Molecular Therapy
Nucleic Acids
Original Article

Recurrence-Associated Long Non-coding RNA Signature for Determining the Risk of Recurrence in Patients with Colon Cancer

Meng Zhou,[1,4] Long Hu,[2,4] Zicheng Zhang,[2] Nan Wu,[3] Jie Sun,[1] and Jianzhong Su[1]

[1]School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou 325027, China; [2]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China; [3]Department of Orthopedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing 100730, China

Patients with colon cancer are often faced a high risk of disease recurrence within 5 years of treatment that is the major cause of cancer mortality. Reliable molecular markers were required to improve the most effective personalized therapy. Here, we identified a recurrence-associated six-lncRNA (long non-coding RNA) signature (*LINC0184*, *AC105243.1*, *LOC101928168*, *ILF3-AS1*, *MIR31HG*, and *AC006329.1*) that can effectively distinguish between high and low risk of cancer recurrence from 389 patients of a discovery dataset, and validated its robust performance in four independent datasets comprising a total of 906 colon cancer patients. We found that the six-lncRNA signature was an independent predictive factor of disease recurrence in multivariate analysis and was superior to the performance of clinical factors and known gene signature. Furthermore, *in silico* functional analysis showed that the six-lncRNA-signature-associated coding genes are significantly enriched in proliferation and angiogenesis, cell death, as well as critical cancer pathways that could play important roles in colon cancer recurrence. Together, the six-lncRNA signature holds great potential for recurrence risk assessment and personalized management of colon cancer patients.

INTRODUCTION

Colon cancer is one of the most common cancers and remains one of the leading causes of cancer death worldwide, with more than 2.2 million new cases and 1.1 million deaths by 2030.[1,2] In China, a significant upward trend in the incidence rate and mortality of colon cancer was observed especially in younger patients in recent years.[3] The colectomy combined with chemotherapy and radiation therapy is a current standard treatment for colon cancer. Despite continuous developments in treatment, earlier detection, and management leading to reductions in the incidence and mortality overall and improvement in overall survival of colon cancer, about ~30%–50% of patients relapsed within 5 years of treatment.[4,5] Thus, reliable and robust molecular markers in addition to the current clinical and pathological factors for determining the risk of recurrence is urged to improve the most effective personalized therapy for patients with colon cancer.

Long non-coding RNAs (lncRNAs) were arbitrarily defined as ncRNA transcripts of greater than 200 nt with no or little protein-coding potential.[6,7] Studies about ncRNA biology have shown that lncRNAs are involved in numerous biological processes by function as an important player of the gene regulatory network on transcriptional, posttranscriptional, and epigenetic levels.[8–10] Growing evidence shows that lncRNAs are an emerging hallmark of cancer,[11] and their aberrant expression contributes to the cancer occurrence, progression, and prognosis.[12,13] Several known lncRNAs of *HOTAIR*, *CCAT*, *MALAT-1*, and *H19* have been found to be involved in the diagnosis, invasion, metastasis, and prognosis of colon cancer.[14] Several studies have already reported lncRNA-focus expression signature for predicting overall survival of patients with colon cancer. For example, Hu and colleagues[15] identified a six-lncRNA signature to improve prognosis prediction of colorectal cancer. Another two-lncRNA signature also was identified to predict survival of patients with colon adenocarcinoma.[16] However, the predictive significance of lncRNAs in risk assessment of recurrence has not already been performed on large patient cohorts.

In this study, we performed a systematic analysis of lncRNA expression profiles and clinical data on a large colon cancer cohort of 1,480 patients to identify a robust and reproducible lncRNA expression signature predictive for colon cancer recurrence.

RESULTS

Identification of Recurrence-Associated lncRNAs in Patients with Colon Cancer

Here, the GSE39582 dataset from the Marisa et al.[40] study, which is the largest patient dataset enrolled in this study, contains 179

Received 1 April 2018; accepted 21 June 2018;
https://doi.org/10.1016/j.omtn.2018.06.007.

[4]These authors contributed equally to this work.

Correspondence: Jie Sun, School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou 325027, China.
E-mail: suncarajie@hotmail.com
Correspondence: Jianzhong Su, School of Ophthalmology & Optometry and Eye Hospital, School of Biomedical Engineering, Wenzhou Medical University, Wenzhou 325027, China.
E-mail: sujz@wibe.ac.cn

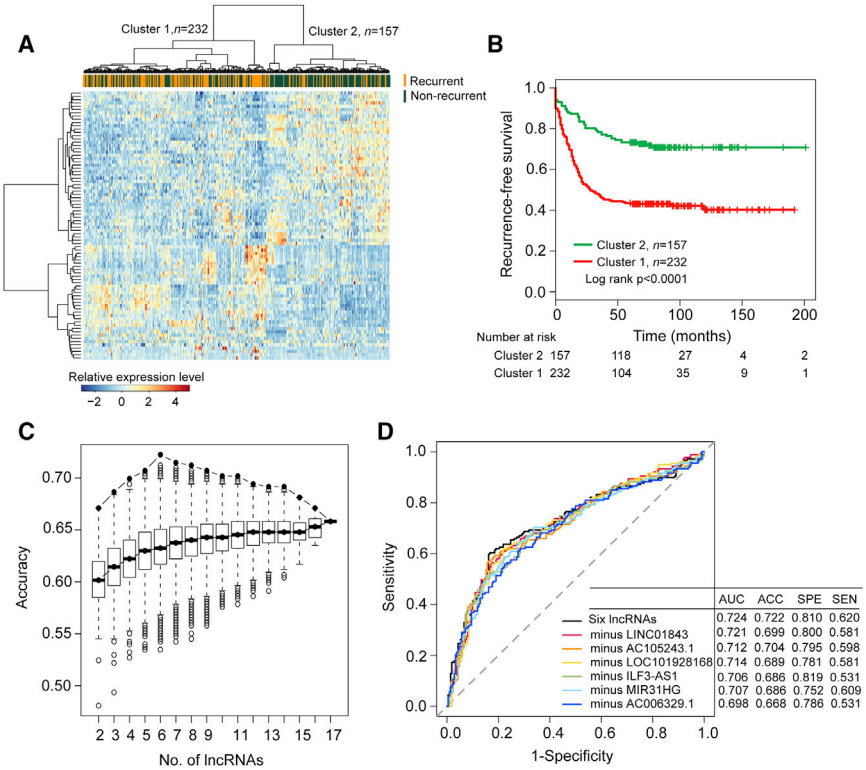


Figure 1. Identification of the Six-lncRNA Signature for Recurrence Risk Prediction in the Discovery Dataset

(A) Unsupervised clustering of patients based on the expression pattern of 82 differentially expressed lncRNAs. (B) Kaplan-Meier estimates of the recurrence-free survival of patients in the two sample clusters are based on 82 differentially expressed lncRNAs. (C) Boxplot of the predicted accuracy of each combination constructed by a specific number of recurrence-associated lncRNAs. (D) Receiver operating characteristic (ROC) curves for the six-lncRNA signature and other six-minus-one lncRNA signatures.

patients with recurrence and 395 recurrence-free patients. In order to identify lncRNAs with close associations with cancer recurrence, 179 patients with recurrence and 210 recurrence-free patients (alive >5 years without any evidence of recurrence) in the Marisa et al.[40] dataset were selected to form a discovery dataset (n = 389). Then, lncRNA expression profiles of 179 patients with recurrence and 210 recurrence-free patients in the discovery dataset were compared to determine whether there was a significant difference in lncRNA expression pattern between patients with and without recurrence. In total, 317 lncRNAs with their coefficient of expression variation greater than 0.1 were identified as variant lncRNAs. Using significance analysis of microarrays (SAM) method with a false discovery rate (FDR)-adjusted p value <0.05 for these 317 variant lncRNAs, 82 lncRNAs were differentially expressed between patients with and without recurrence. Of them, 49 lncRNAs were found to be down-regulated and 33 to be upregulated in recurrent patients. We conducted unsupervised hierarchical clustering analysis on 389 samples of discovery dataset using the set of 82 differentially expressed lncRNAs. As showed in Figure 1A, there was the distinctive expression pattern for 82 differentially expressed lncRNAs that distinguished recurrent patient samples from non-recurrent patient samples ($p < 0.0001$, χ^2 test; Figure 1A). Kaplan-Meier analysis and log rank test revealed the significant difference in recurrence-free survival (RFS) time between the two sample clusters ($p < 0.0001$, log rank test; Figure 1B). Therefore, these differentially expressed lncRNAs were considered as candidate recurrence-associated lncRNAs.

Random forest supervised classification algorithm was applied for further analysis for these candidate recurrence-associated lncRNAs. After five iteration procedures, 17 lncRNAs mostly related to the recurrence classification were identified according to the permutation important score and were selected as recurrence-associated lncRNAs. Clustering analysis of the 17 recurrence-associated lncRNAs clearly separated 389 samples of discovery dataset into the recurrence-like patient cluster and non-recurrence-like patient cluster ($p < 0.0001$, χ^2 test; Figure S1A). Furthermore, there was a significant difference in RFS time between the recurrence-like patient cluster and non-recurrence-like patient cluster ($p < 0.0001$, log rank test; Figure S1B).

