Identification of miRNA signatures for kidney renal clear cell carcinoma using the tensor-decomposition method

Ka-Lok Ng [1, 2] and Y-h Taguchi [3,*]

[1] Department of Bioinformatics and Medical Engineering Asia University, Taichung, Taiwan; ppiddi@gmail.com
[2] Department of Medical Research, China Medical University Hospital China Medical University, Taiwan
[3] Department of Physics Chuo University, 1-13-27 Kasuga Bunky-ku, Tokyo 112-8551, Japan; tag@granular.com
* Correspondence: tag@granular.com, ORCID ID: 0000-0003-0867-8986

Abstract:

Purpose Cancer is a highly complex disease caused by multiple genetic factors. MicroRNA (miRNA) and mRNA expression profiles are useful for identifying prognostic biomarkers for cancer. The kidney renal clear cell carcinoma (KIRC) was selected for our analysis, because KIRC accounts for more than 70% of all renal malignant tumor cases.

Methods Traditional methods of identifying cancer prognostic markers may not be accurate. Tensor decomposition (TD) is a useful method uncovering the underlying low-dimensional structures in the tensor. TD-based unsupervised feature extraction method was applied to analyze mRNA and miRNA expression profiles. Biological annotations of the prognostic miRNAs and mRNAs were examined by utilizing pathway and oncogenic signature databases, i.e. DIANA-miRPath and MSigDB.

Results TD identified the miRNA signatures and the associated genes. These genes were found to be involved in cancer-related pathways and 23 genes were significantly correlated with the survival of KIRC patients. We demonstrated that the results are robust and not highly dependent upon the database we selected. Compare to the t-test, we shown that TD achieves a much better performance in selecting prognostic miRNAs and mRNAs.

Conclusion These results suggest that integrated analysis using the TD-based unsupervised feature extraction technique is an effective strategy for identifying prognostic signatures in cancer studies.

Keywords: cancer biomarkers, diagnostic markers, prognostic markers, microRNA signatures, kidney cancer, tensor decomposition

Acknowledgments

Dr. Ka-Lok Ng is funded by the Ministry of Science and Technology, Taiwan (MOST), grant number MOST 108-2221-E-468-020, and also supported by the Asia University, grant numbers 107-asia-02 and 107-asia-09. Dr. Y-h Taguchi is supported by Kakenhi 19H05270 and 17K00417. We would like to thank Editage (www.editage.com) for English language editing.

1. Introduction

Cancer is a highly complicated and heterogeneous disease. It is the result of a loss of cell cycle control (Vargas-Rondon, Villegas, & Rondon-Lagos, 2017), which is due to accumulation of genetic mutations, gene duplication (Hanahan & Weinberg, 2011), and aberrant epigenetic regulation (Feinberg & Vogelstein, 1983; Rouhi, Mager, Humphries, & Kuchenbauer, 2008). Genetic mutations involving activation of proto-oncogenes to oncogenes (OCG) and inactivation of tumor-suppressing genes (TSG) may cause cancer by alternating transcription factors (TF), such as the *p53* and *ras* oncoproteins, which in turn control the expression of other genes. Gene duplication causes an elevated level of its protein product and thus favor the proliferation of cancer cells. MicroRNAs (miRNAs) are a class of small non-coding RNAs that bind to the messenger RNA (mRNA) and induce either its cleavage or impede translation repression. Several studies have indicated that abnormal miRNA expression is associated with carcinogenesis (Medina & Slack, 2008). miRNAs induce cancers by acting as oncogenes (OCG) and tumor suppressor genes (TSG). An miRNA that targets the mRNA of a TSG would induce loss of the protective effect of the TSG (Medina & Slack, 2008; Zhang, Dahlberg, & Tam, 2007). Although there have been many advancements in cancer therapy and diagnosis, many patients are unable to recover or experience recurrence after treatment. Accordingly, miRNA expression profiles are useful for identifying prognostic biomarkers for cancer diagnosis. For instance, dysregulated miRNAs were identified in urothelial carcinoma of the bladder (Inamoto et al., 2018). Recent studies also suggested that miRNAs could be used as a prognostic biomarker for patients with pancreatic adenocarcinoma (Shi et al., 2018; Yu, Feng, & Cang, 2018). Furthermore, by utilizing meta-analysis, it was reported that a panel of eight-miRNA signatures could serve as an effective marker for predicting overall survival in bladder cancer patients (Zhou et al., 2015). In this study, we selected kidney renal clear cell carcinoma (KIRC) for our analysis. KIRC is the most common cancer subtype of all renal malignant tumors, accounting for more than 70% of the cases (Zhang et al. 2013). Several studies have identified a few miRNA signatures that are associated with the overall survival of KIRC patients (Lokeshwar et al., 2018; Luo et al., 2019; Xie et al., 2018).

Typical data structures in bioinformatics are difficult to analyze because of the small number of samples with many variables. Supervised feature extraction are effective methods for reducing the number of features. If supervised learning is applied, overfitting can occur. Regularization (sparse modeling) attempts to minimize the number of features by restricting the sum of coefficients attributed to features and penalizes the use of additional variables. The disadvantage of regularization is that we must select the values of parameters that balance the prediction accuracy and the number of variables. There are two major issues with supervised feature extraction methods: (i) class labels may not always be true and (ii) there may be more class labels present in the dataset. However, unsupervised methods such as principal component analysis (PCA) are often used to generate a smaller number of variables through the linear combination of original variables. The problem with this approach is that the linear combination of many variables often prevents us from interpreting the newly generated variables. An unsupervised methodology that is suitable for the dimension reduction problems is tensor decomposition (TD)-based unsupervised feature extraction (FE) (Y. Taguchi, 2017; Y. Taguchi & Ng, 2018; Y.-h. Taguchi, 2019a, 2019b, 2019c; Y.-h. Taguchi & T. Turki, 2019; Y. H. Taguchi, 2017a, 2017b, 2017c, 2018a, 2018b, 2018c, 2019; Y. H. Taguchi & T. Turki, 2019). This method allows selection of a smaller number of variables effectively and stably.

