Thoracic Cancer Open Access

Thoracic Cancer ISSN 1759-7706

ORIGINAL ARTICLE

lncRNAs classifier to accurately predict the recurrence of thymic epithelial tumors

Yongchao Su[1], Yongbing Chen[2], Zuochun Tian[1], Chuangang Lu[1], Liang Chen[3] & Ximiao Ma[4]

1 Department of Thoracic Surgery, Sanya Central Hospital, Sanya, China
2 Department of Thoracic Surgery, The Second Affiliated Hospital of Soochow University, Suzhou, China
3 Department of Respiratory Medicine, Sanya Central Hospital, Sanya, China
4 Department of Thoracic Surgery, Haikou People's Hospital, Haikou, China

Keywords
lncRNAs classifier; prediction; recurrence-free survival; thymic epithelial tumors.

Correspondence
Yongchao Su, Department of Thoracic Surgery, Sanya Central Hospital, Sanya 572000, Hainan Province, China.
Tel: 86+15008059632
Fax: 86+15008059632
Email: 15008059632@139.com

Received: 30 January 2020;
Accepted: 30 March 2020.

doi: 10.1111/1759-7714.13439

Thoracic Cancer **11** (2020) 1773–1783

Abstract

Background: Long non-coding RNAs (lncRNAs), which have little or no ability to encode proteins, have attracted special attention due to their potential role in cancer disease. In this study we aimed to establish a lncRNAs classifier to improve the accuracy of recurrence prediction for thymic epithelial tumors (TETs).

Methods: TETs RNA sequencing (RNA-seq) data set and the matched clinicopathologic information were downloaded from the Cancer Genome Atlas. Using univariate Cox regression and least absolute shrinkage and selection operator (LASSO) analysis, we developed a lncRNAs classifier related to recurrence. Functional analysis was conducted to investigate the potential biological processes of the lncRNAs target genes. The independent prognostic factors were identified by Cox regression model. Additionally, predictive ability and clinical application of the lncRNAs classifier were assessed, and compared with the Masaoka staging by receiver operating characteristic (ROC) analysis and decision curve analysis (DCA).

Results: Four recurrence-free survival (RFS)-related lncRNAs were identified, and the classifier consisting of the identified four lncRNAs was able to effectively divide the patients into high and low risk subgroups, with an area under curve (AUC) of 0.796 (three-year RFS) and 0.788 (five-year RFS), respectively. Multivariate analysis indicated that the lncRNAs classifier was an independent recurrence risk factor. The AUC of the lncRNAs classifier in predicting RFS was significantly higher than the Masaoka staging system. Decision curve analysis further demonstrated that the lncRNAs classifier had a larger net benefit than the Masaoka staging system.

Conclusions: A lncRNAs classifier for patients with TETs was an independent risk factor for RFS despite other clinicopathologic variables. It generated more accurate estimations of the recurrence probability when compared to the Masaoka staging system, but additional data is required before it can be used in clinical practice.

Introduction

Thymic epithelial tumors (TETs), which arise from the epithelial cells of the thymus, represent the most common neoplasms in the anterior mediastinum, but are among the rarest of all cancers.[1] According to the 2015 World Health Organization (WHO) classification, TETs are divided into thymoma (A, A/B, B1, B2, B3 subtypes) and thymic carcinoma (TC) (C subtypes) based on the tumor cell morphology, degree of atypia, and extent of the thymocyte component.[2] It has been reported that five-year median survival is about 66% in TC and reaches up to 90% in thymoma.[3] Surgical resection is considered the cornerstone curative treatment. However, local recurrence or distant metastasis may occur in some patients even after complete resection which render major obstacles to long-term survival in TETs.[4] Although various systemic treatment options exist for patients with locally advanced or metastatic disease, none are curative.

The prognosis of patients with TETs is largely determined by the histological type, which is complex. This complexity has led to the present lack of uniform measurement standards. Use of the WHO classification and Masaoka staging system at diagnosis were reported by Marx *et al.* to be the main prognostic factors for recurrence and patient survival.[5] In general, most type A and AB thymomas have low malignant potential, whereas types B1, B2, and B3 thymomas are more aggressive, with B3 thymoma having the greatest tendency for intrathoracic spread. On the contrary, TC is a highly aggressive tumor with frequent lymphatic and hematogenous metastasis.[6] However, its prognostic significance in guiding further treatment is controversial.[7] Hence, identifying reliable and accurate predictive markers to screen out which subset of patients with TETs is vulnerable to develop recurrence is urgently needed.

As previous genome studies have revealed, more than 90% of the human genome is actively transcribed into non-coding RNA (ncRNAs).[8] Conventionally, this ncRNA family is loosely classified into two groups based on molecular size: small ncRNA (eg, microRNA; less than 200 nt in length) and long non-coding RNA (lncRNA; more than 200 nt in length).[9] Unlike protein-coding RNAs, the expression patterns of the lncRNAs are more specific. A large number of studies have reported the diverse biological functions of lncRNAs, such as tumorigenesis, tumor progression, as well as metastasis.[10]

LncRNA are potential new cancer biomarkers, and represent a large number of potential molecular drivers in human cancer disease.[11] Over the past few years, a classifier comprising multiple lncRNAs has been reported in several studies to be able to evaluate the prognostic factors in various cancers, including gastric cancer, cervical carcinoma, head and neck cancer, and laryngeal squamous cell carcinoma.[12–15] However, a lncRNA classifier that can predict the recurrence-free survival (RFS) outcome of TETs has not as yet been determined.

In the current study, by mining the expression data of lncRNAs in The Cancer Genome Atlas (TCGA), we identified lncRNAs that were significantly related to recurrence outcome, and then developed a multiple lncRNAs classifier. We assessed the predictive ability and clinical application of the lncRNAs classifier, and compared it with the WHO classification. In addition, we evaluated the prediction effect of the lncRNAs classifier in clinical subgroups (thymoma and TC).

Methods

Collection of publicly available data from TCGA

TETs RNA sequencing (RNA-seq) dataset and relevant clinical information including age, sex, height, weight, race, sample initial weight, tumor site, mutation count, WHO histological types, Masaoka staging, and RFS time were downloaded from the publicly available TCGA database (https://gdc.cancer.gov/). A total of 114 patients with complete follow-up data were extracted, which had been recorded before 20 December 2019. The clinical endpoint was RFS, defined as the time from final surgical excision to recurrence. Patients not having a recurrence or those patients who died without recurrence were censored at the time of last follow-up. All the data was obtained from TCGA, and informed consent was obtained from the patients before our study commenced.