Development of a lncRNA Signature for Recurrence Risk Stratification in the Discovery Dataset

To obtain an optimal lncRNA combination for recurrence risk stratification by considering a balance between classification accuracy and the number of lncRNAs, we used a support vector machine (SVM) and 5-fold cross-validation to access the classification accuracies for each combination of 131,054 combinations constructed by specific number of recurrence-associated lncRNAs (k = 2, 3, …, 17) in the discovery dataset. The above-mentioned analysis revealed that the combination of six lncRNAs (*LINC0184*, *AC105243.1*, *LOC101928168*, *ILF3-AS1*, *MIR31HG*, and *AC006329.1*) provided the greatest predictive ability with the highest accuracy rate of 72.2% and an area under the curve (AUC) of 0.724 (Figures 1C and 1D). Hierarchical clustering of six lncRNAs clearly separated patients of discovery dataset into two major patient groups with significantly different recurrence risk status ($p < 0.0001$, χ^2 test) and RFS ($p < 0.0001$, log rank test) (Figures S2A and S2B). Moreover, we found that all of the six lncRNAs are significantly associated with the patient's RFS by univariate Cox proportional hazard regression in the discovery dataset (Table 1). Therefore, the combination containing six lncRNAs was selected as the final lncRNA signature for predicting the risk of recurrence.

Table 1. List of Six lncRNAs in the Signature Predictive of Recurrence in the Discovery Dataset

Ensembl or RefSeq ID	Gene Symbol	Genomic Location (GRCh38)	Coefficient	Hazard Ratio	95% CI	p Value
ENSG00000251169	*LINC01843*	chromosome 5: 134,506,552-134,509,229 forward strand	−0.488	0.614	0.465–0.81	0.001
ENSG00000261780	*AC105243.1*	chromosome 18: 73,324,941-73,349,878 forward strand	0.249	1.283	1.091–1.51	0.003
NR_120523.1	*LOC101928168*	chromosome 7: 29,125,033-29,128,172 forward strand	−0.249	0.780	0.668–0.91	0.002
ENSG00000267100	*ILF3-AS1*	chromosome 19: 10,651,862-10,653,844 reverse strand	−0.254	0.776	0.639–0.942	0.010
ENSG00000171889	*MIR31HG*	chromosome 9: 21,455,642-21,559,669 reverse strand	0.545	1.724	1.431–2.076	<0.001
ENSG00000232445	*AC006329.1*	chromosome 7: 101,308,346-101,310,985 forward strand	−0.605	0.546	0.398–0.75	<0.001

To further test whether all of the six lncRNAs in the final lncRNA signature are essential for its predictive ability, we constructed all possible six-minus-one lncRNA signatures by deleting one lncRNA at a time and performed comparison analysis of predictive ability for original six-lncRNA signature and other six-minus-one lncRNA signatures using the SVM and 5-fold cross-validation in the discovery dataset. The comparison showed that none of the six-minus-one lncRNA signatures had a higher predictive accuracy and AUC than the original six-lncRNA signature (Figure 1D). This indicates that all six lncRNAs were essential for the final lncRNA signature for recurrence risk stratification. Finally, a recurrence risk score model was developed from the six-lncRNA signature (hereafter referred to as RRLnc6 score) using a linear combination of the expression level of six lncRNAs, weighted by the regression coefficients derived from the multivariate Cox regression as follows: RRLnc6 score = (−0.13066) * *LINC01843* + (0.00474) * *AC105243.1* + (−0.20691) * *LOC101928168* + (−0.04769) * *ILF3-AS1* + (0.23582) * *MIR31HG* + (−0.35645) * *AC006329.1*.

Predictive Performance of the Six-lncRNA Signature for Recurrence Risk in the Discovery Dataset and GSE39582 Dataset

To investigate the effectiveness of the six-lncRNA signature for clinical recurrence risk prediction, we calculated the RRLnc6 score for each patient in the discovery dataset and compared it between recurrent patients and non-recurrent patients. The RRLnc6 score distribution is significantly different between recurrent patients and non-recurrent patients in the discovery dataset and GSE39582 dataset, and the median RRLnc6 score was significantly higher among patients who developed recurrence compared with patients who did not (0.226 versus −0.092, $p < 0.0001$ for discovery dataset; 0.226 versus −0.013, $p < 0.0001$ for GSE39582 dataset, Mann-Whitney U test) (Figures 2A and 2B).

By using the median value (0.046) of the RRLnc6 score distribution in the discovery as risk cutoff, those patients with an RRLnc6 score of 0.046 or higher were included in the group of patients at high risk of disease recurrence (hereafter referred to as high-risk group), and those with an RRLnc6 score lower than 0.046 were included in the group at low risk of disease recurrence (hereafter referred to as the low-risk group). When the six-lncRNA signature was applied to the discovery dataset, we classified 389 patients of the discovery dataset into the high-risk group (n = 195) and low-risk group (n = 194) based on risk cutoff. The predicted low-risk group had significantly longer RFS than the predicted high-risk group (5-year RFS 69.6% versus 41.5%; $p < 0.0001$, log rank test) (Figure 2C). The same risk cutoff derived from the discovery dataset classified 574 patients of the GSE39582 dataset into the high-risk (n = 291) and low-risk groups (n = 283) with significantly different RFS (5-year RFS 56.5% versus 77.1%, $p < 0.0001$, log rank test) (Figure 2D). Furthermore, the univariate Cox regression analyses also showed that the hazard ratios (HRs) of the predicted high-risk group versus low-risk group for RFS were 2.461 ($p < 0.0001$; 95% confidence interval [CI]: 1.807–3.353) in the discovery dataset and 2.300 ($p < 0.0001$; 95% CI: 1.689–3.132) in the GSE39582 dataset (Table 2).

External Validation of the Six-lncRNA Signature for Predicting Recurrence Risk with Three Independent Microarray Datasets

The robustness and reproducibility of the six-lncRNA signature for determining the risk of recurrence were further examined using three independent microarray datasets from GEO database. We first evaluated the association of RRLnc6 score with the RFS in the univariate analysis and found that the RRLnc6 score remained highly associated with RFS in all tested GEO independent datasets (HR 1.767, 95% CI: 0.997–3.134, $p = 0.051$ for GSE14333; HR 1.831, 95% CI: 1.067–3.141, $p = 0.028$ for GSE17538; HR 3.173, 95% CI: 1.130–8.906, $p = 0.028$ for GSE33113) (Table 2). In addition, the distribution of the RRLnc6 score varied significantly between recurrent patients and non-recurrent patients, and patients who developed recurrence had significantly higher RRLnc6 score than those who did not in all tested independent datasets (median 0.187 versus −0.053, $p = 0.048$ for GSE14333; 0.125 versus −0.046,

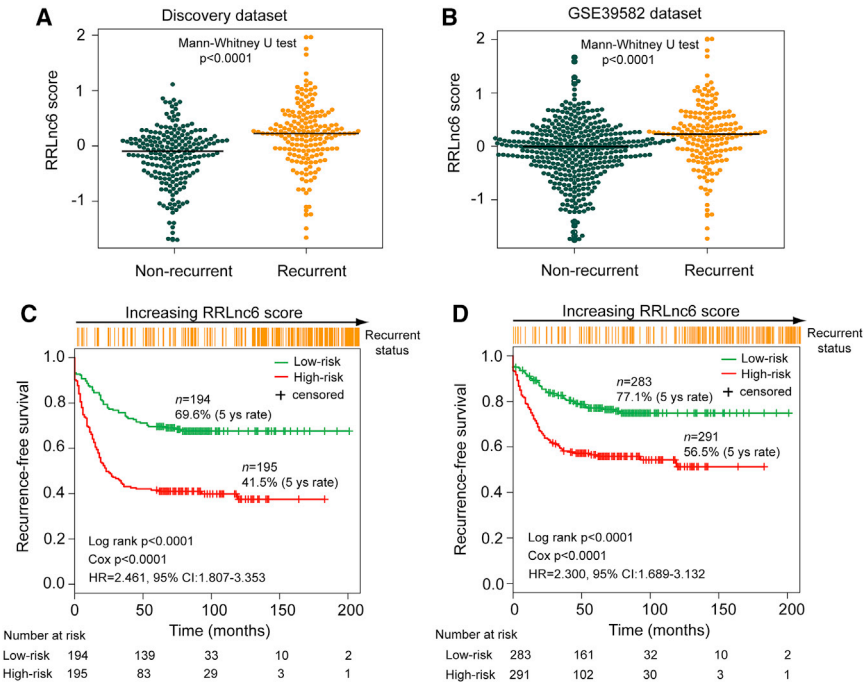


Figure 2. Performance Assessment of the Six-lncRNA Signature in the Discovery and GSE39582 Datasets

Distribution of RRLnc6 score in recurrent patients and non-recurrent patients in the discovery dataset (A) and the GSE39582 dataset (B). Kaplan-Meier estimates of the recurrence-free survival of patients with low or high risk predicted by the six-lncRNA signature in the discovery dataset (C) and the GSE39582 dataset (D).

p = 0.058 for GSE17538; 0.131 versus −0.127, p = 0.025 for GSE33113, Mann-Whitney U test) (Figures 3A–3C).

