Saudi Journal of Biological Sciences 30 (2023) 103819

HOSTED BY

Contents lists available at ScienceDirect

Saudi Journal of Biological Sciences

journal homepage: www.sciencedirect.com

Original article

Identification of key biomarkers and associated pathways of pancreatic cancer using integrated transcriptomic and gene network analysis

Majji Rambabu [a,1], Nagaraj Konageni [a,1], Karthick Vasudevan [a], K R Dasegowda [a], Anand Gokul [b], Sivaraman Jayanthi [c], Karunakaran Rohini [d,e,*]

[a] Department of Biotechnology, REVA University, Bengaluru, Karnataka, India
[b] Department of Computer Science, University of Southern California, Los Angeles, CA, USA
[c] Department of Biotechnology, Vellore Institute of Technology, Vellore, Tamil Nadu, India
[d] Department of Bioinformatics, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, India
[e] Unit of Biochemistry, Faculty of Medicine, AIMST University, Semeling, Bedong, Malaysia

ARTICLE INFO

Article history:
Received 17 June 2023
Revised 11 September 2023
Accepted 21 September 2023
Available online 26 September 2023

Keywords:
Pancreatic cancer
FASTQC
RNA-seq
GSEA
Cytoscape
DESeq2
Enrichment map

ABSTRACT

Pancreatic cancer shows malignancy around the world standing in 4th position for causing death globally. This cancer is majorly divided into exocrine and neuroendocrine where exocrine pancreatic ductal adenocarcinoma is observed to be nearly 85% of cases. The lack of diagnosis of pancreatic cancer is considered to be one of the major drawbacks to the prognosis and treatment of pancreatic cancer patients. The survival rate after diagnosis is very low, due to the higher incidence of drug resistance to cancer which leads to an increase in the mortality rate. The transcriptome analysis for pancreatic cancer involves dataset collection from the ENA database, incorporating them into quality control analysis to the quantification process to get the summarized read counts present in collected samples and used for further differential gene expression analysis using the DESeq2 package. Additionally, explore the enriched pathways using GSEA software and represented them by utilizing the enrichment map finally, the gene network has been constructed by Cytoscape software. Furthermore, explored the hub genes that are present in the particular pathways and how they are interconnected from one pathway to another has been analyzed. Finally, we identified the *CDKN1A, IL6*, and *MYC* genes and their associated pathways can be better biomarker for the clinical processes to increase the survival rate of of pancreatic cancer.

1. Introduction

Pancreatic cancer is the fourth most common cause of cancer death. This cancer occurs when a cell in the pancreas is damaged, causing the malignant (cancer) cell to start growing out of control (Kirby et al., 2016). GLOBOCAN 2018 estimates of pancreatic cancer incidence and mortality trends show a substantial increase in both incidence 77.7% (3,56,358 new cases) and mortality 79.9% (3,45,181 deaths) from 2018 to 2040. This is due to the inability to enhance preventative and treatment methods, and minor influence of preclinical and clinical research on patient outcomes over the last 50 years. There is a need to reduce both pancreatic cancer incidence and mortality through therapy development and adoption of primary and secondary preventative studies (Casolino et al., 2021). Pancreatic cancer has two subgroups i.e., exocrine and neuroendocrine. Exocrine pancreatic ductal adenocarcinoma is the most observed kind of pancreatic cancer in around 85% of cases (Lu et al., 2017). The risk of developing pancreatic cancer increases with age 65 or older (Kirby et al., 2016) . Pancreatic cancer is caused by a series of inherited and acquired genetic events. It is heavily influenced by inherited genetic alterations, both high and low penetrance. Patients with hereditary mutations in the pancreas may be more responsive to specifically targeted medicines, allowing for individual treatment (Chen et al., 2017). Patients with this type of cancer locally advanced have a median survival

* Corresponding author at: Unit of Biochemistry, Faculty of Medicine, AIMST University, Malaysia.
E-mail addresses: majji.rambabu@reva.edu.in (M. Rambabu), 2102446@reva.edu.in (N. Konageni), karthick.vasudevan@reva.edu.in (K. Vasudevan), dasegowda.kr@reva.edu.in (K R Dasegowda), anandgok@usc.edu (A. Gokul), jayanthi.s@vit.ac.in (S. Jayanthi), rohini@aimst.edu.my (K. Rohini).
[1] Authors with Equal Contribution.

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.sjbs.2023.103819
1319-562X/

time between 8 and 12 months, those with distant metastases on the other hand have a worse prognosis with a median survival time of 3 to 6 months (Lu et al., 2017). The accumulation of these genetic changes disrupts important signaling networks, resulting in the development of a malignant phenotype. The average number of non-synonymous somatic mutations in pancreatic ductal cancer patients ranges from 26 to 101, affecting important signaling pathways and contributing to pancreatic ductal adenocarcinoma cells malignant activity. Cell growth regulators, cell interaction with the extracellular environment, and DNA repair mechanisms are among the cellular pathways affected by somatic mutations in pancreatic ductal adenocarcinoma. Structural changes have been detected by single nucleotide variants which are rearranged, scattered, stable and unstable. In over 100 pancreatic ductal adenocarcinomas four distinct mutational changes were discovered Age, APOBEC cytidine deaminases, *BRCA1* and *BRCA2* alterations, and mismatch repair deficiencies are linked (Felsenstein et al., 2018). The early detection of malignancies is critical for developing therapeutic methods that could cure the disease, enhance disease-free survival and improve patient's quality of life. When needed systemic chemotherapy will be used to precede definitive, often curative therapy, such as surgery, ablation procedures, or three-dimensional intensity-modulated radiation. RNA-seq (high-throughput RNA sequencing) promises a complete view of the transcriptome, allowing for complete annotation and quantification of all genes and their isoforms across datasets. RNA-seq allows for the analysis of novel transcripts and has higher resolution, a wider detection range, and lower technical variability than microarrays (Corchete et al., 2020). DNA, RNA, and protein measurement in biological materials is increasingly commonplace. The generated data is rapidly accumulating and analyzing it allows researchers to find new biological functions, genotype-phenotype correlations, and disease causes (Reimand et al., 2019). RNA expression analysis has become a standard tool in biomedical research, obtaining biological information from data. Gene Set Enrichment Analysis (GSEA) is a strong analytical tool for evaluating gene expression data. This method is used for focusing on gene sets, which are collections of genes with similar biological functions, GSEA reveals a lot of biological pathways that are in common (Zhang and Zhong, 2018).

