A miRNA- and mRNA-seq-Based Feature Selection Approach for Kidney Cancer Biomakers

Cancer Informatics
Volume 19: 1–6
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1176935120908301

Shinuk Kim

Department of Civil Engineering, Sangmyung University, Cheonan, Republic of Korea.

ABSTRACT: Microarray data sets have been used for predicting cancer biomarkers. Yet, replication of the prediction has not been fully satisfied. Recently, new data sets called deep sequencing data sets have been generated, with an advantage of less noise in computational analysis. In this study, we analyzed the kidney miRNA and mRNA sequence data sets for predicting cancer markers using 5 different statistical feature selection methods. In the results, we obtained 3 mRNA- and 27 miRNA-based cancer biomarkers to compare with the normal samples. In addition, we clustered the kidney cancer subtypes using a nonnegative matrix factorization method and obtained significant results of survival analysis from the 2 separate groups including miRNA-342 and its target eukaryotic translation initiation factor 5A (*EIF5A*).

KEYWORDS: mRNA- and miRNA-seq, feature selection, NMF clustering, survival analysis

RECEIVED: January 17, 2020. **ACCEPTED:** February 1, 2020.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Research Foundation of Korea (NRF-2017R1A2B4010684) granted by the Ministry of Science and ICT.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Shinuk Kim, Department of Civil Engineering, Sangmyung University, E514, Sangmyungdae-gil 34, Cheonan 31066, Chungnam, Republic of Korea. Email: kshinuk@smu.ac.kr

Introduction

Recently, next-generation sequencing data sets have been wildly used for genomic analysis[1-4] because of technical advantage of free from the probe-specific hybridization of microarray. MicroRNA (miRNA) and mRNA sequencing data sets have been applied to various diseases including kidney renal cell carcinoma.[5-7] Feature selection (FS) methods are essential for building a model such as classification and/or clustering to better predict biomarkers for cancer classifiers. Even though a number of previous studies have been attempted to suggest new FS methods, FS methods are still limited. In this study, we focused on FS methods to identify biomarkers using 5 different statistical methods: information gain,[8,9] gain ratio,[10,11] and symmetrical uncertainty,[12] Spearman rank correlation, and Pearson linear correlation. Information gain, gain ratio, and symmetrical uncertainty used the probability of the classes, whereas the Spearman rank correlation and linear correlation methods used expression values. Technically, each statistical method has its own unique advantage in predicting cancer biomarkers and a disadvantage of losing some critical information. To overcome the drawback of a single statistical method, we selected the overlapping biomarkers from 5 different statistical methods so that information obtained from gene expression was enhanced or enriched. In this study, mRNA/miRNA features were obtained to distinguish between tumors and normal specimens as well as between tumor clustering specimens.

Materials and Methods

Materials

We initially downloaded miRNA and mRNA of kidney cancer data sets from UCSC (https://genome-cancer.ucsc.edu/) in June 2017, uploaded in January 2015. The miRNA was generated from an illuminaHiSeq-miRNASeq platform, whereas the mRNA was generated from an illuminaHiSeq-RNASeqV2 platform: the former included 326 samples and 1046 genes, and the latter included 606 samples and 20 530 genes. The miRNA and mRNA sequencing data sets were matched by patient samples for both tumors and normal specimens. We obtained 71 normal and 255 tumor samples of both miRNA (with 202 genes) and mRNA (with 13 268 genes) after removing unreadable data sets.

Methods

In this study, we tested 5 different statistical methods to identify biomarkers for distinguishing tumors from normal specimens, information gain,[8,9] gain ratio,[10,11] symmetrical uncertainty,[12] Spearman rank correlation, and Pearson linear correlation from the R package, FSselector (https://cran.r-project.org/web/packages/FSelector/index.html).

Here, we briefly introduce the individual advantages of the 5 statistical methods. First, information gain is derived from the information content of a code $-\sum p_i \log(p_i)$, where p_i is the probability of i.

Information gain can also denote the difference between the entropy of 2 classes. Entropy, $H(x)$, is defined as:

$$H(X) = -E(\log P(X)) = -\sum_{x \in \chi} p(x) \log p(x).$$

where X is a finite set.[11,13] The weaker is the entropy, the stronger are the classifiers. Gain ratio[10,11,14] represents the ratio of information gain to split information. Therefore, the algorithm first is processed to split the samples with possible size:

$$S(X) = \sum_{i=1}^{n} p(i) \times H(p(i)),$$
$$G(X) = H(X) - S(X).$$

Table 1. Feature selection using 5 statistical methods.