To confirm that the six-lncRNA signature had similar predictive value in independent datasets, we applied the same recurrence risk score model and risk cutoff derived from the discovery dataset to three independent GEO datasets, classifying patients of each independent dataset into high-risk and low-risk groups. As for the discovery dataset, RFS was significantly different between the high-risk group and the low-risk group (p = 0.048 for GSE14333; p = 0.027 for GSE17538; p = 0.021 for GSE33113, log rank test) (Figures 3D–3F). The 5-year RFS rate of the high-risk group was 70.9%, 63.4%, and 66.1% in the GSE14333 dataset, GSE17538 dataset, and GSE33113 dataset, respectively, whereas the corresponding rates in the low-risk group were 81.4%, 76.5%, and 87.4%, respectively. Moreover, there are significantly more patients with recurrence in the high-risk group than in the low-risk group in all three independent datasets (29.2% versus 15.8%, p = 0.024 for GSE14333; 35.5% versus 20.6%, p = 0.028 for GSE17538; 31.7% versus 10.4%, p = 0.026 for GSE33113, χ^2 test).

Further Validation of the Six-lncRNA Signature for Predicting Recurrence Risk with an Independent RNA-Sequencing Dataset

The predictive performance of the six-lncRNA signature was further tested in the independent RNA-sequencing (RNA-seq) dataset based on the Illumina HiSeq platform from The Cancer Genome Atlas (TCGA) database. Unfortunately, we found that only five lncRNAs (*LINC0184*, *AC105243.1*, *ILF3-AS1*, *MIR31HG*, and *AC006329.1*) of the six-lncRNA signature of the discovery dataset were covered in TCGA dataset. So, the recurrence risk score model based only on these five lncRNAs without re-estimating parameters was used to predict recurrence risk and RFS for TCGA dataset, which perhaps reduces the predictive power. As shown in Figure 4A, risk score of recurrent patients are marginally significantly higher than non-recurrent patients (median 0.135 versus 0.004, p = 0.069, Mann-Whitney U test). The median risk score cutoff point obtained from TCGA dataset classified 391 patients into the high-risk group (n = 196) and the low-risk group (n = 195). Patients in the high-risk group had marginally significantly shorter RFS compared with those in the low-risk group (5-year RFS 55.7% versus 65%; p = 0.059, log rank test) (Figure 4B). The HR of predicted high-risk group versus low-risk group for RFS was 1.499 (p = 0.061; 95% CI: 0.981–2.289).

Independence of the Six-lncRNA Signature from Other Clinical Factors

To assess whether the predictive power of the six-lncRNA signature is independent of other clinical factors, we included the RRLnc6 score in a multivariate Cox regression analysis together with age, gender (male/female), and stage (IV/II). After multivariable adjustment by clinical factors, we found that both the RRLnc6 score (HR 2.043, 95% CI: 1.488–2.806, p < 0.0001 for the discovery dataset; HR 1.933, 95% CI: 1.410–2.650, p < 0.0001 for GSE39582; HR 1.673, 95% CI: 0.943–2.968, p = 0.079 for GSE 14333; HR 1.816, 95% CI: 1.052–3.132, p = 0.032 for GSE17538) and tumor stage (HR 2.210, 95% CI: 1.609–3.035, p < 0.0001 for the discovery dataset; HR 2.398, 95% CI: 1.749–3.287, p < 0.0001 for GSE39582; HR 3.455, 95% CI: 1.893–6.304, p = 0.0001 for GSE 14333; HR 1.650, 95% CI: 0.959–2.836, p = 0.070 for GSE17538) maintained significant or marginally significant correlation with RFS in four datasets (Table 2). In the testing dataset GSE33113, the RRLnc6 score (HR 3.499, 95% CI: 1.228–9.975, p = 0.019) still was significantly associated with RFS even if the tumor stages III and IV were missing (Table 2).

Then the stratification analysis was performed based on tumor stage. When stratified by tumor stage, patients of all datasets were stratified into two subgroups where stages I and II were included in the

Table 2. Univariate and Multivariate Cox Proportional Hazard Regression Analysis of Recurrence-free Survival in Each Dataset

	Univariate Analysis			Multivariate Analysis		
Variables	HR	95% CI	p Value	HR	95% CI	p Value
Discovery Dataset (n = 389)						
RRLnc6 score (high/low)	2.461	1.807–3.353	<0.0001	2.043	1.488–2.806	<0.0001
Age	1.002	0.990–1.014	0.746	1.003	0.991–1.016	0.593
Gender (male/female)	1.281	0.950–1.729	0.105	1.316	0.975–1.776	0.073
Stage (III, IV/I, II)	2.588	1.900–3.524	<0.0001	2.210	1.609–3.035	<0.0001
GEO: GSE39582 Dataset (n = 574)						
RRLnc6 score (high/low)	2.300	1.689–3.132	<0.0001	1.933	1.410–2.650	<0.0001
Age	0.999	0.988–1.010	0.921	1.001	0.990–1.013	0.828
Gender (male/female)	1.287	0.953–1.737	0.099	1.330	0.985–1.797	0.063
Stage (III, IV/I, II)	2.714	1.994–3.696	<0.0001	2.398	1.749–3.287	<0.0001
GEO: GSE14333 Dataset (n = 226)						
RRLnc6 score (high/low)	1.767	0.997–3.134	0.051	1.673	0.943–2.968	0.079
Age	0.981	0.961–1.001	0.069	0.988	0.968–1.009	0.251
Gender (male/female)	1.101	0.629–1.924	0.737	0.965	0.543–1.718	0.905
Stage (III, IV/I, II)	3.683	2.032–6.675	<0.0001	3.455	1.893–6.304	0.0001
GEO: GSE17538 Dataset (n = 200)						
RRLnc6 score (high/low)	1.831	1.067–3.141	0.028	1.816	1.052–3.132	0.032
Age	0.980	0.962–0.999	0.043	0.979	0.959–0.999	0.038
Gender (male/female)	0.750	0.441–1.276	0.289	0.645	0.368–1.130	0.125
Stage (III ,IV/I, II)	1.612	0.945–2.749	0.080	1.650	0.959–2.836	0.070
GEO: GSE33113 Dataset (n = 89)						
RRLnc6 score (high/low)	3.173	1.13–8.906	0.028	3.499	1.228–9.975	0.019
Age	0.980	0.946–1.015	0.254	0.971	0.936–1.007	0.118
Gender (male/female)	0.978	0.386–2.480	0.963	0.893	0.347–2.300	0.815

early-stage subgroup (n = 662) and stages III and IV included in the late-stage subgroup (n = 423). Patients with early-stage disease (I, II) had substantially lower RRLnc6 score compared with those with late-stage disease (III, IV) (median −0.036 versus 0.110; $p < 0.0001$, Mann-Whitney U test) (Figure 5A). Furthermore, there is a good correlation between the RRLnc6 score and tumor stage. The RRLnc6 score among patients with the different stage is also significantly different (median −0.168 for stage I, −0.006 for stage II, 0.092 for stage III, and 0.140 for stage IV; $p < 0.0001$, Kruskal-Wallis test) (Figure 5B). Further investigation showed that the RRLnc6 score of patients without recurrence was significantly lower than those with recurrence in both stage subgroups (median −0.111 versus 0.205, $p < 0.0001$ for early-stage subgroup; 0.051 versus 0.213, $p = 0.002$ for late-stage subgroup, Mann-Whitney U test) (Figures 5C and 5D). When the same recurrence risk score model and risk cutoff derived from the discovery dataset were applied to patients with early-stage disease and late-stage disease, the six-lncRNA signature could classify patients into high-risk and low-risk groups with significantly different RFS in both stage subgroups (5-year RFS 70.5% versus 85.7%, $p < 0.0001$ for early-stage subgroup; 49.5% versus 64.2%, $p = 0.003$ for late-stage subgroup, log rank test) (Figures 5E and 5F). Taken together, these results demonstrated that the six-lncRNA signature was an independent prognostic factor associated with disease recurrence and RFS of patients with colon cancer.

