Supplementary Methods

Patients and samples

At the discovery stage, we retrospectively collected paired cancer and adjacent normal tissues from 119 ESCC patients with at least 5 years of follow up. The end of follow-up time for the patients was Dec. 2012 and the median follow-up time was 32.2 months. The clinical endpoints were defined as: died from ESCC before the time of follow-up or being alive at the time of follow-up (at least 5 years after surgery). The exclusion criteria of patients included: received radiotherapy or chemotherapy before surgery; died from reasons rather than ESCC; tumor or normal tissues too small for the assay (<50ug); RNA quality inadequate for microarray assay; tumor cell percentage of tumor tissue less than 60% or normal tissue contaminated by tumor tissues (Supplementary Figure 1). The tissue samples were snap frozen in liquid nitrogen shortly after resection and stored at -80℃ till RNA extraction. A small part from every sample was paraffin embedded, sectioned, and H&E stained using routine methods for pathological examination. The tumor histology was independently confirmed by two pathologists (F-XL and S-SS). The percentage of cancer cells was 60 or more in all tumor samples and no cancer cells were found in the normal tissues. To validate the prognostic signature, we enrolled an independent cohort of 60 ESCC patients who underwent surgery at the Cancer Institute and Hospital, CAMS from Jan. 2008 to Dec. 2008. The paired cancer and normal tissues of these patients were tested by the same microarray as the 119 patients. The inclusion and exclusion criterion of these patients was also same as the 119 patients (Supplementary Figure 1). For all samples, clinical and pathologic information (age, gender, pathology, differentiation, TNM stage, co-morbidities, post-operative complications and survival time after surgery) was available (Supplementary Table 1). The TNM stage was based on the American Joint Committee on Cancer staging manual (seventh edition). The surgical procedures the patients received were chosen according to the tumor location, mediastinal lymph nodes, and the general performance status of them. All patients enrolled in the study received R0 resection of the tumor with microscopically negative resection margins. The study was approved by the medical ethics committee of Cancer Institute and Hospital, CAMS.

RNA extraction

Total RNA was extracted using the Trizol reagent (Invitrogen) and purified with mirVana miRNA Isolation Kit (Ambion, Austin, TX, USA) according to the manufacturer's protocol. The purity and concentration of RNA were determined by OD260/280 using spectrophotometer (NanoDrop ND-1000). RNA integrity was determined by 1% formaldehyde denaturing gel electrophoresis.

RNA amplification and labeling.

cDNA labeled with a fluorescent dye (Cy5 and Cy3-dCTP) was produced by Eberwine's linear RNA amplification method and subsequent enzymatic reaction.[1] In detail, double-stranded cDNAs

(containing a T7 RNA polymerase promoter sequence) were synthesized from 1 ug total RNA using the CbcScript reverse transcriptase with cDNA synthesis system according to the manufacturer's protocol (Capitalbio) with the T7 Oligo (dT) and T7 Oligo (dN). After completion of double-stranded cDNA (dsDNA) synthesis using DNA polymerase and RNase H, the dsDNA products were purified using PCR NucleoSpin Extract II Kit (MN) and eluted with 30 ul elution buffer. The eluted double-stranded cDNA products were vacuum evaporated to 16 ul and subjected to 40 ul in vitro transcription reactions at 37℃℃for 14 hr using a T7 Enzyme Mix. The amplified cRNA was purified using the RNA Clean-up Kit (MN). Klenow enzyme labeling strategy was adopted after reverse transcription using CbcScript II reverse transcriptase. Briefly, 2 ug amplified RNA was mixed with 4 ug random nanomer, denatured at 65℃℃for 5min, and cooled on ice. Then, 5 ul of 4×first-strand buffer, 2 ul of 0·1M DTT, and 1·5 ul CbcScript II reverse transcriptase were added. The mixtures were incubated at 25℃℃for 10 min, then at 37℃℃for 90 min. The cDNA products were purified using a PCR NucleoSpin Extract II Kit (MN) and vacuum evaporated to 14 ul. The cDNA was mixed with 4 ug random nanomer, heated to 95℃℃for 3 min, and snap cooled on ice for 5 min. Then, 5 ul Klenow buffer, dNTP, and Cy5-dCTP or Cy3-dCTP (GE Healthcare) were added to final concentrations of 240 M dATP, 240 M dGTP, 240 M dTTP, 120 M dCTP, and 40 M Cy-dCTP. 1·2 ul Klenow enzyme was then added, and the reaction was performed at 37℃℃for 90 min. Labeled cDNA was purified with a PCR NucleoSpin Extract II Kit (MN) and resuspended in elution buffer. Labeled controls and test samples labeled with Cy5-dCTP and Cy3-dCTP were dissolved in 80 ul hybridization solution containing 3×SSC, 0·2% SDS, 5×Denhardt's solution and 25% formamide.

