Verification of Existing Immunohistochemical Biomarkers for Diagnosis of Clear Cell Renal Cell Carcinoma by Bioinformatics and Machine Learning Approaches

การตรวจสอบตัวบ่งชี้ชีวภาพวิธีอิมมูนโนฮีสโทเคมีที่ใช้อยู่ในปัจจุบันสำหรับการวินิจฉัยมะเร็งเซลล์ไตชนิด clear cell renal cell carcinoma ด้วยเทคนิคชีวสารสนเทศและการเรียนรู้ของเครื่อง

Chanita Panwoon (ชนิตา ปานวุ่น)* Dr.Wunchana Seubwai (ดร.วันชนะ สืบไวย)**
Sakkarn Sangkhamanon (สักการ สังฆมานนท์)***

ABSTRACT

Immunohistochemical biomarkers may be necessary to support pathologists to make an accurate diagnosis in some unusual morphologic features of renal cancer, especially the most common subtype, clear cell renal cell carcinoma (ccRCC). However, no study has approved these biomarkers' efficacy for diagnosing ccRCC. Therefore, this study aimed to verify the efficacy of eight existing ccRCC biomarkers, including PAX8, CD10, CAIX, Vimentin, CK7, CD117/KIT, AMACR and GATA3, using machine learning and bioinformatics techniques. The expression profiles of eight existing ccRCC biomarkers were retrieved from the Gene Expression Omnibus (GEO), consisted of 72 ccRCC samples with matching normal tissues. The dataset was devided into two groups, consisted of Training set (70%) and Test set (30%). Then, eight existing ccRCC biomarkers were inputted into nine machine learning algorithms to build and evaluate prediction models for diagnosing ccRCC. Expression levels and clinical association of candidate existing ccRCC biomarkers for machine learning were further analyzed by UALCAN based on data from The Cancer Genome Atlas (TCGA). The Navies Bayes (NB) model was the most efficient model with an accuracy of 83.3%, sensitivity of 81%, specificity of 86% and area under curve (AUC) of 0.878. In the decision tree (DT) model, only PAX8 was used to diagnose ccRCC. DT model was efficient with an accuracy of 76.1%, sensitivity of 72%, specificity of 86% and AUC of 0.791. In conclusion, two existing ccRCC biomarkers, including GATA3 from NB and PAX8 from DT, were the most potential diagnostic markers for ccRCC. Furthermore, a combination of bioinformatics and machine learning might be a tool for verifying the efficacy of diagnostic biomarkers for renal cancer and other cancers.

บทคัดย่อ

ตัวบ่งชี้ทางชีวภาพที่ใช้ในเทคนิคอิมมูโนฮิสโตเคมี มีความจำเป็นในการช่วยให้พยาธิแพทย์สามารถวินิจฉัยมะเร็งไตที่มีลักษณะทางสัณฐานวิทยาไม่ชัดเจนได้อย่างแม่นยำ โดยเฉพาะมะเร็งไตชนิด clear cell renal cell carcinoma (ccRCC) ที่สามารถพบได้บ่อยที่สุด อย่างไรก็ตามยังไม่มีการศึกษาพิสูจน์ประสิทธิภาพของตัวบ่งชี้ทางชีวภาพดังกล่าว การศึกษานี้จึงมีวัตถุประสงค์ในการตรวจสอบประสิทธิภาพของตัวบ่งชี้ทางชีวภาพสำหรับวินิจฉัย ccRCC จำนวน 8 ชนิด ได้แก่ PAX8 CD10 CAIX Vimentin CK7 CD117/KIT AMACR และ GATA3 โดยการใช้เทคนิคการเรียนรู้ของเครื่องและชีวสารสนเทศ การแสดงออกของตัวบ่งชี้ทางชีวภาพทั้ง 8 ชนิด ถูกนำมาจากฐานข้อมูล Gene Expression Omnibus (GEO)

*Student, Master of Science Program in Pathology, Department of Pathology, Faculty of Medicine, Khon Kaen University
**Assistant Professor, Department of Forensic Medicine, Faculty of Medicine, Khon Kaen University
***Assistant Professor, Department of Pathology, Faculty of Medicine, Khon Kaen University

