mRNA and long-noncoding RNA signatures for improving the prognosis prediction of cutaneous skin melanoma efficiently

S.-X. ZI[1], Y.-M. XING[2], C.-Z. WANG[3]

[1]Department of Dermatology, Emergency General Hospital, Beijing, China
[2]Department of Oncology and Metabolism, University of Sheffield, School of Medicine Beech Hill Road, Sheffield, United Kingdom
[3]National Laboratory of Biomacromolecules, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China

Abstract. – OBJECTIVE: Non-coding RNAs occupy a significant fraction of the human genome, and their biological significance during the pathological process is proved. More and more lncRNAs are reported in cancer research.

MATERIALS AND METHODS: To investigate the non-coding RNA's biological relevance with cutaneous skin melanoma, we first compared the survival analysis by combining the most differentially expressed mRNA and non-coding RNA expression values.

RESULTS: The result showed that the abundantly expressed mRNAs and lncRNAs have significant effects on the survival of patients. Compared to the mRNAs, these lncRNAs have more impact on the progress of cutaneous skin melanoma. Thus, we combined the two types of RNA factors having significant effects as risk factors to construct the diagnosis model, and the survival analysis confirmed the robustness of the diagnosis model.

CONCLUSIONS: In summary, a list of eight lncRNA and five mRNA expression signatures can be used to improve the prognosis prediction of cutaneous skin melanoma, as well as help us to understand the pathogenic mechanism and provide a hint for targeting therapy.

Key Words:
Non-coding RNAs, Melanoma, Skin, Tumor.

Introduction

Those untranslated transcripts longer than 200 bp are named as long non-coding RNAs (lncRNAs). They possess many structural characteristics same as mRNAs except conserved open reading frame. As backed by a lot of emerging evidence, those non-coding RNAs occupy a significant fraction of the human genome. Many human diseases are reported to have aberration on lncRNAs because of their involvement in DNA methylation, histone modification, and chromatin remodeling.

Melanoma, cancer with long-term and rapid rising incidence, was responsible for 5% of the new cancer cases in America in 2018[1] and is the culprit of approximately 80% of skin cancer-related deaths[2]. The prognosis for most patients with this disease is poor, having a 5-year survival rate of around 16-80%[1-3]. Effective treatment options, which were brought about by novel insights relating to mutations that drive tumorigenesis and immune escape mechanisms of these tumors have revolutionized treatment[4]. By targeting biomarkers or checkpoints, we can improve the overall survival rates in patients[4,5]. Thus, identifying suitable prognosis biomarkers for skin cutaneous melanoma (SKCM) could benefit the diagnosis and treatment of this cancer.

As primary sources of prognosis biomarkers for SKCM, messenger RNA (mRNA) and the long non-coding RNAs (lncRNAs) have been investigated in many researches[6-8]. Some early research[9] reported that mRNAs are involved in developing this cancer. From the view of epigenetic mechanisms, cancer-related lncRNAs and some abnormally expressed ones were also proved to have potential roles in the metastasis of melanoma. Branca et al[10] showed that SKCM is predictable through genomic biomarkers through a meta-analysis of 204 tumors. Winnepenninckx et al[11,12] identified 254 genes whose

Corresponding Author: Shaoxia Zi, MD; e-mail: zishaoxia@zlhospital.org

expression was associated with metastatic dissemination of SKCM. Wang et al[6] characterized the role of messenger RNA signatures in melanoma tumorigenesis and metastasis in 2017 by employing 74 train sets and 58 test sets from the Gene Expression Omnibus database. Along with the increase of samples of SKCM in The Cancer Genome Atlas (TCGA)[13], more thorough investigations that look at the potential involvement of mRNAs in melanoma tumorigenesis and metastasis are required.

This study conducted RNA-sequencing data analyses with 479 SKCM samples from TCGA projects. We firstly compared the diagnosis values of mRNAs and lncRNAs. Then we screened the abundantly expressed lncRNAs and mRNAs to construct a prognosis model. We compared their diagnosis performance and analyzed their relationship. Finally, our results showed that eight lncRNA and five mRNA signatures could be used to calculate the risk score of SKCM.

Materials and Methods

Expression and Clinical Data of Patients

All sample information was downloaded from the TCGA database (Figure 1A), which has a collection of 459 SKCM patients[13]. The RNA-seq data included protein-coding genes and lncRNAs. Altogether, 21,257 mRNA and 14,371 lncRNAs profiles were acquired for all the patients and were normalized as Fragments per Kilobase Million (FPKM). Related follow-up information was also obtained from TCGA.