2. Materials and Methods

2.1 Tensors and tensor decomposition (TD)

Tensor [17] is a mathematical structure for storing datasets associated with more than two properties. If we measure miRNA and mRNA expression for the samples, we cannot avoid storing these two measurements into two separate matrices. However, by using tensor we can store these

two datasets into a tensor, because tensors can have more than two suffixes, which matrices do not have.

TD [17] is a mathematical trick that can approximate tensors as the summation of series whose terms are expressed via the outer product of vectors, each of which represent individual property (in this specific example, these vectors correspond to mRNAs, miRNAs, and samples) .

2.2. Tensor decomposition method

The miRNAseq and mRNAseq expression data for KIRC were retrieved from the TCGA Data Portal Research Network (https ://gdcportal.nci.nih.gov/).

TD is a natural extension of matrix factorization, and is regarded as a generalization of the singular value decomposition (SVD) method. It is a useful technique uncovering the underlying low-dimensional structures in the tensor. There are two popular tensor decomposition algorithms: canonical polyadic decomposition (CPD) and Tucker decomposition (Rabanser, Shchur, & Günnemann, 2017). The rank decomposition method, CPD, is to express a tensor as the sum of a finite number of rank-one tensors. The Tucker decomposition decomposes a tensor into a so-called core tensor and multiple matrices.

TD-based unsupervised FE was applied to analyze mRNA and miRNA expression profiles. Let $x_{ij}^{(mRNA)}$ denote the expression profiles of the *i*th mRNA (*i* = 1, …*N*) of the *jth* sample (*j* = 1, … *M*), whereas $x_{kj}^{(miRNA)}$ denotes the expression profiles of the *k*th miRNA (*k* = 1, …*K*) of the *j*th sample (*j* = 1, … *M*). Both x_{ij} and x_{kj} will be standardized such that they are associated with zero mean and unit variance. Next, we generated a case II type I tensor, that is,

$$x_{ijk} = x_{ij(mRNA)} * x_{kj(miRNA)} \quad (1)$$

x_{ijk} is subjected to Tucker decomposition as follows:

$$x_{ijk} = \sum_{l_1=1}^{N} \sum_{l_2=1}^{M} \sum_{l_3=1}^{K} G(l_1, l_2, l_3)\, u_{l_1 i} u_{l_2 j} u_{l_3 k} \quad (2)$$

where $G \in R^{N\times M\times K}$ is the core tensor and $u_{l_1 i} \in R^{N\times N}$, $u_{l_2 j} \in R^{M\times M}$ and $u_{l_3 k} \in R^{K\times K}$ are singular value matrices that are orthogonal. Because Tucker decomposition is not unique, we have to specify how Tucker decomposition was derived. In particular, we chose higher-order singular value decomposition (HOSVD). Given that x_{ijk} is too large to apply TD, we generated a case II type II tensor, which is given by:

$$x_{ik} = \sum_{j=1}^{M} x_{ijk} \quad (3)$$

By applying SVD, we can get $u_{l_1 i}$ and $u_{l_3 k}$ as

$$x_{ik} = \sum_{l=l_1=l_3=1}^{\min(N,K)} \lambda_l\, u_{l_1 i} u_{l_3 k} \quad (4)$$

Then, we can also obtain two $u_{l_2 j}$ that correspond to miRNA and mRNA expression:

$$u_{l_1 j}^{mRNA} = \sum_{i=1}^{N} x_{ij} u_{l_1 i},\ u_{l_3 j}^{miRNA} = \sum_{k=1}^{K} x_{kj} u_{l_3 k}, \quad (5)$$

Selection of genes can be determined using the following quantities,

$$p_i = p_{\chi^2}\left[> \left(\frac{u_{l_1 i}}{\sigma_{l_1}}\right)^2\right], p_k = p_{\chi^2}\left[> \left(\frac{u_{l_3 k}}{\sigma_{l_3}}\right)^2\right] \quad (6)$$

where p_{χ^2}[>x] is the cumulative probability that the argument is greater than *x* in a χ^2 distribution.

σ_{l_1} and σ_{l_3} denote the standard deviations for $u_{l_1 i}$ and $u_{l_3 k}$, respectively. After the P-values are adjusted by means of the Benjamini–Hochberg (BH) criterion, miRNAs and mRNAs that are associated with adjusted P-values less than 0.01 are selected as those showing differences in expression

between controls (normal tissues) and treated samples (tumors).

2.3 mRNA and miRNA expression

Expression profiles of the mRNA and miRNA were retrieved from TCGA. The samples consisted of 253 kidney tumors and 71 normal kidney tissues (M = 324). The number of mRNAs measured was N = 19536, and the number of measured miRNAs was K = 825.

Another dataset was downloaded from GEO with GEO ID GSE16441, and two files, GSE16441-GPL6480_series_matrix.txt.gz (for mRNA) and SE16441-GPL8659_series_matrix.txt.gz (for miRNA) were used. A total of N = 33698 mRNAs and K = 319 miRNAs were measured for 17 patients and 17 healthy controls (M = 34).

2.4 Analysis of the correlation between miRNA and gene expression

Correlations between $u_{l_1 j}^{mRNA}$ and $u_{l_3 j}^{miRNA}$ ($l_1 = l_3$= 2) were quantified by the Pearson's correlation coefficient (*PCC*). The *PCC* and P-values were calculated using the *corr.function* and *cor.test* function in the R software, respectively.