Array hybridization

DNA in hybridization solution was denatured at 95℃ for 3 min prior to loading onto microarray. Array hybridization was performed in a CapitalBio BioMixerTM II Hybridization Station overnight at a rotation speed of 8 rpm and a temperature of 42℃ and washed with two consecutive solutions (0·2% SDS, 2× SSC at 42℃ for 5 min, and 0·2× SSC for 5 min at room temperature).

Microarray fabrication

The Agilent human lncRNA+mRNA Array v2·0 was designed with four identical arrays per slide (4 x 180K format). Each array contained probes interrogating about 39,000 human lncRNAs and about 32,000 human mRNAs. Each RNA was detected by two probe repeats. The array also contained 4974 Agilent control probes.

Filtering procedure for lncRNAs in the microarray

The probes with same sequence were merged into one, thereafter, resulted in 35,025 unique probes. To obtain the map from the probe to annotated lncRNAs, the UCSC data base, GENCODE(V13) data base and lincRNAs from Cabili et al[2] were taken as the reference annotation (totally 13812 long intergenic and 6528 antisense non-coding RNAs). Then, we employed the blast program to map the probes uniquely to the annotated lncRNA sequences, and 8900 lncRNAs with at least one unique

probe were retrieved. For each of the 8900 lncRNAs, the median of the expression values of the probes mapped to it was used as its expression value. If part of the probes mapped to a lncRNA have missing values, the rest of the probes mapped to it was taken to calculate the median expression value. The expression value of a lncRNA was defined missing value when all the probes mapped to had missing value.

Missing value imputation using random forest unsupervised learning.

Random Jungle[3] was used to impute missing values by unsupervised learning. It began by filling a rough value of the missing data. Then, a forest including 10,000 trees ran and the proximities were computed. The missing values were estimated based on the proximities between the sample and non-missing value samples. The forest was constructed iteratively and the missing values were re-estimated iteratively. The number of iterations was set to 5.

Random forest supervised classification algorithm

In the random forest supervised classification algorithm, an iteration procedure was implemented to narrow down the gene set in which the 1/3 least important lncRNAs were discarded at each iteration step. Ten thousand trees were grown at each step, and the square root of the number of input lncRNAs at each step was set to the size of randomly chosen lncRNAs at each node of single classification tree. Because the number of good-prognostic and poor-prognostic patients were not equal, the class weights were adjusted accordingly. The generalization error was estimated on the out-of-bag samples. Finally, 9 lncRNAs were selected (Figure 1C).

Quantitative RT-PCR

In qRT-PCR, the reverse transcription (RT) reactions were carried out with Reverse Transcriptase (SuperScript III, Invitrogen) according to the manufacture's instruction. Around 3ug total RNAs were added to each reaction. Quantitative PCR reactions were then performed on ABI 7900 in a 10ul system. The reactions were incubated at 95℃ for 5 min, followed by 40 cycles of 95℃ for 15s, and 60℃ for 40s. All quantitative PCR reactions were performed in triplicate. The Ct value of each candidate lncRNA was then normalized to the expression value of GAPDH. Relative expression levels of the lncRNAs were calculated using $2^{-\Delta Ct}$. The sequences of primers used in qRT-PCR of the lncRNAs are listed in Supplementary Table 2.

