Original Article

Machine learning predicts metastatic progression using novel differentially expressed lncRNAs as potential markers in pancreatic cancer

Hasan Alsharoh[1*]

[1] "Iuliu Haţieganu" University of Medicine and Pharmacy, Cluj-Napoca, Romania

*Corresponding author: hasanalsharoh@gmail.com

Abstract

Pancreatic cancer (PC) is associated with high mortality overall. Recent literature has focused on investigating long noncoding RNAs (lncRNAs) in several cancers, but studies on their functions in PC are lacking. To identify significantly altered expression of lncRNA in PC, I collected information from The Cancer Genome Atlas (TCGA) and extracted RNA-sequencing (RNA-seq) transcriptomic profiles of pancreatic carcinomas and performed differential gene expression analysis. Out of 60,660 gene transcripts shared between 151 PC patients, I identified 38 lncRNAs that were significantly differentially expressed. To further investigate the functions of these genes, gene set enrichment analysis (GSEA) was performed on the population lncRNA panel. GSEA results revealed enrichment of several terms implicated in proliferation. To assess the contribution of these lncRNAs to metastatic progression, I used different ML algorithms, including logistic regression (LR), support vector machine (SVM), random forest classifier (RFC) and eXtreme Gradient Boosting Classifier (XGBC). Explicitly using significantly differentiated lncRNA genes and hyperparameter tuning, in addition to reducing bias through the synthetic minority oversampling technique, the accuracy of the ML models improved. Regardless, out of the four algorithms,

both SVM and RFC were able to predict metastatic progression with 76% accuracy. To the best of my knowledge, this is the first study of its kind to identify this lncRNA panel to differentiate between nonmetastatic PC and metastatic PC, with many novel lncRNAs previously unmapped to PC. The ML accuracy score reveals important involvement of the detected RNAs. Based on these findings, I suggest further investigations of this gene panel *in vitro* and *in vivo*, as they could be targeted for improved outcomes in PC patients, as well as assist in the diagnosis of metastatic progression based on RNA-seq data of primary pancreatic tumors.

1. Introduction

Pancreatic cancer (PC) is one of the deadliest cancers, with an overall five-year survival between 7.2 and 10% according to the literature[1,2]. Evidence suggests that PC is often diagnosed in the late stages of tumorigenesis, likely contributing to its high mortality rate[3]. Recent literature has provided increasing evidence regarding the involvement of long noncoding RNAs (lncRNAs) in the development, invasiveness, angiogenic potential, chemotherapeutic resistance and metastatic capacity of PC[4].

LncRNAs are RNA molecules characterized by having an arbitrary lower cutoff of 200 nucleotides that have been shown not to code for proteins post-transcriptionally [4,5]. LncRNAs have been shown to play complex roles in biological processes in various tissues, with possible implications in DNA repair, cellular proliferation, and human diseases, which made them a common target for recent literature to investigate in cancer [6]. lncRNAs have further been used as biomarkers for overcoming chemoresistance, as well as for the diagnosis of several cancers, including PC [7-10].

Emerging research has been able to provide evidence regarding the use of lncRNAs for improved diagnostic accuracy, prognosis prediction, and treatment adjustment using various methods, including machine learning (ML) techniques[8-10]. Literature regarding the utilization of ML algorithms has been rapidly rising, with literature urging more rapid use of such algorithms in oncology to increase diagnostic accuracy or to further improve on the available algorithms [11-13].

In this study, I aimed to investigate potential lncRNAs involved in the metastatic progression of PC based on RNA-sequencing (RNA-seq) data. To achieve this objective, I collected publicly available data from the cancer genome atlas (TCGA) for 172 patients and filtered the data according to predefined inclusion and exclusion criteria, which resulted in 151 PC records. PC records were further categorized according to their TNM staging, and tumor data were separated into tumors with metastatic activity (TMAs) and tumors without metastatic activity (TWAs). Using bioinformatics analytic techniques, I

identified 125 differentially expressed genes (DEGs) among 60,660 genes involved in this study, many of which were novel. I further assessed the functions of this global gene panel using a multiparametric approach.

Finally, I extracted lncRNA counts from the RNA-seq data from the PC population and further characterized 38 novel lncRNAs that were significantly differentially expressed. To further evaluate their involvement, I used 4 ML algorithms to predict and distinguish between TMAs and TWAs. These algorithms included multivariate logistic regression (LR), support vector machine (SVM), random forest classifier (RFC), and eXtreme Gradient Boosting Classifier (XGBC). I used several techniques to further reduce the bias within the included sample as described in the methodology.

Training and evaluation of the ML algorithms was performed by separating the dataset from the 38 DEGs into a training set and a testing set to eventually evaluate the performance of each of the models. Out of all the ML algorithms, SVM and RFC were able to predict TMAs and TWAs with 76% accuracy using the 38 lncRNA data, suggesting important implications for the specified set of lncRNAs in PC. To the best of my knowledge, this is the first study to identify the involvement of this specific lncRNA panel in PC, with many novel lncRNAs lacking any studies performed on which.

The results of this research have important clinical implications, as the novelty of the lncRNAs requires further comprehensive validation and *in vitro* and *in vivo* investigations. The accuracy shown by the ML model suggests that these novel lncRNAs could be used as biomarkers and further targeted for improved diagnosis and outcome in PC patients.

2. Methods

2.1. Data acquisition

TCGA database was used for data collection and is available at https://www.cancer.gov/tcga. Exploration of TCGA-PAAD project data to acquire pancreatic RNA-seq data was performed on 25/10/2023. File filters applied included a) Data Category: transcriptome profiling; b) Data Type: Gene Expression Quantification; c) Experimental strategy: RNA-Seq; d) Access: open. The case filters applied included the following: a) primary site: pancreas; b) project: TCGA-PAAD; and c) disease type: ductal and lobular neoplasms, adenomas and adenocarcinomas.

The inclusion criteria were that for each RNA-seq dataset to be of similar structure, for the predefined PC tumors mentioned in the filters, or regardless of age and gender. Primary tumors, regardless of metastatic stage, were also included. Exclusion criteria included defects in dataset structure, RNA-seq for tumor adjacent tissues, or those that had undergone prior therapy to a potential previous malignancy. I also excluded records with annotations specifying that tumor data were incorrectly labeled in terms of whether the tumor was neoplastic.

Further categorization was performed for the acquired data using Excel sheets. For TNM subgroup analysis, tumors with staging data were categorized into tumors with metastatic activity, which included those classified as M1, MX/M0 and N1 or above, and tumors without metastatic activity, which included those classified as M0N0. Acquired data were also filtered to include only lncRNA gene expression quantification. This subgrouping was performed prior to DGEA to assess differentially expressed genes between TWAs and TMAs.

2.2. Data analysis

Bioinformatics analysis was conducted on the data following matching the subjects to the study's inclusion and exclusion criteria. Python v3.11 (available at https://www.python.org/) was used in an Anaconda jupyter lab environment [14,15]. To restructure the dataset up for the study population RNA-seq datasets and to import the data into Python, the glob module was used [16]. Data manipulation was performed using pandas library v1.5.3 [17]. Libraries such as numpy and scipy were also utilized for data processing [18,19].

Differential gene expression analysis (DGEA) was performed using PyDESeq2, an R package implemented in Python that has been suggested to be reliable and comparable to the R package[20]. The DEGs were matched to gene symbols and further visualized using the matplotlib[21], seaborn[22], and sanbomics[23] packages. PyDeseq2 calculates the significance of genes using the Wald test, performs count normalization using the trimmed mean of M values (TMM), similar to DESeq2, and relies on the statsmodels library[24,25]. Using count normalization has been shown to have higher accuracy than TPM (transcripts per million) and FPKM (fragments per kilobase of transcript per million fragments mapped)[26]. A further description of the package is available elsewhere[20]. Significant differentiation after adjustment of p values was considered at $p<0.05$ and an absolute log2-fold change (log2FC) of >0.5.

A heatmap of the DEGs was made through the matplotlib[21] package as well. Pearson's correlation coefficient was calculated and mapped for all gene transcript data.

2.3. Gene set and ontology enrichment analysis

Gene set enrichment analysis (GSEA) is a method of interpreting gene-wide expression profiles[27]. GSEA was performed using the GSEApy v1.0.6 package, a Rust implementation of GSEA in python, used for performing computation of RNA-seq count data to evaluate predefined gene sets in association with different phenotypes. I ranked expression data using the prerank function available in the package. The

accuracy of this package has been previously proven, and the method to use it is described extensively elsewhere[28].

Enrichment was performed for several gene collections from MSigDB available at (https://www.gsea-msigdb.org/) and miRTarBase 2017[29]. Gene sets and collections that were evaluated for enrichment were c2.cp.kegg.v2023.1.Hs.symbols, c3.mir.v2023.1.Hs.symbols, c3.tft.v2023.1.Hs.symbols, c4.cgn.v2023.1.Hs.symbols, c5.go.bp.v2023.1.Hs.symbols, c5.go.cc.v2023.1.Hs.symbols, c5.go.mf.v2023.1.Hs.symbols, c5.hpo.v2023.1.Hs.symbols, c6.all.v2023.1.Hs.symbols, h.all.v2023.1.Hs.symbols, and miRTarBase_2017.

Gene Ontology (GO) is a detailed resource with annotations of gene and gene product functions [30,31]. It provides the potential to describe gene functions by assigning them to specific terms in which the genes are linked, detailing their relationships with each other. GO term enrichment was performed through GSEApy, and the results were extracted through tools available in said package. GO graph was made after extracting enriched GO terms and the source identifiers were insert into AmiGO[32].

The false discovery rate (FDR) was considered significant when FDR<0.05. Visualization of GSEA results was performed using tools from GSEApy. Data collected from GSEA results included terms, FDR, enrichment and negative enrichment scores, as well as matched genes. The minimum matching size for gene sets when performing GSEA for the global gene panel was set to 150. However, for the lncRNA panel, the minimum matching size was set to 3, as there were few enriched gene sets.

2.4. Machine learning models

I employed multivariate LR, SVM, RFC, and XGBC to predict metastatic risk for the population based on the lncRNA gene count data from TCGA. DEGs were extracted from DGEA for use as sole predictors of metastatic progression in the study population. Analysis of the models' accuracy was performed using packages from the scipy, scikit-learn, and matplotlib libraries.

To train the ML algorithms, data were categorized into a training set (70% of the data) and a testing set (30%). A random state number was set for all the implemented ML models to dictate a specific seed of randomness during the analysis to maintain reproducibility. For binary classification, TNM stage of IIa or below was designated "0" and considered the TWM for the ML algorithms, while TNM stage IIb or above was designated "1" and considered the TMA. The testing sets were hidden from the ML algorithms to evaluate the predictive capacity performance following model training.

