Integrative Analysis of Gene Expression and MicroRNA Expression in Preprocessed TCGA Data for Skin Cutaneous Melanoma (SKCM)

Introduction

Melanoma is a type of skin cancer that develops when melanocytes (the cells that give the skin its tan or brown color) start to grow out of control. So, Melanoma develops in the cells (melanocytes) that produce melanin which is the pigment that gives our skin its color. Melanoma can also form in your eyes and, rarely, inside your body, such as in your nose or throat. The exact cause of all melanomas isn't clear, but exposure to ultraviolet (UV) radiation from sunlight or tanning lamps and beds increases your risk of developing melanoma. Cells in nearly any part of the body can become cancer, and can then spread to other areas of the body. Melanoma is much less common than some other types of skin cancers. But melanoma is more dangerous because it's much more likely to spread to other parts of the body if not caught and treated early. The risk of melanoma seems to be increasing in people under 40, especially women. Knowing the warning signs of skin cancer can help ensure that cancerous changes are detected and treated before the cancer has spread. Melanoma can be treated successfully if it is detected early. The first melanoma signs and symptoms often are: a change in an existing mole or The development of a new pigmented or unusual-looking growth on the skin. Melanoma doesn't always begin as a mole. It can also occur on otherwise normal-appearing skin. Melanoma occurs when something goes wrong in the melanin-producing cells (melanocytes) that give color to your skin. It's likely that a combination of factors, including environmental and genetic factors, causes melanoma. Still, doctors believe exposure to ultraviolet (UV) radiation from the sun and from tanning lamps and beds is the leading cause of melanoma. UV light doesn't cause all melanomas, especially those that occur in places on your body that don't receive exposure to sunlight. This indicates that other factors may contribute to your risk of melanoma.

In this TCGA study, we are aiming to provide a large-scale integrative view of melanoma across gene expression and mirna analysis to indicate the hidden factors that may cause melanoma. Also, to indicate which differentially expressed genes and microRNA (transcripts) are up or downregulated in the control and case (survival and diseased) patients to represent the significant deferentially expressed genes and transcripts that may be used as biomarkers for melanoma skin cancer.

RNAseq Methodology

The analysis is done using R programming language in one script in R studio. The aim of this mirna analysis is to indicate which differentially expressed microRNA (transcripts) are upregulated or downregulated in the control and case (survival and diseased) patients to represent the significant deferentially expressed microRNAs or transcripts that may be used as biomarkers for melanoma skin cancer. In order to do this differential expression analysis, DESeq2 Bioconductor package is installed and used to represent the differentially expressed transcripts and to measure the log2fold change of these transcripts. Fold change (FC) is a measure describing the degree of quantity change between final and original value. The log2(fold-change) is the log-ratio of a gene's or a transcript's expression values in two different conditions. While comparing two conditions each feature you analyze gets (normalized) expression values. This value can be zero and thus lead to undefined ratios. If we use log2(fold change), fold changes lower than 1 become negative, while those greater than 1 become positive. DESeq2 is a tool for differential gene expression analysis of RNA-seq data. It is a new version of DESeq and can detect more differentially expressed genes (DEGs) than DESeq2. However, it also seems to allow more false positives. The DESeq2 algorithm uses the negative binomial distribution, the Wald, and the Likelihood Ratio Tests. DESeq2 performs an internal normalization where geometric mean is calculated for each mirna across all samples. The counts for a mirna in each sample is then divided by this mean. The median of these ratios in a sample is the size factor for that sample. This procedure corrects for library size and RNA composition bias, which can arise for example when only a small number of miRNAs are very highly expressed in one experiment condition but not in the other. This tool takes as input a table of raw counts, here in our data we have the mirna data which is this table of row counts that contain the columns as the samples which are 452 samples and the rows as the miRNAs and here we have 1046 miRNAs. The count table has to be associated with a phenodata or meta data file describing the experimental groups, here in our dataset, the metadata file is the melanoma that includes the columns as the factors and the most important column here is the vital status that includes the diseased and survival patients, and the rows of melanoma metadata include the samples, so the data and meta data are the input of deseq2. In deseq2, columns of the data must be the same as rows of metadata. So, by R code I have done data preprocessing on files of data(mirna) and metadata(melanoma) to intersect the data and meta data in order to make the columns of data same as the rows of metadata and by the same order to apply

deseq2. Also, patients (samples) name are dotted in data and dashed in meta data so I made the same dashed in both in order to be able to perform deseq2 without any error. In addition, the deseq2 package require the count data values to be integers so I converted the data values to integers and renamed the rows of the data. Data preprocessing also included removing NA from vital status column that indicates the diseased and survival patients as this column is our design matrix in the input of deseq2. After creating the deseq of dataset object, the run of pipeline of differential expression steps is done so firstly running the command of deseq2 that specify how many conditions do we want to compare according to the phenotypic table (vital status column in the metadata) so we have 2 conditions the first is diseased and the second is living. After that, the contrast is specified the to make a result object based on two specific conditions. Then the most important step after differential expression analysis is to choose the statistical significant differentially expressed transcripts of miRNAs (DETs) based on the p-adjusted value less than 0.05 and biological significance based on the fold change more than 1.2, these results in 3 significant miRNAs that will be reported in the results as a biomarker that could affect melanoma and detect it as the over and down differential expression of these miRNA is significant in melanoma skin cancer. Finally, I did a principle component analysis on vital status and gender 2 columns of meta data for dimensionality reduction of data and I plot all the plots needed for visualization of data in this analysis and I will report all these plots in the results section. These plots are

1) histogram for visualization of distribution of the data
2) Box plot to scale the data by log2 transformation for better visualization, the +1 at the end of command is to avoid the infinity at log the values equal to zero
3) QQ plot for testing the normality of the data
4) Volcano plot for visualizing the differentially expressed miRNAs
5) Heat map to visualize the significant differentially expressed miRNAs
6) Finally, PCA for multidimensional data visualization for dimensionality reduction (packages used for PCA are tidyverrse, broom, ggbiplot and devtools).
7) Gene enrichment analysis, I tried many tools for miRNA enrichment analysis but all give me no results for of the 3 significant miRNAs as the tools may not have pathways in their options so I couldn't know the interaction of these significant miRNA in which pathways, also, g-profiler has no results. So, instead we did the enrichment analysis on genes (exp file) not on miRNA.

Results of miRNA analysis:

1) Histogram in Fig. (1) shows a visualization of distribution of the data, the result of histogram plot shows that the data is right skewed, so, it is not normally distributed.