Given that the expression level of lncRNAs is relatively low compared with non-coding RNA, it is likely that some lncRNAs have not been analyzed during the sequencing procedure of lncRNAs. Considering this possibility, we defined lncRNAs as being expressed abundantly when their expression level was above 0 and occurred more than 50% in the total samples. The final expression level of lncRNAs was represented as $\log_2(x + 1)$ of the original expression level.

Construction and confirmation of a lncRNAs signature

First, moderated t-statistics method and Benjamini-Hochberg procedure were used to identify distinct differential lncRNAs between normal tissues and TETs tissues, with $P < 0.05$ and the false discovery rate (FDR) <0.05 for filtration. Next, univariate Cox regression analysis was used to select RFS-related lncRNAs that were statistically significant ($P < 0.01$). After primary filtering, a least absolute shrinkage and selection operator (LASSO) analysis was established to select candidate lncRNAs with penalty parameters tuning adjusted by 10 times cross validation.[16] After layers of screening, the eligible lncRNAs were constructed as a classifier. The risk score formula was generated by integrating the RFS-related lncRNA, weighted by their respective LASSO regression coefficients. According to this formula, each patient's risk score was calculated, and patients were divided into high or low risk groups on the basis of the optimal cutoff point, which was adopted in the maximum sensitivity and specificity by using a receiver operating characteristic (ROC) curve (time-independent). The survival differences between high and low risk groups were further compared by Kaplan-Meier analysis with a log-rank test. Stratified analysis based on clinical characteristics (thymoma vs. TC, Masaoka stages I–II vs. Masaoka stages III–IV) was conducted to evaluate the discrimination ability of the lncRNAs classifier.

Function prediction of the prognostic lncRNAs

In TCGA dataset, according to their expression level, Pearson correlation algorithm is performed between the identified lncRNAs and the protein-coding genes (mRNAs). The correlation coefficient > 0.4, $P < 0.001$ are considered significant correlation. The potential biological processes of the lncRNAs target genes were investigated by using Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG). Database for Annotation, Visualization and Integrated Discovery (DAVID) is a common bioinformatics tool (http://david.abcc.ncifcrf.gov/, version 6.8),[17] which is used to explore the biological functions of the selected lncRNAs. *P*-values corrected with a false discovery rate (FDR) <0.05 for GO analysis and KEGG pathways are considered remarkably enriched functional annotations.

Prognostic value and clinical usefulness of lncRNAs classifier

We used univariate and multivariate Cox regression analysis to identify clinical risk parameters associated with RFS. Furthermore, we used ROC analysis to investigate and compare the discrimination ability of the lncRNAs classifier with WHO classification and Masaoka staging. Finally, decision curve analysis (DCA) was used to evaluate the clinical usefulness and net benefit of the lncRNAs classifier, and compared with WHO classification and Masaoka staging.[18]

Statistical analysis

Categorical variables are provided as proportions (%). Continuous variables are described as medians (interquartile ranges [IQRs]) if the distribution was non-normal, and as means (standard deviations [SDs]) if the distribution was normal. After classifying the patients with cancer recurrence, we calculated the best cutoff values of number of mutation count, which was a point when the Youden index (sensitivity +specificity-1) reached the maximum value through receiver operating curve (ROC) analysis.

If there were missed values in some of the potential predictors, these missing data would be imputed, as a complete case analysis would improve the statistical power and reduce the potentially biased results.[19] Multiple imputation (MI) was used to interpolate the missing data as the missing data were considered missing at random after analyzing patterns of them.[20] We used Markov chain Monte Carlo (MCMC) function to perform MI, and selected five iterations to account for possible simulation errors.

The area under curve (AUC) of ROC analysis is between 0.5 and 1. The prediction ability is low when value of AUC is 0.5–0.6, moderate when AUC is 0.6–0.7, and high when AUC is above 0.7.

LASSO algorithm was conducted with “glmnet” packages, and ROC analysis was done with “timeROC” and “survivalROC” packages. DCA was performed with the “stdca.R".

SPSS statistics 22.0 and R software (R version 3.5.2) were used to conduct the statistical analysis. A two-sided $P < 0.05$ was considered to be statistically significant.

Results

Demographic parameters and OS outcome of TETs patients

In the current study, 114 TETs patients with available lncRNAs data and clinicopathological characteristics were included. The basic clinicopathological features of these TETs patients are summarized in Table 1. The median follow-up time was 37.68 months (from 0.46 to 149.87 months). Of all the 114 TETs patients, 19 patients (16.7%) developed recurrence during follow-up. The estimated three-year and five-year RFS rates were 85.5% (78.3%–92.8%) and 81.4% (72.4%–90.4%), respectively.

Construction and confirmation of a lncRNAs signature

Based on the primary filter criteria mentioned in the methods section, we obtained a list of 63 different lncRNAs (Supplementary Material 1). Then, using univariate Cox regression analysis, we identified 15 prognostic related lncRNAs (Supplementary Material 2). Finally, LASSO algorithm was used to shrink and pick out the RFS-related lncRNAs (Fig 1), which were ADAMTS9-AS1, HSD52, LINC00968 and LINC01697, to build a lncRNAs-based classifier.

In order to better investigate the value of the lncRNAs classifier in predicting RFS, a risk score was established, with the coefficients weighted by a LASSO Cox regression model. The risk score was generated as follows: risk score = (0.0028 expression level of ADAMTS9-AS1) + (0.1311 expression level of HSD52) + (0.0115 expression level of LINC00968) + (0.0044 expression level of LINC01697). Using ROC curve to generate the optimal cutoff value for the risk score, patients were divided into high and low risk groups. As shown in Fig 2, patients with a high risk score were more likely to die and had shorter

Table 1 Characteristics of study population with number of missing values (*n* = 114)

Variable	Category	No. (%) or median (IQR)	Missing values (%)
Age (years)		58.5 (17–84)	0 (0)
Sex	Female	55 (48.2)	0 (0)
	Male	59 (51.8)	
Height (cm)		166 (139–194)	21 (18.4)
Weight (g)		77 (44–155)	19 (16.7)
Race	White	95 (83.3)	2 (1.8)
	Black or African American	6 (5.3)	
	Asian	11 (9.6)	
Sample initial weight (g)		20 (125–1170)	0 (0)
Tumor site	Thymus	87 (76.3)	0 (0)
	Anterior mediastinum	27 (23.7)	
WHO histological types	A–B3 type	104 (91.2)	0 (0)
	C type	10 (8.8)	0 (0)
Masaoka staging	I–II	91	
	III–IV	23	0 (0)
Mutation count		9 (1–644)	1 (0.9)

IQR, interquartile range.

