**INDIAN INSTITUTE OF TECHNOLOGY MANDI**
Himachal Pradesh, INDIA - 175005

# Diagnosis of Lung Cancer Using Multi-Omics

Krishna Sai (B19005)
B19005@students.iitmando

Under the supervision of
Dr. Shyam K Masakapalli
School of Biosciences and Bioengineering

# Abstract

Non-small cell lung cancer (NSCLC) remains a leading cause of cancer death globally. More accurate and reliable diagnostic methods/biomarkers are urgently needed. Joint application of metabolomics and transcriptomics technologies possesses the high efficiency of identifying key metabolic pathways and functional genes in lung cancer patients. In this study, we performed an metabolomics analysis of 82 subjects in which 41 are patients and 41 are healthy controls; we observed metabolites in both plasma and serum samples; 27 identified metabolites from plasma samples and 13 identified metabolites from plasma samples were significantly different between NSCLC patients and healthy controls; 27 metabolites were chosen as combinational potential biomarkers for NSCLC using plasma data and 13 metabolites were chosen as combinational potential biomarkers for NSCLC using serum data. Potential diagnostic implications of the metabolic characteristics in NSCLC was studied. The metabolomics results were further verified by transcriptomics analysis of NSCLC patients lung tissue samples and adjacent peritumoral tissues. This analysis identified genes with significantly different expressions in cancer cells compared to normal controls, which in turn defined pathways implicated in the metabolism of the compounds revealed by metabolomics analysis. we built a connected network of metabolites and genes, which shows a good correspondence between the transcriptome analysis and the metabolites selected for diagnosis. In conclusion, this work provides idea that the metabolic biomarkers identified may be used for NSCLC diagnosis and screening. Comprehensive analysis of metabolomics and transcriptomics data offered a validated and comprehensive understanding of metabolism in NSCLC.

# Contents

**Contents** ii

**1 Introduction** 1

1.1 Biomarkers. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

**2 Metabolomics Methodology** 2

2.1 Dataset Description . . . . . . . . . . . . . . . . . . . . . . . . . 2

2.2 Proposed Architecture . . . . . . . . . . . . . . . . . . . . . . . 3

2.3 Biomarker Selection . . . . . . . . . . . . . . . . . . 3

2.4 Machine Learning Algorithm . . . . . . . . . . . . 4

**3 Transcriptomics** 5

3.1 Dataset Description . . . . . . . . . . . . . . . . . . . . . . . . . .5

3.2 Proposed Architecture. . . . . . . . . . . . . . . . . . . . . . . . . . . . . .6

3.2.1 Data upload. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

3.2.2 Quality Control. . . . . . . . . . . . . . . . . . . . . . . . . . 6

3.2.3 Mapping. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

3.2.4 Counting the number of reads per annotated gene. . . . . . 6

3.2.5 Analysis of differential gene expression. . . . . . . . . . . . .6

3.2.6 Functional Enrichment Analysis . . . . . . . . . . . . . . . . . .8

**4 Conclusion** 10

**5 Limitations** 11

**6 References** 12

Chapter 1

# Introduction

Lung cancer has been a threat in the field of medicine for a long time. According to recent studies, an estimation of 235,760 new cases will be diagnosed, and 131,880 people will die from lung cancer in 2021 in the US. The same survey results reported 228,820 diagnoses and 135,720 deaths in 2020. So, without any doubt, the number of patients is increasing at an alarming rate every year. Patients with invasive lung and bronchus cancer were identified from the SEER 18-registry database, covering 28% of the US population. Lung cancer patients can be recovered if they are diagnosed early. Life-loss years vary from 6.16 for Stage I cancer to 16.21 for Stage IV.

## 1.1 Biomarkers

Metabolomics aims to comprehensively analyze wide arrays of metabolites in biological samples. In order to be used as diagnostic markers for biological conditions, including diseases and responses to chemical treatment, metabolite measurements are of fundamental regulatory importance. It is a useful area of research for the study of disease detection. In this sense, the identification and detection of diseases can greatly benefit from the use of biomarkers. Cancer is one of the most prevalent illnesses afflicting humans today. Assume that the metabolites are identified and are either absent or present in amounts that are acceptable in healthy individuals. In that case, it will significantly affect the detection of cancer.

