Bioinformatics Diploma

Course (CIT-658-Sp20)

Integrative Bioinformatics Analysis of MicroRNA and Gene Interactions For Revealing Potential Biomarkers of Non-Small Cell Lung Cancer

Authors:

1. Ghada Hamed Akl, ID: 1922029

2. Amira Ahmed Atef, ID: 1922031

3.Aya Mohamed Attia ID: 202002809

4.Ahmed Badr, ID:

Under supervision of

Dr. Mohamed Hamed

1.Abstract:

Non-small cell lung cancer (NSCLC), is one of the most frequent primary subtype of lung cancer, It consists of two histological subgroups: lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). NSCLC accounts for an estimated 85% of lung cancer cases around the world and it has low survival rate due to diagnosis often made during late stages of the cancer that lead to inefficacy of treatments methods. That aim of this study is to explore the differences between LUAD and LUSC and to identify novel diagnostic and prognostic biomarkers for early detection of each subtype. DESeq2 package was used to identify differentially expressed miRNAs and genes between normal and tumor samples obtained from The Cancer Genome Atlas (TCGA) in LUAD and LUSC and visualized using volcano plot and heatmap, A total of 554 DEGs and 29 DEMs were identified. Functional enrichment analysis of the resulted DEGs and DEMs was performed using FUNRICH software to investigate gene functions, revealing that DEGs were significantly enriched in different signaling pathways, Cell Growth and Maintenance. The Gene–miRNA Interaction Networks in LUAD and LUSC were predicted using GeneMANIA in the Cytoscape showing 7 most significantly common genes between the two subtypes of lung cancer. Whereas, Kaplan–Meier survival analyses was used to investigate the factors affecting survival rates, ANXA10 and CLEC3B were significantly related to survival in LUAD and LUSC respectively. Differential expression of DNA-methylation showing total of 74 upregulated in LUAD and 15 upregulated in LUSC. These findings may improve our understanding of the different molecular mechanisms between lung adenocarcinoma and squamous cell carcinoma and may be used in improving diagnosis strategies, treatment and improvement of the survival rate of NSCLC.

Keywords: Non-small cell lung cancer, TCGA, Adenocarcinoma, Squamous, miRNA-Seq, RNA-Seq, R packages.

2. Introduction

Lung cancer is considered as one of the widely spread malignancies in both men and women worldwide. It is the second most common cancer and the leading cause of morbidity and mortality, with approximately 2 million new cases in 2018 making up 18.4% of the total cancer deaths [1,2]. Non-small-cell lung cancer (NSCLC) is the major pathological subtype of lung cancer, The standard procedure for NSCLC diagnosis is through biopsy or imaging tests that mostly happens at advanced stages of cancer, and despite the progress of treatment methods, they are often less effective in advanced stages compared with early interventions which decrease the overall 5-year survival rate of the patients to only 20%[3,4]. NSCLC consists of two histologic subgroups that comprise the majority of lung cancers: lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC)[5]. LUAD and LUSC originate from different cells and have major differences in their molecular and biological characteristics, in addition to their therapeutic strategies[6].There are several aberrations reported associated with LUAD that are considered as a marker for the treatment of LUAD like mutations in epidermal growth factor receptor ((EGFR) and fusion or rearrangement of anaplastic lymphoma kinase (ALK)[7] but these abnormalities do not occur in LUSC, making the treatments available for the adenocarcinoma ineffective to the squamous carcinoma[8].

Therefore, it is of vital importance to investigate the differences of the two major subtypes of lung cancer for deeper understanding and identification of novel diagnostic and prognostic biomarkers for early detection and improvement of the survival rate of NSCLC.

The Cancer Genome Atlas (TCGA) database has numerous high-throughput sequencing data for various types of cancer including mRNA and microRNA(miRNA) data of lung cancer. Many cancer-related genes, specific biomarkers, and signaling pathways were predicted based on the analysis of these data [9]. The study of the expression profiles of these mRNA and miRNA of LUAD and LUSC and the development of miRNA–mRNA interaction networks can assist in exploring the molecular differences between the two subtypes.

In the present study, We aim to perform Integrative analysis to identify the differentially expressed miRNAs (DEMs) and genes (DEGs) in NSCLC. Additionally, we conducted Functional enrichment analyses and constructed mRNA–miRNA interaction networks. We also performed survival analysis and methylation analysis on candidate biomarker miRNAs.

3. Materials and Methods

3.1. Data Downloading

The raw count data used in our integrative analysis were downloaded from the TCGA Data Portal (https://portal.gdc.cancer.gov/), we chose two types of lung cancer, LUAD and LUSC, which had transcriptome data available for both cancer and normal tissue samples. The two classes of phenotypes we used were "primary tumor" and "solid tissue normal". The number of retrieved samples for each cancer type and subtype are listed in Table 1. The inclusion criteria was adjusted each time according to the performed analysis type. The downloaded data for miRNA-seq (miRNA Expression Quantification) was an open-access raw count table, the data was presented as a .txt file for each sample.
Alternatively, the RNA-seq (HTSeq-Counts) data was an open-access raw count table that was compressed in a zip file, which needed an extra step for unzipping each file in all folders. Along with the mentioned data files, a sample sheet was also retrieved, which contains information about each of the hundreds of samples that will be needed in the analysis later. Clinical data of each type of lung cancer was obtained from GDC Data Portal using R TCGAbiolinks package[10],to evaluate the associations between molecular markers and survival of patients with lung cancer.

Regarding the DNA-methylation analysis, two sets of data have also been retrieved from TCGA, one for each type of NSCLC. The parameters have been set to select the methylated data which are represented as a .txt file, each containing the "Beta value ". Owing to the heavy size of the methylation data, only 10 samples for each case have been selected. A higher number has been attempted but caused an overload on the machine, so we stuck to this sample range for determining the differentially expressed CG islands. The LUAD samples had 4 Primary Tumor samples and 6 Solid Tissue Normal, while the LUSC samples had an even number of 5 for each sample type.

	LUAD RNA-Seq	LUAD miRNA-Seq	LUSC RNA-Seq	LUSC miRNA-Seq
No. of Samples	594	567	551	523
Primary Tumor	533	519	502	478
Solid Tissue Normal	59	46	49	45
Recurrent Tumor	2	2	--	--

Table (1): Number of retrieved samples for each cancer type and subtype

3.2 Differential Expression Screening

The differential expressions of mRNA and miRNA were processed using R DESeq2 package[11] and selected according to the criterion of adjusted p.values less than 0.05 and |log2FC| > 2.0 as reported in previous studies[12,13].

3.3 Visualization of DEGs and DEMs

Data visualization has been performed using RStudio (R version 4.0.3). Implementing diverse kinds of plotting was maintained to gain maximum perspective of the resulted data. While the MA Plot showed the mean of normalized counts versus the log2FoldChange for all tested genes, the Volcano Plot showed the statistical significance (P.value) versus the magnitude of change (FoldChange). Both of which are scatter plots that represent, in our case, the differential expression of genes. Last, but not least, Heatmaps act as a histogram of all the values in the matrix (value versus frequency) and how they correspond to the specified heatmap color range, hence visualizing hierarchical clustering where data values are transformed to color scale.[14]

3.4 Target Gene Prediction and Functional Enrichment Analysis

Functional enrichment analysis was done by FunRich[14,15,16]software on DEGS and DEMs for both LUAD and LUSC to identify the most significant and expressed genes in terms of cellular component, molecular biology, biological process, biological pathways, protein domains, site of expression, transcriptional factors, clinical phenotype and finally COSMIC analysis (Catalogue Of Somatic Mutations In Cancer).
Additionally, one of the tools in FUNRICH analysis of DEMS was used to predict the genes that are possibly affected by miRNAs that were produced from the differential expression analysis.

3.5. Construction of The Gene–miRNA Interaction Network

To identify the common genes between the generated DEGS and Dems, several Databases were used, such as: TFmir, TFmir2, Mirtarbase, String and GEPIA tools. Moreover, through adding the Get-multimir function in R as well as the GeneMANIA in the Cytoscape app (Cytoscape: Network data integration, analysis and visualization v.3.7.1 download) [17], a network between the DEGS of Adenocarcinoma and squamous Carcinoma were constructed. Therefore we were able to identify the shared genes through merging this network with other networks for Lung cancer from NEDx.

3.6 Survival Analysis for differentially expressed genes

Survival data of LUAD and LUSC were retrieved from the TCGA portal using "TCGAbiolinks library" [18]. These data were used to build survival models with Kaplan–Meier survival analyses based on differences between survival curves and log-rank p values utilizing survfit function in the R Survival package.
The normalized expression matrix of differentially expressed genes was obtained by the DESeq2 package. Another online database for survival analysis was also used to build survival models[19].

3.7 Identification of Differentially Methylated Regions

A broadly investigated epigenetic mark in biology is DNA methylation which is closely associated with modifications on the genomic level. These changes have been linked to cell differentiation, in particular at promotors and enhancers [20]. Bisulfite conversion is an indicator of the occurred DNA methylation as it transforms the methylated information into DNA sequence information that can be read by next-generation sequencing platforms [21]. The resulted data of the sequencing is represented as methylation counts that are then analyzed to determine which regions have been differentially expressed. This is beneficial when comparing two biological conditions, in our case, Solid Tissue Normal and Primary Tumor patients. The selection criteria for identifying dmrs was based on the LogFoldChange (LFC > log2(1.5)) as well as the significance level (t.pval <0.05). A Volcano plot and heat-map of methylation-driven genes of LUSC and LUAD were constructed as a form of visualization of the differentially expressed CG islands2

4. Results

4.1 Identification of DEMs and DEGs in LUAD and LUSC

DEMs were filtered out by relevant parameters (padj<0.05 and |log2FC| > 2.0). After screening, 18 DEMs were identified (17 downregulated and 1 upregulated) in LUAD samples, while in LUSC samples 12 DMEs were identified, all of which were downregulated. (Table. 2)

For DEGs analysis in LUAD, the gene expression data of 594 LUAD samples we downloaded (59 controls vs 533 tumors) from TCGA-LUAD project. By setting a series of thresholds (padj<0.05 and |log2FC| > 2.0), 449 DEGs were identified, 17 were upregulated and 432 were downregulated (Supplementary Table S1). Likewise, we downloaded the RNA-seq data of 551 LUSC samples (49 controls vs. 502 tumors) and identified 147 DEGs, of which 16 were upregulated and 131 were downregulated in LUSC. (Supplementary Table S2). were downregulated in LUSC. (Supplementary Table S2).