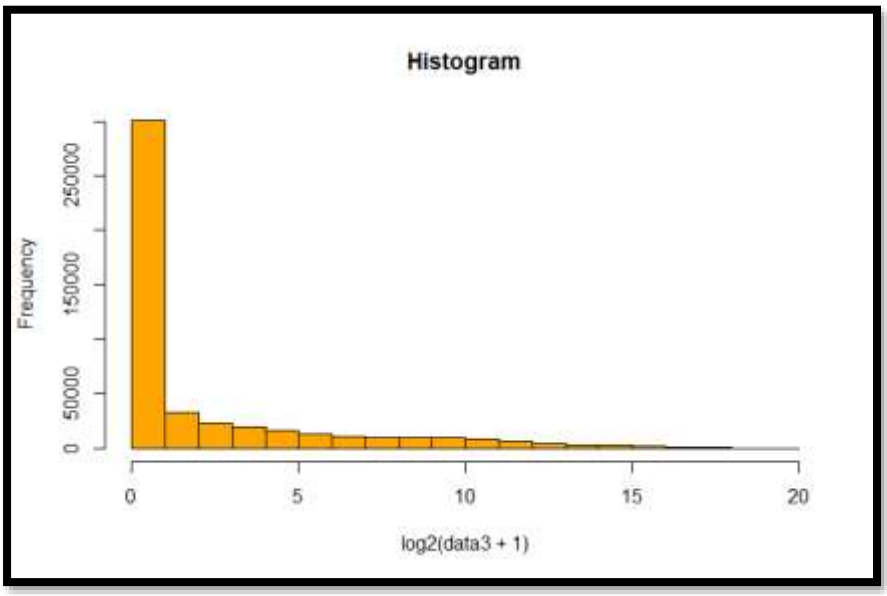


Fig. (1): *Showing the histogram for the miRNA expression data in a right skewed pattern.*

2) Box plot to scale the data by log2 transformation for better visualization shown in Fig. (2) , the +1 at the end of command is to avoid the infinity at log the values equal to zero, the result is the same that the data is not normally distributed.

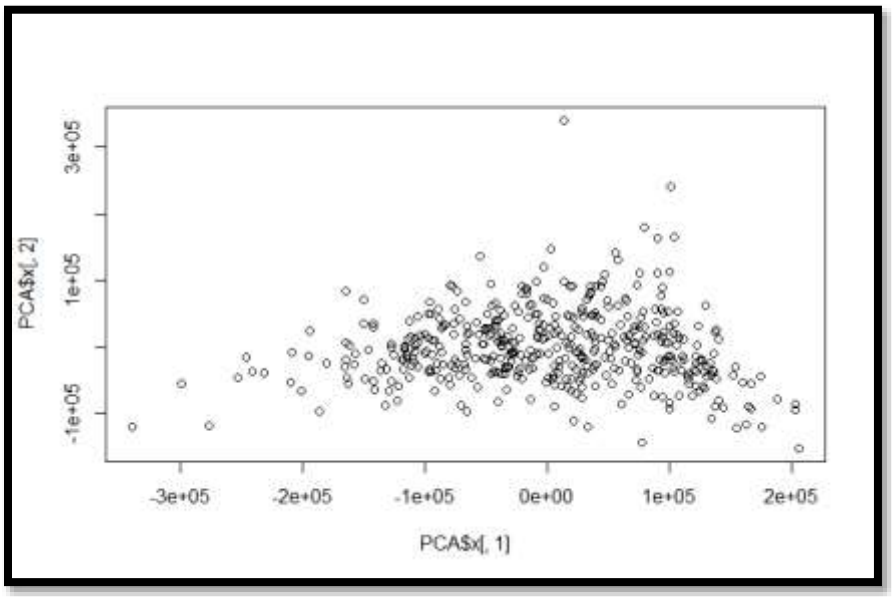


Fig. (2): *Showing the box plot for the miRNA.*

3) QQ plot shown in Fig. (3) is for testing the normality of the data, the results show that the data is not normally distributed as not all the dots are fitted on the line of the QQ norm.

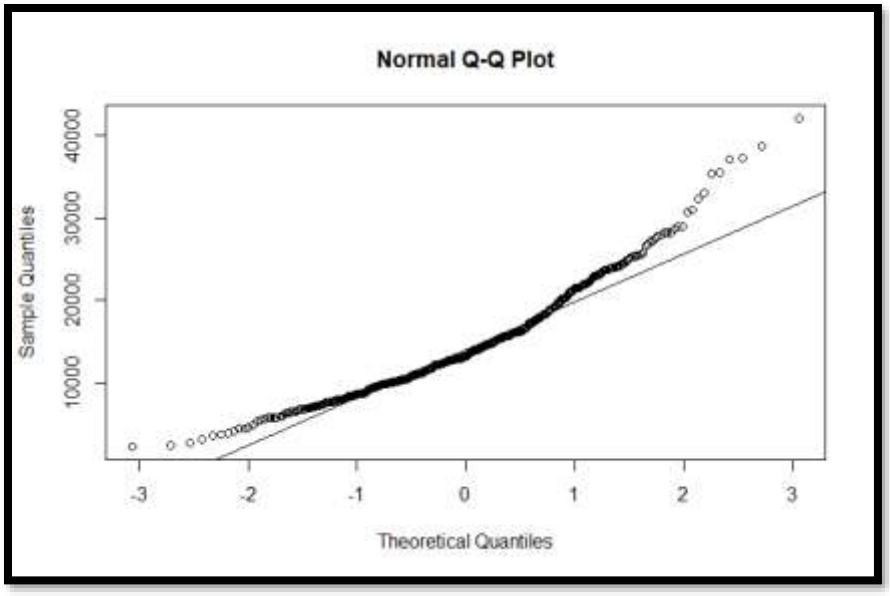


Fig. (3): *Showing the QQ plot for the miRNA showing that the data is not normally distributed.*

4) Volcano plot for visualizing the differentially expressed miRNAs in Fig. (4), the blue color indicates the upregulated differentially expressed miRNAs and the red color indicates the downregulated differentially expressed miRNAs. The total was 3 significant miRNAs that are differentially expressed (significant with p-adjusted value <0.05 and log2foldchange >1.2 (for positive FC upregulated and p-adjusted value <0.05 and log2FC <-1.2 for negative downregulated miRNAs). Here we have 2 miRNAs are down regulated and 1 is upregulated.

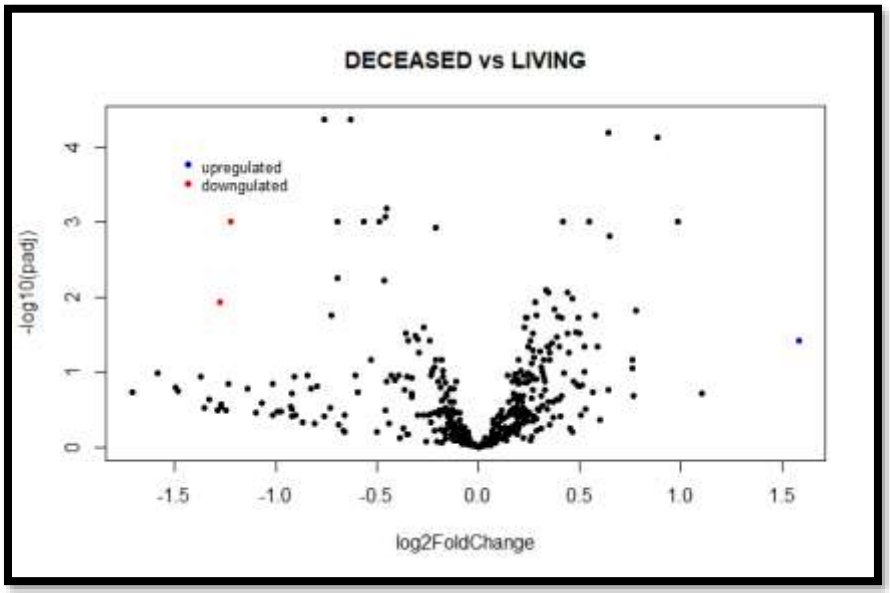


Fig. (4): *Showing the volcano plot for the miRNA with 2 red dots miRNAs down regulated and one blue dot miRNA up regulated.*

5) Heat map in Fig. (5) was generated to visualize the significant differentially expressed miRNAsfor p-adjusted <0.05 and log2FC >1.2, the result is 3 significant differentially expressed miRNAs, these significant micro RNAs could be used later as a biomarker as said before in the methodology section for detection and affection of melanoma by the down and up regulation of these significant differentially expressed miRNAs. But here in our data, we don't have case and control we have all are cases but some are dead and some are survived so I couldn't interpret the results of these 3 significant miRNAs as totally detection or affection of melanoma.

