

SCIENTIFIC REPORTS



OPEN

Identification of a six-lncRNA signature associated with recurrence of ovarian cancer

Kai Yang¹, Yan Hou^{1,2}, Ang Li¹, Zhenzi Li¹, Wenjie Wang¹, Hongyu Xie¹, Zhiwei Rong¹, Ge Lou³ & Kang Li¹

Ovarian cancer (OvCa) is the leading cause of death among all gynecological malignancies, and recurrent OvCa is almost always incurable. In this study, we developed a signature based on long non-coding RNAs (lncRNAs) associated with OvCa recurrence to facilitate personalized OvCa therapy. lncRNA expression data were extracted from GSE9891 and GSE30161. LASSO (least absolute shrinkage and selection operator) penalized regression was used to identify an lncRNA-based signature using the GSE9891 training cohort. The signature was then validated in GSE9891 internal and GSE30161 external validation cohorts. The Database for Annotation, Visualization and Integrated Discovery (DAVID) was used to explore the possible functions of identified lncRNAs. A six-lncRNA signature (RUNX1-IT1, MALAT1, H19, HOTAIRM1, LOC100190986 and AL132709.8) was identified in the training cohort and validated in internal and external validation cohorts using the LASSO method ($P < 0.05$). This signature was also independent of other clinical factors according to multivariate and sub-group analyses. The identified lncRNAs are involved in cancer-related biological processes and pathways. We selected a highly reliable signature based on six lncRNAs associated with OvCa recurrence. This six-lncRNA signature is a promising method to personalize ovarian cancer therapy and may improve patient quality of life quality according to patients' condition in the future.

Ovarian cancer (OvCa) is a common gynecological malignancy and the commonest cause of gynecological cancer-associated deaths in developed countries¹. It is estimated that there will be 22,280 new cases and 14,240 deaths attributed to OvCa in the United States in 2016². Although there is a high initial response rate to standard surgery and chemotherapy, most OvCa patients will develop recurrence within 18 months after standard first-line treatment³. More seriously, recurrent OvCa usually develops into platinum-resistant disease and is almost always incurable⁴. Therefore, stratification of patients to identify high-risk patients may provide more effective treatment strategies and personalized therapies.

Long non-coding RNA (lncRNA) is a class of non-coding RNAs that are longer than 200 nucleotides in length⁵. Increasing studies have showed that abnormal expression of lncRNAs is associated with human cancers and that some play important roles in a variety of biological processes in cancer. Currently, several lncRNA-based signatures have been identified as prediction of patient survival in several cancers, such as gastric cancer⁶, lung cancer⁷, breast cancer⁸, colorectal cancer⁹ and esophageal squamous cell cancer¹⁰. Recent studies also indicated that lncRNAs were associated with OvCa recurrence and survival¹¹. For example, lncRNAs CCAT2, HOTAIR, AB073614, and ANRIL have been demonstrated to be associated with poor prognosis of OvCa¹²⁻¹⁵. A recent study identified an eight-lncRNA signature associated with overall survival (OS) of OvCa based on The Cancer Genome Atlas (TCGA)¹⁶.

In this study, we used LASSO (least absolute shrinkage and selection operator) penalized regression to identify a six-lncRNA signature associated with OvCa recurrence based on a training cohort. Then we validated it in internal and external validation cohorts, analyzed it in sub-groups of OvCa patients, and demonstrated this signature was independent from other clinical factors. We also analyzed the correlation between the signature and OS of OvCa patients. Furthermore, we found that these lncRNAs were involved in biological processes (e.g.,

¹Department of Epidemiology and Biostatistics, School of Public Health, Harbin Medical University, Harbin, 150086, P.R. China. ²Key Laboratory of Cardiovascular Medicine Research, Harbin Medical University, Ministry of Education, Harbin, 150086, P.R. China. ³Department of Gynecology Oncology, the Tumor Hospital, Harbin Medical University, Harbin, 150086, P.R. China. Kai Yang and Yan Hou contributed equally to this work. Correspondence and requests for materials should be addressed to K.L. (email: likang@ems.hrbmu.edu.cn) or G.L. (email: louge@ems.hrbmu.edu.cn)