Furthermore, hyperparameter tuning was performed to improve the predictive accuracy of the model. This was done through the GridSearchCV and BayesianSearchCV modules. Fivefold cross-validation was set as a parameter, and data regularization was done through L2 method. The inverse of the regularization strength (or penalty values) was set according to the optimal values found by the search modules specified above. To identify the best parameters, values were also tested over 50 iterations. Moreover, the synthetic minority oversampling technique (SMOTE) was performed to artificially increase TWM population numbers to reduce bias, which has proven to be a powerful tool in improving ML accuracy and addressing imbalanced samples[33].

These methods of standardization were performed for all ML algorithms used. ML algorithms used were also provided by the scikit-learn and XGBoost libraries. All of the algorithms consist of supervised machine learning algorithms, and are commonly used for classifications of tumors[34,35]. Further, L2 regularization has been considered to provide improved accuracy of the ML algorithms[36].

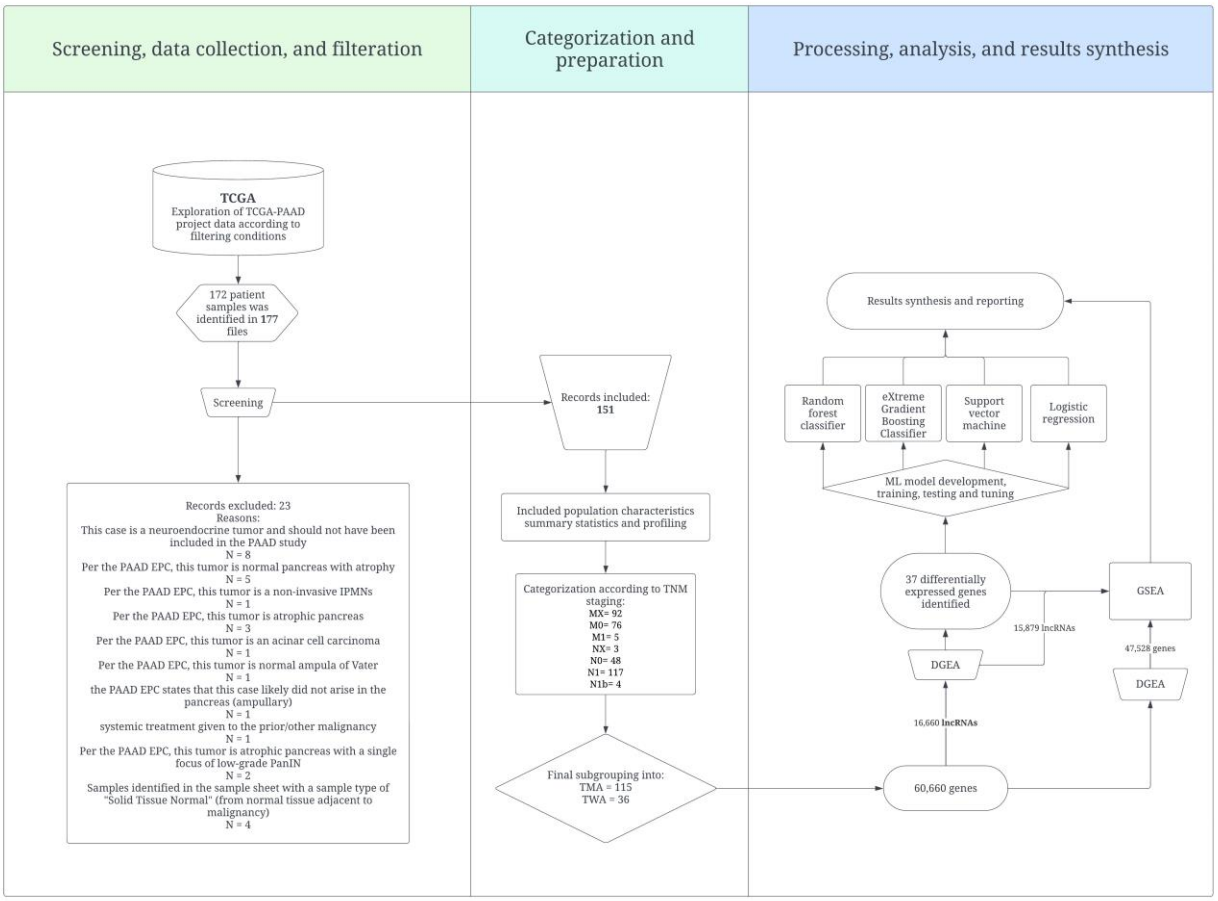
3. Results

3.1. Primary characteristics of the study population

Of the 179 retrieved records, 23 were excluded for the following annotations: a) "This case is a neuroendocrine tumor and should not have been included in the PAAD study" (n = 8); b) "Per the PAAD

161 EPC, this tumor is a normal pancreas with atrophy" (n = 5); c) "Per the PAAD EPC, this tumor is an atrophic
162 pancreas" (n = 3); d) "Per the PAAD EPC, this tumor is a noninvasive IPMN" (n = 1); e) "Per the PAAD EPC,
163 this tumor is an acinar cell carcinoma" (n = 1); f) "Per the PAAD EPC, this tumor is a normal ampula of
164 Vater" (n = 1); g) "The PAAD EPC states that this case likely did not arise in the pancreas (ampullary)" (n =
165 1); h) "Systemic treatment given to the prior/other malignancy" (n = 1); i) "Per the PAAD EPC, this tumor
166 is an atrophic pancreas with a single focus of low-grade PanIN" (n = 2); "Samples identified in the sample
167 sheet with a sample type of "Solid Tissue Normal" (from normal tissue adjacent to malignancy)" (

168 According to the flow diagram found in **Figure 1.** A total of 151 patient records were included. **Table**
169 **1** summarizes the characteristics of the cohort. Notably, 115 records were classified as TMAs, while 36
170 were classified as TWAs. Of the TMAs, 116 were diagnosed as TNM stage IIb, and 8 were diagnosed as
171 stage III and IV. For the TWAs, 26 were at TNM stage IIa.



172

173 **Figure 1.** Flow diagram of the study. Created with Lucidchart, www.lucidchart.com. TCGA: The Cancer
174 Genome Atlas; PAAD: Pancreatic adenocarcinoma; TMA: Tumor with metastatic activity; TWA: Tumor
175 without metastatic activity; DGEA: Differential gene expression analysis; GSEA: Gene set enrichment
176 analysis; ML: Machine learning.

Table 1. Population primary characteristics	
General Characteristics	
Average age	64.62209
Confidence	1.173259
STDEV	10.92365
Max age	88
Min age	35
Males	80
females	70
Pancreas, NOS	14

Head of pancreas	112
Body of pancreas	11
Tail of pancreas	11
Infiltrating duct carcinoma, NOS	133
Adenocarcinoma, NOS	16
Included patient records characteristics	
TMA	115
TWA	36
MX	78
M0	68
M1	5
NX	1
N0	39
N1	108
N1b	3
Staging	
I	0
III	3
IIb	106
IIa	23
IV	5

STDEV; Standard deviation; **NOS**: Not otherwise specified; **TMA**: Tumor with metastatic activity; **TWA**: Tumor without metastatic activity

177

178 The age range of the total patient sample was between 35 and 88 years old (mean = 64.66 ± 10.91).
179 Ninety-four were males, and 78 were females. When reported, 143 had infiltrating duct carcinoma, and
180 16 had adenocarcinoma as the primary diagnosis. Eight had neuroendocrine tumors but were excluded.
181 Seventeen pancreatic tumors had no specified location, 125 were pancreatic head lesions, 15 were
182 pancreatic body lesions, and 13 were pancreatic tail lesions.

183

184 The RNA-seq data included 60,660 gene expression profiles for each of the included patient and
185 control samples. Transcriptomic profiling was performed for the same genes in all patient samples. Of the

available transcripts, 16,901 were lncRNAs. After removing lncRNAs with 0 values among all patients, 15,879 lncRNAs remained. All details regarding the included samples are available in **Supplementary Material 1.**

3.2. DGEA and GSEA of all gene transcripts

A total of 60,660 gene transcripts were filtered following PyDESeq2 analysis, and unavailable values were dropped, resulting in 47,528 transcripts. DGEA revealed 125 differentially expressed genes, as shown in **Table 2,** and the top differentially expressed genes are shown in **Figure 2**. Notably, ADH7, SERPINB13, MIR205HG, NTS, and LINC01300 were the most downregulated genes, with log2FC values of -3.42295, -3.4189, -3.12513, -3.02808, and -2.72096, respectively. The most upregulated genes were PAX7, AC010789.1, TMPRSS15, DEFA6, and DEFA5 and had log2FC values of 3.149596, 3.506053, 3.538356, 3.594891, and 4.800701, respectively.

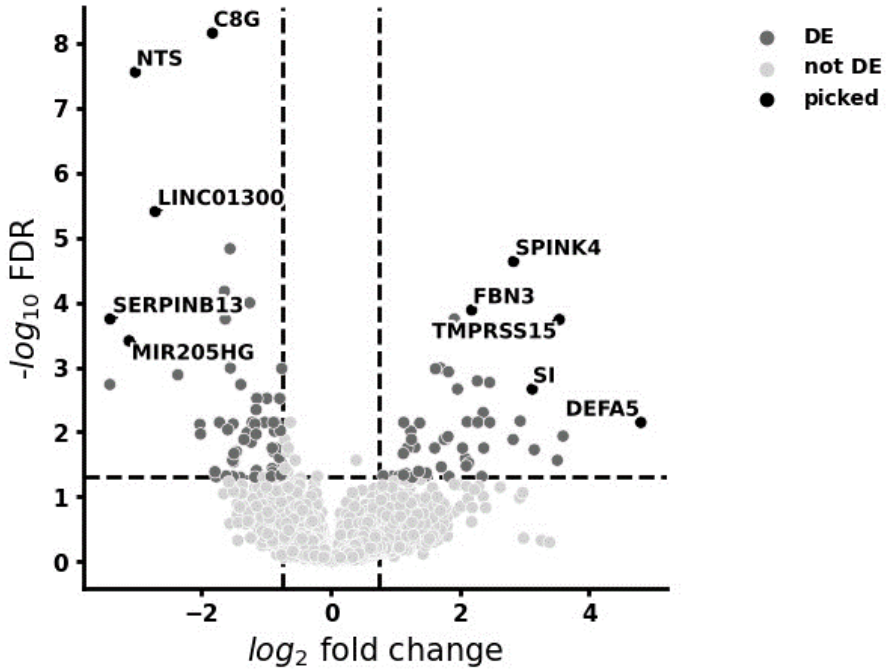


Figure 2. Differentially expressed genes in PC. Absolute log2FC>0.5 and adjusted p value<0.05 were considered as the significance thresholds.