GENES SELECTED USING mRNA–SEQ					
GENE	INFORMATION GAIN	GAIN RATIO	SYMMETRIC UNCERTAINTY	RANKING CORRELATION	LINEAR CORRELATION
SPAG4	44	48	46		50
PVT1	41		44	50	50
C1orf226	45	48	50		48
GENES SELECTED USING miRNA–SEQ WITH TARGET *SPAG4* (BOLD GENES) FROM TARGETSCAN AND COMPARED WITH A PREVIOUS STUDY.[16]					
hsa.mir.106b		*hsa.mir.210*	***hsa.mir.199a.2***	*hsa.mir.429*	
hsa.mir.155		*hsa.mir.224*	*hsa.mir.199b*	***hsa.mir.452***	
hsa.mir.15a		***hsa.mir.25***[16]	*hsa.mir.200b*[16]	***hsa.mir.584***[17]	
hsa.mir.181a.1		***hsa.mir.28***	*hsa.mir.200c*[16]	***hsa.mir.629***	
hsa.mir.181b.1		***hsa.mir.362***	*hsa.mir.21*[16]	***hsa.mir.93***	

Split information is calculated by:

$$SI(X) = -\sum_{i=1}^{n} p(i) \times \log_2(p(i)).$$

Finally, gain ratio is written as:

$$GR(X) = \frac{G(X)}{SI(X)}.$$

The attribute with the maximum gain ratio is selected as the splitting attribute. As both information gain and gain ratio are used as univariate attributes, the advantage of the method is its short computational time with independent classifiers. In addition, correlation-based methods (such as Spearman rank and linear correlations) are used for multivariate attributes, with the computational time being slower than the information gain with dependent features.[10]

The advantage of symmetric uncertainty (SU)[12] is the reduction in the number of comparisons because $SU(x, y) = SU(y, x)$, where x and y are independent variables.[12]

$SU(X,Y) = 2[IG(X|Y)/(H(X)+H(Y))]$, where $IG(X|Y)$ is noted as $H(X) - H(X|Y)$, which is the information gain of feature X, which is an independent attribute; and Y describes the class. $H(X)$ and $H(Y)$ are entropy factors of features X and Y, respectively.[15]

The advantage of the Spearman rank correlation method is that it is a nonparametric (distribution-free) statistical method that measures the strength of association between variables. The Pearson correlation method is a parametric statistical model computed by covariance of the 2 variables divided by the product of their standard deviation.[15]

As the normal sample size is much smaller than the tumor sample size, all computational processes are based on balanced sample sizes. In the processes, we randomly selected the tumor samples to match the normal samples and executed the 5 statistical methods 50 times. Consequently, we selected a total of 2500 features from each method.

Results

mRNA/miRNA features of tumor vs normal tissues

We compared tumor vs normal tissues and aggregated the genes selected more than 40 times out of 50 runs and discovered information from 4 out of the 5 methods in Table 1. In each method, the probability of a hypergeometric test was less than 5.7e−28.

Human sperm–associated antigen 4 (*SPAG4*) was strongly suggested as a potential cancer marker.[18,19] Knaup et al[19] discovered that *SPAG4* was upregulated in human renal clear cells, and *SPAG4* knockdown reduced the growth of renal tumors in vitro.

Recently, Cui et al[20] discovered that plasmacytoma variant translocation 1 (*PVT1*) was related to well-known cancer region 8q24. Many studies have revealed that non-protein coding RNA, which is roughly divided into 2 groups based on size, plays important roles in cancer. One involves the short noncoding group that consists of less than 200 nucleotides in length, whereas the long noncoding group is made up of more than 200 nucleotides.[21] As most studies have focused on short noncoding RNA, such as miRNA,[22] it is well understood compared with long noncoding genes. Therefore, our findings on the long noncoding gene *PVT1* determined that it is a meaningful candidate oncogene. Chromosome 1 open reading frame 226 (*C1orf226*) is a protein-coding gene.