The use of genome-wide approaches to gene expression as a component of the biomarker discovery process enables a thorough search for genes with cancer-associated gene expression for further biomarker development as well as the generation of hypotheses regarding the biological mechanisms underlying many of the observed gene expression differences. The study of transcriptional regulation and gene expression in various biological contexts is known as transcriptomics. The entire "transcriptome," or set of expressed genes, in a given tissue at a given time can now be quantitatively assessed thanks to scientific advancements. These approaches provide a powerful tool for understanding complex biological systems and for developing novel biomarkers

Chapter 2

# Metabolomics Methodology

We worked with some lung cancer patients using the metabolomic biomarkers found in blood samples from their plasma and serum. In order to identify the most important metabolomic biomarkers, we analyzed 158 metabolites. Based on the specific metabolites, we distinguished between healthy individuals and lung cancer patients. Using the Agglomerative Hierarchical Clustering Technique, we also discovered the relationships and hierarchical differences between the metabolites. Finally, we assessed the accuracy of our methods for locating lung cancer patients. We used the Shapiro-Wilk Test during the feature selection process to determine whether the distribution of each feature in our dataset was normally distributed. Then, we verified that the variances in our dataset remained homogeneous (or equal). In order to do so, we used Levene's Test for features without normal distribution and Bartlett's Test for features with normal distribution [9–11]. In order to determine the most dominant metabolites from a large number of lists of those, Student's t-Test [12] for features with homoscedasticity (or equal variances) and Kruskal-Wallis Test for the features with heteroscedasticity (or unequal variances) were both used. With plasma and serum samples, we found the most prevalent metabolites.

## 2.1 Dataset Description

Oliver Fiehn created the dataset that we used for our research with the study ID: ST000392. It was created using the GC-TOF-MS time-of-flight mass spectrometry method. Serum and Plasma made up the entirety of the samples that were gathered. Both samples had 82 participants, and the dataset contained information on 158 metabolites for each of them. Each group of subjects had 41 cancer patients and 41 control subjects; 20 of the subjects were men, and the remaining 60 were women. Tables 1 and 2 provide summaries of each of these.

.

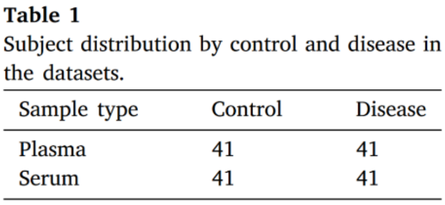


**Table 1**
Subject distribution by control and disease in the datasets.

| Sample type | Control | Disease |
|---|---|---|
| Plasma | 41 | 41 |
| Serum | 41 | 41 |

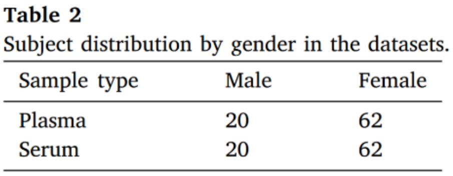
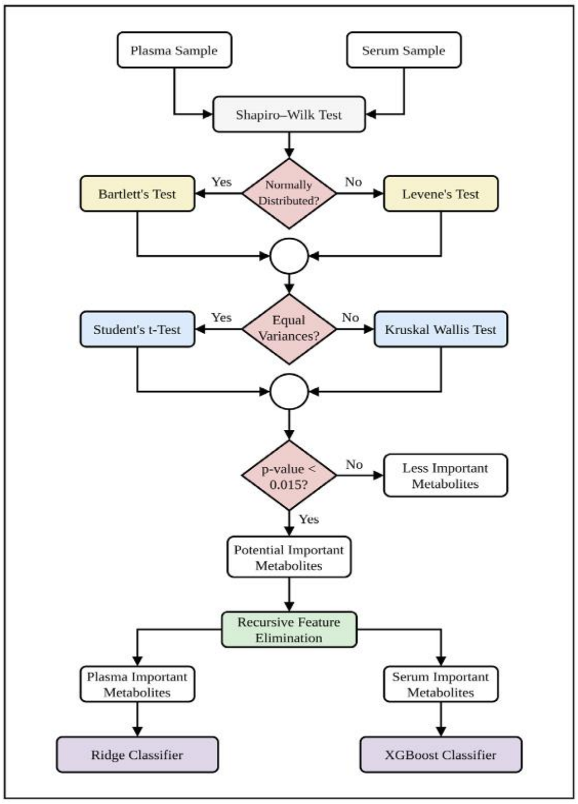


**Table 2**
Subject distribution by gender in the datasets.

| Sample type | Male | Female |
|---|---|---|
| Plasma | 20 | 62 |
| Serum | 20 | 62 |

## 2.2 Proposed Architecture

There is a flow diagram given as Fig. 1. This diagram is a summary of our different approaches with Plasma and Serum samples to make it easier to think and go through with them, which we will go deeper with the details of the approaches in the next sessions.

