

https://dspace.org/

Analysing RNA-sequence data for pancreatic ductal adenocarcinoma tissue samples to identify potential biomarkers

Jamal, Khadija Sanober

2023-09
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/41905
Jamal, Khadija Sanober. (2023). Analysing RNA-sequence data for pancreatic ductal adenocarcinoma tissue samples to identify potential biomarkers. [Master's dissertation, University of the Witwatersrand, Johannesburg]. https://hdl.handle.net/10539/41905

Downloaded from a DSpace repository. https://dspace.org/



UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Analysing RNA-Sequence Data for Pancreatic Ductal Adenocarcinoma Tissue Samples to Identify Potential Biomarkers

by

Khadija Sanober Jamal

(1601727)

Dissertation

Submitted in fulfilment of the requirements for the degree

Master of Science

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Professor Mandeep Kaur

04 September 2023

Declaration:

I, Khadija Sanober Jamal (1601727), am a student registered for the degree of Master of Science (Dissertation) in the academic year 2023.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the research submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise and acknowledged.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature:  Date: 04/09/2023

Copyright © 2023 University of The Witwatersrand

All rights reserved.

Abstract

Pancreatic ductal adenocarcinoma (PDAC) accounts for approximately 90% of pancreatic cancer and is the fourth leading cause of death with a five-year survival rate of less than 10%. Patients are asymptomatic until detection is observed at a metastatic stage, hence contributing massively towards the high mortality rate. This study was conducted to explore PDAC and its two main subtypes, the classical and basal-like subtype, in an in-depth level via bioinformatic analysis. Bioinformatics is a computational approach to evaluate biological data by analysing omics data including genomic expression and proteomic sequences. A workflow consisting of programmes and web-tools was used to analyse PDAC RNA-sequence data. The sample sets were grouped according to tumour, stage, and subtype. The workflow began with quality control using FastQC and Trimmomatic. Alignment of sequencing files and counts were done through HISAT2 and HTSeq. The main component of this workflow was differential gene expression analysis to identify differentially expressed genes (DEGs), statistically significant genes, per compared conditions. WGCNA was used for co-expression analysis to identify the hub genes involved in regulating the biological network. Lastly, *in-silico* validation was done by using available web tools to support the findings of this workflow. The identified tumour genes included *S100A11*, *PKM*, *GPRC5A,* LAMC2 and *ITGA2*, which may represent as universal biomarkers as sample extraction was performed from data generated from individuals belonging to 8 different countries. *KRT13* and *IL6* were identified in the advanced stage and their role in cancer progression have been explored in this current study. The basal-like subtype had *CAV1*, *DCVLD2* and *TGFB2* genes that contribute to treatment resistance. The common dysregulated genes in the basal-like subtype and advanced stage were analysed to evaluate the link between subtype and stage which included *WNT3A*, *TP63*, *KRT13* and *IGF2BP*. Co-expression analysis revealed hub genes for tumour (*KIF4A*, *SPAG5*, *RRM2* and *AURKA*), basal-like subtype (*BUB1*, *DEPDC1* and *KIF14*) and classical-subtype (*PTPRN* and *CAMK2B*). Through a machine learning model, recall, precision and accuracy scores per sample conditions for the DEGs were all above 94%. These potential biomarkers all have significant roles in promoting cancer progression, aggression and resistance. Hence, these may serve as a less invasive screening method for PDAC as DEGs were classified based on tissue or blood (extracellular vesicle) biomarkers. However, further wet laboratory validation is required for these biomarkers.

Acknowledgements

I would like to thank Prof Mandeep Kaur for providing guidance, mentorship and being an incredible supervisor during my degree. I would love to thank University of The Witwatersrand and School of Molecular and Cell Biology for giving me this wonderful opportunity. I thank all the members of the Integrative Cancer Biology Research Laboratory for their support and encouragement and the National Research foundation, South Africa (Bursary: 149649) for providing funding. I also acknowledge the research funding support through Self-Initiated Research grant provided to Prof M Kaur by the South African Medical Research Council, Cape Town. Final thanks to my dear family and friends for providing consistent support and love throughout this degree as well as my faith in God, Allah SWT, for blessing me with an immense amount of strength, courage and determination to complete this degree.

Study output

Poster Presentation:

Jamal KS and Kaur M. Analysing RNA-sequence Data for Pancreatic Ductal Adenocarcinoma to Identify Potential Biomarkers. Molecular Biosciences Research Thrust (MBRT). 01 December 2022.

Table of Contents

Declaration: ..... *I*

Abstract ..... *II*

Acknowledgements ..... *III*

Study output ..... *III*

List of Figures ..... *VI*

List of Tables ..... *XVI*

Nomenclature ..... *XVII*

1. Introduction *1*

1.1. Epidemiology of PDAC **1**

1.2. The pancreas **1**

1.3. Molecular analysis of PDAC and its development **2**

1.4. Molecular genetics of PDAC **4**

1.4.1. The *KRAS* signalling pathway 4

1.4.2. PI3K/AKT signalling pathway 4

1.4.3. SMAD4 and TGF-β signalling pathway 6

1.5. PDAC subtypes **7**

1.6. Stages of PDAC **9**

1.7. Cancer therapy treatment and resistance **11**

1.7.1. Gemcitabine 11

1.7.2. FOLFIRINOX 12

1.7.3. CCL2/CCR2 inhibitor (PF04136309) + FOLFIRINOX combination 12

1.8. Transcriptomic analysis **13**

1.9. Transcriptomics and PC Diagnosis **14**

1.10. Introduction to the present study **15**

2. Aim and objectives *16*

2.1. Aim **16**

2.2. Objectives **16**

3. Materials and Method *17*

3.1. Acquisition of RNA-Seq Data **18**

3.2. FASTQ file format **18**

3.3. Pre-processing of raw data: Quality control **19**

3.3.1. FastQC 19

3.3.2. Trimmomatic 21

3.4. HISAT2: Read alignment **22**
3.5. HTSeq count: Generating count data **24**
3.6. DESeq2: Differential gene expression (DE) analysis **24**
3.7. Machine learning **28**
3.8. WGCNA: Co-expression analysis **30**
3.9. ClusterProfiler: Functional enrichment process **33**
3.10. *In-silico* validation **33**
3.10.1. Kaplan Meier Survival analysis 33
3.10.2. Gene expression atlas 34
3.10.3. GEPIA 34
3.10.4. BBCancer 34
4. Results ***35***
4.1. FastQC analysis **35**
4.2. DESeq2 analysis **35**
4.3. WGCNA: Co-expression analysis of Dataset 2 ..... **57**
5. Discussion ***82***
5.1. DESeq2 analysis **83**
5.2. WGCNA co-expression analysis **90**
6. Conclusion ***98***
7. Future research ***101***
8. Limitations of the study ***102***
9. References ***103***
10. Appendices ***114***
10.1. Appendix A: Fastq Files **114**
10.2. Appendix B: Individual survival curves **136**

List of Figures

Figure 1 | A diagrammatic representation of PC metastasis. Steps 1 to 5 display the process towards cancer cell dissemination which involves detachment of malignant epithelial cells from the membrane. Following detachment, invasion of stroma and bloodstream occur (step 2-3). Step 4 shows dissemination of cancer cells (DCC), subsequently leading to invasion of nearby organs. During this process, there is a release of various pro-metastatic signals that promote an optimal environment to seed cancer growth and clonal amplification of stromal cells (step 8-9). Picture taken from Thomas et al., 2020 [4]. 2

Figure 2 | A pattern of genetic lesions that promote the progression towards PDAC from PanIN-1A. PanIN-3, a higher grade, consists of loss of function for genes that have tumour suppressor roles. Collectively, the dysregulation causes significant increase in cell proliferation, migration and invasion. The genes represented in green are activated throughout the progression process whereas the genes in red display loss of function. Image taken from Bardeesy et al., 2002 [13]. 3

Figure 3 | A diagrammatic representation of the PI3K/AKT signalling pathway. In the presence of a ligand, class 1 PI3Ks produce lipids that modulate multiple signalling pathways including protein kinases and small GTPases. There are four class I isoforms of PI3Ks in a mammalian system (p110s), consisting of a heterodimeric complex. Class I_A PI3K has three catalytic subunits, p110α, β and δ. Each subunit is bound to a Src Homology 2 domain that has a p85 regulatory region. Of these three p110s, p110α gene has a high mutational level when compared to β and δ, dominantly in different types of cancers. Due to it being the main mutated isoform, p110α massively contributes to the hyperactivation of class I/PI3K signalling pathway. Excessive activation of protein kinases and small GTPases results in an increase in cellular growth, translation, proliferation, survival and migration. Image taken from Baer et al., 2015 [19]. 5

Figure 4 | The SMAD4 and TGF-β signalling pathway with the presence of S100A11. TGF-β ligand binds to its transmembrane receptor. The mediated signals allow for the expression of Protein kinase C alpha (PKCα) to promote the phosphorylation of S100A11. Once phosphorylated, S100A11 translocate into the nucleus to compete and release Sp1/3 from its complex. Free Sp1/3 induces the expression of p21. Additionally, S100A11 forms a complex with SMAD2/3 and SMAD4 to activate p21 expression for cellular growth inhibition. Image taken from Zhang et al., 2021 [21]. 7

Figure 5 | The methodological steps for an RNA-seq experiment. Analysis begins from extraction of sample, purification, ligation, amplification and sequencing. Image taken from Sudhagar et al., 2018 [46]. 14

Figure 6 | A flowchart summary of the workflow used in this study on publicly available RNA sequence files. The workflow began with quality control using FastQC and Trimmomatic. Alignment was done through HISAT2 and a count matrix was produced from HTSeq-count. This matrix was analysed with DESeq2, WGCNA and a machine learning approach.

ClusterProfiler indicated the biological roles of the significant genes and lastly, the workflow concluded with in-silico analysis steps to identify potential biomarkers. 17

Figure 7 | A graphical representation of a good per base quality and per sequence GC content graphs. The example graphs show the expected FastQC quality control graphs. a) A good base per quality graph. The box and whisker plots remain in the green region between 28-40 Phred quality scores. If the box and whisker plots as in seen the red region, it would indicate bad quality scores. Therefore, incorrect base calls were performed during sequencing. b) A good per sequence GC content graph. The bell-like distribution peak is observed in the optimal range of 40% - 59% GC content. If the curve was shifted below 40% or above 60%, this would display a low coverage rate due to low clonal amplification during sequencing. Image taken from [52]. 20

Figure 8 | The format of a count matrix used in DESeq analysis. Columns consisted of sample numbers whereas rows showed the gene name/number. The matrix contained count values of each gene per sample. 27

Figure 9 | A summary procedure for WGCNA. Following network production, modules are detected and viewing is done using Cytoscape. 31

Figure 10 | A PCA for the RNA-sequence counts for each dataset. Dataset 1: N = non tumorous and P = PDAC. a) prefilter PCA, P = 15 and N = 8, with outliers seen in region PCA1 < 0 and PCA2 < -20 and b) post filter PCA plot, P = 10 and A = 6. Dataset 2 (1): A = adjacent normal tissue (39) and P = PDAC (140). Outliers in regions PCA2 >50 were removed. b) post filtering PCA plot, P = 116 and N = 29. Dataset 2 (2) C = classical subtype and L = Basal-like subtype. Dataset 2 (3) E = early stage and L = late stage. Plots produced by R-Studio DESeq2 Bioconductor package 3.13. 36

Figure 11| Graph showing the number of upregulated and downregulated DEGs (y-axis) for each dataset (x-axis). Green bars represent the upregulated genes whereas orange bars are the downregulated genes. Dataset 2 produced no DEGs. Plot produced by Python Google collab. 37

Figure 12 | Venn diagrams showing the common upregulated tumour DEGs observed in both dataset 1 and 3 for the same comparison condition. a) common upregulated DEGs (587) and b) common downregulated DEGs (255). Diagram produced by Python Google collab. 38

Figure 13 | The highly significant upregulated tumour DEGs observed dataset 1 and 2 for the tumour samples and the corresponding log_2FC. Plots was produced by R-Studio DESeq2 Bioconductor package 3.13. 39

Figure 14 | Enrichment for all of the upregulated significant common DEGs in the tumour sample mapped against the gene ratio, displaying the percentage of DEGs in the given GO term. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set per GO term. The dot colour shows the statistical significance of gene whereby red is highly significant and blue is significant. a) An enrichGO classification for the cellular

component profiles, b) An enrichGO molecular function, c) An enrichGO biological process and d) An enrich KEGG functional process for the DEGs. Plots produced by R ClusterProfiler version 4.5.1.....41

Figure 15 | Average Kaplan-Meier survival curve for the common upregulated tumour DEGs, following the removal of outliers and insignificant P-value graphs. Graph show the probability of survival versus the time (months). HR serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Individual DEG graphs can be seen in Appendix C. Graph produced by Kaplan-Meier plotter version 2022.08.23.....42

Figure 16 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated tumour DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. There was no expression level for: SOX11 and SLC16A3. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home). 43

Figure 17 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (*Fig.* 16) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).....44

Figure 18 | The common highly significant upregulated genes observed in the late stage of PDAC. The genes are mapped with the corresponding log2FC. Plots was produced by R-Studio DESeq2 Bioconductor package 3.13. 45

Figure 19 | Enrichment for all of the upregulated significant DEGs in the late stage of PDAC samples mapped against the gene ratio, displaying the percentage of DEGs in the given GO term. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set per GO term. The dot colour shows the statistical significance of gene whereby red is highly significant and blue is significant. a) An enrichGO classification for the cellular component profiles, b) An enrichGO molecular function, c) An enrichGO biological process and d) An enrich KEGG functional process for the DEGs. Plots produced by R ClusterProfiler version 4.5.1.....48

Figure 20 | Average Kaplan-Meier survival curve for the upregulated late stage DEGs, following the removal of outliers and insignificant P-value graphs. Graph show the probability of survival versus the time (months). HR serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows

the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Individual DEG graphs can be seen in Appendix C. Graph produced by Kaplan-Meier plotter version 2022.08.23.....48

Figure 21 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated late stage DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).....49

Figure 22 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (Fig. 21) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).....50

Figure 23 | The highly common significant upregulated genes observed in the basal-like subtype of PDAC. The genes are mapped with the corresponding log2FC. Plots was produced by R-Studio DESeq2 Bioconductor package 3.13.....51

Figure 24 | Enrichment for all of the upregulated significant DEGs in the basal-like subtype of PDAC samples mapped against the gene ratio, displaying the percentage of DEGs in the given GO term. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set per GO term. The dot colour shows the statistical significance of gene whereby red is highly significant and blue is significant. a) An enrichGO classification for the cellular component profiles, b) An enrichGO molecular function, c) An enrichGO biological process and d) An enrich KEGG functional process for the DEGs. Plots produced by R ClusterProfiler version 4.5.1.....54

Figure 25 | Average Kaplan-Meier survival curve for the upregulated basal-like subtype DEGs, following the removal of outliers and insignificant P-value graphs. Graph show the probability of survival versus the time (months). HR serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Individual DEG graphs can be seen in Appendix C. Graph produced by Kaplan-Meier plotter version 2022.08.23.....54

Figure 26 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated basal-like DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).....55

Figure 27 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (Fig. 26) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).....56

Figure 28 | WGCNA analysis for the upregulated tumour samples containing DEGs. a) Network topology for the scale independence and mean connectivity analysis of modules consisting of various soft thresholding powers ranging between the values of 1 to 20. The left panel shows the scale-free topology fit (y-axis) as a function of the soft thresholding power (x-axis). The right panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). The power of 3 was selected for downstream analysis as it is the lowest power for which the scale-free topology fit index reaches 0.90. b) Clustering dendrogram of genes among modules via hierarchical clustering of adjacency-based dissimilarity. Plots produced by R WGCNA version 1.71. 57

Figure 29 | A Cytoscape network view of the selected co-expressed gene cluster modules seen in the up-regulated tumour samples, represented in nodes, following the cut off threshold value of 0.91. SPAG5 had the highest connection with 5 nodes. Genes with a connection of less than 2 were discarded. Network produced by Cytoscape version 3.9.1. 58

Figure 30 | GO and KEGG enrichment for all of the upregulated tumour subtype gene clusters. a) A compareCluster biological theme comparison for the enriched GO functional cellular component profiles, displaying functional consensus and differences observed in each upregulated tumour gene cluster. b) A compareCluster biological theme comparison for the enriched GO functional molecular function profiles. c) A compareCluster biological theme comparison for the enriched GO functional biological process profiles. d) A compareCluster biological theme comparison for the enriched KEGG functional process profiles. The x-axis represents the gene clusters whereas the y-axis shows the appropriate labels. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set. The dot colours display the statistical significance whereby red is highly significant and blue is significant. Plot produced by R ClusterProfiler version 4.5.1. 60

Figure 31 | Average Kaplan-Meier survival curve for the average tumour upregulated DEGs selected based on the co-expressed network following the removal of outliers and insignificant P-value graphs. Individual survival curves can be seen in Appendix C. Graph show the probability of survival versus the time (months). HR serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by a drug study that is compared to the hazard reduction for the control of the study. Logrank P values represents the P-values for each plot, showing the statistical significance. The black line in each plot shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graphs produced by Kaplan-Meier plotter version 2022.08.23. 61

Figure 32 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated tumour co-expressed DEGs with significant P-value survival curves.

The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).....62

Figure 33 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (Fig. 33) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).....62

Figure 34 | WGCNA analysis for the upregulated normal samples containing DEGs. a) Network topology for the scale independence and mean connectivity analysis of modules consisting of various soft thresholding powers ranging between the values of 1 to 20. The left panel shows the scale-free topology fit (y-axis) as a function of the soft thresholding power (x-axis). The right panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). The power of 7 was selected for downstream analysis as it is the lowest power for which the scale-free topology fit index reaches 0.90. b) Clustering dendrogram of genes among modules via hierarchical clustering of adjacency-based dissimilarity. Plots produced by R WGCNA version 1.71. 64

Figure 35 | A Cytoscape network view of the selected co-expressed gene cluster modules seen in the up-regulated normal samples, represented in nodes, following the cut off threshold value of 0.99. Network produced by Cytoscape version 3.9.1. 65

Figure 36 | GO and KEGG enrichment for all of the upregulated normal subtype gene clusters. a) A compareCluster biological theme comparison for the enriched GO functional cellular component profiles, displaying functional consensus and differences observed in each upregulated normal gene cluster. b) A compareCluster biological theme comparison for the enriched GO functional molecular function profiles. c) A compareCluster biological theme comparison for the enriched GO functional biological process profiles. d) A compareCluster biological theme comparison for the enriched KEGG functional process profiles. The x-axis represents the gene clusters whereas the y-axis shows the appropriate labels. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set. The dot colours display the statistical significance whereby red is highly significant and blue is significant. Plot produced by R ClusterProfiler version 4.5.1. 66

Figure 37 | Average Kaplan-Meier survival curves for the normal upregulated DEGs selected based on the co-expressed network. Individual survival curves can be seen in Appendix B, Fig. 58. Graph depicts the probability of survival versus the time (months). HR is a measurement of the magnitude for the differences between the two curves in each plot. It shows the hazard reduction obtained by a drug study which is compared to the hazard reduction for the control of the study. Logrank P values represents the P-values for each plot, showing the statistical significance. The black line displays the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graphs produced by Kaplan-Meier plotter version 2022.08.23. 67

Figure 38 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated normal co-expressed DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. It can be observed that the expression level ranges from 0 to only 6 TPM values which is due it being upregulated in the normal samples and not in PC cell lines. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home). 68

Figure 39 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (Fig. 38) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).....68

Figure 40 | WGCNA analysis for the upregulated basal-like subtype samples containing differentially expressed genes. a) Network topology for the scale independence and mean connectivity analysis of modules consisting of various soft thresholding powers ranging between the values of 1 to 20. The left panel shows the scale-free topology fit (y-axis) as a function of the soft thresholding power (x-axis). The right panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). The power of 6 was selected for downstream analysis as it is the lowest power for which the scale-free topology fit index reaches close to 0.90. b) Cluster dendrogram of genes among modules via hierarchical clustering of adjacency-based dissimilarity. Plots produced by R WGCNA version 1.71.....70

Figure 41 | A Cytoscape network view of the selected co-expressed gene cluster modules seen in the up-regulated basal-like subtype samples, represented in nodes, following the cut off threshold value of 0.95. Network produced by Cytoscape version 3.9.1. 70

Figure 42 | GO and KEGG enrichment for all of the upregulated basal-like subtype gene clusters. a) A compareCluster biological theme comparison for the enriched GO functional cellular component profiles, displaying functional consensus and differences observed in each upregulated basal-like subtype gene cluster. b) A compareCluster biological theme comparison for the enriched GO functional molecular function profiles. c) A compareCluster biological theme comparison for the enriched GO functional biological process profiles. d) A compareCluster biological theme comparison for the enriched KEGG functional process profiles. The x-axis represents the gene clusters whereas the y-axis shows the appropriate labels. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set. The dot colours display the statistical significance whereby red is highly significant and blue is significant. Plot produced by R ClusterProfiler version 4.5.1.....72

Figure 43 | Average Kaplan-Meier survival curves for the basal-like upregulated DEGs selected based on the co-expressed network. Individual survival curves can be seen in Appendix C. Graphs show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by a drug study that is compared to the hazard reduction for the control of the study. Logrank P values represents the P-values for each

plot, showing the statistical significance. The black line in each plot shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graphs produced by Kaplan-Meier plotter version 2022.08.23.....73

Figure 44 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated basal-like co-expressed DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).....74

Figure 45 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (Fig. 44) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).....74

Figure 46 | WGCNA analysis for the up-regulated classical subtype samples containing differentially expressed genes. a) Network topology for the scale independence and mean connectivity analysis of modules consisting of various soft thresholding powers ranging between the values of 1 to 20. The left panel shows the scale-free topology fit (y-axis) as a function of the soft thresholding power (x-axis). The right panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). The power of 7 was selected for downstream analysis as it is the lowest power for which the scale-free topology fit index reaches close to 0.90. b) Cluster dendrogram of genes among modules via hierarchical clustering of adjacency-based dissimilarity. Plots produced by R WGCNA version 1.71.....76

Figure 47 | A Cytoscape network view of the selected co-expressed gene cluster modules seen in the upregulated classical subtype samples, represented in nodes, following the cut off threshold value of 0.975. Network produced by Cytoscape version 3.9.1. 76

Figure 48 | GO and KEGG enrichment for all of the upregulated classical subtype gene clusters. a) A compareCluster biological theme comparison for the enriched GO functional cellular component profiles, displaying functional consensus and differences observed in each upregulated classical subtype gene cluster. b) A compareCluster biological theme comparison for the enriched GO functional molecular function profiles. c) A compareCluster biological theme comparison for the enriched GO functional biological process profiles. d) A compareCluster biological theme comparison for the enriched KEGG functional process profiles. The x-axis represents the gene clusters whereas the y-axis shows the appropriate labels. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set. The dot colours display the statistical significance whereby red is highly significant and blue is significant. Plot produced by R ClusterProfiler version 4.5.1.....78

Figure 49 | Average Kaplan-Meier survival curves for the classical up-regulated DEGs selected based on the co-expressed network. Individual survival curves can be seen in Appendix C. Graphs show the probability of survival versus the time (months). The hazard

ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by a drug study that is compared to the hazard reduction for the control of the study. Logrank P values represents the P-values for each plot, showing the statistical significance. The black line in each plot shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graphs produced by Kaplan-Meier plotter version 2022.08.23.....79

Figure 50 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated classical subtype co-expressed DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).....80

Figure 51 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (Fig. 50) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).....81

Figure 52 | Untrimmed FastQC files for dataset 1. The box and whisker plots show the quality scores across all bases per position in read (bp) and score value (y-axis). The graph has three colour coded regions, corresponding to the Y-axis scores. A good base call will have a high score value which is placed in the green region going towards the yellow region. A bad base call will have a lower value in the red region. It can be seen that each untrimmed Fastq file had a decrease in base call score towards the end of the cycle.. Graph produced by FastQC (command line and Python).....125

Figure 53 | FastQC files for trimmed samples used in dataset 1. The box and whisker plots show the quality scores across all bases per position in read (bp) and score value (y-axis). The graph has three colour coded regions, corresponding to the Y-axis scores. A good base call will have a high score value which is placed in the green region going towards the yellow region. A bad base call will have a lower value in the red region. All trimmed sample files were used for downstream analysis as each sample produced a satisfactory quality control plot. Graph produced by FastQC (command line and Python).....136

Figure 54 | Kaplan-Meier survival curves for each common upregulated tumour DEGs. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEGs with insignificant P values were discarded: SLC411 and ULBP2. KM plotter did not produce a plot for: BLACAT1. Graph produced by Kaplan-Meier plotter version 2022.08.23. 138

Figure 55 | Kaplan-Meier survival curves for each upregulated advanced stage DEGs. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEGs with insignificant P values were discarded: CLCA2, PPBP, ERMN, FGF11, ITGB3, LOC101927354, KCNQ5, NDUFA4L2, SENA3D and CDH13. KM plotter did not produce a plot for: LINC01468. Graph produced by Kaplan-Meier plotter version 2022.08.23.....139

Figure 56 | Kaplan-Meier survival curves for each upregulated basal-like subtype DEGs. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEG with an insignificant P value was discarded: CLCA2. KM plotter did not produce a plot for: MIG7 and MIR205HG. Graph produced by Kaplan-Meier plotter version 2022.08.23. 141

Figure 57 | Kaplan-Meier survival curves for each tumour upregulated DEGs selected based on the co-expressed network. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graph produced by Kaplan-Meier plotter version 2022.08.23. 142

Figure 58 | Kaplan-Meier survival curves for each normal upregulated DEGs selected based on the co-expressed network. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEG with an insignificant P value was discarded: SYCN. Graph produced by Kaplan-Meier plotter version 2022.08.23. 143

Figure 59 | Kaplan-Meier survival curves for each Basal-like subtype upregulated DEGs selected based on the co-expressed network. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by

drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEG with an insignificant P value was discarded: USH1G. Graph produced by Kaplan-Meier plotter version 2022.08.23.....144

Figure 60 | Kaplan-Meier survival curves for each classical subtype upregulated DEGs selected based on the co-expressed network. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEGs with insignificant P values were discarded: CPB1 and CELA2B. Graph produced by Kaplan-Meier plotter version 2022.08.23.....144

List of Tables

Table 1 | The molecular subtypes of PC based on Raphael and Moffit classification and methodology analysis.....8

Table 2 | Cancer derived cells and signalling pathways that dominate the paracrine and autocrine signals, its functions and involvement in PC metastasis.....10

Table 3 | The BBCancer and GENT2 expression values (TPM) for the selected tumour DEGs. 44

Table 4 | The BBCancer and GENT2 expression values (TPM) for the selected late stage DEGs.....50

Table 5 | The BBCancer and GENT2 expression values (TPM) for the selected basal-like DEGs.....56

Table 6 | The BBCancer and GENT2 expression values (TPM) for the selected tumour co-expressed DEGs. 63

Table 7 | The BBCancer and GENT2 expression values (TPM) for the selected normal co-expressed DEGs. 69

Table 8 | The BBCancer and GENT2 expression values (TPM) for the selected basal-like co-expressed DEGs. 75

Table 9 | The BBCancer and GENT2 expression values (TPM) for the selected classical subtype co-expressed DEGs. 81

Nomenclature

5-FU	Fluorouracil
ABC	ATP-Binding Cassette
ADM	Acinar-to-Ductal Metaplasia
AJCC	American Joint Committee on Cancer
APC/C	Anaphase Promoting Complex/Cyclosome
AURKA	Aurora-A Kinase
BAM	Binary Alignment/Map
BH	Benjamini-Hochberg
BUB1	Budding Uninhibited by Benzimidazole 1
CAFs	Cancer-Associated Fibroblasts
CAMK2B	Calcium/Calmodulin Dependent Protein Kinase II Beta
CAV1	Caveolin-1
CCL2	Monocyte chemoattractant protein-1 MCP-1
cDNA	Complementary DNA
CSC	Cancer Stem Cell Signalling
DCBLD2	Discoidin, CUB And LCCL Domain Containing 2
DCC	Dissemination of Cancer Cells
Dck	Deoxycytidinekinase
DEG	Differentially Expressed Gene
DEPDC1	Domain Containing 1
ECM	Extracellular Matrix
EGF	Epidermal Growth Factor
EMBL	European Molecular Biology Laboratory
EMT	Epithelial to Mesenchymal Transition
ENA	European Nucleotide Archives
EVs	Extracellular Vesicles
FdUM	Fluorodeoxyuridine monophosphate
FM	Ferragina Manzini
GDC	National Cancer Institute Genomic Data Commons
GDP	Guanosine Triphosphate
GEM	Gemcitabine, 2′ ,2′- Difluorodeoxycytidine, Dfdc
GENT2	Gene Expression Database of Normal and Tumour Tissues 2

GEPIA	Gene Expression Profiling Interactive Analysis
GLM	Generalized Linear Models
GO	Gene Ontology
GPRC5A	G Protein-Coupled Receptor Class C Group 5 Member
GTP	Guanosine Diphosphate
GTPases	Guanosine Triphosphatases
HIF-1α	Hypoxia-inducible Factor
HISAT2	Hierarchical Indexing for Spliced Alignment of Transcripts
HR	Hazard Ratio
HT	High-Throughput
HTML	HyperText Markup Language
IGF2BP	Insulin-Like Growth Factor-2 mRNA-Binding Proteins
ITGA2	Integrin Subunit Alpha 2
JAK	Janus Kinases
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KIF14	Kinesin Family Member 14
KIF4A	Kinesin Family Member 4A
KM	Kaplan-Meier
LAMB	Laminin Subunit Beta 3
LAMC2	Laminin Subunit Gamma 2
lncRNA	Long Non-Coding RNA
Log_2FC	Log_2 Fold Change
MDR	Multidrug Resistance
miRNA	Micro-RNA
MMP2	Matrix Metalloproteinase-2
mRNA	Messenger RNA
MYT1L	Myelin Transcription Factor 1
NGS	Next Generation Sequencing
P63	Tumour Protein 63
PanINs	Pancreatic Intraepithelial Neoplasia
PC	Pancreatic Cancer
PCA	Principal Component Analysis
PDAC	Pancreatic Ductal Adenocarcinoma
PI3K/ AKT	Phosphatidylinositol 3-Kinase /Protein Kinase B

PKMs	Pyruvate Kinases
PTKs	Protein Tyrosine Kinases
PTPRN	Protein Tyrosine Phosphatase Receptor Type N
RAM	Random Access Memory
RAS/MAPK	Rat Sarcoma Viral Oncogene Family/Mitogen-Activated Protein Kinase
RFC	Random Forest Classifier
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
RRM2	Ribonucleotide Reductase Subunits M1
rRNA	Ribo-Depletion
RT-PCR	Reverse Transcription Polymerase Chain Reaction
S100A1	Calcium Binding Protein
SAM	Sequence Alignment/Map Format
SCG3	Secretogranin III
SE	Single End
Shh	Sonic hedgehog
siRNA	Small Interfering RNA
SMAD4	Mothers Against Decapentaplegic Homolog 4
SPAG5	Sperm-Associated Antigen 5
TCGA	The Cancer Genome Atlas
TGF-β	Transforming Growth Factor β
TGFβ2	Transforming Growth Factor Beta 2
TME	Tumour Microenvironment
TNM	Tumour, Node and Distant Metastasis
TOM	Topological Overlap Matrix
TPM	Transcripts Per Million
USCS	University of California Santa Cruz
VST	Variancestabilizingtransformation
WGCNA	Weighted Gene Co-Expression Network Analysis
Wnts	Wingless-Related Integration S

1. Introduction

1.1. Epidemiology of PDAC

Cancer is a disease that is responsible for the highest mortality rate with nearly 10 million deaths in 2020 [1]. Cancer cells acquire unique capabilities that allow for the transformation to a neoplastic growth state from the normalcy. These capabilities are crucial factors towards the formation of malignant tumours. There are commonalties between the types of cancers, however, each cancer has a distinctive form based on the type of organs and cells involved. Cancer has been defined as an unregulated growth of cells with excessive division via bypassing the checkpoints in the cell cycle [2]. The irregular functioning of cellular signalling pathways contributes to a metastatic state. Pancreatic cancer (PC) is one of the most lethal malignancies with an overall survival rate of less than 11% [3]. Metastatic patients have a 5-year survival rate of less than 3% [4]. In 2019, approximately 450, 000 new cases were reported and over 430, 000 deaths caused by PC alone, highlighting a close relationship between the incidence and mortality rate [5]. According to the National Cancer Registry South Africa, there was a 0.53% and 0.58% incidence rate in females and males, respectively [6].

There are various types of PC, however, pancreatic ductal adenocarcinoma (PDAC) accounts for approximately 90% of pancreatic cancers [7]. PDAC is the fourth leading cause of death with a five-year survival rate of less than 10% [7]. A vital factor towards a high mortality rate is that patients are asymptomatic until detection at an advanced stage whereby distant metastasis has taken place [8].

1.2. The pancreas

The pancreas is a multifunctional organ that contains two distinct cell populations, namely the exocrine and endocrine cells. The exocrine portion, compromising of 85% of the pancreas, secretes zymogens (digestive enzymes) to promote digestion and absorption in the gastrointestinal tract. The endocrine portion, located in the Islet of Langerhans cells, secretes insulin and glucagon into the bloodstream [9]. Acinar cells act as a functional unit of the exocrine portion that holds various responsibilities including storage, regulation and the production of enzymes via an increase in intercellular calcium [10]. However, acinar cells have high degree of plasticity. During environmental stress, these cells undergo a process known as acinar-to-ductal metaplasia (ADM). This process enables trans-differentiation of acinar cells

into epithelial-like cells, rendering them more susceptible to proto-oncogenic hits. This causes mutations in proto-oncogenes such as *KRAS* and therefore leading towards the formation of PDAC [11].

1.3. Molecular analysis of PDAC and its development

The exact cause of PC has yet to be clearly defined, however, the causative factors that promote the development and complexity of this disease include environmental, dietary, metabolic, genetic predisposition and epigenetic factors [12]. Additionally, a process known as epithelial to mesenchymal transition (EMT) is seen microscopically in an early stage of PDAC. *Fig. 1* shows the type of cells involved in the development of a pro-metastatic niche.

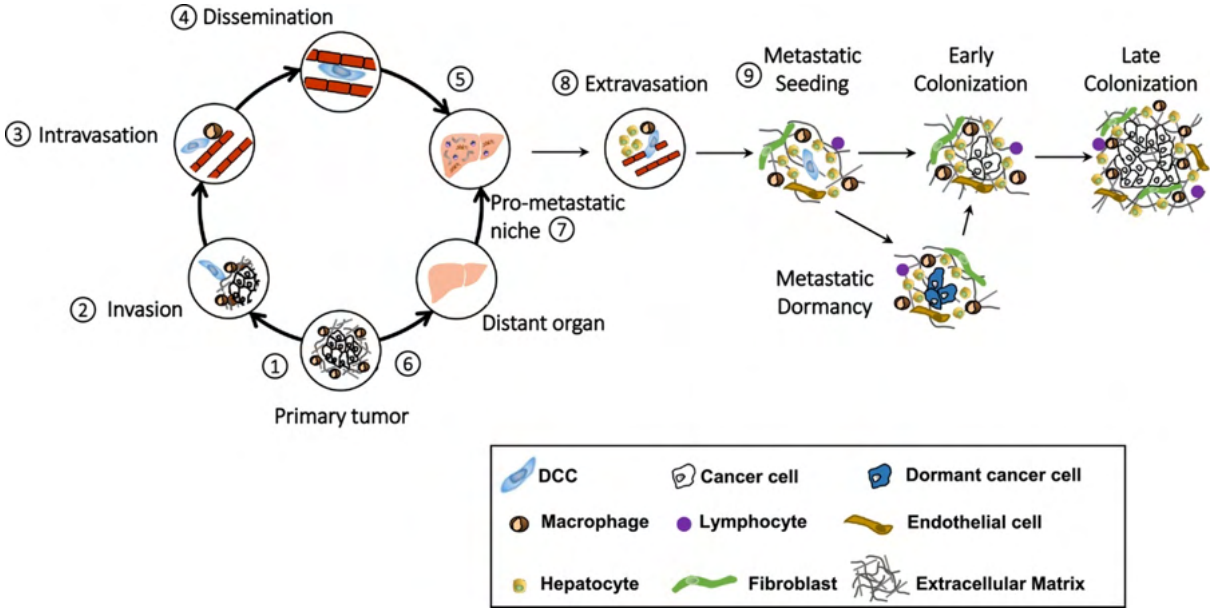


Figure 1 | A diagrammatic representation of PC metastasis. Steps 1 to 5 display the process towards cancer cell dissemination which involves detachment of malignant epithelial cells from the membrane. Following detachment, invasion of stroma and bloodstream occur (step 2-3). Step 4 shows dissemination of cancer cells (DCC), subsequently leading to invasion of nearby organs. During this process, there is a release of various pro-metastatic signals that promote an optimal environment to seed cancer growth and clonal amplification of stromal cells (step 8-9). Picture taken from Thomas *et al*., 2020 [4].

The development of PDAC has been shown to occur through a pattern of genetic lesions. These lesions include signature cancer causing genes such as breast cancer gene 1/2 (*BRCA1/2*), tumour protein p53 gene (*TP53),* Kirsten rat sarcoma viral oncogene homolog (*KRAS*), cyclin-dependent kinase inhibitor 2A (*CDKN2A*) and mothers against decapentaplegic homolog 4/

Deleted in Pancreatic Cancer-4 (*SMAD4/DPC4*) [13]. Mutations in these genes coincidentally correlate to the stage of disease. It has been accepted that PDAC arises via a series of precursor lesions called pancreatic intraepithelial neoplasia (PanINs) and intraductal papillary mucinous neoplasms, showing that genetic mutations have a carcinogenic role when combined towards the formation of PDAC. PanINs, the progenitors of PDAC, are abnormal ducts that are formed through the activation or loss of function for certain genes. The level of PanINs depends on the degree of cellular and nuclear atypia, as seen in *Fig. 2* [14]. A normal duct displays low cuboidal cells and a single cell layer. PanIN-1A consists of elongated cells and produce mucin with an activation of *KRAS* whereas PanIN-1B show papillary architecture with no nuclear atypia and loss of function for *INK4A*. PanIN-2 is characterized by low-grade dysplasia with nuclear abnormalities with loss of function for *TP53* gene[13]. PanIN-3 has high-grade dysplasia due to budding into the lumen and severe nuclear atypia. This linear progression leads to the formation of PDAC with invasive growth and an increase in stromal heterogeneity [14]. During this formation, there is a continuous activation of *KRAS* and *ERBB2/EGFR* genes.

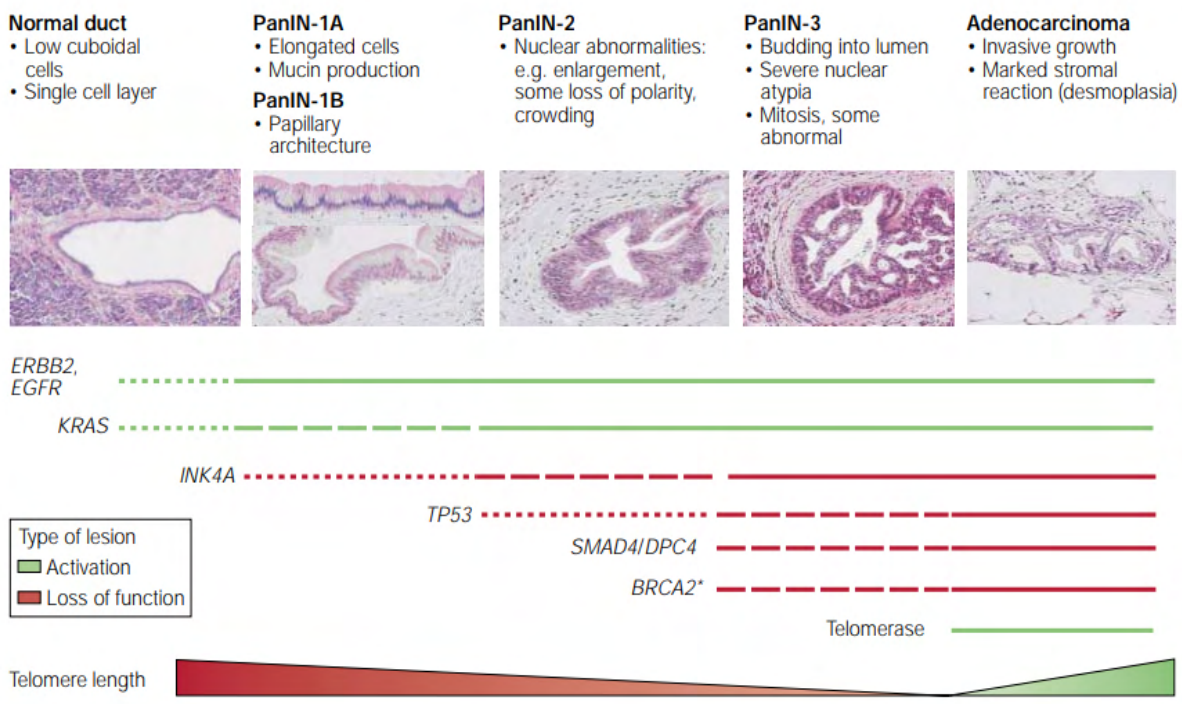


Figure 2 | A pattern of genetic lesions that promote the progression towards PDAC from PanIN-1A. PanIN-3, a higher grade, consists of loss of function for genes that have tumour suppressor roles. Collectively, the dysregulation causes significant increase in cell proliferation, migration and invasion. The genes represented in green are activated throughout the progression process whereas the genes in red display loss of function. Image taken from Bardeesy *et al*., 2002 [13].

1.4. Molecular genetics of PDAC

1.4.1. The *KRAS* signalling pathway

The main oncogenic driver for PDAC is known to be alterations in the *KRAS* gene as mutations are observed in 90% of PDAC patients [15]. This mutation is seen in the earliest genetic alterations as it is found in low grade PanINs. It is known to reprogram pancreatic cells into duct-like structures that are capable of progression to an advanced state [16]. The *KRAS* gene encodes for the K-Ras protein that activates the rat sarcoma viral oncogene family/mitogen-activated protein kinase (RAS/MAPK) signalling pathway. The RAS family has a role in mediating cell growth, differentiation and survival. RAS is a protein that belongs to the guanosine triphosphatases (GTPases) family. It serves as a binary switch by controlling the entry of signal transduction molecules and responsible for the activation and inactivation of GDP [17]. The inactivation state is mediated by GTPase-activating proteins (GAPs) which induces hydrolysis of guanosine diphosphate (GTP) to guanosine triphosphate (GDP). When an appropriate stimulus is received, such as the binding of the epidermal growth factor (EGF), there is a decrease in the strength of affinity in the KRAS/GDP complex. This allows for the replacement of GDP to GTP which has a stronger affinity with KRAS due to high concentration levels that cause a conformational change, subsequently activating KRAS [17]. This activation causes pro-tumourigenic changes in the microenvironment due to the targeted pathways that promote migration, cell survival, and proliferation of cancerous cells. These pathways include RAF-MEK-ERK pathway, PI3K-AKT-mTOR pathway and RalGDS/p38MAPK [18]. The mutations seen in KRAS prevent the interaction with GAPs which leaves KRAS bound to GTP. This results in a constant activation state for the targeted signalling pathways. Despite being shown to be the main cause of the development to pancreatic neoplasia, targeting KRAS directly has not produced positive results at a clinical level and hence, deemed as undruggable [18]. Therefore, the sequential tuning of the downstream pathways that contribute towards the invasion, migration and proliferation of KRAS has been considered as a more targeted approach.

1.4.2. PI3K/AKT signalling pathway

Approximately, 50% of PDAC patients display an activation in the Phosphatidylinositol 3-kinase / protein kinase B (PI3K/AKT) signalling pathway that is linked to poor prognosis [19]. Upon binding with an appropriate ligand, such as epidermal growth factors or insulin-like growth factors, the receptor tyrosine kinase recruits PI3K for the phosphorylation of the

phosphatidylinositol. This results in the activation of AKT via the recruitment of phosphatidylinositol to phosphorylate several effectors including AKT. These effectors have a role in promoting translation, proliferation, cell survival and migration, as seen in *Fig. 3*. The mutations in the *pi3kca* coding region that codes for p110α coding region for p85 contributes to constant activation of the PI3K pathway. These mutations and elevated levels of AKTs are observed in PDAC patients. Therefore, PI3K pathway has been suggested as a therapeutic target for PDAC [19].

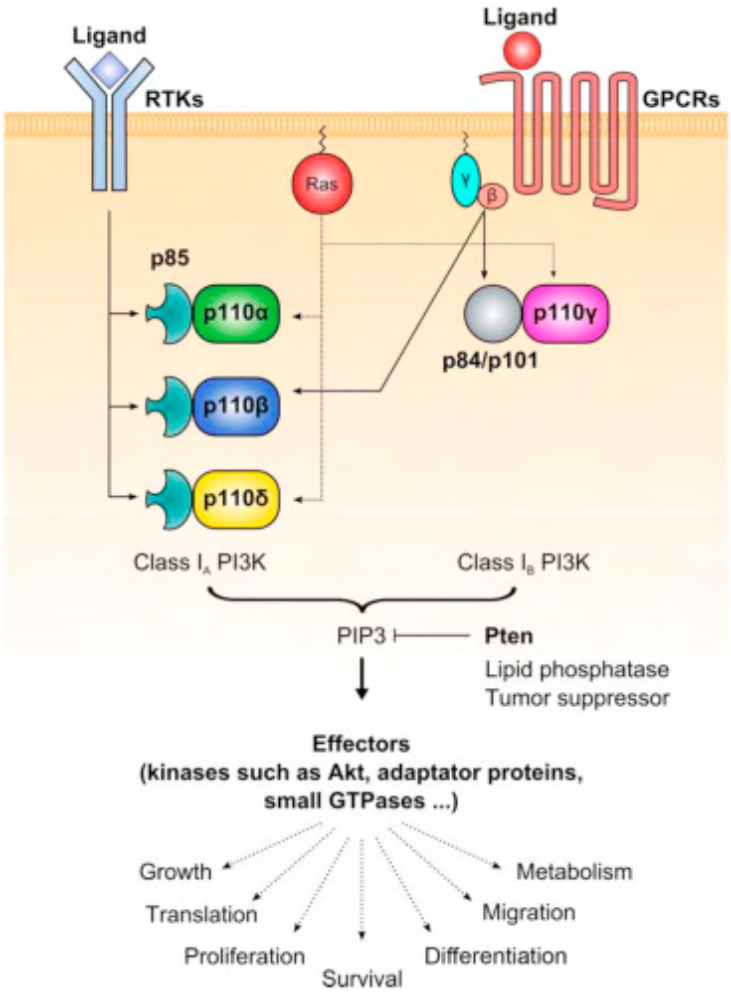


Figure 3 | A diagrammatic representation of the PI3K/AKT signalling pathway. In the presence of a ligand, class 1 PI3Ks produce lipids that modulate multiple signalling pathways including protein kinases and small GTPases. There are four class I isoforms of PI3Ks in a mammalian system (p110s), consisting of a heterodimeric complex. Class I_A PI3K has three catalytic subunits, p110α, β and δ. Each subunit is bound to a Src Homology 2 domain that has a p85 regulatory region. Of these three p110s, p110α gene has a high mutational level when compared to β and δ, dominantly in different types of

cancers. Due to it being the main mutated isoform, p110α massively contributes to the hyperactivation of class I/PI3K signalling pathway. Excessive activation of protein kinases and small GTPases results in an increase in cellular growth, translation, proliferation, survival and migration. Image taken from Baer *et al*., 2015 [19].

1.4.3. SMAD4 and TGF-β signalling pathway

The transforming growth factor β (TGF-β) has pleiotropic roles via the regulation of physiological processes. SMAD4 protein is a modulator for the signalling pathway [20]. Genetic alterations have been observed in 60% PDAC patients with a deletion or mutation in the SMAD4 gene which maps to chromosome 18q21. This pathway is initiated upon the binding of TGF-β on its transmembrane receptor, causing phosphorylation of SMAD2/3 and inducing a heterodimeric complex with SMAD4. The presence of S100A11 protein further promotes the formation of this complex [21]. Once formed, it is translocated into the nucleus to activate its target genes, such as cyclin-dependent kinase inhibitor 1 (p21), for the regulation of cell proliferation, cell cycle arrest, and apoptosis. Furthermore, phosphorylation of S100A11 allows it to translocate into the nucleus to compete with Sp1/3 for the expression of p21, as seen in *Fig. 4* [21]. In PDAC, the role of this pathway alters depending on the stage of disease. In the early stage, the TGF-β/SMAD4 signalling pathway has a tumour suppressive role whereby cell cycle arrest and apoptosis is dominant [22]. Once mutated, the signalling activity has an impact on cancer development. Hence, the late stage of PDAC consists of high TGF-β/SMAD4 signalling activity but with an inverse role. Here, the dysregulation of this pathway causes stromal proliferation and EMT. This revealed that TGF-β/SMAD4 signalling pathway has both growth-inhibitory and growth-promoting effects depending on the stage of PDAC [22].

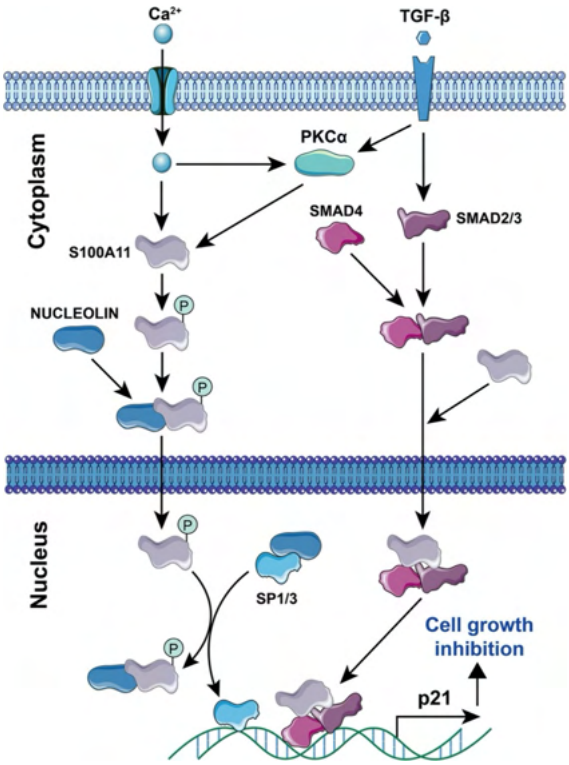


Figure 4 | The SMAD4 and TGF-β signalling pathway with the presence of S100A11. TGF-β ligand binds to its transmembrane receptor. The mediated signals allow for the expression of Protein kinase C alpha (PKCα) to promote the phosphorylation of S100A11. Once phosphorylated, S100A11 translocate into the nucleus to compete and release Sp1/3 from its complex. Free Sp1/3 induces the expression of p21. Additionally, S100A11 forms a complex with SMAD2/3 and SMAD4 to activate p21 expression for cellular growth inhibition. Image taken from Zhang *et al*., 2021 [21].

1.5. PDAC subtypes

Molecular subtypes of a cancer serve as a guide in preclinical and clinical therapeutic developments as it allows for an optimal prediction of patient response. The identification of subtypes reduces the delay in response to treatment as well as cost due to unresponsive or ineffective treatments. The subtypes are distinguished through molecular profiling and histopathological analysis.

There are two main PDAC classifications, seen in *Table 1*, namely the classical subtype and the basal-like subtype [23]. These subtypes have been defined based on malignant epithelial cell gene expression profiles and have been linked to the treatment outcome [24].

According to Moffit *et al*., 2015, the basal-like and classical are tumour specific subtypes that correlate to the survival period [25]. This study included a large number of microarray data consisting of both primary and metastatic PDAC tumours, *Table 1,* to assist in the classification of subtype. The basal-like had a poor median survival rate of 11 months whereas the classical subtype had a survival period of 19 months. Furthermore, for the one-year survival period, basal-like subtype displayed a 44% and classical subtype had 70% survival rate. It had been observed that there was high gene expression levels of laminin and keratins in the basal-like subtype [25]. Interestingly, these gene expression levels observed in the basal-like PDAC coincided with the expression levels observed in previously classified basal-like breast and bladder cancer subtype. Additionally, there was an enrichment in genes linked with GATA6 which is known to aid in epithelial cell differentiation in the classical subtype (88.5%) when compared to the basal-like (11.5%) [25].

Raphael *et al*., 2017 conducted multi-omics sequencing to understand the differences in gene expression to conclude the class of subtype. The results showed different expression levels in long noncoding (lncRNA) for the two, including cancer-associated genes (*UCA1, HNF1A-AS1, NORAD/*LINC00657*)* [26]. Additionally, the lncRNA *EVADR* had an abundant expression with a two-fold increase in the classical subtype and has a role in regulating the *FOXA2* expression via the recruitment of SMAD2/3. *DEANR1* and lncRNA *GATA6-AS1* also had a two-fold increase in classical subtype. The classical subtype had a high copy number gains of *GATA6* whereas loss of expression in the basal-like/squamous subtype was observed due to high DNA methylation levels. This displays a selective pressure on the subtype of cancer. Lastly, there were more mutations observed in the *TP53* region for the basal-like subtype [26]. Therefore, subtypes are distinguished by mRNA, miRNA and DNA methylation levels which results in a classical or basal-like classification.

Table 1 | The molecular subtypes of PC based on Raphael and Moffit classification and methodology analysis.

Reference	**Classification**	**Source**	**Methodology**
Moffit [25]	Classical: high adhesion-associated, ribosomal and epithelial gene expression and elevated GATA6 expression	61 metastatic PDAC tumours, 145 primary PDAC tumours,	Gene expression analysis and RNA sequencing.

	Basal-like: high expression levels of laminins and keratins.	46 pancreas and 88 adjacent normal samples	
Raphael [26]	Basal-like: upregulation of TP53 mutations and micro- RNA and DNA methylation. Classical: High expression of RTK and MTORC signalling	150 pancreatic cancers	Genomic, transcriptomic and proteomic profiling.

The severity of the basal-like subtype has been explored via single cell transcriptomic sequencing analysis on epithelial cells of six different classical PDAC patients [27]. Stromal heterogeneity contributes massively towards the difficulty in classifying the subtype. It has been revealed that there are four cell clustered regions in the classical subtype, however, one region displayed the phenotype of a basal-like subtype, indicating the level of widespread of this aggressive subtype [27]. The presence of this intra-heterogeneity promotes inaccurate classification due to the coexistence of these subtypes. Therefore, this current study aims to understand the differences in the classical and basal-like subtype via RNA-sequencing by analysing the differentially expressed genes based on the expression levels observed between the subtypes. Each subtype will have unique sets of genetic alternations, exclusively displaying the differences which may assist in correct classification and tailored treatment.

1.6. Stages of PDAC

The American Joint Committee on Cancer (AJCC) describes the cancer stage through the use of tumour, node and distant metastasis (TNM) staging system. In stage I, the invasive mass is approximately two cm (centimetres) in dimension, however, there is no signs of invasion of lymph nodes. Stage II has a dimension of more than four cm, growth beyond the pancreas and regional spread in the lymph nodes [28]. The lymph nodes play an important role in advancement of cancer as they are found between the pancreas and duodenum. This close proximity allows the nodes to be a common site for metastasis. Hence, in stage III, there is a tumour involvement in the celiac, superior mesenteric and hepatic artery. Lastly, distant metastasis occurs in stage IV whereby cancerous cells migrate to organs including the liver, gastrointestinal tract and lungs. Patients are often diagnosed at a late stage due to being asymptomatic. Late diagnosis is the leading factor for a high mortality rate in PDAC. As the

main organs are largely infected by cancerous cells, this creates an unresectable environment [28].

A molecular analysis of cancer cells in different stages had shown a link between the stage and genetic alterations which have roles in maintaining a malignant phenotype, promote invasion and migration [4]. These alterations were often observed in tumour suppressor genes and oncogenes, resulting in an invasive cancer. Via gene expression profiling, metastatic cancer cells display far more gene alterations when compared to non-metastatic cells [4]. In PC, the process from a primary tumour advancing to a metastatic stage is driven by the paracrine and autocrine pathways that contribute towards colonizing distant tissues. The early stage is dominated by the paracrine growth signal whereas the autocrine signals, including positive growth factors, dominate the late stage [29]. As the paracrine signals dominate the early stage, these cancer cell derived signals (*Table* 2) assist in the advancement to a metastatic stage. The autocrine signals observed in the late stage have roles in chemoresistance, EMT and metastasis, *Table 2.*

Table 2 | Cancer derived cells and signalling pathways that dominate the paracrine and autocrine signals, its functions and involvement in PC metastasis.

Signalling pathway/protein	**Target**	**Function**
Paracrine signals [4].		
Sonic hedgehog (Shh)	Fibroblasts	Myofibroblast proliferation
Transforming growth factor β (TGF-β)	Macrophages, T cells and fibroblasts	EMT, cancer-associated fibroblasts (CAF) activation, immune suppression
Matrix metalloproteinase 7	Stromal matrix	Promote metastasis
Hyaluronic acid	Cancer cells	Role in proliferation and invasion
Colony stimulating factor 1	Macrophages	Invasion
Chemokine ligand	Granulocytes	Liver pro-metastatic liver niche and immune suppression
Versican-Wnt/ β catenin	Macrophages	Metastatic growth

Autocrine signals		
Annexin A2/semaphorin 3D	Blood vessels	Promotes metastasis [30]
Dual specificity phosphatase 28	Patelet derived growth factor A	Chemo-resistance and migration [31]
TGF-β	Stromal fibroblasts	EMT, CAFs, growth and invasion [32]

1.7. Cancer therapy treatment and resistance

Systemic chemotherapy has been proposed to be a viable option as patients are diagnosed at an advanced stage. Due to the late stage, surgical resection is not recommended as metastasis has taken place. There are two main treatments namely, gemcitabine (GEM) and FOLFIRINOX.

1.7.1. Gemcitabine

Gemcitabine is often used in combination with other therapies, however, has a limited success rate due to drug resistance. The pathway is initiated through the entry of GEM (2′ ,2′-difluorodeoxycytidine, dFdC, a nucleoside analogue) [33]. Via a nucleoside transporter, GEM proceeds though the membrane for a two- step phosphorylation process [33]. A rate-limiting enzyme known as deoxycytidinekinase (dCK) phosphorylates GEM to form a monophosphate (dFdCMP). The second phosphorylation is involved in producing GEM diphosphate (dFdCDP) and triphosphate (dFdCTP). As an active form, dFdCTP is stochastically incorporated into the DNA strand during replication and therefore hindering DNA synthesis due to the termination of DNA chain elongation [34]. However, there is a high GEM intrinsic and extrinsic resistance and leads to a response rate of less than 20% [34]. Many factors contribute towards resistance including the amplification of the heterogenous stroma which promotes deamination of dFdCMP into difluorodeoxyuridine. Difluorodeoxyuridine is not a nucleoside substrate analogue and therefore undergoes degradation. Additionally, hypoxia-inducible factor-1α (HIF-1α) promotes the EMT process that contributes towards reducing nucleoside transporters. This prevents GEM from passing though the membrane and therefore targeted for degradation [33].

1.7.2. FOLFIRINOX

FOLFIRINOX (folinic acid/leucovorin, 5-fluorouracil, irinotecan, oxaliplatin) is a chemotherapy regimen that used in metastatic and advanced PC. The chemotherapy agents used in this regimen all contribute towards decreases the symptoms of PC and with a significantly longer overall survival period and response rate. Once in a cell, a conversion of fluorouracil (5-FU) to fluorodeoxyuridine monophosphate (FdUMP) occurs to produce a complex with an enzyme, thymidylate synthase [35]. This complex inhibits the production of deoxythymidine monophosphate which is vital for DNA synthesis and repair. Hence, a reduction in this compound induces strand breakage and acts a pyrimidine analog by incorporating a uracil instead of a thymine base. The severe damage caused by 5-FU overwhelms the DNA repair machinery, leading to death in proliferating cells[35]. Leucovorin enhances the toxicity of 5-FU and increasing the binding affinity to thymidylate synthase [36]. Irinotecan inhibits topoisomerase 1 which is involved in preventing the supercoiling of a DNA strand , hence limiting DNA replication [37]. Lastly, oxaliplatin induces cross-link formation which leads to a decrease in DNA replication and synthesis [38].

FOLFIRINOX has a higher survival period when compared to GEM (53.5:35.5 months), however, FOLFIRINOX is a highly toxic drug combination that is only given to certain patients with good health status [39].

1.7.3. CCL2/CCR2 inhibitor (PF04136309) + FOLFIRINOX combination

Monocyte chemoattractant protein-1 MCP-1 (CCL2) is a chemokine that is highly expressed on monocytes as a respond to proinflammatory conditions. CCL2 binds to its receptor, CCR2 (a G-protein coupled receptor), which is expressed on multiple different types of cells including natural killer cells [40]. Interestingly, CCL2 can be the causative factor for tumour growth by regulating various signalling pathways involved in cell proliferation, migration and invasion. Furthermore, high CCL2 expression has been associated to relapse and cancer progression to a metastatic stage. Notable cancer causing signalling pathways would include the phosphatidylinositol 3-kinase (PI3K)/AKT, mitogen-activated protein kinase (MAPK)/p38 and Janus kinase (JAK)/STAT3 [41]. Hence, CCL2/CCR2 inhibitor in combination with FOLFIRINOX has shown a higher survival rate when compared to monotherapy (29% vs 19%) [42].

1.8. Transcriptomic analysis

Transcriptome analysis is the study of the transcriptome which consists of ribonucleic acid (RNA) sequences that are obtained through high-throughput (HT) sequencing techniques. This analysis is commonly used when comparing different sample conditions such as healthy versus disease states. It is applied in diagnoses, personalized medicine, biomarker and drug discovery [43]. RNA sequencing (RNA-Seq) is a ubiquitous tool that quantifies and determines the sequence of an RNA sample through the use of next generation sequencing (NGS) methods in order to understand the complexity and dynamic nature of the transcriptome. It is frequently used for differential gene expression analysis via the quantification of the abundance level or relative transcript level changes under different conditions [44]. When designing an experiment, the number of biological and technical replicates, depth and breadth of sequencing and coverage are to be considered. When compared to previous sequencing techniques, such as Sanger sequencing and micro-array methods, RNA-Seq has a higher coverage, sensitivity and resolution [43]. Transcriptomic analysis would include information regarding messenger RNA (mRNA), long non-coding RNA (lncRNA) and small RNAs such as micro-RNA (miRNA). The purification step will determine the type of RNA present in the sequencing step.

The selection of sequencing library depends on the biological experiment which is based on the type of RNA isolation. The common types consist of poly-A tail selection and ribo-depletion. Poly-A selection is used for gene expression profiling, alternative splicing and detection of variants. This process is done by selectively removing the RNA strands that do not contain the 3' poly-adenylated tail [45]. This selection also removes precursor mRNAs and promote the enrichment of mRNAs in the sample. Once the total RNA has been extracted, the strands are denatured for a short period to allow for the preparation of magnetic beads. The addition of poly-T oligos, 30 thymine bases, are covalently attached to the magnetic beads which then bind to the poly-A tails of the strands. As the remaining types of RNAs do not contain a poly-A tail, it is washed away. Via water heating, the magnetic beads are removed from the purified mRNA strands [45]. Ribo-depletion (rRNA) removes the ribosomal RNAs and allows for the enrichment of non-coding RNAs. This process includes the use of RNase for the degradation of DNA-RNA heteroduplexes. This removes cytoplasmic and mitochondrial rRNAs. rRNA depletion is used as eukaryotic cells contain an abundant level of ribosomal RNA, >80%. Therefore, rRNA depletion is performed to ensure a higher coverage of significant RNAs [45].

An RNA-Seq workflow begins with sample preparation and library generation. This includes RNA extraction from sample of interest, ploy-A tail and mRNA purification. Once purified, the mRNA strands are fragmented, and reverse transcription is done to produce complementary DNA (cDNA) strands. The cDNA is used for the preparation of adapter-ligated sequencing libraries, corresponds to the extracted RNA transcripts, for clonal amplification. Following parallel deep sequencing with an NGS platform, the raw read data is analysed using computational tools to provide biological insights as seen in *Fig. 5* [43].

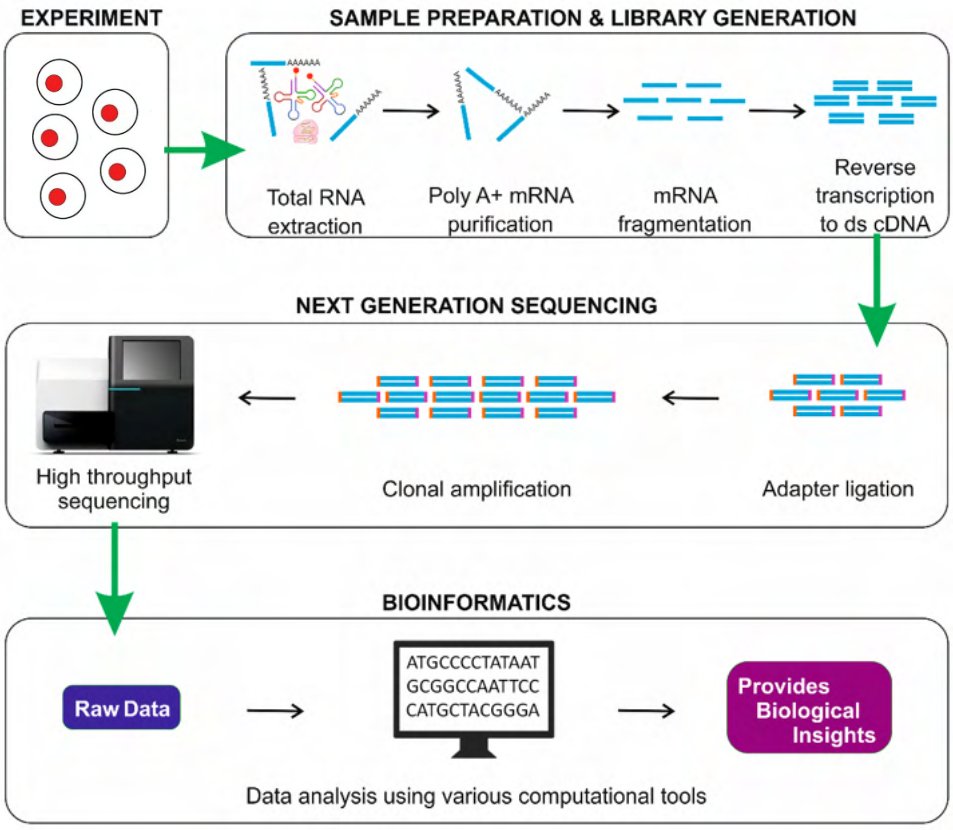


Figure 5 | The methodological steps for an RNA-seq experiment. Analysis begins from extraction of sample, purification, ligation, amplification and sequencing. Image taken from Sudhagar *et al*., 2018 [46].

