Identification of IGFBP3 and LGALS1 as potential secreted biomarkers for clear cell renal cell carcinoma based on bioinformatics analysis and machine learning

Wunchana Seubwai[1,2,A–F], Sakkarn Sangkhamanon[3,A,E,F], Xuhong Zhang[4,A,E,F]

[1] Department of Forensic Medicine, Faculty of Medicine, Khon Kaen University, Thailand
[2] Center for Translational Medicine, Faculty of Medicine, Khon Kaen University, Thailand
[3] Department of Pathology, Faculty of Medicine, Khon Kaen University, Thailand
[4] Luddy School of Informatics, Computing and Engineering, Indiana University, Bloomington, USA

A – research concept and design; B – collection and/or assembly of data; C – data analysis and interpretation; D – writing the article; E – critical revision of the article; F – final approval of the article

Advances in Clinical and Experimental Medicine, ISSN 1899–5276 (print), ISSN 2451–2680 (online) *Adv Clin Exp Med*. 2025;34(10):1649–1659

Address for correspondence
Wunchana Seubwai
E-mail: wunchanas@yahoo.com

Funding sources
The study was supported by a grant from the Fundamental Fund of Khon Kaen University, under the National Science, Research and Innovation Fund (NSRF), Thailand (grant No. 179490).

Conflict of interest
None declared

Acknowledgements
We would like to acknowledge Dr. Dylan Southard for editing the manuscript via the KKU Publication Clinic, Khon Kaen University, Thailand.

Received on June 27, 2024
Reviewed on September 22, 2024
Accepted on October 2, 2024

Published online on January 14, 2025

Cite as
Seubwai W, Sangkhamanon S, Zhang X. Identification of IGFBP3 and LGALS1 as potential secreted biomarkers for clear cell renal cell carcinoma based on bioinformatics analysis and machine learning. *Adv Clin Exp Med*. 2025;34(10):1649–1659. doi:10.17219/acem/194036

DOI
10.17219/acem/194036

Copyright
Copyright by Author(s)
This is an article distributed under the terms of the Creative Commons Attribution 3.0 Unported (CC BY 3.0) (https://creativecommons.org/licenses/by/3.0/)

Abstract

Background. Clear cell renal cell carcinoma (ccRCC) is the most common subtype of renal cell carcinoma (RCC). Due to the lack of symptoms until advanced stages, early diagnosis of ccRCC is challenging. Therefore, the identification of novel secreted biomarkers for the early detection of ccRCC is urgently needed.

Objectives. This study aimed to identify novel secreted biomarkers for diagnosing ccRCC using bioinformatics and machine learning techniques based on transcriptomics data.

Material and methods. Differentially expressed genes (DEGs) in ccRCC compared to normal kidney tissues were identified using 3 transcriptomics datasets (GSE53757, GSE40435 and GSE11151) from the Gene Expression Omnibus (GEO). Potential secreted biomarkers were examined within these common DEGs using a list of human secretome proteins from The Human Protein Atlas. The recursive feature elimination (RFE) technique was used to determine the optimal number of features for building classification machine learning models. The expression levels and clinical associations of candidate biomarkers identified with RFE were validated using transcriptomics data from The Cancer Genome Atlas (TCGA). Classification models were then developed based on the expression levels of these candidate biomarkers. The performance of the models was evaluated based on accuracy, evaluation metrics, confusion matrices, and ROC-AUC (receiver operating characteristic-area under the ROC curve) curves.

Results. We identified 44 DEGs that encode potential secreted proteins from 274 common DEGs found across all datasets. Among these, insulin-like growth factor binding protein 3 (IGFBP3) and lectin, galactoside-binding, soluble, 1 (LGALS1) were selected for further analysis using the RFE technique. Both IGFBP3 and LGALS1 showed significant upregulation in ccRCC tissues compared to normal tissues in the GEO and TCGA datasets. The results of the survival analysis indicated that patients with higher expression levels of these genes exhibited shorter overall and disease-free survival times (OS and DFS). Decision tree and random forest models based on IGFBP3 and LGALS1 levels achieved an accuracy of 98.04% and an AUC of 0.98.

Conclusions. This study identified IGFBP3 and LGALS1 as promising novel secreted biomarkers for ccRCC diagnosis.

Key words: TCGA, GEO, machine learning, clear cell renal cell carcinoma, bioinformatics

Background

Renal cell carcinoma (RCC) represents the most common type of kidney cancer, accounting for approx. 90% of all cases.[1] Globally, RCC is the 14th most commonly diagnosed malignancy, with over 400,000 new cases reported annually.[2] Smoking, alcohol consumption, obesity, and high blood pressure are associated risk factors for RCC.[3] Renal cell carcinoma is often asymptomatic in its early stages, with 60% of cases being discovered incidentally during imaging studies for unrelated conditions. When symptomatic, patients may present with a triad of flank pain, hematuria and an abdominal mass, although this classic presentation is relatively uncommon. Systemic symptoms, including fever, weight loss and paraneoplastic syndromes, may result from advanced disease.[4,5] The treatment of RCC has undergone significant evolution over the past few decades. Surgical resection is the standard of care for patients with localized RCC, while targeted therapies and immunotherapy have been promising treatment options for advanced and metastatic RCC.[6,7]

There are 3 common pathological RCC subtypes, including clear cell RCC (ccRCC), which makes up 70–80% of cases; papillary RCC, which comprises 10–15%; and chromophobe RCC, which accounts for 5%.[8] Clear cell RCC is the primary cause of death in kidney cancer patients due to its asymptomatic nature in the early stages and resistance to chemotherapy and radiotherapy.[9] Early detection of ccRCC is challenging, relying on a combination of imaging techniques and histological examination. Therefore, identifying novel secreted biomarkers is crucial for its effective diagnosis.

Transcriptomics data, encompassing the complete set of all RNAs transcribed by specific tissues or cells, are widely used to identify novel biomarkers and promising drug targets in many diseases, including cancers.[10] Public transcriptomics databases such as The Cancer Genome Atlas (TCGA)[11,12] and the Gene Expression Omnibus (GEO)[13,14] have become invaluable resources for researchers in this field. A combination of bioinformatics and machine learning approaches to analyze public transcriptomics data has emerged as a pivotal approach to cancer research, offering unprecedented opportunities to identify novel biomarkers and potential drug targets for various cancers,[15,16] such as colorectal cancer,[17] pancreatic cancer[18] and breast cancer.[19] However, our latest review found no reports identifying potential secreted biomarkers for ccRCC using bioinformatics and machine learning approaches on public transcriptomics datasets.

In this study, bioinformatics and machine learning analysis were used to identify novel secreted biomarkers for ccRCC diagnosis using transcriptomics datasets from the GEO and TCGA databases. Differentially expressed genes (DEGs) were identified by comparing ccRCC tissues with normal kidney tissues, and potentially secreted proteins among the common DEGs were further analyzed. The optimal number of features for building machine learning models was determined using the recursive feature elimination (RFE) technique. Subsequently, classification models were developed based on the expression levels of candidate-secreted biomarkers. Finally, the expression levels and clinical associations of these candidate biomarkers were validated using additional transcriptomic data from the TCGA database.

Objectives

This research aimed to discover novel secreted biomarkers for diagnosing ccRCC by integrating bioinformatics and machine learning techniques with public transcriptomics data.

Materials and methods

Transcriptomics datasets

Three microarray datasets of ccRCC and normal kidney tissues, including GSE11151, GSE40435 and GSE53757, were obtained from the GEO database (https://www.ncbi.nlm.nih.gov/geo). The datasets GSE11151 and GSE53757 were generated using the Affymetrix Human Genome U133 Plus 2.0 Array platform (Thermo Fisher Scientific, Waltham, USA), which was utilized for transcriptional profiling, while GSE40435 was based on the Illumina HumanHT-12 V4.0 expression BeadChip platform (Illumina Inc., San Diego, USA).

