Statistical Analysis of Gene Expression Data of Lungs

Kritika Verma
kritikaverma3107@gmail.com
Cluster Innovation Centre, University of Delhi, Delhi - 110007

ABSTRACT

Gene expression analysis is a fundamental tool in bioinformatics, with the aim of understanding the genetic mechanisms that regulate the behavior of cells and organisms. GEO (Gene Expression Omnibus) GEO is a public functional genomics data repository which is used to access datasets of gene expression which were analysed using GEOquery library of R and GEO2R tool. One of the key approaches used in gene expression analysis is hypothesis testing, which allows researchers to compare different samples and determine if there are significant differences in gene expression levels between them. A statistical hypothesis is a statement of the parameters of a population or populations. The hypothesis testing process involves formulating a null hypothesis, which assumes that there is no difference between the samples, and an alternative hypothesis, which states that there is a significant difference between the samples. Several factors need to be considered when performing hypothesis testing for gene expression analysis, including the type of experiment, the sample size, the choice of statistical test, and significance level.In this paper, the dataset obtained is of lung tumor and normal lung cells which has been analysed on basis of criteria like, tissue type, gender and smoking history. In conclusion, hypothesis testing is a powerful tool for gene expression analysis that allows researchers to identify genes that are deferentially expressed between different samples. However, some changes have to be made to ensure reliable results.

Keywords: Alternative hypothesis, bioinformatics, gene expression, GEOquery, R.

INTRODUCTION

The study of gene expression analysis has become increasingly important in the field of genetics and genomics. By analyzing gene expression data, researchers can gain insight into the molecular mechanisms underlying various biological processes, as well as the development of diseases. The advent of high-throughput technologies, such as microarrays and RNA seq, has made it possible to generate large amounts of gene expression data. However, the analysis of such data requires specialized tools and methods. One such tool is the GEO database, which contains a vast amount of publicly available gene expression data. Using the GEO database, researchers can retrieve data for various experiments, perform statistical analysis, and test hypotheses. In this paper, we will discuss the process of gene expression analysis using the GEO database and hypothesis testing. Specifically, we will describe how to retrieve gene expression data from the GEO database using the GEOquery package in R. After, preprocessing the data hypothesis testing is perform on it, giving analysis of real life datasets. Using a dataset series for GEO[1][2], of gene expression level in lung tissue normal and tumorous, we will work on hypothesis testing of expression levels of various genes.

METHODOLOGY

1.1 Retrieving Data from GEOquery Using R

GEO database provides datasets in various classes namely platform, samples, series and datasets[3].

1.1.1 Platform

A list of elements in the array is stored in platforms. A distinctive and stable GEO accession number (GPLxxx) is given to each Platform record. A Platform may make reference to a large number of Samples that were contributed by numerous submitters.

1.1.2 Samples

A Sample record describes the conditions under which an individual Sample was handled, the manipulations it underwent, and the abundance measurement of each element derived from it according to Davis et al[3]. Each Sample record is assigned a unique and stable GEO accession number (GSMxxx). A Sample entity must reference only one Platform and may be included in multiple Series.

1.1.3 Series

A series contains relevant and/or inter-related samples. It also has a unique accession code in the form (GSEXXXXXX) which can be used for accessing it on the GEO website, GEO2R as well as with R.

1.1.4 Datasets

Accorging to Davis et al[3], GEO DataSets (GDSxxx) are curated sets of GEO Sample data which represents a collection of biologically and statistically comparable GEO Samples and forms the basis of GEO's suite of data display and analysis tools. Samples within a GDS refer to the same Platform, that is, they share a common set of probe elements.

These four classes can be accessed an analysed using R's library GEOquery which can be used to obtain the samplevalues, GSEmatrix which can be used to plot a heatmap of the sample using pheatmap library, data summary consisting of mean, median, etc, and metadata about the expression values.

