INDIAN INSTITUTE OF TECHNOLOGY, MANDI

BY516: INTRODUCTION TO "OMICS" AND SYSTEMS ANALYSIS

Diagnosis of Lung Cancer Using Multi-Omics

By Krishna Sai (B19005)

Under the supervision of

Dr. Shyam Kumar Masakapalli

Dr. Tulika Srivatsava

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.

Metabolomics aims to comprehensively analyse wide arrays of metabolites in biological samples. Metabolite measurements bear fundamental regulatory importance to be used as diagnostic markers for biological conditions, including diseases and response to chemical treatment. It is a beneficial field of study in the field of disease detection. In this regard, biomarkers can play a crucial role in disease detection and identification. One of the most common diseases faced with the earth today is cancer. Suppose the metabolites are recognized, either not present or present up to a tolerable amount in healthy cases. In that case, it will have a huge impact on the identification of cancer.

The use of genome-wide gene expression approaches as part of the biomarker discovery process allows for a broad search for genes with cancer-associated gene expression for further biomarker development, and it can generate hypotheses as to the biological processes responsible for many of the observed gene expression differences. Transcriptomics is the study of how our genes are regulated and expressed in different biological settings. Technical advances now enable quantitative assessment of all expressed genes (i.e., the entire ‘transcriptome’) in a given tissue at a given time. These approaches provide a powerful tool for understanding complex biological systems and for developing novel biomarkers

I had worked with some lung cancer patients using those metabolomic biomarkers present in Plasma and Serum samples of blood of those patients. I had analysed 158 metabolites to find out the most significant metabolomic biomarkers. I had classified a person as a normal or a lung cancer patient based on the specific metabolites. I had also found the hierarchical differences and relations between the metabolites using the Agglomerative Hierarchical Clustering Technique. Eventually, I had evaluated our approaches in terms of accuracy to identify lung cancer patients. In feature selection, for looking into the distribution of each feature in our dataset, I had used Shapiro–Wilk Test to check if the features were normally distributed. Then, I checked if our dataset-maintained homogeneity (or equality) of variances. For that, I used Bartlett’s Test for the features with normal distribution and Levene’s Test for the features without normal distribution [9–11]. Finally, 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 used and using the test statistics and p-Value, to obtain the most dominant metabolites from a large number of lists of those. I obtained the most dominant metabolites in the case of Plasma and Serum Samples.

I also worked with the Transcriptomics data which I have obtained from the NCBI under the Accession number GSE186872 which is the RNA-seq study of SRSF2 over-expressing lung cancer cell lines, this study inspected transcriptome difference between clinical hyper-progressive samples to immunotherapy-response samples based public datasets. Spliceosome factors SRSF2 were found most enriched in neoantigen-related splicing events. Similar splicing pattern and increased cell invasion were found in two lung cancer cell lines after SRSF2 over-expressed. This study shows that dysregulation of SRSF2 -related splicing might be an important feature of immune hyper-progress occurrence.

Metabolomics Methodology

2.1 Dataset Description

The dataset we used in our study was produced under the study ID: ST000392 by Oliver Fiehn. It was produced by the time-of-flight mass spectrometry GC-TOF-MS technique. All the samples collected were of two types, Plasma and Serum. Both samples contained 82 subjects, for which data of 158 metabolites were given in the dataset. Among the subjects, the number of cancer patients and control patients was 41 each, and 20 of the subjects were male, and the remaining 60 subjects were female. Each of these are summarized in Tables 1 and 2.

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

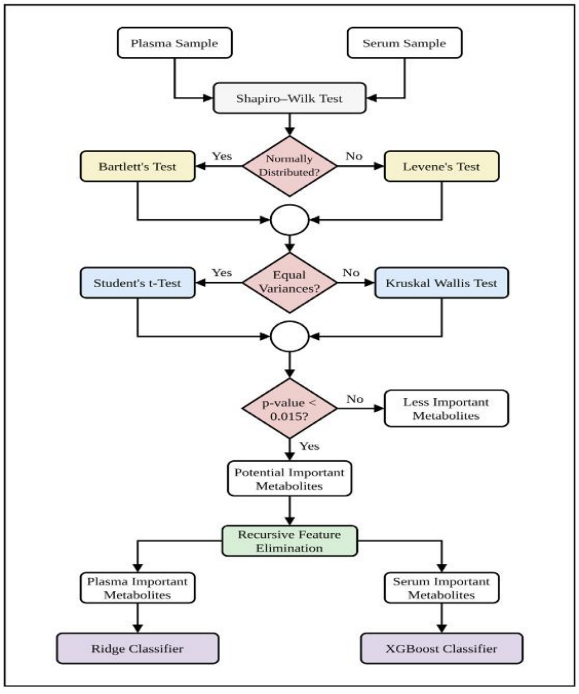
Sample type	Control	Disease
Plasma	41	41
Serum	41	41

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.