2.5. Biological function analysis

We evaluated the biological significance of the set of differentially expressed miRNAs and their correlated mRNAs. Biological annotations of the prognostic miRNAs and mRNAs were examined by employing the DIANA-miRPath (Vlachos et al., 2015) and MSigDB (Liberzon et al., 2015) databases, respectively.

3. Results

We applied TD-based unsupervised FE to the KIRC dataset retrieved from TCGA. It was found that $u_{l_1 j}^{mRNA}$ and $u_{l_3 j}^{miRNA}$ ($l_1 = l_3$=2) varied between the normal and tumor samples. The t-test derived P-values were 7.10×10^{-39} for mRNA and 2.13×10^{-71} for miRNA, respectively. In order to see if u_{2j}^{mRNA} and u_{2j}^{miRNA} are significantly correlated, we computed the *PCC* between them, which was 0.905 (P = 1.63×10^{-121}), indicating that they are highly correlated.

The results of the miRNA signatures and their significant correlated genes are shown in Table 1. A total of 11 miRNAs and 72 genes were identified. To determine if these miRNAs and mRNAs are significantly correlated, we computed the *PCC* for all $11 \times 72 = 792$ pairs. Among them, 353 pairs were positively correlated and 358 pairs were negatively correlated (P-values were less than 0.01 after correcting with the BH criterion). Therefore, 90% of pairs are significantly correlated. Moreover, we could successfully identify significantly correlated pairs of miRNAs and mRNAs. We noted that among the predicted 11 miRNAs, one miRNA (miR-155) matched the result reported by Lokeshwar et al. (Lokeshwar et al., 2018).

Table 1. The results of the miRNA signatures and genes of KIRC patients based on the TD analysis.

miRNA ID					
hsa-mir-210	hsa-mir-891a	hsa-mir-155	hsa-mir-200c	hsa-mir-141	hsa-mir-508
hsa-mir-122	hsa-mir-514-3	hsa-mir-514-1	hsa-mir-514-2	hsa-mir-184	
Gene symbol					
ACTG1	ADAM6	AIF1L	ALDOA	ALDOB	ANGPTL4
APLP2	APP	AQP1	AQP2	ASS1	ATP1A1

ATP1B1	ATP5A1	ATP5B	B2M	C3	C4A
C7	CA12	CCND1	CD74	CDH16	COL4A1
COL4A2	CP	CYFIP2	ENO1	FN1	FTL
GAPDH	GATM	GNB2L1	GPX3	HLA-A	HLA-B
HLA-C	HLA-DRA	HSD11B2	HSP90AA1	HSPA8	IGFBP3
IGFBP5	ITM2B	KNG1	LDHA	LDHB	LOC96610
NDRG1	NDUFA4L2	NNMT	P4HB	PCK1	PEBP1
PLIN2	PLVAP	PODXL	RGS5	SERPINA1	SLC12A1
SLC12A3	SOD2	SPARC	SPP1	TGFBI	TMBIM6
TMSB10	UBC	UMOD	VEGFA	VIM	VWF

Next, in order to evaluate the biological significance of selected mRNAs, we determined the top 10 oncogenic signatures of the 72 genes reported by MSigDB (Table 2).

Table 2. The top 10 oncogenic signatures of the 72 genes reported by the MSigDB. #Genes (K): the number of genes in each overexpressed gene set. # Genes in overlap (*k*): overlaps with genes selected via the TD-based unsupervised FE method. .

Gene Set Name [# Genes (K)]	**Description**	**#Genes in overlap (*k*)**	**p-value**	**FDR q-value**
CAMP_UP.V1_UP [200]	Genes up-regulated in primary thyrocyte cultures in response to cAMP signaling pathway activation by thyrotropin (TSH).	7	9.97 e-8	1.88 e-5
SNF5_DN.V1_DN [168]	Genes down-regulated in MEF cells (embryonic fibroblasts) with knockout of SNF5 [Gene ID=6598] gene.	6	7.64 e-7	7.22 e-5
ESC_V6.5_UP_ LATE.V1_UP [188]	Genes up-regulated during the late stages of differentiation of embryoid bodies from V6.5 embryonic stem cells.	6	1.47 e-6	9.27 e-5
ESC_V6.5_UP_ EARLY.V1_DN [175]	Genes down-regulated during the early stages of differentiation of embryoid bodies from V6.5 embryonic stem cells.	5	1.98 e-5	8.54 e-4
ESC_J1_UP_ LATE.V1_UP [189]	Genes up-regulated during the late stages of differentiation of embryoid bodies from J1 embryonic stem cells.	5	2.86 e-5	8.54 e-4
SIRNA_EIF4GI_UP [95]	Genes up-regulated in MCF10A cells vs knockdown of the EIF4G1 [Gene ID=1981] gene by RNAi.	4	3.11 e-5	8.54 e-4
P53_DN.V1_DN [193]	Genes down-regulated in the NCI-60 panel of cell lines with mutated TP53 [Gene ID=7157].	5	3.16 e-5	8.54 e-4
MEL18_DN.V1_UP [141]	Genes up-regulated in DAOY cells (medulloblastoma) upon knockdown of PCGF2 [Gene ID=7703] gene by RNAi.	4	1.45 e-4	3.42 e-3
LTE2_UP.V1_UP [188]	Genes up-regulated in MCF-7 cells (breast cancer) positive for ESR1	4	4.33 e-4	8.51 e-3

	[Gene ID=2099] MCF-7 cells (breast cancer) and long-term adapted for estrogen-independent growth.			
RPS14_DN.V1_UP [190]	Genes up-regulated in CD34+ hematopoietic progenitor cells after knockdown of RPS14 [Gene ID=6208] by RNAi.	4	4.5 e-4	8.51 e-3

The results of the top 10 REACTOME pathways reported by MSigDB are summarized in Table 3.