Table 2. Differentially expressed genes found in the global gene sample

ENSEMBL ID	Symbol	log2FoldChange	Rank	Adjusted p-value
ENSG00000196344	ADH7	-3.42295	-4.80123	0.001852
ENSG00000197641	SERPINB13	-3.4189	-5.41464	0.00018
ENSG00000230937	MIR205HG	-3.12513	-5.2119	0.000392
ENSG00000133636	NTS	-3.02808	-7.0417	2.79E-08
ENSG00000253595	LINC01300	-2.72096	-6.20744	3.95E-06
ENSG00000196427	NBPF4	-2.37012	-4.89494	0.001312
ENSG00000122133	PAEP	-2.21459	-5.13911	0.00054
ENSG00000137975	CLCA2	-2.02911	-4.3518	0.007622
ENSG00000241794	SPRR2A	-2.01833	-4.2515	0.010747
ENSG00000176919	C8G	-1.83355	-7.32595	6.96E-09
ENSG00000285722	AC207130.1	-1.7953	-3.80052	0.04115
ENSG00000162951	LRRTM1	-1.77758	-3.70984	0.049095
ENSG00000075673	ATP12A	-1.72492	-4.41944	0.007122
ENSG00000273143	DUSP5-DT	-1.64896	-5.64949	6.75E-05
ENSG00000230916	MTCO1P53	-1.63588	-5.3881	0.000181
ENSG00000170477	KRT4	-1.62228	-3.73908	0.047978
ENSG00000258010	AC016705.1	-1.59722	-4.30607	0.009204
ENSG00000086570	FAT2	-1.58451	-5.08612	0.00067

ENSG00000214711	CAPN14	-1.56223	-5.95983	1.48E-05
ENSG00000101197	BIRC7	-1.55642	-4.98114	0.001031
ENSG00000110680	CALCA	-1.52146	-3.93287	0.02735
ENSG00000130822	PNCK	-1.51657	-3.71624	0.04873
ENSG00000166558	SLC38A8	-1.51515	-4.3587	0.007531
ENSG00000015592	STMN4	-1.50729	-3.91404	0.02896
ENSG00000205426	KRT81	-1.50078	-4.00887	0.021539
ENSG00000154975	CA10	-1.46245	-4.02791	0.020145
ENSG00000016602	CLCA4	-1.40473	-3.69932	0.049966
ENSG00000124466	LYPD3	-1.39559	-4.79298	0.001855
ENSG00000228705	LINC00659	-1.34601	-4.17454	0.013082
ENSG00000134339	SAA2	-1.29552	-4.26585	0.010255
ENSG00000108786	HSD17B1	-1.2613	-6.32798	2.43E-06
ENSG00000121552	CSTA	-1.25807	-5.5557	0.000101
ENSG00000116014	KISS1R	-1.23273	-4.14973	0.014369
ENSG00000204882	GPR20	-1.21769	-4.39539	0.007122
ENSG00000184564	SLITRK6	-1.17582	-3.70529	0.049585
ENSG00000253522	MIR3142HG	-1.17478	-4.3625	0.007531
ENSG00000255129	TTC12-DT	-1.15786	-4.24636	0.01081
ENSG00000233828	MIR4280HG	-1.15767	-4.56131	0.004522

ENSG00000132746	ALDH3B2	-1.15254	-3.81836	0.039435
ENSG00000181652	ATG9B	-1.14738	-4.6546	0.003036
ENSG00000115008	IL1A	-1.12498	-3.77886	0.043629
ENSG00000177627	C12orf54	-1.02545	-4.3938	0.007122
ENSG00000180739	S1PR5	-0.99067	-4.65304	0.003036
ENSG00000272948	AP001412.1	-0.92415	-3.72701	0.048613
ENSG00000167971	CASKIN1	-0.90985	-3.8412	0.036673
ENSG00000278743	AC087239.1	-0.90938	-4.0769	0.01772
ENSG00000175189	INHBC	-0.90458	-3.79432	0.041787
ENSG00000272906	AL353708.3	-0.88897	-4.40552	0.007122
ENSG00000178445	GLDC	-0.88519	-4.03342	0.020145
ENSG00000268041	ERFL	-0.8732	-4.28133	0.009738
ENSG00000254266	PKIA-AS1	-0.86706	-4.38841	0.007122
ENSG00000117407	ARTN	-0.8143	-4.08225	0.01772
ENSG00000204963	PCDHA7	-0.79643	-3.97143	0.024672
ENSG00000286810	AL513128.3	-0.793	-4.66358	0.003036
ENSG00000268403	AC132192.2	-0.78124	-4.29421	0.00953
ENSG00000277218	AL139123.1	-0.77132	-3.75731	0.046252
ENSG00000102466	FGF14	-0.76382	-3.83777	0.036813
ENSG00000100162	CENPM	-0.76299	-3.89051	0.031239

ENSG00000232573	RPL3P4	-0.76122	-4.96717	0.00105
ENSG00000237181	PRKAR1B-AS1	-0.75848	-4.09162	0.017711
ENSG00000233901	LINC01503	-0.73675	-3.75264	0.046672
ENSG00000267710	EDDM13	-0.71591	-4.18507	0.013075
ENSG00000196420	S100A5	-0.70536	-3.75049	0.046672
ENSG00000287575	AL390755.3	-0.70456	-3.84387	0.036651
ENSG00000227256	MIS18A-AS1	-0.66983	-3.80296	0.041147
ENSG00000263412	NFE2L1-DT	-0.66855	-4.03071	0.020145
ENSG00000158292	GPR153	-0.6636	-3.71762	0.04873
ENSG00000270426	AC099343.2	-0.66056	-4.09624	0.017609
ENSG00000269961	ERBIN-DT	-0.62896	-4.4063	0.007122
ENSG00000270659	AC079610.1	-0.55231	-3.93098	0.02735
ENSG00000109684	CLNK	0.811049	3.743661	0.047532
ENSG00000007171	NOS2	0.9501	3.727137	0.048613
ENSG00000168004	PLAAT5	1.046959	3.713844	0.04873
ENSG00000217275		1.103167	3.945237	0.026898
ENSG00000244675	AC108676.1	1.121008	4.006455	0.021539
ENSG00000249574	AC226118.1	1.122255	3.758216	0.046252
ENSG00000165186	PTCHD1	1.127601	4.425499	0.007122
ENSG00000204710	SPDYC	1.164409	3.774913	0.043911

ENSG00000133317	LGALS12	1.185546	4.074521	0.01772
ENSG00000110195	FOLR1	1.236702	4.28444	0.009738
ENSG00000179766	ATP8B5P	1.248188	4.18952	0.013025
ENSG00000243910	TUBA4B	1.26142	3.699883	0.049966
ENSG00000231106	LINC01436	1.269281	3.713818	0.04873
ENSG00000079841	RIMS1	1.301664	4.104695	0.017223
ENSG00000254872	LINC02688	1.359219	3.809804	0.040421
ENSG00000077935	SMC1B	1.37522	4.3749	0.007278
ENSG00000047936	ROS1	1.462054	3.733667	0.048592
ENSG00000250337	PURPL	1.478326	3.784362	0.043081
ENSG00000211951	IGHV2-26	1.607862	4.072645	0.01772
ENSG00000113722	CDX1	1.619806	4.955568	0.001058
ENSG00000261409		1.673178	4.087027	0.01772
ENSG00000095627	TDRD1	1.695616	4.99184	0.001031
ENSG00000275874	PICSAR	1.709471	3.859728	0.034709
ENSG00000138823	MTTP	1.75384	4.194027	0.012975
ENSG00000109182	CWH43	1.779912	4.179552	0.013082
ENSG00000286734	AC133530.1	1.81407	4.219397	0.011788
ENSG00000159251	ACTC1	1.820261	4.924839	0.00118
ENSG00000248635		1.82577	4.389008	0.007122

ENSG00000124237	C20orf85	1.830513	3.714347	0.04873
ENSG00000070019	GUCY2C	1.911694	5.38153	0.000181
ENSG00000185105	MYADML2	1.961126	4.745671	0.002179
ENSG00000179914	ITLN1	2.039036	4.069503	0.01773
ENSG00000130700	GATA5	2.082396	3.959788	0.025605
ENSG00000264404	LINC02675	2.097143	3.871155	0.03347
ENSG00000189052	CGB5	2.11069	4.450157	0.006997
ENSG00000198788	MUC2	2.134906	3.90208	0.030102
ENSG00000142449	FBN3	2.180393	5.489955	0.000131
ENSG00000250376		2.232802	4.650635	0.003036
ENSG00000151365	THRSP	2.253533	4.419036	0.007122
ENSG00000115850	LCT	2.267295	4.842754	0.001634
ENSG00000198842	STYXL2	2.278382	4.441746	0.007079
ENSG00000205076	LGALS7	2.337392	3.713775	0.04873
ENSG00000166869	CHP2	2.35771	4.533106	0.005018
ENSG00000113196	HAND1	2.366848	4.07674	0.01772
ENSG00000091138	SLC26A3	2.457377	4.823641	0.001724
ENSG00000282122	IGHV7-4-1	2.461323	4.384127	0.007122
ENSG00000016490	CLCA1	2.822016	4.177146	0.013082
ENSG00000122711	SPINK4	2.828123	5.854698	2.34E-05

ENSG00000228674	PPIAP59	2.932048	4.462678	0.006788
ENSG00000090402	SI	3.119419	4.749043	0.002179
ENSG00000009709	PAX7	3.149596	4.052701	0.018812
ENSG00000224817	AC010789.1	3.506053	3.93045	0.02735
ENSG00000154646	TMPRSS15	3.538356	5.363727	0.000184
ENSG00000164822	DEFA6	3.594891	4.227086	0.011582
ENSG00000164816	DEFA5	4.800701	4.413106	0.007122

GSEA was subsequently performed, with libraries investigated available in **Supplementary Materials 2**. There were many gene sets enriched with the genes, as many genes were included in the study's gene panel. Notably, several GO terms were enriched, as well as some terms from miRTarBase 2017, as shown in **Figure 3 A and B**. FDR values were significant for the enriched terms (FDR<0.01).