Next, training and test sets were randomly generated to build the diagnosis model. The training set includes 230 clinician samples, while the training set includes 229 clinician samples (Figures 1B and 1C).

Definition of the mRNA and lncRNAs Abundantly Expressed

The expression levels of mRNAs and lncRNAs vary significantly in and among patients. Thus, we take those mRNAs and lncRNAs expressed in 50% of samples as abundantly expressed ones. With this step, the number of mRNAs was reduced to 17,912, and the number of lncRNAs was decreased to 6,567. That means 15.73% of mRNAs and 54.30% lncRNAs are not generally expressed or are not detected by sequencing technology. Thus, this step would make sure the robustness of the following analysis.

Identification and Selection of Prognostic-Related mRNA and lncRNAs

Firstly, a correlation analysis was performed between the overall survival time and expression level of mRNAs and lncRNAs from the training set containing 230 samples. For this analysis, the Survival package was applied under the computing environment R[14]. Those mRNAs and lncRNAs having a significant correlation with the

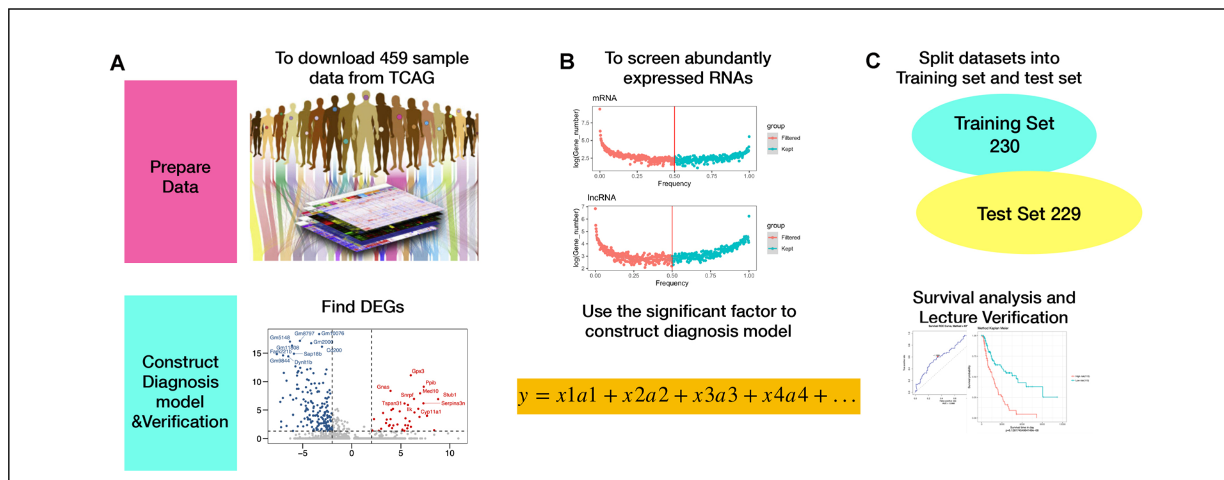

Figure 1. Flowchart to screen RNA signatures to construct the prognosis model of SKCM. **A,** First, data are processed through downloading from TCGA and filtering abundantly expressed RNA signatures. **B-C,** Afterward, we randomly split the 459 patients into a training set with 230 samples and a test set with 229 samples. Then We get differently expressed genes with univariable cox regression analysis, and two formulas to calculate the risk score were constructed based on the two different signatures. Further survival analysis and a literature search were performed to verify the results.

survival time were identified as possible prognostic-related factors (Figure 1).

Secondly, Rbsurv packge[15-17] was used to perform a robust likelihood-based survival analysis. We could further screen possible prognostic and lncRNAs from the set of possible prognostic-related mRNA and lncRNAs mentioned above. In the training set, the possible prognostic related mRNA and lncRNA genes acquired from the last step were fitted using an unviable cox regression model. Then, we used all samples and the test set to evaluate the estimation parameters.

Thirdly, this process was repeated 10 times, which means each gene has 10 log likelihoods. We selected mRNA and lncRNA genes with the most significant average log-likelihood value, and a series of the predictive model was constructed. Finally, an optimal predictive model was selected with the lowest Akaike information criterion (AIC) value. As a result, five prognosis-related mRNAs and five prognoses-related lncRNAs and mRNAs were strictly selected.

Establishment and Validation of the Risk Score Formula

Integrating all chosen mRNAs, a risk formula was established based on the estimated regression coefficients for the training set. With this formula, the risk scores for all patients from the training set could be calculated, and then these patients can be grouped into two groups: the high-risk group and the low-risk group. Additional receiver operating characteristic (ROC) curve was drawn with survival ROC package in R environment, and then an optimal cut-off was set. The acquired high-risk and low-risk groups are assessed by the Kaplan-Meier estimate and compared using a log-rank test. To further validate the formula, we also applied it to fit in the test and complete dataset.