RFS time than patients with a low risk score (18.2 vs. 115.1 months, HR = 6.3, 95% CI: 2.6–15.6, P = 0.005). The lncRNAs classifier had a superior prediction effect, with an AUC of 0.796 (three years RFS) and an AUC of 0.788 (five years RFS) (Fig 2c). Additionally, when stratification analysis was performed according to WHO classification (thymoma vs. TC) and Masaoka stages (I–II vs. III–IV), the lncRNAs classifier seemed to remain a clinically and statistically significant prognostic model (Fig 3).

Functional enrichment analysis of lncRNAs

To investigate the potential function of the four lncRNAs, a total of 519 protein-coding genes (mRNAs) were significantly correlated with at least one of the four lncRNAs (Pearson coefficient > 0.4, P < 0.001), which were considered eligible for pathway enrichment (Supplementary Material 3). The four lncRNAs were mainly related to extracellular matrix structural constituent, glycosaminoglycan binding, growth factor binding and so on (Fig 4a). The KEGG pathway analysis revealed that the four lncRNAs related target genes (519 protein coding genes) were mainly enriched in spliceosome, cell cycle, DNA replication and so on (Fig 4b).

Prognostic value and clinical usefulness of lncRNAs classifier

Using univariate Cox analysis, we identified that the Masaoka staging and the lncRNAs classifier were associated with RFS (Table 2). Multivariate Cox analysis continued to verify that the lncRNAs classifier was an independent risk factor for RFS, regardless of other clinicopathologic variables. In addition, time-independent ROC analysis uncovered WHO classification had a low prediction value, with an AUC of 0.567 (three years RFS) and an AUC of 0.533 (five years RFS) (Fig S1A). Time-independent ROC analysis revealed that Masaoka stages had a moderate prediction effect, with an AUC of 0.629 (three years RFS) and an AUC of 0.665 (five years RFS) (Fig S1B). Further, to evaluate the discrimination ability of the lncRNAs classifier, we compared the lncRNAs classifier with WHO classification

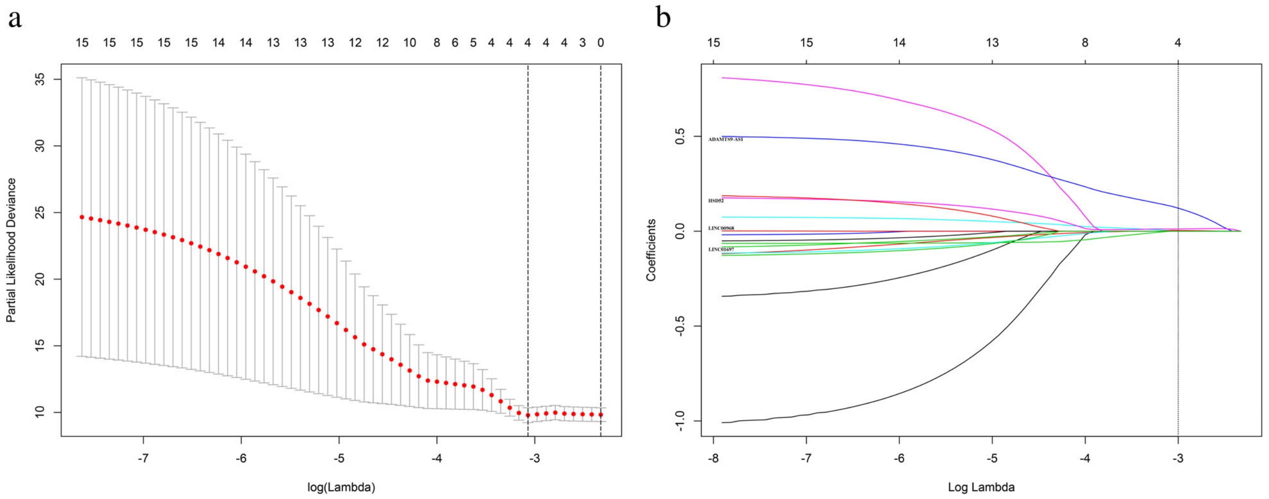

Figure 1 Four lncRNAs selected by LASSO Cox regression analysis. (**a**) The two dotted vertical lines are drawn at the optimal values by minimum criteria (left) and 1 - s.e. criteria (right). Details are provided in Methods. (**b**) LASSO coefficient profiles of the 15 lncRNAs. A vertical line is drawn at the optimal value by minimum criteria and results in four nonzero coefficients. Four lncRNAs—ADAMTS9-AS1, HSD52, LINC00968 and LINC01697—with coefficients 0.0028, 0.1311, 0.0115, 0.0044, respectively, were selected in the LASSO Cox regression model.

