OPEN ACCESS

ORIGINAL ARTICLE

LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma

Jiagen Li,[1] Zhaoli Chen,[1] Liqing Tian,[2] Chengcheng Zhou,[1] Max Yifan He,[1] Yibo Gao,[1] Suya Wang,[1] Fang Zhou,[1] Susheng Shi,[3] Xiaoli Feng,[3] Nan Sun,[1] Ziyuan Liu,[1] Geir Skogerboe,[2] Jingsi Dong,[1] Ran Yao,[1] Yuda Zhao,[1] Jian Sun,[1] Baihua Zhang,[1] Yue Yu,[1] Xuejiao Shi,[1] Mei Luo,[1] Kang Shao,[1] Ning Li,[1] Bin Qiu,[1] Fengwei Tan,[1] Runsheng Chen,[2] Jie He[1]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2013-305806).

For numbered affiliations see end of article

Correspondence to
Dr Jie He, Department of Thoracic Surgery, Cancer Institute and Hospital, Chinese Academy of Medical Sciences, Panjiayuannanli No 17, Chaoyang District, Beijing 10021, The People's Republic of China; prof.hejie@gmail.com and Runsheng Chen, Bioinformatics Laboratory and Laboratory of Noncoding RNA, Institute of Biophysics, Chinese Academy of Sciences, Datun Rd No. 15, Chaoyang District, Beijing, 100101, The People's Republic of China; crs@sun5.ibp.ac.cn

JL, ZC and LT contributed equally.

Received 2 August 2013
Revised 2 January 2014
Accepted 13 January 2014
Published Online First
12 February 2014

CrossMark

To cite: Li J, Chen Z, Tian L, *et al. Gut* 2014;**63**:1700–1710.

ABSTRACT

Background Oesophageal cancer is one of the most deadly forms of cancer worldwide. Long non-coding RNAs (lncRNAs) are often found to have important regulatory roles.

Objective To assess the lncRNA expression profile of oesophageal squamous cell carcinoma (OSCC) and identify prognosis-related lncRNAs.

Method LncRNA expression profiles were studied by microarray in paired tumour and normal tissues from 119 patients with OSCC and validated by qRT-PCR. The 119 patients were divided randomly into training (n=60) and test (n=59) groups. A prognostic signature was developed from the training group using a random Forest supervised classification algorithm and a nearest shrunken centroid algorithm, then validated in a test group and further, in an independent cohort (n=60). The independence of the signature in survival prediction was evaluated by multivariable Cox regression analysis.

Results LncRNAs showed significantly altered expression in OSCC tissues. From the training group, we identified a three-lncRNA signature (including the lncRNAs ENST00000435885.1, XLOC_013014 and ENST00000547963.1) which classified the patients into two groups with significantly different overall survival (median survival 19.2 months vs >60 months, p<0.0001). The signature was applied to the test group (median survival 21.5 months vs >60 months, p=0.0030) and independent cohort (median survival 25.8 months vs >48 months, p=0.0187) and showed similar prognostic values in both. Multivariable Cox regression analysis showed that the signature was an independent prognostic factor for patients with OSCC. Stratified analysis suggested that the signature was prognostic within clinical stages.

Conclusions Our results suggest that the three-lncRNA signature is a new biomarker for the prognosis of patients with OSCC, enabling more accurate prediction of survival.

Significance of this study

What is already known about this subject?

- Long non-coding RNAs (lncRNAs) have important regulatory roles in cancer formation and development.
- Some lncRNAs have been found to be associated with the survival of patients of various cancers.
- The tumour node metastasis staging system which relies on anatomical and pathological features has limitations in the prognosis of patients with oesophageal squamous cell carcinoma (OSCC).
- In many cancers, miRNA and mRNA prognostic signatures, which robustly predict the survival of patients, have been identified, but whether the lncRNA signature might also predict survival of patients with cancer remains unknown.

What are the new findings?

- LncRNA expression profile in OSCC tissues is profoundly different from that in normal oesophageal epithelial tissues.
- A three-lncRNA signature was identified which can reliably predict the survival of patients with OSCC.
- Like mRNAs and miRNAs, the lncRNA signature could be used as a biomarker for the prognosis of patients with cancer.

How might it impact on clinical practice in the foreseeable future?

- The lncRNA signature might help to predict the survival of patients with OSCC more accurately in clinical practice than previously possible.

INTRODUCTION

Oesophageal cancer ranks as the world's sixth most deadly cancer.[1] It has two major histological types: adenocarcinoma and squamous cell carcinoma (OSCC). In China, over 90% of the cases of oesophageal cancer are OSCC, which is the fourth most prevalent cancer of the country.[2] OSCC is a highly aggressive malignancy with poor prognosis. Better understanding of the genetic and molecular disorders of the disease is the key to early diagnosis, appropriate treatment and improved prognosis of patients with OSCC.

Long non-coding RNAs (lncRNAs) are transcripts longer than 200 nucleotides not translated into proteins.[3] [4] In recent years, lncRNAs have attracted increasing scientific interest and are believed to be implicated in diverse biological processes,[5] by promoting or repressing transcription,[6] or by acting as modulators of mRNA translation.[7] LncRNAs affect the transcription of numerous genes located throughout the genome,[6] the regulatory mechanisms being diverse and complex. Some lncRNAs regulate the transcription of nearby genes in *cis*, while others act in *trans*. Some lncRNAs regulate transcription through epigenetic pathways, while others interact directly with RNA polymerases or transcription factors.[8] The well-known lncRNA HOTAIR is overexpressed in breast cancer where it induces genome-wide retargeting of polycomb repressive complex 2 (PRC2).[9] This results in altered histone H3K27 methylation and gene expression, which further promotes cancer invasiveness and metastasis.[9] A large number of human lncRNAs have been identified, but their characteristics and functions remain largely unknown.[10]

An increasing number of studies have suggested deregulation of lncRNAs in cancers,[9] [11] [12] and reports on lncRNA expression profiles in specific cancers are beginning to be published. Studies on lncRNA expression profiles in five pairs of liver cancer and normal tissues,[13] six pairs of renal clear cell carcinoma and corresponding normal tissues,[14] and one glioblastoma tissue with one normal brain tissue from an age-matched donor[15] found large numbers of lncRNAs significantly deregulated in cancer tissues. A clear understanding of the alterations in lncRNA expression occurring in cancers will require larger-scale studies than those yet reported and as far as we know, our study is the first to employ more than 100 sample pairs. Microarray assay is a popular and reliable method of profiling lncRNA expression. Compared with RNA sequencing, microarray has the advantages of low cost, ‘lower technical variation and better detection sensitivity for low-abundance transcripts’ and the ability to quantify antisense single-exon lncRNAs.[16]

For most solid cancers, including OSCC, clinical stage of the cancer is still the main predictor of survival for patients who have received surgery, but it does not provide an accurate prediction. Cancers are heterogeneous at the molecular and genetic levels,[17] [18] and patients of the same stage and who have received similar treatment, may nonetheless have quite different clinical outcomes. A number of studies have shown that messenger RNAs (mRNAs) and microRNAs (miRNAs) can be powerful predictors of survival in patients with cancer, particularly those mRNA or miRNA signatures consisting of multiple markers.[19] [20] However, up to now, whether an lncRNA signature might have similar prognostic power to that of mRNA and miRNA signatures for patients with cancer is not known.

This study reports the first examination of lncRNA expression profiles in paired tumour and normal tissues in a large cohort of more than 100 patients with OSCC. We identified a three-lncRNA signature with the ability to predict the overall survival of patients with OSCC and validated its prognostic value in an independent cohort of 60 patients.