Probe	Gene symbol	Chromosomal location	C ^a	C ^b	P ^b	HR ^b
220918_at	RUNX1-IT1	Chr21q22.12	0.1078	0.5589	<0.0001	1.749
223940_x_at	MALAT1	Chr11q13.1	0.0751	0.1514	0.0410	1.163
224646_x_at	H19	Chr11p15.5	0.1083	0.1625	0.0004	1.176
228642_at	HOTAIRM1	Chr7p15.2	0.1110	0.2776	0.0011	1.320
235167_at	LOC100190986	Chr16p12.2	-0.0155	-0.3217	0.0061	0.725
242856_at	AL132709.8	Chr14q32.31	0.0195	0.2975	0.0016	1.346

Table 1. Overall information of six prognostic lncRNAs associated with DFS in GSE9891 training cohort (n = 100). ^aDerived from the LASSO penalized regression in 100 patients of GSE9891 training cohort.

^bDerived from the univariate Cox proportional hazards regression in 100 patients of GSE9891 training cohort. Abbreviations: C Coefficient, P P value, HR Hazard Ratio.

cell adhesion, inflammatory response and immune response) and pathways (e.g., ECM-receptor interaction, focal adhesion and cell adhesion molecules) in cancers. Thus, our six-lncRNA signature may be a promising method to stratify OvCa patients and identify those at high-risk in the future.

Results

Demographic and clinical characteristics. The detailed demographic and clinical characteristics are listed in Supplementary Table S1. A total of 311 OvCa patients were included in our study, including 100 in the GSE9891 training cohort, 157 in the GSE9891 internal validation cohort and 54 in the GSE30161 external validation cohort. The median ages (ranges) of the three cohorts were 58 (23–80), 60 (33–80), and 62 (38–84) years, respectively. Seventy-four (74%), 111 (71%), and 48 (89%) patients relapsed during follow-up, respectively. The tumor stage, tumor grade and histology type are also summarized in Supplementary Table S1.

Identification of lncRNA signature and generation of risk score. To identify lncRNAs associated with OvCa recurrence, LASSO penalized regression was performed using lncRNA expression data. This method can select an optimal subset of lncRNAs without collinearity by imposing a penalty and shrinking most regression coefficients to zero. After 100 times of 10-fold cross validation, the optimal tuning parameter lambda1 was 16.9841 in our study. As a result, the regression coefficients of six lncRNAs were not zero when lambda1 was 16.9841, and we selected these six lncRNAs as signatures associated with OvCa recurrence (Table 1). We also conducted univariate cox regression for the six lncRNAs, in which the regression coefficients were consistent with that in LASSO penalized regression, and all six lncRNAs were statistically significant (P < 0.05). Of these six lncRNAs, five were positively associated with DFS (high expression of these lncRNAs led to a high-risk score and shorter survival) and one was negatively associated with DFS (the high expression of this lncRNA led to a low-risk score and longer survival).

To identify low- and high-risk patients for OvCa recurrence, we developed a prognostic model based on the expression of six lncRNAs and their regression coefficient in LASSO penalized regression as follows: Risk score = (0.1078*RUNX1-IT1) + (0.0751*MALAT1) + (0.1083*H19) + (0.111*HOTAIRM1) - (0.0155*LOC100190986) + (0.0195*AL132709.8). The patients were then divided into low- and high-risk groups according to the median risk score value (2.9232). As a result, the patients in the low-risk group had a better survival outcome than the high-risk group, as shown in Fig. 1a (P < 0.0001). The area under the curve (AUC) of time-dependent ROC curves for the risk score was 0.813 at three years (Fig. 1c). These results demonstrated a good performance of our six-lncRNA signature in predicting DFS for OvCa patients. Risk scores and relative expression levels of all patients are shown in Fig. 1b,d.

Validation of the six-lncRNA signature in the GSE9891 internal validation and entire cohorts. To confirm the ability of the six-lncRNA signature in predicting DFS for OvCa patients, we validated it in the GSE9891 internal validation and entire cohorts. The same risk formula and cutoff value were used to calculate risk scores and divide the patients into low- and high-risk groups. The results of the two cohorts were consistent with the GSE9891 training cohort. Patients with higher risk scores had poorer DFS. The differences of survival curves between the two groups were statistically significant in two cohorts (Fig. 2a and Supplementary Fig. S1a). The AUCs of the two cohorts were 0.665 and 0.697, respectively (Fig. 2c and Supplementary Fig. S1c). Risk scores and relative expression levels of all patients in the two cohorts are separately shown in Fig. 2b,d and Supplementary Fig. S1b,d.