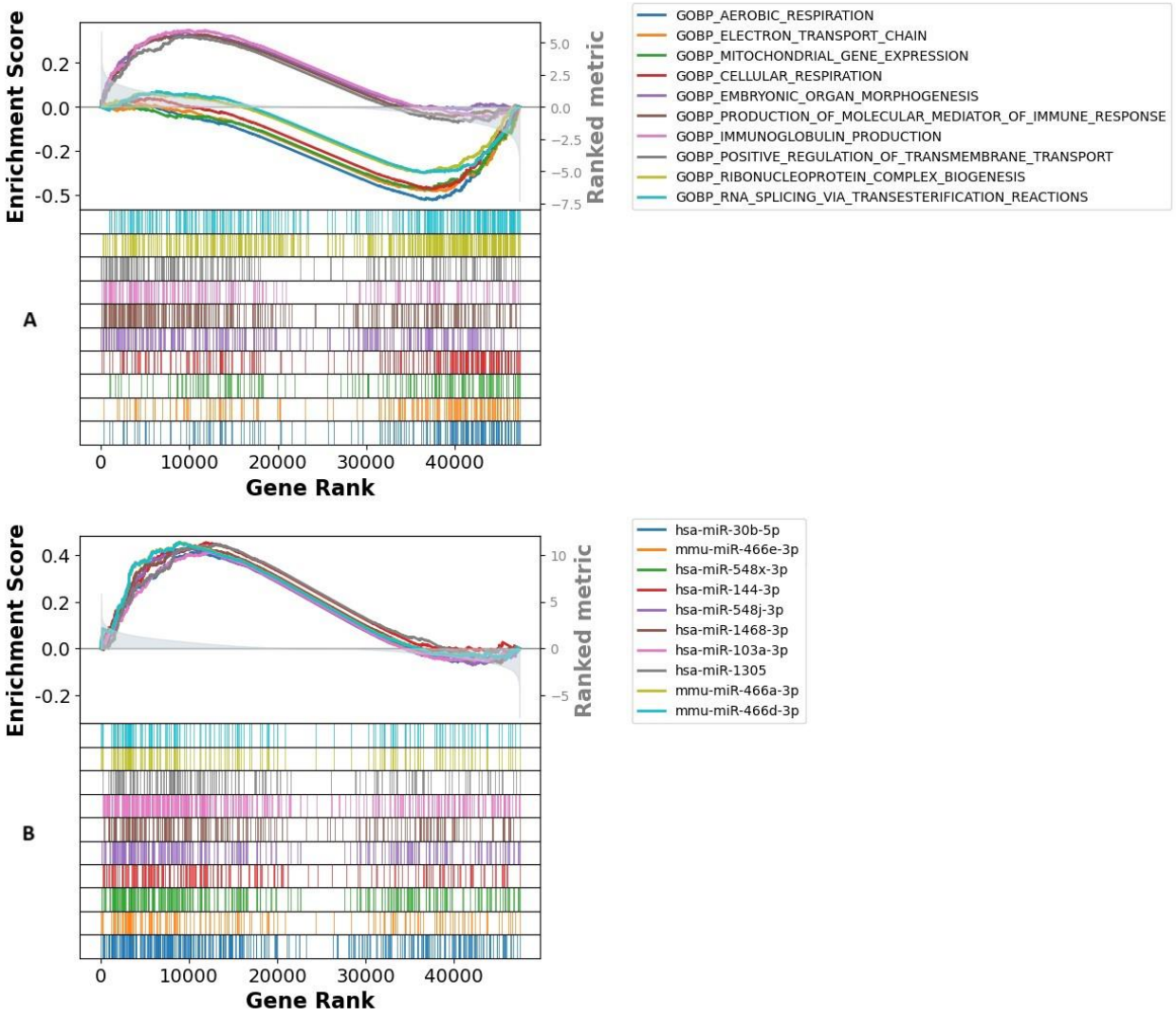


Fig.3 A. GOBP (GO biological process) term enrichment. Upregulated genes had a lower rank, and downregulated genes had a higher rank. The enrichment score correlates with the number of genes from the gene panel enriching the gene set with significantly differentiated expression. More genes enriching this term are downregulated in this study due to the enrichment score reaching -0.5 since these genes have a higher density of higher ranked genes. **B.** miRTarBase_2017 term enrichment. Upregulated genes had a lower rank, and downregulated genes had a higher rank. The enrichment score correlates with the number of genes from the gene panel enriching the gene set with significantly

differentiated expression. Here, the gene set was more enriched with the upregulated genes from the gene panel.

3.3. lncRNA DGEA, correlations, and GSEA

Further subgroup analysis was performed for lncRNAs in PC, which returned 16,901 gene expression values, for which PyDeseq2 was also used to analyze DEGs. Dropping the 0-sum, duplicate, and unavailable values retrieved 15,568 lncRNAs. Of the lncRNA panel, 38 lncRNAs were significantly differentially expressed (shown in **Figure 4**).

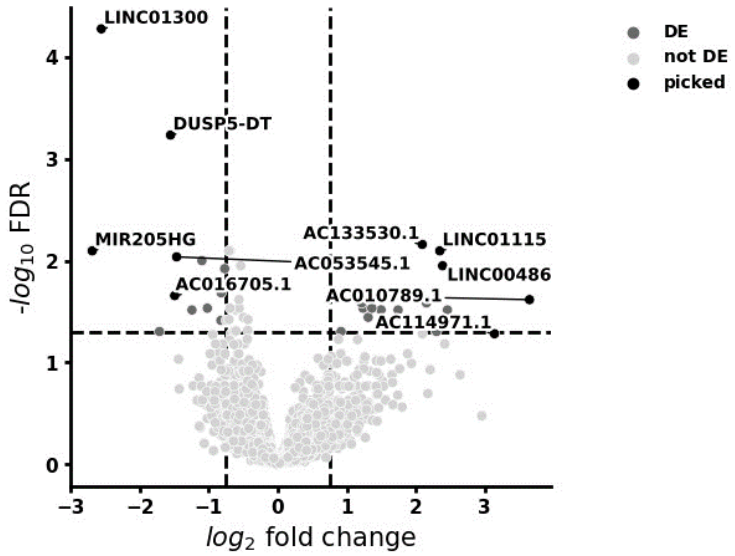


Figure 4. Differentially expressed LncRNA. Absolute log2FC>0.5 and adjusted p value<0.05 were considered as the significance thresholds.

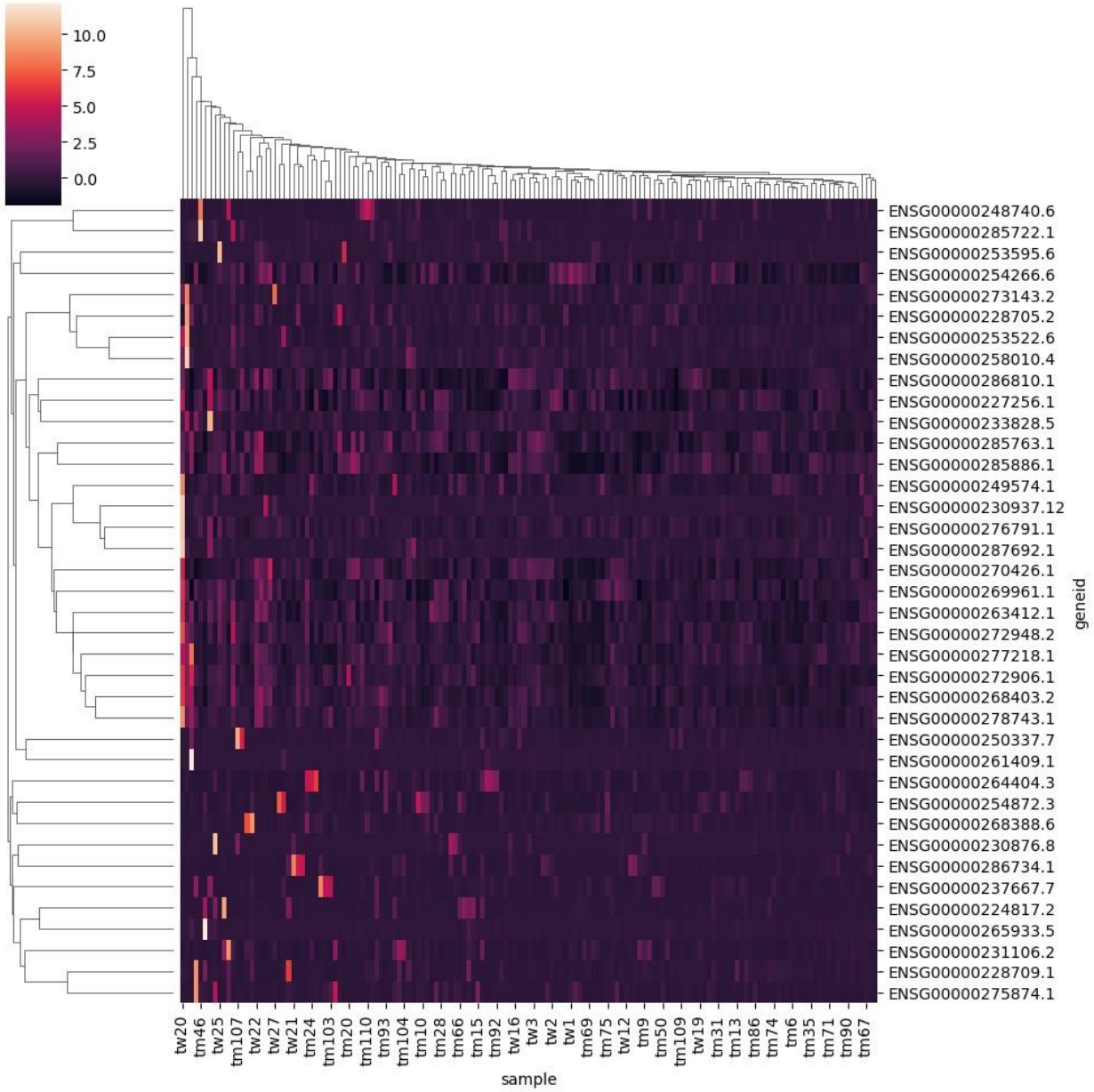
Interestingly, the most downregulated genes were LINC01300, DUSP5-DT, AL513128.3, MIR205HG, and AC132192.2, with Log2FC values of -2.55682, -1.55378, -0.70877, -2.68894, and -0.68868, respectively. The most upregulated genes were AC010789.1, LINC00486, ENSG00000261409 (referred to as RF00019), LINC01115, and AC133530.1, with log2FC values of 2.154221, 1.214608, 3.647081, 1.705921, and 2.388161, respectively. Results of DGEA on the lncRNAs are shown in **Table 3**.

Table 3. DGEA of lncRNAs in PC.