ซึ่งประกอบด้วยตัวอย่างจากเนื้อเยื่อ ccRCC และเนื้อเยื่อปกติจำนวนกลุ่มละ 72 ตัวอย่าง ข้อมูลถูกแบ่งออกเป็นสองกลุ่ม ได้แก่ Training set (70%) และ Test set (30%) และทดสอบโดยใช้อัลกอริธึมการเรียนรู้ของเครื่องจำนวน 9 ชนิด เพื่อสร้างแบบจำลองสำหรับประเมินประสิทธิภาพของตัวบ่งชี้ชีวทางภาพแต่ละชนิด จากนั้นการแสดงออกของตัวบ่งชี้ทางชีวภาพที่มีประสิทธิภาพในการวินิจฉัย ccRCC จะถูกยืนยันและหาความสัมพันธ์กับคุณลักษณะทางคลินิก โดยใช้ข้อมูลจากฐานข้อมูล The Cancer Genome Atlas (TCGA) ผลการวิเคราะห์พบว่า Navies Bayes (NB) มีประสิทธิภาพสูงสุดด้วยความแม่นยำร้อยละ 83.3 ความไวร้อยละ 81 ความจำเพาะร้อยละ 86 และพื้นที่ใต้เส้นโค้งเท่ากับ 0.878 ส่วน decision tree (DT) ซึ่งใช้ตัวบ่งชี้ทางชีวภาพเพียง 1 ชนิด (PAX8) ในการวินิจฉัย ccRCC มีความแม่นยำร้อยละ 76.1 ความไวร้อยละ 72 ความจำเพาะร้อยละ 86 และพื้นที่ใต้เส้นโค้งเท่ากับ 0.791 โดยสรุปตัวบ่งชี้ชีวทางภาพชนิด GATA3 ที่ได้จาก NB และ PAX8 ที่ได้จาก DT เป็นตัวบ่งชี้ทางชีวภาพที่มีประสิทธิภาพในการวินิจฉัย ccRCC นอกจากนี้การผสมผสานระหว่างเทคนิคการเรียนรู้ของเครื่องและชีวสารสนเทศสามารถใช้เป็นเครื่องมือในการทดสอบประสิทธิภาพของตัวบ่งชี้ชีวภาพในการวินิจฉัยมะเร็งไตหรือมะเร็งชนิดอื่นๆ

Keyword: Clear cell renal cell carcinoma, Biomarker, Machine learning
คำสำคัญ: มะเร็งเซลล์ไตชนิด clear cell ตัวบ่งชี้ทางชีวภาพ การเรียนรู้ของเครื่อง

Introduction

Renal cell carcinoma (RCC) originated from the epithelium of the renal tubules. RCC is responsible for about 2% of all cancer diagnoses and deaths globally, and the incidence of cancer is expected to rise (Ngan et al., 2001). Approximately 85-90% of all cases were clear cell renal cell carcinoma (ccRCC), followed by papillary renal cell carcinoma (PRCC) and chromophobe renal cell carcinoma (chRCC), respectively (Cheville et al., 2003). RCC diagnosis required ancillary investigations, immunohistochemistry (IHC) is the most widely used. Currently, there are approximately eight biomarkers which are routinely used in diagnostic pathology service including Paired-box gene 8 (PAX8), Cluster of differentiation 10 (CD10), Carboxinic anhydrase IX (CAIX), Vimentin, Cytokeratin 7 (CK7), Receptor tyrosine kinase (CD117/KIT), **α**-Methylacyl coenzyme A racemase (AMACR), GATA binding protein 3 (GATA3) (Table 1) (Akgul & Cheng, 2020). However, the diagnostic performance of these biomarkers has not been evaluated. The purpose of this study was to assess the efficacy of these biomarkers in distinguishing ccRCC from normal kidney tissue.

In recent years, public genomics, and transcriptomics databases, such as Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA), have provided a powerful tool for elucidating key genetic alterations in carcinogenesis in recent years, and have been widely used to discover promising biomarkers for cancer diagnosis, new therapeutic targets, and prognosis prediction. Furthermore, to avoid a limited sample size in study, bioinformatics techniques have become increasingly commonly used in cancer research (Lu et al., 2020). Machine learning (ML) is increasingly being used in medical research, especially utilized to diagnose and prediction various diseases (Cruz & Wishart, 2006). One of the supervised

learning models is classification, which aims to develop or create a model that can be utilized for classification (Kabir & Ludwig, 2019).