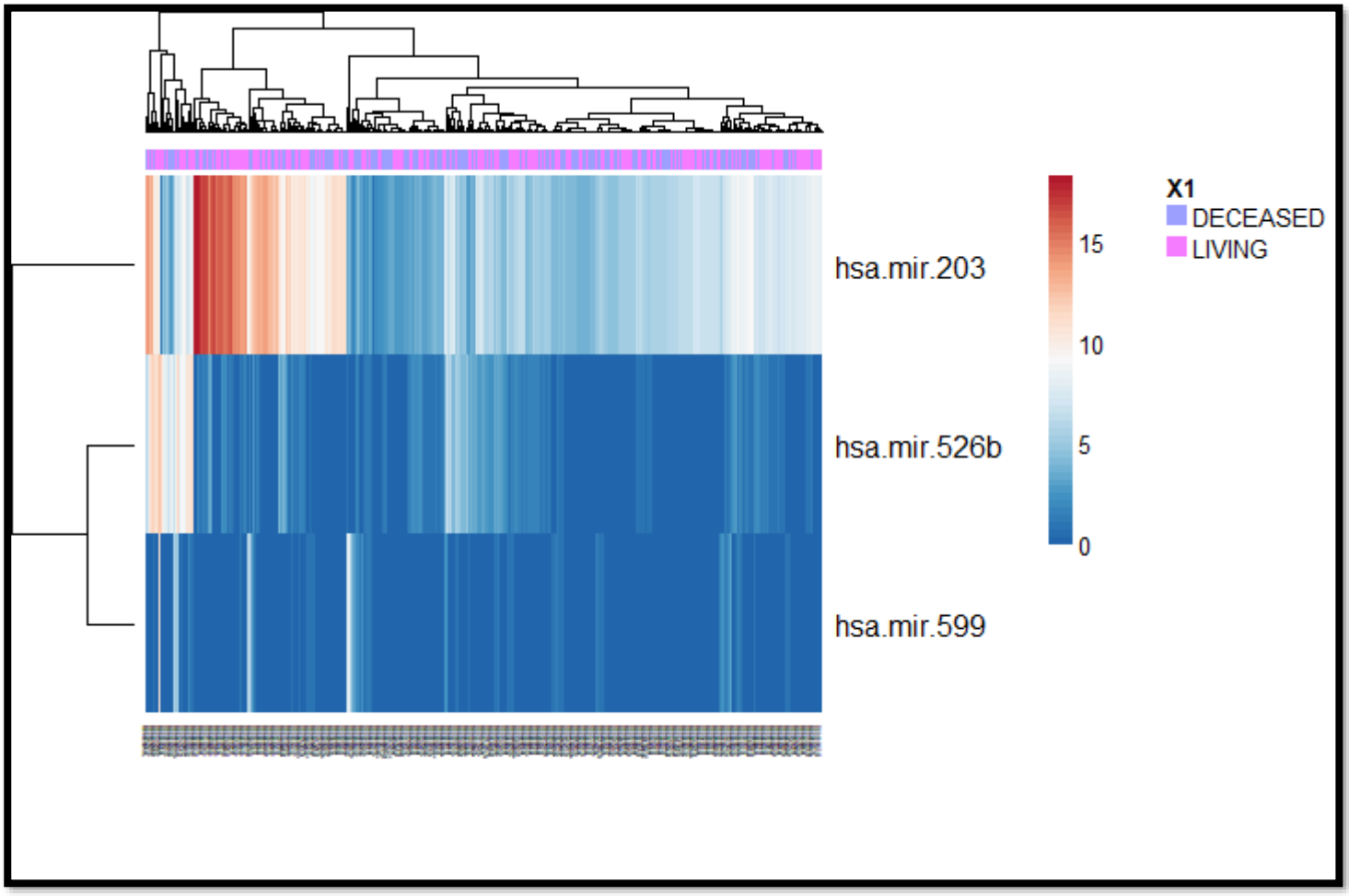


Fig. (5): *Showing the heat map the 3 significant miRNAs that were found in our analysis.*

6) PCA (principal component analysis) for multidimensional data visualization for dimensionality reduction plot results the variability of data is each plot, one for the vital status diseased and living shown in Fig. (6), and one for the gender male and female shown in Fig. (7), the results of the pca shows 37% variance of pc1 and 13% variance of pc2 captured of the total variability of data and it is obvious that pc1 capture more variability of data.

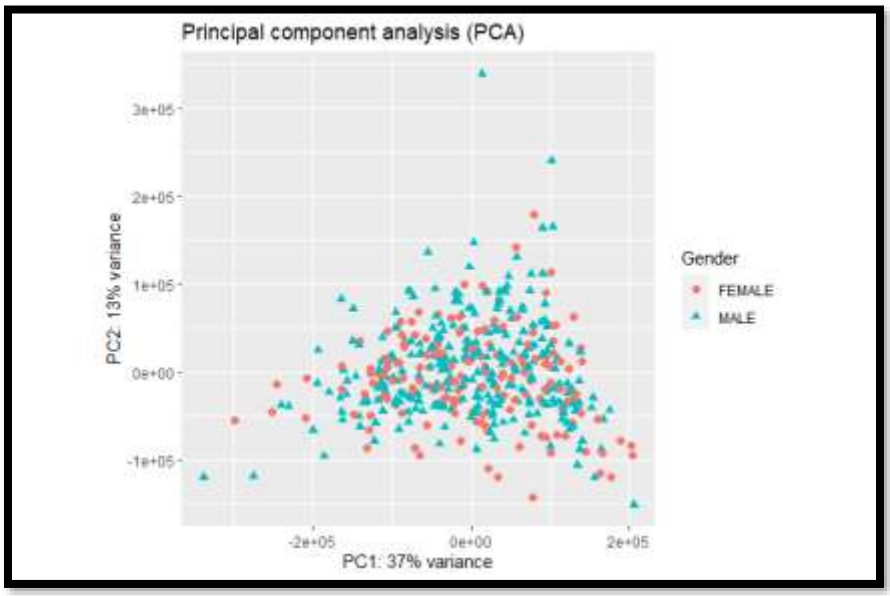
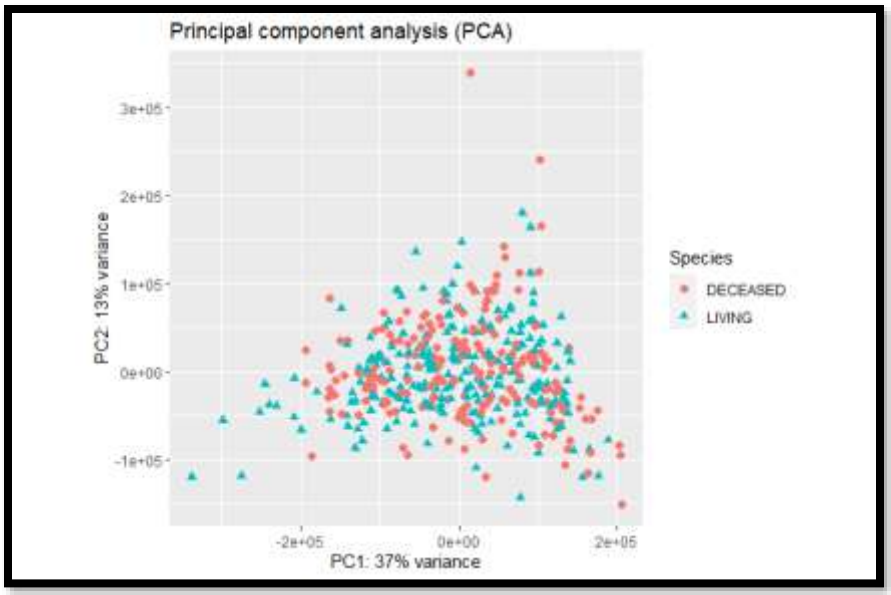