ENSEMBL ID	symbol	log2FoldChange	Rank	Adjusted p-value
ENSG00000253595	LINC01300	-2.55682	-5.79229	5.25E-05
ENSG00000273143	DUSP5-DT	-1.55378	-5.24801	0.000582
ENSG00000286810	AL513128.3	-0.70877	-4.56204	0.007988
ENSG00000230937	MIR205HG	-2.68894	-4.5098	0.007988
ENSG00000268403	AC132192.2	-0.68868	-4.48214	0.007988
ENSG00000287692	AC053545.1	-1.46588	-4.42416	0.009158
ENSG00000233828	MIR4280HG	-1.09978	-4.37975	0.00999
ENSG00000269961	ERBIN-DT	-0.53815	-4.31064	0.011198
ENSG00000272906	AL353708.3	-0.76963	-4.27697	0.011946
ENSG00000254266	PKIA-AS1	-0.82096	-4.11833	0.020627
ENSG00000258010	AC016705.1	-1.49399	-4.08738	0.022009
ENSG00000270426	AC099343.2	-0.56239	-4.03685	0.024114
ENSG00000253522	MIR3142HG	-1.02076	-3.95022	0.029262
ENSG00000263412	NFE2L1-DT	-0.5603	-3.93248	0.029262
ENSG00000277218	AL139123.1	-0.69149	-3.90873	0.029262
ENSG00000227256	MIS18A-AS1	-0.61208	-3.88033	0.03036
ENSG00000228705	LINC00659	-1.24398	-3.8544	0.030461
ENSG00000285886	AC211476.6	-0.70691	-3.76746	0.037816
ENSG00000272948	AP001412.1	-0.82452	-3.75478	0.038514
ENSG00000278743	AC087239.1	-0.74556	-3.74815	0.038514
ENSG00000285763	AL358777.1	-0.60768	-3.67036	0.048983
ENSG00000276791	AC092117.1	-0.63072	-3.66641	0.048983
ENSG00000285722	AC207130.1	-1.71493	-3.65656	0.049573
ENSG00000265933	LINC00668	2.298864	3.643841	0.049573
ENSG00000268388	FENDRR	0.916365	3.644596	0.049573
ENSG00000231106	LINC01436	1.30986	3.795489	0.035968
ENSG00000248740	LINC02428	2.458109	3.852876	0.030461
ENSG00000275874	PICSAR	1.742661	3.862257	0.030461
ENSG00000250337	PURPL	1.495827	3.887343	0.03036
ENSG00000228709	LINC02575	1.235504	3.908659	0.029262
ENSG00000254872	LINC02688	1.363374	3.920829	0.029262
ENSG00000264404	LINC02675	2.154221	3.993045	0.025979
ENSG00000249574	AC226118.1	1.214608	3.998806	0.025979
ENSG00000224817	AC010789.1	3.647081	4.04573	0.024114
ENSG00000261409		1.705921	4.224935	0.013912
ENSG00000230876	LINC00486	2.388161	4.312419	0.011198
ENSG00000237667	LINC01115	2.347361	4.509832	0.007988
ENSG00000286734	AC133530.1	2.092716	4.689599	0.006905

228

229 Moreover, since the number of DEGs was feasible, to further visualize the relationship between
230 these lncRNAs, each was correlated to the rest, and Pearson's correlation coefficients for all the lncRNAs
231 were extracted. The results are visualized in **Figure 5**. A table of all Pearson's correlation coefficients can
232 be found in **Supplementary Material 3**.



233

Figure 5. Hierarchical clustering heatmap of lncRNAs amongst the sample population. The color gradient in the legend refers to Pearson's correlation coefficient. The dendrogram linkage is based on the correlation strength. Geneid: ENSEMBL ID. tw: TWAs; tm: TMAs.

GSEA and GO analyses were subsequently performed for all the lncRNA data. Due to the lack of studies on the genes of these transcripts, there was no significant enrichment in most databases. Notably, a few terms were enriched from the MSigDB c3.tft.v2023.1.Hs.symbols collection, which is focused on transcription factors. The results of the term enrichment for the top 10 terms in this collection are shown in **Figure 6**, and the results for insignificant term enrichment for other collections and databases can be found in **Supplementary Material 3.**

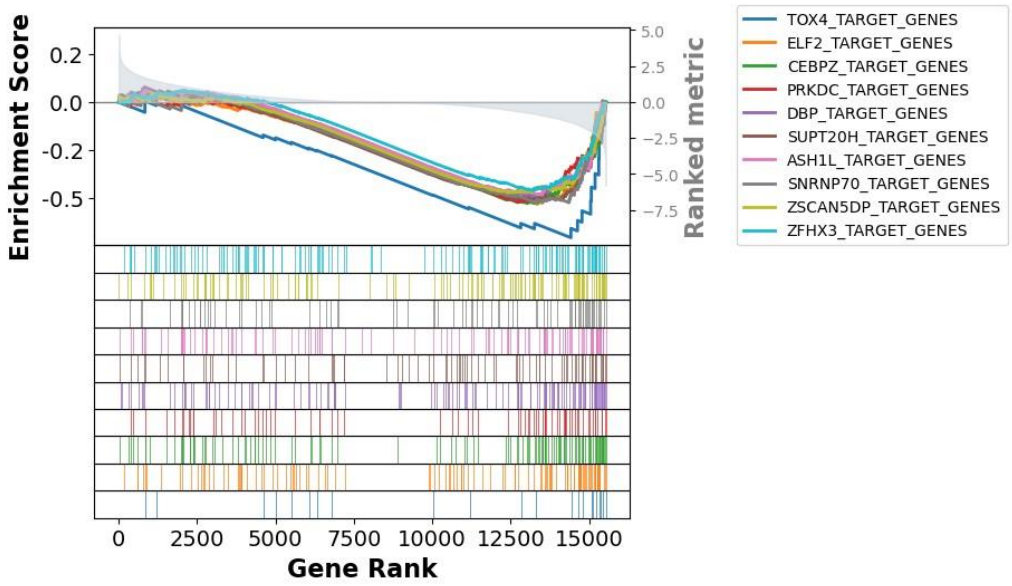


Figure 6. GSEA of lncRNA data. Terms are more significantly enriched with downregulated genes.

3.4. ML model prediction of PC metastatic potential according to lncRNA gene expression

Following the training and testing of each of the ML models, optimizations were performed to find the highest possible accuracy obtainable while reducing bias. Therefore, SMOTE was implemented in all

the ML algorithms. Reducing sample imbalances improved the predictive accuracy of the utilized algorithms.

Following SMOTE implementation and thorough hyperparameter tuning, LR demonstrated an accuracy score of 73.91% when distinguishing between TMAs and TWAs when tested, as well as an F1 score of 82.57% and a recall of 90.63%. Regardless, the area under the curve (AUC) for LR was 0.63, which was relatively low. **Figure 7 A and B** show the receiver operating characteristic (ROC) curve and for logistic regression following the implementation of SMOTE and the precision-recall (PR) curve. **C** shows the weight of each lncRNA (feature) in assisting the regression

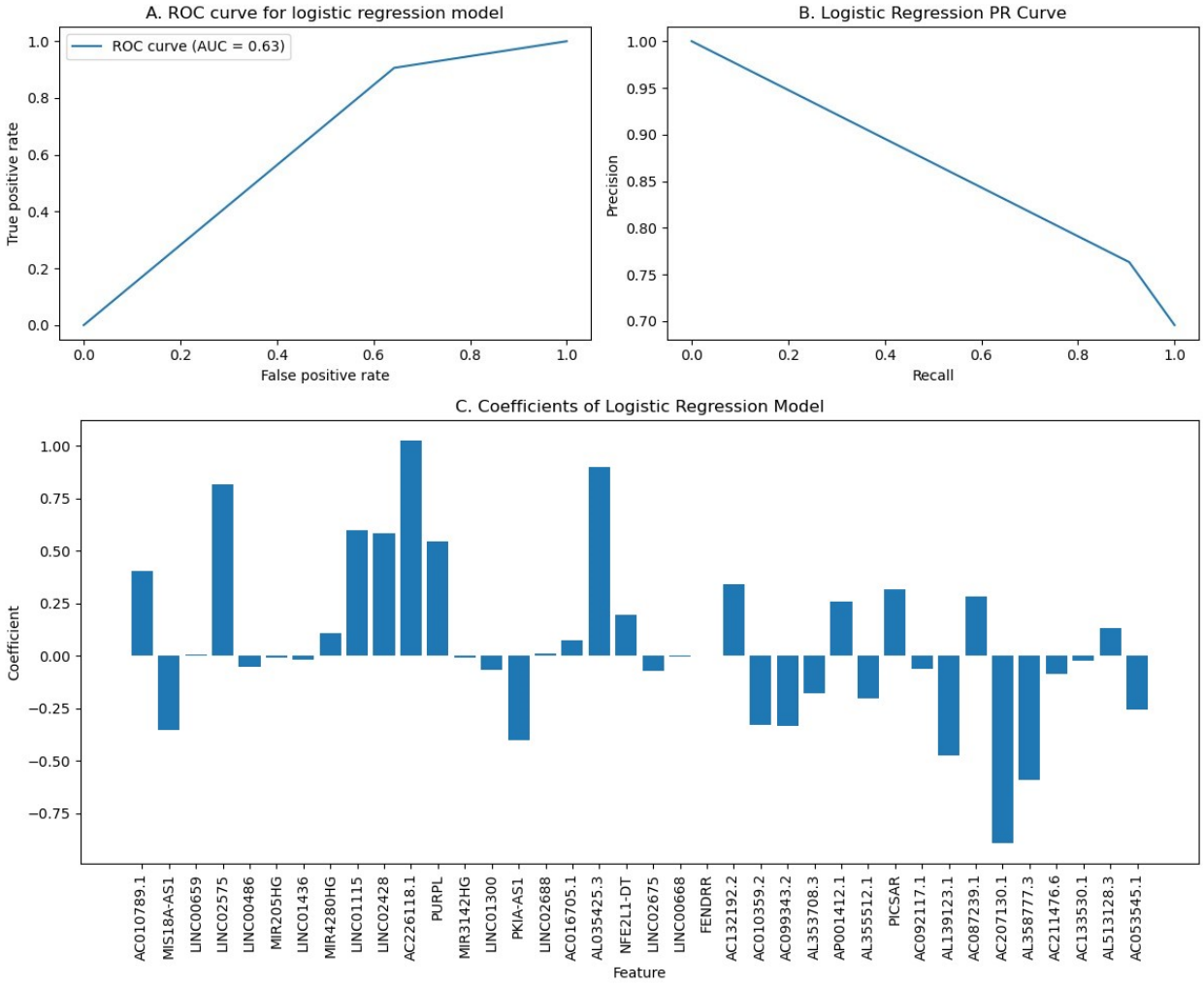


Fig.7 A. The LR model showed an AUC = 0.63, demonstrating relatively weak classification performance, despite the good accuracy of detecting PC cases at TNM stage IIb or above. **B.** LR model accuracy of predicting positive values in comparison to the true positive rate (recall). **C.** Weights of each of the differentially expressed lncRNAs allowing the LR model to differentiate between nonmetastatic tumors and metastatic tumors.