1.9. Transcriptomics and PC Diagnosis

Although there has been significant progress in therapeutic options, the desired results have yet to be achieved in terms of high survival and remission rates. A crucial factor towards an increased survival rate is the correct diagnosis of cancer. Transcriptomic analysis has assisted in PC research via different types of sequencing techniques. Single cell sequencing identified the differences observed in the heterogeneity of the cellular compositions in primary and

metastatic lesions from 10 PDAC patients [47]. Additionally, it defined the state of subtype which is used to determine the prognosis of disease and clinical outcome.

Transcriptomic analysis of the heterogeneity in 17 PDAC patients revealed long-term and short-term survival rates based on the gene expression levels calculated through computational tools [48]. This was done via analysing the RNA expression levels using a microarray approach to determine the clinical outcome of patients based on their response level to treatments. This study used an endoscopic ultrasound–guided fine-needle aspiration method to include all patients as well as fresh RNA samples directly from the source to reduce degraded or inefficient sample usage. Affymetrix microchips was used and a principal component analysis was performed to group the patients based on their expression profiles. This analysis utilized the clinical data and linked it to the short-term or long-term survival rates [48]. This was followed by heat maps to determine all of the differentially expressed genes between the short-term and long-term survival patients. Gene ontology analysis revealed the biological roles of the identified genes in cancer and how it contributes towards a short-term survival term. Furthermore, transcriptomic analysis predicted individual patient drug sensitivity rates for anticancer drugs [48]. This shows the important role that bioinformatic analysis can play in personalized medicine.

1.10. Introduction to the present study

An *in-silico* approach is used to identify potential biomarkers for various comparison conditions in PDAC. As in-house sequencing was not performed, the datasets used in this study was downloaded from well-established databases. These databases contain sequencing information from different research studies, allowing for a multi-purpose usage of the sequencing data. Through the use of existing sequencing files, an RNA-sequence workflow was created to identify potential biomarkers along with their *in-silico* validation. The comparison conditions included PDAC versus normal, early versus late stage, cancer versus treatment and lastly, classical subtype versus the basal-like subtype. The aforementioned comparisons were done to predict potential biomarkers for early diagnosis and prognosis of PDAC as well as treatment efficiency with high survival rates.

2. Aim and objectives

2.1. Aim

The aim of this project was to analyse RNA-seq PDAC data by performing downstream analysis. This was done with the use of bioinformatic tools for the identification of differentially expressed genes that would represent potential biomarkers. This *in silico* approach was performed with the following objectives:

2.2. Objectives

2.2.1. Obtain publicly available RNA sequence datasets for the comparison conditions.

2.2.2. Pre-process and analyse the downloaded data using an RNA-Sequence workflow with the following programmes:

a. Perform a quality control test using FastQC and Trimmomatic to assess sequence quality, GC content quantification and removal of adapter and low-quality sequences.

b. Alignment of sequencing reads to the human reference genome using the bioinformatics tool HISAT2.

c. Generating count data via the bioinformatic tool HTSeq-count following alignment.

d. Perform differential gene expression analysis using the R Bioconductor tool DESeq2 to produce the differentially expressed genes for each comparison condition.

e. Perform co-expression analysis using WGCNA, R Bioconductor tool, and visualisation of network with Cytoscape.

f. Perform functional enrichment using the R Bioconductor tool Cluster-Profiler for the identification of molecular pathways.

g. Produce Kaplan-Meier survival plots on the selected differentially expressed genes, representing an *in-silico* validation step.

h. Validation of the signature genes in existing databases.

i. Evaluation of the signature genes in PDAC.

3. Materials and Method

The project design for this study is seen in *Fig. 6.* The two datasets followed the workflow for the identification of potential biomarkers relevant to compared conditions.

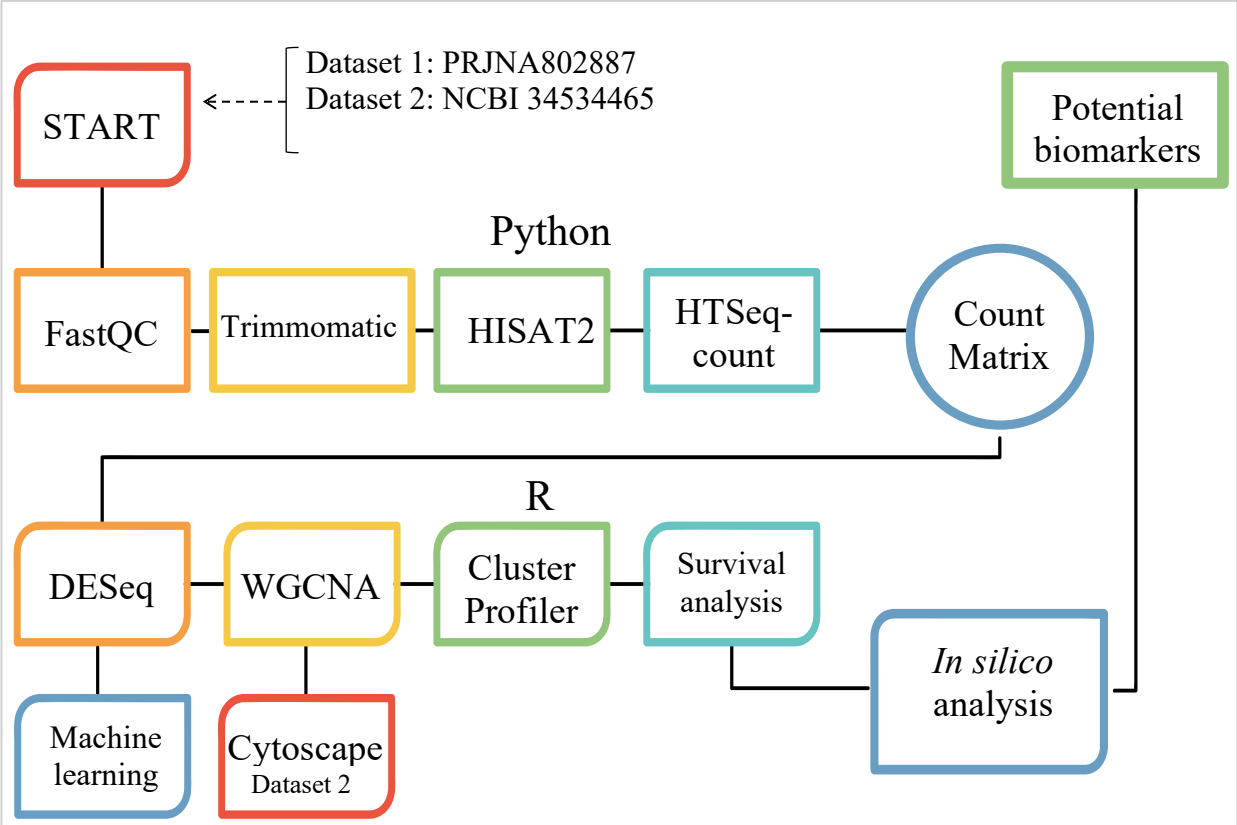


Figure 6 | A flowchart summary of the workflow used in this study on publicly available RNA sequence files. The workflow began with quality control using FastQC and Trimmomatic. Alignment was done through HISAT2 and a count matrix was produced from HTSeq-count. This matrix was analysed with DESeq2, WGCNA and a machine learning approach. ClusterProfiler indicated the biological roles of the significant genes and lastly, the workflow concluded with *in-silico* analysis steps to identify potential biomarkers.

The workflow, presented in *Fig. 6*, consists of two main parts sectioned based on the programming languages used. Python was utilized via the command line for simultaneous analysis of each file which followed a sequential order that required manual checkpoints. The script indicated the location of files, processed and produced an output. The end line for Python was the generation of a raw count matrix with HTSeq2. The count matrix was analysed in R Studio consisting of different packages including DESeq2, WGCNA and ClusterProfiler.

3.1. Acquisition of RNA-Seq Data

There are several open-source databases that were formed to aid in bioinformatic research including the European molecular biology laboratory (EMBL). The European Nucleotide Archives (ENA) is a subdivision of EMBL that holds core publicly resource data such as annotated DNA and RNA sequences obtained through HT sequencing techniques (https://www.ebi.ac.uk/ena/browser/).

Dataset 1 were downloaded from ENA for different comparisons. Sequencing of single-end RNA files for PDAC (n = 12) and normal adjacent (n = 12) samples were downloaded to predict biomarkers that contribute towards the aggression and/or progression of cancer (PRJNA802887, https://www.ebi.ac.uk/ena/browser/view/PRJNA802887).

The dataset 2 consisted of a larger sample number (n = 140 tumours and n = 39 normal) that was obtained from the National Cancer Institute Genomic Data Commons (GDC) Data Portal. GDC is a robust platform that provides users with web-based access to HT sequencing data from cancer genomics studies including the Cancer Genome Atlas (TCGA). This platform enables users to submit or download high performance data in either raw or pre-processed files (counts) following alignment. The GDC database, accessible at https://portal.gdc.cancer.gov/, has five navigation panels including a repository button which is used to filter the required data of interest. The dataset study and supplementary material is accessible at https://pubmed.ncbi.nlm.nih.gov/34534465/. The count datasets were downloaded and evaluated for stage and subtype analysis in DESeq2.

3.2. FASTQ file format

Datasets 1 was in a fastq file format. Each sample file has a 'fa.gz' extension. This represents a FASTQ file format that has relevant information such as a unique identifier, sequence and quality scores that is organized in four-line blocks [49]. A greater than character (>) designates the beginning of a new sequence and contains an accession number/identifier. The sequence identifier contains information about the sequencing procedure including the machine and library design [49]. The actual sequence is shown on the second line containing letters that correspond to a base call. The separator, also known as the place holder, is used for additional comments associated with the sequencing reads. A plus sign is commonly used to retain the four-line FASTQ structure. Lastly, the fourth line represents the quality scores for each base call. These scores are reported using the Phred +33 encoded scale that are ASCII characters

represented numerical quality scores. A FASTQ file is an output of next generation sequencing techniques such as Illumina sequencing [49].

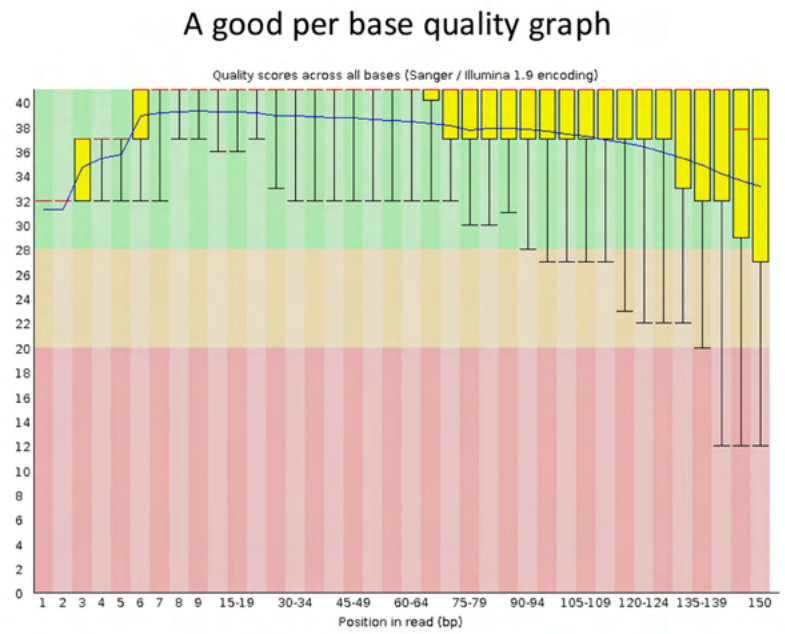
Dataset 2 files were not available in fastq file format, hence a count matrix (.cvs format) was used. Quality control was not done for the sequencing files, however, as the samples were submitted to The Cancer Genome Atlas (TCGA), a rigorous quality control criteria is followed by TCGA. Sample files accepted by TCGA undergo quality control protocols to remove bad sequencing files that may not accurately represent the base of a sequencing template [50].

3.3. Pre-processing of raw data: Quality control

3.3.1. FastQC

The FASTQ files undergo a quality control test to determine the quality of the sequences. This vital step is performed prior to downstream analysis as poor sequencing files are removed to avoid inconsistent results. FastQC dashboard is written in a combination of a command line and Python. FastQC [51] (version 0.11.9), developed by Barbraham Bioinformatics, is common tool used to analyse the quality of high throughput sequence data. FastQC outputs a graphical representation of various modules such as the basic statistics, per base sequence quality, per sequence quality scores, per base content, per base GC content, over representation of sequence, K-mer content and adapter content [51]. The per base sequence quality diagram displays an overview of the quality scores across all bases [51]. Each position in read base pair (bp) has a box and whisker plot to show the minimum, lower quartile, median, upper quartile and the maximum values. The quality scores are seen on the Y-axis whereas the position in read (bp) is placed on the X-axis. The graph itself has three colour coded regions which corresponds to the Y-axis scores. A good base call will have a high score value which is placed in the green region going towards the yellow region, as seen in *Fig. 7a*. A bad base call will have a lower value in the red region [51]. It is expected to see a drop in base call as the cycle progresses towards the end of the run. This represents the ‘phasing’, an incorrect removal of the nucleotide blocker after imaging.

a.



b.

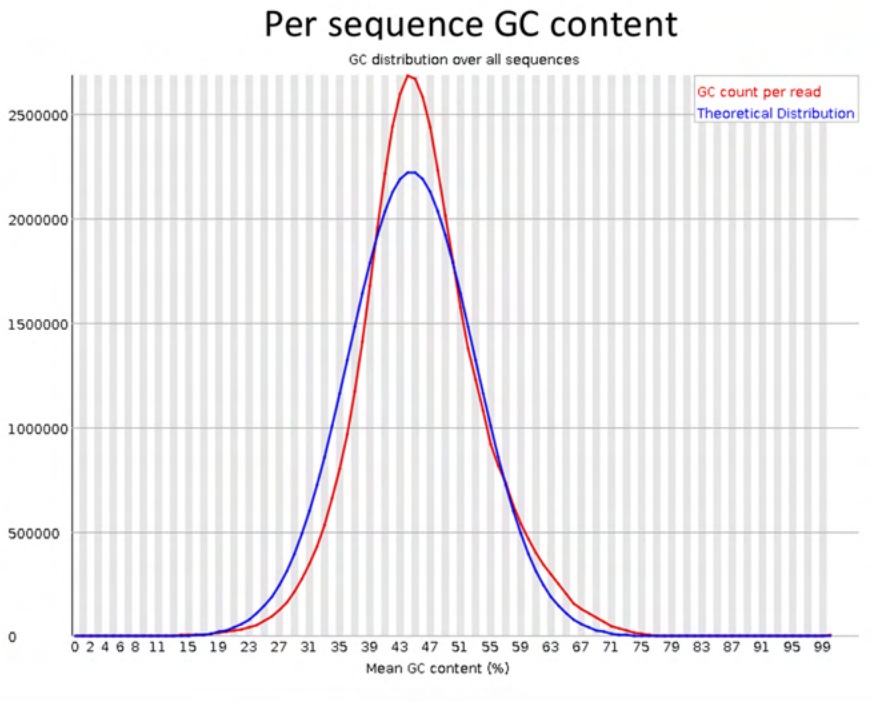


Figure 7 | A graphical representation of a good per base quality and per sequence GC content graphs. The example graphs show the expected FastQC quality control graphs. a) A good base per quality graph. The box and whisker plots remain in the green region between 28-40 Phred quality scores. If the box and whisker plots as in seen the red region, it would indicate bad quality scores. Therefore, incorrect base calls were performed during sequencing. b) A good per sequence GC content graph. The bell-like distribution peak is observed in the optimal range of 40% - 59% GC content. If the curve was shifted

below 40% or above 60%, this would display a low coverage rate due to low clonal amplification during sequencing. Image taken from [52].

The per base GC content is another important module as it determines the coverage rate. An optimal GC content ranges between 40%-60% whereby a high value displays high thermal stability and low GC content shows low thermal stability [51]. GC content above 60% is not ideal as it is thermodynamically unfastened. This requires more energy to separate the strands and therefore decreasing the clonal amplification and low coverage rate. Hence the percentage of GC content is evaluated prior downstream analysis. Graphically, the X-axis has a mean GC content (%) of values from 1-100. There are two bell-like distribution curves, the GC count per read and a theoretical distribution. The bell peak is expected to be within the optimal GC content range, *Fig. 7b* [51]. Sequencing files with bad base calls and per sequence GC content was removed from downstream analysis. Simultaneous Fastq analysis for all files was done via Python by downloading FastQC as a non-interactive mode (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/INSTALL.txt) for the script. The file location and output directories were specified and HyperText Markup Language (HTML) reports were generated per file. The fastq files that produced a good per base quality graph and GC content values were inputted into Trimmomatic for the detection and removal of low quality reads and adapter sequences.

3.3.2. Trimmomatic

The adapter content determines the presence of adapter sequences that were used to ligate the strands onto the flow cell. These sequences are not present in the human reference genome and are removed to prevent low alignment rates due to the presence of unaligned reads. This problem is circumvented through the use of a bioinformatic tool Trimmomatic [53] (version 0.39), a read trimming tool through the command line. Poor quality and technical sequences (adapter sequences) affect the consistency of the results. Trimmomatic contains an additional folder consisting of the adapter sequences present for both single-end and paired-end reads. This adapter folder contains 'TruSeq3-SE.fa' files which indicates the type of adapter sequences, single-end files (SE) [53]. It removes the adapter section by identifying the universal adapter sequence present in the reads used in next generation sequencing (NGS) data. Furthermore, it evaluates GC content, removes low quality leading and trailing reads. This process is repeated until high quality reads are achieved [53]. Therefore, this package was used in Python, via the command line, to analyse the fastq files and produce trimmed files with the

extension of '.fastq.gz'. The ILLUMINACLIP step allowed for the cutting of the adapter sequences and other Illumina specific bases. Following Trimmomatic, a read alignment step was performed using HISAT2 for the alignment of the fastq reads to the human reference genome using Python.

3.4. HISAT2: Read alignment

Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT2) [54](version 2.2.1) is a bioinformatic tool used to align DNA and RNA sequences to a reference genome. As RNA-sequencing experiments are commonly based on patient data, a human reference genome is considered. HISAT2 is a fast and highly sensitive splice aligner that accounts for multi-exon spanning reads and pseudogenes with a moderate random access memory (RAM) usage [55]. HISAT2 is about 50 times faster than TopHat2 and marginally faster than STAR (consumes high memory space) with the use of less memory due to the usage of local indices. Additionally, HISAT2 has a sensitivity rate of 97.3% when compared to STAR (96.3%) and Tophat2 (79.05%). The precision rate of HISAT2 and STAR are 93.1% and 91.8%, respectively [54].

This rapid alignment of sequence is due to the implementation of a generalized compressed suffix array, an extension of Burrows-Wheeler transform, and a Ferragina Manzini (FM) index [55]. Additionally, to provide a higher alignment rate with greater coverage, HISAT2 incorporates both global and local indices. The global index spans across the entire genome and 14.5 million small variants [56]. The local or small indices consists of 64,000 bps that collectively makeup the three billion bases of the entire genome with overlapping regions of 1024 bps regions. For a local index, chromosome 1 would have bps from 1 to 56,000 with a second index running from 55,000 to 111,000 bps, representing an overlapping boundary between 55,000-56,000 bps. These multiple indices decrease the ambiguity of the read mapping which allows for a unique alignment [55].

RNA-Seq reads contain large gaps due to the removal of intronic regions as the region undergoes splicing. The cDNA will have low alignment rates due to the absence of these regions as the reference genome contains both introns and exons. Additionally, exons are relatively small in bps when compared to introns (< 200 bps :11,000 bps) hence a 100 bp read may have more than one exon [57]. HISAT2 divides the exon spanning reads into three categories, long-anchored reads (> 15 bps), intermediate-anchored reads (8-15 bps) and short-

anchored reads (1-7 bps) [55]. The sizes of each category differ to allow for a relatively easy alignment with unique mapping locations and reduce the number of unmapped reads.

As mentioned, HISAT2 requires a reference genome for indexing. The human reference genome, hg38.fa, was downloaded from University of California Santa Cruz (UCSC) genome browser, (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/). The reference genome was assembled via Python using the 'hisat2-build' command into its eight genome '.ht2' files. Following QC, the fastq files and the genome file were used to produce a sequence alignment/map format (SAM).

HISAT2 produced a summary statistics output on the command line. Each sample file contained information about the overall alignment rate, total paired and unpaired rates. Concurrently, a SAM tab-delimited human readable file is produced at the located output folder. SAM files have a fixed format that is used for storing sequencing data following alignment. A SAM file has two main sections, a header (@) and a section with the alignment data itself. Each file contains 11 tab-delimited compulsory fields namely QNAME, FLAG, RNAME, POS, MAPQ, CIGAR, RNEXT, PNEXT, TLEN, SEQ and QUAL [58]. The query template name is stored under the QNAME. RNAME provides details with regards to which chromosome the sequencing read was aligned too. POS gives the relative genomic coordinates of the aligned first base to the chromosome. MAPQ and QUAL field contain information about the sequence quality scores (Phred + 33 scale) and value of alignment. RNEXT and PNEXT are used in paired-end experiments and has information about the positional relationship between the first and second read pair. The flag field has the outcome of alignment encoded in bit notation to reduce the size of the SAM file. The sequencing read is provided in the SEQ field. Lastly, the CIGAR field contains the alignment itself given in a CIGAR string notation that is relative to the reference genome [58].

The SAM files were converted into binary alignment/map (BAM) files, non-human readable files, using SAMtools [59] (version 1.7). The BAM files were compressed versions of SAM files that were used to produce sorted BAM and index BAM files with SAMtools that was inputted into HTSeq for the generation of count matrix data. Additionally, the use of indexing reduced the memory usage whereby the external table of indexes was used to navigate to unique regions without loading the entire sequence [58].

3.5. HTSeq count: Generating count data

BAM index and sorted files were further analysed to produce a count matrix. HTSeq [60] (version 0.11.1) is a Python framework that has the ability to process the SAM/BAM files following alignment. A core component of HTSeq is working with both genomic coordinates and intervals. HTSeq count is used for differential gene expression analysis. Along with a SAM/BAM file, it requires a gene transfer file (GTF) format. A GTF (GRCh38) file contains gene models, genomic individual features, annotations and relative positions about the gene structure. The GTF file was downloaded from NCBI (https://www.ncbi.nlm.nih.gov/genome/guide/human/).

The use of HTSeq count allows for unambiguous mapping. Hence, reads mapped to a single gene is considered whereas multiple mapping of reads or overlapping reads are discarded. This is performed to avoid incorrect identification of differentially expressed genes. A user may set a desired overlapping resolution mode from the three modes available namely, 'union', 'intersection-strict' and 'intersection-nonempty'. The default and recommended mode 'union' was used. HTSeq overlayed the GTF file on the BAM files using the 'htseq-count' command to identify the position of mapped reads. This generated a count matrix in a tab-delimited format consisting of the first column containing all the ENSEMBL gene identification names and the first row with the sample names along with numerical values to represent the number of mapped reads [60]. The raw count matrix was inputted into R Studio DESeq2 to identify the differentially expressed genes per sample condition.

3.6. DESeq2: Differential gene expression (DE) analysis

The use of high-throughput count data often includes the analysis for the detection of differentially expressed genes between comparison conditions. To determine the systemic changes between the comparison conditions require quantification and statistical calculations which can be conducted by the R Bioconductor package DESeq2 [61] (version 1.37.4). This package conducts differential gene expression by testing through the use of a negative binomial generalized linear models (GLM) and empirical Bayes technique for the shrinkage estimation of log fold change (LFC) [61] .

Genes with low read counts display a strong variance of LFC estimation which is reduced through the use of the empirical Bayes technique. The weakly expressed genes show a larger difference than the strongly expressed genes which is a direct result of using data with low

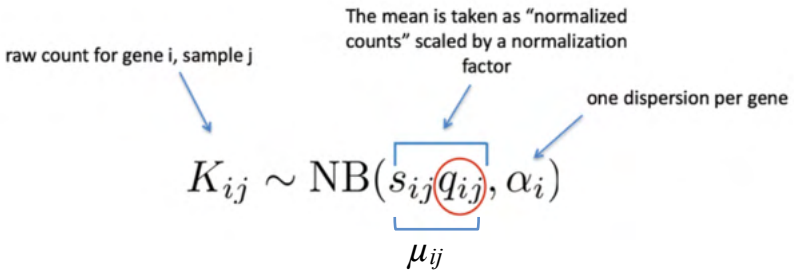
counts associated with high noise, representing a heteroskedastic data. This problem is circumvented through the use of a stronger shrinkage LFC estimation towards a zero for genes with low count values. Additionally, DESeq2 performs two rounds of GLM for an even spread such that strong LFC estimates are not observed with genes that display weak expression [61].

DESeq2 requires minimal input data including the count matrix (*K*) with a single row for each gene ($i = 1,..,n$) and one column for each sample represented in $j = 1,..,m$. This produces a count matrix, *Kij,* containing the number of reads that have been mapped to the human reference genome [62]. The function 'design' serves as an indication of modelling or grouping the samples in batches in order to measure the effect of condition with regards to the control through the differences calculated. These functions are inputted in DESeq2 with the function 'DESeqDataSetFromMatrix'. It is to be noted that DESeq2 requires unnormalized integer data obtained from an RNA-Seq experiment [62]. Additionally, the remaining main functions include 'DESeq', 'results', 'vst 'and 'plotPCA' for this study.

The function 'DESeq' performs internal default analysis based on the negative binomial distribution [61]. This includes the estimation of size factors, estimation of dispersion and lastly, the negative binomial GLM fitting and Wald statistics, returning a table consisting of baseMean, log2 fold changes (log_2FC**)**, P-values and p-adjusted values using the 'results' function [62].

The estimate size factor uses a median ratio factor. The estimation of dispersion for a DESeqDataSet obtains dispersion estimates for the negative binomial distributed data containing a 'fit type' argument. This shows the relationship between the mean and variance. For each gene, the dispersion estimate is calculated, maximising the Cox Reid adjusted likelihood to alleviate problems one may encounter such as bias, inconsistent or variance estimates [62]. A trend line is captured which displays the dispersion mean relationship. This is fitted to the maximum likelihood estimates which is followed by the final dispersion parameter used for downstream analysis. Lastly, the Wald test for the GLM coefficients determines the significance of coefficient in the negative binomial GLM by using the size factor and dispersion estimate values for the calculation of the log_2 fold change and effect size [63].

Once the dispersion has been estimated, the count value of every *i* is fitted into the GLM model consisting of the following formula:



As mentioned previously, K_{ij} represents the raw count matrix whereby i = gene and j = sample, inputting two parameters (μ_{ij} and α_i). With the use of the fitted mean (μ_{ij}) and a gene-specific dispersion (α_i) parameter, for every i, j is modelled using the negative binomial distribution [62].

The mean consists of a normalization/size factor (s_j) that is multiplied by the quantity (q_{ij}) in the following equation:

$$\mu_{ij} = s_j q_{ij}$$

Following the fitting of model, a coefficient estimation is performed for each sample group (the compared conditions seen below) and its standard errors with the equation:

$$\log_2 (q_{ij}) = x_j.\beta_i$$

The coefficients βi provides the $\log_2$FC for gene i for each j of the model matrix (x) [62]. The $\log_2$FC value display the change in expression level for the sample set.

Following the GLM fitting and WALD statistics, the results were wrapped up using varianceStabilizingTransformation (VST) for downstream methods such as the construction of a principal component analysis (PCA) plot. The function 'VST' calculates the variance stabilizing transformation from the fitted dispersion-mean relationship and perform a normalization step to produce a more homoscedastic data [61]. This allows for a constant variance and prevent random noise domination [62]. VST requires an object (DESeqDataSetFromMatrix'), blind and fit type. The function 'blind' was set to 'TRUE' in order to compare samples in an unbiased manner.

Additionally, DESeq2 has the ability for detecting outliers. An outlier would represent a gene that has a count value in a single digit for all the samples excluding one sample that would have a drastic change in digits and therefore considered an outlier. This is an important step to remove genes with low counts and outliers to avoid inconsistent results that would have a

negative impact on downstream analysis. The aim is to find both upregulated and downregulated genes that are consistent across all samples. Hence, DESeq provides two responses for the flagged genes [61]. If there is an outlier that is seen in less than six samples of the same conditions, the gene is removed from all samples. However, a more favoured approach is that if the outlier is present in more than seven samples, the count is trimmed to a mean value of the gene [61].

DESeq2 required an input data including the count matrix in a table format as seen in *Fig. 8,* along with a meta data file called 'colData' containing sample information with a group column containing the compared character. The HTSeq count values were inputted and analysed accordingly with the DESeq function.

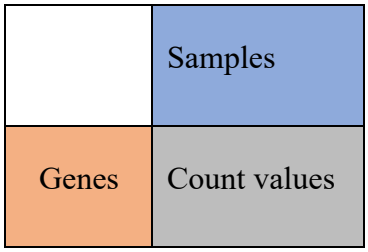


Figure 8 | The format of a count matrix used in DESeq analysis. Columns consisted of sample numbers whereas rows showed the gene name/number. The matrix contained count values of each gene per sample.

The comparison conditions included the following:

Dataset 1:

1. Normal versus PDAC

Dataset 2:

1. Normal versus PDAC
2. Basal-like versus Classical subtypes
3. Early stage versus Late stage

A prefiltering step was performed whereby genes with count values of less than 10 were discarded as it provides minimal to no information and contributes to the heteroskedasticity of data [63]. A principal component analysis (PCA) plot was produced for each comparison condition as it shows the clusters of the condition based on the similarity observed between

them. The plot allowed for a graphical representation of outliers that were removed, through sub-setting, in order to analyse samples that display unique clustering per condition. Following sample clean-up, the sample conditions were compared accordingly. For each comparison, a reference group was selected and the results function produced a statistical table including the log_2FC values. The log_2FC values were ordered according to their values whereby a fold change of more than 1 was represented as the upregulated differentially expressed genes (DEGs), and fold change of less than -1 were the downregulated DEGs. Genes with log_2FC of 0 were discarded as it displayed no change in gene expression level between the compared conditions. Furthermore, the Benjamini-Hochberg (BH) adjusted P-values were used as it reduces the false outcome rate [62]. The top significant genes with BH-adjusted P-values and q-values of less than 0.05 was selected for further analyses to show a 95% confidence rate in the selected DEGs. DESeq2 was performed in R-Studio (version 1.4.1717).

3.7. Machine learning

To determine the sensitivity and specificity of the selected DEGs that may be represented as potential biomarkers, an external database was to be used (Oncomine, https://www.oncomine.com/) to serve as a validation step. However, the free version of Oncomine was removed permanently, hence a machine learning approach was taken to calculate the required values.

Each comparison condition had a set of significant DEGs. A machine learning approach was performed for prediction accuracy in biomarker discovery. The following was done for each compared condition in Python using the Scikit-learn [64]:

Pre-processing the data matrix for machine learning:

The DEGs obtained from DESeq2 consisted of a data matrix with various columns including the names of DEGs and count values. The count values for each of the DEGs from the initial raw count data matrix were extracted and transposed. Transposing produced a table with the first column as sample IDs, second column with DEG names (X, count values) and lastly, a phenotype column (y). The phenotype column had two binary numbers corresponding to the sample condition whereby tumour was assigned a ‘1’ and normal was assigned a ‘0’. This was followed by merging the normal and tumour tables through the ‘concatenation’ function.

X = matrix with count values and sample IDs

y = phenotype (0 and 1 values)

Random Forest Classifier

A Random Forest Classifier (RFC) is used for solving regression and classification problems. It consists of a 3-step process including shuffle and split, training of model and testing the trained model. This algorithm is a collection of trees whereby each tree is the ensemble consisting of sample data extracted from a sub training set. It uses the average of the sub samples to improve the prediction accuracy and to control an overfitting situation [65].

A training test is done by splitting and shuffling the rows in the data matrix. CLF 'fit'(X_train, y_train) is used for model training on the training set (70% of data) to build a forest tree. An uppercase 'X' represents the matrix of count values for each DEG per sample ID whereas a lowercase 'y' is the assigned binary numbers, also known as the classification label.

The next step involves 'predict (X)' with the parameter X and return y, the predicted classes. This is done to see how well the algorithm can predict the y phenotype for the X on the untrained test data (30% of data) [66].

Once done, the recall, specificity score, accuracy score, F-1 score and precision score is calculated with the y-tested and y-predicted values. Cross validation was done for each comparison as each score was performed 5 times and an average with a standard deviation was calculated.

Recall score

The recall score computes ability of the classifier to find all the positive samples (good value = 1 and bad value = 0) and is calculated with the following equation:

$$\frac{tp}{tp + fn}$$

tp = number of true positives

fn = number of false negatives

Precision score

The precision score displays the ability of the classifier to correctly identify a sample such that a negative sample may not be labelled as a positive, good value = 1 and bad value = 0.

$$\frac{tp}{tp + fp}$$

tp = number of true positives

fp = number of false positives

F1/ balanced F-score

F1 score is a numerical average or harmonic mean of the recall score and the precision score to give a value ranging from 0 (bad value) to 1 (good value). The score is calculated by:

$$2 * \frac{precision * recall}{precision + recall}$$

In a Random Forest Classifier, an external validation step is required. This displays how well the trained system can predict a phenotype based on the count values in an independent study/data matrix and therefore, dataset 1 was used as the external validation step with the same DEG names. The 'predict (X)' with the parameter X consisted of the entire data matrix (100%) for dataset 1 along with the y_test, representing the phenotype (cancer or normal). This showed how well the algorithm can predict the correct phenotype based on the trend observed in the trained data from dataset 2.

3.8. WGCNA: Co-expression analysis

Weighted Gene Co-expression Network Analysis (WGCNA) has a variety of usages including construction of networks and detection of gene modules to identify the hub proteins within a cluster that are expressed together [67]. WGCNA has been used in analysing big data from different cancers, disorders and other mammalian organisms [67]. The authors of WGCNA had recommended a sample size of at least 15 or more samples [68]. Hence, the applicable sample set was dataset 2 count matrix as it consisted of 140 initial tumour samples. WGCNA is an available R package ('WGCNA' version 1.71).

WGCNA describes a correlation pattern amongst a sample set by identifying the clusters, also known as modules, of highly correlated dysregulated genes in a sample set. This correlation

network is performed through mathematical calculations. The network is produced based on the correlation that could be described as an n by m matrix $X = [x_{il}]$ whereby the row corresponds to the network nodes (i = 1, . ., n) and the columns (l = 1, . ., m) represents the sample measurements [67].

In RNA-Seq analysis, the first step consists of constructing a gene expression network on the data produced by DE analysis as the identified differentially expressed genes are significant to the compared conditions whereas genes with no significance or extensively low counts are discarded. Once the DEGs were extracted from the original count matrix with the corresponding count values for all samples, normalization was performed using variance stabilizing transformation. The normalized matrix was then inputted in WGCNA analysis, following transposing of matrix, for all compared conditions. An overview of a WCGNA methodology used in RNA-Seq is shown in *Fig. 9*.

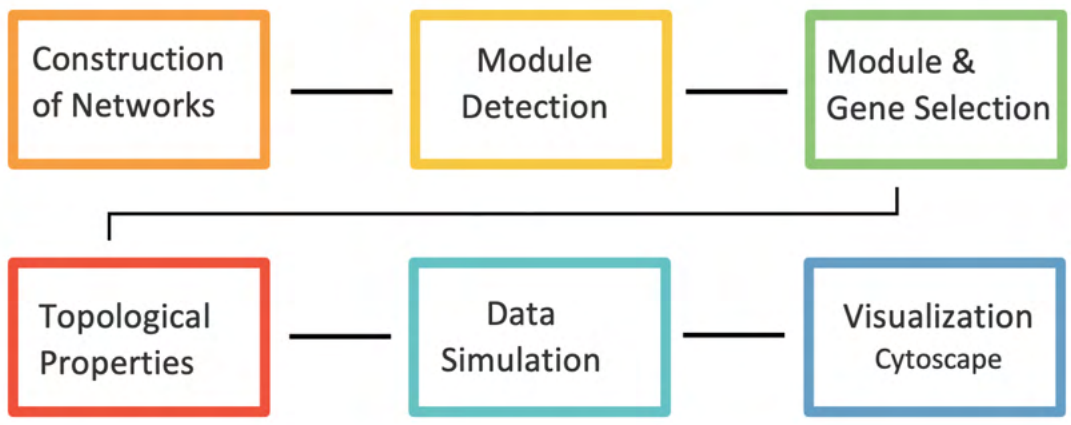


Figure 9 | A summary procedure for WGCNA. Following network production, modules are detected and viewing is done using Cytoscape.

This analysis begins with a network construction which is done through the use of an adjacency matrix (a_{ij}), an n by n matrix with binary numbers (0,1). a_{ij} represents the strength of the network connection between nodes i and j. For the calculations of the adjacency matrix, a co-expression similarity (s_{ij}) is done to show the value of correlation coefficient between the nodes. Additionally, there are two types of thresholds available in the WGCNA package for co-expression analysis. The threshold procedure consists of converting the similarity matrix into a weighted or unweighted adjacency. A hard threshold or a discontinuous is used in

constructing unweighted networks and use absolute values whereas a soft threshold supresses low correlations in a soft manner to avoid loss of information in a weighted network [67].

A weighted network uses a soft threshold whereby the co-expression similarity is raised to a power:

$$a_{ij} = s^{\beta}{}_{ij},\ \beta \geq 1$$

The scale independence plot contains information about the soft threshold (x-axis) and the scale free topology model fit (R^2, y-axis). The soft threshold (β) value is calculated by plotting a *log(p(k))* against *log(k)*. *K* represents the level of connectivity and *p(k)* is the frequency distribution. R^2 is the regression correlation value that is used for measuring the correlation of highly connected genes to lowly connected genes [69]. Hence, the value lies between 0-1 whereby values closer to 1 will contain highly connected genes in the dataset, representing the hub genes that regulates the entire biological network with low background noise [69]. The recommended R^2 should be above 0.8 with the corresponding soft threshold value [69].

Following network construct and selection of threshold power, a module detection is performed to define the interconnected clusters. The adjacency matrix is transformed into a topological overlap matrix (TOM), a similarity measure [69]. TOM assigns genes to a high topology if there is a common neighbourhood and is calculated by:

$$TOM_{ij} = \frac{\overbrace{\sum_{u} a_{iu} a_{uj}}^{l_{ij}} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

TOM_{ij} represents the nodes between *i* and *j* which is defined by l_{ij} and k_i. l_{ij} shows the number of nodes that connect to both *i* and *j* whereas k_i /k_j is the node connectivity for *i* and *j* with adjacency values of $0 \leq a_{ij} \geq 1$, respectively. However in WGCNA, a TOM dissimilarity matrix is recommended to allow for distinctive gene modules [69]:

$$DistTOM_{ij} = 1 - TOM_{ij}$$

The $DistTOM_{ij}$ is used to form a gene dendrogram containing various coloured clusters and branches to show the highly connected genes [69].

The next step was to calculate a module eigengene of the expression matrix with the corresponding modules identified in the $DistTOM_{ij}$. Each module colour had a hub gene which is considered as the representative of that module. The module itself contains sets of highly co-expressed genes [69].

The last step included visualization of the biological network through R Cytoscape [70] (version: 3.9.1) which required information produced by the scale independence plot, TOM graph and eigengene. Visualization was done by using the function 'exportNetworkToCytoscape' that produced node and edge files. This created a biological network whereby nodes represent gene names and a connective line to display the interaction.

3.9. ClusterProfiler: Functional enrichment process

ClusterProfiler is an R Bioconductor package used for analysing and visualizing functional profiles by interpreting omics data. This package supports overexpression, gene set enrichment and biological theme analysis and various annotated ontologies such as Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways [71]. One of the useful utilities provided are functions including compareCluster that is used for comparisons between gene clusters. Furthermore, this package has the ability to explore both coding and non-coding data with corresponding gene annotations [71]. ClusterProfiler [71] (version 4.5.1) was used in this study for each comparison condition. It was used to classify the genes based on the cellular component, molecular function, biological process and KEGG pathway analysis for all of the upregulated DEGs per sample set which consisted of the output from DESeq2 for dataset 2. The BH-adjusted P-value of less than 5% was considered to reduce the false outcome rate for each GO and KEGG setting.

3.10. *In-silico* validation

3.10.1. Kaplan Meier Survival analysis

A survival curve is used to describe the relationship between the probability of survival with gene expression levels against time, representing a mathematical model for survival data. A Kaplan-Meier (KM) curve is produced by a KM plotter (https://kmplot.com/analysis/index.php?p=service&cancer=pancancer_rnaseq). This plotter

contains information about survival data for various cancers including PDAC [72]. KM curves have visual representations of survival with a high or low gene expression at a respective time intervals. The plot contains hazard ratios (HR) and P-values to show the significance of the data for the selected genes. Furthermore, multiple genes can be constructed in one plot with the overall mean expression values. This *in-silico* validation step is calculated using manually curated clinical data. The top significant genes for each comparison condition had survival curves to support or reject the findings following the selection of 'pan-cancer' option and PC type on KM plotter [73]. This was done for each gene as well as a mean plot for the significant KM genes.

3.10.2. Gene expression atlas

The gene expression atlas is a database that contains information regarding gene and protein expression data for various species including *Homo Sapiens* that could be visualized diagrammatically. It is an open-source database (https://www.ebi.ac.uk/gxa/home) that is publicly available and created with curated sources [74]. This database has various options to explore, including a query for a specific gene's expression in diseases such as cancer. The 'biological condition' search bar is used to provide the type of disease and cell line. Hence, a user can select the type of cell line, disease and organism. This produced a boxplot per comparison condition with an expression value, transcripts per million (TPM), for each DEG.

3.10.3. GEPIA

Gene Expression Profiling Interactive Analysis (GEPIA) (http://gepia.cancer-pku.cn/) is a web-based tool that was created to manage data produced by large consortiums [75]. It contains gene expression analysis for over 9000 tumour samples and 8000 normal samples. GEPIA provides several interactive features such as box plotting, survival curves and correlation analysis. The DEGs with high expression seen in the gene atlas database were used to produce box expression plots for both PDAC and normal samples to aid as an additional validation step.

3.10.4. BBCancer

BBCancer (http://bbcancer.renlab.org/explore) is another expression atlas that is used for blood-based screening for early diagnosis and for the development of new biomarkers [76]. Blood base screening includes different types of RNAs and extracellular vesicles (EVs) that have been extracted from cancer patients. The available blood screening for PDAC was expression abundance for EVs. EVs are produced by cancer cells that have a vital role in

intercellular communication to promote cell growth and survival by delivering relevant proteins and metabolites for cancer progression and metastasis [77]. DEGs with boxplots were used in BBCancer for the corresponding value in PDAC.

3.10.5. GENT2

Gene Expression database of Normal and Tumor tissues 2 (GENT2) (http://gent2.appex.kr/gent2/) is a platform for gene expression patterns for normal and tumour tissue samples for different types of cancers. It provides prognostic data and log_2FC for a gene per cancer type, obtained via statistical analysis. This step was done to compare the value of the log_2FC for the DEGs found in this study and the GENT2 database.

3.10.6. Script information

Detailed scripts are available on GitHub (https://github.com/KhadijaSJ/MSc_codes).

4. Results

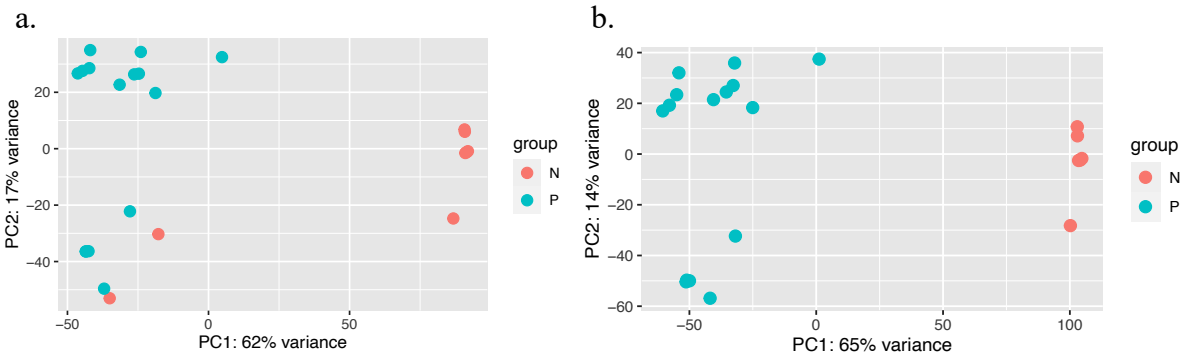
4.1. FastQC analysis

The FastQC file for each sample can be seen in Appendix B. Samples that did not produce a satisfactory base call graph in the green/yellow regions were discarded to avoid the inclusion of bad sequencing files in downstream analysis.

4.2. DESeq2 analysis

Following quality control, read alignment and count generations, the data matrix was inputted into DESeq2 for the identification of differentially expressed genes. The PCA plots (*Fig. 10*) were made to remove outliers to prevent inconsistent results and evaluation of data samples.

Dataset 1



Dataset 2

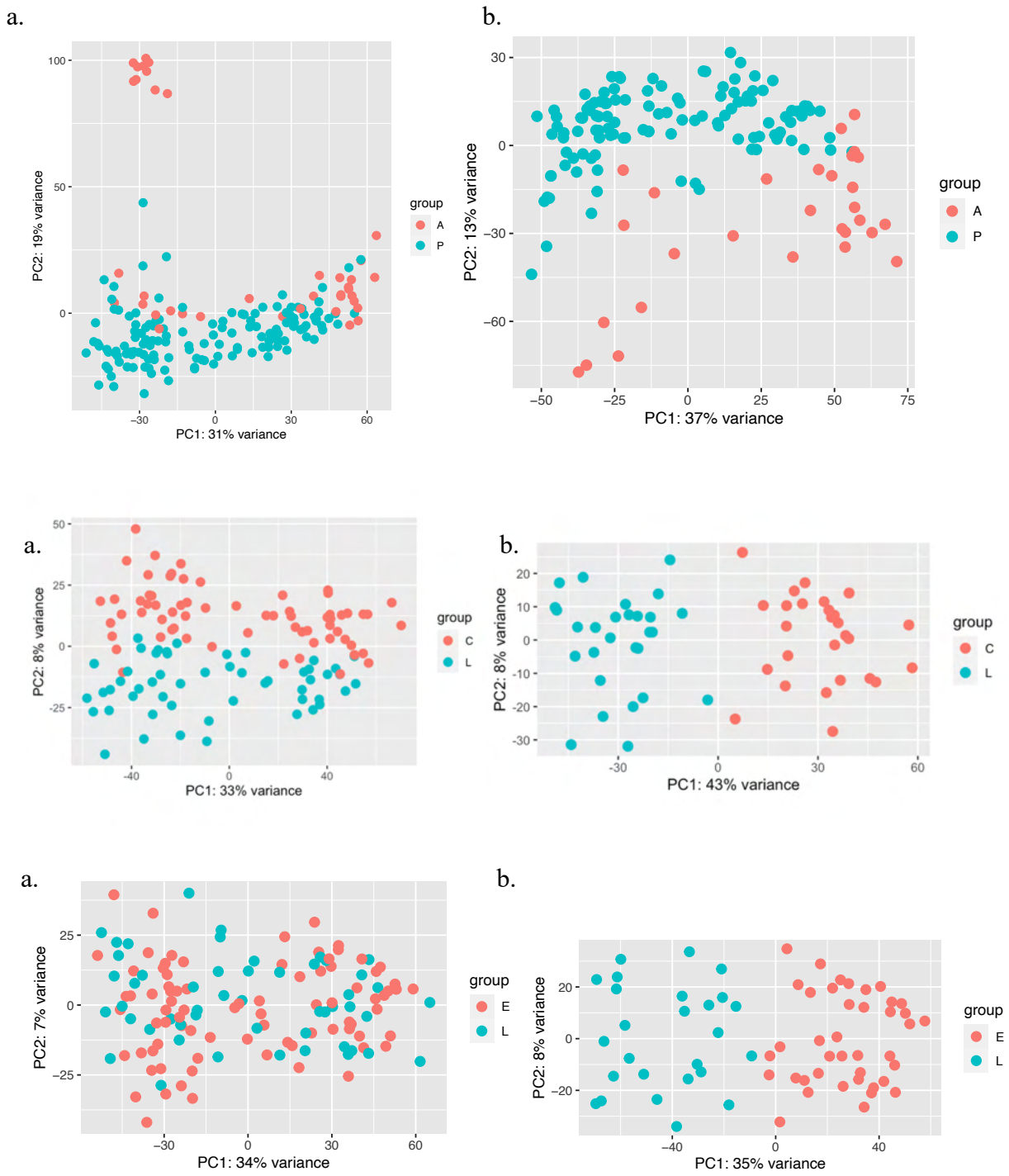


Figure 10 | A PCA for the RNA-sequence counts for each dataset. Dataset 1: N = non tumorous and P = PDAC. a) prefilter PCA, P = 15 and N = 8, with outliers seen in region PCA1 < 0 and PCA2 < -20 and b) post filter PCA plot, P = 10 and A = 6. Dataset 2 (1): A = adjacent normal tissue (39) and P = PDAC (140). Outliers in regions PCA2 >50 were removed. b) post filtering PCA plot, P = 116 and N = 29. Dataset 2 (2) C = classical subtype and L = Basal-like subtype. Dataset 2 (3) E = early stage and L = late stage. Plots produced by R-Studio DESeq2 Bioconductor package 3.13.

Dataset 1 produced a good PCA plot following clean-up, however, the sample numbers had reduced to an extent whereby co-expression analysis could not be performed. Dataset 2 was downloaded in a count data matrix format, hence pre-processing and alignment had been performed already and inputted for co-expression analysis due to a large sample set. Following the visualisation of the PCA plots per sample condition, the number of upregulated and downregulated DEGs were identified through DESeq2, as seen in *Fig. 11*.

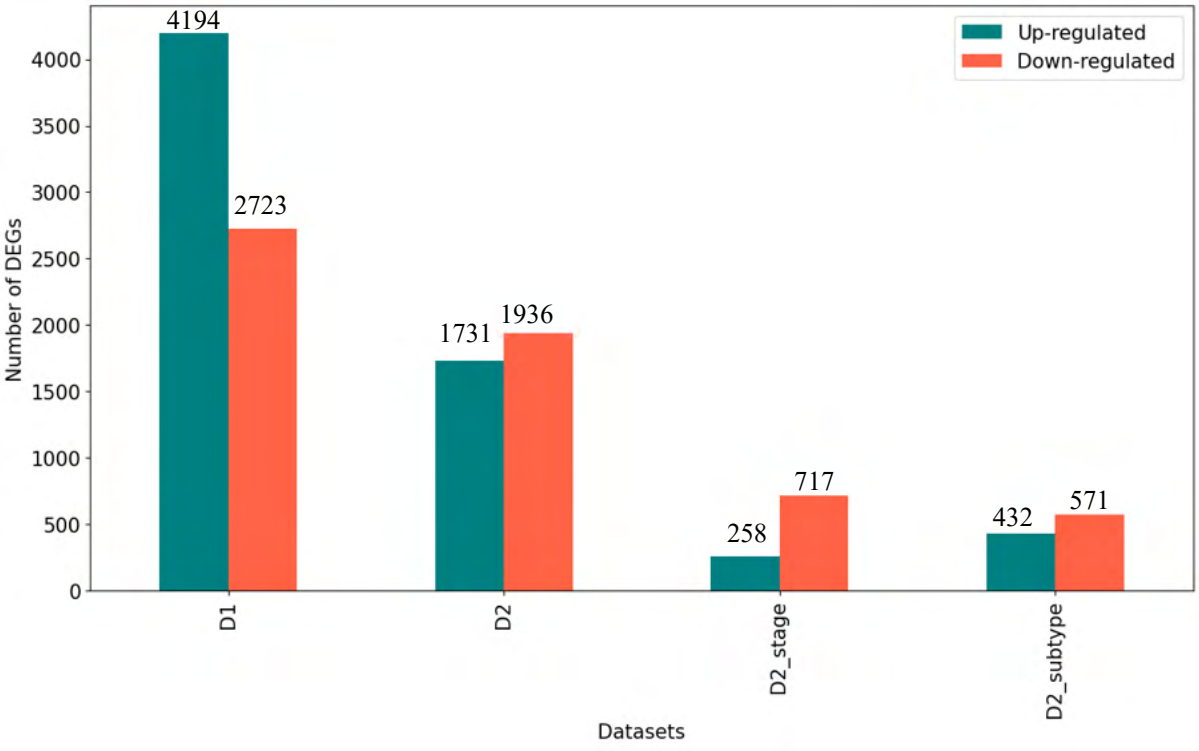


Figure 11| Graph showing the number of upregulated and downregulated DEGs (y-axis) for each dataset (x-axis). Green bars represent the upregulated genes whereas orange bars are the downregulated genes. Dataset 2 produced no DEGs. Plot produced by Python Google collab.

4.2.1. Tumour versus normal DESeq analysis

Dataset 1 and 3 both contained PDAC and normal samples, and therefore a Venn diagram was made to identify the common DEGs present in the two datasets (*Fig. 12*) . As the two datasets were sequenced independently, the common identified DEGs may represent universal biomarkers, irrespective of race, gender and ethnicity.

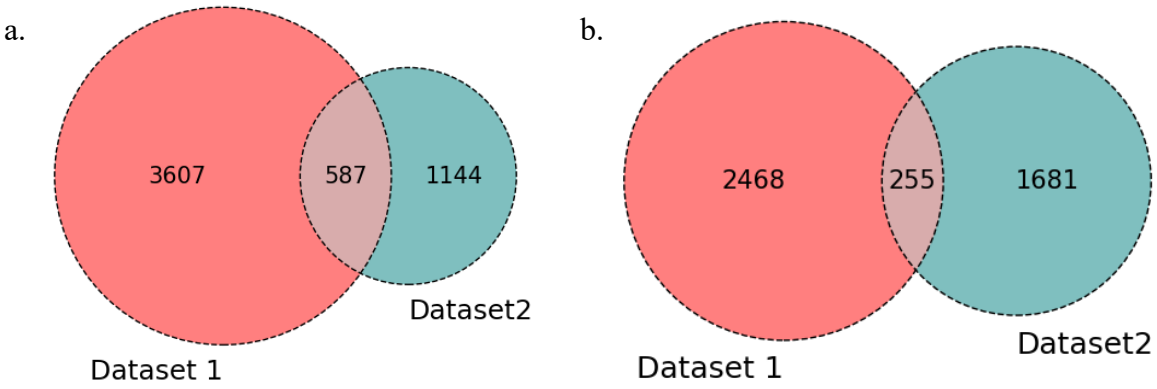


Figure 12 | Venn diagrams showing the common upregulated tumour DEGs observed in both dataset 1 and 3 for the same comparison condition. a) common upregulated DEGs (587) and b) common downregulated DEGs (255). Diagram produced by Python Google collab.

The DEGs had a criteria of P values < 5%. A selection was performed again for the top 5% of the significant DEGs. This provides a further 95% interval confidence rate in the selected DEGs. The significant DEGs, seen in *Fig. 13*, were taken for downstream analysis. DESeq2 analysis was focused on the upregulated genes to understand how gene amplification contributes to high protein expression levels and the roles in cell growth, proliferation and metastasis.

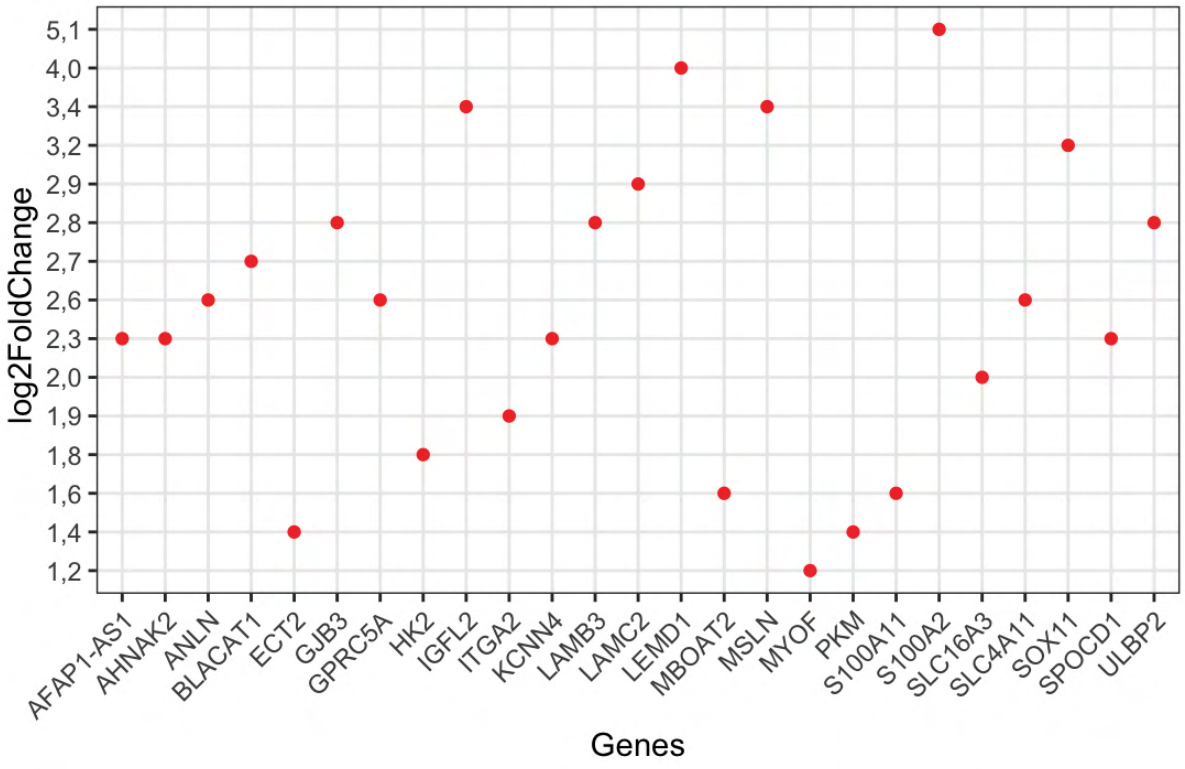
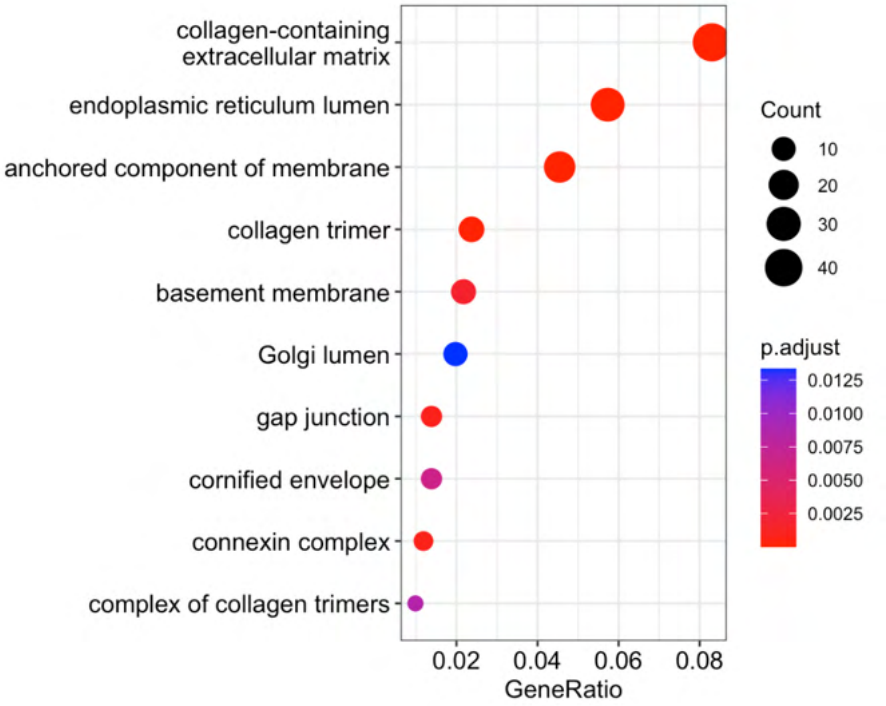


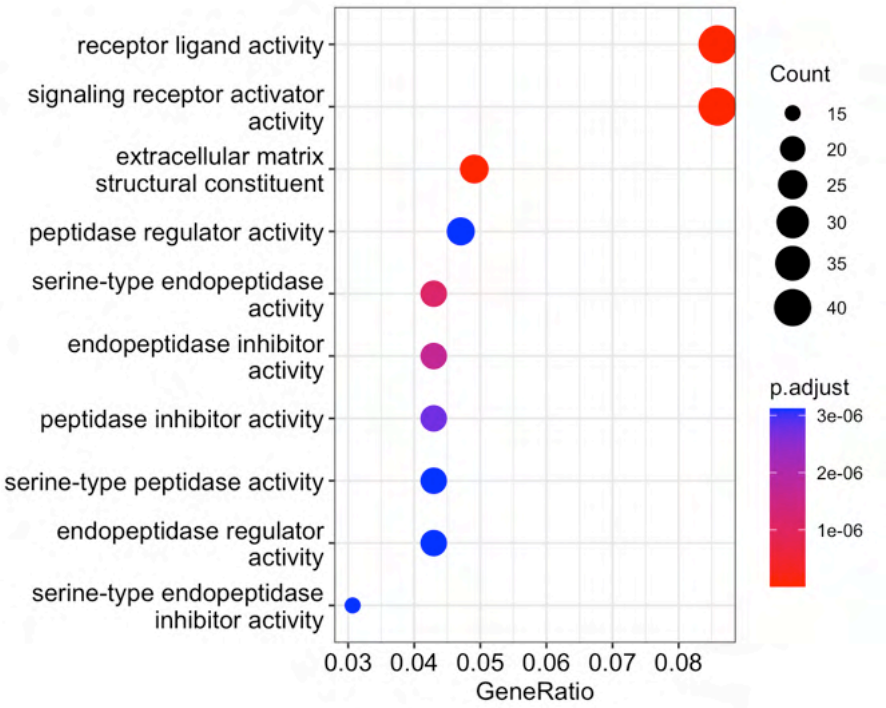
Figure 13 | The highly significant upregulated tumour DEGs observed dataset 1 and 2 for the tumour samples and the corresponding log_2FC. Plots was produced by R-Studio DESeq2 Bioconductor package 3.13.

Following the identification and selection of the DEGs, ClusterProfiler was used to determine the enrichment classification for the cellular component, molecular function, biological processes and lastly, KEGG pathways. This shows the link between the DEGs and its role in the biological system. The graphs in *Fig. 14* show the corresponding functions per counts (the number of genes in a circle).

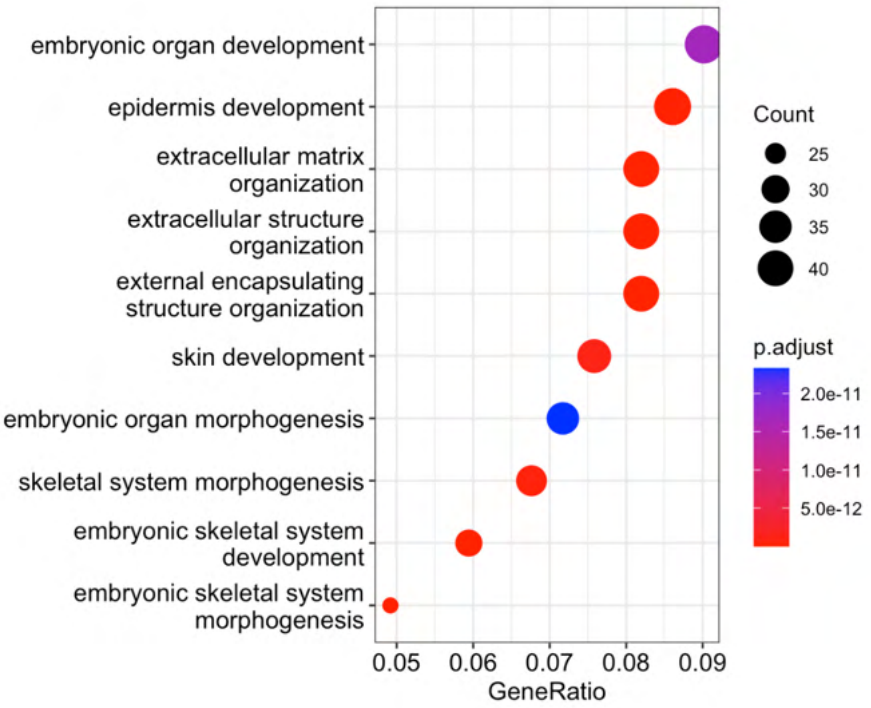
a.


collagen-containing extracellular matrix
endoplasmic reticulum lumen
anchored component of membrane
collagen trimer
basement membrane
Golgi lumen
gap junction
cornified envelope
connexin complex
complex of collagen trimers
0.02
0.04
0.06
0.08
GeneRatio
Count
10
20
30
40
p.adjust
0.0125
0.0100
0.0075
0.0050
0.0025

b.


receptor ligand activity
signaling receptor activator activity
extracellular matrix structural constituent
peptidase regulator activity
serine-type endopeptidase activity
endopeptidase inhibitor activity
peptidase inhibitor activity
serine-type peptidase activity
endopeptidase regulator activity
serine-type endopeptidase inhibitor activity
0.03
0.04
0.05
0.06
0.07
0.08
GeneRatio
Count
15
20
25
30
35
40
p.adjust
3e-06
2e-06
1e-06

c.



d.

