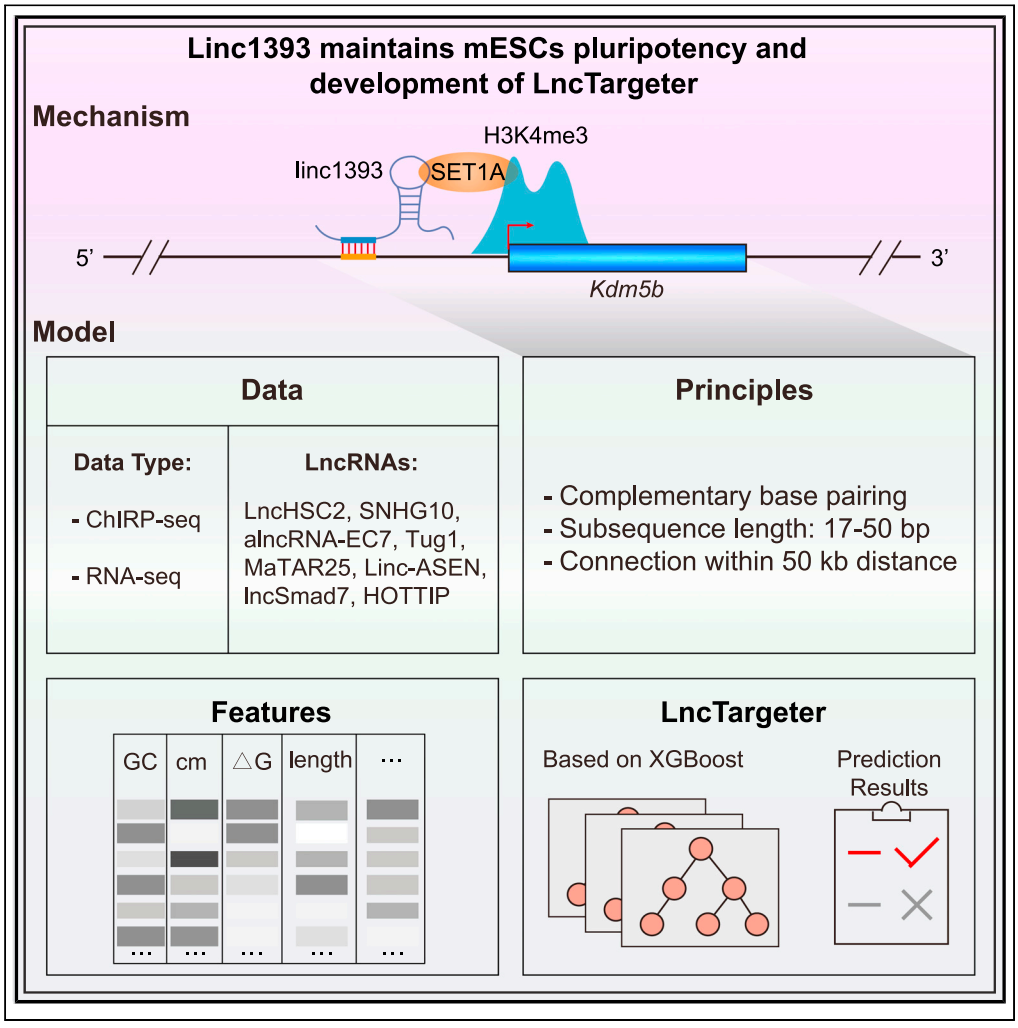
iScience

CellPress
OPEN ACCESS

Article

Network characterization linc1393 in the maintenance of pluripotency provides the principles for lncRNA targets prediction



Weibo Hou, Ming Zong, Qi Zhao, ..., Meiling Gao, Jianzhong Su, Qingran Kong

gaoml@wmu.edu.cn (M.G.)
sujz@wmu.edu.cn (J.S.)
kqr721726@163.com (Q.K.)

Highlights

Nuclear lncRNA linc1393 maintains pluripotency in mESC

Linc1393 establishes H3K4me3 by recruiting SET1A to activate pluripotent genes

LncTargeter is a promising tool for predicting the regulatory networks of lncRNAs

Hou et al., iScience 26, 107469
August 18, 2023
https://doi.org/10.1016/j.isci.2023.107469

iScience

CellPress
OPEN ACCESS

Article

Network characterization linc1393 in the maintenance of pluripotency provides the principles for lncRNA targets prediction

Weibo Hou,[1,5] Ming Zong,[1,4,5] Qi Zhao,[2,5] Xu Yang,[1] Jiaming Zhang,[1,2] Shuanghui Liu,[1] Xuanwen Li,[1] Lijun Chen,[1] Chun Tang,[3] Xinyu Wang,[3] Zhixiong Dong,[1] Meiling Gao,[3,*] Jianzhong Su,[2,3,*] and Qingran Kong[1,6,*]

SUMMARY

Long non-coding RNAs (lncRNAs) have been implicated in diverse biological processes. However, the functional mechanisms have not yet been fully explored. Characterizing the interactions of lncRNAs with chromatin is central to determining their functions but, due to precise and efficient approaches lacking, our understanding of their functional mechanisms has progressed slowly. In this study, we demonstrate that a nuclear lncRNA linc1393 maintains mouse ESC pluripotency by recruiting SET1A near its binding sites, to establish H3K4me3 status and activate the expression of specific pluripotency-related genes. Moreover, we characterized the principles of lncRNA–chromatin interaction and transcriptional regulation. Accordingly, we developed a computational framework based on the XGBoost model, LncTargeter, to predict the targets of a given lncRNA, and validated its reliability in various cellular contexts. Together, these findings elucidate the roles and mechanisms of lncRNA on pluripotency maintenance, and provide a promising tool for predicting the regulatory networks of lncRNAs.

INTRODUCTION

A larger portion of the genome can produce thousands of non-coding RNAs (ncRNAs).[1] Among ncRNAs, long non-coding RNAs (lncRNAs) are non-protein coding transcripts longer than 200 nucleotides.[2] LncRNAs can be classified into two broad categories based on intracellular localization: cytoplasmic lncRNAs and nuclear lncRNAs.[1,3,4] In general, cytoplasmic lncRNAs regulate gene expression via mRNA degradation or mediation of translational repression at the post-transcriptional level. Meanwhile, most lncRNAs tend to be enriched in the nuclear fraction.[4–6] Nuclear lncRNAs can repress or promote neighboring genes via *cis* activities or distantly located genes (near their binding sites) via *trans* activities and the regulation of chromatin epigenetic status, acting as molecular scaffolds to connect functionally related proteins to precise genomic loci.[6–10] Thus, genome-wide detection of binding sites and target genes is essential to revealing the complex regulatory networks of nuclear lncRNAs.

LncRNA binding sites are focal, sequence-specific, and abundant in the genome.[11,12] Various techniques have been developed to localize lncRNAs on chromatin. These methods include chromatin isolation by RNA purification (ChIRP), capture hybridization analysis of RNA targets (CHART), and RNA affinity purification (RAP-DNA), all of which rely on complementary sequences capturing a specific RNA followed by deep sequencing to identify chromatin targets.[12–15] However, these methods allow for analysis of the binding sites of only a single lncRNA at a time. Furthermore, a strategy for mapping global RNA interactions with DNA by deep sequencing (GRID-seq), which uses a bivalent linker to ligate RNA to DNA *in situ* on fixed nuclei, has been reported, but many less abundant lncRNAs may escape detection and this strategy is not accessible to most users, as it relies on commercial sequencing and complex technological systems.[16] Consequently, lncRNA binding sites, and thus their target genes, require more efficient and accessible identification methods.

Presently, lncRNAs are known to play diverse functional roles in various biological pathways, including disease progression and developmental regulation, and are particularly true in embryonic stem cells (ESCs), where subtle changes in lncRNA expression can lead to the loss of pluripotency.[4,10,17–23] For example, lncRNA Panct1 can maintain mouse ESC (mESC) identity by regulating TOBF1 recruitment to Oct-Sox

[1]Oujiang Laboratory, Zhejiang Provincial Key Laboratory of Medical Genetics, Key Laboratory of Laboratory Medicine, Ministry of Education, School of Laboratory Medicine and Life Sciences, Wenzhou Medical University, Wenzhou, Zhejiang, China

[2]Oujiang Laboratory, Zhejiang Lab for Regenerative Medicine, Vision and Brain Health, Wenzhou Medical University, Wenzhou, Zhejiang, China

[3]School of Optometry and Ophthalmology and Eye Hospital, Wenzhou Medical University, Wenzhou, Zhejiang, China

[4]College of Life Science, Northeast Agricultural University, Harbin, Heilongjiang, China

[5]These authors contributed equally

[6]Lead contact

*Correspondence: gaoml@wmu.edu.cn (M.G.), sujz@wmu.edu.cn (J.S.), kqr721726@163.com (Q.K.)