Further validation of the six-lncRNA signature in the GSE30161 external validation cohort. We further validated our findings in GSE30161 external validation cohort. The same risk formula was used to calculate risk scores for every OvCa patient. The median risk score value in GSE30161 (2.2672) was used to divide the patients into low- and high-risk scores. Similar to the result in GSE9891, the patients in low-risk group also had a better survival outcome than those in the high-risk group, as shown in Fig. 3a (P = 0.0114). The AUC of time-dependent ROC curves for the risk score was 0.711 at three years (Fig. 3c). Risk scores and relative expression levels of all patients are also shown in Fig. 3b,d.

Sub-group analysis of six-lncRNA signature in GSE9891 and GSE30161. We then analyzed the six-lncRNA signature in sub-groups of OvCa patients. As shown in Fig. 4, the differences of K-M survival curves between the low- and high-risk patients were statistically significant (P < 0.05) for late-stage (Fig. 4b), low-grade

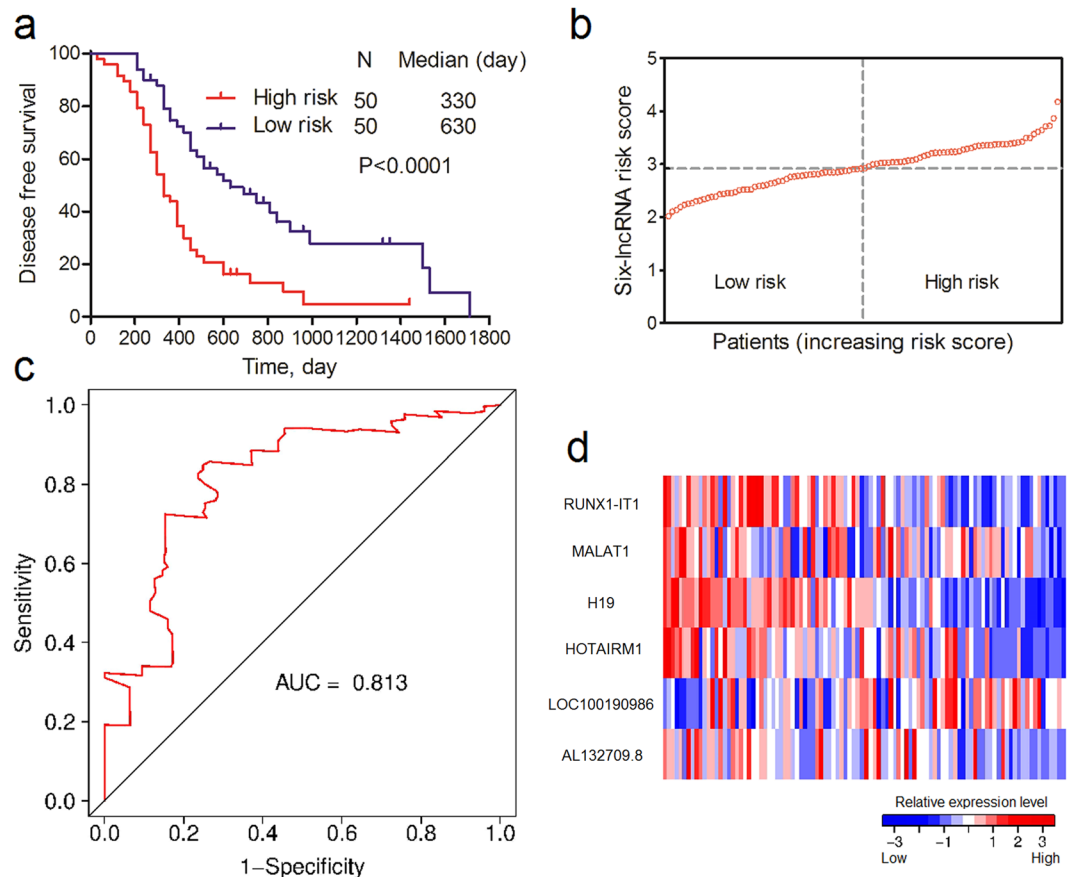


Figure 1. Association between the six-lncRNA signature and DFS of OvCa patients in GSE9891 training cohort. **(a)** K-M curve of DFS between low- and high-risk patients. **(b)** Risk scores of each patient in the GSE9891 training cohort (sorted by risk score). **(c)** Time-dependent ROC curve analysis of the DFS prediction based on the risk score with three years as the time point. **(d)** Expression heat map of the six lncRNAs in OvCa patients in the GSE9891 training cohort (sorted by risk score).