In this study, we integrated bioinformatic and machine learning to verify these existing biomarkers with array datasets from the Gene Expression Omnibus (GEO) to classify clear cell RCC (ccRCC) samples and normal samples. Then, we use UALCAN to analyze expression and correlation with survival rates of interesting genes.

Table 1 The description of eight existing biomarkers

Genes	Description	Staining areas	Reference
PAX8	- PAX8 is a nephric-lineage transcription factor and associated with organogenesis of the thyroid gland, kidney, and Müllerian system. - High expression in kidney and ovarian carcinomas.	Nuclear	(Tacha et al., 2011), (Truong & Shen, 2011)
CD10	- CD10 is a cell-surface glycoprotein expressed in a variety of tissues and malignancies. - The CD10 show positive stain in clear cell RCC and papillary RCC.	Cytoplasmic and membrane	(Shen et al., 2012)
CAIX	- Carbonic anhydrase IX is a transmembrane enzyme that regulates cell proliferation, adhesion, and invasion. - Expression in many carcinomas of the kidney, endometrium, stomach, cervix, breast, lung, and liver and in brain tumors, neuroendocrine tumors, and mesotheliomas.	Membrane	(Truong & Shen, 2011)
Vimentin	- Vimentin is a mesenchymal marker, which expressed by most types of RCC, often in a diffuse fashion. - This is useful in narrowing down the differential diagnosis of metastatic RCC	Cytoplasm	(Truong & Shen, 2011)
CK7	- Cytokeratin 7 (CK7) is a type II basic low molecular weight cytokeratin. - Present in simple epithelia in a variety of organs, including all epithelia in the female genital tract.	Cytoplasmic and membrane	(Rabban et al., 2010)

Table 1 The description of eight existing biomarkers (Cont.)

Genes	Description	Staining areas	Reference
CD117/KIT	- C-kit (CD117) is a proto-oncogene encoding a tyrosine transmembrane receptor. - CD117 expression in many tumors such as gastrointestinal stromal tumors, leukemia, lung carcinoma, mesothelioma, neuroendocrine carcinoma, serous ovarian carcinoma, and melanoma.	Cytoplasm	(Truong & Shen, 2011)
AMARC	- α-Methylacyl coenzyme A racemase (AMACR) is a mitochondrial enzyme mediating oxidation of fatty acids.	Cytoplasm	(Truong & Shen, 2011)
GATA3	- GATA3 is a transcription factor of the GATA family. - GATA3 is a multifunctional transcription factor that has a role in the development and function of breast ductal epithelial cells, urothelia, epidermis, some skin adenxa, and T-cell subsets.	Nuclear	(Miettinen et al., 2014)

Remarks: Paired-box gene 8 (PAX8), Cluster of differentiation 10 (CD10), Carboxinic anhydrase IX (CAIX), Vimentin, Cytokeratin 7 (CK7), Receptor tyrosine kinase (CD117/KIT), α-Methylacyl coenzyme A racemase (AMACR), GATA binding protein 3 (GATA3)

Material and Methods

Transcriptomic data sources and Data processing

The gene expression dataset of ccRCC and normal tissues were retrieved from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) (Edgar et al., 2002) and The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov/) (Tomczak et al., 2015). GSE53757 dataset (Table 1) based on GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array (Table 2). GEO data were analyzed using the R package, including GEOquery (Davis & Meltzer, 2007) and limma (Ritchie et al., 2015). Expression profiles and correlation with the survival rate of each candidate genes in kidney renal clear cell carcinoma were analyzed using UALCAN (http://ualcan.path.uab.edu/) based on TCGA data (Chandrashekar et al., 2017).

Table 2 The information of array dataset

Dataset	Description
Samples	Clear cell renal cell carcinoma tissue versus matched normal kidney tissue
Sample Sizes (Normal/Tumor)	72/72
Platforms	GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array
Reference	Von Roemeling CA, Radisky DC, Marlow LA, Cooper SJ et al. Neuronal pentraxin 2 supports clear cell renal cell carcinoma by activating the AMPA-selective glutamate receptor-4. Cancer Res 2014 Sep 1;74(17):4796-810.