https://doi.org/10.1016/j.isci.2023.107469

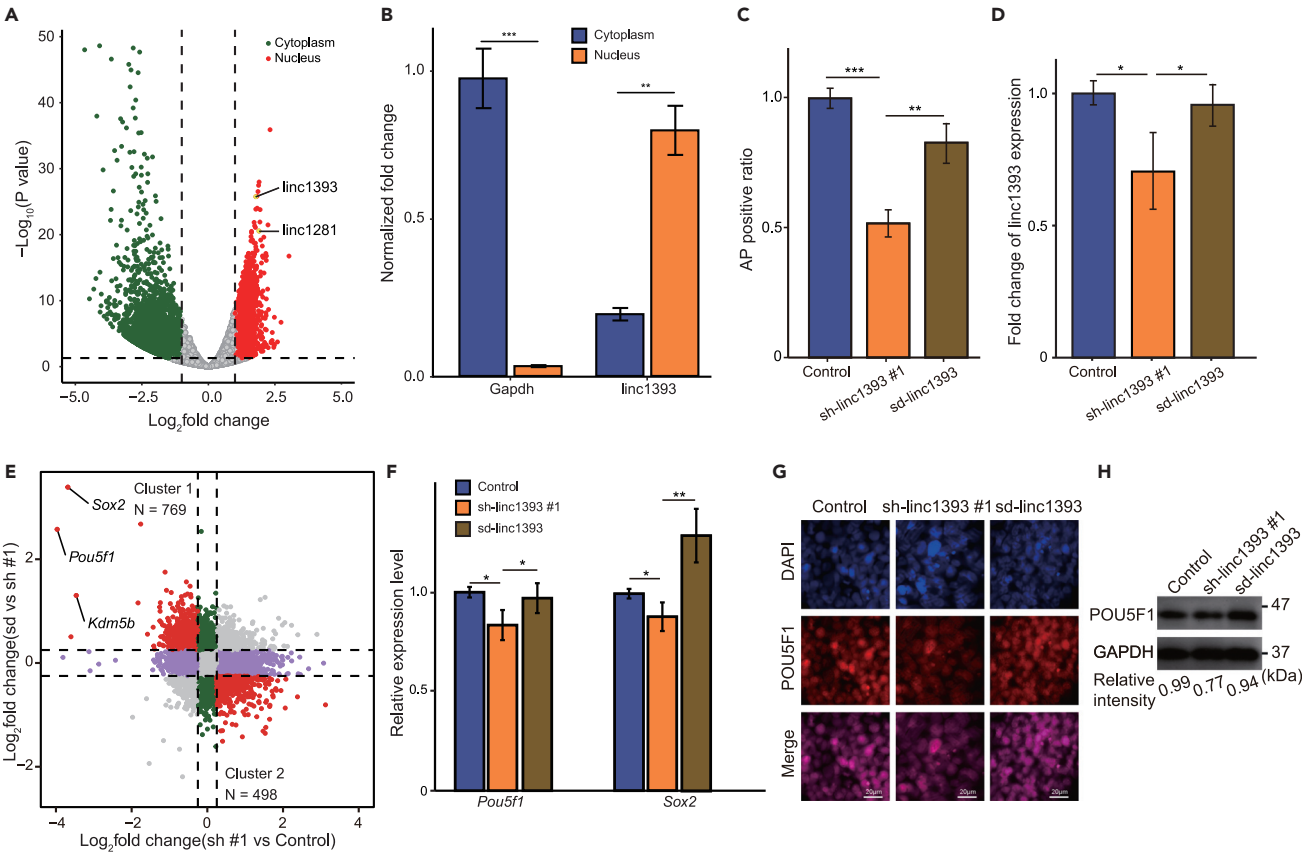


Figure 1. The nuclear lncRNA linc1393 maintains mESC pluripotency

(A) RNA-seq analysis of lncRNAs expressed in the nucleus and cytoplasm of mESCs, and volcano plot showing differentially expressed lncRNAs. Red and green dots represent nucleus-enriched (FC > 2) and cytoplasm-enriched (FC < −2) lncRNAs with $p < 0.05$.

(B) qPCR analysis of linc1393 expression in the nucleus and cytoplasm of mESCs. Data are means ± SEM from three independent experiments, $^{**}p < 0.01$, $^{***}p < 0.001$.

(C) AP positive ratios of mESCs in the control, sh-linc1393 and sd-linc1393 groups. Data are mean ± SEM. $^{**}p < 0.01$, $^{***}p < 0.001$.

(D) Linc1393 expression levels in the control, sh-linc1393 and sd-linc1393 groups. Data are mean ± SEM. $^{*}p < 0.05$.

(E) Scatterplot showing $\log_2$-fold changes in the expression levels of all genes following linc1393 knockdown (x axis) or linc1393 rescue (y axis). Red dots indicate genes regulated by linc1393.

(F) qPCR analysis of *Pou5f1* and *Sox2* expression levels in the control, sh-linc1393 and sd-linc1393 groups. Data are mean ± SEM. $^{*}p < 0.05$, $^{***}p < 0.01$.

(G) Immunostaining analysis of OCT4 expression in the control, sh-linc1393 and sd-linc1393 groups. Scale bars: 20 μm.

(H) Western blot of OCT4 expression in the control, sh-linc1393 and sd-linc1393 groups.

motifs in the early G1 phase.[24] However, the functional mechanisms through which many lncRNAs contribute to pluripotency have not yet been fully explored. In this study, we screened the nuclear lncRNAs in mESCs, and demonstrated that linc1393 maintains mESC pluripotency through recruitment of SET1A near its binding sites, to establish the H3K4me3 modification and activate the expression of specific pluripotent genes. Furthermore, we developed a computational framework based on the XGBoost model, LncTargeter, to predict the binding sites and target genes of a given lncRNA.

RESULTS

Linc1393 knockdown affects mESC pluripotency

First, we employed publicly available nuclear and cytoplasmic RNA-seq data (GEO: GSE58757) from mESCs to identify lncRNAs highly enriched in the nucleus (Figure 1A; FC > 2, $p < 0.05$); linc1393 and linc1281 with remarkably higher expression in the nucleus compared to the cytoplasm were selected.[25] Depletion of linc1393 or linc1281 using specific small interfering RNAs (siRNAs) in mESCs resulted in a significant decrease of the alkaline phosphatase (AP) positive ratio, compared to the cells transfected with control siRNA, especially for linc1393 (Figures S1A and S1B; $p < 0.001$). Linc1393 knockdown led to a

decrease of *Pou5f1* mRNA in mESCs by quantitative polymerase chain reaction (qPCR) (Figure S1C; $p < 0.05$), suggesting that linc1393 may be important for mESC pluripotency.[26] Then, we performed cellular fractionation followed by RNA extraction and qPCR analysis to confirm the nuclear localization of linc1393 (Figure 1B; $p < 0.01$). Furthermore, mESCs were stably transfected with control short hairpin RNA (shRNA), two shRNAs targeting linc1393 (sh-linc1393) or sh-linc1393 with targeting site deleted (sd-linc1393). The knockdown leaded to no significant cell apoptosis determined by flow cytometry (Figure S1D), but mESC differentiation (Figure S1E). We also found that the AP positive ratio of the sh-linc1393 cell line was significantly reduced compared with the control, but markedly rescued in the sd-linc1393 group (Figures 1C and S1F; $p < 0.01$). To clarify the transcriptome regulated by linc1393 in mESCs, RNA sequencing (RNA-seq) was performed on mESCs transfected with control shRNA, sh-linc1393 or sd-linc1393, and unsupervised hierarchical clustering could separate the groups distinctly (Figures S1G and S1H). Downregulated expression of linc1393 due to sh-linc1393, and upregulated expression due to sd-linc1393, was verified (Figures 1D and S1I). Differential expression analysis revealed that the expression levels of 769 genes (Cluster 1), including *Pou5f1* and *Sox2,* were decreased by sh-linc1393 and rescued by sd-linc1393, and also the expression levels of 498 genes (Cluster 2) were upregulated by sh-linc1393 and downregulated by sd-linc1393 (Figure 1E; Table S1). The decrease and rescue of *Pou5f1* and *Sox2* expressions were also found by another sh-linc1393 using RNA-seq data (Figure S1J; Table S1). The decrease and rescue of the expression of *Pou5f1* and *Sox2* were confirmed by qPCR (Figures 1F and S1K; $p < 0.05$), and immunostaining and Western blot were also verified the expression changes of *Pou5f1* in protein level (Figures 1G and 1H). In addition, we further checked the role of linc1393 in another mESC line cultured in LIF/KSR without 2i. RNA-seq data also showed the downregulated expressions of pluripotency-related genes, such as *Pou5f1* and *Sox2* (Figures S1L and S1M; Table S1). Collectively, these results indicate that linc1393 is highly enriched in the nucleus of mESCs and participates in mESC pluripotency.

Global interactions of linc1393 with chromatin in mESCs

The function of nuclear lncRNA involves interaction with chromatin.[7,27] Thus, we mapped the genome-wide binding sites of linc1393 using ChIRP-seq to establish interaction networks of linc1393. In total, 11,429 binding sites were detected, which were mainly enriched in promoters, indicating a role of linc1393 in the regulation of gene transcription (Figures 2A and 2B; Table S2). To explore the principles of lncRNA–chromatin interaction, the linc1393 sequence was segmented into fixed subsequences based on a sliding window of 10–100 bp with a step size of 1 bp, and by sequence alignment analysis, we found that 79% of the binding sites identified using ChIRP-seq are complementary to the subsequences (Figure 2C), and that the lengths of most subsequences mediating interactions were 17–50 bp (Figure 2D). Furthermore, motif analysis of the binding sites revealed eight highly significantly enriched DNA motifs, and we found eight core subsequences, which were identified by mapping of ChIRP-seq reads onto the linc1393 sequence, were complementary to those motifs (Figures 2E and 2F), suggesting that these core subsequences may mainly mediate the interaction of nuclear lncRNA with chromatin by direct base pairing. Together, these results reveal that complementary base pairing may be the primary mechanism driving the global interactions of nuclear lncRNA with chromatin.

Regulatory network of linc1393 in mESC pluripotency

To clarify the regulatory network of linc1393 target genes via its binding sites, we calculated the distance from the transcription start site (TSS) of each gene in Cluster 1 and Cluster 2 to the nearest binding site of linc1393, and found that the distances were significantly shorter for Cluster 1 than Cluster 2 (Figure 3A; $p < 0.001$). Furthermore, analysis of public high-resolution High-throughput Chromosome Conformation Capture (Hi-C) data for mESCs (GEO: GSE96107) showed that Cluster 1 genes had stronger interactions with their nearest binding sites than Cluster 2 genes (Figure 3B; $p < 0.001$). The expression of Cluster 1 genes was downregulated with the knockdown of linc1393, implying that linc1393 may promote the activities of genes near its binding sites in mESCs. Further analysis revealed that the expression levels of Cluster 1 genes, located within 50 kb of the nearest binding sites of linc1393, were more significantly affected than those genes 50 kb away from the binding sites, but this pattern was not observed for Cluster 2 genes (Figures 3C and S2A; $p < 0.001$), indicating that nuclear lncRNAs are connected to their directly regulated genes through binding sites located within 50 kb. Therefore, we identified 528 genes located within 50 kb of the nearest binding site in Cluster 1 as putative direct target genes of linc1393. To show the potential interaction between linc1393 and its predicted targets, we performed aggregate peak analysis (APA) using Hi-C data from mESC, and the strong interactions were verified (Figure S2B; P2LL = 1.397). To test the principles, we generated a mESCs line stably transfected with sh-linc1393 and targeting site-depleted linc1393

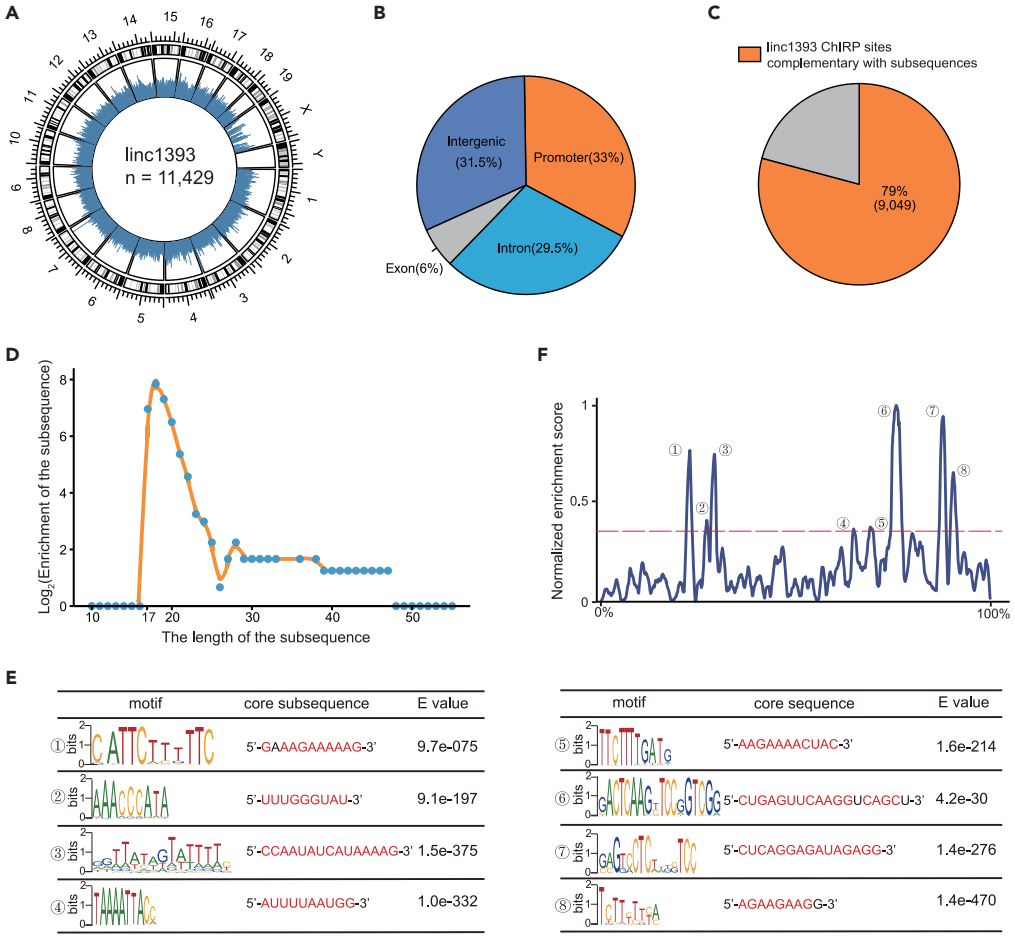


Figure 2. Interactions of linc1393 with chromatin through core subsequence-mediated complementary base pairing