PATIENTS AND METHODS

Patients and samples

We retrospectively collected paired cancer and adjacent normal tissues from 119 patients with OSCC with follow-up information (minimum of 5 years) and examined the lncRNA expression profile of the tissues by microarray analysis. All patients had surgically proven primary OSCC and received oesophagectomy (R0 resection) at the Cancer Institute and Hospital of the Chinese Academy of Medical Sciences (CAMS) between December 2005 and December 2007. Samples were obtained with informed consent. To validate the prognostic signature, we enrolled an independent cohort of 60 patients with OSCC who underwent surgery at the Cancer Institute and Hospital, CAMS between January 2008 and December 2008 and examined the lncRNA expression level of their paired tumour and normal tissues using the same microarray assay as used for the original 119 patients. Details of the patient enrolment procedure are given in online supplementary methods and figure S1; clinical and pathological information of the patients is shown in online supplementary table S1. The study was approved by the medical ethics committee of the Cancer Institute and Hospital, CAMS.

RNA extraction, amplification, labelling and array hybridisation

Total RNA was first extracted from the tumour and normal tissues (see online supplementary methods) and used to produce labelled cDNA (see online supplementary methods). Array hybridisation using the labelled cDNA was performed in a CapitalBio BioMixerTM II hybridisation station (see online supplementary methods).

All the experimental procedures were done blinded to the clinical and pathological information and to the survival information of the patients.

Microarray processing and statistical analysis

LncRNA expression profiling was performed using the Agilent human lncRNA+mRNA array V.2.0 platform. After a filtering procedure, 8900 human lncRNAs (annotated by GENCODE (V13) database, lincRNAs from Cabili *et al*,[21] and the University of California Santa Cruz database) were selected for the following analysis (see online supplementary methods). First, quantile normalisation of the microarray data (containing the 8900 lncRNAs and all mRNAs in the microarray) of all 119 paired tumour–normal samples was carried out. Then, the data was log 2-scale transformed. Missing values were imputed using the random Forest unsupervised classification algorithm (see online supplementary methods). The data of the 60 sample pairs in the independent cohort were processed independently in the same way.

Hierarchical clustering of the lncRNA profiles was performed using cluster 3.0.[22] The normalised expression values of the lncRNAs were centred on the median before performing unsupervised hierarchical clustering. Clustering was done with complete linkage and centred Pearson correlation.

On the whole, lncRNAs have lower expression level than mRNAs. The average expression level of lncRNAs (after quantile normalisation and log 2 transformation) for the 119 paired tumour-normal samples was 5.93, while that of mRNAs was 10.19. In this study, we were only concerned with the lncRNAs with high and median expression values. LncRNAs with average expression value lower than five in both tumour and normal tissues of the 119 patients were deleted. Further, lncRNAs with invariable expression level (coefficient of variance <0.03) in 119 paired tissues were also filtered out. Finally, 4874 lncRNAs were left for further analysis.

For prognostic signature analysis, the 119 patients were first assigned into groups with good (47 patients) or poor prognosis (72 patients) according to an expected survival time of >5 or <5 years. They were then randomly divided into a training set (n=60) and a test set (n=59) using the random_shuffle function from C++ standard template library.

The 909 lncRNAs differentially expressed between tumour and normal tissues with absolute fold change >2 (false discovery rate adjusted p value of Student's t test <0.10 for all) in the 60 patients of the training set were selected from the 4874 lncRNAs (figure 1A,B). To reduce the influence of heterogeneity among different patients, the expression level of tumour minus normal was used for the following analysis.

Using random Forest supervised classification algorithm, nine lncRNAs mostly related to the prognostic classification were selected among the 909 lncRNAs (figure 1C) according to the permutation important score by the software Random Jungle (see online supplementary methods).[23]

There were $2^9-1=511$ combinations of the nine lncRNAs and we developed a signature for each combination from the training set using the nearest shrunken centroid algorithm. For each combination, two centroids ('good' and 'poor') were created using the mean gene expression profile of the lncRNAs based on the patients with good prognosis and those with poor prognosis, respectively. Then, the Euclid distances between all samples and the two centroids were calculated. If $d_{ig}<d_{ip}$ (d_{ig} is the Euclid distance between sample i and the centroid 'good', d_{ip} is that between sample i and the centroid 'bad'), sample i was predicted as 'good' (low-risk group); otherwise predicted as 'poor' (high-risk group) (figure 1D).

After the construction of all 511 signatures, we compared their classification accuracies in the training set. Because the sample size was not balanced between the 'good' and 'poor' groups, the classification accuracy was defined as the average of classification accuracy of the group with good prognosis and that of the group with poor prognosis. First, for signatures constructed by specific number of lncRNAs (k=1, 2, …, 9), the one with the highest classification accuracy was selected for each k (figure 1E). One of these selected signatures was then defined as the final signature, considering a balance between classification accuracy and the number of lncRNAs.

Quantitative RT-PCR

Quantitative RT-PCR (qRT-PCR) was performed to validate the microarray results. The reverse transcription reactions were carried out with reverse transcriptase (SuperScript III, Invitrogen) and quantitative PCR reactions were then performed on ABI 7900 (see online supplementary methods and supplementary table S2).

RESULTS

LncRNA expression profiles display significant differences between OSCC tissues and adjacent normal tissues

We first compared the lncRNA expression profiles of OSCC tissues and adjacent normal tissues using unsupervised hierarchical clustering in 119 patients. In total, 6389 lncRNAs with a coefficient of variance >0.10 were selected from the 8900 lncRNAs for clustering analysis. Hierarchical clustering of these 6389 lncRNAs based on centred Pearson correlation clearly separated OSCC tissues from normal tissues (figure 2). Only 12 samples (six tumour samples and six normal samples) were misclassified by the clustering analysis. Among all the lncRNAs, 799 showed at least a twofold change in the OSCC tissues compared with the normal tissues (355 being upregulated and 444 downregulated).

Derivation of a three-lncRNA prognostic signature from the training set

We next explored the association between lncRNA expression and the overall survival of patients with OSCC. A three-lncRNA signature including ENST00000435885.1, XLOC_013014 (annotated by Cabili *et al*[21]) and ENST00000547963.1) was selected from the training set considering a balance between accuracy and the number of lncRNAs (figure 1E). The expression level of the three lncRNAs measured by microarray was verified by qRT-PCR (see online supplementary results and supplementary figure S2). In this signature, the 'good' and 'poor' centroids were (−2.11, −1.35, 3.38) and (−0.57, −2.50, 2.38), which represented the average expression level of the three lncRNAs for the patients with good and poor prognosis, respectively. The signature was defined as follows:

$$d_{ig}=\sqrt{(E_1^i+2.11)^2+(E_2^i+1.35)^2+(E_3^i-3.38)^2}$$

$$d_{ip}=\sqrt{(E_1^i+0.57)^2+(E_2^i+2.5)^2+(E_3^i-2.38)^2}$$

where $E_1^iE_2^iE_3^i$ denoted the expression level of ENST00000435885.1, XLOC_013014, ENST00000547963.1 for sample i, respectively. A patient was classified as 'low risk' if $d_{ig}<d_{ip}$ according to the patient's three-lncRNA expression value and as 'high risk' if not.

A three-lncRNA signature predicts survival of patients with OSCC

With the three-lncRNA signature, patients of the training group were divided into a high-risk group (n=33) or a low-risk group (n=27). Patients with the high-risk signature had significantly shorter overall survival than those with the low-risk signature (median survival 19.2 months vs >60 months, p<0.0001) (figure 3A,D). There was no significant difference in clinical and pathological characteristics between high- and low-risk group patients (table 1).

The three-lncRNA signature was then tested for its prognostic value in the test group of 59 patients. The same model and criteria as those derived from the training group classified 25 and 34 patients of the test group into the high-risk and low-risk groups, respectively. As in the training group, the overall survival time of the high-risk group patients was significantly shorter than that of low-risk group patients (median survival 21.5 months vs >60 months, p=0.0030) (figure 3B,E). The two groups of patients differed significantly in N stage (p=0.0290), tumour node metastasis (TNM) stage (p=0.0378) and arrhythmia (p=0.0055), but not in other clinical and pathological factors (table 1).

To validate the prognostic value of the three-lncRNA signature, we used the lncRNA expression values and survival data of an independent cohort of 60 patients. The patients of the independent cohort were classified as high-risk (37 patients) or low-risk (23 patients) according to their three-lncRNA signature (median survival 25.8 months vs >48 months, p=0.0187) (figure 3C,F). The two groups of patients did not differ significantly in clinical and pathological characteristics (table 1).