Comparison with Other Clinical Factors and Known Gene Signatures

To compare the predictive value of the six-lncRNA signature with that of clinical factors and known gene signatures used for recurrence risk stratification including age, gender, stage, and 13-mRNA signature, which is the latest mRNA-based signature that outperformed other gene classifiers (herein after referred to as 13mSig),[17] we performed receiver operator characteristic (ROC) curves analysis on all patients of four datasets. As shown in Figure 6A, the AUC of the six-lncRNA signature was 0.634, which was significantly larger than that of age (AUC = 0.550; $p = 0.003$) and gender (AUC = 0.518; $p < 0.0001$). When compared with the stage and 13mSig, the AUC of the six-lncRNA signature was much the same as that of the stage (0.634 versus 0.631; $p = 0.850$) and 13mSig (0.634 versus 0.645; $p = 0.638$). Further comparison of Kaplan-Meier survival curves showed that high-risk patients predicted by the six-lncRNA signature had a worse prognosis compared with those with

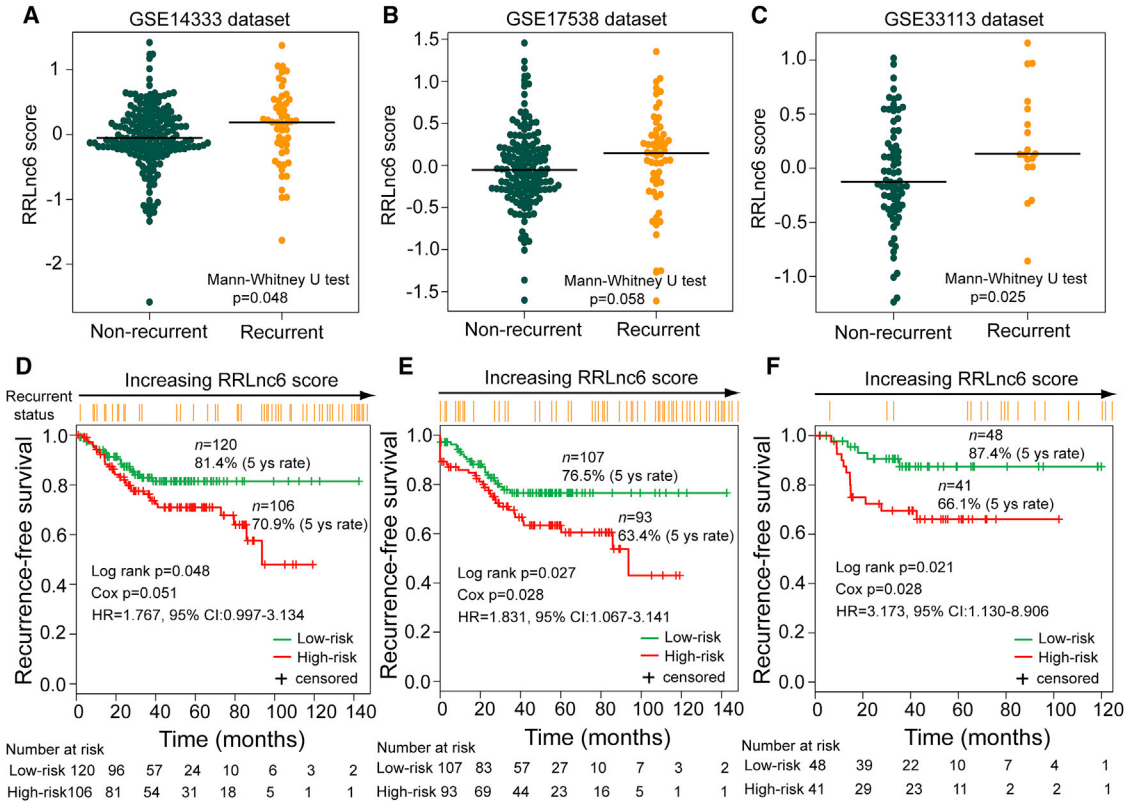


Figure 3. Independent Validation of the Six-lncRNA Signature in Three Independent Datasets

The distribution of RRLnc6 scores for recurrent patients and non-recurrent patients in the GSE14333 dataset (A), the GSE17538 dataset (B), and the GSE33113 dataset (C). Kaplan-Meier estimates of the recurrence-free survival of patients with low or high risk predicted by the six-lncRNA signature in the GSE14333 dataset (D), the GSE17538 dataset (E), and the GSE33113 dataset (F).

high-risk scores predicted by 13mSig and low-risk patients predicted by the six-lncRNA signature had a better prognosis compared with those with low-risk scores predicted by 13mSig (Figure 6B). In addition, the 5-year RFS rate of high-risk patients predicted by the six-lncRNA signature is 61.4%, which is lower than that (64%) of high-risk patients predicted by the 13mSig, whereas the corresponding rates for low-risk group predicted by the six-lncRNA signature is 78.5%, which is higher than that (77.6%) of the low-risk group predicted by the13mSig. These results indicated that the six-lncRNA signature had an equivalent or better predictive ability than stage and 13mSig.

Functional Implication of the Six-lncRNA Signature

We further investigate the potential functional mechanisms behind the altered expression of lncRNAs in the signature using "guilt by association." We first measured the expression correlation between lncRNAs in the signature and protein-coding genes (PCGs), and obtained lncRNA-correlated PCGs, which significantly co-expressed with that of at least one of the six lncRNAs in the signature. Then function enrichment analysis was performed for lncRNA-correlated PCGs. Gene Ontology (GO) analysis of lncRNA-correlated PCGs revealed a number of significantly enriched terms that can be categorized into four functional clusters including cell proliferation and angiogenesis, ATP metabolic process, cell death, and leukocyte differentiation (Figure 7A). Focusing on the biological pathways involved in lncRNA-correlated PCGs, we found 12 significantly enriched pathways, most of which are linked to tumor-promoting function, including Proteoglycans in cancer, focal adhesion, oxidative phosphorylation, extracellular matrix (ECM)-receptor interaction, regulation of actin cytoskeleton, pathways in cancer, cyclic AMP (cAMP) signaling pathway, and peroxisome proliferator-activated receptor (PPAR) signaling pathway (Figure 7B). These results indicated that an altered expression of lncRNAs in the signature may be involved in colon cancer biology through disrupting the balance of the lncRNA-related PCGs regulatory network to affect known critical biological pathways involved in cancer progression.

DISCUSSION

Although recent advances in large-scale sequencing and analyses have provided novel insights into the biology of colon cancer,[18] unfortunately, a large number of patients after curative surgery still faced a high risk of disease recurrence which is the major cause of cancer mortality. Colon cancer is a highly heterogeneous disease characterized by distinct genetic, epigenetic, and clinical properties.[19] Patients

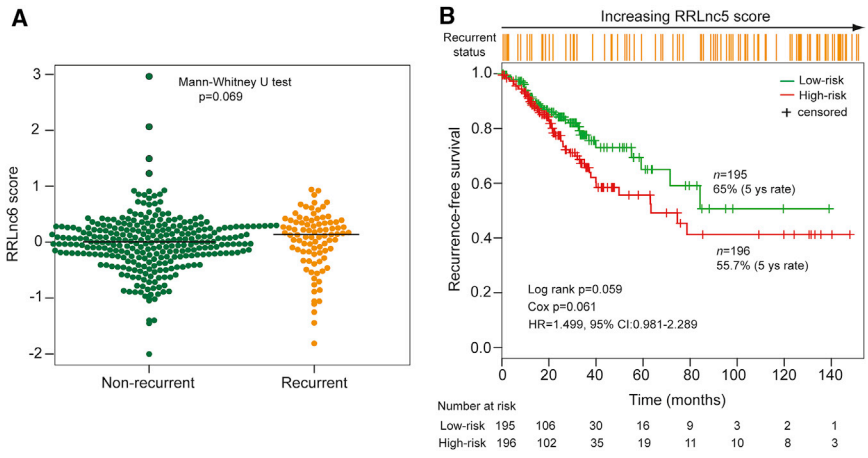


Figure 4. Independent Validation of the Six-lncRNA Signature in the TCGA Dataset

(A) The distribution of RRLnc5 score for recurrent patients and non-recurrent patients in the TCGA dataset. (B) Kaplan-Meier estimates of the recurrence-free survival of patients with low or high risk predicted by the five-lncRNA signature in the TCGA dataset.

with similar clinical and pathologic features often showed different recurrence risk and diverse clinical outcomes. Thus, traditional prognostic factors based on anatomical and pathological features, such as TNM staging system, tumor grade, lymphatic and vascular invasion, and so on, have revealed their limitations and insufficiency in the recurrence prediction of patients with colon cancer. During the past years, molecular profiles have demonstrated the potential as predictive and prognostic biomarkers to improve diagnosis, management, and treatment of cancer patients.[20–22] Several molecular signatures have been identified to predict recurrence of colon cancer, most of which are focused on mRNA expression.[17,23–25] However, up to now, whether a lncRNA signature might have similar predictive value in recurrence risk stratification to that of mRNA-focus expression signatures or patients with colon cancer is not known.