Differentially expressed gene analysis

The DEGs were identified by comparing ccRCC and normal kidney tissues using the GEO2R (https://www.ncbi.nlm.nih.gov/geo/geo2r) with an adjusted $p < 0.05$ and absolute log fold-changes ≥1.0 as criteria. GEO2R is a web-based tool provided by the Gene Expression Omnibus (GEO) for analyzing gene expression data. It allows researchers to compare 2 or more groups of samples to identify differentially expressed genes. Data visualization was performed using a volcano plot in RStudio (https://rstudio.com). Additionally, Venn diagrams (http://bioinformatics.psb.ugent.be/webtools/Venn) were generated to display common DEGs across the 3 transcriptomics datasets.

Gene expression analysis

Gene expression analysis was conducted on the GSE40435 dataset retrieved from the GEO database using the GEOquery package in R (R Foundation for Statistical Computing, Vienna, Austria).[20] Expression data were transformed using a base-2 logarithmic scale to normalize the distribution.

Identification of potential secreted biomarkers in common DEGs

Potential secreted biomarkers in ccRCC were identified based on overlapping genes between common DEGs and a list of 1,665 secreted proteins from The Human Protein Atlas (https://www.proteinatlas.org).[21,22]

Feature selection

The recursive feature elimination, based on the random forest classifier, was employed to select the minimal set of genes needed to create classification models. The feature-selection process was conducted using the scikit-learn library (https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html). Feature subsets of sizes 44, 20, 10, 5, 3, 2, and 1 were used for training and evaluation. The performance of the random forest classifier for each gene subset was assessed using several evaluation metrics, including accuracy, precision, recall, and F1-score.

Machine learning for classification

Seven supervised machine learning algorithms, including decision trees, random forests, logistic regression, K-nearest neighbors, Gaussian naive Bayes (GNB), support vector machines, and multilayer perceptrons (MLPs), were used to develop classification models based on the selected potential secreted DEGs. The Python scikit-learn library (https://scikit-learn.org) was used to implement these algorithms. The transcriptomics data were split into training and test sets. The training set was used to develop models with 7 machine learning algorithms, and their performance was evaluated on the test set. GridSearchCV (https://scikit-learn.org/1.5/modules/generated/sklearn.model_selection.GridSearchCV.html) was employed to optimize hyperparameters for each model. The classification performance of each model was assessed using accuracy, precision, recall, F1-score, confusion matrix, and receiver operating characteristic (ROC) curves.

Validation of biomarkers gene expression and clinical association

The expression levels of potential biomarkers in ccRCC and normal kidney tissue were validated using the TCGA dataset, which includes 523 ccRCC samples and 100 normal kidney samples, employing Gene Expression Profiling Interactive Analysis (GEPIA) (http://gepia.cancer-pku.cn).[23,24] Additionally, the correlation between the expression levels of potential biomarkers and the survival of ccRCC patients was analyzed using GEPIA.

Statistical analyses

In the GEPIA, the significant difference between the 2 groups was compared using Student's t-test. The correlation between gene expression and both overall survival (OS) and disease-free survival (DFS) in ccRCC patients was evaluated using Kaplan–Meier analysis, accompanied by a log-rank test and hazard ratio (HR) calculation. Statistical significance was considered to be $p < 0.05$. In the box plots, the central line represents the median value of the data. The boxes extended from the 1st quartile (Q1) to the 3rd quartile (Q3), representing the interquartile range (IQR). The whiskers extended to the most extreme data points within 1.5 times the IQR from Q1 and Q3. Data points beyond this range were considered outliers.

Results

Identification of common DEGs in ccRCC

Differentially expressed genes were identified by comparing ccRCC and normal kidney tissues from 3 GEO datasets: GSE53757, GSE40435 and GSE11151. The selection criteria were an adjusted $p < 0.05$ and an absolute log fold change ≥1.0. Based on these criteria, 2,917 DEGs were identified in GSE11151, 1,521 in GSE40435 and 6,665 in GSE53757. Specifically, GSE11151 had 1,180 upregulated and 1,737 downregulated genes; GSE40435 had 680 upregulated and 841 downregulated genes; and GSE53757 had 3,124 upregulated and 3,541 downregulated genes (Fig. 1, Table 1).

Table 1. Number of differentially expressed genes (DEGs) in 3 ccRCC datasets

GEO accession	DEGs	Upregulated genes	Downregulated genes
GSE11151	2,917	1,180	1,737
GSE40435	1,521	680	841
GSE53757	6,665	3,124	3,541

ccRCC – clear cell renal cell carcinoma.

We further identified the common DEGs using a Venn diagram (Fig. 2). There were 274 common DEGs across all 3 datasets (GSE11151, GSE40435 and GSE53757). This identification of common DEGs could potentially help to identifying potential candidate biomarkers for ccRCC.

Identification of potential secreted DEGs in ccRCC

The list of human secretome proteins from The Human Protein Atlas, which includes 1,665 human-secreted proteins, was used to identify which DEGs potentially encode secreted proteins. A Venn diagram was used to determine the secreted proteins among the 274 common DEGs (Fig. 3). We identified 44 DEGs that potentially encode secreted proteins, including *ADM*, *ANGPT2*, *ANGPTL4*, *ANXA1*, *ANXA2*, *APOC1*, *C1QA*,

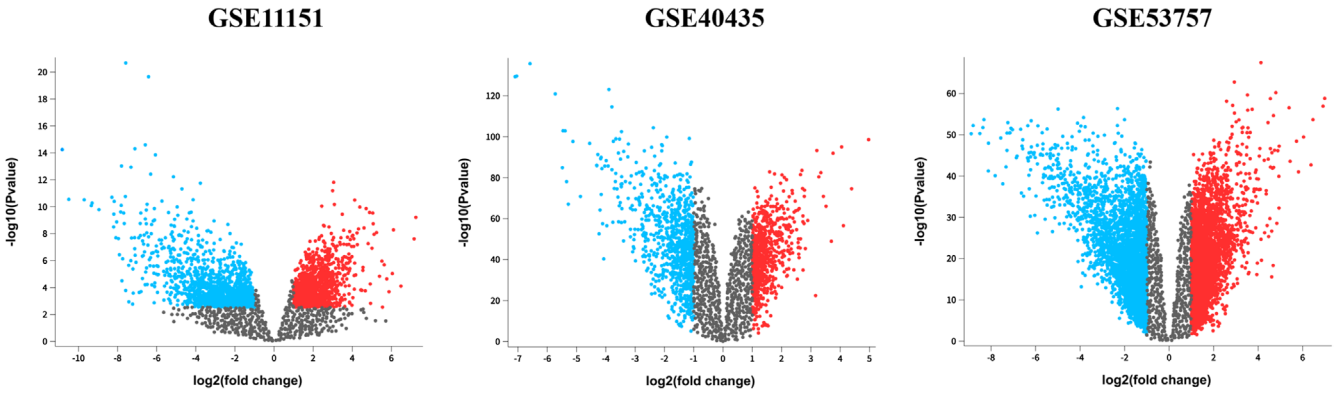

Fig. 1. Volcano plots demonstrating differentially expressed genes in 3 Gene Expression Omnibus (GEO) datasets (GSE11151, GSE40435 and GSE53757). The X-axis represents the log2 fold change, and the Y-axis represents the negative logarithm (base 10) of the p-value. Significantly upregulated genes are indicated in red, downregulated genes in blue, and nonsignificant genes in grey.