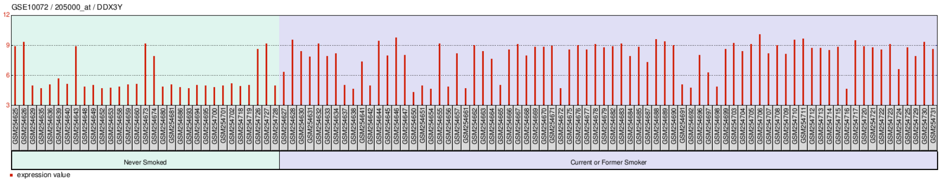
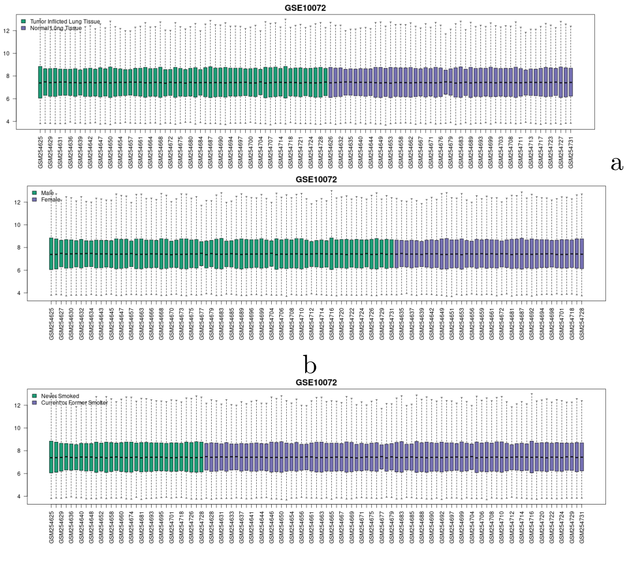


fig 1.1 Gene expression of DDX3Y of non-smokers and smokers.



a

b

c

fig 1.2 The distribution of the values of the selected Samples for a) Tumor and Normal Tissue b) Men and Women c) Non-smokers and current/former smokers

1.2 Analysing the Data

We used GEO2R a tool used for analysis for GEO series, samples and other classes. The series GSE10072 was used for the analysis which contained gene

expression data of 107 people for normal lung tissues as well as tumor lung tissues. The data was then split into subcategories and sample mean of 3 genes was calculated for the whole data and for each subcategory.
a)Normal tissues and tumorous tissues
b) Gene expression in women and men
c) People who have never smoked and current/former smoker

The catagories were not further divided in current and former smoker due to the data set decreasing furthermore.

1.3 Hypothesis Testing

A python[4] program was created to generate the sample mean of 30 random sample. The null hypothesis was formulated on the basis of the sample mean.

$$H_0 : \mu = Samplemean$$

$$H_1 : \mu ! = Samplemean$$

The hypothesis is tested using the z test and t test.

1.3.1 z test

Z-test is a statistical method to determine whether the distribution of the test statistics can be approximated by a normal distribution. It is the method to determine whether two sample means are approximately the same or different when their variance is known and the sample size is large (should be ¿= 30).

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

where,
X¯: mean of the sample. Mu: mean of the population. Sigma: Standard deviation of the population and n: sample size.