Supervised machine learning classifiers

Nine types of supervised machine learning classifiers, including Decision Tree (DT), Deep Learning (DL), Fast Large Margin (FLM), Generalized Linear Model (GLM), Gradient Boosted Trees (GBT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM), were used to create models for predicting diagnostic markers using the Array dataset. To evaluate the model's performance, the following metrics, including accuracy, sensitivity, specificity, and area under the curve (AUC) value, were determined. All machine learning classifiers were implemented using RapidMiner (version 9.9) (https://rapidminer.com/) (Jungermann, 2009) with the default parameters.

In this study, samples were divided into two group including training set (70%), were utilized to run models, and test set (30%), was used to assess model performance. The eight existing RCC biomarkers consist of PAX8, CD10, CAIX, Vimentin, CK7, CD117/KIT, AMACR, GATA3 were input into nine machine learning algorithms to build a prediction model for predicting diagnostic marker and evaluated predictive model. The workflow of this study is shown in Figure 1.

Results

In this study, the classifiers were developed to identify ccRCC. Samples from the GSE53757 dataset were divided into a training set (70%) and a test set (30%). The performance of nine supervised machine learning classifiers, including DT, DL, FLM, GLM, GBT, LR, NB, RF, and SVM, were evaluated based on the expression profile of eight candidate genes. A summarization of the performance of each model was shown in Table 3.

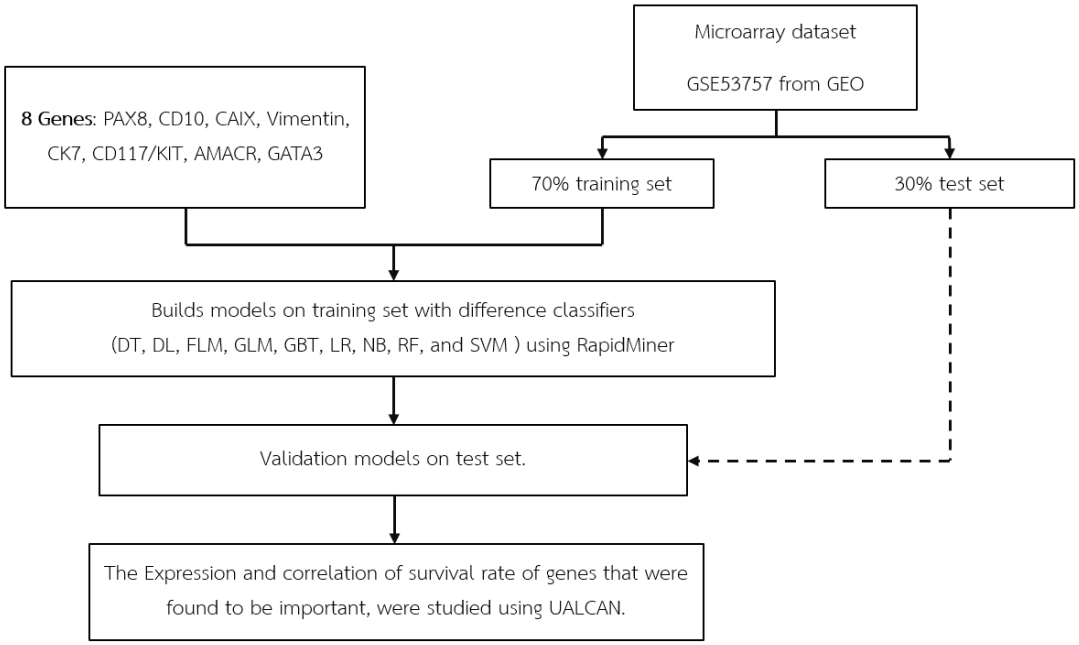

Figure 1 The workflow of this study. Gene Expression Omnibus (GEO), Decision Tree (DT), Deep Learning (DL), Fast Large Margin (FLM), Generalized Linear Model (GLM), Gradient Boosted Trees (GBT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM)

The NB model was the best performance (Table 3), with an overall predictive accuracy of 0.833, AUC of 0.878, a sensitivity of 0.810, and specificity of 0.860. However, the performance of other models gave satisfactory value as well. The LR had an accuracy second to the NB at 0.811. The models of DL and FLM had the same accuracy of 0.789. As for the AUC value of RF, DL, and DT were less than NB at 0.833, 0.830, 0.791, respectively. Specificity shown high value in almost all models, except in SVM.