Survival prediction by the three-lncRNA signature is independent of clinical and pathological factors

To assess whether the survival prediction ability of the three-lncRNA signature is independent of other clinical or pathological factors of the patients with OSCC, multivariable

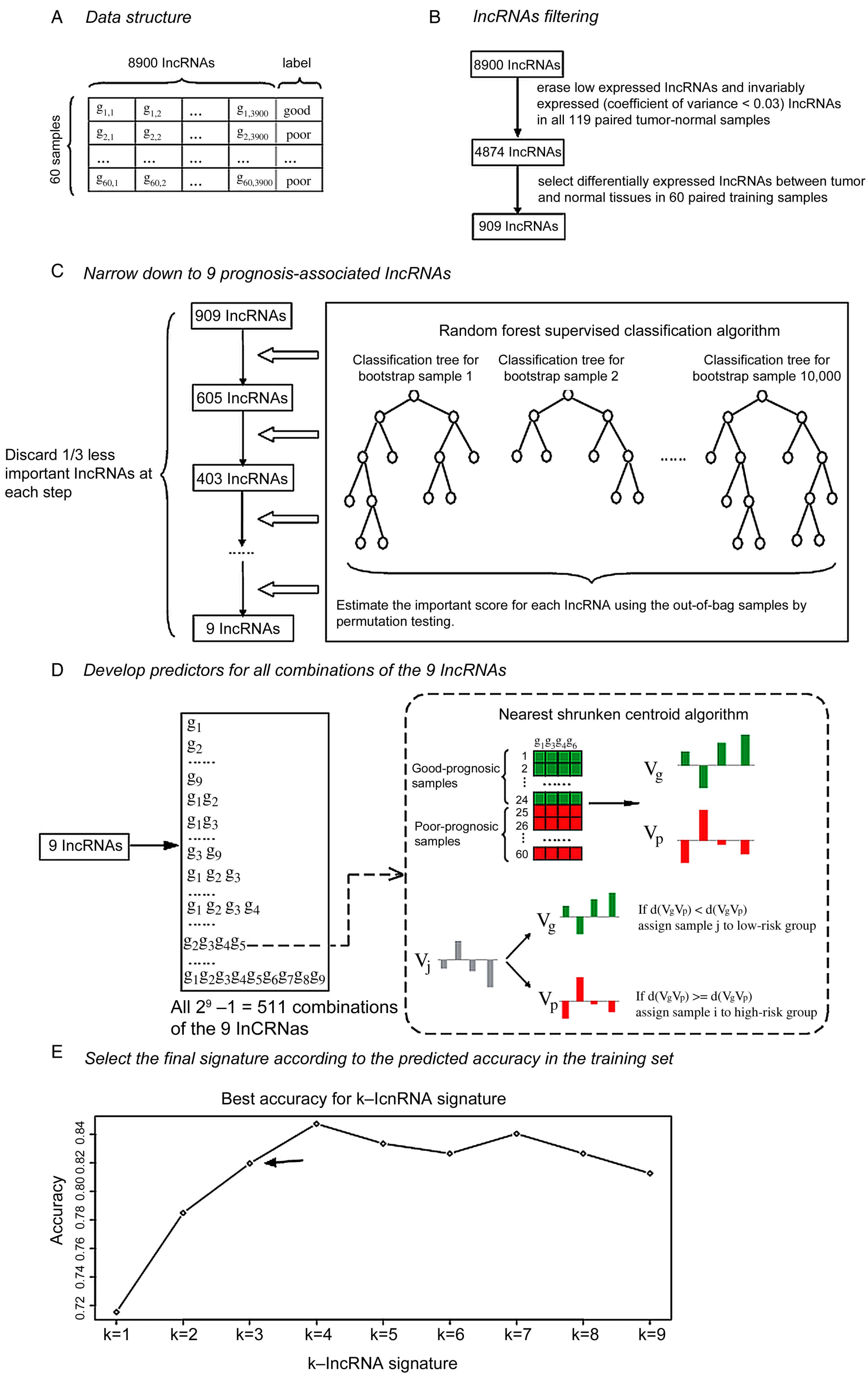
A Data structure
8900 lncRNAs
label
60 samples
$g_{1,1}$ $g_{1,2}$ … $g_{1,3900}$ good
$g_{2,1}$ $g_{2,2}$ … $g_{2,3900}$ poor
… … … … …
$g_{60,1}$ $g_{60,2}$ … $g_{60,3900}$ poor
B lncRNAs filtering
8900 lncRNAs
erase low expressed lncRNAs and invariably expressed (coefficient of variance < 0.03) lncRNAs in all 119 paired tumor-normal samples
4874 lncRNAs
select differentially expressed lncRNAs between tumor and normal tissues in 60 paired training samples
909 lncRNAs
C Narrow down to 9 prognosis-associated lncRNAs
Discard 1/3 less important lncRNAs at each step
909 lncRNAs
605 lncRNAs
403 lncRNAs
……
9 lncRNAs
Random forest supervised classification algorithm
Classification tree for bootstrap sample 1
Classification tree for bootstrap sample 2
Classification tree for bootstrap sample 10,000
……
Estimate the important score for each lncRNA using the out-of-bag samples by permutation testing.
D Develop predictors for all combinations of the 9 lncRNAs
9 lncRNAs
g_1
g_2
……
g_9
g_1g_2
g_1g_3
……
$g_3\ g_9$
$g_1\ g_2\ g_3$
……
$g_1\ g_2\ g_3\ g_4$
……
$g_2g_3g_4g_5$
……
$g_1g_2g_3g_4g_5g_6g_7g_8g_9$
All $2^9-1=511$ combinations of the 9 lnCRNas
Nearest shrunken centroid algorithm
$g_1g_3g_4g_6$
Good-prognosic samples
1
2
⋮
24
……
Poor-prognosic samples
25
26
⋮
60
……
V_g
V_p
V_j
V_g
If $d(V_gV_p) < d(V_gV_p)$ assign sample j to low-risk group
V_p
If $d(V_gV_p) >= d(V_gV_p)$ assign sample i to high-risk group
E Select the final signature according to the predicted accuracy in the training set
Best accuracy for k–lcnRNA signature
Accuracy
0.72
0.74
0.76
0.78
0.80
0.82
0.84
k=1
k=2
k=3
k=4
k=5
k=6
k=7
k=8
k=9
k–lncRNA signature

Figure 1 Identification of the long non-coding RNA (lncRNA) signature in the training set. (A) After microarray processing, the microarray data was described by an 60×8900 matrix with a 'good' or 'poor' label column. (B) After two filtering procedures, 909 lncRNAs remained for further analysis. (C) Selection process for the nine lncRNAs with highest classification power for patient survival. A random Forest supervised classification algorithm was used to narrow down the number of lncRNAs by several iterative steps, in which one-third of the least important lncRNAs were discarded at each step according to their importance score. (D) Development of prognostic classifier for all combinations ($N=2^9-1=511$) of the nine lncRNAs using the nearest shrunken centroid algorithm. V_g and V_p are the mean expression profiles of the lncRNA combination ($g_1\ g_3\ g_4\ g_6$) for good-prognostic samples and poor-prognostic samples, respectively. V_i is the expression profile of sample i. The Euclid distances $d(V_i,V_g)$ and $d(V_i,V_p)$ are used to classify sample i into a low- or high-risk group. (E) The procedure for identifying the final signature. The accuracies of all 511 signatures were calculated and the nine highest accuracies for k=1, 2, …, 9 are shown in the plot. The signature containing three lncRNAs was selected as the final signature.