Fig. (6): *Showing the PCA plot for miRNA expression data based on gender according to meta data.*

Fig. (7): *Showing the PCA plot for miRNA expression data based on vital status according to meta data.*



Gene Expression Analysis Methodology

Filtration:

The gene expression data containing 20531 genes and 473 patient were analyzed in a single R script on R Studio. Initially, all missing (NA) data was removed from both the gene expression data and the meta data, specifically the vital status, as the gene expression analysis was based on deceased and survived patients. A single patient with missing data was found in the meta data, but none were found in the gene expression data. To intersect the meta data patients and the patients in the expression data, the "-" in the patient names in the meta data were substituted with "." to match the patient names in the gene expression data.

Exploring Data distribution:

The gene expression data was analyzed for normal distribution using statistical techniques. A histogram was created to visualize the distribution of the data and assess its normality. The results of the histogram were further confirmed by conducting quantile-quantile (QQ) plots and box plots, which are commonly used techniques to analyze the distribution of data. The use of these statistical techniques enabled us to determine the normality of the gene expression data and make reliable inferences about the underlying biological processes being studied.

Differential expression:

Differential expression was performed using DEseq2, the same tool used in miRNA analysis. To ensure the efficiency of DEseq2, the columns of the expression data and the rows of the meta data (sample information) were ordered, and the data values were converted to integers as DEseq2 requires integer data. The conditions "deceased" and "alive" were identified, and the analysis was run using the meta data column containing these conditions. The results were exported in a CSV file and three additional files for the gProfiler and GSEA tools, to perform Gene Enrichment Analysis and pathway analysis using two different methods.

Once the results from DEseq2 were obtained, a volcano plot was generated to visualize the most significantly differentially expressed genes in deceased patients by showing the upregulated and downregulated genes. The significance was calculated based on genes with p-values less than 0.05 and log fold changes greater than 1.5. A heat map was also generated to provide further visualization of the significant expressed genes.

PCA:

To investigate the underlying structure of the sample population, a principal component analysis (PCA) was performed using various features present in the meta data, including gender, vital status, and others. The PCA was executed using an R script. This analysis aimed to draw insights into the sample population structure and identify any potential relationships among the features. The results of the PCA could be useful in determining the relevance of each feature in characterizing the sample population, and guiding further studies.

Enrichment analysis:

The significantly differentially expressed genes were further analyzed using Gene Set Enrichment Analysis (GSEA) and pathway analysis. The gene names obtained from the previous differential expression analysis were utilized as input for gprofiler to identify the most significant pathways that are correlated with the genes of interest. Subsequently, the files were modified to meet the requirements for GSEA analysis, which was then performed using the GSEA tool. The objective of this analysis was to gain a deeper understanding of the biological processes and functions associated with the significant differentially expressed genes.

Gene Expression Analysis Results

Data distribution:

The Histogram in Fig. (8) showed that the gene expression data doesn't follow the normal distribution and that it is presented in right skewed pattern.

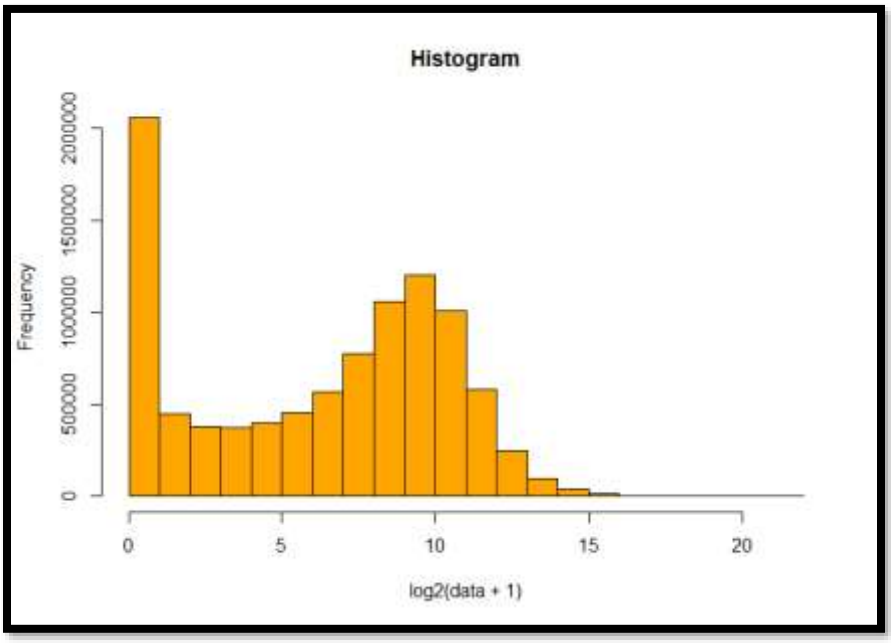


Fig. (8): *Showing the histogram for the gene expression data in a right skewed pattern.*

This was further confirmed by the quantile-quantile (QQ) plot in Fig. (9), where the observed values were not aligned with the expected normal distribution line. Additionally, the box plot in Fig. (10) exhibited a zigzag pattern, further supporting the conclusion that the data was not normally distributed.