Results

Possible Prognostic Related mRNAs and LncRNAs for SKCM

21,257 mRNA and 14,371 lncRNAs expression profiles were acquired for 459 patients from TCGA (Figure 2A). Some genes whose expressions were not detected may result in the technology variation. Thus, we filtered out those mRNAs and lncRNAs with no expression information in 50% of samples of the 459 patients. After this step, we got 17,912 mRNAs and 6,567 lncRNAs. They are supposed to be abundantly expressed (Figure 2A).

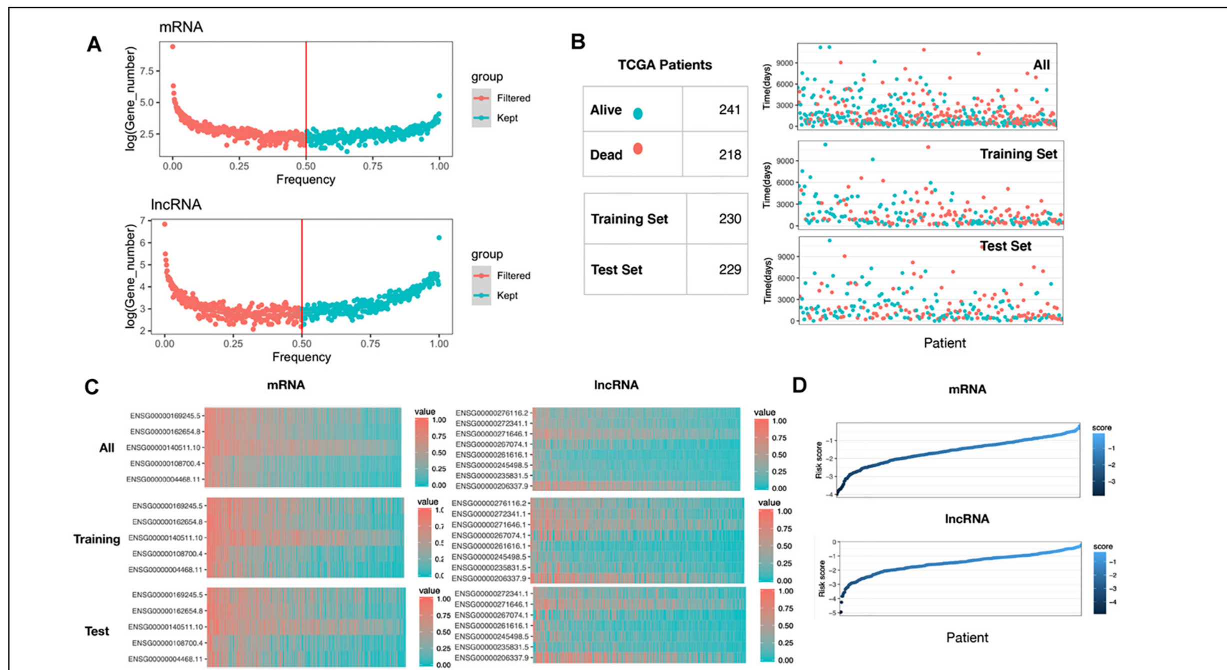

Figure 2. Identification of the signatures as possible risk factors. **A,** The distribution of genes with different abundance. The red line indicts genes express in half samples. 5-mRNA-based risk scores. **B,** The distribution of patients with different statuses. Some patients are dead while another is alive. All and the spliced two datasets have a similar distribution. **C,** Expression pattern of the RNA signatures, which are used as risk predictors. **D,** The risk scores predicted with two RNA signatures are similar

Next, the univariable cox regression analysis was performed on a randomly generalized training set, including expression data of 230 patients. We acquired a set of differentially expressed genes, including 3,254 mRNA genes and 1,223 lncRNAs (Figure 2B, p <0.05). These are used as possible prognostic factors. Twenty mRNA genes with the lowest p values (p <0.0000002) and twenty lncRNAs (p <0.00013) were shown in Table I.

To check whether these factors could be used as prognostic factors, we take those patients with high expression of these 20 mRNAs and lncRNAs as high-risk ones and patients with low expression of these 20 mRNAs and lncRNAs as low-risk ones to perform survival analysis. The results showed that samples with higher gene expression levels have a terrible survival probability (p <0.5). From the survival analysis, it could be found that the 20 mRNAs may perform better than the 20 lncRNAs in diagnosing SKCM. Among the mRNA genes, *RARRES3* is reported to be mildly expressed in tumor cells[18,19]. *GBP4* mRNAs in SKCM are associated with a favorable prognosis and may act as potential prognostic biomarkers[20]. Moreover, *PSMB9* is also related to the SKCM[21]. Among the lncRNA genes, Thus, our analysis results are robust and reliable.