Multiple imputation of Markov chain Monte Carlo (MCMC) for missing value of adjuvant in Cox regression analysis

For our data, the probability of missing adjuvant therapy information could be dependent on fully observed clinical factors like age, N stage and TNM stage, and is independent of the unobserved covariable (adjuvant therapy). Thus the missing at random mechanism should be suitable for our data.[4, 5] Here we used the Multiple Imputation procedure in SAS, which is popularly used for missing at random data.[4, 6, 7]

For the training set, adjuvant therapy information of ten out of 60 patients was missing, and we created ten imputations by multiple Markov chains. At first, we did univariable Cox regression analysis for adjuvant therapy. The results of seven out of ten imputations showed that adjuvant

therapy was not significantly associated with survival (p>0.05), which were consistent with the combining inference result (p=0.1694) (Supplementary Table 3). Then, we did multivariable Cox regression analysis. As that of the univariable analysis, the results of seven in ten imputations showed that adjuvant therapy was not an independent prognostic factor (p>0.05 by stepwise regression, the p-value cutoffs of entry and stay were both set to 0.1), and the combining inference result was similar (p=0.3694 by full model) (Supplementary Table 3).

For the combined test and independent cohort, adjuvant therapy information for 20 of the 119 patients was missing and we performed the same procedure as in the training set. In univariable Cox regression analysis, adjuvant therapy was significantly associated with survival in all of the ten imputations (p<0.05), and also in the combining inference result (p=0.0077) (Supplementary Table 3). In multivariable Cox regression analysis, the results of seven out of ten imputations showed that adjuvant therapy was an independent prognostic factor (p<0.05 by stepwise regression, the p-value cutoffs of entry and stay were both set to 0.1), and so did the combining inference result (p=0.0406 by full model) (Supplementary Table 3).

In Table 2, we reported the result of one from the ten imputations that has similar result with the combining inference of the ten imputations.

ROC analysis

The patients could be classified to “good” or “poor” prognostic groups according to the survival time being longer than 5 years or not. We compared the survival prediction abilities for training and test set among 3 factors: TNM stage (I-II vs III), the 3-lncRNA signature (low-risk vs high risk), and the combination of the two factors.

Next, we constructed prognostic score models for the two factors and the combined model by following the method of Liu N et al.[8] In the prognostic score models, the coefficients of low-risk in the signature and I-II stages in TNM were set to 1, and the coefficients of low-risk in the signature and I-II stages in TNM were set to the hazard ratio in univariable Cox regression. A cumulative risk score was calculated for each patient in training and test set and was used to perform receiver operating characteristic (ROC) analysis. In the comparison of area under the ROC (AUROC) among the 3 models, the bootstrap test was used with 10,000 trials.

References

1 Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W*, et al.* Performance comparison of one-color and two-color platforms within the MicroArray Quality Control (MAQC) project. Nat Biotechnol 2006;**24**:1140-50.

2 Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A*, et al.* Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011;**25**:1915-27.

3 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.

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

5 Ibrahim JG, Molenberghs G. Missing data methods in longitudinal studies: a review. Test (Madr) 2009;**18**:1-43.
6 Yuan YC. Multiple imputation for missing data: Concepts and new developments. Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, 2000:267.
7 Harel O, Zhou XH. Multiple imputation: review of theory, implementation and software. Stat Med 2007;**26**:3057-77.
8 Liu N, Chen NY, Cui RX, Li WF, Li Y, Wei RR*, et al.* Prognostic value of a microRNA signature in nasopharyngeal carcinoma: a microRNA expression analysis. Lancet Oncol 2012;**13**:633-41.