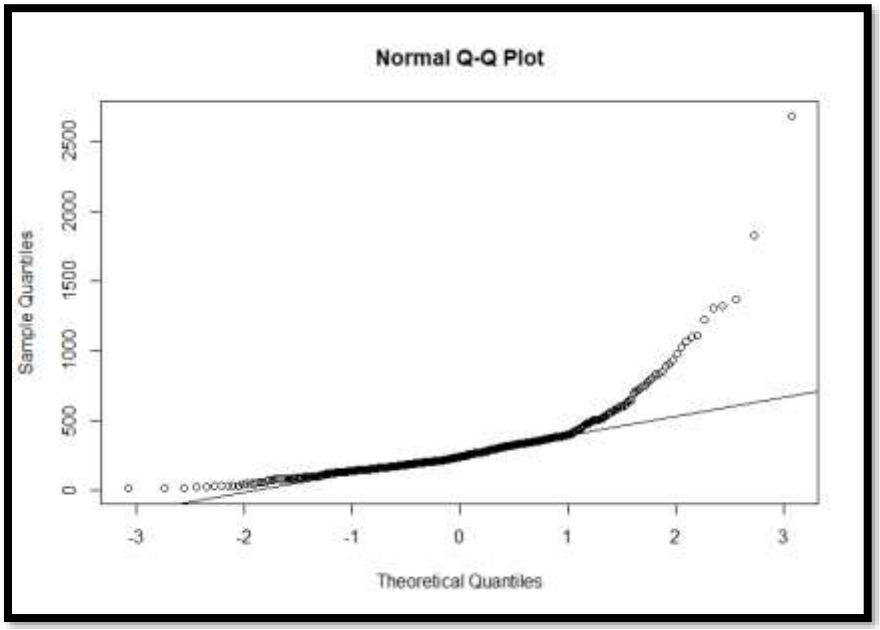


Fig. (9): *Showing the QQ plot for the gene expression data with data points not aligned to the normalization line.*

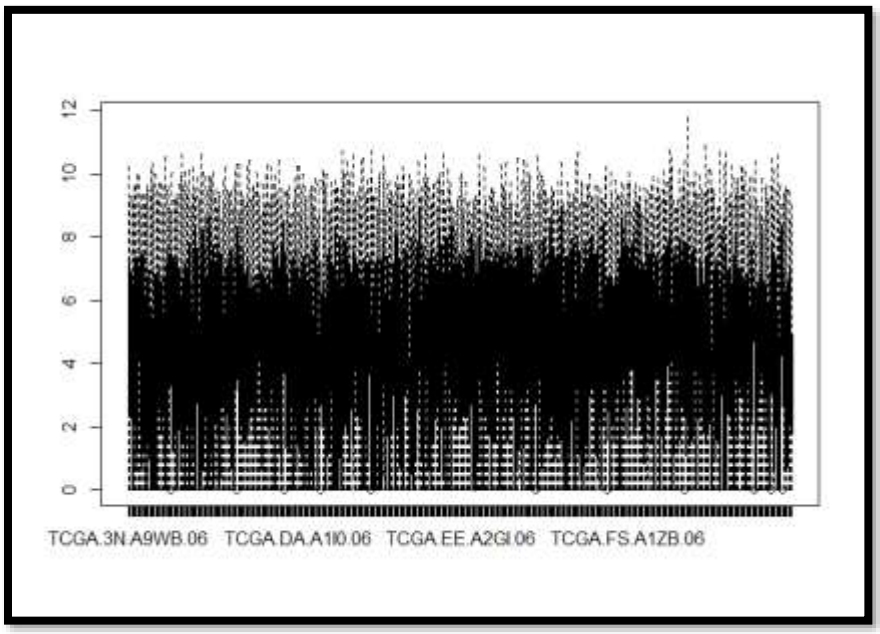


Fig. (10): *Showing the box plot for the gene expression data.*

Volcano Plot & heat map:

The results from the DEseq2 showed 64 genes that were significantly expressed based on the p-value and the log fold change in the deceased state in patients with SKCM with 41 genes upregulated and 23 genes downregulated. This was visualized using the volcano plot shown in Fig. (11) highlighting the genes that are most significantly differentially expressed as the red dots represent the up regulated genes and the blue dots the down regulated genes. A heat map shown in Fig. (12) was also generated that showed the most differentially expressed genes.

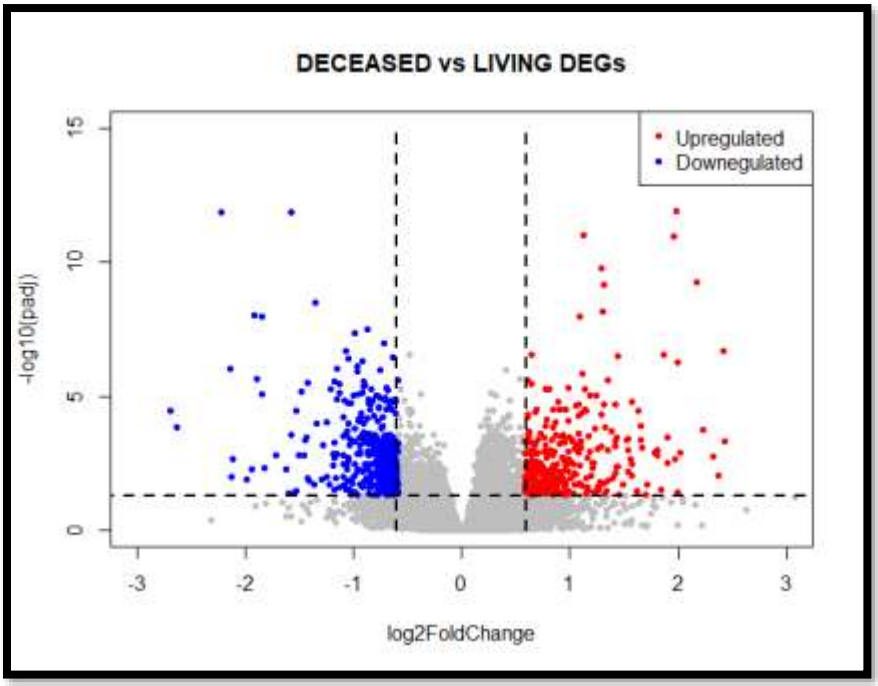


Fig. (11): *Showing the volcano plot for most differentially expressed genes.*

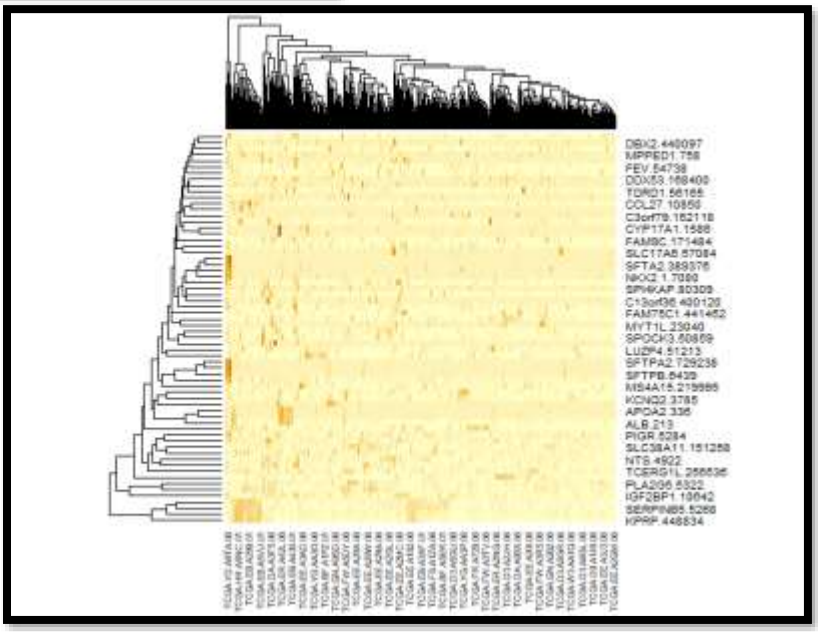


Fig. (12): *Showing the heat map plot for most differentially expressed genes.*

PCA:

The gene expression data underwent principal component analysis (PCA) without any filtration. The first principal component (PC1) accounted for 24% of the variance, while the second principal component (PC2) accounted for 19% of the variance. To investigate the potential effect of vital status and gender on gene expression, meta data was used to classify samples based on these two factors. However, the PCA plots based on vital status and gender (Fig. (13) and (14), respectively) revealed no clear correlation between these factors and gene expression. The data points in both plots were randomly distributed, indicating a lack of a specific pattern.