Table I. Twenty mRNA and lncRNAs with the lowest p-values.

	Ensemble ID	Symbol	nloglik	AIC
mRNA	ENSG00000108700.4	CCL8	462.25	926.51*
	ENSG00000004468.11	CD38	460.72	925.43*
	ENSG00000140511.10	HAPLN3	456.83	919.65*
	ENSG00000162654.8	GBP4	455.39	918.78*
	ENSG00000169245.5	CXCL10	453.91	917.81*
	ENSG00000145649.7	GZMA	453.22	918.45
	ENSG00000240065.6	PSMB9	452.38	918.75
	ENSG00000133321.9	RARRES3	451.84	919.68
	ENSG00000137496.16	IL18BP	451.05	920.1
	ENSG00000143185.3	XCL2	449.39	918.77
	ENSG00000138755.5	CXCL9	449.02	920.04
	ENSG00000179344.15	HLA-DQB1	448.89	921.78
	ENSG00000105374.8	NKG7	448.84	923.68
	ENSG00000239713.6	APOBEC3G	448.82	925.63
	ENSG00000092010.13	PSME1	446.83	923.67
	ENSG00000275302.1	CCL4	446.81	925.62
	ENSG00000162645.11	GBP2	446.75	927.5
	ENSG00000198502.5	HLA-DRB5	446.59	929.18
	ENSG00000128284.18	APOL3	445.15	928.3
lncRNA	ENSG00000245498.5	RP11-677M14.7	497.62	997.24*
	ENSG00000235831.5	BHLHE40-AS1	491.74	987.47*
	ENSG00000271646.1	RP11-326I11.3	489.04	984.08*
	ENSG00000272341.1	RP1-151F17.2	487.47	982.94*
	ENSG00000267074.1	RP11-1094M14.5	486.98	983.95*
	ENSG00000206337.9	HCP5	485.62	983.24*
	ENSG00000276116.2	FUT8-AS1	484.38	982.76*
	ENSG00000261616.1	RP11-6O2.3	482.07	980.14*
	ENSG00000178977.3	LINC00324	481.93	981.86
	ENSG00000205537.2	RP11-89H19.1	480.1	980.2
	ENSG00000224429.6	LINC00539	480.09	982.17
	ENSG00000257924.1	RP11-493L12.5	479.67	983.33
	ENSG00000237352.2	RP11-145M4.3	478.86	983.72
	ENSG00000237372.2	UNQ6494	478.82	985.63
	ENSG00000227531.1	RP11-202G18.1	478.68	987.37
	ENSG00000246526.2	RP11-539L10.2	478.58	989.15
	ENSG00000237775.1	DDR1-AS1	474.1	982.2
	ENSG00000275557.1	RP11-353N4.6	473.98	983.95
	ENSG00000256262.1	USP30-AS1	473.1	984.2

*Based on the 5 mRNA and 8 lncRNA, a signature-based risk model was constructed based on corresponding Cox coefficients to investigate the prognosis of SKCM comprehensively.

Screening Possible Prognostic Factors and Constructing Calculation Model for Risk Score of SKCM

A robust likelihood-based on survival analysis was performed, and according to the AIC values, we got eight lncRNAs and five mRNAs as signatures that could be used to construct the risk models, respectively (Table I). Among these mRNAs, CCL8 has been demonstrated to have a regulatory role during the progression of melanoma[22]. CD38 is reported to play a role in murine and human lung tumorigenesis[23]. Furthermore, *HAPLN3* also has an increased expression in cancer cells[24]. In summary, all the mRNA signatures identified in this study are reliable prognostic factors related to mRNAs for SKCM (Figure 2C).

To comprehensively investigate the possibility of using 5 significant mRNA factors for the prognosis of SKCM, a 5-mRNA and 8-lncRNA signature-based risk were constructed based on corresponding Cox coefficients.

The risk scores for the test sets were calculated using the two formulas we got from the training set. The survival time of SKCM patients was adversely associated with their risk scores (Figure 2D).