(A) Circos plot showing the binding sites of linc1393 across the genome by ChIRP-seq.

(B) Pie chart illustrating the genomic distribution of binding sites of linc1393 relative to RefSeq annotations. Promoters represent regions ±0.5 kb around the TSS.

(C) The percentage of linc1393 ChIRP-seq reads complementary to various subsequences.

(D) Line plot showing the distribution of the length of the subsequence complementary to the binding sites of linc1393.

(E) Motif analysis of the binding sites of linc1393 revealing eight highly significantly enriched DNA motifs complementary to subsequences of linc1393.

(F) Identification of core subsequences by mapping ChIRP-seq reads onto linc1393.

with deletion of core subsequence 3 (dd-linc1393), which was predicted to directly regulate *Kdm5b*, an evidenced pluripotency-associated gene important for productive transcriptional elongation of pluripotency-associated genes, including *Tet1*, *Fubp1* and *Mcm4*, through prevention of cryptic transcription.[28] The Hi-C data showed a strong interaction between *Kdm5b* and the binding site mediated by core subsequence 3 (Figure 3D). Definitely, RNA-seq analysis demonstrated that a small subset of the target genes of linc1393 was not rescued in the dd-linc1393 group, including *Kdm5b* (Figures 3E and S2C). Immunostaining further confirmed that the expression of KDM5B was decreased in sh-linc1393 mESCs, and rescued by sd-linc1393, but not dd-linc1393 (Figure 3F; $p < 0.001$), showing that nuclear lncRNA could regulate the expression of specific genes located near its binding sites by core subsequence-mediated complementary base pairing. In accordance with the previous study,[28] we found significant decreases in full-length (intron-spanning) transcriptions of *Tet1*, *Fubp1*, and *Mcm4* when the expression of *Kdm5b* was downregulated in the sh-linc1393 and dd-linc1393 groups, which were rescued in the sd-linc1393 group (Figures 3G, S2D, and S2E). Taken together, these observations reveal that linc1393 could maintain mESC pluripotency through activation of the specific pluripotency-associated genes nearby its binding sites.

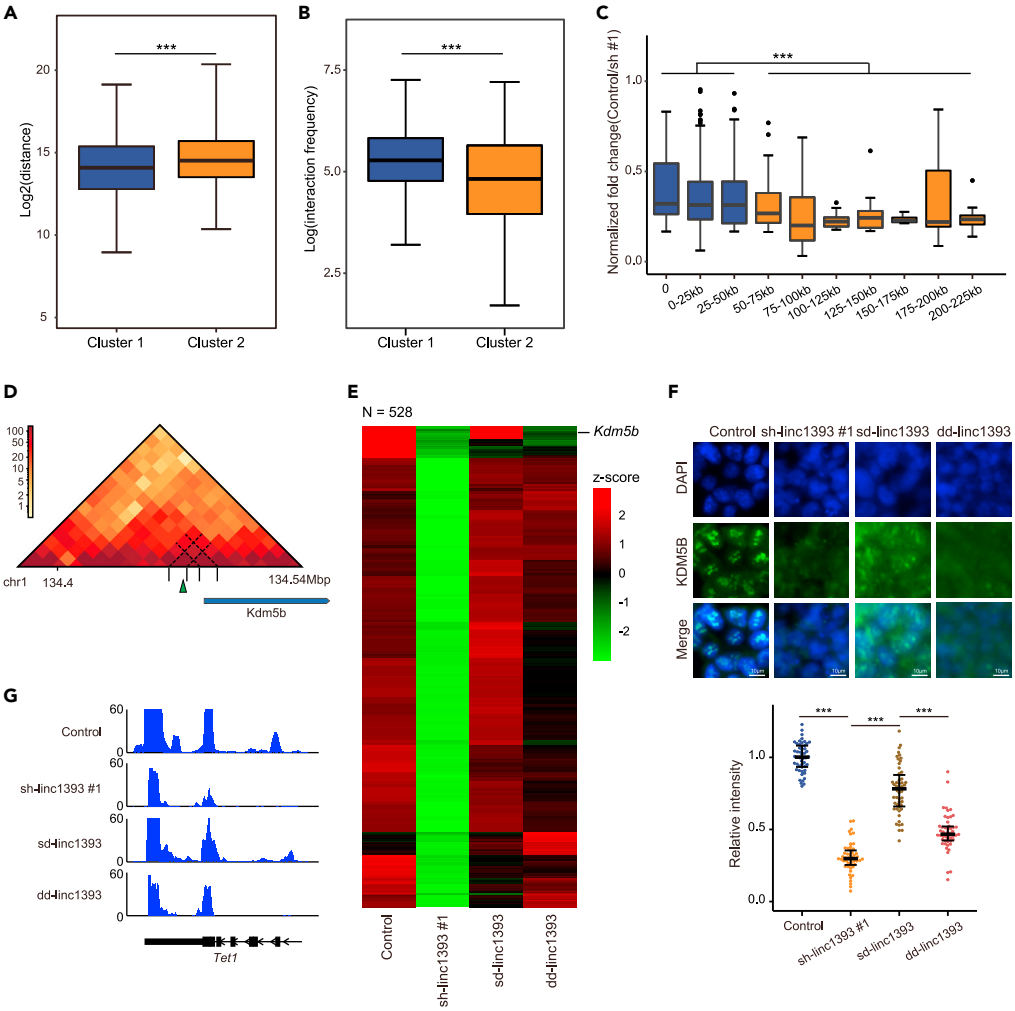


Figure 3. Linc1393 maintains mESC pluripotency by activating the specific pluripotency-associated genes nearby its binding sites

(A and B) Boxplots illustrating the distance (A) and interaction frequency (B) between Cluster1 or Cluster2 genes and their corresponding linc1393 binding sites. ***$p < 0.001$.

(C) Boxplot indicating fold changes in the expression of Cluster 1 genes with distance to the nearest binding site. ***$p < 0.001$.

(D) High-resolution Hi-C analysis of mESCs at the *Kdm5b* locus. Green triangle indicates the linc1393 binding site.

(E) Heatmap showing the expression levels of direct target genes of linc1393 in the control, sh-linc1393, sd-linc1393 and dd-linc1393 groups. Mean values of two biological replicates were scaled and are represented as Z scores.

(F) Immunostaining and quantification analysis of KDM5B in control, sh-linc1393, sd-linc1393 and dd-linc1393 groups. One representative image is shown. ***$p < 0.001$. Scale bars: 10 μm.

(G) Genome browser view of RNA-seq signals at the *Tet1* locus in control, sh-linc1393, sd-linc1393 and dd-linc1393 groups.

Genome-wide profile of H3K4me3 status established through the interaction of linc1393 with SET1A in mESCs

To investigate the mechanism through which linc1393 affects the activation of specific pluripotency-associated genes, we performed pull-down assays with biotinylated linc1393, followed by mass spectrometry (MS) to search for potential linc1393-interacting proteins. A specific band was observed in the linc1393 pull-down group and 77 proteins were identified (Figure 4A; Table S3). Gene Ontology (GO) analysis of the candidate linc1393-interacting proteins showed that they were enriched in the process related to H3K4-specific methyltransferase activity (Figure 4B), including the histone methyltransferases SET1A and KMT2A. RPISeq prediction[29] showed that SET1A had a stronger interaction with linc1393 than KMT2A (Figure 4C). SET1B is a paralog