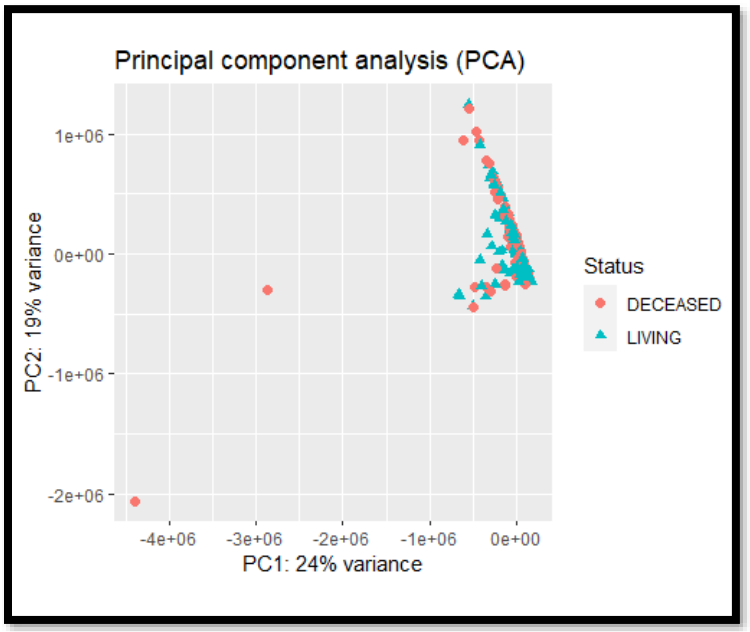
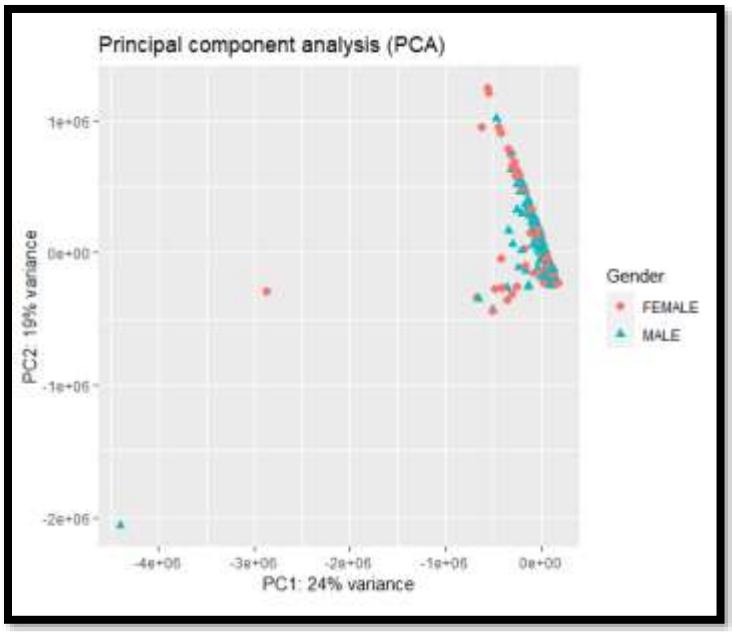


Fig. (12): *Showing the PCA plot for gene expressions based on vital status.*

Fig. (13): *Showing the PCA plot for gene expressions based on gender.*



Enrichment analysis:

(1) gprofiler

GProfiler is a web-based platform that conducts gene set enrichment analysis (GSEA) on a given list of genes. The tool utilizes gene-to-function mapping algorithms to identify statistically significant functional terms, such as biological pathways and processes that are over-represented among the input genes.
In the current study, we applied GProfiler to analyze the differentially expressed genes in a cutaneous melanoma (SKCM) dataset. The results showed significant enrichment of the surfactant metabolism pathway and the colony-stimulating factor receptor A (CSF2RA) gene in the Reactome database, with no hits in the KEGG database. These findings suggest a correlation between the differentially expressed genes in SKCM and pulmonary diseases, as both pathways are associated with lung-related disorders.

source	Term_name	Adjusted_p_value
REAC	Diseases associated with surfactant metabolism	2.21148E-06
REAC	Defective CSF2RA causes SMDP4	0.000238463
REAC	Defective CSF2RB causes SMDP5	0.000238463
REAC	Surfactant metabolism	0.000400591
REAC	Diseases of metabolism	0.016917285

Table (1): *Showing the gprofiler results.*

(2) GSEA software

Gene Set Enrichment Analysis (GSEA) is a computational approach that evaluates the overrepresentation of a pre-defined set of genes in a given sample or dataset. The method compares the distribution of genes between two distinct biological conditions and determines whether the set of genes exhibits statistically significant differences in expression.
GSEA was used to compare gene expression in two groups of samples - deceased and living individuals. To perform GSEA, we separated their data into two separate files and combined them into a format suitable for analysis. We also prepared a phenotype file, which is required by the GSEA tool. The results showed that 27 out of 50 gene sets were enriched in the deceased group. Specifically, the Human Gene Sets HALLMARK_MITOTIC_SPINDLE (Fig. 14) and HALLMARK_G2M_CHECKPOINT (Fig. 15) were among the enriched gene sets.

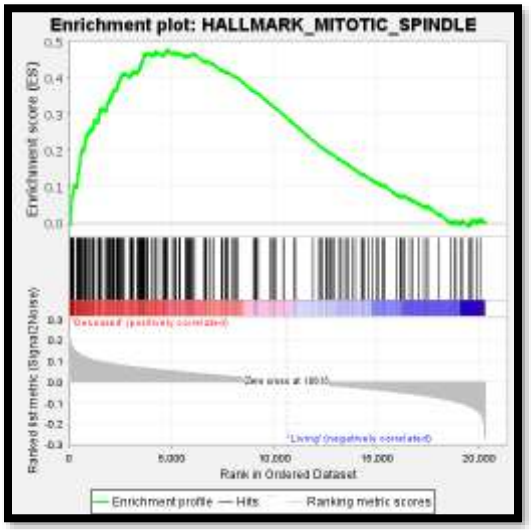


Fig. (14): *GSEA results showing the Up regulation of the mitotic spindle pathway in deceased state compared to living.*