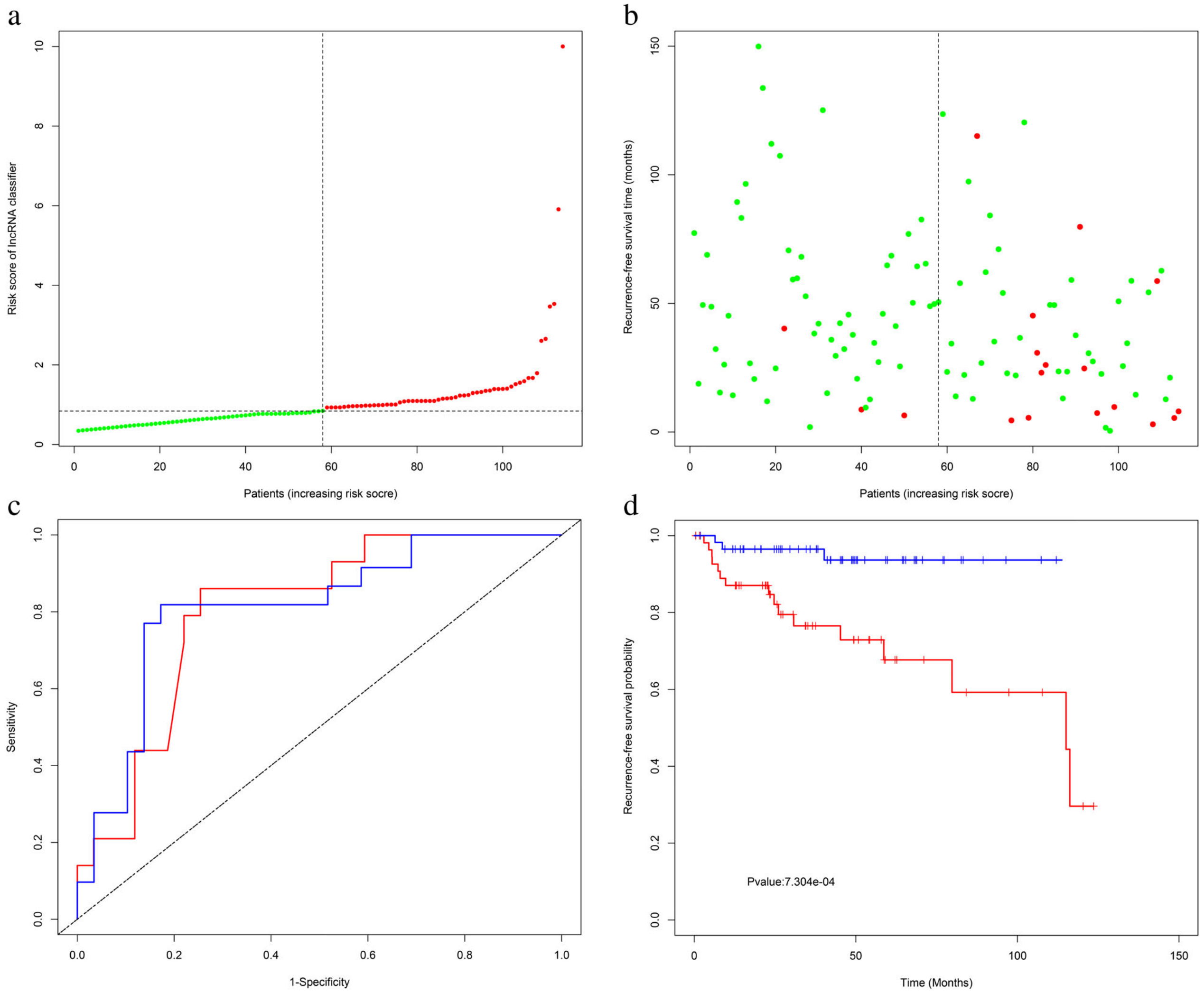

Figure 2 Development of lncRNAs signature for prediction of survival in TETs patients. (**a** and **b**) Distribution of lncRNAs-based classifier risk score (●) Live (●) Dead. (**c**) Time-independent ROC curves with AUC values to evaluate predictive efficacy of lncRNA signature risk score (——) AUC of 3 years of RFS:0.796 (——) AUC of 5 years of RFS:0.788. (**d**) Kaplan-Meier estimates of patients' survival status and time using the optimal lncRNA signature risk score cutoff which divided patients into low and high risk groups. (——) High risk (——) low risk.

and Masaoka staging. ROC analysis indicated that the lncRNAs classifier (AUC of three-year 0.796, and AUC of five-year 0.788) was better than WHO classification (AUC of three-year 0.567, and AUC of five-year 0.533) (Fig S2) and Masaoka stages (AUC of three-year 0.629, and AUC of five-year 0.665) in predicting RFS (Fig 5). Finally, DCA was used to compare the clinical usability of the lncRNAs classifier to that of traditional WHO classification and Masaoka staging. Based on a continuum of potential thresholds for death (x-axis) and the net benefit of using the lncRNAs classifier to risk-stratify patients (y-axis) relative to assuming all patients will recur, DCA graphically revealed that the lncRNAs classifier was superior to the traditional WHO classification (Fig S3) and Masaoka staging (Fig 6).

Discussion

Analyzing TETs RNA sequencing (RNA-seq) data set and the relevant clinical parameters of 114 TETs patients from TCGA, we identified four lncRNAs related to RFS. On the basis of these lncRNAs, we developed a lncRNAs classifier which could accurately categorize patients into high and low risk status. Additionally, the lncRNAs classifier effectively predicted recurrence probability, with a three-year AUC of 0.796 and five-year AUC of 0.788, which possessed better predictive ability and clinical usability than Masaoka staging.

TETs is a heterogeneous group, comprising different subsets with distinct outcomes. This heterogeneity may be ascribed to differences in the biologic behaviors of tumors. Traditional prognostic factors are not helpful in predicting