Table 3. The top 10 oncogenic signatures of the 72 genes reported by the MSigDB. #Genes (K): the number of genes in each overexpressed gene set. # Genes in overlap (k): overlaps with genes selected by TD-based unsupervised FE method.

Gene Set Name [# Genes (K)]	Description	# Genes in overlap (k)	p-value	FDR q-value
REACTOME_REGULATION_OF_INSULIN_LIKE_GR_GROWTH_FACTOR_IGF_TRANSPORT_AND_UPTAK TAKE_BY_INSULIN_LIKE_GROWTH_FACTOR_BIN BINDING_PROTEINS_IGFBPS [124]	Regulation of insulin-like growth factor (IGF) transport and uptake by insulin-like growth factor binding proteins (IGFBPs)	12	9.03 e-18	1.35 e-14
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_ NE_SYSTEM [856]	Cytokine signaling within the immune system	18	1.85 e-14	1.39 e-11
REACTOME_RESPONSE_TO_ELEVATED_PLATELET LET_CYTOSOLIC_CA2PLUS [132]	Response to elevated platelet cytosolic Ca^{2+}	9	3.42 e-12	1.71 e-9
REACTOME_SIGNALING_BY_INTERLEUKINS [631]	Signaling by interleukins	13	1.53 e-10	4.86 e-8
REACTOME_INNATE_IMMUNE_SYSTEM [1104]	Innate immune system	16	1.62 e-10	4.86 e-8
REACTOME_PLATELET_ACTIVATION_SIGNALING ING_AND_AGGREGATION [260]	Platelet activation, signaling and aggregation	9	1.45 e-9	3.63 e-7
REACTOME_ENDOSOMAL_VACUOLAR_PATHWAY [11]	Endosomal/Vacuolar pathway	4	3.63 e-9	7.78 e-7
REACTOME_GLUCONEOGENESIS [34]	gluconeogenesis	5	5.22 e-9	9.79 e-7
REACTOME_POST_TRANSLATIONAL_PROTEIN_MO _MODIFICATION [1429]	Post-translational protein modification	16	6.56 e-9	1.09 e-6
REACTOME_DISEASE [1075]	Disease	14	1.02 e-8	1.53 e-6

These results suggest that the selected 72 mRNAs are likely related to oncogenesis. In order to further confirm if these 72 mRNAs are related to kidney cancer, we checked if these genes were linked to

survival rates (Table 4). Among 72 mRNAs, 23 were significantly correlated with the survival of kidney cancer patients. This also highlights the effectiveness of our analysis.

Table 4. Survival analysis of KIRC using OncoLnc (Anaya, 2016) (Kaplan plots are provided in the supplementary materials)

Gene	Cox Coeff.	P-value	FDR Corrected	Rank	Median Expression	Mean Expression	Kaplan plot Low (%)	Kaplan plot High (%)	Kaplan plot P-value
VWF	-0.3	1.90E-04	1.41E-03	2253	23278.72	25958.34	50	50	3.99E-03
VEGFA	0.25	2.90E-03	1.19E-02	4064	31629.77	35072.27	70	30	2.32E-02
TMBIM6	-0.2	5.00E-03	1.82E-02	4583	27241.35	28733.19	40	60	1.34E-02
PODXL	-0.4	8.00E-06	1.22E-04	1092	6659.17	7271.06	50	50	1.36E-06
PLVAP	-0.2	7.10E-03	2.39E-02	4946	15470.76	17515.66	50	50	4.10E-04
PLIN2	-0.3	6.20E-04	3.56E-03	2902	18947.56	22839.08	50	50	1.71E-05
PCK1	-0.3	1.00E-04	8.58E-04	1931	1120.74	3037.73	50	50	7.84E-06
NDRG1	-0.2	1.20E-02	3.61E-02	5506	50127.14	51689.99	60	40	2.72E-02
ITM2B	-0.3	6.00E-04	3.47E-03	2880	34751.8	36807.63	50	50	1.36E-02
HSPA8	-0.3	1.20E-03	5.90E-03	3363	17668.96	18139.95	40	60	1.04E-02
HLA-DRA	-0.2	3.80E-03	1.46E-02	4304	29068.65	32924.27	20	80	4.22E-02
GATM	-0.3	4.20E-04	2.61E-03	2683	5433.14	6800.94	50	50	3.09E-04
CYFIP2	-0.5	2.20E-09	4.32E-07	82	3482.26	4051.73	50	50	9.88E-08
CDH16	-0.2	4.40E-03	1.65E-02	4430	4093.23	4940.33	50	50	1.14E-03
CCND1	-0.2	3.00E-03	1.22E-02	4068	17278.68	19256.81	50	50	2.85E-04
ATP5B	-0.2	1.10E-02	3.37E-02	5360	11450.7	13211.83	30	70	2.59E-03
ATP5A1	-0.2	2.20E-03	9.54E-03	3812	7988.24	9278.65	50	50	2.86E-02
ATP1B1	-0.3	1.50E-03	7.03E-03	3514	18741.07	21002.32	50	50	3.90E-02
ATP1A1	-0.3	4.90E-05	4.98E-04	1634	12917.72	15392.31	40	60	2.34E-02
AQP1	-0.3	4.30E-05	4.52E-04	1580	16717.87	19036.22	50	50	3.11E-08
APP	-0.4	1.90E-05	2.36E-04	1329	32137.14	33051.3	50	50	1.33E-06
ALDOB	-0.3	4.40E-05	4.61E-04	1587	467.22	3374.03	50	50	3.27E-06
AIF1L	-0.2	1.50E-03	7.03E-03	3510	1984.01	2798.36	60	40	2.80E-02

We also evaluated the identified 11 miRNAs by DIANA-mirpath. Table 5 shows the enriched disease-related KEGG pathways (P-value < 0.05). The renal cell carcinoma pathway is identified with a significant P-value equal to 0.01613.