We also analyzed the weight of each gene in different models to determine the significance of that gene (Table 4). It was found that the most effective NB model had the highest weight gene GATA3 at 0.396, followed by CD10, VIM, CAIX, and PAX8 at 0.122, 0.090, 0.090, and 0.058, respectively

GATA3 and PAX8 were selected for further analysis. The expression and clinical correlation of GATA3 and PAX8 were determined using UALCAN based on RNA-Seq dataset from the TCGA database, consisting of 72 normal tissues and 533 ccRCC tissues. The GATA3 gene was highly expressed in normal tissues compared to the ccRCC samples at p-value < 0.001 (Figure 2A). However, no relationships between GATA3 expressions and clinicopathological findings, including age, sex, tumor staging, tumor type and metastasis status, and the cumulative survival of the patients were noted (Figure 2B). Similar findings were found in PAX8, down-regulated expression of PAX8 was observed in ccRCC patients' tissues with normal tissues (p-value < 0.001) (Figure 3A). In addition, there is no significant association between

PAX8 expression and clinical findings of ccRCC patients (Figure 3B). For differential diagnosis of ccRCC and normal samples, GATA3 and PAX8 shown high-performance values of 85.8% and 73.9% accuracy, respectively. GATA3 had 86.0% of both sensitivity and specificity, whereas PAX8 had 81.0% of sensitivity and 67.0% of specificity (Table 5).

Table 3 The performance of each classifier is represented as a heat map. Accuracy, classification Error, AUC, precision recall, F Measure, sensitivity, and specificity are among the metrics

Model	Accuracy	Classification Error	AUC	Precision	Recall	F Measure	Sensitivity	Specificity
DT	0.761	0.239	0.791	0.803	0.720	0.748	0.720	0.860
DL	0.789	0.211	0.830	0.883	0.670	0.748	0.670	0.860
FLM	0.789	0.211	0.735	0.883	0.670	0.748	0.670	0.910
GLM	0.786	0.214	0.760	0.833	0.720	0.764	0.720	0.910
GBT	0.569	0.431	0.679	0.633	0.330	0.429	0.330	0.910
LR	0.811	0.189	0.788	0.883	0.710	0.775	0.710	0.810
NB	0.833	0.167	0.878	0.850	0.810	0.828	0.810	0.860
RF	0.764	0.236	0.833	0.817	0.670	0.729	0.670	0.810
SVM	0.500	0.500	0.500	0.500	1.000	0.667	1.000	0.000

Remarks: Decision Tree (DT), Deep Learning (DL), Fast Large Margin (FLM), Generalized Linear Model (GLM), Gradient Boosted Trees (GBT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM)

Discussion

Accurate diagnosis of ccRCC may provide the opportunity for appropriate treatment. Therefore, developing a novel diagnostic biomarker for ccRCC is needed as it can assist in the existing clinical management of this cancer. This study combined bioinformatics analysis and machine learning were used to determine the effective biomarkers for classifying ccRCC based on eight candidate ccRCC markers.

Table 4 The weight of each gene in different models

Attribute	NB	GLM	LR	FLM	DL	DT	RF	GBT	SVM
AMACR	0.029	0.007	0.113	0.076	0.129	0.029	0.015	0.085	2.26E-06
CAIX	0.090	0.172	0.133	0.182	0.226	0.020	0.068	0.019	9.34E-07
CD10	0.122	0.262	0.281	0.317	0.148	0.018	0.067	0.056	5.35E-06
CD117	0.031	0.006	0.017	0.088	0.048	0.017	0.019	0.012	2.53E-06
GATA3	0.396	0.227	0.159	0.049	0.184	0.015	0.103	0.054	1.71E-06
KRT7	0.028	0.034	0.111	0.141	0.060	0.037	0.037	0.025	8.44E-08
PAX8	0.058	0.278	0.204	0.039	0.180	0.232	0.142	0.077	2.60E-06
VIM	0.090	0.014	0.087	0.184	0.041	0.012	0.023	0.029	1.83E-06

Remarks: Decision Tree (DT), Deep Learning (DL), Fast Large Margin (FLM), Generalized Linear Model (GLM), Gradient Boosted Trees (GBT), Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), and Support Vector Machine (SVM), Paired-box gene 8 (PAX8), Cluster of differentiation 10 (CD10), Carboxinic anhydrase IX (CAIX), Vimentin, Cytokeratin 7 (CK7), Receptor tyrosine kinase (CD117/KIT), α-Methylacyl coenzyme A racemase (AMACR), GATA binding protein 3 (GATA3).