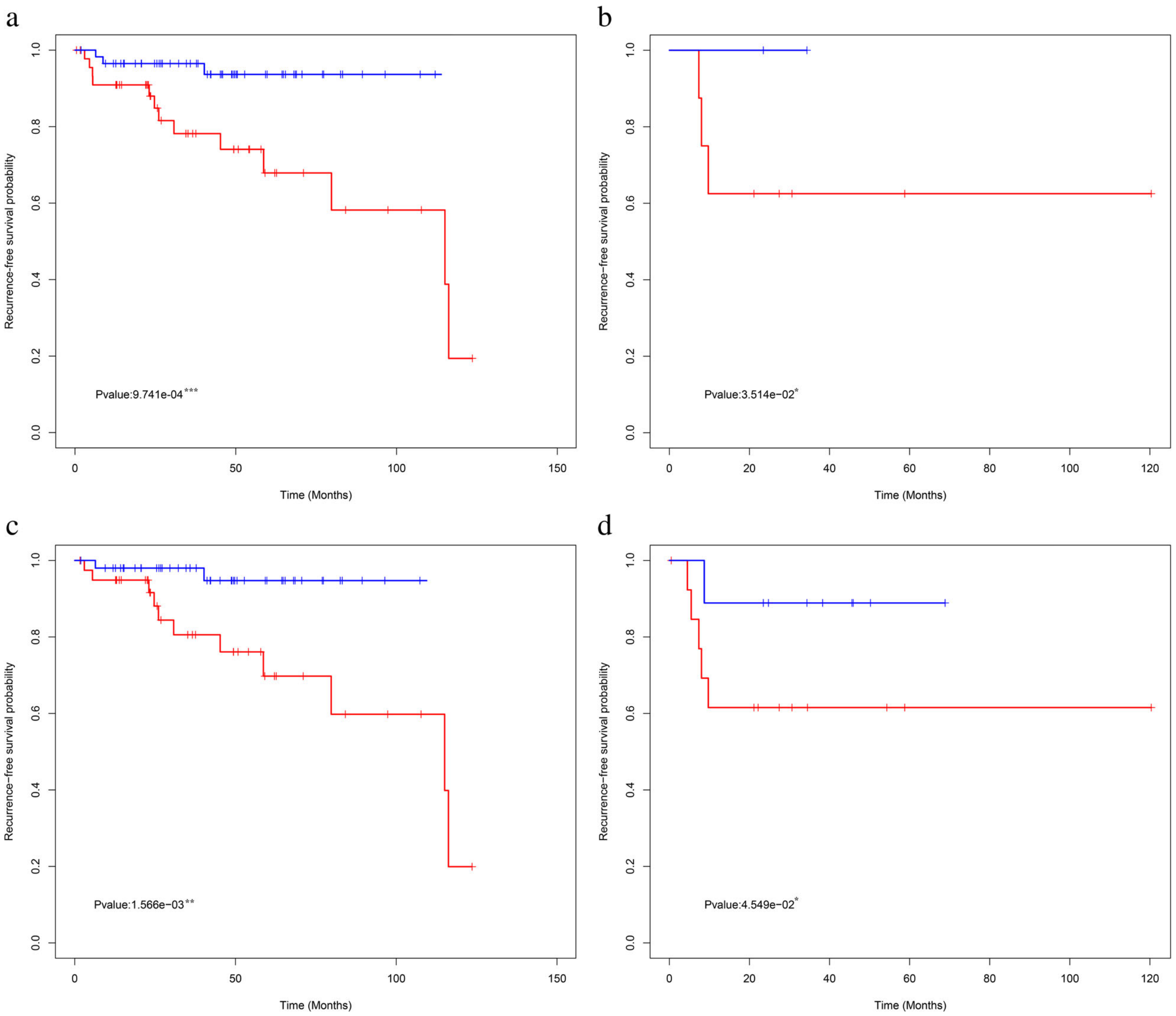

Figure 3 Kaplan-Meier survival analysis according to the lncRNAs classifier stratified by WHO classification and Masaoka stages. (**a**) Thymoma, (——) High risk (——) low risk (**b**) TC; (——) High risk (——) low risk (**c**) I–II, (——) High risk (——) low risk and (**d**) III–IV (——) High risk (——) low risk. *P*-values were calculated using the log-rank test.* $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

which patients with TETs will develop recurrence. Molecular investigation of TETs could provide information for predicting recurrence and for triaging the patients who may require and benefit from adjuvant therapy. Hence, in the current study, using TCGA database containing large-scale lncRNAs expression data, we aimed to identify RFS-related lncRNAs and establish a lncRNAs classifier, which may be more valuable for TETs patients to optimize tailored treatment in the era of precision medicine.

To our knowledge, this is the first study to construct a lncRNAs classifier consisting of ADAMTS9-AS1, HSD52, LINC00968 and LINC01697, for predicting recurrence probability in patients with TETs. It could effectively classify patients into a high risk group with shorter RFS and low risk group with longer RFS. Functional analysis suggested that lncRNAs target genes both participated in various biological processes and pathways in patients with TETs. Using stratified analysis, the lncRNA classifier appears to show a perfect discrimination ability in either the thymoma subgroup or TC subgroup, or Masaoka stages I–II subgroup or Masaoka stages III–IV subgroup. Additionally, we affirmed that the lncRNA classifier was an independent predictor, regardless of other clinicopathologic factors. In this study, in terms of the discrimination ability of the model, the performance of the lncRNA classifier in predicting recurrence ability was superior to the

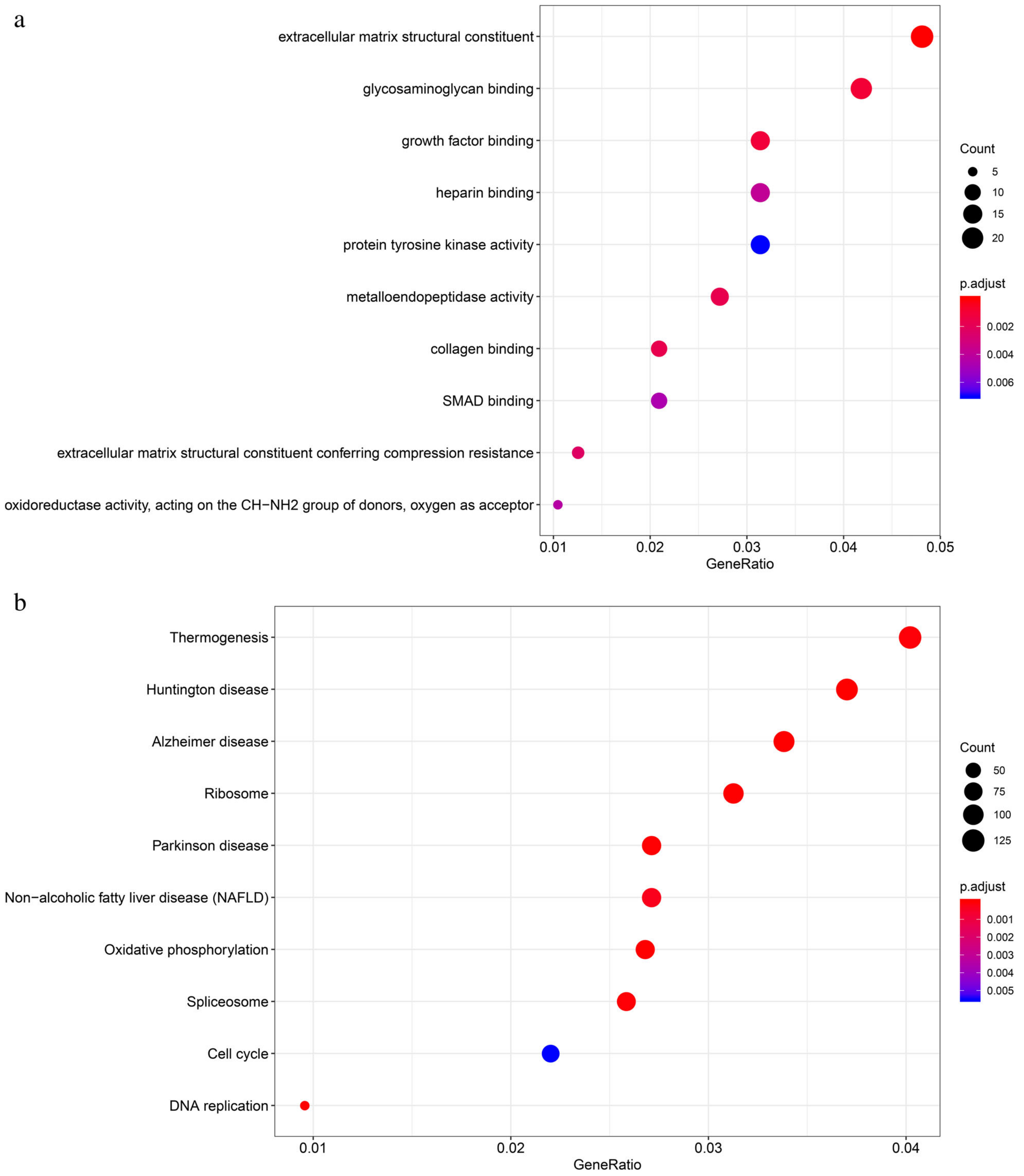

Figure 4 Functional annotation of the prognostic lncRNAs. (**a**) Significantly enriched using the co-expressed mRNAs of the lncRNAs in GO analysis and (**b**) KEGG pathway analysis.