Consequently, in this study, we identified and validated a novel molecular signature, consisting of six lncRNAs (LINC0184, AC105243.1, LOC101928168, ILF3-AS1, MIR31HG, and AC006329.1) that can distinguish between colon cancer patients with high and low risks of cancer recurrence. For our predictive lncRNA selection for the signature, we performed an integrative computational and statistical strategy. We first determined the altered expression pattern of lncRNAs between patients with and without recurrence as candidate recurrence-associated lncRNAs. Then we further employed a random forest supervised classification algorithm and SVM to identify six predictive lncRNAs from those candidate recurrence-associated lncRNAs. Finally, we developed a molecular prognostic signature for recurrence risk with selected predictive lncRNAs. Moreover, we were able to validate the predictive value of the six-lncRNA signature in four additional datasets of patients with colon cancer, and demonstrated the necessity of all six predictive lncRNAs in the signature for recurrence risk prediction.

Although the TNM staging system is widely accepted to predict the prognosis and guide treatment decisions for patients with colon patients, there are critical limitations for the TNM staging system in the clinic because of molecular heterogeneity of colon cancer.[26] Our multivariate and stratified analysis suggested that the six-lncRNA signature not only was independent of clinicopathological factors, but also showed the ability in predicting recurrence for patients with the similar clinical stage. More recently, some studies have reported some mRNA-focus expression signature to identify patients at high risk of recurrence. Our comparative analysis demonstrated that the six-lncRNA signature has an equivalent or better ability in distinguishing patients at low risk versus high risk of tumor recurrence compared with mRNA-based signature. As previously reported, tumor-specific lncRNA protected from the RNases present in body fluids are contained in membranous particles released by tumor cells, and their expression is stable and easily detectable in plasma or other body fluids by qRT-PCR.[27] Moreover, lncRNA expression may be a better indicator of tumor status relative to mRNAs because their function is closely associated with their transcript abundance.[28,29] Given these advantages, combined with their more tissue- and cancer-type-specific manner relative to mRNAs, the six-lncRNA signature may be more easily applicable in clinics, which need to be evaluated in prospective cohorts and clinical trial.

Despite that increasing research has identified a huge number of lncRNAs in mammals using high-throughput experimental technologies, only a very small fraction of lncRNAs were well functionally characterized, and the functions of most lncRNAs remain largely unknown.[30] Among six lncRNAs in the signature, two lncRNAs (*ILF3-AS1* and *MIR31HG*) have been validated as potential prognostic markers in several cancers. lncRNA *ILF3-AS1* has been shown to be upregulated in melanoma, which promotes cell proliferation, migration, and invasion via negatively regulating miR-200b/a/429.[31] A recent study confirmed the oncogenic role of *MIR31HG* (also known as *LncHIFCAR*) by regulating the *HIF-1* transcriptional network, and revealed the potential utility of *MIR31HG* as an independent adverse prognostic predictor for the cancer progression.[32] For the remaining four lncRNAs in the signature, to our knowledge, no associations have been reported between these lncRNAs and cancer. Therefore, we performed *in silica* analysis to gain functional insight of the identified six-lncRNA signature. Functional analysis of lncRNAs based on co-expressed mRNAs with lncRNAs revealed a number of highly enriched tumor-promoting-related biological processes and pathways, including Proteoglycans in cancer,[33] focal adhesion,[34] oxidative phosphorylation,[35] ECM-receptor interaction,[36] regulation of actin cytoskeleton,[37] pathways in cancer,

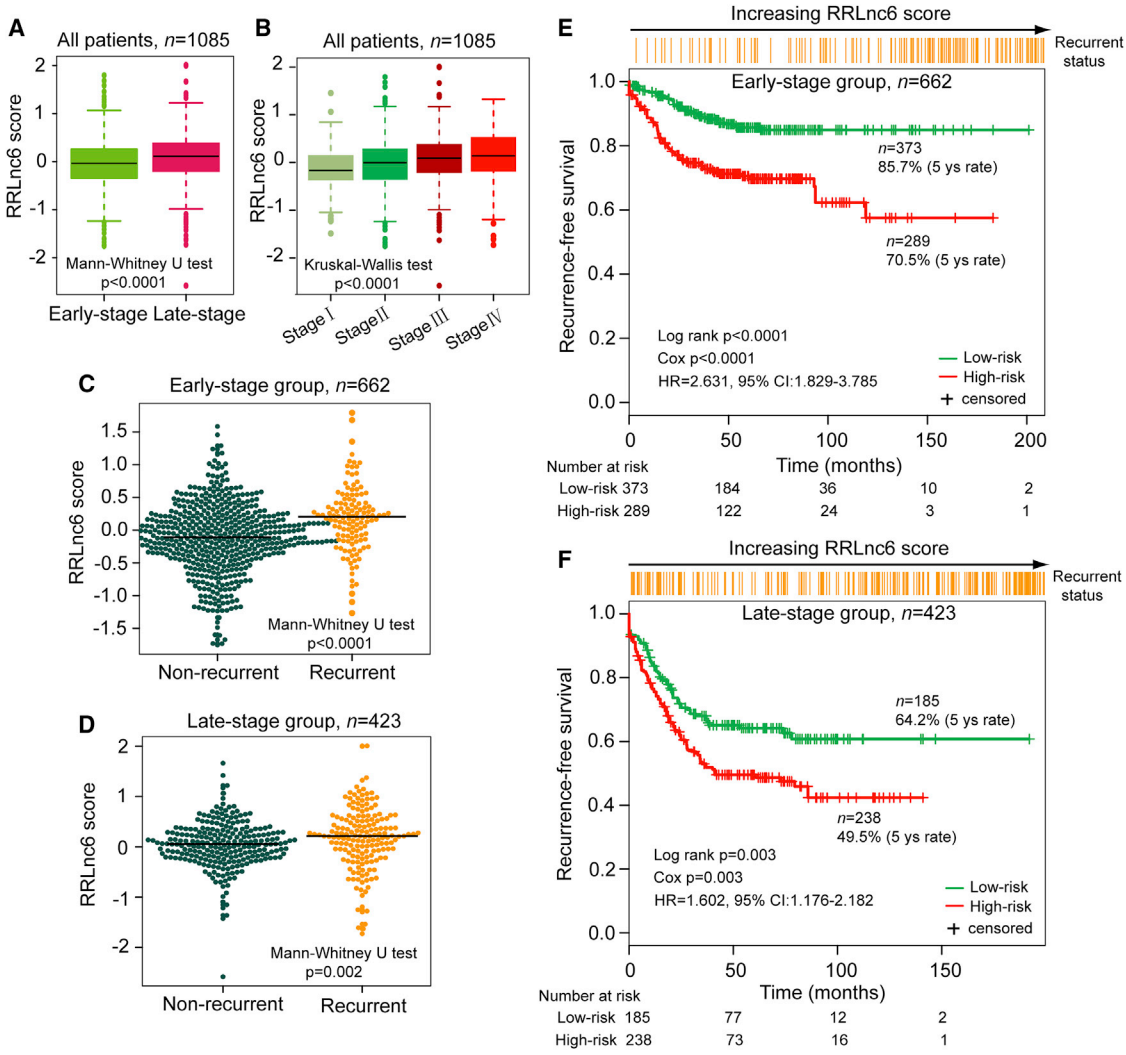


Figure 5. Risk Prediction of the Six-lncRNA Signature for Patients Stratified by Tumor Stage

(A) The distribution of RRLnc6 scores of early-stage patients and late-stage patients. (B) The distribution of RRLnc6 scores in patients from stages I–IV. (C) The distribution of RRLnc6 scores in recurrent patients and non-recurrent patients for the early-stage subgroup. (D) The distribution of RRLnc6 scores in recurrent patients and non-recurrent patients for the late-stage subgroup. (E) Kaplan-Meier estimates of the recurrence-free survival of patients with low or high risk predicted by the six-lncRNA signature in the early-stage subgroup. (F) Kaplan-Meier estimates of the recurrence-free survival of patients with low or high risk predicted by the six-lncRNA signature in the late-stage subgroup.

cAMP signaling pathway,[38] and PPAR signaling pathway.[39] Thus, our *in silica* analysis of the six-lncRNA signature suggested that variation in lncRNA expression might be involved in new critical biological processes and pathways involved in cancer progression, further supporting the relevance of the six-lncRNA signature in colon cancer recurrence.

In conclusion, we first determined differentially expression pattern of lncRNAs between patients with recurrence and recurrence-free patients, and identified recurrence-associated lncRNAs. Then we constructed a lncRNA expression signature composed of six recurrence-associated lncRNAs (*LINC0184*, *AC105243.1*, *LOC101928168*, *ILF3-AS1*, *MIR31HG*, and *AC006329.1*) and generated a recurrence risk score model (RRLnc6 score) that can effectively stratify colon cancer patients into groups with low and high risks of disease recurrence in the discovery dataset and three independent datasets. Moreover, the six-lncRNA signature is an independent predictive marker of disease recurrence and is superior to the performance of clinical factors and known gene signatures. With further prospective studies, the six-lncRNA signature not only holds great potential as a novel tool for recurrence risk assessment and personalized management of colon cancer patients, but also may present new insights into the mechanisms underlying colon cancer recurrence.