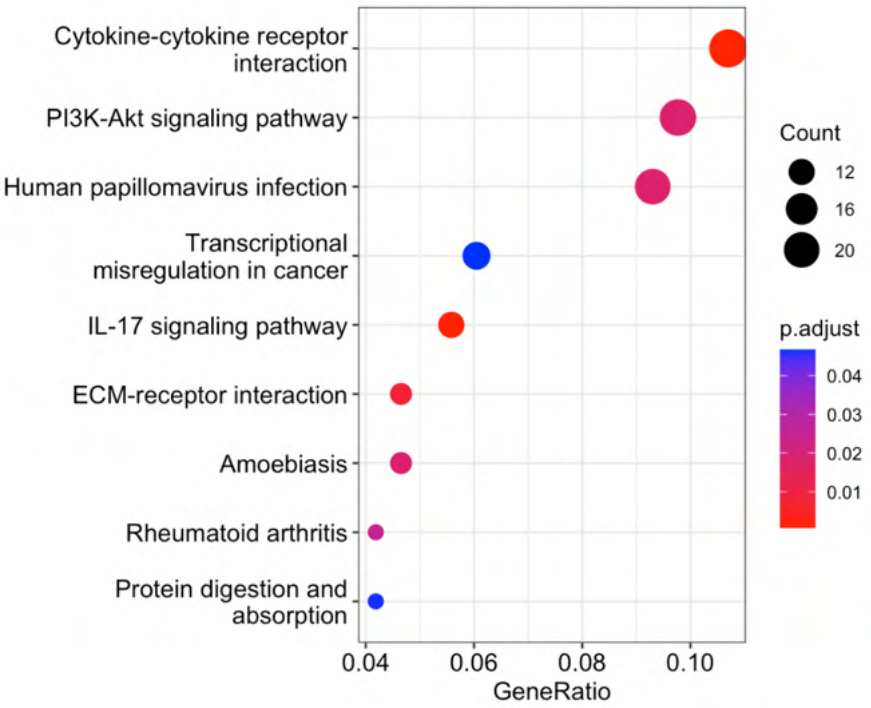


Figure 14 | Enrichment for all of the upregulated significant common DEGs in the tumour sample mapped against the gene ratio, displaying the percentage of DEGs in the given GO term. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set per GO term. The dot colour shows the statistical significance of gene whereby red is highly significant and blue is significant. a) An enrichGO classification for the cellular component profiles, b) An enrichGO molecular function,

c) An enrichGO biological process and d) An enrich KEGG functional process for the DEGs. Plots produced by R ClusterProfiler version 4.5.1.

Kaplan- Meier survival curves were generated for all of the selected DEGs. This was performed as the first *in-silico* validation step. Survival graphs with insignificant P-values were excluded from downstream analysis, stated in Appendix B, *Fig. 54*. Additionally, DEGs with curves that did not support the compared conditions were discarded. An average graph with supportive curves and significant P-values is seen in *Fig. 15*. Based on this graph, the DEGs had a low survival rate when upregulated with a highly significant P-value in the PDAC sample.

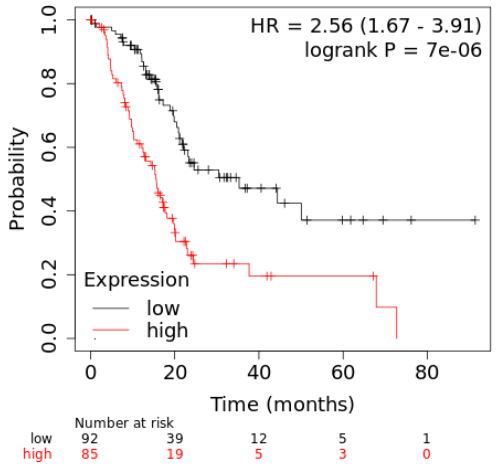


Figure 15 | Average Kaplan-Meier survival curve for the common upregulated tumour DEGs, following the removal of outliers and insignificant P-value graphs. Graph show the probability of survival versus the time (months). HR serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Individual DEG graphs can be seen in Appendix C. Graph produced by Kaplan-Meier plotter version 2022.08.23.

The Kaplan-Meier plots condensed the list of the DEGs. This was followed by the use of an Expression Atlas map. The Gene Expression Atlas, a resource of EMBL-EBI (https://www.ebi.ac.uk/), was made to visualise gene and protein expression levels in multiple species. This open access database includes information about RNA expression levels in biological systems for different types of diseases. Expression levels per gene can be observed

in a map for a disease, in this case, PDAC. The DEGs with supportive survival curves were used to produce an expression map, *Fig. 16,* representing the next *in-silico* validation step.

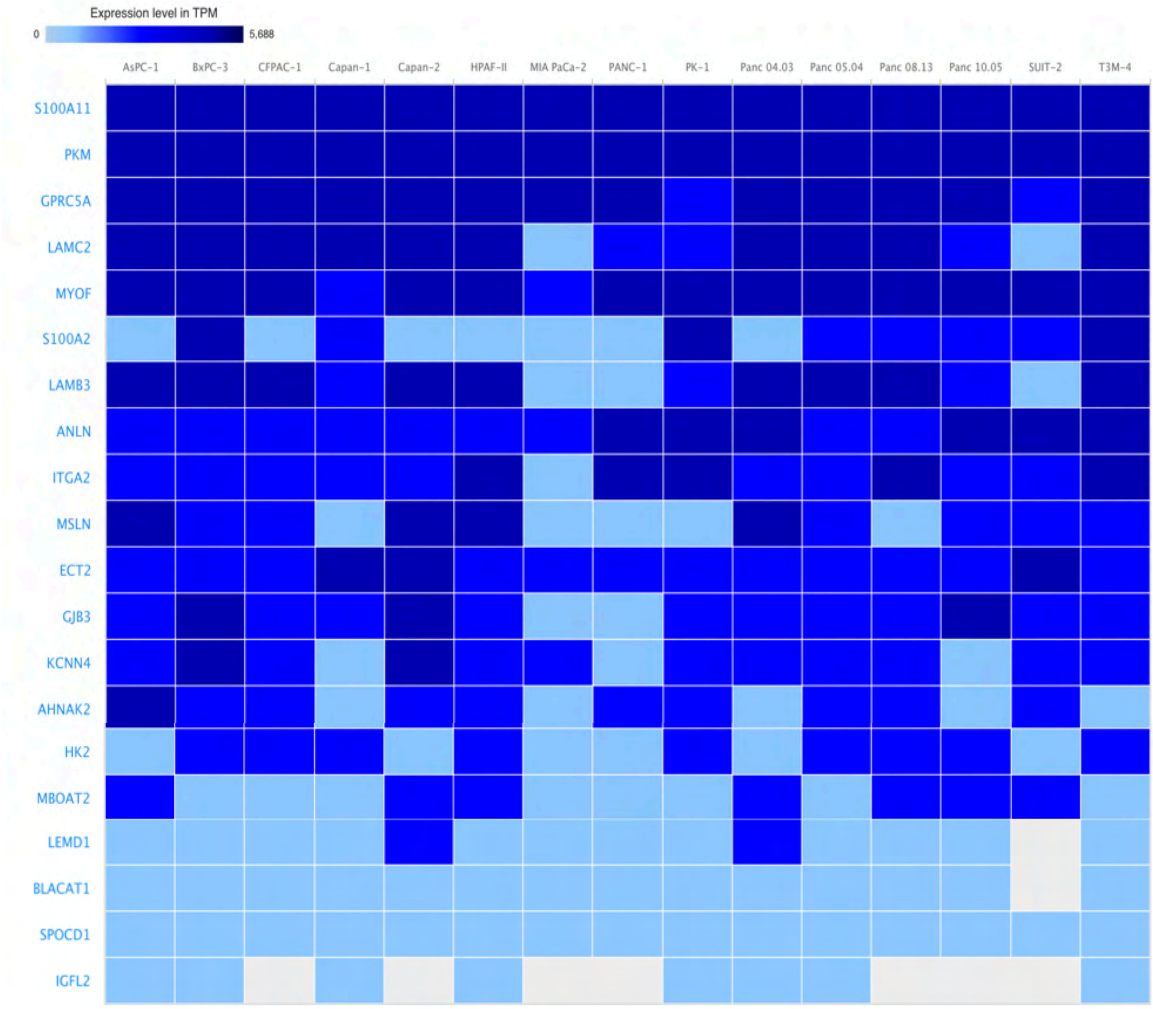
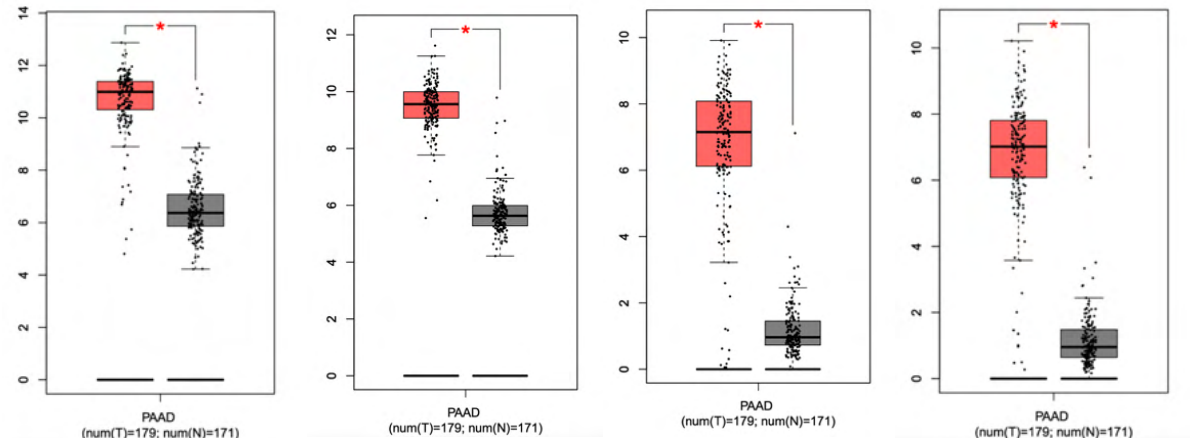


Figure 16 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated tumour DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. There was no expression level for : SOX11 and SLC16A3. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).

The DEGs with high expression levels in the map, *Fig. 16*, were used to produce expression box plots in from the GEPIA cancer database, *Fig. 17*. This validation step shows the gene expression level in a PDAC patient sample relative to the adjacent normal.



S100A11 **PKM** **GPR5A** **LAMC2**

Figure 17 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (*Fig.* 16) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).

Lastly, *Table 3* shows the identified signature genes and its expression values in the extracellular vesicles and the log_2FC from different databases, BBCancer and GENT2, respectively. The expression values per gene show if an upregulation (positive value) or downregulation (negative value) was observed. An upregulation in the extracellular vesicles may indicate migration whereas the GENT2 log_2FC values support the log_2FC output in this study (*Fig. 13*).

Table 3 | The BBCancer and GENT2 expression values (TPM) for the selected tumour DEGs.

GENE	BBCancer (Extracellular vesicles)	GENT2 Log_2FC
GPRC5A	- 0.12	2.56
ITGA2	2.48	1.32
LAMB3	- 0.12	1.87
LAMC2	- 0.27	2.08
LEMD1	- 0.66	1.35
PKM	2.36	1.07

S100A11	3.67	1.29

4.2.2. Early versus late stage DESeq2 analysis

The tumour samples were divided based on the stage of diseases whereby stage I and II were classified as the early stage and stage III and IV were classified as the late stage. DESeq2 analysed the count matrix to produce the upregulated 5% of DEGs with significant P-values observed in the late stage of PDAC, *Fig. 18*.

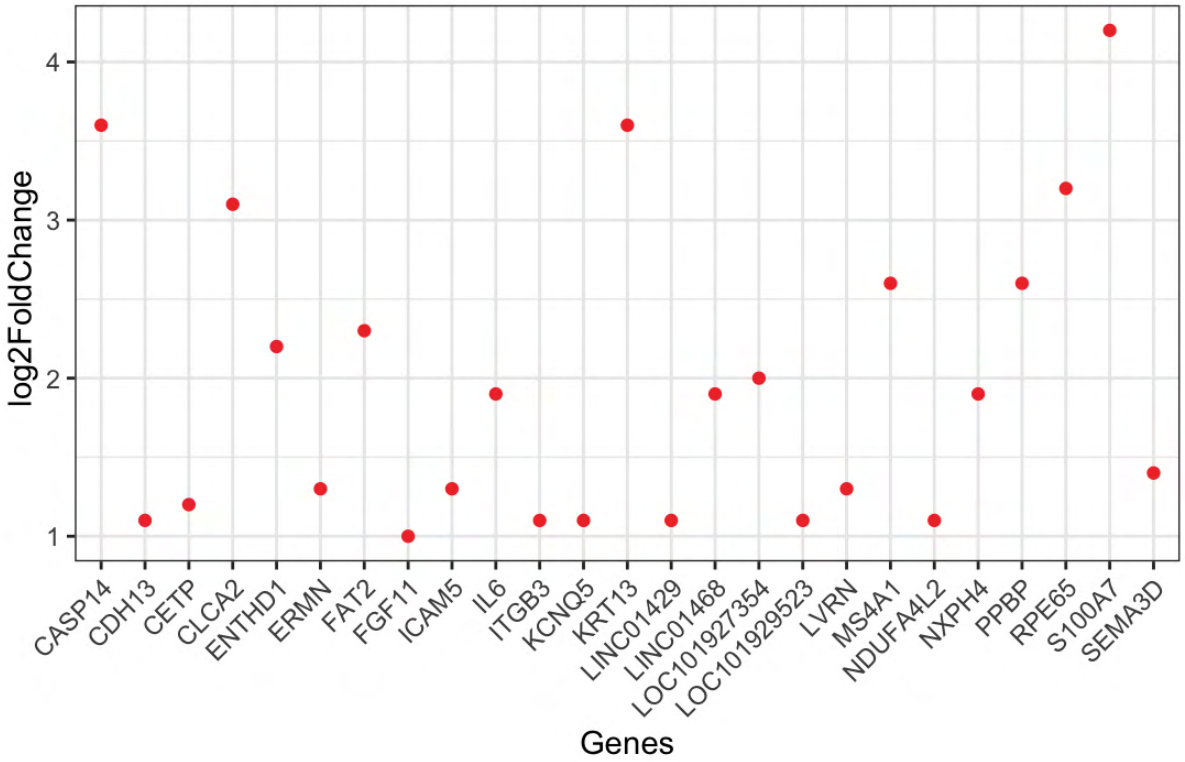
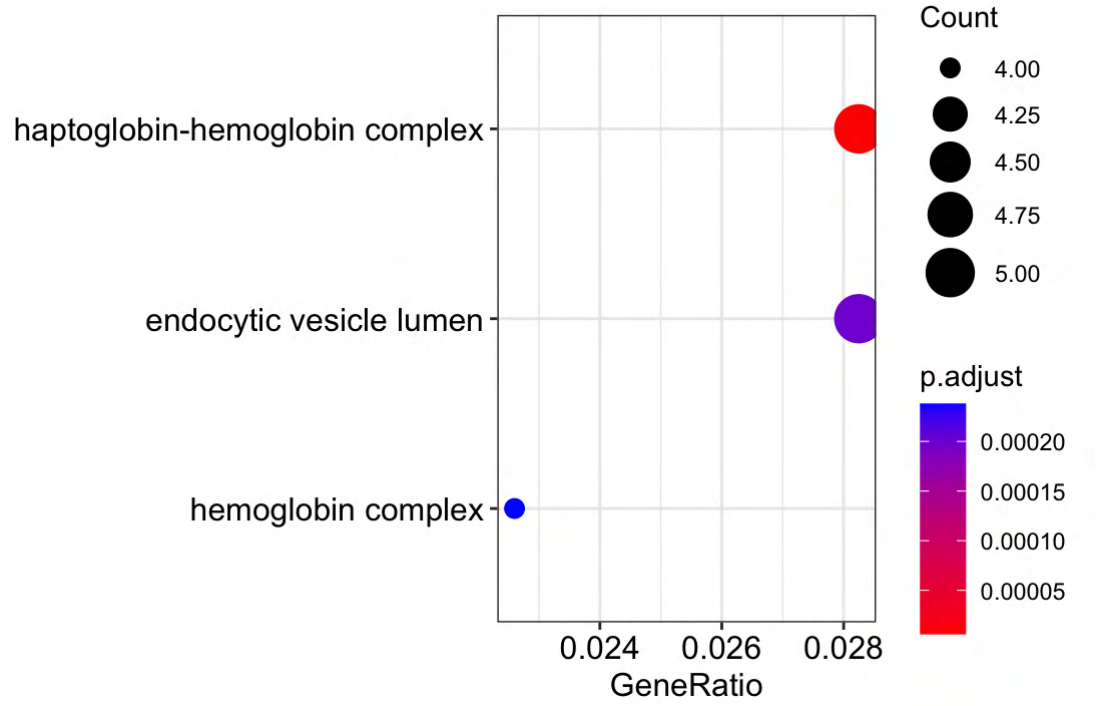


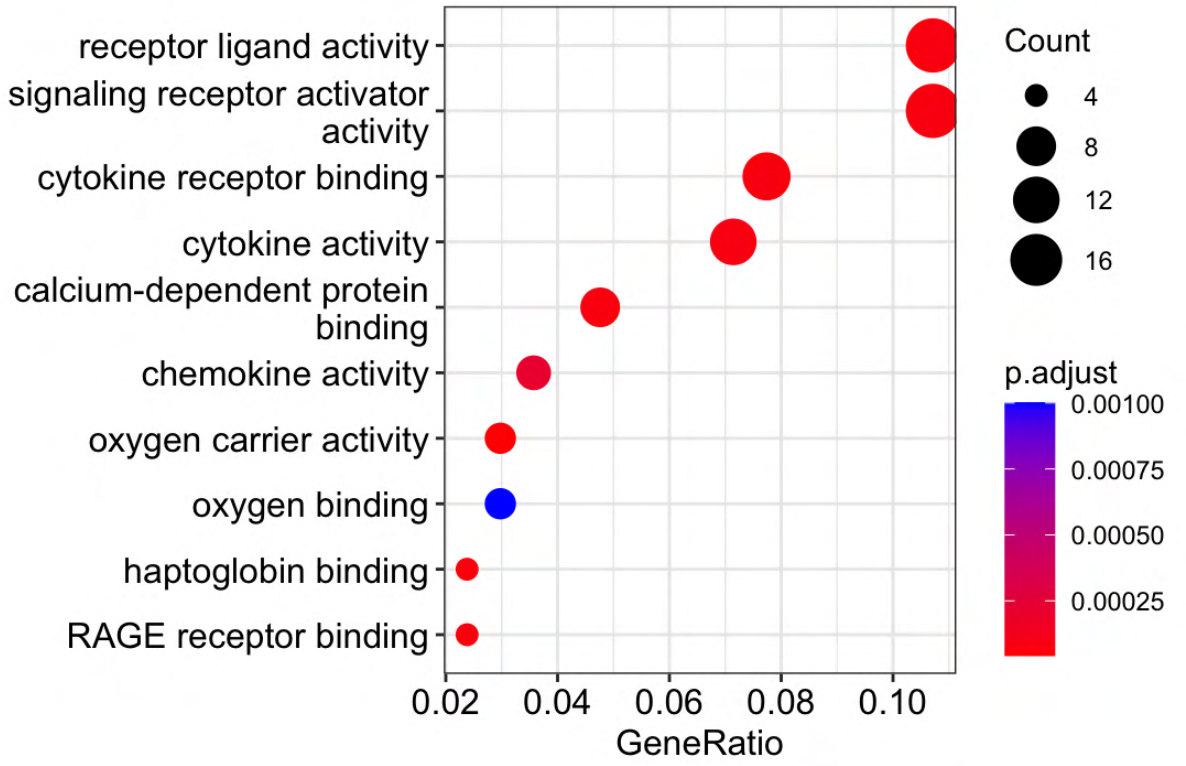
Figure 18 | The common highly significant upregulated genes observed in the late stage of PDAC. The genes are mapped with the corresponding log2FC. Plots was produced by R-Studio DESeq2 Bioconductor package 3.13.

ClusterProfiler was then used to determine the enrichment classification for the cellular component, molecular function, biological processes and lastly, KEGG pathways. This shows the link between the DEGs and its role in the biological system. The graphs in *Fig. 19* show the corresponding functions per counts (the number of genes in a circle).

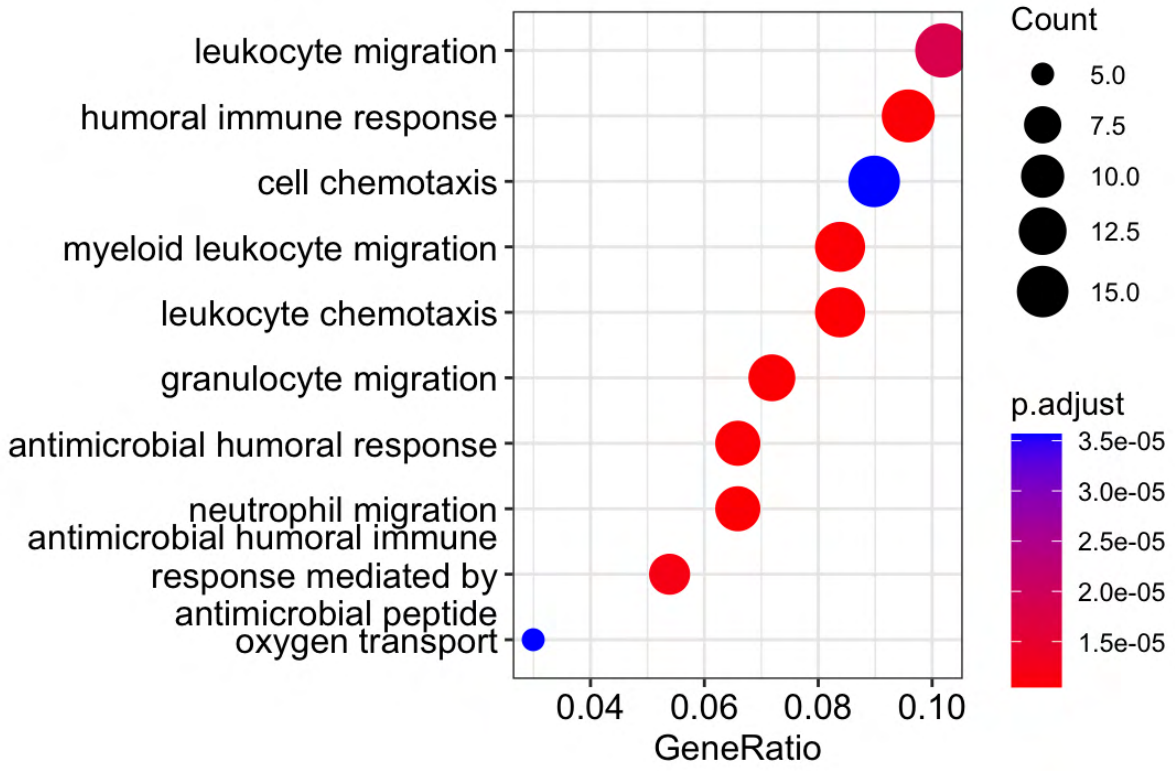
a.


haptoglobin-hemoglobin complex
endocytic vesicle lumen
hemoglobin complex
0.024
0.026
0.028
GeneRatio
Count
4.00
4.25
4.50
4.75
5.00
p.adjust
0.00020
0.00015
0.00010
0.00005

b.


receptor ligand activity
signaling receptor activator activity
cytokine receptor binding
cytokine activity
calcium-dependent protein binding
chemokine activity
oxygen carrier activity
oxygen binding
haptoglobin binding
RAGE receptor binding
0.02
0.04
0.06
0.08
0.10
GeneRatio
Count
4
8
12
16
p.adjust
0.00100
0.00075
0.00050
0.00025

c.


leukocyte migration
humoral immune response
cell chemotaxis
myeloid leukocyte migration
leukocyte chemotaxis
granulocyte migration
antimicrobial humoral response
neutrophil migration
antimicrobial humoral immune response mediated by antimicrobial peptide
oxygen transport
0.04
0.06
0.08
0.10
GeneRatio
Count
5.0
7.5
10.0
12.5
15.0
p.adjust
3.5e-05
3.0e-05
2.5e-05
2.0e-05
1.5e-05

d.

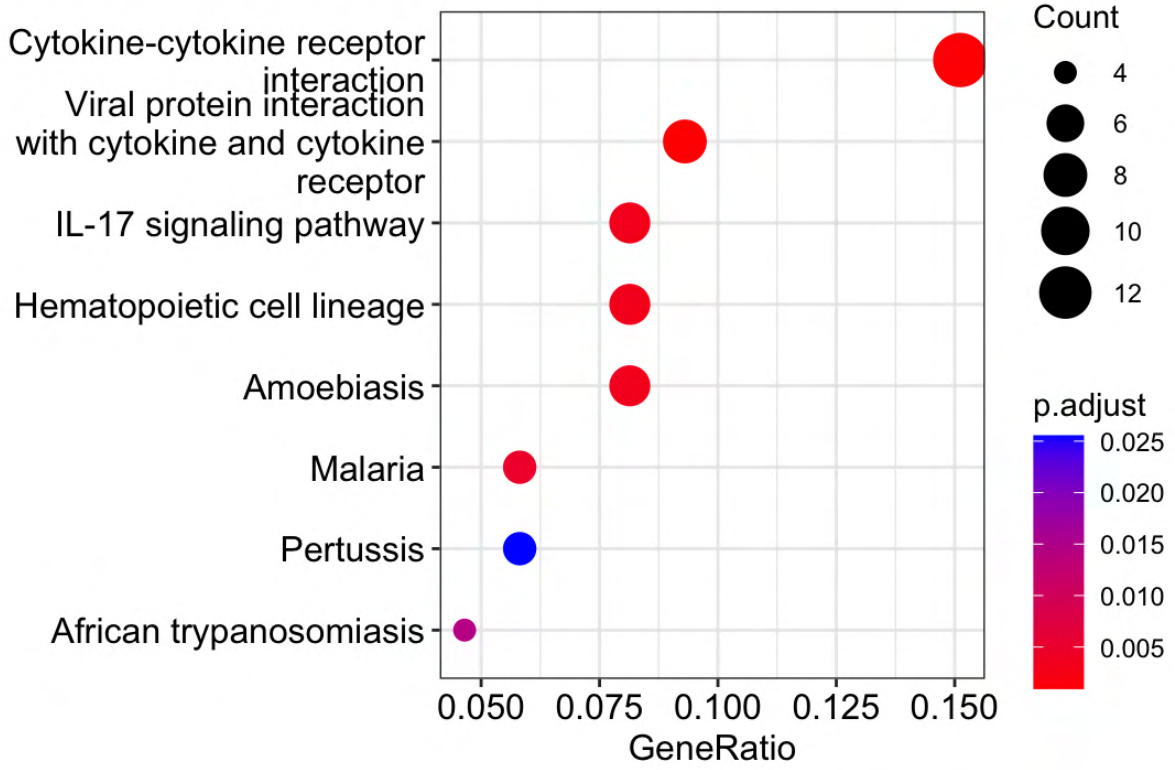

Cytokine-cytokine receptor interaction
Viral protein interaction with cytokine and cytokine receptor
IL-17 signaling pathway
Hematopoietic cell lineage
Amoebiasis
Malaria
Pertussis
African trypanosomiasis
0.050
0.075
0.100
0.125
0.150
GeneRatio
Count
4
6
8
10
12
p.adjust
0.025
0.020
0.015
0.010
0.005

Figure 19 | Enrichment for all of the upregulated significant DEGs in the late stage of PDAC samples mapped against the gene ratio, displaying the percentage of DEGs in the given GO term. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set per GO term. The dot colour shows the statistical significance of gene whereby red is highly significant and blue is significant. a) An enrichGO classification for the cellular component profiles, b) An enrichGO molecular function, c) An enrichGO biological process and d) An enrich KEGG functional process for the DEGs. Plots produced by R ClusterProfiler version 4.5.1.

Kaplan- Meier survival curves were generated for all of the selected DEGs. This was performed for the late stage of PDAC. Survival graphs with insignificant P-values were excluded from downstream analysis. Additionally, DEGs with curves that did not support the compared conditions were discarded. An average graph with supportive curves and significant P-values is seen in *Fig. 20*. Based on this graph, the DEGs had a much lower survival rate when upregulated in the late stage with a highly significant P-value (0.00013).

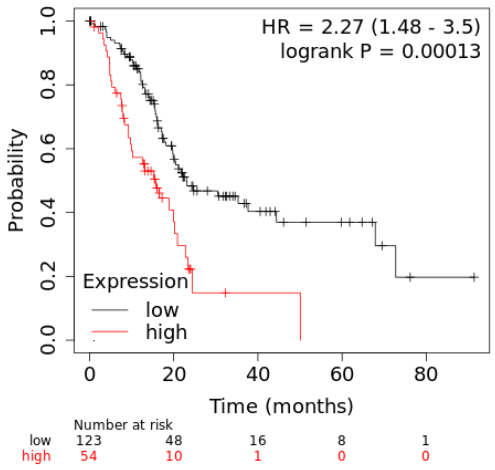


Figure 20 | Average Kaplan-Meier survival curve for the upregulated late stage DEGs, following the removal of outliers and insignificant P-value graphs. Graph show the probability of survival versus the time (months). HR serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Individual DEG graphs can be seen in Appendix C. Graph produced by Kaplan-Meier plotter version 2022.08.23.

Next, the Expression Atlas showed the levels per gene in a map for the DEGs in the late stage of PDAC. The DEGs with supportive survival curves were used to produce an expression map, *Fig. 21.*

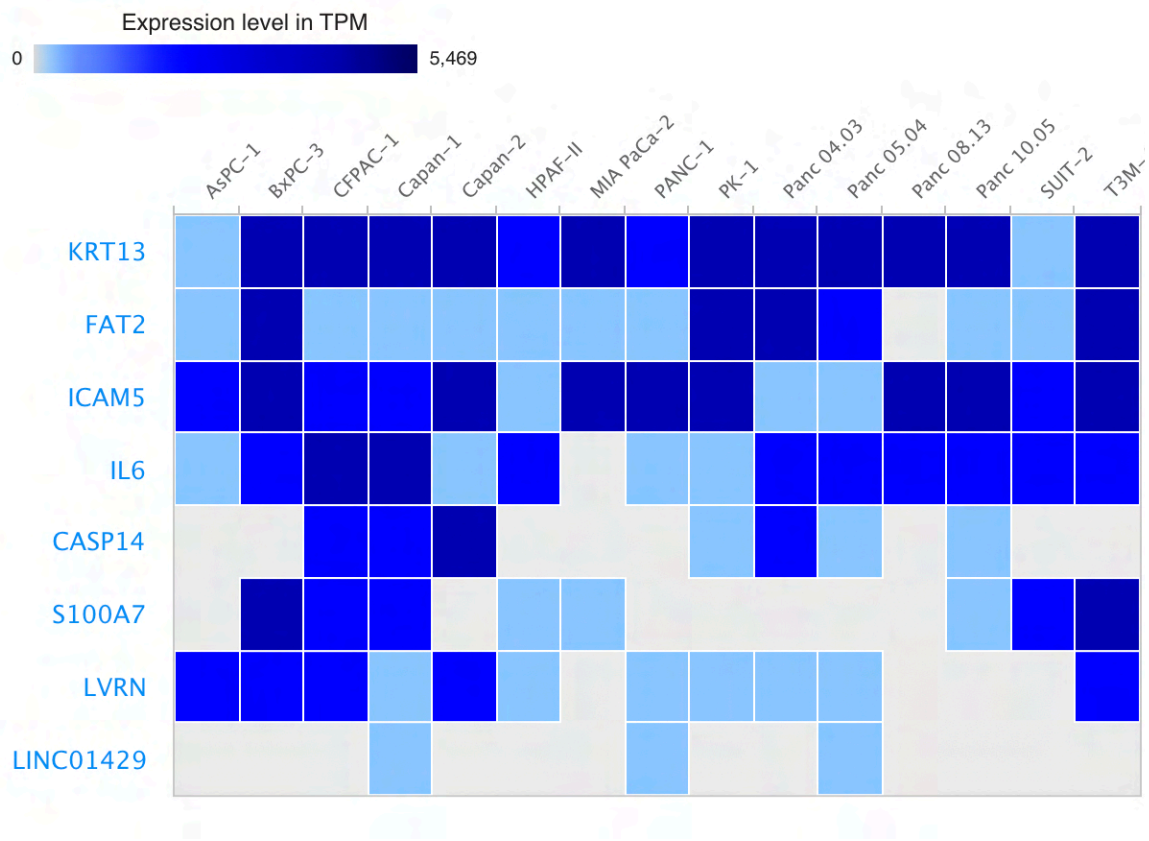
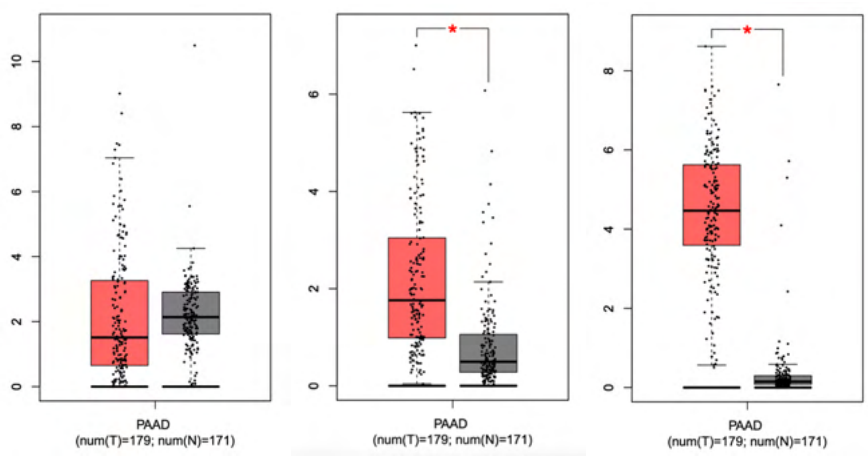


Figure 21 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated late stage DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).

The late stage DEGs with high expression levels in the map, *Fig. 21* were used to produce expression box plots in from the GEPIA cancer database, *Fig. 22*. This validation step shows the gene expression level in a PDAC patient sample relative to the adjacent normal.



KRT13 IL6 MMP9

Figure 22 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (*Fig. 21*) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).

Table 4 shows the identified signature genes in the late stage and its expression values in the extracellular vesicles and the log_2FC from different databases, BBCancer and GENT2, respectively. An upregulation in the extracellular vesicles may indicate migration whereas the GENT2 log_2FC values support the log_2FC output in this study (*Fig. 18*).

Table 4 | The BBCancer and GENT2 expression values (TPM) for the selected late stage DEGs.

GENE	**BBCancer (Extracellular vesicles)**	**GENT2 Log_2FC**
KRT13	- 0.87	0.70
IL6	- 1.18	- 0.44
S100A7	-	0.47

4.2.3. Classical subtype versus basal-like subtype

The last comparison of DESeq2 involved the classical subtype versus the basal-like subtype. The tumour samples were divided based on the subtype. This step allowed for an in-depth analysis as to why the basal-like display resistance to treatment and observed in the metastatic stage of disease. Once again, the top 5% of the significant DEGs were selected based on the P-values. *Fig. 23* show the genes with the corresponding log_2FC values.

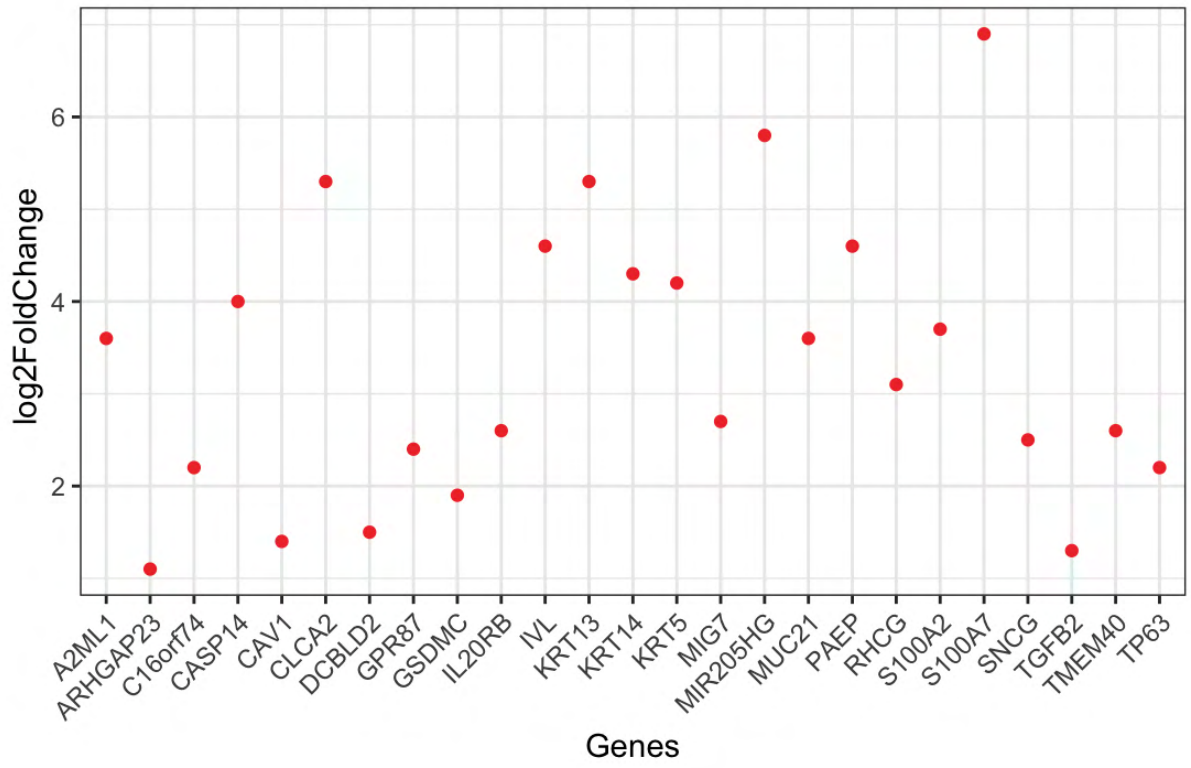
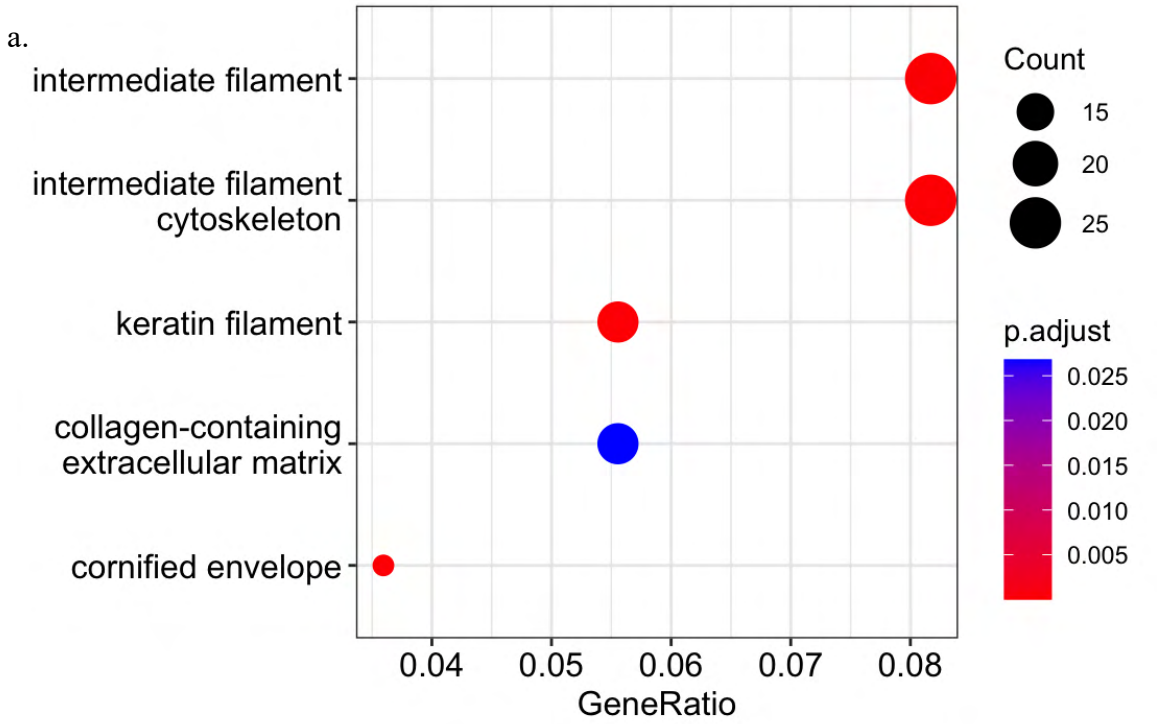


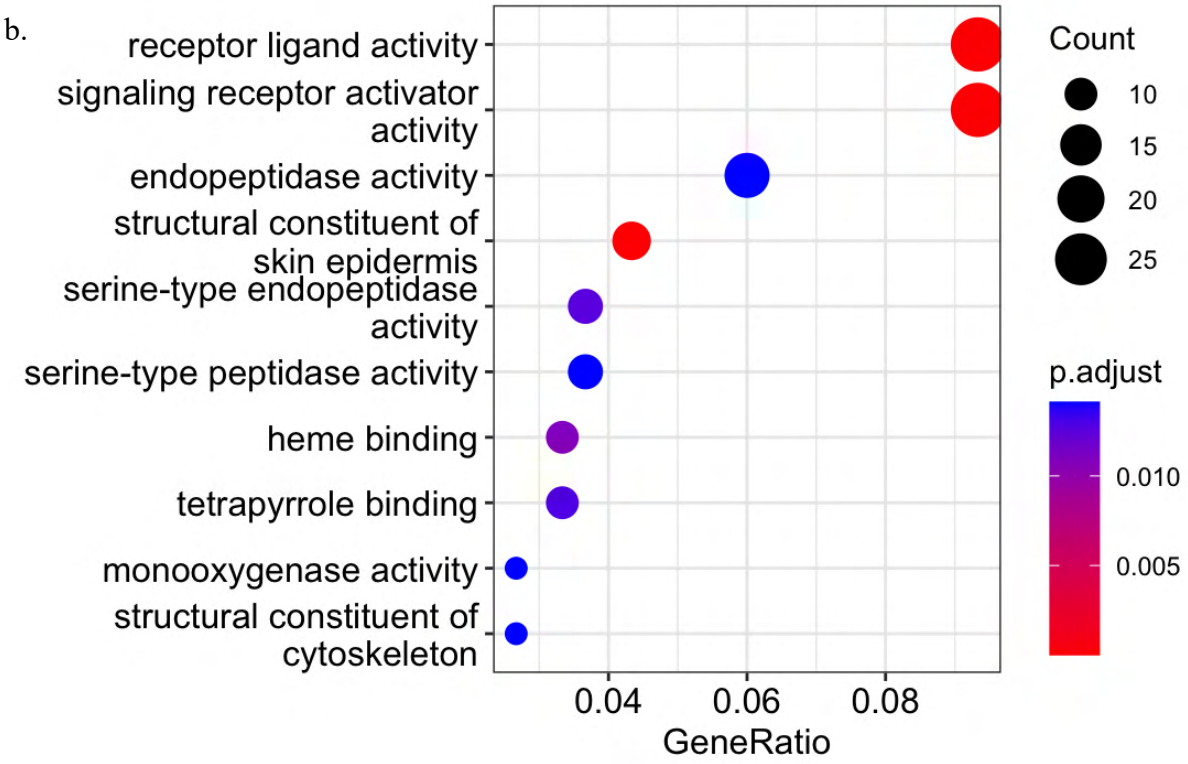
Figure 23 | The highly common significant upregulated genes observed in the basal-like subtype of PDAC. The genes are mapped with the corresponding log2FC. Plots was produced by R-Studio DESeq2 Bioconductor package 3.13.

ClusterProfiler was used to determine the enrichment classification for the cellular component, molecular function, biological processes and lastly, KEGG pathways. This shows the link between the DEGs and its role in the biological system. The graphs in *Fig. 24* show the corresponding functions per counts (the number of genes in a circle).

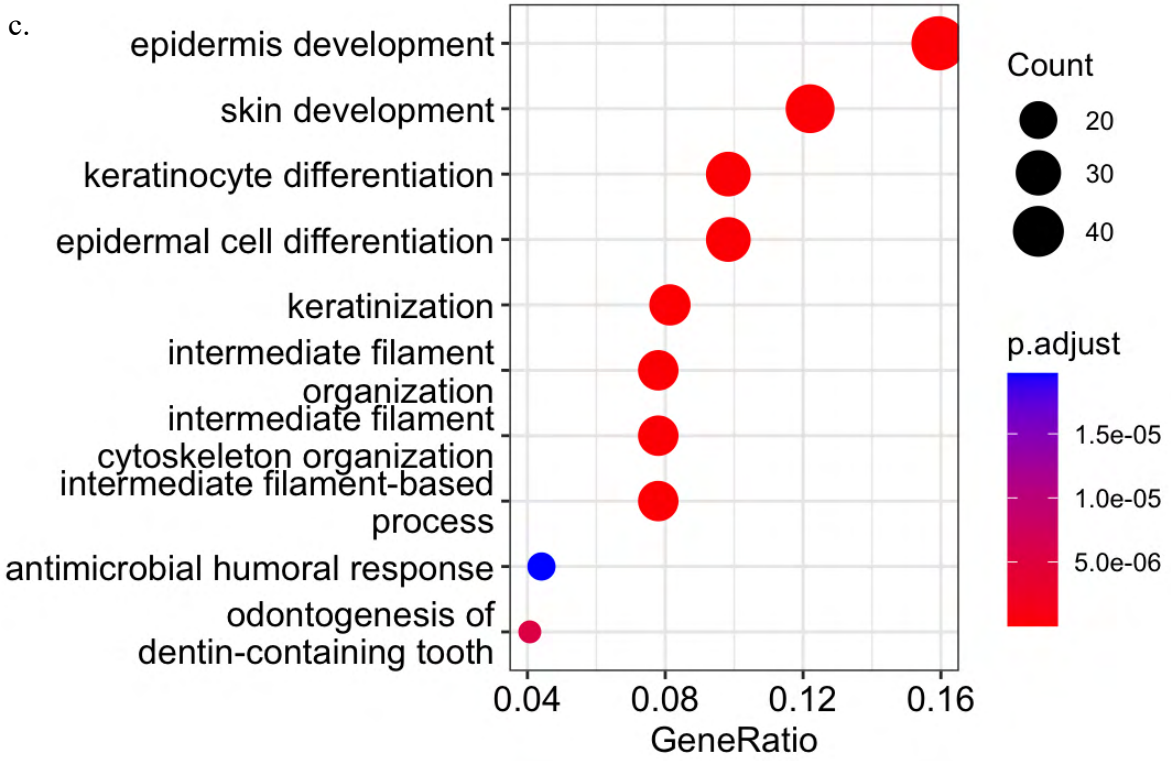
a.


intermediate filament
intermediate filament cytoskeleton
keratin filament
collagen-containing extracellular matrix
cornified envelope
0.04
0.05
0.06
0.07
0.08
GeneRatio
Count
15
20
25
p.adjust
0.025
0.020
0.015
0.010
0.005

b.


receptor ligand activity
signaling receptor activator activity
endopeptidase activity
structural constituent of skin epidermis
serine-type endopeptidase activity
serine-type peptidase activity
heme binding
tetrapyrrole binding
monooxygenase activity
structural constituent of cytoskeleton
0.04
0.06
0.08
GeneRatio
Count
10
15
20
25
p.adjust
0.010
0.005

c.


epidermis development
skin development
keratinocyte differentiation
epidermal cell differentiation
keratinization
intermediate filament organization
intermediate filament cytoskeleton organization
intermediate filament-based process
antimicrobial humoral response
odontogenesis of dentin-containing tooth
0.04
0.08
0.12
0.16
GeneRatio
Count
20
30
40
p.adjust
1.5e-05
1.0e-05
5.0e-06

d.

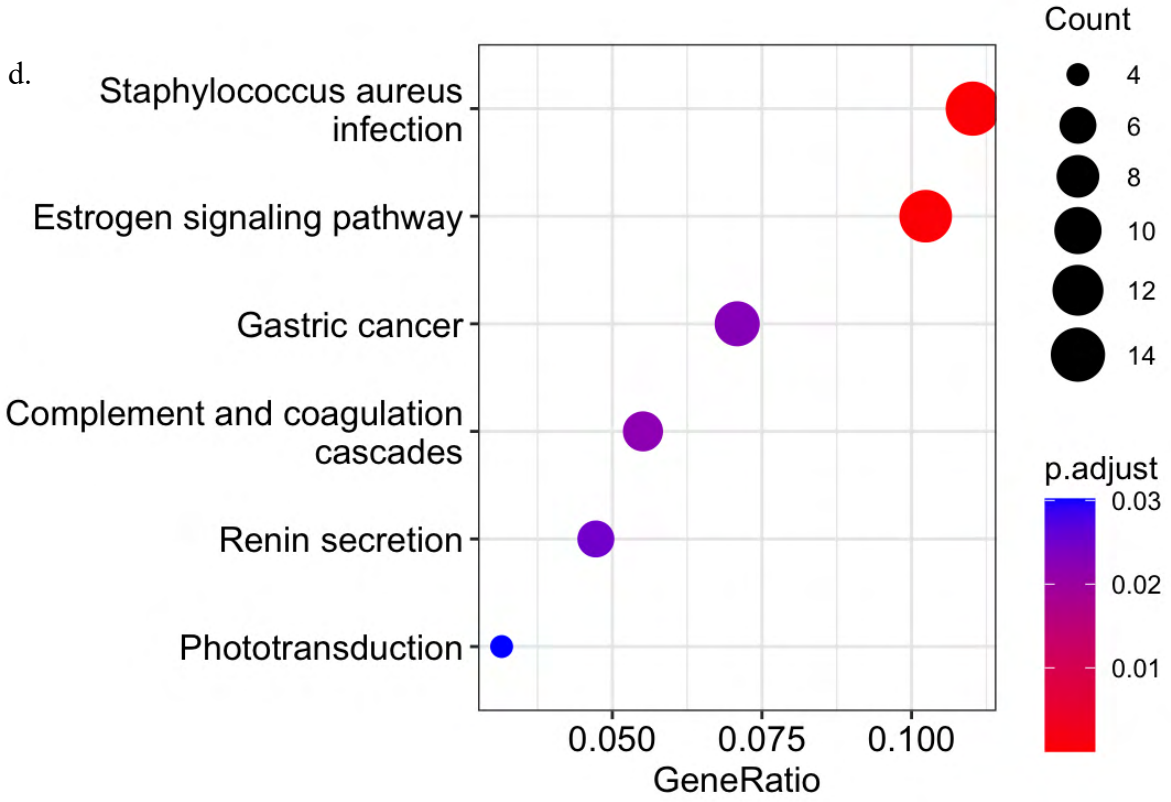

Staphylococcus aureus infection
Estrogen signaling pathway
Gastric cancer
Complement and coagulation cascades
Renin secretion
Phototransduction
0.050
0.075
0.100
GeneRatio
Count
4
6
8
10
12
14
p.adjust
0.03
0.02
0.01

Figure 24 | Enrichment for all of the upregulated significant DEGs in the basal-like subtype of PDAC samples mapped against the gene ratio, displaying the percentage of DEGs in the given GO term. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set per GO term. The dot colour shows the statistical significance of gene whereby red is highly significant and blue is significant. a) An enrichGO classification for the cellular component profiles, b) An enrichGO molecular function, c) An enrichGO biological process and d) An enrich KEGG functional process for the DEGs. Plots produced by R ClusterProfiler version 4.5.1.

Kaplan-Meier survival curves were generated for all of the selected DEGs for the basal-like subtype of PDAC. Survival graphs with insignificant P-values were excluded from downstream analysis. Additionally, DEGs with curves that did not support the compared conditions were discarded. An average graph with supportive curves and significant P-values is seen in *Fig. 25*. Based on this graph, the DEGs had a much lower survival rate when upregulated in the basal-like subtype with a highly significant P-value (0.0.001).

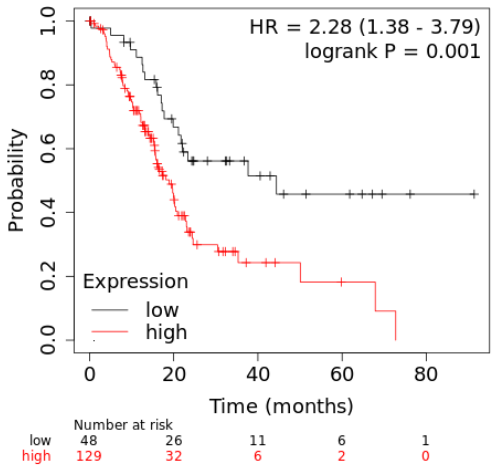


Figure 25 | Average Kaplan-Meier survival curve for the upregulated basal-like subtype DEGs, following the removal of outliers and insignificant P-value graphs. Graph show the probability of survival versus the time (months). HR serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Individual DEG graphs can be seen in Appendix C. Graph produced by Kaplan-Meier plotter version 2022.08.23.

The Expression Atlas showed the levels per gene in a map for the DEGs in the basal-like subtype of PDAC. The DEGs with supportive survival curves (*Fig. 25*) were used to produce an expression map, *Fig. 26.*

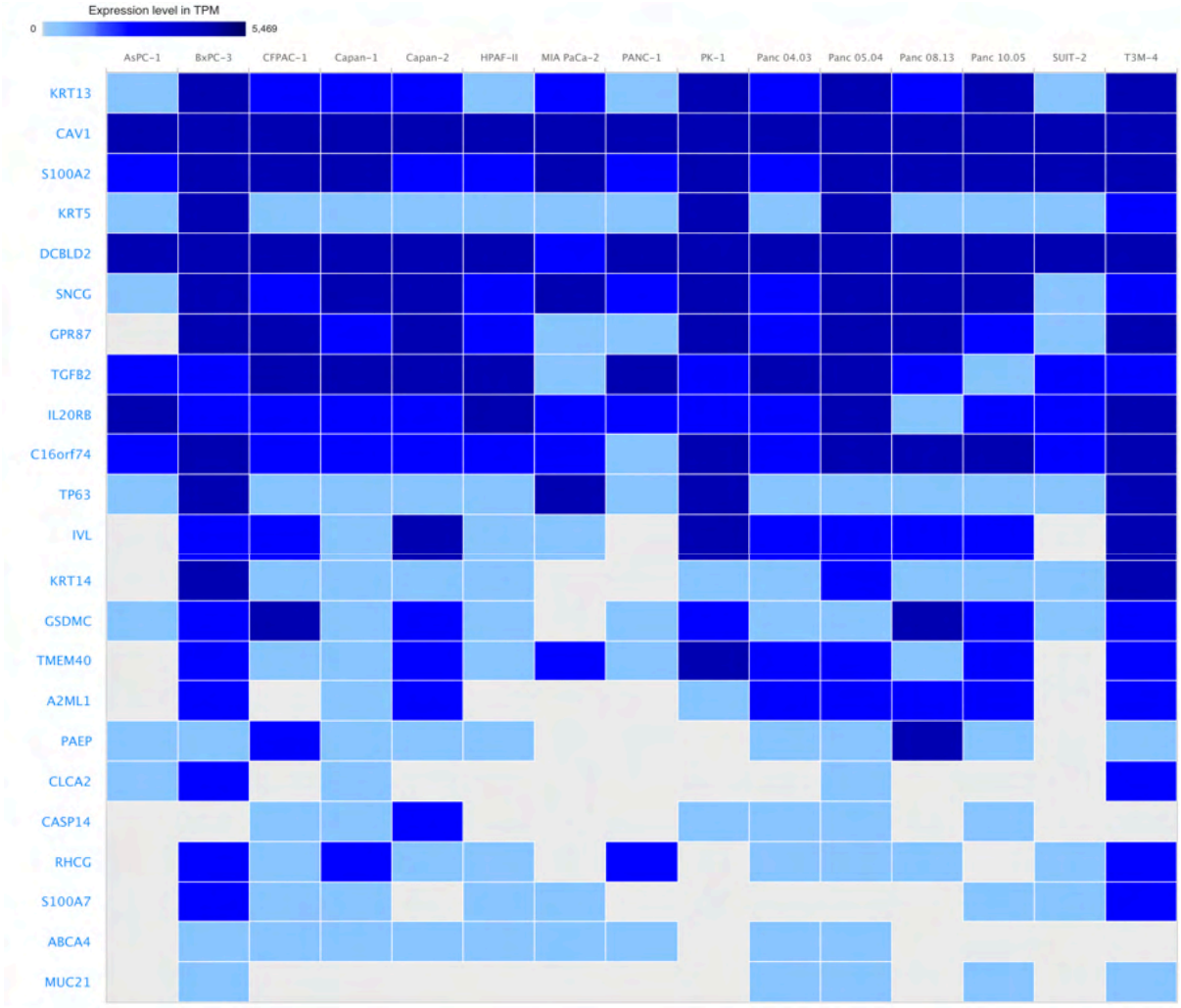
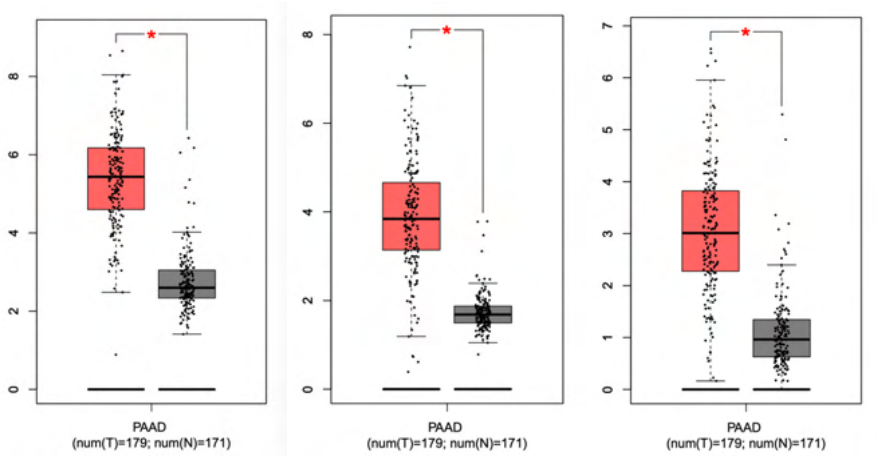


Figure 26 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated basal-like DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).

The basal-like subtype DEGs with high expression levels in the map, *Fig. 26*, were used to produce expression box plots in from the GEPIA cancer database, *Fig. 27*. This validation step shows the gene expression level in a PDAC patient sample relative to the adjacent normal.



CAV1 DCBLD2 TGFB2

Figure 27 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (*Fig. 26)* plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).

Table 5 shows the identified signature genes in the basal-like subtype and its expression values in the extracellular vesicles and the log2FC from different databases, BBCancer and GENT2, respectively. An upregulation in the extracellular vesicles may indicate migration whereas the GENT2 log_2FC values support the log_2FC output in this study (*Fig. 23*).

Table 5 | The BBCancer and GENT2 expression values (TPM) for the selected basal-like DEGs.

GENE	**BBCancer (Extracellular vesicles)**	**GENT2 Log_2FC**
DCBLD2	- 0.13	0.62
CAV1	1.15	0.23
TGFB2	0.27	0.61

4.3. WGCNA: Co-expression analysis of Dataset 2

Co-expression analysis was performed on dataset 2 as it met the sample number requirements for WCGNA analysis. The output of DESeq2, upregulated and downregulated DEGs were used to extract the count values from the original raw count data matrix which was then normalised through variance stabilising transformation. DEGs upregulated in the tumour samples are downregulated in normal samples and DEGs downregulated in the tumour samples are upregulated in the normal samples. Therefore, co-expression was performed on the upregulated tumour and normal samples separately. This allows for the identification of DEGs that are being co-expressed in a specific condition.

4. Tumour upregulated DEGs

In co-expression analysis, a soft threshold value (*Fig.28a*) is required to construct a cluster dendrogram (*Fig.28b*). A soft threshold value close to a 1 (y-axis, R^2) in the scale free topology model represents highly connected DEGs with low background noise, hence value selection was made based on this criteria.

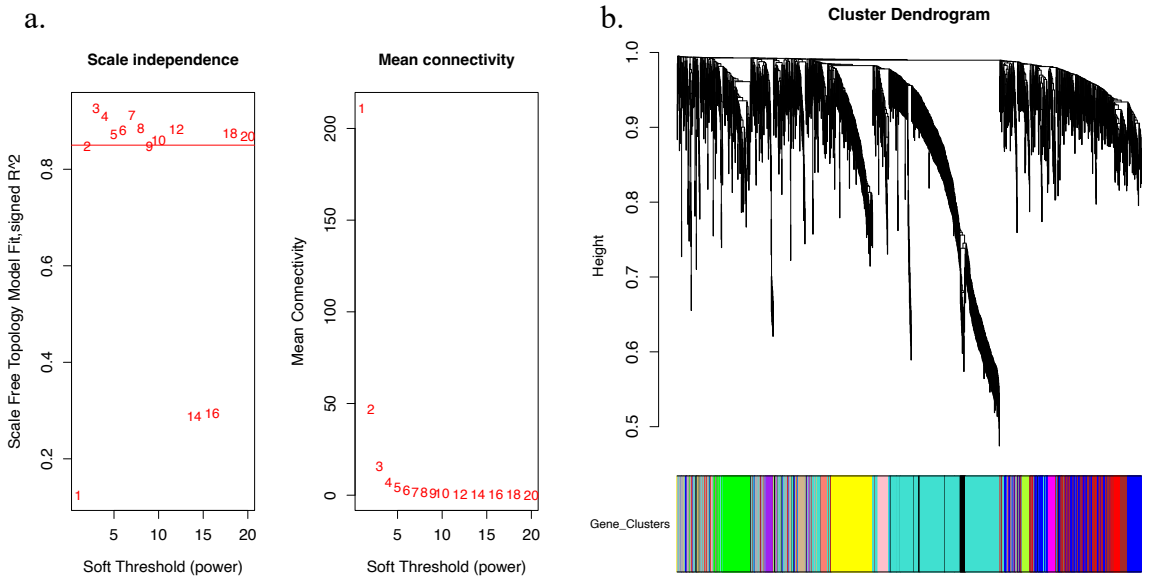


Figure 28 | WGCNA analysis for the upregulated tumour samples containing DEGs. a) Network topology for the scale independence and mean connectivity analysis of modules consisting of soft thresholding powers ranging between the values of 1 to 20. The scale independence plot consists of a y-axis which represents the scale-free topology fit as a function of the soft thresholding power (x-axis). The mean connectivity plot displays the degree (y-axis) as a function of the soft-thresholding power (x-axis). The power of 3 was selected for downstream analysis as it is the lowest power for which the

scale-free topology fit index reaches 0.90. b) Clustering dendrogram of genes among modules via hierarchical clustering of adjacency-based dissimilarity. Plots produced by R WGCNA version 1.71.

A network plot was made for the DEGs that are co-expressed in the same biological condition, *Fig. 29*. The lines between the nodes (genes) may represent potential interactions with the genes.

