



# Characterization of super-enhancer-associated functional lncRNAs acting as ceRNAs in ESCC

Qiu-Yu Wang<sup>1,2</sup>, Liu Peng<sup>1</sup>, Yang Chen<sup>1,3</sup>, Lian-Di Liao<sup>1,3</sup>, Jia-Xin Chen<sup>2</sup>, Meng Li<sup>2</sup>, Yan-Yu Li<sup>2</sup>, Feng-Cui Qian<sup>2</sup>, Yue-Xin Zhang<sup>2</sup>, Fan Wang<sup>2</sup>, Chun-Quan Li<sup>2</sup> , De-Chen Lin<sup>4</sup>, Li-Yan Xu<sup>3</sup> and En-Min Li<sup>1</sup> 

1 The Key Laboratory of Molecular Biology for High Cancer Incidence Coastal Chaoshan Area, Shantou University Medical College, Shantou, China

2 School of Medical Informatics, Harbin Medical University, Daqing, China

3 Institute of Oncologic Pathology, Medical College of Shantou University, Shantou, China

4 Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA

## Keywords

cancer hallmark; CeRNA; long noncoding RNA; super-enhancer

## Correspondence

C.-Q. Li, School of Medical Informatics, Daqing Campus, Harbin Medical University, Daqing 163319, China  
Fax: 86-459-8153035  
Tel: +86-15004591078  
E-mail: lcqbio@163.com

D.-C. Lin, Department of Medicine, Cedars-Sinai Medical Center, Samuel Oschin Comprehensive Cancer Institute, 8700 Beverly Blvd, Los Angeles, CA 90048, USA  
Fax: +1-310-423-7182  
Tel: +1-310-423-7736  
E-mail: dchlin11@gmail.com

L.-Y. Xu, Institute of Oncologic Pathology, Medical College of Shantou University, Shantou 515041, China  
Fax: +86 754 88900847  
Tel: +86 754 88900460  
E-mail: lyxu@stu.edu.cn

E.-M. Li, The Key Laboratory of Molecular Biology for High Cancer Incidence Coastal Chaoshan Area, Shantou University Medical College, Shantou 515041, China  
Fax: +86 754 88900847  
Tel: +86 754 88900413  
E-mail: nmli@stu.edu.cn

Long noncoding RNAs (lncRNAs) have important regulatory roles in cancer biology. Although some lncRNAs have well-characterized functions, the vast majority of this class of molecules remains functionally uncharacterized. To systematically pinpoint functional lncRNAs, a computational approach was proposed for identification of lncRNA-mediated competing endogenous RNAs (ceRNAs) through combining global and local regulatory direction consistency of expression. Using esophageal squamous cell carcinoma (ESCC) as model, we further identified many known and novel functional lncRNAs acting as ceRNAs (ce-lncRNAs). We found that most of them significantly regulated the expression of cancer-related hallmark genes. These ce-lncRNAs were significantly regulated by enhancers, especially super-enhancers (SEs). Landscape analyses for lncRNAs further identified SE-associated functional ce-lncRNAs in ESCC, such as HOTAIR, XIST, SNHG5, and LINC00094. THZ1, a specific CDK7 inhibitor, can result in global transcriptional downregulation of SE-associated ce-lncRNAs. We further demonstrate that a SE-associated ce-lncRNA, LINC00094 can be activated by transcription factors TCF3 and KLF5 through binding to SE regions and promoted ESCC cancer cell growth. THZ1 downregulated expression of LINC00094 through inhibiting TCF3 and KLF5. Our data demonstrated the important roles of SE-associated ce-lncRNAs in ESCC oncogenesis and might serve as targets for ESCC diagnosis and therapy.

Qiu-Yu Wang and Liu Peng contributed equally to this work as first authors

## Abbreviations

Ce-lncRNAs, lncRNAs acting as ceRNAs; CeRNAs, competing endogenous RNAs; ChIP-seq, chromatin immunoprecipitation sequencing; ESCC, esophageal squamous cell carcinoma; GloceRNA, global and local regulatory direction consistency of expression of ceRNAs; lncRNAs, long noncoding RNAs; OS, overall survival; PCGs, protein-coding genes; qRT-PCR, quantitative real-time PCR; SEs, super-enhancers; siRNA, small interfering RNA; TEs, typical enhancers.

(Received 24 February 2020, revised 1 May 2020, accepted 20 May 2020, available online 20 June 2020)

doi:10.1002/1878-0261.12726

## 1. Introduction

Long noncoding RNAs (lncRNAs) participate in a wide range of biological and cellular processes through mechanisms including modulation of chromatin structure, scaffolding, mRNA stability, or other transcriptional and post-transcriptional processes (Flynn and Chang, 2014; Gupta *et al.*, 2010; Schmitt and Chang, 2016; Vance and Ponting, 2014). Although some lncRNAs have well-characterized biological functions, the vast majority of this class of molecules remains functionally uncharacterized (Batista and Chang, 2013; Du *et al.*, 2013; Hosono *et al.*, 2017; Li *et al.*, 2018; Prensner *et al.*, 2011, 2013; Zhang *et al.*, 2018a,d). Accumulating evidence predicts that a large number of lncRNAs may act as competing endogenous RNAs (ceRNAs) to sponge miRNAs, resulting in the derepression of miRNA targets (Conte *et al.*, 2017; Karreth and Pandolfi, 2013; Paci *et al.*, 2014; Salmena *et al.*, 2011; Tay *et al.*, 2014; Zhou *et al.*, 2016). The ceRNA mechanisms might be general acting in downstream regulation of lncRNAs (Paci *et al.*, 2014; Poliseno *et al.*, 2010). Thus, it is of great interest to uncover functional lncRNAs through characterizing lncRNAs acting as ceRNAs (ce-lncRNAs). Indeed, studies demonstrated that previously uncharacterized lncRNAs could be functionalized, partly through the identification of their ceRNA interactors, and presented a framework for the prediction and validation of ceRNA interactions (Cesana *et al.*, 2011; Conte *et al.*, 2017). Especially, Paci *et al.* proposed a novel and useful computational approach to identify lncRNAs to act as ceRNAs through calculating the difference between Pearson and partial correlation coefficients (Paci *et al.*, 2014). Based on the approach, they effectively explored miRNA decoy mechanism in gene regulatory circuitry using expression data from breast invasive carcinoma.

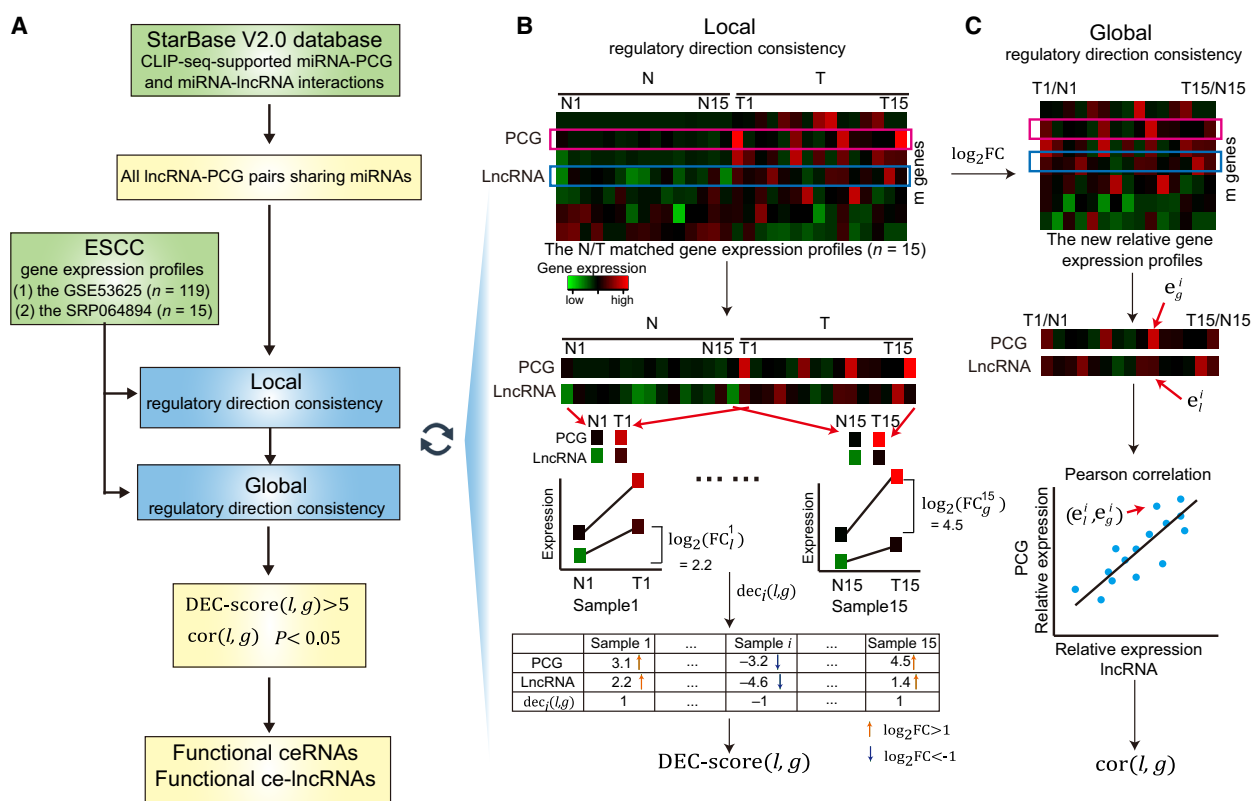
Enhancers are cis-acting DNA segments that control cell type-specific gene expression. Locally clustered enhancers form super-enhancers (SEs), which are enriched for binding of a large number of transcription factors and play prominent roles in control of gene expression program and cell identity (Amaral and Banister, 2014; Chipumuro *et al.*, 2014; Hnisz *et al.*, 2013; Whyte *et al.*, 2013). Importantly, SEs exhibit

much stronger lineage and tissue specificity compared with typical enhancers (TEs) (Hnisz *et al.*, 2013). Because SEs are frequently identified near protein-coding genes (PCGs) or noncoding RNAs that are important for controlling cell identity and differentiation, characterizing the function of SEs provides an opportunity to quickly identify key nodes driving diseases and biological processes (Hnisz *et al.*, 2013, 2015; Jiang *et al.*, 2019; Qian *et al.*, 2019; Tang *et al.*, 2019). Recently, some enhancer databases were developed, including SEdb (Jiang *et al.*, 2019), db-SUPER (Khan and Zhang, 2016), SEA (Wei *et al.*, 2016), and ENdb (Bai *et al.*, 2020). These databases provided a large number of SE/TE regions and related annotation information for various tissue/cell types. SE-associated upstream and downstream regulatory analysis can be further performed using the SEanalysis and KnockTF tool, which characterized SE-associated genes and transcription factors binding to target SEs (Feng *et al.*, 2020; Qian *et al.*, 2019). Studies have shown that a large number of novel noncoding RNAs are capable of being driven by SEs/TEs (Duan *et al.*, 2016; Hnisz *et al.*, 2013; Huang *et al.*, 2019; Jiang *et al.*, 2018b; Miao *et al.*, 2018; Peng *et al.*, 2019; Wood *et al.*, 2018; Xiang *et al.*, 2014; Xie *et al.*, 2018; Zhang *et al.*, 2017). Especially, a few SE-associated lncRNAs have well-characterized functions in cancer (Jiang *et al.*, 2018b; Peng *et al.*, 2019; Xie *et al.*, 2018), which reveals upstream regulatory mechanisms of lncRNAs. For example, SE-associated lncRNA LINC01503 was recently reported to promote the oncogenic phenotype of esophageal squamous cell carcinoma (ESCC) cells and was further identified as a squamous cell carcinoma-specific lncRNA (Xie *et al.*, 2018). SE-associated lncRNA HCCL5 activated by transcription factor ZEB1 can promote the malignancy of hepatocellular carcinoma (Peng *et al.*, 2019). Co-activation of SE-driven lncRNA CCAT1 by TP63 and SOX2 promotes squamous cancer progression (Jiang *et al.*, 2018b). However, whether and how functional lncRNAs are regulated by SE-associated genes is incompletely understood, due to the technical challenges in systematic characterization of SEs and functional lncRNAs. Since ce-lncRNAs have high expression level, they might be controlled SEs/TEs, which perform important functions through regulating ce-lncRNAs to

driver a large of downstream target genes. Ce-lncRNAs might appear to be a potential oncogenic downstream effector of SEs.

Here, we developed a two-stage computational approach, termed GloceRNA, for the identification of functional ce-lncRNAs through combining global and local regulatory direction consistency of expression of ceRNAs (Fig. 1). We used normal/tumor (N/T) matched samples to improve prediction of functional ceRNAs. Especially, GloceRNA can measure the differential expression consistency of the lncRNA-PCG pair at single sample level, which can effectively evaluate possibility of ceRNAs significantly appearing in some local samples. Using ESCC as a model, GloceRNA identified many known and novel functional

ce-lncRNAs. We demonstrated that GloceRNA robustly predicted ce-lncRNAs in multiple ESCC datasets, and the predicted ce-lncRNAs strongly regulated the expression of a large number of cancer hallmark genes. Moreover, we experimentally validated that some new predicted ce-lncRNAs were highly associated with ESCC, including LINC00094, LINC00338, SNHG10, and MF12-AS1. Furthermore, we found that ce-lncRNAs were significantly regulated by enhancers, especially SEs. We further demonstrated that a novel SE-driven ce-lncRNA – LINC00094 – promoted the growth and survival of ESCC cells. Lastly, we showed that TCF3 and KLF5 cooperatively regulated the express of LINC00094 through activation of its SE and promoter. Our study improved the original



**Fig. 1.** Schematic overview of the GloceRNA method. (A) Flow diagram of GloceRNA. The lncRNA-PCG pairs sharing miRNA target sites are first established using CLIP-seq-supported miRNA-PCG and miRNA-lncRNA interactions. Next, GloceRNA calculates the local and global regulatory direction consistency of each lncRNA-PCG pair. Finally, GloceRNA tests whether each lncRNA-PCG pair meets the local and global direction consistency criteria. A lncRNA-PCG pair sharing miRNAs will be identified as a functional ceRNA if it meets the two direction consistency criteria. The related lncRNA will be identified as a functional ce-lncRNA. (B) Schematic overview of local regulatory direction consistency of expression of ceRNAs. (C) Schematic overview of global regulatory direction consistency of expression of ceRNAs. N, normal; T, tumor.  $DEC\text{-}score(l, g)$ : local regulatory direction consistency score of the lncRNA-PCG pair, which can effectively evaluate possibility of ceRNAs significantly appearing in samples.  $dec_i(l, g)$ : the expression consistency score of the lncRNA-PCG pair in the sample  $i$ , which represents the regulatory direction consistency of expression at single sample level.  $cor(l, g)$ : global regulatory direction consistency score of the lncRNA-PCG pair, which is calculated using Pearson correlation coefficient of lncRNA-PCG.  $P$ :  $P$  value of Pearson correlation coefficient.  $\log_2(FC_l^i)$  and  $\log_2(FC_g^i)$ :  $\log_2 FC$  value of gene expression of lncRNA  $l$  and PCG  $g$  in sample  $i$ , which represent the relative gene expression level of tumor minus normal.  $e_l^i = \log_2(FC_l^i)$  and  $e_g^i = \log_2(FC_g^i)$ .

ceRNA identification methods, by using local regulatory direction consistency of expression strategy in N/T matched samples and emphasizing identification and analysis of functional ce-lncRNAs in ESCC.

## 2. Materials and methods

### 2.1. Genome-wide gene expression profiles of ESCC

Four datasets for genome-wide gene expression profiles of ESCC were used in the study, including: (a) the GSE53625 ( $n = 119$ ) dataset; (b) the SRP064894 dataset ( $n = 15$ ); (c) the TCGA ESCC dataset ( $n = 80$ ); (d) the GSE53625 ( $n = 60$ ) dataset. The clinical and pathological characteristics of patients in all datasets were provided in Table S1 and Appendix S1. The GSE53625 dataset included two independent experimental subdatasets for gene expression profiles: GSE53625 ( $n = 119$ ) and GSE53625 ( $n = 60$ ). The GSE53625 ( $n = 119$ ) dataset contained the 119 N/T matched samples. The GSE53625 ( $n = 60$ ) dataset contained the 60 N/T matched samples. These data were downloaded from Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53625>). The expression profiles were performed using the agilent human lncRNA+mRNA Array v2.0 (4\*180k) (Li *et al.*, 2014a). To obtain maps from probes to annotated lncRNAs, we employed the blast program to map probes uniquely to the annotated lncRNA sequences. GENCODE (V19) and Ensembl 75 database were used as the reference annotation, and 8900 lncRNAs with at least unique probes mapped to it was used as its expression value. The SRP064894 dataset, which was generated by us, included the 15 N/T matched samples (Li *et al.*, 2017). RNA sequencing was performed using the Illumina HiSeq 2500 (Illumina, San Diego, CA, USA). Sequencing reads were mapped to the human genome assembly (NCBI Build 37) using TOPHAT (v2.0.6). The expression profiles with the lncRNAs and PCGs were extracted by using EASYRNASEQ (1.6.0). The TCGA ESCC dataset included the ESCC samples of 80 patients (Cancer Genome Atlas Research *et al.*, 2017).