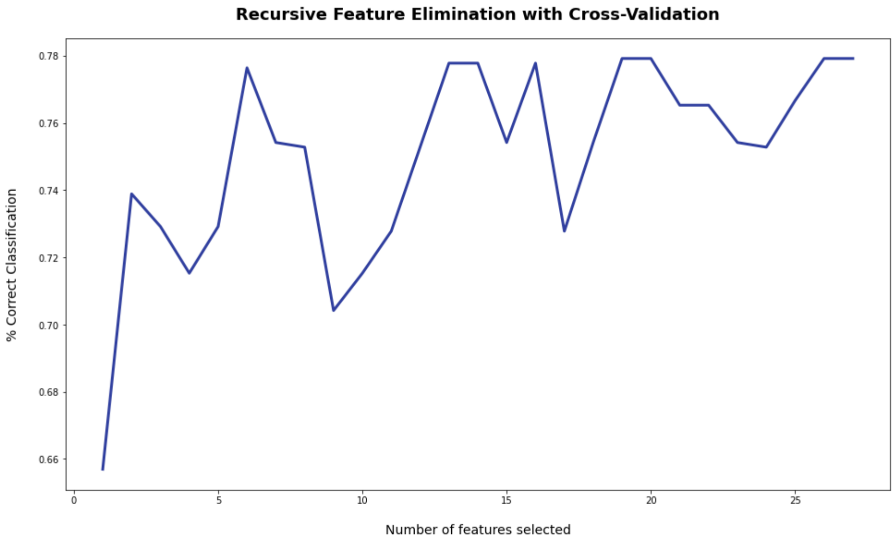


## 2.3 Biomarker Selection

We employed our methodology on the dataset prepared by Oliver Fiehn. Going through all these steps, we obtained some important metabolites based on the ***p***-value of the features, using Student's t-Test and Kruskal–Wallis Test. We marked a metabolite as a potential for which a ***p***-Value < 0.015 was found. There were 26 such metabolites in the plasma sample. It was 16 in the case of the Serum sample.

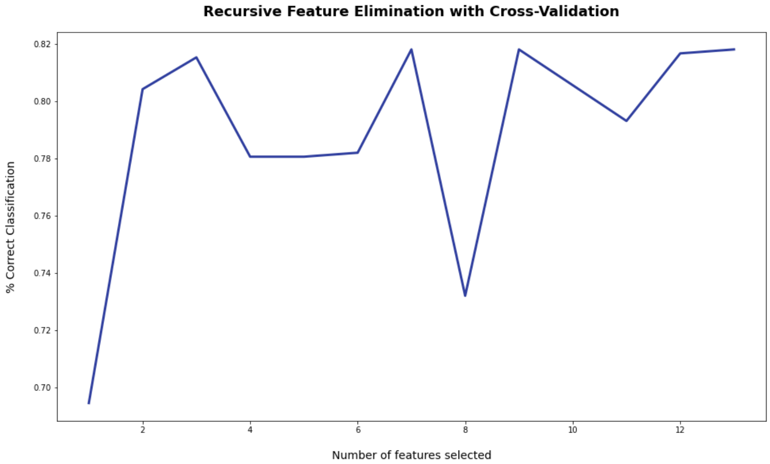
**Potential Biomarkers of Plasma: 27**

*uric acid, tryptophan, taurine, pyruvic acid, pyrophosphate, phosphoethanolamine , phenol, nornicotine , N-methylalanine , methionine sulfoxide, maltotriose , maltose, malic acid, lactic acid, lactamide , hypoxanthine, glycerol-3-galactoside, glutamine, citrulline, benzoic acid, aspartic acid, asparagine, alpha-ketoglutarate, adenosine-5-monophosphate, 5-methoxytryptamine, 5-hydroxynorvaline NIST, 3-phosphoglycerate*



**Optimal number of features after using Recursive Feature Elimination: 19**
*'Uric acid', 'tryptophan', 'taurine', 'pyruvic acid', 'pyrophosphate', 'phosphoethanolamine', 'phenol', 'nornicotine', 'methionine sulfoxide', 'maltose', 'lactic acid', 'hypoxanthine', 'glutamine', 'citrulline', 'aspartic acid', 'adenosine-5-monophosphate', '5-methoxytryptamine', '5-hydroxynorvaline NIST', '3-phosphoglycerate'*

**Potential Biomarkers of Serum: 13**
*uric acid, threonine, taurine, phenylalanine, phenol, N-methylalanine , malic acid, lactic acid, inosine, glutamic acid, deoxypentitol , cholesterol, aspartic acid*



**Optimal number of features after using Recursive Feature Elimination: 7**
*['threonine', 'taurine', 'phenol', 'N-methyl alanine', 'inosine', 'cholesterol', 'aspartic acid']*