	LUAD RNA-Seq	LUAD miRNA-Seq	LUSC RNA-Seq	LUSC miRNA-Seq
Original Number of Samples	594	567	551	523
No. of genes (dds Input)	34,125	1,881	34,125	1,881
Resulted DE	449	18	147	12
Up regulated	17	1	16	---
Down regulated	432	17	131	12

Table (2): List of the Differentially Expressed mRNAs and miRNAs

4.2 Identification of Differentially Expressed miRNAs And Genes

4.2.1 MA Plot

This type of scatter plot helps us visualize the reproducibility between the two types of samples we have ("Solid Tissue Normal" and "Primary Tumor").

The genes that are significantly differentially expressed (DE) are colored in blue to be easily identified. Genes with similar expression values in both normal and patient samples are clustered around M=0 value which are the genes expressed with no significant differences in between treatments (from normal to tumor). Points away from the M=0 line indicate genes with significant expression. The upregulated genes are scattered above the M=0 line ,while the downregulated genes are present below the M=0 line.

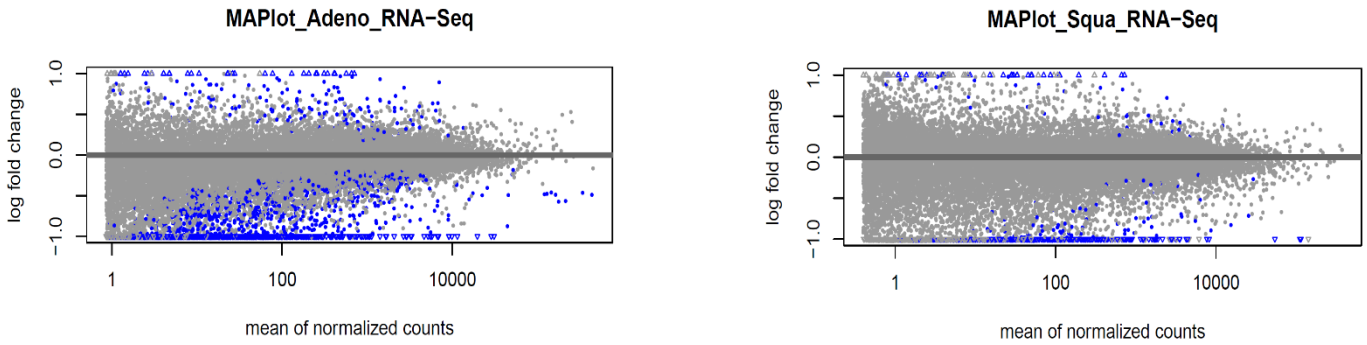


Figure (1): MA Plot of LUAD and LUSC RNA-Seq

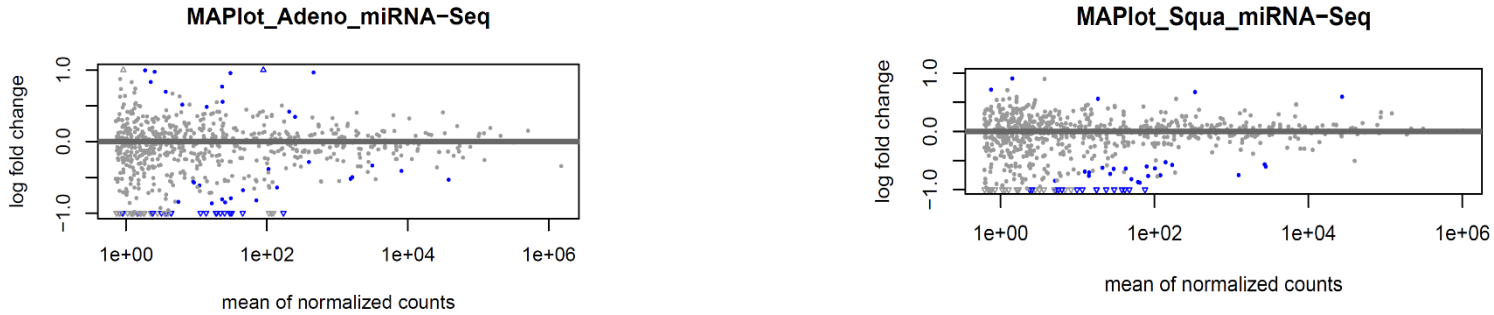


Figure (2): MA Plot of LUAD and LUSC miRNA-Seq

4.2.2 Volcano Plot

Volcano plot was efficient in Identifying the differences between the large datasets of LUAD and LUSC by highlighting the most statistically significant genes. The parameters used were log2FoldChange > 2 and padj < 0.05, which allows visualization through coloring the respective gene that falls in these criteria.

Typically, the most upregulated genes lie towards the right of the zero value of x-axis (logFoldChange), whilst the most downregulated genes are towards the left. Additionally, the most statistically significant genes are usually scattered towards the top as we can see in Figure (3, 4).

Depending on the log2FoldChange value (positive or negative) we can determine which of the differentially expressed results are upregulated and which are downregulated. When comparing the number of upregulated genes or miRNAs with their respective plot, it can be observed that the colored spots in red and blue on the right side of the volcano plot match the aforementioned number of upregulated genes. It can be concluded that the case would be similar with the downregulated genes and miRNAs. Therefore, the volcano Plot can be possibly a reliable visualization method for identifying the significant DE genes and miRNAs.

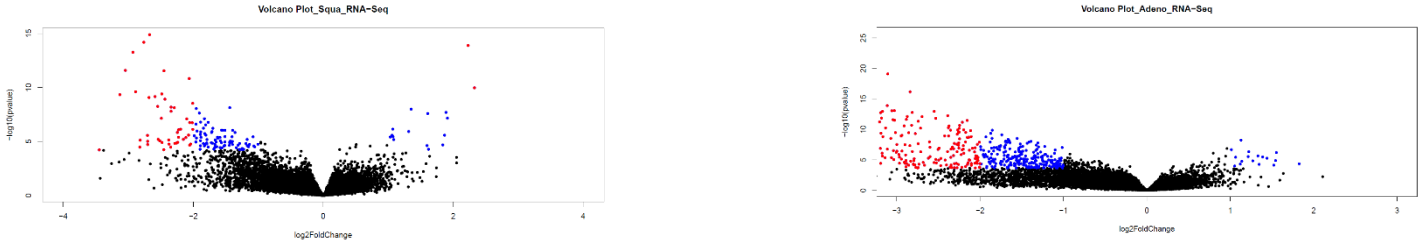


Figure (3): Volcano Plot of LUAD and LUSC RNA-Seq

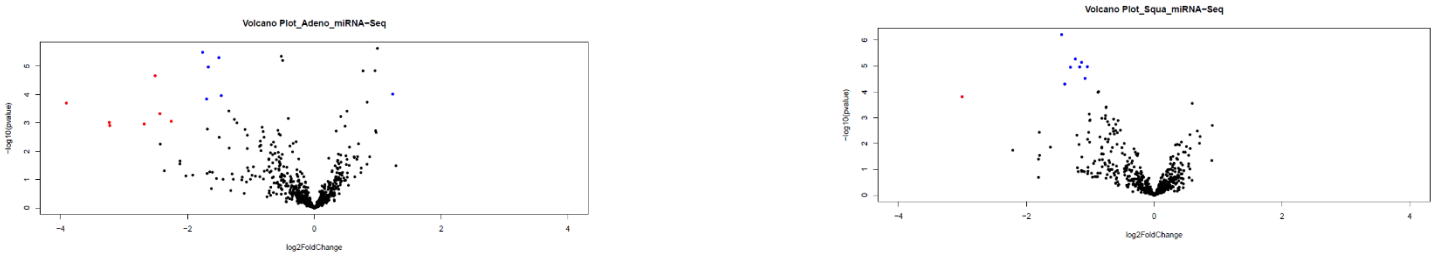


Figure (4): Volcano Plot of LUAD and LUSC miRNA-Seq

4.2.3 Heatmap

Construction of the four heatmaps has been done using the pheatmap package. Annotation colors have been established and a scaling of data was done to give a color scale for each of the map's values. According to the data range, miRNA had a scale of (-3:3), while RNA-Seq had a rage of (-6:6) which could have been due to the value difference of each of the two groups. In miRNA-Seq, the first 100 samples only were selected to demonstrate the relation between the value versus the frequency of this gene (Figure 5). Alternatively, 150 samples were chosen for the RNA-Seq analysis data, considering the observed higher number of genes presented in this section compared with the former, the miRNA-Seq (Figure 6). Moreover, you can determine from the bar at top which of these belong to Normal (Control) patients and which is of patients with Primary Tumor (Case).

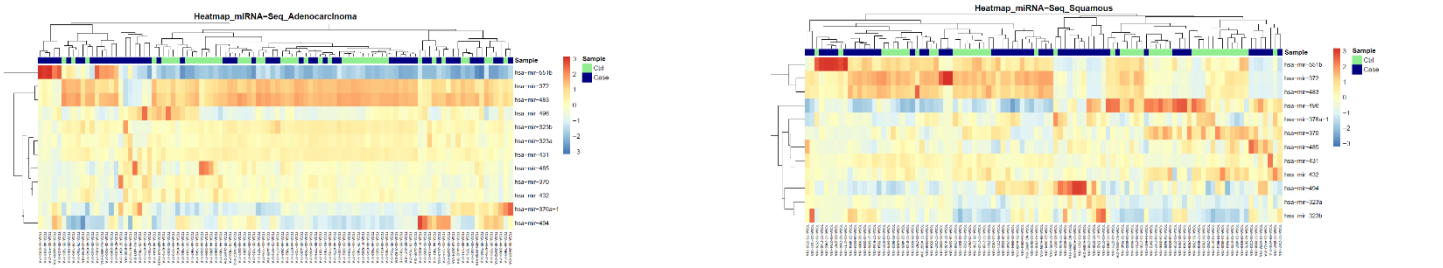


Figure (5): A Heatmap of LUAD and LUSC miRNA-Seq

Figure (6): A Heatmap of LUAD and LUSC RNA-Seq

4.3. Target Gene Prediction and Functional Enrichment Analysis

The parameters that were used to identify significant genes are No. of genes in the dataset, No. of genes in the background dataset, Percentage of genes, Fold enrichment, P-value (Hypergeometric test), Bonferroni method, BH method, Q-value (Storey-Tibshirani method), genes mapped (from input data set). We chose which GO terms that are highly expressed by our data based on these parameters.