The GSE53625 ( $n = 119$ ) and SRP064894 datasets were used as identifying functional lncRNA-mediated ceRNAs and evaluating the robust of results. Further, the TCGA ESCC dataset was used as independent data to test the expression correlation of functional lncRNA-mediated ceRNA pairs predicted by GloceRNA. Using the dataset, we also compared the expression correlation between functional lncRNA-

mediated ceRNA pairs and other potential ceRNA pairs sharing miRNAs. The GSE53625 ( $n = 119$ ) and GSE53625 ( $n = 60$ ) collected the survival information of patients. Therefore, they were used as survival analysis of functional ce-lncRNAs and ceRNA pairs. Although the TCGA ESCC dataset also included the survival time of patients. However, the survival analysis were not performed in the study because survival time was too short for most of patients (the average survival time (day) = 193; 63% patients with survival time < 50 days, see Appendix S1).

### 2.2. CLIP-seq-supported miRNA-mRNA interactions

Cross-linking and Argonaute (Ago) immunoprecipitation coupled with high-throughput sequencing (CLIP-seq) could identify the genome-wide interaction of miRNAs and their targets (37). The starBase V2.0 database is designed for decoding interaction network via integrating large-scale CLIP-seq (HITS-CLIP, PAR-CLIP, iCLIP, CLASH) data (Li *et al.*, 2014b). MiRNA targets of starBase V2.0 were predicted by five target predicted algorithms, including TargetScan, miRanda, Pictar, PITA, and RNA22. In this study, we downloaded CLIP-seq-supported miRNA-lncRNA and miRNA-PCG interactions from starBase V2.0 database. In total, we obtained 423 975 miRNA-PCG interactions with 386 miRNAs and 13 802 PCGs and 10 212 miRNA-lncRNA interactions with 277 miRNAs and 1127 lncRNAs. All lncRNAs and PCGs, which can be assigned to HGNC symbol names, were used to the following ceRNA identification. The lncRNA-PCG pairs sharing at least one miRNA were computed through considering CLIP-seq-supported miRNA-PCG and miRNA-lncRNA interactions from starBase V2.0 database. These pairs were used as identification of functional lncRNA-mediated ceRNAs.

### 2.3. Identification of functional lncRNA-mediated ceRNAs

We developed a computational approach, called GloceRNA, which aims to identify functional lncRNA-mediated ceRNAs through combining global and local regulatory direction consistency of expression about ceRNAs (Fig. 1A). Notably, we first computed all lncRNA-PCG pairs sharing miRNAs from CLIP-seq-supported miRNA-PCG and miRNA-lncRNA interactions from starBase V2.0 database. These pairs were used as the following identification of functional ceRNAs. Next, we calculated the local and global regulatory direction consistency of each lncRNA-PCG pair.



Finally, GloceRNA tested whether each lncRNA-PCG pair meets the local and global direction consistency criteria. A lncRNA-PCG pair sharing miRNAs will be identified as a functional ceRNA if it meets the criteria. The related lncRNA will be identified as a functional ce-lncRNA.

We used N/T matched samples to evaluate local regulatory direction consistency of a potential lncRNA-PCG ceRNA pair (Fig. 1B). We found that gene expression profiles with N/T matched samples are available for ESCC and many other diseases. Based on ceRNA principle, the increase of lncRNA expression in the lncRNA-PCG ceRNA pair tends to lead to increase of the PCG expression, which means that the expression direction of ceRNA pair tends to be consistent. In N/T matched samples or even a single N/T matched sample, the expression direction of ceRNA pair also tends to be consistent. That is, for a pair of N/T samples from the same patient, a ceRNA pair usually displays consistently upregulated (or downregulated) in expression direction. Therefore, we used N/T matched samples to improve prediction of functional ceRNAs through capturing the local regulatory direction consistency information of expression. Suppose we have an ESCC expression profile dataset with  $n$  N/T matched samples and  $m$  genes (lncRNAs and PCGs) (Fig. 1B, Top panel). For a lncRNA-PCG pair with lncRNA  $l$  and PCG  $g$ , the local regulatory direction consistency, called DEC score( $l, g$ ), can be measured using the expression level of lncRNA  $l$  and PCG  $g$ . We first compute the  $\log_2 FC$  value of gene expression for the lncRNA  $l$  and PCG  $g$  in a N/T matched sample  $i$  (Fig. 1B, Middle panel) as follows:

$$\log_2(FC_l^i) = \log_2(y_l^i) - \log_2(x_l^i), \quad (1)$$

$$\log_2(FC_g^i) = \log_2(y_g^i) - \log_2(x_g^i), \quad (2)$$

where  $y_l^i$  is the tumor expression value of lncRNA  $l$  in the N/T matched sample  $i$ , and  $x_l^i$  is the normal expression value of the lncRNA in the N/T matched sample  $i$ . Similarly,  $y_g^i$  and  $x_g^i$  are the tumor and normal expression values of PCG  $g$  in the N/T matched sample  $i$ . The  $\log_2(FC_l^i)$  and  $\log_2(FC_g^i)$  values represent the relative gene expression level of tumor minus normal. Next, we used the  $\log_2 FC$  values of the lncRNA  $l$  and PCG  $g$  to compute the differential expression consistency score  $dec_i(l, g)$  of the lncRNA-PCG pair at single sample level (Fig. 1B, Bottom panel). When two  $\log_2 FC$  values of lncRNA and PCG are larger than 1 (i.e.,  $\log_2(FC_l^i) > 1$  and  $\log_2(FC_g^i) > 1$ ), the lncRNA-PCG pair will be defined as consistently upregulated

(+1) in differential expression direction. On the contrary, the pair is defined consistently downregulated (−1) if all two values  $< -1$ . Therefore, the regulatory direction consistency of expression at single sample level was calculated as follows:

$$dec_i(l, g) = \begin{cases} 1, & \text{if } \log_2(FC_l^i) > 1 \text{ and } \log_2(FC_g^i) > 1 \\ -1, & \text{if } \log_2(FC_l^i) < -1 \text{ and } \log_2(FC_g^i) < -1 \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where  $dec_i(l, g)$  is the expression consistency score of the lncRNA-PCG pair in the sample  $i$ . The 1 and −1 represent that the lncRNA-PCG pair is consistently upregulated or downregulated. For example, when  $\log_2(FC_l^i) = 3.1$  and  $\log_2(FC_g^i) = 2.2$ , the value of  $dec_i(l, g)$  is 1, which means that the lncRNA-PCG pair is consistently upregulated (see Fig. 1B bottom panel and Table S2 for more examples). Finally, for a lncRNA-PCG pair, we computed local regulatory direction consistency, called DEC score( $l, g$ ), through counting sum of all consistently up/downregulated samples across all  $n$  samples as follows:

$$DEC - score(l, g) = \sum_{i=1}^n |dec_i(l, g)|, \quad (4)$$

DEC score( $l, g$ ) represents sample number that meets the differential expression direction consistency at single sample level (see Table S2 for an example of calculating DEC score( $l, g$ )). DEC score( $l, g$ ) can be used to effectively evaluate possibility of ceRNAs significantly appearing in some local samples. The higher value of DEC score( $l, g$ ) is, the more samples meet that when a lncRNA is upregulated (or downregulated) in single ESCC sample, the corresponding PCG is also upregulated (or downregulated) in the same sample. Compared with global measures, the DEC score focused on mining ceRNA signals in the local samples since some strong ceRNA relationships may only exist in some patients due to cancer heterogeneity. In order to stably capture the local feature rather than global information, the cutoff of DEC score needs to be set appropriately. If the cutoff is set too small (e.g.,  $< 3$ ), we think that the result of 'local' regulatory direction consistency may be not stable due to random probability of regulatory direction consistency. On the contrary, too large cutoff (e.g., greater than half of the total number of samples) may lead to too strict, which makes DEC score tend to capture global rather than local information. In order to stably capture the local feature in two datasets, we think that the candidate cutoffs can be set as  $> 3, 4, 5, 6, 7$ , or  $8$ , which may be more

appropriate. We tested these cutoffs in two ESCC datasets (Table S3). In order to stably capture the local feature, and keep balance between local feature, number and similarity of ceRNAs in two datasets, the cutoff was set as  $> 5$  in the paper. When DEC score( $l, g$ )  $> 5$ , the pair is considered as meeting local regulatory direction consistency of ceRNAs.

For each lncRNA-PCG pair sharing miRNAs, global regulatory direction consistency was further computed based on the relative gene expression profiles (Fig. 1C, Top panel). Notably, we first converted the gene expression profiles ( $m \times 2n$  matrix) into a new gene expression profiles with relative expression level ( $m \times n$  matrix) (Fig. 1C, Top panel). The  $\log_2FC$  values were used to represent the relative gene expression level of tumor minus normal in the new gene expression profiles. For example, the expression value of the lncRNA  $l$  and PCG  $g$  in the sample  $i$  of the new gene expression dataset is  $\log_2(FC_l^i)$  and  $\log_2(FC_g^i)$ . Next, we used the Pearson correlation coefficient to evaluate the global regulatory direction consistency, called  $\text{cor}(l, g)$ , of the lncRNA-PCG pair based on relative expression values ( $\log_2FC$ ) across all samples of the dataset (Fig. 1C, Middle panel).

$$\text{cor}(l, g) = \frac{\sum_{i=1}^n (e_l^i - \bar{e}_l) (e_g^i - \bar{e}_g)}{\sqrt{\sum_{i=1}^n (e_l^i - \bar{e}_l)^2} \sqrt{\sum_{i=1}^n (e_g^i - \bar{e}_g)^2}}, \quad (5)$$

where  $e_l^i = \log_2(FC_l^i)$  and  $e_g^i = \log_2(FC_g^i)$ . The  $e_l^i$  and  $e_g^i$  are the relative expression levels of lncRNA  $l$  and PCG  $g$  in sample  $i$ . The  $\bar{e}_l$  and  $\bar{e}_g$  are the average value of the relative expression levels of lncRNA  $l$  and PCG  $g$  across all samples. The  $\text{cor}(l, g)$  can be used to effectively evaluate possibility of ceRNAs through measuring expression correlation of the lncRNA-PCG pair across all samples. The statistical significance of  $\text{cor}(l, g)$ , termed  $P$ , was calculated using the significance  $P$  value of the Pearson correlation coefficient. The Pearson correlation coefficient was adopted by many ceRNA studies and have been proved to be effective for identification of ceRNAs (Paci *et al.*, 2014; Wang *et al.*, 2015; Xu *et al.*, 2015). These existing studies used the absolute expression level of genes, whereas the relative expression levels of genes ( $\log_2FC$ ) were considered by previous studies to be able to reduce the influence of heterogeneity among different ESCC patients (Li *et al.*, 2014a). Therefore, instead of the absolute expression level, we computed Pearson correlation coefficient by using the relative expression level.

Finally, a lncRNA-PCG pair sharing miRNAs will be defined as functional ceRNA relationship in ESCC

if it meets the following criteria: (a) DEC score( $l, g$ )  $> 5$ ; (b)  $\text{cor}(l, g) > 0$  and  $P < 0.05$ . The above method was applied to all CLIP-seq-supported lncRNA-PCG pairs sharing miRNAs in starBase V2.0 database, and all functional ceRNAs meeting the criteria were identified. The lncRNAs identified in functional lncRNA-mediated ceRNAs were defined as functional ce-lncRNAs.

#### 2.4. The traditional ceRNA identification methods

Traditionally, a lncRNA-PCG pair sharing miRNAs will be defined as functional ceRNA relationship based on the following criteria: (a) Expression correlation of lncRNA-PCG pair; (b) Shared miRNAs; and (c) Differentially expression level of lncRNAs/PCGs. Although most of studies identify ceRNAs based on the three criteria, different combinations of them exist. Therefore, we used six different combinations for fair comparison with our method, including SAM(0.01)+Cor, Limma(0.01)+Cor, SAM(0.05)+Cor, SAM(0.01)+Hyper+Cor, Limma(0.01)+Hype+Cor, and SAM(0.05)+Hype+Cor. Notably, Pearson correlation coefficient (Cor) between a lncRNA-PCG pair is usually used to identify whether lncRNA-PCG pair is co-expressed. All lncRNA-PCG pairs with  $\text{Cor} > 0$  and  $\text{FDR} < 0.05$  were identified as candidate ceRNA pairs. The differentially expressed genes are identified using the SAM or Limma method with  $\text{FDR} < 0.01$  or  $0.05$ . A hypergeometric test is used to compute significance of shared miRNAs for each possible lncRNA-PCG pair. All  $P$  values were subject to FDR correction. For example, Limma(0.01)+Hype+Cor represents that ceRNAs meet significance of expression correlation (+Cor) and share miRNAs (+Hyper) between lncRNA-PCG pairs, with differentially expression of lncRNAs/PCGs based on Limma  $\text{FDR} < 0.01$  (+Limma(0.01)). Limma(0.01)+Cor represents that ceRNAs meet significance of expression correlation between lncRNA-PCG pairs, with Limma  $\text{FDR} < 0.01$ , but not use the ‘Shared miRNAs’ criterion with only needing to share at least one miRNA.

#### 2.5. Degree and betweenness centrality

The most elementary characteristic of a node is its degree, which represents how many links the node has to other nodes (Barabasi and Oltvai, 2004). Betweenness centrality is a measure of a node’s centrality in a network and is equal to the number of shortest paths from each node to all others that pass through this node. It reflects the amount of control that a node

exerts over the interactions of other nodes in the network.

## 2.6. Analysis of lncRNA-related cancer hallmarks

Hanahan and Weinberg (2011) have proposed that cancer cells acquire a number of hallmark biological capabilities during the multistep process of tumor pathogenesis (Hanahan and Weinberg, 2011). We used these hallmarks for analysis of lncRNA-related cancer hallmarks, including ‘Activating Invasion’, ‘Disrupting Cellular Energetics’, ‘Angiogenesis’, ‘Enabling Replicative Immortality’, ‘Genome Instability’, ‘Resisting Cell Death’, ‘Sustaining proliferative signaling’, ‘Tumor-Promoting Inflammation’, ‘Evading Growth Suppressors’, and ‘Avoiding Immune Destruction’. To obtain cancer hallmark genes, we firstly corresponded cancer hallmark to the Gene Ontology (GO) terms according to the study of Hnisz *et al.* (2013). Secondly, the genes annotated to these GO terms were downloaded from the databases MsigDB V6.1 (Subramanian *et al.*, 2005) and bioMart (Ensembl v91). Thirdly, for each GO term, the union of their related genes obtained from the two databases was used as the annotated genes of the GO term. The result showed that all cancer hallmarks can correspond to 31 GO terms with the annotated genes. Finally, these GO terms were used as proxies for the characteristic hallmark capabilities that are thought to be acquired in cancers.

To test whether ce-lncRNAs can control broad cancer-related hallmarks, we investigated ce-lncRNAs in the context of cancer hallmarks. On the one hand, we mapped all ce-lncRNA-related PCGs identified by GloceRNA from the GSE53625 ( $n = 119$ ) and SRP064894 datasets to cancer hallmarks and used hypergeometric test to calculate the enrichment significance of each cancer hallmark GO terms. On the other hand, we explored hallmark functions associated with each ce-lncRNA. Notably, for each ce-lncRNA, we used the ce-lncRNA-related PCGs from two datasets to identify the enriched hallmark GO terms. The enrichment significance was calculated using hypergeometric test.

## 2.7. Survival analysis

A clear understanding of the alterations in lncRNA expression occurring in cancers will require larger-scale studies. The GSE53625 ( $n = 119$ ) and GSE53625 ( $n = 60$ ) datasets were used as survival analysis of functional ce-lncRNAs and ceRNA pairs. The clinical and survival information of patients in the two datasets was provided in Table S1 and Appendix S1. For a

lncRNA (or PCG), the relationship between lncRNA (or PCG) expression and prognosis of ESCC patients was explored by Kaplan–Meier analysis (Li *et al.*, 2019). The mean value of gene expression was used as cutoff to classify patients into high- and low-risk groups. The statistical significance was assessed using the log-rank test by calculating the  $P$  values. For a ceRNA pair, an average expression of the corresponding lncRNA and PCG was calculated for each patient. Then, we used the average expression level of the ceRNA pair as the ‘pair expression’ to evaluate the association between survival and the ceRNA pair. Similarly, the mean value of ‘pair expression’ was used as cutoff to classify patients into high- and low-risk groups. The statistical significance was assessed using the log-rank test by calculating the  $P$  values. The lncRNA, PCG, and the ceRNA pair with  $P < 0.05$  were defined as significant. We used the same ‘mean value’ strategy as the cutoff to classify patients into high and low-risk groups in the GSE53625 ( $n = 119$ ) and GSE53625 ( $n = 60$ ) datasets. All analyses were performed on the R 2.13.2 framework.

## 2.8. Chromatin immunoprecipitation sequencing data analysis

Chromatin immunoprecipitation sequencing (ChIP-seq) files have been obtained from our previous studies with GEO database (GEO ID: GSE76861 and GSE106563) (Jiang *et al.*, 2017, 2018b). H3K27ac ChIP-seq was sequenced in six ESCC cell lines, including KYSE140, TT, KYSE510, KYSE70, TE5, and TE7. H3K27ac ChIP-seq reads were mapped using BOWTIE ALIGNER (v0.12.9) to hg19 human reference genome (Langmead *et al.*, 2009). MACS (model-based analysis of ChIP-seq) (v1.4.2) was used to identify enhancer enrichment regions (Zhang *et al.*, 2008). The corresponding wiggle files were generated using read pileups and were normalized using reads per million (rpm) by dividing tag counts by the total number of reads. We converted wiggle files into bigwig files using WIGTOBIGWIG tool (<http://hgdownload.cse.ucsc.edu/admin/exe/>) and visualized them using INTEGRATIVE GENOMICS VIEWER (<http://www.broadinstitute.org/igv/home>). ROSE software was used to identify potential SE regions as ‘python ROSE main.py -g hg19 -i \*\*\*\*\*.gff -r \*\*\*\*\* cas.sort.bam -c \*\*\*\*\* input.sort.bam -o \*\*\*\*\* -s 12500 -t 2000’ (Hnisz *et al.*, 2013). Briefly, H3K27ac peaks that occurred within  $\pm 1$  kb of transcription start sites were subtracted. ROSE stitched enhancers within 12.5 kb together. It separated SEs from TEs through ranking H3K27ac signal of them. Finally, a threshold was defined according to

the geometric inflection point to distinguish between TE and SE. Both SEs and TEs were assigned to the overlap, proximal, and closest genes to the center of the stitched enhancer. If lncRNAs appeared in the overlap, proximal, or closest genes of SEs or TEs, they were considered as SE/TE-associated lncRNAs. If SE/TE-associated lncRNAs belong to ce-lncRNAs in ESCC, we considered them as SE/TE-associated ce-lncRNAs in ESCC.