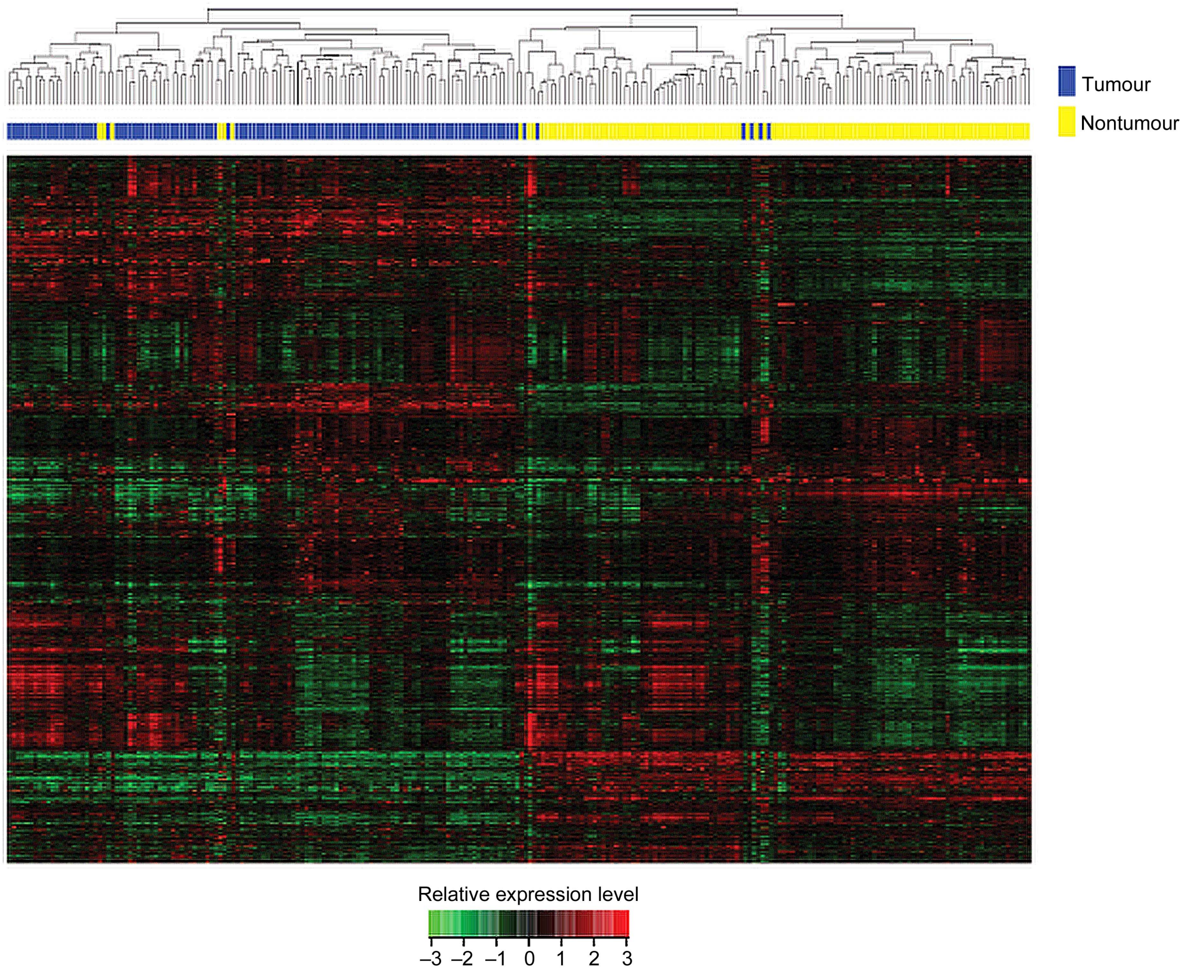

Figure 2 Unsupervised hierarchical clustering of the 119 pairs of tissues. The normalised expression data of the 6389 lncRNAs with coefficient of variance >0.10 was used for clustering analysis. Hierarchical clustering clearly separated tumour (blue bar) and normal (yellow bar) samples. Only six tumour samples and six normal samples were misclassified.

Cox regression analysis was performed using a stepwise variable selection method. Selected covariables included age, sex, tobacco use, alcohol use, tumour location, tumour grade, T stage, N stage, TNM stage, postoperative complications, adjuvant therapy and the lncRNA signature. Because adjuvant therapy information was missing for some of the patients, we used the multiple imputation method of Markov chain Monte Carlo to impute the missing value of adjuvant therapy in the Cox regression analysis (see details in online supplementary methods and supplementary table S3).[24–26] The results from the training set showed that the high-risk three-lncRNA signature (HR=8.486, 95% CI 3.550 to 20.284, p<0.0001), older age (HR=2.366, 95% CI 1.191 to 4.701, p=0.0140) and postoperative anastomotic leak (HR=5.805, 95% CI 1.605 to 21.000, p=0.0073) was significantly correlated with poor overall survival of the patients with OSCC (table 2). Combined test and independent datasets showed that the three-lncRNA signature (HR=2.203, 95% CI 1.330 to 3.649, p=0.0022), adjuvant therapy (HR=2.328, 95% CI 1.299 to 4.172, p=0.0045) and age (HR=1.674, 95% CI 1.033 to 2.713, p=0.0365) were independent prognostic factors for patients with OSCC (table 2). The results of the multivariable Cox regression analysis thus indicated that the predictive ability of the three-lncRNA signature is independent of other clinical and pathological factors for the survival of patients with OSCC.

The three-lncRNA signature has prognostic value within clinical stages

We next carried out a stratified analysis in TNM stage II and III patients to evaluate whether the three-lncRNA signature could predict survival of patients within the same clinical stage. Log-rank test of stage II patients in both the training group (p<0.0001, figure 4A) and the combination of test and independent cohort (p=0.0257, figure 4B) showed that the signature could classify stage II patients with OSCC into high- and low-risk groups. For patients with stage III OSCC, the three-lncRNA signature showed similar prognostic value in the training (p=0.0104, figure 4C) and the combined test and independent (p=0.0105, figure 4D) datasets. Because of limited sample size (n=10), the stratified analysis was not performed for stage I patients.

Survival prediction power: comparison of TNM stage and the three-lncRNA signature

To compare the sensitivity and specificity in survival prediction between TNM stage and the three-lncRNA signature, we performed receiver operating characteristic (ROC) analysis (see online supplementary methods).[20] We also constructed a prognostic model combining the two factors and compared the predictive ability. In the training set, predictive ability of both three-lncRNA signature and the combined model were significantly better than TNM stage alone (p=0.0268, p=0.0006,