4.3.1. Enrichment Analysis of DEGs for LUSC

Using FUNRICH software in enrichment analysis, the dataset identifies 106 genes and converts them to ID Entrez. The first part is to map these genes to the chromosomes which belong to, alternative names, gene symbol and description of its function. (supplementary table3). GO terms which are significant with mRNA are:

biological process: -

Number of genes in the dataset (which are available in Biological process database) :102 genes
The most enriched biological process is Cell Growth and Maintenance with P-value "0.03" and gene percentage 16.7% as shown in fig. (7.1)

protein domains: -

Number of genes in the dataset (which are available in Protein domain database) :80 genes
The most enriched protein domain is the responsible for signal peptide with P-value " < 0.001 " and gene percentage 53.8% as shown in fig.(7.2)

site of expression: -

Number of genes in the dataset (which are available in Site of expression database) :104 genes
This analysis shows the sites of expression of most significant genes are Gastric juice with P-value "0.001" and Gene percentage 9.6% and Saliva with P-value "0.012" and Gene percentage 16.3% as shown in fig. (7.3)

transcriptional factors: -

Number of genes in the dataset (which are available in Transcription factor database) :84 genes

The only transcriptional factor that was significant enough is FOXA1 with P-value ” .024” and Gene Percentage “16.7%” as shown in fig.(7.4)

-FOXA1 expression is increased in a rat model of ALI (acute lung injury), in which it is suggested to function in alveolar type II epithelial cell apoptosis. (20)

-FOXA1 is necessary for H2O2-induced apoptosis in A549 cells. (21)

-FOXA1 may participate in the onset or progression of lung diseases not limited to cancer.

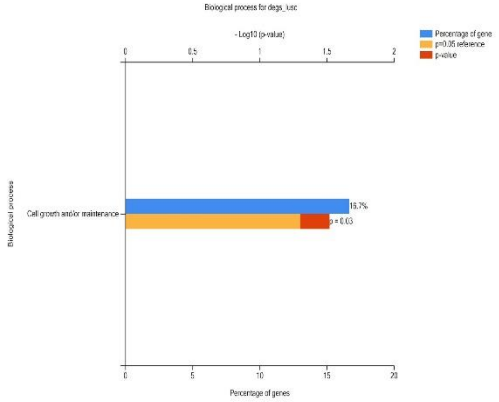


Figure (7.1): Biological process

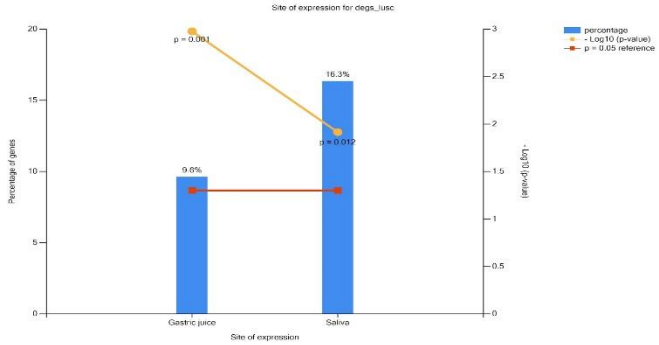


Figure (7.2): Protein Domains

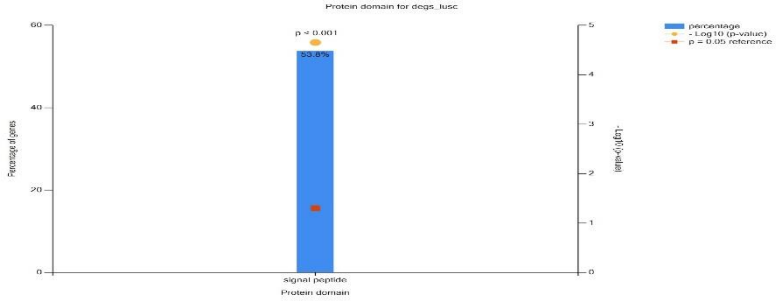


Figure (7.3): Sites of expression

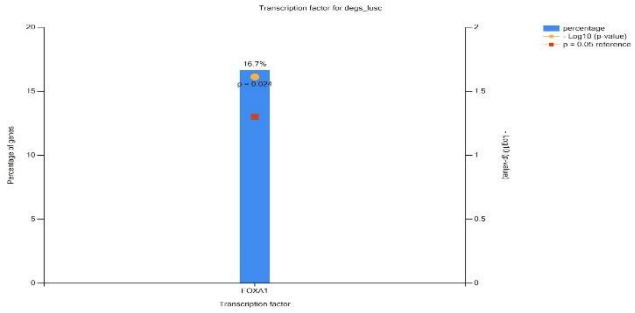


Figure (7.4): Transcription Factors

COSMIC analysis: -

COSMIC analysis used for exploring the impact of somatic mutations in human cancer and in our analysis we found that the skin genes is the most site with mutations with P-value “0.039” and Gene Percentage 76.9% as shown in fig. (7.5)

Interacting data: -

interacting genes are shown in red the genes from our data that have interactions with other genes, as shown in fig. (7.6)

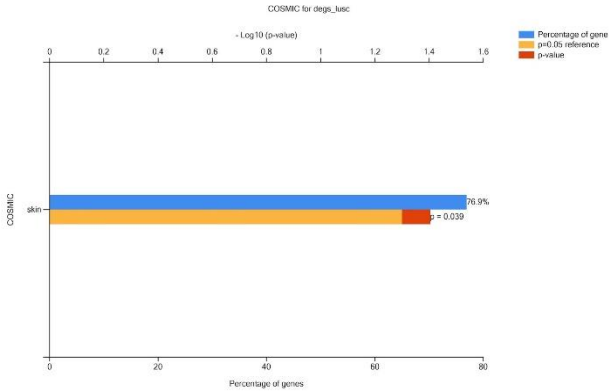


Figure (7.5): Sites of expression

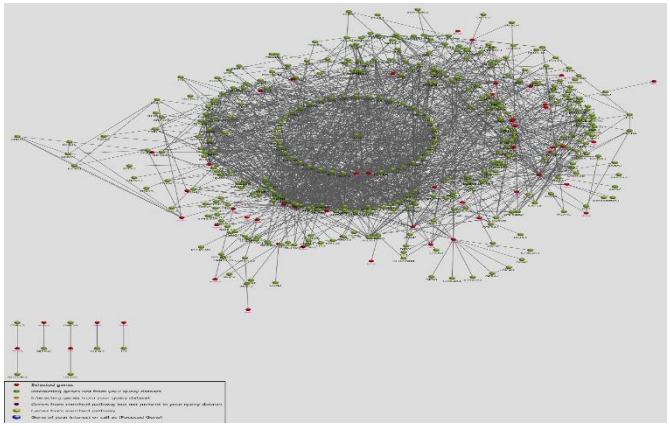


Figure (7.6): Transcription Factors

4.3.2. enrichment analysis of DEMs for LUSC

We used the same parameters as in enrichment analysis of DEGs of LUSC and the same software (FunRich)
Visualization of the GO terms that are most significant with miRNA with P-value<0.05 : -

Cellular component: -

This analysis shows the most significant Cellular Components that are related to miRNAs of LUSC extracted by R
Nucleus with very high significance "P-value < 0.001" and high Gene Percentage of 49.9%
Cytoplasm with P-value "0.069" which is not very significant but with high Gene Percentage of 43.5, fig.(8)

Molecular function: -

The significance of this GO term is majorly related to TFs
Transcriptional factor activity with low P-value "<0.001" and Gene percentage 8.6%
Transcriptional regulatory activity with low P-value "<0.001" and Gene percentage 7.1%
ubiquitin-specific protease activity with P-value "0.017" and Gene percentage 3.4%, as shown in fig.(8.1)
-Overexpression of USP22 was observed in 49.0% (99/202) of NSCLC tissues; higher USP22 immunostaining was found to be associated with enhanced angiogenesis and recurrence of NSCLC.
-USP22 plays critical roles in the malignancy and progression of NSCLC and provides rationales for targeting USP22, which induces broad anti-cancer activities, as a novel therapeutic strategy for NSCLC patients.

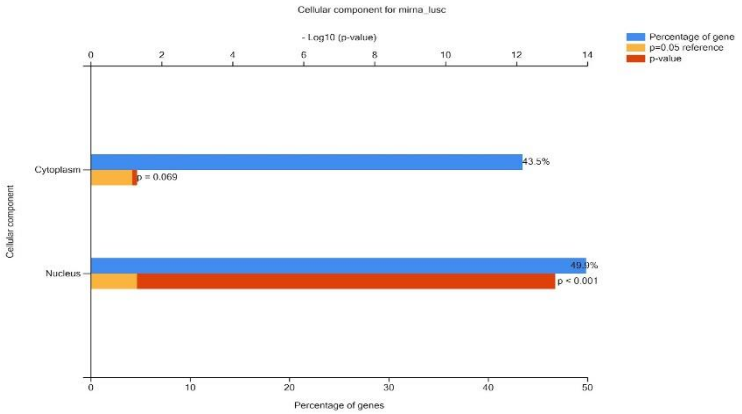


Figure (8): Cellular components

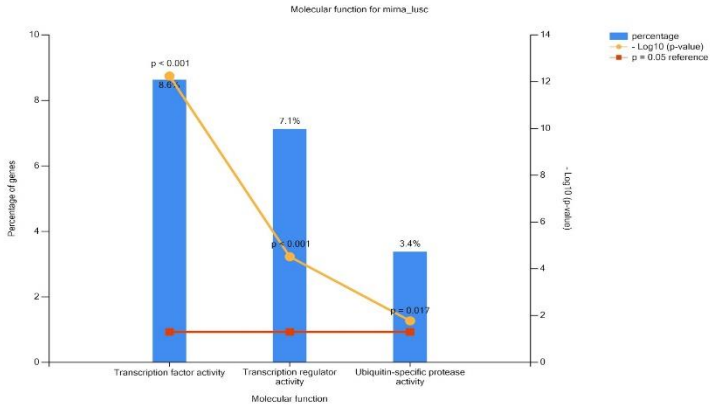


Figure (8.1): Molecular functions

Biological process:

The most significant term is related to **Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism** with high P-value "<0.001" and Gene percentage 21.9%, as shown in fig. (8.2)

biological pathways: -

This is the most enriched GO term and most significant for LUSC miRNA.

All of the pathways have very low P-value of "0.001" and high Gene Percentage around 35%, as shown in fig. (8.2) We can see the major enriched pathways are related to signal activity that play a key role in the pathogenesis of various forms of human cancer. PI3K is enriched two times in the graph and signaling events mediated by AKT with high significance. Many downstream regulators of PI3K pathway have become targets for cancer treatment with encouraging results up to date. Indeed, numerous targeted agents directly against the PI3K pathway have already reached the clinical stage either as single agents or in combination with conventional chemotherapy or other targeted therapies, presenting a much better toxicity profile compared to conventional chemotherapy.

protein domains: -

The most significant protein domain is HLH(helix-loop-helix) with low P-value "<0.001" and Gene percentage 2%, as shown in figure(8.3) and The inhibitor of differentiation/DNA-binding (ID) is a member of the helix–loop–helix (HLH) transcription factor family, and plays a role in tumorigenesis, invasiveness and angiogenesis.

genetic mutations of ID family members were identified in lung cancer. functional enrichment analysis results suggested that ID1/2/4 were significantly enriched in 'regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism' for biological process, 'transcription factor activity' for molecular function and 'HLH domain' for protein domain. expression of ID family members may affect the occurrence and prognosis of lung cancer, and may be related to cell metabolism and transcriptional regulation. The potential role of the BHLHB3 protein as a tumor suppressor for lung cancer.