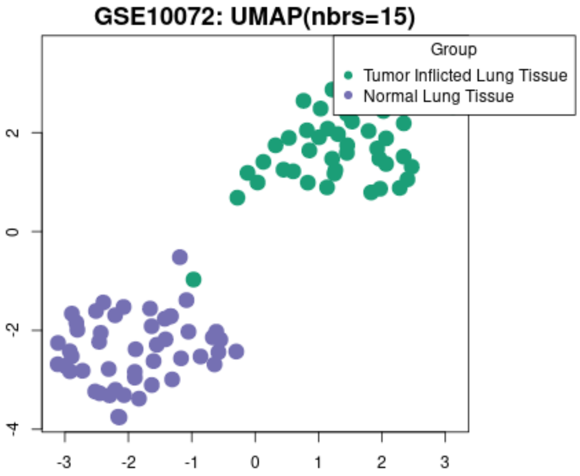
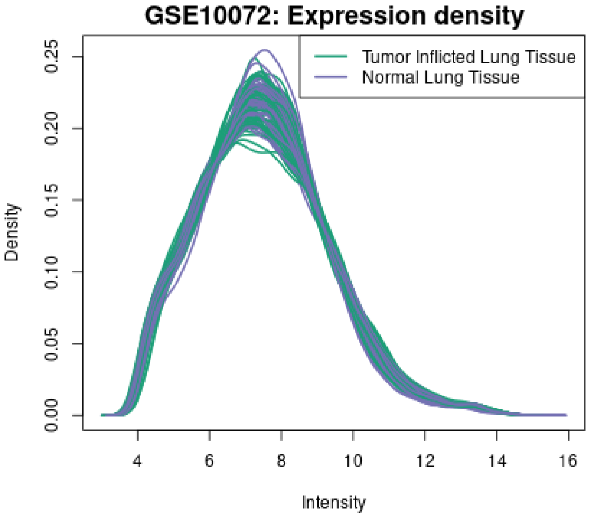
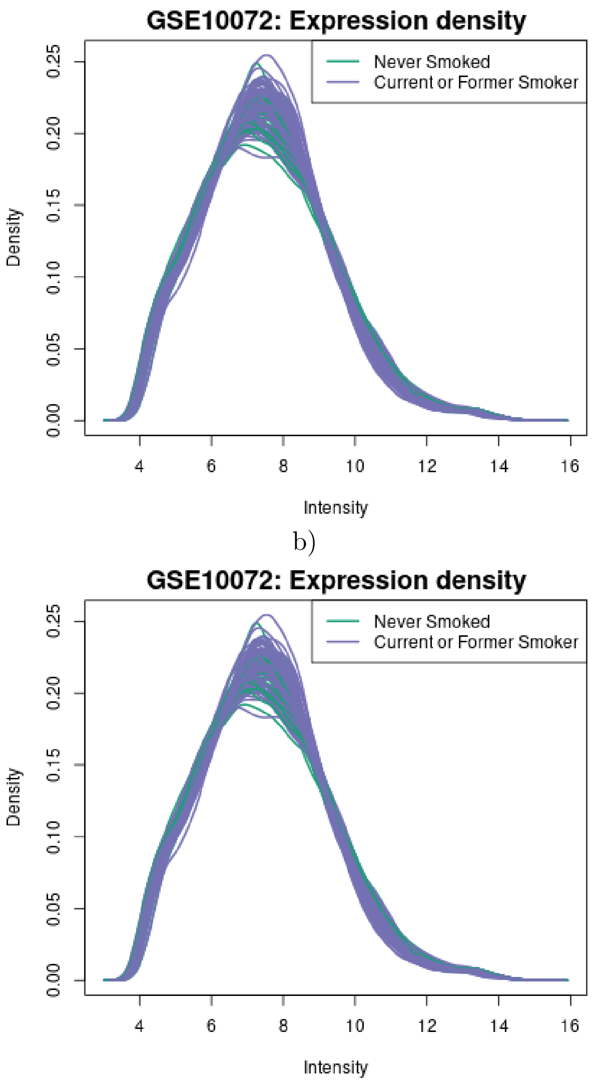


fig 1.3 Uniform Manifold Approximation and Projection (UMAP) plot generated for tumor and normal tissues.



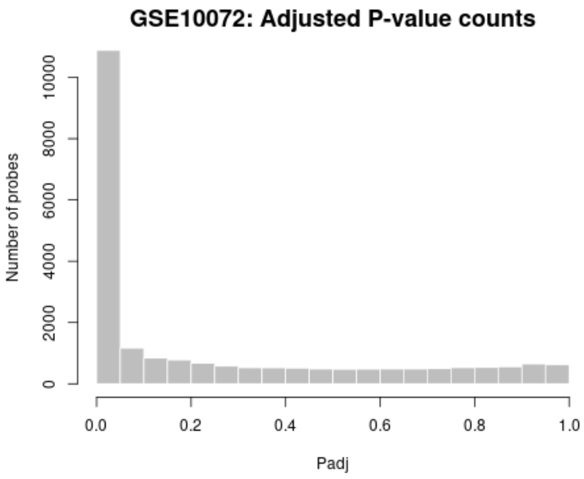
a)



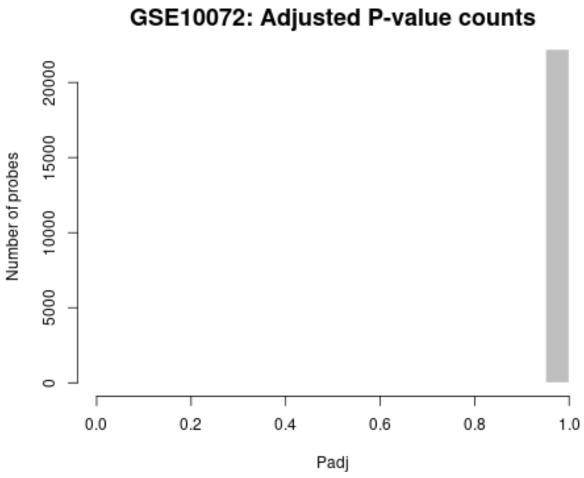
b)

c)