Tumor clustering

Analysis of mRNA tumor samples. Among the 255 tumor samples of mRNA-seq, we separated the samples using nonnegative matrix factorization (NMF)[23] methods between groups (k)

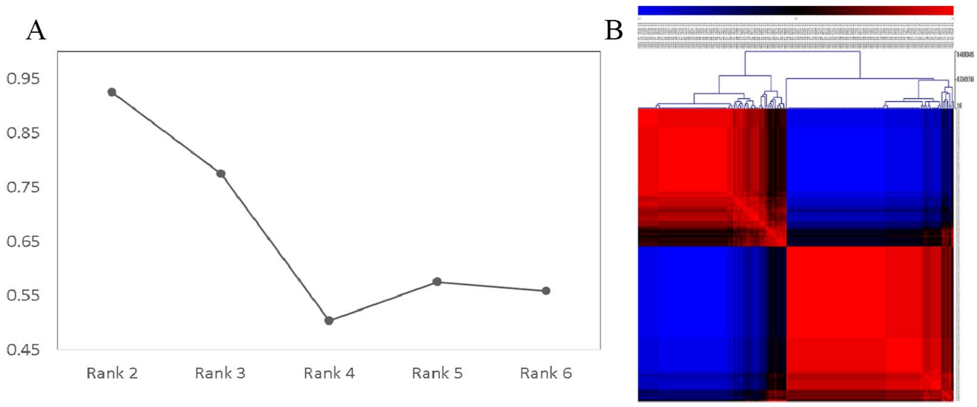

Figure 1. (A) Cophenetic coefficients of rank 2 to rank 6. (B) NMF clustering with $k=2$. NMF indicates negative matrix factorization.

Table 2. Genes selected from 5 statistical methods using based on 2 groups.

GENE	INFORMATION GAIN	GAIN RATIO	SYMMETRIC UNCERTAINTY	RANKING CORRELATION	LINEAR CORRELATION
UTP14C	50	50	50	50	50
USP8	47	48	50	50	50
FBXL6	50	45	49	50	50

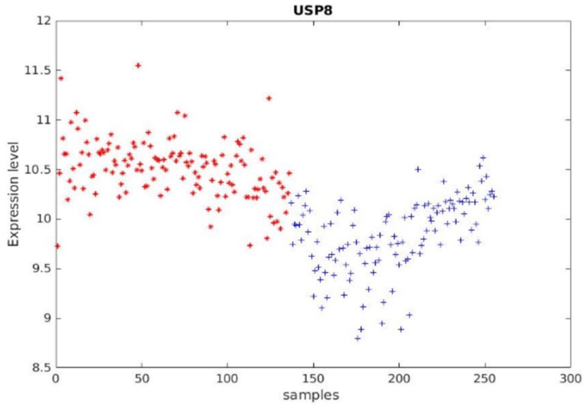

Figure 2. Comparison of expression levels of *USP8* between cluster 1 (red dots) and cluster 2 (blue dots).

$k=2$ to $k=6$. When $k=2$, cophenetic coefficients are the most ideal, with a value of 0.92, in Figure 1. When k values are equal to 3, 4, 5, and 6, cophenetic coefficients are 0.77, 0.50, 0.57, and 0.56, respectively.

A total of 255 tumor samples were separated into 2 groups of 119, denoted as cluster 1, and 136, denoted as cluster 2. We selected features using those 2 separated tumor groups with 5 statistical methods. The selected features were *UTP14C, USP8*, and *FBXL6*, which were found from the 5 different methods with 40 appearances out of 50 runs in Table 2. Figure 2 shows *USP8* expression levers between 2 classes, with t-test rejected the null hypothesis with $P \cong 2.08\text{e}{-25}$.

Analysis of miRNA tumor samples. Among the 255 tumor samples of miRNA-seq, 8 had missing information. Therefore, we used a final total of 246 samples. We tested the NMF method from $k=2$ to 6 for the cluster. The data were separated into 2 groups, because the highest cophenetic coefficient was obtained as 0.93 when k was equal to 2. When k values are equal to 3, 4, 5, and 6, cophenetic coefficients are 0.79, 0.68, 0.61, and 0.48, respectively. One group denoted as cluster 1 consisted of 124 samples, whereas the other group denoted as cluster 2 consisted of 122 samples. The Kaplan-Meier survival analysis[24] is shown in Figure 3 with significance (<<.001) using IBM SPSS Statistics 20.

A total of 27 miRNA genes, presented in Table 3, were shown to be differentially expressed between the 2 clusters using the 5 statistical methods with 50 runs for each method. The selected genes were more robust than those of the single method because all the genes were selected in all iterations and methods.