Patients with high-risk scores predicted by mRNA signatures tend to have high CD38 and CXCL10 expression and low CCL8, HAPLN3 and GBR4 expression. Patients in the low-risk group have significantly longer life spans than those in the high-risk group. It also showed that the predictive ability of the risk score is higher than any of the identified mRNAs. ROC analysis was performed for the 5-mRNA risk scores to compare the specificity and sensitivity of survival prediction. The area under the curve (AUC) was 0.718, and the cut-off point selected was -1.316 (Figure 3A). With this cut-off point, the patients

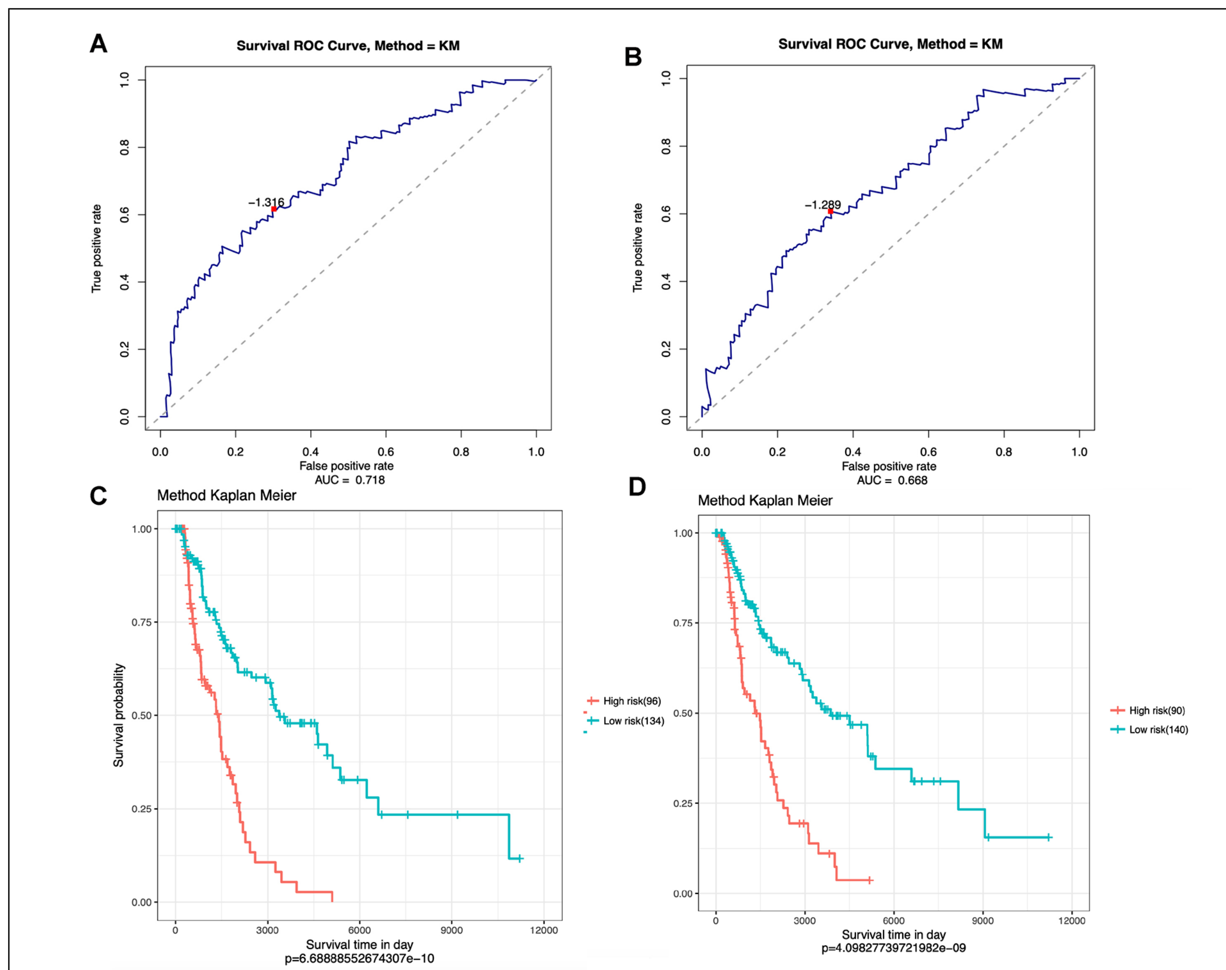

Figure 3. Application of risk score based on five mRNA signatures in the train set. **A-B,** ROC analysis of the sensitivity and specificity of the survival time by the 5-mRNA/8-lncRNA signature-based risk score. **C-D,** Kaplan-Meier estimates patients' survival status and time using the median mRNA/lncRNA risk score cut-off in the train set.

were further divided into a high risk-group and a low-risk group, which revealed a significant difference. The Kaplan-Meier curve and log-rank test further indicated a substantial difference in survival time between the high-risk and low-risk groups (Figure 3C).