GSE – GEO accession number. A p-value < 0.05 was considered statistical significance.

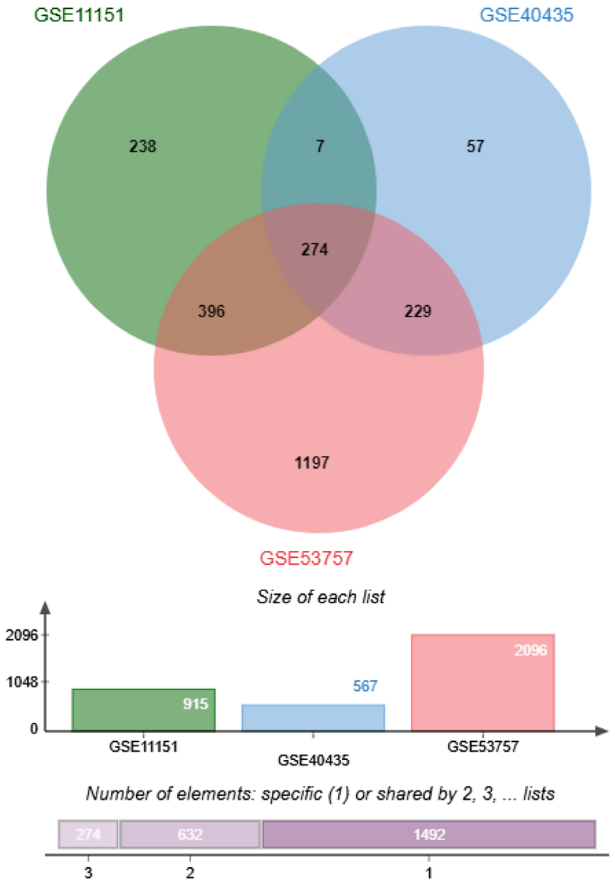

Fig. 2. Common DEGs in ccRCC across 3 Gene Expression Omnibus (GEO) datasets. The Venn diagram displays the overlap of DEGs among the 3 datasets. Numbers within the sections indicate the count of DEGs specific to 1 dataset or shared among multiple datasets. DEGs in GSE11151, GSE40435 and GSE53757 are represented in green, blue and red, respectively. The bar chart shows the total number of DEGs in each dataset

DEGs – differentially expressed genes; ccRCC – clear cell renal cell carcinoma; GSE – GEO accession number.

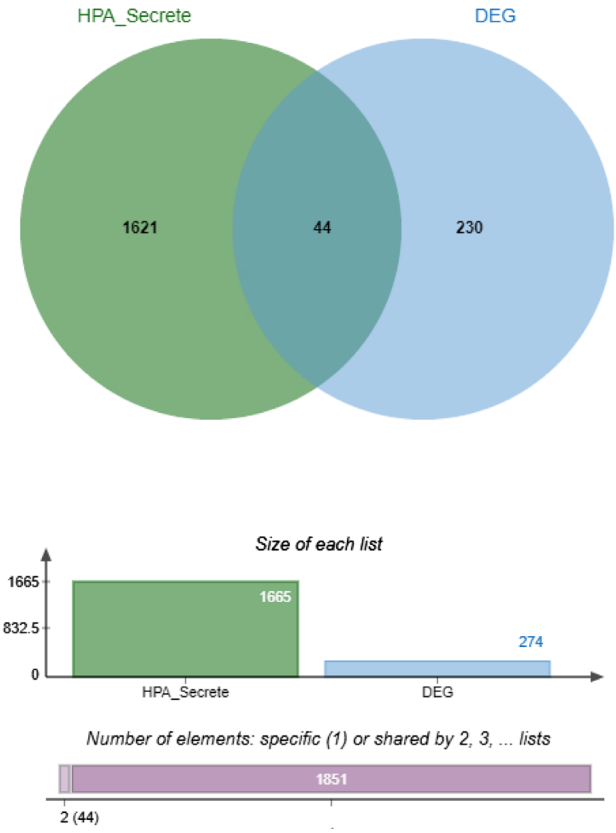

Fig. 3. Potentially secreted DEGs in ccRCC. The Venn diagram illustrates the overlap between the secreted gene list from The Human Protein Atlas (HPA_Secrete) and the common DEGs in ccRCC. The green circle represents the HPA_Secrete, with 1,665 genes. The blue circle represents the common DEGs in ccRCC, with 274 genes. The intersection of the 2 circles shows 44 potential secreted DEGs in ccRCC

DEGs – differentially expressed genes; ccRCC – clear cell renal cell carcinoma.

C1QC, *C3*, *CCL20*, *CD14*, *CHSY1*, *COL4A1*, *CTHRC1*, *CXCL10*, *CXCL9*, *EMILIN2*, *GNLY*, *GZMA*, *GZMH*, *IGFBP3*, *INHBB*, *ISG15*, *LAMA4*, *LAMC1*, *LGALS1*, *LOX*, *LY86*, *LY96*, *LYZ*, *NPTX2*, *OLFML2B*, *PLA2G7*, *PTHLH*, *RNASE6*, *RNASET2*, *SPARC*, *SRGN*, *STC2*, *TIMP1*, *TNFAIP6*, *VASH1*, *VEGFA*, and *VWF*. The overlap of the DEGs with the secretome database indicates that these 44 genes are strong candidates for secreted biomarkers for ccRCC.

Feature selection

The 44 DEGs that potentially encode secreted proteins in ccRCC were subjected to feature selection using the RFE technique based on the RandomForestClassifier. Various numbers of gene sets (1, 2, 3, 5, 10, 20, and 44 genes) were selected during the feature selection process. The RFE results indicated that sets of 3 and 5 genes resulted in the greatest accuracy (97.5%) (Table 2). However, we were able to achieve an accuracy of 96.3% using only 2 genes, namely insulin-like growth factor binding protein 3 (*IGFBP3*) and lectin, galactoside-binding, soluble, 1 (*LGALS1*), which we used to construct classification models.

The expression of IGFBP3 and LGALS1 in ccRCC patients based on the GEO dataset

The expression levels of *IGFBP3* and *LGALS1* were measured in adjacent non-tumor renal tissues and ccRCC tissues using the GSE40435 dataset. The pair plot revealed distinct clustering by tissue type, indicating a potential correlation between the expression levels of these genes and tissue classification (Fig. 4A). Additionally, the box plot showed that both *IGFBP3* and *LGALS1* expression levels were significantly elevated in ccRCC tissues compared to adjacent non-tumor renal tissues (Fig. 4B). These findings suggested that *IGFBP3* and *LGALS1* may serve as potential biomarkers for distinguishing ccRCC tissues from adjacent non-tumor renal tissues.