In addition, Table 3 includes miRNAs' target mRNA genes that are demonstrated in Table 2 using TargetScan (http://www.targetscan.org).

Analysis using common samples of mRNA and miRNA. As the tumor samples contained different miRNA- and mRNA-seq-based clusters, we selected the common samples from each of the 2 groups in Figure 4 (67 and 76 common samples) to identify enhanced biomarkers.

Table 4 presents mRNA genes selected from more than 40 appearances out of 50 runs for all 5 statistical methods comparing cluster 1 with cluster 2. All 5 selected genes rejected the null hypothesis based on t-test with P-values, and presented in supporting file. We also selected 30 miRNA genes and described the target mRNA from TargetScan with values less than the Pearson coefficient. Interestingly, *LIFR* is underexpressed in

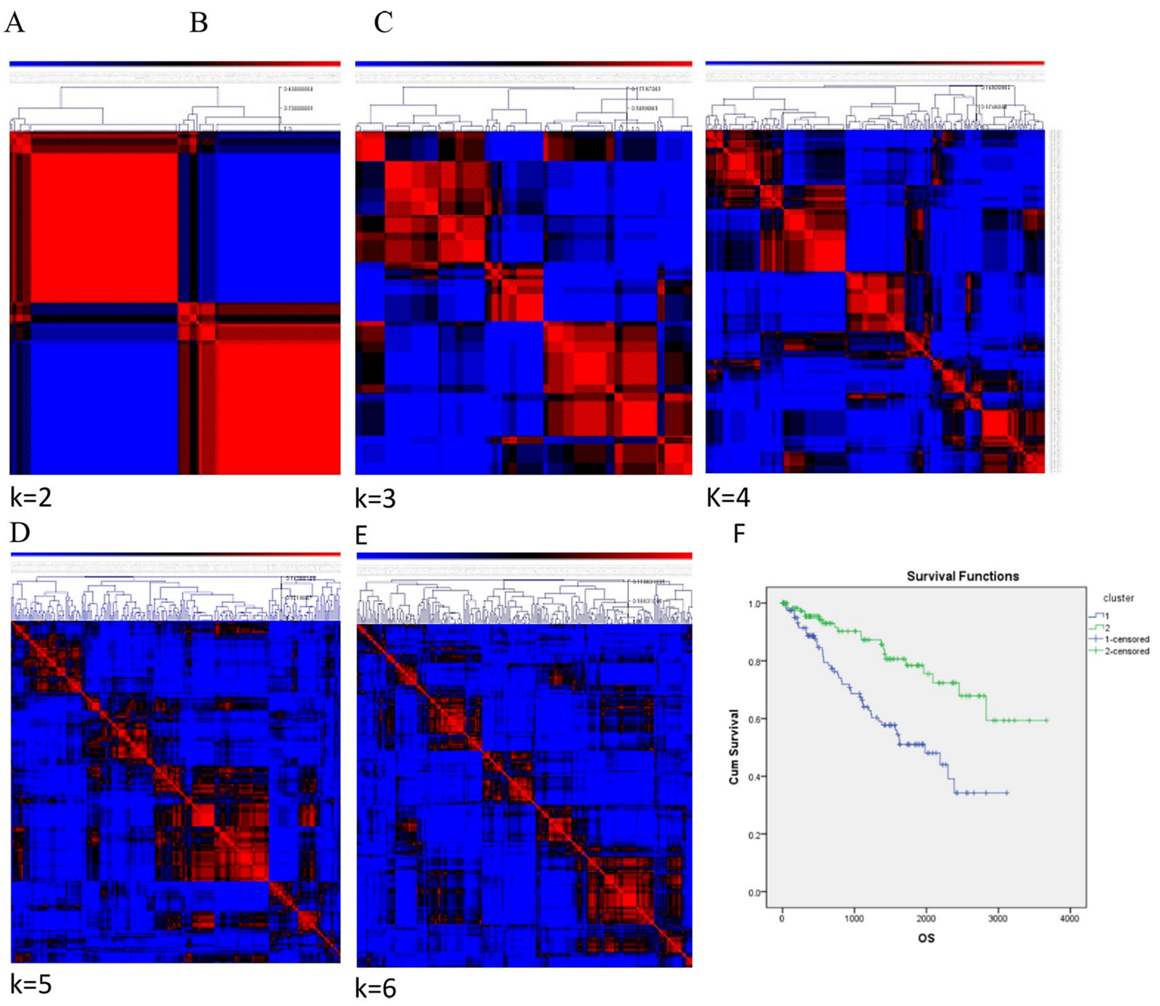

Figure 3. NMF clustering with $k=2$ to 6 (A-E) and (F) Kaplan-Meier survival analysis of tumor cluster used by miRNA-seq with $k=2$. NMF indicates negative matrix factorization.