For the SVM model, SMOTE implementation, and hyperameter tuning also improved the predictive potential of the algorithm, which, on testing, returned an accuracy of 76.09%, with a true positive rate of 84.51% and a recall of 93.75%. **Figure A and B** show the ROC curve as well as the PR curve of the SVM model.

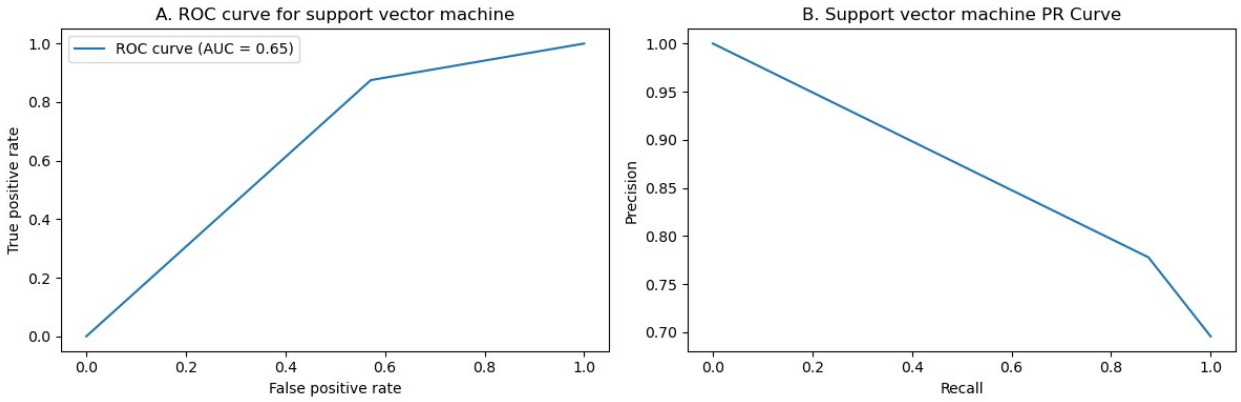


Fig. 8 A. The SVM algorithm showed an AUC = 0.65, demonstrating modest accuracy of detecting PC cases at TNM stage IIb or above and distinguishing them from less metastatic stages. **B.** SVM model accuracy of predicting positive values in comparison to its recall capacity.

RFC was one of the most accurate models; after hyperparameter tuning, it returned an accuracy of 76.09% and an F1 score of 81.96%, with a recall of 78.13%. Most importantly, the AUC for this model was 0.75, showing good performance in classifying the tumors. Regardless, the gene panel consisting of 38 genes allowed the ML algorithms to discern advanced TNM stages from relatively early TNM stages in PC. **Figure 9 A and B** also show the RFC model accuracy and PR curve.

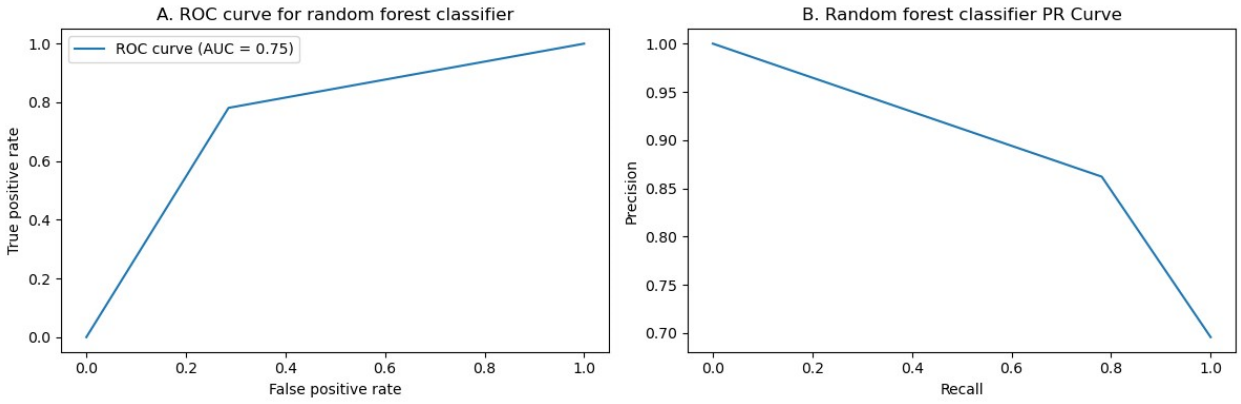
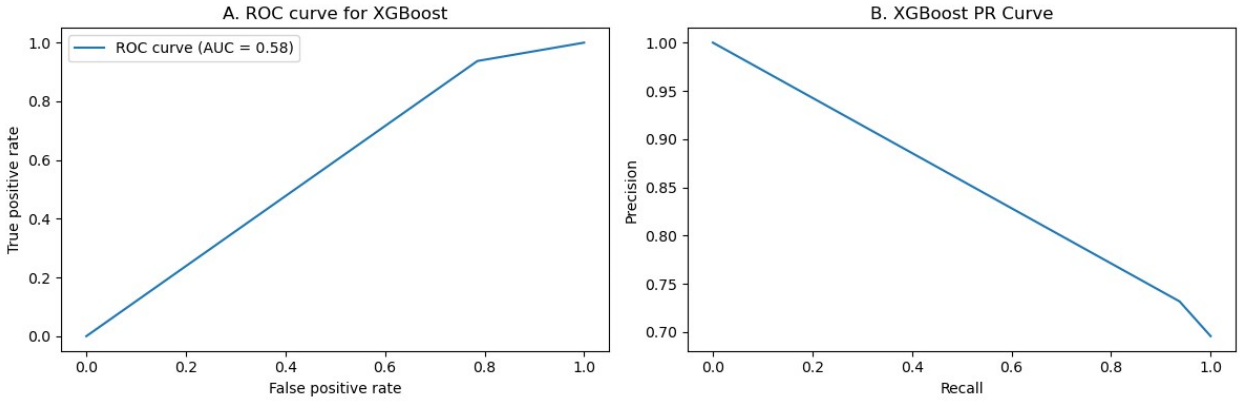


Fig.9 A. The RFC ROC AUC was 0.75, demonstrating acceptable accuracy of detecting PC cases among the other ML algorithms when using the differentially expressed lncRNA counts data, regardless of hyperparameter tuning. **B.** The RFC PR curve showed good recall, albeit with low precision.

As for XGBC, the model showed 71.73% accuracy; This specific model had the most inconsistency in predicting tumor types following each randomization. **Figure 10 A and B** show the low AUC and its PR curve. Data regarding the evaluation of the ML algorithms are available in **Supplementary Material 4**.



4. Discussion

Despite advances in diagnostics and therapeutics, PC remains a very challenging condition to treat, with consistently high mortality rates and limited available treatments[37,38]. Recently, research has focused on identifying prognostic markers for PC, and preclinical studies have identified several prognostic lncRNA

signatures[8,39-41]. LncRNAs have been further suggested to have implications in diagnosis, drug resistance, and therapeutics in PC[4]. However, as most patients are often diagnosed at advanced stages of disease, mutational burdens show complex relationships with lncRNA regulation[4]. Therefore, as the literature suggests, these relationships must be investigated to adjust treatment modalities. This becomes even more crucial in the latter stages of PC.

This study aimed to provide details regarding DEGs in PC first and then to further analyze differentially expressed lncRNA and assess the diagnostic potential of these lncRNAs during the transition from stage IIa and stage IIb and above. These lncRNAs were extracted after performing DGEA to extract 38 gene transcripts from the global RNA-seq gene panel among 151 patient samples. The diagnostic potential of lncRNAs was assessed using supervised ML techniques to predict metastatic transition. I employed four ML techniques with established accuracy in prediction: LR[42], SVM[43], RFC[43] and XBGC[44].

DGEA of the global gene panel revealed 125 DEGs, many of which were previously uninvestigated. Of the downregulated DEGs, ADH7 was hypothesized to have implications when mutated in pancreatic injury[45]. NTS was also associated with PC[46]. However, SERPINB13 and MIR205HG were previously unexplored in PC but had been discussed in other cancers and were implicated in poor clinical outcomes[47,48]. No studies are available regarding LINC01300, which warrants further investigation. For the upregulated DEGs, PAX7 was previously reported to have some relationship with cancers, yet studies regarding this specific gene transcript are lacking [49]. For DEFA6 and DEFA5, a report suggested a link between them and clinical outcomes in colorectal cancer[50]. While there were no studies regarding AC010789.1 and TMPRSS15 in PC, some studies linked the potential implications of these genes with other cancers[51,52].

GSEA for the global gene panel revealed several enriched pathways. For example, GO enrichment revealed that the gene panel significantly enriched pathways relevant in the regulation of aerobic

311 respiration (GO:1903715), electron transport carrier chain (GO:0022900), and mitochondrial gene
312 expression and translation into RNA transcripts (GO:0140053). Notably, of the miRTarBase enriched
313 pathways, mir-30b-5p microRNA (miRNA) was previously linked to PC[53,54]. While miR-548x-3p has not
314 been studied regarding its function in cancer, miR-144-3p was previously implicated in PC[55,56].
315 Additionally, mir-548j-3p had no studies documenting its relationship with cancer. For miR-1468-3p,
316 some studies have suggested it as a biomarker for non-small cell lung cancer and prostate cancer[57,58].

317 Following the filtering of the global RNA-seq gene panel to lncRNAs exclusively, DGEA revealed 38
318 differentially expressed lncRNAs, many of which were novel. LINC01300 and MIR205HG, as previously
319 described, in addition to DUSP5-DT and AL513128.3, had no studies in PC, with the latter two lacking any
320 studies on which. In contrast, one report regarding AC132192.2 indicated its relevance in prostate
321 cancer[59]. For the upregulated lncRNAs, AC010789.1, as previously stated, had a report regarding its
322 function in colorectal cancer[52,60]. LINC00486, RF00019, LINC01115, and AC133530.1 all lack validation
323 studies in PC, but other reports indicate involvement in several diseases, including cancer[61-64].

324 As these novel lncRNAs lack studies regarding their functions, GSEA of the selected MSigDB collections
325 returned no significant enrichment but in one transcription factor collection. Notably, the most enriched
326 pathway described genes containing one or more binding regions for a transcription factor that regulates
327 cell fate and controls cell cycle progression from the mitotic phase to interphase, known as TOX high
328 mobility group box family member 4 (TOX4)[65,66]. Interestingly, lncRNAs enriching this path were primarily
329 downregulated.