Table 5. The top 10 enriched KEGG pathways predicted by DIANA-mirpath for the 11 identified miRNAs (P-values are corrected). The full list can be obtained from http://snf-515788.vm.okeanos.grnet.gr/#mirnas=hsa-miR-210-3p;hsa-miR-210-5p;hsa-miR-891a-3p;hsa-miR-891a-5p;hsa-miR-200c-5p;hsa-miR-200c-5p;hsa-miR-141-5p;hsa-miR-141-3p;hsa-miR-122-3p;hsa-miR-122-5p;hsa-miR-155-3p;hsa-miR-155-5p;hsa-miR-508-3p;hsa-miR-508-5p;hsa-miR-514a-3p;hsa-miR-514a-5p;hsa-miR-184&methods=Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase;Tarbase&selection=0

KEGG pathway	P-value	#genes	#miRNAs
Chronic myeloid leukemia	5.90E-08	39	6
Proteoglycans in cancer	3.67E-06	72	8
Prostate cancer	2.58E-05	43	7
Pathways in cancer	3.10E-05	128	10
Pancreatic cancer	3.94E-05	32	5
Glioma	9.09E-05	28	5
Hepatitis B	9.11E-05	47	5
Small cell lung cancer	0.0002621	38	5
Non-small cell lung cancer	0.0002975	24	4
Colorectal cancer	0.0002975	28	7
Endometrial cancer	0.0007913	23	6
Viral carcinogenesis	0.0007913	59	8
Bladder cancer	0.001004	20	5
Melanoma	0.01584	25	5
Renal cell carcinoma	0.01613	27	5
Hepatitis C	0.02652153	44	6

4. Discussion

The top signature in Table 2 is related to the cAMP signaling pathway. Targeting the cAMP pathway is an effective treatment for kidney cancer (Piazzon, Maisonneuve, Guilleret, Rotman, & Constam, 2012; Torres & Harris, 2014). The second signature in Table 2 is the *Snf5* gene expression profile of a murine model (Mouse Embryonic Fibroblast (MEF) cells) that closely resembles that of human SNF5-deficient rhabdoid tumors (pediatric soft tissue sarcoma that arises in the kidney, the liver, and the peripheral nerves) (Isakoff et al., 2005). Impairment of the SWI/SNF chromatin remodeling complex plays an important role in the development and aggressiveness of clear cell renal cell carcinoma (Sarnowska et al., 2017). The sixth signature in Table 2 comes from a study of the effects of knockdown of the gene family of eukaryotic translation initiation factors (EIF) by RNAi in MCF10A cells. EIF3b is a promising prognostic biomarker and a potential therapeutic target for patients with clear cell renal cell carcinoma (Zang et al., 2017), and EIF4GI is a target for cancer therapeutics (Jaiswal, Koul, Palanisamy, & Koul, 2019).

The top pathway in Table 3 is the 'Pathway of regulation of IGF activity by IGFBP'. Studies show that insulin-like growth factors (IGFs) and insulin play a stimulatory role for renal cancer cells (Braczkowski, Bialozyt, Plato, Mazurek, & Braczkowska, 2016; Solarek, Koper, Lewicki, Szczylik, & Czarnecka, 2019). Patients with IGF-1 receptor overexpression have a 70% increased risk of death (Tracz, Szczylik, Porta, & Czarnecka, 2016). Moreover, this overexpression has been shown to increase kidney cancer risk in middle-aged male smokers (Major, Pollak, Snyder, Virtamo, & Albanes, 2010). The second pathway in Table 3 is 'Cytokine Signaling in Immune system'. Cytokines are important biomolecules that play essential roles in tumor formation (Lee & Rhee, 2017) and they are therapeutic targets (Doehn, Kausch, Melz, Behm, & Jocham, 2004; Macleod et al., 2015). The IL-6 cytokine family can serve as useful diagnostic and prognostic biomarkers. In fact, IL-6 is a potential target in cancer therapy (Kaminska, Czarnecka, Escudier, Lian, & Szczylik, 2015; Unver & McAllister, 2018). Ishibashi et al., reported that IL-6 suppresses the expression of the cytokine signaling-3 (*SOCS3*) gene, and is associated with poor prognosis of kidney cancer patients (Ishibashi et al., 2018).

Table 4 shows the significant relationships between the predicted 23 mRNAs and the patients' survival rates. For some of the 23 genes, patients cannot be divided equally based on expression of considered genes in order to get significant P-values for the Kaplan-Meier plots. A majority of the mRNAs (15 out of 23) are associated with P-values less than 0.05 with 50/50 divisions based on the level of gene expression. Among the 16 KEGG pathways predicted by DIANA-mirpath (Table 5), 14 are directly related to cancers, except for Hepatitis B and Hepatitis C. Therefore, we correctly identified miRNA signatures that are cancer-related.

In order to validate the robustness of our findings, we employed an independent dataset to confirm that our results are independent of datasets to some extent. The alternative dataset was downloaded from GEO (GSE16441). The procedures applied to analyze the GEO dataset are similar to those applied to the dataset obtained from TCGA. The only difference is the number of samples, miRNAs, and mRNAs. After repeating the same procedures, we realized that $u_{l_1 j}^{mRNA}$ and $u_{l_3 j}^{miRNA}$ ($l_1 = l_3 = 2$) also varied between normal and tumor samples (Fig 1). P-values computed by the t-test were 6.74×10^{-22} for mRNA and 2.54×10^{-18} for miRNA. In order to ascertain whether u_{2j}^{mRNA} and u_{2j}^{miRNA} are significantly correlated, we calculated the *PCC* between them, which was 0. 931 (p-value $= 1.58 \times 10^{-15}$), indicating that they are highly correlated.