In conclusion, the combination of RNA-seq, GSEA, and network biology represents a powerful translational approach to combat pancreatic cancer. Using several methods like Precision Medicine through RNA-seq, Early Detection and Biomarker Discovery, Uncovering Biological Pathways with GSEA, Network Biology and Personalized Treatment and Improved Outcome**s.** By leveraging these advanced technologies, we can advance our understanding of the disease, identify novel therapeutic targets, and develop tailored treatment strategies that have the potential to reduce both the incidence and mortality of pancreatic cancer. This interdisciplinary approach offers hope to patients and researchers alike in the ongoing fight against this devastating disease.

2. Material and methods

2.1. Data collection and preprocessing analysis

Pancreatic cancer samples were collected from the ENA database with project ID PRJNA316672 which consists of 14 samples which have two conditions Sensitive and Resistant (Rastrojo et al., 2019). which were sequenced using Illumina HiSeq 2000 platform (Pan and Ma, 2020). The raw data obtained from the ENA database are further assessed to evaluate the quality of the data. Here we use FASTQC tool to check the quality of each dataset where we observed. (Albrecht et al., 2021)(de Sena Brandine and Smith, 2019). We also generated a summarized report of the read quality of all the samples using MultiQC software (Ewels et al., 2016).

2.2. Mapping and quantification

After aligning each of the datasets to the reference genome where we get the alignment data in the Sequence Alignment Map (SAM) format which is basically a text-based file used for storing an alignment (Srivastava et al., 2020). Further, it is converted from SAM to BAM format which provides binary versions of most of the same data by using SAM tools. The BAM alignment files are sorted to reduce memory usage and are designed to compress reasonably for downstream analysis(Oliva et al., 2021) (Danecek et al., 2021). The mapped reads were subjected to transcript quantification using the FeatureCounts program implemented by the Subread package (Li and Dewey, 2011). FeatureCounts program is used to assign mapped reads to genomic features such as genes, and exons which are specified in the reference file. FeatureCounts takes two input files: one or more sorted bam files and a GTF reference file(hg38.gtf) which is downloaded from the ENSEMBLE database (Fraser et al., 2021). A count file is generated where the number of reads is mapped to individual transcripts in the form of read counts where genes are present. This file is used to identify DEGs in the group of samples for further analysis (Liao et al., 2014).

2.3. Differential expression analysis

Differential expression analysis shows the genes with significant changes used in the experimental conditions. DESeq2 is a Bioconductor package used in R studio software for analyzing RNA-Seq data for Differential Expression analysis which uses negative binomial generalized linear models to identify statistically significant DEGs (Stupnikov et al., 2021)(Michael et al., 2013). Initially, genes were filtered based on the false discovery rate adjusted p-value < 0.05. Finally, upregulated and downregulated genes were obtained based on log2-fold change values (Love et al., 2014). Various plots such as PCA plot, Heatmap, and Volcano plot were generated the depict the gene expression results.

2.4. GSEA software

Gene Set Enrichment Analysis (GSEA) is one of the most popular bioinformatics tools which can determine the gene sets within given biological groups (Croken et al., 2014). For building GSEA Enrichment Map three different files that include hallmark geneset (gmt), expression profile (gct), and class file (cls) were prepared (Suárez-Fariñas et al., 2010)(Joly et al., 2020). The hallmark geneset file (h.all.v7.5.1.symbols.gmt was downloaded from MSig database) which has all human genes that are expressed in cancer, expression matrix contains the particular cancer expression profile, such as pancreatic cancer and class file.

2.5. Enrichment map and network analysis

Enrichment Map is an open-source and freely available plugin for Cytoscape which is used for network visualization of enriched pathways. An enrichment map can be generated based on the files obtained from GSEA results by using expression profile, hallmark geneset, enrichment sets, and setting parameters like p-value and overlap coefficient (Reimand et al., 2019). Enrichment map by showing pathway as a network in the form of nodes that are interconnected with edges that shares the common genes in respective pathways. Furthermore, analyzing the individual gene interactions associated with the different pathways are selected and a network

is built by annotating the genes with the string database using Cytoscape software (Merico et al., 2011).

3. Results

3.1. Quality control and alignment

Raw reads from the ENA database were first analyzed using the FASTQC tool to assess the quality of each individual sample. This analysis generated two output files: an HTML file and a zip file. Subsequently, we employed the multiqc tool to visualize all the samples collectively, taking a comparative approach to evaluate the overall quality of the samples in a single multiqc report (Brown et al., 2017). The results, which indicate the overall quality of all the samples, can be found in Table 1.

To determine the alignment rate for all the samples, we mapped the reads to a reference genome, and the resulting alignment rates are presented in Table 1.

Table 1
Overall Alignment results for all samples obtained by mapping to reference genome.

Samples	Total reads	Mapped reads	Alignment rate
SRR3308934	145,813,311	140,620,181	79.08%
SRR3308935	170,374,408	164,692,812	79.04%
SRR3308936	122,307,806	117,412,308	78.47%
SRR3308937	192,629,201	184,442,211	77.25%
SRR3308938	122,490,233	117,240,881	78.38%
SRR3308939	109,014,852	105,112,686	79.41%
SRR3308940	106,132,174	99,617,792	78.61%
SRR3308941	110,530,413	105,502,365	78.23%
SRR3308942	120,991,201	115,929,512	78.18%
SRR3308943	121,794,062	115,813,634	79.84%
SRR3308944	123,407,950	117,153,721	77.68%
SRR3308945	119,691,024	114,309,916	80.00%
SRR3308946	114,785,394	110,731,575	80.42%
SRR3308947	124,070,713	118,840,038	77.57%

3.2. Identification of differentially expressed genes (DEGs) in response to pancreatic cancer

3.2.1. PCA Plot and Heatmap

Principal Component Analysis (PCA) is a statistical technique harnessed for the purpose of highlighting variation and revealing prominent patterns within a provided dataset. In our analysis, we employed the DESeq2 package to conduct PCA. The PCA plot visually represents the clustering patterns between two conditions, specifically the "Resistant" and "Sensitive" conditions, based on a dataset containing a total of 14 samples. Each condition comprises seven samples. Notably, for this analysis, we focused on the top 700 most variable genes (Son et al., 2018), as detailed in Fig. 1. A heatmap is a valuable visualization tool used to depict differentially expressed genes within distinct sample groups. It enables the identification of statistically significant alterations in gene expression across hundreds to thousands of genes, each of which is associated with various treatment conditions. In this heatmap, colors are employed to represent diverse sets of values using a continuous color map (Carroll et al., 2020). On the X-axis of the heatmap, you will find the treatment conditions labeled by sample IDs, while on the Y-axis, you will see the gene names. The colors in the heatmap correspond to the level of gene expression, ranging from high to low, and are determined by the values within a defined range, typically between −2 to 2, as depicted in Fig. 2.