## 2.9. Identification of transcription factors binding to SEs of ce-lncRNAs

Identification of transcription factors that were predicted to bind to SEs of lncRNAs was based on motif scanning in SE regions associated with ce-lncRNAs. More than 3000 DNA binding motifs for 695 transcription factors are compiled from the TRANSFAC database (Matys *et al.*, 2006) and MEME suite (Bailey *et al.*, 2009), based on the following collections: JASPAR CORE 2014 vertebrates (Mathelier *et al.*, 2014), Jolma2013 (Jolma *et al.*, 2013), Homeodomains (Berger *et al.*, 2008), UniPROBE (Robasky and Bulyk, 2011), and Wei2010 (Wei *et al.*, 2010). For each of six ESCC cell line, we obtained the genomic regions of the constituents of SEs associated with ce-lncRNAs. According to these regions, we extracted their corresponding sequence from hg19 human reference genome using the *getfasta* function of BEDTOOLS (v2.25.0) (Quinlan and Hall, 2010) and followed by motif scanning with FIMO (Find Individual Motif Occurrences) at a *P* value threshold of  $10^{-4}$  (Grant *et al.*, 2011). Transcription factors having at least two significant DNA binding sequence motif instances in the SEs of each ce-lncRNA were identified. For each of identified transcription factor, we computed unique lncRNAs regulated by it through merging relationships between transcription factors and SE-associated ce-lncRNAs for all six ESCC cell lines. All transcription factors were finally ranked according to number of lncRNAs significantly regulated by them.

## 2.10. Gene expression profile for the effects of THZ1 inhibition for lncRNAs and related PCGs

Gene expression profiles for the effects of THZ1 inhibition were performed in our groups. The data can be downloaded from NCBI GEO database (GSE number: GSE76860). The detailed experimental descriptions were provided in our previous published paper (Jiang *et al.*, 2017). Briefly, whole-transcriptome RNA sequencing was performed before/after THZ1 treatment in TE7 and KYSE510 cells using illumina HiSeq

2000. The RNA-seq results were involved in gene expression level of either THZ1 or DMSO at indicated time points at 2, 4, 6, and 8 h, which were computed using FPKM through mapping reads to human reference genome. We filtered genes according to FPKM, and those active genes with FPKM > 1 were considered in following analyses.

## 2.11. Construction of THZ1-sensitive ceRNA networks

Firstly, we used gene expression profiles for the effects of THZ1 inhibition to compute fold changes of the expression level for SE/TE-associated ce-lncRNAs. If the expression level of SE/TE-associated ce-lncRNAs decreased over 1.5-fold at 12 h compared with DMSO, we defined them as ‘THZ1-sensitive SE/TE-ce-lncRNAs’. A total of 42 unique THZ1-sensitive SE/TE-ce-lncRNAs were identified in TE7 and KYSE510 cells. Secondly, we obtained the 26 shared THZ1-sensitive SE/TE-ce-lncRNAs in both cell lines. Based on these ce-lncRNAs, we extracted the first neighbor nodes in ESCC ceRNA network, and thus, the related PCGs associated with THZ1-sensitive SE/TE-ce-lncRNAs were obtained. Finally, a subnetwork of ESCC ceRNA network, called THZ1-sensitive ceRNA networks, was constructed through extracting the subgraph using THZ1-sensitive SE/TE-ce-lncRNAs and their related PCGs. The nodes in the subnetwork are THZ1-sensitive SE/TE-ce-lncRNAs or their related PCGs, and edges are the ceRNA relationships between them.

## 2.12. Cell culture and RNA interference

Cell lines used in this study and related cell culture information has been described previously (Long *et al.*, 2018). The KYSE150, KYSE510, and TE3 human esophageal squamous carcinoma cell lines were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (HYCLONE, Logan, UT, USA). ESCC cell line KYSE450 was cultured in Dulbecco’s modification of Eagle’s medium Dulbecco (DMEM) medium (Thermo Fisher Scientific, Waltham, MA, USA). All media were supplemented with 10% FBS (Thermo Fisher Scientific), penicillin-G ( $100 \text{ units}\cdot\text{mL}^{-1}$ ), and streptomycin ( $100 \mu\text{g}\cdot\text{mL}^{-1}$ ). Cells were incubated at  $37^\circ\text{C}$  in a humidified atmosphere containing 5%  $\text{CO}_2$ .

In functional assays, KYSE150, KYSE450, and TE3 cells were seeded into 6-well plates or 12-well plates and cultured for 12–24 h until 70–80% confluence. ESCC cells were transfected with 25 or 50 nm small interfering RNA (siRNA) using DharmaFECT™ Transfection Reagents (Dharmacon, Waltham, MA, USA) or



Lipofectamine 3000 (Invitrogen, Carlsbad, CA, USA) according to the manufacturer's instructions. The LINC00094, LINC00338, SNHG10, MFI2-AS1, and a negative control (NC) siRNAs were synthesized by Dharmacon. The TCF3 and KLF5 siRNAs were synthesized by GenePharma (Suzhou, China). The siRNA target sequence for lncRNAs and two transcription factors' mRNAs is described in Table S4.

### 2.13. RNA extraction and qRT-PCR

Total RNA from ESCC cells were extracted using TRIzol (Invitrogen) according to the manufacturer's protocol. The purity and concentration of RNA were determined by OD<sub>260/280</sub> using a NanoDrop ND-2000 spectrophotometer (Agilent, Santa Clara, CA, USA), and 1 µg of total RNA was reverse transcribed into cDNA using PrimeScript RT reagent Kit with gDNA Eraser (TaKaRa, Otsu, Japan) in accordance with the manufacturer's instructions. Quantitative real-time PCR (qRT-PCR) was performed by SYBR Premix Ex Taq (TaKaRa) using a 7500 Real-Time PCR System (Applied Biosystems, Waltham, MA, USA). Primers for quantitative real-time PCR are shown in Table S5. β-Actin was measured as an internal control and used for normalization. RNA expression was normalized against the relative value from the NC control group. qRT-PCR was performed in triplicate and repeated at least three times.

### 2.14. Wound healing assay

KYSE150, KYSE450, and TE3 cells were transfected with siRNAs targeting lncRNAs, and then, cells were starved in serum-free medium for 12 h after being transfected for 36 h. Circles 3 mm in diameter were marked on the bottom of each dish to identify the areas for image capture and ensure that measurements were taken at the same locations. A wound was made by scraping the cell monolayer with a 200-µL pipette tip. ESCC cells were maintained in RPMI-1640 medium or DMEM medium with 2.5% FBS. Images were captured at 0 and 36 h using a Leica DMI3000B inverted phase-contrast microscope (Leica Microsystems GmbH, Wetzlar, Germany). The wound closure rate was calculated from six images, using IMAGEJ (National Institutes of Health, Bethesda, MD, USA) analysis. Each experiment was performed in triplicate.

### 2.15. Transwell assay

Transwell assay was performed as described previously (Zhang *et al.*, 2018c). KYSE150, KYSE450, and TE3

cells were starved in serum-free medium for 12 h after being transfected. A total of  $5 \times 10^4$  cells were plated in medium without serum in the upper well of a transwell chamber of a 24-well transwell with 8-µm pores (BD Biosciences, San Jose, CA, USA), placed in a bottom chamber containing medium supplemented with 10% FBS. After 48 h, the membranes were fixed ice-cold methanol and stained with hematoxylin solution, and migration was quantified by counting 10 random fields under a Leica DMI3000B inverted phase-contrast microscope (400×). The migration cell numbers were counted with IMAGEJ. Each experiment was performed in triplicate.

### 2.16. Colony formation assay

Colony formation assay was performed as described previously (Zeng *et al.*, 2017). Briefly, transfected cells were trypsinized and counted with a cell counter (Bio-Rad, Hercules, CA, USA). Then, cells were plated at a density of 1000 cells per well in 6-well plates and incubated for 14 days at 37 °C with 5% CO<sub>2</sub>. After washing with 4 °C precooled PBS twice, cultures were fixed with ice-cold methanol for 15 min and stained with hematoxylin for 30 min. Colonies were photographed by ChemiDoc Touch (Bio-Rad). Each experiment was performed in triplicate.

### 2.17. Western blotting

ESCC cells were lysed with Laemmli sample buffer (Bio-Rad), heated for 10 min at 95°C. Western blotting was performed using SDS/PAGE. Proteins were transferred to PVDF membrane (Millipore, Billerica, MA, USA), which were then blocked for 1 h with 5% skim milk in TBST (20 mM Tris, 137 mM NaCl, 0.1% Tween-20). Membranes were incubated with primary antibody [1 : 1000 anti-KLF5 (Santa Cruz Biotechnology, Delaware Ave, Santa Cruz, CA, USA; sc-398470) and anti-TCF3 (Cell Signaling Technology, Danvers, MA, USA; Cat#4865) and anti-β-actin (Santa Cruz Biotechnology; sc-47778)] overnight in 4 °C. After three washes with TBST, membranes were incubated with secondary HRP-conjugated antibody [1 : 5000 (mouse; Santa Cruz Biotechnology; sc-516102) and 1 : 2000 (rabbit; Cell Signaling Technology; cat# 31,460)] for 1 h at room temperature. Signals were detected with ChemiDoc Touch (Bio-Rad).

### 2.18. Chromatin immunoprecipitation

Chromatin immunoprecipitation analysis was performed as described previously (Jiang *et al.*, 2018b). In

brief, ESCC cells KYSE150, KYSE510, and TE3 were treated with THZ1 (100 nM, 12 h), and then,  $1 \times 10^7$  cells were cross-linked with 1% formaldehyde solution (Thermo Fisher Scientific) and neutralized by 1.25 M glycine. Cross-linked cells were lysed and sonicated (Covaris E220, Woburn, MA, USA) to release 100–200 bp fragments. Anti-KLF5 (Santa Cruz Biotechnology; sc-398470x), anti-TCF3 (Santa Cruz Biotechnology; sc-166411x), or normal IgG was added to each sonicated chromatin and incubated at 4 °C overnight. Then, these complexes were conjugated to Dynabeads protein A/G magnetic beads (Invitrogen) for 4 h at 4 °C. After incubation, DNA was eluted from immunoprecipitate complexes, reverse cross-linked, and purified with QIAquick PCR purification kit (QIAGEN, Hilden, Germany).

The purified DNA was analyzed by real-time PCR with the use of LINC00094 super-enhancer-specific primers. Primers for ChIP-PCR were shown in Table S5. Relative enrichment was normalized to input. IgG antibody was used as a negative control.

### 2.19. Statistical analysis

Results are analyzed by SPSS software, 13.0 (SPSS, Chicago, IL, USA) or R 3.1.2 for windows. Where indicated, statistical analysis was performed by calculating means and SD. Graphs about biological experiments were mainly made by GRAPHPAD PRISM 6 (GraphPad, San Diego, CA, USA). Differences between groups were evaluated with the Student's *t*-test.  $P < 0.05$  was considered to be statistically significant. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ . Graphs about bioinformatics were mainly made by R 3.1.2.

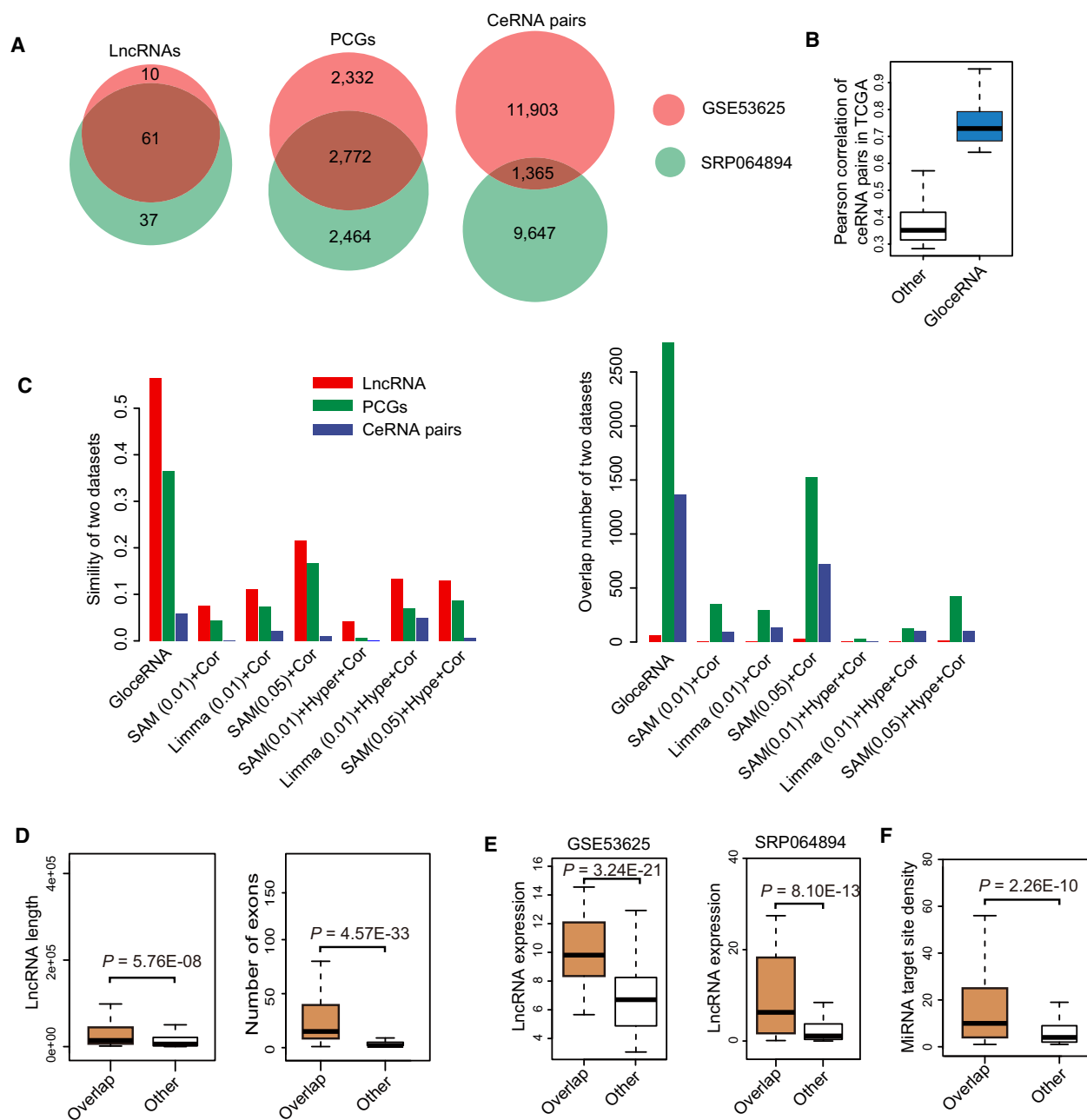
## 3. Result

### 3.1. Genome-wide identification of ce-lncRNAs using GloceRNA

To systematically identify functional ce-lncRNAs, we developed a two-stage identification method, termed GloceRNA, which integrated miRNA target sequences and gene expression profile information of lncRNAs and PCGs in large-scale N/T matched samples (see Materials and methods). Our hypothesis is that functional ceRNAs display expression direction consistency in local matched samples and at global gene expression level cross all samples. Briefly, lncRNA-PCG pairs sharing miRNA target sites were first established through using CLIP-seq-supported miRNA-PCG and miRNA-lncRNA interactions (Fig. 1A). Next, each

lncRNA-PCG pair sharing miRNAs was tested using two measures DEC score( $l, g$ ) and cor( $l, g$ ) and identified as a functional ceRNA relationship if it meets the following criteria: (a) local regulatory direction consistency of expression at single sample level (DEC score( $l, g$ ) > 5) (Fig. 1A,B); (b) global regulatory direction consistency of expression (cor( $l, g$ ) > 0 and  $P < 0.05$ ) (Fig. 1A,C). Finally, the related lncRNAs in functional ceRNAs were identified as functional ce-lncRNAs. The GloceRNA method has two advantages. On the one hand, using a new measure DEC score( $l, g$ ), local differential expression consistency between lncRNAs and PCGs can be effectively considered through computing regulatory direction consistency of expression at single N/T matched sample level, which can effectively evaluate possibility of ceRNAs appearing in parts of samples. On the other hand, global expression consistency of a lncRNA-PCG pair is tested through applying Pearson correlation coefficient to all samples, which can effectively evaluate possibility of ceRNAs through measuring relative expression correlation of the lncRNA-PCG pair cross all samples. Therefore, our method not only considered regulatory direction consistency of expression at the global level but also mined hidden regulatory direction information of ceRNAs from single and local some N/T matched samples.