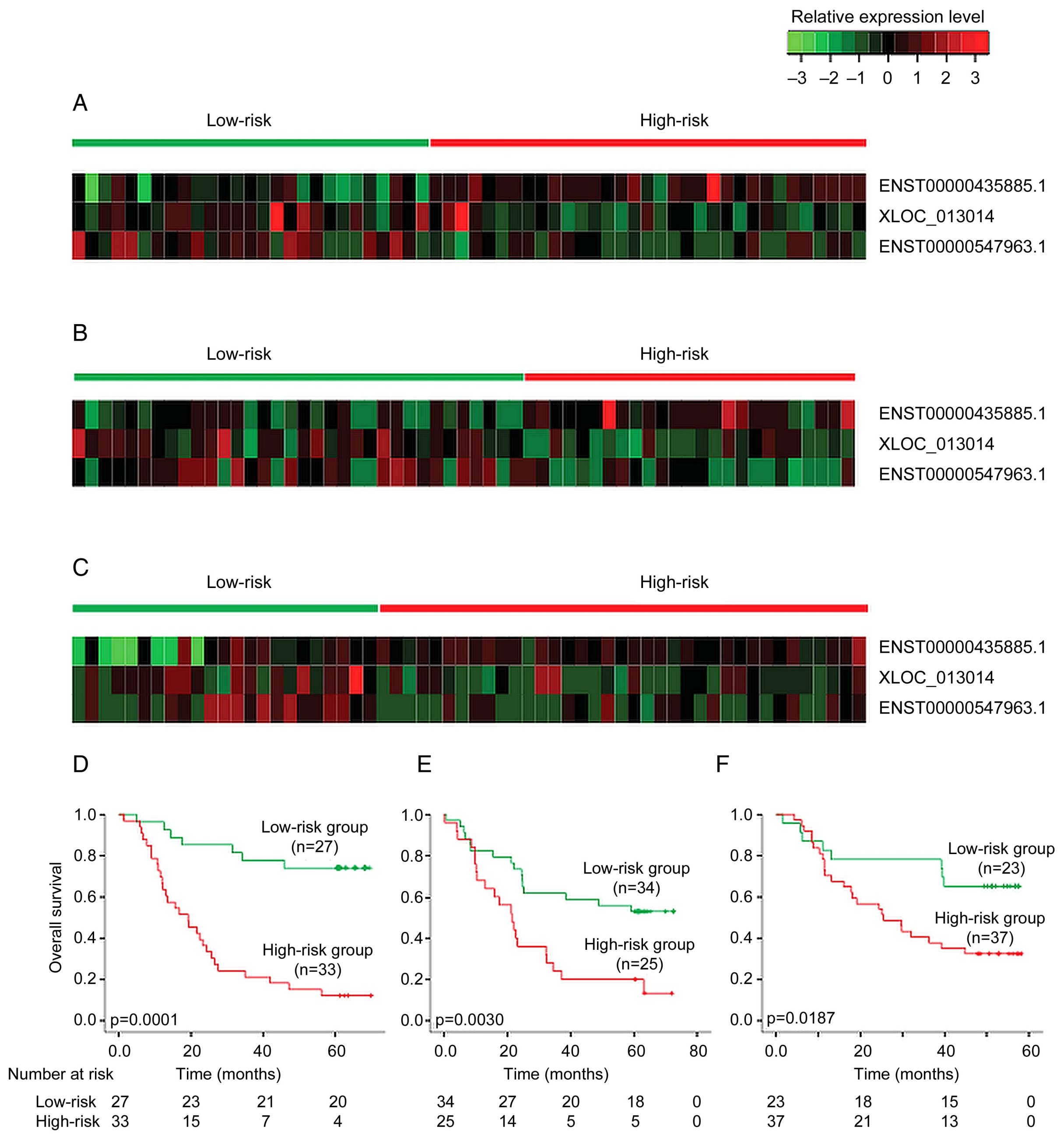

Figure 3 The three-lncRNA signature predicts overall survival of patients with OSCC. Heat maps (A–C) of the relative expression level (tumour minus normal) after z-score transformation for each lncRNA, and Kaplan–Meier survival curves (D–F) of patients classified into high- and low-risk groups using the three-lncRNA signature. p Values were calculated by log-rank test. (A, D) Training set, 60 patients. (B, E) Test set, 59 patients. (C, F) Independent cohort, 60 patients. OSCC, oesophageal squamous cell carcinoma.

respectively, figure 5A). In the test set, no significantly different predictive ability between the TNM stage and the signature was found. The combined model had a higher area under the ROC curve than the TNM stage (0.71 vs 0.63, figure 5B); however, the difference was not significant (p=0.1256), probably owing to limited sample size. ROC analysis was not performed for the independent cohort because the follow-up period of these patients was <5 years.

All three lncRNAs of the signature are essential for its prognostic value

To confirm that all of the three lncRNAs of the signature are required for its prognostic value, we constructed all possible 'signatures' containing from one to three lncRNAs (a total of seven signatures). The prognostic value of all signatures with fewer than three lncRNAs was evaluated by log-rank test in the training, test and independent datasets and compared with the original three lncRNA signature. The comparison showed that none of the signatures with fewer than three lncRNAs was consistently associated with patient survival in all three groups of patients (see online supplementary table S4). This indicates that all three lncRNAs are essential for the prognostic power of the signature.

Functional enrichment analysis of genes correlated with the signature lncRNAs

We next sought to explore the potential role of the lncRNAs of the prognostic signature in OSCC tumorigenesis and

Table 1 Clinical and pathological characteristics of patients with OSCC with high- or low-risk lncRNA signature in the three datasets

	Training set (n=60)			Test set (n=59)			Independent set (n=60)		
Characteristics	High-risk group (n=33)	Low-risk group (n=27)	p Value	High-risk group (n=25)	Low-risk group (n=34)	p Value	High-risk group (n=37)	Low-risk group (n=23)	p Value
Age, median (IQR)	59.0 (11.0)	55.0 (17.5)	0.7976*	62.0 (12.0)	59.0 (9.5)	0.3834*	62.0 (13.0)	58.0 (11.0)	0.8231
Gender, male	26 (78.8)	23 (85.2)	0.7391	21 (84.0)	28 (82.4)	1.0000	28 (75.7)	20 (87.0)	0.3404
Tobacco use, yes	19 (57.6)	20 (74.1)	0.1825	18 (72.0)	23 (67.6)	0.7197	18 (48.6)	16 (69.6)	0.1119
Alcohol use, yes	20 (60.6)	16 (59.3)	0.9156	16 (64.0)	22 (64.7)	0.9554	18 (48.6)	14 (60.9)	0.3562
Tumour location			0.2460			0.5411			0.3780
Upper	7 (21.2)	2 (7.4)		1 (4.0)	4 (11.8)		5 (13.5)	1 (4.3)	
Middle	15 (45.5)	17 (63.0)		17 (68.0)	20 (58.8)		18 (48.6)	10 (43.5)	
Lower	11 (33.3)	8 (29.6)		7 (28.0)	10 (29.4)		14 (37.8)	12 (52.2)	
Tumour grade			0.5977			0.3126			0.4270
Well differntiated	8 (24.2)	6 (22.2)		4 (16.0)	5 (14.7)		4 (10.8)	5 (21.7)	
Moderately differentiated	17 (51.5)	17 (63.0)		10 (40.0)	20 (58.8)		21 (56.8)	13 (56.5)	
Poorly differentiated	8 (24.2)	4 (14.8)		11 (44.4)	9 (26.5)		12 (32.4)	5 (21.7)	
T stage			0.2524			0.1632			0.2271
T1	1 (3.0)	2 (7.4)		1 (4.0)	4 (11.8)		1 (2.7)	3 (13.0)	
T2	3 (9.1)	2 (7.4)		4 (16.0)	11 (32.4)		5 (13.5)	2 (8.7)	
T3	17 (51.5)	19 (70.4)		15 (60.0)	11 (32.4)		31 (83.8)	17 (73.9)	
T4	12 (36.4)	4 (14.8)		5 (20.0)	8 (23.5)		0	1 (6.3)	
N stage			0.1350			0.0290			0.7255
N0	11 (33.3)	16 (59.3)		6 (24.0)	21 (61.8)		16 (43.2)	13 (56.5)	
N1	18 (54.5)	8 (29.6)		9 (36.0)	7 (20.6)		14 (37.8)	6 (26.1)	
N2	1 (3.0)	2 (7.4)		7 (28.0)	3 (8.8)		6 (16.2)	3 (13.0)	
N3	3 (9.1)	1 (3.7)		3 (12.0)	3 (8.8)		1 (2.7)	1 (4.3)	
TNM stage			0.1106			0.0378			0.5552
I	0	2 (7.4)		0	4 (11.8)		2 (5.4)	2 (8.7)	
II	10 (30.3)	12 (44.4)		8 (32.0)	17 (50.0)		17 (45.9)	13 (56.5)	
III	23 (69.7)	13 (48.1)		17 (68.0)	13 (38.2)		18 (48.6)	8 (34.8)	
Tumour clearance			N/A			N/A			N/A
R0	33 (100)	27 (100)		25 (100)	34 (100)		37 (100)	23 (100)	
R1/R2	0	0		0	0		0	0	
Postoperative complication									
Pneumonia	1 (3.0)	1 (3.7)	1.0000	6 (24.0)	4 (11.8)	0.2970	2 (5.4)	1 (4.3)	1.0000
Anastomotic leak	3 (9.1)	1 (3.7)	0.6199	3 (12.0)	4 (11.8)	1.0000	1 (2.7)	0	1.0000
Arrhythmia	11 (33.3)	5 (18.5)	0.2481	9 (36.0)	2 (5.9)	0.0055	10 (27.0)	6 (26.1)	1.0000
Adjuvant therapy			0.6209			0.2585			0.5196
Yes	20 (60.6)	13 (48.1)		16 (64.0)	20 (58.8)		23 (62.2)	12 (52.2)	
No	8 (24.2)	9 (33.3)		1 (4.0)	6 (17.6)		11 (29.7)	10 (43.5)	
Unknown	5 (15.2)	5 (18.5)		16 (27.1)	8 (35.5)		3 (8.1)	1 (4.3)	
Median survival (months)	19.2	>60	<0.0001†	21.5	>60	0.0030†	25.8	>48	0.0187†