Our study compared nine ML algorithms using GEO data and found that NB and DT can better predict the ccRCC patients. The highest accuracy was found in NB algorithms. The weight score from NB model, the gene with the highest weight was GATA3, followed by CD10, VIM, CAIX, and PAX8. However, NB model uses multiple genes to classify ccRCC. In contrast, the DT model produces results that are easy to understand in terms of the predictor variables and target. In this present study, DT model also got adequate accuracy performance. DT model used only one gene, namely PAX8, to distinguish ccRCC from normal samples.

To confirm the reliability of the analyzed results, the expression levels and clinical correlation of selected genes, including GATA3 and PAX8, were further examined. The results elucidated strong evidence to support the analysis that both GATA3 and PAX8 were significantly down- regulated in ccRCC when compared with normal controls even though the analysis was performed on different databases and different platforms.

GATA3 is a transcription factor of the GATA family. GATA3 plays multiple roles such as regulation of MUC1/EMA genes, that involved in differentiation of breast epithelium, related to development of T-cell, skin, hair, trophoblast and blood vessel (Miettinen et al., 2014). In the surgical pathology field, GATA3 is commonly used as a marker for breast and urothelial carcinomas. From the previous report, GATA3 is useful biomarker of breast cancer, 80-90% of primary and metastatic mammary carcinomas has GATA3 expression (Cimino-Mathews et al., 2013). GATA3 also showed high level expression in urothelial tumors by cDNA array. GATA3 protein expression may be utility in the differentiation of urothelial carcinomas from other genitourinary neoplasms in the differential diagnosis (Higgins et al., 2007). In addition,

GATA3 activation may inhibit IL6/STAT3 signaling inhibition, thereby reducing migration in both renal and renal carcinoma cells (Shi et al., 2020).

Table 5 The performance values of PAX8 and GATA3.

	PAX8 based on Decision tree		GATA3 based on Naive bayes	
	Value	SD	Value	SD
Accuracy	0.739	0.159	0.858	0.099
Classification error	0.261	0.159	0.142	0.099
AUC	0.708	0.178	0.861	0.095
Precision	0.720	0.159	0.870	0.120
Recall	0.810	0.207	0.860	0.129
F-measure	0.753	0.161	0.859	0.095
Sensitivity	0.810	0.207	0.860	0.129
Specificity	0.670	0.211	0.860	0.129

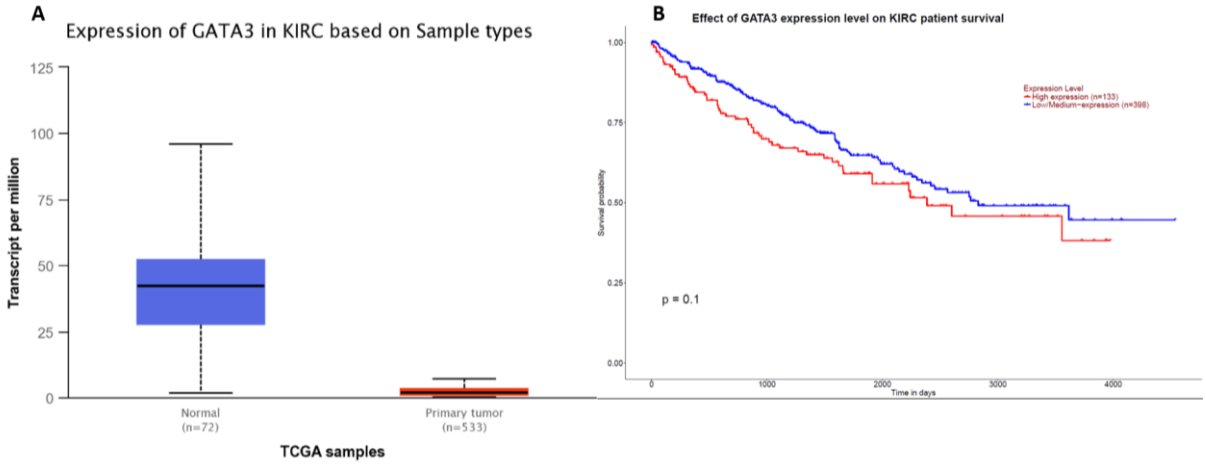

Figure 2 Expression levels of GATA3 and correlation between GATA3 expression and cumulative survival rate in ccRCC based on TCGA data. (A) Expression of GATA 3 in ccRCC sample versus normal, $p < 0.001$. (B) The Kaplan Meier plot showed effect of GATA 3 gene expression on patient survival, $p = 0.1$. kidney renal clear cell carcinoma (KIRC)