Table 3. Genes selected from cluster 1 vs cluster 2 used by miRNA-seq, and their target mRNAs were identified from TargetScan.

	UTP14C	*USP8*	*FBXL6*
hsa.mir.101.1	○		
hsa.mir.1301			
hsa.mir.130b	○		
hsa.mir.142	○		
hsa.mir.146b	○		○
hsa.mir.155			
hsa.mir.16.1			
hsa.mir.16.2		○	○
hsa.mir.191		○	○
hsa.mir.193b	○		
hsa.mir.29b.1	○		○
hsa.mir.29b.2	○		
hsa.mir.301a	○		
hsa.mir.331	○	○	○
hsa.mir.339	○		
hsa.mir.342	○		
hsa.mir.34a	○		
hsa.mir.3607	○	○	○
hsa.mir.3647		○	○
hsa.mir.3653	○		○
hsa.mir.425	○	○	○
hsa.mir.484	○		
hsa.mir.590	○	○	○
hsa.mir.671		○	○
hsa.mir.766			
hsa.mir.9.1	○		
hsa.mir.9.2	○		

(Continued)

Table 3. (Continued)

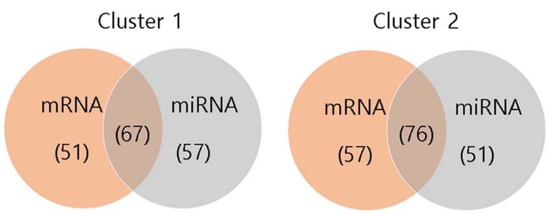

Figure 4. Overlapping samples between miRNA and mRNA clusters.

cluster 1; otherwise, *TAF10, NUDT1, B3GNTL1*, and eukaryotic translation initiation factor 5A (*EIF5A*) are overexpressed in cluster 1.

Figure 5 shows Kaplan-Meier survival analysis, which interpreted that the mortality of patients in cluster 1 occurred at a much faster rate than patients in cluster 2. We calculated the Pearson correlation coefficient of selected miRNA and mRNA from cluster 1. We only considered correlation values less than

Table 4. Genes selected from common samples of both mRNA- and miRNA-seq data sets with Pearson correlation *P*-values (<–.2).

GENES SELECTED FROM mRNA-SEQ						
GENE	INFORMATION GAIN	GAIN RATIO	SYMMETRIC UNCERTAINTY	RANKING CORRELATION	LINEAR CORRELATION	*P*-VALUE
TAF10	50	49	50	50	50	1.4e–30
NUDT1	50	50	50	50	50	2.2e–28
B3GNTL1	44	47	50	50	50	4.1e–27
LIFR	48	50	50	47	50	8.0e–29
EIF5A	50	50	50	50	50	1.0e–25
GENES SELECTED FROM miRNA-SEQ WITH THEIR TARGET MRNAS						
hsa.mir.101.1		hsa.mir.142	hsa.mir.18a –.242 (*NUDT1*)	hsa.mir.331 –.205 (*TAF10*)	hsa.mir.365.1 –.201 (*B3GNT*)	hsa.mir.590
hsa.mir.101.2		hsa.mir.146b	hsa.mir.193b	hsa.mir.339 –.2014 (*TAF10*)	hsa.mir.365.2	hsa.mir.625 –.2532 (*TAF10*)
hsa.mir.1301		hsa.mir.155	hsa.mir.21 –.224 (*NUT1*) –.256 (*B3GNT*)	hsa.mir.342 –.288 (*B3GNT*) –.306 (*EIF5A*)	hsa.mir.425	hsa.mir.671 –.255 (*TAF10*) –.210 (*NUDT1*) –.250 (*B3GNT*) –.251 (*EIF5A*)
hsa.mir.130b –.216 (*TAF10*) –.229 (*EIF5A*)		hsa.mir.16.1	hsa.mir.29b.1 –.225 (*NUDT1*)	hsa.mir.34a	hsa.mir.454 –.305 (*TAF10*) –.227 (*NUDT1*) –.274 (*B3GNT*) –.205 (*EIF5A*)	hsa.mir.9.1
hsa.mir.139		hsa.mir.16.2	hsa.mir.29b.2 –.217 (*NUDT1*)	hsa.mir.3647	hsa.mir.484	hsa.mir.9.2

Abbreviation: eukaryotic translation initiation factor 5A.