All TFs that are shown in figure(4.13) is highly significant with P-value "<0.001"

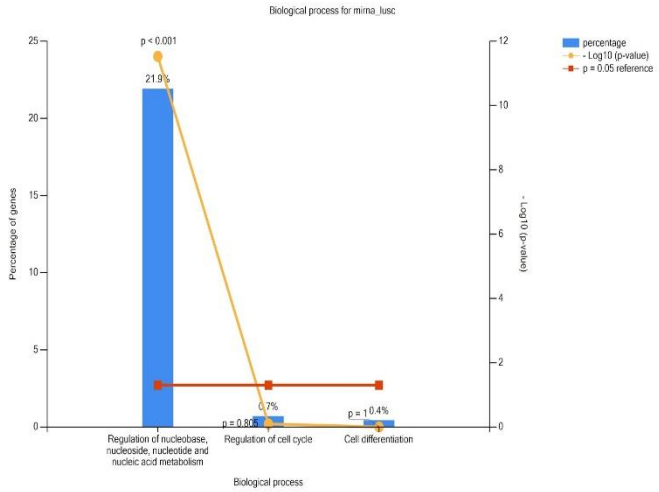


Figure (8.2): Cellular components

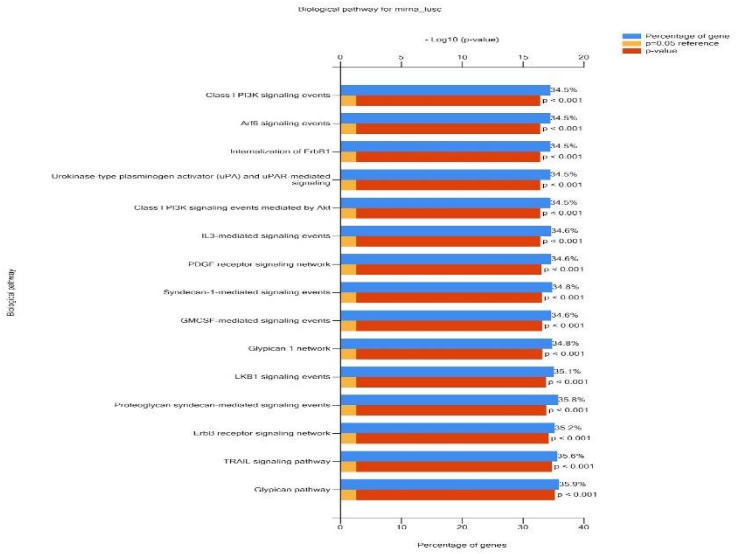


Figure (8.3): protein domains

site of expression: -

The group of DEMs is expressed in different sites is the body and majorly in the **lung** with significant P-value "<0.001" and high Gene Percentage of 61.4%, as shown in fig. (8.5)

transcriptional factors: -

All TFs that are shown in fig(8.4) is highly significant with P-value "<0.001"

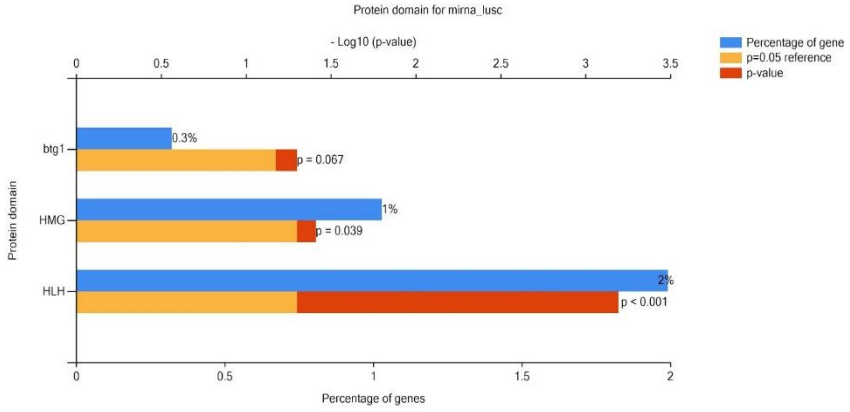


Figure (8.4): Site of expression

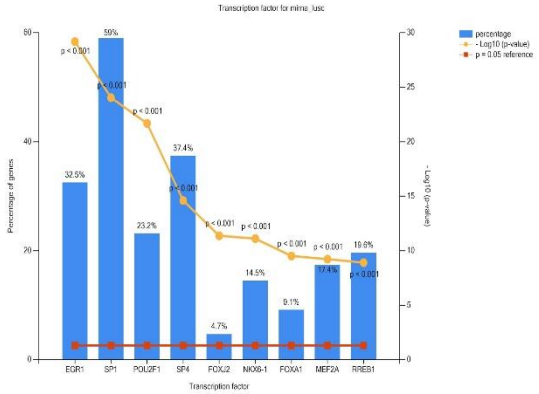


Figure (8.5): Transcription factor

COSMIC analysis:

The Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CGC) is an expert-curated description of the genes driving human cancer that is used as a standard in cancer genetics across basic research, medical reporting and pharmaceutical development and these genes are enriched in our analysis with "P-value <0.001", as shown in fig. (8.6).The Endometrium scores a high Gene percentage of 71.9 in COSMIC analysis and this strengthens the potential of a relation between the genes that may play a role in lung cancer and uterine cancer. Uterine metastasis from lung adenocarcinoma is uncommon and difficult to differentiate from primary uterine cancer. The possibility of lung cancer metastasis should be considered in patients who have adenocarcinoma on biopsy of uterine lesions.

Prediction of Targeted genes by miRNA:

A simple venn diagram between the predicted genes from miRNA data and differentially expressed genes from RNAseq data ((supplementary table. 5). as shown in fig. (8.7), shows that there are five common genes (CLIC5, KIRREL3, NMNAT2, NOL4, TMEM100) for further investigation and enrichment analysis.

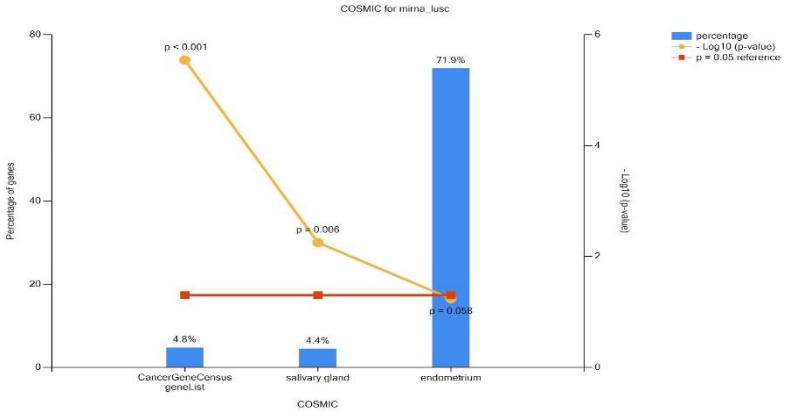


Figure (8.6): Cosmic analysis

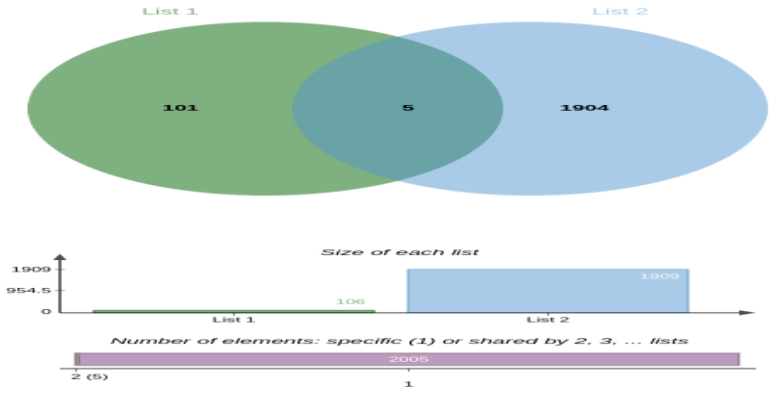


Figure (8.7): prediction of targeted genes by miRNA

4.3.3 Enrichment analysis of DEMs for LUAD

Visualization of the GO terms that are most significant with miRNA with P-value<0.05 : -

Cellular Component: -

This analysis shows the most significant Cellular Components that are related to miRNAs of LUAC extracted by R Nucleus with very high significance P-value " < 0.001" and high Gene Percentage of 51.3% as shown in fig.(9) and this is the case with the enrichment analysis of miRNas of LUSC also.

molecular function: -

The significance of this GO term is majorly related to TFs, Transcriptional factor activity with low P-value "<0.001" and Gene percentage 9.8% , Transcriptional regulatory activity with low P-value "0.081" and Gene percentage 6.9% fig. (9.1), Transcriptional activity is also highly enriched in LUSC.

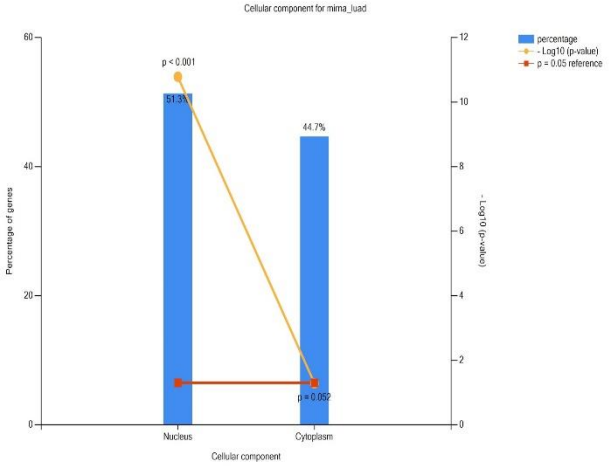


Figure (9)

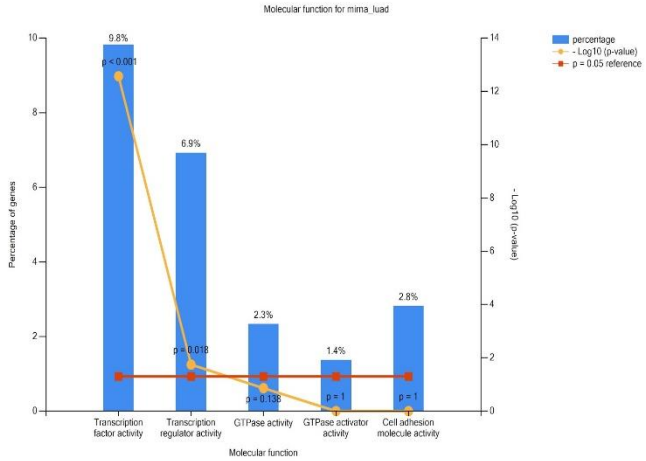


Figure (9.1)

Biological process:

The most significant term is related to **Regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism** with high P-value "<0.001" and Gene percentage 22.8%, figure(9.2)
The significance of this GO term is correlated with the DEMs of LUSC too. functional enrichment analysis results suggested that ID1/2/4 were significantly enriched in 'regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism' for biological process, 'transcription factor activity' for molecular function and 'HLH domain' for protein domain.

biological pathways: -

This is the most enriched GO term and most significant for LUSC miRNA.
All of the pathways have very low P-value of "0.001" and high Gene Percentage 36% , as shown in figure(9.3)
As in enriched biological pathways in LUSC data also in LUAD pathways, the signaling pathways is the most significant and plays a role in tumori-genesis. The PI3K/AKT/mTOR signaling cascade plays key roles in the tumori-genesis of lung cancer and also contributes to EGFR TKI resistance. Targeted agents against this pathway are currently being investigated in early clinical trials for lung cancer.

