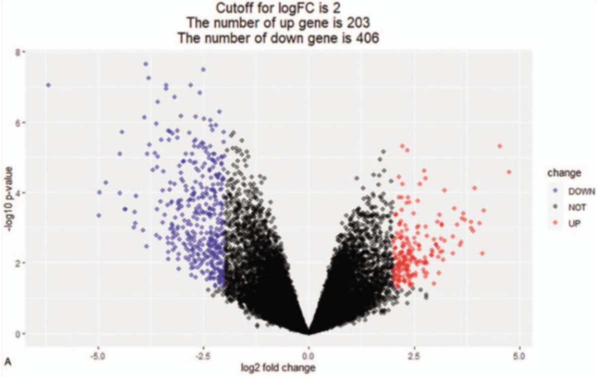
Creating Volcano Plot and Heatmap with Hierarchical Clustering from microarray data using python

In the study, 'Analysis of genes associated with prognosis of lung adenocarcinoma based on GEO and TCGA databases', the investigators aimed to assess potential biomarkers for the prognosis of lung adenocarcinoma (LUAD) by the analysis of gene expression microarrays (Yu and Tian). LUAD is the most common primary lung cancer seen in the United States and falls under the category of on-small cell lung cancer (NSCLC) (Myers and Wallen). Uncovering the underlying molecular mechanisms involved in LUAD is of vital importance for therapy and prevention of the cancer. Biomarker discovery is crucial for testing efficacy during drug discovery and development. They also allow physicians to make more precise diagnoses and prescribe personalized treatments. This study downloaded gene expression data from the GSE118370 dataset from the Gene Expression Omnibus (GEO) database, which included six LUAD samples and 6 normal lung samples. The data was screened using bioinformatics tools including volcano plot and hierarchical clustering for differentially expressed genes (DEGs) between normal and disease samples. In addition, Gene Ontology (GO) terminology, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, and protein–protein interaction (PPI) analysis were also performed to screen for DEGs and other biological pathways. For this project, I used the limma package in R to get the DEGs and perform log transformation on the preprocessed data. I also attempted to recreate the Volcano plot and the Hierarchical cluster.

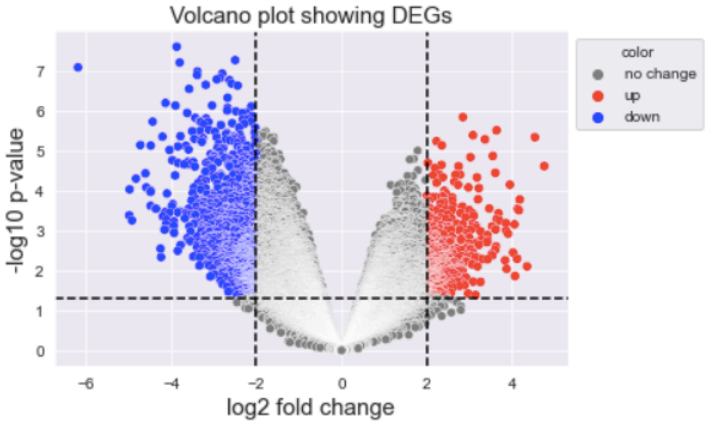
R script: I used the limma package in R to calculate DEGs from the microarray data. I used GEOquery to directly import the data from the GEO database. I used the eBayes() function to calculate the statistics for the following parameters to select DEGs: |log2fold change| > 2 and $P < 0.05$. I also normalized the data in R to generate a matrix with the log2fold change (logFC) values, p-values, adjusted p-value. I used this data later for further analysis to create the plots.

Volcano plots are scatter plots that help us in visualizing changes in the dataset composed of replicate data. Here, they have used volcano plots to plot the log2 fold change of gene expression on the x axis and the corresponding negative log10 p-values on the y axis. In the paper, the DEGs are defined as genes having expression data which corresponds to |log2fold change| > 2 and $P < 0.05$. I used pandas to create a dataframe from the data I obtained from the R script for all the genes, and created subsets for the upregulated DEGs and for the downregulated DEGs. I used a function map_color to add a new column 'color' to the dataframe according to the subset the gene belongs to. I used sns.scatterplot to create the volcano plot and used matplotlib.pyplot to plot vertical and horizontal lines to show the DEG cut-offs. I was able to get a similar plot as in the paper. However, the resulting number of DEGs that they report is much less than the genes that I got corresponding to these parameters. I was only able to get a comparable number of DEGs when I added another parameter for the adjusted p-value as <0.05. I created a second volcano plot to express this data, which was much different from the plot shown in the paper.