The performance of the classification model based on the expression levels of IGFBP3 and LGALS1

We evaluated the perfo rmance of 7 supervised machine learning algorithms, including Decision Trees, Random Forests, Logistic Regression, K-nearest Neighbors, Gaussian Naive Bayes, Support Vector Machines, and Multilayer Perceptrons, using the expression levels of *IGFBP3*

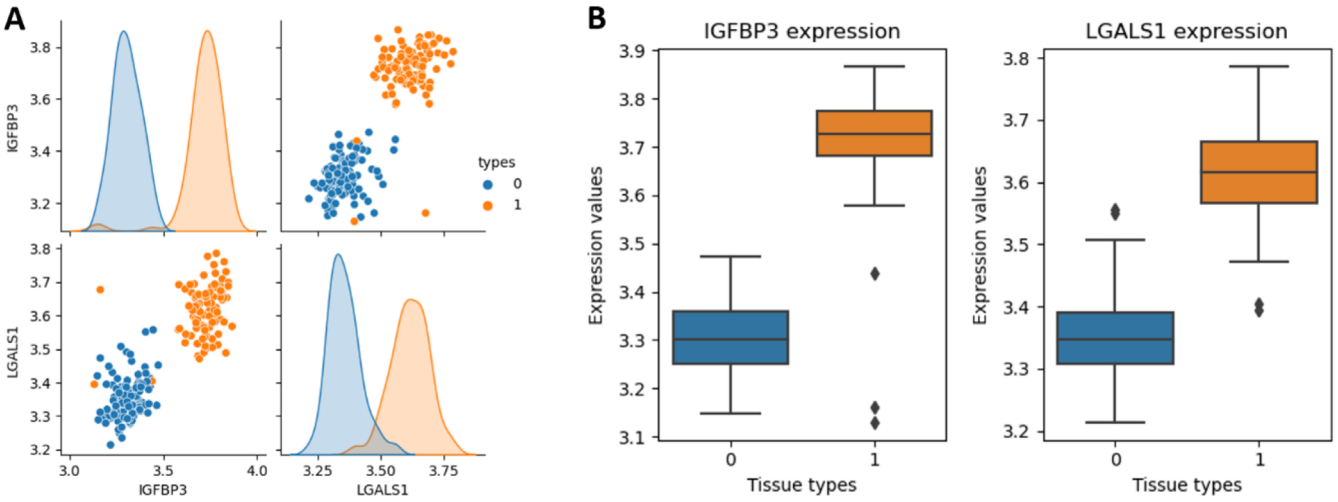

Fig. 4. Expression levels of IGFBP3 and LGALS1 in adjacent non-tumor renal tissues and ccRCC tissues. A. A pair plot displays the distribution and correlation of IGFBP3 and LGALS1 expression levels in adjacent non-tumor renal tissues (blue, Class 0) and ccRCC tissues (orange, Class 1); B. Box plots display the expression levels of IGFBP3 and LGALS1 between adjacent non-tumor renal tissues (Class 0) and ccRCC tissues (Class 1). In the box plots, the central line indicated the median, the box represented the interquartile range (IQR; Q1 to Q3) and the whiskers extended to 1.5 times the IQR from the quartiles. Outliers were plotted as individual points

IGFBP-3 – insulin-like growth factor binding protein 3; LGALS1 – lectin, galactoside-binding, soluble, 1; ccRCC – clear cell renal cell carcinoma.

Table 2. Performance of different feature sets in the feature selection process using RFE

Features	Accuracy [%]	Precision		Recall		F1-score	
		normal	ccRCC	normal	ccRCC	normal	ccRCC
44	96.3	0.96	0.97	0.98	0.94	0.97	0.96
20	96.3	0.96	0.97	0.98	0.94	0.97	0.96
10	96.3	0.96	0.97	0.98	0.94	0.97	0.96
5	97.5	0.96	1.00	1.00	0.94	0.98	0.97
3	97.5	0.96	1.00	1.00	0.94	0.98	0.97
2	96.3	0.96	0.97	0.98	0.94	0.97	0.96
1	91.4	0.93	0.89	0.91	0.91	0.92	0.90

RFE – recursive feature elimination; ccRCC – clear cell renal cell carcinoma.

Table 3. Classification performance of optimized decision tree and random forest models based on *IGFBP3* and *LGALS1*

Model	Accuracy [%]	AUC	Precision [%]		Recall [%]		F1-score [%]	
			normal	ccRCC	normal	ccRCC	normal	ccRCC
DT	98.04	0.98	0.96	1.00	1.00	0.96	0.98	0.98
RF	98.04	0.98	0.96	1.00	1.00	0.96	0.98	0.98

IGFBP3 – insulin-like growth factor binding protein 3; LGALS1 – lectin, galactoside-binding, soluble, 1; DT – decision tree; RF – random forest; AUC – area under the curve; ccRCC – clear cell renal cell carcinoma.

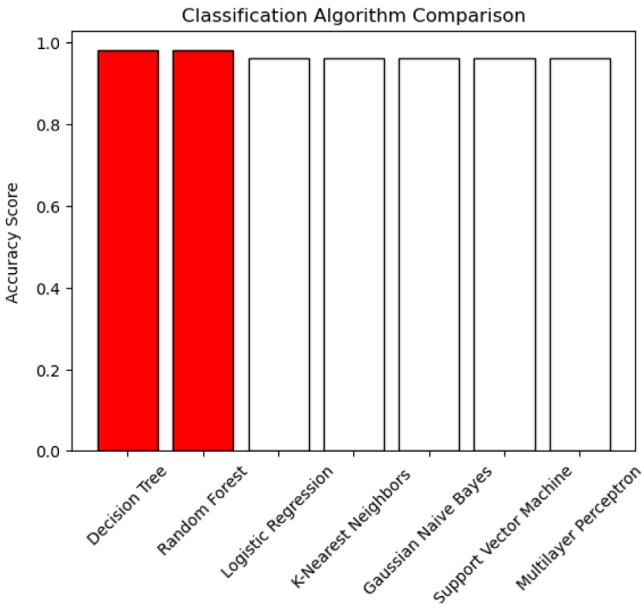

Fig. 5. Classification performance of 7 supervised machine learning algorithms. The bar chart demonstrates the accuracy scores of 7 supervised machine learning algorithms for classifying ccRCC. The decision tree and random forest algorithms achieved the highest accuracy scores, indicated by the red bars, while the other algorithms are represented by white bars

ccRCC – clear cell renal cell carcinoma.

and *LGALS1*. All algorithms demonstrated high accuracy, ranging from 96% to 98%. The Decision Tree and Random Forest models achieved the highest accuracy scores (Fig. 5). Consequently, these 2 models were selected for further optimization using GridSearchCV. After optimization, both models exhibited high performance, with an accuracy of 98.04% and an area under the ROC curve (AUC) of 0.98 (Table 3, Fig. 6).