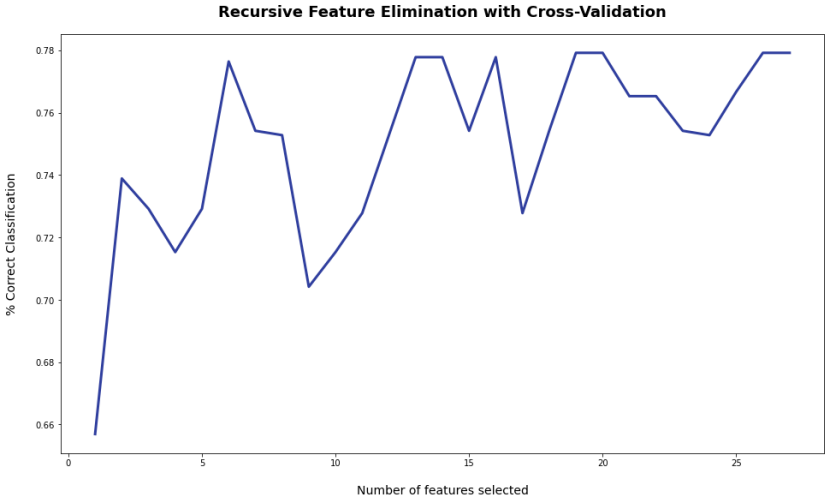


2.3 Biomarker Selection

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

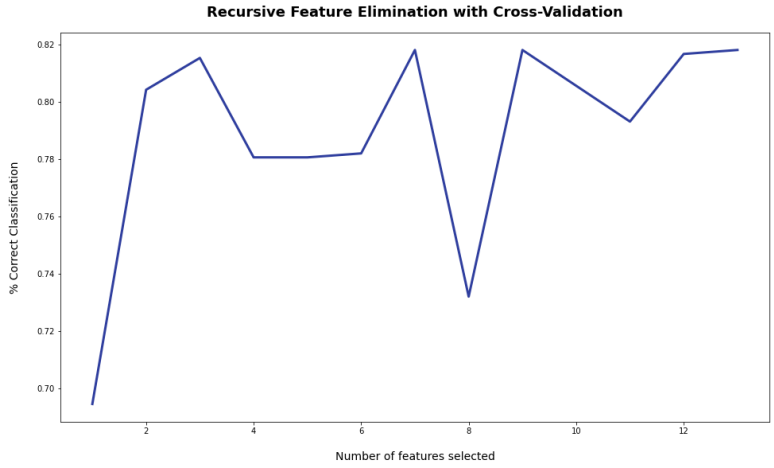


Optimal number of features after using Recursive Feature Elimination: 19

'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.

Transcriptomics Methodology

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: RNA was isolated from 1-2x106 cells stored in TRIzol reagent following manufacturer's instruction. RNA libraries were prepared for sequencing using standard Illumina protocols

This Data compares the SRSF2 Gene Over expression in lung cancer patients in two different Cell lines 1. A549 Cell Line, 2. H1650 Cell Line. This Data contains 12 SRA samples 6 from one cell line and other 6 from another cell line, in each cell line they have taken 3 control and 3 lung cancer patients and sequenced their transcriptome, and verified for the SRSF2 gene Over expression.

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

During sequencing, errors are introduced, such as incorrect nucleotides being called. These are due to the technical limitations of each sequencing platform. Sequencing errors might bias the analysis and can lead to a misinterpretation of the data. Adapters may also be present if the reads are longer than the fragments sequenced and trimming these may improve the number of reads mapped. We used FastQC to create a report of sequence quality, MultiQC (Ewels *et al.* 2016) To aggregate generated reports and Cutadapt (Marcel 2011) to improve the quality of sequences via trimming and filtering.