3.2.2. MA Plot

MA plots are used to visualize the log fold-change values (on the y-axis) plotted against the mean expression values (on the x-axis) for comparisons between two conditions. Each data point in the plot corresponds to a specific gene. In this particular plot (Fig. 3), genes that exhibit a significantly adjusted p-value of<0.05 (Love et al., 2014).

3.2.3. Dispersion plot and Volcano Plot

A mean dispersion plot was constructed to visualize the dispersion values on the y-axis and the mean of normalized counts on the x-axis for an RNA-seq experiment, as demonstrated in Fig. 4. A volcano plot, on the other hand, is a type of scatterplot that illustrates the relationship between statistical significance (p-value) and the magnitude of change (fold change) (Fig. 5). Another commonly used comparison between two treatment conditions involves plotting the adjusted P-value against the log fold change. In our analysis, we have identified and presented the top five upregulated genes in Table 2 and the top five downregulated genes in Table 3.

3.2.4. GSEA and Enrichment Map

An enrichment map serves as a visual representation tool that organizes similar gene sets into a network structure. In this visualization (Supplementary Fig. 1), each gene set is represented as a node, and the shared genes between these sets are depicted as edges connecting the nodes. The grouping of substantially similar gene sets naturally arranges the nodes. The coloration within the map corresponds to the expression levels of genes in different pathways. Specifically, nodes colored in red represent upregulated pathways, while those in blue represent downregulated pathways. These color assignments are based on the enrichment scores provided in Supplementary Table 2. Moreover, we have highlighted the most frequently recurring genes and their associated pathways in Table 4.

3.3. Identification of gene network analysis for the enriched genes using Cytoscape

3.3.1. String Network

The String App is a valuable Cytoscape plugin designed to facilitate the visualization of gene associations. It aids in the retrieval of functionally enriched genes that participate in various pathways (Fitts et al., 2016). To delve into this network, we focused on the top three genes CDKN1A, IL6, and MYC that are expressed in at least 10 different pathways. These genes were incorporated into a String network. In this String network, we included the top 50 most highly expressed genes within the pathways identified in the enrichment map. Among these genes, CDKN1A, MYC, and IL6 emerged as hub genes, being expressed in at least 10 pathways. Furthermore, we constructed a combined String network specifically highlighting the most commonly recurring genes: CDKN1A, MYC, and IL6, which is depicted in Fig. 6 (Dey et al., 2023). Additionally, a separate String network was created for CDKN1A, IL6, and MYC, shedding light on their unique gene-gene interactions with other genes, as shown in Supplementary Fig. 2. In essence, the comprehensive gene network was established to encompass all the pathways, associated genes, and their intricate gene-gene interactions (Mishra et al., 2020).

4. Discussion

Pancreatic cancer samples were collected from the ENA database which consists of 14 samples with two conditions Sensitive and Resistant. The quality assessment of each sample was performed using FASTQC, and a comprehensive overview of the sample quality can be found in Supplementary Table 1. In Table 1, you can see the results of the overall alignment to the reference genome for each sample. To identify Differentially Expressed Genes,

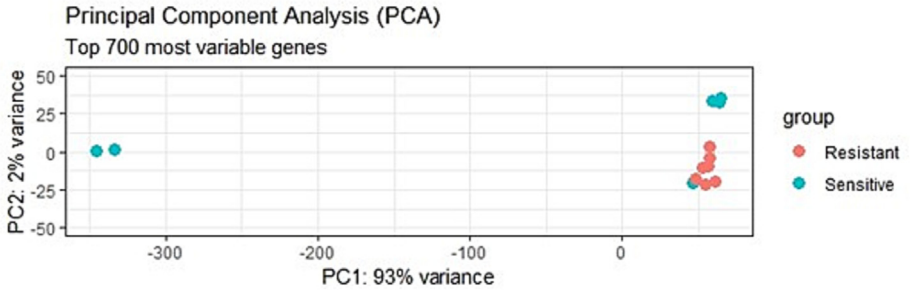

Fig. 1. The PCA plot illustrates the clustering of 14 samples, categorizing them into two conditions: ‘‘Sensitive” and ‘‘Resistant,” with each condition comprising seven samples. This clustering is based on the expression levels of the top 700 most variable genes. In the plot, the grouping of conditions is visually represented by color coding, where ‘‘light blue” signifies the ‘‘Sensitive” condition, and ‘‘dark orange” signifies the ‘‘Resistant” condition.