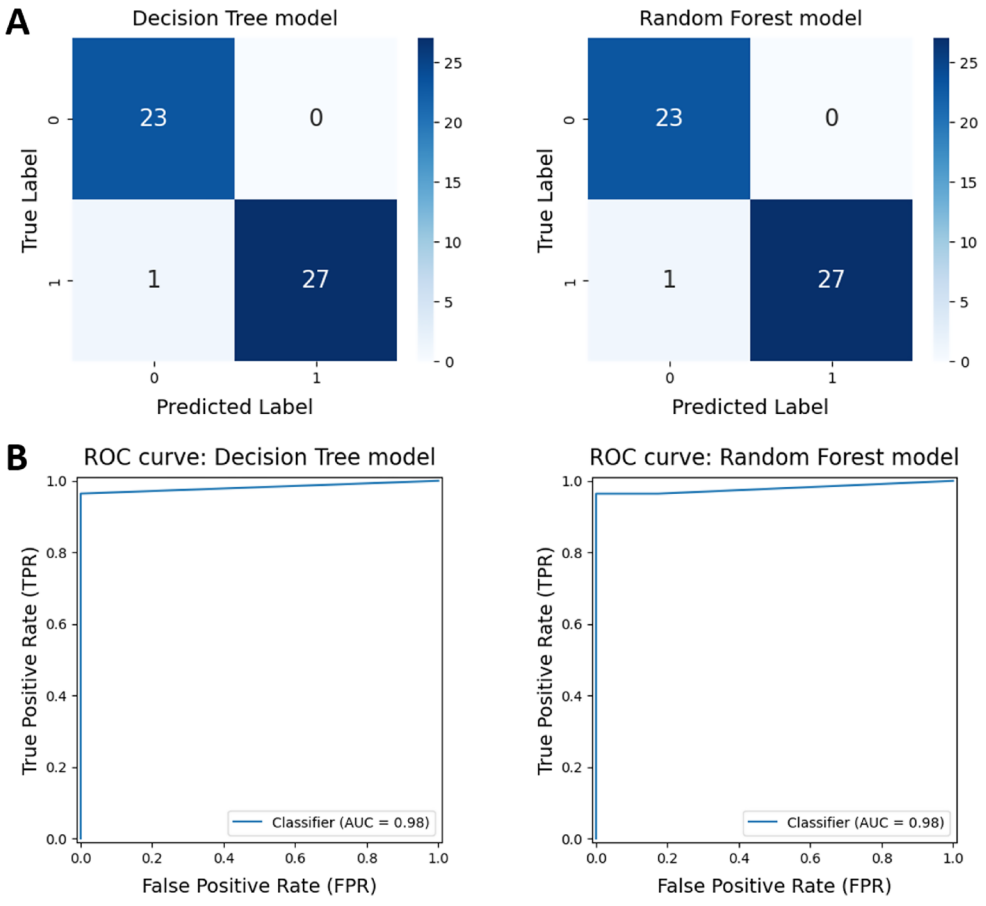

Fig. 6. Results from hyperparameter tuning of decision tree and random forest models. A. The confusion matrices show the performance of the decision tree and random forest models after hyperparameter tuning; B. The ROC curves for the decision tree and random forest models. Both models achieved a high AUC score of 0.98, indicating excellent performance in distinguishing between the adjacent non-tumor renal tissues (Class 0) and ccRCC tissues (Class 1)

ROC – receiver operating characteristic; AUC – area under the curve; ccRCC – clear cell renal cell carcinoma.

Validation of potential secreted biomarkers expression and clinical association based on the TCGA dataset

The expression levels of *IGFBP3* and *LGALS1* were confirmed using the TCGA dataset using the GEPIA online tool. Both *IGFBP3* and *LGALS1* were differentially expressed in kidney renal clear cell carcinoma (KIRC) samples, showing significantly higher expression compared to normal kidney tissue (Fig. 7).

The results of the survival analysis based on the TCGA dataset indicated that high IGFBP3 expression levels were associated with significantly reduced OS and DFS in ccRCC patients. Similarly, we found that high expression of *LGALS1* correlated with a trend of decreased OS and significantly affected DFS in ccRCC patients (Fig. 8A,B).

Discussion

Our bioinformatics analyses of 3 ccRCC datasets from the GEO identified 274 common DEGs. We then used the list of secreted proteins from the Human Protein Atlas to identify 44 potential secreted biomarkers for ccRCC.

The RFE technique, based on the RandomForestClassifier, highlighted a smaller subset of 2 genes that provided high classification accuracy, including *IGFBP3* and *LGALS1*. Decision Tree and Random Forest models based on the expression levels of *IGFBP3* and *LGALS1* demonstrated particularly high classification accuracy, underscoring their potential in diagnosing ccRCC patients.

Currently, several potential secreted biomarkers for diagnosing ccRCC have been identified. Carbonic anhydrase IX (CA9) is considered one of the promising biomarkers for ccRCC. Serum levels of CA9 were significantly higher in ccRCC patients than in those with non-CCRCC and benign tumours.[25] A similar finding was reported in 2018, in which plasma CA9 was evaluated in patients with ccRCC compared with patients with benign tumors and healthy controls.[26] However, the diagnostic performance of secreted CA9 in ccRCC remains unclear. Yang et al. identified 3 potential serum biomarkers for ccRCC using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. These biomarkers demonstrated a mean sensitivity of 88.38% and a mean specificity of 91.67%.[27] In 2017, Raf kinase inhibitor protein and phosphor Raf kinase inhibitor were also identified as potential urinary biomarkers for ccRCC using a proteomics technique with an AUC of 0.93.[28] In addition, Bao et al. identified hub genes associated with ccRCC from GEO dataset (GSE47352). They found that hub genes could distinguish ccRCC from paired normal tissue with an AUC ranging from 0.517 to 0.945.[29] Compared to the performance of currently established biomarkers for ccRCC diagnosis, the present study used a combination of bioinformatics and machine learning algorithms based on the expression levels of *IGFBP3* and *LGALS1* to achieve a notably higher diagnostic accuracy of 98.04% and an AUC of 0.98. Our results demonstrated the value of machine learning in achieving higher accuracy and consistency, which could lead to improved early detection and patient outcomes.

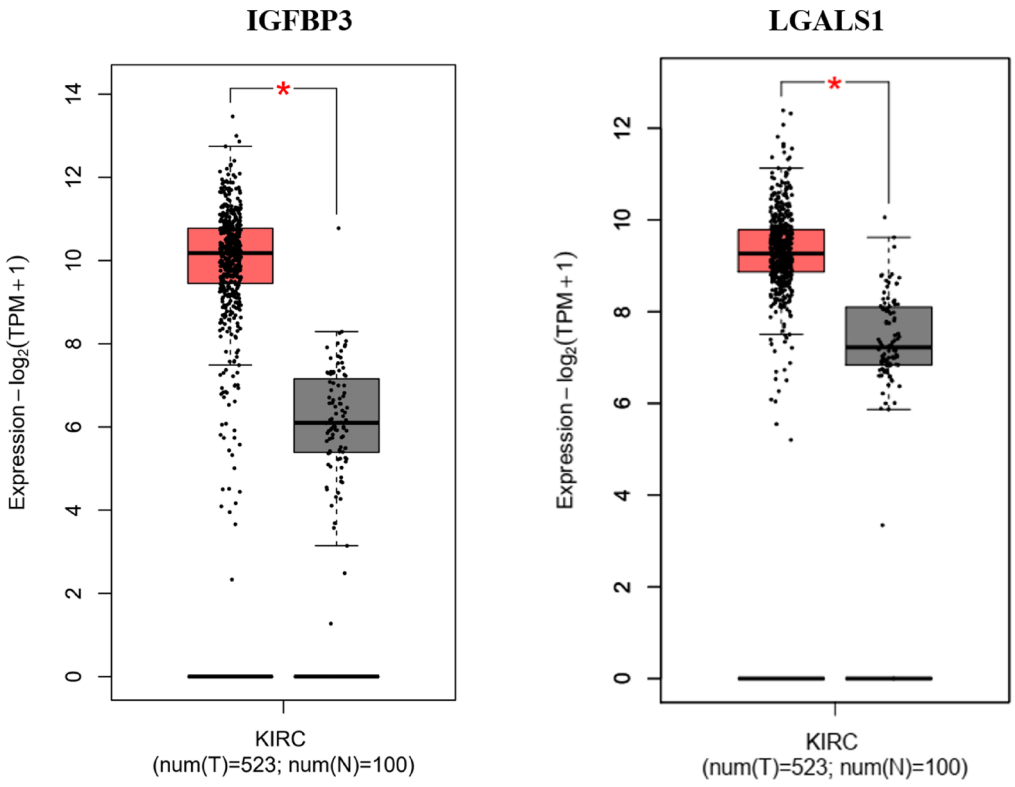

Fig. 7. Expression of IGFBP3 and LGALS1 in the The Cancer Genome Atlas (TCGA) dataset. Box plots display the expression levels of IGFBP3 and LGALS1 in kidney renal clear cell carcinoma (KIRC; red plot) tissues and non-tumor renal tissues (gray plot). In the box plots, the central line indicated the median, the box represented the interquartile range (IQR; Q1 to Q3) and the whiskers extended to 1.5 times the IQR from the quartiles. Outliers were shown as individual points

IGFBP-3 – insulin-like growth factor binding protein 3; LGALS1 – lectin, galactoside-binding, soluble, 1. *p < 0.05.