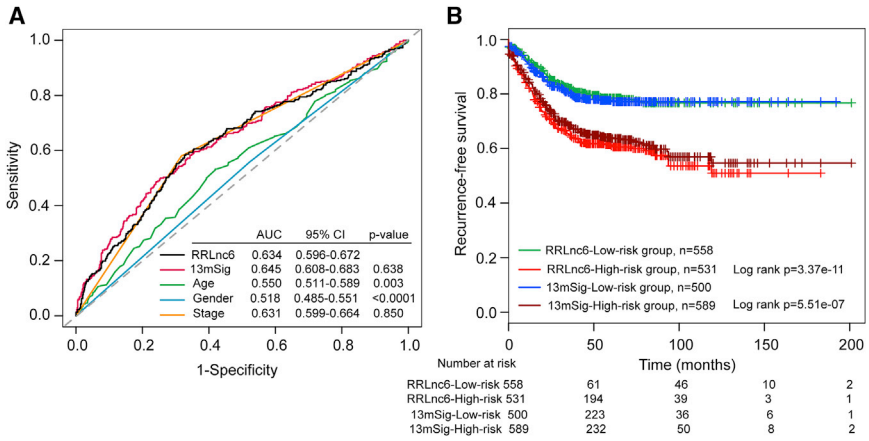


Figure 6. Comparison of Sensitivity and Specificity for Recurrence Risk Prediction by the Six-lncRNA Signature, Other Clinical Factors, and Known Gene Signature

(A) Receiver operating characteristic (ROC) curves of the six-lncRNA signature, other clinical factors, and 13mSig. (B) Comparison of recurrence-free survival differences in high- and low-risk groups predicted by 13mSig and the six-lncRNA signature.

MATERIALS AND METHODS

Colon Cancer Sample Datasets

We retrospectively collected clinical data and microarray data of colon cancer patient samples from five publicly available datasets from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/) and TCGA database (https://cancergenome.nih.gov/). After removing patient samples without recurrence information, a total of 1,480 colon cancer patient samples were enrolled from four GEO datasets and one TCGA dataset in this study, including 574 patients from the Marisa et al.[40] study (the accession number is GEO: GSE39582, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?Ac=GSE39582), 226 patients from the Jorissen et al.[41] study (the accession number is GEO: GSE14333 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?Ac=GSE14333), 200 patients from the Smith et al.[42] study (the accession number is GEO: GSE17538, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?Ac=GSE17538), 89 patients from the Sousa et al.[43] study (the accession number is GEO: GSE33113, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?Ac=GSE33113), and 391 patients from the National Cancer Institute Genomic Data Commons (NCI GDC) Data Portal. The median ages (range) and RFS time (range) of the five datasets were 68.1 years (22–97.0 years) and 44.5 months (0–201 months), 67 years (26–92 years) and 38.5 months (0.9–142.6 months), 66 years (23–94 years) and 39 months (0–142.6 months), 73 years (34–95 years) and 39.5 months (1.8–120 months), and 67 years (31–90 years) and 21.3 months (0–147.9 months), respectively. Of these patients, 179 (31.18%), 50 (22.12%), 55 (27.5%), 18 (20.22%), and 89 (22.76%) patients relapsed during follow-up, respectively. Detailed clinical characteristics of colon cancer patients are listed in Table S1.

Acquisition and Processing of lncRNA Expression Profiles

The raw data files (.CEL format) of four colon cancer patient datasets profiled by Affymetrix Human Genome U133 Plus 2.0 (Affymetrix HG-U133 Plus 2.0) were downloaded directly from the GEO database and were normalized using robust multichip average method by R 'affy' package for background correction, quantile normalization, and log2 transformation.[44] lncRNA expression profiles of four colon cancer patient datasets were obtained by superimposing data of the Affymetrix HG-U133 Plus 2.0 platform based on the NetAffx annotation files (HG-U133 Plus 2.0 Annotations, CSV format, release 36, 7/12/16) of the probe sets and the annotation files of RefSeq (release 79) and GENCODE (release 25) according to a previous study:[45] (1) we first extracted probe sets with RefSeq IDs that were labeled as "NR_" and annotated with "long non-coding RNA" in the RefSeq database (release 79); (2) we then extracted probe sets with Ensembl gene IDs, which were annotated as "long non-coding RNA" in the GENCODE project (release 25); and (3) finally, expression data of 2,466 unique lncRNAs corresponding to 3,431 probe sets of four colon cancer patient datasets were used for further analysis. To account for heterogeneity of multiple microarray datasets in systematic measurement, we standardized each dataset independently by the Z score transformation to scale expression intensities of each probe.[46]

Identification of Recurrence-Associated lncRNA Signature

To identify recurrence-associated lncRNAs, we compared expression differences between patients with recurrence and those without recurrence for >5 years for each individual lncRNA using variance filtering and SAM method. Differentially expressed lncRNAs (coefficient of variation [CV] value >10% and false discovery rate (FDR) adjusted $p < 0.05$) were defined as candidate recurrence-associated lncRNAs. Then recurrence-associated lncRNAs were identified from these candidate recurrence-associated lncRNAs using random forest supervised classification algorithm with an iteration procedure in which the one-third least important lncRNAs were discarded according to the permutation important score at each iteration step. Finally, SVM and 5-fold cross-validation were used to determine the best combination of the selected recurrence-associated lncRNAs as an optimal lncRNA signature for recurrence prediction by comparing classification accuracies of each combination in the discovery dataset.

Statistical Analyses

A risk score model for recurrence risk stratification was constructed using the sum of expression values of lncRNAs in the optimal lncRNA signature weighted by the estimated regression coefficients in the multivariate Cox regression analysis as described previously.[47,48] The risk score model was calculated for each patient and classified each patient into a low- or high-risk group using

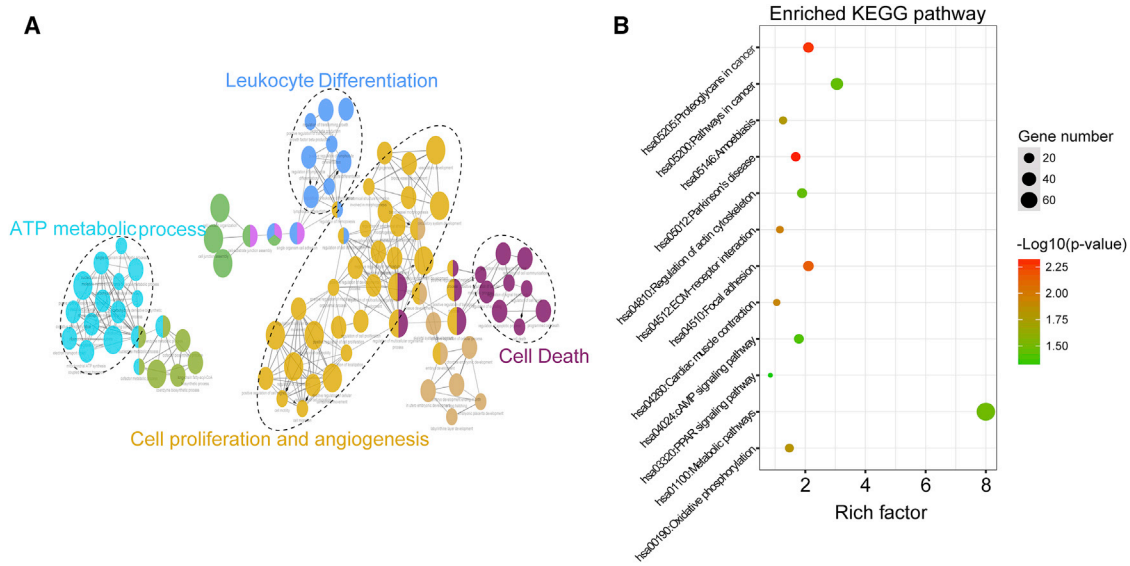


Figure 7. Functional Analysis of the Six-lncRNA Signature

(A) Functionally grouped network with enriched GO terms as nodes linked based on kappa score threshold of 0.4. Node size represents the term enrichment significance. (B) The most significantly enriched KEGG pathways. The node size represents the number of genes in the pathways, and the color represents the pathway enrichment significance.

the median risk score of the discovery dataset as a cutoff. Patients belonging to the low-risk group have a lower risk of recurrence and long-term RFS, and those belonging to the high-risk group have a higher risk of recurrence and short-term RFS. Kaplan-Meier survival curves and log rank tests were used to assess the differences in RFS of the predicted high-risk and low-risk groups. Univariate and multivariate analyses with Cox proportional hazards regression for RFS were performed on the clinical variables including age, gender (male versus female), stage (IV, III versus I, II), and risk score. HRs and 95% CIs were calculated. Hierarchical clustering of both patients and lncRNAs was performed with R software using the metric of Euclidean distance and complete linkage. The χ^2 test was used to evaluate the significance of differences in recurrence status between clustered patient groups. All statistical analyses were performed using R software and Bio-conductor.