Since we have previously characterized several important lncRNAs in ESCC, for which we also generated RNA-seq data from patient samples (Jiang *et al.*, 2018b; Li *et al.*, 2017; Xie *et al.*, 2018; Zhang *et al.*, 2018c), we next applied GloceRNA to this cancer type. Using internal dataset (15 paired tumor and normal samples), 13 268 ceRNA pairs were identified, involving 98 lncRNAs and 5236 PCGs. To evaluate the robustness of this result, we analyzed another large-scale transcriptomic dataset (119 paired tumor and normal samples). Strikingly, in this independent cohort, 61 out of 71 lncRNAs (85.91%) were significantly shared with our internal dataset ( $P = 0$ , hypergeometric test, Fig. 2A). The result showed that the ceRNA networks derived from different datasets shared similar lncRNAs. Moreover, we found that the overlaps of PCGs ( $P = 0$ ) and ceRNA pairs ( $P = 0$ ) between the two cohorts were also highly significant statistically (Fig. 2A), highlighting the consistency and robustness of our method. We further found that similarities of nodes and edges were obviously different although the overlaps were highly statistically significant. The overlaps between lncRNAs, as well as PCGs, were much larger than those between ceRNA pairs (Fig. 2A). This suggests that the ceRNA networks derived from different ESCC datasets might more tend



**Fig. 2.** Identification and analysis of functional ce-lncRNAs in ESCC. (A) Venn diagram showing the overlap of lncRNAs (left), PCGs (middle), and ceRNA pairs (right) between both ESCC datasets (GSE53625 and SRP064894). (B) Box plots of expression correlation of ceRNA pairs in TCGA ESCC samples. The bars represent expression of ceRNA pairs (blue) and all background pairs (white). (C) Comparison of results between our method and other methods. Left panel shows overlap similarity of lncRNAs (red), PCGs (green) and ceRNA pairs (blue). Right panel shows overlap number of lncRNAs (red), PCGs (green), and ceRNA pairs (blue). Traditionally, a lncRNA-PCG pair sharing miRNAs will be defined as functional ceRNA relationship based on the following criteria: (a) Expression correlation of lncRNA-PCG pair (Cor); (b) Shared miRNAs (Hyper); and (c) Differentially expression level of lncRNAs/PCGs (SAM or Limma). We used six different combinations of them for fair comparison with our method, including SAM(0.01)+Cor, Limma(0.01)+Cor, SAM(0.05)+Cor, SAM(0.01)+Hyper+Cor, Limma(0.01)+Hyper+Cor, and SAM(0.05)+Hyper+Cor. Box plots of ce-lncRNAs are displayed according to (D) Length (left) and number (right) of exons, (E) expression level, and (F) number of miRNA target sites. GSE53625 represents the GSE53625 ( $n = 119$ ) dataset.

to share similar nodes compared with edges. In other words, the nodes in the ESCC ceRNA network may be more conservative than regulatory relationships between nodes in different patients and datasets. In the ESCC tissues of different patients, those lncRNAs, which perform their ceRNA functions, may usually be about the same. Moreover, they tend to regulate the similar terminal target PCGs. However, despite significantly sharing some ceRNA regulatory paths in the different patients, these ce-lncRNAs may adopt many different regulatory paths to transmit signals and implement the regulation for the same target PCGs.

Next, we compared the GloceRNA with other ceRNA identification methods, including SAM and Limma (Fig. 2C, Fig. S1), and GloceRNA displayed markedly higher consistency and stability compared with either SAM or Limma. To further test the performance of GloceRNA, we analyzed the TCGA ESCC datasets. Because TCGA did not have full matched samples, ceRNAs cannot be identified directly using our method. Alternatively, we computed Pearson correlation of the expression ceRNAs. Indeed, we observed that ceRNA pairs identified by our methods were significantly higher co-expressed than others (Fig. 2B).

We next focused on the 61 ce-lncRNAs shared in two datasets. We observed that transcripts for ce-lncRNAs were longer than other lncRNAs ( $P = 5.76E-8$ , Wilcoxon rank-sum test, Fig. 2D). Moreover, ce-lncRNAs had more exons per transcript than other lncRNAs ( $P = 4.57E-33$ , Wilcoxon rank-sum test, Fig. 2D). These observations support previous findings that lncRNAs with longer transcripts and a greater number of exons would be expected to have a higher probability of forming sequence structures that harbor miRNA target sites (Wang *et al.*, 2015). In addition, these 61 ce-lncRNAs were expressed higher and contained more miRNA target sites than other lncRNAs (Fig. 2E,F), again consistent with known features of ceRNAs (Wang *et al.*, 2015).

### 3.2. The topological network analysis identifies novel functional ce-lncRNAs in ESCC

Using internal dataset (15 paired tumor and normal samples), 13 268 ceRNA pairs were identified, involving 98 lncRNAs and 5236 PCGs. We next investigated the 1365 ceRNA pairs shared in two datasets, involving 40 lncRNAs and 1004 PCGs (Right panel, Fig. 2A). To understand this complex regulatory network, we applied the topology theory in biology, wherein biological molecules sharing components within the network are predicted to be more

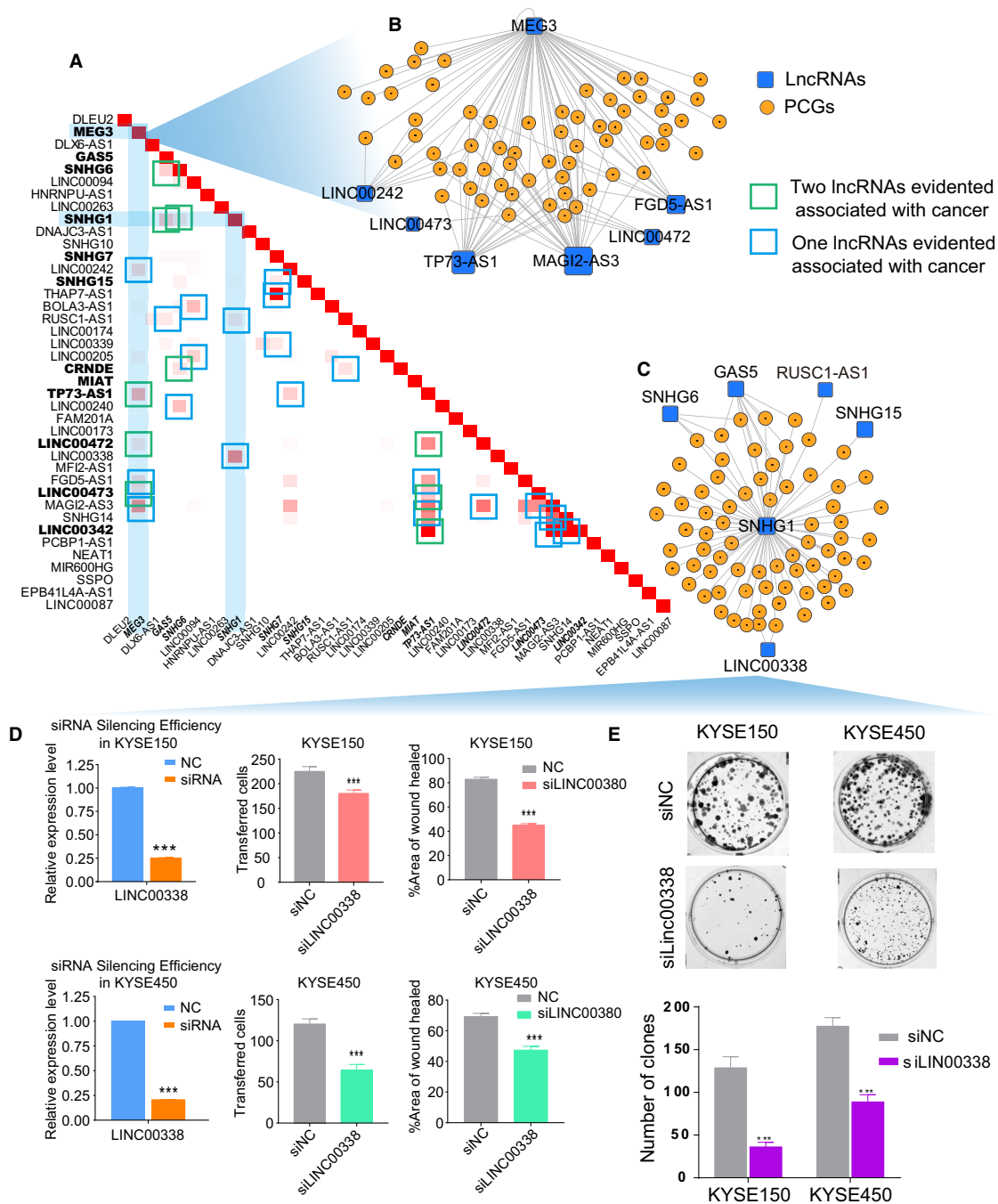
biologically functionally similar. Specifically, we computed the shared PCGs of these lncRNAs in a pairwise manner. Importantly, a few known functional lncRNAs in cancer biology were validated by this method (Fig. 3A). For example, the lncRNA MEG3 shared multiple PCGs with six other lncRNAs, of which three (TP73-AS1, LINC00472 and LINC00473) (Chen *et al.*, 2018; Mazor *et al.*, 2019; Shen *et al.*, 2015) were also confirmed to be of biological significance in cancer (Fig. 3B). On the other hand, we proposed that the functions of poorly characterized lncRNAs may be predicted on the basis of sharing PCGs with known lncRNAs (i.e., guilt-by association). To address this hypothesis, we tested LINC00338, an uncharacterized lncRNA which shared PCGs with SNHG1 (Fig. 3C). SNHG1 contributes to cell growth and survival in several cancer types, and we also found it connected with other known cancer-associated lncRNAs, such as GAS5 and SNHG6 (Fig. 3C). To probe the biological function of LINC00338 in ESCC, we examined the effect of LINC00338 knockdown and found that silencing of this lncRNA potently reduced the proliferation, migration and clonogenicity of ESCC cells (Fig. 3D,E). These data demonstrate that our topological network analysis is capable of identifying both known and novel functional ce-lncRNAs.

### 3.3. Ce-lncRNAs control broad cancer-related hallmarks

Next, we investigated ce-lncRNAs in the context of cancer hallmarks. We collected ten cancer hallmarks and their associated genes based on 31 GO terms (Fig. S2A, Appendix S2). Through mapping all ce-lncRNAs-related PCGs identified by GloceRNA, we found that seven of ten hallmarks, which corresponded to 18 GO terms, were significantly enriched ( $P < 0.05$ , hypergeometric test, Fig. 4A, Fig. S2B). The 'Evading Growth' hallmark displayed the most significant enrichment, followed by 'Resisting Cell Death', 'Genome Instability', and 'Angiogenesis and Activating Invasion' (hypergeometric test, Fig. 4A). We next explored hallmark functions associated with each ce-lncRNA through enrichment analysis and revealed that a total of 449 pairs were enriched in the 10 hallmark GO terms (Fig. 4B red and yellow part, Appendix S2). Eighty-nine out of 109 ce-lncRNAs were significantly associated with at least cancer hallmark (Fig. 4C). Notably, up to 51 ce-lncRNAs were significantly enriched in the 'Cell proliferation' and 'Cell\_cycle' terms (Fig. 4B top panel, Fig. S2C).

On the basis of the number of enriched GO terms, LINC00094, a novel lncRNA with unknown functions,



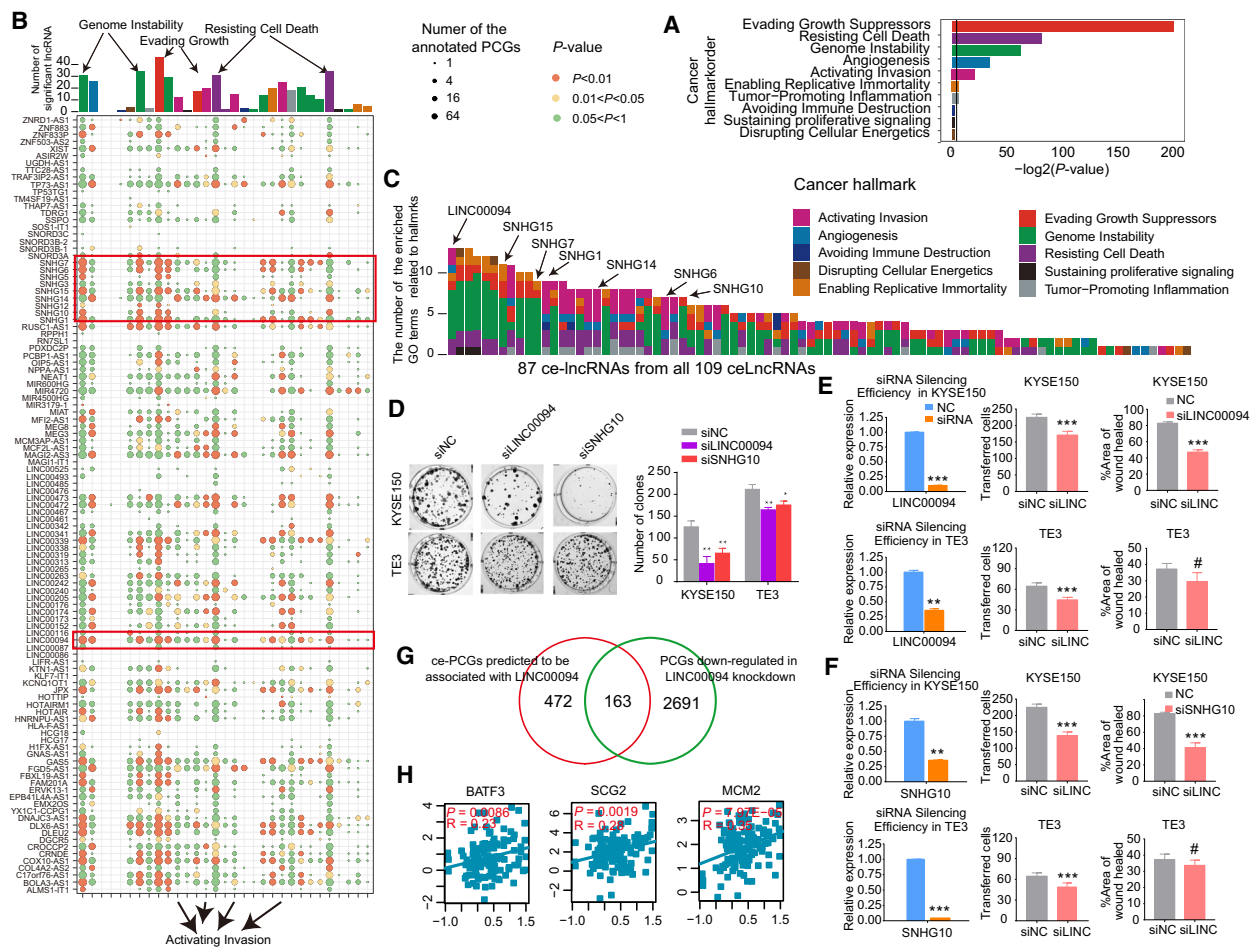


**Fig. 3.** The functional ce-lncRNAs in the conservative ceRNA network. (A) Similarity between lncRNAs based on the conservative ceRNA network. For a lncRNA-lncRNA pair, similarity is tested through computing the shared PCGs of two lncRNAs. (B, C) The subnetworks related to MEG3 or SNHG1, respectively. The two subnetworks were extracted from the conservative ceRNA network through considering MEG3 or SNHG1 as the center. Similarities between lncRNAs were tested through computing the shared PCGs of two lncRNAs in the network, which have been provided in (A). (D) Wound healing assay, transwell migration assays, and (E) colony formation assay were performed to determine the effect of LINC00038 on proliferation, migration, and clonogenicity. Mean+s.d. are shown,  $n = 3$ . \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .

was top ranked, and it was enriched in several cancer hallmarks, including ‘Evading Growth’, ‘Resisting Cell Death’, ‘Genome Instability’, ‘Angiogenesis’, and ‘Activating Invasion’ (Fig. 4C). To test this, we silenced this lncRNA and observed that LINC00094 knockdown significantly inhibited proliferation, migration and clonogenicity in ESCC cells (Fig. 4D,E). More importantly, RNA-seq data showed that the downregulated PCGs upon LINC00094 knockdown significantly overlapped with those predicted by

GloceRNA, strongly validating our method ( $P = 7.97E-05$ , hypergeometric test, Fig. 4G). Some of LINC00094 target PCGs have well-known functions in cancer, including BATF3, SCG2, and MCM2. Expectedly, their expression levels were significantly correlated with LINC00094 ( $P = 7.97E-05$ , Pearson correlation coefficient test, Fig. 4H).

In addition to LINC00094, we also noted that small nucleolar RNA host genes (SNHG), including SNHG15, SNHG7, SNHG1, SNHG14, SNHG6,



**Fig. 4.** The ceRNA network controls broad cancer-associated hallmarks. (A) The cancer hallmarks enriched by ce-lncRNA-related PCGs in the ceRNA network. The colors of bars correspond to different cancer hallmarks. (B) The summary bubble-bar plot shows the functional enrichment results of each ce-lncRNA based on their related PCGs. The top bars show the number of significantly enriched ce-lncRNAs in each GO term. The ten different colors of bars correspond to ten different cancer hallmarks. The bubble size indicates the number of the annotated ce-lncRNA-related PCGs in each term, and different colors correspond to different  $P$  values. (C) The number of the significantly enriched GO terms for each ce-lncRNA. Only lncRNAs with at least one enriched GO terms are displayed. These lncRNAs are ranked by number of the GO terms. The colors correspond to different cancer hallmarks. (D) Colony formation assay and (E, F) wound healing assay, and transwell migration assays were performed to determine the effect of LINC00094 and SNHG10 on proliferation, migration, and clonogenicity. Mean  $\pm$  SD are shown,  $n = 3$ . \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ , #not significant. (G) Venn diagram showing the overlap between ce-PCGs predicted to be associated with LINC00094 and PCGs downregulated in LINC00094 knockdown. (H) Gene expression correlation between LINC00094 and several representative ce-PCGs with cancer hallmark, including BATF3, SCG2, and MCM2. LINC00094 expression level was significantly highly co-expressed with them based on Pearson correlation coefficient test.