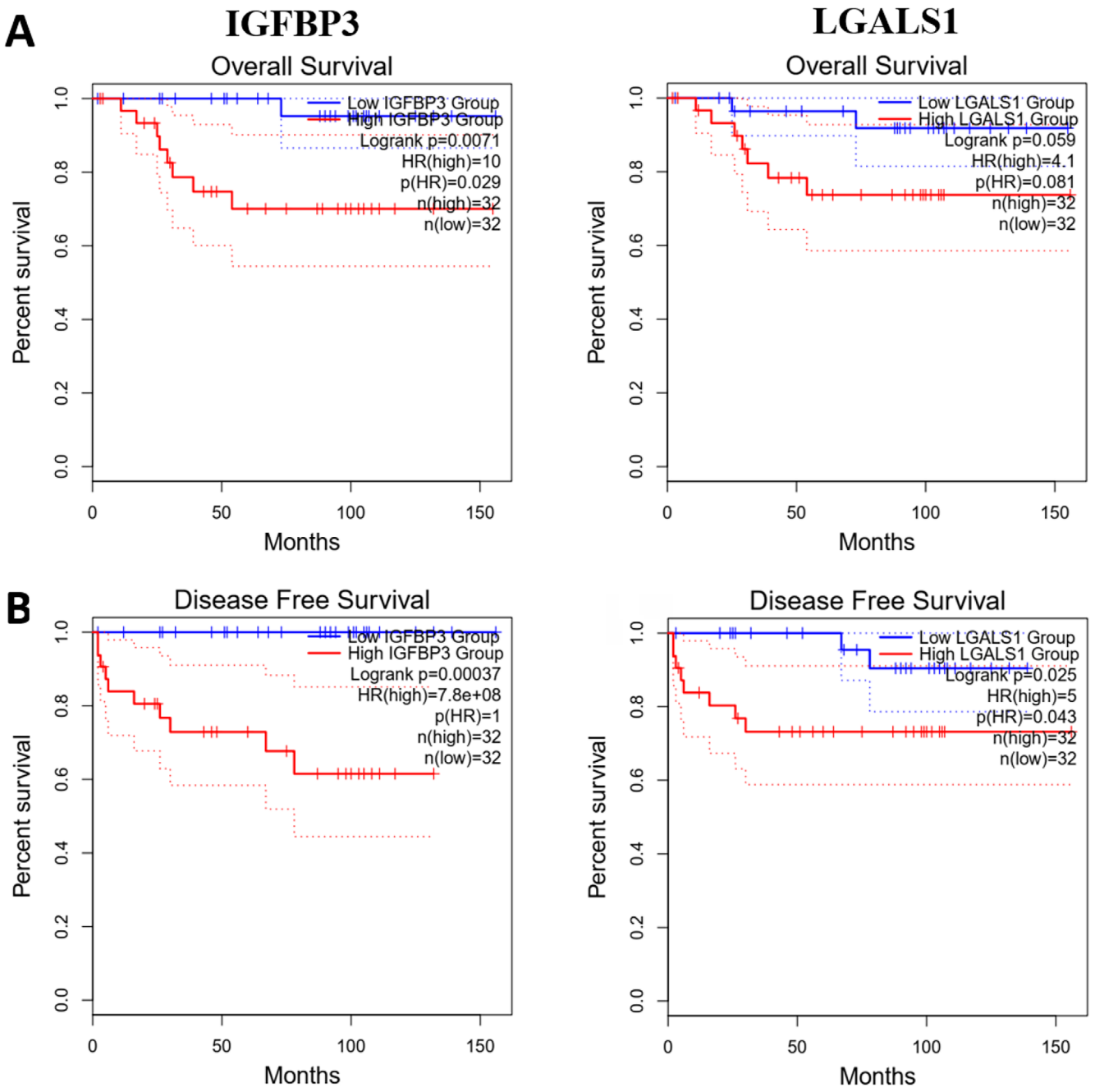

Fig. 8. Impact of IGFBP3 and LGALS1 expression on survival in ccRCC patients. The overall survival (A) and the disease-free survival (B) of ccRCC patients based on IGFBP3 and LGALS1 expression levels. Patients were divided into high- and low-expression groups

IGFBP-3 – insulin-like growth factor binding protein 3; LGALS1 – lectin, galactoside-binding, soluble, 1; ccRCC – clear cell renal cell carcinoma; HR – hazard ratio.

Transcriptomic data from the GEO and TCGA databases revealed high expression levels of *IGFBP3* and *LGALS1* in ccRCC tissue compared to normal kidney tissue. In addition, high expression of these genes was associated with shorter OS and DFS in ccRCC patients, underscoring their potential utility as diagnostic markers and prognostic indicators.

In cancer cells, *IGFBP3* regulates cell proliferation and apoptosis through both IGF-dependent and independent mechanisms. *IGFBP3* depletion suppresses glioma cell growth by inducing DNA damage and apoptosis. Furthermore, suppression of *IGFBP3* markedly increased the survival of brain-tumor-bearing mice.[30] Suppression of the IGFBP3-AKT/STAT3/MAPK-Snail signaling pathway by cyclovirobuxine resulted in a reduction of cell viability, proliferation, angiogenesis, migration, and invasion in ccRCC cells.[31] Overexpression of *IGFBP3* has been reported in several cancers, including breast cancer and nasopharyngeal carcinoma.[32,33] *IGFBP3* expression is associated with adverse outcomes such as metastasis, poor responses to chemoradiotherapy and decreased survival rates in cancer patients[33–35] Moreover, serum *IGFBP3*

is an independent prognostic risk factor in esophageal squamous cell carcinoma and esophagogastric junction adenocarcinoma.[36,37] Overexpression of *IGFBP3* has also been reported in ccRCC. A study by Braczkowski et al. demonstrated *IGFBP3* overexpression in ccRCC compared to adjacent non-cancerous kidney tissues using a quantitative reverse transcription polymerase chain reaction (RT-qPCR) assay.[38] The distribution of *IGFBP3* genotypes was significantly associated with the histological grade and clinical stage of ccRCC patients.[39] This information suggests that *IGFBP3* could serve as a diagnostic and prognostic biomarker for ccRCC.

LGALS1, also known as galectin-1, is involved in various processes associated with cancer development and progression, including tumor transformation, cell cycle regulation, apoptosis, adhesion, migration, and inflammation.[40,41] Huang et al. reported that the suppression of *LGALS1* led to reduced cell invasion, clonogenic ability, epithelial–mesenchymal transition, and angiogenesis in renal cancer cell lines by upregulating C-X-C chemokine receptor type 4 through nuclear factor kappa B (NF-κB) activation.[42] Similarly, a report from 2014 highlighted that *LGALS1* plays a critical role in promoting the migration and invasion of ccRCC cells by activating the hypoxia-inducible factors/mammalian target of rapamycin signaling pathway.[43] Overexpression of *LGALS1* is correlated with tumor aggressiveness, including growth, cell migration, invasion, metastasis, and poor prognosis in several cancers such as hepatocellular carcinoma (HCC), upper urinary urothelial carcinoma, ovarian cancer, and squamous cervical cancer.[44–48] The potential of *LGALS1* as a serum biomarker has also been demonstrated in several cancers. Elevated plasma levels of galectin-1 have been found in pancreatic cancer,[49] classical Hodgkin lymphoma[50] and serous ovarian carcinoma. High serum levels of galectin-1 are associated with metastasis in epithelial ovarian cancer[51] and colorectal cancer.[52] In ccRCC, *LGALS1* expression was significantly associated with higher clinical grade and stage[53] and favorable outcomes from anti-PD1 treatment.[54] These results indicate the potential of using *LGALS1* as a prognostic marker and therapeutic target in ccRCC patients.

