**Differential Gene Expression Analysis**

*Main Contributor: Ziming Wei  
Email Address:* [*zimingwei@hsph.harvard.edu*](mailto:zimingwei@hsph.harvard.edu)

*Harvard ID: 41515244*

**Introduction**

Differential Gene Expression Analysis is an important step of showing how genes are differently expressed between cancer patients and normal individuals. The genes that are most differently expressed would be filtered out and do a Gene Ontology Enrichment to find out what Gene ontology pathways the over or ender expressed genes belongs to. The gene that are differently expressed in the dataset would also be a good starting point for building a Breast Cancer Gene network. The filtered-out genes are passed to build a gene network to see the correlation among the genes. Hubs would be found out in the network and to see if the high connectivity hubs overlap with the genes that are highly differently expressed. The result would also provide a comparison with the feature importance got from the Machine Learning Prediction section to check the validity of gene importance in Breast Cancer.

**Data**

The data we used to conduct the differential gene expression analysis is the raw count data of RNA expression obtained from micro-arrays. The RNA micro-array raw count data is from the GEO data base on ncbi ([*https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45827*](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45827))[1]. The reason why we use un-normalized data is because raw count data is important for DESeq2’s statistical model to hold.[2] The raw RNA micro-array data obtained are 178 .cel files containing gene expression data for each patient. 23 samples are removed from the dataset due to duplication. 155 samples are remained and passed into differential gene expression analysis. All of the samples in the dataset are annotated, and since differential gene expression analysis needs raw count data, only duplication removal is done before the data was passed to the task.

Table

Description automatically generated

*Table 1: Raw RNA micro-array count data, columns are sample ids and indexes are gene probe ids. There are 54675 rows and 155 columns in this table (More rows are not shown due to page limit)*

**Methods**

The 155 .cel files are read into R through *affy* *(version 3.15)* package [3], combining the separated data files into an integrated data frame (Table 1), the columns are sample names, and the index are representing gene probe ids. DESeq (version 1.36.0) [2] was selected as the package to do differential gene expression analysis. Two files are passed to DESeq, one is the raw count data shown in Table 1, and the other one is the annotation table (Table 2), which indicates the labels of the 155 samples. The output of DESeq is in the format of DESeqDataSet, which is later converted into a data frame in R. The output contains gene probe ids, its Log2Fold Change, and the corresponding p-value (Table 3). The gene probe ids are then translated into gene official symbols which are easier to conduct and compare with subsequent analyzes.

Table

Description automatically generated

*Table 2: Annotation table, containing samples and their phenotype (Breast Cancer Status). There are 155 rows in total. (More rows are not shown due to page limit)*

**Results and Discussion**

We could see here from the outcome table (Table 3) that the most significant gene that is differently expressed between cancer samples and normal samples is FRG1BP, which has a log2fold change of -3.895 and with p-value much smaller than 1e-20. This indicates that the expression of FRG1BP gene is under expressed in breast cancer samples compare to the normal samples. After search for detail information of FRG1BP, we found that the gene is coding for promoter and enhancers [4]. It’s under expressed in cancer samples, which may indicate that FRB1BP might be a promoter of a cancer repressor gene. We could also see from the table that gene LEP is also under expressed in the cancer samples. This makes sense since LEP gene provides instructions for making leptin [5], which is released by fat cells, while fat cells accumulate in breast. Thus, the under expression of LEP genes might indicate a mutation in fat cells thus leads to breast cancer. These findings proves our hypothesis that the genes differentially expressed in cancer samples compare to normal sample are possibly more related to breast cancer formation.

A screenshot of a computer

Description automatically generated with low confidence

*Table 3: Outcome Table, containing gene official symbol, their log2fold change between cancer samples and non-cancer samples and their corresponding p-values.*

**Reference**

Gruosso T, Mieulet V, Cardon M, Bourachot B et al. Chronic oxidative stress promotes H2AX protein degradation and enhances chemosensitivity in breast cancer patients. EMBO Mol Med 2016 May;8(5):527-49. PMID: 27006338 [1]

Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014). <https://doi.org/10.1186/s13059-014-0550-8>. [2]

Gautier L, Cope L, Bolstad BM, Irizarry RA (2004). “affy—analysis of Affymetrix GeneChip data at the probe level.” Bioinformatics, 20(3), 307–315. ISSN 1367-4803, doi: 10.1093/bioinformatics/btg405. [3]

Wan C, Borgeson B, Phanse S, Tu F, Drew K, Clark G, Xiong X, Kagan O, Kwan J, Bezginov A, Chessman K, Pal S, Cromar G, Papoulas O, Ni Z, Boutz DR, Stoilova S, Havugimana PC, Guo X, Malty RH, Sarov M, Greenblatt J, Babu M, Derry WB, Tillier ER, Wallingford JB, Parkinson J, Marcotte EM, Emili A. Panorama of ancient metazoan macromolecular complexes. Nature. 2015 Sep 17;525(7569):339-44. doi: 10.1038/nature14877. Epub 2015 Sep 7. PMID: 26344197; PMCID: PMC5036527. [4]

Dubern B, Clement K. Leptin and leptin receptor-related monogenic obesity. Biochimie. 2012 Oct;94(10):2111-5. doi: 10.1016/j.biochi.2012.05.010. Epub 2012 May 22. PMID: 22627381. [5]