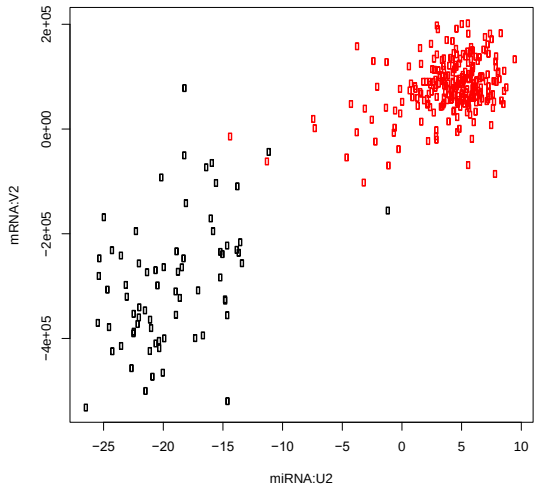

Fig 1 Scatter plot between $u_{l_1 j}^{mRNA}$ (vertical axis) and $u_{l_3 j}^{miRNA}$ (horizontal axis). Black (red) open circle corresponds to normal (tumor) tissue.

Next, we checked if the selected miRNAs and mRNAs were common between the TCGA and GEO datasets. We identified three miRNAs – hsa-miR-141, hsa-miR-210, and hsa-miR-200c, which are listed in Table 1. On the other hand, 209 genes were identified. After restricting genes included in both TCGA and GEO datasets, we evaluated the overlap as the confusion matrix (Table 6).

Table 6. Confusion matrix between genes selected in TCGA and GEO dataset.

		GEO	
		Not selected	Selected
TCGA	Not selected	17209	160
	Selected	60	11

The P-value determined using the Fisher exact test was 8.97×10^{-11} and the odds ratio was 19.7. Therefore, the coincidence between selected genes in the TCGA and GEO datasets is significant and the results obtained for TCGA are robust and not highly dependent upon specific samples.

To test the superiority to the conventional method, we applied the t-test to the TCGA and GEO datasets. After applying the t-test, P-values were calculated and adjusted based on the BH criterion. Then, 13,895 genes and 399 miRNAs for TCGA and 12,152 genes and 78 miRNAs for GEO were associated with adjusted P-values less than 0.01. Relative to the TD method, the t-test identified a larger number of genes and miRNAs using the P-values as criteria. If the top ranked (small enough or restricted) number of genes and miRNAs was selected by the t-test, the coincidence between TCGA and GEO might be compatible. Therefore, we selected the same number of genes and miRNAs by the t-test as those selected by TD. Only one miRNA and no genes were common between the TCGA and GEO datasets. Therefore, we determined that the t-test could identify less coincident sets of genes and miRNAs between TCGA and GEO. In conclusion, this strongly suggests that the proposed method is superior to the t-test.

5. Conclusions

In this study, we applied the TD-based unsupervised FE method to the KIRC miRNA expression and gene expression data. The TD-based method can identify miRNA signatures with differential expression between normal tissues and tumors as well as significant correlations between the gene expression data. Selected mRNAs and miRNAs are not only mutually correlated, but are also significantly related to various aspects of cancers. This suggests that integrated analysis performed by TD-based unsupervised FE is an effective strategy, despite its simplicity to identify biologically significant pairs of miRNAs and mRNAs, which is not easy by other strategies.

Supplementary Materials: Supplementary figures. The results of the Kaplan-Meier plots of the 23 KIRC survival-associated genes by using OncoLnc.

Author Contributions: Ka-Lok Ng foresee the research, prepared the data, writing—original draft preparation, review and editing. Y-h Taguchi performed the formal analysis, writing—original draft preparation, review and editing.

Conflicts of Interest: The authors declare no conflict of interest.

References

Anaya, J. (2016). OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. *PeerJ Computer Science, 2*, e67. doi:10.7717/peerj-cs.67

Braczkowski, R., Bialozyt, M., Plato, M., Mazurek, U., & Braczkowska, B. (2016). Expression of insulin-like growth factor family genes in clear cell renal cell carcinoma. *Contemp Oncol (Pozn), 20*(2), 130-136. doi:10.5114/wo.2016.58720

Doehn, C., Kausch, I., Melz, S., Behm, A., & Jocham, D. (2004). Cytokine and vaccine therapy of kidney cancer. *Expert Rev Anticancer Ther, 4*(6), 1097-1111. doi:10.1586/14737140.4.6.1097

Feinberg, A. P., & Vogelstein, B. (1983). Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature, 301*(5895), 89-92. doi:10.1038/301089a0

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell, 144*(5), 646-674. doi:10.1016/j.cell.2011.02.013

Inamoto, T., Uehara, H., Akao, Y., Ibuki, N., Komura, K., Takahara, K., . . . Azuma, H. (2018). A Panel of MicroRNA Signature as a Tool for Predicting Survival of Patients with Urothelial Carcinoma of the Bladder. *Dis Markers, 2018*, 5468672. doi:10.1155/2018/5468672

Isakoff, M. S., Sansam, C. G., Tamayo, P., Subramanian, A., Evans, J. A., Fillmore, C. M., . . . Roberts, C. W. (2005). Inactivation of the Snf5 tumor suppressor stimulates cell cycle progression and cooperates

with p53 loss in oncogenic transformation. *Proc Natl Acad Sci U S A, 102*(49), 17745-17750. doi:10.1073/pnas.0509014102