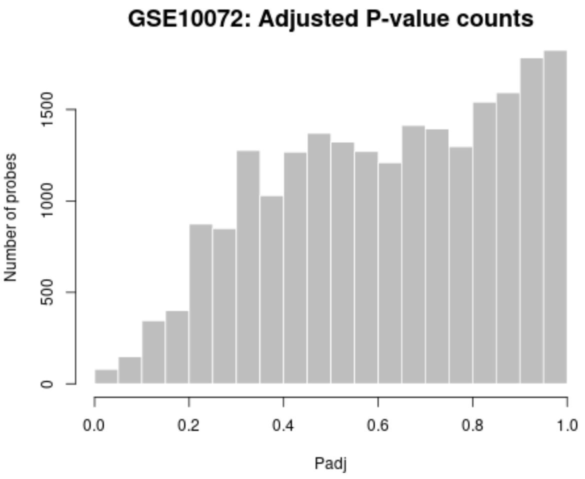
fig 1.4 The distribution of the values of the selected Samples a) Tumor and Normal Tissue b) Men and Women c) Non-smokers and current/former smokers.


GSE10072: Adjusted P-value counts
Number of probes
0
2000
4000
6000
8000
10000
0.0
0.2
0.4
0.6
0.8
1.0
Padj

a)


GSE10072: Adjusted P-value counts
Number of probes
0
5000
10000
15000
20000
0.0
0.2
0.4
0.6
0.8
1.0
Padj

b)



c)

fig 1.5 Use to view the distribution of the P-values in the analysis results for a) Tumor and Normal Tissue b) Men and Women c) Non-smokers and current/former smokers. The P-value here is the same as in the Top differentially expressed genes table and computed using all selected contrasts.

1.3.1 t test

The t-test can be used when the population standard deviations are not known and the sample size is smaller (less than 30). The two sample t-statistic calculation depends on given degrees of freedom, df = n1 + n2 – 2.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

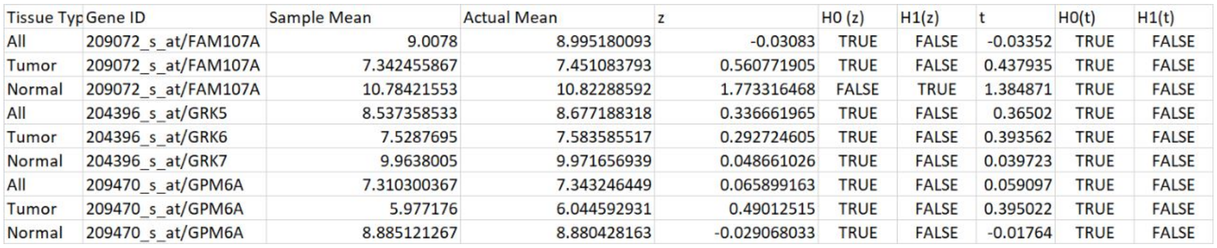
where,
X¯: mean of the sample. Mu: mean of the population. s: Standard deviation of the sample and n: sample size

RESULTS

After testing the null hypothesis with z test as well t test at the significance level,

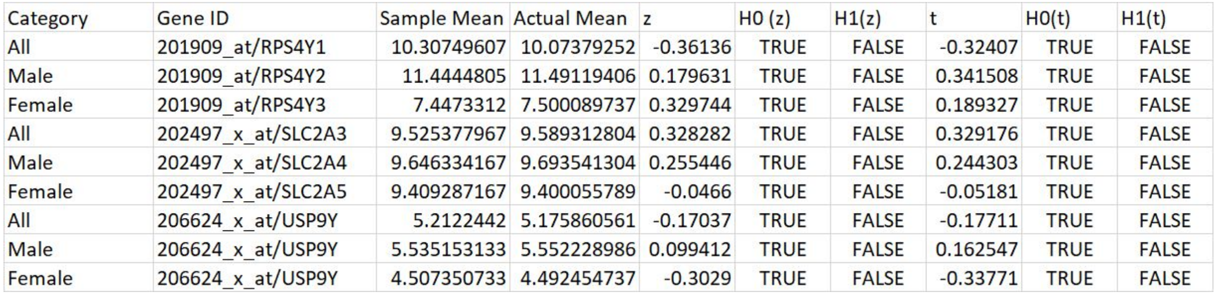
$$\alpha = 0.05$$

we get the result shown in the following tables.