## 2.4 Machine Learning Algorithm

After getting the most potential Biomarker I had developed a machine Learning Algorithm, trained the Model with the data available (GCMS DATA) with 41 people as control and 41 as cancer and used 92% of our whole data for training data and 8% of the data as testing data
Giving the more weightage to the most prominent metabolites we have filtered using different statistical techniques such as ridge classifier for plasma samples and XGBoost for serum samples
Accuracy score for training data (plasma sample - ridge classifier): 76.71
Accuracy score for testing data (plasma sample - ridge classifier): 55.56
Accuracy score for training data (serum sample - xgboost classifier): 100.00
Accuracy score for testing data (serum sample - xgboost classifier): 77.78
Later from the transcriptomics approach we get the most prominent genes which are responsible for these metabolites formation and we also train our model with the transcriptomics data and integrate both the data to the model and it now checks both the samples and gives us the output that the peroson is Diagnosed with Lung Cancer or not with the Accuracy of the Prediction. It also Displays the most prominent Biomarkers which helps in treatment.

Chapter 3

# Transcriptomics Methodology

I also used transcriptomics data from the RNA-seq study of SRSF2 over-expressing lung cancer cell lines, accession number GSE186872, which I got from the NCBI. In this study, the transcriptome difference between clinically hyper-progressive samples and samples that responded to immunotherapy was examined using publicly available datasets. Neoantigen-related splicing events were found to have the highest concentration of SRSF2 spliceosome factors. Following SRSF2 over-expression, two lung cancer cell lines showed a similar splicing pattern and increased cell invasion. This study demonstrates that dysregulation of SRSF2-related splicing may be a key aspect of the occurrence of immune hyper-progress.

## 3.1 Dataset Description

The dataset we used in our study was produced under the GO Accession Number GSE186872, that deals with the RNA-seq study of SRSF2 over-expressing lung cancer cell lines, This an RNA-seq Data of lung Tissue.
*Instrument:* Illumina NovaSeq 6000
*Strategy:* RNA-Seq
*Organism:* Homo sapiens
*Source:* TRANSCRIPTOMIC
*Selection:* cDNA
*Layout:* PAIRED

*Construction protocol: lIn accordance with the manufacturer's instructions, RNA was isolated from 1-2x106 cells kept in TRIzol reagent.* In order to prepare RNA libraries for sequencing, standard Illumina procedures were followed.

This information contrasts the overexpression of the SRSF2 gene in two different cell lines from lung cancer patients. 1. A549 Cell Line, 2. H1650 Cell Line. This data consists of 12 SRA samples, six from one cell line and the remaining six from another. In each cell line, they took three healthy individuals and three lung cancer patients, sequenced their transcriptomes, and checked for the overexpression of the SRSF2 gene.

## 3.2 Proposed Architecture

### 3.2.1 Data Upload

Upload the data to the galaxy server, upload 12 SRA files in to the server in the FASTQ format (our reads are Paired end).

#### 3.2.2 Quality Control

Errors are introduced during sequencing, such as incorrect nucleotide calls. These are brought on by the distinct sequencing platforms' technical constraints. Sequencing mistakes may skew the analysis and cause the data to be incorrectly interpreted. If the reads are longer than the sequenced fragments, adapters might also be present; trimming these may increase the number of reads mapped. We used Cutadapt (Marcel 2011) to enhance the quality of the sequences through trimming and filtering, MultiQC (Ewels et al. 2016) to aggregate generated reports, and FastQC to generate a report of sequence quality.

#### 3.2.3 Mapping

We must first identify the sequences' genomic origin in order to determine which genes they belong to in order to make sense of the reads. Aligning or "mapping" the reads to the reference is the process used when there is a reference genome for the organism. This can be compared to putting together a jigsaw puzzle, but sadly not every piece is unique. The authors of this study used lung cells from Homo sapiens. Therefore, we should map the quality-controlled sequences to the Hoo sapience Gch38 reference genome.

#### 3.2.4 Counting the number of reads per annotated gene

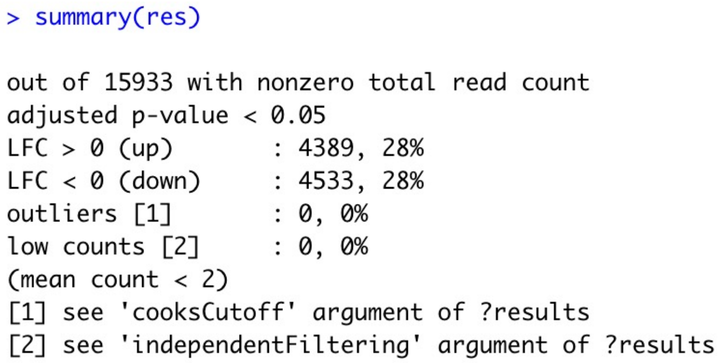
Quantifying the number of reads per gene, or more precisely the number of reads mapping to each gene's exons, is a crucial first step in comparing the expression of individual genes under various circumstances (such as with or without PS depletion). There are two main read counting tools: featureCounts (Liao et al. 2013) or HTSeq-count (Anders et al. 2015). Additionally, STAR enables read counts while mapping; the outcomes are the same as those of HTSeq-count. While most analyses can be completed with this output, featureCounts offers greater customization for read counting. We now use featureCounts to count the number of reads per annotated gene since we selected to use the tutorial's featureCounts flavor.
(Note: The process described above is the standard methodology for RNA sequence data analysis, but I started analyzing the RNA SEQ data using R programming and libraries from BIOCONDUCTOR after receiving the raw gene count data from the GEO Data set.)