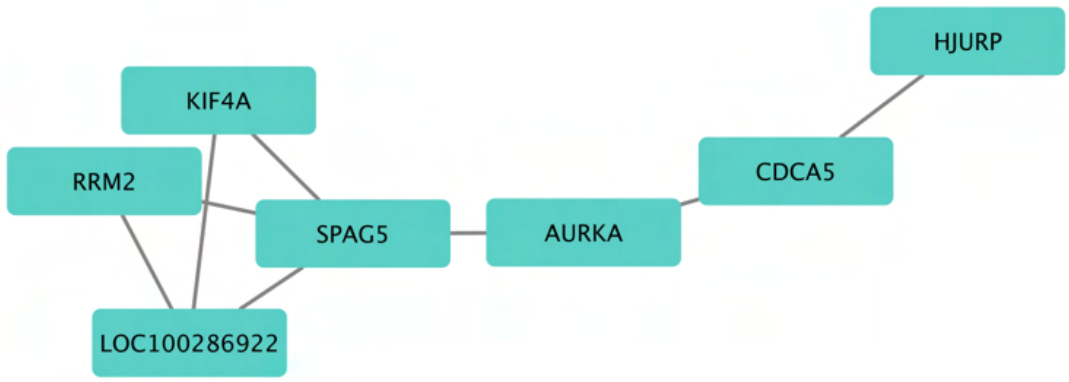
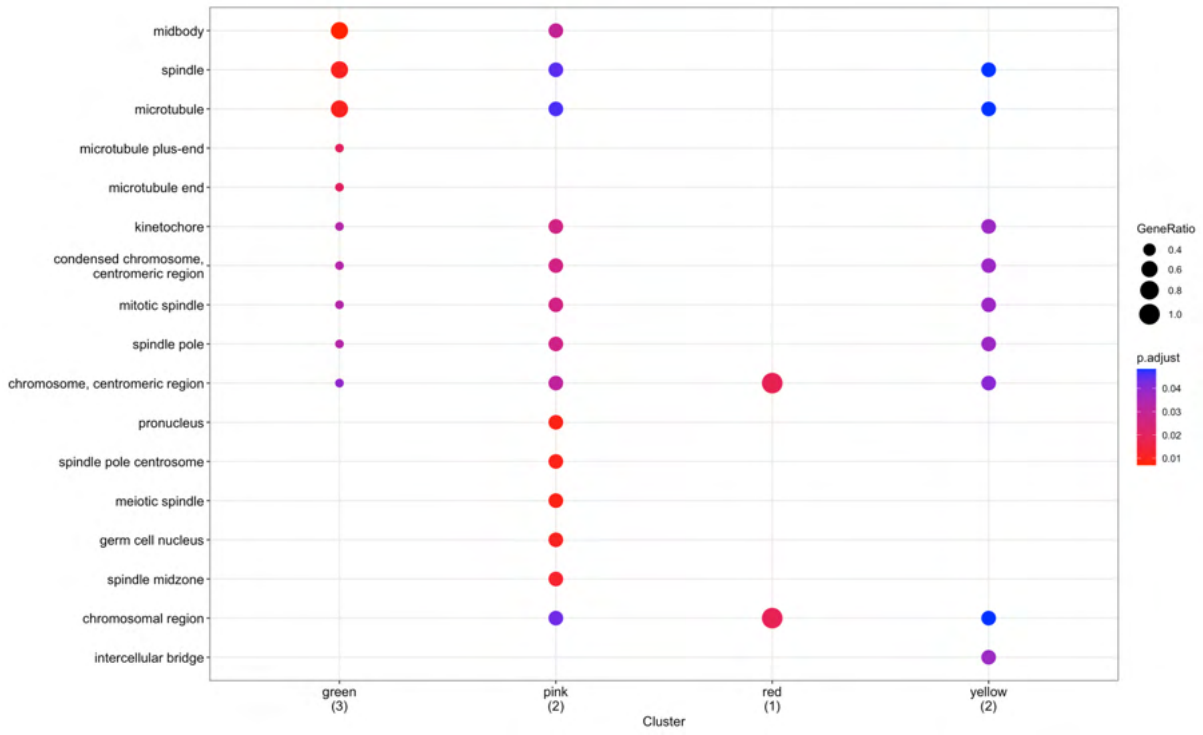


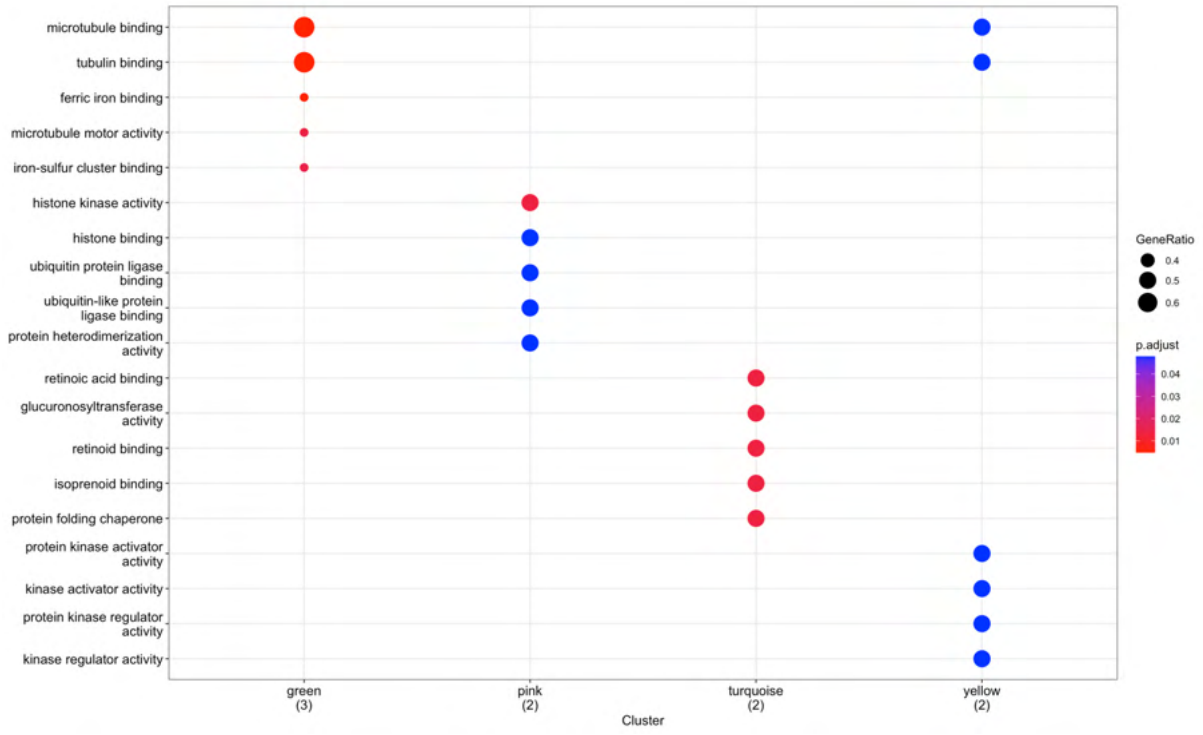
Figure 29 | A Cytoscape network view of the selected co-expressed gene cluster modules seen in the up-regulated tumour samples, represented in nodes, following the cut off threshold value of 0.91. SPAG5 had the highest connection with 5 nodes. Genes with a connection of less than 2 were discarded. Network produced by Cytoscape version 3.9.1.

ClusterProfiler was used to determine the enrichment compareCluster biological theme classification for the cellular component, molecular function, biological processes and lastly, KEGG pathways (*Fig. 30*). Each cluster on the x-axis represents the cluster module seen in *Fig. 28b*. A cluster consists of the DEGs that are co-expressed. The plots show the DEGs and its role in the biological system (y-axis).

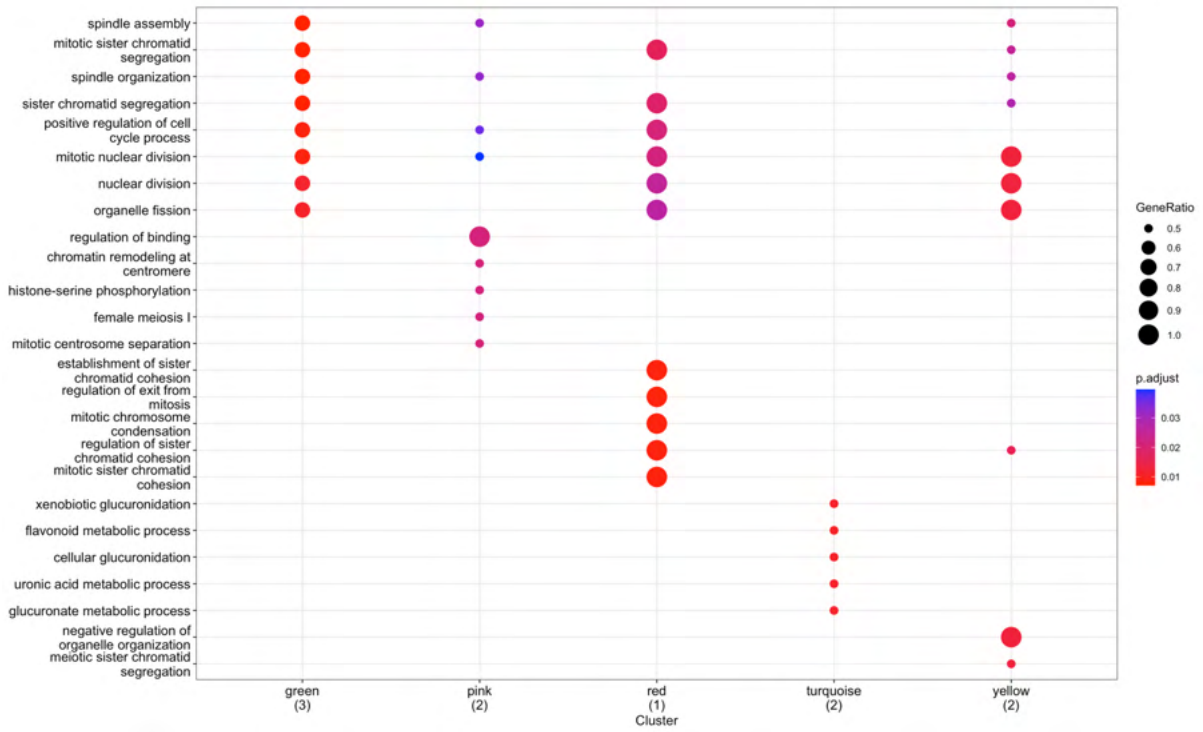
a


midbody
spindle
microtubule
microtubule plus-end
microtubule end
kinetochore
condensed chromosome, centromeric region
mitotic spindle
spindle pole
chromosome, centromeric region
pronucleus
spindle pole centrosome
meiotic spindle
germ cell nucleus
spindle midzone
chromosomal region
intercellular bridge
green (3)
pink (2)
red (1)
yellow (2)
Cluster
GeneRatio
0.4
0.6
0.8
1.0
p.adjust
0.04
0.03
0.02
0.01

b.


microtubule binding
tubulin binding
ferric iron binding
microtubule motor activity
iron-sulfur cluster binding
histone kinase activity
histone binding
ubiquitin protein ligase binding
ubiquitin-like protein ligase binding
protein heterodimerization activity
retinoic acid binding
glucuronosyltransferase activity
retinoid binding
isoprenoid binding
protein folding chaperone
protein kinase activator activity
kinase activator activity
protein kinase regulator activity
kinase regulator activity
green (3)
pink (2)
turquoise (2)
yellow (2)
Cluster
GeneRatio
0.4
0.5
0.6
p.adjust
0.04
0.03
0.02
0.01

c.



d.

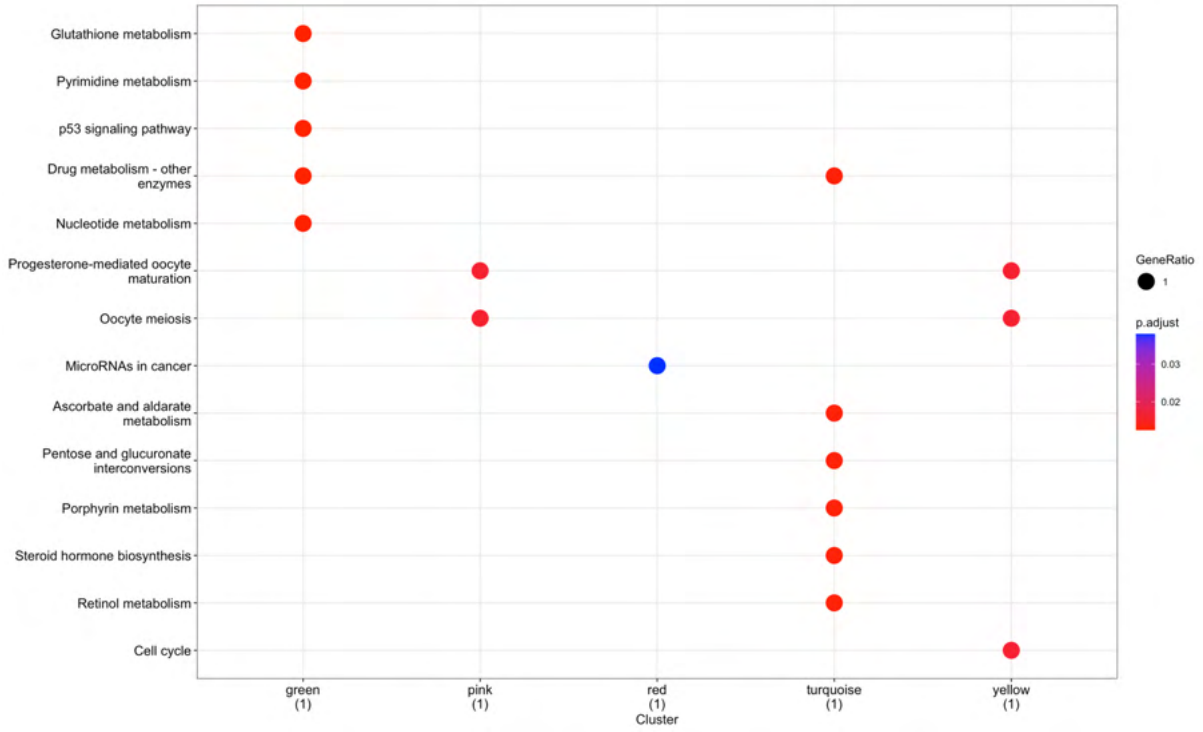


Figure 30 | GO and KEGG enrichment for all of the upregulated tumour subtype gene clusters. a) A compareCluster biological theme comparison for the enriched GO functional cellular component profiles, displaying functional consensus and differences observed in each upregulated tumour gene

cluster. b) A compareCluster biological theme comparison for the enriched GO functional molecular function profiles. c) A compareCluster biological theme comparison for the enriched GO functional biological process profiles. d) A compareCluster biological theme comparison for the enriched KEGG functional process profiles. The x-axis represents the gene clusters whereas the y-axis shows the appropriate labels. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set. The dot colours display the statistical significance whereby red is highly significant and blue is significant. Plot produced by R ClusterProfiler version 4.5.1.

Kaplan-Meier survival curves were generated for all of the selected co-expressed DEGs for the tumour sample. Survival graphs with insignificant P-values were excluded from downstream analysis. Additionally, DEGs with curves that did not support the compared conditions were discarded. An average graph with supportive curves and significant P-values is seen in *Fig. 31*. Based on this graph, the DEGs had a much lower survival rate when upregulated in the tumour sample with a highly significant P-value (0.00032).

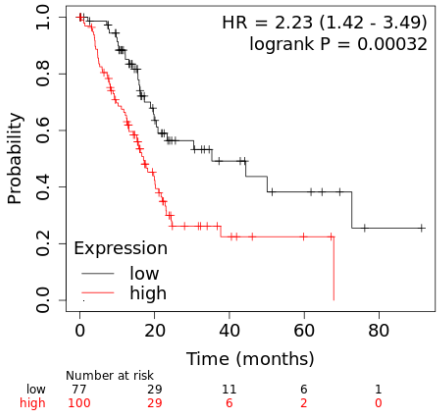


Figure 31 | Average Kaplan-Meier survival curve for the average tumour upregulated DEGs selected based on the co-expressed network following the removal of outliers and insignificant P-value graphs. Individual survival curves can be seen in Appendix C. Graph show the probability of survival versus the time (months). HR serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by a drug study that is compared to the hazard reduction for the control of the study. Logrank P values represents the P-values for each plot, showing the statistical significance. The black line in each plot shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graphs produced by Kaplan-Meier plotter version 2022.08.23.

The Expression Atlas showed the levels per gene in a map for the co-expressed DEGs in PDAC. The DEGs with supportive survival curves (*Fig. 31*) were used to produce an expression map, *Fig. 32.*

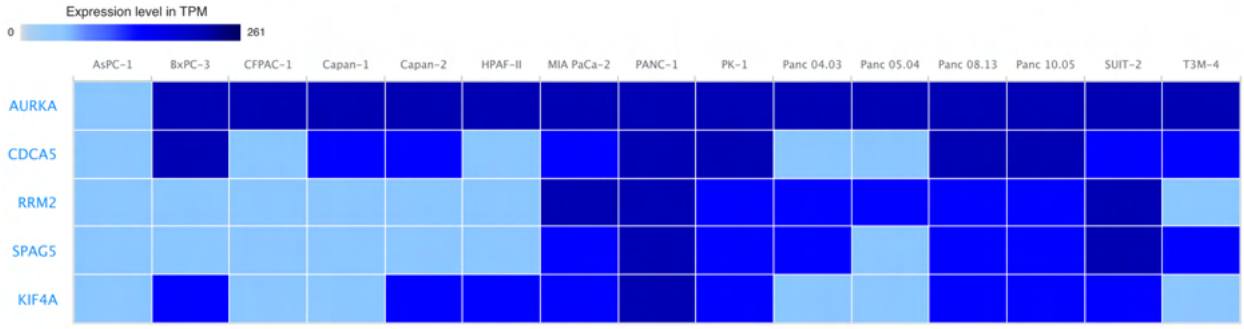


Figure 32 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated tumour co-expressed DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).

The tumour co-expressed DEGs with high expression levels in the map, *Fig. 32*, were used to produce expression box plots in from the GEPIA cancer database, *Fig. 33*. This validation step shows the gene expression level in a PDAC patient sample relative to the adjacent normal.

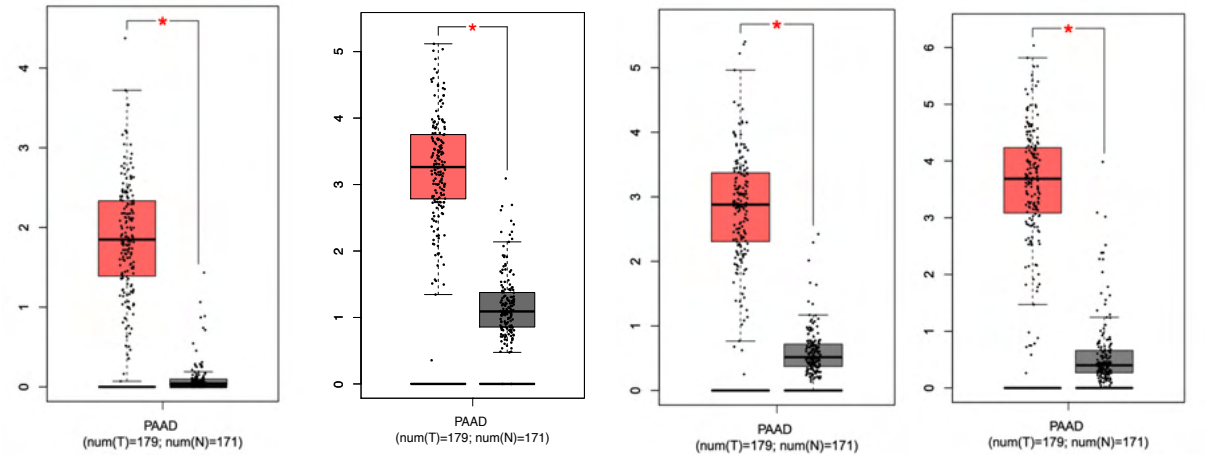


Figure 33 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (*Fig. 33*) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).

Table 6 shows the identified signature genes in PDAC and its expression values in the extracellular vesicles and the log_2FC from different databases, BBCancer and GENT2, respectively. An upregulation in the extracellular vesicles may indicate migration whereas the GENT2 values indicate the expression level in terms of the log_2FC. A negative value represents a downregulation and a positive value shows the upregulation level.

Table 6 | The BBCancer and GENT2 expression values (TPM) for the selected tumour co-expressed DEGs.

GENE	**BBCancer (Extracellular vesicles)**	**GENT2 Log_2FC**
AURKA	0.34	1.01
RRM2	-0.22	1.22
SPAG5	0.029	0.94
KIF4A	0.11	1.07
CDCA5	-0.14	1.06

4.3.1. Normal upregulated DEGs

The normal upregulated samples were included to evaluate the difference in gene expression level observed in a normal condition when compared to a tumourigenic environment. The soft threshold value (*Fig.34a*) was calculated and used to construct a cluster dendrogram (*Fig.34b*). A soft threshold value close to a 1 (y-axis, R^2) in the scale free topology model represents highly connected DEGs with low background noise, hence value selection was 7 for the normal sample set.

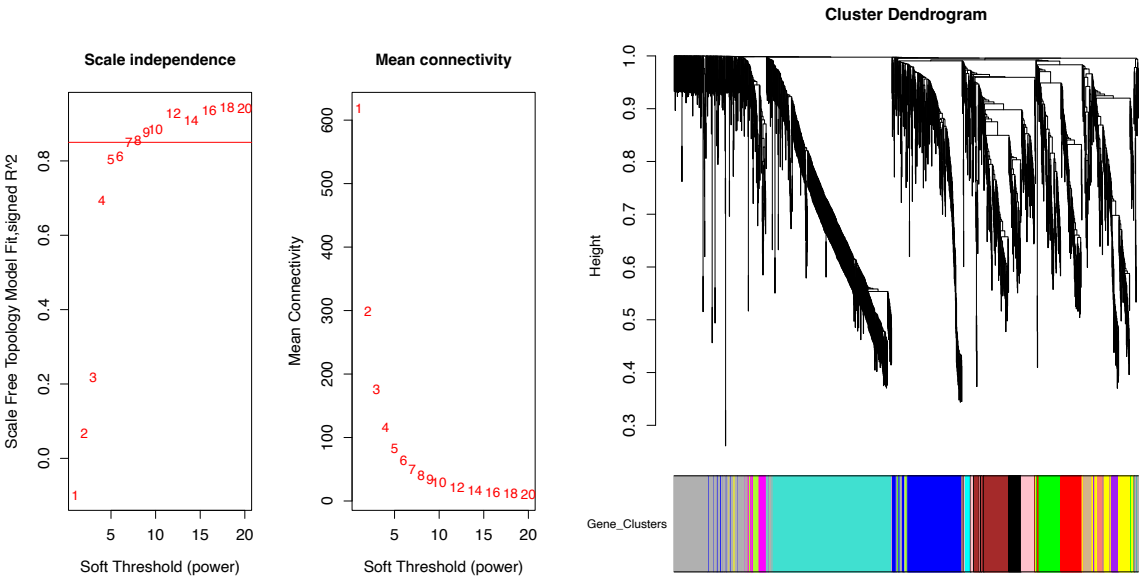


Figure 34 | WGCNA analysis for the upregulated normal samples containing DEGs. a) Network topology for the scale independence and mean connectivity analysis of modules consisting of soft thresholding powers ranging between the values of 1 to 20. The scale independence plot consists of a y-axis which represents the scale-free topology fit as a function of the soft thresholding power (x-axis). The mean connectivity plot displays the degree (y-axis) as a function of the soft-thresholding power (x-axis). The power of 7 was selected for downstream analysis as it is the lowest power for which the scale-free topology fit index reaches 0.90. b) Clustering dendrogram of genes among modules via hierarchical clustering of adjacency-based dissimilarity. Plots produced by R WGCNA version 1.71.

A network plot was made for the DEGs that are co-expressed in the same biological condition, *Fig. 35*. The lines between the nodes (genes) may represent potential interactions with the genes.

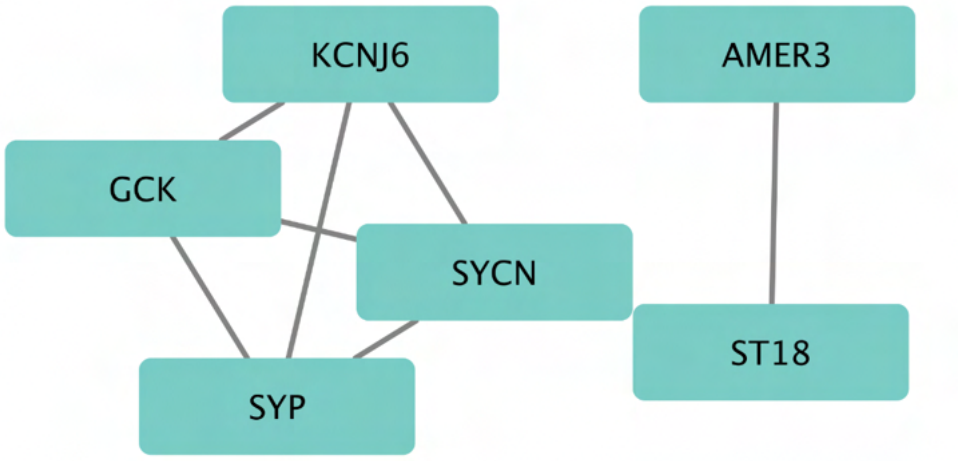
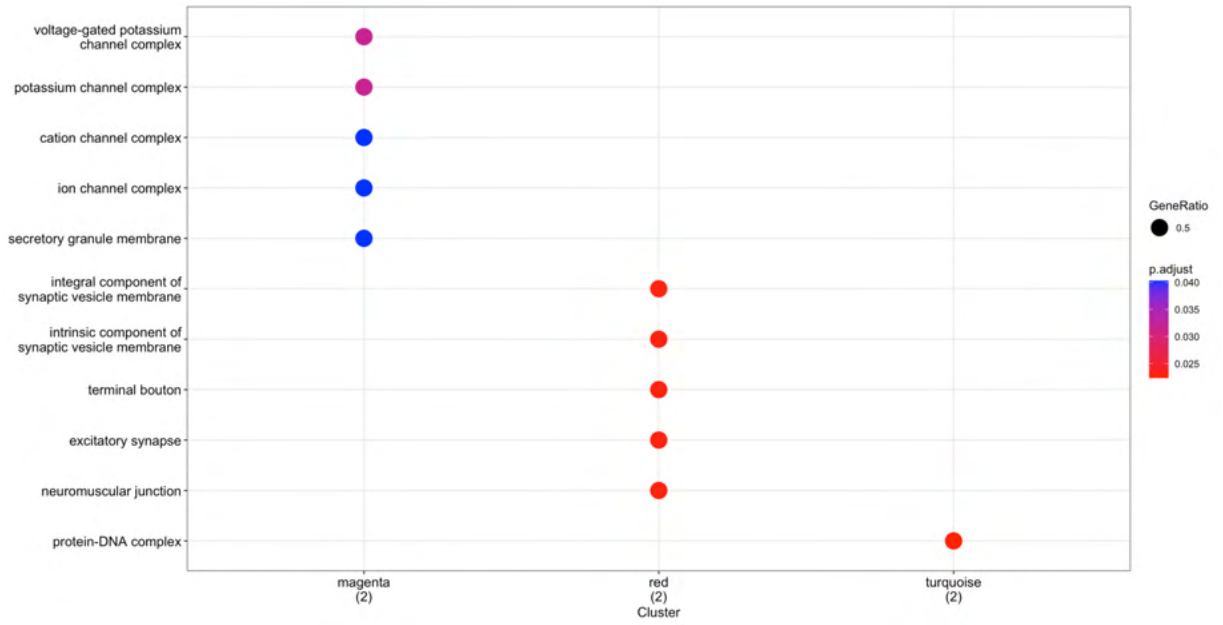


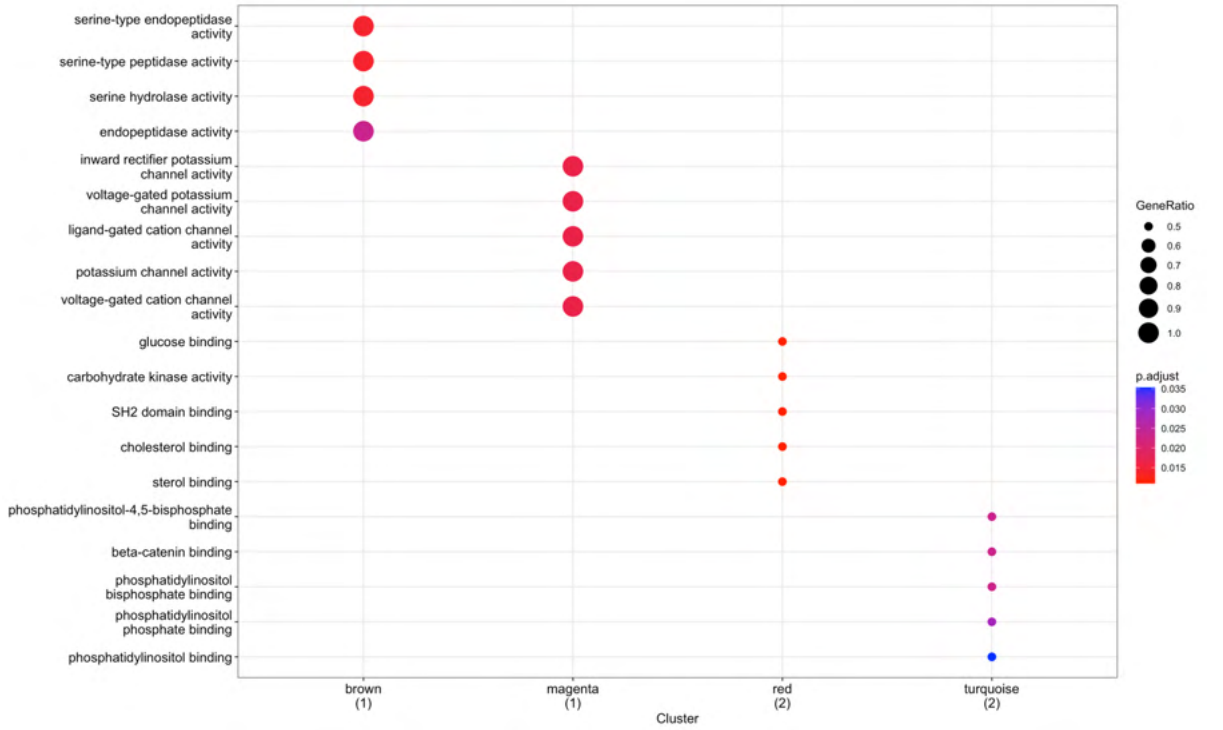
Figure 35 | A Cytoscape network view of the selected co-expressed gene cluster modules seen in the up-regulated normal samples, represented in nodes, following the cut off threshold value of 0.99. Network produced by Cytoscape version 3.9.1.

ClusterProfiler was used to determine the enrichment compareCluster biological theme classification for the cellular component, molecular function, biological processes and lastly, KEGG pathways (*Fig. 36*) for all of the significant genes. A cluster consists of the DEGs that are co-expressed. The plots show the DEGs and its role in the biological system (y-axis).

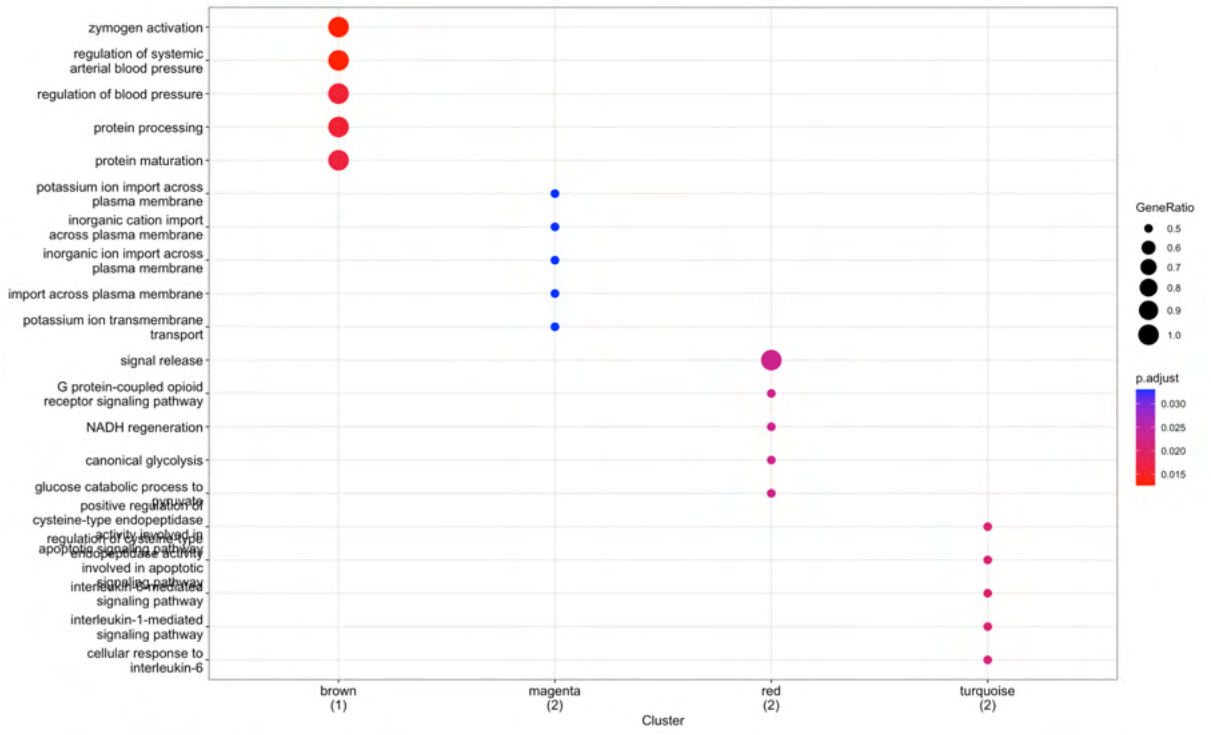
a.



b.



c.



d.

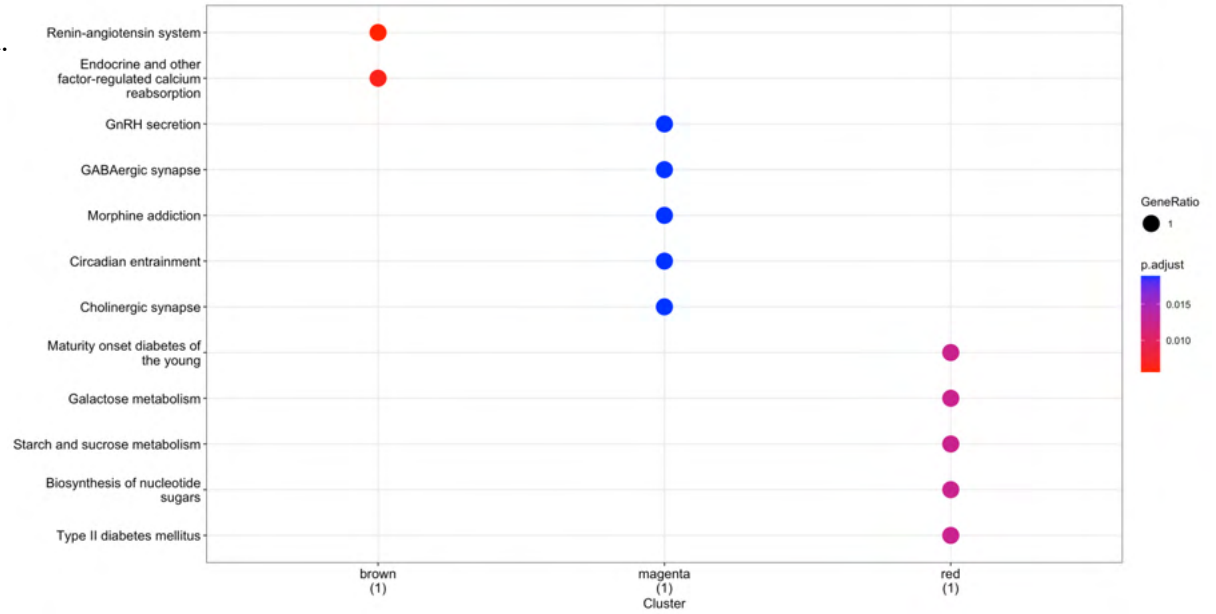


Figure 36 | GO and KEGG enrichment for all of the upregulated normal gene clusters. a) A compareCluster biological theme comparison for the enriched GO functional cellular component profiles, displaying functional consensus and differences observed in each upregulated normal gene cluster. b) A compareCluster biological theme comparison for the enriched GO functional molecular function profiles. c) A compareCluster biological theme comparison for the enriched GO functional biological process profiles. d) A compareCluster biological theme comparison for the enriched KEGG functional process profiles. The x-axis represents the gene clusters whereas the y-axis shows the appropriate labels. The gene ratio, represented in dot size, shows the percentage of the total genes in

the given set. The dot colours display the statistical significance whereby red is highly significant and blue is significant. Plot produced by R ClusterProfiler version 4.5.1.

Kaplan-Meier survival curves were generated for all of the selected co-expressed DEGs in the normal sample set. Survival graphs with insignificant P-values were excluded from downstream analysis. Additionally, DEGs with curves that did not support the compared conditions were discarded. An average graph with supportive curves and significant P-values is seen in *Fig. 37*. These DEGs were upregulated in the normal condition whereas downregulated in the cancer condition. As seen in the KM plot, when upregulated, the probability of survival at 80 months increases to a 0.4. When downregulated, the survival rate decreases to a 0.01 at 75 months with a P-value of 0.0022.

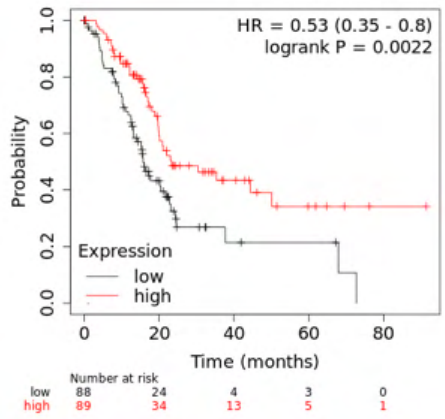


Figure 37 | Average Kaplan-Meier survival curves for the normal upregulated DEGs selected based on the co-expressed network. Individual survival curves can be seen in Appendix B, *Fig. 58*. Graph depicts the probability of survival versus the time (months). HR is a measurement of the magnitude for the differences between the two curves in each plot. It shows the hazard reduction obtained by a drug study which is compared to the hazard reduction for the control of the study. Logrank P values represents the P-values for each plot, showing the statistical significance. The black line displays the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graphs produced by Kaplan-Meier plotter version 2022.08.23.

The Gene Expression Atlas showed the expression levels per gene in a map for the co-expressed DEGs in the normal sample set. The DEGs with supportive survival curves (*Fig. 37*) were used to produce an expression map, *Fig. 38.*

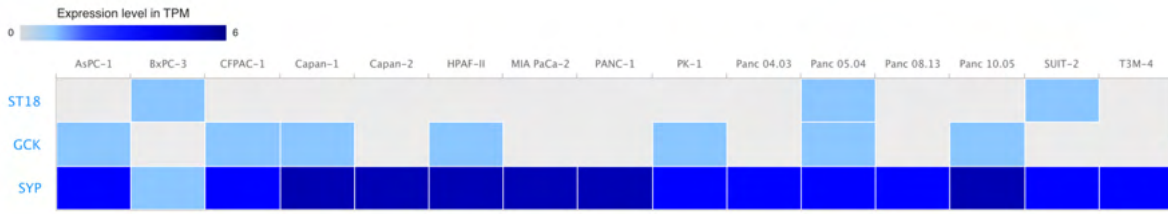


Figure 38 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated normal co-expressed DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. It can be observed that the expression level ranges from 0 to only 6 TPM values which is due it being upregulated in the normal samples and not in PC cell lines. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).

The normal co-expressed DEGs with high expression levels in the map, *Fig. 38*, were used to produce expression box plots from the GEPIA cancer database, *Fig. 39*. This validation step shows the gene expression level in a PDAC patient sample relative to the adjacent normal.

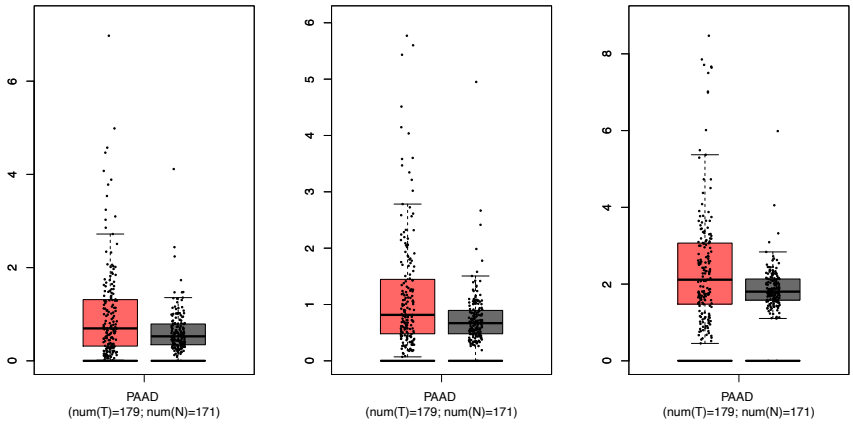


Figure 39 | Box plots for genes with TPM expression levels for the PC cell lines observed in the expression atlas graph (*Fig. 38*) plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).

Table 7 shows the identified signature genes in normal sample set and its expression values in the extracellular vesicles and the $\log_2$FC from different databases, BBCancer and GENT2, respectively. An upregulation in the extracellular vesicles may indicate migration whereas the

GENT2 values indicate the expression level in terms of the log_2FC. A negative value represents a downregulation and a positive value shows the upregulation level. These genes were upregulated in the normal samples and it can be observed that there are negative expression levels in both BBCancer and GENT2 cancer databases for pancreatic cancer.

Table 7 | The BBCancer and GENT2 expression values (TPM) for the selected normal co-expressed DEGs.

GENE	**BBCancer (Extracellular vesicles)**	**GENT2 Log_2FC**
ST18	N/A	-0.90
GCK	-3.43	-0.51
SYP	N/A	-1.89

4.3.2. Basal-like subtype upregulated DEGs

The soft threshold value (*Fig. 40a*) was calculated and used to construct a cluster dendrogram (*Fig.40b*). A soft threshold value close to a 1 (y-axis, R^2) in the scale free topology model represents highly connected DEGs with low background noise, hence value selection was 6 for the basal-like subtype condition.

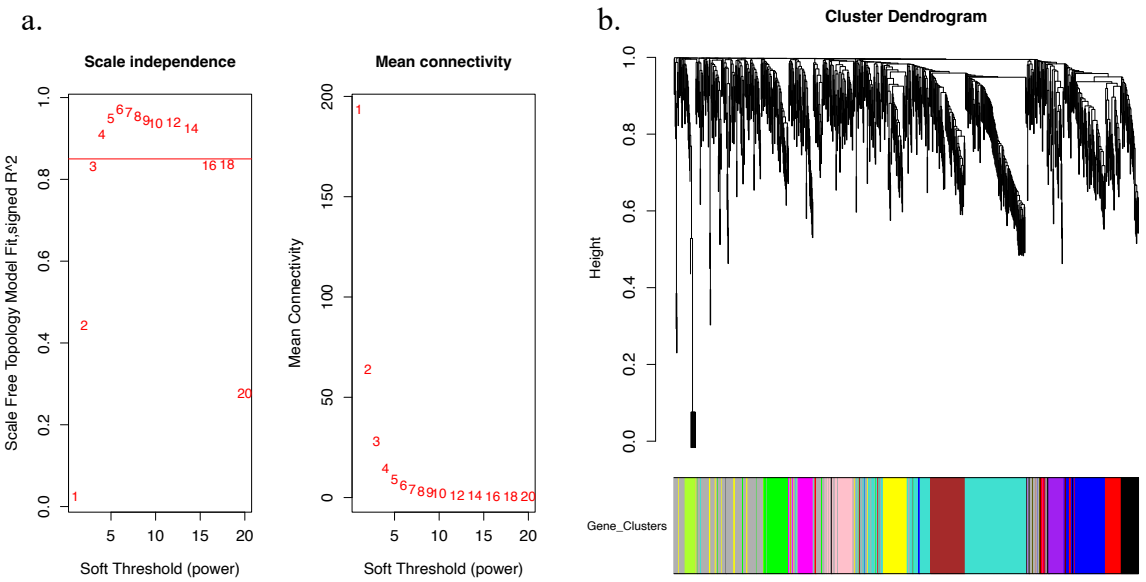


Figure 40 | WGCNA analysis for the upregulated basal-like subtype samples containing differentially expressed genes. a) Network topology for the scale independence and mean connectivity analysis of modules consisting of soft thresholding powers ranging between the values of 1 to 20. The scale independence plot consists of a y-axis which represents the scale-free topology fit as a function of the soft thresholding power (x-axis). The mean connectivity plot displays the degree (y-axis) as a function of the soft-thresholding power (x-axis). The power of 6 was selected for downstream analysis as it is the lowest power for which the scale-free topology fit index reaches close to 0.90. b) Cluster dendrogram of genes among modules via hierarchical clustering of adjacency-based dissimilarity. Plots produced by R WGCNA version 1.71.

A network plot was made for the DEGs that are co-expressed in the same biological condition- the basal-like subtype, *Fig. 35*. The lines between the nodes (genes) may represent potential interactions with the genes.

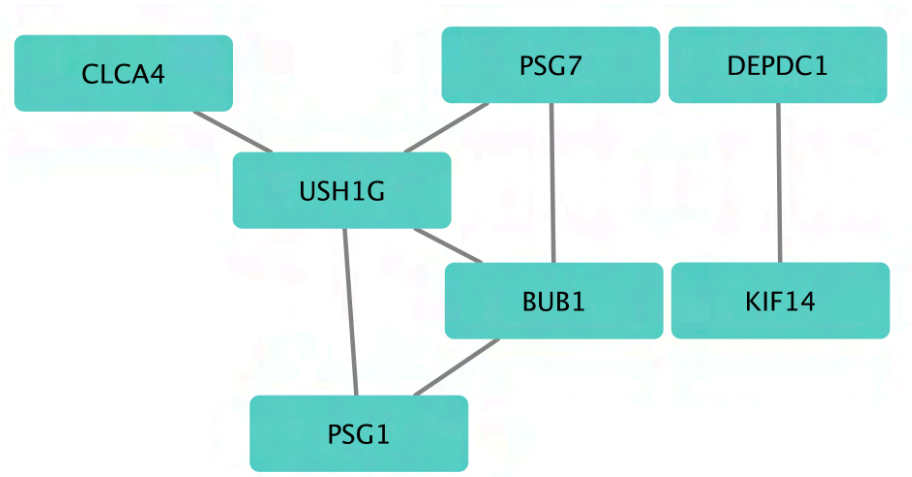
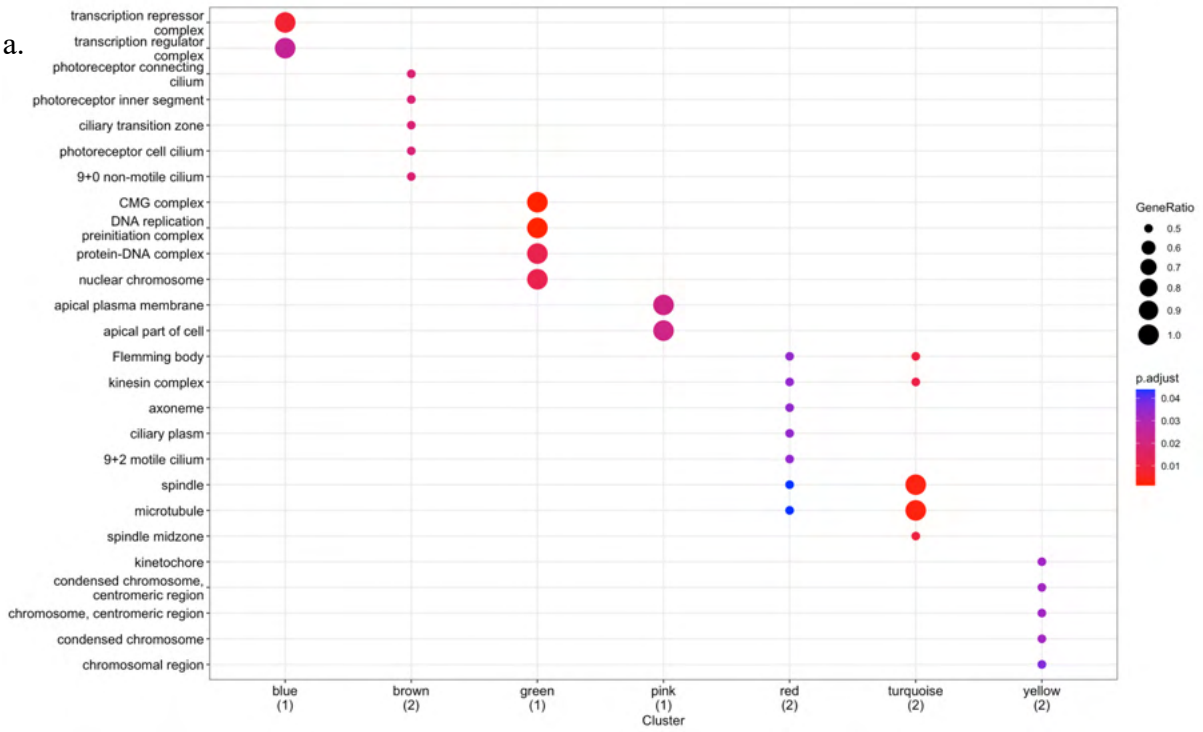


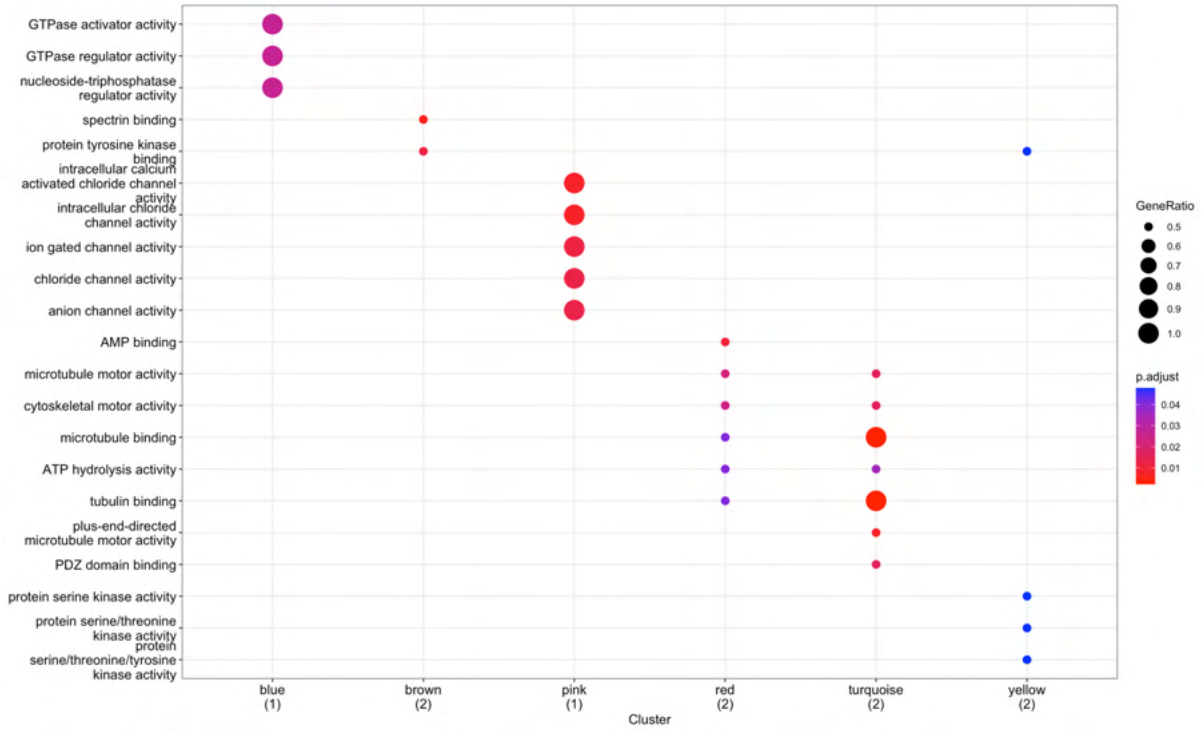
Figure 41 | A Cytoscape network view of the selected co-expressed gene cluster modules seen in the up-regulated basal-like subtype samples, represented in nodes, following the cut off threshold value of 0.95. Network produced by Cytoscape version 3.9.1.

ClusterProfiler was used to determine the enrichment compareCluster biological theme classification for the cellular component, molecular function, biological processes and lastly, KEGG pathways (*Fig. 42*) for the basal-like subtype. Each cluster on the x-axis represents the cluster module seen in *Fig. 40b*. A cluster consists of the DEGs that are co-expressed. The plots show the DEGs and its role in the biological system (y-axis).

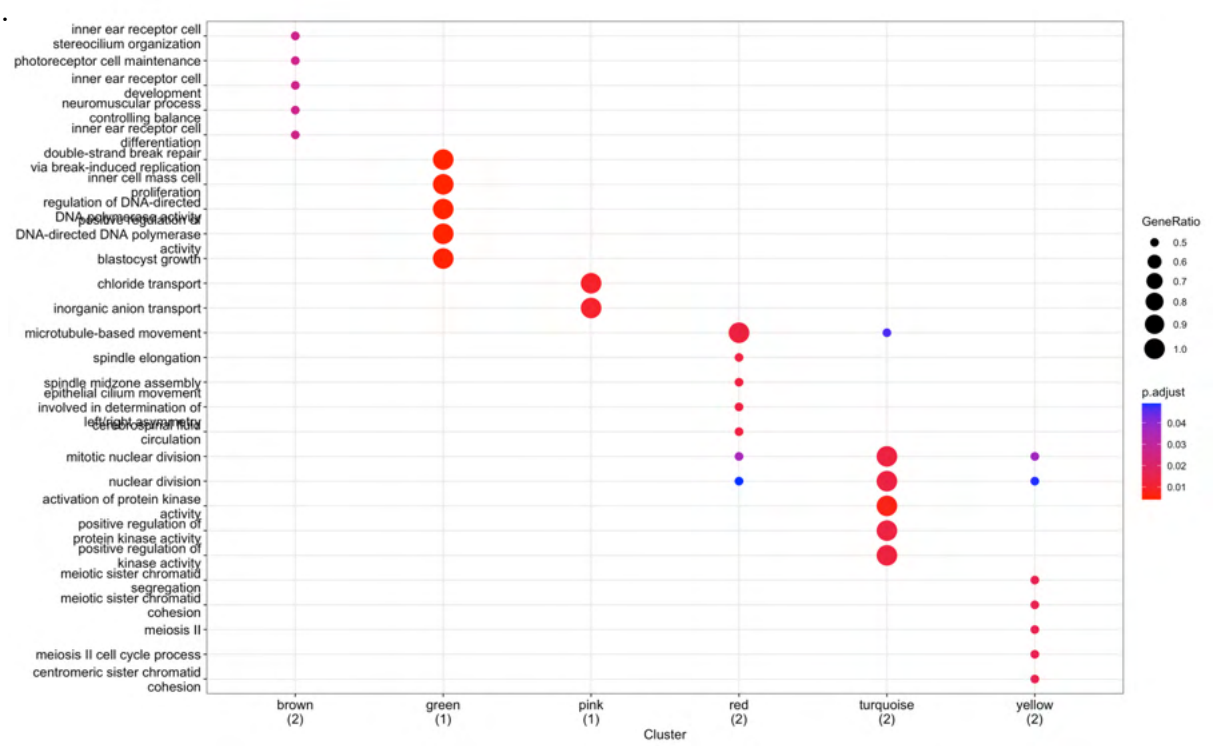
a.


transcription repressor complex
transcription regulator complex
photoreceptor connecting cilium
photoreceptor inner segment
ciliary transition zone
photoreceptor cell cilium
9+0 non-motile cilium
CMG complex
DNA replication preinitiation complex
protein-DNA complex
nuclear chromosome
apical plasma membrane
apical part of cell
Flemming body
kinesin complex
axoneme
ciliary plasm
9+2 motile cilium
spindle
microtubule
spindle midzone
kinetochore
condensed chromosome, centromeric region
chromosome, centromeric region
condensed chromosome
chromosomal region
blue (1)
brown (2)
green (1)
pink (1)
red (2)
turquoise (2)
yellow (2)
Cluster
GeneRatio
0.5
0.6
0.7
0.8
0.9
1.0
p.adjust
0.04
0.03
0.02
0.01

b.


GTPase activator activity
GTPase regulator activity
nucleoside-triphosphatase regulator activity
spectrin binding
protein tyrosine kinase binding
intracellular calcium activated chloride channel activity
intracellular chloride channel activity
ion gated channel activity
chloride channel activity
anion channel activity
AMP binding
microtubule motor activity
cytoskeletal motor activity
microtubule binding
ATP hydrolysis activity
tubulin binding
plus-end-directed microtubule motor activity
PDZ domain binding
protein serine kinase activity
protein serine/threonine kinase activity
protein serine/threonine/tyrosine kinase activity
blue (1)
brown (2)
pink (1)
red (2)
turquoise (2)
yellow (2)
Cluster
GeneRatio
0.5
0.6
0.7
0.8
0.9
1.0
p.adjust
0.04
0.03
0.02
0.01

c.



d.

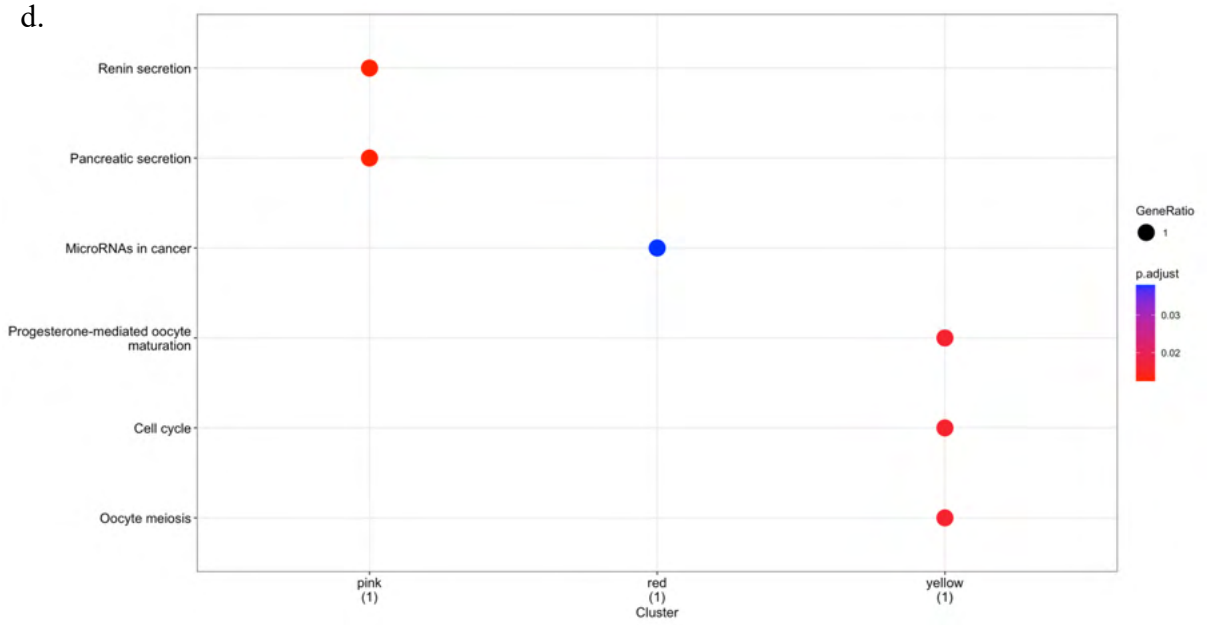


Figure 42 | GO and KEGG enrichment for all of the upregulated basal-like subtype gene clusters. a) A compareCluster biological theme comparison for the enriched GO functional cellular component profiles, displaying functional consensus and differences observed in each upregulated basal-like subtype gene cluster. b) A compareCluster biological theme comparison for the enriched GO functional molecular function profiles. c) A compareCluster biological theme comparison for the enriched GO functional biological process profiles. d) A compareCluster biological theme comparison for the enriched KEGG functional process profiles. The x-axis represents the gene clusters whereas the y-axis

shows the appropriate labels. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set. The dot colours display the statistical significance whereby red is highly significant and blue is significant. Plot produced by R ClusterProfiler version 4.5.1.

Kaplan-Meier survival curves were made for all of the selected co-expressed DEGs for the basal-like subtype. Survival graphs with insignificant P-values were excluded from downstream analysis. Additionally, DEGs with curves that did not support the compared conditions were discarded. An average graph with supportive curves and significant P-values is seen in *Fig. 43*. Based on this graph, the DEGs had a much lower survival rate when upregulated in this subtype with a highly significant P-value (0.00024).

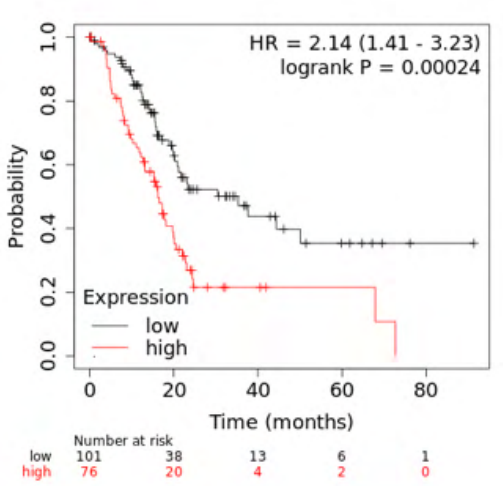


Figure 43 | Average Kaplan-Meier survival curves for the basal-like upregulated DEGs selected based on the co-expressed network. Individual survival curves can be seen in Appendix C. Graphs show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by a drug study that is compared to the hazard reduction for the control of the study. Logrank P values represents the P-values for each plot, showing the statistical significance. The black line in each plot shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graphs produced by Kaplan-Meier plotter version 2022.08.23.

The Gene Expression Atlas showed the levels per gene in a map for the co-expressed DEGs for the basal-like subtype. The DEGs with supportive survival curves (*Fig. 43*) were used to produce an expression map, *Fig. 44*.



Figure 44 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated basal-like co-expressed DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).

The tumour co-expressed DEGs with high expression levels in the map, *Fig. 44*, were used to produce expression box plots in from the GEPIA cancer database, *Fig. 45*. This validation step shows the gene expression level in a PDAC patient sample relative to the adjacent normal.

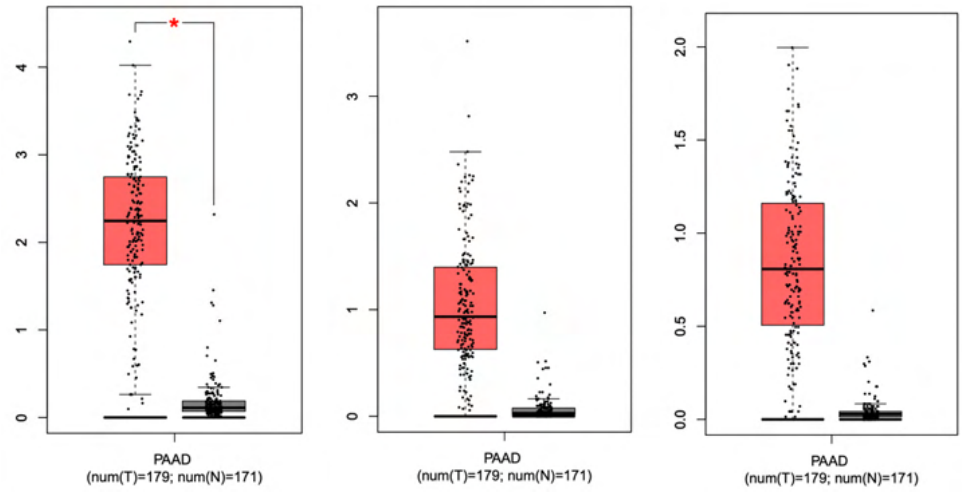


Figure 45 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (*Fig. 44)* plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).

Table 8 shows the identified signature genes in basal-like subtype and its expression values in the extracellular vesicles and the log_2FC from different databases, BBCancer and GENT2, respectively. An upregulation in the extracellular vesicles may indicate migration whereas the GENT2 values indicate the expression level in terms of the log_2FC. A negative value represents a downregulation and a positive value shows the upregulation level.

Table 8 | The BBCancer and GENT2 expression values (TPM) for the selected basal-like co-expressed DEGs.

GENE	BBCancer (Extracellular vesicles)	GENT2 Log_2FC
BUB1	0.28	0.70
DEPDC1	-0.22	0.73
KIF14	0.43	0.97

4.3.3. Classical up-regulated DEGs

The soft threshold value (*Fig.46a*) was calculated and used to construct a cluster dendrogram (*Fig.46b*). A soft threshold value close to a 1 (y-axis, R^2) in the scale free topology model represents highly connected DEGs with low background noise, hence value selection was 7 for the classical subtype condition.

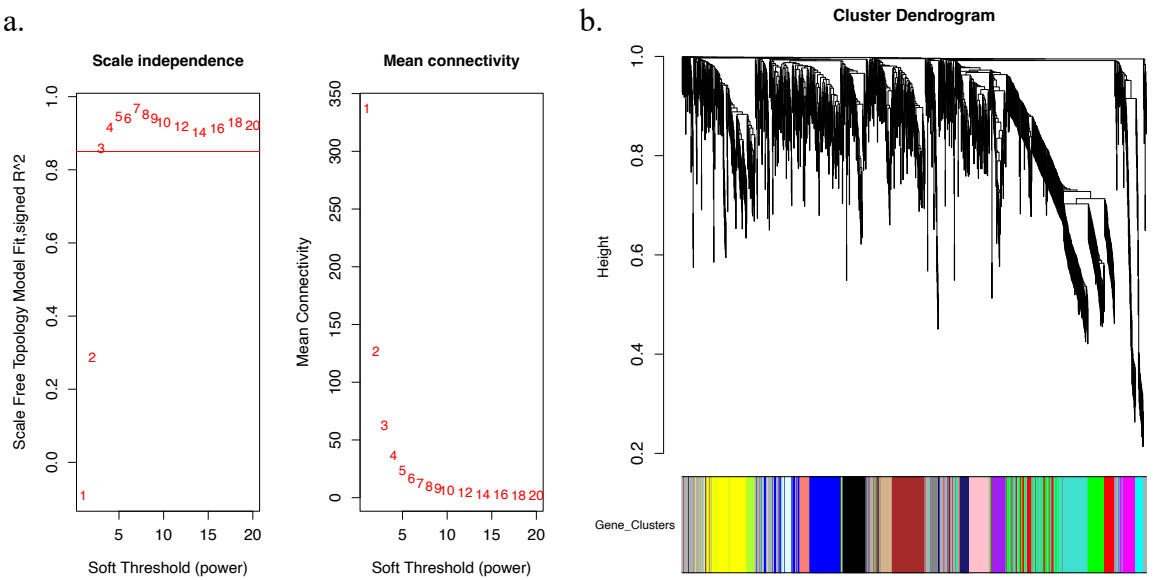


Figure 46 | WGCNA analysis for the up-regulated classical subtype samples containing differentially expressed genes. a) Network topology for the scale independence and mean connectivity analysis of modules consisting of soft thresholding powers ranging between the values of 1 to 20. The scale independence plot consists of a y-axis which represents the scale-free topology fit as a function of the soft thresholding power (x-axis). The mean connectivity plot displays the degree (y-axis) as a function of the soft-thresholding power (x-axis). The power of 7 was selected for downstream analysis as it is the lowest power for which the scale-free topology fit index reaches close to 0.90. b) Cluster dendrogram of genes among modules via hierarchical clustering of adjacency-based dissimilarity. Plots produced by R WGCNA version 1.71.