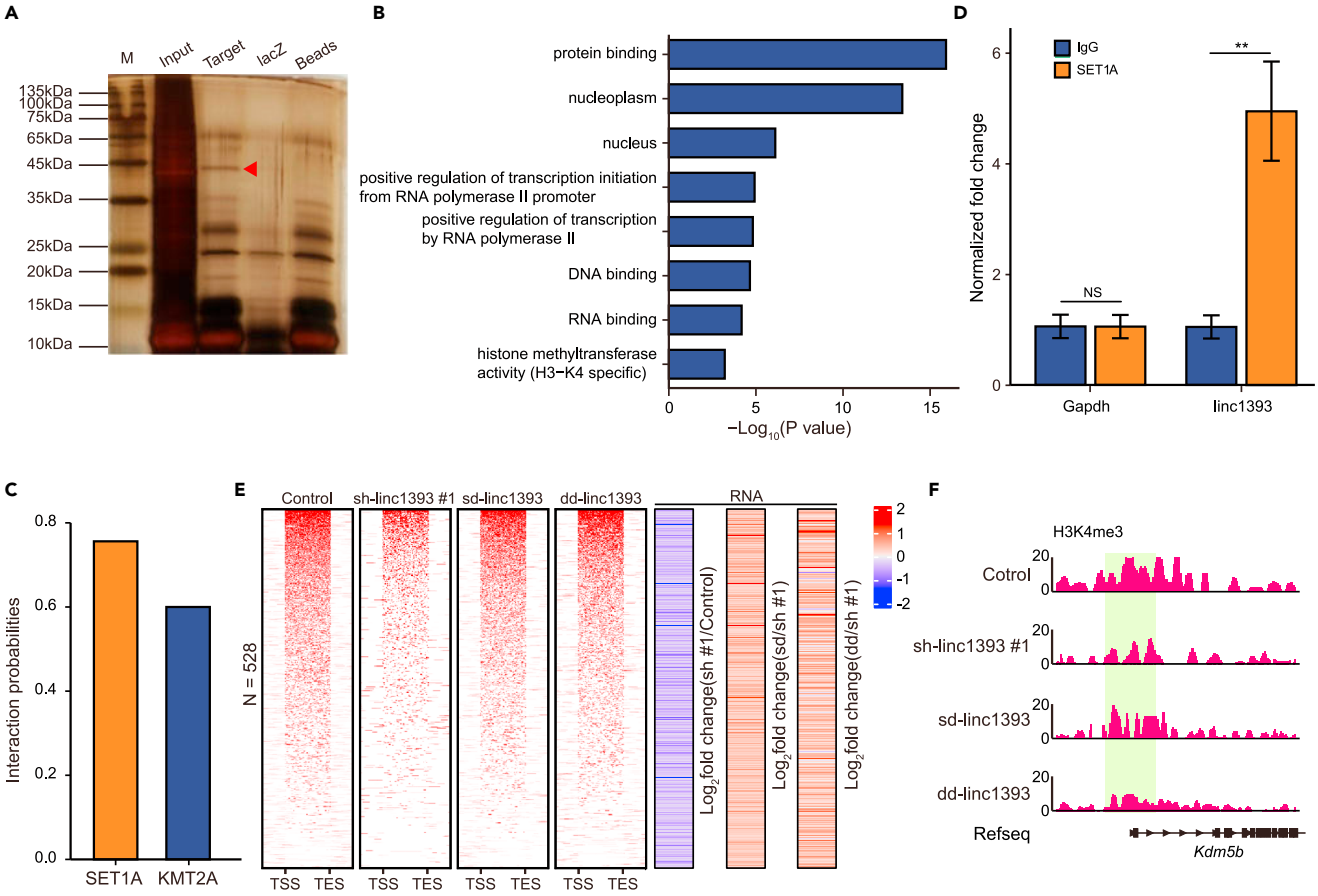


Figure 4. Linc1393 establishes H3K4me3 at the specific promoters of pluripotency-associated genes by interaction with SET1A

(A) RNA pull-down assays with biotinylated linc1393 in mESCs visualized by sliver staining.

(B) Gene Ontology (GO) analysis illustrating the biological processes enriched among potential linc1393-interacting proteins identified in mESCs by LC-MS. p-value derived random sampling of the whole genome combined with Fisher's inverse chi-square method.

(C) RNA–protein interaction prediction via the RPISeq-Platform (SVM classifier).

(D) RNA binding protein immunoprecipitation (RIP) of mESCs. Fold changes in enrichment relative to the IgG control are shown. Error bars indicate the mean ± SEM. **$p < 0.01$.

(E) Heatmap (left) showing H3K4me3 signals ranked based on their relative changes in the control, sh-linc1393, sd-linc1393 and dd-linc1393 groups. Heatmap (right) comparing gene expression fold changes between the sh-linc1393 and control, sd-linc1393 and sh-linc1393, and dd-linc1393 and sh-linc1393 groups.

(F) Genome browser view of H3K4me3 signals at the *Kdm5b* locus in the control, sh-linc1393, sd-linc1393 and dd-linc1393 groups.

of SET1A and also directs H3K4 methyltransferase. Multiple sequence alignment analysis showed the linc1393-interacting protein identified by MS was SET1A, but not SET1B (Figure S3A). And, the interaction of SET1B with linc1393 was weaker, comparing to SET1A (Figure S3B). Further, the specificity of the linc1393-SET1A interaction was verified by an RNA immunoprecipitation (RIP) assay in mESCs (Figure 4D; $p < 0.01$), which suggested that linc1393 may recruit SET1A to activate the expression of specific genes by establishing active histone modifications. H3K4me3 is the key histone marker of active promoters.[30] Thus, we used Cleavage Under Targets and Tagmentation (CUT&Tag) to generate genome-wide H3K4me3 maps for mESCs in the control, sh-linc1393, sd-linc1393 and dd-linc1393 groups. Changes in the expression of direct target genes of linc1393 and their associated H3K4me3 signals were strongly correlated in the control, sh-linc1393, sd-linc1393 and dd-linc1393 groups (Figure 4E; N = 528). Notably, the enrichments of H3K4me3 were markedly decreased with linc1393 knockdown at the loci of target genes of linc1393, but were recovered in the sd-linc1393 group (Figure 4E). In the dd-linc1393 group, most of the loss of the H3K4me3 enrichments caused by sd-linc1393 could be rescued (Figure 4E), except for the enrichment at the promoter region of *Kdm5b* (Figure 4F). We also found *Set1a* was overexpressed in the sh-linc1393 mESC (Figure S3C). In order to check the role of *Set1a* in the linc1393-knockdown mESC background, we overexpressed *Set1a* in the sh-linc1393 mESC (Figure S3C).

Differential expression analysis revealed that the expressions of pluripotency-related genes were not significantly rescued, such as *Pou5f1* and *Sox2* (Figures S3D and S3E; Table S5), and verified by qPCR (Figure S3F), suggesting SET1A regulates pluripotency in a linc1393-dependent manner. These results demonstrate that core subsequence-mediated complementary base pairing allows linc1393 to recruit SET1A to its occupancy locus, establishing H3K4me3 and thus activating the expression of pluripotency-related genes to maintain mESC pluripotency.

Development of LncTargeter for prediction of lncRNA targets

Because nuclear lncRNA functions via its binding sites and target genes, we established a new algorithm that integrates the principles of lncRNA–chromatin interaction and transcriptional regulation to identify the regulatory network of a given nuclear lncRNA. The following principles must be considered: (1) nuclear lncRNA and chromatin interact through core subsequence-mediated complementary base pairing; (2) the lengths of subsequences are generally in the range of 17–50 bp; (3) the connections of nuclear lncRNAs with their target genes through binding sites within 50 kb. Accordingly, we developed a computational strategy, LncTargeter, that integrates these principles to predict the binding sites and target genes of a given lncRNA using ChIRP-seq and RNA-seq data of 8 different lncRNAs in humans and mice obtained from the publicly accessible National Center for Biotechnology Information Gene Expression Omnibus (NCBI GEO) repository (Figures S4A and S4B). The predictive results of the approach used here with 10-fold cross-validation, and XGBoost exhibited the highest performance (Figure 5A; area under the curve [AUC] = 0.808). Moreover, the XGBoost model had the best Acc, Sn, Pr, F1 and Sp (Figure 5B). The importance of individual features was determined using the standard feature importance calculation function. Among the extracted features, the GC content of the predicted binding sequence was the most significant (Figures 5C and 5D; $p < 0.001$). The regulated genes of the eight lncRNAs independently identified through RNA-seq analysis, as well as those within 50 kb of the nearest binding site, were presumed to be target genes and included in the positive dataset. Random genes were included in the negative dataset. These datasets were used to train the models. Notably, XGBoost exhibited the best performance among models (Figures 5E and 5F; AUC = 0.904), and was selected as the ensemble model for this study. Taken together, our evaluation results suggest that LncTargeter can precisely and efficiently predict the binding sites and target genes of a given lncRNA.

Validation and application of LncTargeter

For further validation of LncTargeter, we predicted the binding sites and target genes of linc1281 using LncTargeter, as this lncRNA may impact mESC pluripotency based on the knockdown test results (Figures S1A and S1B), and validated the predictions with its ChIRP-seq data in mESCs. In total, 11,007 binding sites and 339 target genes were predicted (Figure 6A), and for those with scores more than 0.9, over 75% of the predicted binding sites overlapped with those identified from ChIRP-seq data (Figure 6B; Table S4). Additionally, eight core subsequences were predicted to mediate the interactions of linc1281 with chromatin, and the predicted binding sites with high correlations with the binding sites identified by ChIRP-seq were observed (Figure 6C). Analysis of public high-resolution Hi-C data for mESCs (GEO: GSE96107) showed strong interactions between predicted binding sites and corresponding target genes (Figure 6D). Moreover, the target genes were associated with stem cell maintenance and transcriptional regulation of pluripotent stem cells (Figures 6E and 6F), suggesting a regulatory role of linc1281 in mESC pluripotency. We further applied LncTargeter to LIMIT, a cancer immunogenic lncRNA that can activate the guanylate-binding protein (GBP) gene cluster.[31] In this analysis, 14,931 binding sites and 248 target genes were predicted (Figure S5A). Functional enrichment analysis revealed that the predicted target genes were associated with the immune system, immune response and cell cycle, in accordance with the capacity of LIMIT to activate transcription of HSF1 and MHC-I machinery (Figure S5C).[31] Through analysis of Hi-C data from the DLD-1 cell line (GEO: GSE160235), we identified significant and strong interactions between the predicted binding sites and corresponding target genes, including GBP2 (Figures S5B and S5D). To further test the accuracy of LncTargeter, we predicted the targets of LINC01271, a human homologous lncRNA of MaTAR25 in mice, and by alignment analysis, we found 76% of the predicted binding sites were homologous with MaTAR25 binding sites checked by ChIRP-seq. Finally, the MaTAR25-regulated gene TNS1 was also found to strongly interact with a binding site of LINC01271 predicted by LncTargeter (Figure S6). As expected, these results highlight that LncTargeter is powerful for predicting lncRNA regulatory networks.