SNHG5, SNHG12, and SNHG10, were significantly enriched to most of hallmarks (Fig. 4C). Moreover, their related PCGs were with high number of annotated genes (Fig. 4B). Interestingly, multiple small nucleolar RNA host genes were recently frequently reported in cancers (Damas *et al.*, 2016; Dong *et al.*, 2018; Guo *et al.*, 2018; Jiang *et al.*, 2018a; Shan *et al.*, 2018; Sun *et al.*, 2017; Xu *et al.*, 2018; Zhu *et al.*, 2019). In ESCC, we found that knockdown of SNHG10, an uncharacterized lncRNA, reduced proliferation, migration, and clonogenicity in KYSE150 and TE3 cells (Fig. 4D,F). These results suggest that our hallmark enrichment analysis of ce-lncRNAs may be used to identify additional functional lncRNAs in cancer biology.

### 3.4. Survival analysis of ce-lncRNAs

An increasing number of studies have suggested that lncRNAs acting as ceRNAs can be powerful predictors of survival in cancer patients (Wang *et al.*, 2015; Xu *et al.*, 2015). We next explored the relationship between ce-lncRNA expression and prognosis of ESCC patients by Kaplan–Meier analysis and log-rank test. Eight of 61 (11.26%) ce-lncRNAs were identified with  $P < 0.05$  (Fig. 5A). Five of them were associated with cancer hallmarks (Fig. 5B). Three ce-lncRNAs including LINC00094, LINC00205, and SNHG6 exhibited higher degree/betweenness in the ceRNA network and more numbers of hallmarks than most of other ce-lncRNAs (Fig. 5C). For the lncRNA LINC00094, patients with high lncRNA expression have significantly shorter overall survival (OS) than those with the low expression (Fig. 5D). These lncRNAs were all enriched to ‘Evading Growth’, a hallmark most significantly enriched by ce-lncRNA network (Figs 4A and 5B). These data suggest that these three lncRNAs might have potential biological significance in ESCC.

Further exhaustive survival analysis was performed on each ceRNA pair (i.e., a pair of lncRNA and PCG) to test their prognostic value. We observed that the ceRNA pairs identified by GloceRNA were more associated with ESCC prognosis than random pairs ( $P = 1.16E-38$ , Wilcoxon rank-sum test), with the ceRNA pairs in the topological ceRNA network being more associated (Fig. 5E, ‘overlap’ in Left panel). Moreover, the ceRNA pairs annotated to functional pathways were more associated with prognosis than others ( $P < 0.01$ , Wilcoxon rank-sum test, Fig. 5E, Right panel). Specifically, a total of 31 lncRNA-PCG pairs were significantly associated with ESCC prognosis (Fig. 5F). As an example for the ceRNA pair of

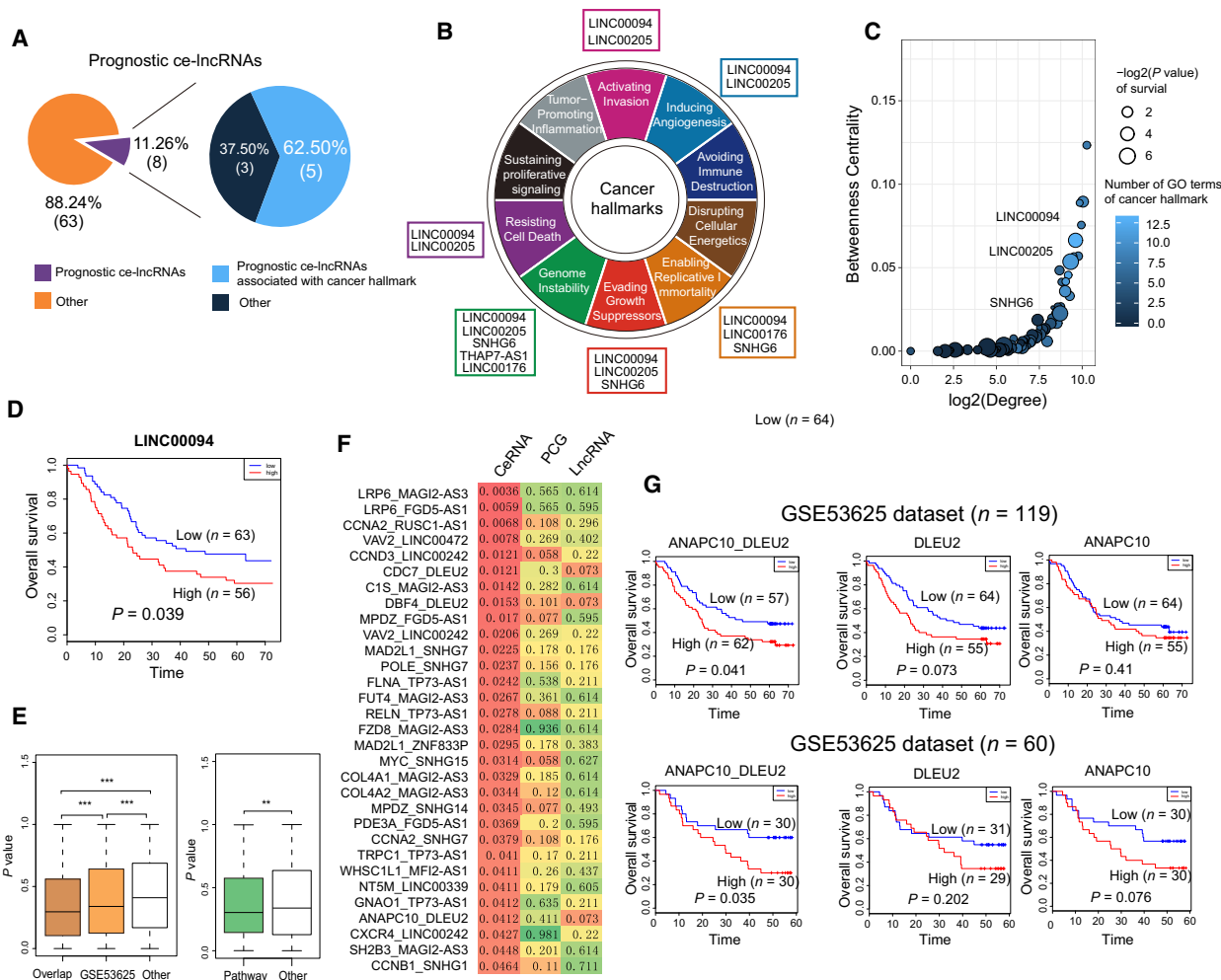
ANAPC10-DLEU2, patients with high expression have significantly shorter OS than those with the low expression in the cohort of 119 patients (the GSE53625  $n = 119$  dataset) (Fig. 5G, Top panel), which was validated in another independent ESCC cohort (GSE53625  $n = 60$  dataset) (Fig. 5G, Bottom panel). These data indicate that functional ce-lncRNAs and ceRNA pairs have prognostic value in ESCC.

### 3.5. SEs play key roles in the regulation of ce-lncRNAs

Although the biological functions of a few ce-lncRNAs have been characterized, the upstream regulatory mechanisms of this class of RNAs are largely unknown. Recent studies have shown that a large number of novel noncoding RNAs can be driven by SEs/TEs, which are important for controlling cell identity and cell type-specific processes (Duan *et al.*, 2016; Hnisz *et al.*, 2013; Huang *et al.*, 2019; Jiang *et al.*, 2018b; Miao *et al.*, 2018; Peng *et al.*, 2019; Wood *et al.*, 2018; Xiang *et al.*, 2014; Xie *et al.*, 2018). To explore the epigenomic mechanisms regulating the expression of our ce-lncRNAs, we characterized active cis-regulatory elements in six ESCC cell lines using H3K27ac ChIP-seq data (Jiang *et al.*, 2018b). We identified SEs and TEs using ROSE software (Hnisz *et al.*, 2013) and found that that most of ce-lncRNAs identified by GloceRNA (102/109, 93%) were associated with SEs/TEs in multiple ESCC cell lines (Fig. 6A,C).

Focusing on SE-associated lncRNAs, we determined that 37 out of 109 (33.94%) ce-lncRNAs were assigned to SEs (some examples displayed in Fig. S3), exhibiting 3-fold enrichment than total lncRNAs ( $P = 1.59E-14$ , hypergeometric test, Fig. 6B). Expectedly, SE-associated ce-lncRNAs were expressed at higher levels than TE-associated ce-lncRNAs in TCGA ESCC samples ( $P = 1.20E-16$ , Wilcoxon rank-sum test, Fig. 6D). Moreover, SE-associated ce-lncRNAs had higher prognostic value than TE-associated ce-lncRNAs (Fig. S4). These data imply that SE-associated ce-lncRNAs might be of more biological importance.

We next correlated the expression level, the topological interactive degree, and cancer hallmark analysis of SE-associated ce-lncRNAs. Importantly, we observed that SE-associated ce-lncRNAs with higher topological degree were strongly associated with expression level and the number of cancer hallmarks enriched (Fig. 6E). For example, LINC00094 had the 3rd strongest topological degree, was enriched in the largest numbers of hallmarks, and was expressed at 9th of all ce-lncRNAs. The well-established lncRNA NEAT1, a



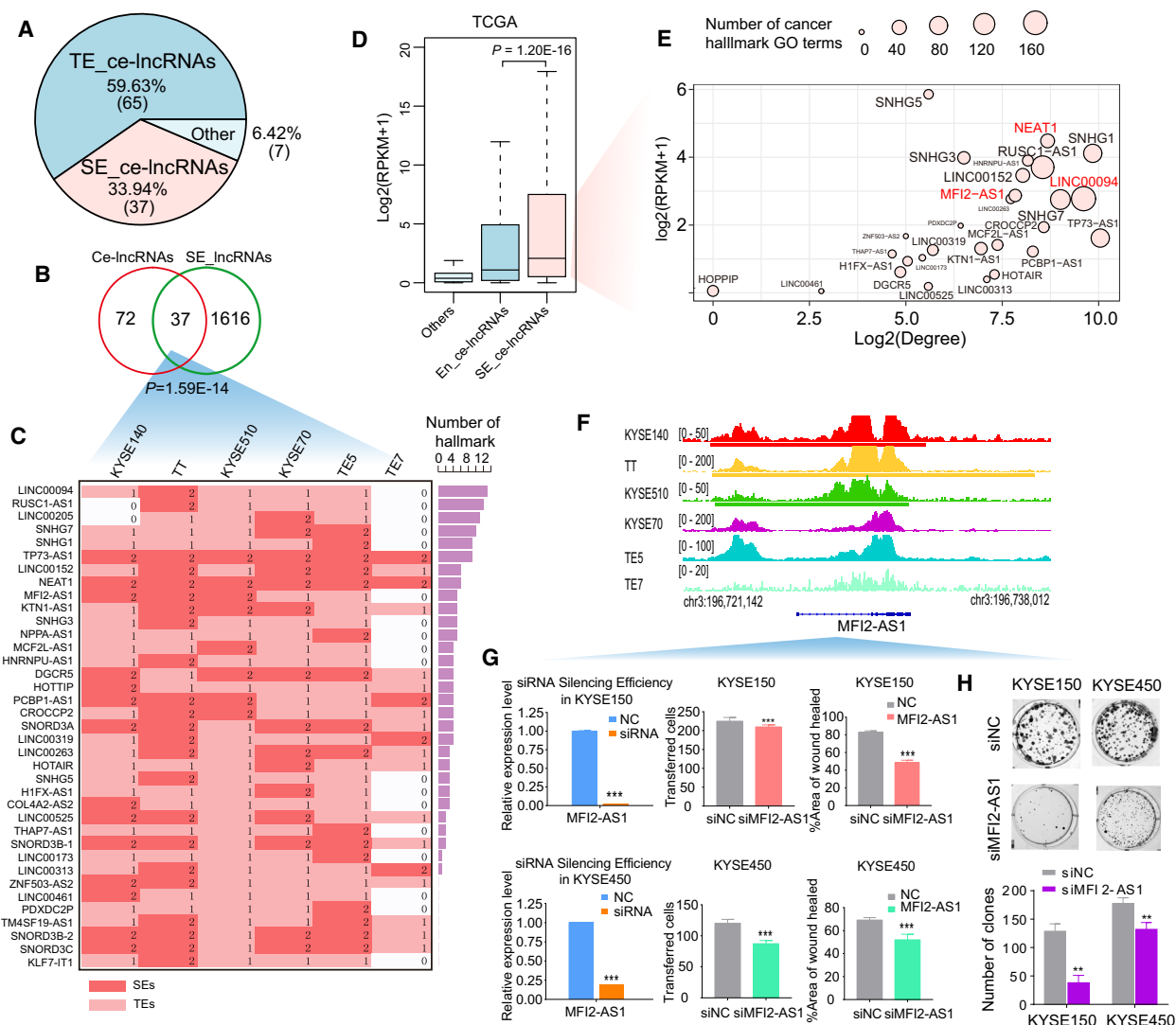
**Fig. 5.** Prognostic analysis of ce-lncRNAs. (A) The pie charts show the proportion of prognostic ce-lncRNAs (covered by cancer hallmarks). (B) Diagram of the ten hallmarks of cancer adapted from Hanahan and Weinberg (2011). The prognostic ce-lncRNAs were assigned to hallmark categories based on GO terms enriched by ce-lncRNA-related PCGs. (C) The summary bubble plot showing the relationships between topological feature and number of hallmark GO terms of SE-associated lncRNAs. X- and y-axis represent degree and betweenness of ce-lncRNAs in the ceRNA network. The bubble size indicates number of hallmark GO terms. (D) Kaplan–Meier survival curves of patients with ESCC classified into high- and low-risk groups based on the signature of lncRNA LINC00094. (E) Left panel: box plots of the lncRNA-PCG ceRNA pairs from pairs in the conservative ceRNA network (overlap), pairs identified by GSE53625 (n = 119), as well as other random pairs. Right panel: box plots of the lncRNA-PCG ceRNA pairs annotated to functional pathways. P values were calculated using Wilcoxon rank-sum test. \*P < 0.05, \*\*P < 0.01, \*\*\*P < 0.001. (F) The lncRNA-PCGs ceRNA pairs that distinguish ESCC patients better than the corresponding single gene. This means that the ceRNA pair using the log-rank test with P values < 0.05 was identified significant, whereas individual lncRNA and PCG were not significant. Color was related to P values. (G) Kaplan–Meier survival curves of ESCC patients the GSE53625 (n = 119) and GSE53625 (n = 60) datasets that were classified into high- and low-risk groups based on ceRNA pair signature, as well as their corresponding lncRNA and PCG.

SE-associated ce-lncRNAs, was also top ranked in terms of expression level, the topological degree and cancer hallmark enrichment. We next explored whether we could identify novel functional ce-lncRNAs by this integrative analysis. We selected a new SE-associated ce-lncRNA (MFI2-AS1), whose SEs appeared in multiple cell lines, was confirmed by us as functionally oncogenic lncRNAs (Fig. 6F–H).