3.2.3 Mapping

To make sense of the reads, we need to first figure out where the sequences originated from in the genome, so we can then determine to which genes they belong. When a reference genome for the organism is available, this process is known as aligning or "mapping" the reads to the reference. This is equivalent to solving a jigsaw puzzle, but unfortunately, not all pieces are unique. In this study, the authors used *Homo Sapiens lung* cells. We should therefore map the quality-controlled sequences to the reference genome of *Hoo sapience Gch38.*

3.2.4 Counting the number of reads per annotated gene

To compare the expression of single genes between different conditions (*e.g.*, with or without PS depletion), an essential first step is to quantify the number of reads per gene, or more specifically the number of reads mapping to the exons of each gene. Two main tools are available for read counting: **HTSeq-count** (Anders *et al.* 2015) Or **featureCounts** (Liao *et al.* 2013). Additionally, STAR allows to count reads while mapping: its results are identical to those from **HTSeq-count**. While this output is sufficient for most analyses, **featureCounts** offers more customization on how to count reads. As we chose to use the featureCounts flavour of the tutorial, we now run **featureCounts** to count the number of reads per annotated gene.

(Note: The above process describes the original methodology for the RNA sequence Data Analysis but from the GEO Data set I have got the Raw gene count data so from that step I started analysing the RNA SEQ data using R Programming and libraries from BIOCONDUCTOR)

3.2.5 Analysis of the differential gene expression

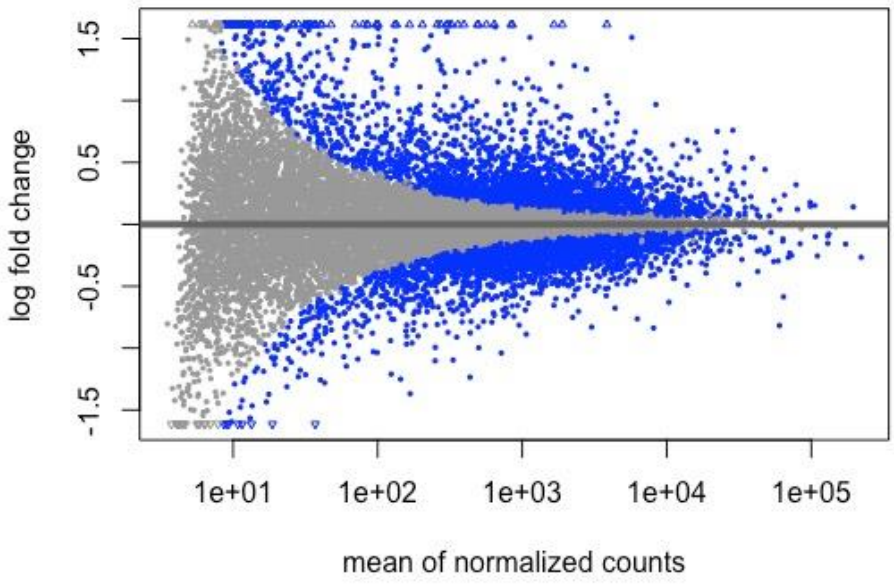
For the Differential Gene Expression, we used the RAW GENE COUNT dataset. First, import the count data and metadata directly from the web. Set up the DESeq data set, run the DESeq2 pipeline. Next, get results for the SRSF2 Over expressing versus control, and reorder them by p-value. Call summary on the results object to get a sense of how many genes are up or down-regulated

Res contains the genes id's, Base Mean, Log2foldchange, Pvalue and Padj Value.

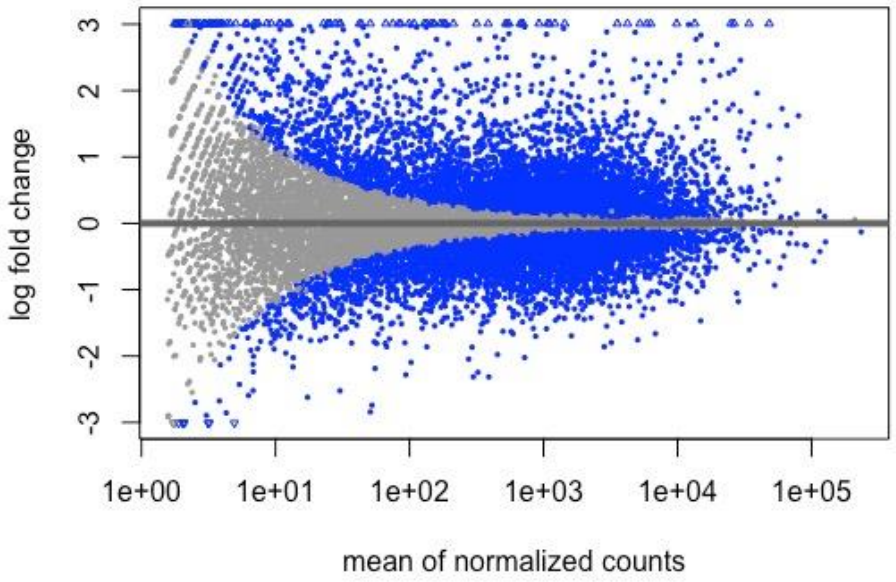
```
> summary(res)

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
```

From the Differential Gene Expression, we have plotted the MA plot for both the cells lines when compared with control.