WHO classification and Masaoka stages. Remarkably, DCA results showed that TETs recurrence-related treatment decision based on the nomogram led to more net benefit than treatment decisions based on the WHO classification or Masaoka stages, or treating either all patients or none. Taken together, the present lncRNA classifier would

Table 2 Univariate and multivariate Cox regression analysis for prediction of RFS

		Univariate analysis		Multivariate analysis	
Factors	Subgroup	HR (95%CI)	*P*-value	HR (95%CI)	*P*-value
Age		0.99 (0.96–1.03)	0.692	NA	NA
Sex	Female	1			
	Male	0.60 (0.24–1.50)	0.276	NA	NA
Height		0.99 (0.95–1.04)	0.750	NA	NA
Weight		1.00 (0.97–1.02)	0.704	NA	NA
Race	White	1			
	Black or African American	2.64 (0.60–11.67)	0.200	NA	NA
	Asian	0.33 (0.04–2.50)	0.281	NA	NA
Sample initial weight		1.00 (0.99–1.00)	0.320	NA	NA
Tumor site	Thymus	1			
	Anterior mediastinum	1.73 (0.65–4.56)	0.271	NA	NA
WHO histological types	A–B3 type	1			
	C type	2.15 (0.62–7.47)	0.230	NA	NA
Masaoka staging	I–II	1			
	III–IV	1.42 (1.12–2.04)	0.03*	1.21 (0.92–1.78)	0.10
Mutation count	<9	1			
	≥9	0.58 (0.22–1.52)	0.265	NA	NA
LncRNA classifier		1.17 (1.08–1.28)	<0.001*	1.17 (1.04–1.31)	0.008*

NA, not available. These variables were eliminated in the multivariate Cox regression model, so the HR and *P*-values were not available. *$P < 0.05$. CI, confidence intervals; HR, hazard ratio; RFS, recurrence-free survival.

be clinically useful for clinicians in tailoring recurrence-associated treatment decisions.

We identified four RFS-related lncRNAs that were significantly different among the TETs patients. Among lncRNAs, ADAMTS9-AS1, LINC00968 and LINC01697 have been previously reported to be associated with cancers, such as bladder cancer, esophageal squamous cell carcinoma (ESCC), breast cancer, ovarian cancer and non-small cell lung cancer (NSCLC).[21–25] Wang *et al.*[21] investigated the regulatory network of lncRNAs as competing endogenous RNAs (ceRNA) in bladder urothelial carcinoma (BUC) based on gene expression data derived from TCGA, which confirmed ADAMTS9-AS1 was a potential prognostic biomarker for BUC patients and was validated using gene expression profiling interactive analysis (GEPIA). Additionally, by using the differential co-expression method, Li *et al.*[22]

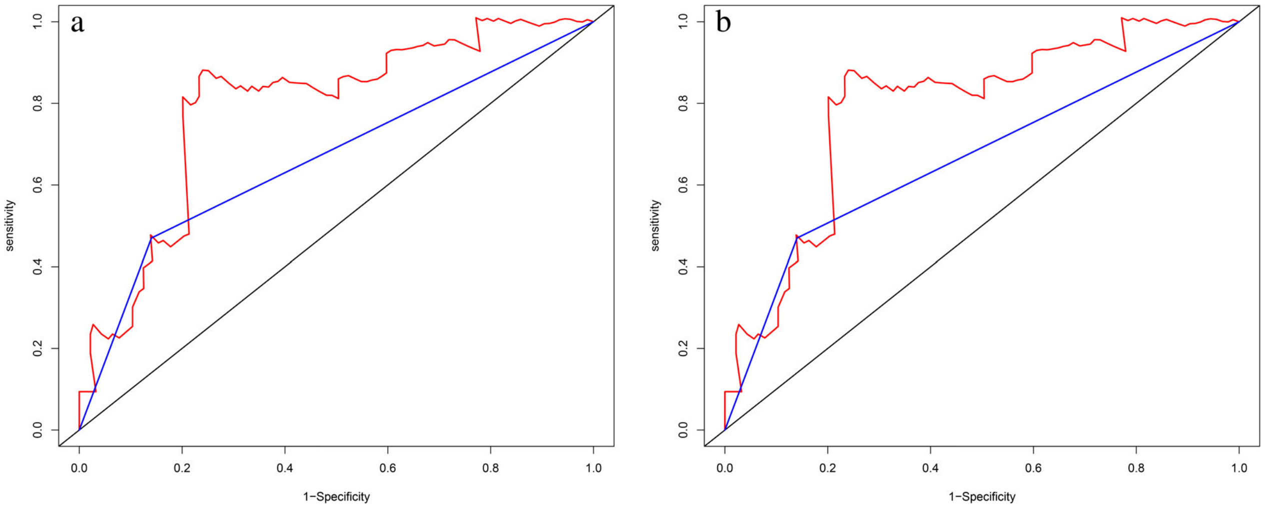

Figure 5 ROC curves compare the prognostic accuracy of the lncRNAs classifier with Masaoka staging in predicting (**a**) three-year RFS probability (——) LncRNA.classifier: 0.788 (——) Masaoka.staging: 0.665 and (**b**) five-year RFS probability. (——) LncRNA.classifier: 0.788 (——) Masaoka.staging: 0.665.