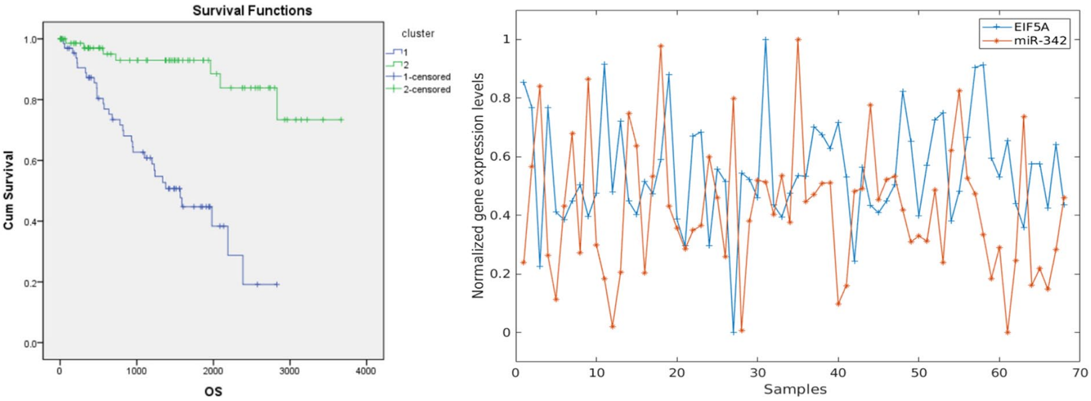

Figure 5. (A) Kaplan-Meier survival analysis used by common samples of both miRNA- and mRNA-seq. (B) Plot for inhibition between miRNA-342 and *EIF5A* (eukaryotic translation initiation factor 5A) used by cluster 1.

−.2 because the miRNA inhibited its target mRNA, as presented in the Table 4.

We uncovered pairs of miRNA and mRNA that had significantly different relations compared with the 2 clusters. We present a graph of hsa.mir.342 and its target gene *EIF5A*, discovered by calculating the Pearson correlation (−.3058) used by cluster 1 data sets.

Conclusions

In this study, we tested 5 different statistical methods for selecting enhanced significant cancer biomarkers using miRNA and mRNA sequence data sets of kidney cancer. We presented 3 mRNA and 27 miRNA markers for predicting cancer compared with the normal samples. In addition, we clustered the kidney tumors samples using miRNA and mRNA data sets independently and obtained 2 separate groups from both miRNAs and mRNAs. After matching the cluster samples, a total of 67 samples were contained in one group called cluster 1, and 76 samples were contained in cluster 2. According to the Kaplan-Meier analysis, the subtypes of kidney cancer were strongly related to mortality. We suggest the 5 strong candidate genes *TAF10, NUDT1, B3GNTL1, LIFR*, and *EIF5A* and 30 miRNAs that are differentially expressed between 2 subtypes in tumor samples related to mortality. Our enhanced methods discovered *B3GNT1*, whereas the rest of them were presented in https://www.proteinatlas.org.[25] In addition, we discovered 21 pairs of miRNAs and their target mRNAs including miR-342 and its target *EIF5A*.

Acknowledgement

The authors thank Dohyun Kim for performing the technical algorithm.

Author Contributions

SK designed and performed research, analyzed data, and wrote the paper.