### 3.6. THZ1 inhibits SEs associated ce-lncRNAs

To further investigate the regulation dynamics of SEs on these ce-lncRNAs, we examined the transcriptomic data upon CDK7 inhibition (THZ1), which we have previously shown to preferentially reduce the activity of SEs over TEs (Jiang *et al.*, 2017). The effects of THZ1 inhibition for lncRNAs and related PCGs were

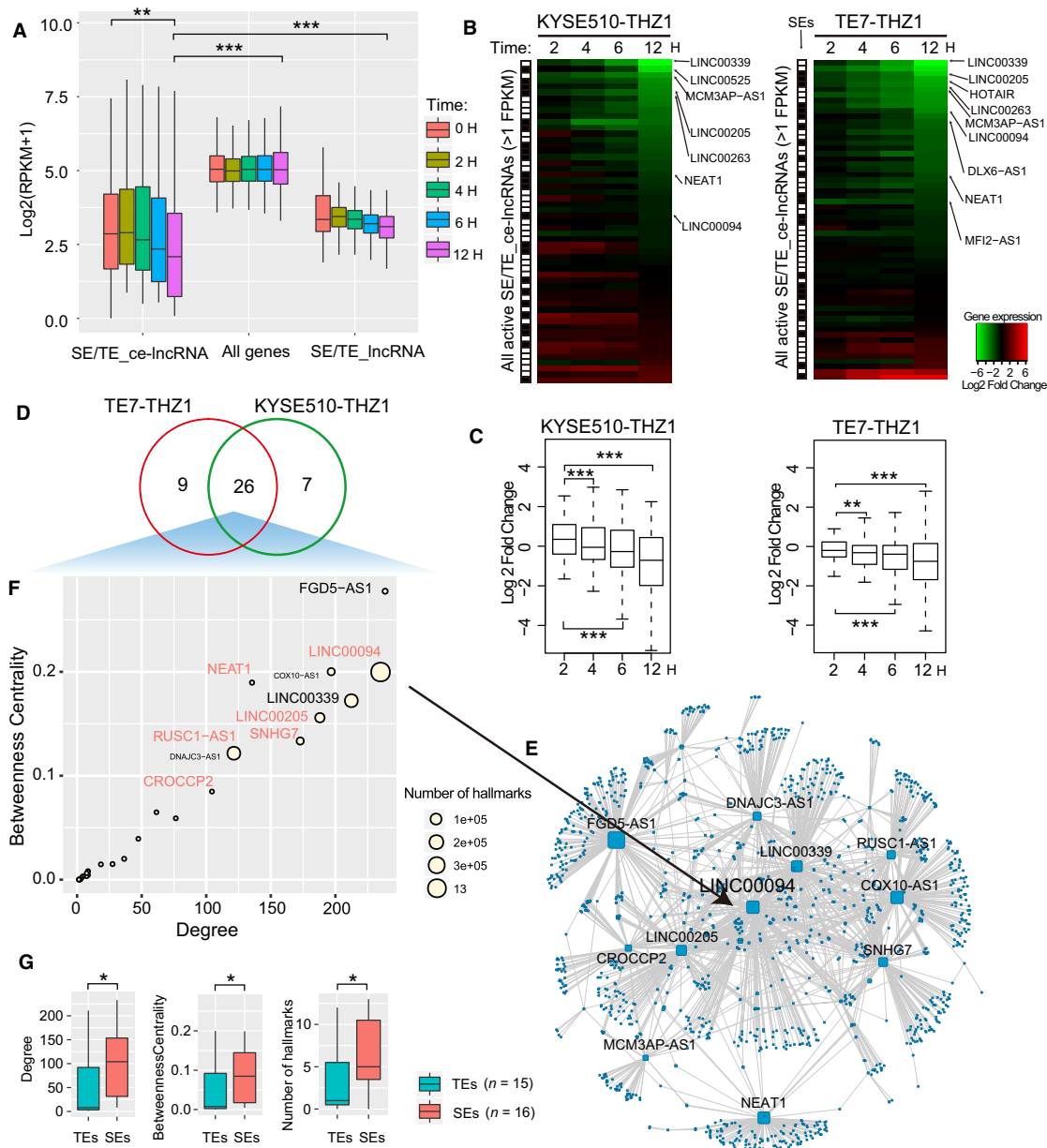




**Fig. 6.** Global overview of SE/TE-associated ce-lncRNAs. (A) Pie chart displays SE- and TE-associated ce-lncRNAs. (B) Venn diagram showing the overlap between ce-lncRNAs and SE-associated lncRNAs. SE-associated lncRNAs are union of lncRNAs associated with SEs appearing in six cell lines. (C) The distribution of SEs/TEs in six cell types about the SE-associated ce-lncRNAs. The colored grids represent the TE-associated (red) or SE-associated (light red) ce-lncRNAs involved in each cell type. The ce-lncRNAs at least appearing in a cell line are listed on the left. The bars on the right show number of GO terms of cancer hallmarks. (D) Box plots of expression (reads per kilobase of exon per million mapped reads, RPKM) from TE- and SE-associated ce-lncRNAs, as well as other nonenhancer ce-lncRNAs. The expression of lncRNAs was obtained from TCGA, and all disease samples of ESCC were considered. *P* values were calculated using Wilcoxon rank-sum test. (E) The summary bubble plot showing the relationships between topological feature, expression, and number of hallmark GO terms of SE-associated lncRNAs. *X*- and *y*-axis represent degree in the ceRNA network and expression of lncRNAs. The bubble size indicates number of hallmark GO terms. (F) H3K27ac chromatin immunoprecipitation sequencing (ChIP-seq) binding profiles of representative SE-associated ce-lncRNAs in six cell lines. (G) Wound healing assay, transwell migration assays, and (H) colony formation assay were performed to determine the effect of MFI2-AS1 on proliferation, migration, and clonogenicity. Mean ± SD are shown, *n* = 3. \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001.

examined using our previous published gene expression profile (GSE76860) for THZ1 treatment in TE7 and KYSE510 cells (see Materials and methods). The data were involved in gene expression levels associated with either THZ1 or DMSO at indicated time points at 2,

4, 6, and 8 h. We found that although SE/TE-associated lncRNAs and all background PCGs did not display significant downregulation by THZ1, THZ1 resulted in global downregulation of SE/TE-associated ce-lncRNAs at 12 h relative to 0 h (Fig. 7A, Fig. S5).



**Fig. 7.** Inhibition of THZ1 for SE/TE-associated ce-lncRNAs. (A) Boxplot of expression of enhancer associated ce-lncRNAs upon either DMSO or THZ1 (50 nM) at indicated time points. (B) Heatmap showing expression changes ( $\text{log}_2$  fold changes) of all active TE/TE-associated ce-lncRNAs upon either DMSO or THZ1 (50 nM) at indicated time points. (C) Box plots of  $\text{log}_2$  fold changes in global lncRNA expression in KYSE510 and TE7 cells treated with either DMSO or THZ1 (50 nM) at indicated time points. (D) Venn diagram showing the overlap between SE/TE-associated ce-lncRNAs from KYS510 and TE7 cells which decreased over 1.5-fold at 12 h. The overlapped lncRNAs were defined as THZ1-sensitive SE/TE-ce-lncRNAs. (E) A THZ1-sensitive ceRNA network that is constructed using THZ1-sensitive SE/TE-ce-lncRNAs and their related PCGs. (F) The summary bubble plot showing the relationships between topological feature and number of hallmark GO terms of SE-associated lncRNAs. X- and y-axis represent degree and betweenness of THZ1-sensitive SE/TE-ce-lncRNAs in the THZ1-sensitive ceRNA network. The bubble size indicates number of hallmark GO terms. (G) Comparison between SE-associated and SE-associated THZ1-sensitive ce-lncRNAs, including degrees and betweenness in the THZ1-sensitive ceRNA network, as well as the number of cancer hallmark GO terms.

We observed that about half of SE/TE-associated ce-lncRNAs were downregulated by THZ1 at 12 h (Fig. 7B,C). We termed this group (SE/TE-associated ce-lncRNAs which decreased over 1.5-fold at 12 h) as 'THZ1-sensitive SE/TE-ce-lncRNAs', which comprised 42 ce-lncRNAs and 26 of them were shared in both cell lines (Fig. 7D).

Focusing on these 26 ce-lncRNAs, we found that their paired PCGs were also highly sensitive to THZ1 treatment (Fig. S6). We next similarly constructed a THZ1-sensitive SE/TE-ceRNA topological network (Fig. 7E) and computed degrees and betweenness of the network for each ce-lncRNAs. Betweenness is equal to the number of shortest paths from a node to all others that pass through this node, which reflects the ability of control that a node exerts in the network. SE-associated lncRNAs displayed significantly higher topological importance (degrees and betweenness) than TE-associated lncRNAs (Fig. 7F,G). Moreover, they regulated significantly more cancer hallmark pathways than TE-associated lncRNAs (Fig. 7F,G). Some of these SE-associated lncRNAs including LINC00094, LINC00205, and RUSC1-AS1, were shown in Fig. 6E.

### 3.7. KLF5 and TCF3 regulated LINC00094 through binding to its SE regions

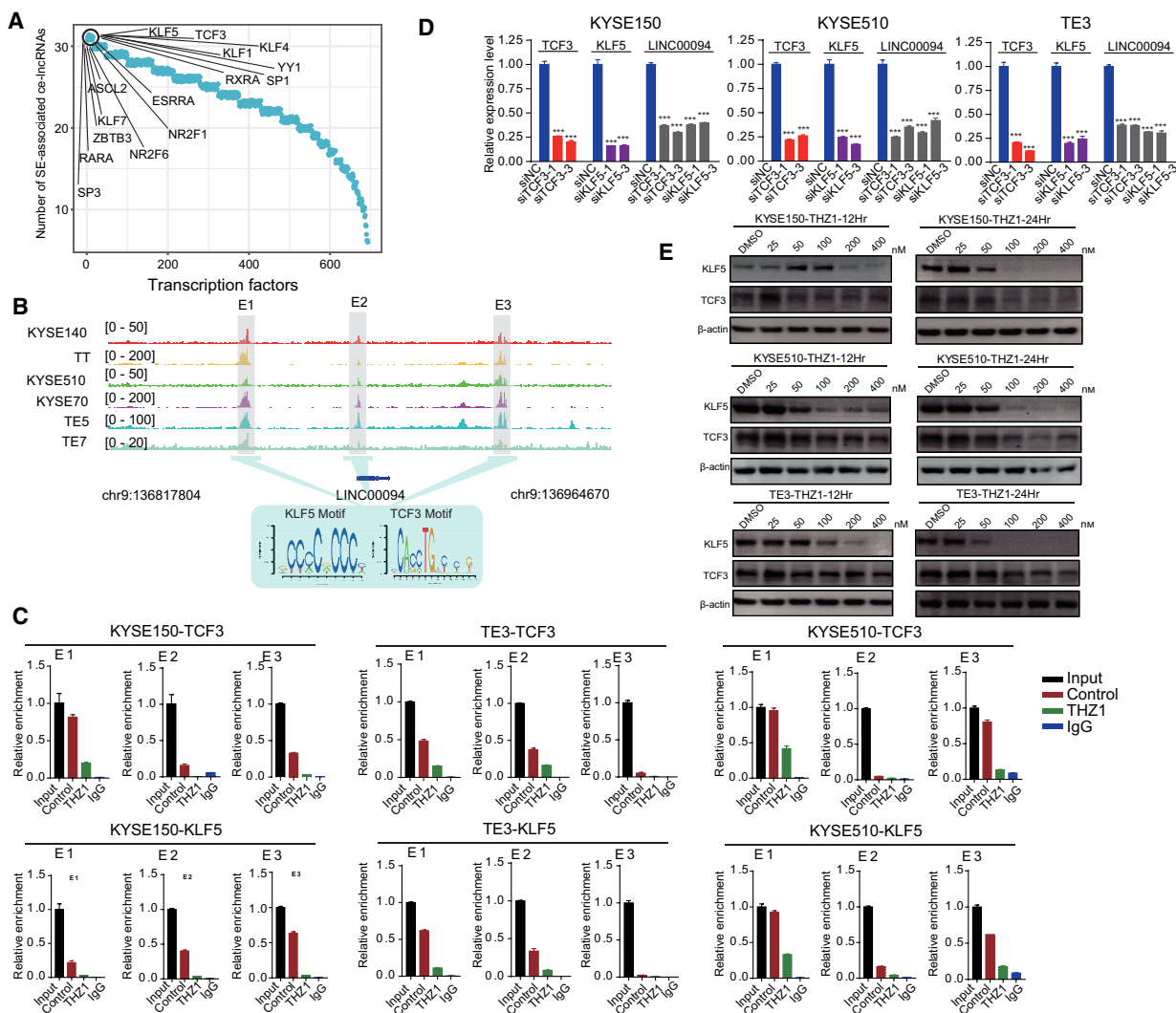
Master transcription factors play key roles in regulating the activity of SEs. To identify such transcription factors responsible for the regulation of SE-associated ce-lncRNAs, we analyzed the frequency of TF binding motifs within SE regions associated with ce-lncRNAs via FIMO software (Grant *et al.*, 2011) from the TRANSFAC database (Matys *et al.*, 2006) and MEME suite (Bailey *et al.*, 2009). We ranked transcription factors according to number of SE-associated ce-lncRNAs significantly regulated by them. 16 transcription factors that regulated most numbers of SE-associated ce-lncRNAs were identified (Fig. 8A).

Next, because of the functional importance of LINC00094 for ESCC, we focused on this ce-lncRNAs to validate the motif analysis results, which predicted the binding of TCF3 and KLF5 to SEs (E1, E2, and E3) of LINC00094 (Fig. 8B). To validate this, CHIP-qPCR was performed and their enrichment was confirmed at all three SE regions (E1, E2, and E3) (Fig. 8C). Furthermore, we confirmed that THZ1 can inhibit the interaction of TCF3 and KLF5 with the SEs (E1, E2, and E3) (Fig. 8C). More importantly, knockdown of TCF3 or KLF5 significantly downregulated expression of LINC00094 (Fig. 8D, Fig. S7). We also observed decreased expression of TCF3 and

KLF5 in a dose-dependent manner upon THZ1 treatment (Fig. 8E). To further explore the specific mechanism by which signaling pathway this TF-lncRNA axis regulates, we extracted LINC00094-related PCGs identified by GloceRNA and annotated these PCGs to KEGG pathways using the iSubpathwayMiner software package (Li *et al.*, 2009, 2013). Then, the pathways significantly enriched by LINC00094-related PCGs were identified using hypergeometric test with FDR corrected  $P < 0.05$ . We found that signaling pathways and cancer pathways were significantly enriched, including 'PI3K-Akt signaling pathway', 'Pathways in cancer', 'Cell cycle', and 'ErbB signaling pathway'. In these pathways, 'PI3K-Akt signaling pathway' contained many LINC00094-related PCGs (Fig. S8A). Notably, LINC00094 regulated 19 PCGs in the pathway (Fig. S8B). Especially, we found that the core nodes within the 'PI3K-Akt signaling pathway' such as PIK3CA and AKT3 can be regulated by LINC00094 (Fig. S8C). These data demonstrate that TCF3 and KLF5 occupy the SEs of LINC00094, thereby activating its transcription and related downstream signaling pathways in ESCC cells.

## 4. Discussion

Esophageal squamous cell carcinoma is the predominant histological type of esophageal cancer and is considered one of the most common and leading aggressive malignancies with poor prognosis (Jemal *et al.*, 2011). In China, over 90% of the cases of esophageal cancer are ESCC, which is the fourth most prevalent cancer of the country (Yang *et al.*, 2005; Zhao *et al.*, 2010). Recently, researchers have determined the genomic landscape of ESCC and identified a number of driver events (Agrawal *et al.*, 2012; Gao *et al.*, 2014; Lin *et al.*, 2014; Song *et al.*, 2014). However, genetic alterations of drug targets are infrequent in patients with ESCC (Agrawal *et al.*, 2012; Gao *et al.*, 2014; Lin *et al.*, 2014; Song *et al.*, 2014). Clearly, alternative molecular approaches are needed to further elucidate the pathogenesis of ESCC for developing more innovative and effective regimens. It has now become widely accepted that mammalian genomes encode numerous lncRNAs. Nonetheless, the functional roles of most of these transcripts remain obscure and their upstream/downstream regulatory mechanisms are largely unknown. To systematically pinpoint functional lncRNAs involved in ESCC pathogenesis, we constructed a putative lncRNA-mediated ceRNA network by integrating lncRNA and PCG expression based on high-throughput RNA sequencing and microarray data. Based on



**Fig. 8.** KLF5 and TCF3 bind to SE regions for regulation of LINC00094. (A) The ranked transcription factors according to number of SE-associated ce-lncRNAs significantly regulated by transcription factors. (B) H3K27ac ChIP-seq signals at the LINC00094 locus in six ESCC cell lines. Three constituent enhancers (E1, E2, and E3) within the SE were labeled in grey shadings. TCF3 and KLF5 motif occupy at E1, E2, and E3 enhancer loci. (C) ChIP-qPCR experiments measuring TCF3 and KLF5 binding enrichment on the LINC00094 SEs segments (divided into enhancer 1, E1; enhancer 2, E2 and enhancer 3, E3) upon treating with THZ1 (100 nM, 12 h). Two pairs of primers were designed for each SE segments, which has the better enrichment was finally selected. (D) Relative RNA expression of LINC00094 upon knockdown of TCF3 or KLF5 in KYSE150, KYSE510, and TE3 cells. (E) Western blotting analysis of KLF5 and TCF3 expression in KYSE150, KYSE510 and TE3 cells which were treated with either THZ1 or DMSO at indicated time points and indicated concentrations. Bars of D represent mean  $\pm$  SD of three experimental replicates. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$ .  $P$  values were determined using  $t$ -test.

bioinformatic and experimental approaches, we identified many known and novel functional ce-lncRNAs and found that most of them acted as a ceRNAs to regulate the expression of broad cancer-related hallmark genes in ESCC. Interestingly, these lncRNAs acting as ceRNAs were significantly regulated by enhancers, especially SEs. Ce-lncRNAs have recently been observed to be regulated by SEs. However, ce-lncRNAs targeted by SEs have not been discovered

thus far in ESCC, and the regulation of SEs on ce-lncRNAs has not been studied.