Patients with high-risk scores predicted by lncRNA signatures tend to have high BHLHE40-AS1 expression and low RP11-677M14.7, RP11-326I11.3, RP1-151F17.2, RP11-1094M14.5, HCP5, FUT8-AS1 and RP11-6O2.3 expression. The area under the curve (AUC) here was 0.67, and the cut-off point selected was -1.289 (Figure 3B). With this cut-off point, the patients were further divided into a high risk-group and a low-risk group, which revealed a significant difference. The Kaplan-Meier curve and log-rank test further indicated a substantial difference in survival time between the high-risk and low-risk groups (Figure 3D), being more significant than that with the above five mRNAs.

Further Survival Prediction Over the Test Set and Whole Set Showed the Robustness of These Prognostic Indicators

The above results showed that prognostic indicators both have good performances over the training set. Here, further validations were made in the complete set and test set. Using the same formula and selecting optimal cut-off points based on mRNA signatures on the complete set, a total of 184 patients were recognized as high-risk ones, and a total of 275 patients were identified as low-risk ones (Figure 4A). In comparison, 198 patients were recognized as high-risk ones with

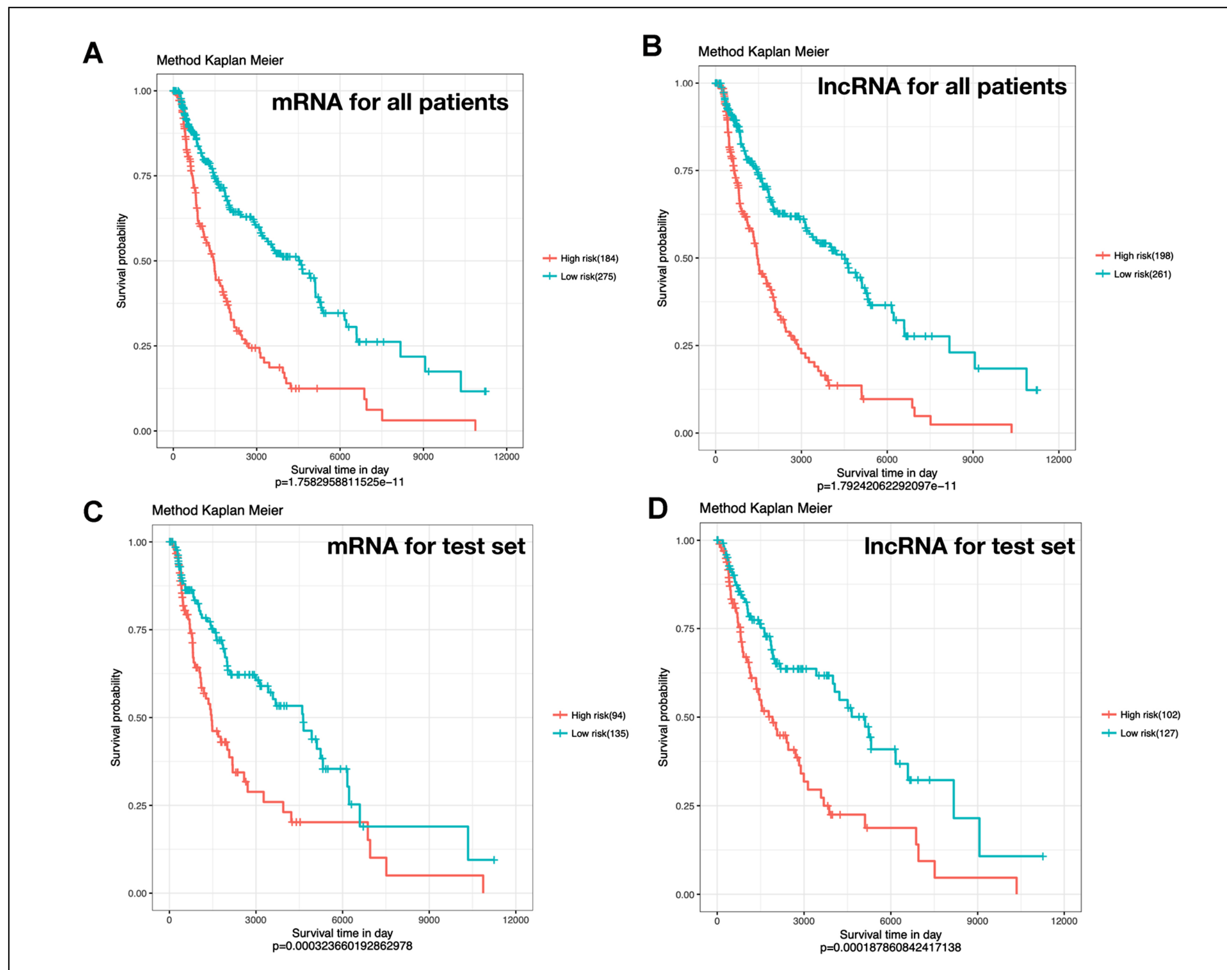

Figure 4. Application of risk score based on five mRNA signatures to predict the survival probability of SKCM patients. **A-B,** Kaplan-Meier estimates patients' survival status and time using the median mRNA/lncRNA risk score cut-off in the whole dataset. A total of 184 patients were predicted as high-risk ones, while a total of 275 patients were predicted as low risk. The log test showed that the two groups have a significant survival probability difference. **C-D,** Kaplan-Meier estimates patients' survival status and time using the median mRNA/lncRNA risk score cut-off in the test set. For 249 patients in the test dataset, the low-risk ones have higher survival probabilities.

lncRNA signatures (Figure 4B). For the test set, 94 patients were high-risk, 135 were low-risk, and 102 patients were recognized as high-risk (Figures 4C and 4D). In both sets, the mRNA signatures could distinguish the patients with survival rates very well ($p < 0.0003$).

Discussion

In this study, a list of 5 mRNAs and 8 lncRNAs were identified as effective indicators for the survival of patients with SKCM based on an analysis of the training set containing 230 SKCM patients. Two formulas based on the 5 mRNAs and 8 lncRNAs were developed independently to predict the survival probability of SKCM patients, which were validated within the training dataset and the whole dataset. Patients with high-risk scores tend to have lower survival probabilities, while those with low-risk scores tend to live longer.