#### 3.2.5 Analysis of the differential gene expression

We used the RAW GENE COUNT dataset for the Differential Gene Expression analysis. Import the metadata and count data first, straight from the internet. Run the DESeq2 pipeline after setting up the DESeq data set. Rearrange the results by p-value after obtaining the results for the SRSF2 Over expressing versus control. To determine the number of genes that are up- or down-regulated, call summary on the results object.

The gene IDs, Base Mean, Log2foldchange, Pvalue, and Padj Value are all included in the resource.

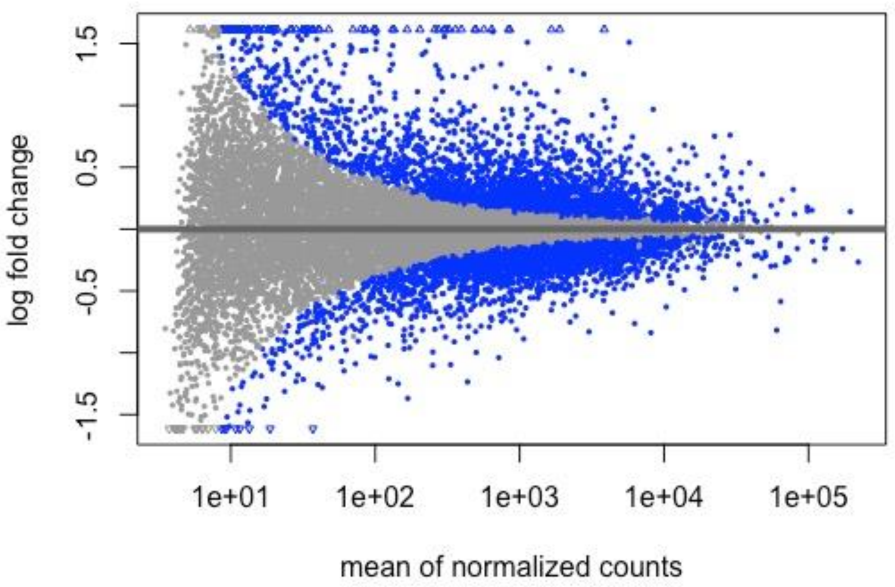
.