Data are shown as n (%). p Values are calculated by χ^2 test or Fisher's exact test, unless otherwise stated.
*Student's t test.
†Log-rank test. N/A: p values are not calculated because all patients received R0 resection.
OSCC, oesophageal squamous cell carcinoma; TNM, tumour node metastasis.

development. For this purpose, we examined the correlation between their expression values and those of the mRNAs in the original group of 119 patients and summarised the genes correlated with the three lncRNAs. The expression level of 292 protein coding genes was positively correlated (Pearson correlation coefficient >0.60) with that of at least one of the three signature lncRNAs. The 292 genes clustered most significantly in ectoderm development and epithelial cell differentiation in gene ontology (GO) biological process enrichment analysis[27 28] (see online supplementary table S5). The same analysis of the 1572 genes negatively correlated with at least one of the three signature lncRNAs (Pearson correlation coefficient <−0.40) returned GO term cell cycle regulation and ubiquitin-protein ligase activity regulation (see online supplementary table S6). These results suggest that the lncRNAs of the signature may positively regulate genes which affect the development and differentiation of oesophageal epithelial cells and repress genes which affect cell cycle and ubiquitin-protein ligase activity.

DISCUSSION

In this study, we examined the lncRNA profiles of OSCC tissues and paired adjacent normal tissues and identified a

Table 2 Univariable and multivariable Cox regression analysis of the lncRNA signature and survival in the training set (n=60) and in the combined test and independent cohort (n=119)

		Univariable analysis		Multivariable analysis	
		HR (95% CI)	**p Value**	**HR (95% CI)**	**p Value**
Training set					
Age	>60/≤60	1.595 (0.821 to 3.098)	0.1680	2.366 (1.191 to 4.701)	0.0140
Gender	Female/male	1.233 (0.561 to 2.707)	0.6022		
Tobacco use	Y/N	0.693 (0.357 to 1.346)	0.2790		
Alcohol use	Y/N	0.896 (0.464 to 1.732)	0.7445		
Tumour location	Upper, middle/lower	1.249 (0.602 to 2.591)	0.5504		
Tumour grade	Moderately differentiated, poorly/well differentiated	1.569 (0.685 to 3.592)	0.2863		
T	T3, T4/T1, T2	0.767 (0.319 to 1.845)	0.5540		
N	N1, N2, N3/N0	1.960 (0.974 to 3.943)	0.0592		
TNM	III/I, II	2.506 (1.202 to 5.226)	0.0143		
Pneumonia	Y/N	1.050 (0.144 to 7.672)	0.9614		
Anastomotic leak	Y/N	2.716 (0.829 to 8.892)	0.0987	5.805 (1.605 to 21.000)	0.0073
Arrhythmia	Y/N	1.416 (0.706 to 2.837)	0.3271		
Adjuvant therapy	Y/N	1.501 (0.849 to 2.652)	0.1625		
LncRNA signature	High risk/low risk	6.578 (2.837 to 15.252)	<0.0001	8.486 (3.550 to 20.284)	<0.0001
Test+independent cohort					
Age	>60/≤60	1.724 (1.072 to 2.774)	0.0246	1.674 (1.033 to 2.713)	0.0365
Gender	Female/male	1.283 (0.714 to 2.306)	0.4045		
Tobacco use	Y/N	0.788 (0.488 to 1.272)	0.3295		
Alcohol use	Y/N	0.866 (0.539 to 1.390)	0.5501		
Tumour location	Upper, middle/lower	1.184 (0.719 to 1.951)	0.5065		
Tumour grade	moderately differentiated, poorly/well differentiated	0.982 (0.502 to 1.919)	0.9571		
T	T3, T4/T1, T2	1.237 (0.716 to 2.183)	0.4458		
N	N1, N2, N3/N0	2.214 (1.346 to 3.640)	0.0017		
TNM	III/I, II	2.031 (1.258 to 3.278)	0.0037		
Pneumonia	Y/N	1.507 (0.721 to 3.152)	0.2759		
Anastomotic leak	Y/N	0.942 (0.343 to 2.589)	0.9085		
Arrhythmia	Y/N	0.976 (0.558 to 1.705)	0.9311		
Adjuvant therapy	Y/N	2.227 (1.241 to 3.997)	0.0073	2.328 (1.299 to 4.172)	0.0045
LncRNA signature	High risk/low risk	2.412 (1.464 to 3.975)	0.0005	2.203 (1.330 to 3.649)	0.0022

TNM, tumour node metastasis.

three-lncRNA signature which was closely related to the prognosis of patients with OSCC. The prognostic value of this signature was verified in the test set of 59 patients and in an independent cohort of 60 patients.

In recent years, an increasing number of lncRNAs have been identified and associations between lncRNAs and various diseases have been reported.[29] The roles of lncRNAs in cancer development are increasingly being studied.[9 30 31] However, the involvement of lncRNAs in OSCC has not been reported. Here, we present the first report on differential lncRNA expression in a cohort of 119 patients with OSCC. Through an analysis of tumour and normal tissues, we found that many lncRNAs were differently expressed in OSCC tissues compared with adjacent normal tissues, indicating that lncRNAs may have critical roles in OSCC tumorigenesis.

Our finding of a three-lncRNA signature in OSCC suggests that lncRNAs can be powerful predictors for survival of patients with cancer. The correlation of lncRNA expression levels with the prognosis of patients with cancer has recently been reported for several malignancies, such as hepatocellular carcinoma,[13] breast cancer[9] and colorectal cancer.[30] In our study, the three-lncRNA signature identified in the training group showed similar prognostic value in both the test group and the independent cohort. Thus, we believe that the prognostic power of the signature has a solid basis in patients with OSCC. This is a pioneering study of the association between lncRNA expression and the survival of patients with cancer. Our findings are important because we show that lncRNA has a similar prognostic power to those of mRNA or miRNA for patients with cancer. Moreover, according to Du and colleagues in their recent report, the function of lncRNAs is more closely associated with their expression level compared with mRNAs as they do not encode proteins.[16]

For the statistical analysis of high-throughput biological data, the 'curse-of-dimensionality' problem (small sample size combined with a very large number of genes) is very common. In this work, we tried to reduce the effects of the 'curse-of-dimensionality' problem. At first, 909 lncRNAs differentially expressed between tumour and normal samples were filtered out and then subjected to random Forest supervised classification in order to further narrow down the number of lncRNAs associated with prognosis. The random sampling and ensemble strategies used in random Forest classification enable it to achieve accurate predictions while running efficiently on 'curse-of-dimensionality' datasets. In random Forest classification, the measures of gene importance are used to filter the original gene set iteratively, resulting in good performance in feature selection.