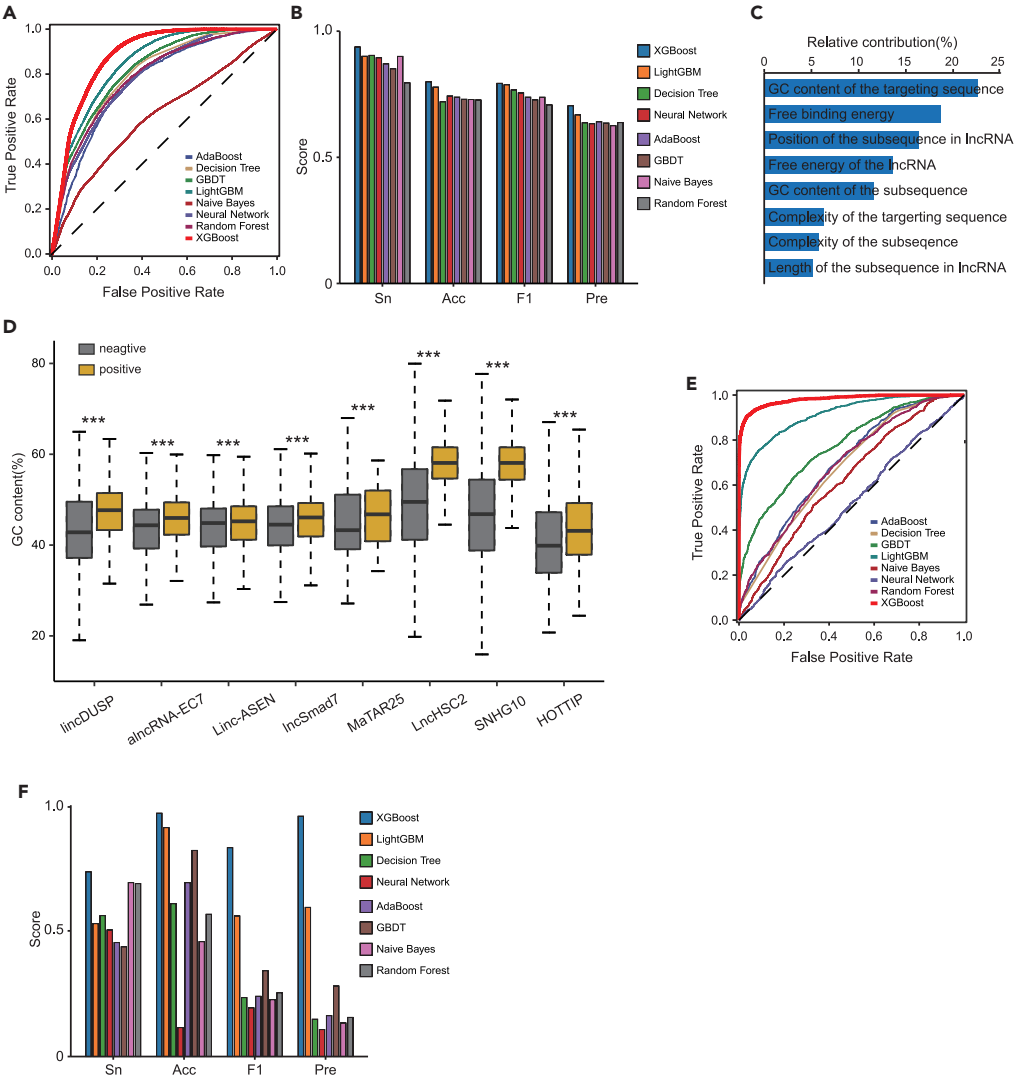


Figure 5. Summary of LncTargeter results

(A and E) ROC curves indicating predictive performance for binding sites (A) and target genes (E) on the eight lncRNAs, shown in decreasing order of AUROC values.

(B and F) Comparison of the results obtained using validation data from GBDT, XGBoost, LightGBM, AdaBoost, RF, LR, NB and DTs for predicting binding sites (B) and target genes (F) in terms of Sn, Acc, F1 and Pre.

(C) Comparison of the importance of features used to construct the LncTargeter model.

(D) Boxplot showing the GC contents of the target sequences of eight lncRNAs used as negative and positive datasets to construct the LncTargeter model. ***p < 0.001.

DISCUSSION

LncRNAs have been demonstrated to act as regulatory factors in concert with chromatin-modifying machinery.[5,6] Although various techniques have been developed to localize lncRNAs on chromatin,[12–14,16] our understanding of the associated regulatory networks has progressed slowly due to the lack of a precise and efficient approach to map genome-wide lncRNA interactions. In the study, by revealing the critical role of linc1393 in the maintenance of pluripotency, we explored the principles of lncRNA-chromatin interactions and transcriptional regulation of target genes, and developed a computational framework LncTargeter to predict lncRNA targets. Deciphering the regulatory network of lncRNA may reveal new avenues of RNA biology.

Here, we present the functional characterization of linc1393, a nuclear lncRNA expressed in mESCs that plays an important role in pluripotency maintenance. We demonstrate that linc1393 acts in coordination with SET1A

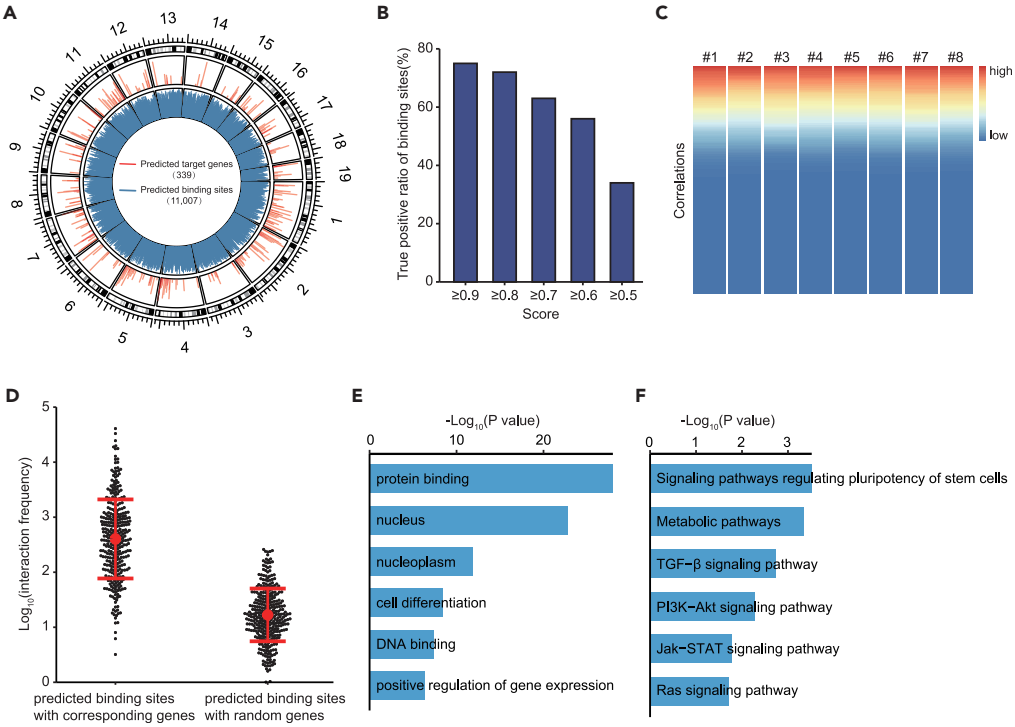


Figure 6. Validation of LncTargeter through prediction of the binding sites and target genes of linc1281

(A) Circos plot showing the predicted binding sites and target genes of linc1281 throughout the genome.

(B) Percentages of predicted linc1281 binding sites that overlap with those identified from ChIRP-seq data at various prediction score thresholds.

(C) Heatmap showing the Jaccard correlations of the binding sites corresponding to eight predicted core subsequences and the binding sites of linc1281 identified by ChIRP-seq.

(D) Dot plot showing interaction intensities between predicted binding sites of linc1281 and target genes obtained from Hi-C data.

(E and F) GO (E) and KEGG (F) analyses of the predicted target genes of linc1281.

to establish H3K4me3 and activate the expression of pluripotency-related genes, revealing the mechanism through which linc1393 activates *Kdm5b* to support productive transcriptional elongation of pluripotency-associated genes. But it's worth noting that KDM5B is H3K4 demethylase. It is possible that a negative feedback loop exists, involving with linc1393, KDM5B, and SET1A, that serves to fine-tune pluripotency regulation. Specifically, the study shows that linc1393 regulates the expression of KDM5B, which may interfere with or abolish the effect of SET1A if both proteins come into contact. Thus, further studies are needed to demonstrate the mechanism of this feedback loop. Interestingly, linc1393 function was observed with overexpression of a mutant version of linc1393, in which the subsequence that binds DNA through base pairing has been scrambled. The role played by linc1393 as a regulatory factor of this specific chromatin locus might reflect a general mechanism underlying the action of lncRNA.

Determining where lncRNAs bind on the genome is crucial to characterize their functions.[11–16,32] lncRNA can interact with a chromatin locus in various ways, including direct interaction with DNA through canonical Watson-Crick base pairing, as a targeting module directing complexes to specific chromatin loci, and indirect interaction mediated by proteins acting as a scaffold for multiple complexes.[5] ChIRP-seq and RNA-seq data have enabled genome-wide analysis of linc1393 occupancy and regulation. Notably, our findings highlighted that (1) complementary base pairing may be the primary mechanism for the interactions between nuclear lncRNA and chromatin mediated by its core subsequence; (2) the lengths of subsequences are mainly in the range of 17–50 bp; and (3) the connections of nuclear lncRNAs with their target genes through binding sites within 50 kb.

Identification of the signatures of linc1393-chromatin interactions is a promising method for establishing lncRNA regulatory networks. However, very few studies have provided models for such analysis. Our

LncTargeter approach, which can precisely and efficiently predict nuclear lncRNA targets, fills that gap. The predicted results of LncTargeter for linc1281 were validated through experimental detection. In addition, a direct regulatory effect of LIMIT on GBP2, an experimentally validated target gene of LIMIT, was predicted using LncTargeter.[31] Therefore, the reliability of LncTargeter has been validated in various cellular contexts, showing that it is a promising tool for predicting the regulatory networks of lncRNAs related to diverse biological processes, such as complex traits and human disease.

Limitations of the study

In the study, we revealed that linc1393 coordinated with SET1A to establish H3K4me3 and activate the expression of pluripotency-related genes, such as *Kdm5b*, to maintain mESC pluripotency. A first limitation of the study lacks the evidence to demonstrate how the core pluripotency genes such as *Oct4* and *Sox2* are activated by linc1393; that may be helpful to understand the critical role of linc1393 in the maintenance of pluripotency. By revealing the functional role of linc1393, we explored the principles of lncRNA-chromatin interactions and transcriptional regulation of target genes, and developed a computational framework LncTargeter to predict lncRNA targets. However, one potential limitation of LncTargeter is the limited number of datasets available to train the model. Thus, expanding the technique to use additional omics data related to lncRNA in future work is of particular importance.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Plasmid construction
  - Cell transfection and alkaline phosphatase staining
  - RNA interference
  - RNA-binding protein immunoprecipitation
  - RNA pull-down and liquid chromatography-mass spectrometry
  - Immunofluorescence staining
  - RNA extraction, reverse transcription and qPCR analysis
  - Feature extraction
  - Classification algorithms and evaluation metrics
  - RNA-seq
  - RNA-seq data analysis
  - CUT&Tag
  - CUT&Tag data processing
  - ChIRP-seq data processing and analysis
  - Hi-C data processing and analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107469.

ACKNOWLEDGMENTS

We thank Ling Shuai from Nankai University for the GT1-1 mESC line. This work was supported by the Basic Public Welfare Research Program of Zhejiang Province (LY22C120001).

AUTHOR CONTRIBUTIONS

M.L.G., J.Z.S., and Q.R.K. conceived and designed the experiments; W.B.H., X.W.L., X.Y.W., and L.J.C. acquired the data and performed the computational analysis; Experiments were performed by M.Z., Q.Z., and

X.Y. with contributions from J.M.Z., C.T., and S.H.L.; Z.X.D. and M.L.G provided experimental support; W.B.H., J.Z.S., and Q.R.K. co-wrote and edited the manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no conflict of interest.

Received: November 4, 2022
Revised: June 7, 2023
Accepted: July 21, 2023
Published: July 26, 2023

REFERENCES

1. St Laurent, G., Wahlestedt, C., and Kapranov, P. (2015). The Landscape of long noncoding RNA classification. Trends Genet. *31*, 239–251. https://doi.org/10.1016/j.tig.2015.03.007.

2. Batista, P.J., and Chang, H.Y. (2013). Long noncoding RNAs: cellular address codes in development and disease. Cell *152*, 1298–1307. https://doi.org/10.1016/j.cell.2013.02.012.

3. Chen, L.L. (2016). Linking Long Noncoding RNA Localization and Function. Trends Biochem. Sci. *41*, 761–772. https://doi.org/10.1016/j.tibs.2016.07.003.