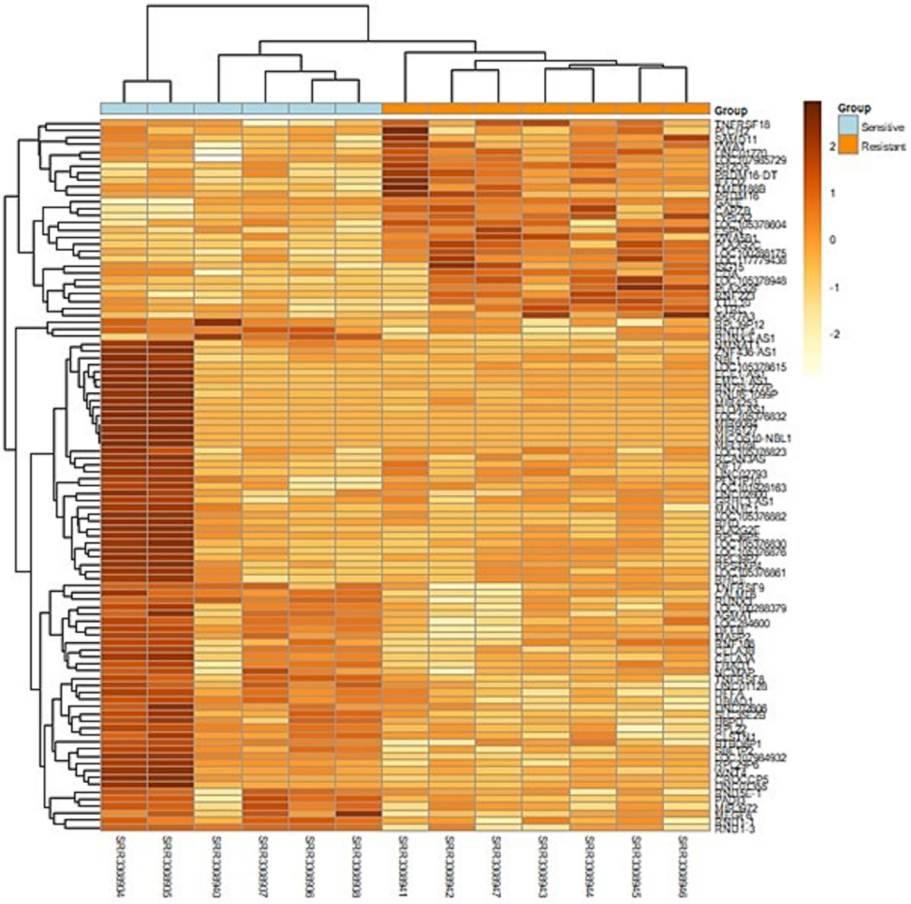

Fig. 2. The heatmap visually presents the differential expression of genes (DEGs) within the dataset. It effectively distinguishes between upregulated genes (with a log2 fold change of ≥ 2 and a significance level of $P < 0.05$), represented by a dark brown color, and downregulated genes (with a log2 fold change of ≤ -2 and $P < 0.05$), represented by a light orange color. To enhance the interpretation of the data, genes that share similar expression patterns have been clustered together using hierarchical clustering techniques.

we utilized the DESeq2 program for transcriptome analysis. DESeq2 generates various plots that enable us to visualize the gene expression patterns across different samples. Notably, the PCA plot serves as a visualization tool to capture the complexity of high-dimensional data (Lever et al., 2017). The PCA plot effectively demonstrates the clustering patterns among the samples, distinguished by distinct colors. Importantly, this plot reaffirmed the clear separation between normal and pancreatic cancer samples, validating the formation of distinct clusters (Son et al., 2018).

The normalized differences in expression patterns are used to compute a distance matrix with the help of PCA. In a PCA plot, the X-axis and Y-axis represent a mathematical modification of these distances that allows data to be shown in two dimensions that is PC1 versus PC2 with variances of 93% and 2% respectively as shown in Fig. 1. If log2fc>=2 and $P < 0.05$ indicates the upregulated genes and log2FC<=2 and $P < 0.05$ indicates downregulated genes represented by dark brown and light orange colour respectively in the heatmap as shown in (Fig. 2). The expression heatmap can be useful for determining how different all relevant genes expression in between sample groups, while the expression plot can be used to explore the expression levels between sample groups by looking at the top significant genes or selecting individ-

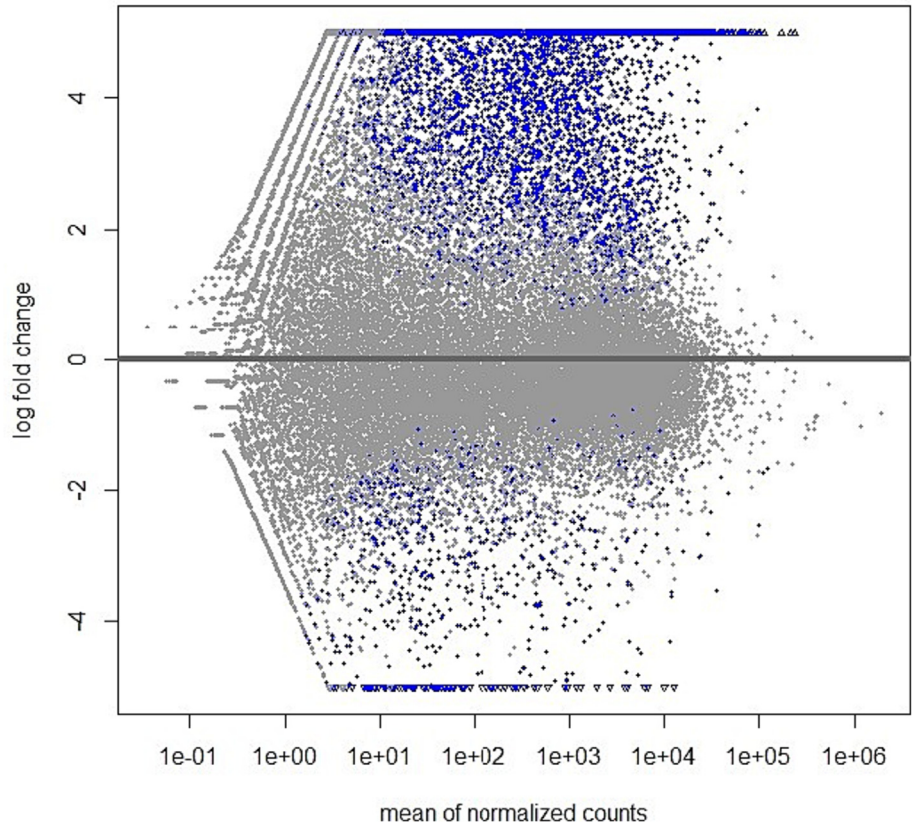

Fig. 3. MA plot describes the scattering of differentially expressed genes where blue data points falling above and below zero value indicates upregulation and down regulation of genes respectively, whereas grey data points show non-significant genes.