MiRNAs can mediate ceRNA interaction. If sample matched miRNA, lncRNA, and PCG expression profiles are available, expression correlation of the lncRNA-miRNA-PCG triplet can be calculated. Especially, Paci *et al.* developed an effective measure, called sensitivity correlation, to calculate the difference between Pearson and partial correlation coefficients



for identification of ceRNAs. However, in the ESCC study, we did not measure expression correlation of the lncRNA-miRNA-PCG triplet because our miRNA expression profiles are unavailable. We also found that for many diseases, it is difficult to obtain the sample matched lncRNA, miRNA, and PCG expression profiles. Instead, we focus on prediction of functional ce-lncRNAs using N/T matched samples. The functional ce-lncRNAs are predicted using GloceRNA based on merging global and local expression associated with ceRNAs. Especially, using a new measure  $dec_f(l, g)$ , expression direction consistency between lncRNAs and PCGs can be effectively considered at single sample level. Suppose that a lncRNA-PCG ceRNA pair is true. Then, when the expression level of lncRNA in the pair increases in the tumor sample of a patient compared with normal sample, expression of the corresponding PCG should also tend to increase. Therefore, for a pair of N/T samples from the same patient, a ceRNA pair usually displays consistently upregulated (or downregulated) in expression direction. We used  $dec_f(l, g)$  to measure consistency at single sample level, which displayed hidden regulatory direction information of ceRNAs from single N/T matched samples. Based on  $dec_f(l, g)$ , we further counted sample number of up/downregulated differential expression consistency across all samples, defined as DEC score, for obtaining local regulatory direction consistency. DEC score can evaluate possibility of ceRNAs significantly appearing in parts of samples through testing times of consistency at single sample level across all samples. We demonstrated that our methods robustly predicted ce-lncRNAs in multiple ESCC datasets, and the predicted ce-lncRNAs strongly regulated cancer hallmarks. Moreover, we experimentally validated that some new ce-lncRNAs predicted by GloceRNA were highly associated with oncogenic functions of ESCC, including LINC00094, LINC00338, SNHG10 and MFI2-AS1. Especially, a SE-associated ce-lncRNA, LINC00094, can promote ESCC cancer cell growth through being activated by TFs binding to SEs. Taken together, if the N/T matched data are available, GloceRNA can provide some useful predictions through effectively using N/T matched samples. GloceRNA thus has potential to complement the existing ceRNA identification methods, as the effective use of N/T matched data and focusing on functional ce-lncRNAs in ESCC.

We found that most of them significantly regulated the expression of cancer-related hallmark genes. These ce-lncRNAs were significantly regulated by enhancers, especially SEs. Landscape analyses for lncRNAs further identified SE-associated functional ce-lncRNAs in ESCC, such as HOTAIR, XIST, SNHG5, and

LINC00094. THZ1, a specific CDK7 inhibitor, can result in global transcriptional downregulation of SE-associated ce-lncRNAs. We further demonstrate that a SE-associated ce-lncRNA, LINC00094 can be activated by transcription factors TCF3 and KLF5 through binding to SE regions and promoted ESCC cancer cell growth. THZ1 downregulated expression of LINC00094 through inhibiting TCF3 and KLF5. Our data demonstrated the important roles of SE-associated ce-lncRNAs in ESCC oncogenesis and might serve as targets for ESCC diagnosis and therapy.

Efforts to interpret the functional consequences of SEs have mainly focused on the regulation of PCGs, although in a few cases lncRNA regulation was studied. Recent report demonstrated that master transcription factors TP63 and SOX2 promote SCC tumorigenesis such as ESCC through lineage specifically regulating a lncRNA mediated by SEs. We defined a new class of lncRNA, SE-associated ce-lncRNA, and performed a thorough investigation of its functional relevance in ESCC cancer cells. Some SE-associated ce-lncRNAs with high degree/betweenness were highly associated with cancer hallmarks, including lncRNAs reported in cancer (e.g., NEAT1, HOTAIR, XIST, and SNHG5). Two novel SE-associated ce-lncRNAs (LINC00094 and MFI2-AS1) was identified and validated by us as functionally oncogenic lncRNAs. Our previous studies showed that the unbiased high-throughput small-molecule inhibitor screening discover a highly potent anti-ESCC compound, THZ1, a specific CDK7 inhibitor. Targeting SE-associated coding gene activation by THZ1 shows powerful antineoplastic properties against ESCC cells (Jiang *et al.*, 2017). Furthermore, we found that THZ1 resulted in global downregulation of SE/TE-ce-lncRNAs. Furthermore, 26 THZ1-sensitive SE/TE-ce-lncRNAs in both cell lines were identified by us and the related THZ1-sensitive ceRNA network was extracted. In the network, SE-associated lncRNAs displayed significantly higher topological importance than TE-associated lncRNAs. Moreover, they significantly regulated more cancer hallmark pathways than TE-associated lncRNAs, such as LINC00094, LINC00205, and RUSC1-AS1. Our findings support recent studies suggesting that SEs can function as important regulators of lncRNAs. SEs play important roles by ce-lncRNAs.

In process of analysis, we found an important functional ce-lncRNA, LINC00094. The enrichment analysis showed that the SE-associated lncRNA was closely related to more than number of cancer hallmarks than other lncRNAs (Fig. 4). LINC00094 parted in core cancer hallmark of ESCC ceRNAs such as 'Evading

Growth' and 'Genome Instability', and its overexpression was highly associated with poor clinical outcome in ESCC patients (Fig. 5D). In all eight significant prognostic ce-lncRNAs, LINC00094 was with highest degree, betweenness in the ceRNA network and related to most numbers of hallmarks (Fig. 5C). LINC00094 was strongly inhibited by THZ1 and located at the center of the THZ1-sensitive ceRNA network (Fig. 7E,F). Krüppel-like transcription factors (KLF) play important roles in development and cancer. KLF4 is a master transcription factor for maintaining the pluripotency of embryonic stem cells (Takahashi and Yamanaka, 2006). It has been reported that KLF5 is highly expressed in multiple cancer types and promotes cancer cell proliferation, migration and survival (Ben-Porath *et al.*, 2008; Chia *et al.*, 2015; Jia *et al.*, 2016; Nandan *et al.*, 2008; Qin *et al.*, 2015; Zhang *et al.*, 2018b). Especially, KLF5 activates cell identity genes and cancer genes in squamous cell carcinomas (Nakaya *et al.*, 2014). KLF5 is also able to occupy the lncRNA RP1 promoter to enhance RP1 expression, which plays an oncogenic role in breast cancer (Jia *et al.*, 2019). All the evidence indicates the importance of KLF5 activation in human cancer. We confirmed that transcription factors TCF3 and KLF5 occupied the SE constituents of LINC00094, thereby activating its transcription in ESCC cells. THZ1 decreased expression of TCF3 and KLF5 and inhibited the occupancy of TCF3 and KLF5 (Fig. 8). These results demonstrate that TCF3 and KLF5 can occupy the SEs of LINC00094, thereby activating its transcription in ESCC cells. THZ1 down-regulated expression of LINC00094 through inhibiting TCF3 and KLF5.

GloceRNA successfully predicted many ce-lncRNAs and experimentally validated some new functional ce-lncRNAs. However, our study has also some limitations. For example, we integrated large-scale CLIP-seq (HITS-CLIP, PAR-CLIP, iCLIP, CLASH) from the starBase database to obtain enough experimental miRNA-lncRNA and miRNA-mRNA interactions. Although these 'big' data provided the comprehensive high-quality information, the datasets used were based on different biological sources such as patients, cell lines, and some did not come from squamous cells and cancer cells. With the accumulation of esophageal squamous cell data, use of cell type-specific data would be helpful for more accurately identifying ceRNAs and ce-lncRNAs. Furthermore, there is still much room for improvement in the usability and stability of GloceRNA. For example, although GloceRNA displayed higher stability for identification of ceRNA pairs compared with other state-of-the-art methods, there is still

much room for improvement in the stability of ceRNA pairs. Also, the current version of GloceRNA must input the N/T matched data. Therefore, GloceRNA was still unavailable for input of data with non-matched samples. Some 'single sample' strategies of expression profiling analysis may be useful for improving the ability of our method to identify ceRNAs in nonmatched data in the future (Li *et al.*, 2020; Liu *et al.*, 2016, 2017). In addition, in the current version of GloceRNA, the cutoff of DEC score needs to be manually set and adjusted. The different flexible strategies for setting the cutoff of DEC score, as well as automatic parameter adjustment, would facilitate the identification of functional ce-lncRNAs. The current strategy for setting the cutoff in the paper is simple and intuitive, and setting the same cutoff in two dataset can also penalize the dataset with small sample size, in which the higher proportion of samples need to meet regulatory direction consistency. We think that other strategies for setting cutoffs may also be effective. For example, the different cutoffs can be selected such as setting the cutoffs according to proportion of samples meeting regulatory direction consistency. With advances in our identification strategy and the accumulation of genomic/transcriptomic profiling data, performance of GloceRNA would continue to improve.

## 5. Conclusion

In summary, we focus on prediction of functional ce-lncRNAs using N/T matched samples. We developed the GloceRNA method for identification of functional ce-lncRNAs based on merging global and local regulatory direction consistency of expression associated with ceRNAs. The ce-lncRNAs unique to squamous cell carcinomas have not been studied extensively. GloceRNA identified many known and novel functional ce-lncRNAs which regulated the expression of a large number of cancer hallmark genes. Interestingly, we identified novel SE-associated ce-lncRNAs in ESCC. Among them, we identified a SE mediated mechanism for the upregulation of a novel oncogenic lncRNA, LINC00094, in ESCC. Considering this gene's ESCC-specific nature, its association with poor patient survival, and its oncogenic functions, LINC00094 represents a potential biomarker and/or therapeutic target in this group of deadly cancers.

## Acknowledgements

We thank for the members of Daqing Campus, Harbin Medical University, Medical College of Shantou

University, and Cedars-Sinai Medical Center for helpful discussions. This work has been supported by the National Natural Science Foundation of China [81772532, 61601150, 81572341, 81602630] (in part); The National Key R&D Program of China (2018YFC1313101); Natural Science Foundation of Heilongjiang Province [JJ2016ZR1232/F2016031]; Yu Weihai Outstanding Youth Training Fund of Harbin Medical University.

## Conflict of interest

The authors declare no conflict of interest.

## Author contributions

E-ML, Q-YW, L-YX, D-CL, and C-QL conceived and devised the study. Q-YW and LP designed experiments and analysis. LP, Q-YW, YC, L-DL, J-XC, ML, Y-YL, F-CQ, Y-XZ, and FW performed the experiments. Q-YW, J-XC, ML, Y-YL, F-CQ, Y-XZ, FW, and C-QL performed bioinformatics and statistical analysis. Q-YW, LP, YC, L-DL, J-XC, ML, Y-YL, F-CQ, Y-XZ, and FW analyzed the data. Q-YW, LP, E-ML, L-YX, D-CL, J-XC, and C-QL wrote the manuscript.

## Data availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

## References

- Agrawal N, Jiao Y, Bettgowda C, Hutfless SM, Wang Y, David S, Cheng Y, Twaddell WS, Latt NL, Shin EJ *et al.* (2012) Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. *Cancer Discov* **2**, 899–905.
- Amaral PP and Bannister AJ (2014) Re-place your BETs: the dynamics of super enhancers. *Mol Cell* **56**, 187–189.
- Bai X, Shi S, Ai B, Jiang Y, Liu Y, Han X, Xu M, Pan Q, Wang F, Wang Q *et al.* (2020) ENdb: a manually curated database of experimentally supported enhancers for human and mouse. *Nucleic Acids Res* **48**, D51–D57.
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW and Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202–W208.
- Barabasi AL and Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113.
- Batista PJ and Chang HY (2013) Long noncoding RNAs: cellular address codes in development and disease. *Cell* **152**, 1298–1307.
- Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A and Weinberg RA (2008) An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* **40**, 499–507.
- Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell* **133**, 1266–1276.
- Cancer Genome Atlas Research Network, Analysis Working Group: Asan University, BC Cancer Agency, Brigham and Women's Hospital, Broad Institute, Brown University, Case Western Reserve University, Dana-Farber Cancer Institute, Duke University, Greater Poland Cancer Centre *et al.* (2017) Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175.
- Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A and Bozzoni I (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **147**, 358–369.
- Chen Z, Lin S, Li JL, Ni W, Guo R, Lu J, Kaye FJ and Wu L (2018) CRTC1-MAML2 fusion-induced lncRNA LINC00473 expression maintains the growth and survival of human mucoepidermoid carcinoma cells. *Oncogene* **37**, 1885–1895.
- Chia NY, Deng N, Das K, Huang D, Hu L, Zhu Y, Lim KH, Lee MH, Wu J, Sam XX *et al.* (2015) Regulatory crosstalk between lineage-survival oncogenes KLF5, GATA4 and GATA6 cooperatively promotes gastric cancer development. *Gut* **64**, 707–719.
- Chipumuro E, Marco E, Christensen CL, Kwiatkowski N, Zhang T, Hatheway CM, Abraham BJ, Sharma B, Yeung C, Altabef A *et al.* (2014) CDK7 inhibition suppresses super-enhancer-linked oncogenic transcription in MYCN-driven cancer. *Cell* **159**, 1126–1139.
- Conte F, Fiscon G, Chiara M, Colombo T, Farina L and Paci P (2017) Role of the long non-coding RNA PVT1 in the dysregulation of the ceRNA-ceRNA network in human breast cancer. *PLoS One* **12**, e0171661.
- Damas ND, Marcatti M, Come C, Christensen LL, Nielsen MM, Baumgartner R, Gylling HM, Maglieri G, Rundsten CF, Seemann SE *et al.* (2016) SNHG5 promotes colorectal cancer cell survival by counteracting STAU1-mediated mRNA destabilization. *Nat Commun* **7**, 13875.
- Dong H, Wang W, Chen R, Zhang Y, Zou K, Ye M, He X, Zhang F and Han J (2018) Exosome-mediated transfer of

- lncRNASNHG14 promotes trastuzumab chemoresistance in breast cancer. *Int J Oncol* **53**, 1013–1026.
- Du Z, Fei T, Verhaak RG, Su Z, Zhang Y, Brown M, Chen Y and Liu XS (2013) Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol* **20**, 908–913.
- Duan Q, Mao X, Xiao Y, Liu Z, Wang Y, Zhou H, Zhou Z, Cai J, Xia K, Zhu Q *et al.* (2016) Super enhancers at the miR-146a and miR-155 genes contribute to self-regulation of inflammation. *Biochim Biophys Acta* **1859**, 564–571.
- Feng C, Song C, Liu Y, Qian F, Gao Y, Ning Z, Wang Q, Jiang Y, Li Y, Li M *et al.* (2020) KnockTF: a comprehensive human gene expression profile database with knockdown/knockout of transcription factors. *Nucleic Acids Res* **48**, D93–D100.
- Flynn RA and Chang HY (2014) Long noncoding RNAs in cell-fate programming and reprogramming. *Cell Stem Cell* **14**, 752–761.
- Gao YB, Chen ZL, Li JG, Hu XD, Shi XJ, Sun ZM, Zhang F, Zhao ZR, Li ZT, Liu ZY *et al.* (2014) Genetic landscape of esophageal squamous cell carcinoma. *Nat Genet* **46**, 1097–1102.
- Grant CE, Bailey TL and Noble WS (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018.
- Guo T, Wang H, Liu P, Xiao Y, Wu P, Wang Y, Chen B, Zhao Q, Liu Z and Liu Q (2018) SNHG6 acts as a genome-wide hypomethylation trigger via coupling of miR-1297-mediated S-adenosylmethionine-dependent positive feedback loops. *Cancer Res* **78**, 3849–3864.
- Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076.
- Hanahan D and Weinberg RA (2011) Hallmarks of cancer: the next generation. *Cell* **144**, 646–674.
- Hnisz D, Abraham BJ, Lee TI, Lau A, Saint-Andre V, Sigova AA, Hoke HA and Young RA (2013) Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947.
- Hnisz D, Schuijers J, Lin CY, Weintraub AS, Abraham BJ, Lee TI, Bradner JE and Young RA (2015) Convergence of developmental and oncogenic signaling pathways at transcriptional super-enhancers. *Mol Cell* **58**, 362–370.
- Hosono Y, Niknafs YS, Prensner JR, Iyer MK, Dhanasekaran SM, Mehra R, Pitchiaya S, Tien J, Escara-Wilke J, Poliakov A *et al.* (2017) Oncogenic role of THOR, a conserved cancer/testis long non-coding RNA. *Cell* **171**, 1559–1572, e1520.
- Huang S, Li X, Zheng H, Si X, Li B, Wei G, Li C, Chen Y, Chen Y, Liao W *et al.* (2019) Loss of super-enhancer-regulated circrna Nfix induces cardiac regeneration after myocardial infarction in adult mice. *Circulation* **139**, 2857–2876.
- Jemal A, Bray F, Center MM, Ferlay J, Ward E and Forman D (2011) Global cancer statistics. *CA Cancer J Clin* **61**, 69–90.
- Jia L, Zhou Z, Liang H, Wu J, Shi P, Li F, Wang Z, Wang C, Chen W, Zhang H *et al.* (2016) KLF5 promotes breast cancer proliferation, migration and invasion in part by upregulating the transcription of TNFAIP2. *Oncogene* **35**, 2040–2051.
- Jia X, Shi L, Wang X, Luo L, Ling L, Yin J, Song Y, Zhang Z, Qiu N, Liu H *et al.* (2019) KLF5 regulated lncRNA RP1 promotes the growth and metastasis of breast cancer via repressing p27kip1 translation. *Cell Death Dis* **10**, 373.
- Jiang H, Li T, Qu Y, Wang X, Li B, Song J, Sun X, Tang Y, Wan J, Yu Y *et al.* (2018a) Long non-coding RNA SNHG15 interacts with and stabilizes transcription factor Slug and promotes colon cancer progression. *Cancer Lett* **425**, 78–87.
- Jiang Y, Jiang YY, Xie JJ, Mayakonda A, Hazawa M, Chen L, Xiao JF, Li CQ, Huang ML, Ding LW *et al.* (2018b) Co-activation of super-enhancer-driven CCAT1 by TP63 and SOX2 promotes squamous cancer progression. *Nat Commun* **9**, 3619.
- Jiang YY, Lin DC, Mayakonda A, Hazawa M, Ding LW, Chien WW, Xu L, Chen Y, Xiao JF, Senapedis W *et al.* (2017) Targeting super-enhancer-associated oncogenes in oesophageal squamous cell carcinoma. *Gut* **66**, 1358–1368.
- Jiang Y, Qian F, Bai X, Liu Y, Wang Q, Ai B, Han X, Shi S, Zhang J, Li X *et al.* (2019) SEdb: a comprehensive human super-enhancer database. *Nucleic Acids Res* **47**, D235–D243.
- Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell* **152**, 327–339.
- Karreth FA and Pandolfi PP (2013) ceRNA cross-talk in cancer: when ce-bling rivalries go awry. *Cancer Discov* **3**, 1113–1121.
- Khan A and Zhang X (2016) dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Res* **44**, D164–D171.
- Langmead B, Trapnell C, Pop M and Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25.
- Li C, Han J, Yao Q, Zou C, Xu Y, Zhang C, Shang D, Zhou L, Zou C, Sun Z *et al.* (2013) Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic Acids Res* **41**, e101.
- Li CQ, Huang GW, Wu ZY, Xu YJ, Li XC, Xue YJ, Zhu Y, Zhao JM, Li M, Zhang J *et al.* (2017) Integrative