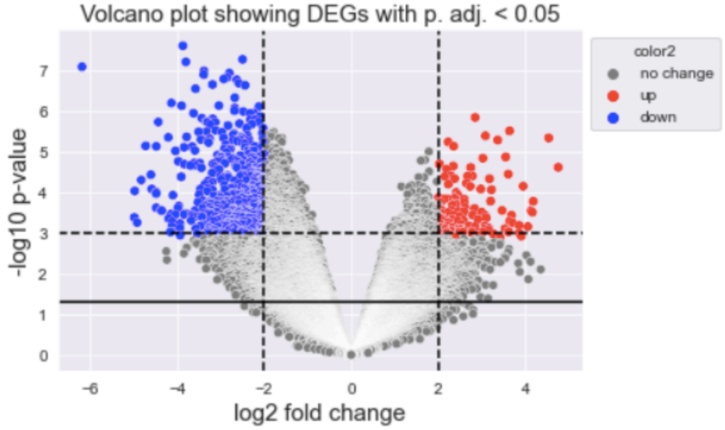
Volcano plot to be recreated:



Plots created:



Plot1: Volcano plot with 1598 DEGs; 518 upregulated, and 1080 downregulated genes. Red: upregulation with |log2fold change | >2, P<.05; blue: downregulation with |log2fold change | >2, P<.05; black: unchanged genes.



Plot2: Volcano plot with 637 DEGs; 123 upregulated and 514 downregulated genes. Red: upregulation with |log2fold change | >2, P<.05, adj. p<0.05; blue: downregulation with |log2fold change | >2, P<.05, adj. p<0.05; grey: unchanged genes.

The two vertical lines in plot 1 show the log2FC value cutoff at x= 2 and x= -2, and the horizontal line shows the p-value cutoff at y = -log10(0.05). According to the paper, all the genes that fall outside these limits should be DEGs. From the subsets I created using these parameters, there should be 518 up regulated and 1080 down regulated genes. However, in the paper they only report 203 upregulated genes and 406 downregulated genes but the DEGs expressed in my volcano plot with 1,598 DEGs match the ones in their volcano plot. In fact, I got comparable values of 123 upregulated genes and 514 down regulated genes only when I added an additional cutoff for the adjusted p-value to be less than 0.05. These values correspond to the values I got from using the limma package. To show the difference in DEGs using the different cut-off values, I created two volcano plots with the different cut off values for the p-value and the adjusted p-value. In the plot 2, the solid line y = 1.3 is the cut-off for p-value while the dashed line y = 3 is the cut-off for adjusted p-value.

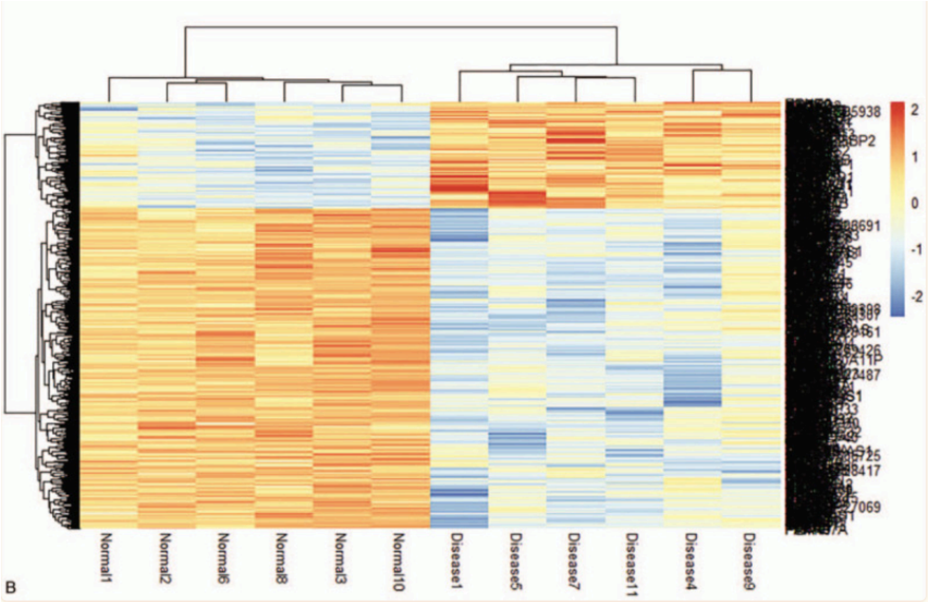
Hierarchical clusters are used to visualize data matrices by drawing rows and columns corresponding to the rows and columns in the data matrix, with the color of a cell corresponding to their value in the matrix (Engle et al). In this paper, the researchers plotted the 609 DEGs in a heatmap showing their expression values and clustered them according to their expression values across samples. They used an agglomerative cluster on both axes. The y axis has the gene clusters and the x axis has the sample clusters. The dendrograms on each side show the relation between the different clusters. To calculate this, the Euclidean distance metric was used the formula for which is as follows:

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Fig1: Euclidean distance formula, cited from Google search

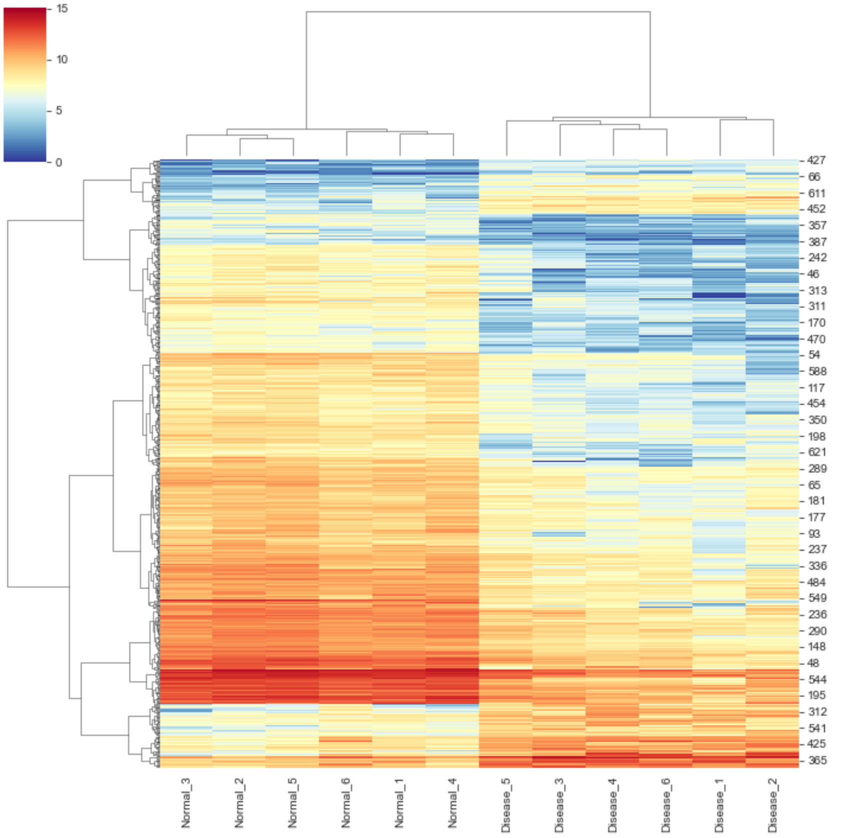
Here p_i is gene 1 in sample 1 and q_i is gene 2 in sample 2 and the program goes through all the genes iteratively calculating the distances between the genes and building the clusters. To plot the heatmap, I used the Affymetrix data provided in the GEO dataset as well as the processed data I obtained from using the limma package. I plotted all the 637 DEGs and used sns.clustermap to create the heatmap.

Heatmap with hierarchical cluster to be recreated:



Plot created: (on next page)

I was able to exactly replicate the clustering on the x axis, however the clustering on the y-axis in my plot is slightly different from the original plot. This is because I used 637 DEGs as compared to the 609 mentioned in the paper. Due to this, my heatmap is longer and the clustering on the y-axis is different from the original plot. I decided to get rid of the labels on the y-axis since they were not adding any information to the plot and only made it more confusing. Since I plotted the expression data on the heatmap, the scale goes from 0 to 15. However, in the original plot the scale goes from -2 to 2; which is not explained anywhere in the paper.



Plot 3: Heatmap of the 637 DEGs with |log2fold change | >2, P<.05. Red: higher expression; blue: lower expression.

Reference:

1. Yu Y, Tian X. Analysis of genes associated with prognosis of lung adenocarcinoma based on GEO and TCGA databases. Medicine (Baltimore). 2020 May;99(19):e20183. doi: 10.1097/MD.0000000000020183. PMID: 32384511; PMCID: PMC7220259.
2. Myers DJ, Wallen JM. Lung Adenocarcinoma. [Updated 2022 Jun 21]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2022 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK519578/
3. Engle, S., Whalen, S., Joshi, A. *et al.* Unboxing cluster heatmaps. *BMC Bioinformatics* **18** (Suppl 2), 63 (2017). https://doi.org/10.1186/s12859-016-1442-6
4. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). “limma powers differential expression analyses for RNA-sequencing and microarray studies.” *Nucleic Acids Research*, **43**(7), e47. doi: 10.1093/nar/gkv007.
5. Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss. 03021
6. Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. Nature 585, 357-362 (2020). DOI: 10.1038/s41586-020-2649-2. (Publisher link).
7. McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).
8. J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.