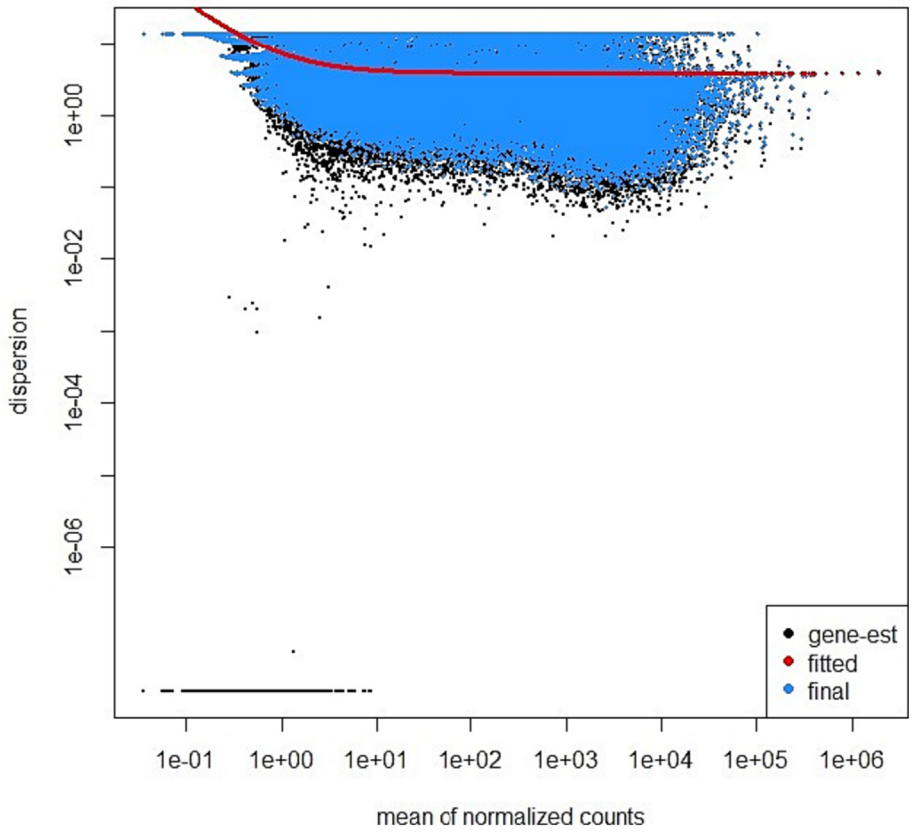

Fig. 4. Dispersion plot describes genes which are expressed based on p-value.

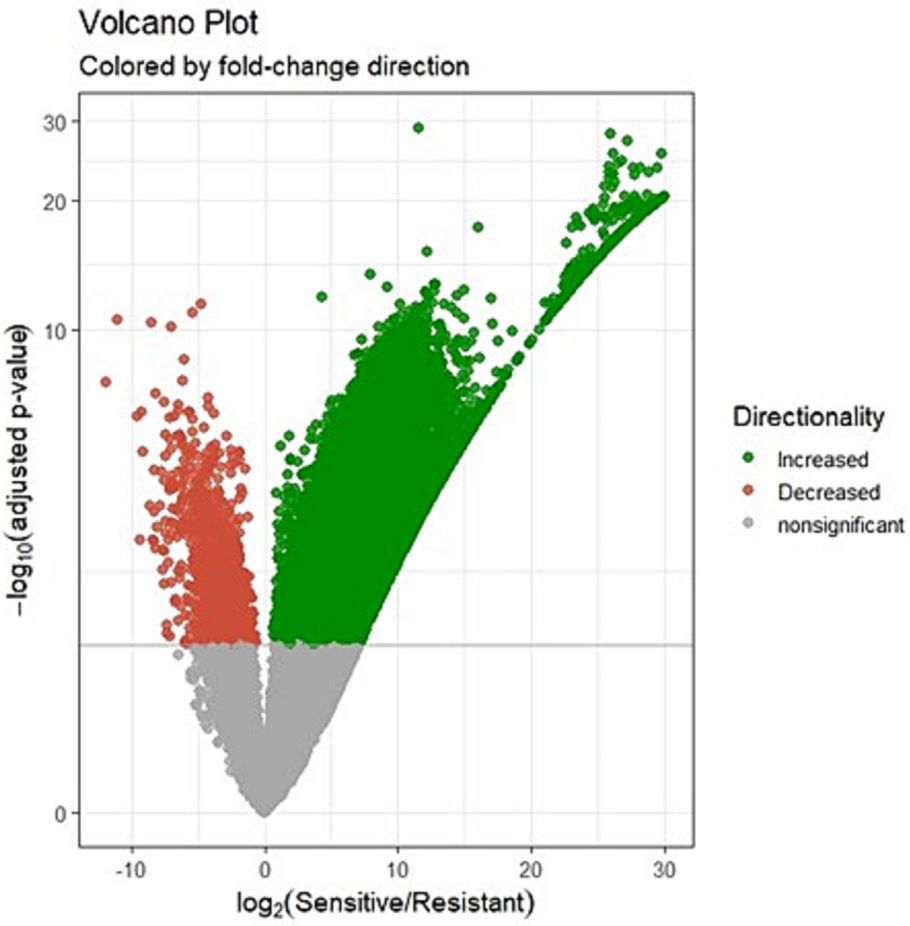

Fig. 5. Volcano Plot generated using a DESeq2 dataset, with base-10 log and base 2-fold change and P-value threshold of 0.05. In the plot the genes are colored if they pass thresholds for FDR and log fold change green indicates the upregulated genes and red colour indicates the downregulated genes, below the central line white colour indicates non-significant genes.

Table 2
Top 5 Upregulated genes as shown below table, identified the significantly differentially expressed genes using the parameters: FDR corrected P-value < 0.05 and fold change > 0 which shows the top five upregulated genes.

Genes	baseMean	log2FoldChange	lfcSE	stat	pvalue	Padj
LINC01128	1128.416	0.876969	0.361293	2.427307	0.015211	0.047632
SLC35E2B	16416.02	1.728116	0.555628	3.110204	0.00187	0.00746
CALML6	259.8582	1.926081	0.675873	2.84977	0.004375	0.015879
MEGF6	38658.97	2.175747	0.690112	3.152747	0.001617	0.006551
RPL22	5114.045	1.294715	0.400175	3.235368	0.001215	0.005069

Table 3
Top 5 Downregulated genes as shown below the table, Identified the significantly differentially expressed genes using the parameters: FDR corrected P-value < 0.05 and fold change < 0 which shows the top five downregulated genes.

Genes	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
SAMD11	200.163	−2.47936	0.910586	−2.72282	0.006473	0.022532
LOC100288175	310.4647	−1.77218	0.600263	−2.95234	0.003154	0.011855
RNF223	1387.049	−2.57033	1.030341	−2.49464	0.012609	0.040478
TTLL10-AS1	29.76435	−1.61312	0.638882	−2.52491	0.011573	0.037545
TTLL10	106.5688	−1.91649	0.611598	−3.13357	0.001727	0.006946

ual genes of interest. MA plots represent log fold-change value along the y-axis versus the mean expression value on the x-axis between the two conditions in the form of data points which was applied on a significantly adjusted P-value < 0.05 considered in (Fig. 3). The data points with above the zero thresholds indicate significant genes are upregulated and below zero indicates a high level of downregulated genes. Dispersion plot describes the expressed genes based on P-value; the plot indicates the data points shows variation of dispersion along with mean of normalized counts (Yoon and Nam, 2017) as shown in (Fig. 4). The dispersion smoothly decreases for genes with higher expression and eventually reaches an asymptote, which can be considered as the

Table 4
Most common repeated genes and its pathways.