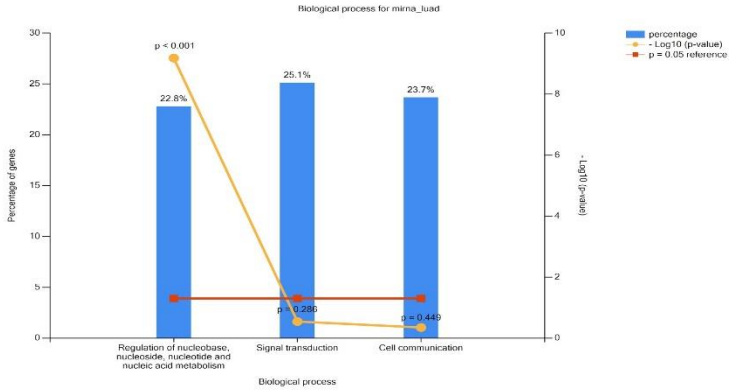


Figure (9.2)

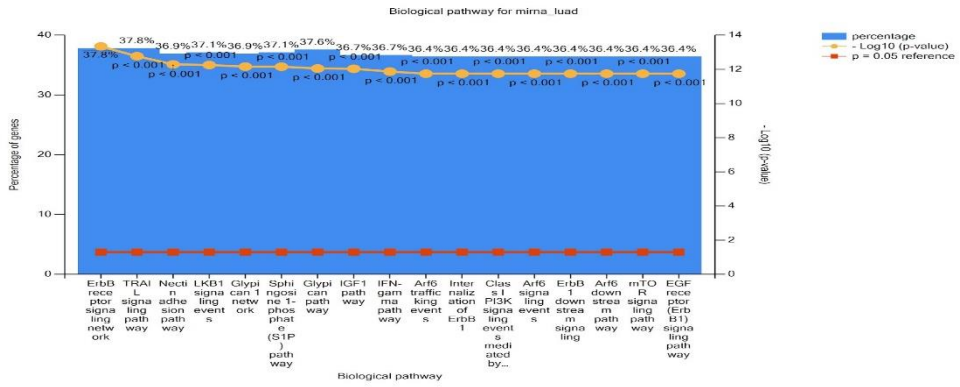


Figure (9.3)

protein domains: -

ZnF_C2H2 is significant with P-value"0.008" and Gene Percentage 8.5%, as shown in figure(9.4) .Zinc finger proteins are the largest transcription factor family in human genome, and play a vital role in cancer biology, So drugs targeting specific C2H2 ZNF protein expression or activity can be developed for therapeutic strategy against tumors in a specific stage of cancer progression.

site of expression: -

The group of DEMs is expressed in different sites is the body and majorly in the **lung** with significant P-value "<0.001" and high Gene Percentage of 62.4%, as shown in figure(9.5)

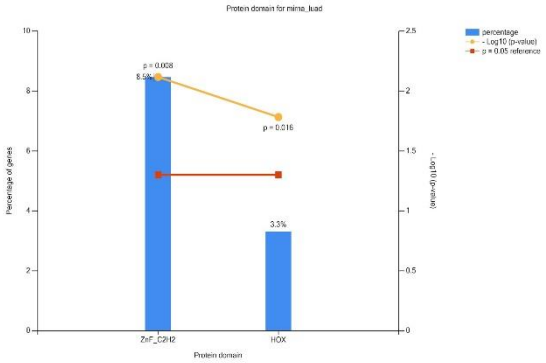


Figure (9.4)

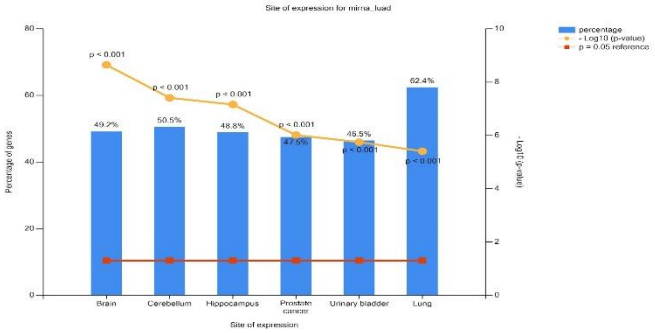
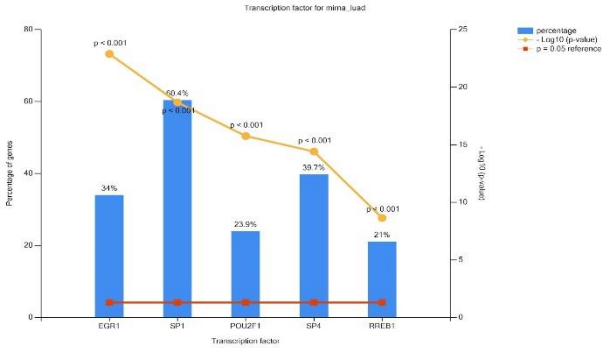


Figure (9.5)

transcriptional factors: -

All TFs that are shown in figure(9.6) is highly significant with P-value "<0.001"



COSMIC analysis: -

The Catalogue of Somatic Mutations in Cancer (COSMIC) **Cancer Gene Census** (CGC) is an expert-curated description of the genes driving human cancer that is used as a standard in cancer genetics across basic research, medical reporting and pharmaceutical development and these genes are enriched in our analysis with P-value" <0.001", as shown in figure(9.7)

Prediction of Targeted genes by miRNA: -

Comparing the predicted genes of DEMs from both LUSC and LUAD(Supplementary table 6) by applying a simple venn diagram, shows that there are 564 genes that are common.

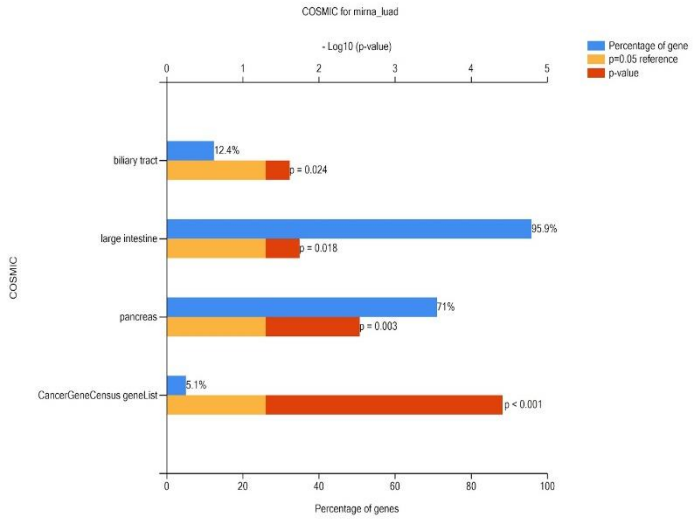


Figure (9.7)

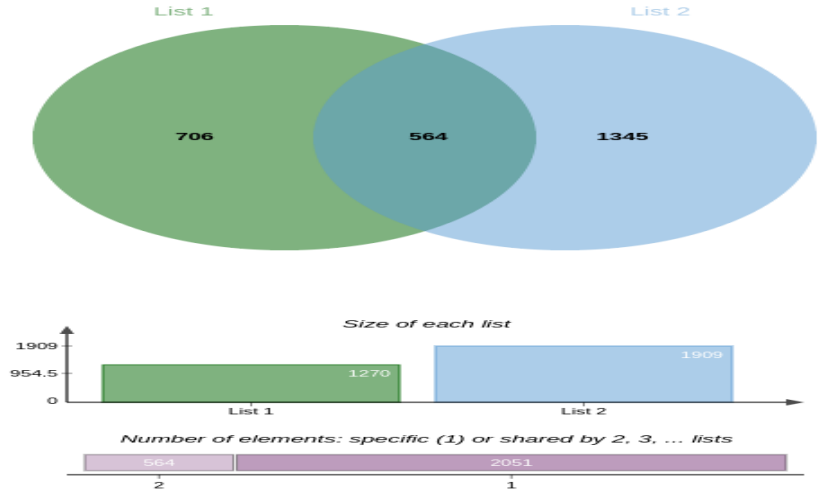


Figure (9.8)

4.4. Integrative Analysis of miRNA And mRNA Expression

4.4.1. Identification of common genes between DEGs and genes controlled by DEMs:

After using many databases: TFmir, TFmir2, mirTarbase and Tarbase using get-multimir on R and they all give null or no interactions between Dems and Degs. To analyse these results, we tried to add the miRna names individually on mirtarbase and show the target genes for this miRNA and it was found that the resulting genes were not found in our DEGS list which confirmed that there is no interaction between DEMs and DEgs.

4.4.2. Identification of Common Genes or Most Significant Genes Using GEPIA2 Tools.

GePia tools compared the list of DEGS to the dataset in TCGA for each carcinoma (AdenoCarcinoma and SquamousCarcinoma) and showed the top 10 significant genes for tumor and normal cells using z score.