4. Fico, A., Fiorenzano, A., Pascale, E., Patriarca, E.J., and Minchiotti, G. (2019). Long non-coding RNA in stem cell pluripotency and lineage commitment: functions and evolutionary conservation. Cell. Mol. Life Sci. *76*, 1459–1471. https://doi.org/10.1007/s00018-018-3000-z.

5. Statello, L., Guo, C.J., Chen, L.L., and Huarte, M. (2021). Gene regulation by long non-coding RNAs and its biological functions. Nat. Rev. Mol. Cell Biol. *22*, 96–118. https://doi.org/10.1038/s41580-020-00315-9.

6. Gil, N., and Ulitsky, I. (2020). Regulation of gene expression by cis-acting long non-coding RNAs. Nat. Rev. Genet. *21*, 102–117. https://doi.org/10.1038/s41576-019-0184-5.

7. Sun, Q., Hao, Q., and Prasanth, K.V. (2018). Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. Trends Genet. *34*, 142–157. https://doi.org/10.1016/j.tig.2017.11.005.

8. Tsai, M.C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. Science *329*, 689–693. https://doi.org/10.1126/science.1192002.

9. Najafi, S., Tan, S.C., Raee, P., Rahmati, Y., Asemani, Y., Lee, E.H.C., Hushmandi, K., Zarrabi, A., Aref, A.R., Ashrafizadeh, M., et al. (2022). Gene regulation by antisense transcription: A focus on neurological and cancer diseases. Biomed. Pharmacother. *145*, 112265. https://doi.org/10.1016/j.biopha.2021.112265.

10. Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z., and Shen, F. (2019). Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. Int. J. Mol. Sci. *20*, 5573. https://doi.org/10.3390/ijms20225573.

11. Kuo, C.C., Hänzelmann, S., Sentürk Cetin, N., Frank, S., Zajzon, B., Derks, J.P., Akhade, V.S., Ahuja, G., Kanduri, C., Grummt, I., et al. (2019). Detection of RNA-DNA binding sites in long noncoding RNAs. Nucleic Acids Res. *47*, e32. https://doi.org/10.1093/nar/gkz037.

12. Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol. Cell *44*, 667–678. https://doi.org/10.1016/j.molcel.2011.08.027.

13. Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973. https://doi.org/10.1126/science.1237973.

14. Simon, M.D., Wang, C.I., Kharchenko, P.V., West, J.A., Chapman, B.A., Alekseyenko, A.A., Borowsky, M.L., Kuroda, M.I., and Kingston, R.E. (2011). The genomic binding sites of a noncoding RNA. Proc. Natl. Acad. Sci. USA *108*, 20497–20502. https://doi.org/10.1073/pnas.1113536108.

15. Chu, C., Spitale, R.C., and Chang, H.Y. (2015). Technologies to probe functions and mechanisms of long noncoding RNAs. Nat. Struct. Mol. Biol. *22*, 29–35. https://doi.org/10.1038/nsmb.2921.

16. Li, X., Zhou, B., Chen, L., Gou, L.T., Li, H., and Fu, X.D. (2017). GRID-seq reveals the global RNA-chromatin interactome. Nat. Biotechnol. *35*, 940–950. https://doi.org/10.1038/nbt.3968.

17. Li, Y.P., and Wang, Y. (2015). Large noncoding RNAs are promising regulators in embryonic stem cells. J Genet Genomics *42*, 99–105. https://doi.org/10.1016/j.jgg.2015.02.002.

18. Flynn, R.A., and Chang, H.Y. (2014). Long noncoding RNAs in cell-fate programming and reprogramming. Cell Stem Cell *14*, 752–761. https://doi.org/10.1016/j.stem.2014.05.014.

19. Fiorenzano, A., Pascale, E., Patriarca, E.J., Minchiotti, G., and Fico, A. (2019). LncRNAs and PRC2: Coupled Partners in Embryonic Stem Cells. Epigenomes *3*, 14. https://doi.org/10.3390/epigenomes3030014.

20. Guttman, M., Donaghey, J., Carey, B.W., Garber, M., Grenier, J.K., Munson, G., Young, G., Lucas, A.B., Ach, R., Bruhn, L., et al. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature *477*, 295–300. https://doi.org/10.1038/nature10398.

21. Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., and Young, R.A. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. USA *110*, 2876–2881. https://doi.org/10.1073/pnas.1221904110.

22. Sahu, M., and Mallick, B. (2019). Modulation of specific cell cycle phases in human embryonic stem cells by lncRNA RNA decoys. J. Mol. Recognit. *32*, e2763. https://doi.org/10.1002/jmr.2763.

23. Long, Y., Wang, X., Youmans, D.T., and Cech, T.R. (2017). How do lncRNAs regulate transcription? Sci. Adv. *3*, eaao2110. https://doi.org/10.1126/sciadv.aao2110.

24. Chakraborty, D., Paszkowski-Rogacz, M., Berger, N., Ding, L., Mircetic, J., Fu, J., Iesmantavicius, V., Choudhary, C., Anastassiadis, K., Stewart, A.F., and Buchholz, F. (2017). lncRNA Panct1 Maintains Mouse Embryonic Stem Cell Identity by Regulating TOBF1 Recruitment to Oct-Sox Sequences in Early G1. Cell Rep. *21*, 3012–3021. https://doi.org/10.1016/j.celrep.2017.11.045.

25. Tan, J.Y., Sirey, T., Honti, F., Graham, B., Piovesan, A., Merkenschlager, M., Webber, C., Ponting, C.P., and Marques, A.C. (2015). Extensive microRNA-mediated crosstalk between lncRNAs and mRNAs in mouse embryonic stem cells. Genome Res. *25*, 655–666. https://doi.org/10.1101/gr.181974.114.

26. Loh, Y.H., Wu, Q., Chew, J.L., Vega, V.B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., et al. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nat. Genet. *38*, 431–440. https://doi.org/10.1038/ng1760.

27. Rinn, J.L. (2014). lncRNAs: linking RNA to chromatin. Cold Spring Harb. Perspect. Biol. *6*, a018614. https://doi.org/10.1101/cshperspect.a018614.

28. Xie, L., Pelz, C., Wang, W., Bashar, A., Varlamova, O., Shadle, S., and Impey, S. (2011). KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription. EMBO J. *30*, 1473–1484. https://doi.org/10.1038/emboj.2011.91.

29. Muppirala, U.K., Honavar, V.G., and Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. BMC Bioinf. *12*, 489. https://doi.org/10.1186/1471-2105-12-489.

30. Xhabija, B., and Kidder, B.L. (2019). KDM5B is a master regulator of the H3K4-methylome in stem cells, development and cancer. Semin. Cancer Biol. *57*, 79–85. https://doi.org/10.1016/j.semcancer.2018.11.001.

31. Li, G., Kryczek, I., Nam, J., Li, X., Li, S., Li, J., Wei, S., Grove, S., Vatan, L., Zhou, J., et al. (2021). LIMIT is an immunogenic lncRNA in cancer immunity and immunotherapy. Nat. Cell Biol. *23*, 526–537. https://doi.org/10.1038/s41556-021-00672-3.

32. Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol. Cell *40*, 939–953. https://doi.org/10.1016/j.molcel.2010.12.011.

33. Zhang, J.M., Hou, W.B., Du, J.W., Zong, M., Zheng, K.L., Wang, W.J., Wang, J.Q., Zhang, H., Mu, Y.S., Yin, Z., et al. (2020). Argonaute 2 is a key regulator of maternal mRNA degradation in mouse early embryos. Cell Death Discov. *6*, 133. https://doi.org/10.1038/s41420-020-00368-x.

34. Lorenz, R., Bernhart, S.H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26. https://doi.org/10.1186/1748-7188-6-26.

35. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25. https://doi.org/10.1186/gb-2009-10-3-r25.

36. Luo, M., Jeong, M., Sun, D., Park, H.J., Rodriguez, B.A.T., Xia, Z., Yang, L., Zhang, X., Sheng, K., Darlington, G.J., et al. (2015). Long non-coding RNAs control hematopoietic stem cell function. Cell Stem Cell *16*, 426–438. https://doi.org/10.1016/j.stem.2015.02.002.

37. Alvarez-Dominguez, J.R., Knoll, M., Gromatzky, A.A., and Lodish, H.F. (2017). The Super-Enhancer-Derived alncRNA-EC7/Bloodlinc Potentiates Red Blood Cell Development in trans. Cell Rep. *19*, 2503–2514. https://doi.org/10.1016/j.celrep.2017.05.082.

38. Chang, K.C., Diermeier, S.D., Yu, A.T., Brine, L.D., Russo, S., Bhatia, S., Alsudani, H., Kostroff, K., Bhuiya, T., Brogi, E., et al. (2020). MaTAR25 lncRNA regulates the Tensin1 gene to impact breast cancer progression. Nat. Commun. *11*, 6438. https://doi.org/10.1038/s41467-020-20207-y.

39. Maldotti, M., Lauria, A., Anselmi, F., Molineris, I., Tamburrini, A., Meng, G., Polignano, I.L., Scrivano, M.G., Campestre, F., Simon, L.M., et al. (2022). The acetyltransferase p300 is recruited in trans to multiple enhancer sites by lncSmad7. Nucleic Acids Res. *50*, 2587–2602. https://doi.org/10.1093/nar/gkac083.

40. Long, J., Badal, S.S., Ye, Z., Wang, Y., Ayanga, B.A., Galvan, D.L., Green, N.H., Chang, B.H., Overbeek, P.A., and Danesh, F.R. (2016). Long noncoding RNA Tug1 regulates mitochondrial bioenergetics in diabetic nephropathy. J. Clin. Invest. *126*, 4205–4218. https://doi.org/10.1172/JCI87927.

41. Lee, H.C., Kang, D., Han, N., Lee, Y., Hwang, H.J., Lee, S.B., You, J.S., Min, B.S., Park, H.J., Ko, Y.G., et al. (2020). A novel long noncoding RNA Linc-ASEN represses cellular senescence through multileveled reduction of p21 expression. Cell Death Differ. *27*, 1844–1861. https://doi.org/10.1038/s41418-019-0467-6.

42. Lan, T., Yuan, K., Yan, X., Xu, L., Liao, H., Hao, X., Wang, J., Liu, H., Chen, X., Xie, K., et al. (2019). LncRNA SNHG10 Facilitates Hepatocarcinogenesis and Metastasis by Modulating Its Homolog SCARNA13 via a Positive Feedback Loop. Cancer Res. *79*, 3220–3234. https://doi.org/10.1158/0008-5472.CAN-18-4044.

43. Luo, H., Zhu, G., Xu, J., Lai, Q., Yan, B., Guo, Y., Fung, T.K., Zeisig, B.B., Cui, Y., Zha, J., et al. (2019). HOTTIP lncRNA Promotes Hematopoietic Stem Cell Self-Renewal Leading to AML-like Disease in Mice. Cancer Cell *36*, 645–659.e8. https://doi.org/10.1016/j.ccell.2019.10.011.

44. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

45. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

46. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

47. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

48. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.

49. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.