Functional Enrichment Analysis

Functional enrichment analysis of GO and Kyoto encyclopedia of genes and genomes (KEGG) pathway was performed to determine significantly enriched GO terms and KEGG pathways of genes correlated with the six-lncRNA signature using the ClueGO plugin (version 2.3.3) in Cytoscape limited in biological processes[49] and DAVID Bioinformatics Resources (https://david.ncifcrf.gov/, version 6.8).[50] Functional map and clusters of enriched GO terms were obtained and visualized using a two-sided hypergeometric test with Bonferroni stepdown correction and kappa score threshold of 0.4, and limited in the GO level intervals 3–8 with minimum gene 20 and $p \leq 0.05$. Biological pathways with $p < 0.05$ was considered as significant using functional annotation chart options with the whole human genome as background.

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and one table and can be found with this article online at https://doi.org/10.1016/j.omtn.2018.06.007.

AUTHOR CONTRIBUTIONS

J. Sun and J. Su designed the study. M.Z., L.H., Z.Z., and N.W. performed data analysis. M.Z. and J. Sun drafted the manuscript. All authors read and approved the final manuscript.

CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (61602134), the Special Foundation for Key Basic Research of Wenzhou Institute of Biomaterials and Engineering, CAS, China (WIBEZD2017009-05), and the Scientific Research Foundation for Talents of Wenzhou Medical University (QTJ18023 and QTJ18024). The funders had no roles in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

1. Siegel, R.L., Miller, K.D., Fedewa, S.A., Ahnen, D.J., Meester, R.G.S., Barzi, A., and Jemal, A. (2017). Colorectal cancer statistics, 2017. CA Cancer J. Clin. *67*, 177–193.

2. Arnold, M., Sierra, M.S., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2017). Global patterns and trends in colorectal cancer incidence and mortality. Gut *66*, 683–691.

3. Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., and He, J. (2016). Cancer statistics in China, 2015. CA Cancer J. Clin. *66*, 115–132.

4. Schmoll, H.J., Van Cutsem, E., Stein, A., Valentini, V., Glimelius, B., Haustermans, K., Nordlinger, B., van de Velde, C.J., Balmana, J., Regula, J., et al. (2012). ESMO Consensus Guidelines for management of patients with colon and rectal cancer. a personalized approach to clinical decision making. Ann. Oncol. *23*, 2479–2516.

5. Miller, K.D., Siegel, R.L., Lin, C.C., Mariotto, A.B., Kramer, J.L., Rowland, J.H., Stein, K.D., Alteri, R., and Jemal, A. (2016). Cancer treatment and survivorship statistics, 2016. CA Cancer J. Clin. *66*, 271–289.

6. Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

7. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

8. Quinn, J.J., and Chang, H.Y. (2016). Unique features of long non-coding RNA biogenesis and function. Nat. Rev. Genet. *17*, 47–62.

9. Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nat. Struct. Mol. Biol. *20*, 300–307.

10. Hung, T., and Chang, H.Y. (2010). Long noncoding RNA in genome regulation: prospects and mechanisms. RNA Biol. *7*, 582–585.

11. Gutschner, T., and Diederichs, S. (2012). The hallmarks of cancer: a long non-coding RNA point of view. RNA Biol. *9*, 703–719.

12. Spizzo, R., Almeida, M.I., Colombatti, A., and Calin, G.A. (2012). Long non-coding RNAs and cancer: a new frontier of translational research? Oncogene *31*, 4577–4587.

13. Nana-Sinkam, S.P., and Croce, C.M. (2011). Non-coding RNAs in cancer initiation and progression and as novel biomarkers. Mol. Oncol. *5*, 483–491.

14. Luo, J., Qu, J., Wu, D.K., Lu, Z.L., Sun, Y.S., and Qu, Q. (2017). Long non-coding RNAs: a rising biotarget in colorectal cancer. Oncotarget *8*, 22187–22202.

15. Hu, Y., Chen, H.Y., Yu, C.Y., Xu, J., Wang, J.L., Qian, J., Zhang, X., and Fang, J.Y. (2014). A long non-coding RNA signature to improve prognosis prediction of colorectal cancer. Oncotarget *5*, 2230–2242.

16. Xue, W., Li, J., Wang, F., Han, P., Liu, Y., and Cui, B. (2017). A long non-coding RNA expression signature to predict survival of patients with colon adenocarcinoma. Oncotarget *8*, 101298–101308.

17. Tian, X., Zhu, X., Yan, T., Yu, C., Shen, C., Hu, Y., Hong, J., Chen, H., and Fang, J.Y. (2017). Recurrence-associated gene signature optimizes recurrence-free survival prediction of colorectal cancer. Mol. Oncol. *11*, 1544–1560.

18. Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

19. Linnekamp, J.F., Wang, X., Medema, J.P., and Vermeulen, L. (2015). Colorectal cancer heterogeneity and targeted therapy: a case for molecular disease subtypes. Cancer Res. *75*, 245–249.

20. Hanash, S.M., Baik, C.S., and Kallioniemi, O. (2011). Emerging molecular biomarkers—blood-based strategies to detect and monitor cancer. Nat. Rev. Clin. Oncol. *8*, 142–150.

21. Zhang, F., Ren, C., Lau, K.K., Zheng, Z., Lu, G., Yi, Z., Zhao, Y., Su, F., Zhang, S., Zhang, B., et al. (2016). A network medicine approach to build a comprehensive atlas for the prognosis of human cancer. Brief. Bioinform. *17*, 1044–1059.

22. Zhou, M., Zhao, H., Wang, X., Sun, J., and Su, J. (2018). Analysis of long noncoding RNAs highlights region-specific altered expression patterns and diagnostic roles in Alzheimer's disease. Brief. Bioinform. Published online April 17, 2018, https://doi.org/10.1093/bib/bby021.

23. Wang, L., Shen, X., Wang, Z., Xiao, X., Wei, P., Wang, Q., Ren, F., Wang, Y., Liu, Z., Sheng, W., et al. (2015). A molecular signature for the prediction of recurrence in colorectal cancer. Mol. Cancer *14*, 22.

24. Xu, G., Zhang, M., Zhu, H., and Xu, J. (2017). A 15-gene signature for prediction of colon cancer recurrence and prognosis based on SVM. Gene *604*, 33–40.

25. Alajez, N.M. (2016). Large-scale analysis of gene expression data reveals a novel gene expression signature associated with colorectal cancer distant recurrence. PLoS ONE *11*, e0167455.

26. Li, J., Yi, C.-H., Hu, Y.-T., Li, J.-S., Yuan, Y., Zhang, S.-Z., Zheng, S., and Ding, K.F. (2016). TNM staging of colorectal cancer should be reconsidered according to weighting of the T stage: verification based on a 25-year follow-up. Medicine (Baltimore) *95*, e2711.

27. Huarte, M. (2015). The emerging role of lncRNAs in cancer. Nat. Med. *21*, 1253–1261.

28. Du, Z., Fei, T., Verhaak, R.G., Su, Z., Zhang, Y., Brown, M., Chen, Y., and Liu, X.S. (2013). Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat. Struct. Mol. Biol. *20*, 908–913.

29. Hauptman, N., and Glavač, D. (2013). Long non-coding RNA in cancer. Int. J. Mol. Sci. *14*, 4655–4669.

30. Guo, X., Gao, L., Wang, Y., Chiu, D.K., Wang, T., and Deng, Y. (2016). Advances in long noncoding RNAs: identification, structure prediction and function annotation. Brief. Funct. Genomics *15*, 38–46.

31. Chen, X., Liu, S., Zhao, X., Ma, X., Gao, G., Yu, L., Yan, D., Dong, H., and Sun, W. (2017). Long non coding RNA ILF3-AS1 promotes cell proliferation, migration, and invasion via negatively regulating miR-200b/a/429 in melanoma. Biosci. Rep. *37*, BSR20171031.

32. Shih, J.W., Chiang, W.F., Wu, A.T.H., Wu, M.H., Wang, L.Y., Yu, Y.L., Hung, Y.W., Wang, W.C., Chu, C.Y., Hung, C.L., et al. (2017). Long noncoding RNA LncHIFCAR/MIR31HG is a HIF-1α co-activator driving oral cancer progression. Nat. Commun. *8*, 15874.

33. Afratis, N., Gialeli, C., Nikitovic, D., Tsegenidis, T., Karousou, E., Theocharis, A.D., Pavão, M.S., Tzanakakis, G.N., and Karamanos, N.K. (2012). Glycosaminoglycans: key players in cancer cell biology and treatment. FEBS J. *279*, 1177–1197.