(Fig. 4c), high-grade (Fig. 4d) OvCa patients in the GSE9891 and low-grade (Fig. 4e) OvCa patients in the GSE30161. For the early-stage patients in GSE9891 (Fig. 4a) and high-grade patients in GSE30161 (Fig. 4f), the differences of K-M survival curves were not statistically significant, but patients in the low-risk group still tended to have a better DFS. The ROCs of these sub-groups were in Supplementary Fig. S2.

Independence of six-lncRNA signature and other clinical factors. To assess whether the six-lncRNA signature was independent of other clinical factors, univariate and multivariate cox regression analyses were conducted in each patient cohort including risk scores for the six lncRNAs, age, tumor stage, tumor grade and histology type. In univariate cox regression, risk score and tumor stage were significantly associated with DFS. After adjusting for age, tumor stage, tumor grade and histology type, risk score still maintained a significant correlation with DFS in all GSE9891 and GSE30161 cohorts (Table 2).

Correlation between the six-lncRNA signature and OS. In addition to the analyses of association between the signature and OvCa recurrence, we also explored the correlation between the signature and OS. In accordance with the results of DFS, patients with higher risk scores had poorer OS. The differences of survival curves between the two groups were statistically significant in GSE9891 training, GSE9891 entire and GSE30161 external validation cohorts (Supplementary Fig. S3a,c,d). The differences of K-M survival curves were not statistically significant in GSE9891 internal validation cohort ($P = 0.0612$), but patients in the low-risk group tended to have a better OS (Supplementary Fig. S3b). After adjusting for age, tumor stage, tumor grade and histology subtype, risk score still maintained a significant correlation with OS in GSE9891 training and entire cohorts. The patients in the low-risk group in GSE9891 internal validation and GSE30161 external validation cohorts still tended to have a better OS ($P < 0.08$) (Supplementary Table S2).

Functional characteristics of six lncRNAs. We further analyzed the possible functions associated with the six identified lncRNAs in OvCa by GO and KEGG functional enrichment analysis using the DAVID tool. Using Satterthwaite t-test and FDR correction, 3814 genes were found to be differentially expressed between the low- and high-risk groups. These DEGs were considered as genes associated with six lncRNAs. After functional annotation in DAVID, 68 GO biological process (BP) terms, 27 GO cellular component (CC) terms,