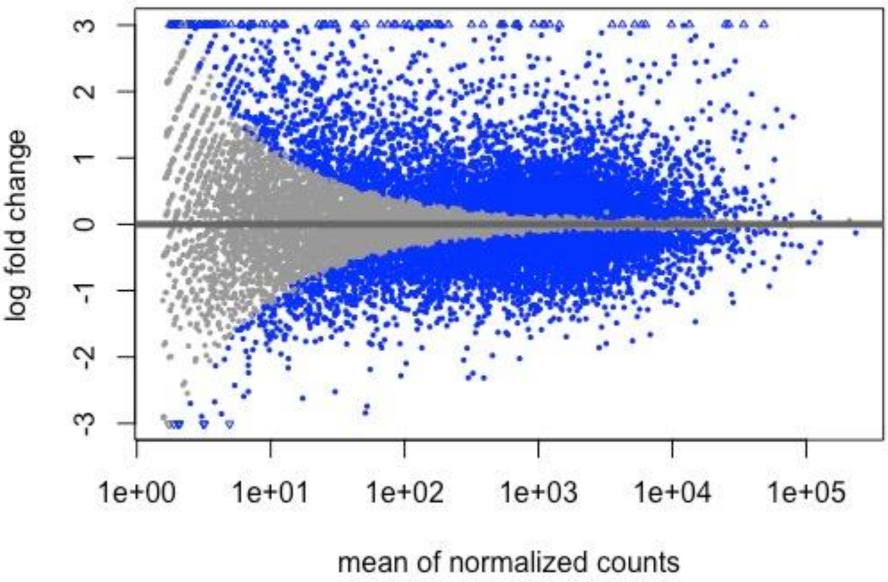
```
> summary(res)

out of 15933 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 4389, 28%
LFC < 0 (down)     : 4533, 28%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 2)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

We created the MA plot for both cell lines based on the differential gene expression when compared to the control.



*A459 Cell line (Blue Colour Indicates the Differentially expressed Genes)*

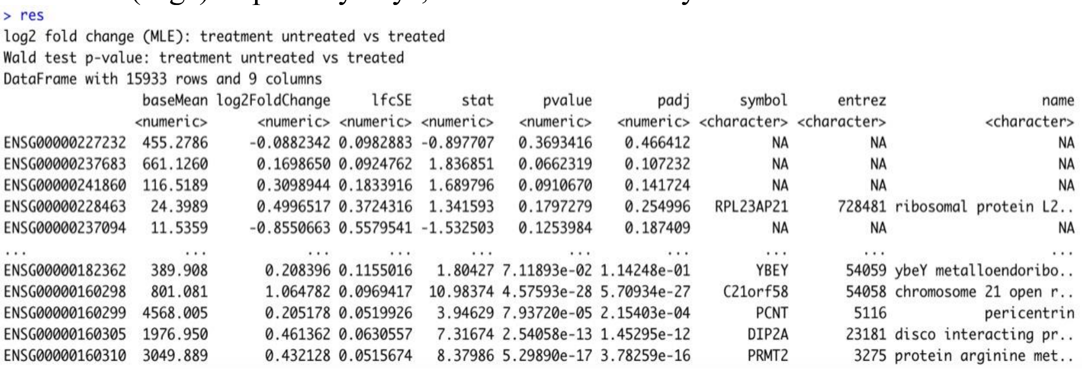


*H1650 Cell line (Blue Colour Indicates the Differentially expressed Genes)*

After applying the DESeq function to the dataset, we obtained 15933 Gene's data. Since we need to extract the most potentially useful genes for our analysis, we would like to do so using a fold change > 2 (or  1/2). Now, we will only choose genes with a fold change (FC) of greater than 2 or less than 0.5. We filter for abs(log2FC)>1abs(log2FC)>1 (Which implies FC > 2 or FC  0.5) because the DESeq2 output file contains log2FClog2FC rather than FC itself.
From all these filters we have filtered 1354 important Genes out 16,599 Genes.

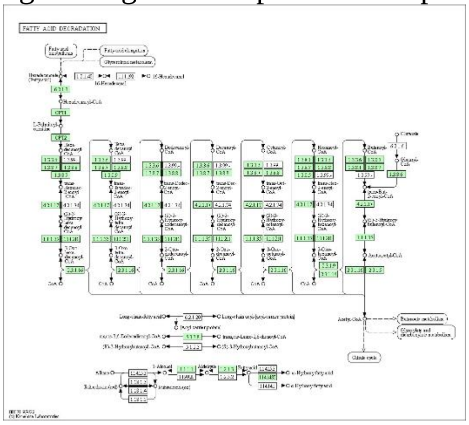
### 3.2.6 Functional Enrichment Analysis of DE Genes

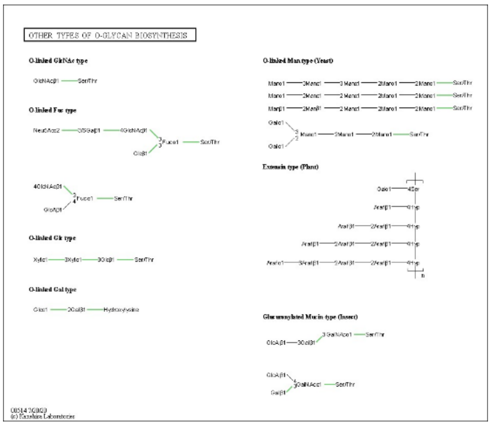
Since we mapped and counted against the Ensembl annotation, our results only have information about Ensembl gene IDs. But our pathway analysis downstream will use KEGG pathways, and genes in KEGG pathways are annotated with Entrez gene IDs. Here we're using the organism package ("org") for Homo sapiens ("Hs"), organized as an AnnotationDbi database package ("db") using Entrez Gene IDs ("eg") as primary keys, From all these analyses our res data look like this:

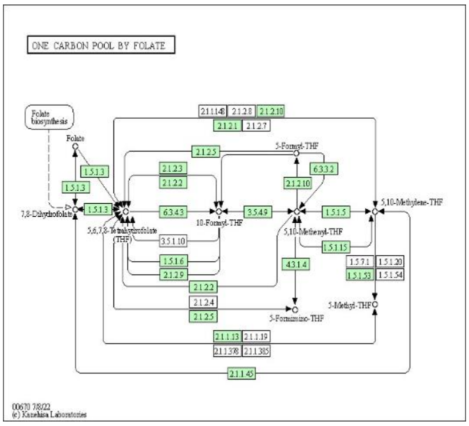


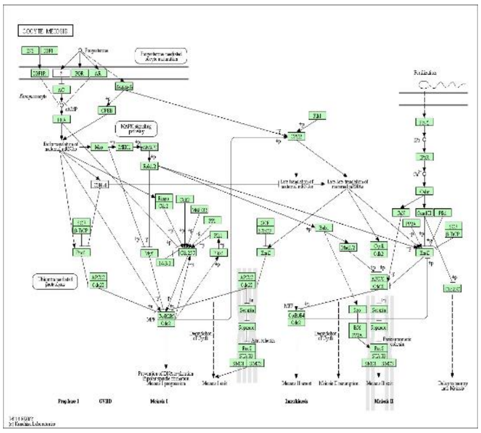
```
> res
log2 fold change (MLE): treatment untreated vs treated
Wald test p-value: treatment untreated vs treated
DataFrame with 15933 rows and 9 columns
                 baseMean log2FoldChange     lfcSE      stat      pvalue       padj      symbol      entrez                   name
                <numeric>      <numeric> <numeric> <numeric>   <numeric>  <numeric> <character> <character>            <character>
ENSG00000227232  455.2786     -0.0882342 0.0982883 -0.897707   0.3693416   0.466412          NA          NA                     NA
ENSG00000237683  661.1260      0.1698650 0.0924762  1.836851   0.0662319   0.107232          NA          NA                     NA
ENSG00000241860  116.5189      0.3098944 0.1833916  1.689796   0.0910670   0.141724          NA          NA                     NA
ENSG00000228463   24.3989      0.4996517 0.3724316  1.341593   0.1797279   0.254996   RPL23AP21      728481 ribosomal protein L2..