A network plot was made for the DEGs that are co-expressed in the same biological condition- the classical subtype, *Fig. 47*. The lines between the nodes (genes) may represent potential interactions with the genes.

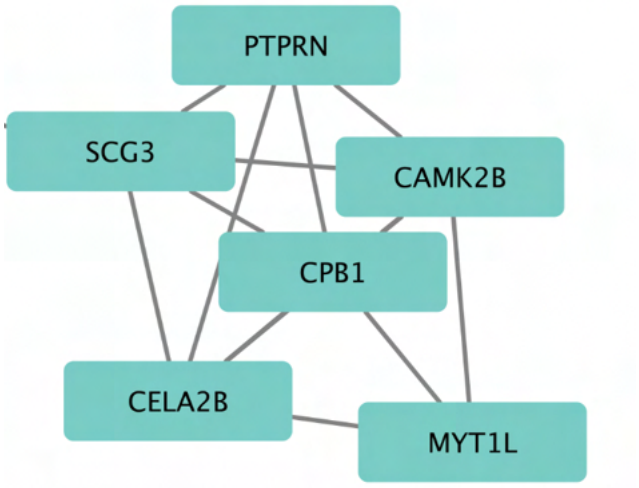
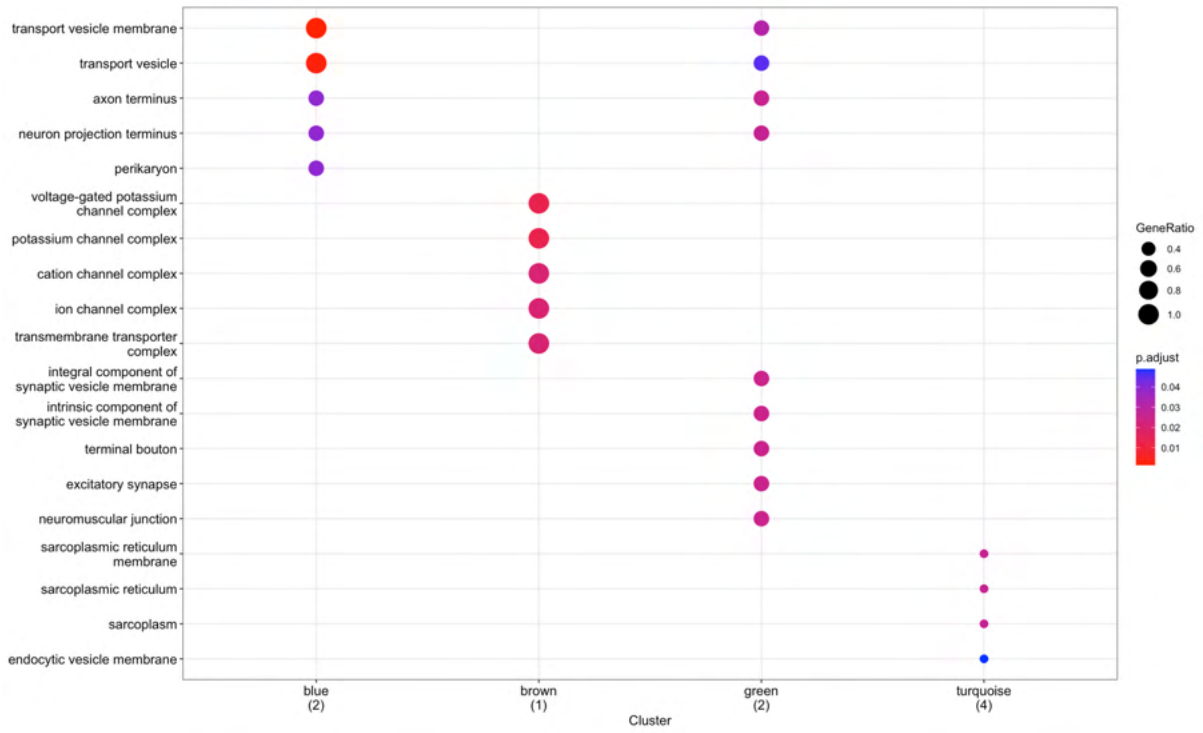


Figure 47 | A Cytoscape network view of the selected co-expressed gene cluster modules seen in the upregulated classical subtype samples, represented in nodes, following the cut off threshold value of 0.975. Network produced by Cytoscape version 3.9.1.

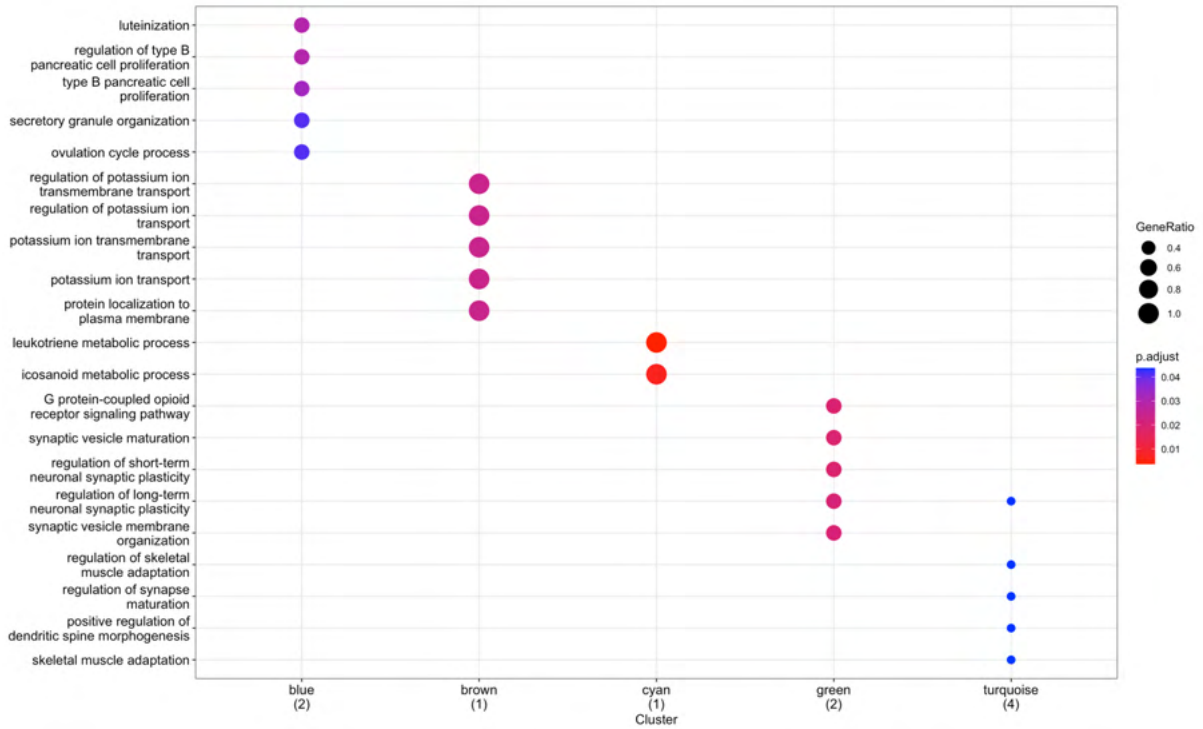
ClusterProfiler was used to determine the enrichment compareCluster biological theme classification for the cellular component, molecular function, biological processes and lastly, KEGG pathways (*Fig. 48*) for the classical subtype. Each cluster on the x-axis represents the

cluster module seen in *Fig. 40b*. A cluster consists of the DEGs that are co-expressed. The plots show the DEGs and its role in the biological system (y-axis).

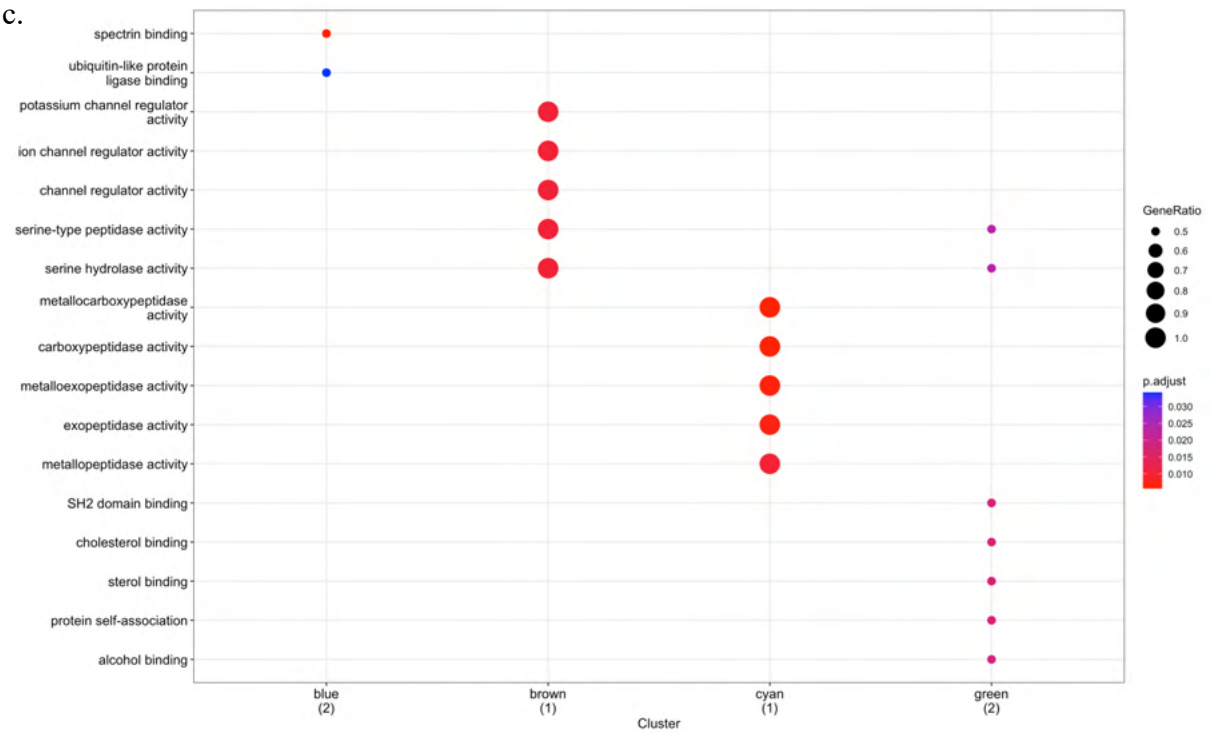
a.



b.



c.



d.

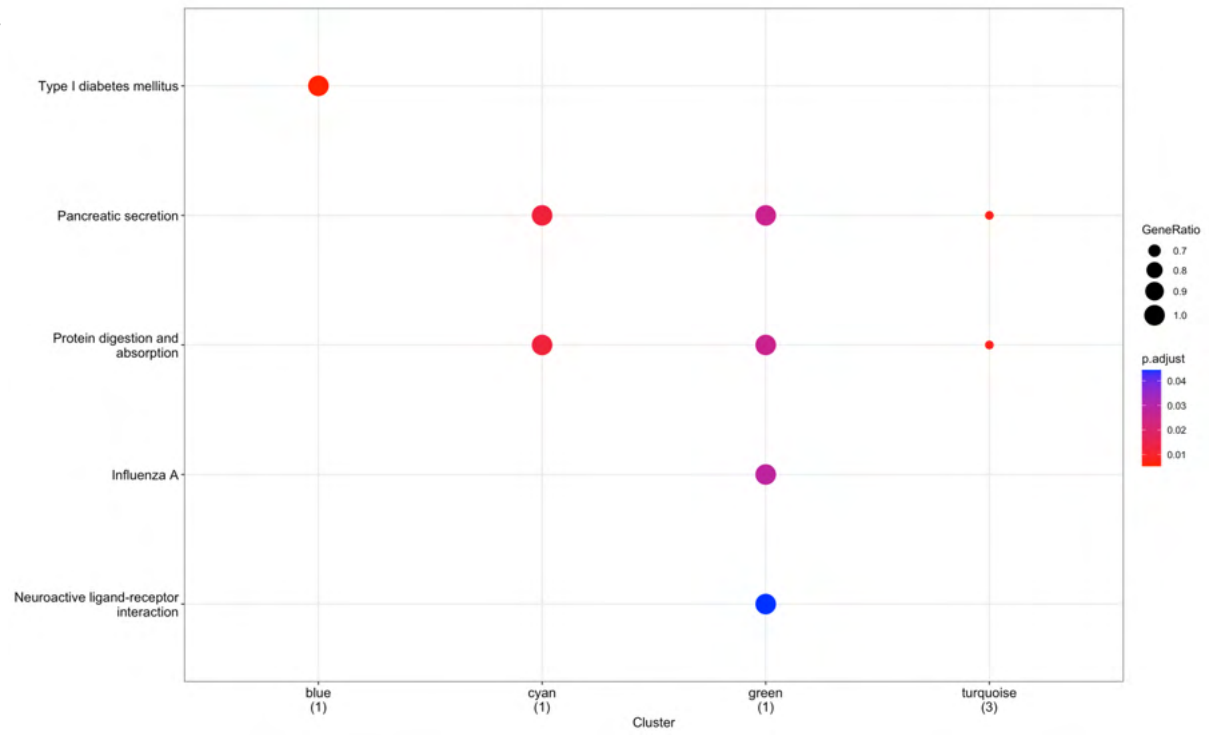


Figure 48 | GO and KEGG enrichment for all of the upregulated classical subtype gene clusters. a) A compareCluster biological theme comparison for the enriched GO functional cellular component profiles, displaying functional consensus and differences observed in each upregulated classical subtype

gene cluster. b) A compareCluster biological theme comparison for the enriched GO functional molecular function profiles. c) A compareCluster biological theme comparison for the enriched GO functional biological process profiles. d) A compareCluster biological theme comparison for the enriched KEGG functional process profiles. The x-axis represents the gene clusters whereas the y-axis shows the appropriate labels. The gene ratio, represented in dot size, shows the percentage of the total genes in the given set. The dot colours display the statistical significance whereby red is highly significant and blue is significant. Plot produced by R ClusterProfiler version 4.5.1.

Kaplan-Meier survival curves were made for all of the selected co-expressed DEGs for the classical subtype. Survival graphs with insignificant P-values were excluded from downstream analysis. Additionally, DEGs with curves that did not support the compared conditions were discarded. An average graph with supportive curves and significant P-values is seen in *Fig. 49*. Based on this graph, the DEGs had a much lower survival rate when upregulated in this subtype with a significant P-value (0.0016).

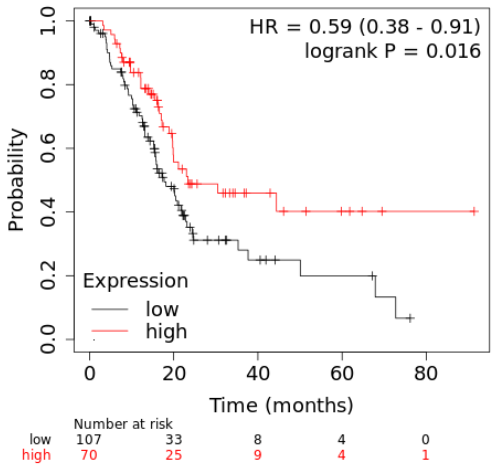


Figure 49 | Average Kaplan-Meier survival curves for the classical up-regulated DEGs selected based on the co-expressed network. Individual survival curves can be seen in Appendix C. Graphs show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by a drug study that is compared to the hazard reduction for the control of the study. Logrank P values represents the P-values for each plot, showing the statistical significance. The black line in each plot shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graphs produced by Kaplan-Meier plotter version 2022.08.23.

The Expression Atlas showed the levels per gene in a map for the co-expressed DEGs for the classical subtype. The DEGs with supportive survival curves (*Fig. 49*) were used to produce an expression map, *Fig. 50.*

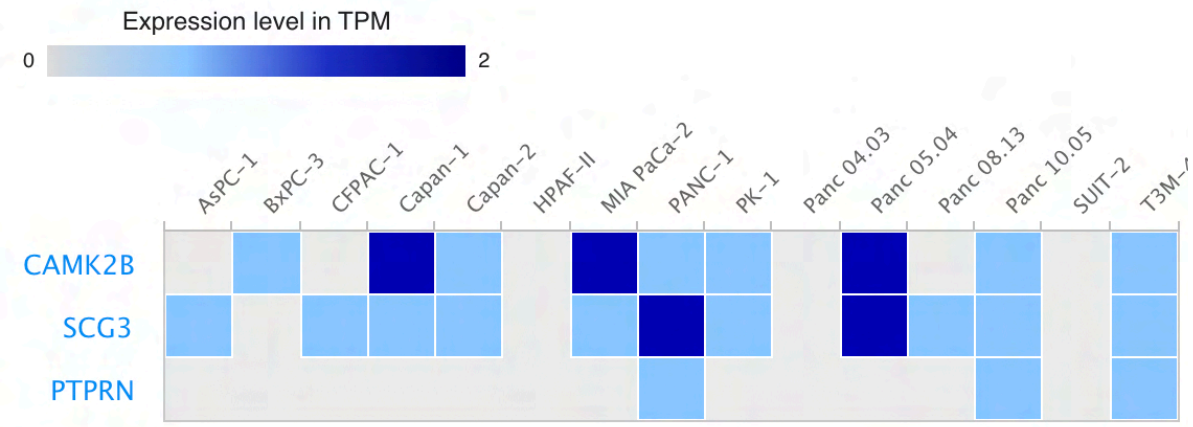
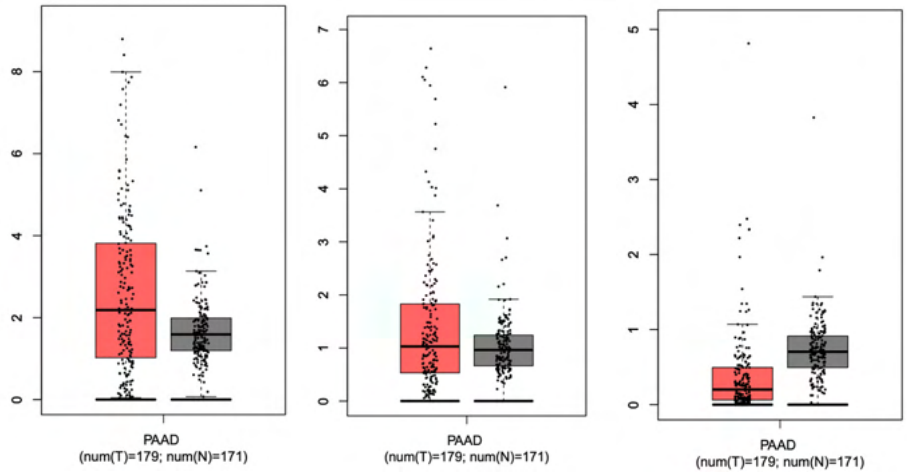


Figure 50 | RNA-Seq mRNA baseline gene expression atlas graph for 15 PC human cell lines for selected upregulated classical subtype co-expressed DEGs with significant P-value survival curves. The expression level per cell line is seen in transcripts per million (TPM) whereby the blue gradient shows the level of expression. High expression is seen towards a dark shade and low expression with a light shade. Graph produced by Expression Atlas - EMBL-EBI (https://www.ebi.ac.uk/gxa/home).

The tumour co-expressed DEGs with high expression levels in the map, *Fig. 50*, were used to produce expression box plots in from the GEPIA cancer database, *Fig.51*. This validation step shows the gene expression level in a PDAC patient sample relative to the adjacent normal.



SCG3 CAMK2B MYT1L

Figure 51 | Box plots for genes with high TPM expression levels for the PC cell lines observed in the expression atlas graph (*Fig. 50)* plotted with log2(TPM + 1) for log-scale. Red box plots are for the tumour samples (n = 179) and grey box plots represents the normal samples (n = 171) for PDAC. Plots produced by GEPIA (http://gepia.cancer-pku.cn/detail.php).

Table 9 shows the identified signature genes in PDAC and its expression values in the extracellular vesicles and the log_2FC from different databases, BBCancer and GENT2, respectively. An upregulation in the extracellular vesicles may indicate migration whereas the GENT2 values indicate the expression level in terms of the log_2FC. A negative value represents a downregulation and a positive value shows the upregulation level. It can be noted that these DEGs do not have a positive expression values in the extracellular vesicles, indicting the absence of migration caused by them. These DEGs were upregulated in the classical subtype, which is observed in the early stage, hence migration has yet to take place.

Table 9 | The BBCancer and GENT2 expression values (TPM) for the selected classical subtype co-expressed DEGs.

GENE	**BBCancer (Extracellular vesicles)**	**GENT2 Log_2FC**
SCG3	-0.92	-0.74
PTPRN	-0.52	-2.62
CAMK2B	-1.48	-0.27
MYT1L	-1.47	-0.22

4.3.4. Machine learning

A machine learning approach was used on all the DEGs per sample condition to serve as a step for calculating the recall score, precision score, F1 and accuracy score. This was done to show how well the model can predict that a sample belongs to its category; normal or PDAC based on the count values of the significant DEGs. As the insignificant genes were discarded based on P-values, it is expected to see high scores for the selected DEGs due to their statistical significance, hence predicting potential biomarkers. *Table 10* contains the score values per condition. Additionally, dataset 1 was used to test the trained system from dataset 2 which represented the external validation step for RFC.

Table 10 | The average scores for recall, precision, F1 and accuracy values (n=5) and standard deviation with dataset 1, the external validation step for RFC.

PDAC - Dataset 2	1	2	3	4	5	Ave (n=5)	Std
Recall_score	0,97	1,00	1,00	1,00	1,00	0,99	0,01
Precision_score	0,93	0,93	0,93	0,91	1,00	0,94	0,03
F1_score	0,96	0,96	0,96	0,94	1,00	0,96	0,02
Accuracy_score	0,94	0,92	0,92	0,89	1,00	0,93	0,04
PDAC - External Validation (Dataset 1)							
Recall_score	0,40	0,67	0,53	0,40	0,47	0,49	0,11
Precision_score	0,57	0,52	0,57	0,57	0,62	0,57	0,03
F1_score	0,57	0,67	0,64	0,57	0,64	0,62	0,04
Accuracy_score	1,00	0,67	0,80	1,00	1,00	0,89	0,15

5. Discussion

The application of bioinformatics methods and interpretation of large data is done via computational tools by combining mathematics, physics and biology to have an in-depth understanding of diseases. Hence, in cancer research, applying bioinformatics can be useful for biomarker discovery. This study was performed to identify potential biomarkers in PDAC, its subtype and stage of disease. Several bioinformatic tools and web tools were utilized to create an RNA-Seq analysis workflow which ended off with *in-silico* validation steps to provide corroborative evidence supporting the accuracy of this workflow. The various comparative conditions were included in the analysis to understand the molecular basis of PDAC. As mentioned previously, PDAC is the fourth leading cause of death with a low survival rate. The main contributing factor towards the high mortality rate is due to diagnoses at an advanced stage and PDAC subtype, thereby reducing the overall response rate. Therefore, by analysing and identifying DEGs unique to the compared conditions, diagnostic rates may improve subsequently aiding in effective therapeutic approaches.

Quality control was a vital step that was included for assuring the use of good sequencing files. FastQC and Trimmomatic assisted in removing bad sequencing files and low-quality regions. Quality control plots are seen in Appendix A. It can be observed that untrimmed files had bad base call values as the box and whisker plots are commonly seen in the red region (*Fig. 52*). When compared to the trimmed Fastq files, the box and whisker plots are located in the green to yellow region, indicating a good base call (*Fig. 53*). Therefore, the trimmed Fastq files produced good quality control plots. As mentioned previously, Trimmomatic has the ability to remove the adapter sequences whereby it locates these universal sequences by through the package. Additionally, Trimmomatic filtered out low quality reads that may interfere with alignment rate and downstream analysis [78].

Following quality control, files were inputted in DESeq2 for differential gene expression analysis. PCA plots were made to identify outliers and the variance observed in each dataset. *Fig. 10* shows the PCA plots for each dataset. Dataset 1 and 2 produced satisfactory plots following filtering of samples (*Fig. 10*). Each condition had unique clustering to allow for optimal downstream analysis and DEG identification as seen in *Fig. 11.*

5.1. DESeq2 analysis

5.1.1. Identified signature genes in the tumour samples

Through the combination of dataset 1 and dataset 2, the common differentially expressed genes were identified and may represent a universal biomarker. Dataset 1 study was conducted in patients from Japan whereas dataset 2 contained samples from multiple different countries (Canada, India, USA, Russia, Poland, China, Iraq, Serbia, Bulgaria and Denmark). Hence, these sample sets include different race, gender, ethnicity as well as patients with pre-existing diseases. The intersecting DE genes from D1 and D2 were selected for further analysis with a total of 588 upregulated genes and 256 downregulated DEGs as seen in *Fig. 12.* The DEGs were identified based on the P-value whereby genes with P-values of more than 5% were discarded (insignificant). Next, the significant DEGs (P-value of less than 5%) were condensed by selecting the highly significant P-values to provide a further 95% confidence interval rate. These DEGs were plotted in *Fig. 13* and represented the selected genes for further analysis. The upregulated DEGs were analysed to understand how gene amplification contributes towards the formation of PDAC and progression to a metastatic stage. Survival curves were generated for each of the DE genes (Appendix C) to detect and remove outliers and genes with insignificant P-values. The plots were combined on the Kaplan-Meier plotter to determine the

average survival rates with high or low gene expression (*Fig. 15*) for the upregulated DEGs. It was seen that there was a significant difference between the survival rates with high and low expression whereby high gene expression leaded to a decline in the probability of survival to a 0.02 within the time frame of 60 to 80 months and P-value of 7×10^{-6}.

Expression atlas or expression heat map was generated as secondary validation step whereby it produced a map of the selected genes in TPM values per pancreatic cancer cell line (*Fig. 16*). A few genes were highly expressed in all 15 pancreatic cancer cell lines namely, *S100A11*, *PKM*, *GPRC5A* and *LAMC2*. Gene *S100A11* had the highest expression with value of up to 5688 TPM.

The Calcium Binding Protein (S100A11) has been linked to various diseases including cancers via activating multiple signalling pathways such as the Wnt/B catenin, TGF-B1/SMAD, PI3K/AKT, Nf-kB signalling pathways [21]. S100A11 protein can interact with P53 (regulator of tumour growth), ANXA2 (repairment of cancer cells) and ANXA1 (promote ER to endosome cholesterol transport) [21]. Through the interaction with oncogenes and its respective pathways, S100A1 has a major role in tumour progression, cell proliferation, migration and invasion when highly expressed, therefore considered as an oncogene. High expression has been observed in PDAC, breast, colorectal, prostate and lung cancer. Li *et al*., 2021 evaluated the role of S100A1 expression PC tissue samples. S100A11 was linked to the pathological stage, infiltration rate, resistance to drugs and a shorter survival rate, hence suggested as a prognostic biomarker for PC due to its tumorigenic roles [79]. *S100A11* had a fold change (log_2FC) of 1.6 in the tumour samples as seen in *Fig. 13*. KEGG (*Fig. 14d*) analysis showed a high gene ratio with a significant P-value for the PI3K/AKT signalling pathway, showing that S100A11 upregulates the PI3K/AKT signalling pathway. Furthermore, receptor ligand activity was seen in the GO molecular function which is involved in the activation of the PI3K/AKT signalling pathway. As mentioned previously, PI3K/AKT has a tumorigenic role via the mutation of the p110α subunit which promotes cell proliferation and invasion [80]. Additionally, *S100A11* had the highest expression in the extravesicles (EV), 3.67, potentially could be a blood biomarker (*Table. 3*).

Pyruvate kinases (PKMs) have a role in the cell metabolism rate. *PKM* gene produces two isoforms, PKM1 and PKM2. PKM2, a metabolic regulator, promotes anabolic metabolism and an increase in lactate production for tumour cell growth [81]. It has a role in activating angiogenesis. Under a hypoxic condition, PKM2 and p63 translocate into the nucleus, resulting

in the transcription of VEGF-A. An upregulation of VEGF-A increases angiogenesis [81]. *PKM* was upregulated with a log_2FC of 1.4 (*Fig. 13*), present in the EV (2.36) and a GENT2 log_2FC of 1.04 (*Table. 3*). PKM is an oncogene that is highly expressed in PDAC with a poor survival rate [82]. Inhibition of *PKM2* caused ATP depletion which resulted in a decrease in cell proliferation, migration and induction of cell death [82].

The G protein-coupled receptor class C group 5 member (GPRC5A) is encoded by the *GPRC5A* gene. This protein is located on the membrane as well as in the endoplasmic reticulum (ER) [83]. It is involved in the interaction of retinoic acid, has a role in cell growth, and G proteins. In PC, there is a low expression of GPRC5A in the normal adjacent samples, however, highly expressed with a significant P-value (9.20×10^{-5}) in the tumour samples and an overall low survival rate (0.0026) [84]. Additionally, it enhances the resistance to GEM treatment along with promoting cell migration [84]. Studies have shown that cell lines with a mutated p53, an oncogene, exhibited high expression of GPRC5A and a wild type p53 with low GPRC5A expression. This suggested a crosstalk between GPRC5A and mutant p53. GPRC5A had a log_2FC of 2.6 in the tumour samples (*Fig. 13*) and has a GO cellular component role in the ER and an anchored component of the membrane (*Fig. 14a*). The log_2FC from GENT2 and this study are in close proximity, hence may act as a supporting for the gene expression observed in the tumour sample. GPRC5A is a potential prognostic biomarker for PDAC as an upregulation has been observed to be linked the pathological stage, larger tumour size and grade and therefore may determine the survival rate [84].

Integrin subunit alpha 2 (ITGA2), laminin subunit beta 3 (LAMB3), and laminin subunit gamma 2 (LAMC2) were all upregulated in the tumour sample with P-values in the same range. ITGA2, LAMB3 and LAMC2 had a log_2FC of 1.9, 2.8, 2.9 (*Fig. 13*), respectively. Islam *et al*., 2021 had performed bioinformatic analysis on PDAC samples and identified these three genes as interconnected hub protein coding genes that were linked to the pathological stages of PDAC [85]. These proteins were exclusively expressed in PDAC cell, showing the role in tumour progression. ITGA2 mediates cell-to-extracellular matrix (ECM) attachment that alters the stromal behaviour and ECM remodelling (seen in the KEGG pathway, *Fig. 14d*) [85]. This alteration promotes cellular transformation and enhancement of a tumorigenic environment. Additionally, it promotes a desmoplastic environment by activating cancer-associated fibroblasts (CAFs) [85]. It has been reported that LAMC2 is overexpressed in several cancers with a pivotal role in chemoresistance. Through reverse transcription polymerase chain

reaction (RT-qPCR), 114 PDAC patient samples were analysed. Along with high expression levels, LAMC2 was linked to poor survival rate and therefore a prognostic biomarker for survival [86]. EMT is a contributing factor to drug resistance, including GEM treatment. To determine the role of LAMC2 in drug resistance, small interfering RNA (siRNA) was used to silence the genomic region in PC cells. The inhibition of LAMC2 significantly decreased the EMT phenotype, increased the sensitivity of GEM and induced apoptosis of PC cells [86]. Furthermore, there was a decrease in cancer cell colony formation, migration and invasion. Overexpression of LAMC2 caused a significant increase in the mRNA expression levels ATP-binding cassette (ABC) transporters including ABCA1 [86].

5.1.2. Identified signature genes in the advanced stage of PDAC

PDAC patients are often asymptomatic and therefore diagnosed at an advanced stage. As metastasis takes place, systematic chemotherapy is the viable option, however, with low response rate. Analysing the DEGs expressed in the advanced stage may assist in the correct diagnoses and prognosis of disease. Survival curves were produced for the top 5% of DEGs in *Fig. 20* and outliers were removed. The overall survival curve for these genes shows that with high expression, the probability of survival comes down to 0.01 in 50 months with a P-value of 0.00013. The gene expression atlas showed elevated expression levels for *KRT13* and *IL6* for most of the PC cell lines (*Fig. 21*) and box plots showing a high difference in tumour expression when compared to the normal samples (*Fig. 22, KRT13* had a low significant differences between tumour normal).

Keratin 13 is encoded by the *KRT13* gene. Keratins are fibrous proteins that are involved in the structural framework of epithelial cells [87]. In breast cancer MCF-7 cells, overexpression of KRT13 enhanced cell proliferation, invasion and migration in mice. Whereas knockdown of KRT13 inhibited cell growth and aggressive behaviour. Via molecular analysis, KRT13 induced stemness in MCF-7 cells by increasing the expression of stemness related genes including c-Myc [87]. To unveil the interaction between KRT13 and c-Myc, loss and gain of c-Myc function in both overexpression and knockdown of KRT13 showed a direct relationship. In overexpressed KRT13, siRNA-mediated c-Myc suppression decreased the stemness and growth whilst overexpression of c-Myc in knowdown-KRT13 increased self-renewal, growth and EMT process. Hence, KRT13 has been proposed to be a biomarker for breast cancer [87]. In PDAC, the upregulation of KRT13 has not been thoroughly investigated, however, there

was a significant increase in log_2FC for the advanced stage of PDAC with a value of 3.60 (*Fig. 18*) and a TPM expression value of up to 5400 in numerous PC cell lines (*Fig. 21*).

Cytokine interleukin 6 (IL-6), encoded by the *IL6* gene, is part of the immune response to infections and has a cytokine receptor binding activity for the GO plots (*Fig. 19*) . Interestingly, tumour cells produce IL-6 that has both intrinsic and extrinsic functions in the tumour microenvironment (TME) for sustenance of a tumorigenic state [88]. Upon binding with IL-6R, IL-6 dimerizes with gp130 to activate its receptor Janus kinases (JAK). JAK phosphorylates both gp130 and itself. These sites then act as a docking station for the signal transducer and activator of transcription 3 (STAT3) which undergoes a tyrosine phosphorylation, allowing STAT to translocate into the nucleus and activate its target genes. The targeted genes include Bcl-2, c-Myc, Cyclin Ds, MMP2 and MMP9 [89]. Matrix metalloproteinase-2 (MMP2) and -9 have notable roles in cancer metastasis as these proteins degrade the ECM to promote tumour invasion [90]. The IL-6/JAK/STAT3 signalling pathway is overexpressed in many cancers including PDAC as well as promoting drug resistance. In PC, cytokines play a major role in metastasis as it mediates signals between cancer cells, the TME and CAFs [91]. There is an elevated level of IL-6 in PDAC patients when compared to healthy patients and high expression was observed in metastatic PDAC patients [91]. When compared to the early stage, IL-6 was upregulated in the advanced stage with a log_2FC of 1.8 (*Fig. 18*). Furthermore, MMP-9 had a log_2FC of 1.09 in the advanced stage. These findings are supported by literature as IL-6 activates the JAK/STAT3 signalling pathway which in turn activates its target gene, *MMP-9*. As this signalling cascade is observed in an advanced stage whereby metastasis has taken place, IL-6 may potentially serve as a biomarker for the advanced stage of PDAC. IL-6 also activates the expression of cyclin Ds. Cyclin Ds are known to promote cell cycle progression via phosphorylation of substrates to push the cycle past the gap phase 1 checkpoint to gain survival advantages and cell proliferation [92].

5.1.3. Identified signature genes in the basal-like subtype of PDAC

The basal-like subtype DEGs were evaluated to understand the genes involved in promoting the severity of the disease that is linked to low survival rates and drug resistance. Based on the gene expression atlas plot, *CAV1*, *DCBLD2* and *TGFB2* had high expression in most of the PC cell lines (*Fig. 26*). The overall survival curve for these genes shows that with high expression, the probability of survival comes down to 0.01 in 75 months with a P-value of 0.001 (*Fig. 25*). The box plots for these three genes display high TPM expression in PDAC tumour samples,

depicted in *Fig. 27*. Additionally, KRT13 and S100A2 had high expression values too. These genes were also identified in the tumour and advanced stage samples, discussed above.

Caveolin-1 (CAV1) is a scaffold protein that is responsible for multiple roles throughout the body. CAV1 is part of the caveolae which are small sacs in the membrane and plays a structural role, supported by the GO plots (*Fig. 24a*) [93]. CAV1 expression varies depending on the type of cancer to either act as a tumour suppressor or promoter of metastasis. Furthermore, CAV1 expression has been linked to therapy resistance as it promotes stromal heterogeneity and tumour-stroma communication [94]. In PDAC, stromal heterogeneity is one of the main contributors to therapy resistance as it promotes deamination of the nucleoside analogue (mentioned previously). In PDAC, high expression has been associated with tumour growth size and stage. As CAV1 was upregulated with a $\log_2$FC of 1.42 (*Fig. 23*) and EV value of 1.15 (*Table.* 5) in the basal-like subtype, this may suggest its role in subtype selection. To investigate the role of CAV1 in PDAC, silencing resulted disrupting its interaction with focal adhesion kinases which reduces cell adhesion and survival. Whereas, overexpression increased GEM therapy resistance [94]. Additionally, CAV1 expression is linked to metabolism and cholesterol homeostasis [93].

The Discoidin, CUB and LCCL Domain Containing 2 (DCBLD2) protein has a role in the negative regulation of cell growth and healing. DCBLD2 overexpression increases tumour development and EMT. Along with these roles, high expression promotes 5-FU drug resistance in colorectal cancer cells [95]. DCBLD2 was upregulated in the basal-like subtype with a $\log_2$FC of 1.5 (*Fig. 23*). Via transcriptomic analysis, DCBLD2 has been reported to play an oncogenic role in PDAC and therefore a prognostic and diagnostic biomarker [96].

The transforming Growth Factor Beta 2 (TGFB2) protein partakes in the TGF-β pathway that is known to activate target genes such as p21 for regulating cell cycle progression [20]. It plays a growth promoting role with dominant features such as stromal proliferation, EMT and cell cycle progression in the advanced stage of PDAC. The basal-like is often observed in the late stage [22]. This shows the role of TGFB2 in both progression and subtype of disease as it was upregulated with a $\log_2$FC of 1.28 (*Fig. 23*) and EV values of 0.27 (*Table. 5*).

The aforementioned genes are involved in drug resistance for the basal-like sample and may be investigated further with regards to their role in drug mechanism, metabolism and resistance.

5.1.4. Identified common signature genes in the basal-like subtype and advanced stage samples.

As mentioned previously, the basal-like subtype often displays the stage of disease as it is commonly observed in the advanced stage and linked to drug resistance. To understand the genes involved in both stage and subtype, the upregulated DEGs that intersected with both the advanced stage and basal-like subtype were evaluated with a total of 79 commonly expressed genes. Amongst the intersected significant genes, there was notable genes involved in cancer progression, migration and invasion. These genes are *WNT3A*, *TP63*, *KRT13*, *IGF2BP* and *IL1A*. KRT13 and its function in PDAC has been previously described.

The Wingless-related integration site (Wnts) are secreted during embryonic development to regulate cell cycle, mobility and differentiation. The canonical Wnt/β-catenin signalling pathway plays a pivotal role in regulating cell cycle progression, apoptosis, genetic stability and stem cell renewal [97]. This pathway is activated upon binding of the Wnt ligand to its respective transmembrane protein, consisting of frizzled and LRP5/6, for the activation of Dishevelled. This allows for the stabilization, accumulation and translocation of β-catenin from the cytoplasm into the nucleus for the activation of its target genes [16]. Wnts have a vital role in cancer stem cell signalling (CSC) as high expression of Wnt correlates to high CSC activity [98]. The CSC population is known to induce self-renewal and differentiation capabilities in several cancers including PDAC. A study had revealed the role of Wnt in PDAC using a Wnt antagonist, DKK1, and an anti-frizzled antibody. This experiment delayed the formation of PDAC [99]. Additionally, inducing Wnt expression levels in xenograft mice models had stimulated the progression of PanINs, leading to the formation of PDAC. Interestingly, Wnt expression is not found in normal pancreatic tissues and therefore suggesting the role of Wnt in tumour progression via the activation of CSC signalling pathways. In the basal-like subtype, Wnt had a $\log_2$FC of 1.02 and a $\log_2$FC of 1.01 in the advanced stage. As the $\log_2$FC values lies in close proximity to one another, this suggests the role of Wnt in progression as well as result in a severe subtype of PDAC. Furthermore, the upregulation of Wnt in the basal-like may indicate a role in drug resistance.

Tumour protein 63 (p63) is encoded by the *TP63* gene. It is a transcription factor that regulates gene expression levels for genes involved in the growth and structural development. Elevated expression levels of p63 have been observed in PDAC, lung, ovarian, colon and breast cancer [100, 101]. Under a normal condition, p63 induces cell death, suppresses metastasis and cell

cycle due to its tumour suppressor features, however, the functions are altered when overexpressed [102]. Its isoform, ΔNp63, has high expression levels with a dominant role of cell proliferation and survival. Furthermore, there is a direct correlation between ΔNp63 and poor clinical responses [102]. Upregulation of ΔNp63 is an identifier for the basal-like subtype and act as a master regulator to enhance the aggressiveness of the disease. Furthermore, ΔNp63 interacts with its target genes including S100A2 and IL1A (a pro-inflammatory cytokine) to drive tumour progression and angiogenesis [101]. S100A2 and IL1A were common genes identified with an average log_2FC of 2.49 and 1.05, respectively. This shows the role of cross talk between several genes that contribute towards the formation of the basal-like subtype in an advanced stage.

The insulin-like growth factor-2 mRNA-binding proteins (IGF2BP) are oncofoetal proteins that occur in embryogenesis or tumorous proteins. In cancer, it originates within the tumorous cells and enter the bloodstream, making it a potential tumour marker [103]. It has been reported that IGF2BPs have various roles such as metabolism, cell proliferation, migration and differentiation. In normal tissues, there is a low expression level, however, are re-expressed in cancerous cells along with c-Myc and KRAS. The mechanism has yet to be clearly understood in cancer, however, through transcriptomic analysis of 177 PDAC samples, it had been shown that IGF2BP has a tumour promoting role in PDAC [104]. Another study had shown the IGF2BP promotes cell proliferation by binding and stabilizing GLUT1, a glucose transporter, for cell survival [105]**.** Following DESeq analysis, *IGF2BP* had an overall log_2FC of 1.36 and therefore suggesting the role of *IGF2BP* gene in tumour progression and subtype.

5.2. WGCNA co-expression analysis

Co-expression analysis was done on dataset 2 as it fulfilled the sample number requirement. The purpose of co-expression is to identify novel genes that regulate the entire biological system in a specific condition. These novel genes are involved in different signalling pathways that are known to have an important role in cancer. Despite producing a biological network for each compared condition, genes with supportive survival curves were selected and taken into consideration. Co-expression was performed by calculating a threshold values per sample condition. This produced a scale independence and mean connectivity plot. The cluster dendrograms represent coloured modules containing co-expressed genes. *Fig 28,34, 40* and *46* all show the graphs and its selected threshold value used for downstream analysis.

5.2.1. Signature co-expressed upregulated DEGs in the tumour samples

Fig. 29 represents the co-expressed DEGs in the tumour samples. The network consisted of *KIF4A*, *SPAG5*, *RRM2*, *AURKA* and *CDCA5*. The average survival curve (*Fig. 31)* shows that with high expression, the probability of survival is close to 0 by 65 months with a highly significant P-value of 0.00032. Furthermore, the gene expression atlas and box plots for each gene display high expression levels in tumour samples (*Fig. 32* and *33*). KIF4A had the largest difference in expression level when compared to the normal samples in the box plots.

The kinesin family member 4A (KIF4A) protein is associated with condensing the chromosome arms, integrity and the central spindle during mitosis [106]. An abnormal amount of KIF4A results excessive spindle microtubule assembly which is often observed in several cancers such as PDAC, lung, colorectal, cervical, breast and prostate cancer [107]. It facilitates cell proliferation through inducing p21, an oncogene, to push the cell cycle to bypass the gap phase 1 and 2 check points. This allows for the division of cancerous cells, leading to a metastatic environment [106]. In PDAC, KIF4A upregulation was determined by analysing 86 patient samples. Western blotting and siRNA experiments were performed to determine the role of KIF4A in PDAC. Through silencing of *KIF4A*, tumour growth and metastasis had decreased significantly and therefore suggesting the role in PDAC development and progression *in vitro* and *in vivo* [108]. This new therapeutic target has been found to be co-expressed with SPAG5 in this project.

The sperm-associated Antigen 5 (SPAG5), a mitotic spindle protein, is a new oncogene that has been found to be upregulated is various cancers. It is involved in multiple pathways such as the FOXM1/MMP2 and Wnt/β-catenin signalling pathway [109]. The role of MMP2 and Wnt proteins in tumour development and progression has been describe above. In breast cancer, SPAG5 promotes the upregulation of the Wnt/β-catenin pathway and transcriptional activity of TCF4, a tumour suppressor protein that regulates gene expression. Furthermore, inhibition of Wnt reversed the effects of SPGA5, thus showing a relationship between SPAG5 and Wnt [110]. A protein-protein network for ovarian cancer, following the identification of DEGs, had shown the top 20 genes that has tumorigenic roles including KIF4A, CDCA5 and SPGA5 [111]. These co-expressed genes were also identified in *Fig. 29*. The expression levels of SPAG5 in PDAC has yet to be investigated experimentally, however, this study has identified the link between SPAG5 and KIF4A. This suggests that the co-expressed genes, SPAG5 and KIF4A may be potential biomarkers for PDAC.

Ribonucleotide reductase subunits M1 (RRM2) and Aurora-A kinase (AURKA) were also co-expressed within the network. RRM2 protein levels have been investigated in PDAC tissue samples whereby there was an over expression at a protein and mRNA level. This protein has a role in cell cycle progression, DNA repair and synthesis correlated with an overall low survival rate [112]. Interestingly, inhibition of AURKA expression in PDAC induced necrosis like death of cancerous cells. This was accompanied by reduced colony formation, invasion and migration. AURKA has a regulatory role in the mitotic cell division and high expression has been linked to poor prognosis of several cancers [113]. Oral administration of AURKA inhibitors had decreased tumour growth in mice models with PDAC along with a reduce in phosphorylating AURKA and GSK3β, a regulator of the Wnt/β catenin pathway [114]. These findings all suggest that SPAG5 may potentially be a biomarker for PDAC as three other co-expressed genes (KIF4A, RRM2 and AURKA) all have tumorigenic roles in PDAC. Studies have not yet reported a link between SPAG5, RRM2 and AURKA in PDAC. Due to the tumorigenic roles exerted by both proteins and co-expressed, the relationship could be further investigated.

Fig. 30 shows the cellular component, molecular function and biological processes of *KIF4A*, *SPAG5*, *AURKA* and *RMM2*. As seen in the *Fig. 30c*, these genes have roles in the spindle role, centromere, condensation, microtubule activity, protein kinase activity, spindle assembly, mitotic centromere separation and nuclear division. These functions are supported by literature, mentioned above, whereby each protein promotes the cell cycle progression through different activities. Furthermore, the KEGG pathway (*Fig. 30d*) includes cell cycle progression as well as steroid hormone biosynthesis, suggesting the potential role of cholesterol in PDAC.

5.2.2. Signature of co-expressed upregulated DEGs in the normal samples

Co-expression was performed on the normal sample set (*Fig. 35*) to see how the DEGs differ from the tumour set. As these DEGs (*ST18*, *GCK*, *SYP*) were upregulated in the normal, it is downregulated in the tumour samples. Hence, the *in-silico* validation steps which included the KM plot (*Fig. 37*), gene expression atlas plot (*Fig. 38*) and GEPIA box plots (*Fig. 39*) display low expression levels in PDAC when compared to the normal samples. KM plot shows that the upregulation of these DEGs correlate to a higher survival rate in PDAC patients whereas downregulation displays the inverse relationship, *Fig. 37*. Furthermore, *Table. 7* has negative expression levels for each of the DEGs in PDAC.

ST18 C2H2C-type zinc finger transcription factor (ST18) is a transcription factor located in the nucleus. It has several functions including cytokine-mediated signalling pathways and negative regulation of cell proliferation [115]. Cytokine signalling pathways play important roles in the human body as it required for several cellular functioning by activating intracellular cascades from the immune system to hormonal pathways. In breast cancer, *ST18* gene was revealed to be a transcriptional regulator by acting as a tumour suppressor gene. When overexpressed in MCF-7 cells, there was a reduction in colony formation and tumour growth in xenograft mouse models [116].

Glucokinase, encoded by the *GCK* gene, is primarily produced in the pancreas and has a major role in glucose metabolism to promote glucose homeostasis throughout the body [117]. Glucokinase acts as a glucose sensor in the β-cells which secrete insulin. Loss of function has been linked to several diabetic conditions, however, the loss of role in cancer has yet to be investigated [117]. *GCK* was upregulated in the normal sample but had a negative expression level in the tumour samples and extracellular vesicles, *Table. 7*.

Lastly, synaptophysin (SYP) was upregulated in the normal samples. The *SYP* gene encodes for a membrane protein of synaptic small vesicles in both brain and endocrine cells, also observed in GO plot (*Fig. 36a*) [118]. In glioma patients, the expression level of SYP links to the survival rate. Via the examination of over 600 patient data from TCGA in combination with a machine learning-based analysis, a decreased SYP expression level was associated with a higher tumour grade which resulted in a low survival rate. Conversely, patients with high SYP levels had an increase in survival rate. This revealed the prognostic role of SYP in glioma patients [119].

The upregulated DEGs in the normal samples all have roles in promoting efficient cellular signalling and homeostasis. Loss of function of these DEGs observed in other tumours are linked to promote tumour growth and lower survival periods.

5.2.3. Signature of co-expressed upregulated DEGs in the basal-like samples

The gene expression atlas (*Fig. 44*) showed three genes that had high expression levels observed in most of the PC cell lines namely *BUB1*, *DEPDC1* and *KIF14* with supportive boxplots (*Fig. 45*). Additionally, these genes had shown a correlation with high expression and low survival rates in the survival curves, as seen in *Fig. 43*.

The budding uninhibited by benzimidazole 1 (BUB1) is encoded by the *BUB1* gene [120]. This serine/threonine kinase protein plays a vital role in regulating the cell cycle via the recruitment of various proteins responsible for spindle checkpoints. During a cell division for both mitosis and meiosis, the correct attachment of a kinetochore allows for a proper cell division into two separate cells in anaphase. This is monitored by several checkpoints such as BUB1. Hence, unattachment of the kinetochore will be flagged and there will be a delay in anaphase [121]. This delay is done via signals emitted from the spindle checkpoint proteins. BUB1 initiates a signalling pathway that forms a complex to prevent the activity of the anaphase promoting complex/cyclosome (APC/C). The inhibition of this complex delays the anaphase process to allow for the correct attachment of the kinetochore, therefore regulating the cell cycle [121]. Furthermore, the functions are seen in the GO and KEGG pathways for cell division, serine/threonine kinase activity, mitotic and meiotic segregation, *Fig. 42*. *Table 8* depicts the expression value of EV and GENT2FC of 0.28 and 0.70, respectively.

When dysregulated, *BUB1* may have a dual function depending on the type of cancer. It can either act as an oncogene or a tumour suppressor gene in many cancers such as breast, PDAC, lung and prostate cancer [122]. The boxplots for *BUB1*, *Fig. 45*, shows that there is a high significant expression in PDAC when compared to the normal samples. Coincidentally, evidence has suggested the role of a dysregulated cell cycle as a hallmark of PDAC. A study had shown high expression levels of BUB1 in PDAC tissue samples via immunohistochemical staining of 99 PDAC and 71 normal pancreatic tissues. There was a direct correlation of tumour size and short survival rate with an upregulation of *BUB1*, suggesting that BUB1 could potentially be a prognostic biomarker for PDAC and serve as a therapeutic target [123].

Domain Containing 1 (DEPDC1) is a protein that regulates mitotic progression [124]. The expression of DEPDC1 has been associated with several diseases including various types of cancer. It is a prognostic marker for colorectal cancer [125]. The GO annotation includes GTPase activity which is also seen in *Fig. 42b.* Recent research has identified DEPDC1 as a tumour-related protein due to being upregulated in cancers and its tumorigenic properties. High expression was observed in breast cancer tissue samples via immunohistochemical staining [126]. The overexpression of DEPDC1 in MCF-7 cells promoted proliferation by regulating the cell cycle that increased the migration and invasion rate. Furthermore, DEPDC1 upregulated the PI3K/AKT signalling pathway [126]. *Fig. 3* showed the role of PI3K/AKT/mTOR in growth, translations, proliferation and survival for cancer cells. As

mentioned previously, the PI3K/AKT pathway is activated in 50% of PDAC patients. However, the function of DEPDC1 in PDAC has yet to be investigated but due to the linkage seen between DEPDC1 and PI3K/AKT pathway, it can be suggested as a potential biomarker.

Lastly, the Kinesin Family Member 14 (KIF14) is part of the microtubule motor proteins. Their functionality varies from vesicle transportation, cytokinesis and chromosomal segregation [127]. This is also seen in the GO plots (*Fig. 42*). An upregulation of KIF14 has been seen in retinoblastoma, lung and breast cancer cell lines and therefore, posed as a potential oncogene. The metastatic role has been investigated in gastric cancer. The upregulation of KIF14 had a direct link in colony formation, invasion, migration and TNM staging whereas knockdown of KIF14 reversed the roles [128]. Additionally, high expression was linked to the progression towards an advanced stage of gastric cancer. Upon knockdown, it was noted that there was a decrease in AKT phosphorylation. Inhibiting AKT phosphorylation prevents the activation of the PI3K/AKT pathway. These findings had concluded that KIF14 has a metastatic role in gastric cancer with poor prognosis [128]. KIF14 was one of the ten hub genes identified in PDAC that had been linked to progression of cancer through promoting the cell cycle, however, a wet lab study has yet to be conducted [129].

5.2.4. Signature of co-expressed upregulated DEGs in the classical subtype samples

Following co-expression analysis, the upregulated DEGs with thresholds towards a 1 displayed higher interconnection. Based on the network in *Fig. 47*, *PTPRN*, *SCG3*, *CAMK2B* and *MYT1L* DEGs were highly significant. The selected genes were expressed in the classical subtype which is known to have a higher response rate to therapy and therefore, a good survival rate. The survival curves for the selected genes show that elevated expression leads to a high survival rate whereas low expression leads to a low survival rate (*Fig. 49*.). DE analysis showed that these genes are being downregulated in the basal-like subtype and as mentioned before, basal-like show more resistance to treatment with low survival rates.

Protein Tyrosine Phosphatase Receptor Type N **(**PTPRN) is encoded by the *PTPRN* gene. This protein is part of the protein tyrosine phosphatase (PTPs) family which are signalling molecules that regulate multiple cellular pathways such as cell cycle, growth, differentiation and most importantly, oncogenic transformation [130]. PTPRN is involved in vesicle-mediated secretory processes in the pancreatic islets. The GO and KEGG diagrams, *Fig. 48*, display the

role in transport vesicles, regulation of type B pancreatic cell proliferation, synaptic vesicle maturation and type 1 diabetes mellitus. High expression of PTPRN has been linked to insulinoma and type 1 diabetes mellitus [131]. This may suggest that patients with type 1 diabetes or insulinoma may have a higher chance of developing PDAC. In PDAC, there is a balanced feedback loop mediated by protein tyrosine kinases (PTKs) and PTPs to activate or inhibit its target genes [132]. The role of PTPRN has not yet been evaluated directly in PDAC, however, in lung adenocarcinoma, PTPRN has been linked to poor prognosis and metastasis via the EMT process by upregulating the MEK/ERK and PI3K/AKT signalling pathways. Additionally, upregulation of PTPRN decreased the natural killer cell population that is responsible for exerting natural cytotoxicity against cancerous cells and infections [130].

CAMK2B was co-expressed with PTPRN, MYT1L and SCG3. Calcium/calmodulin dependent protein kinase II beta (CAMK2B) belongs to the serine protein kinase family. CAMK2B has a structural role in dendritic and synapse formation along with regulating the sarcoplasmic reticulum transport and the migration of developing neurons [133]. In *Fig. 48*, the GO and KEGG processes include sarcoplasmic reticulum, neuron projection, positive regulation of the dendritic spine morphogenesis and neuroactive ligand-receptor interaction. In PDAC, the role of CAMK2B has yet to be elucidated, however, it has been expressed in liver cancer. An inhibitor of CAMK2B, KN-62, had shown suppression effects on the hypoxia-inducible factor (HIF)-1α which has a role in promoting a hypoxic niche for cancerous cells via angiogenesis in hepatoma cells. Furthermore, there was a decrease in cancer cell proliferation and growth [134]. Additionally, high expression was shown to reduce cycle arrests and apoptosis. This relationship has been observed in colon, breast, prostate and gastric cancer [134]. Despite being linked to decrease cell cycle arrests and apoptosis in other cancers, the role of CAMK2B needs to be investigated in PDAC. The survival curve for CAMK2B (Appendix B, *Fig. 60*) show that with elevated expression levels, there is a high survival rate for PDAC patients. This may indicate that in a classical subtype, there is an inverse role of CAMK2B and when inhibited, there is an increase cell migration and proliferation along with a decrease in survival rate.

The functions of Secretogranin III (SCG3) and Myelin transcription factor 1 (MYT1L) in PDAC have not been investigated thoroughly, however, high expression has been linked to favourable outcomes with low malignancy in patients with glioblastoma [135, 136].

PTPRN, CAMK2B, SCG3 and MYT1L expression levels in PDAC have yet to be investigated and may present as a classical subtype diagnostic biomarkers as each survival curve show high survival rate with high expression. However, these genes are not expressed in the EV (*Table 9*) and therefore, not a blood biomarker.

5.3. Machine learning analysis

The RF model consists of classifications and multiple regression trees which are trained with bootstrap samples and aggregated model results to prevent the model form overfitting and data generalization. As the trees grow, the new information is included and the impurity is reduced. The new information comes from randomly shuffling the data provided for the model. In Alzheimer's disease, RF identified potential biomarkers and three brain areas that are affected by disease and classified separately from normal patients [64]. The scores produced in this study can be seen in *Table 10*. The precision score for all sample conditions were above 90%, hence the model correctly predicted the sample class/phenotype, i.e., the PDAC sample node was predicted to be part of the PDAC sample without the use of the phenotype column. Furthermore, the model had good values for the recall and accuracy scores. This supports that the identified potential biomarkers are associated PDAC. Furthermore, an external validation step was performed to test the trained system. Dataset 1 was used to validate and produced an average precision score of 0.57. Both datasets were taken from independent studies and pre-processed differently. Despite having differences in the small sample number (N = 6, P = 10) when compared to dataset 2, sample extraction from a different geographical location as well as methodology, the model rate for the correct phenotype prediction and accuracy rates was 57% and 100%, respectively. The low scored values may improve by the inclusion of additional datasets with larger sample numbers as an external study had performed RFC on datasets with over 1000 tumour samples in breast cancer [137]. This would provide further validation for the tested system and display trends that are distinctively observed in PDAC.

5.4. The role of cholesterol in PDAC

Several DEGs mentioned have roles in cholesterol homeostasis. As these genes are dysregulated in the tumour samples, thus depicting the role of cholesterol in PDAC development and progression. As mentioned previously, LAMC2 was upregulated in the tumour samples and is known to regulate the expression of ATP-binding cassette (ABC) transporters including ABCA1 in PDAC, and IL6 is known to induce expression of ABCA1.

Increasing studies have shown the role of cholesterol in cancer and its involvement in multidrug resistance (MDR). The ABC family has various biological roles and regulates multiple pathways. ABC transporters export various molecules such as hormones, sugars, lipids and vitamins [138]. However, elevated expression levels are observed in many cancers. ABCA1 protein is a lipid transporter that is produced by many tissues including the liver tissue. Its normal function exports cholesterol and phospholipids out of the cells for cellular lipid homeostasis.

Notably, high cholesterol levels have been observed in many cancers. High cholesterol causes an increase in lipid rafts that are known to function cellular signalling pathways, adhesion, differentiation and apoptosis [139]. The presence of lipid rafts provides hub signals for inducing EMT. Signalling pathways include the Wnt/β-catenin, Shh and TGF-β pathways, which have been previously described. Furthermore, PKM and IL6 (upregulated in the advanced stage) contribute towards a hypoxic niche. An increase in hypoxia has been shown to increase cholesterol, which in turn supports EMT proteins. Hence, it has been suggested that hypoxia-induced EMT may have a role in facilitating cholesterol synthesis. This has been supported by experiments whereby hypoxic induced cells promoted the coalesces of lipid rafts and therefore, activating the EMT signalling partway to aid in invasion and metastasis [139]. Additionally, cholesterol has a role of regulating the expression of MDR efflux transporters as it is a transcriptional regulator. Drug resistance was also observed in the hypoxic-induced cells as it influences the activity of MDR transporters for the regulation of anti-apoptotic proteins and cancer cell senescence [139].

The DEGs mentioned in the PDAC samples have been linked to the synthesis of cholesterol due to the proteins and signalling pathways that promote EMT and drug resistance. This shows that various upregulated DEGs, when expressed, have a connection in promoting the synthesis of cholesterol in PDAC. Hence, the role of cholesterol in PDAC must be explored.

6. Conclusion

Biomarker discovery plays a pivotal role in correct diagnosis and prognosis of a disease. PDAC is a type of a deadly cancer with high mortality rate linked to poor prognosis. Bioinformatic analysis was performed in this study to aid in improving the diagnosis and prognosis of PDAC. The stage and subtype of disease also plays a major part in the high mortality rate. Hence, this study was focused on identifying potential biomarkers in PDAC, its stage and subtype. DE and

co-expression analysis was done on datasets that met the required program conditions. The common DEGs from datasets 1 and 2 were identified to represent universal biomarkers as patient samples were extracted from various countries. The significant tumour DEGs included *S100A11*, *PKM*, *GPRC5A* and *ITGA2*. *S100A11* and *GPRC5A* that have roles in infiltration and progression of cancer as well as resistance to treatments. *PKM* increases proliferation and migration via activating angiogenesis. *ITGA2* allows for ECM remodelling, enhancement of cellular transformation and tumorigenic environment. These DEGs are also blood biomarkers due to their presence in the EVs. Lastly, LAMC2 supports EMT phenotype, GEM resistance, decreased apoptosis and increased ABCA1 expression. The advanced stage had two DEGs, *KRT13* and *IL6*. These DEGs have tumorigenic properties as well target c-Myc and cyclin Ds to push the cell cycle pass checkpoints. This shows the role in progressing the cancer to a metastatic stage. The basal-like showed DEGs (*CAV1*, *DCVLD2* and *TGFB2*) involved in both GEM and 5-FU resistance along with promoting EMT phenotype and stromal heterogeneity. Furthermore, the common DEGs present in the basal-like and advance stage would assist in understanding the link between subtype and stage. This included *WNT3A*, *TP63*, *KRT13* and *IGF2BP*. The signalling pathways activated by these DEGs promote cancer stem cell pathways, angiogenesis and stimulates PanIN progression towards an adenocarcinoma. Co-expression analysis for the tumour sample had *KIF4A*, *SPAG5*, *RRM2* and *AURKA* DEGs that interacted with the Wnt pathway. *KIF4A* and *RRM2* had a role in bypassing cell cycle checkpoints, DNA synthesis and repair. Co-expression of the basal-like subtype revealed *BUB1*, *DEPDC1* and *KIF14* DEGs. *BUB1* expression is a hallmark for PDAC and acts as an oncogene by promoting cell cycle progression. *DEPDC1* is upregulates the PI3K/AKT signalling pathway for migration and invasion. Lastly, *KIF14* expression is used for the TNM staging and a hub gene for PDAC. The classical subtype had *PTPRN* which increases EMT phenotype and decrease natural killer cells and *CAMK2B* that is known to create a hypoxic niche in liver cancer, leading to angiogenesis.

Table 11 | A summary of the identified DEGs in this study for PDAC, its subtype and stage of disease for biomarker usage.

DEGs	**PDAC biomarker**	**Blood biomarker (extracellular vesicles)**
S100A11	Prognostic biomarker [140]	YES
PKM	Potential biomarker	YES

GPRC5A	Prognostic biomarker [84]	NO
ITGA2	Potential diagnostic and prognostic biomarker [85]	YES
LAMB3	Potential diagnostic and prognostic biomarker [85]	NO
LAMC2	Potential diagnostic and prognostic biomarker [85]	NO
Advanced stage of PDAC		
KRT13	Potential new biomarker	NO
IL-6	Inflammatory biomarker [141]	NO
Basal-like subtype of PDAC		
CAV1	Prognostic biomarker [142]	YES
DCBLD2	Diagnostic biomarker [96]	NO
TGFB2	Prognostic biomarker [143]	YES
Basal-like subtype and advanced stage of PDAC		
Wnt	Prognostic biomarker [144]	NO
p63	Prognostic biomarker [145]	NO
IGF2BP	Potential new biomarker	YES
KRT13	Potential new biomarker	NO
Co-expressed DEGs in PDAC		
KIF4A	Potential new biomarker	YES
SPAG5	Potential new biomarker	YES
RRM2	Prognostic biomarker [112]	NO
AURKA	Prognostic biomarker [146]	YES
Co-expressed DEGs in basal-like PDAC		
BUB1	Prognostic biomarker [147]	YES
DEPDC1	Potential new biomarker	NO

KIF14	Prognostic biomarker [148]	YES
Co-expressed DEGs in classical PDAC		
PTPRN	Potential new biomarker	NO
CAMK2B	Potential new biomarker	NO
SCG3	Potential new biomarker	NO

This workflow used in this study was able to identify several existing biomarkers as well as new potential biomarkers (*Table. 11*). These potential biomarkers have significant roles in promoting cancer progression, aggression and resistance. Hence, may serve as a less invasive method for early diagnosis and improvement in treatments to increase the patient survival rate.