Ishibashi, K., Koguchi, T., Matsuoka, K., Onagi, A., Tanji, R., Takinami-Honda, R., . . . Kojima, Y. (2018). Interleukin-6 induces drug resistance in renal cell carcinoma. *Fukushima J Med Sci, 64*(3), 103-110. doi:10.5387/fms.2018-15

Jaiswal, P. K., Koul, S., Palanisamy, N., & Koul, H. K. (2019). Eukaryotic Translation Initiation Factor 4 Gamma 1 (EIF4G1): a target for cancer therapeutic intervention? *Cancer Cell Int, 19*, 224. doi:10.1186/s12935-019-0947-2

Kaminska, K., Czarnecka, A. M., Escudier, B., Lian, F., & Szczylik, C. (2015). Interleukin-6 as an emerging regulator of renal cell cancer. *Urol Oncol, 33*(11), 476-485. doi:10.1016/j.urolonc.2015.07.010

Lee, M., & Rhee, I. (2017). Cytokine Signaling in Tumor Progression. *Immune Netw, 17*(4), 214-227. doi:10.4110/in.2017.17.4.214

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst, 1*(6), 417-425. doi:10.1016/j.cels.2015.12.004

Lokeshwar, S. D., Talukder, A., Yates, T. J., Hennig, M. J. P., Garcia-Roig, M., Lahorewala, S. S., . . . Lokeshwar, V. B. (2018). Molecular Characterization of Renal Cell Carcinoma: A Potential Three-MicroRNA Prognostic Signature. *Cancer Epidemiol Biomarkers Prev, 27*(4), 464-472. doi:10.1158/1055-9965.Epi-17-0700

Luo, Y., Chen, L., Wang, G., Xiao, Y., Ju, L., & Wang, X. (2019). Identification of a three-miRNA signature as a novel potential prognostic biomarker in patients with clear cell renal cell carcinoma. *J Cell Biochem, 120*(8), 13751-13764. doi:10.1002/jcb.28648

Macleod, L. C., Tykodi, S. S., Holt, S. K., Wright, J. L., Lin, D. W., Tretiakova, M. S., . . . Gore, J. L. (2015). Trends in Metastatic Kidney Cancer Survival From the Cytokine to the Targeted Therapy Era. *Urology, 86*(2), 262-268. doi:10.1016/j.urology.2015.05.008

Major, J. M., Pollak, M. N., Snyder, K., Virtamo, J., & Albanes, D. (2010). Insulin-like growth factors and risk of kidney cancer in men. *Br J Cancer, 103*(1), 132-135. doi:10.1038/sj.bjc.6605722

Medina, P. P., & Slack, F. J. (2008). microRNAs and cancer: an overview. *Cell Cycle, 7*(16), 2485-2492. doi:10.4161/cc.7.16.6453

Piazzon, N., Maisonneuve, C., Guilleret, I., Rotman, S., & Constam, D. B. (2012). Bicc1 links the regulation of cAMP signaling in polycystic kidneys to microRNA-induced gene silencing. *J Mol Cell Biol, 4*(6), 398-408. doi:10.1093/jmcb/mjs027

Rabanser, S., Shchur, O., & Günnemann, S. (2017). Introduction to Tensor Decompositions and their Applications in Machine Learning.

Rouhi, A., Mager, D. L., Humphries, R. K., & Kuchenbauer, F. (2008). MiRNAs, epigenetics, and cancer. *Mamm Genome, 19*(7-8), 517-525. doi:10.1007/s00335-008-9133-x

Sarnowska, E., Szymanski, M., Rusetska, N., Ligaj, M., Jancewicz, I., Cwiek, P., . . . Sarnowski, T. J. (2017). Evaluation of the role of downregulation of SNF5/INI1 core subunit of SWI/SNF complex in clear cell renal cell carcinoma development. *Am J Cancer Res, 7*(11), 2275-2289.

Shi, X. H., Li, X., Zhang, H., He, R. Z., Zhao, Y., Zhou, M., . . . Qin, R. Y. (2018). A Five-microRNA Signature for Survival Prognosis in Pancreatic Adenocarcinoma based on TCGA Data. *Sci Rep, 8*(1), 7638. doi:10.1038/s41598-018-22493-5