Sl No	Genes	Number of occurences	Pathways
1	**CDKN1A**	10	E2F_TARGETS, P53_PATHWAY, PI3K_AKT_MTOR_SIGNALING, INTERFERON_GAMMA_RESPONSE, MTORC1_SIGNALING, APOPTOSIS, INFLAMMATORY_RESPONSE, TNFA_SIGNALING_VIA_NFKB, MYOGENESIS, HYPOXIA
2	**IL6**	10	EPITHELIAL_MESENCHYMAL_TRANSITION, IL6_JAK_STAT3_SIGNALING, ALLOGRAFT_REJECTION, APOPTOSIS, UV_RESPONSE_UP, INFLAMMATORY_RESPONSE, TNFA_SIGNALING_VIA_NFKB, COMPLEMENT, INTERFERON_GAMMA_RESPONSE, HYPOXIA
3	**MYC**	10	MYC_TARGETS_V2, WNT_BETA_CATENIN_SIGNALING, MYC_TARGETS_V1, UV_RESPONSE_DN, IL2_STAT5_SIGNALING, E2F_TARGETS, G2M_CHECKPOINT, ESTROGEN_RESPONSE_EARLY, TNFA_SIGNALING_VIA_NFKB, INFLAMMATORY_RESPONSE

biological variability that is present in the dataset (Love et al., 2014). We create a volcano plot at the conclusion of the DESeq2 programme to identify the genes that are up and down-regulated based on P-value and log fold-change. This pot shows the comparison between two distinct circumstances with a cluster of data-points (Yoon and Nam, 2017) Commonly using the negative base -10 log and base2 log fold change. The extreme values of the log fold-change along the x-axis show more significant differences, with data points closer to 0 denoting genes with equal or identical mean expression levels. A larger dispersion suggests that there is a greater difference in gene expression between the two group conditions as shown in (Fig. 5). Top 5 Upregulated significantly differentially expressed genes were identified using parameters: FDR corrected P-value < 0.05 and fold change > 0 and Top 5 Downregulated significantly expressed genes were identified using parameters: FDR corrected P-value < 0.05 and fold change < 0 using DESeq2 as shown in (Table 2 & 3) respectively.

Further we predict the pathways for the dataset using GSEA software with the help of hallmark symbols. Where we can get different pathways for the given dataset when we annotate the genes with condition to hallmark symbols. Next, we visualize the pathways using Enrichment Map software, In the enrichment map out of all 50 nodes, only 4 nodes are shared the common genes which are represented by edges in between them. Where gene set are represented as nodes and overlapped genes between them are represented by edges and colour indicates the expression level of genes in different pathways as shown in (Supplementary Fig. 1). Top 50 pathways were identified for the given data set based on the enrichment score with False Discovery Rate q-value <= 1 as shown in the (Table 4), then we observed the greatest number of occurrences of genes and its associated pathways as shown in (Supplementary Table 2). Next in our study is Gene network analysis based on the gene network that has generated from the data. Using StringApp, Cytoscape plugin for visualization and analysis of string network (Rambabu et al., 2017). The imported genes in this plugin start searching the STRING database and annotating the targeted gene network by representing the nodes as genes, and edges give the similar functional activity present between the genes (Otasek et al., 2019). The top three genes (*CDNK1A, IL6, MYC*) that are expressed in 10 different pathways were selected and incorporated into string network as shown in the (Fig. 6). A String network

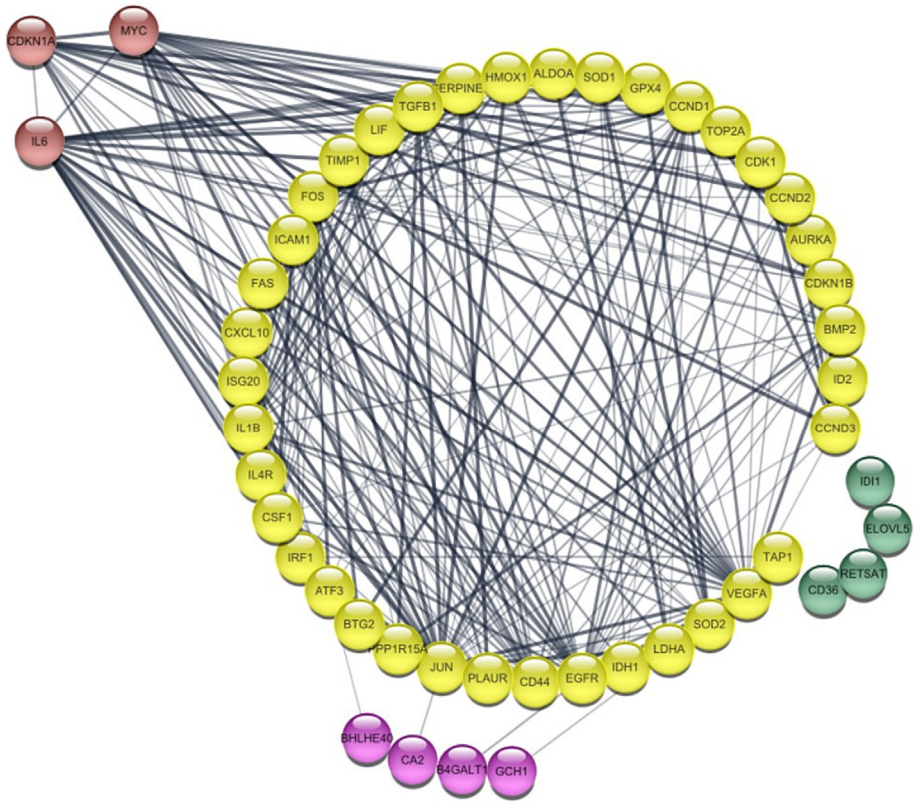

Fig. 6. A string network of the hub genes CDKN1A, MYC, IL6 that are associated with maximum (10) pathways.

was constructed using the top 50 highest expressed genes from the enriched pathways in the enrichment map analysis. Among these, CDKN1A, MYC, and IL6 stood out as hub genes, as they were found to be expressed in at least 10 different pathways. A combined String network was then established specifically for the most frequently recurring genes, namely CDKN1A, MYC, and IL6, and this network is visualized in Fig. 6. Additionally, a separate String network was created to focus exclusively on the interactions among CDKN1A, IL6, and MYC, showcasing the unique gene-gene interactions between these key genes. This separate network is presented in Supplementary Fig. 2. While these genes were identified as hub genes, it's worth noting that previous research has indeed associated CDKN1A, IL6, and MYC with the development and progression of pancreatic cancer. For instance, CDKN1A has been found to play a tumor-suppressive role in pancreatic cancer by inhibiting cell cycle progression and promoting cellular senescence(Xiao et al., 2020).