7. Future research

This study was performed to contribute towards gaining a deeper understanding of PDAC in the research field specifically in terms of genes being expressed in the tumour samples, stage and subtype to assist in biomarker discovery for improvements in terms of diagnosis and prognosis of disease. Through computational analysis, DEGs relevant to a compared condition were analysed along with *in-silico* validation research. However, a wet laboratory component would provide substantial support for the accuracy of this workflow and its results. This workflow narrowed down the search for the prime genes involved in tumour progression and drug resistance. This would be validated further with the use of RT-PCR (expression profiles of genes) and western blots to detect the presence of these proteins in PDAC patient's tumour and blood samples through the use of polyacrylamide gels and immunoblotting which utilizes enzymes, primary and secondary antibodies [149]. Lastly, as genes involved in drug resistance were identified, silencing or downregulating their expression in PC treated cell lines may be performed to see improvements in the efficacy of treatment and understanding the role of cholesterol in EMT and drug resistance. Hence, *in-vitro* and possibly *in- vivo* analysis would provide a strong foundation for the output of this workflow to establish these genes as biomarkers for PDAC diagnosis and prognosis. Thereafter, further validation of these biological biomarkers can be done with the use of patients' samples to determine their sensitivity and specificity rate.

8. Limitations of the study

This study had used available RNA-sequence data from open-source repositories. There was variability present due to these being patient data. A filtering step was performed to produce unique clustering per comparison condition. The removal of samples may remodel the biological information depending on the expressed genes. Furthermore, pre-existing diseases, disorders or intake of medications were not taken into consideration. These factors may have an impact on the expressed genes at the time of sequencing. This can be rectified by performing in-house sequencing on selected PDAC patients by taking account of the type of diseases, disorders, medications, and clinical characteristics for grouping. Analysis can be done per group to reduce the variance in the sample sets and therefore providing unique clustering in the PCA plots per compared condition.

Furthermore, datasets 1 and 2 differ in file format. Dataset 1 was downloaded as raw fastq files whereas dataset 2 was a count matrix file obtained from TCGA. The datasets used in this study were acquired through a rigorous search from multiple databases (EBI/EMBL) and published articles, showing the lack of resources and gap between PDAC and bioinformatic analysis. This current study was aimed to contribute towards the understanding of PDAC in the medical research field by analysing available and suitable sequencing files. Dataset 1 had used HISAT2 for alignment whereas dataset 2 used Bowtie2. Bowtie2 is a fast aligner, however, the limitation comes when short reads are present and may not be aligned correctly. HISAT2 had a 99.08% alignment sensitivity rate and Bowtie2 had a 97.68% rate, respectively [56]. Although the differences in alignment does not vary significantly, this variability can be prevented by performing in-house sequencing on PDAC patients and pre-processed in the same structure. A larger dataset was not available in a fastq file format for PDAC, hence the dissimilarity in the pre-processing of files does display a limitation in this study.

9. References

1. Sung, Hyuna, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. 2021. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* 71: 209–249. https://doi.org/10.3322/caac.21660.
2. Matthews, Helen K., Cosetta Bertoli, and Robertus A. M. de Bruin. 2022. Cell cycle control in cancer. *Nature Reviews Molecular Cell Biology* 23. Nature Publishing Group: 74–88. https://doi.org/10.1038/s41580-021-00404-3.
3. Herremans, Kelly M., Andrea N. Riner, Miles E. Cameron, Kelley L. McKinley, Eric W. Triplett, Steven J. Hughes, and Jose G. Trevino. 2022. The oral microbiome, pancreatic cancer and human diversity in the age of precision medicine. *Microbiome* 10: 93. https://doi.org/10.1186/s40168-022-01262-7.
4. Thomas, Stacy K., Jesse Lee, and Gregory L. Beatty. 2020. Paracrine and cell autonomous signalling in pancreatic cancer progression and metastasis. *EBioMedicine* 53: 102662. https://doi.org/10.1016/j.ebiom.2020.102662.
5. Rawla, Prashanth, Tagore Sunkara, and Vinaya Gaduputi. 2019. Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World Journal of Oncology* 10: 10–27. https://doi.org/10.14740/wjon1166.
6. Kotze, Urda Karin. 2022. A prospective, longitudinal, observational study to assess the health-related quality of life of patients with pancreatic ductal adenocarcinoma in the South African context. Master Thesis, Faculty of Health Sciences.
7. Tang, Sijia, Lei Shi, Breona T. Luker, Channen Mickler, Bhavana Suresh, Gregory B. Lesinski, Daping Fan, Yuan Liu, and Ming Luo. 2022. Modulation of the tumor microenvironment by armed vesicular stomatitis virus in a syngeneic pancreatic cancer model. *Virology Journal* 19: 32. https://doi.org/10.1186/s12985-022-01757-7.
8. Hidalgo, Manuel. 2010. Pancreatic Cancer. *New England Journal of Medicine* 362. Massachusetts Medical Society: 1605–1617. https://doi.org/10.1056/NEJMra0901557.
9. Slack, J.M. 1995. Developmental biology of the pancreas. *Development* 121: 1569–1580. https://doi.org/10.1242/dev.121.6.1569.
10. Leung, Po Sing, and Siu Po Ip. 2006. Pancreatic acinar cell: its role in acute pancreatitis. *The International Journal of Biochemistry & Cell Biology* 38: 1024–1030. https://doi.org/10.1016/j.biocel.2005.12.001.
11. Orth, Michael, Philipp Metzger, Sabine Gerum, Julia Mayerle, Günter Schneider, Claus Belka, Maximilian Schnurr, and Kirsten Lauber. 2019. Pancreatic ductal adenocarcinoma: biological hallmarks, current status, and future perspectives of combined modality treatment approaches. *Radiation Oncology* 14: 141. https://doi.org/10.1186/s13014-019-1345-6.
12. Oberstein, Paul E., and Kenneth P. Olive. 2013. Pancreatic cancer: why is it so hard to treat? *Therapeutic Advances in Gastroenterology* 6: 321–337. https://doi.org/10.1177/1756283X13478680.
13. Bardeesy, Nabeel, and Ronald A. DePinho. 2002. Pancreatic cancer biology and genetics. *Nature Reviews Cancer* 2. Nature Publishing Group: 897–909. https://doi.org/10.1038/nrc949.
14. Crnogorac-Jurcevic, Tatjana, Claude Chelala, Sayka Barry, Tomohiko Harada, Vipul Bhakta, Sam Lattimore, Stipo Jurcevic, Mary Bronner, Nicholas R. Lemoine, and Teresa A. Brentnall. 2013. Molecular Analysis of Precursor Lesions in Familial Pancreatic Cancer. *PLOS ONE* 8. Public Library of Science: e54830. https://doi.org/10.1371/journal.pone.0054830.

15. Helal, Clara, Marine Valéry, Michel Ducreux, Antoine Hollebecque, and Cristina Smolenschi. 2022. FGFR2 fusion in metastatic pancreatic ductal adenocarcinoma: Is there hope? *European Journal of Cancer* 176. Elsevier: 168–170. https://doi.org/10.1016/j.ejca.2022.09.016.
16. Morris, John P., Sam C. Wang, and Matthias Hebrok. 2010. KRAS, Hedgehog, Wnt and the twisted developmental biology of pancreatic ductal adenocarcinoma. *Nature Reviews Cancer* 10. Nature Publishing Group: 683–695. https://doi.org/10.1038/nrc2899.
17. Huang, Lamei, Zhixing Guo, Fang Wang, and Liwu Fu. 2021. KRAS mutation: from undruggable to druggable in cancer. *Signal Transduction and Targeted Therapy* 6. Nature Publishing Group: 1–20. https://doi.org/10.1038/s41392-021-00780-4.
18. Eser, S., A. Schnieke, G. Schneider, and D. Saur. 2014. Oncogenic KRAS signalling in pancreatic cancer. *British Journal of Cancer* 111. Nature Publishing Group: 817–822. https://doi.org/10.1038/bjc.2014.215.
19. Baer, Romain, Célia Cintas, Nicole Therville, and Julie Guillermet-Guibert. 2015. Implication of PI3K/Akt pathway in pancreatic cancer: When PI3K isoforms matter? *Advances in Biological Regulation* 59. The PI3-Kinase/Akt Pathway: From Signaling to Diseases: 19–35. https://doi.org/10.1016/j.jbior.2015.05.001.
20. Zhao, Ming, Lopa Mishra, and Chu-Xia Deng. 2018. The role of TGF-β/SMAD4 signaling in cancer. *International Journal of Biological Sciences* 14: 111–123. https://doi.org/10.7150/ijbs.23230.
21. Zhang, Linqiang, Tingting Zhu, Huilai Miao, and Bin Liang. 2021. The Calcium Binding Protein S100A11 and Its Roles in Diseases. *Frontiers in Cell and Developmental Biology* 9.
22. Ahmed, Sunjida, Azore-Dee Bradshaw, Shweta Gera, M. Zahidunnabi Dewan, and Ruliang Xu. 2017. The TGF-β/Smad4 Signaling Pathway in Pancreatic Carcinogenesis and Its Clinical Significance. *Journal of Clinical Medicine* 6: 5. https://doi.org/10.3390/jcm6010005.
23. Collisson, Eric A., Peter Bailey, David K. Chang, and Andrew V. Biankin. 2019. Molecular subtypes of pancreatic cancer. *Nature Reviews Gastroenterology & Hepatology* 16. Nature Publishing Group: 207–220. https://doi.org/10.1038/s41575-019-0109-y.
24. Espinet, Elisa, Lukas Klein, Ellen Puré, and Shiv K. Singh. 2022. Mechanisms of PDAC subtype heterogeneity and therapy response. *Trends in Cancer*. https://doi.org/10.1016/j.trecan.2022.08.005.
25. Moffitt, Richard A., Raoud Marayati, Elizabeth L. Flate, Keith E. Volmar, S. Gabriela Herrera Loeza, Katherine A. Hoadley, Naim U. Rashid, et al. 2015. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature Genetics* 47: 1168–1178. https://doi.org/10.1038/ng.3398.
26. Raphael, Benjamin J., Ralph H. Hruban, Andrew J. Aguirre, Richard A. Moffitt, Jen Jen Yeh, Chip Stewart, A. Gordon Robertson, et al. 2017. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* 32. Elsevier: 185-203.e13. https://doi.org/10.1016/j.ccell.2017.07.007.
27. Juiz, Natalia, Abdessamad Elkaoutari, Martin Bigonnet, Odile Gayet, Julie Roques, Rémy Nicolle, Juan Iovanna, and Nelson Dusetti. 2020. Basal-like and classical cells coexist in pancreatic cancer revealed by single-cell analysis on biopsy-derived pancreatic cancer organoids from the classical subtype. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 34: 12214–12228. https://doi.org/10.1096/fj.202000363RR.
28. Park, Min Young, Sang Hyun Shin, Ki Byung Song, DaeWook Hwang, Jae Hoon Lee, Young-Joo Lee, and Song Cheol Kim. 2019. Validation of the eighth edition of the

American Joint Committee on Cancer staging system and proposal of an improved staging system for pancreatic ductal adenocarcinoma. *Annals of Hepato-Biliary-Pancreatic Surgery* 23: 46–55. https://doi.org/10.14701/ahbps.2019.23.1.46.
29. Nicolson, Garth L. 1993. Cancer Progression and Growth: Relationship of Paracrine and Autocrine Growth Mechanisms to Organ Preference of Metastasis. *Experimental Cell Research* 204: 171–180. https://doi.org/10.1006/excr.1993.1022.
30. Foley, Kelly, Agnieszka A. Rucki, Qian Xiao, Donger Zhou, Ashley Leubner, Guanglan Mo, Jennifer Kleponis, et al. 2015. Semaphorin 3D autocrine signaling mediates the metastatic role of annexin A2 in pancreatic cancer. *Science Signaling* 8. American Association for the Advancement of Science: ra77–ra77. https://doi.org/10.1126/scisignal.aaa5823.
31. Lee, Jungwhoi, Jungsul Lee, Jeong Hun Yun, Chulhee Choi, Sayeon Cho, Seung Jun Kim, and Jae Hoon Kim. 2017. Autocrine DUSP28 signaling mediates pancreatic cancer malignancy via regulation of PDGF-A. *Scientific Reports* 7. Nature Publishing Group: 12760. https://doi.org/10.1038/s41598-017-13023-w.
32. Ungefroren, Hendrik. 2021. Autocrine TGF-β in Cancer: Review of the Literature and Caveats in Experimental Analysis. *International Journal of Molecular Sciences* 22: 977. https://doi.org/10.3390/ijms22020977.
33. Liang, Chen, Si Shi, Qingcai Meng, Dingkong Liang, Shunrong Ji, Bo Zhang, Yi Qin, Jin Xu, Quanxing Ni, and Xianjun Yu. 2017. Complex roles of the stroma in the intrinsic resistance to gemcitabine in pancreatic cancer: where we are and where we are going. *Experimental & Molecular Medicine* 49. Nature Publishing Group: e406–e406. https://doi.org/10.1038/emm.2017.255.
34. Gu, Zongting, Yongxing Du, Xueping Zhao, and Chengfeng Wang. 2021. Tumor microenvironment and metabolic remodeling in gemcitabine-based chemoresistance of pancreatic cancer. *Cancer Letters* 521: 98–108. https://doi.org/10.1016/j.canlet.2021.08.029.
35. Casale, Jarett, and Preeti Patel. 2022. Fluorouracil. In *StatPearls*. Treasure Island (FL): StatPearls Publishing.
36. Hegde, Vishwajit S., and Shivaraj Nagalli. 2022. Leucovorin. In *StatPearls*. Treasure Island (FL): StatPearls Publishing.
37. Reyhanoglu, Gizem, and Travis Smith. 2022. Irinotecan. In *StatPearls*. Treasure Island (FL): StatPearls Publishing.
38. Quinn, Bridget A., Nathaniel A. Lee, Timothy P. Kegelman, Praveen Bhoopathi, Luni Emdad, Swadesh K. Das, Maurizio Pellecchia, Devanand Sarkar, and Paul B. Fisher. 2015. Chapter Eight - The Quest for an Effective Treatment for an Intractable Cancer: Established and Novel Therapies for Pancreatic Adenocarcinoma. In *Advances in Cancer Research*, ed. Paul B. Fisher and Kenneth D. Tew, 127:283–306. Academic Press. https://doi.org/10.1016/bs.acr.2015.04.009.
39. Conroy, Thierry, Florence Castan, Anthony Lopez, Anthony Turpin, Meher Ben Abdelghani, Alice C. Wei, Emmanuel Mitry, et al. 2022. Five-Year Outcomes of FOLFIRINOX vs Gemcitabine as Adjuvant Therapy for Pancreatic Cancer: A Randomized Clinical Trial. *JAMA Oncology* 8: 1571–1578. https://doi.org/10.1001/jamaoncol.2022.3829.
40. PubChem. 2023. CCL2 - C-C motif chemokine ligand 2 (human). https://pubchem.ncbi.nlm.nih.gov/gene/CCL2/human. Accessed March 13.
41. Fei, Liyang, Xiaochen Ren, Haijia Yu, and Yifan Zhan. 2021. Targeting the CCL2/CCR2 Axis in Cancer Immunotherapy: One Stone, Three Birds? *Frontiers in Immunology* 12.
42. Tu, Megan M., Hany A. Abdel-Hafiz, Robert T. Jones, Annie Jean, Katelyn J. Hoff, Jason E. Duex, Ana Chauca-Diaz, et al. 2020. Inhibition of the CCL2 receptor, CCR2,

enhances tumor response to immune checkpoint therapy. *Communications Biology* 3. Nature Publishing Group: 1–12. https://doi.org/10.1038/s42003-020-01441-y.
43. Kukurba, Kimberly R., and Stephen B. Montgomery. 2015. RNA Sequencing and Analysis. *Cold Spring Harbor protocols* 2015. NIH Public Access: 951. https://doi.org/10.1101/pdb.top084970.
44. Stark, Rory, Marta Grzelak, and James Hadfield. 2019. RNA sequencing: the teenage years. *Nature Reviews Genetics* 20. Nature Publishing Group: 631–656. https://doi.org/10.1038/s41576-019-0150-2.
45. Chen, Li, Ruirui Yang, Tony Kwan, Chao Tang, Stephen Watt, Yiming Zhang, Guillaume Bourque, et al. 2020. Paired rRNA-depleted and polyA-selected RNA sequencing data and supporting multi-omics data from human T cells. *Scientific Data* 7. Nature Publishing Group: 376. https://doi.org/10.1038/s41597-020-00719-4.
46. Sudhagar, Arun, Gokhlesh Kumar, and Mansour El-Matbouli. 2018. Transcriptome Analysis Based on RNA-Seq in Understanding Pathogenic Mechanisms of Diseases and the Immune System of Fish: A Comprehensive Review. *International Journal of Molecular Sciences* 19: 245. https://doi.org/10.3390/ijms19010245.
47. Lin, Wei, Pawan Noel, Erkut H. Borazanci, Jeeyun Lee, Albert Amini, In Woong Han, Jin Seok Heo, et al. 2020. Single-cell transcriptome analysis of tumor and stromal compartments of pancreatic ductal adenocarcinoma primary tumors and metastatic lesions. *Genome Medicine* 12: 80. https://doi.org/10.1186/s13073-020-00776-9.
48. Duconseil, Pauline, Marine Gilabert, Odile Gayet, Celine Loncle, Vincent Moutardier, Olivier Turrini, Ezequiel Calvo, et al. 2015. Transcriptomic Analysis Predicts Survival and Sensitivity to Anticancer Drugs of Patients with a Pancreatic Adenocarcinoma. *The American Journal of Pathology* 185: 1022–1032. https://doi.org/10.1016/j.ajpath.2014.11.029.
49. Zhang, Hongen. 2016. Overview of Sequence Data Formats. In *Statistical Genomics: Methods and Protocols*, ed. Ewy Mathé and Sean Davis, 3–17. Methods in Molecular Biology. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-3578-9_1.
50. National Institutes of Health The Cancer Genome Atlas (TCGA). 2023. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v11.p8. Accessed March 8.
51. Brown, Joseph, Meg Pirrung, and Lee Ann McCue. 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.
52. FastQC Tutorial & FAQ. 2023. https://rtsf.natsci.msu.edu/genomics/tech-notes/fastqc-tutorial-and-faq/. Accessed March 8.
53. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
54. Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 12. Nature Publishing Group: 357–360. https://doi.org/10.1038/nmeth.3317.
55. Kim, Daehwan, Ben Langmead, and Steven L. Salzberg. 2014. HISAT: Hierarchical Indexing for Spliced Alignment of Transcripts. bioRxiv. https://doi.org/10.1101/012591.
56. Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* 37. Nature Publishing Group: 907–915. https://doi.org/10.1038/s41587-019-0201-4.

57. Sakharkar, Meena Kishore, Vincent T. K. Chow, and Pandjassarame Kangueane. 2004. Distributions of exons and introns in the human genome. *In Silico Biology* 4: 387–393.
58. Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
59. Danecek, Petr, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, et al. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10: giab008. https://doi.org/10.1093/gigascience/giab008.
60. Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31: 166–169. https://doi.org/10.1093/bioinformatics/btu638.
61. Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15: 550. https://doi.org/10.1186/s13059-014-0550-8.
62. Love, Michael, Constantin Ahlmann-Eltze, Forbes Kwame, and Simon Anders. 2022. Package 'DESeq2.'
63. Piper, Meeta Mistry, Radhika Khetani, Mary. 2017. Gene-level differential expression analysis with DESeq2. *Introduction to DGE - ARCHIVED*. October 13.
64. Song, Minseok, Hyeyoom Jung, Seungyong Lee, Donghyeon Kim, and Minkyu Ahn. 2021. Diagnostic Classification and Biomarker Identification of Alzheimer's Disease with Random Forest Algorithm. *Brain Sciences* 11. Multidisciplinary Digital Publishing Institute: 453. https://doi.org/10.3390/brainsci11040453.
65. Breiman, Leo. 2001. Random Forests. *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.
66. sklearn.ensemble.RandomForestClassifier. 2023. *scikit-learn*. https://scikit-learn/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Accessed March 13.
67. Langfelder, Peter, and Steve Horvath. 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 559. https://doi.org/10.1186/1471-2105-9-559.
68. WGCNA package: Frequently Asked Questions. 2022. https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/faq.html. Accessed December 31.
69. Zhang, Bin, and Steve Horvath. 2005. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* 4. De Gruyter. https://doi.org/10.2202/1544-6115.1128.
70. Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* 13: 2498–2504. https://doi.org/10.1101/gr.1239303.
71. Wu, Tianzhi, Erqiang Hu, Shuangbin Xu, Meijun Chen, Pingfan Guo, Zehan Dai, Tingze Feng, et al. 2021. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* 2: 100141. https://doi.org/10.1016/j.xinn.2021.100141.
72. Nagy, Ádám, Gyöngyi Munkácsy, and Balázs Győrffy. 2021. Pancancer survival analysis of cancer hallmark genes. *Scientific Reports* 11: 6047. https://doi.org/10.1038/s41598-021-84787-5.
73. Zwyea, S., L. Naji, and S. Almansouri. 2021. Kaplan-Meier plotter data analysis model in early prognosis of pancreatic cancer. *Journal of Physics: Conference Series* 1853. IOP Publishing: 012033. https://doi.org/10.1088/1742-6596/1853/1/012033.

74. Papatheodorou, Irene, Nuno A Fonseca, Maria Keays, Y Amy Tang, Elisabet Barrera, Wojciech Bazant, Melissa Burke, et al. 2018. Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Research* 46: D246–D251. https://doi.org/10.1093/nar/gkx1158.
75. Tang, Zefang, Chenwei Li, Boxi Kang, Ge Gao, Cheng Li, and Zemin Zhang. 2017. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research* 45: W98–W102. https://doi.org/10.1093/nar/gkx247.
76. Zuo, Zhixiang, Huanjing Hu, Qingxian Xu, Xiaotong Luo, Di Peng, Kaiyu Zhu, Qi Zhao, Yubin Xie, and Jian Ren. 2020. BBCancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers. *Nucleic Acids Research* 48: D789–D796. https://doi.org/10.1093/nar/gkz942.
77. Zhang, Xinyi, Dianfeng Liu, Yongjian Gao, Chao Lin, Qingwu An, Ye Feng, Yangyang Liu, Da Liu, Haoming Luo, and Dongxu Wang. 2021. The Biology and Function of Extracellular Vesicles in Cancer Development. *Frontiers in Cell and Developmental Biology* 9.
78. Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
79. Li, Xiaomin, Ning Qiu, and Qijuan Li. 2021. Prognostic Values and Clinical Significance of S100 Family Member's Individualized mRNA Expression in Pancreatic Adenocarcinoma. *Frontiers in Genetics* 12: 758725. https://doi.org/10.3389/fgene.2021.758725.
80. Yang, Jing, Ji Nie, Xuelei Ma, Yuquan Wei, Yong Peng, and Xiawei Wei. 2019. Targeting PI3K in cancer: mechanisms and advances in clinical trials. *Molecular Cancer* 18: 26. https://doi.org/10.1186/s12943-019-0954-x.
81. Zahra, Kulsoom, Tulika Dey, Ashish, Surendra Pratap Mishra, and Uma Pandey. 2020. Pyruvate Kinase M2 and Cancer: The Role of PKM2 in Promoting Tumorigenesis. *Frontiers in Oncology* 10.
82. James, Andrew D., Daniel A. Richardson, In-Whan Oh, Pishyaporn Sritangos, Thomas Attard, Lisa Barrett, and Jason I. E. Bruce. 2020. Cutting off the fuel supply to calcium pumps in pancreatic cancer cells: role of pyruvate kinase-M2 (PKM2). *British Journal of Cancer* 122. Nature Publishing Group: 266–278. https://doi.org/10.1038/s41416-019-0675-3.
83. PubChem. 2022. GPRC5A - G protein-coupled receptor class C group 5 member A (human). https://pubchem.ncbi.nlm.nih.gov/gene/GPRC5A/human. Accessed November 17.
84. Qian, Xuetian, Chengfei Jiang, Shanshan Shen, and Xiaoping Zou. 2021. GPRC5A: An emerging prognostic biomarker for predicting malignancy of Pancreatic Cancer based on bioinformatics analysis. *Journal of Cancer* 12: 2010–2022. https://doi.org/10.7150/jca.52578.
85. Islam, Shajedul, Takao Kitagawa, Byron Baron, Yoshihiro Abiko, Itsuo Chiba, and Yasuhiro Kuramitsu. 2021. ITGA2, LAMB3, and LAMC2 may be the potential therapeutic targets in pancreatic ductal adenocarcinoma: an integrated bioinformatics analysis. *Scientific Reports* 11. Nature Publishing Group: 10563. https://doi.org/10.1038/s41598-021-90077-x.
86. Okada, Yasuyuki, Naoki Takahashi, Tetsuji Takayama, and Ajay Goel. 2021. LAMC2 promotes cancer progression and gemcitabine resistance through modulation of EMT and ATP-binding cassette transporters in pancreatic ductal adenocarcinoma. *Carcinogenesis* 42: 546–556. https://doi.org/10.1093/carcin/bgab011.

87. Yin, Lijuan, Qinlong Li, Stefan Mrdenovic, Gina Chia-Yi Chu, Boyang Jason Wu, Hong Bu, Peng Duan, et al. 2022. KRT13 promotes stemness and drives metastasis in breast cancer through a plakoglobin/c-Myc signaling pathway. *Breast Cancer Research* 24: 7. https://doi.org/10.1186/s13058-022-01502-6.
88. Johnson, Daniel E., Rachel A. O'Keefe, and Jennifer R. Grandis. 2018. Targeting the IL-6/JAK/STAT3 signalling axis in cancer. *Nature Reviews Clinical Oncology* 15. Nature Publishing Group: 234–248. https://doi.org/10.1038/nrclinonc.2018.8.
89. Mauer, Jan, Jesse L. Denson, and Jens C. Brüning. 2015. Versatile functions for IL-6 in metabolism and cancer. *Trends in Immunology* 36. Elsevier: 92–101. https://doi.org/10.1016/j.it.2014.12.008.
90. Li, Hai, Zhenwei Qiu, Feng Li, and Chunlei Wang. 2017. The relationship between MMP-2 and MMP-9 expression levels with breast cancer incidence and prognosis. *Oncology Letters* 14: 5865–5870. https://doi.org/10.3892/ol.2017.6924.
91. van Duijneveldt, Gemma, Michael D.W. Griffin, and Tracy L. Putoczki. 2020. Emerging roles for the IL-6 family of cytokines in pancreatic cancer. *Clinical Science (London, England : 1979)* 134: 2091–2115. https://doi.org/10.1042/CS20191211.
92. Musgrove, Elizabeth A., C. Elizabeth Caldon, Jane Barraclough, Andrew Stone, and Robert L. Sutherland. 2011. Cyclin D as a therapeutic target in cancer. *Nature Reviews Cancer* 11. Nature Publishing Group: 558–572. https://doi.org/10.1038/nrc3090.
93. Nwosu, Zeribe Chike, Matthias Philip Ebert, Steven Dooley, and Christoph Meyer. 2016. Caveolin-1 in the regulation of cell metabolism: a cancer perspective. *Molecular Cancer* 15: 71. https://doi.org/10.1186/s12943-016-0558-7.
94. Ketteler, Julia, and Diana Klein. 2018. Caveolin-1, cancer and therapy resistance. *International Journal of Cancer* 143: 2092–2104. https://doi.org/10.1002/ijc.31369.
95. Xie, Pan, Fu-Qiang Yuan, Ma-Sha Huang, Wei Zhang, Hong-Hao Zhou, Xi Li, and Zhao-Qian Liu. 2021. DCBLD2 Affects the Development of Colorectal Cancer via EMT and Angiogenesis and Modulates 5-FU Drug Resistance. *Frontiers in Cell and Developmental Biology* 9: 669285. https://doi.org/10.3389/fcell.2021.669285.
96. Feng, Zengyu, Kexian Li, Yulian Wu, and Chenghong Peng. 2021. Transcriptomic Profiling Identifies DCBLD2 as a Diagnostic and Prognostic Biomarker in Pancreatic Ductal Adenocarcinoma. *Frontiers in Molecular Biosciences* 8: 659168. https://doi.org/10.3389/fmolb.2021.659168.
97. S, Patel, Alam A, Pant R, and Chattopadhyay S. 2019. Wnt Signaling and Its Significance Within the Tumor Microenvironment: Novel Therapeutic Insights. *Frontiers in immunology* 10. Front Immunol. https://doi.org/10.3389/fimmu.2019.02872.
98. Jd, Holland, Klaus A, Garratt An, and Birchmeier W. 2013. Wnt signaling in stem and cancer stem cells. *Current opinion in cell biology* 25. Curr Opin Cell Biol. https://doi.org/10.1016/j.ceb.2013.01.004.
99. Zhan, T., N. Rindtorff, and M. Boutros. 2017. Wnt signaling in cancer. *Oncogene* 36. Nature Publishing Group: 1461–1473. https://doi.org/10.1038/onc.2016.304.
100. Steurer, Stefan, Claudia Riemann, Franziska Büscheck, Andreas M. Luebke, Martina Kluth, Claudia Hube-Magg, Andrea Hinsch, et al. 2021. p63 expression in human tumors and normal tissues: a tissue microarray study on 10,200 tumors. *Biomarker Research* 9: 7. https://doi.org/10.1186/s40364-021-00260-5.
101. Somerville, Tim D.D., Yali Xu, Koji Miyabayashi, Hervé Tiriac, Cristian R. Cleary, Diogo Maia-Silva, Joseph P. Milazzo, David A. Tuveson, and Christopher R. Vakoc. 2018. TP63-Mediated Enhancer Reprogramming Drives the Squamous Subtype of Pancreatic Ductal Adenocarcinoma. *Cell reports* 25: 1741-1755.e7. https://doi.org/10.1016/j.celrep.2018.10.051.

102. Melino, G. 2011. p63 is a suppressor of tumorigenesis and metastasis interacting with mutant p53. *Cell Death & Differentiation* 18. Nature Publishing Group: 1487–1499. https://doi.org/10.1038/cdd.2011.81.
103. Chism, Stanley E., Robert C. Burton, and Noel L. Warner. 1978. Immunogenicity of oncofetal antigens: A review. *Clinical Immunology and Immunopathology* 11: 346–373. https://doi.org/10.1016/0090-1229(78)90059-4.
104. Glaß, Markus, Patrick Michl, and Stefan Hüttelmaier. 2020. RNA Binding Proteins as Drivers and Therapeutic Target Candidates in Pancreatic Ductal Adenocarcinoma. *International Journal of Molecular Sciences* 21: 4190. https://doi.org/10.3390/ijms21114190.
105. Huang, Shan, Zheng Wu, Yunyun Cheng, Wenzhen Wei, and Linlin Hao. 2019. Insulin-like growth factor 2 mRNA binding protein 2 promotes aerobic glycolysis and cell proliferation in pancreatic ductal adenocarcinoma via stabilizing GLUT1 mRNA. *Acta Biochimica Et Biophysica Sinica* 51: 743–752. https://doi.org/10.1093/abbs/gmz048.
106. Hou, Ping-Fu, Tao Jiang, Fang Chen, Pei-Cong Shi, Hai-Qing Li, Jin Bai, and Jun Song. 2018. KIF4A facilitates cell proliferation via induction of p21-mediated cell cycle progression and promotes metastasis in colorectal cancer. *Cell Death & Disease* 9. Nature Publishing Group: 1–15. https://doi.org/10.1038/s41419-018-0550-9.
107. Xue, Dan, Pu Cheng, Mengjiao Han, Xiyong Liu, Lijun Xue, Chenyi Ye, Ke Wang, and Jian Huang. 2018. An integrated bioinformatical analysis to evaluate the role of KIF4A as a prognostic biomarker for breast cancer. *OncoTargets and therapy* 11: 4755–4768. https://doi.org/10.2147/OTT.S164730.
108. Chen, Jing, Cui-Cui Zhao, Fei-Ran Chen, Guo-Wei Feng, Fei Luo, and Tao Jiang. 2021. KIF4A Regulates the Progression of Pancreatic Ductal Adenocarcinoma through Proliferation and Invasion. *BioMed Research International* 2021. Hindawi: e8249293. https://doi.org/10.1155/2021/8249293.
109. He, Ji, Andrew R. Green, Yan Li, Stephen Y. T. Chan, and Dong-Xu Liu. 2020. SPAG5: An Emerging Oncogene. *Trends in Cancer* 6: 543–547. https://doi.org/10.1016/j.trecan.2020.03.006.
110. Jiang, Jue, Juan Wang, Xin He, Wenqi Ma, Lei Sun, Qi Zhou, Miao Li, and Shanshan Yu. 2019. High expression of SPAG5 sustains the malignant growth and invasion of breast cancer cells through the activation of Wnt/β-catenin signalling. *Clinical and Experimental Pharmacology & Physiology* 46: 597–606. https://doi.org/10.1111/1440-1681.13082.
111. Gui, Ting, Chenhe Yao, Binghan Jia, and Keng Shen. 2021. Identification and analysis of genes associated with epithelial ovarian cancer by integrated bioinformatics methods. *PLOS ONE* 16. Public Library of Science: e0253136. https://doi.org/10.1371/journal.pone.0253136.
112. Klimaszewska-Wiśniewska, Anna, Karolina Buchholz, Izabela Neska-Długosz, Justyna Durślewicz, Dariusz Grzanka, Jan Zabrzyński, Paulina Sopońska, Alina Grzanka, and Maciej Gagat. 2021. Expression of Genomic Instability-Related Molecules: Cyclin F, RRM2 and SPDL1 and Their Prognostic Significance in Pancreatic Adenocarcinoma. *Cancers* 13: 859. https://doi.org/10.3390/cancers13040859.
113. Du, Ruijuan, Chuntian Huang, Kangdong Liu, Xiang Li, and Zigang Dong. 2021. Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy. *Molecular Cancer* 20: 15. https://doi.org/10.1186/s12943-020-01305-3.
114. Xie, Yangchun, Shan Zhu, Meizuo Zhong, Manhua Yang, Xiaofan Sun, Jinbao Liu, Guido Kroemer, et al. 2017. Inhibition of Aurora Kinase A Induces Necroptosis in Pancreatic Carcinoma. *Gastroenterology* 153: 1429-1443.e5. https://doi.org/10.1053/j.gastro.2017.07.036.

115. Henry, Cyndi, Anne-Françoise Close, and Jean Buteau. 2014. A Critical Role for the Neural Zinc Factor ST18 in Pancreatic β-Cell Apoptosis*. *Journal of Biological Chemistry* 289: 8413–8419. https://doi.org/10.1074/jbc.M114.554915.
116. Jandrig, Burkhard, Susanne Seitz, Bernd Hinzmann, Wolfgang Arnold, Burkhard Micheel, Konrad Koelble, Reiner Siebert, et al. 2004. ST18 is a breast cancer tumor suppressor gene at human chromosome 8q11.2. *Oncogene* 23. Nature Publishing Group: 9295–9302. https://doi.org/10.1038/sj.onc.1208131.
117. Sternisha, Shawn M., and Brian G. Miller. 2019. Molecular and Cellular Regulation of Human Glucokinase. *Archives of biochemistry and biophysics* 663: 199–213. https://doi.org/10.1016/j.abb.2019.01.011.
118. Cesca, F., P. Baldelli, F. Valtorta, and F. Benfenati. 2010. The synapsins: key actors of synapse function and plasticity. *Progress in Neurobiology* 91: 313–348. https://doi.org/10.1016/j.pneurobio.2010.04.006.
119. Xiao, Zheng, Shun Yao, Zong-ming Wang, Di-min Zhu, Ya-nan Bie, Shi-zhong Zhang, and Wen-li Chen. 2021. Multiparametric MRI Features Predict the SYP Gene Expression in Low-Grade Glioma Patients: A Machine Learning-Based Radiomics Analysis. *Frontiers in Oncology* 11: 663451. https://doi.org/10.3389/fonc.2021.663451.
120. Long, Zeling, Tong Wu, Qunyan Tian, Luke A. Carlson, Wanchun Wang, and Gen Wu. 2021. Expression and prognosis analyses of BUB1, BUB1B and BUB3 in human sarcoma. *Aging (Albany NY)* 13: 12395–12409. https://doi.org/10.18632/aging.202944.
121. Cairo, Gisela, Anne M. MacKenzie, and Soni Lacefield. 2020. Differential requirement for Bub1 and Bub3 in regulation of meiotic versus mitotic chromosome segregation. *Journal of Cell Biology* 219: e201909136. https://doi.org/10.1083/jcb.201909136.
122. Zhu, Li-Jing, Yan Pan, Xiao-Ying Chen, and Pan-Fei Hou. 2020. BUB1 promotes proliferation of liver cancer cells by activating SMAD2 phosphorylation. *Oncology Letters* 19: 3506–3512. https://doi.org/10.3892/ol.2020.11445.
123. Piao, Junjie, Lianhua Zhu, Jie Sun, Nan Li, Bing Dong, Yang Yang, and Liyan Chen. 2019. High expression of CDK1 and BUB1 predicts poor prognosis of pancreatic ductal adenocarcinoma. *Gene* 701: 15–22. https://doi.org/10.1016/j.gene.2019.02.081.
124. Mi, Yan, Chundong Zhang, Youquan Bu, Ying Zhang, Longxia He, Hongxia Li, Huifang Zhu, Yi Li, Yunlong Lei, and Jiang Zhu. 2015. DEPDC1 is a novel cell cycle related gene that regulates mitotic progression. *BMB Reports* 48: 413–418. https://doi.org/10.5483/BMBRep.2015.48.7.036.
125. DEPDC1 protein expression summary - The Human Protein Atlas. 2023. https://www.proteinatlas.org/ENSG00000024526-DEPDC1. Accessed January 8.
126. Zhao, Huishan, Mingwei Yu, Laijian Sui, Benjiao Gong, Bo Zhou, Jian Chen, Zhaohua Gong, and Cuifang Hao. 2019. High Expression of DEPDC1 Promotes Malignant Phenotypes of Breast Cancer Cells and Predicts Poor Prognosis in Patients With Breast Cancer. *Frontiers in Oncology* 9.
127. PubChem. 2023. KIF14 - kinesin family member 14 (human). https://pubchem.ncbi.nlm.nih.gov/gene/KIF14/human. Accessed January 8.
128. Yang, Zhongyin, Chen Li, Chao Yan, Jianfang Li, Min Yan, Bingya Liu, Zhenggang Zhu, Yingli Wu, and Qinlong Gu. 2019. KIF14 promotes tumor progression and metastasis and is an independent predictor of poor prognosis in human gastric cancer. *Biochimica Et Biophysica Acta. Molecular Basis of Disease* 1865: 181–192. https://doi.org/10.1016/j.bbadis.2018.10.039.
129. Zhou, Zhou, Yian Cheng, Yinan Jiang, Shi Liu, Meng Zhang, Jing Liu, and Qiu Zhao. 2018. Ten hub genes associated with progression and prognosis of pancreatic carcinoma identified by co-expression analysis. *International Journal of Biological Sciences* 14: 124–136. https://doi.org/10.7150/ijbs.22619.

130. Song, Xinyue, Xue Jiao, Han Yan, Lifeng Yu, Longyang Jiang, Ming Zhang, Lianze Chen, et al. 2021. Overexpression of PTPRN Promotes Metastasis of Lung Adenocarcinoma and Suppresses NK Cell Cytotoxicity. *Frontiers in Cell and Developmental Biology* 9.
131. PTPRN protein tyrosine phosphatase receptor type N [Homo sapiens (human)] - Gene - NCBI. 2022. https://www.ncbi.nlm.nih.gov/gene/5798. Accessed November 24.
132. Ruckert, Mariana Tannús, Pamela Viani de Andrade, Verena Silva Santos, and Vanessa Silva Silveira. 2019. Protein tyrosine phosphatases: promising targets in pancreatic ductal adenocarcinoma. *Cellular and Molecular Life Sciences* 76: 2571–2592. https://doi.org/10.1007/s00018-019-03095-4.
133. Küry, Sébastien, Geeske M. van Woerden, Thomas Besnard, Martina Proietti Onori, Xénia Latypova, Meghan C. Towne, Megan T. Cho, et al. 2017. De Novo Mutations in Protein Kinase Genes CAMK2A and CAMK2B Cause Intellectual Disability. *American Journal of Human Genetics* 101: 768–788. https://doi.org/10.1016/j.ajhg.2017.10.003.
134. He, Qi, and Zhenyu Li. 2021. The dysregulated expression and functional effect of CaMK2 in cancer. *Cancer Cell International* 21: 326. https://doi.org/10.1186/s12935-021-02030-7.
135. Wang, Yi, Nan Ji, Junmei Wang, Jingli Cao, Deling Li, Yang Zhang, and Liwei Zhang. 2021. SCG3 Protein Expression in Glioma Associates With less Malignancy and Favorable Clinical Outcomes. *Pathology and Oncology Research* 27: 594931. https://doi.org/10.3389/pore.2021.594931.
136. Melhuish, Tiffany A., Izabela Kowalczyk, Arkadi Manukyan, Ying Zhang, Anant Shah, Roger Abounader, and David Wotton. 2018. Myt1 and Myt1l transcription factors limit proliferation in GBM cells by repressing YAP1 expression. *Biochimica Et Biophysica Acta. Gene Regulatory Mechanisms* 1861: 983–995. https://doi.org/10.1016/j.bbagrm.2018.10.005.
137. Wenric, Stephane, and Ruhollah Shemirani. 2018. Using Supervised Learning Methods for Gene Selection in RNA-Seq Case-Control Studies. *Frontiers in Genetics* 9.
138. Adamska, Aleksandra, and Marco Falasca. 2018. ATP-binding cassette transporters in progression and clinical outcome of pancreatic cancer: What is the way forward? *World Journal of Gastroenterology* 24: 3222–3238. https://doi.org/10.3748/wjg.v24.i29.3222.
139. Abdulla, Naaziyah, C. Theresa Vincent, and Mandeep Kaur. 2021. Mechanistic Insights Delineating the Role of Cholesterol in Epithelial Mesenchymal Transition and Drug Resistance in Cancer. *Frontiers in Cell and Developmental Biology* 9.
140. Wu, Yu, Qi Zhou, Fangyue Guo, Mingming Chen, Xufeng Tao, and Deshi Dong. 2021. S100 Proteins in Pancreatic Cancer: Current Knowledge and Future Perspectives. *Frontiers in Oncology* 11.
141. Kim, Hyoung Woo, Jong-chan Lee, Kyu-hyun Paik, Jingu Kang, Jaihwan Kim, and Jin-Hyeok Hwang. 2017. Serum interleukin-6 is associated with pancreatic ductal adenocarcinoma progression pattern. *Medicine* 96: e5926. https://doi.org/10.1097/MD.0000000000005926.
142. Chatterjee, Moumita, Edgar Ben-Josef, Dafydd G. Thomas, Meredith A. Morgan, Mark M. Zalupski, Gazala Khan, Charles Andrew Robinson, et al. 2015. Caveolin-1 is Associated with Tumor Progression and Confers a Multi-Modality Resistance Phenotype in Pancreatic Cancer. *Scientific Reports* 5: 10867. https://doi.org/10.1038/srep10867.
143. Zhao, J., Y. Liang, Q. Yin, S. Liu, Q. Wang, Y. Tang, and C. Cao. 2016. Clinical and prognostic significance of serum transforming growth factor-beta1 levels in patients with pancreatic ductal adenocarcinoma. *Brazilian Journal of Medical and Biological Research* 49: e5485. https://doi.org/10.1590/1414-431X20165485.

144. Dart, Dafydd A., Damla E. Arisan, Sioned Owen, Chunyi Hao, Wen G. Jiang, and Pinar Uysal-Onganer. 2019. Wnt-11 Expression Promotes Invasiveness and Correlates with Survival in Human Pancreatic Ductal Adeno Carcinoma. *Genes* 10. Multidisciplinary Digital Publishing Institute: 921. https://doi.org/10.3390/genes10110921.
145. Taherian, Mehran, Hua Wang, and Huamin Wang. 2022. Pancreatic Ductal Adenocarcinoma: Molecular Pathology and Predictive Biomarkers. *Cells* 11. Multidisciplinary Digital Publishing Institute: 3068. https://doi.org/10.3390/cells11193068.
146. Gomes-Filho, Sandro Mascena, Edmilson Ozorio Dos Santos, Ester Risério Matos Bertoldi, Luiza Coimbra Scalabrini, Vitor Heidrich, Bianca Dazzani, Elena Levantini, Eduardo Moraes Reis, and Daniela Sanchez Bassères. 2020. Aurora A kinase and its activator TPX2 are potential therapeutic targets in KRAS-induced pancreatic cancer. *Cellular Oncology (Dordrecht)* 43: 445–460. https://doi.org/10.1007/s13402-020-00498-5.
147. Piao, Junjie, Lianhua Zhu, Jie Sun, Nan Li, Bing Dong, Yang Yang, and Liyan Chen. 2019. High expression of CDK1 and BUB1 predicts poor prognosis of pancreatic ductal adenocarcinoma. *Gene* 701: 15–22. https://doi.org/10.1016/j.gene.2019.02.081.
148. Klimaszewska-Wiśniewska, Anna, Izabela Neska-Długosz, Karolina Buchholz, Justyna Durślewicz, Dariusz Grzanka, Anna Kasperska, Paulina Antosik, Jan Zabrzyński, Alina Grzanka, and Maciej Gagat. 2021. Prognostic Significance of KIF11 and KIF14 Expression in Pancreatic Adenocarcinoma. *Cancers* 13: 3017. https://doi.org/10.3390/cancers13123017.
149. Mahmood, Tahrin, and Ping-Chang Yang. 2012. Western Blot: Technique, Theory, and Trouble Shooting. *North American Journal of Medical Sciences* 4: 429–434. https://doi.org/10.4103/1947-2714.100998.

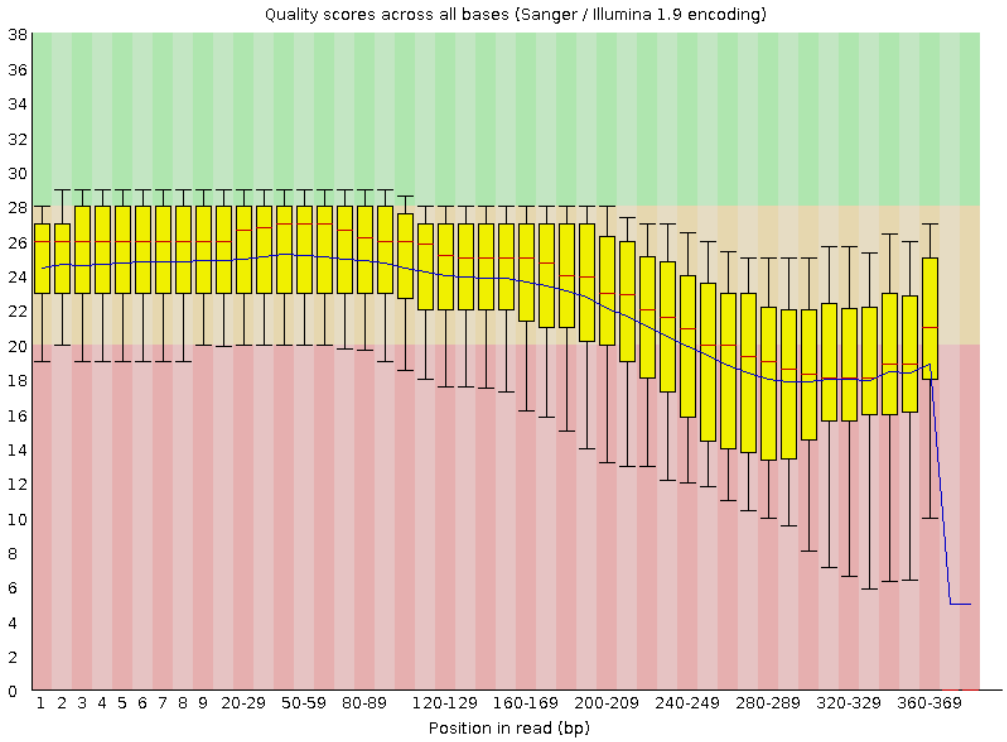
10. Appendices

10.1. Appendix A: Fastq Files

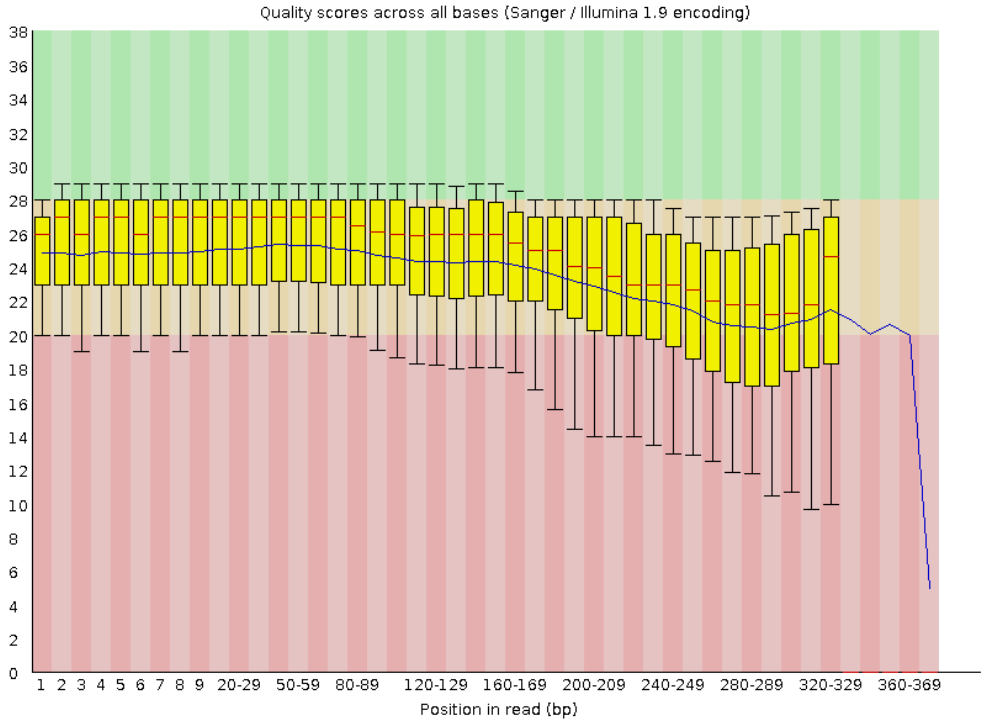
10.1.1. FastQC files for dataset 1

Untrimmed Fastq files:

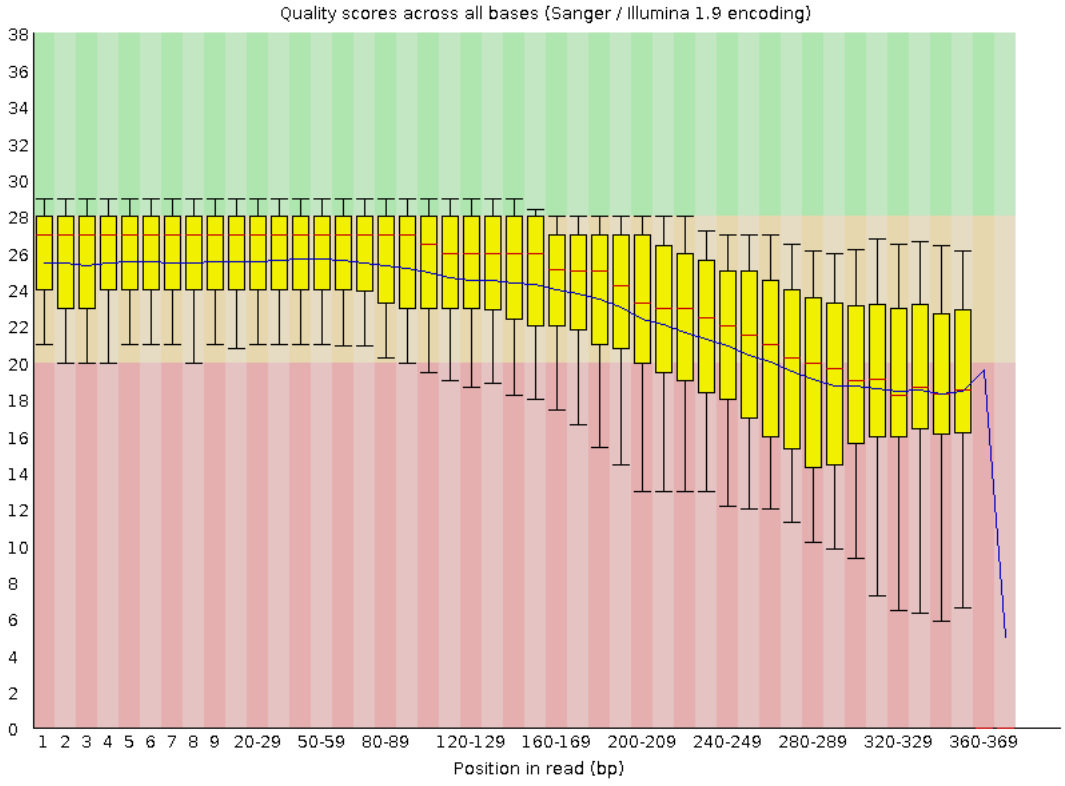
SRR17867010.fastq.gz



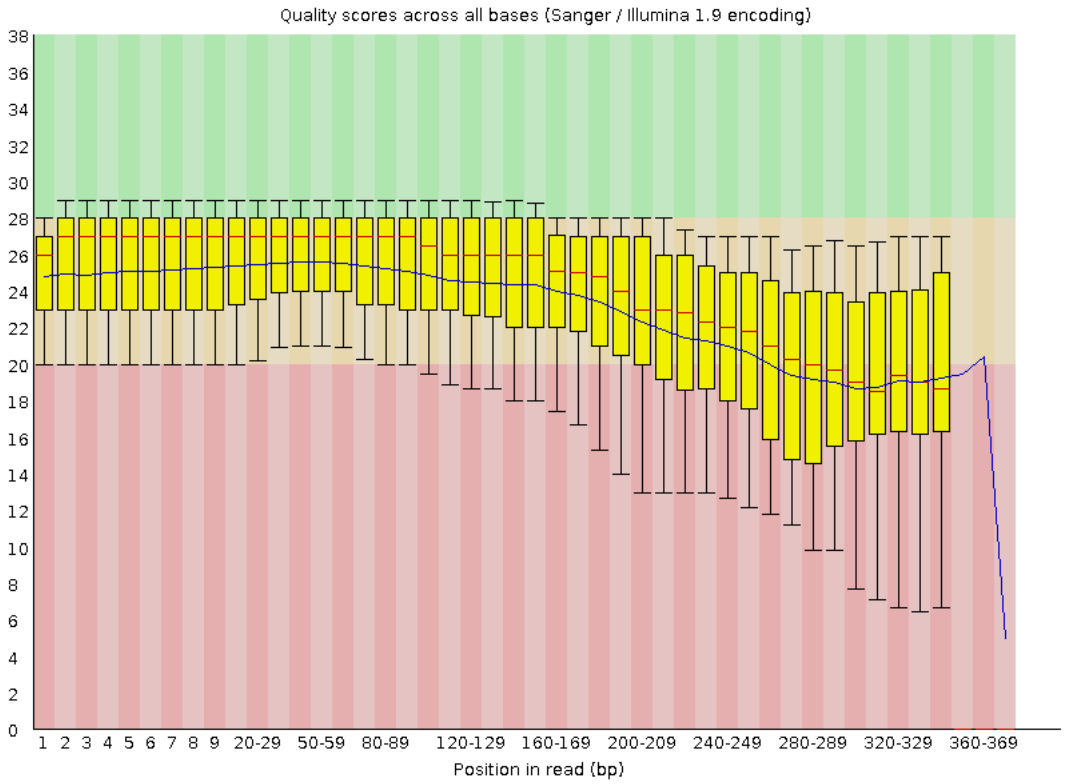
SRR17867012.fastq.gz



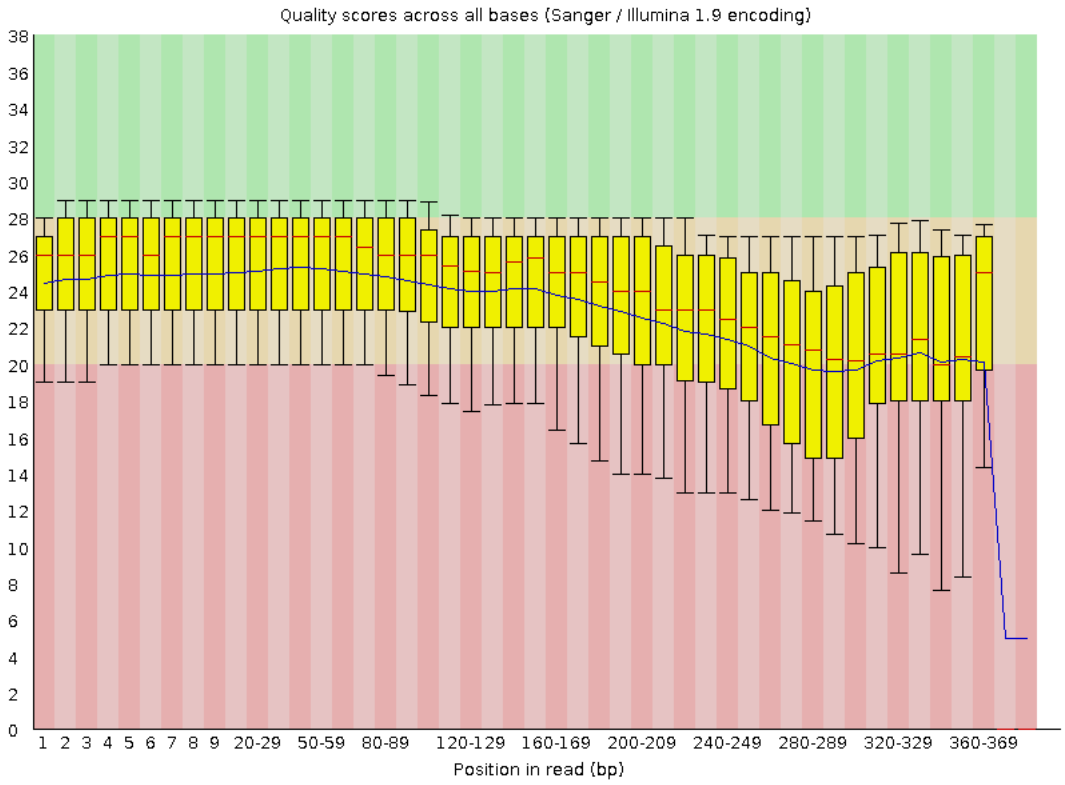
SRR17867013.fastq.gz



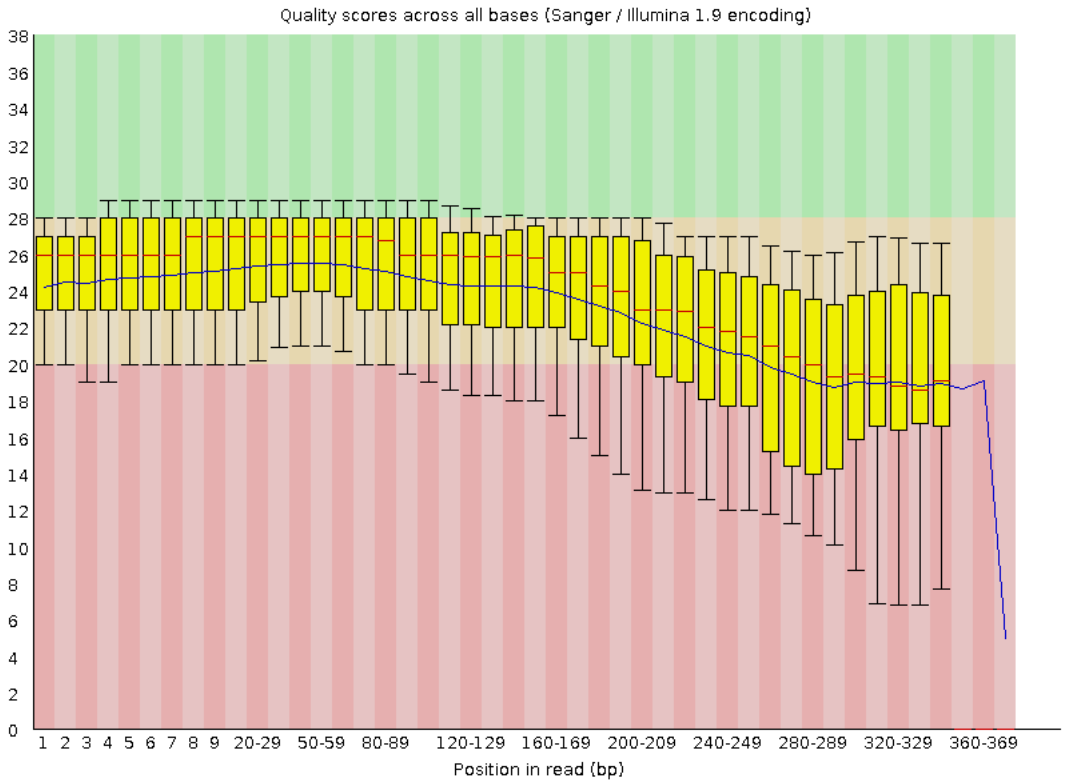
SRR17867014.fastq.gz



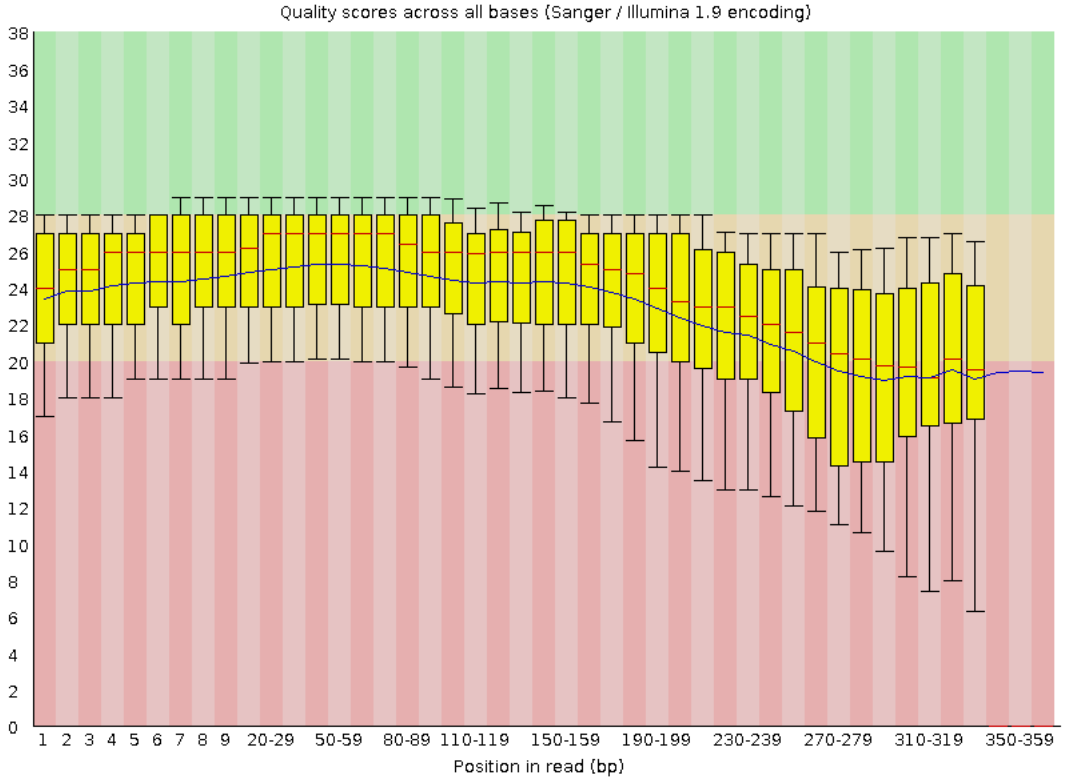
SRR17867015.fastq.gz



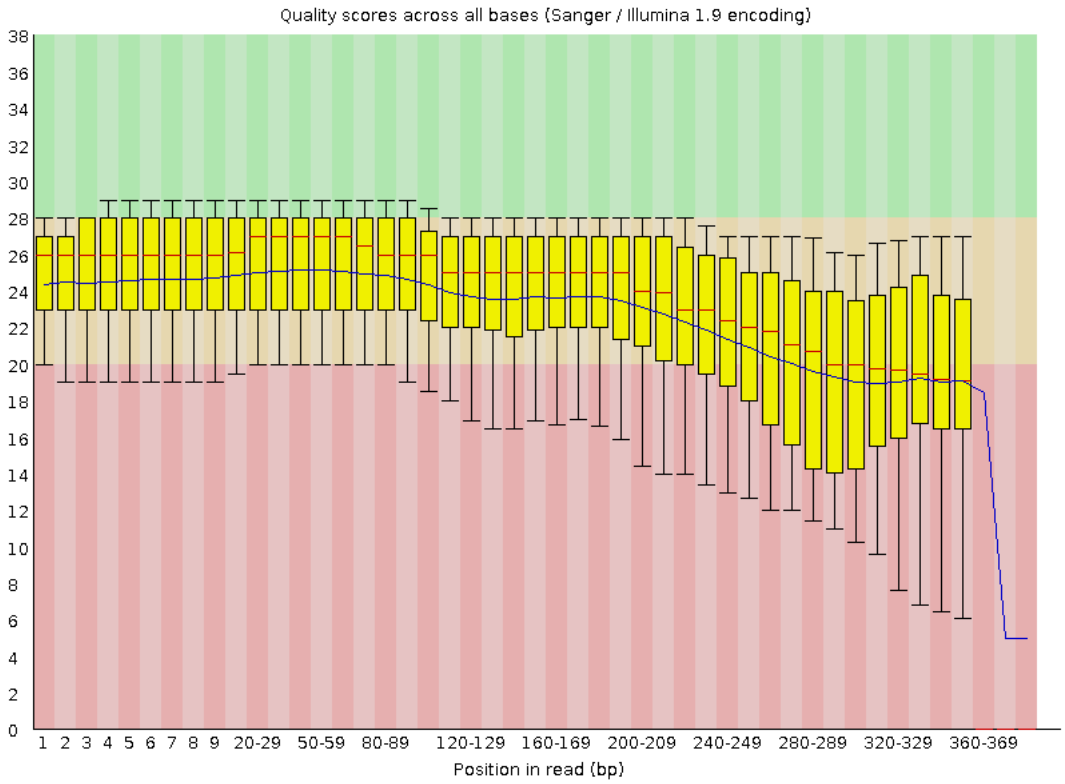
SRR17867016.fastq.gz



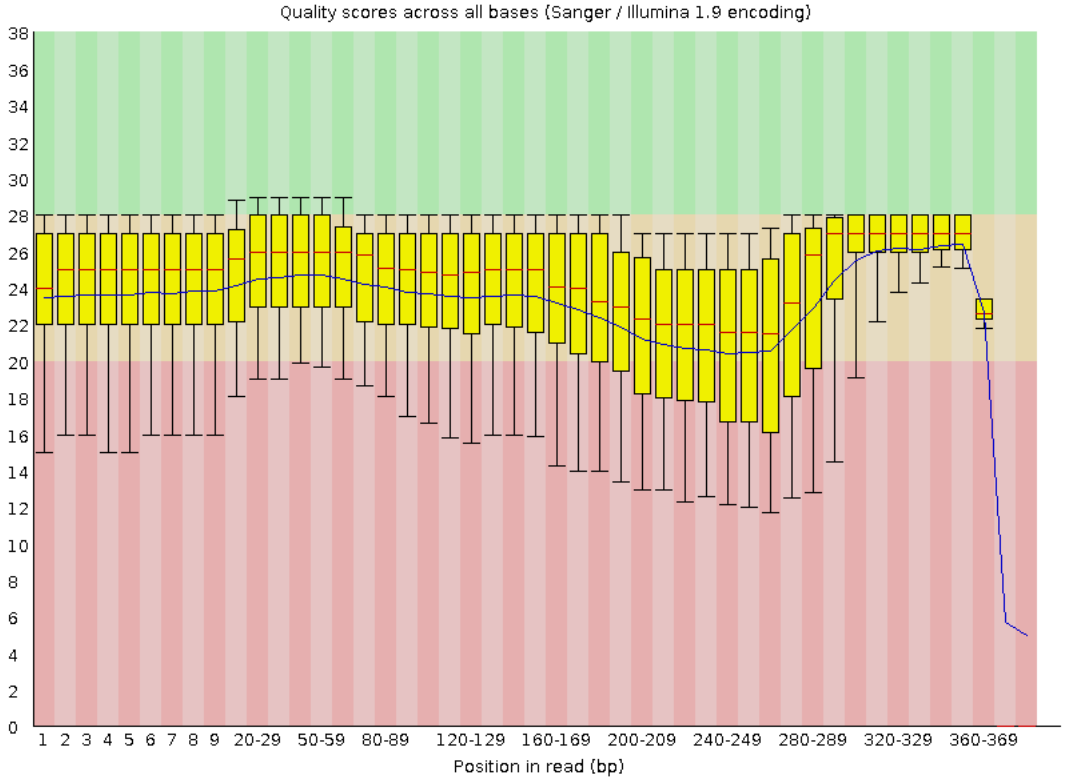
SRR17867017.fastq.gz



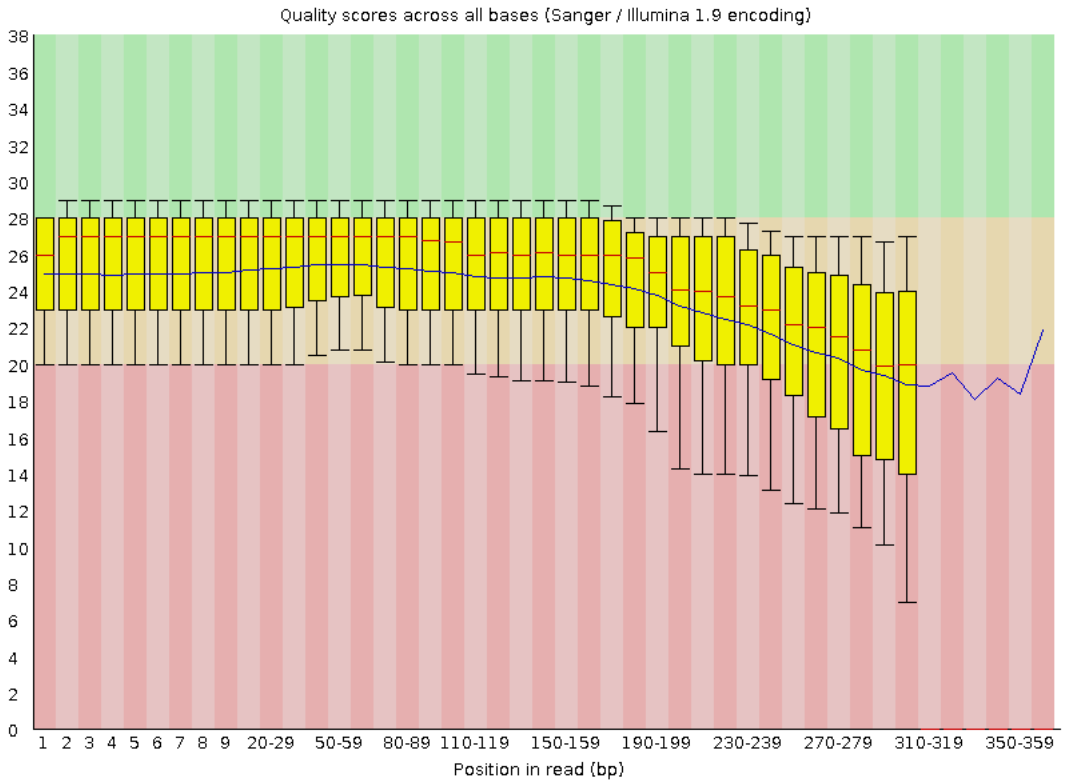
SRR17867017.fastq.gz



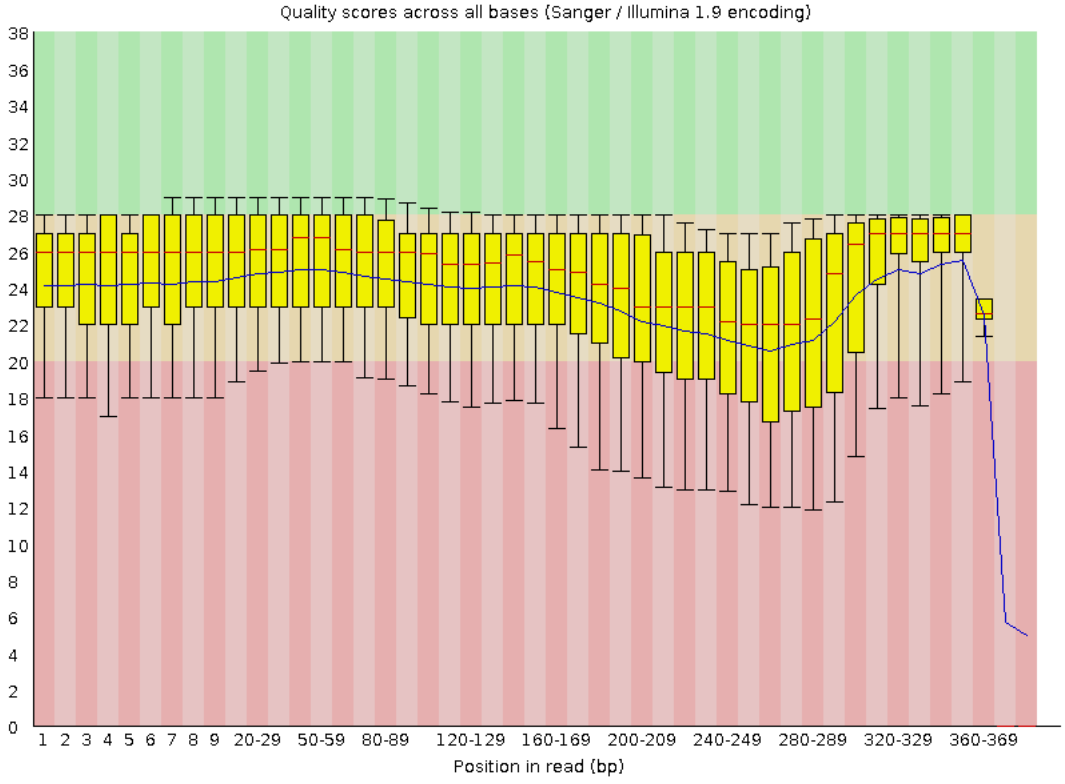
SRR17867019.fastq.gz



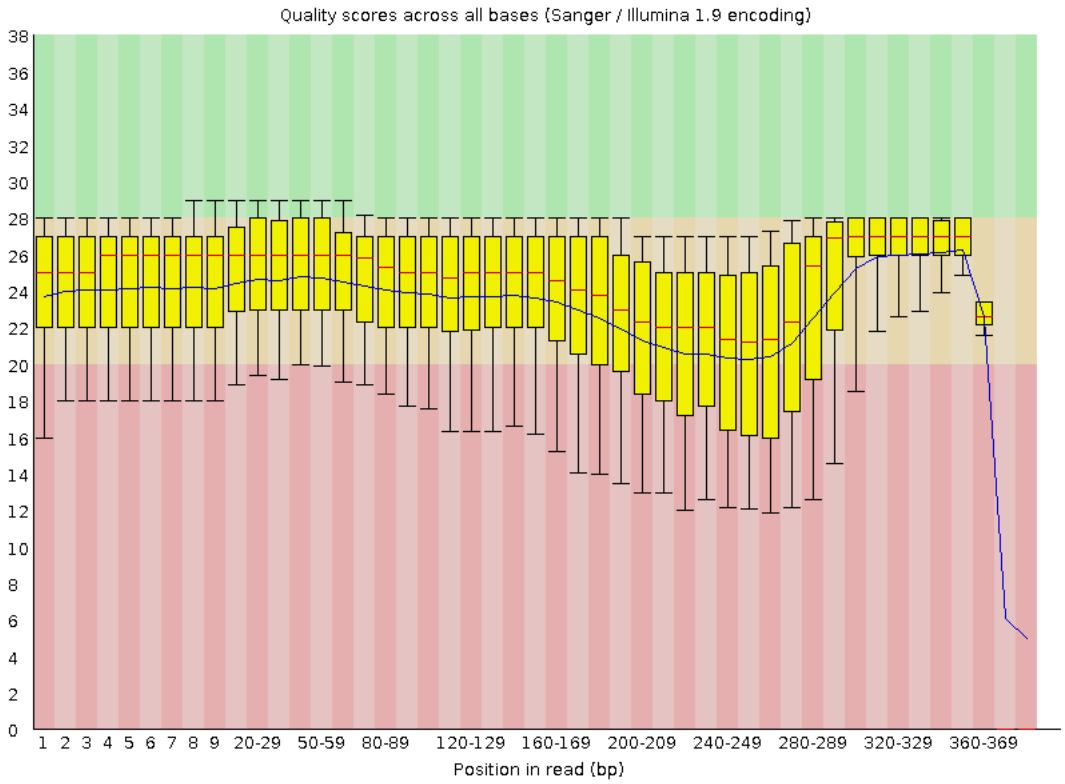
SRR17867020.fastq.gz



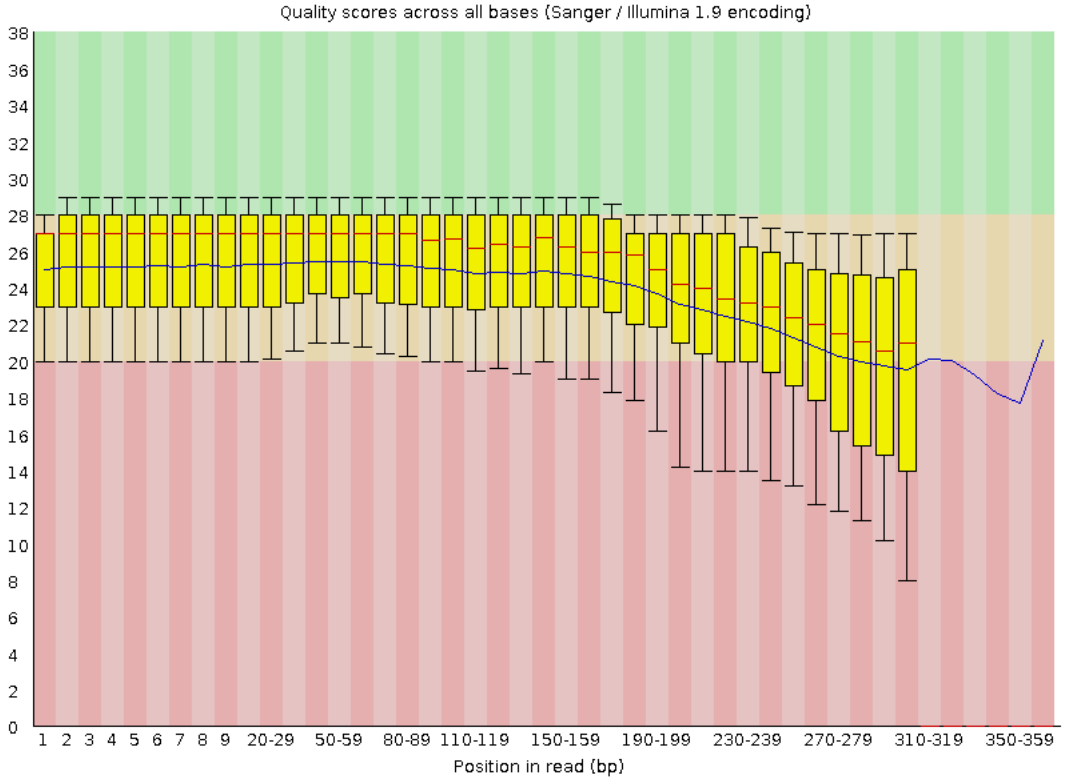
SRR17867021.fastq.gz



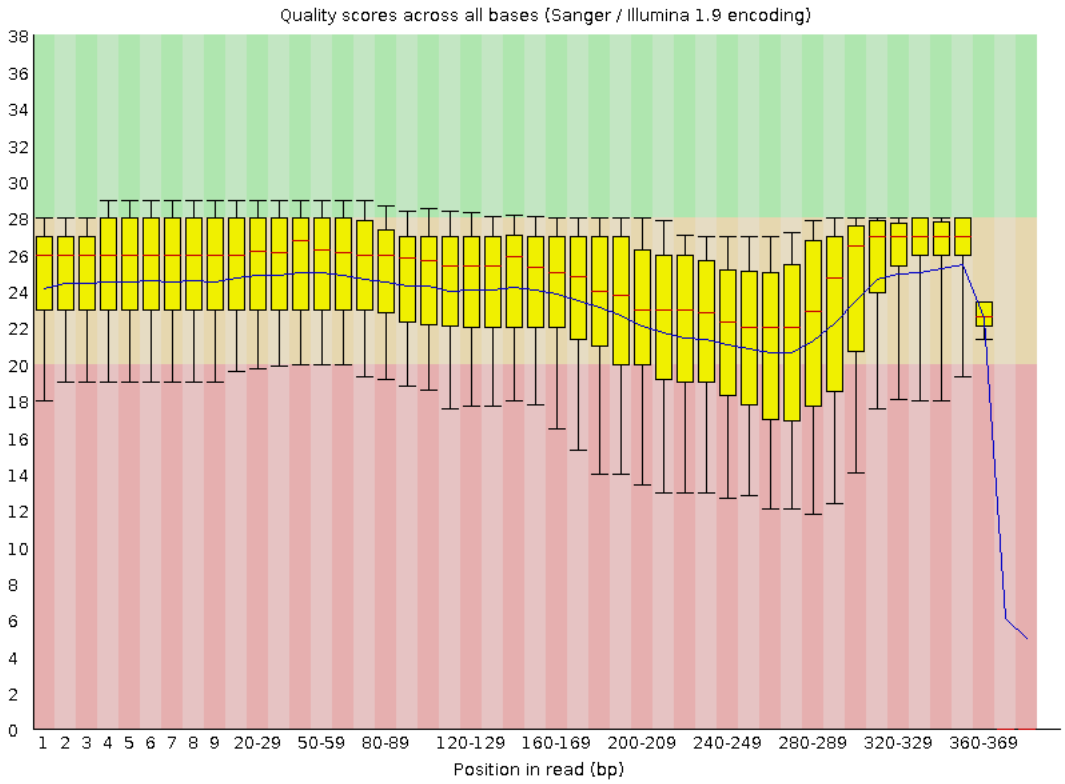
SRR17867022.fastq.gz



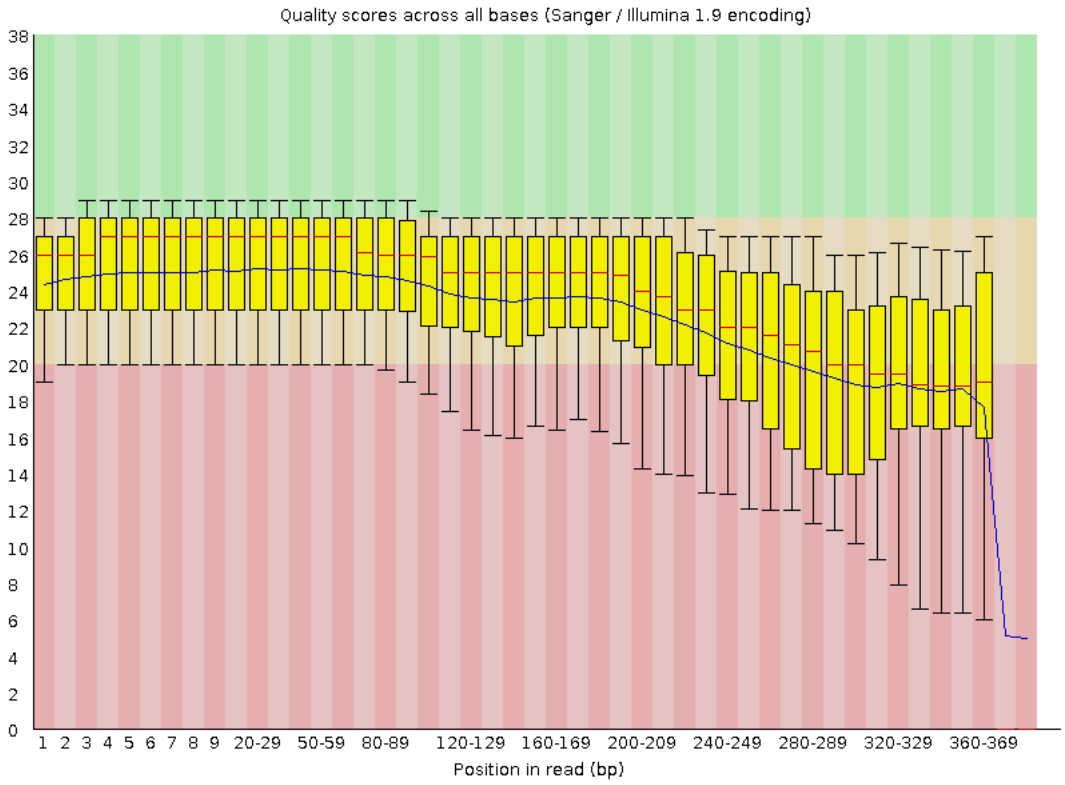
SRR17867023.fastq.gz



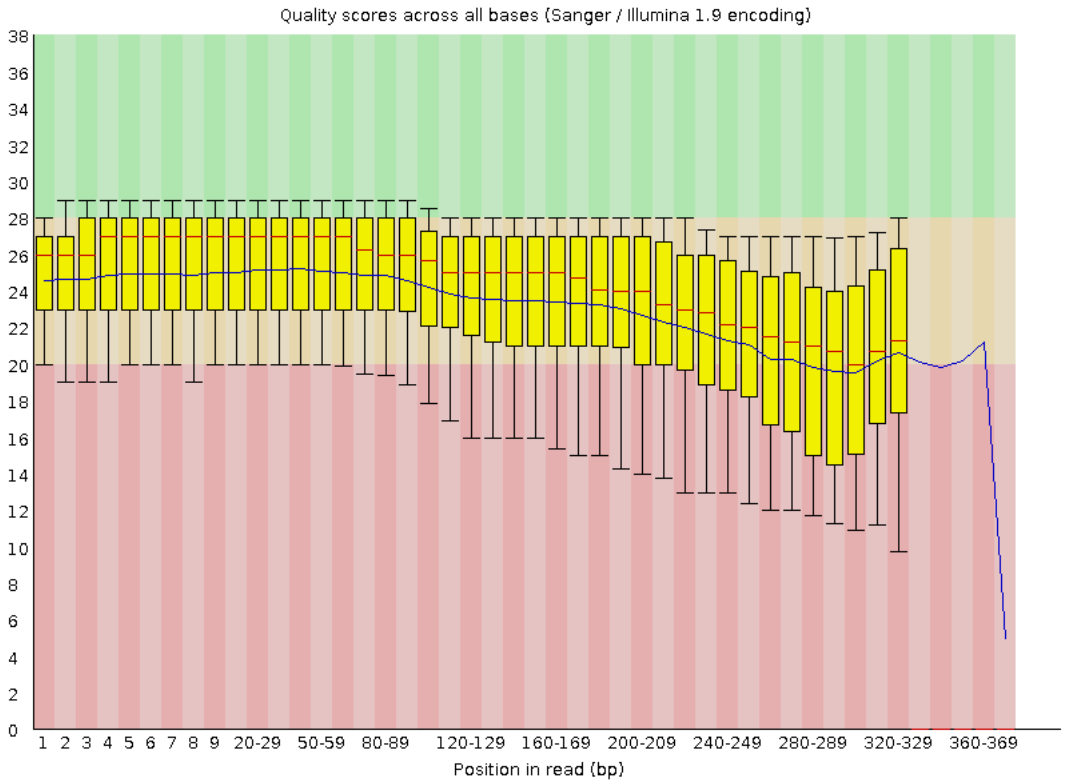
SRR17867024.fastq.gz



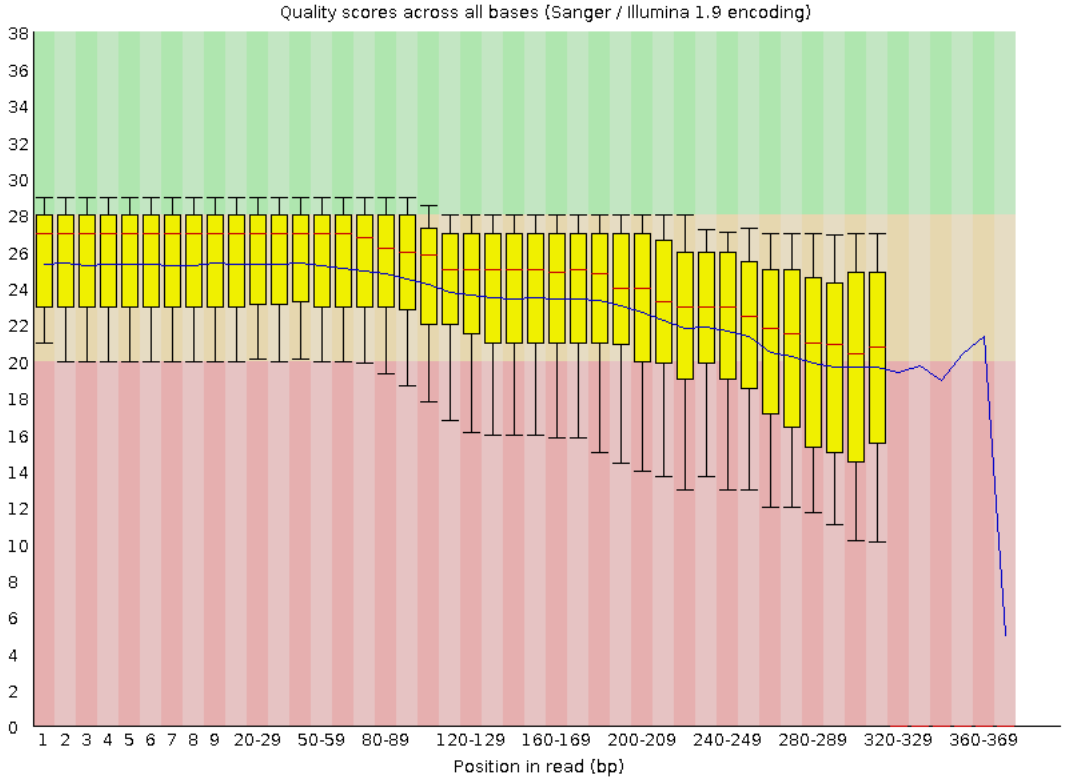
SRR17867025.fastq.gz



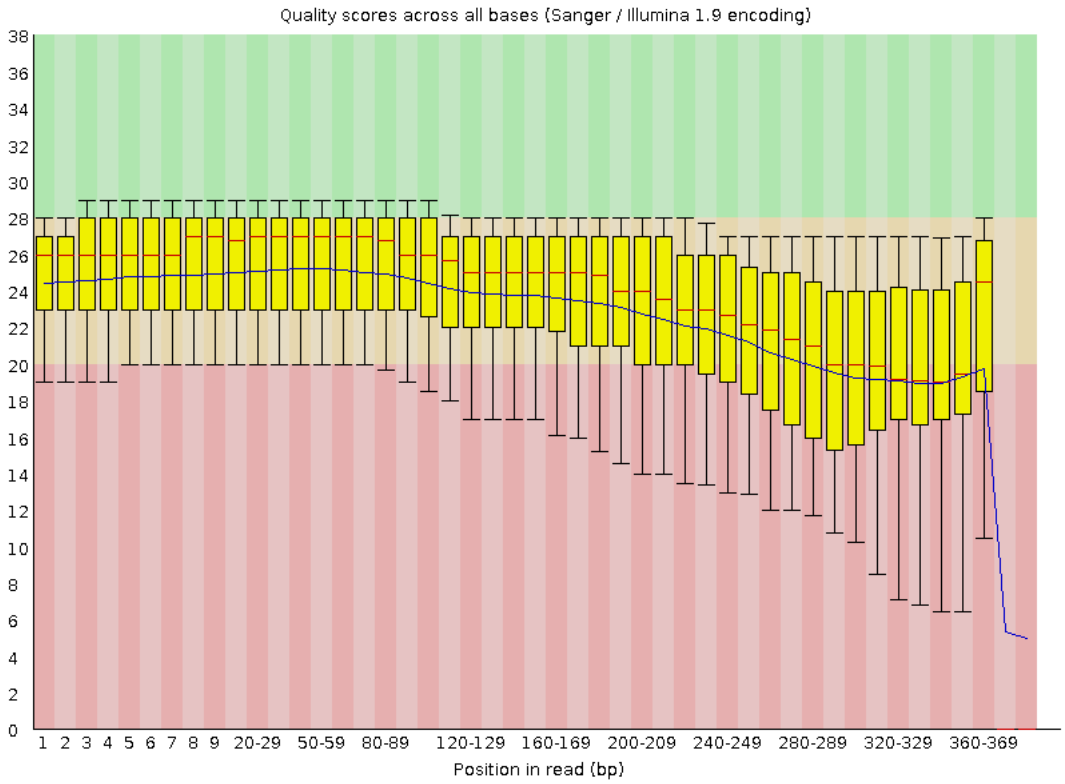
SRR17867026.fastq.gz



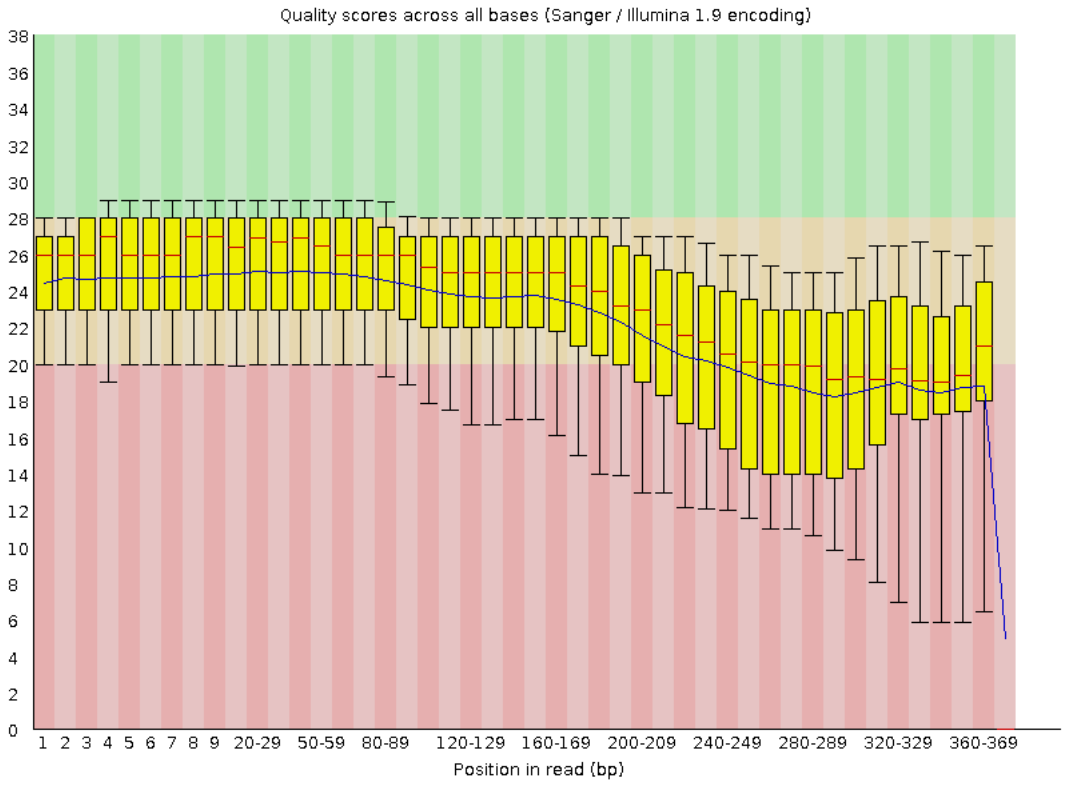
SRR17867027.fastq.gz



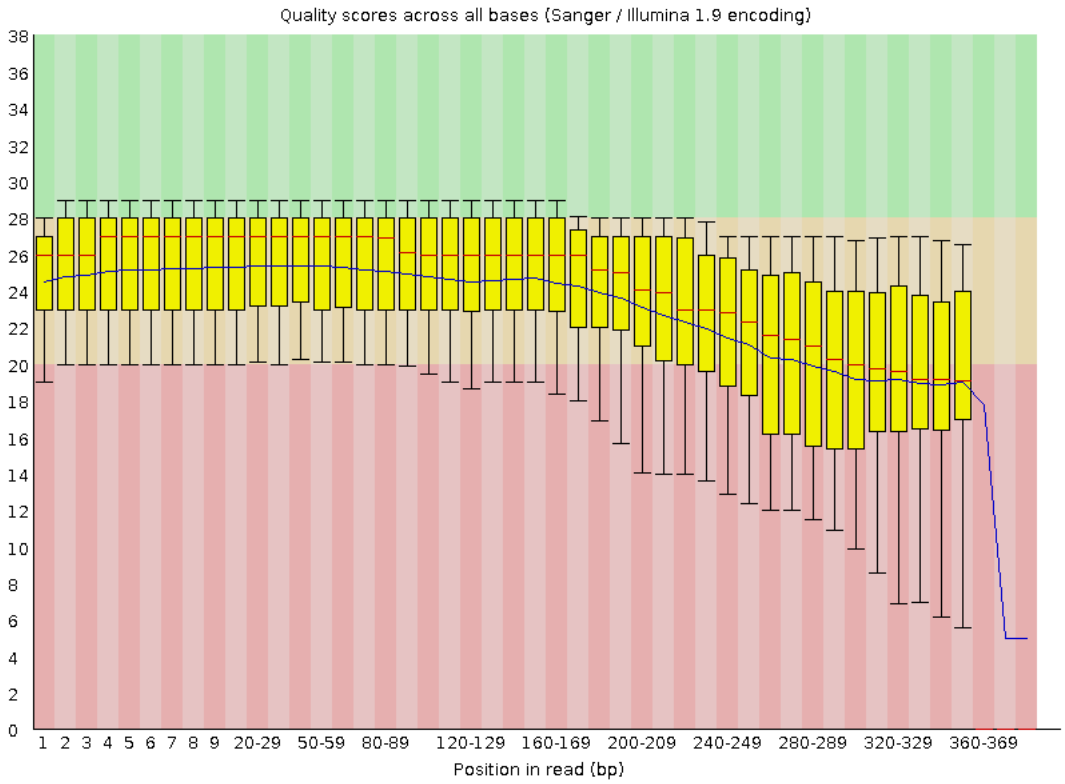
SRR17867028.fastq.gz



SRR17867030.fastq.gz



SRR17867031.fastq.gz



SRR17867032.fastq.gz

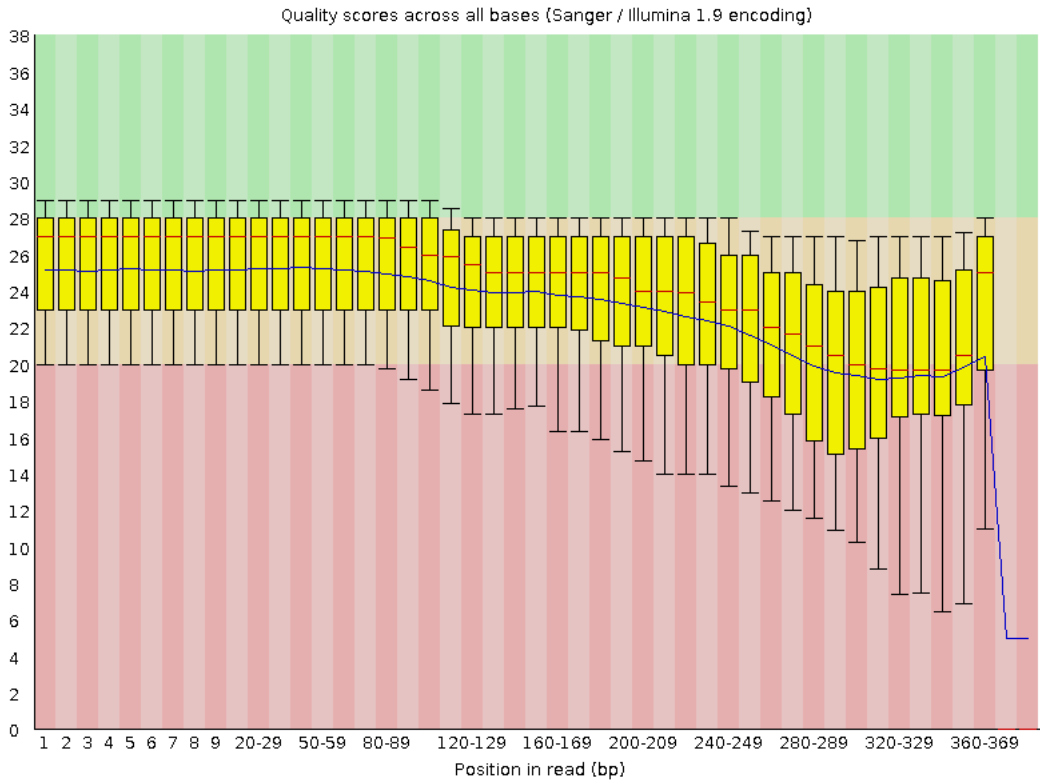
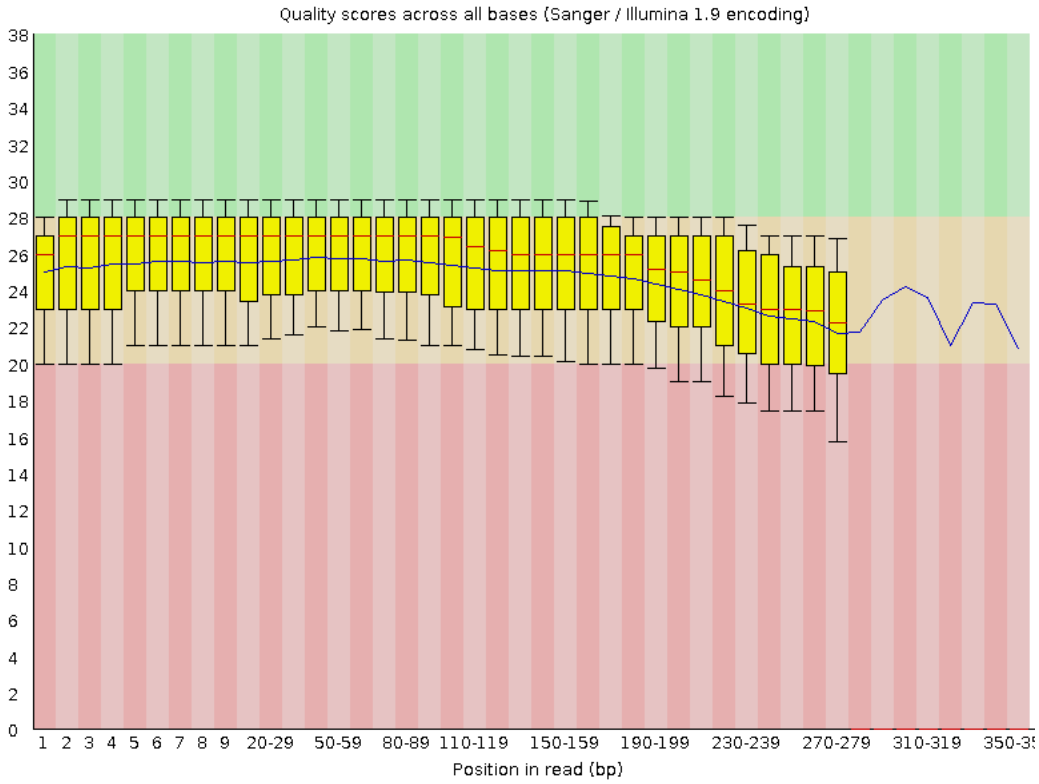


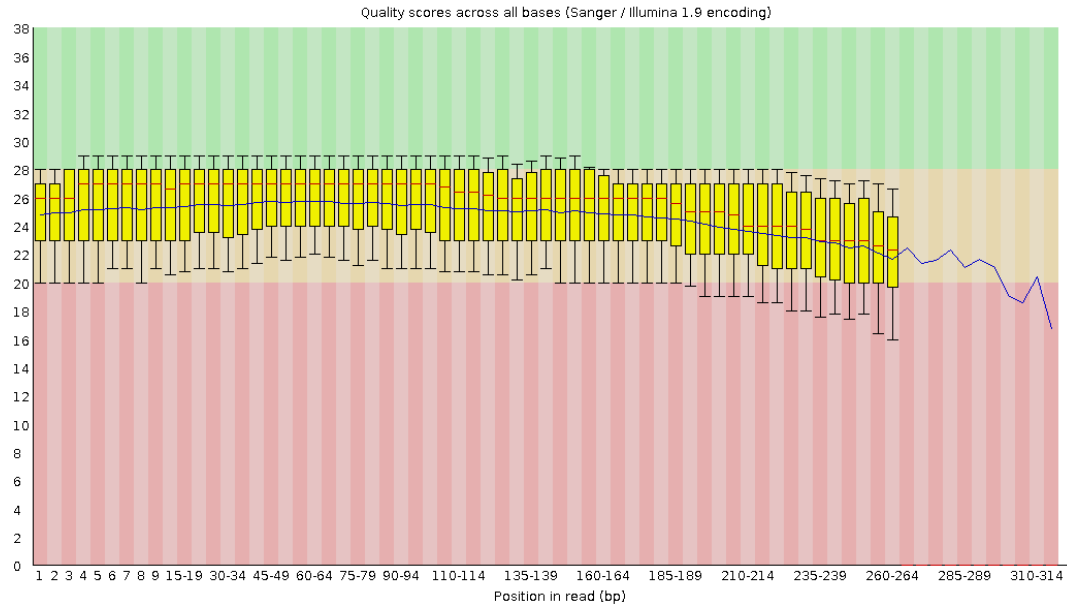
Figure 52 | Untrimmed FastQC files for dataset 1. The box and whisker plots show the quality scores across all bases per position in read (bp) and score value (y-axis). The graph has three colour coded regions, corresponding to the Y-axis scores. A good base call will have a high score value which is placed in the green region going towards the yellow region. A bad base call will have a lower value in the red region. It can be seen that each untrimmed Fastq file had a decrease in base call score towards the end of the cycle.. Graph produced by FastQC (command line and Python).

Trimmed Fastq files:

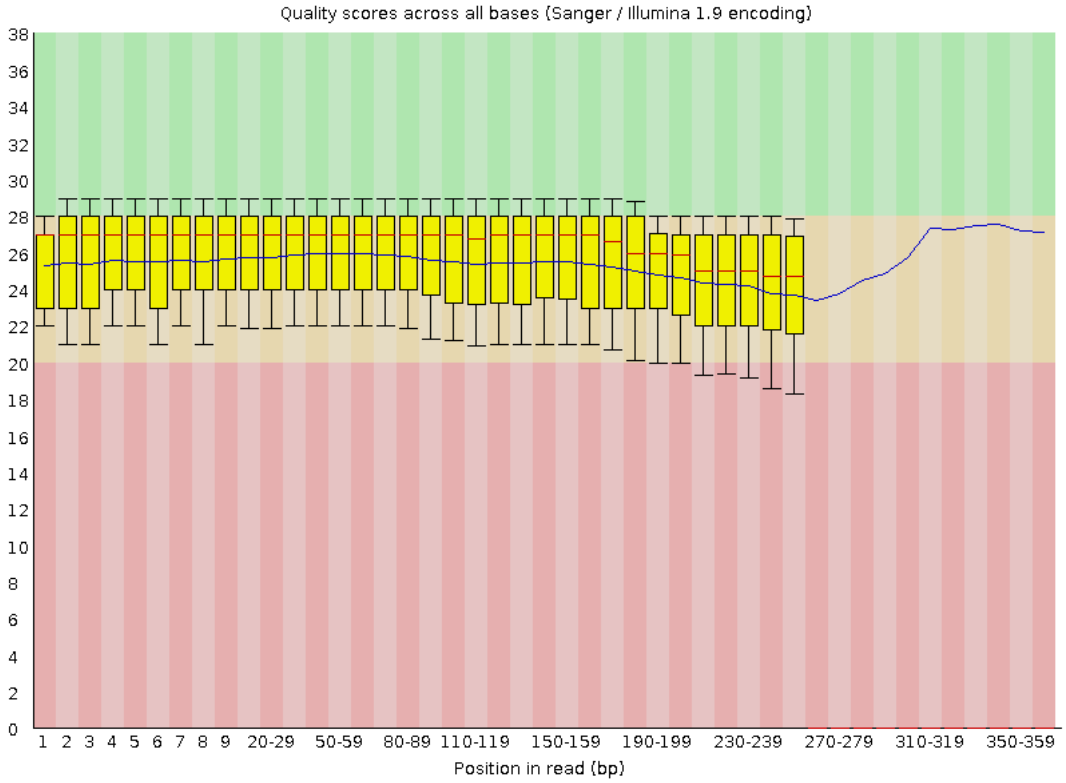
SRR17867010.fastq.gz



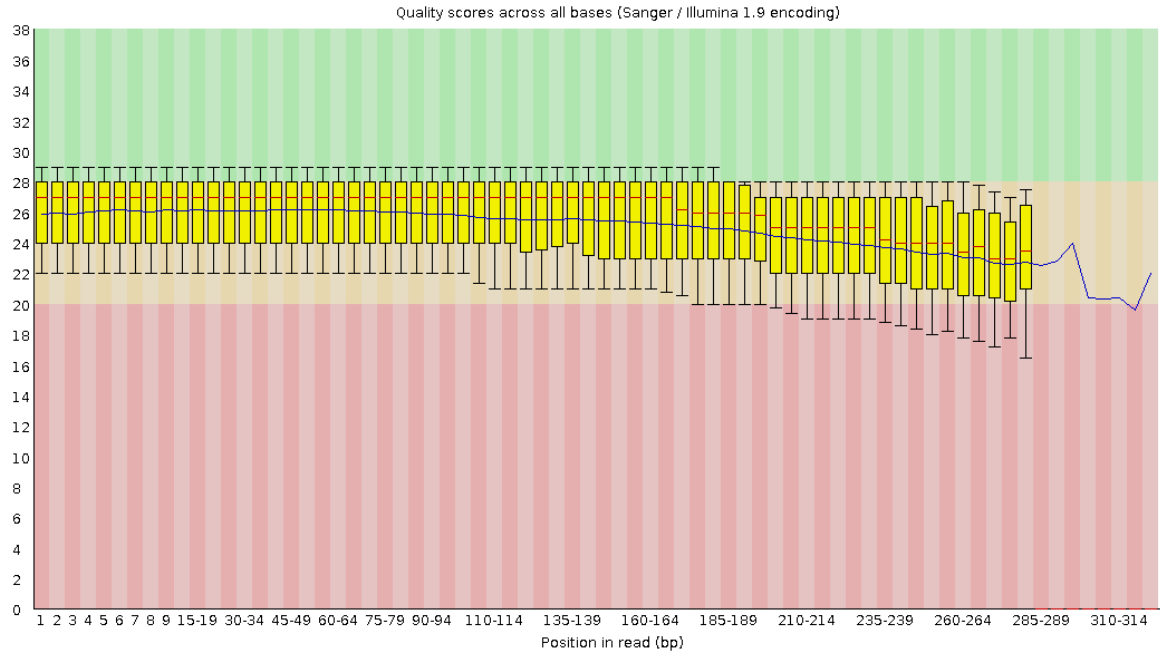
SRR17867011.fastq.gz



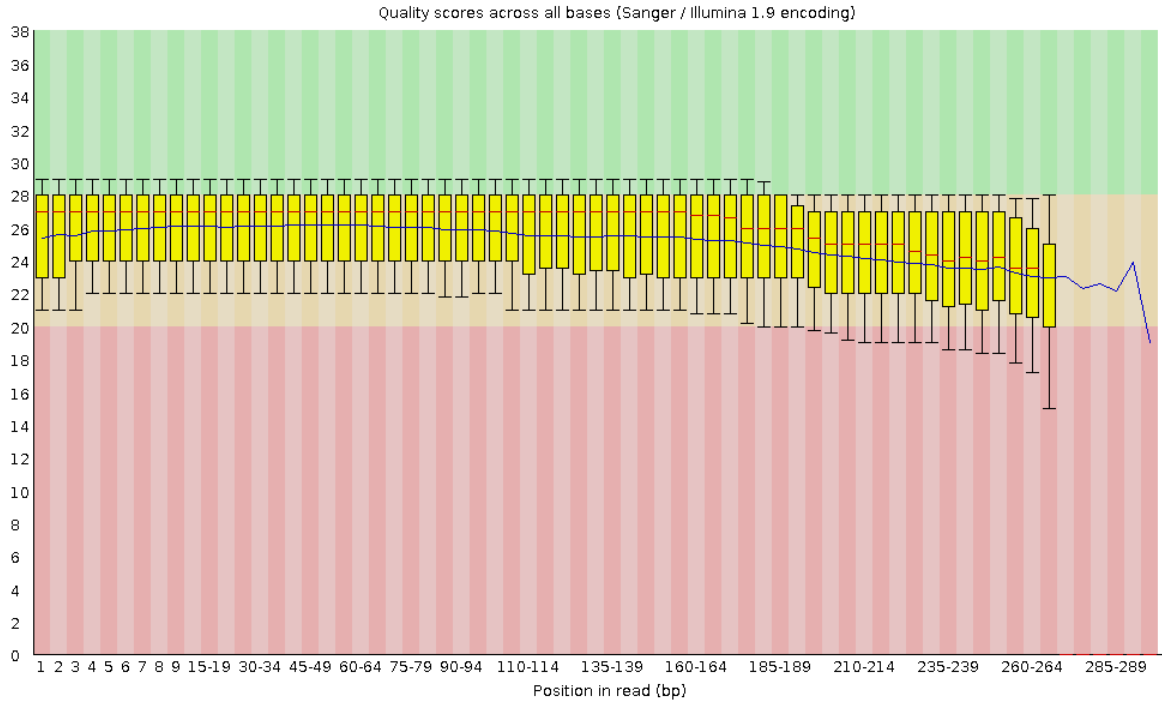
SRR17867012.fastq.gz



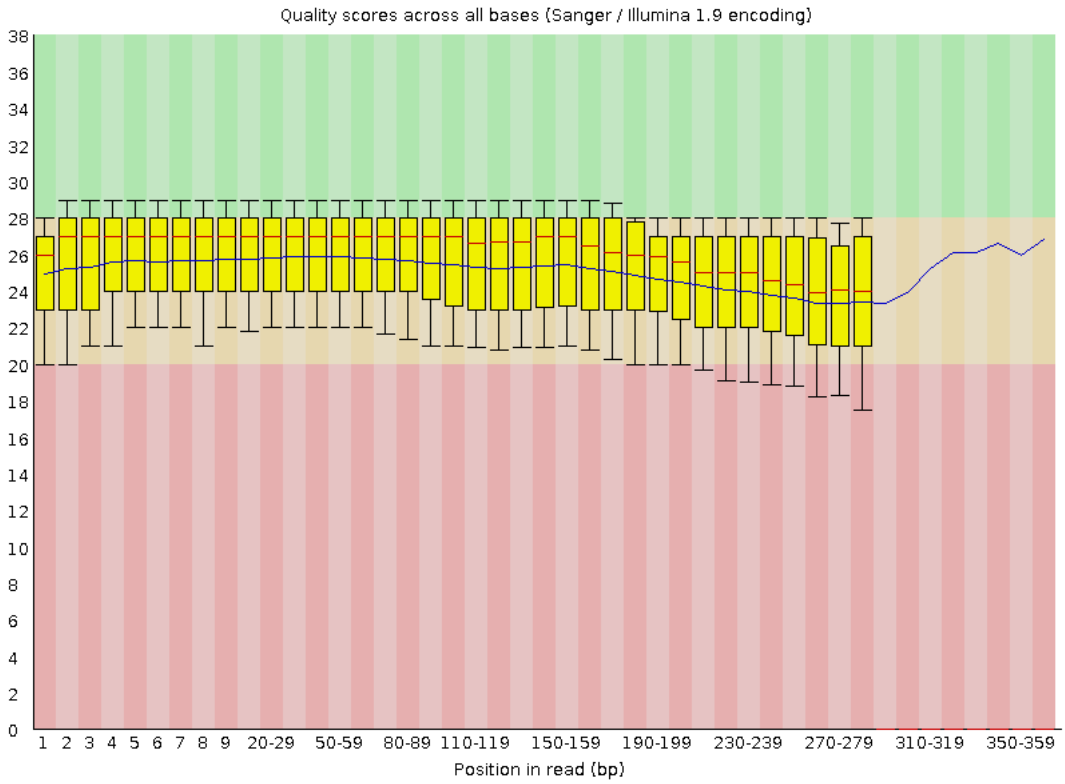
SRR17867013.fastq.gz



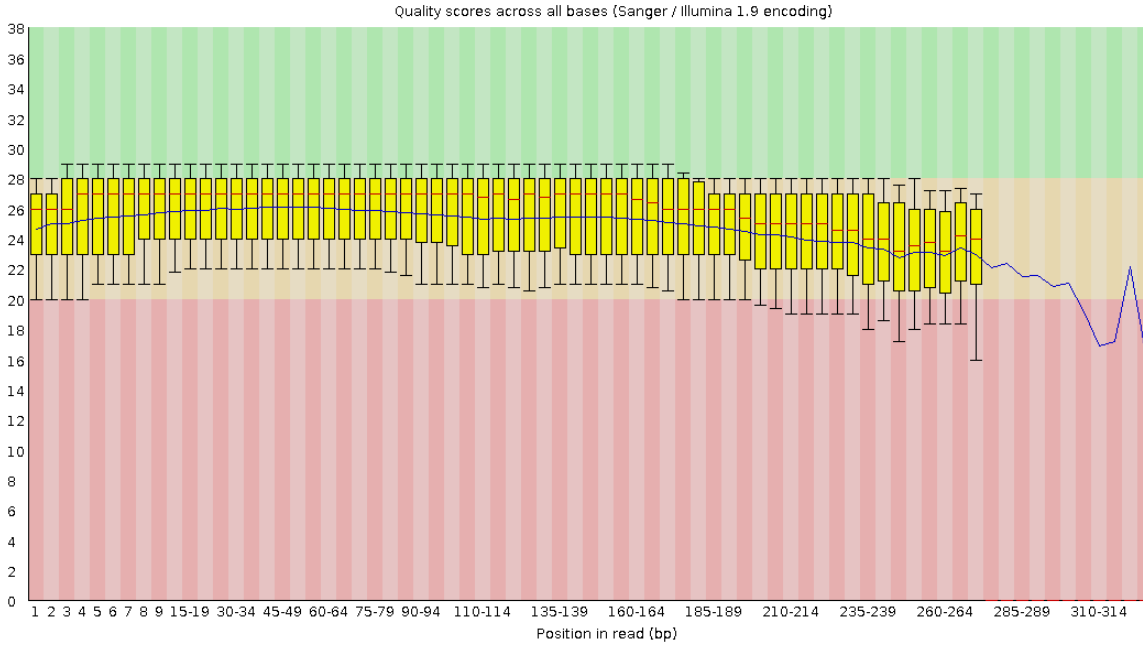
SRR17867014.fastq.gz



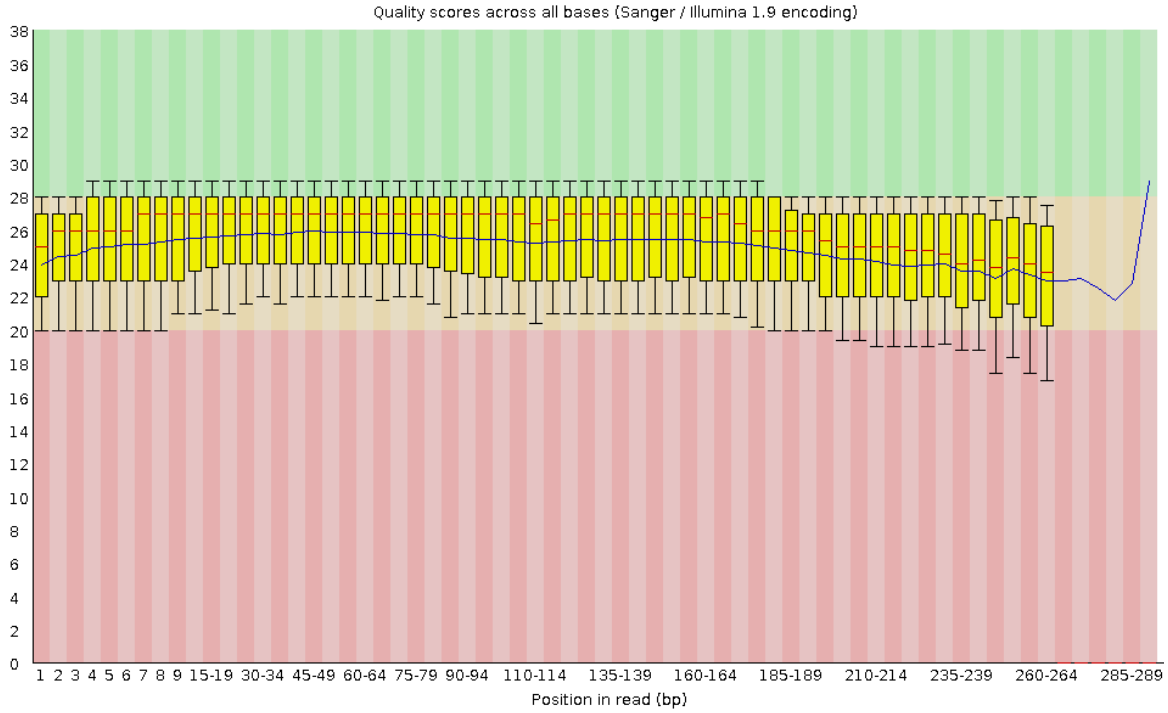
SRR17867015.fastq.gz



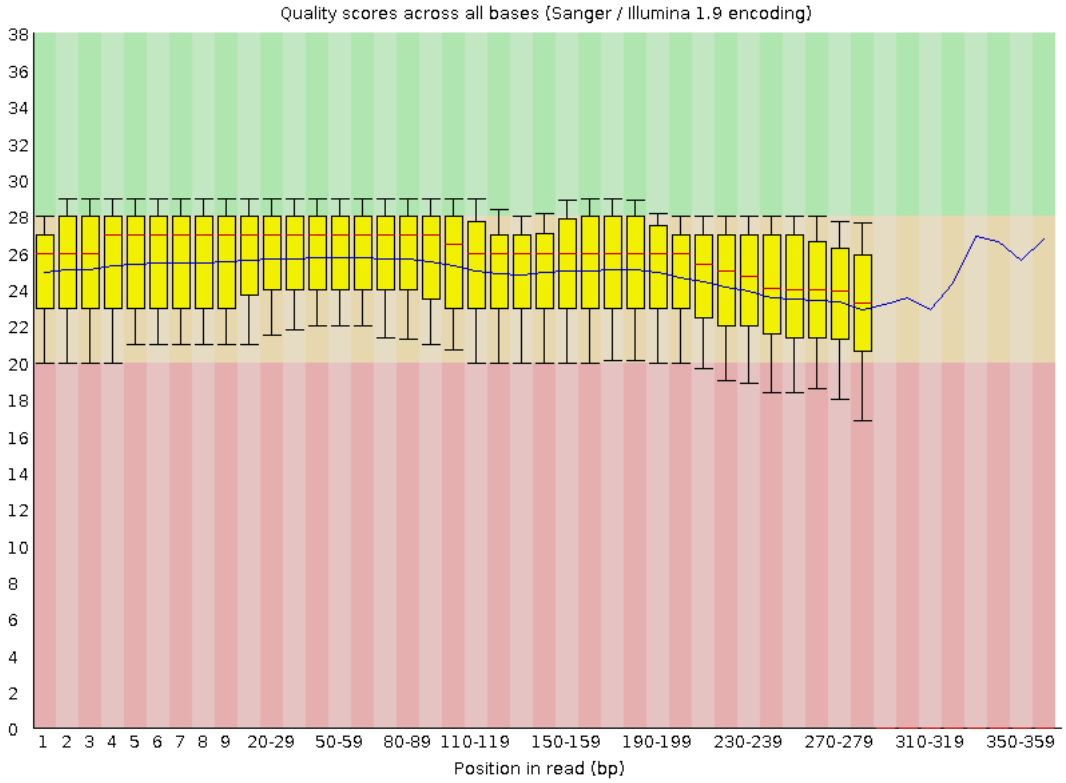
SRR17867016.fastq.gz



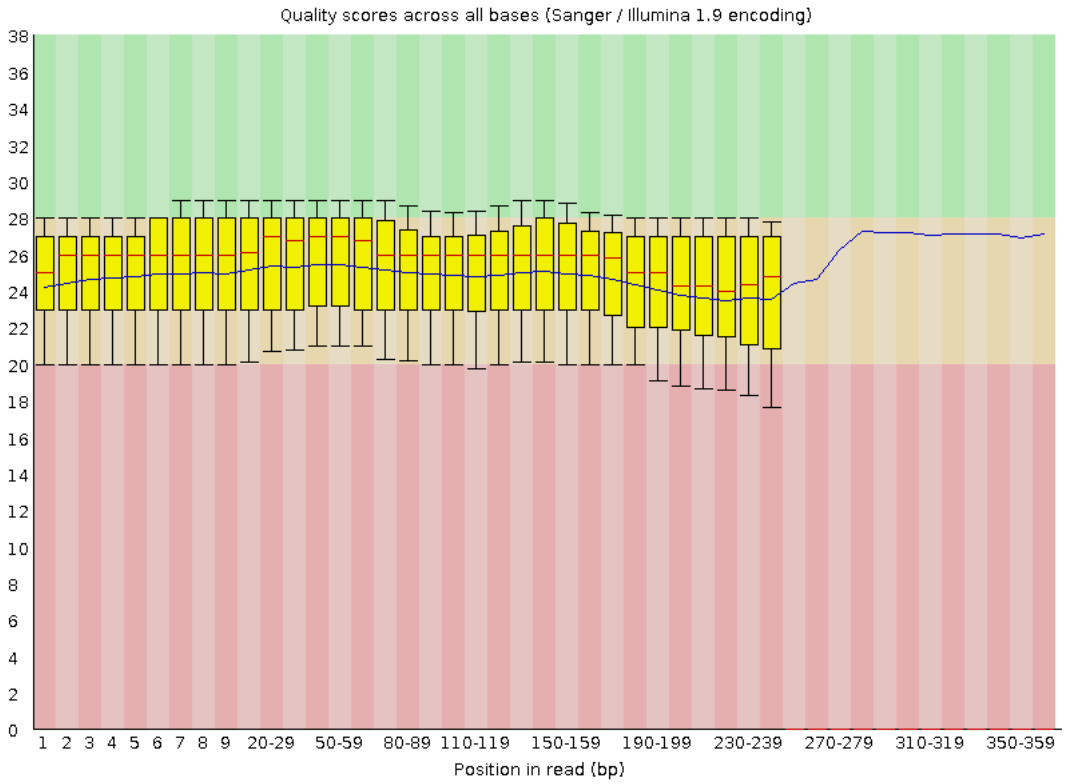
SRR17867017.fastq.gz



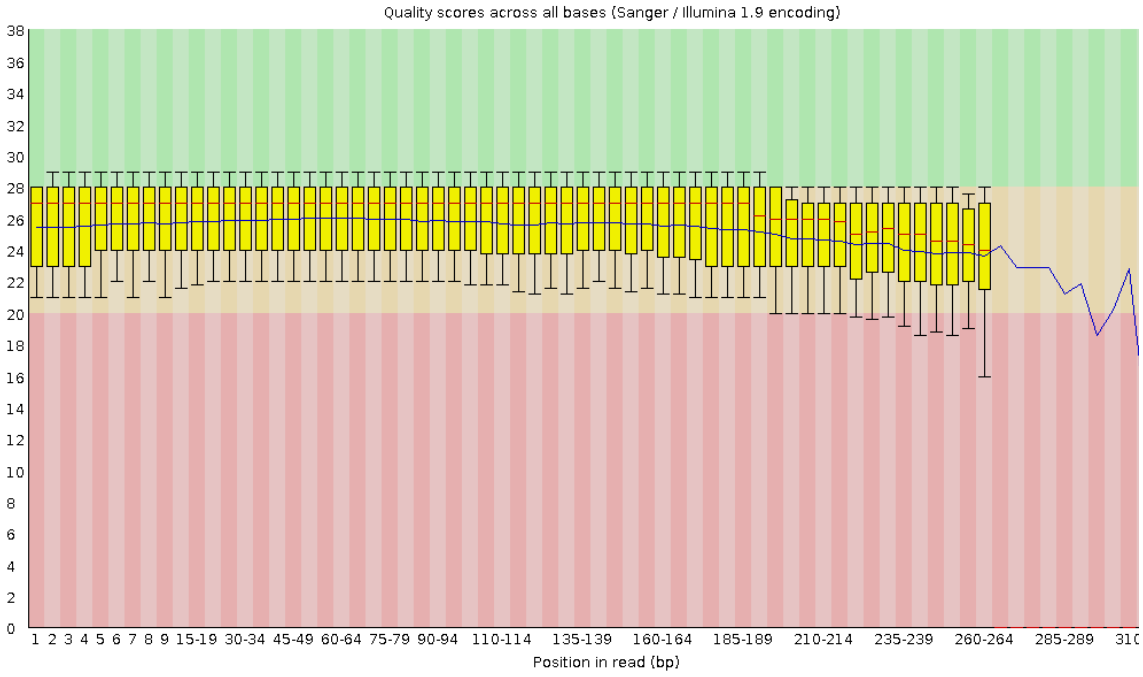
SRR17867018.fastq.gz



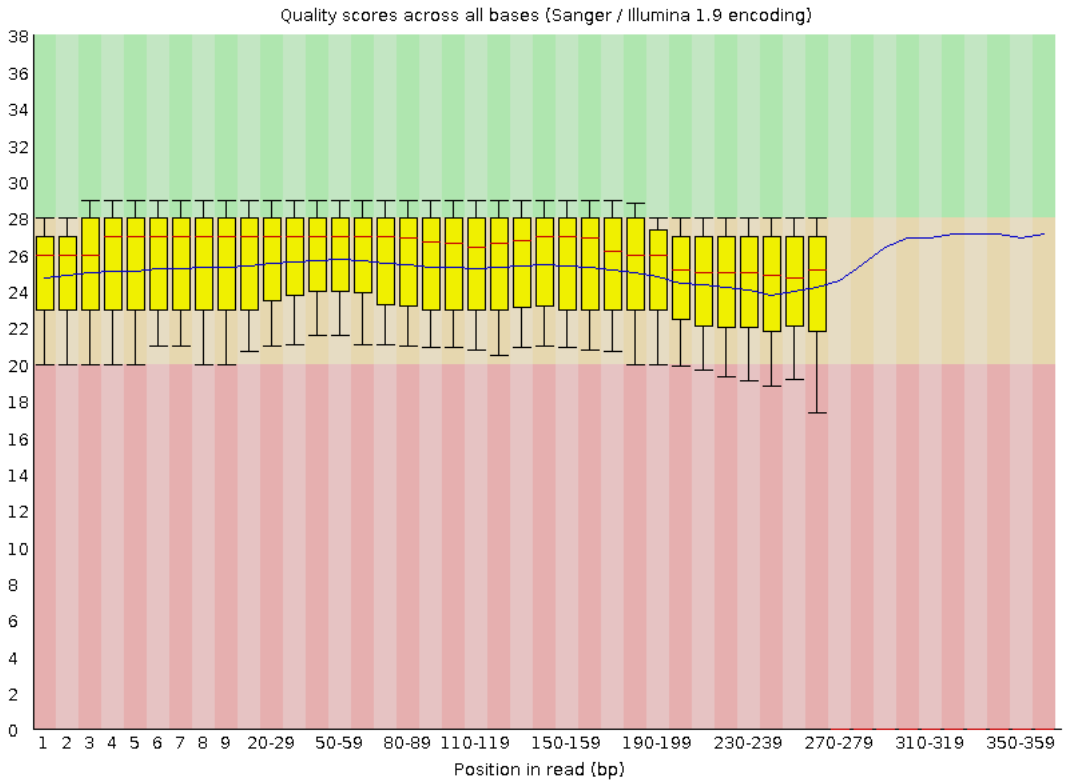
SRR17867019.fastq.gz



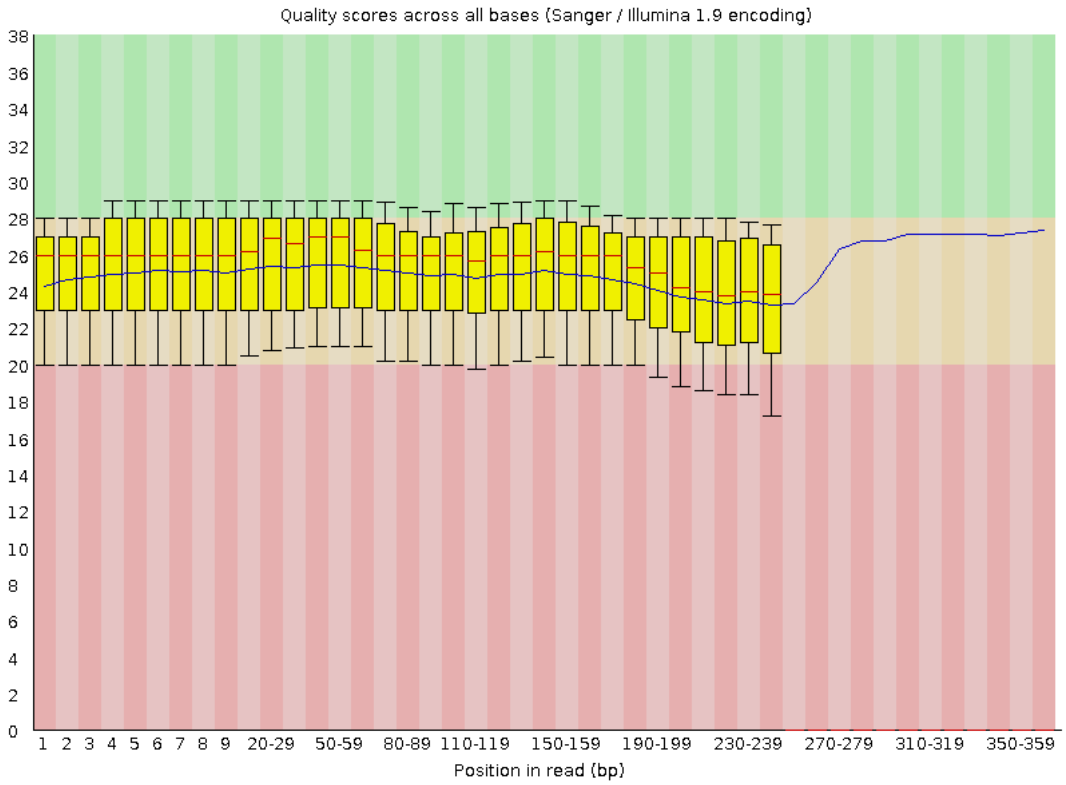
SRR17867020.fastq.gz



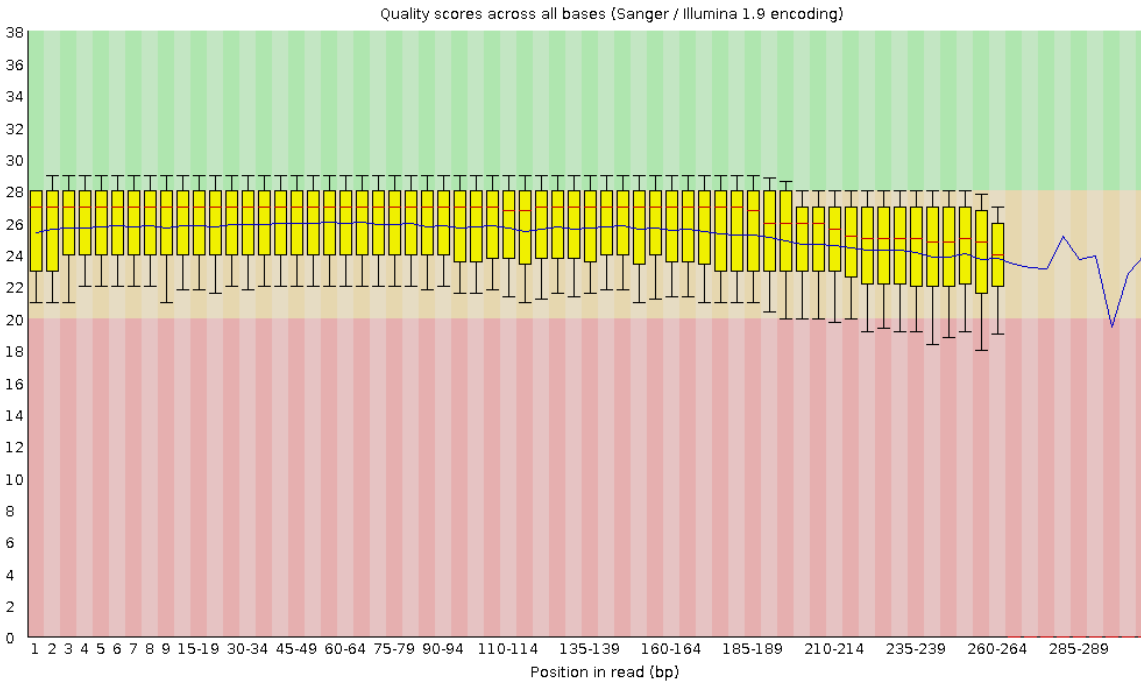
SRR17867021.fastq.gz



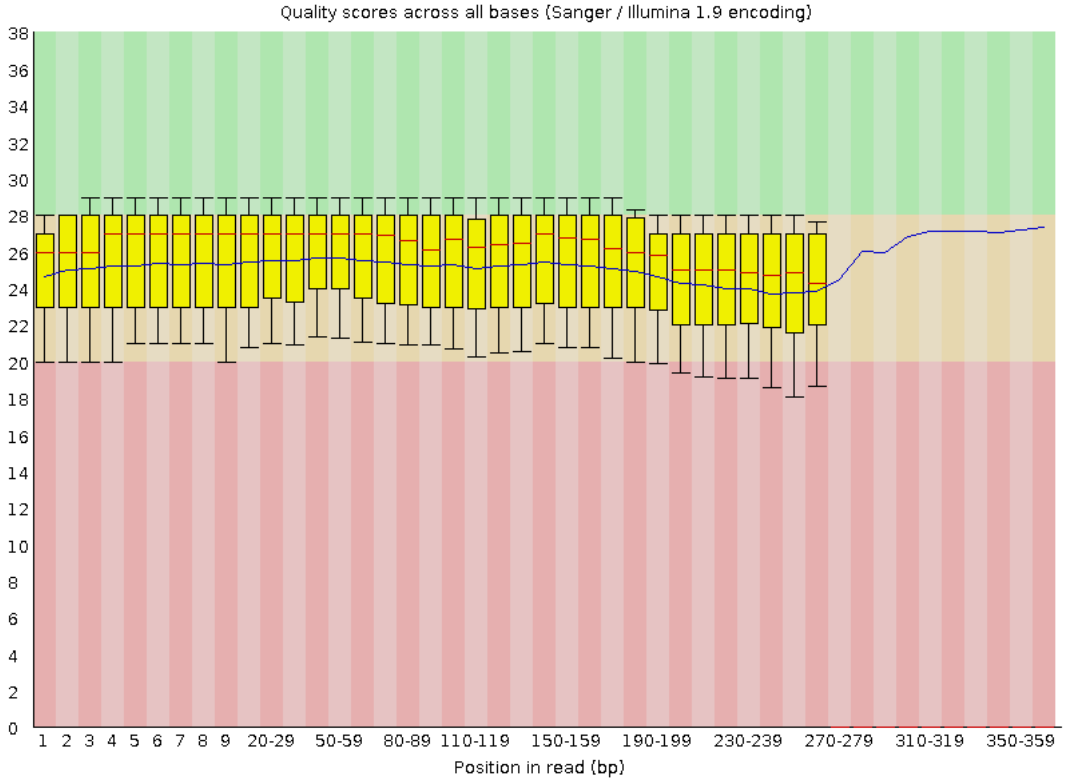
SRR17867022.fastq.gz



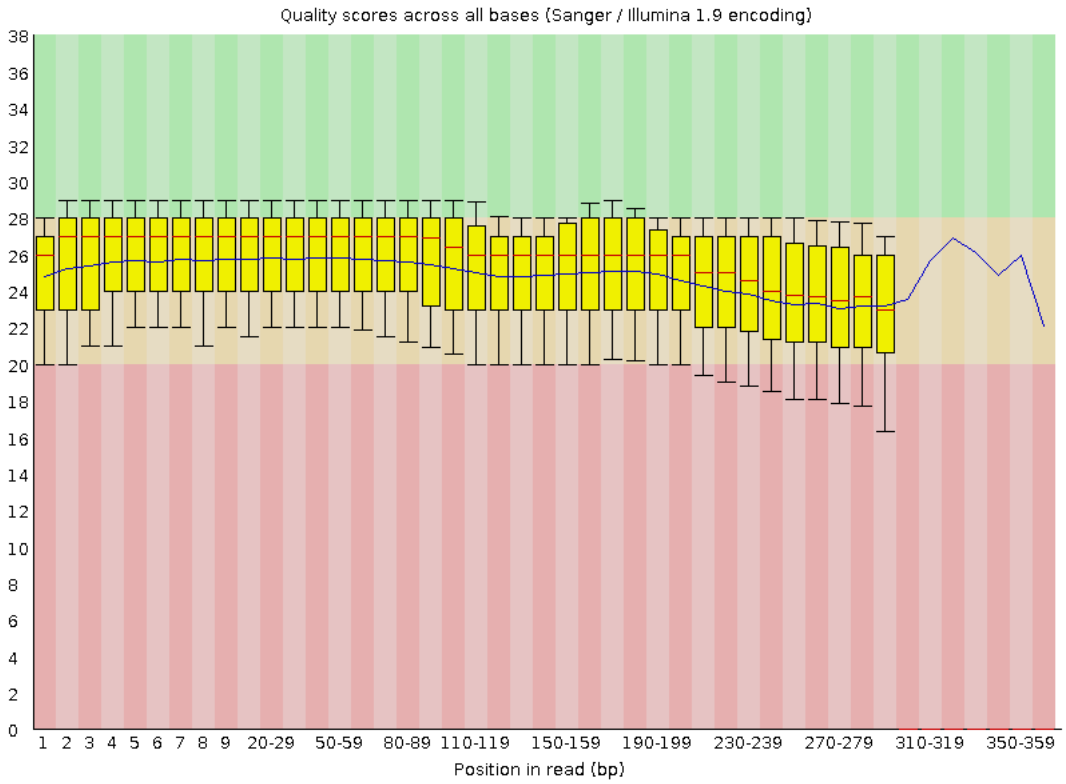
SRR17867023.fastq.g



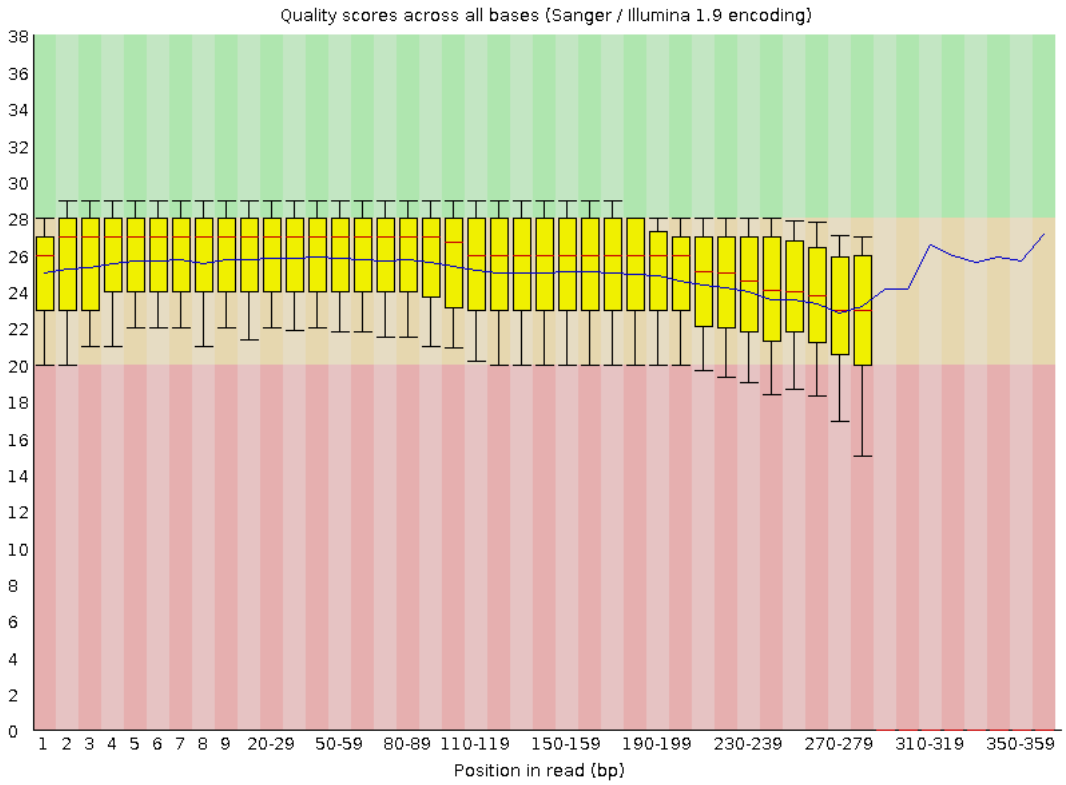
SRR17867024.fastq.gz



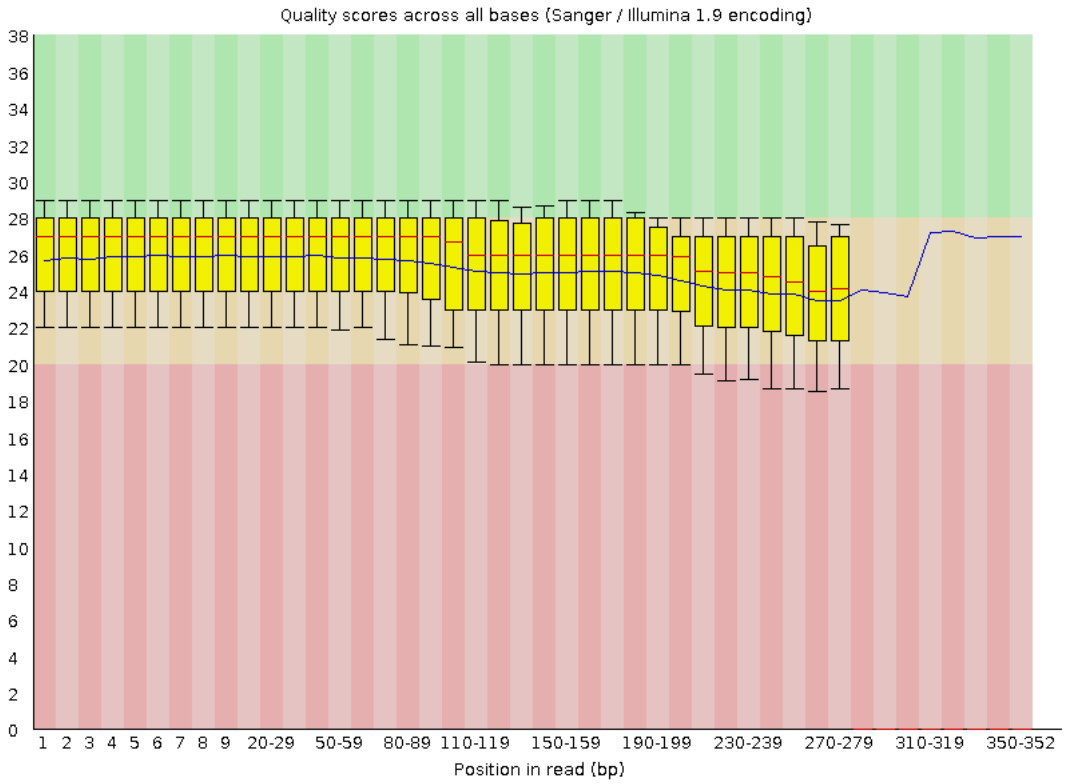
SRR17867025.fastq.gz



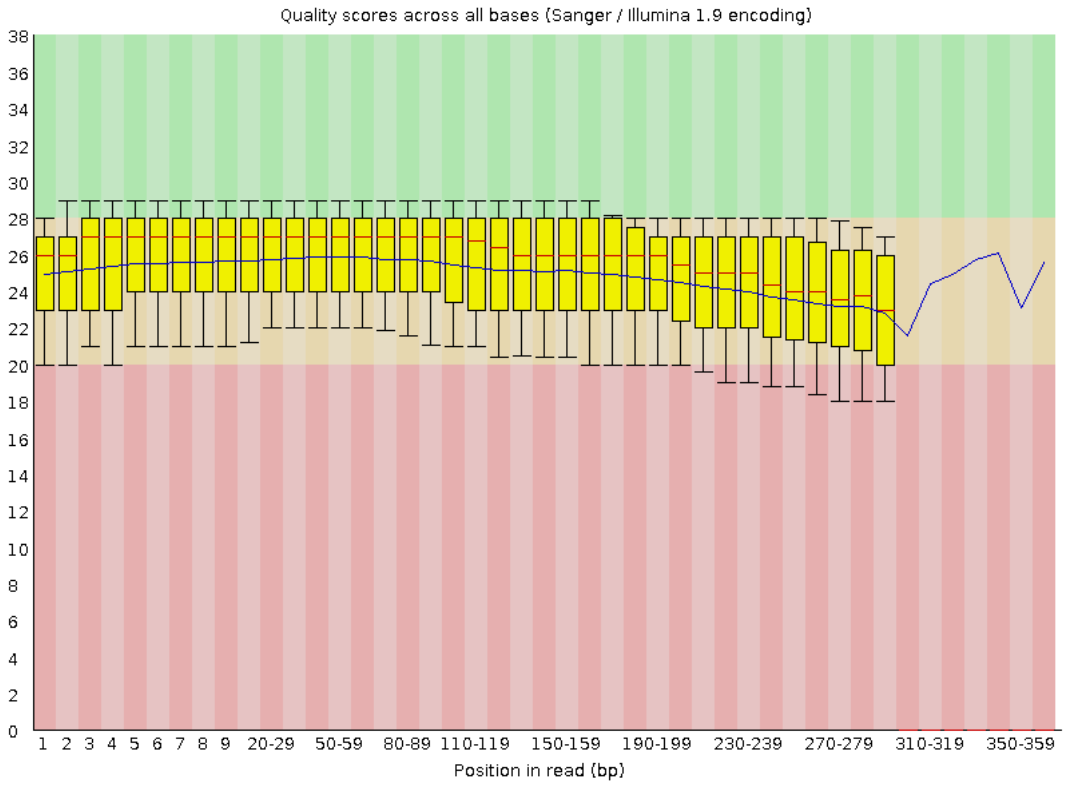
SRR17867026.fastq.gz



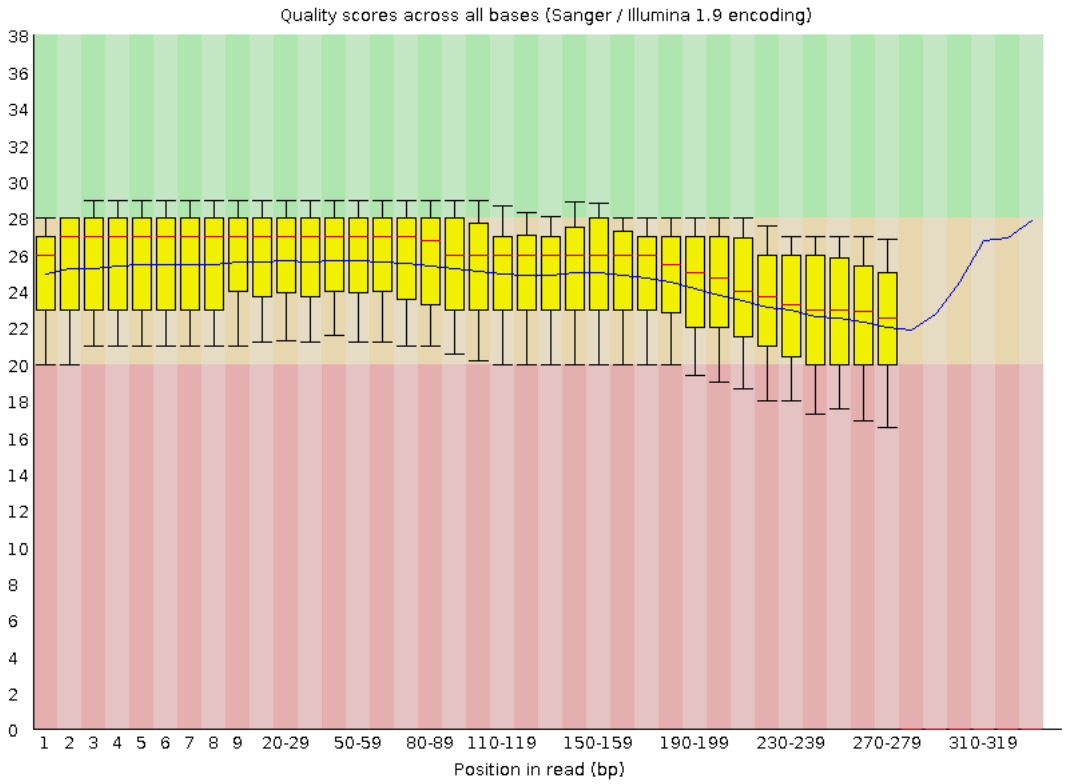
SRR17867027.fastq.gz



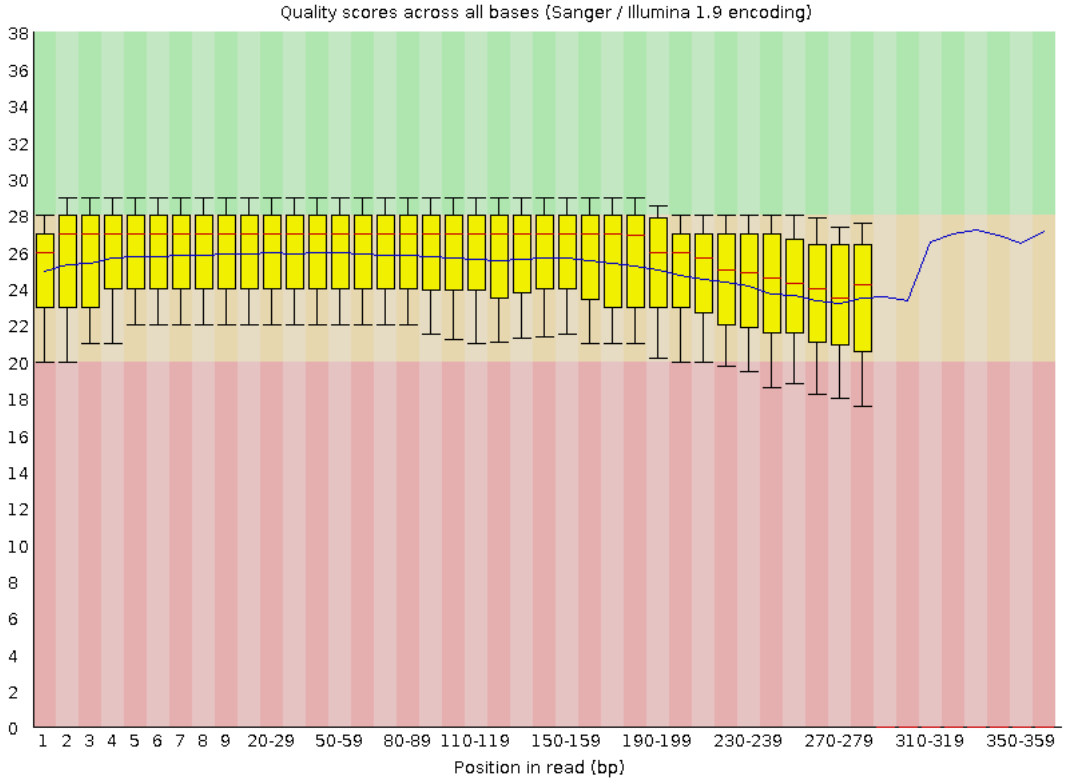
SRR17867028.fastq.gz



SRR17867030.fastq.gz



SRR17867031.fastq.gz



SRR17867032.fastq.gz

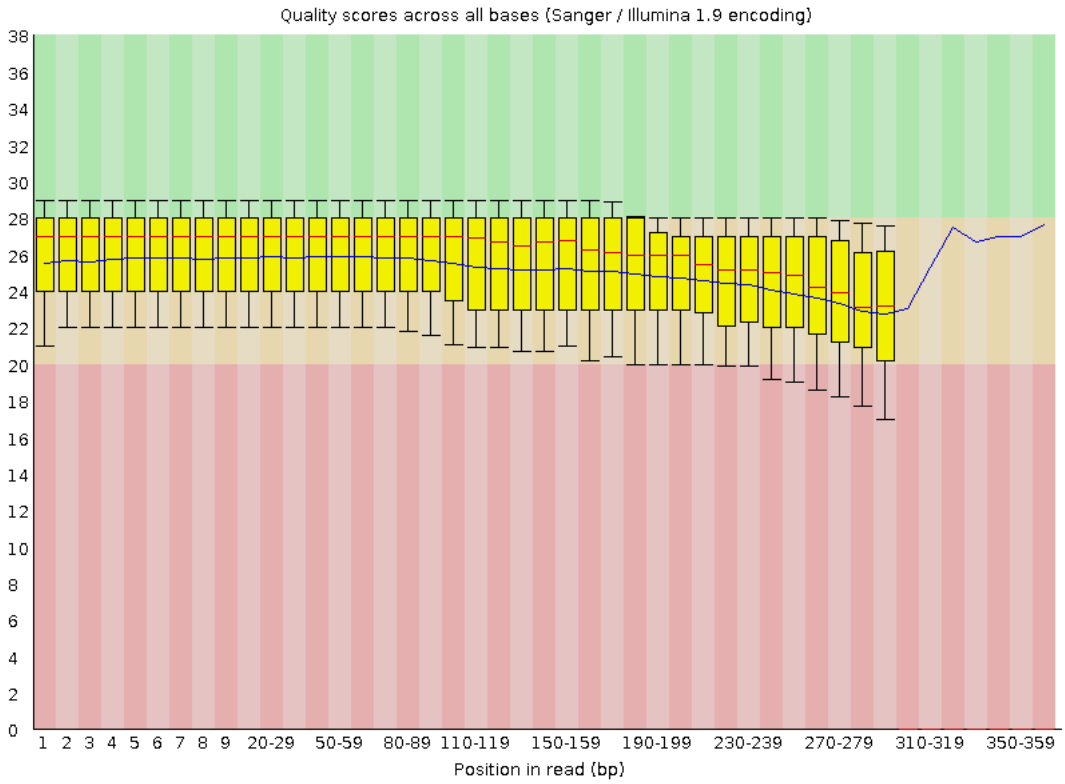
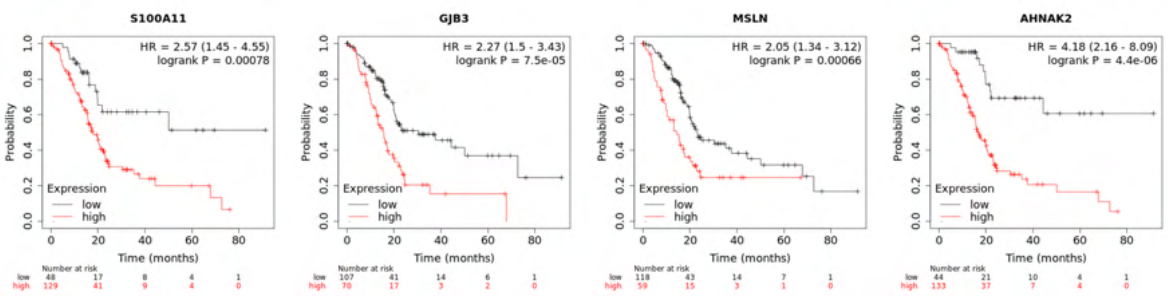


Figure 53 | FastQC files for trimmed samples used in dataset 1. The box and whisker plots show the quality scores across all bases per position in read (bp) and score value (y-axis). The graph has three colour coded regions, corresponding to the Y-axis scores. A good base call will have a high score value which is placed in the green region going towards the yellow region. A bad base call will have a lower value in the red region. All trimmed sample files were used for downstream analysis as each sample produced a satisfactory quality control plot. Graph produced by FastQC (command line and Python).

10.2. Appendix B: Individual survival curves

10.2.1. Normal versus tumour samples

Common up-regulated tumour survival curves



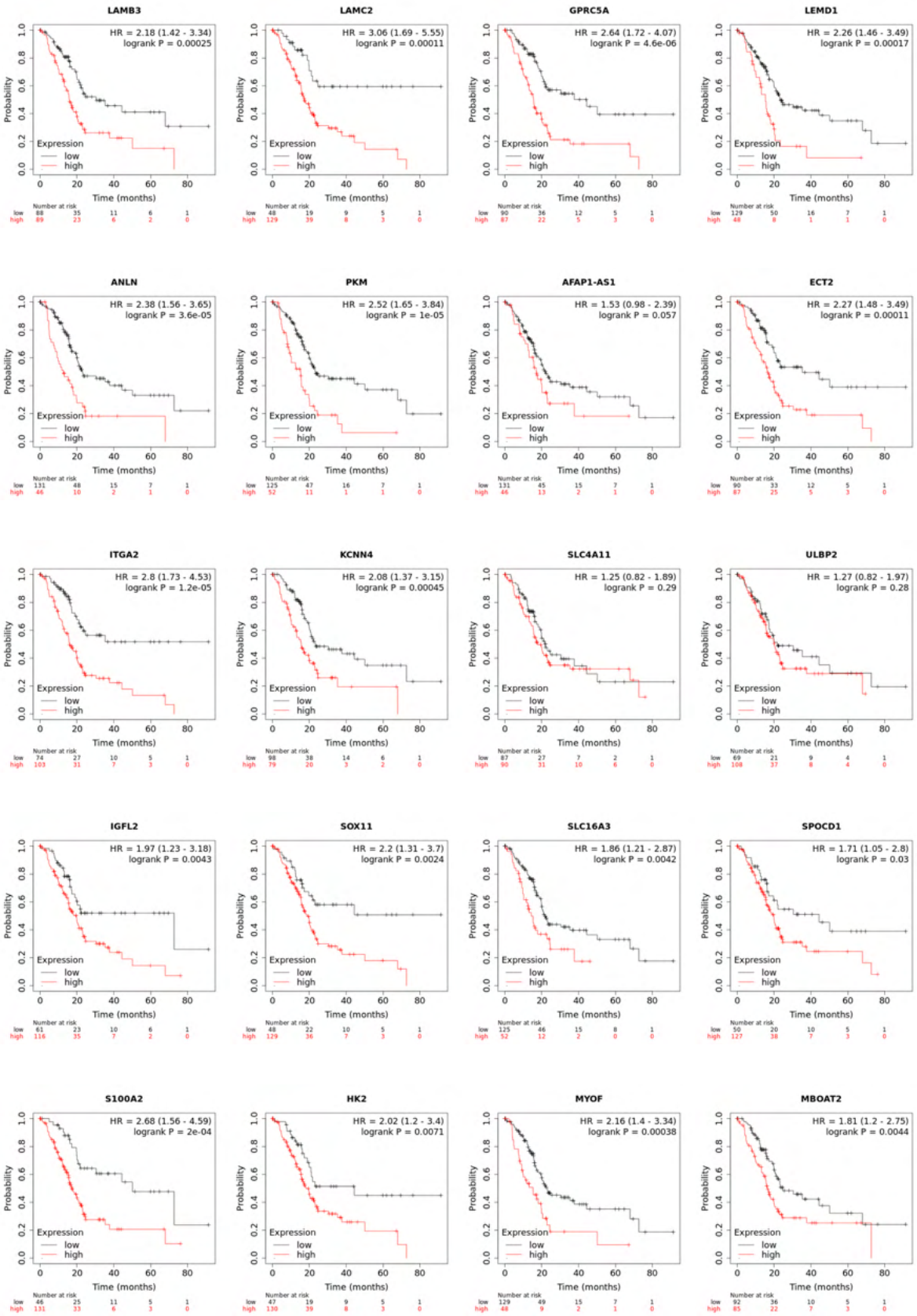
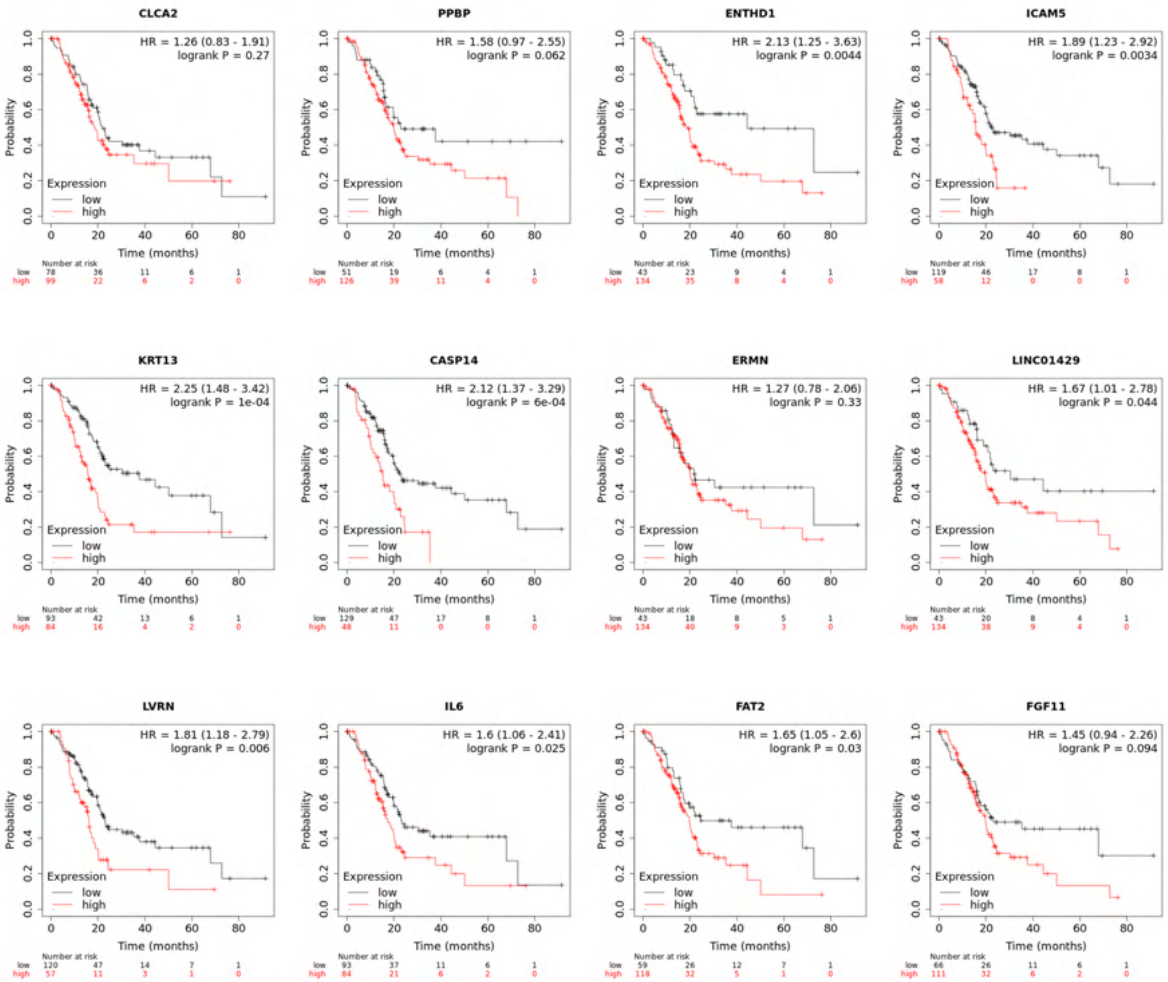

LAMB3
HR = 2.18 (1.42 - 3.34)
logrank P = 0.00025
LAMC2
HR = 3.06 (1.69 - 5.55)
logrank P = 0.00011
GPRC5A
HR = 2.64 (1.72 - 4.07)
logrank P = 4.6e-06
LEMD1
HR = 2.26 (1.46 - 3.49)
logrank P = 0.00017
ANLN
HR = 2.38 (1.56 - 3.65)
logrank P = 3.6e-05
PKM
HR = 2.52 (1.65 - 3.84)
logrank P = 1e-05
AFAP1-AS1
HR = 1.53 (0.98 - 2.39)
logrank P = 0.057
ECT2
HR = 2.27 (1.48 - 3.49)
logrank P = 0.00011
ITGA2
HR = 2.8 (1.73 - 4.53)
logrank P = 1.2e-05
KCNN4
HR = 2.08 (1.37 - 3.15)
logrank P = 0.00045
SLC4A11
HR = 1.25 (0.82 - 1.89)
logrank P = 0.29
ULBP2
HR = 1.27 (0.82 - 1.97)
logrank P = 0.28
IGFL2
HR = 1.97 (1.23 - 3.18)
logrank P = 0.0043
SOX11
HR = 2.2 (1.31 - 3.7)
logrank P = 0.0024
SLC16A3
HR = 1.86 (1.21 - 2.87)
logrank P = 0.0042
SPOCD1
HR = 1.71 (1.05 - 2.8)
logrank P = 0.03
S100A2
HR = 2.68 (1.56 - 4.59)
logrank P = 2e-04
HK2
HR = 2.02 (1.2 - 3.4)
logrank P = 0.0071
MYOF
HR = 2.16 (1.4 - 3.34)
logrank P = 0.00038
MBOAT2
HR = 1.81 (1.2 - 2.75)
logrank P = 0.0044
Probability
Time (months)
Expression
low
high
Number at risk

Figure 54 | Kaplan-Meier survival curves for each common upregulated tumour DEGs. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEGs with insignificant P values were discarded: SLC411 and ULBP2. KM plotter did not produce a plot for: BLACAT1. Graph produced by Kaplan-Meier plotter version 2022.08.23.

10.2.2. Early versus advanced stage



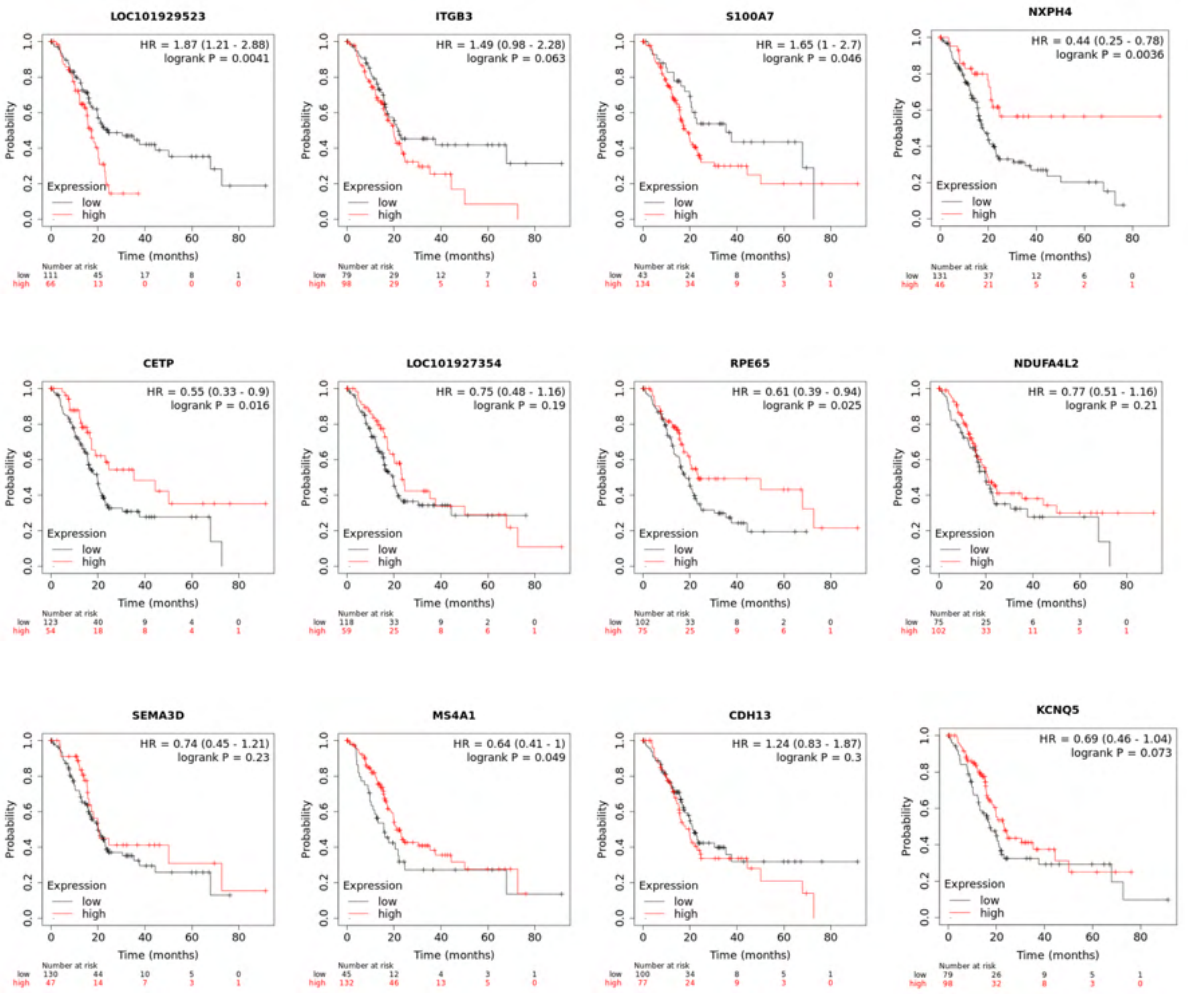
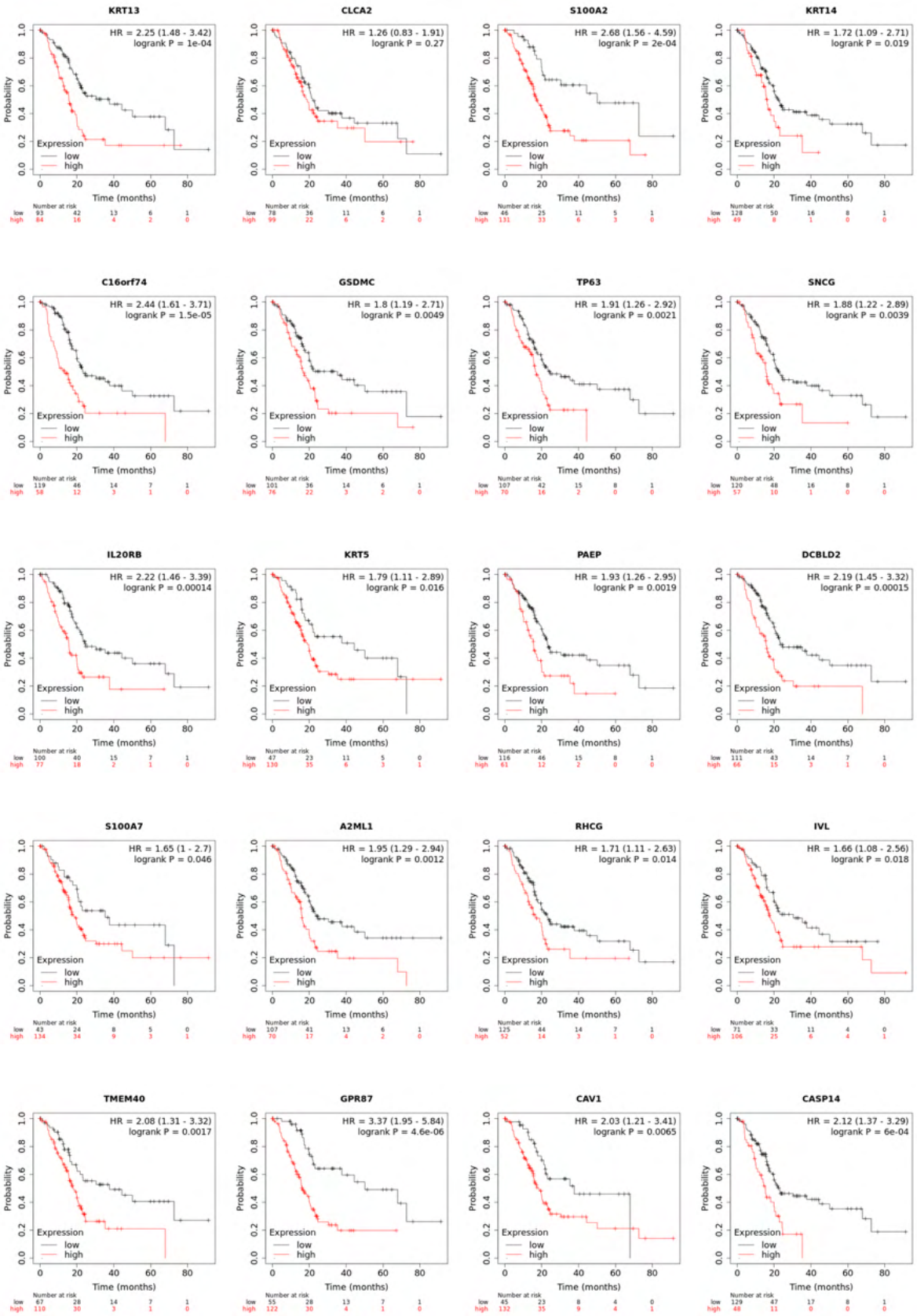


Figure 55 | Kaplan-Meier survival curves for each upregulated advanced stage DEGs. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEGs with insignificant P values were discarded: CLCA2, PPBP, ERMN, FGF11, ITGB3, LOC101927354, KCNQ5, NDUFA4L2, SENA3D and CDH13. KM plotter did not produce a plot for: LINC01468. Graph produced by Kaplan-Meier plotter version 2022.08.23.

10.3. Classical subtype versus basal-like subtype.


KRT13
HR = 2.25 (1.48 - 3.42)
logrank P = 1e-04
CLCA2
HR = 1.26 (0.83 - 1.91)
logrank P = 0.27
S100A2
HR = 2.68 (1.56 - 4.59)
logrank P = 2e-04
KRT14
HR = 1.72 (1.09 - 2.71)
logrank P = 0.019
C16orf74
HR = 2.44 (1.61 - 3.71)
logrank P = 1.5e-05
GSDMC
HR = 1.8 (1.19 - 2.71)
logrank P = 0.0049
TP63
HR = 1.91 (1.26 - 2.92)
logrank P = 0.0021
SNCG
HR = 1.88 (1.22 - 2.89)
logrank P = 0.0039
IL20RB
HR = 2.22 (1.46 - 3.39)
logrank P = 0.00014
KRT5
HR = 1.79 (1.11 - 2.89)
logrank P = 0.016
PAEP
HR = 1.93 (1.26 - 2.95)
logrank P = 0.0019
DCBLD2
HR = 2.19 (1.45 - 3.32)
logrank P = 0.00015
S100A7
HR = 1.65 (1 - 2.7)
logrank P = 0.046
A2ML1
HR = 1.95 (1.29 - 2.94)
logrank P = 0.0012
RHCG
HR = 1.71 (1.11 - 2.63)
logrank P = 0.014
IVL
HR = 1.66 (1.08 - 2.56)
logrank P = 0.018
TMEM40
HR = 2.08 (1.31 - 3.32)
logrank P = 0.0017
GPR87
HR = 3.37 (1.95 - 5.84)
logrank P = 4.6e-06
CAV1
HR = 2.03 (1.21 - 3.41)
logrank P = 0.0065
CASP14
HR = 2.12 (1.37 - 3.29)
logrank P = 6e-04
Probability
Expression
low
high
Time (months)
Number at risk

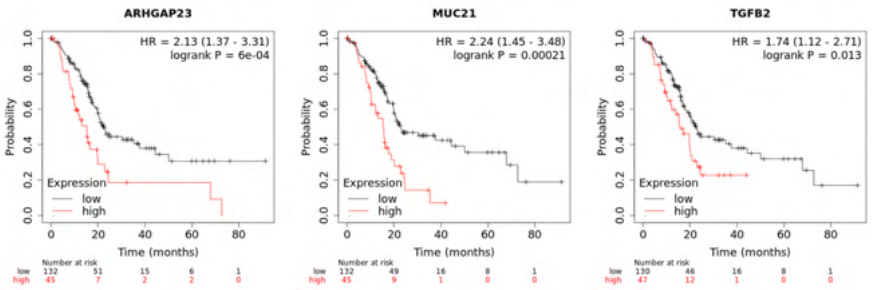
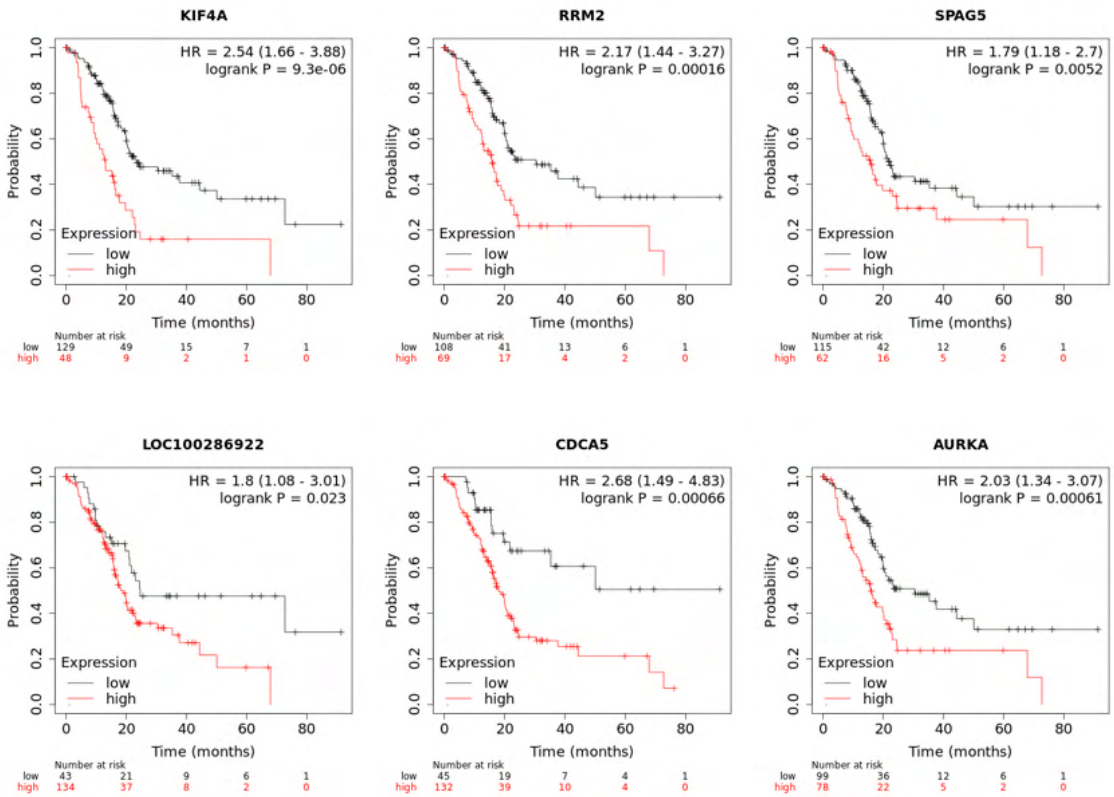


Figure 56 | Kaplan-Meier survival curves for each upregulated basal-like subtype DEGs. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEG with an insignificant P value was discarded: CLCA2. KM plotter did not produce a plot for: MIG7 and MIR205HG. Graph produced by Kaplan-Meier plotter version 2022.08.23.

10.4. WGCNA: Co-expression analyses of dataset 2

10.4.1. Tumour upregulated DEGs



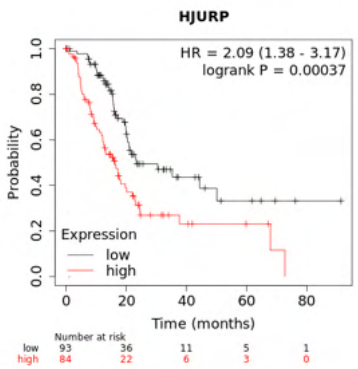


Figure 57 | Kaplan-Meier survival curves for each tumour upregulated DEGs selected based on the co-expressed network. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. Graph produced by Kaplan-Meier plotter version 2022.08.23.

10.4.2. Normal upregulated DEGs

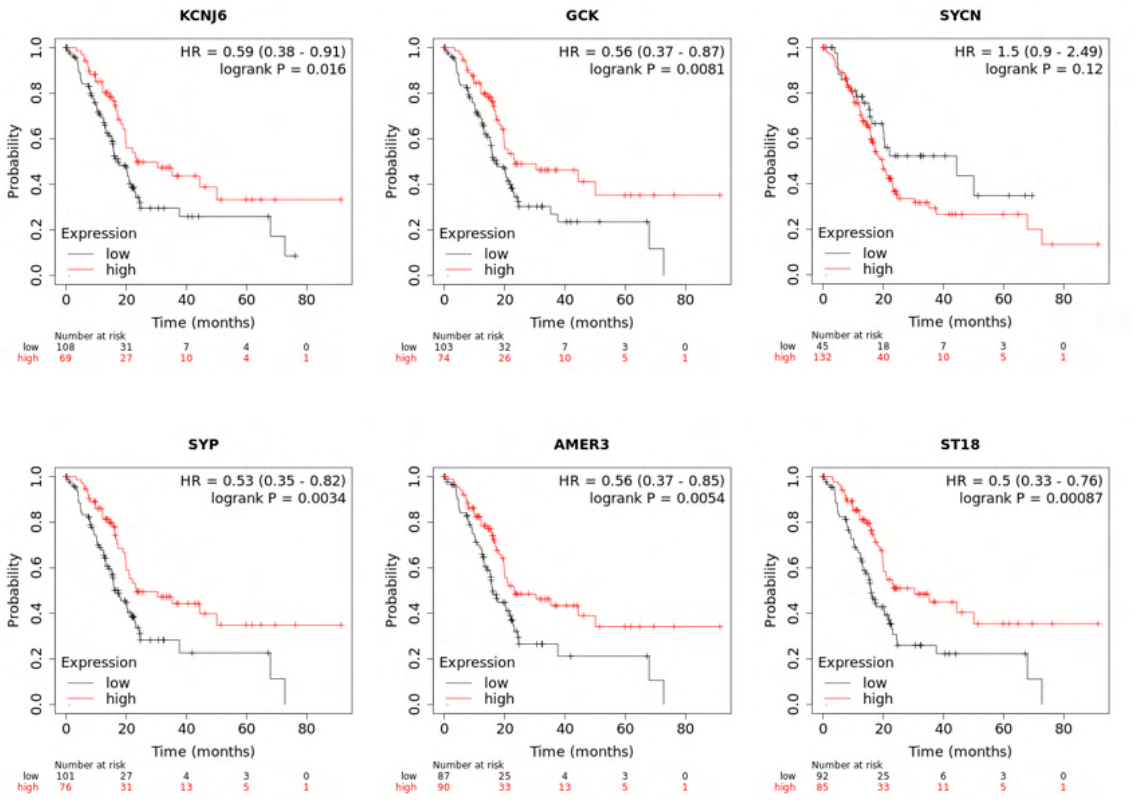


Figure 58 | Kaplan-Meier survival curves for each normal upregulated DEGs selected based on the co-expressed network. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEG with an insignificant P value was discarded: SYCN. Graph produced by Kaplan-Meier plotter version 2022.08.23.

10.4.3. Basal-like subtype upregulated DEGs

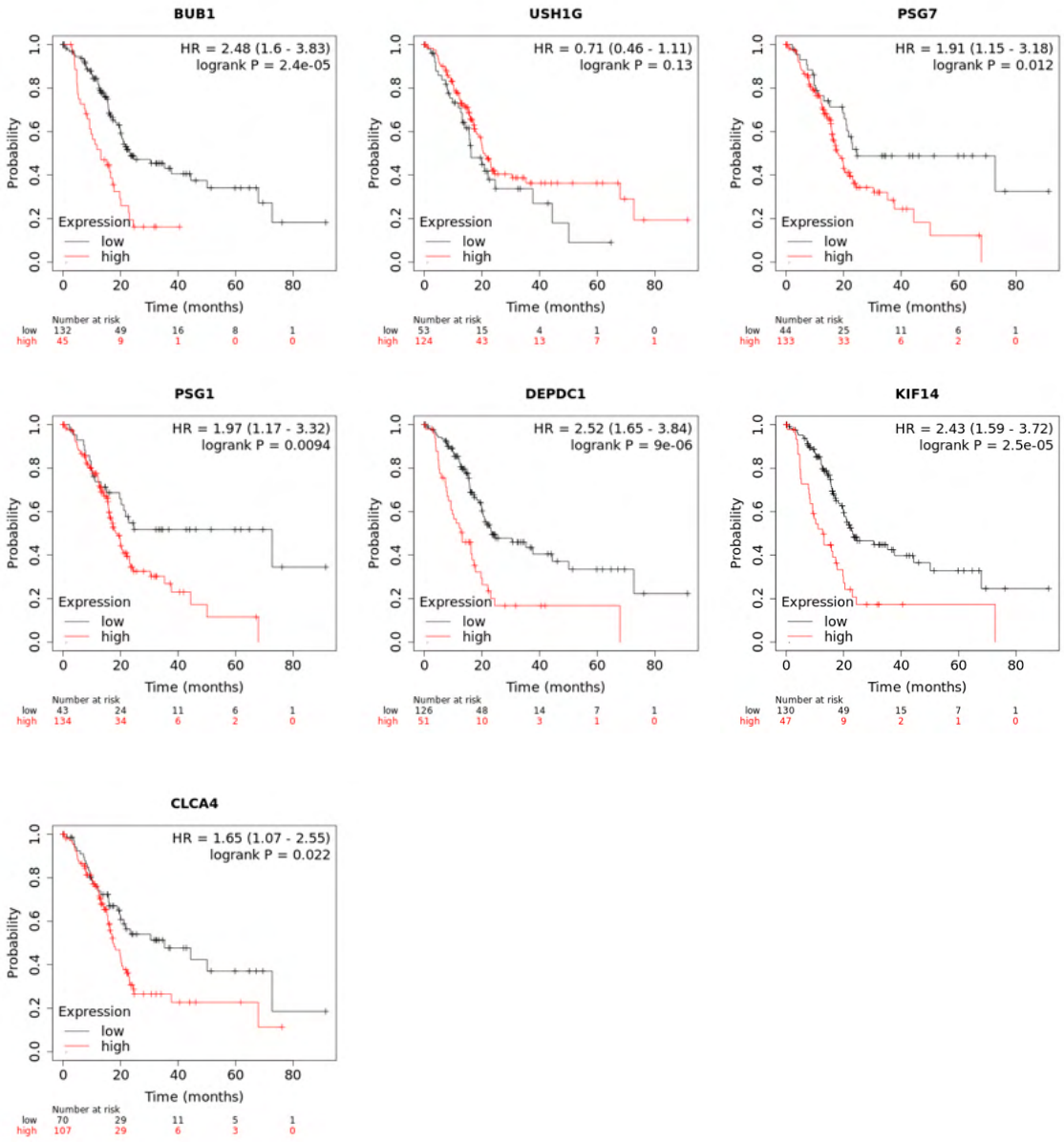


Figure 59 | Kaplan-Meier survival curves for each Basal-like subtype upregulated DEGs selected based on the co-expressed network. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEG with an insignificant P value was discarded: USH1G. Graph produced by Kaplan-Meier plotter version 2022.08.23.

10.4.4. Classical subtype upregulated DEGs

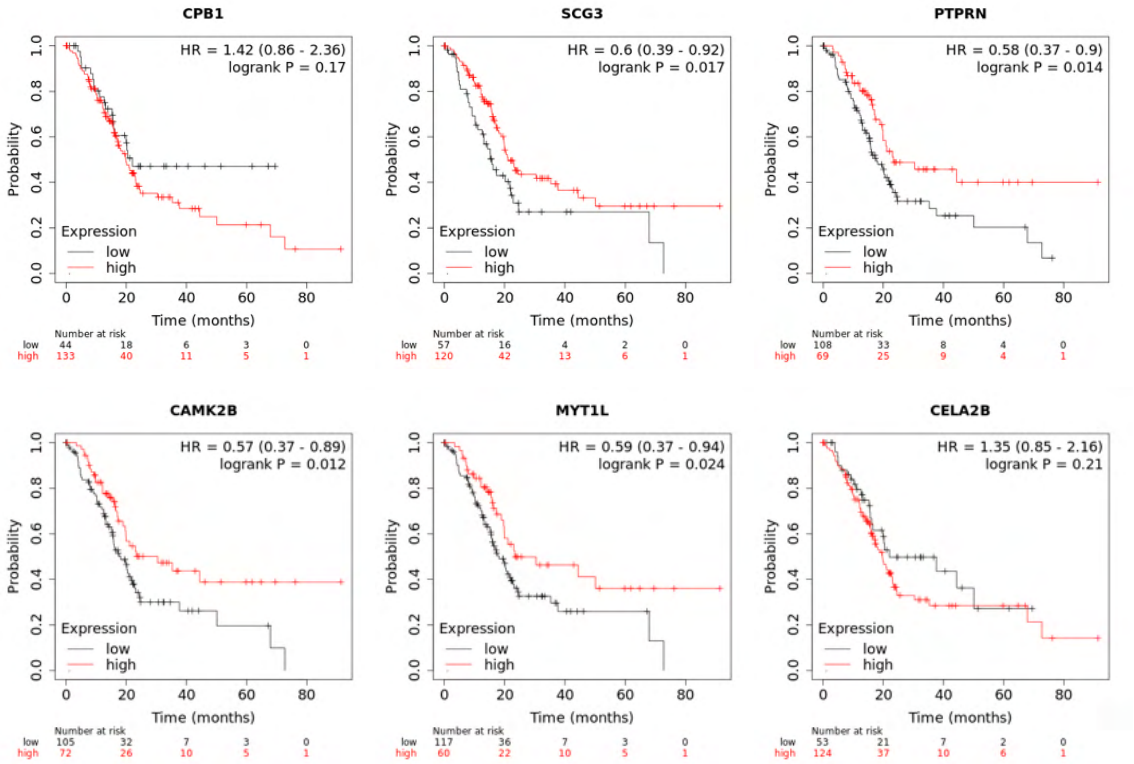


Figure 60 | Kaplan-Meier survival curves for each classical subtype upregulated DEGs selected based on the co-expressed network. Graph show the probability of survival versus the time (months). The hazard ratio (HR) serves as a measurement of the magnitude for the difference between the two curves in each plot. HR shows the hazard reduction obtained by drug studies that is compared to the hazard reduction for the control of the study. Logrank P value represents the average P-value, showing the statistical significance. The black line shows the probability of survival with low gene expression whereas the red line displays the probability of survival with high gene expression. DEGs with

insignificant P values were discarded: CPB1 and CELA2B. Graph produced by Kaplan-Meier plotter version 2022.08.23.

Jamal_1601727_MSc_Plagiarism-1.pdf

ORIGINALITY REPORT

9%	6%	7%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	downloads.hindawi.com Internet Source	1%
2	pubmed.ncbi.nlm.nih.gov Internet Source	<1%
3	www.science.gov Internet Source	<1%
4	"Pancreatic Cancer", Springer Science and Business Media LLC, 2018 Publication	<1%
5	rcastoragev2.blob.core.windows.net Internet Source	<1%
6	bmcplantbiol.biomedcentral.com Internet Source	<1%
7	maayanlab.cloud Internet Source	<1%
8	Lin Shi, Shi-mei Zhao, Yu Luo, An-wen Zhang, Li-hua Wei, Zheng-yi Xie, Yuan-yuan Li, Wei Ma. "MiR-375: A Prospective Regulator in Medullary Thyroid Cancer Based on	<1%