330 To further explore the significance of the identified 38 lncRNAs, ML algorithms were employed to
331 predict the metastatic state of cancer (designated “0” for stages IIa or below and “1” for stages IIb and
332 above). Of all the algorithms, RFC showed superior accuracy to the other algorithms, showing an AUC of
333 0.75 and an accuracy of over 76%. While there is much to be understood regarding the functions of the

identified lncRNA panel, the accuracy shown by RFC reveals important aspects about the involvement of these lncRNAs in PC. This finding warrants further *in vitro* and *in vivo* investigations.

For most of the identified lncRNA panels, this was the first study to uncover their involvement in PC. Regardless, there are many clinical implications for the findings discussed here. The results of this study suggest that the identified lncRNAs could be further utilized to assess the metastatic potential of PC, as well as aid in drug development, since these lncRNAs can be used as drug targets. Since their involvement allowed the prediction and distinction between TNM stages, further investigation of their functions seems crucial.

Despite the significant findings, this study is not without limitations. First, DEGA was performed for a large number of data, which likely raised data noise. Second, TWAs used as controls were low in number, as most samples had a stage IIb diagnosis, and SMOTE was necessary to utilize for the ML algorithms to reduce bias. Third, there was a lack of normal tissue control samples, which makes it difficult to provide more accurate assessments of the nature of these genes. Last, there might have been biases in the TCGA data from incorrect measurements or sequencing, potentially skewing the results of the RNA-seq data. All of these findings indicate that the findings of this study should be further validated and interpreted with caution.

Regardless, the presence of evidence regarding some of the identified novel lncRNAs indicates the strength of the rigorous methods used in this study. This further adds to the implications of the findings discussed here and the importance of future research to address these novel lncRNAs as potential markers of metastatic progression in PC.

5. Conclusion

DGEA utilized in this study identified a set of 38 novel lncRNAs that could contribute to metastatic progression in PC. GSEA was unable to provide sufficient information to further describe the functions of these lncRNA, due to the scarcity of available data relevant to the genes identified. Since different ML algorithms were able to predict metastatic PC with acceptable accuracy and the RFC model predicted PC with 76% accuracy based on the 38 lncRNA DEG panel, it is likely that these genes participate in the metastatic progression of PC, warranting further investigation.

The significance and importance of this study is represented by the identified novel lncRNA gene set. Metastatic PC lacks sufficient studies regarding the involvement of lncRNAs in tumor proliferation and progression, especially those that use ML algorithms with proven accuracy. This is the first study of its kind to use this methodology to reveal the discussed gene set in PC to distinguish between early-stage and advanced PC. Regardless, more studies are needed to identify the role these genes play in PC metastasis and other cancers.

Based on the findings of this study, I suggest further research to take place into the role of these genes. *In vitro* and *in vivo* experiments must be conducted to further elucidate the functions these genes may take part in. The accuracy of the ML algorithms to determine PC metastatic potential reveals that these genes could be added to diagnostic methods if their clinical manifestations are confirmed by future studies.

6. Data availability statement

All raw data acquired from TCGA, in addition to all analyses performed on said data and source code utilized to perform the analyses mentioned in the methodology, are available at the link https://github.com/hasanalsharoh/PanC.

7. References

1. Hu, J.X., Zhao, C.F., Chen, W.B., Liu, Q.C., Li, Q.W., Lin, Y.Y., and Gao, F. (2021). Pancreatic cancer: A review of epidemiology, trend, and risk factors. World J Gastroenterol *27*, 4298-4321. 10.3748/wjg.v27.i27.4298.

2. Partyka, O., Pajewska, M., Kwaśniewska, D., Czerw, A., Deptała, A., Budzik, M., Cipora, E., Gąska, I., Gazdowicz, L., Mielnik, A., et al. (2023). Overview of Pancreatic Cancer Epidemiology in Europe and Recommendations for Screening in High-Risk Populations. Cancers *15*. 10.3390/cancers15143634.

3. Andersson, R., Haglund, C., Seppänen, H., and Ansari, D. (2022). Pancreatic cancer – the past, the present, and the future. Scandinavian Journal of Gastroenterology *57*, 1169-1177. 10.1080/00365521.2022.2067786.

4. Bin, W., Yuan, C., Qie, Y., and Dang, S. (2023). Long non-coding RNAs and pancreatic cancer: A multifaceted view. Biomedicine & Pharmacotherapy *167*, 115601. https://doi.org/10.1016/j.biopha.2023.115601.

5. Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227. 10.1038/nature07672.

6. Kore, H., Datta, K.K., Nagaraj, S.H., and Gowda, H. (2023). Protein-coding potential of non-canonical open reading frames in human transcriptome. Biochem Biophys Res Commun *684*, 149040. 10.1016/j.bbrc.2023.09.068.

7. Aswathy, R., and Sumathi, S. (2023). Defining new biomarkers for overcoming therapeutical resistance in cervical cancer using lncRNA. Mol Biol Rep. 10.1007/s11033-023-08864-w.

8. Zhang, N., Yu, X., Sun, H., Zhao, Y., Wu, J., and Liu, G. (2023). A prognostic and immunotherapy effectiveness model for pancreatic adenocarcinoma based on cuproptosis-related lncRNAs signature. Medicine (Baltimore) *102*, e35167. 10.1097/md.0000000000035167.

9. Wang, T., Ji, M., Liu, W., and Sun, J. (2023). Development and validation of a novel DNA damage repair-related long non-coding RNA signature in predicting prognosis, immunity, and drug sensitivity in uterine corpus endometrial carcinoma. Comput Struct Biotechnol J *21*, 4944-4959. 10.1016/j.csbj.2023.10.025.

10. Zhao, Y., Song, Y., Zhang, Y., Ji, M., Hou, P., and Sui, F. (2023). Screening protective miRNAs and constructing novel lncRNAs/miRNAs/mRNAs networks and prognostic models for triple-negative breast cancer. Mol Cell Probes *72*, 101940. 10.1016/j.mcp.2023.101940.

11. Collins, G.S., Whittle, R., Bullock, G.S., Logullo, P., Dhiman, P., de Beyer, J.A., Riley, R.D., and Schlussel, M.M. (2023). OPEN SCIENCE PRACTICES NEED SUBSTANTIAL IMPROVEMENT IN PROGNOSTIC MODEL STUDIES IN ONCOLOGY USING MACHINE LEARNING. J Clin Epidemiol. 10.1016/j.jclinepi.2023.10.015.

12. Rasti, P., Wolf, C., Dorez, H., Sablong, R., Moussata, D., Samiei, S., and Rousseau, D. (2019). Machine Learning-Based Classification of the Health State of Mice Colon in Cancer Study from Confocal Laser Endomicroscopy. Sci Rep *9*, 20010. 10.1038/s41598-019-56583-9.

13. Sharma, A.N., Shwe, S., and Mesinkovska, N.A. (2022). Current state of machine learning for non-melanoma skin cancer. Arch Dermatol Res *314*, 325-327. 10.1007/s00403-021-02236-9.

14. Anaconda (2016). (Anaconda Software Distribution).

15. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al. (2016). Jupyter Notebooks – a publishing format for reproducible

419 computational workflows. In Positioning and Power in Academic Publishing: Players, Agents and
420 Agendas, (IOS Press), pp. 87-90. 10.3233/978-1-61499-649-1-87.

421 16. glob — Unix style pathname pattern expansion. (2023).

422 17. team, T.p.d. (2023). pandas-dev/pandas: Pandas.

423 18. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser,
424 E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357-362.
425 10.1038/s41586-020-2649-2.

426 19. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E.,
427 Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific
428 computing in Python. Nature Methods *17*, 261-272. 10.1038/s41592-019-0686-2.

429 20. Boris, M., Maria, T., Vincent, C., and Mathieu, A. (2022). PyDESeq2: a python package for bulk
430 RNA-seq differential expression analysis. bioRxiv, 2022.2012.2014.520412. 10.1101/2022.12.14.520412.

431 21. Hunter, J.D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering
432 *9*, 90-95. 10.1109/MCSE.2007.55.

433 22. Waskom, M.L. (2021). seaborn: statistical data visualization. Journal of Open Source Software *6*,
434 3021. 10.21105/joss.03021.

435 23. Sanborn, M. (2023). sanbomics.

436 24. Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with
437 python. In 61. (Austin, TX), pp. 10-25080.

438 25. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and
439 dispersion for RNA-seq data with DESeq2. Genome biology *15*, 1-21.

26. Zhao, Y., Li, M.-C., Konaté, M.M., Chen, L., Das, B., Karlovich, C., Williams, P.M., Evrard, Y.A., Doroshow, J.H., and McShane, L.M. (2021). TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository. Journal of Translational Medicine *19*, 269. 10.1186/s12967-021-02936-w.

27. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545-15550. 10.1073/pnas.0506580102.

28. Fang, Z., Liu, X., and Peltz, G. (2023). GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics *39*, btac757. 10.1093/bioinformatics/btac757.

29. Chou, C.H., Shrestha, S., Yang, C.D., Chang, N.W., Lin, Y.L., Liao, K.W., Huang, W.C., Sun, T.H., Tu, S.J., Lee, W.H., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res *46*, D296-d302. 10.1093/nar/gkx1067.

30. The Gene Ontology, C., Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., et al. (2023). The Gene Ontology knowledgebase in 2023. Genetics *224*, iyad031. 10.1093/genetics/iyad031.

31. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics *25*, 25-29. 10.1038/75556.

32. Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., the Ami, G.O.H., and the Web Presence Working, G. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics *25*, 288-289. 10.1093/bioinformatics/btn615.

462 33. Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority
463 over-Sampling Technique. J. Artif. Int. Res. *16*, 321–357 , numpages = 337.

464 34. Garg, S., and Raghavan, B. (2023). Comparison of machine learning algorithms for the
465 classification of spinal cord tumor. Irish Journal of Medical Science (1971 -). 10.1007/s11845-023-03487-
466 3.

467 35. Bruno, V., Betti, M., D'Ambrosio, L., Massacci, A., Chiofalo, B., Pietropolli, A., Piaggio, G.,
468 Ciliberto, G., Nisticò, P., Pallocca, M., et al. (2023). Machine learning endometrial cancer risk prediction
469 model: integrating guidelines of European Society for Medical Oncology with the tumor immune
470 framework. Int J Gynecol Cancer. 10.1136/ijgc-2023-004671.

471 36. Gutman, R., Aronson, D., Caspi, O., and Shalit, U. (2023). What drives performance in machine
472 learning models for predicting heart failure outcome? Eur Heart J Digit Health *4*, 175-187.
473 10.1093/ehjdh/ztac054.

474 37. Wall, N.R., Fuller, R.N., Morcos, A., and De Leon, M. (2023). Pancreatic Cancer Health Disparity:
475 Pharmacologic Anthropology. Cancers (Basel) *15*. 10.3390/cancers15205070.