50. Zhang, S., Übelmesser, N., Josipovic, N., Forte, G., Slotman, J.A., Chiang, M., Gothe, H.J., Gusmao, E.G., Becker, C., Altmüller, J., et al. (2021). RNA polymerase II is required for spatial chromatin reorganization following exit from mitosis. Sci. Adv. *7*, eabg8205. https://doi.org/10.1126/sciadv.abg8205.

51. Li, J., Fang, K., Choppavarapu, L., Yang, K., Yang, Y., Wang, J., Cao, R., Jatoi, I., and Jin, V.X. (2021). Hi-C profiling of cancer spheroids identifies 3D-growth-specific chromatin interactions in breast cancer endocrine resistance. Clin. Epigenetics *13*, 175. https://doi.org/10.1186/s13148-021-01167-6.

52. Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., and Cavalli, G. (2017). Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell *171*, 557–572.e24. https://doi.org/10.1016/j.cell.2017.09.043.

53. Servant, N., Varoquaux, N., Lajoie, B.R., Viara, E., Chen, C.J., Vert, J.P., Heard, E., Dekker, J., and Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. Genome Biol. *16*, 259. https://doi.org/10.1186/s13059-015-0831-x.

54. Akdemir, K.C., and Chin, L. (2015). HiCPlotter integrates genomic data with interaction matrices. Genome Biol. *16*, 198. https://doi.org/10.1186/s13059-015-0767-1.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit mAb Oct-4A	Cell Signaling Technology	Cat#C30A3C1; RRID:AB_10547892
Nanog Antibody (pAb)	ACTIVE MOTIF	Cat#61419; RRID:AB_2750953
Human/Mouse/Rat SOX2 Antibody	R&D	Cat#AF2018; RRID:AB_355110
Tri-Methyl-Histone H3 (Lys4)	CST	Cat# 9727
Chemicals, peptides, and recombinant proteins		
PD-0325901	MedChemExpress	Cat#HY-10254
CHIR-99021	MedChemExpress	Cat#HY-10182
mLIF Medium Supplement	Millipore	Cat#ESG1106
Critical commercial assays		
TruePrep DNA Library Prep Kit V2 for Illumina	Vazyme	Cat#TD502
Hyperactive Universal CUT&Tag Assay Kit for Illumina	Vazyme	Cat#TD903
Lipofectamine™ 3000 Transfection Kit	Invitrogen	Cat#L3000008
TB TB Green® Fast qPCR Mix	TaKaRa	Cat#RR430A
BCIP/NBT Alkaline Phosphatase Color Development Kit	Beyotime	Cat#C3206
Deposited data		
RNA-Seq sequencing data	This paper	CRA011822
ChIRP-Seq sequencing data	This paper	CRA011822
Experimental models: Cell lines		
Mouse:Passage 30 mES cells	Laboratory of Shuai Ling	mES Cell Line: GT1-1
Oligonucleotides		
siRNA targeting sequence: Linc1393 #1: GCAGAGACUUCCCGCUUAATT	This paper	N/A
siRNA targeting sequence: Linc1393 #2: GCUUGGCAACAACUCUUUATT	This paper	N/A
Primer:Gapdh Forward:TGCCCAGAACATCATCCCT	This paper	N/A
Primer:Linc1393 Forward:GAGGGGCTTGAGGCTTGG	This paper	N/A
Primer:Oct4 Forward:GCTCACCCTGGGCGTTCT	This paper	N/A
Primer:Sox2 Forward:GGGGCAGCGGCGTAAGAT	This paper	N/A
Primer:Set1a Forward:GCAAGGTCCTACCA TCAGCA	This paper	N/A
Recombinant DNA		
Mouse raptor: pLKO.1-copGFP-2A-PURO-linc1393	This paper	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
fastp	https://github.com/OpenGene/fastp	Version 0.22.0
HISAT2	http://daehwankimlab.github.io/hisat2/	Version 2.2.1
featureCounts	https://subread.sourceforge.net/	Version 2.0.3
DESeq2	https://github.com/mikelove/DESeq2	Version 1.34.0
Macs2	https://github.com/macs3-project/MACS	Version 2.2.7.1
Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2	Version 2.5.0
BEDTools	https://bedtools.readthedocs.io/	Version 2.30.0
deepTools	https://deeptools.readthedocs.io/	Version 3.5.1
SAMtools	https://www.htslib.org/	Version 1.1
Picard	https://github.com/broadinstitute/picard/releases/tag/2.26.11	Version 2.20.4
ImageJ	NIH	https://imagej.nih.gov/ij/
LncTargeter	This paper	https://github.com/summus-kong/LncTargeter

RESOURCE AVAILABILITY

Lead contact

Further information and request for resources and reagents should be directed to and would be fulfilled by the Lead Contact, Qingran Kong (kqr721726@163.com).

Materials availability

Unique materials generated in this study are available upon completing the materials transfer agreement.

Data and code availability

- The datasets in this study ave been deposited in the Genome Sequence Archive (https://bigd.big.ac.cn/gsa/) under accession no. CRA011822.
- The source code generated during this study is available via GitHub repository at https://github.com/summus-kong/LncTargeter.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The GT1-1 mESC line (gift from Ling Shuai, Nankai University) was cultured in high glucose Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal FBS (Gibco), 10% KSOR (Gibco), 1% sodium pyruvate (Gibco), 1% NEAA (Gibco), 100 mmol/L β-mercaptoethanol (Sigma), 10^6 U/L mLIF (Millipore), 3 μM CHIR-99021 and 1 μM PD-0325901. When the degree of cell confluences was more than 80%, it was passaged in an appropriate proportion. All mESCs were cultured at 37°C under 5% CO2 and saturated humidity.

METHOD DETAILS

Plasmid construction

shRNA targeting linc1393 was synthesized (Azenta GENEWIZ), annealed and cloned into the EF1α-pBUD-EGFP vector. Vectors expressing control linc1393, sh-linc1393 were synthesized with the targeting site deleted, and linc1393 with core subsequence 3 deleted (Azenta/GENEWIZ), and then cloned into the U6-pBUD-EGFP plasmid digested with Spe I and Kpn I. All vectors were verified through Sanger sequencing.

Cell transfection and alkaline phosphatase staining

The shRNA sequences of linc1393 are as follows (5′-3′): sh-linc1393 #1: (GCAGAGACTTCCCGCTTAAGT); sh-linc1393 #2: (GGACAGATACTGCGATTTACA). mESCs were cultured to approximately 50% confluence,

and Lipofectamine 3000 (ThermoFisher) was used for transfection of shRNAs according to the manufacturer's instructions. After 24 h, the mESC medium containing 200 μg/mL Zeocin (Invitrogen) was replaced, incubation was continued for 3 days, and cells were collected from the sh-linc1393, sd-linc1393 and dd-linc1393 group. AP activity was monitored using a commercial AP detection kit (Beyotime) according to the manufacturer's instructions.

RNA interference

All siRNAs used in this study were purchased from GenePharma. The siRNA sequences of linc1393 and linc1281 are as follows (5′–3′): si-linc1393-1: (GCAGAGACUUCCCGCUUAATT); si-linc1393-2: (GCUUGGC AACAACUCUUUATT); si-linc1281-1: (TGAGTTCAAATCCCAGCAACC); and si-linc1281-2: (AGGTAATCTGT GATCCCTTTAGGC). siRNAs were transfected into mESCs at a final concentration of 10 μM using Lipofectamine 3000 (ThermoFisher) according to the manufacturer's instructions, and cells were collected 72 h after transfection.

RNA-binding protein immunoprecipitation

Protein G magnetic beads were coated with 5 μg of the primary antibody (SET1A; Abcam, IgG; Millipore) in RIP wash buffer (50 mM Tris–HCl, pH 7.4, 150 mM sodium chloride, 1 mM $MgCl_2$, and 0.05% NP-40) containing 5% bovine serum albumin (BSA), for 3 h at 4°C with rotation. Then, we collected the cells and added 100 μL of RIP lysis buffer (150 mM NaCl, 50 mM Tris-HCl, pH 7.4, 1 mM ethylenediaminetetraacetic acid [EDTA], 0.1% sodium dodecyl sulfate [SDS], 1% NP-40, 0.5% sodium deoxycholate, 0.5 mM dithiothreitol, 1 mM phenylmethylsulfonyl fluoride cocktail) and 10 μL of RNase inhibitor (AM2694; Ambion), followed by incubation on ice for 10 min. Next, 90 μL of the supernatant was transferred after centrifugation at 14,000 rpm for 10 min at 4°C. Then, 900 μL RIP immunoprecipitation buffer (860 μL premixed RIP wash buffer, 35 μL 0.5 M EDTA, and 5 μL RNase inhibitor) was added, and the mixture was incubated with rotation overnight at 4°C. The residual 10 μL of RIP lysate was used as the input. Beads were washed and treated with proteinase K at 55°C for 30 min to digest proteins, followed by RNA extraction from the supernatant. Then, linc1393 was quantified through qPCR.

RNA pull-down and liquid chromatography-mass spectrometry

To identify linc1393-interacting proteins in mESCs, biotin RNA pull-down was performed, followed by liquid chromatography-mass spectrometry (LC-MS). We performed pull-down assays with biotinylated linc1393 or antisense RNA using the Pierce Magnetic RNA-Protein Pull-Down Kit (ThermoFisher) according to the manufacturer's instructions. For LC-MS analysis, the lyophilized peptide fractions were re-suspended in double-distilled water containing 0.1% formic acid, and 2 μL aliquots of this suspension were loaded into a nanoViper C18 (Acclaim PepMap 100, 75 μm × 2 cm) trap column. Online chromatographic separation was performed on the Easy nLC 1200 system (ThermoFisher). The trapping and desalting procedures were conducted in a volume of 20 μL comprised of 100% solvent A (0.1% formic acid). Then, an elution gradient of 5–38% solvent B (80% acetonitrile, 0.1% formic acid) over 60 min was used for the analytical column (Acclaim PepMap RSLC, 75 μm × 25 cm C18-2 μm 100 Å). The DDA (data-dependent acquisition) mass spectral technique was used to acquire tandem MS data on a ThermoFisher Q Exactive mass spectrometer fitted with a Nano Flex ion source. Data were acquired using an ion spray voltage of 1.9 kV with an interface heater temperature of 275°C. For a full mass spectrometric survey, the target value was 3×10^6, the scan range was 350–2,000 m/z at a resolution of 70,000, and the maximum injection time was 100 ms. For the MS2 scan, only spectra with a charge state of 2–5 were selected for fragmentation through higher-energy collision dissociation, with a normalized collision energy of 28. The MS2 spectra were acquired using the ion trap in rapid mode with an automatic gain control target of 8,000 and maximum injection time of 50 ms. Dynamic exclusion was set to 25 s.