ORCID iD

Shinuk Kim https://orcid.org/0000-0002-0065-2581

REFERENCES

1. Alkhateeb A, Rezaeian I, Singireddy S, Cavallo-Medved D, Porter LA, Rueda L. Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer. *Cancer Inform*. 2019;18:1176935119835522. doi:10.1177/1176935119835522.
2. Andres-Leon E, Nunez-Torres R, Rojas AM. miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci Rep*. 2016;6:25749. doi:10.1038/srep25749.
3. Zhan C, Yan L, Wang L, et al. Identification of reference miRNAs in human tumors by TCGA miRNA-seq data. *Biochem Biophys Res Commun*. 2014;453:375-378. doi:10.1016/j.bbrc.2014.09.086.
4. Lin KH, Huang MY, Cheng WC, et al. RNA-seq transcriptome analysis of breast cancer cell lines under shikonin treatment. *Sci Rep*. 2018;8:2672. doi:10.1038/s41598-018-21065-x.
5. Pal SK, He M, Tong T, et al. RNA-seq reveals aurora kinase-driven mTOR pathway activation in patients with sarcomatoid metastatic renal cell carcinoma. *Mol Cancer Res*. 2015;13:130-137. doi:10.1158/1541-7786.MCR-14-0352.
6. Li P, Conley A, Zhang H, Kim HL. Whole-transcriptome profiling of formalin-fixed, paraffin-embedded renal cell carcinoma by RNA-seq. *BMC Genomics*. 2014;15:1087. doi:10.1186/1471-2164-15-1087.
7. Zwiener I, Frisch B, Binder H. Transforming RNA-Seq data to improve the performance of prognostic gene signatures. *PLoS ONE*. 2014;9:e85150.
8. Kent JT. Information gain and a general measure of correlation. *Biometrika*. 1983;70:163-173.
9. Johnsson J. Contracting: how to gain the information edge over HMOs. *Hospitals*. 1992;66:44, 6, 8.
10. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23:2507-2517. doi:10.1093/bioinformatics/btm344.
11. Krishnaiah PR, Kanal LN, eds. *Classification Pattern Recognition and Reduction of Dimensionality*. vol. 2. Amsterdam, The Netherlands: North-Holland; 1982.
12. Ali SI, Shahzad W, eds. A feature subset selection method based on symmetric uncertainty and ant colony optimization. 2012 International Conference on Emerging Technologies (ICET); October 8-9, 2012; Islamabad, Pakistan. New York, NY: IEEE.
13. Fan R, Zhong M, Wang S, et al. Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genet Epidemiol*. 2011;35:706-721. doi:10.1002/gepi.20621.
14. Koyama S, Kostal L. The effect of interspike interval statistics on the information gain under the rate coding hypothesis. *Math Biosci Eng*. 2014;11:63-80. doi:10.3934/mbe.2014.11.63.
15. Hauke J, Kossowski T. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaest Geogr*. 2011;30:87-93.
16. Youssef YM, White NM, Grigull J, et al. Accurate molecular classification of kidney cancer subtypes using microRNA signature. *Eur Urol*. 2011;59:721-730. doi:10.1016/j.eururo.2011.01.004.
17. Ueno K, Hirata H, Shahryari V, et al. Tumour suppressor microRNA-584 directly targets oncogene Rock-1 and decreases invasion ability in human clear cell renal cell carcinoma. *Br J Cancer*. 2011;104:308-315. doi:10.1038/sj.bjc.6606028.
18. Kennedy C, Sebire K, de Kretser DM, O'Bryan MK. Human sperm associated antigen 4 (*SPAG4*) is a potential cancer marker. *Cell Tissue Res*. 2004;315:279-283. doi:10.1007/s00441-003-0821-2.
19. Knaup KX, Monti J, Hackenbeck T, et al. Hypoxia regulates the sperm associated antigen 4 (*SPAG4*) via HIF, which is expressed in renal clear cell carcinoma and promotes migration and invasion in vitro. *Mol Carcinog*. 2014;53:970-978. doi:10.1002/mc.22065.
20. Cui M, You L, Ren X, Zhao W, Liao Q, Zhao Y. Long non-coding RNA *PVT1* and cancer. *Biochem Biophys Res Commun*. 2016;471:10-14. doi:10.1016/j.bbrc.2015.12.101.
21. Fang Y, Fullwood MJ. Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics Proteomics Bioinformatics*. 2016;14:42-54. doi:10.1016/j.gpb.2015.09.006.
22. Adhami M, MotieGhader H, Haghdoost AA, Afshar RM, Sadeghi B. Gene co-expression network approach for predicting prognostic microRNA biomarkers in different subtypes of breast cancer. *Genomics*. 2020;112:135-143. doi:10.1016/j.ygeno.2019.01.010.
23. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999;401:788-791. doi:10.1038/44565.
24. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. *Int J Ayurveda Res*. 2010;1:274-278. doi:10.4103/0974-7788.76794.
25. Uhlen M, Fagerberg L, Hallstrom BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347:1260419. doi:10.1126/science.1260419.