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



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

After the DESeq Function to the dataset, we have got 15933 Gene's data from that we need the most potential genes for our analysis so we would like to extract the most differentially expressed genes with a fold change > 2 (or < 1/2). We will now select only the genes with a fold change (FC) > 2 or FC < 0.5. Note that the DESeq2 output file contains log2FC*log2FC*, rather than FC itself, so we filter for abs(log2FC)>1*abs(log2FC)>1* (Which implies FC > 2 or FC < 0.5).

From all these filters we have filtered 1354 important Genes out 16,599 Genes.

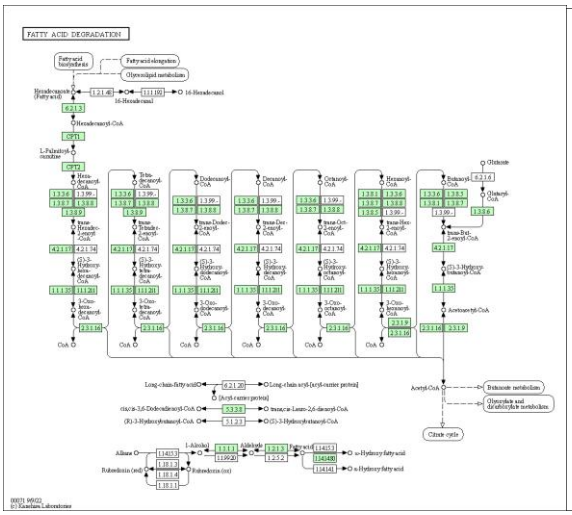
3.2.6 Functional Enrichment Analysis of DE Genes

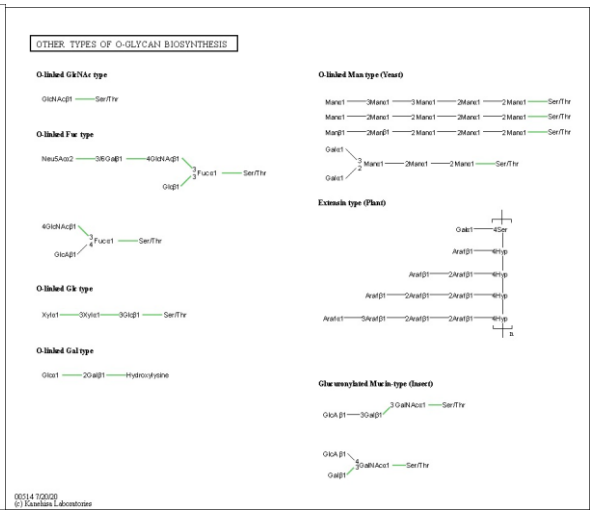
Since we mapped and counted against the Ensembl annotation, our results only have information about Ensembl gene IDs. But our pathway analysis downstream will use KEGG pathways, and genes in KEGG pathways are annotated with Entrez gene IDs. Here we're using the organism package ("org") for Homo sapiens ("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:

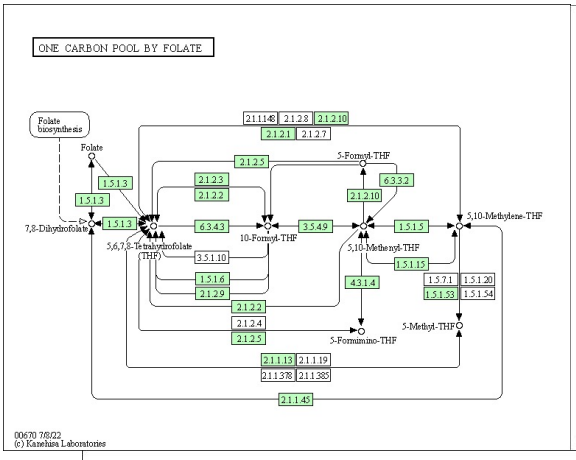
```
> res
log2 fold change (MLE): treatment untreated vs treated
Wald test p-value: treatment untreated vs treated
DataFrame with 15933 rows and 9 columns
                 baseMean log2FoldChange     lfcSE      stat      pvalue       padj      symbol      entrez                  name
                <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..
```