The most significant gene in tumor cells for squamous carcinoma was “AJAPI” which target the hsa-miR-494-3p which is one of the miRNAs targeted by this gene on mirtarbase database. As this gene is an up-regulated gene so it is targeting a down-regulated miRNA fig. (10),table (3) Besides, the most significant genes in tumor cells are CPLX2, ERVH48-1 and OLFM4 as shown in fig. (11)

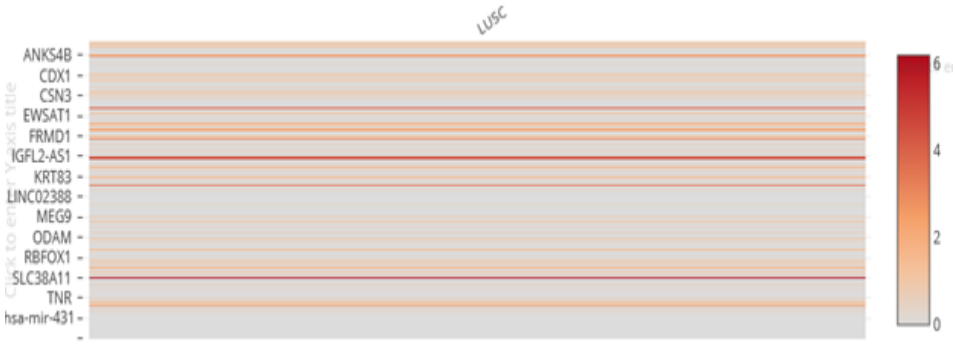


Figure (10)

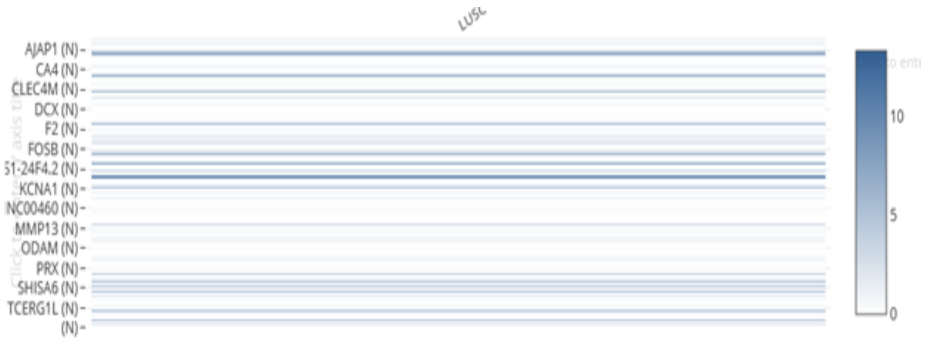


Figure (11)

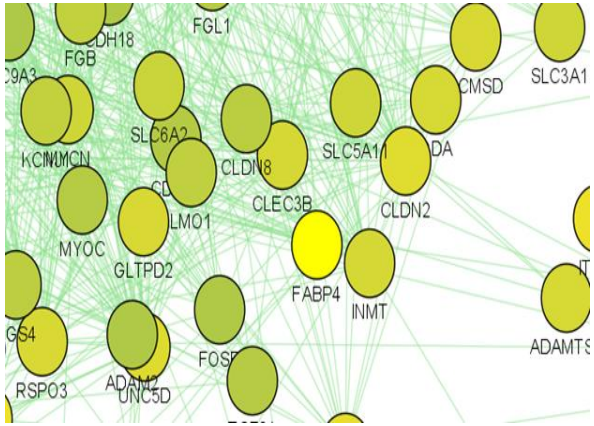
Z score (N)	Z score `(T)	Up regulated	down regulated
0.3	0.1	AJAP1	hsa-miR-494-3p
5.8	0.4	CA4	
1	0.1	CLEC4M	
0	0	DCX	
0	0	F2	
7.8	3.1	FOSB	
0	0.1	GS1-24F4.2	
0	0	KCNA1	
0.3	0.5	LINC00460	
0.3	2.2	MMP13	
2.2	0.3	ODAM	
4.9	1.2	PRX	
0.2	0.1	SHISA6	
0	0	TCERG1L	

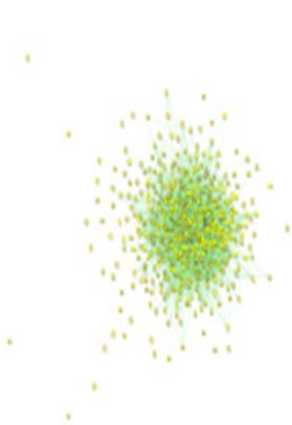
Z score (N)	Z score (T)	Up regulated	down regulated
0	0	APOC3	
6	6.3	CBR1	
0.2	0.1	CPLX2	
0.1	0.1	ERVH48-1	
1.5	0.4		GFRA1
0	0	ITIH6	
0	0	LINC00648	
0	0	MTND4P24	
0	0.1	OLFM4	
0.5	1.3	PROC	
0	0	RNU5A-1	
0	0	SNORA7B	
0	0	TMPRSS15	

.4.3. Constructing A Network Between DEGs of Both Types of Non-Small Cell Lung Cancer Using GeneMANIA and Cytoscape.

As shown in fig. (12) the genemmania generated a network between target genes in adenocarcinoma and squamous carcinoma which was visualized on cytoscape. The most significant genes are in yellow color. It also showed the log score and annotation type as shown in (Table 3).

AKR1C4	-0.525258
SOX11	-0.358255
KRT83	-0.546889
IL20	-0.264193
LBP	-0.502534
B4GALNT2	-0.179903
CYP2A6	-0.449918
CTCFL	-0.230523
SLC3A1	-0.315997
ASCL1	-0.420392
TF	-0.395696
KNG1	-0.483594
CLEC3B	-0.273351





4.5. Survival Analysis for DEGs in LUAD and LUSC

Survival analysis was first applied on the clinical data to determine which factors influence the survival in lung cancer patients by conduction of Kaplan-Meier plot through the survfit function. In LUAD (figure 13(A)) females seem to have a poor survival probability shown in two very similar trends with males till approximately the 3000-day mark, however, gender alone does not significantly affect the prognosis of survival as The p-value is non-significant($p > 0.01$) and that is validated when we studied the patients at risk. Additionally, the KP plot shows that most of the patients die or are censored before the 2000-day mark. Based on tumor stages the p-value is small (log-rank $p < 0.001$) proving that the tumor stages differ between survival times (Figure 13(B))

To investigate the clinical impact of the significantly expressed genes, we performed Kaplan–Meier survival analysis on six genes utilizing an online database “GEPIA2" on six genes. Three of those genes: ANXA10, SLC13A2, and PADI3, are correlated with worse overall survival in LUAD patients and considered risk factors (all $P < 0.05$; HRs, 0.76–1.5). In contrast, the other three genes: REG4, RN7SK, and SLC5A8 (P=0.2, 0.84 and 0.75 respectively), may improve the survival of patients. (Figure 14)

In LUSC, based on the survival probabilities of the relative clinical information, tumor stages are diverse in their survival rate with log-rank P value = 0.052. Even though the gender does not affect the survival significantly ($p > 0.01$), it is shown that males have a lower survival probability than females (Figure 15) Three DEGS are shown to have an impact on the poor prognosis of patients with lung cancer, which are CLEC3B, MCEMP1, and CLDN18 (all $p < 0.001$). The remaining three genes did not indicate worse survival GPD1, CPS1, and PRX (p ranges from 0.11 to 0.21).(Figure 16)

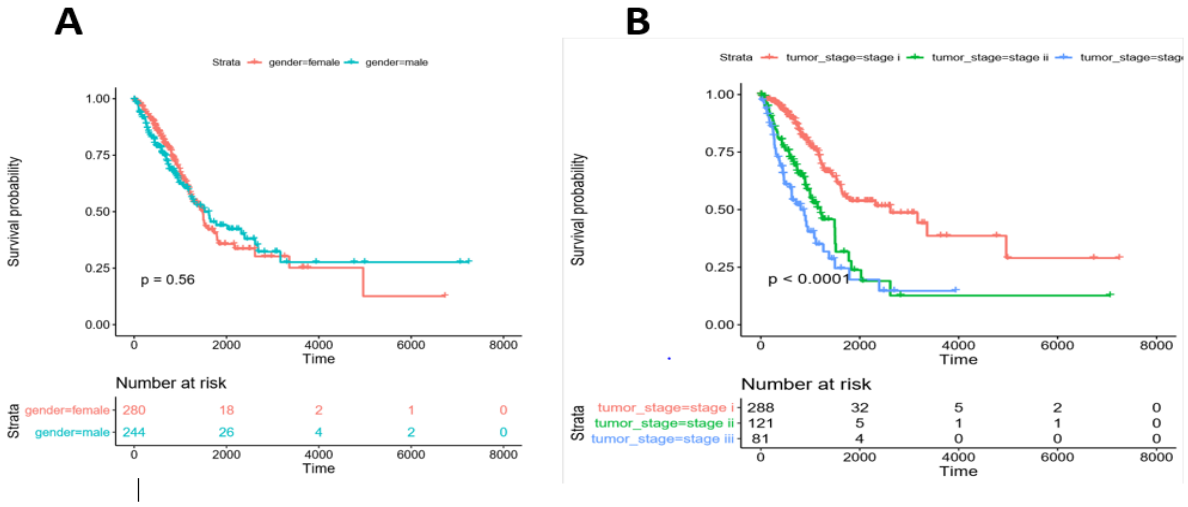


Figure 13.