34. Gómez Del Pulgar, T., Cebrián, A., Fernández-Aceñero, M.J., Borrero-Palacios, A., Del Puerto-Nevado, L., Martínez-Useros, J., Marín-Arango, J.P., Caramés, C., Vega-Bravo, R., Rodríguez-Remírez, M., et al. (2016). Focal adhesion kinase: predictor of tumour response and risk factor for recurrence after neoadjuvant chemoradiation in rectal cancer. J. Cell. Mol. Med. *20*, 1729–1736.

35. Maiuri, M.C., and Kroemer, G. (2015). Essential role for oxidative phosphorylation in cancer progression. Cell Metab. *21*, 11–12.

36. Jinka, R., Kapoor, R., Sistla, P.G., Raj, T.A., and Pande, G. (2012). Alterations in cell-extracellular matrix interactions during progression of cancers. Int. J. Cell Biol. *2012*, 219196.

37. Fife, C.M., McCarroll, J.A., and Kavallaris, M. (2014). Movers and shakers: cell cytoskeleton in cancer metastasis. Br. J. Pharmacol. *171*, 5507–5523.

38. Löffler, I., Grün, M., Böhmer, F.D., and Rubio, I. (2008). Role of cAMP in the promotion of colorectal cancer cell growth by prostaglandin E2. BMC Cancer *8*, 380.

39. Tachibana, K., Yamasaki, D., Ishimoto, K., and Doi, T. (2008). The role of PPARs in cancer. PPAR Res. *2008*, 102737.

40. Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.C., Schiappa, R., Guenot, D., Ayadi, M., et al. (2013). Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. *10*, e1001453.

41. Jorissen, R.N., Gibbs, P., Christie, M., Prakash, S., Lipton, L., Desai, J., Kerr, D., Aaltonen, L.A., Arango, D., Kruhøffer, M., et al. (2009). Metastasis-associated gene expression changes predict poor outcomes in patients with Dukes stage B and C colorectal cancer. Clin. Cancer Res. *15*, 7642–7651.

42. Smith, J.J., Deane, N.G., Wu, F., Merchant, N.B., Zhang, B., Jiang, A., Lu, P., Johnson, J.C., Schmidt, C., Bailey, C.E., et al. (2010). Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology *138*, 958–968.

43. de Sousa E Melo, F., Colak, S., Buikhuisen, J., Koster, J., Cameron, K., de Jong, J.H., Tuynman, J.B., Prasetyanti, P.R., Fessler, E., van den Bergh, S.P., et al. (2011). Methylation of cancer-stem-cell-associated Wnt target genes predicts poor prognosis in colorectal cancer patients. Cell Stem Cell *9*, 476–485.

44. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249–264.

45. Zhang, X., Sun, S., Pu, J.K., Tsang, A.C., Lee, D., Man, V.O., Lui, W.M., Wong, S.T., and Leung, G.K. (2012). Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiol. Dis. *48*, 1–8.

46. Zhou, M., Xu, W., Yue, X., Zhao, H., Wang, Z., Shi, H., Cheng, L., and Sun, J. (2016). Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma. Oncotarget *7*, 29720–29738.

47. Zhou, M., Zhang, Z., Zhao, H., Bao, S., Cheng, L., and Sun, J. (2018). An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme. Mol. Neurobiol. *55*, 3684–3697.

48. Zhou, M., Zhang, Z., Zhao, H., Bao, S., and Sun, J. (2018). A novel lncRNA-focus expression signature for survival prediction in endometrial carcinoma. BMC Cancer *18*, 39.

49. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics *25*, 1091–1093.

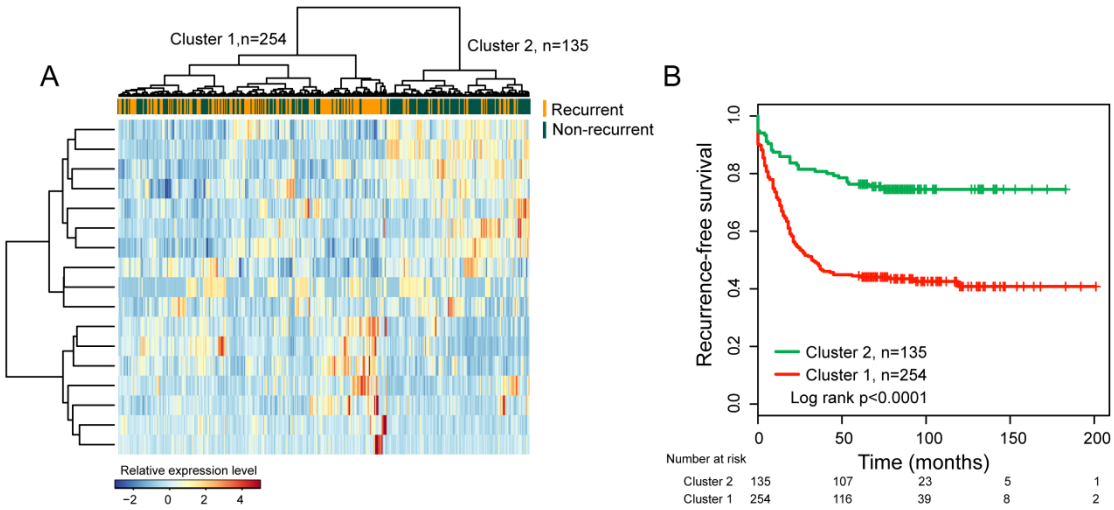
50. Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

OMTN, Volume *12*

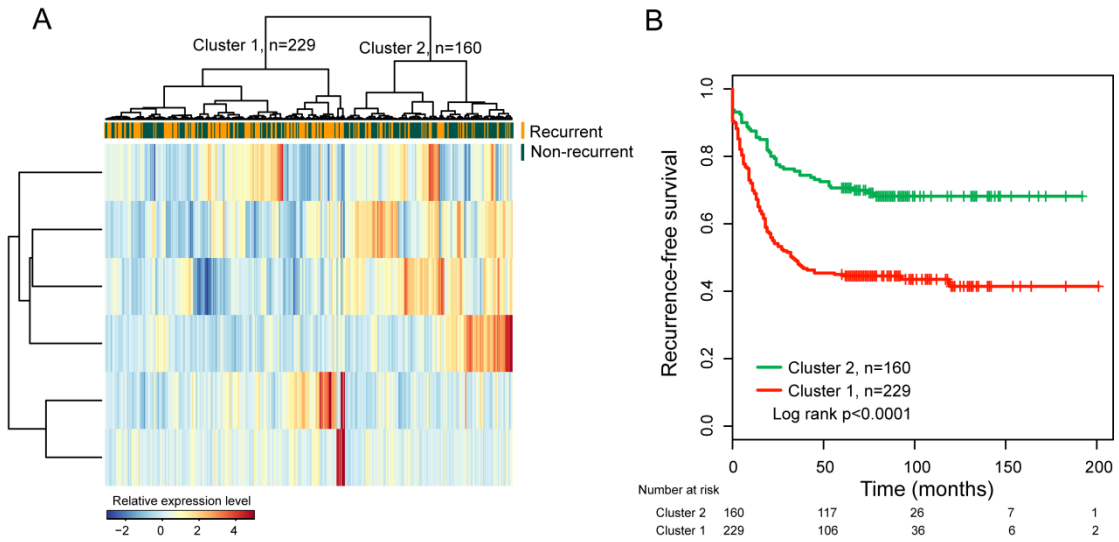
Supplemental Information

Recurrence-Associated Long Non-coding RNA Signature for Determining the Risk of Recurrence in Patients with Colon Cancer

Meng Zhou, Long Hu, Zicheng Zhang, Nan Wu, Jie Sun, and Jianzhong Su



Supplemental Figure S1. (A) Unsupervised clustering of patients based on the expression pattern of 17 recurrence-associated lncRNAs. (B) Kaplan-Meier estimates of the recurrence-free survival of patients in the two sample clusters based on 17 recurrence-associated lncRNAs.



Supplemental Figure S2. (A) Unsupervised clustering of patients based on the expression pattern of six lncRNAs in the signature. Recurrence. (B) Kaplan-Meier estimates of the recurrence-free survival of patients in the two sample clusters based on six lncRNAs in the signature.

Supplemental Table S1. Clinical characteristics of patients with colon cancer in the six datasets

Characteristics	Discovery dataset	GSE39582 dataset	GSE14333 dataset	GSE17538 dataset	GSE33113 dataset	TCGA dataset
No. of patients	389	574	226	200	89	391
Age (years)						
Range	30.2-97	22-97	26-92	23-94	34-95	31-90
Median	68	68.1	67	66	73	67
Gender						
Female	175	257	120	98	47	181
Male	214	317	106	102	42	210
Stage						
I/II	196	304	135	134	89	227
III/IV	193	266	91	66	0	155
Unknown		4				9
Recurrence						
Yes	179	179	50	55	18	89
No	210	395	176	145	71	302
Recurrence-free survival (months)						
Range	0-201	0-201	0.9-142.6	0-142.6	1.8-120	0-147.9
Median	65	44.5	38.5	39	39.5	21.3