476 38. de Jesus, V.H.F., Mathias-Machado, M.C., de Farias, J.P.F., Aruquipa, M.P.S., Jácome, A.A., and
477 Peixoto, R.D. (2023). Targeting KRAS in Pancreatic Ductal Adenocarcinoma: The Long Road to Cure.
478 Cancers (Basel) *15*. 10.3390/cancers15205015.

479 39. Sun, Y., Yao, L., Man, C., Gao, Z., He, R., and Fan, Y. (2023). Development and validation of
480 cuproptosis-related lncRNAs associated with pancreatic cancer immune microenvironment based on
481 single-cell. Front Immunol *14*, 1220760. 10.3389/fimmu.2023.1220760.

482 40. Wang, H., Ding, Y., He, Y., Yu, Z., Zhou, Y., Gong, A., and Xu, M. (2023). LncRNA UCA1 promotes
483 pancreatic cancer cell migration by regulating mitochondrial dynamics via the MAPK pathway. Arch
484 Biochem Biophys *748*, 109783. 10.1016/j.abb.2023.109783.

485 41. Zhang, R., Wang, X., Ying, X., Huang, Y., Zhai, S., Shi, M., Tang, X., Liu, J., Shi, Y., Li, F., et al.
486 (2023). Hypoxia-induced long non-coding RNA LINC00460 promotes p53 mediated proliferation and
487 metastasis of pancreatic cancer by regulating the miR-4689/UBE2V1 axis and sequestering USP10. Int J
488 Med Sci *20*, 1339-1357. 10.7150/ijms.87833.

489 42. Tsai, C.W., Chang, W.S., Yueh, T.C., Wang, Y.C., Chin, Y.T., Yang, M.D., Hung, Y.C., Mong, M.C.,
490 Yang, Y.C., Gu, J., and Bau, D.T. (2023). The Significant Impacts of Interleukin-8 Genotypes on the Risk of
491 Colorectal Cancer in Taiwan. Cancers (Basel) *15*. 10.3390/cancers15204921.

492 43. Earnest, A., Tesema, G.A., and Stirling, R.G. (2023). Machine Learning Techniques to Predict
493 Timeliness of Care among Lung Cancer Patients. Healthcare (Basel) *11*. 10.3390/healthcare11202756.

494 44. Padwal, M.K., Basu, S., and Basu, B. (2023). Application of Machine Learning in Predicting
495 Hepatic Metastasis or Primary Site in Gastroenteropancreatic Neuroendocrine Tumors. Current
496 Oncology *30*, 9244-9261. 10.3390/curroncol30100668.

497 45. Chiang, C.P., Wu, C.W., Lee, S.P., Chung, C.C., Wang, C.W., Lee, S.L., Nieh, S., and Yin, S.J. (2009).
498 Expression pattern, ethanol-metabolizing activities, and cellular localization of alcohol and aldehyde
499 dehydrogenases in human pancreas: implications for pathogenesis of alcohol-induced pancreatic injury.
500 Alcohol Clin Exp Res *33*, 1059-1068. 10.1111/j.1530-0277.2009.00927.x.

501 46. Kanellopoulos, P., Nock, B.A., Krenning, E.P., and Maina, T. (2020). Optimizing the Profile of
502 [(99m)Tc]Tc-NT(7-13) Tracers in Pancreatic Cancer Models by Means of Protease Inhibitors. Int J Mol Sci
503 *21*. 10.3390/ijms21217926.

504 47. de Koning, P.J., Bovenschen, N., Leusink, F.K., Broekhuizen, R., Quadir, R., van Gemert, J.T.,
505 Hordijk, G.J., Chang, W.S., van der Tweel, I., Tilanus, M.G., and Kummer, J.A. (2009). Downregulation of
506 SERPINB13 expression in head and neck squamous cell carcinomas associates with poor clinical
507 outcome. Int J Cancer *125*, 1542-1550. 10.1002/ijc.24507.

508 48. Xu, Y., Yuan, C., Peng, J., Zhou, L., Lin, Y., Wang, Y., Zhang, J., Ma, J., Yin, W., and Lu, J. (2022).
509 LncRNA MIR205HG expression predicts efficacy of neoadjuvant chemotherapy for patients with locally
510 advanced breast cancer. Genes Dis *9*, 837-840. 10.1016/j.gendis.2021.10.001.

511 49. He, W.A., Berardi, E., Cardillo, V.M., Acharyya, S., Aulino, P., Thomas-Ahner, J., Wang, J.,
512 Bloomston, M., Muscarella, P., Nau, P., et al. (2013). NF-κB-mediated Pax7 dysregulation in the muscle
513 microenvironment promotes cancer cachexia. J Clin Invest *123*, 4821-4835. 10.1172/jci68523.

514 50. Zhao, X., Lu, M., Liu, Z., Zhang, M., Yuan, H., Dan, Z., Wang, D., Ma, B., Yang, Y., Yang, F., et al.
515 (2022). Comprehensive analysis of alfa defensin expression and prognosis in human colorectal cancer.
516 Front Oncol *12*, 974654. 10.3389/fonc.2022.974654.

517 51. Sun, N.K., Huang, S.L., Lu, H.P., Chang, T.C., and Chao, C.C. (2015). Integrative transcriptomics-
518 based identification of cryptic drivers of taxol-resistance genes in ovarian carcinoma cells: Analysis of the
519 androgen receptor. Oncotarget *6*, 27065-27082. 10.18632/oncotarget.4824.

520 52. Duan, W., Kong, X., Li, J., Li, P., Zhao, Y., Liu, T., Binang, H.B., Wang, Y., Du, L., and Wang, C.
521 (2020). LncRNA AC010789.1 Promotes Colorectal Cancer Progression by Targeting MicroRNA-432-
522 3p/ZEB1 Axis and the Wnt/β-Catenin Signaling Pathway. Front Cell Dev Biol *8*, 565355.
523 10.3389/fcell.2020.565355.

524 53. Liu, Y., Xu, G., and Li, L. (2021). LncRNA GATA3-AS1-miR-30b-5p-Tex10 axis modulates
525 tumorigenesis in pancreatic cancer. Oncol Rep *45*. 10.3892/or.2021.8010.

526 54. Chen, K., Wang, Q., Liu, X., Wang, F., Yang, Y., and Tian, X. (2022). Hypoxic pancreatic cancer
527 derived exosomal miR-30b-5p promotes tumor angiogenesis by inhibiting GJA1 expression. Int J Biol Sci
528 *18*, 1220-1237. 10.7150/ijbs.67675.

529 55. Liu, S., Luan, J., and Ding, Y. (2018). miR-144-3p Targets FosB Proto-oncogene, AP-1
530 Transcription Factor Subunit (FOSB) to Suppress Proliferation, Migration, and Invasion of PANC-1
531 Pancreatic Cancer Cells. Oncol Res *26*, 683-690. 10.3727/096504017x14982585511252.

532 56. Yang, J., Cong, X., Ren, M., Sun, H., Liu, T., Chen, G., Wang, Q., Li, Z., Yu, S., and Yang, Q. (2019).
533 Circular RNA hsa_circRNA_0007334 is Predicted to Promote MMP7 and COL1A1 Expression by
534 Functioning as a miRNA Sponge in Pancreatic Ductal Adenocarcinoma. J Oncol *2019*, 7630894.
535 10.1155/2019/7630894.

536 57. Janpipatkul, K., Trachu, N., Watcharenwong, P., Panvongsa, W., Worakitchanon, W.,
537 Metheetrairut, C., Oranratnachai, S., Reungwetwattana, T., and Chairoungdua, A. (2021). Exosomal
538 microRNAs as potential biomarkers for osimertinib resistance of non-small cell lung cancer patients.
539 Cancer Biomark *31*, 281-294. 10.3233/cbm-203075.

540 58. Daniel, R., Wu, Q., Williams, V., Clark, G., Guruli, G., and Zehner, Z. (2017). A Panel of MicroRNAs
541 as Diagnostic Biomarkers for the Identification of Prostate Cancer. Int J Mol Sci *18*.
542 10.3390/ijms18061281.

543 59. Wang, K., Zhong, W., Long, Z., Guo, Y., Zhong, C., Yang, T., Wang, S., Lai, H., Lu, J., Zheng, P., and
544 Mao, X. (2021). 5-Methylcytosine RNA Methyltransferases-Related Long Non-coding RNA to Develop
545 and Validate Biochemical Recurrence Signature in Prostate Cancer. Front Mol Biosci *8*, 775304.
546 10.3389/fmolb.2021.775304.

547 60. Li, R., Gao, X., Sun, H., Sun, L., and Hu, X. (2022). Expression characteristics of long non-coding
548 RNA in colon adenocarcinoma and its potential value for judging the survival and prognosis of patients:
549 bioinformatics analysis based on The Cancer Genome Atlas database. J Gastrointest Oncol *13*, 1178-
550 1187. 10.21037/jgo-22-384.

551 61. Zeng, X., Wang, Y., Liu, B., Rao, X., Cao, C., Peng, F., Zhi, W., Wu, P., Peng, T., Wei, Y., et al.
552 (2023). Multi-omics data reveals novel impacts of human papillomavirus integration on the epigenomic
553 and transcriptomic signatures of cervical tumorigenesis. J Med Virol *95*, e28789. 10.1002/jmv.28789.

554 62. Wang, W.F., Zhong, H.J., Cheng, S., Fu, D., Zhao, Y., Cai, H.M., Xiong, J., and Zhao, W.L. (2023). A
555 nuclear NKRF interacting long noncoding RNA controls EBV eradication and suppresses tumor
556 progression in natural killer/T-cell lymphoma. Biochim Biophys Acta Mol Basis Dis *1869*, 166722.
557 10.1016/j.bbadis.2023.166722.

558 63. Bi, X.-a., Li, L., Xu, R., and Xing, Z. (2021). Pathogenic Factors Identification of Brain Imaging and
559 Gene in Late Mild Cognitive Impairment. Interdisciplinary Sciences: Computational Life Sciences *13*, 511-
560 520. 10.1007/s12539-021-00449-0.

561 64. Gusev, F.E., Reshetov, D.A., Mitchell, A.C., Andreeva, T.V., Dincer, A., Grigorenko, A.P., Fedonin,
562 G., Halene, T., Aliseychik, M., Filippova, E., et al. (2019). Chromatin profiling of cortical neurons identifies
563 individual epigenetic signatures in schizophrenia. Transl Psychiatry *9*, 256. 10.1038/s41398-019-0596-1.

564 65. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y., and Kolpakov, F. (2019). GTRD: a database
565 on gene transcription regulation-2019 update. Nucleic Acids Res *47*, D100-d105. 10.1093/nar/gky1128.

566 66. The UniProt, C. (2023). UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids
567 Research *51*, D523-D531. 10.1093/nar/gkac1052.

568