Loss of CDKN1A function has been associated with increased tumor growth and poorer patient outcomes. IL6 has also been implicated in pancreatic cancer, with studies suggesting that it promotes tumor growth and invasion through its pro-inflammatory and pro-angiogenic effects. Its oncogenic impact partly involves the significant role it plays in the tumor growth, particularly through its epigenetic silencing of CDKN1A (Lian et al., 2018). IL6 has also been shown to contribute to the development of chemoresistance in pancreatic cancer cells. MYC is often overexpressed in pancreatic cancer cells and has been associated with tumor progression and poor prognosis (Hessmann et al., 2016) Research has demonstrated that MYC plays a role in promoting tumor growth, invasion, and resistance to chemotherapy in pancreatic cancer cells. Overall, these three genes have important roles in pancreatic cancer development and progression, and targeting them may be a promising approach for the treatment of this disease.

5. Conclusion

In this current research study, we analyzed the differential expressed genes in pancreatic cancer samples of Sensitive and Resistant conditions. When compared with control tissues, our results showed significant differences in the expression of genes in pancreatic cancer samples. Creating a gene interaction network alongside its linked pathways can be a valuable approach for predicting novel disease biomarkers. This can potentially provide insights into the underlying molecular mechanisms involved in the transition from early-stage to metastatic pancreatic cancer. Such insights could be highly relevant and contribute to the identification of therapeutic targets, ultimately aiding in the development of effective treatments for this disease by comprehending its associated pathways. Based on our report the most common genes are i.e., *CDKN1A, IL6*, and *MYC*, and have a significant role in pancreatic cancer, which are expressed in ten pathways. TNFA_SIGNALING_VIA_NFKB is the most common pathway of expressed genes such as *CDKN1A, IL6*, and *MYC*. These genes can be used for further clinical processes to overcome better treatment methods and drug resistance.

CRediT authorship contribution statement

Majji Rambabu: Conceptualization, Methodology. **Nagaraj Konageni:** Conceptualization, Methodolgy. **Karthick Vasudevan:** Conceptualization, Methodology. **KR Dasegowda:** Visualization. **Anand Gokul:** Visualization. **Sivaraman Jayanthi:** Visualization. **Rohini Karunakaran:** Conceptualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors express deep gratitude to the management of REVA University and AIMST University for all the support, assistance, and consistent encouragement to carry out this work.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.sjbs.2023.103819.

References

Albrecht, S., Sprang, M., Andrade-Navarro, M.A., Fontaine, J.F., 2021. seqQscorer: automated quality control of next-generation sequencing data using machine learning. Genome Biol. 22, 1–20. https://doi.org/10.1186/s13059-021-02294-2.

Brown, J., Pirrung, M., Mccue, L.A., 2017. FQC Dashboard: Integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics 33, 3137–3139. https://doi.org/10.1093/bioinformatics/btx373.

Carroll, J.A., Race, B., Williams, K., Striebel, J., Chesebro, B., 2020. RNA-seq and network analysis reveal unique glial gene expression signatures during prion infection. Mol. Brain 13, 1–26. https://doi.org/10.1186/s13041-020-00610-8.

Casolino, R., Braconi, C., Malleo, G., Paiella, S., Bassi, C., Milella, M., Dreyer, S.B., Froeling, F.E.M., Chang, D.K., Biankin, A.V., Golan, T., 2021. Reshaping preoperative treatment of pancreatic cancer in the era of precision medicine. Ann. Oncol. 32, 183–196. https://doi.org/10.1016/j.annonc.2020.11.013.

Chen, F., Roberts, N.J., Klein, A.P., 2017. Inherited pancreatic cancer. Chin. Clin. Oncol., 6 https://doi.org/10.21037/cco.2017.12.04.

Corchete, L.A., Rojas, E.A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N.C., Burguillo, F.J., 2020. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. Sci. Rep. 10, 1–15. https://doi.org/10.1038/s41598-020-76881-x.

Croken, M.M., Qiu, W., White, M.W., Kim, K., 2014. Gene Set Enrichment Analysis (GSEA) of Toxoplasma gondii expression datasets links cell cycle progression and the bradyzoite developmental program. BMC Genomics 15, 1–13. https://doi.org/10.1186/1471-2164-15-515.

de Sena Brandine, G., Smith, A.D., 2019. Falco: high-speed FastQC emulation for quality control of sequencing data. F1000Res 8, 1874. https://doi.org/10.12688/f1000research.21142.1.

Danecek, Petr Bonfield, James K. Liddle, Jennifer Marshall, John Ohan, Valeriu Pollard, Martin O. Whitwham, Andrew Keane, Thomas McCarthy, Shane A. Davies, Robert M. Li, Heng, 2021. Twelve years of SAMtools and BCFtools. Gigascience 10 (2), 1–4. doi:10.1093/gigascience/giab008.

Dey, H., Vasudevan, K., Doss C, G. P., Kumar, S. U., El Allali, A., Alsamman, A. M., Zayed, H., 2023. Integrated gene network analysis sheds light on understanding the progression of Osteosarcoma. Frontiers in Medicine 10, 1154417. https://doi.org/10.3389/fmed.2023.1154417.

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

Felsenstein, M., Hruban, R.H., Wood, L.D., 2018. New developments in the molecular mechanisms of pancreatic tumorigenesis. Adv. Anat. Pathol. 25, 131–142. https://doi.org/10.1097/PAP.0000000000000172.

Fitts, B., Zhang, Z., Maher, M., Demchak, B., 2016. dot-app: a Graphviz-Cytoscape conversion plug-in. F1000Res 5, 2543. https://doi.org/10.12688/f1000research.9751.1.

Fraser, C.M., Rasko, D.A., Mahurkar, A., Hotopp, C.D., 2021. Host-Microbe Biology FADU : a Quanti fi cation Tool for Prokaryotic Transcriptomic Analyses 1–16.