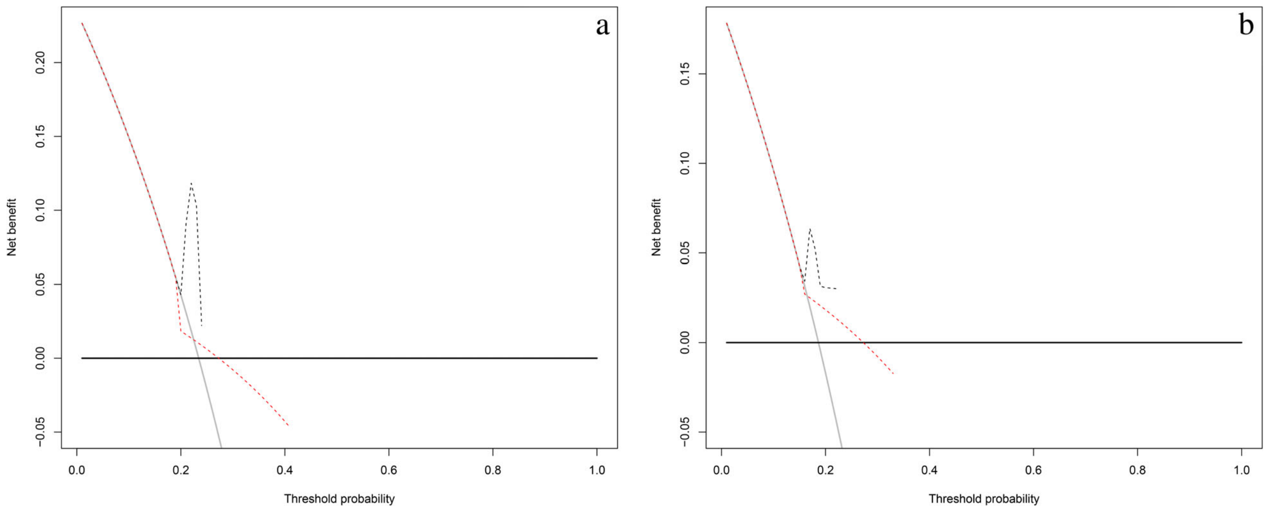

Figure 6 Decision curve analysis for the lncRNAs classifier and Masaoka staging in prediction of (**a**) three-year RFS probability (——) None (——) All (------) LncRNA.classifier (------) Masaoka.staging and (**b**) five-year RFS probability (——) None (——) All (------) LncRNA.classifier (------) Masaoka.staging.

identified two novel lncRNAs, and reported that ADAMTS9-AS1 may serve as a prognostic biomarker for clinical applications of ESCC. They also explored the mechanism of abnormal regulation of lncRNAs and determined that most of the differentially regulated links were modulated by 37 transcription factors. Sun *et al.*[23] found that overexpression of LINC00968 inhibited breast cancer cell proliferation, migration and tube formation abilities in vitro as well as tumor growth in vivo through inhibition of hsa-miR-423-5p, which speculates LINC00968 inhibits the progression of breast cancer through impeding hsa-miR-423-5p-mediated PROX1 inhibition. LINC00968 may be a potential therapeutic target for breast cancer therapy. Yao *et al.*[24] revealed that LINC00968 expression is markedly upregulated in ovarian cancer. Meanwhile, it arrests the cell cycle in the G1 phase by inhibiting the ERK and AKT pathways, thus accelerating ovarian cancer progression. By analyzing transcriptional profiling of LncRNAs, Liu *et al.*[25] uncovered that LINC01697 with accurate diagnosis value for NSCLC, was significantly correlated with NSCLC stage and survival time, and may be potential anti-NSCLC targets for drug development. Hence, further characterization of molecules such as ADAMTS9-AS1, HSD52, LINC00968 and LINC01697 will provide new perspective for the development and progress of TETs, and assist in finding potential therapeutic targets for TETs patients.

In our study, we found WHO histological types was not a significant association with recurrence among patients with TETs, which is in agreement with previous trials.[26,27] Nevertheless, several published trials reported that WHO histological types were an independent risk factor for TETs patients in predicting RFS.[28,29] Masaoka stage was significantly associated with recurrence among patients with TETs in univariate cox analysis, whereas not in multivariate cox analysis. One possible explanation is that a small sample size of TC or Masaoka stages III–IV (advanced stage), including 10 or 23 patients, respectively, is not enough to produce an effect size that is statistically significant. Additionally, several reports have revealed that age is linked with recurrence,[27,30] whereas the effect of age on the prognosis of RFS was not statistically significant. In the future, several large multicenter prospective cohorts with sufficient TC or advanced stage patients are needed to investigate the efficacy of these clinical variables. In addition to these clinical factors, as expected, the lncRNAs classifier was an effective independent prognostic factor for the prediction of patients with TETs.

Although the lncRNA classifier demonstrated impressive performance in TETs recurrence prediction, there are specific limitations associated with our trial. First, the presented lncRNA classifier based only on TCGA database with limited sample sizes for TETs, is not yet suitable for general use prior to validation of the predictive models with external datasets. Therefore, external and multicenter prospective cohorts with large sample sizes (sufficient TC or advanced stage patients) are still needed to validate the clinical application of our classifier.

Second, our choice of factors was limited to those available in our database. On account of the anonymous database, we could not extend our database with variables such as frequently reported completeness of surgical resection and elevated CRP, which have been previously reported to influence recurrence rates and prognosis in patients with TETs.[29,31] Further efforts to incorporate more patient-specific, tumor-specific and molecular factors into multivariate Cox analysis will potentially investigate whether the lncRNA classifier was still an independent risk factor or not.

Third, based on TETs RNA sequencing (RNA-seq) dataset in TCGA, we built a lncRNAs classifier which specifically predicted the recurrence of TETs. Whether the lncRNAs classifier still has prognostic value for other types of cancer should be clarified in future research.

Fourth, we did not explore the underlying biological function and pathways of the prognostic lncRNAs by vitro experiment, so further studies are needed to uncover the related mechanisms.

In conclusion, we built a lncRNAs classifier, based on TCGA database, for predicting recurrence probability in patients with TETs. The lncRNAs classifier is significantly better than Masaoka staging alone in terms of the predictive value and clinical usability. Importantly, our lncRNAs classifier appeared to present good discrimination ability in the thymoma and TC subgroups, as well as Masaoka stages I–II and Masaoka stages III–IV subgroups.

Disclosure

The authors declare that they have no competing interests.

References

1 Engels EA. Epidemiology of thymoma and associated malignancies. *J Thorac Oncol* 2010; **5** (10 Suppl 4): S260–5.

2 Travis WD, Brambilla E, Burke AP, Marx A, Nicholson AG. Introduction to the 2015 World Health Organization classification of tumors of the lung, pleura, thymus, and heart. *J Thorac Oncol* 2015; **10** (9): 1240–2.

3 Scorsetti M, Leo F, Trama A *et al.* Thymoma and thymic carcinomas. *Crit Rev Oncol Hematol* 2016; **99**: 332–50.

4 Hishida T, Nomura S, Yano M e a. Long-term outcome and prognostic factors of surgically treated thymic carcinoma: Results of 306 cases from a Japanese Nationwide database study. *Eur J Cardio-Thorac Surg* 2016; **49** (3): 835–41.