- analyses of transcriptome sequencing identify novel functional lncRNAs in esophageal squamous cell carcinoma. *Oncogenesis* **6**, e297.
- Li C, Li X, Miao Y, Wang Q, Jiang W, Xu C, Li J, Han J, Zhang F, Gong B *et al.* (2009) SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res* **37**, e131.
- Li J, Chen Z, Tian L, Zhou C, He MY, Gao Y, Wang S, Zhou F, Shi S, Feng X *et al.* (2014a) LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with esophageal squamous cell carcinoma. *Gut* **63**, 1700–1710.
- Li JH, Liu S, Zhou H, Qu LH and Yang JH (2014b) starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* **42**, D92–D97.
- Li M, Zhao J, Li X, Chen Y, Feng C, Qian F, Liu Y, Zhang J, He J, Ai B *et al.* (2019). HiFreSP: a novel high-frequency sub-pathway mining approach to identify robust prognostic gene signatures. *Brief Bioinform*, bbz078. <https://doi.org/10.1093/bib/bbz078>
- Li Y, Li X, Yang Y, Li M, Qian F, Tang Z, Zhao J, Zhang J, Bai X, Jiang Y *et al.* (2020) TRlnc: a comprehensive database for human transcriptional regulatory information of lncRNAs. *Brief Bioinform*, bbaa011. <https://doi.org/10.1093/bib/bbaa011>
- Li Z, Zhang J, Liu X, Li S, Wang Q, Di C, Hu Z, Yu T, Ding J, Li J *et al.* (2018) The LINC01138 drives malignancies via activating arginine methyltransferase 5 in hepatocellular carcinoma. *Nat Commun* **9**, 1572.
- Lin DC, Hao JJ, Nagata Y, Xu L, Shang L, Meng X, Sato Y, Okuno Y, Varela AM, Ding LW *et al.* (2014) Genomic and molecular characterization of esophageal squamous cell carcinoma. *Nat Genet* **46**, 467–473.
- Liu X, Chang X, Liu R, Yu X, Chen L and Aihara K (2017) Quantifying critical states of complex diseases using single-sample dynamic network biomarkers. *PLoS Comput Biol* **13**, e1005633.
- Liu X, Wang Y, Ji H, Aihara K and Chen L (2016) Personalized characterization of diseases using sample-specific networks. *Nucleic Acids Res* **44**, e164.
- Long L, Pang XX, Lei F, Zhang JS, Wang W, Liao LD, Xu XE, He JZ, Wu JY, Wu ZY *et al.* (2018) SLC52A3 expression is activated by NF-kappaB p65/Rel-B and serves as a prognostic biomarker in esophageal cancer. *Cell Mol Life Sci* **75**, 2643–2661.
- Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**, D142–D147.
- Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K *et al.* (2006) TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**, D108–D110.
- Mazor G, Levin L, Picard D, Ahmadov U, Caren H, Borkhardt A, Reifengerger G, Lepruvier G, Remke M and Rotblat B (2019) The lncRNA TP73-AS1 is linked to aggressiveness in glioblastoma and promotes temozolomide resistance in glioblastoma cancer stem cells. *Cell Death Dis* **10**, 246.
- Miao Y, Ajami NE, Huang TS, Lin FM, Lou CH, Wang YT, Li S, Kang J, Munkacsy H, Maurya MR *et al.* (2018) Enhancer-associated long non-coding RNA LEENE regulates endothelial nitric oxide synthase and endothelial function. *Nat Commun* **9**, 292.
- Nakaya T, Ogawa S, Manabe I, Tanaka M, Sanada M, Sato T, Taketo MM, Nakao K, Clevers H, Fukayama M *et al.* (2014) KLF5 regulates the integrity and oncogenicity of intestinal stem cells. *Cancer Res* **74**, 2882–2891.
- Nandan MO, McConnell BB, Ghaleb AM, Bialkowska AB, Sheng H, Shao J, Babbitt BA, Robine S and Yang VW (2008) Kruppel-like factor 5 mediates cellular transformation during oncogenic KRAS-induced intestinal tumorigenesis. *Gastroenterology* **134**, 120–130.
- Paci P, Colombo T and Farina L (2014) Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. *BMC Syst Biol* **8**, 83.
- Peng L, Jiang B, Yuan X, Qiu Y, Peng J, Huang Y, Zhang C, Zhang Y, Lin Z, Li J *et al.* (2019) Super-enhancer-associated long noncoding RNA HCCL5 is activated by ZEB1 and promotes the malignancy of hepatocellular carcinoma. *Cancer Res* **79**, 572–584.
- Poliseno L, Salmena L, Zhang J, Carver B, Haveman WJ and Pandolfi PP (2010) A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038.
- Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD *et al.* (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nat Biotechnol* **29**, 742–749.
- Prensner JR, Iyer MK, Sahu A, Asangani IA, Cao Q, Patel L, Vergara IA, Davicioni E, Erho N, Ghadessi M *et al.* (2013) The long noncoding RNA SCHLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. *Nat Genet* **45**, 1392–1398.
- Qian FC, Li XC, Guo JC, Zhao JM, Li YY, Tang ZD, Zhou LW, Zhang J, Bai XF, Jiang Y *et al.* (2019) SEanalysis: a web tool for super-enhancer associated regulatory analysis. *Nucleic Acids Res* **47**, W248–W255.
- Qin J, Zhou Z, Chen W, Wang C, Zhang H, Ge G, Shao M, You D, Fan Z, Xia H *et al.* (2015) BAP1 promotes

- breast cancer cell proliferation and metastasis by deubiquitinating KLF5. *Nat Commun* **6**, 8471.
- Quinlan AR and Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842.
- Robasky K and Bulyk ML (2011) UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **39**, D124–D128.
- Salmena L, Poliseno L, Tay Y, Kats L and Pandolfi PP (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* **146**, 353–358.
- Schmitt AM and Chang HY (2016) Long noncoding RNAs in cancer pathways. *Cancer Cell* **29**, 452–463.
- Shan Y, Ma J, Pan Y, Hu J, Liu B and Jia L (2018) LncRNA SNHG7 sponges miR-216b to promote proliferation and liver metastasis of colorectal cancer through upregulating GALNT1. *Cell Death Dis* **9**, 722.
- Shen Y, Wang Z, Loo LW, Ni Y, Jia W, Fei P, Risch HA, Katsaros D and Yu H (2015) LINC00472 expression is regulated by promoter methylation and associated with disease-free survival in patients with grade 2 breast cancer. *Breast Cancer Res Treat* **154**, 473–482.
- Song Y, Li L, Ou Y, Gao Z, Li E, Li X, Zhang W, Wang J, Xu L, Zhou Y *et al.* (2014) Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91–95.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102**, 15545–15550.
- Sun Y, Wei G, Luo H, Wu W, Skogerbo G, Luo J and Chen R (2017) The long noncoding RNA SNHG1 promotes tumor growth through regulating transcription of both local and distal genes. *Oncogene* **36**, 6774–6783.
- Takahashi K and Yamanaka S (2006) Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676.
- Tang Z, Li X, Zhao J, Qian F, Feng C, Li Y, Zhang J, Jiang Y, Yang Y, Wang Q *et al.* (2019) TRCirc: a resource for transcriptional regulation information of circRNAs. *Brief Bioinform* **20**, 2327–2333.
- Tay Y, Rinn J and Pandolfi PP (2014) The multilayered complexity of ceRNA crosstalk and competition. *Nature* **505**, 344–352.
- Vance KW and Ponting CP (2014) Transcriptional regulatory functions of nuclear long noncoding RNAs. *Trends Genet* **30**, 348–355.
- Wang P, Ning S, Zhang Y, Li R, Ye J, Zhao Z, Zhi H, Wang T, Guo Z and Li X (2015) Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. *Nucleic Acids Res* **43**, 3478–3489.
- Wei GH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR *et al.* (2010) Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* **29**, 2147–2160.
- Wei Y, Zhang S, Shang S, Zhang B, Li S, Wang X, Wang F, Su J, Wu Q, Liu H *et al.* (2016) SEA: a super-enhancer archive. *Nucleic Acids Res* **44**, D172–D179.
- Whyte WA, Orlando DA, Hnisz D, Abraham BJ, Lin CY, Kagey MH, Rahl PB, Lee TI and Young RA (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319.
- Wood CD, Carvell T, Gunnell A, Ojienyi OO, Osborne C and West MJ (2018) Enhancer control of microRNA miR-155 expression in Epstein-Barr virus-infected B cells. *J Virol* **92**, e00716-18.
- Xiang JF, Yin QF, Chen T, Zhang Y, Zhang XO, Wu Z, Zhang S, Wang HB, Ge J, Lu X *et al.* (2014) Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res* **24**, 513–531.
- Xie JJ, Jiang YY, Jiang Y, Li CQ, Lim MC, An O, Mayakonda A, Ding LW, Long L, Sun C *et al.* (2018) Super-enhancer-driven long non-coding RNA LINC01503, regulated by TP63, is over-expressed and oncogenic in squamous cell carcinoma. *Gastroenterology* **154**, 2137–2151, e2131.
- Xu J, Li Y, Lu J, Pan T, Ding N, Wang Z, Shao T, Zhang J, Wang L and Li X (2015) The mRNA related ceRNA-ceRNA landscape and significance across 20 major cancer types. *Nucleic Acids Res* **43**, 8169–8182.
- Xu M, Chen X, Lin K, Zeng K, Liu X, Pan B, Xu X, Xu T, Hu X, Sun L *et al.* (2018) The long noncoding RNA SNHG1 regulates colorectal cancer cell growth through interactions with EZH2 and miR-154-5p. *Mol Cancer* **17**, 141.
- Yang L, Parkin DM, Ferlay J, Li L and Chen Y (2005) Estimates of cancer incidence in China for 2000 and projections for 2005. *Cancer Epidemiol Biomarkers Prev* **14**, 243–250.
- Zeng FM, Wang XN, Shi HS, Xie JJ, Du ZP, Liao LD, Nie PJ, Xu LY and Li EM (2017) Fascin phosphorylation sites combine to regulate esophageal squamous cancer cell behavior. *Amino Acids* **49**, 943–955.
- Zhang E, Han L, Yin D, He X, Hong L, Si X, Qiu M, Xu T, De W, Xu L *et al.* (2017) H3K27 acetylation activated-long non-coding RNA CCAT1 affects cell proliferation and migration by regulating SPRY4 and HOXB13 expression in esophageal squamous cell carcinoma. *Nucleic Acids Res* **45**, 3086–3101.
- Zhang E, He X, Zhang C, Su J, Lu X, Si X, Chen J, Yin D, Han L and De W (2018a) A novel long noncoding

RNA HOXC-AS3 mediates tumorigenesis of gastric cancer by binding to YBX1. *Genome Biol* **19**, 154.

- Zhang X, Choi PS, Francis JM, Gao GF, Campbell JD, Ramachandran A, Mitsuishi Y, Ha G, Shih J, Vazquez F *et al.* (2018b) Somatic superenhancer duplications and hotspot mutations lead to oncogenic activation of the KLF5 Transcription factor. *Cancer Discov* **8**, 108–125.
- Zhang XD, Huang GW, Xie YH, He JZ, Guo JC, Xu XE, Liao LD, Xie YM, Song YM, Li EM *et al.* (2018c) The interaction of lncRNA EZR-AS1 with SMYD3 maintains overexpression of EZR in ESCC cells. *Nucleic Acids Res* **46**, 1793–1809.
- Zhang Y, Liu T, Meyer CA, Eickhout J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137.
- Zhang Y, Pitchiaya S, Cieslik M, Niknafs YS, Tien JC, Hosono Y, Iyer MK, Yazdani S, Subramaniam S, Shukla SK *et al.* (2018d) Analysis of the androgen receptor-regulated lncRNA landscape identifies a role for ARLNC1 in prostate cancer progression. *Nat Genet* **50**, 814–824.
- Zhao P, Dai M, Chen W and Li N (2010) Cancer trends in China. *Jpn J Clin Oncol* **40**, 281–285.
- Zhou M, Wang X, Shi H, Cheng L, Wang Z, Zhao H, Yang L and Sun J (2016) Characterization of long non-coding RNA-associated ceRNA network to reveal potential prognostic lncRNA biomarkers in human ovarian cancer. *Oncotarget* **7**, 12598–12611.
- Zhu L, Zhang X, Fu X, Li Z, Sun Z, Wu J, Wang X, Wang F, Li X, Niu S *et al.* (2019) c-Myc mediated upregulation of long noncoding RNA SNHG12 regulates proliferation and drug sensitivity in natural killer/T-cell lymphoma. *J Cell Biochem* **120**, 12628–12637.

## Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**Fig. S1.** Identification ce-lncRNAs in ESCC. Venn diagram showing the overlap of lncRNAs (left), PCGs (middle) and ceRNA pairs (right) between both ESCC datasets (GSE53625 (n = 119) and SRP064894). Traditionally, a lncRNA-PCG pair sharing miRNAs will be defined as functional ceRNA relationship based on the following criteria: (1) Expression correlation of lncRNA-PCG pair (Cor); (2) Shared miRNAs (Hyper); (3) Differentially expression level of lncRNAs/PCGs (SAM or Limma). We used six different combinations of them to identify ceRNAs, including (A) SAM(0.01)+Cor. (B) Limma(0.01)+Cor. (C) SAM(0.05)+Cor. (D) SAM(0.01)+Hyper+Cor. (E)

Limma(0.01)+Hype+Cor. (F) SAM(0.05)+Hype+-Cor.

**Fig. S2.** The ceRNA network controls broad cancer associated hallmarks. (A) The cancer hallmarks corresponding to GO terms. The colors corresponds to different cancer hallmarks. (B) The cancer hallmarks related GO terms enriched by ce-lncRNA-related PCGs in the ceRNA network. The colors of bars corresponds to different cancer hallmarks. (C) Number of significantly enriched ce-lncRNAs for each cancer hallmark.

**Fig. S3.** H3K27ac ChIP-seq signals at the SE-associated lncRNA locus in six ESCC cell lines.

**Fig. S4.** Box plots of prognostic value associated with the SE-associated ce-lncRNAs, TE-associated ce-lncRNAs, as well as other random pairs.

**Fig. S5.** Inhibition of THZ1 for SE/TE-associated ce-lncRNAs. (A) Boxplot of expression of SE/TE-associated ce-lncRNAs upon either DMSO or THZ1 (50nM) at indicated time points in KYSE510 cells. SE/TE-associated ce-lncRNAs were identified in KYSE510 or in all other five cell lines. (B) Boxplot of expression of SE/TE-associated ce-lncRNAs upon either DMSO or THZ1 (50nM) at indicated time points, which involved in KYSE510 or TE7 cell lines. \*  $P < 0.05$ , \*\*  $P < 0.01$ , \*\*\*  $P < 0.001$ .  $P$  values were determined using Wilcoxon rank-sum test.

**Fig. S6.** Box plots of log2 fold changes in expression of lncRNA associated ce-PCGs in KYS510 and TE7 cells treated with either DMSO or THZ1 (50nM) at indicated time points. \*  $P < 0.05$ , \*\*  $P < 0.01$ , \*\*\*  $P < 0.001$ .  $P$  values were determined using Wilcoxon rank-sum test.

**Fig. S7.** Western blotting detection for the expression of KLF5 and TCF3 in three ESCC cell lines (KYSE150, KYSE510 and TE3) upon silencing of KLF5 and TCF3 by using different siRNA.

**Fig. S8.** The downstream pathway analysis of LINC00094. (A) The pathways significantly enriched by LINC00094-related PCGs in the ceRNA network. Enrichment significance was performed by the iSubpathwayMiner software package using hypergeometric test. The pathways with FDR corrected  $P < 0.05$  were considered as significant. Number of (\*) represents the annotated gene number in the corresponding pathway. (B) LINC00094-related PCGs that were annotated to the ‘PI3K-Akt signaling’ pathways. (C) The ‘PI3K-Akt signaling’ pathways where LINC00094-related PCGs were annotated. The genes (rectangular nodes) mapped by LINC00094-related PCGs were shown with red node labels and borders.

**Table S1.** Clinical and pathological characteristics of patients in four datasets for genome-wide gene expression profiles of ESCC.

**Table S2.** An example of calculating local regulatory direction consistency of a potential lncRNA-PCG ceRNA pair.

**Table S3.** The overlap and similarity of ceRNA pairs and ce-lncRNAs identified in two ESCC datasets.

**Table S4.** siRNA target sequences.

**Table S5.** Primers used in this study.

**Appendix S1.** The clinical and pathological characteristics of patients in ESCC datasets.

**Appendix S2.** Cancer hallmarks and their associated genes based on 31 GO terms.