Hessmann, E., Schneider, G., Ellenrieder, V., Siveke, J.T., 2016. MYC in pancreatic cancer: Novel mechanistic insights and their translation into therapeutic strategies. Oncogene. https://doi.org/10.1038/onc.2015.216.

Joly, J.H., Lowry, W.E., Graham, N.A., 2020. Differential Gene Set Enrichment Analysis: A statistical approach to quantify the relative enrichment of two gene sets. Bioinformatics 36, 5247–5254. https://doi.org/10.1093/bioinformatics/btaa658.

Kirby, M.K., Ramaker, R.C., Gertz, J., Davis, N.S., Johnston, B.E., Oliver, P.G., Sexton, K. C., Greeno, E.W., Christein, J.D., Heslin, M.J., Posey, J.A., Grizzle, W.E., Vickers, S. M., Buchsbaum, D.J., Cooper, S.J., Myers, R.M., 2016. RNA sequencing of pancreatic adenocarcinoma tumors yields novel expression patterns associated with long-term survival and reveals a role for ANGPTL4. Mol. Oncol. 10, 1169–1182. https://doi.org/10.1016/j.molonc.2016.05.004.

Lever, J., Krzywinski, M., Altman, N., 2017. Points of Significance: Principal component analysis. Nat. Methods. https://doi.org/10.1038/nmeth.4346.
Li, B., Dewey, C.N., 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. 12. https://doi.org/10.1186/1471-2105-12-323.
Lian, Y., Yang, J., Lian, Y., Xiao, C., Hu, X., Xu, H., 2018. DUXAP8, a pseudogene derived lncRNA, promotes growth of pancreatic carcinoma cells by epigenetically silencing CDKN1A and KLF2. Cancer Commun. 38. https://doi.org/10.1186/s40880-018-0333-9.
Liao, Y., Smyth, G.K., Shi, W., 2014. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930. https://doi.org/10.1093/bioinformatics/btt656.
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 1–21. https://doi.org/10.1186/s13059-014-0550-8.
Lu, S., Ahmed, T., Du, P., Wang, Y., 2017. Genomic variations in pancreatic cancer and potential opportunities for development of new approaches for diagnosis and treatment. Int. J. Mol. Sci. 18, 1–20. https://doi.org/10.3390/ijms18061201.
Merico, D., Isserlin, R., Bader, G.D., 2011. Visualizing gene-set enrichment results using the cytoscape plug-in enrichment map. Methods Mol. Biol. 781, 257–277. https://doi.org/10.1007/978-1-61779-276-2_12.
Michael, A., Mpimg, L., Anders, S., Huber, W., Heidelberg, E., Love, M.M., 2013. Package ' DESeq2 .'.
Mishra, S., Shah, M.I., Kumar, S.U., Kumar, D.T., Gopalakrishnan, C., Al-Subaie, A.M., Kamaraj, B., 2020. Network analysis of transcriptomics data for the prediction and prioritization of membrane-associated biomarkers for idiopathic pulmonary fibrosis (IPF) by bioinformatics approach. Adv. Protein Chem. Struct. Biol. 123, 241–273. https://doi.org/10.1016/bs.apcsb.2020.10.003.
Oliva, A., Tobler, R., Llamas, B., Souilmi, Y., 2021. Additional evaluations show that specific BWA-aln settings still outperform BWA-mem for ancient DNA data alignment. Ecol. Evol. 11, 18743–18748. https://doi.org/10.1002/ece3.8297.
Otasek, D., Morris, J.H., Bouças, J., Pico, A.R., Demchak, B., 2019. Cytoscape automation: Empowering workflow-based network analysis. Genome Biol. 20, 1–15. https://doi.org/10.1186/s13059-019-1758-4.
Pan, X., Ma, X., 2020. A novel six-gene signature for prognosis prediction in ovarian cancer. Front. Genet. 11, 1–14. https://doi.org/10.3389/fgene.2020.01006.
Rambabu, M., Dass, J.F.P., Jayanthi, S., 2017. Association of claudin family protein in human cancer types: a network approach, Int. J. Bioinformatics Research and Applications.
Rastrojo, A., Corvo, L., Lombraña, R., Solana, J.C., Aguado, B., Requena, J.M., 2019. Analysis by RNA-seq of transcriptomic changes elicited by heat shock in Leishmania major. Sci. Rep. 9, 1–18. https://doi.org/10.1038/s41598-019-43354-9.
Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., Merico, D., Bader, G.D., 2019a. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat. Protoc. 14, 482–517. https://doi.org/10.1038/s41596-018-0103-9.
Son, K., Yu, S., Shin, W., Han, K., Kang, K., 2018a. A simple guideline to assess the characteristics of RNA-Seq Data. Biomed Res. Int. 2018. https://doi.org/10.1155/2018/2906292.
Srivastava, A., Malik, L., Sarkar, H., Zakeri, M., Almodaresi, F., Soneson, C., Love, M.I., Kingsford, C., Patro, R., 2020. Alignment and mapping methodology influence transcript abundance estimation. Genome Biol. 21, 1–29. https://doi.org/10.1186/s13059-020-02151-8.
Stupnikov, A., McInerney, C.E., Savage, K.I., McIntosh, S.A., Emmert-Streib, F., Kennedy, R., Salto-Tellez, M., Prise, K.M., McArt, D.G., 2021. Robustness of differential gene expression analysis of RNA-seq. Comput. Struct. Biotechnol. J. 19, 3470–3481. https://doi.org/10.1016/j.csbj.2021.05.040.
Suárez-Fariñas, M., Lowes, M.A., Zaba, L.C., Krueger, J.G., 2010. Evaluation of the psoriasis transcriptome across different studies by Gene Set Enrichment Analysis (GSEA). PLoS One 5. https://doi.org/10.1371/journal.pone.0010247.
Xiao, B.-D., Zhao, Y.-J., Jia, X.-Y., Wu, J., Wang, Y.-G., Huang, F., 2020. Multifaceted p21 in carcinogenesis, stemness of tumor and tumor therapy. World J. Stem Cells 12, 481–487. https://doi.org/10.4252/wjsc.v12.i6.481.
Yoon, S., Nam, D., 2017. Gene dispersion is the key determinant of the read count bias in differential expression analysis of RNA-seq data. BMC Genomics 18. https://doi.org/10.1186/s12864-017-3809-0.
Zhang, F., Zhong, B.J., 2018. Image retrieval based on interested objects. Tien Tzu Hsueh Pao/Acta Electron. Sin. 46, 1915–1923. https://doi.org/10.3969/j.issn.0372-2112.2018.08.016.