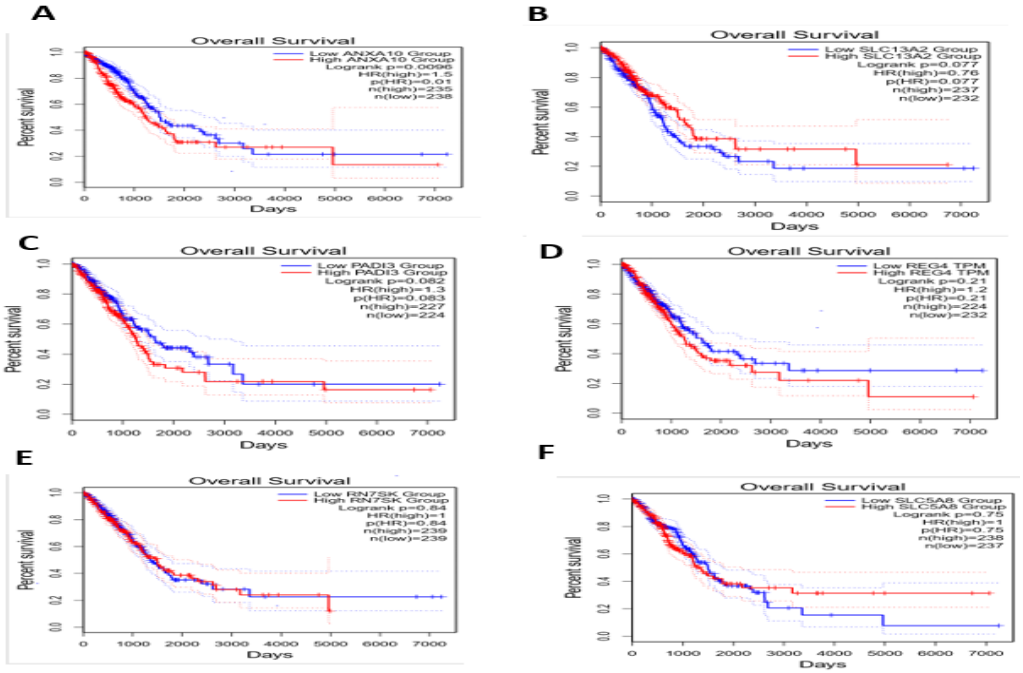


Figure 14

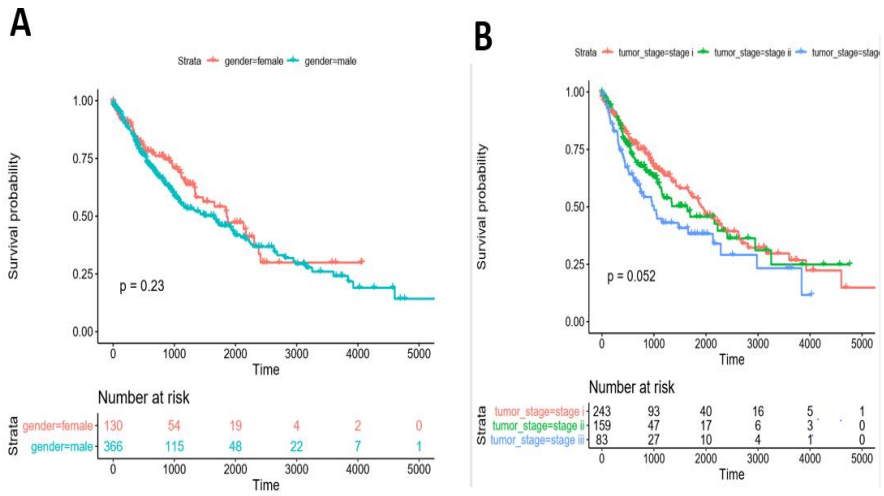


Figure 15

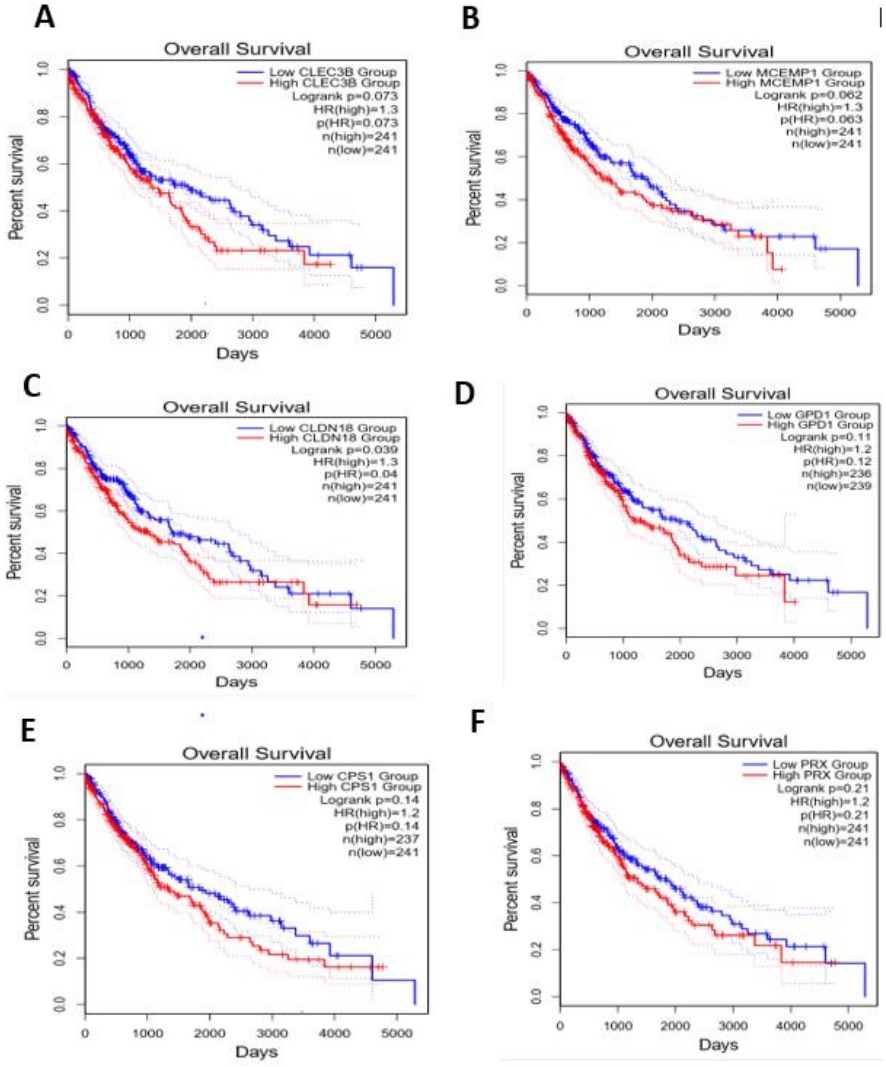


Figure 16

.6. DNA Methylation Analysis of LUAD and LUSC

Differential expression of DNA-methylation data produced from sequencing has been set to the aforementioned parameters, which caused a significant drop in the CG islands number from 394,577 to 89. After that, generating a list expressing the regulation change of these islands showed that 11 has been upregulated in case of LUAD and 74 in LUSC. On the other hand, none of the DE CG islands has been downregulated in LUAD, while 15 have been observed in LUSC. Table 5

Volcano plot showing the differentially methylated genes in each subtype Figure(17) X-axis represents log2 fold change and Y-axis represents –log10 (adjusted *p*-value). In LUSC the differentially methylated are represented in green dots with the hypermethylated on the right side of the plot and the hypomethylated on the left side, while in LUAD most of the methylated genes were hypermethylated.

Heat map of methylation-driven genes was also constructed for LUSC and LUAD Figure (18)

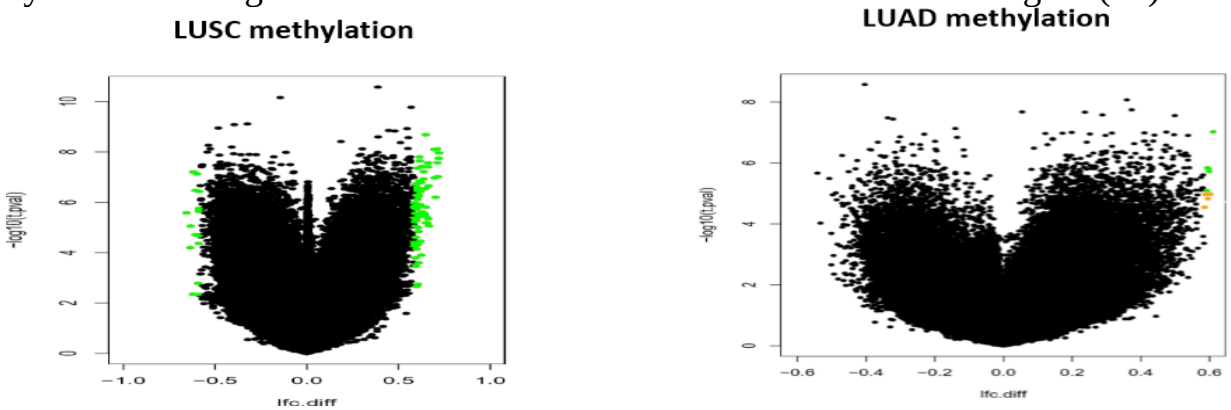


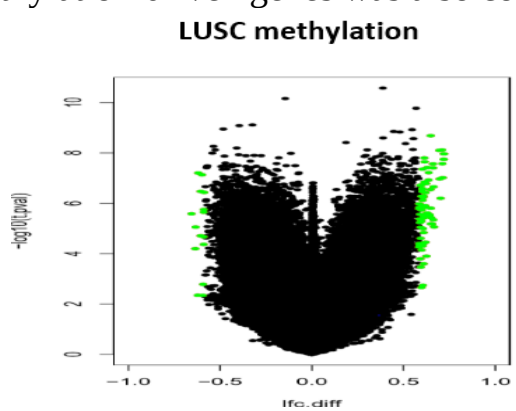
Figure 17

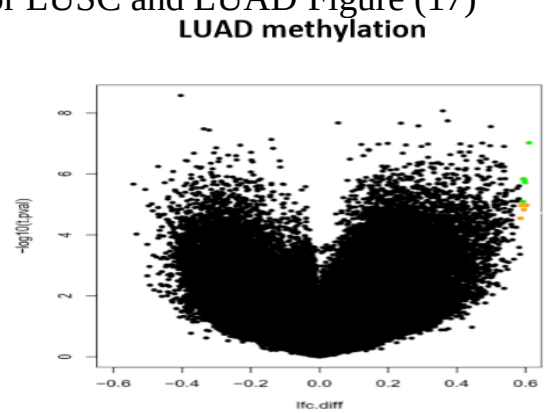
6. DNA Methylation Analysis of LUAD and LUSC

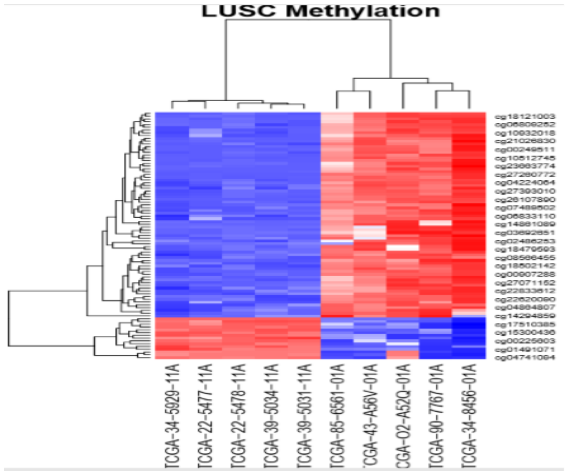
Differential expression of DNA-methylation data produced from sequencing has been set to the aforementioned parameters, which caused a significant drop in the CG islands number from 394,577 to 89. After that, generating a list expressing the regulation change of these islands showed that 11 has been upregulated in case of LUAD and 74 in LUSC. On the other hand, none of the DE CG islands has been downregulated in LUAD, while 15 have been observed in LUSC.(Table 5,6)

Volcano plot showing the differentially methylated genes in each subtype Figure (16)X-axis represents log2 fold change and Y-axis represents –log10 (adjusted *p*-value). In LUSC the differentially methylated are represented in green dots with the hypermethylated on the right side of the plot and the hypomethylated on the left side, while in LUAD most of the methylated genes were hypermethylated.