ENSG00000237094   11.5359     -0.8550663 0.5579541 -1.532503   0.1253984   0.187409          NA          NA                     NA
...                   ...            ...       ...       ...         ...        ...         ...         ...                    ...
ENSG00000182362   389.908       0.208396 0.1155016   1.80427 7.11893e-02 1.14248e-01        YBEY       54059 ybeY metalloendoribo..
ENSG00000160298   801.081       1.064782 0.0969417  10.98374 4.57593e-28 5.70934e-27     C21orf58       54058 chromosome 21 open r..
ENSG00000160299  4568.005       0.205178 0.0519926   3.94629 7.93720e-05 2.15403e-04        PCNT        5116             pericentrin
ENSG00000160305  1976.950       0.461362 0.0630557   7.31674 2.54058e-13 1.45295e-12       DIP2A       23181 disco interacting pr..
ENSG00000160310  3049.889       0.432128 0.0515674   8.37986 5.29890e-17 3.78259e-16       PRMT2        3275 protein arginine met..
```

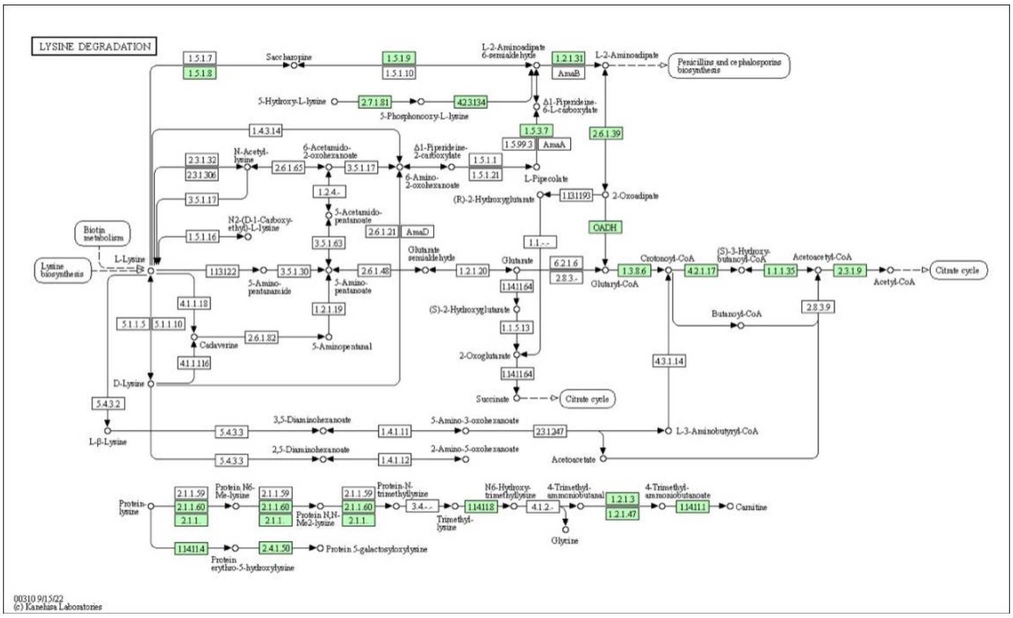
Kegg Pathway Analysis: The gageData package has pre-compiled databases mapping genes to KEGG pathways and GO terms for common organisms. kegg.sets.hs is a named list of 229 elements. Each element is a character vector of member gene Entrez IDs for a single KEGG pathway. we pick the most important upregulated and downregulated genes and plot their pathways. I have picked top 5 upregulated genes and plotted their pathways.




OTHER TYPES OF O-GLYCAN BIOSYNTHESIS
O-linked GlcNAc type
O-linked Man type (Yeast)
O-linked Fuc type
Extensin type (Plant)
O-linked Glc type
O-linked Gal type
Glucosylated Mucin type (Insect)
00514 7/26/22
(c) Kanehisa Laboratories







Chapter 4

# Conclusion

One of the most common fatal diseases in the world is still cancer. Numerous cancer subtypes claim the lives of numerous people each year. According to this study, metabolomic biomarkers discovered from differentially expressed metabolites in lung cancer patients can have a significant impact on the medical field because early and affordable methods of identifying cancer patients are essential to saving lives. We attempted to integrate transcriptomics and metabolomics in this work. We combine the genes from the metabolic studies and the genes from the transcriptomic studies to create a machine learning model for the purpose of diagnosis. This model also aids in selecting the best course of treatment for the patient and monitoring the effectiveness of that treatment.

# Limitations

For the integration of two Omics, the same type of data from the same samples must be used; however, in our study, the data came from two different sources: the metabolomics data came from blood and plasma samples taken from 82 individuals, while the transcriptomics data came from tissue samples taken from different people. The problem is that these two samples cannot be accurately analyzed together; the results would be incorrect if we combined the blood and plasma samples from metabolomics with the tissue samples from transcriptomics.