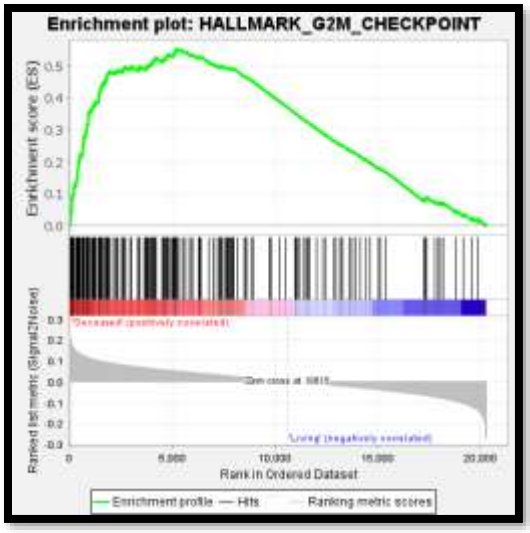


Fig. (15): *GSEA results showing the Up regulation of the G2M pathway in deceased state compared to living.*

Integromics analysis

The rapid decrease in high throughput sequencing cost in parallel with increasing accuracy of the measurement of the other omics such as metabolomics and proteomics push toward the tendency to account for multiple data modalities. However, the computational power required to integrate different omics data represents a challenge due to the normalization method variations for each omic and the probability of having a large number of missing values. Using Multi-Omics Factor Analysis (MOFA) as a statistical framework can address most of these challenges. However, the Application of MOFA analysis does not account for some biases such as batch bias. On the other hand, pooling and contrasting information across studies could give us more comprehensive insights into the complexity underlying biological systems.

Methods

MOFA is a type of matrix factorization and mixed factor category. It differs from PCA in the fact that it can infer an interpretable low-dimensional representation in terms of latent factors which could effectively address the relevant signal in the input data. Furthermore, it requires a relatively large sample size of not less than 15 samples also it deals with missing data.

The experimental analysis started with preprocessing and exploration of different input data and then intersecting the dominant samples in each omic. Also, the model was fitted with the type of input data by adapting the data and model options. For instance, the scale view of the data option was changed to True.

After running the MOFA object, the highly weighted Factors were annotated using the supplemented metadata to investigate if these factors could account for any of the measured metadata. Also, the genes with high weights correlated to the most significant factor were subjected to enrichment analysis using different databases at the Gprofiler website to explore the pathways that could be involved in this factor.

Integromics analysis results

Exploration of the three omics data showed that the range of data varies between different omics as illustrated in Fig. (14 A, B, and C). Consequently, the scale view of the data option of MOFA parameters was changed to True. Also, it was found that the number of intersected samples is equal to 450 as shown in Fig. (14 D).

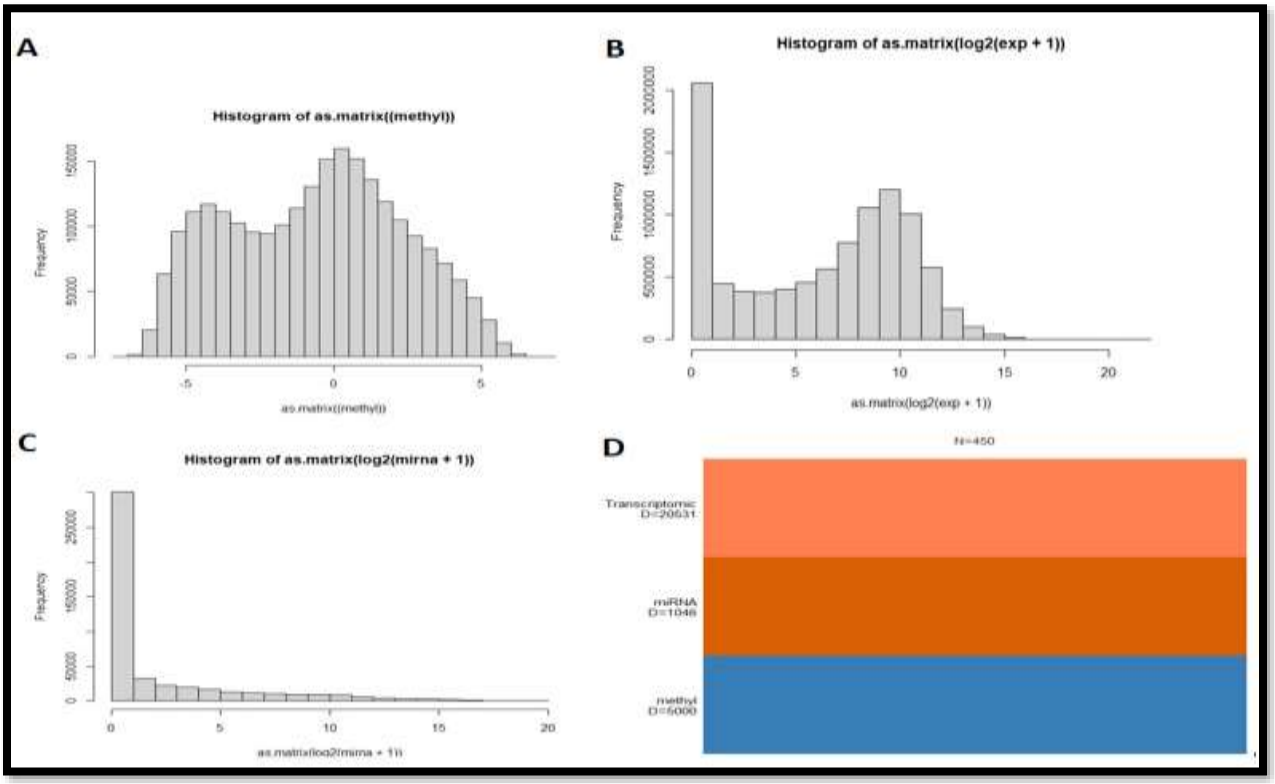


***Fig. (16)**: A, B, and C show the data exploration of different Omics, D shows the number of samples subjected to MOFA analysis.*

The likelihood parameter was also adjusted to Gaussian for the three omics and the group parameter to False. By running the MOFA object, Factors with different variation percentages appeared. As represented in Fig. (17 A) only factor 2 has a moderate effect on transcriptomic and factor 3 showed a similar effect on miRNA while factor 1 obviously has a significant effect on the methylation model. Different variances of factors were recorded in Table (1). Furthermore, the heat map plots illustrated in Fig. (17 B, C, and D) represent the weight of the first five factors on different omics and whether these factors up-regulate or down-regulate the recorded features.

