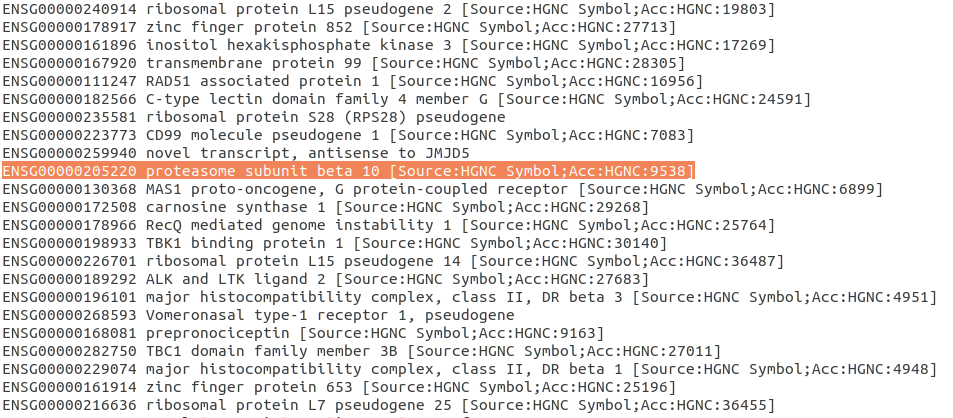
To propose 1 gene as an example, we download the gene descriptions from Biomart and checked if we can find a gene which can be related with the known phenotypes of the disease.



**Figure 27: Gene descriptions of all human genes**

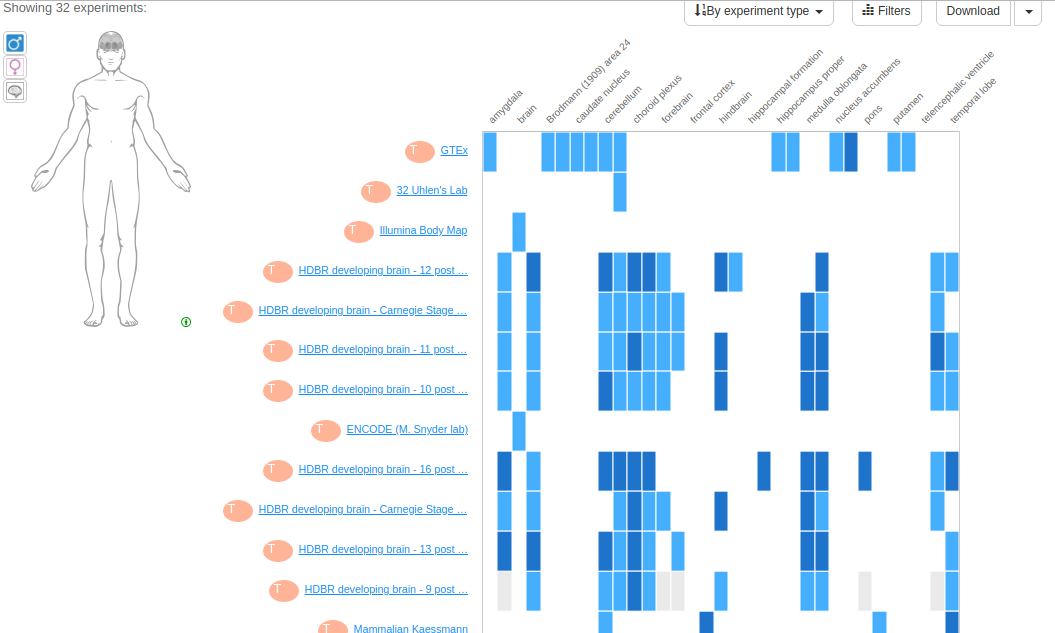


**Figure 28: Obtaining gene descriptions for only our 522 genes**



**Figure 29: Gene descriptions**

Among many genes, ENSG00000205220 is a subunit of proteasome. Also, this protein is expressed in various brain tissues, in many cases it is highly expressed.



**Figure 30: ENSG00000205220 gene expression in brain tissues**

Faulty protein aggregation is a common pathology in Alzheimer’s disease.

So we propose studying ENSG00000205220 gene as a novel approach to treat Alzheimer’s disease.