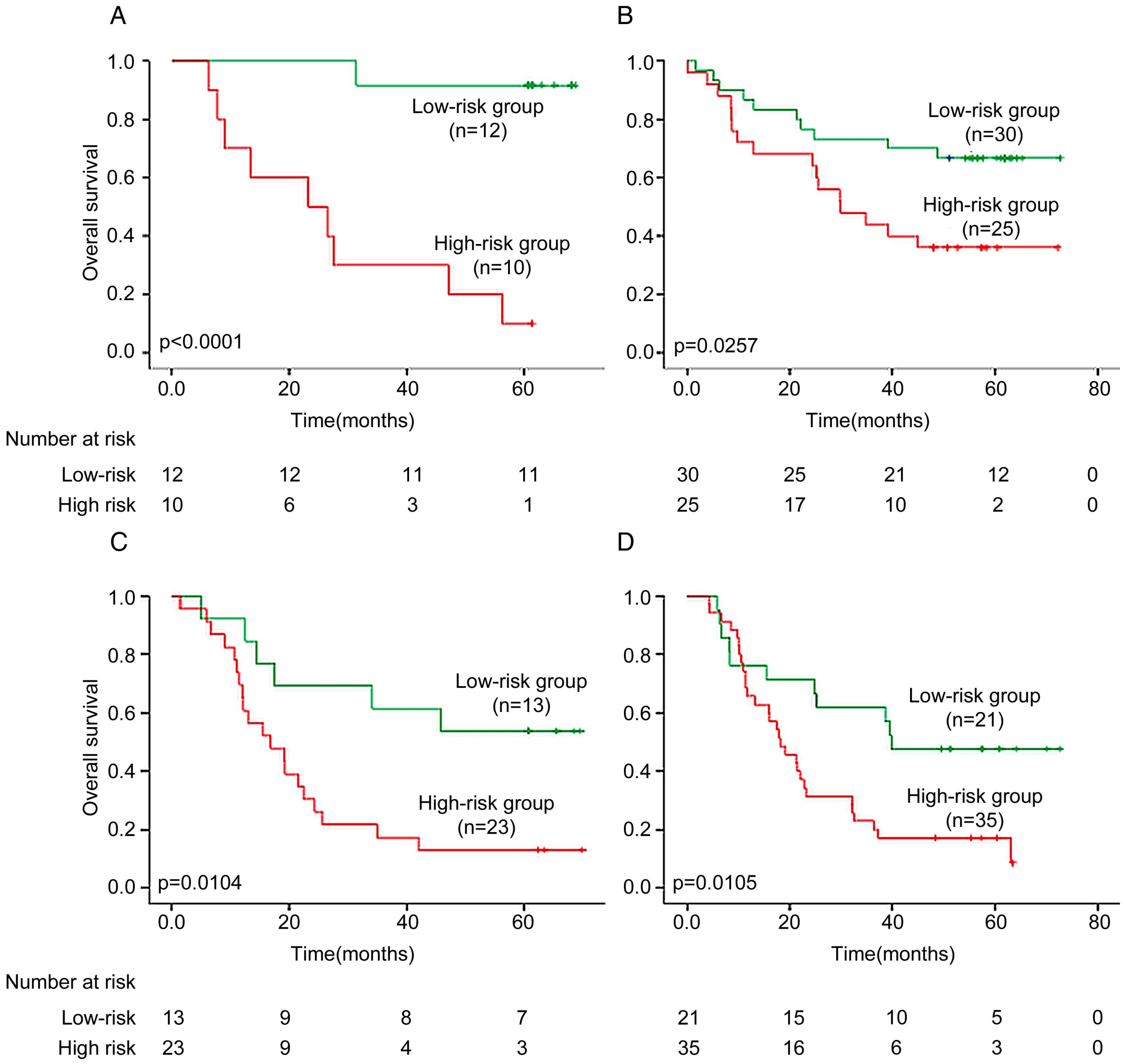

Figure 4 Survival prediction in stage II and III patients. Kaplan–Meier survival curves of stage II and III patients with OSCC classified into high- and low-risk groups based on the three-lncRNA signature. (A) Stage II patients, training set (n=22). (B) Stage II patients, combined test set and independent cohort (n=55). (C) Stage III patients, training set (n=36). (D) Stage III patients, combined test set and independent cohort (n=56). OSCC, oesophageal squamous cell carcinoma.

After the feature selection procedure, we constructed a classifier for each combination of the nine selected lncRNAs using the nearest shrunken centroid algorithm. In this study, we compared the performances of k-lncRNA signatures in the training set for all k=1,2,…,9 and the best accuracies for each k were listed. As shown in figure 1E, the accuracies were similar for k ≥ 3—between 81.3% and 84.7%. Although the signature with k=4 had the highest accuracy, we found that one lncRNA in the signature was redundant (see online supplementary results). Also the prognostic classification and performance of the four-lncRNA and three-lncRNA signatures were similar (see online supplementary results). Thus for the above reasons and the rule of Occam's razor, the signature with k=3 was selected as the final signature.

The current TNM staging system has critical limitations in predicting the survival of patients with OSCC. Thus molecular markers are needed to assist doctors in clinical practice. In the stratified analysis, the three-lncRNA signature showed prognostic value both in stage II and stage III patients. The three-lncRNA signature can classify patients of the same TNM stage into high- and low-risk groups with significantly different survival prospects, indicating that the signature can improve the accuracy of survival prediction. This finding might help doctors to select high-risk patients for adjuvant therapy in addition to traditional surgery, which can improve the outcome of OSCC.

In this study, we have analysed the prognostic value of the three-lncRNA signature. Whether this signature might be used to predict if adjuvant therapy would be of benefit for patients was not evaluated since accurate and complete information about adjuvant therapy after surgery was not available for some patients. Also, as the lncRNA signature was derived from patients who received R0 resection, whether it has prognostic value in suboptimal R1/R2 patients remains unknown. One limitation of our study is the generalisability of the three-lncRNA signature identified. Although this signature was generated and tested in the largest cohort of patients with OSCC by far and the patients enrolled were from different regions of China, datasets from other institutes and other countries are still necessary

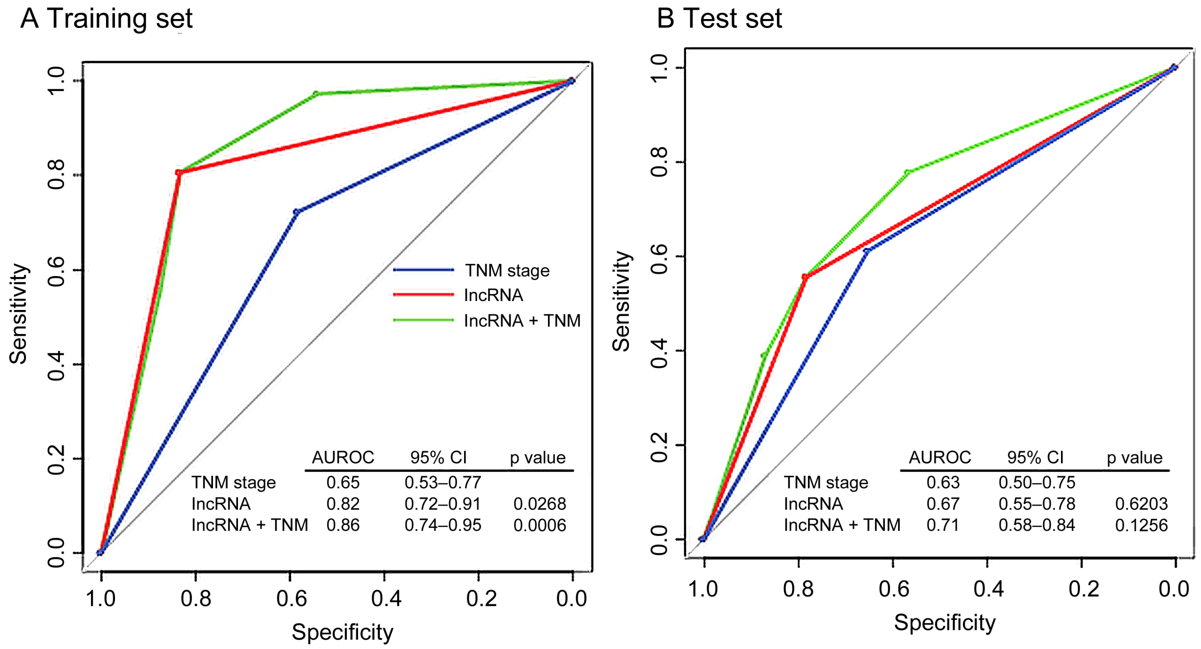

Figure 5 Comparison of sensitivity and specificity for survival prediction by the three-lncRNA signature, TNM stage and combination of the two factors. The three receiver operating characteristics (ROC) curves in the training set (A) and test set (B). p Values show the area under the ROC (AUROC) of TNM stage versus the AUROC of the three-lncRNA signature, or the combination of signature and TNM. TNM, tumour node metastasis.

to verify its generalisability. Its validity should be further tested in prospective cohorts.