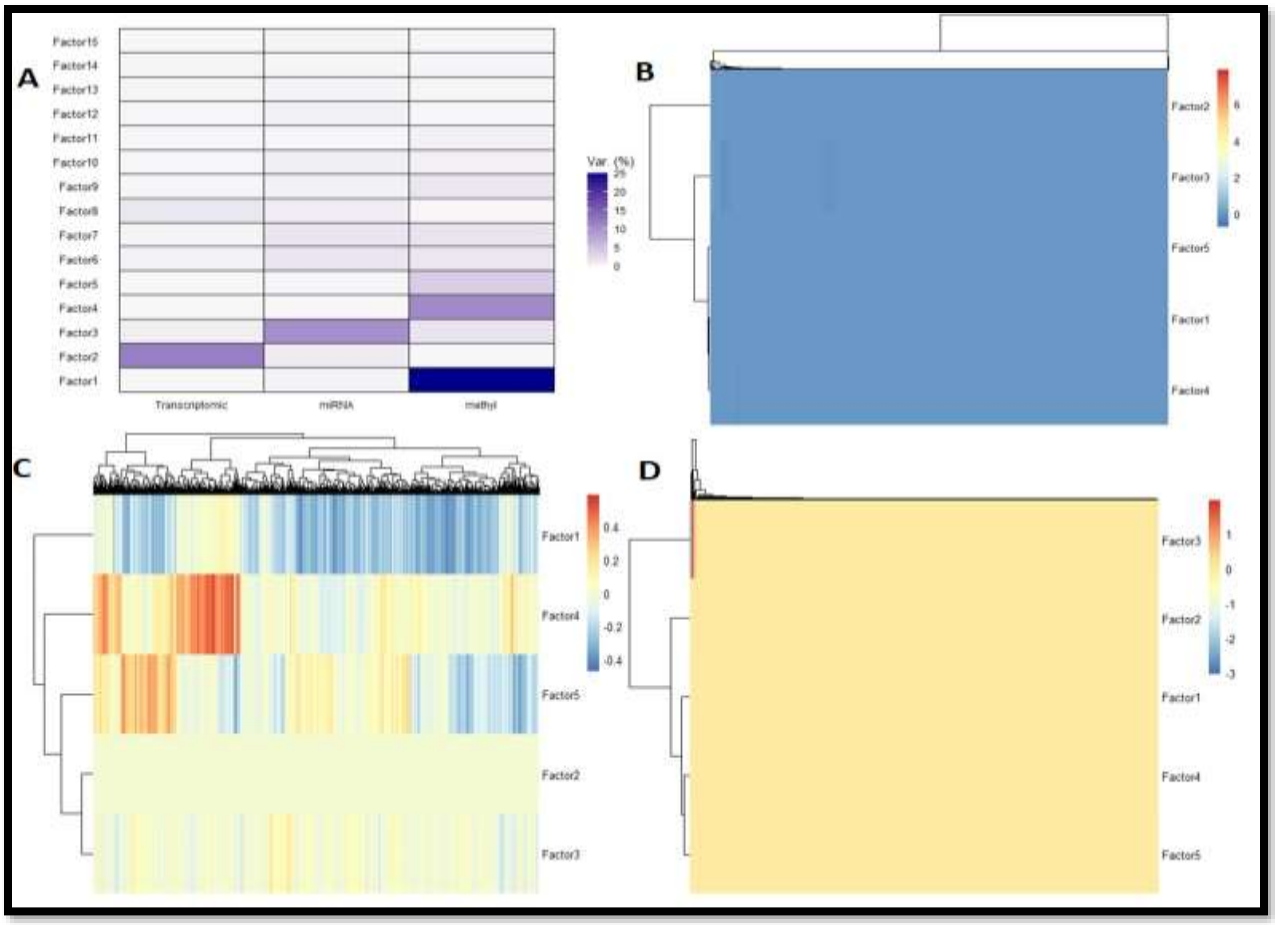


Fig. (17): A; showed variance per each factor, B, C, and D; heat map plots represent the weight of the first five factors on different omics; transcriptomic, methylation, and miRNA respectively.

Table 2: Different variance of the top 10 factors.

	Transcriptomic	miRNA	Methylation
Factor 1	0.009509473	0.3600252	25.02014174
Factor 2	11.92205398	1.15146223	0.000117751
Factor 3	0.784866861	10.29299781	1.777308367
Factor 4	0.009050113	0.01754803	10.74563729
Factor 5	0.008402158	0.08568564	3.939771645
Factor 6	0.500316575	1.70587409	1.569868218
Factor 7	0.245021057	1.62830976	1.805255087
Factor 8	1.534628954	1.0605555	0.03494122
Factor 9	0.047913571	0.60034546	1.590610336
Factor 10	0.093828339	0.74625869	0.635306857

Also, the rank and weight of the high-weight features were represented in Supp. Fig (1, 2, 3, and 4). And to assign the significant factors for each model, the records of metadata file (gender, sample type, tumor tissue site, and vital status) were used to annotate these factors Supp. Fig. (5, 6, 7, 8, 9, 10, 11, and 12). However, there is no clear correlation between these factors and metadata records. Consequently as represented in Supp. Table 1, enrichment analysis for the high-weight features (genes) of factor 2 (transcriptomic model) was conducted and the results revealed that; according to the GO molecular function those genes correlated mostly to three functions (structural constituent of skin epidermis, structural constituent of cytoskeleton, and structural molecule activity) while, according to GO cellular component those genes correlated mostly to keratinocyte differentiation, keratinization, and epidermis development. However, according to GO biological process, those are correlated mainly to keratin filament, intermediate filament, and intermediate filament cytoskeleton. Additionally, according to Reactome databases, the genes play a vital role in the formation of the cornified envelope and keratinization. On the other hand, there is no obvious pathway according to KEGG databases.

References:

1) What is melanoma skin cancer? What Is Melanoma? (n.d.). Retrieved February 5, 2023, from https://www.cancer.org/cancer/melanoma-skin-cancer/about/what-is-melanoma.html

2) NHS. (n.d.). NHS choices. Retrieved February 5, 2023, from https://www.nhs.uk/conditions/melanoma-skin-cancer/

3) Melanoma skin cancer. Melanoma skin cancer | Cancer Research UK. (2020, May 21). Retrieved February 5, 2023, from https://www.cancerresearchuk.org/about-cancer/melanoma

4) Melanoma treatment (PDQ®)–patient version. National Cancer Institute. (n.d.). Retrieved February 5, 2023, from https://www.cancer.gov/types/skin/patient/melanoma-treatment-pdq

5) Tammi, D. M. T. (n.d.). DESEQ2. Retrieved February 5, 2023, from https://bioinformaticshome.com/tools/rna-seq/descriptions/DESeq2.html#gsc.tab=0

6) Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with deseq2. Genome biology. Retrieved February 5, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/

7) Deseq2 manual - bioconductor. (n.d.). Retrieved February 5, 2023, from https://www.bioconductor.org/packages/devel/bioc/manuals/DESeq2/man/DESeq2.pdf/

8) Creative Proteomics. (n.d.). Bioinformatic Fold Change Analysis Service. Creative Proteomics. Retrieved February 5, 2023, from https://www.creative-proteomics.com/services/bioinformatic-fold-change-analysis-service.htm

9) seqprone Junior Member Join Date: Oct 2015 Posts: 9, dpryan Devon Ryan Join Date: Jul 2011 Posts: 3478, & Michael.Ante Senior Member Join Date: Oct 2011 Posts: 127. (n.d.). Header leaderboard ad. SEQanswers. Retrieved February 5, 2023, from https://www.seqanswers.com/forum/general/51314-what-does-the-%E2%80%9Clog2-fold-change-%E2%80%9D-in-the-gene-differential-expression-testing-result

Name	Contribution	
Mohamed Elsayed Elmanzalawi	**Gene expression analysis report + R code**	**Comprehensive evaluation of the Integromics analysis results, and the generation of conclusions.**
Menna Ramadan	**Mirna expression analysis report + R code**	
Mohamed Aboalkasem	**Integromics analysis report + R code**	