Patients with high-risk scores tend to have high CD38, CXCL10, BHLHE40-AS1 expressions and low CCL8, HAPLN3, GBP4, AP000866.1, AC099343.3, AL137003.2, AC015911.3, *HCP5*, FUT8-AS1 and AC036108.3 expressions. The knockdown of CD38 and CXCL10 helps inhibit tumorigenesis[23,25]. C-C chemokine ligand 8 (CCL8) was reported to mediate the migration of CCR5(+) regulatory T cells, and HAPLN3 was reported to have a slight effect on cancer development[24]. Guanylate binding protein 4 GBP4 acts as a critical inflammasome adaptor required for prostaglandin biosynthesis and bacterial clearance by neutrophils[26]. Here, we may have formed a new understanding regarding their cooperative relationships for the survival of SKCM patients.

Although the Kaplan-Meier curves and the log-rank test both showed that the risk scores could efficiently predict the survival probability of SKCM patients, the AUC was 0.718, which is not enough. In addition, the mRNAs employed for prediction are only five. Thus, lncRNAs, miRNAs, and SNP, together with other factors, should be introduced into the prediction formula of risk score, which should better benefit the prognosis predictions of patients.

Conclusions

During further investigation, we noticed limited information on these signatures, especially the information on these lncRNAs. As they perform well in the survival analysis, more wet lab research should be conducted in the following study.

Conflict of Interest

The Authors declare that they have no conflict of interests.

Authors' Contribution

Concept: Shaoxia Zi and Cheng-Zhi Wang; Data Collection and/or Processing: Shaoxia Zi and Yuming Xing.

Funding

This study was supported by the Emergency General Hospital Fund (K201215).

Ethics Approval

TCGA belongs to public databases. Users can download relevant data for free for research and publish relevant articles. Our study is based on open source data, so there are no ethical issues and other conflicts of interest.

Informed Consent

Not required.

ORCID ID

Shaoxia Zi: 0000-0001-8697-5009.

References

1) Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA: Ca-cancer J Clin 2018; 68: 7-30.
2) Pal HC, Hunt KM, Diamond A, Elmets CA, Afaq F. Phytochemicals for the Management of Melanoma. Mini Rev Med Chem 2016; 16: 953-979.
3) Abigail Maciolek C, Buchanan PJ, Bueno RA, Neumeister MW. Subungual melanoma: a review of current treatment. Plast Reconstr Surg 2014; 134: 259-273.
4) Rozeman EA, Dekker TJA, Haanen JBAG, Blank CU. Advanced Melanoma: Current Treatment Options, Biomarkers, and Future Perspectives. Am J Clin Dermatol 2018; 19: 303-317.
5) Zhou AY, Johnson DB. Combinatorial Therapies in Melanoma: MAPK Inhibitors and Beyond. Am J Clin Dermatol 2018; 19: 181-193.
6) Wang S, Fan W, Wan B, Tu M, Jin F, Liu F, Xu H, Han P. Characterization of long non-coding RNA and messenger RNA signatures in melanoma tumorigenesis and metastasis. PloS one 2017; 12: e0172498.
7) Richtig G, Ehall B, Richtig E, Aigelsreiter A, Gutschner T, Pichler M. Function and Clinical Im-

plications of Long Non-Coding RNAs in Melanoma. Int J Mol Sci 2017; 18: 715-719.