The results of our integrated bioinformatics and machine learning analysis indicate that *IGFBP3* and *LGALS1* are promising potential secreted biomarkers for the diagnosis of ccRCC.

Limitations

It is important to acknowledge the limitations of this study. The findings were derived from publicly available datasets from the GEO and the TCGA databases. The selection of these datasets may introduce potential biases, as they may not fully represent the broader patient population. Furthermore, the generalizability of our results may be constrained by variations in sample collection, processing methods and demographic factors across different studies. Accordingly, further research is planned to validate these findings in independent cohorts using serum or urine of ccRCC patients compared to healthy controls, with the objective of ensuring robustness and applicability to clinical settings.

Conclusions

The use of bioinformatics and machine learning enabled the identification of IGFBP3 and LGALS1 as potential secreted biomarkers for ccRCC. The classification models based on IGFBP3 and LGALS1 demonstrated the capacity to effectively differentiate ccRCC patients from healthy controls. Furthermore, the expression levels of IGFBP3 and LGALS1 were found to be useful not only for the diagnosis of ccRCC but also as prognostic biomarkers to predict patient outcomes.

Supplementary data

The Supplementary materials are available at https://doi.org/10.6084/m9.figshare.27154224.v1. The package includes the following files:

Supplementary File 1. DEG analysis from GEO database (GSE11151, GSE40435 and GSE53757).

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Consent for publication

Not applicable.

ORCID iDs

Wunchana Seubwai https://orcid.org/0000-0002-9265-5113
Sakkarn Sangkhamanon https://orcid.org/0000-0002-5203-8700
Xuhong Zhang https://orcid.org/0000-0001-7563-9915

References

1. Bukavina L, Bensalah K, Bray F, et al. Epidemiology of renal cell carcinoma: 2022 update. *Eur Urol*. 2022;82(5):529–542. doi:10.1016/j.eururo.2022.08.019
2. Young M, Jackson-Spence F, Beltran L, et al. Renal cell carcinoma. *Lancet*. 2024;404(10451):476–491. doi:10.1016/S0140-6736(24)00917-6
3. Huang J, Leung DKW, Chan EOT, et al. A global trend analysis of kidney cancer incidence and mortality and their associations with smoking, alcohol consumption, and metabolic syndrome. *Eur Urol Focus*. 2022;8(1):200–209. doi:10.1016/j.euf.2020.12.020
4. Vasudev NS, Wilson M, Stewart GD, et al. Challenges of early renal cancer detection: Symptom patterns and incidental diagnosis rate in a multicentre prospective UK cohort of patients presenting with suspected renal cancer. *BMJ Open*. 2020;10(5):e035938. doi:10.1136/bmjopen-2019-035938