5 Marx A, Chan JK, Coindre JM *et al.* The 2015 World Health Organization classification of tumors of the thymus: Continuity and changes. *J Thorac Oncol* 2015; **10** (10): 1383–95.

6 Kondo K, Yoshizawa K, Tsuyuguchi M *et al.* WHO histologic classification is a prognostic indicator in thymoma. *Ann Thorac Surg* 2004; **77** (4): 1183–8.

7 Feng Y, Lei Y, Wu X *et al.* GTF2I mutation frequently occurs in more indolent thymic epithelial tumors and predicts better prognosis. *Lung Cancer* 2017; **110**: 48–52.

8 Stein LD. Human genome: End of the beginning. *Nature* 2004; **431** (7011): 915–6.

9 Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. *Cancer Discov* 2011; **1** (5): 391–407.

10 Tano K, Akimitsu N. Long non-coding RNAs in cancer progression. *Front Genet* 2012; **3**: 219.

11 Yarmishyn AA, Kurochkin IV. Long noncoding RNAs: A potential novel class of cancer biomarkers. *Front Genet* 2015; **6**: 145.

12 Zhu X, Tian X, Yu C *et al.* A long non-coding RNA signature to improve prognosis prediction of gastric cancer. *Mol Cancer* 2016; **15** (1): 60.

13 Mao X, Qin X, Li L *et al.* A 15-long non-coding RNA signature to improve prognosis prediction of cervical squamous cell carcinoma. *Gynecol Oncol* 2018; **149** (1): 181–7.

14 Cui J, Wen Q, Tan X. An integrated nomogram combining lncRNAs classifier and clinicopathologic factors to predict the recurrence of head and neck squamous cell carcinoma. *Sci Rep* 2019; **9** (1): 17460.

15 Cui J, Wen Q, Tan X. A genomic-clinicopathologic nomogram predicts survival for patients with laryngeal squamous cell carcinoma. *Dis Markers* 2019; **6**: 1–13.

16 Datta S, Le-Rademacher J, Datta S. Predicting patient survival from microarray data by accelerated failure time modeling using partial least squares and LASSO. *Biometrics* 2007; **63** (1): 259–71.

17 Huang W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37** (1): 1–13.

18 Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; **352**: i6.

19 Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: a gentle introduction to imputation of missing values. *J Clin Epidemiol* 2006; **59** (10): 1087–91.

20 Fletcher Mercaldo S, Blume JD. Missing data and prediction: the pattern submodel. *Biostatistics* 2020; **21**(2):236–252.

21 Wang J, Zhang C, Wu Y, He W, Gou X. Identification and analysis of long non-coding RNA related miRNA sponge regulatory network in bladder urothelial carcinoma. *Cancer Cell Int* 2019; **19**: 327.

22 Li Z, Yao Q, Zhao S, Wang Y, Li Y, Wang Z. Comprehensive analysis of differential co-expression patterns reveal transcriptional dysregulation mechanism and identify novel prognostic lncRNAs in esophageal squamous cell carcinoma. *Onco Targets Ther* 2017; **10**: 3095–105.

23 Sun X, Huang T, Zhang C *et al.* Long non-coding RNA LINC00968 reduces cell proliferation and migration and angiogenesis in breast cancer through up-regulation of PROX1 by reducing hsa-miR-423-5p. *Cell Cycle* 2019; **18** (16): 1908–24.

24 Yao N, Sun JQ, Yu L, Ma L, Guo BQ. LINC00968 accelerates the progression of epithelial ovarian cancer via

mediating the cell cycle progression. *Eur Rev Med Pharmacol Sci* 2019; **23** (11): 4642–9.

25 Liu J, Yao Y, Hu Z. Transcriptional profiling of long-intergenic noncoding RNAs in lung squamous cell carcinoma and its value in diagnosis and prognosis. *Mol Genet Genomic Med* 2019; **7** (12): e994.

26 Zhao M, Yin J, Yang X *et al.* Nomogram to predict thymoma prognosis: a population-based study of 1312 cases. *Thoracic Cancer* 2019; **10** (5): 1167–75.

27 Gokmen-Polar Y, Cook RW, Goswami CP *et al.* A gene signature to determine metastatic behavior in thymomas. *PLOS One* 2013; **8** (7): e66047.

28 Yanagiya M, Nitadori JI, Nagayama K, Anraku M, Sato M, Nakajima J. Prognostic significance of the preoperative neutrophil-to-lymphocyte ratio for complete resection of thymoma. *Surg Today* 2018; **48** (4): 422–30.

29 Janik S, Bekos C, Hacker P *et al.* Elevated CRP levels predict poor outcome and tumor recurrence in patients with thymic epithelial tumors: A pro- and retrospective analysis. *Oncotarget* 2017; **8** (29): 47090–102.

30 Li JF, Hui BG, Li X *et al.* Video-assisted thoracic surgery for thymoma: Long-term follow-up results and prognostic factors-single-center experience of 150 cases. *J Thorac Dis* 2018; **10** (1): 291–7.

31 Mou H, Liao Q, Hou X, Chen T, Zhu Y. Clinical characteristics, risk factors, and outcomes after adjuvant radiotherapy for patients with thymoma in the United States: Analysis of the surveillance, epidemiology, and end results (SEER) registry (1988-2013). *Int J Radiat Biol* 2018; **94** (5): 495–502.

Supporting Information

Additional Supporting Informationmay be found in the online version of this article at the publisher's website:

Figure S1 (A) Time-independent ROC curves with AUC values to evaluate predictive efficacy of WHO classification risk score. (B) Time-independent ROC curves with AUC values to evaluate predictive efficacy of Masaoka staging risk score.

Figure S2 ROC curves compare the prognostic accuracy of the lncRNAs classifier with WHO classification in predicting three-year RFS probability (A) and five-year RFS probability.

Figure S3 Decision curve analysis for the lncRNAs classifier and WHO classification in predicting three-year RFS probability (A) and five-year RFS probability.

Supplementary Material 1 Based on the primary filter criteria mentioned in the Methods section, we obtained a list of 63 different lncRNAs.

Supplementary Material 2 Using univariate Cox regression analysis, we identified 15 RFS-related lncRNAs.

Supplementary Material 3 Using the Pearson correlation algorithm, a total of 519 protein-coding genes (mRNAs) were significantly correlated with RFS-related lncRNAs.