8) Leucci E, Coe EA, Marine J-C, Vance KW. The emerging role of long non-coding RNAs in cutaneous melanoma. Pigment Cell Melanoma Res 2016; 29: 619-626.

9) Guo W, Xu T, Lee JJ, Murphy GF, Lian CG. Epigenetic markers in melanoma. Melanoma Manag 2015; 2: 367-382.

10) Branca M, Orso S, Molinari RC, Xu H, Guerrier S, Zhang Y, Mili N. Is nonmetastatic cutaneous melanoma predictable through genomic biomarkers? Melanoma Res 2017; 28: 1-7.

11) Winnepenninckx V, On behalf of the Melanoma Group of the European Organization for R, Treatment of C. Gene Expression Profiling of Primary Cutaneous Melanoma and Clinical Outcome. J Natl Cancer Inst 2006; 98: 472-482.

12) Winnepenninckx V, van den Oord JJ. Gene expression profiling and clinical outcome in melanoma: in search of novel prognostic factors. Expert Rev Anticancer Ther 2007; 7: 1611-1631.

13) Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp oncol 2015; 19: A68-A77.

14) Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. Ann Stat 1982; 10: 1100-1120.

15) Wang JY, Tai JJ. Robust quantitative trait association tests in the parent-offspring triad design: conditional likelihood-based approaches. Ann Hum Genet 2012; 73: 231-244.

16) Renaud G, Stenzel U, Maricic T, Wiebe V, Kelso J. deML: robust demultiplexing of Illumina sequences using a likelihood-based approach. Bioinformatics 2014; 31: 770-772.

17) Kendall WL, Pollock KH, Brownie C. A likelihood-based approach to capture-recapture estimation of demographic parameters under the robust design. Biometrics 1995; 51: 293-308.

18) Hsu TH, Jiang SY, Chang WL, Eckert RL, Scharadin TM, Chang TC. Involvement of RARRES3 in the regulation of Wnt proteins acylation and signaling activities in human breast cancer cells. Cell Death Differ 2015; 22: 801-814.

19) Shyu RY, Jiang SY, Chou JM, Shih YL, Lee MS, Yu JC, Jao SW. RARRES3 expression positively correlated to tumour differentiation in tissues of colorectal adenocarcinoma. Br J Cancer 2003; 89: 146-151.

20) Wang Q, Wang X, Liang Q, Wang S, Xiwen L, Pan F, Li D. Distinct prognostic value of mRNA expression of guanylate-binding protein genes in skin cutaneous melanoma. Oncol lett 2018; 15: 7914-7922.

21) Qian J, Liu H, Wei S, Liu Z, Li Y, Wang LE, Wei Q. Association between putative functional variants in the PSMB9 gene and risk of melanoma--re-analysis of published melanoma genome-wide association studies. Pigment Cell Melanoma Res 2013; 26: 392-401.

22) Barbai T, Fejős Z, Puskas LG, Tímár J, Rásó E. The importance of microenvironment: the role of CCL8 in metastasis formation of melanoma. Oncotarget 2015; 6: 29111-29128.

23) Bu X, Kato J, Hong JA, Merino MJ, Schrump DS, Lund FE, Moss J. CD38 knockout suppresses tumorigenesis in mice and clonogenic growth of human lung cancer cells. Carcinogenesis 2018; 39: 242-251.

24) Shou-Jen K, Su-Yu C, Che L, Szu-Erh C, Hsiu-Ting T, Dar-Ren C. Significant elevation of CLDN16 and HAPLN3 gene expression in human breast cancer. Oncol Rep 2010; 24: 759-763.

25) Gao J, Wu L, Wang Y, Cui S, Duan S, Dong Z, Chen X. Knockdown of Cxcl10 Inhibits Mesangial Cell Proliferation in Murine Habu Nephritis Via ERK Signaling. Cell Physiol Biochem 2017; 42: 2118-2125.

26) Tyrkalska SD, Candel S, Angosto D, Gómez-Abellán V, Martín-Sánchez F, García-Moreno D, Mulero V. Neutrophils mediate Salmonella Typhimurium clearance through the GBP4 inflammasome-dependent production of prostaglandins. Nat comm 2016; 7: 12077-12082.