PAX8 is transcription factor that are essential for the development of kidney, müllerian, and other organs. PAX8 is expressed in normal kidney and most renal neoplasms and has similar expression in RCC, ovarian, endometrial carcinoma, thyroid follicular cells, and thyroid carcinoma (Shen et al., 2012). A previous study found PAX8 may be a specific and sensitive marker for renal cell and ovarian carcinomas. PAX8 showed positive stained 100% in normal kidney and 90% of Renal cell carcinomas. Furthermore, PAX 8 also positive stain in ovarian cancers, thyroid cancer, endometrial cancers, cervical

adenocarcinoma but negative stain in cervical squamous cell carcinomas, lung cancer, cancers of the colon, breast, prostate, liver, testicular, stomach, esophagus, melanoma, gastrointestinal stromal tumors, leiomyosarcoma, and pheochromocytoma (Tacha et al., 2011). PAX8 has shown to be a helpful immunohistochemical marker in surgical pathology, with a wide range of diagnostic uses (Ordóñez, 2012).

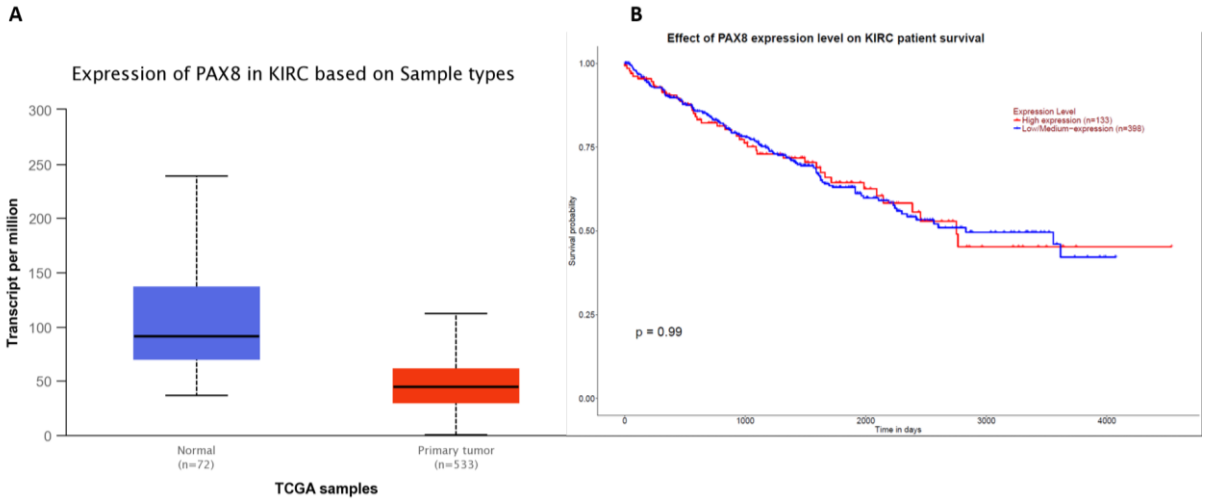

Figure 3 Expression levels of PAX8 and correlation between PAX8 expression and cumulative survival rate in ccRCC based on TCGA data. (A) Expression of PAX8 in ccRCC sample versus normal, $p < 0.001$. (B) The Kaplan Meier plot showed the effect of PAX8 gene expression on patient survival, $P = 0.99$. kidney renal clear cell carcinoma (KIRC)

Conclusion

The present study provided potential diagnostic biomarkers for ccRCC from computational analysis. The reliability of computational analysis results based on the microarray dataset from GEO was confirmed in different databases (TCGA) and different platforms (RNA-seq). The results demonstrated exciting and strong evidence that GATA3 and PAX8 may be used as effective diagnostic biomarkers for ccRCC.