Most lncRNAs are not yet functionally annotated. However, we can infer the possible function of the lncRNAs in OSCC using the mRNA expression data of the same group of patients. Genes whose expression value positively correlated with the three lncRNAs were enriched for the GO biological process term ectoderm development and epithelial cell differentiation, and the negatively correlated genes clustered in cell cycle regulation and ubiquitin-protein ligase activity regulation GO terms. Thus it is a plausible inference that the three lncRNAs associated with survival of patients with OSCC may be involved in the development, differentiation and cell cycle regulation of oesophageal epithelia cells and their deregulation may lead to OSCC tumorigenesis and progress. Some of the ectoderm development and differentiation related genes correlated with the signature lncRNAs have already been reported to have tumour suppressive functions. For instance, ANXA1 gene encodes the Ca^{2+}-dependent phospholipid-binding protein annexin I, which inhibits the cancer related NF-κB signal transduction pathway.[32] Another gene clustered into the same GO term, *PPL*, is also a well-studied gene involved in tumour formation and development. Its protein product periplakin is a component of desmosomes involved in cell–cell junction.[33 34]

In conclusion, our study has shown that the lncRNA expression profile is altered in OSCC tissues compared with normal oesophageal tissues. The three-lncRNA signature we discovered robustly predicts the survival of patients with OSCC. Furthermore, this signature can predict the survival of patients with OSCC within same TNM stages. To our knowledge, it is the first lncRNA signature identified that predicts survival in patients with cancer. Further validation studies in prospective cohorts and in cohorts from different institutions are needed to test the prognostic power of the signature before it is applied clinically. Whether the signature is useful for the prediction of the benefit of adjuvant therapy after surgical resection for patients with OSCC requires study with a sufficient number of patients with clear postoperative adjuvant therapy information.

Author affiliations
![1]Department of Thoracic Surgery, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, The People's Republic of China
[2]Bioinformatics Laboratory and Laboratory of Noncoding RNA, Institute of Biophysics, Chinese Academy of Sciences, Beijing, The People's Republic of China
[3]Department of Pathology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, The People's Republic of China

Contributors JH, RC, JL and ZC: contributed to the concept and design of the study. RC, JL and LT: contributed to interpretation of the data (statistical and computational analysis). JL, ZC and LT: contributed to the writing of the manuscript. JH, GS and ZC: contributed to the review and revision of the manuscript. JL, CZ, YZ, SW, FZ, JS and BZ: contributed to the RNA extraction and array hybridisation. CZ: contributed to the qRT-PCR. SS and XF: contributed to the pathological identification of the samples. MYH, YG, NS and ZL: contributed to the haematoxyloin and eosin staining of the samples. JD, RY, YY, XS and ML: contributed to the collection of clinical and pathological data of the patients. KS, NL, BQ, FT: contributed to the collection of samples. JH is the guarantor of the paper, who accepts full responsibility for the work and the conduct of the study. He has access to the data and controls the decision to publish.

Funding The study was supported by National High Technology Research and Development Program of China (2012AA02A502, 2012AA02A503, 2012AA02A207), International Science and Technology Corporation and Exchange Project (2010DFB30650), National Natural Science Foundation of China (81172336, 81101772, 81201856), Beijing Natural Science Foundation (7141011).

Competing interests None.

Ethics approval Medical ethics committee of the Cancer Institute and Hospital, Chinese Academy of Medical Science.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement The microarray original data, processed data and the clinical and pathoogical data of our study have been submitted to the Gene Expression Omnibus with accession number GSE53625.

REFERENCES

1 Jemal A, Bray F, Center MM, *et al*. Global cancer statistics. *CA Cancer J Clin* 2011;61:69–90.

2 Yang L, Parkin DM, Ferlay J, *et al*. Estimates of cancer incidence in China for 2000 and projections for 2005. *Cancer Epidemiol Biomarkers Prev* 2005;14:243–50.

3 Birney E, Stamatoyannopoulos JA, Dutta A, *et al*. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447:799–816.

4 Clark MB, Johnston RL, Inostroza-Ponta M, *et al*. Genome-wide analysis of long noncoding RNA stability. *Genome Res* 2012;22:885–98.

5 Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 2012;81:145–66.

6 Nagano T, Fraser P. No-nonsense functions for long noncoding RNAs. *Cell* 2011;145:178–81.

7 Yoon JH, Abdelmohsen K, Srikantan S, *et al*. LincRNA-p21 suppresses target mRNA translation. *Mol Cell* 2012;47:648–55.

8 Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature* 2012;482:339–46.

9 Gupta RA, Shah N, Wang KC, *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* 2010;464:1071–6.

10 Tsai MC, Spitale RC, Chang HY. Long intergenic noncoding RNAs: new links in cancer progression. *Cancer Res* 2011;71:3–7.

11 Troy A, Sharpless NE. Genetic "lnc"-age of noncoding RNAs to human disease. *J Clin Invest* 2012;122:3837–40.

12 Brunner AL, Beck AH, Edris B, *et al*. Transcriptional profiling of lncRNAs and novel transcribed regions across a diverse panel of archived human cancers. *Genome Biol* 2012;13:R75.

13 Yang F, Zhang L, Huo XS, *et al*. Long noncoding RNA high expression in hepatocellular carcinoma facilitates tumor growth through enhancer of zeste homolog 2 in humans. *Hepatology* 2011;54:1679–89.

14 Yu G, Yao W, Wang J, *et al*. LncRNAs expression signatures of renal clear cell carcinoma revealed by microarray. *PLoS One* 2012;7:e42377.

15 Han L, Zhang K, Shi Z, *et al*. LncRNA profile of glioblastoma reveals the potential role of lncRNAs in contributing to glioblastoma pathogenesis. *Int J Oncol* 2012;40:2004–12.

16 Du Z, Fei T, Verhaak RG, *et al*. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol* 2013;20:908–13.

17 Gerlinger M, Rowan AJ, Horswell S, *et al*. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 2012;366:883–92.

18 Navin N, Kendall J, Troge J, *et al*. Tumour evolution inferred by single-cell sequencing. *Nature* 2011;472:90–4.

19 Chen HY, Yu SL, Chen CH, *et al*. A five-gene signature and clinical outcome in non-small-cell lung cancer. *N Engl J Med* 2007;356:11–20.

20 Liu N, Chen NY, Cui RX, *et al*. Prognostic value of a microRNA signature in nasopharyngeal carcinoma: a microRNA expression analysis. *Lancet Oncol* 2012;13:633–41.

21 Cabili MN, Trapnell C, Goff L, *et al*. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011;25:1915–27.

22 Eisen MB, Spellman PT, Brown PO, *et al*. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;95:14863–8.

23 Schwarz DF, Konig IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* 2010;26:1752–8.

24 Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: issues and methods. *J Clin Oncol* 2012;30:3297–303.

25 Yuan YC. Multiple imputation for missing data: concepts and new developments. *Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference*, 2000:267.

26 Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. *Stat Med* 2007;26:3057–77.

27 Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;4:44–57.

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

29 Chen G, Wang Z, Wang D, *et al*. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* 2013;41:D983–6.

30 Kogo R, Shimamura T, Mimori K, *et al*. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res* 2011;71:6320–6.

31 Niinuma T, Suzuki H, Nojima M, *et al*. Upregulation of miR-196a and HOTAIR drive malignant character in gastrointestinal stromal tumors. *Cancer Res* 2012;72:1126–36.

32 Zhang Z, Huang L, Zhao W, *et al*. Annexin 1 induced by anti-inflammatory drugs binds to NF-kappaB and inhibits its activation: anticancer effects in vitro and in vivo. *Cancer Res* 2010;70:2379–88.

33 Ruhrberg C, Hajibagheri MA, Parry DA, *et al*. Periplakin, a novel component of cornified envelopes and desmosomes that belongs to the plakin family and forms complexes with envoplakin. *J Cell Biol* 1997;139:1835–49.

34 Straub BK, Boda J, Kuhn C, *et al*. A novel cell-cell junction system: the cortex adhaerens mosaic of lens fiber cells. *J Cell Sci* 2003;116:4985–95.