# References

[1] American Cancer Society. Cancer facts & figures 2021. Atlanta: American Cancer Society; 2021, p. 17.

[2] American Cancer Society. Cancer facts & figures 2020. Atlanta: American Cancer Society; 2020, p. 17.

[3] Howlader N, Forjaz G, Mooradian MJ, Meza R, Kong CY, Cronin KA, et al. The effect of advances in lung-cancer treatment on population mortality. N Engl J Med 2020;383(7):640–9.

[4] Mar J, Arrospide A, Iruretagoiena ML, Clèries R, Paredes A, Elejoste I, et al. Changes in lung cancer survival by TNM stage in the basque country from 2003 to 2014 according to period of diagnosis. Cancer Epidemiol 2020; 65:101668.

[5] Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev 2007;26(1):51–78.

[6] Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L. Metabolite profiling: from diagnostics to systems biology. Nature Rev Mol Cell Biol 2004;5(9):763–9.

[7] Davidson I, Ravi S. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In: European conference on principles of data mining and knowledge discovery. Springer; 2005, p. 59–70.

[8] Royston P. Approximating the Shapiro-Wilk W-test for non-normality. Stat Comput 1992;2(3):117–9.

[9] de Gois G, de Oliveira-Júnior JF, da Silva Junior CA, Sobral BS, de Bodas Terassi PM, Junior AHSL. Statistical normality and homogeneity of a 71-year rainfall dataset for the state of Rio de Janeiro—Brazil. Theor Appl Climatol 2020;141(3):1573–91.

[10] Arsham H, Lovric M. Bartlett's test. 2011.

[11] Brown MB, Forsythe AB. Robust tests for the equality of variances. J Amer Statist Assoc 1974;69(346):364–7.

[12] Kumar N, Shahjaman M, Mollah MNH, Islam SS, Hoque MA. Serum and plasma metabolomic biomarkers for lung cancer. Bioinformation 2017;13(6):202.

[13] Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Amer Statist Assoc 1952;47(260):583–621.

[14] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn 2002;46(1):389–422.

[15] Miyamoto S, Taylor SL, Barupal DK, Taguchi A, Wohlgemuth G, Wikoff WR,et al. Systemic metabolomic changes in blood samples of lung cancer patients identified by gas chromatography time-of-flight mass spectrometry. Metabolites 2015;5(2):192–210.

[16] Masrur T, Hasan MAM. Identification of metabolomic biomarker using multiple statistical techniques and recursive feature elimination. In: 2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2). IEEE; 2019, p. 1–4.

[17] Masrur T, Hasan MAM, Mondal MNI. Metabolomic biomarker identification for lung cancer by combining multiple statistical approaches. In: 2019 international conference on electrical, computer and communication engineering (ECCE). IEEE;2019, p. 1–6.