Acknowledgment

This work was supported by a Postgraduate Study Support Grant of Faculty of Medicine, Khon Kaen University.

References

Akgul M, Cheng L. Immunophenotypic and pathologic heterogeneity of unclassified renal cell carcinoma: A study of 300 cases. Hum Pathol. 2020 Aug; 102: 70-78.

Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BVSK, Varambally S. UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. Neoplasia. 2017 Aug; 19(8): 649-658.

Cheville JC, Lohse CM, Zincke H, Weaver AL, Blute ML. Comparisons of Outcome and Prognostic Features Among Histologic Subtypes of Renal Cell Carcinoma. Am j Surg Pathol 2003 May; 27(5): 612-24.

Cimino-Mathews A, Subhawong AP, Illei PB, Sharma R, Halushka MK, Vang R, Fetting JH, Park BH, Argani P. GATA3 expression in breast carcinoma: Utility in triple-negative, sarcomatoid, and metastatic carcinomas,. Hum Pathol. 2013 Jul; 44(7): 1341-9.

Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Inform. 2007 Feb 11; 2: 59-77.

Davis S, Meltzer PS. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatic. 2007 Jul 15; 23(14): 1846-7.

Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1; 30(1): 207-10.

Higgins JP, Kaygusuz G, Wang L, Montgomery K, Mason V, Zhu SX, Marinelli RJ, Presti JC Jr, van de Rijn M, Brooks JD. Placental S100 (S100P) and GATA3: markers for transitional epithelium and urothelial carcinoma discovered by complementary DNA microarray. Am J Surg Pathol. 2007 May; 31(5): 673-80.

Joseph T. Rabban, Robert A. Soslow and Charles Z. Zaloudek. Chapter 18—Immunohistology of the Female Genital Tract. Diagnostic Immunohistochemistry (Third Edition). W.B. Saunders. 2020: 690-762.

Jungermann F. (2009). Information Extraction with RapidMiner. In Proceedings of the GSCL Symposium'Sprachtechnologie und eHumanities, pp. 50-61. 2009.

Kabir MF, Ludwig SA. Enhancing the Performance of Classification Using Super Learning. Data-Enabled Discov. Appl. 3, 5; 2019.

Lu D, Jiang J, Liu X, Wang H, Feng S, Shi X, Wang Z, Chen Z, Yan X, Wu H, Cai K. Machine Learning Models to Predict Primary Sites of Metastatic Cervical Carcinoma From Unknown Primary. Front Genet. 2020 Dec 21; 11: 614823.

Miettinen M, McCue PA, Sarlomo-Rikala M, Rys J, Czapiewski P, Wazny K, Langfort R, Waloszczyk P, Biernat W, Lasota J, Wang Z. GATA3: a multispecific but potentially useful marker in surgical pathology: a systematic analysis of 2500 epithelial and nonepithelial tumors. Am J Surg Pathol. Am J Surg Pathol. 2014 Jan; 38(1): 13-22.

Ngan KW, Ng KF, Chuang CK. Solid variant of papillary renal cell carcinoma. Chang Gung Med J. 2001 Sep; 24(9): 582-6.

Ordóñez NG. Value of PAX 8 immunostaining in tumor diagnosis: A review and update. Adv Anat Pathol. 2012 May; 19(3): 140-51.

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20; 43(7): e47.

Shen SS, Truong LD, Scarpelli M, Lopez-Beltran A. Role of immunohistochemistry in diagnosing renal neoplasms: When is it really useful? Arch Pathol Lab Med. 2012 Apr; 136(4): 410-7.

Shi Q, Xu R, Song G, Lu H, Xue D, He X, Xia Y. GATA3 suppresses human fibroblasts-induced metastasis of clear cell renal cell carcinoma via an anti-IL6/STAT3 mechanism. Cancer Gene Ther. 2020 Dec; 27(12): 979-982.

Tacha D, Zhou D, Cheng L. Expression of PAX8 in normal and neoplastic tissues: A comprehensive immunohistochemical study. Appl Immunohistochem Mol Morphol. 2011 Jul; 19(4): 293-9.

Tomczak K, Czerwiń́ska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Contemp Oncol (Pozn). 2015; 19(1A): A68-77.

Truong LD, Shen SS. Immunohistochemical diagnosis of renal neoplasms. Arch Pathol Lab Med. 2011 Jan; 135(1): 92-109.