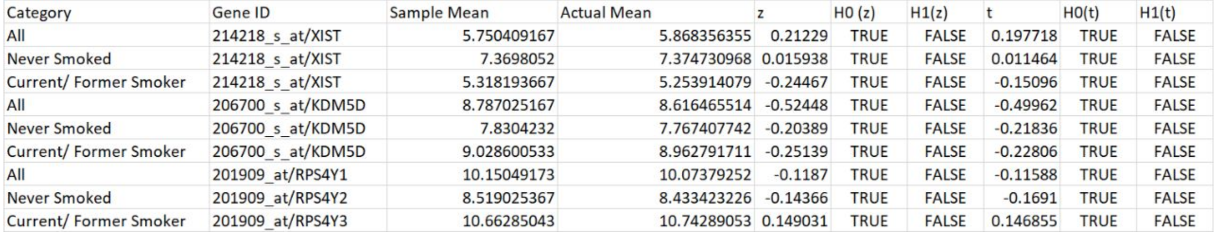
Tissue Typ	Gene ID	Sample Mean	Actual Mean	z	H0 (z)	H1(z)	t	H0(t)	H1(t)
All	209072_s_at/FAM107A	9.0078	8.995180093	-0.03083	TRUE	FALSE	-0.03352	TRUE	FALSE
Tumor	209072_s_at/FAM107A	7.342455867	7.451083793	0.560771905	TRUE	FALSE	0.437935	TRUE	FALSE
Normal	209072_s_at/FAM107A	10.78421553	10.82288592	1.773316468	FALSE	TRUE	1.384871	TRUE	FALSE
All	204396_s_at/GRK5	8.537358533	8.677188318	0.336661965	TRUE	FALSE	0.36502	TRUE	FALSE
Tumor	204396_s_at/GRK6	7.5287695	7.583585517	0.292724605	TRUE	FALSE	0.393562	TRUE	FALSE
Normal	204396_s_at/GRK7	9.9638005	9.971656939	0.048661026	TRUE	FALSE	0.039723	TRUE	FALSE
All	209470_s_at/GPM6A	7.310300367	7.343246449	0.065899163	TRUE	FALSE	0.059097	TRUE	FALSE
Tumor	209470_s_at/GPM6A	5.977176	6.044592931	0.49012515	TRUE	FALSE	0.395022	TRUE	FALSE
Normal	209470_s_at/GPM6A	8.885121267	8.880428163	-0.029068033	TRUE	FALSE	-0.01764	TRUE	FALSE

Table 1.1 Hypothesis testing result for dataset categorized in tumor and normal tissue.



Category	Gene ID	Sample Mean	Actual Mean	z	H0 (z)	H1(z)	t	H0(t)	H1(t)
All	201909_at/RPS4Y1	10.30749607	10.07379252	-0.36136	TRUE	FALSE	-0.32407	TRUE	FALSE
Male	201909_at/RPS4Y2	11.4444805	11.49119406	0.179631	TRUE	FALSE	0.341508	TRUE	FALSE
Female	201909_at/RPS4Y3	7.4473312	7.500089737	0.329744	TRUE	FALSE	0.189327	TRUE	FALSE
All	202497_x_at/SLC2A3	9.525377967	9.589312804	0.328282	TRUE	FALSE	0.329176	TRUE	FALSE
Male	202497_x_at/SLC2A4	9.646334167	9.693541304	0.255446	TRUE	FALSE	0.244303	TRUE	FALSE
Female	202497_x_at/SLC2A5	9.409287167	9.400055789	-0.0466	TRUE	FALSE	-0.05181	TRUE	FALSE
All	206624_x_at/USP9Y	5.2122442	5.175860561	-0.17037	TRUE	FALSE	-0.17711	TRUE	FALSE
Male	206624_x_at/USP9Y	5.535153133	5.552228986	0.099412	TRUE	FALSE	0.162547	TRUE	FALSE
Female	206624_x_at/USP9Y	4.507350733	4.492454737	-0.3029	TRUE	FALSE	-0.33771	TRUE	FALSE

Table 1.2 Hypothesis testing results.