5. Patard JJ, Leray E, Rodriguez A, Rioux-Leclercq N, Guillé F, Lobel B. Correlation between symptom graduation, tumor characteristics and survival in renal cell carcinoma. *Eur Urol*. 2003;44(2):226–232. doi:10.1016/S0302-2838(03)00216-1
6. Chen YW, Wang L, Panian J, et al. Treatment landscape of renal cell carcinoma. *Curr Treat Options Oncol*. 2023;24(12):1889–1916. doi:10.1007/s11864-023-01161-5
7. Ghatalia P, Zibelman MR, Geynisman DM, Plimack ER. Evolving landscape of the treatment of metastatic clear cell renal cell carcinoma. *Clin Adv Hematol Oncol*. 2018;16(10):677–686. PMID:30543598.
8. Hsieh JJ, Purdue MP, Signoretti S, et al. Renal cell carcinoma. *Nat Rev Dis Primers*. 2017;3(1):17009. doi:10.1038/nrdp.2017.9
9. Schiavoni V, Campagna R, Pozzi V, et al. Recent advances in the management of clear cell renal cell carcinoma: Novel biomarkers and targeted therapies. *Cancers (Basel)*. 2023;15(12):3207. doi:10.3390/cancers15123207
10. Yang X, Kui L, Tang M, et al. High-throughput transcriptome profiling in drug and biomarker discovery. *Front Genet*. 2020;11:19. doi:10.3389/fgene.2020.00019
11. McCain J. The cancer genome atlas: New weapon in old war? *Biotechnol Healthc*. 2006;3(2):46–51B. PMID:23424349. PMCID:PMC3571024.
12. Tomczak K, Czerwińska P, Wiznerowicz M. Review The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp Oncol (Pozn)*. 2015;1A:68–77. doi:10.5114/wo.2014.47136
13. Barrett T, Edgar R. Gene Expression Omnibus: Microarray data storage, submission, retrieval, and analysis. *Methods Enzymol*. 2006;411:352–369. doi:10.1016/S0076-6879(06)11019-8
14. Clough E, Barrett T. The Gene Expression Omnibus database. *Methods Mol Biol*. 2016;1418:93–110. doi:10.1007/978-1-4939-3578-9_5
15. Ng S, Masarone S, Watson D, Barnes MR. The benefits and pitfalls of machine learning for biomarker discovery. *Cell Tissue Res*. 2023;394(1):17–31. doi:10.1007/s00441-023-03816-z
16. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13:8–17. doi:10.1016/j.csbj.2014.11.005
17. Hammad A, Elshaer M, Tang X. Identification of potential biomarkers with colorectal cancer based on bioinformatics analysis and machine learning. *Math Biosci Eng*. 2021;18(6):8997–9015. doi:10.3934/mbe.2021443
18. Li C, Zeng X, Yu H, Gu Y, Zhang W. Identification of hub genes with diagnostic values in pancreatic cancer by bioinformatics analyses and supervised learning methods. *World J Surg Onc*. 2018;16(1):223. doi:10.1186/s12957-018-1519-y
19. Alam MS, Sultana A, Sun H, et al. Bioinformatics and network-based screening and discovery of potential molecular targets and small molecular drugs for breast cancer. *Front Pharmacol*. 2022;13:942126. doi:10.3389/fphar.2022.942126
20. Davis S, Meltzer PS. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*. 2007;23(14):1846–1847. doi:10.1093/bioinformatics/btm254
21. Thul PJ, Lindskog C. The Human Protein Atlas: A spatial map of the human proteome. *Protein Sci*. 2018;27(1):233–244. doi:10.1002/pro.3307
22. Pontén F, Jirström K, Uhlen M. The Human Protein Atlas: A tool for pathology. *J Pathol*. 2008;216(4):387–393. doi:10.1002/path.2440
23. Li C, Tang Z, Zhang W, Ye Z, Liu F. GEPIA2021: Integrating multiple deconvolution-based analysis into GEPIA. *Nucl Acids Res*. 2021;49(W1):W242–W246. doi:10.1093/nar/gkab418
24. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucl Acids Res*. 2017;45(W1):W98–W102. doi:10.1093/nar/gkx247
25. Takacova M, Bartosova M, Skvarkova L, et al. Carbonic anhydrase IX is a clinically significant tissue and serum biomarker associated with renal cell carcinoma. *Oncol Lett*. 2013;5(1):191–197. doi:10.3892/ol.2012.1001
26. Lucarini L, Magnelli L, Schiavone N, et al. Plasmatic carbonic anhydrase IX as a diagnostic marker for clear cell renal cell carcinoma. *J Enzyme Inhib Med Chem*. 2018;33(1):234–240. doi:10.1080/14756366.2017.1411350
27. Yang J, Yang J, Gao Y, et al. Identification of potential serum proteomic biomarkers for clear cell renal cell carcinoma. *PLoS One*. 2014;9(11):e111364. doi:10.1371/journal.pone.0111364
28. Papale M, Vocino G, Lucarelli G, et al. Urinary RKIP/p-RKIP is a potential diagnostic and prognostic marker of clear cell renal cell carcinoma. *Oncotarget*. 2017;8(25):40412–40424. doi:10.18632/oncotarget.16341
29. Bao L, Zhao Y, Liu C, et al. The identification of key gene expression signature and biological pathways in metastatic renal cell carcinoma. *J Cancer*. 2020;11(7):1712–1726. doi:10.7150/jca.38379
30. Chen CH, Chen PY, Lin YY, et al. Suppression of tumor growth via IGFBP3 depletion as a potential treatment in glioma. *J Neurosurg*. 2020;132(1):168–179. doi:10.3171/2018.8.JNS181217
31. Liu Y, Lv H, Li X, et al. Cyclovirobuxine inhibits the progression of clear cell renal cell carcinoma by suppressing the IGFBP3-AKT/STAT3/MAPK-Snail signalling pathway. *Int J Biol Sci*. 2021;17(13):3522–3537. doi:10.7150/ijbs.62114
32. McCarthy K, Laban C, McVittie CJ, et al. The expression and function of IGFBP-3 in normal and malignant breast tissue. *Anticancer Res*. 2009;29(10):3785–3790. PMID:19846909.
33. Bao L, Liu H, You B, et al. Overexpression of IGFBP3 is associated with poor prognosis and tumor metastasis in nasopharyngeal carcinoma. *Tumor Biol*. 2016;37(11):15043–15052. doi:10.1007/s13277-016-5400-8
34. Sakata J, Hirosue A, Yoshida R, et al. Enhanced expression of IGFBP-3 reduces radiosensitivity and is associated with poor prognosis in oral squamous cell carcinoma. *Cancers (Basel)*. 2020;12(2):494. doi:10.3390/cancers12020494
35. Yamamoto N, Oshima T, Yoshihara K, et al. Clinicopathological significance and impact on outcomes of the gene expression levels of IGF-1, IGF-2 and IGF-1R, IGFBP-3 in patients with colorectal cancer. *Oncol Lett*. 2017;13(5):3958–3966. doi:10.3892/ol.2017.5936
36. Luo Y, Hong CQ, Huang BL, et al. Serum insulin-like growth factor binding protein-3 as a potential biomarker for diagnosis and prognosis of oesophageal squamous cell carcinoma. *Ann Med*. 2022;54(1):2153–2166. doi:10.1080/07853890.2022.2104921
37. Ding TY, Peng YH, Hong CQ, et al. Serum insulin-like growth factor binding protein 3 as a promising diagnostic and prognostic biomarker in esophagogastric junction adenocarcinoma. *Discov Onc*. 2022;13(1):128. doi:10.1007/s12672-022-00591-1
38. Braczkowski R, Białożyt M, Plato M, Mazurek U, Braczkowska B. Expression of insulin-like growth factor family genes in clear cell renal cell carcinoma. *Contemp Oncol (Pozn)*. 2016;2:130–136. doi:10.5114/wo.2016.58720
39. Safarinejad MR. Insulin-like growth factor binding protein-3 (*IGFBP-3*) gene variants are associated with renal cell carcinoma. *BJU Int*. 2011;108(5):762–770. doi:10.1111/j.1464-410X.2010.10017.x
40. Rabinovich GA. Galectin-1 as a potential cancer target. *Br J Cancer*. 2005;92(7):1188–1192. doi:10.1038/sj.bjc.6602493
41. Cousin J, Cloninger M. The role of galectin-1 in cancer progression, and synthetic multivalent systems for the study of galectin-1. *Int J Mol Sci*. 2016;17(9):1566. doi:10.3390/ijms17091566
42. Huang CS, Tang SJ, Chung LY, et al. Galectin-1 upregulates CXCR4 to promote tumor progression and poor outcome in kidney cancer. *J Am Soc Nephrol*. 2014;25(7):1486–1495. doi:10.1681/ASN.2013070773
43. White NMA, Masui O, Newsted D, et al. Galectin-1 has potential prognostic significance and is implicated in clear cell renal cell carcinoma progression through the HIF/mTOR signaling axis. *Br J Cancer*. 2014;110(5):1250–1259. doi:10.1038/bjc.2013.828
44. Setayesh T, Colquhoun SD, Wan YJY. Overexpression of galectin-1 and galectin-3 in hepatocellular carcinoma. *Liver Res*. 2020;4(4):173–179. doi:10.1016/j.livres.2020.11.001
45. Leung Z, Ko FCF, Tey SK, et al. Galectin-1 promotes hepatocellular carcinoma and the combined therapeutic effect of OTX008 galectin-1 inhibitor and sorafenib in tumor cells. *J Exp Clin Cancer Res*. 2019;38(1):423. doi:10.1186/s13046-019-1402-x
46. Su YL, Luo HL, Huang CC, et al. Galectin-1 overexpression activates the FAK/PI3K/AKT/mTOR pathway and is correlated with upper urinary urothelial carcinoma progression and survival. *Cells*. 2020;9(4):806. doi:10.3390/cells9040806
47. Zhang P, Zhang P, Shi B, et al. Galectin-1 overexpression promotes progression and chemoresistance to cisplatin in epithelial ovarian cancer. *Cell Death Dis*. 2014;5(1):e991–e991. doi:10.1038/cddis.2013.526
48. Punt S, Thijssen VL, Vrolijk J, De Kroon CD, Gorter A, Jordanova ES. Galectin-1, -3 and -9 expression and clinical significance in squamous cervical cancer. *PLoS One*. 2015;10(6):e0129119. doi:10.1371/journal.pone.0129119

49. Martinez-Bosch N, Barranco LE, Orozco CA, et al. Increased plasma levels of galectin-1 in pancreatic cancer: Potential use as biomarker. *Oncotarget*. 2018;9(68):32984–32996. doi:10.18632/oncotarget.26034
50. Ouyang J, Plütschow A, von Strandmann EP, et al. Galectin-1 serum levels reflect tumor burden and adverse clinical features in classical Hodgkin lymphoma. *Blood*. 2013;121(17):3431–3433. doi:10.1182/blood-2012-12-474569
51. Chen L, Yao Y, Sun L, et al. Clinical implication of the serum galectin-1 expression in epithelial ovarian cancer patients. *J Ovarian Res*. 2015;8(1):78. doi:10.1186/s13048-015-0206-7
52. Wu KL, Chen HH, Pen CT, et al. Circulating galectin-1 and 90K/Mac-2BP correlated with the tumor stages of patients with colorectal cancer. *Biomed Res Int*. 2015;2015:306964. doi:10.1155/2015/306964
53. Fang J, Wang X, Xie J, et al. LGALS1 was related to the prognosis of clear cell renal cell carcinoma identified by weighted correlation gene network analysis combined with differential gene expression analysis. *Front Genet*. 2023;13:1046164. doi:10.3389/fgene.2022.1046164
54. Li Y, Yang S, Yue H, et al. Unraveling LGALS1 as a potential immune checkpoint and a predictor of the response to anti-PD1 therapy in clear cell renal carcinoma. *Pathol Oncol Res*. 2020;26(3):1451–1458. doi:10.1007/s12253-019-00710-4