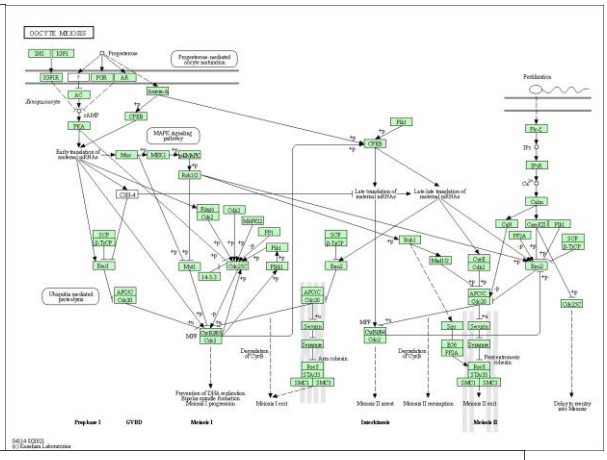
Kegg Pathway Analysis: The gageData package has pre-compiled databases mapping

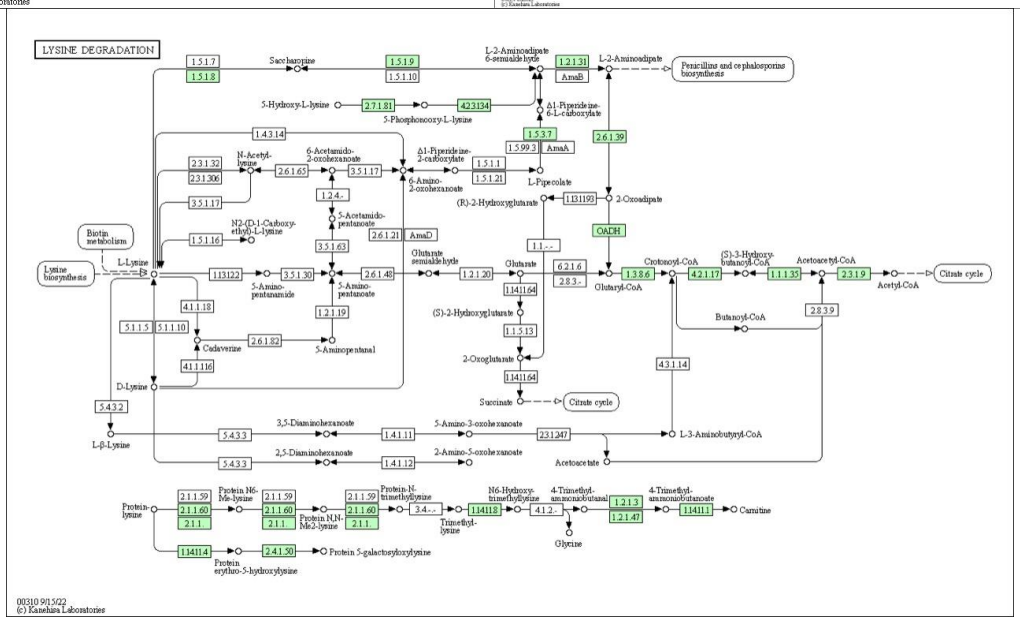
genes to KEGG pathways and GO terms for common organisms. kegg.sets.hs is a named list of 229 elements. Each element is a character vector of member gene Entrez IDs for a single KEGG pathway. we pick the most important upregulated and downregulated genes and plot their pathways. I have picked top 5 upregulated genes and plotted their pathways.











Conclusion:

Cancer continues to be one of the most common deadly diseases in the world. Every year, large numbers of people lose their lives to different subtypes of cancer. This research found that metabolomic biomarkers identified from differentially expressed metabolites in lung cancer patients can have a huge impact on the field of medicine as early and cost-effective measures of identifying cancer patients are crucial to save lives. In this work, we tried to integrate both Transcriptomics and Metabolomics, Genes that are obtained from the transcriptomic studies and the genes which are responsible for the metabolites formation that we have got from the metabolic studies, we then integrate both of them to the machine learning model for the diagnosis purpose, and also helps in giving the patient the right treatment and check the progress of the treatment.

Limitations:

For integrating two Omics we need the data of same kind from the same samples, but in our study, we have got the data from two different sources, Metabolomics data is from blood and plasma samples of 82 people whereas Transcriptomics Data is from the tissue samples of different people. So here the problem arose that one cannot really integrate these two samples for the analysis it would be not accurate if we integrate the tissue samples from transcriptomics to the blood and plasma samples from metabolomics.

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.