343 Solarek, W., Koper, M., Lewicki, S., Szczylik, C., & Czarnecka, A. M. (2019). Insulin and insulin-like growth
344 factors act as renal cell cancer intratumoral regulators. *J Cell Commun Signal, 13*(3), 381-394.
345 doi:10.1007/s12079-019-00512-y
346 Taguchi, Y. (2017, 23-25 Oct. 2017). *One-class Differential Expression Analysis using Tensor Decomposition-based*
347 *Unsupervised Feature Extraction Applied to Integrated Analysis of Multiple Omics Data from 26 Lung*
348 *Adenocarcinoma Cell Lines.* Paper presented at the 2017 IEEE 17th International Conference on
349 Bioinformatics and Bioengineering (BIBE).
350 Taguchi, Y., & Ng, K. (2018, 29-31 Oct. 2018). *[Regular Paper] Tensor Decomposition–Based Unsupervised Feature*
351 *Extraction for Integrated Analysis of TCGA Data on MicroRNA Expression and Promoter Methylation of*
352 *Genes in Ovarian Cancer.* Paper presented at the 2018 IEEE 18th International Conference on
353 Bioinformatics and Bioengineering (BIBE).
354 Taguchi, Y.-h. (2019a). *Multiomics Data Analysis Using Tensor Decomposition Based Unsupervised Feature*
355 *Extraction –Comparison with DIABLO–* (Vol. 11643): Springer.
356 Taguchi, Y.-h. (2019b). *Tensor Decomposition Based Unsupervised Feature Extraction Applied to Bioinformatics.*
357 *Application of Omics, AI and Blockchain in Bioinformatics Research* (J. J. P. T. a. K.-L. Ng Ed.): World
358 Scientifc Publisher.
359 Taguchi, Y.-h. (2019c). *Unsupervised Feature Extraction Applied to Bioinformatics: A PCA Based and TD Based*
360 *Approach*: Springer International.
361 Taguchi, Y.-h., & Turki, T. (2019). Neurological disorder drug discovery from gene expression with tensor
362 decomposition. *Current Pharmaceutical Design*, 704163. doi:10.1101/704163
363 Taguchi, Y. H. (2017a). Identification of candidate drugs using tensor-decomposition-based unsupervised
364 feature extraction in integrated analysis of gene expression between diseases and DrugMatrix
365 datasets. *Sci Rep, 7*(1), 13733. doi:10.1038/s41598-017-13003-0
366 Taguchi, Y. H. (2017b). Tensor decomposition-based unsupervised feature extraction applied to matrix
367 products for multi-view data processing. *PLoS One, 12*(8), e0183933. doi:10.1371/journal.pone.0183933
368 Taguchi, Y. H. (2017c). Tensor decomposition-based unsupervised feature extraction identifies candidate genes
369 that induce post-traumatic stress disorder-mediated heart diseases. *BMC Med Genomics, 10*(Suppl 4),
370 67. doi:10.1186/s12920-017-0302-1
371 Taguchi, Y. H. (2018a). Correction: Tensor decomposition-based unsupervised feature extraction applied to
372 matrix products for multi-view data processing. *PLoS One, 13*(7), e0200451.
373 doi:10.1371/journal.pone.0200451
374 Taguchi, Y. H. (2018b). Tensor decomposition-based and principal-component-analysis-based unsupervised
375 feature extraction applied to the gene expression and methylation profiles in the brains of social
376 insects with multiple castes. *BMC Bioinformatics, 19*(Suppl 4), 99. doi:10.1186/s12859-018-2068-7
377 Taguchi, Y. H. (2018c). Tensor Decomposition-Based Unsupervised Feature Extraction Can Identify the
378 Universal Nature of Sequence-Nonspecific Off-Target Regulation of mRNA Mediated by MicroRNA
379 Transfection. *Cells, 7*(6). doi:10.3390/cells7060054
380 Taguchi, Y. H. (2019). Drug candidate identification based on gene expression of treated cells using tensor
381 decomposition-based unsupervised feature extraction for large-scale data. *BMC Bioinformatics,*
382 *19*(Suppl 13), 388. doi:10.1186/s12859-018-2395-8
383 Taguchi, Y. H., & Turki, T. (2019). Tensor Decomposition-Based Unsupervised Feature Extraction Applied to
384 Single-Cell Gene Expression Analysis. *Front Genet, 10*, 864. doi:10.3389/fgene.2019.00864

385 Torres, V. E., & Harris, P. C. (2014). Strategies targeting cAMP signaling in the treatment of polycystic kidney
386 disease. *J Am Soc Nephrol, 25*(1), 18-32. doi:10.1681/asn.2013040398
387 Tracz, A. F., Szczylik, C., Porta, C., & Czarnecka, A. M. (2016). Insulin-like growth factor-1 signaling in renal
388 cell carcinoma. *BMC Cancer, 16*, 453. doi:10.1186/s12885-016-2437-4
389 Unver, N., & McAllister, F. (2018). IL-6 family cytokines: Key inflammatory mediators as biomarkers and
390 potential therapeutic targets. *Cytokine & Growth Factor Reviews, 41*, 10-17.
391 doi:https://doi.org/10.1016/j.cytogfr.2018.04.004
392 Vargas-Rondon, N., Villegas, V. E., & Rondon-Lagos, M. (2017). The Role of Chromosomal Instability in Cancer
393 and Therapeutic Responses. *Cancers (Basel), 10*(1). doi:10.3390/cancers10010004
394 Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., . . .
395 Hatzigeorgiou, A. G. (2015). DIANA-miRPath v3.0: deciphering microRNA function with
396 experimental support. *Nucleic Acids Res, 43*(W1), W460-466. doi:10.1093/nar/gkv403
397 Xie, M., Lv, Y., Liu, Z., Zhang, J., Liang, C., Liao, X., . . . Li, Y. (2018). Identification and validation of a four-
398 miRNA (miRNA-21-5p, miRNA-9-5p, miR-149-5p, and miRNA-30b-5p) prognosis signature in clear
399 cell renal cell carcinoma. *Cancer Manag Res, 10*, 5759-5766. doi:10.2147/cmar.S187109
400 Yu, Y., Feng, X., & Cang, S. (2018). A two-microRNA signature as a diagnostic and prognostic marker of
401 pancreatic adenocarcinoma. *Cancer Manag Res, 10*, 1507-1515. doi:10.2147/cmar.S158712
402 Zang, Y., Zhang, X., Yan, L., Gu, G., Li, D., Zhang, Y., . . . Xu, Z. (2017). Eukaryotic Translation Initiation Factor
403 3b is both a Promising Prognostic Biomarker and a Potential Therapeutic Target for Patients with
404 Clear Cell Renal Cell Carcinoma. *J Cancer, 8*(15), 3049-3061. doi:10.7150/jca.19594
405 Zhang, W., Dahlberg, J. E., & Tam, W. (2007). MicroRNAs in tumorigenesis: a primer. *Am J Pathol, 171*(3), 728-
406 738. doi:10.2353/ajpath.2007.070070
407 Zhou, H., Tang, K., Xiao, H., Zeng, J., Guan, W., Guo, X., . . . Ye, Z. (2015). A panel of eight-miRNA signature as
408 a potential biomarker for predicting survival in bladder cancer. *J Exp Clin Cancer Res, 34*, 53.
409 doi:10.1186/s13046-015-0167-0
410