Heat map of methylation-driven genes was also constructed for LUSC and LUAD Figure (17)







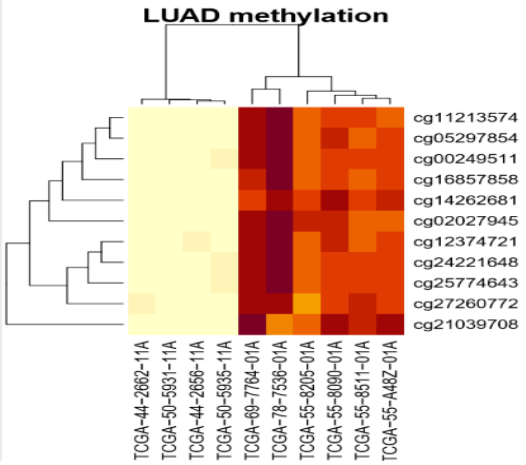


Figure (17) Heat map of Methylated genes

	LUAD	LUSC
No of Samples	10	10
Primary Tumor	6	5
Solid Tissue Normal	4	5
No. of CG islands (After filtering)	394,291	394,577
Resulted DE	11	89
Up-regulated	11	74
Down-regulated	---	15

Table (5): List of the Differentially Expressed CG Islands

	LUAD	LUSC
No of Samples	10	10
Primary Tumor	6	5
Solid Tissue Normal	4	5
No. of CG islands (After filtering)	394,291	394,577
Resulted DE	11	89
Up-regulated	11	74

Down-regulated	---	15

Table (6): List of the Differentially Expressed CG Islands

5. Conclusion

Non-small cell lung cancer early diagnosis and corresponding intervention is considered as the most effective approach for increasing the survival time and decreasing the mortality caused from NSCLC. There is an unrelenting necessity for biomarkers of this type of cancer to be identified, which will promote early detection and hence higher treatment rate. Integrated analysis of bioinformatics methods is the key to fulfill such purpose. In our study, we implemented a variety of techniques to determine Differentially Expressed products of various Next Generation Sequencing (NGS) analysis data. A collection of RNA-Seq (mRNA genes), miRNA-Seq (miRNA), Survival ( ), and DNA-Methylation (CG Islands). The main used analysis platform in our study was RStudio, and several others were combined to give desired results.

Differential expression of RNA-Seq data generated 449 in LUAD (17 upregulated and 432 downregulated), while LUSC had 147 genes (16 upregulated and 131 downregulated). Alternatively, miRNA-Seq resulted in 18 differentiated LUAD miRNAs, one of which was upregulated (hsa-mir-34b) and the rest were downregulated.

Subsequent enrichment analysis of the resulted DE genes and miRNAs on DAVID was performed which allowed close investigation of gene functions, cellular components, biological processes, molecular function, protein domains, site of expression, biological pathways, transcription factors and interacting genes. Using FUNRICH software in analysis of DEGS revealed that they were significantly enriched in different signaling pathways in LUSC. Additionally, COSMIC analysis was used for measuring the impact of somatic mutations in human cancer. Same parameters were applied for the DEGs and DEMs of both lung cancer types, and more detailed view of the results are shown in the supplementary data files.

Prediction of Gene–miRNA Interaction Networks in LUAD and LUSC in the Cytoscape revealed the seven most significantly common genes between the two subtypes of lung cancer. Whereas, Kaplan–Meier survival analysis was used to investigate the factors affecting survival rates. ANXA10 and CLEC3B were significantly related to survival in LUAD and LUSC respectively. Differential expression of DNA-methylation showed a total of 74 upregulated in LUAD and 15 upregulated in LUSC.

All in all, there were some potential biomarkers generated throughout our study, however further analysis is needed to validate which of these genes, miRNAs and CG islands can be relied on for early prognosis. These findings may improve our understanding of the different molecular mechanisms between lung adenocarcinoma and squamous cell carcinoma and may be used in improving diagnosis strategies, treatment and improvement of the survival rate of NSCLC.

6. References

1. F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
2. World Health Organization. Cancer fact sheet, 2018 https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed January 15, 2021.
3. Siegel, R., DeSantis, C., Virgo, K., Stein, K., Mariotto, A., Smith, T., … Ward, E. (2012). Cancer treatment and survivorship statistics, 2012. *CA: A Cancer Journal for Clinicians*, 62, 220–241.
4. Goldstraw P, Ball D, Jett JR, Le Chevalier T, Lim E, Nicholson AG. et al. Non-small-cell lung cancer. *Lancet.* 2011;378(9804):1727–40.
5. Y. Zhang, H. Wang, J. Wang et al., "Global analysis of chromosome 1 genes among patients with lung adenocarcinoma, squamous carcinoma, large-cell carcinoma,
small-cell carcinoma, or non-cancer," *Cancer and Metastasis Reviews*, vol. 34, no. 2, pp. 249–264, 2015.

6. Zhan C, Yan L, Wang L, Sun Y, Wang X, Lin Z, Zhang Y, Shi Y, Jiang W and Wang Q: Identification of immunohistochemical markers for distinguishing lung adenocarcinoma from squamous cell carcinoma. J Thorac Dis. 7:1398–1405. 2015.
7. C.-W. Xu, X.-Y. Cai, Y. Shao et al., "A case of lung adenocarcinoma with a concurrent EGFR mutation and ALK rearrangement: a case report and literature review," *Molecular Medicine Reports*, vol. 12, no. 3, pp. 4370–4375, 2015.
8. Rekhtman N, Paik PK, Arcila ME, Tafe LJ, Oxnard GR, Moreira AL, Travis WD, Zakowski MF, Kris MG and Ladanyi M: Clarifying the spectrum of driver oncogene mutations in biomarker-verified squamous carcinoma of lung: Lack of EGFR/KRAS and presence of PIK3CA/AKT1 mutations. Clin Cancer Res. 18:1167–1176. 2012.
9. The TCGA Legacy, (2018). The TCGA Legacy. *Cell* 173, 281–282. doi: 10.1016/j.cell.2018.03.049
10. Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550.
11. Yang S, Sui J, Liang G. Diagnosis value of aberrantly expressed microRNA profiles in lung squamous cell carcinoma: a study based on the Cancer Genome Atlas. *PeerJ*. 2017; 5: e4101.
12. Wright GW and Simon RM: A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics. 19:2448–2455. 2003.

13. Barter, R. L., & Yu, B. (2018). Superheat: An R package for creating beautiful and extendable heatmaps for visualizing complex data. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, *27*(4), 910–922. https://doi.org/10.1080/10618600.2018.1473780

14. Fonseka, P., Pathan, M., Chitti, S.V., Kang, T. and Mathivanan, S. (2021) FunRich enables enrichment analysis of OMICs datasets. Journal of Molecular Biology. 166747.

15. Pathan, M., Keerthikumar, S., Chisanga, D., Alessandro, R., Ang, C.S., Askenase, P., Batagov, A.O., Benito-Martin, A., Camussi, G., Clayton, A., Collino, F., Di Vizio, D., Falcon-Perez, J.M., Fonseca, P., Fonseka, P., Fontana, S., Gho, Y.S., Hendrix, A., Nolte-'t Hoen, E., Iraci, N., Kastaniegaard, K., Kislinger, T., Kowal, J., Kurochkin, I.V., Leonardi, T., Liang, Y., Llorente, A., Lunavat, T.R., Maji, S., Monteleone, F., Overbye, A., Panaretakis, T., Patel, T., Peinado, H., Pluchino, S., Principe, S., Ronquist, G., Royo, F., Sahoo, S., Spinelli, C., Stensballe, A., Thery, C., van Herwijnen, M., Wauben, M., Welton, J., Zhao, J. and Mathivanan, S. (2017). A novel community driven software for functional enrichment analysis of extracellular vesicles data. J Extracellular Vesicles. 1:1321455.

16. Pathan, M., Keerthikumar, S., Ang, C.S., Gangoda, L., Quek, C.M.J., Williamson, N.J., Mouradov, D., Sieber, O.M., Simpson, R.J., Salim, A., Bacic, A., Hill, A.F., Stroud, D.A., Ryan, M.T., Agbinya, J.A., Mariadasson, J.M., Burgess, A.W. and Mathivanan, S. (2015) FunRich: a standalone tool for functional enrichment analysis. Proteomics.15, 2597-2601.

17. Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44:e71. doi: 10.1093/nar/gkv1507

18. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis, *Nucleic Acids Research*, Volume 47, Issue W1, 02 July 2019, Pages W556–W560, https://doi.org/10.1093/nar/gkz430

19. Müller, F., Scherer, M., Assenov, Y. *et al.* RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol* 20, 55 (2019). https://doi.org/10.1186/s13059-019-1664-9

20. Song L, Zhang B, Feng Y, Luo X, Wei X and Xiao X (2009) A role for forkhead box A1 in acute lung injury. *Inflammation* 32, 322–332)

21. Song L, Wei X, Zhang B, Luo X, Liu J, Feng Y and Xiao X (2009) Role of Foxa1 in regulation of bcl2 expression during oxidative-stress-induced apoptosis in A549 type II pneumocytes. *Cell Stress Chaperones* 14, 417–425)

22. Zhang, K., Yang, L., Wang, J. *et al.* Ubiquitin-specific protease 22 is critical to in vivo angiogenesis, growth and metastasis of non-small cell lung cancer. *Cell Commun Signal* 17, 167 (2019). https://doi.org/10.1186/s12964-019-0480-x

23. Sarris, Evangelos G et al. "The Biological Role of PI3K Pathway in Lung Cancer." *Pharmaceuticals (Basel, Switzerland)* vol. 5,11 1236-64. 20 Nov. 2012, doi:10.3390/ph5111236

24. Suming Xu, Yaoqin Wang, Yanhong Li, Lei Zhang, Chunfang Wang, Xueqing Wu; Comprehensive analysis of inhibitor of differentiation/DNA-binding gene family in lung cancer using bioinformatics methods. *Biosci Rep* 28 February 2020; 40 (2): BSR20193075. doi: **https://doi.org/10.1042/BSR20193075**

25. Falvella, F., Colombo, F., Spinola, M. *et al.* BHLHB3: a candidate tumor suppressor in lung cancer. *Oncogene* 27, 3761–3764 (2008). https://doi.org/10.1038/sj.onc.1211038)

26. Ahmad Z, Raza A, Patel MR. Endometrial metastasis of lung adenocarcinoma: a report of two cases. *Am J Case Rep*. 2015;16:296-299. Published 2015 May 18. doi:10.12659/AJCR.892495

27. Suming Xu, Yaoqin Wang, Yanhong Li, Lei Zhang, Chunfang Wang, Xueqing Wu; Comprehensive analysis of inhibitor of differentiation/DNA-binding gene family in lung cancer using bioinformatics methods. *Biosci Rep* 28 February 2020; 40 (2): BSR20193075. doi: **https://doi.org/10.1042/BSR20193075**

28. Cheng, Haiying et al. "Targeting the PI3K/AKT/mTOR pathway: potential for lung cancer treatment." *Lung cancer management* vol. 3,1 (2014): 67-75. doi:10.2217/lmt.13.72

Jen, Jayu, and Yi-Ching Wang. "Zinc finger proteins in cancer progression." *Journal of biomedical science* vol. 23,1 53. 13 Jul. 2016, doi:10.1186/s12929-016-0269-9