Category	Gene ID	Sample Mean	Actual Mean	z	H0 (z)	H1(z)	t	H0(t)	H1(t)
All	214218_s_at/XIST	5.750409167	5.868356355	0.21229	TRUE	FALSE	0.197718	TRUE	FALSE
Never Smoked	214218_s_at/XIST	7.3698052	7.374730968	0.015938	TRUE	FALSE	0.011464	TRUE	FALSE
Current/ Former Smoker	214218_s_at/XIST	5.318193667	5.253914079	-0.24467	TRUE	FALSE	-0.15096	TRUE	FALSE
All	206700_s_at/KDM5D	8.787025167	8.616465514	-0.52448	TRUE	FALSE	-0.49962	TRUE	FALSE
Never Smoked	206700_s_at/KDM5D	7.8304232	7.767407742	-0.20389	TRUE	FALSE	-0.21836	TRUE	FALSE
Current/ Former Smoker	206700_s_at/KDM5D	9.028600533	8.962791711	-0.25139	TRUE	FALSE	-0.22806	TRUE	FALSE
All	201909_at/RPS4Y1	10.15049173	10.07379252	-0.1187	TRUE	FALSE	-0.11588	TRUE	FALSE
Never Smoked	201909_at/RPS4Y2	8.519025367	8.433423226	-0.14366	TRUE	FALSE	-0.1691	TRUE	FALSE
Current/ Former Smoker	201909_at/RPS4Y3	10.66285043	10.74289053	0.149031	TRUE	FALSE	0.146855	TRUE	FALSE

Table 1.3 Hypothesis testing results for gene expression for non-smoker and former/current smokers.

DISCUSSION

We assumed our data follows normal distribution and since it was smaller in size we also test t distribution. The results show that for the most extent our hypothesis was true for the normal distribution and follows t-distribution for all the data values. The population size was about 107 datapoints, categorizing the dataset diminished the population size further more and there were was much disparity in the population size in categories like men, women, smokers and non-smokers. The data can be compared with and/or used to build a model for prediction gene expression level of certain genes during specific diseases and control situations. This can also be used to detect certain diseases like cancer with the help of just gene expression levels with paramters like age, gender, past health records and many more.

CONCLUSION

Gene expression levels can be analysed by statistics using hypothesis testing which provides an accurate tool for testing if the value of the parameter has changed or not, setting up a model and ensuring the validity of the model. heatmaps can also be analysed to predict statistical patterns and trends in the data. This data can be used to set up machine learning model for predicting and detecting certain diseases with the help of gene expression data.

REFERENCES

1. Home - Geo - NCBI (no date) National Center for Biotechnology Information. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/geo/ (Accessed: April 25, 2023).

2. Lee, M. (2018) Geo accession viewer, National Center for Biotechnology Information. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse10072 (Accessed: April 22, 2023).

3. Davis, S. (2014) Using the GEOquery Package, Using the GEOquery package. Available at: https://bioconductor.org/packages/devel/bioc/vignettes/GEOquery/inst/doc/GEOquery.html (Accessed: April 29, 2023).

4. Lee, M. (2018) Geo accession viewer, National Center for Biotechnology Information. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse10072 (Accessed: April 22, 2023).
5. Shobha Bagai (2022) Scrutinizing String Art Through a Mathematical Lens, PRIMUS, DOI: 10.1080/10511970.2022.2068094

6. About geo2r - geo - NCBI (no date) National Center for Biotechnology Information. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html (Accessed: April 25, 2023).

7. About geo2r - geo - NCBI (no date) National Center for Biotechnology Information. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html (Accessed: April 25, 2023).

8. About geo2r - geo - NCBI (no date) National Center for Biotechnology Information. U.S. National Library of Medicine. Available at: https://www.ncbi.nlm.nih.gov/geo/info/geo2r.html (Accessed: April 25, 2023).