Immunofluorescence staining

For immunofluorescence staining, mESCs were plated onto Matrigel-coated 24-well chambers and washed twice in phosphate-buffered saline (PBS). Cells were fixed for 30 min at room temperature (RT) in 4% paraformaldehyde, followed by permeabilization in 10% Triton X-100 for 1 h at RT. Cells were then blocked in blocking solution (1% BSA in PBS) for 1 h at RT after washing three times (for 5 minutes each time) in washing solution (0.1% Tween-20 and 0.01% Triton X-100 in PBS). Incubations were performed overnight at 4°C using the following antibodies and dilutions in blocking solution: KDM5B (1:100), OCT4 (1:100) and Alexa Fluor 546-conjugated donkey anti-rabbit IgG (1:500). Then, the cells were washed three times in washing

solution and incubated with secondary antibodies (1:1,000) for 1 h at RT. Slides were mounted with VectaShield mounting medium containing DAPI (4′,6-diamidino-2-phenylindole) and imaged on a Leica SP5 upright confocal microscope using 20× air and 60× oil immersion objectives.

RNA extraction, reverse transcription and qPCR analysis

Total RNA was extracted using the TRIzol (Invitrogen) method following the manufacturer's protocol and reverse-transcribed using the M5 Sprint qPCR RT kit (Mei5bio). mESCs were lysed in 50 mL of lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM MgCl2 and 0.15% NP-40) for 10 min on ice. Immediately after lysis, samples were then centrifuge at 300 g at 4°C for 5 min. The supernatant was the cytoplasmic fraction and the precipitate was the nuclear fraction.[33] qPCR was performed using TB Green Premix Ex Taq (TaKaRa) and CFX96 qPCR System (Bio-Rad). The reaction parameters were as follows: 50°C for 2 min, 95°C for 2 min, and 40 two-step cycles of 95°C for 15 s and 60°C for 1 min. Threshold cycle (Ct) values were calculated using Sequence Detection System software, and the amount of the target sequence normalized to the reference sequence was calculated as $2^{-\Delta\Delta Ct}$.

Feature extraction

From the FASTA sequence of the lncRNA, RNAfold,[34] a tool commonly used for the prediction of RNA secondary structure, was employed to determine the secondary structures of RNAs. Based on a sliding window of 17–50 bp with a step size of 1 bp, each lncRNA sequence was segmented into fixed subsequences, and its GC content was then calculated. Then, it was mapped to the genome (Genome assembly: GRCm38.p6, GRCh38.p13) using Bowtie software.[35] The GC content of the binding site (200 bp around the match site) was calculated using Biopython. For each match, the free energy (ΔG) of the strand–strand interaction between RNA and DNA was calculated. Defining a finite sequence ω including bases A, C, T, and G, where ω is a finite sequence of length $|\omega|$ and n is a unique integer such that $4^n + n - 1 \leq |\omega| < 4^{(n+1)} + (n+1) - 1$, topological entropy was calculated for $\omega_1^{(4^n+n-1)}$, i.e., the first $4^n + n - 1$ bases of ω, as:

$$H_{top}(\omega) = \frac{log_4\left(p_{\omega_1^{4^n+n-1}}(n)\right)}{n}$$

Classification algorithms and evaluation metrics

We obtained ChIRP–seq and RNA-seq data of lncHSC2, alncRNA-EC7, MaTAR25, lncSmad7, Tug1, Linc-ASEN, SNHG10 and HOTTIP from the GEO database (http://www.ncbi.nlm.nih.gov/geo/; GEO: GSE63277, GSE97119, GSE142169, GSE154738, GSE77493, GSE77506, GSE128398, GSE119773, GSE120095 and GSE114981).[36–43] To predict the binding sites of a given lncRNA, the subsequences were matched to the human or mouse genome (Genome assembly: GRCh38.p13, GRCm38.p6). The genomic sites matching eight lncRNAs present as peaks in ChIRP-seq data, and were included in positive sets for training the machine learning models; the corresponding negative sets consisted of random loci from the remaining matched genomic sites. To identify the best-performing classification algorithm, eight machine learning models, including gradient boosting decision tree (GBDT), extreme gradient boosting (XGBoost), light gradient boosted machine (LightGBM), adaptive boosting (AdaBoost), random forest (RF), logistic regression (LR), naive Bayes (NB), and decision trees (DTs) were employed as binary classifiers, using eight features including the free energy, sequence length and complexity of the subsequences for lncRNA, and the matching genomic sites. To assess the performance of the models, the metrics of accuracy (Acc), sensitivity (Sn), precision (Pr) and F1-Scores (F1), and the area under the receiver operating characteristic curve (AUROC) were used. These metrics are defined as follows:

$$\mathrm{Acc} = \frac{TP+TN}{TP+TN+FP+TN}$$

$$\mathrm{Sn} = \frac{TP}{TP+FN}$$

$$\mathrm{Pr} = \frac{TP}{TP+FP}$$

$$\mathrm{F1} = 2 \times \frac{Pr \times Sn}{Pr+Sn}$$

$$Sp = \frac{TN}{TN+FP}$$

$$AUROC = \int TPRd(FPR)$$

Where TP, TN, FP, FN, TPR, and FPR indicate true-positives, true-negatives, false-positives, false-negatives, the true-negative rate, and the false-positive rate, respectively. Higher values of Acc, Sn, Pr, F1, and Sp indicate more robust models. The AUROC was obtained from the plot of TPR and FPR at various thresholds. A perfect model has an AUC of 1, and random performance equates to an AUC of 0.5.

RNA-seq

RNA-seq libraries were prepared as previously described. Briefly, two replicates were performed for each group. Cells were washed three times in 1× Dulbecco's PBS solution to avoid possible contamination. cDNA was amplified using Phusion Hot Start II High-Fidelity PCR Master Mix (F-548L; ThermoFisher). Library preparation was performed using the TruePrep DNA Library Prep Kit V2 for Illumina (TD501; Vazyme Biotech) according to the manufacturer's instructions. All libraries were sequenced using the NovaSeq 6000 (Illumina) platform according to the manufacturer's instructions.

RNA-seq data analysis

For RNA-seq analysis of Illumina reads, the FastQC tool was employed. We used Trim Galore software to eliminate low-quality reads, trim adaptor sequences, and low-quality base reads. Then, the HISAT2 program[44] was used to align reads to the mouse or human reference genome (Genome assembly: GRCm38.p6, GRCh38.p13). Gene-level quantification was used to aggregate the raw counts of mapped reads using the featureCounts tool.[45] The expression level of each gene was quantified in terms of normalized fragments per kilobase of transcript per million mapped reads (FPKM). Next, the R package DESeq2[46] was employed for differential gene expression analysis. Kyoto Encyclopedia of Genes and Genomes (KEGG) and GO analyses of the screened genes were performed using the KOBAS online tool (http://kobas.cbi.pku.edu.cn/kobas3/).

CUT&Tag

CUT&Tag was performed using the Hyperactive In-Situ ChIP Library Prep Kit for Illumina (TD901; Vazyme Biotech). Cells were incubated with 10 μL pre-washed ConA beads. Then, 50 μL of antibody buffer and 0.5 μg of the resulting antibody were added, and the mixture was cultured for 2 h at RT. After washing twice with wash buffer containing digitonin (dig-wash buffer), 150 μL of dig-wash buffer containing 0.7 μg secondary antibody was added and the mixture was incubated at RT for 1 h. After washing twice with 800 μL dig-wash buffer, 0.3 μL of pG–Tn5 and 50 μL of dig-300 buffer were added, and the samples were incubated at RT for 1 h followed by washing (twice) with 800 μL dig-wash buffer. We added 300 μL of tagmentation buffer and the samples were incubated at 37°C for 1 h. The reaction was stopped with 10 μL 0.5 M EDTA, 3 μL 10% SDS and 2.5 μL 20 mg/ml proteinase K. After extraction with phenol–chloroform and ethanol precipitation, PCR was performed to obtain libraries under the following cycling conditions: 72°C for 5 min, 98°C for 30 s, 17 cycles of 98°C for 10 s and 60°C for 30 s, followed by a final extension at 72°C for 1 min and holding at 4°C. Post-PCR clean-up was performed through the addition of DNA Clean Beads at 1.2× the PCR product volume (N411; Vazyme Biotech), after which the libraries were incubated with beads for 5 min at RT, washed gently in 80% ethanol, and eluted in 20 μL water. All libraries were sequenced using the Illumina NovaSeq 6000 platform according to the manufacturer's instructions.

CUT&Tag data processing

The paired-end CUT&Tag reads of H3K4me3 were aligned to the mouse reference genome (GRCm38.p6) using Bowtie2 v2.2.2 software, with the following options: -local -very-sensitive -local -no-unal -nomixed -no-discordant -phred33 -I 10 -X 700. This was followed by filtering with SAMtools v1.7 software[47] at MAPQ10. Unmapped and non-uniquely mapped reads were removed. The reproducibility of CUT&Tag data between replicates was assessed through correlation analysis of mapped read counts across the genome. Then, we pooled the biological replicates for each stage and performed downstream analysis. For quantitative analysis, we normalized the read counts by computing the number of reads per kilobase of transcript per million mapped reads (RPKM). RPKM values were calculated using merged replicate BAM files

with the bamCoverage tool of deepTools software.[48] To minimize batch and cell type variations, the RPKM values were further normalized through *Z* score transformation (whole-genome 100-bp bins, excluding outliner regions). The washU epigenome browser (http://epigenomegateway.wustl.edu/browser/) was used to visualize H3K4me3 CUT&Tag data.

ChIRP-seq data processing and analysis

For ChIRP-seq data analysis, low-quality reads and adaptors were trimmed using Trim Galore and clean paired-end reads were mapped to the mouse or human reference genome (Genome assembly: GRCm38.p6, GRCh38.p13) using Bowtie2 v2.2.2 software with the default parameters, followed by filtering with SAMtools v1.7 software. ChIRP peaks were called based on merged replicates and normalized to the input level using MACS2 v2.1.1.20160309 software.[49] Peak calls with a false discovery rate (FDR) $\leq$ 5% were used for downstream analysis.

Hi-C data processing and analysis

Hi-C data were downloaded from the GEO database (GEO: GSE96107, GSE165572 and GSE160235).[50–52] To process these data, we first trimmed the raw data to remove adaptor sequences and low-quality reads. Raw paired-end reads in the Hi-C libraries were aligned, processed and iteratively corrected using HiC-Pro (version 3.1.0) software.[53] Briefly, the sequencing reads were independently aligned to the mouse or human reference genome (Genome assembly: GRCm38.p6, GRCh38.p13) using the Bowtie2 end-to-end algorithm. Unmapped reads, multiply mapped reads and singletons were then discarded, and valid read pairs were binned at a specific resolution based on division of the genome into bins of equal size. We selected a bin size of 10 kb to show local interactions, and for topologically associated domain calling. After detailed statistics were generated, a series of Hi-C interaction frequency heatmaps were drawn using HiCPlotter.[54]

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed using R v4.1.0 software (R Development Core Team, Vienna, Austria). Data are expressed as mean $\pm$ standard error of the mean (SEM). Differences between means were evaluated using the two-tailed Student's *t* test or Wilcoxon rank-sum test. Asterisks indicate significant differences as follows: *$p < 0.05$, **$p < 0.01$, and ***$p < 0.001$.