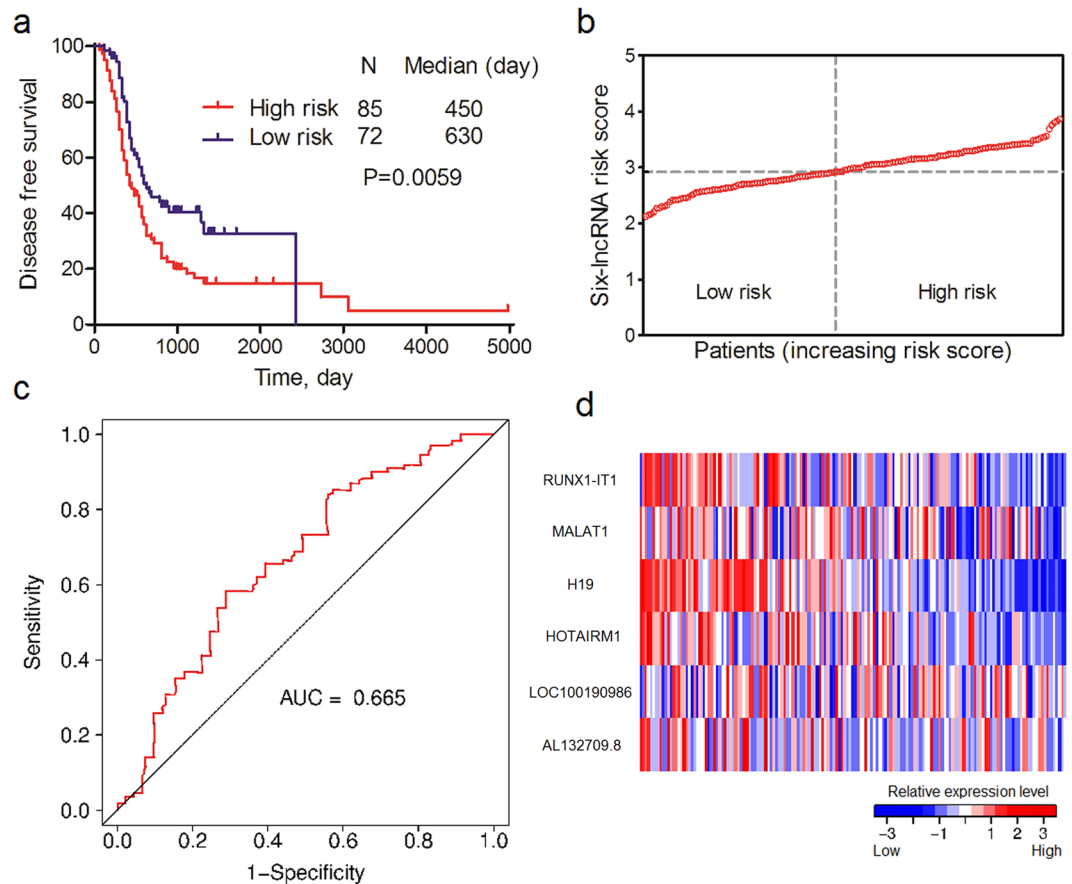


Figure 2. Association between the six-lncRNA signature and DFS of OvCa patients in GSE9891 internal validation cohort. **(a)** K-M curve of DFS between low- and high-risk patients. **(b)** Risk scores of each patient in the GSE9891 internal validation cohort (sorted by risk score). **(c)** Time-dependent ROC curve analysis of the DFS prediction based on the risk score with three years as the time point. **(d)** Expression heat map of six lncRNAs in OvCa patients in the GSE9891 internal validation cohort (sorted by risk score).

18 GO molecular function (MF) terms, and 6 KEGG pathways were significantly enriched with $FDR < 0.05$ (Supplementary Tables S3–S6). The enriched BP terms were involved in some important biological processes in cancer, such as blood vessel development, inflammatory response and immune response (Table 3). The enriched KEGG pathways were also involved in cancer-related pathways, such as ECM-receptor interaction, focal adhesion and cell adhesion molecules (Table 3). An interaction network of significant BP terms with similar function showed the six lncRNAs were mainly associated with inflammatory response, immune system process, cell migration, cell adhesion, angiogenesis and extracellular matrix organization (Fig. 5). Most of the enriched GO terms and KEGG pathways have been found to be closely associated with OvCa initiation and progression, which could increase the credibility of the six-lncRNA signature from a biological perspective.

Discussion

OvCa is the sixth commonest cause of female cancer death in developed countries¹. Approximately 75% of women with OvCa present with late-stage disease, most (78.79% in this study) of whom will develop recurrence^{4,17}. Treatment of OvCa is based on a combination of surgery and chemotherapy, and standard treatment for late-stage OvCa combined platinum-based chemotherapy with cytoreductive surgery can achieve a good result¹⁸. However, for most late-stage patients, cancer recurrence seems to be inevitable. Moreover, recurrent OvCa is often insensitive to chemotherapy and is generally incurable⁴. Thus, it is of great importance to identify low- and high-risk patients to enable improved treatment. Once we can identify high-risk patients, we can adopt a more effective personalized treatment strategy, such as novel drugs and targeted therapies, to delay the recurrence of cancer and improve patient quality of life¹⁸.

In our study, LASSO penalized regression was used to identify six lncRNAs associated with OvCa recurrence. LASSO penalized regression is optimized to high-dimensional data, which can select a few of the most influential variables and avoid collinearity of variables at the same time. These properties are not available in many other univariate and multivariate methods. lncRNA may be superior biomarkers in cancer as it is a new area of molecular biology, does not encode proteins, and may have better specificity¹⁹. In addition to the inherent virtues of LASSO and lncRNAs, we also validated the six-lncRNA signature in internal and external cohorts, and certified

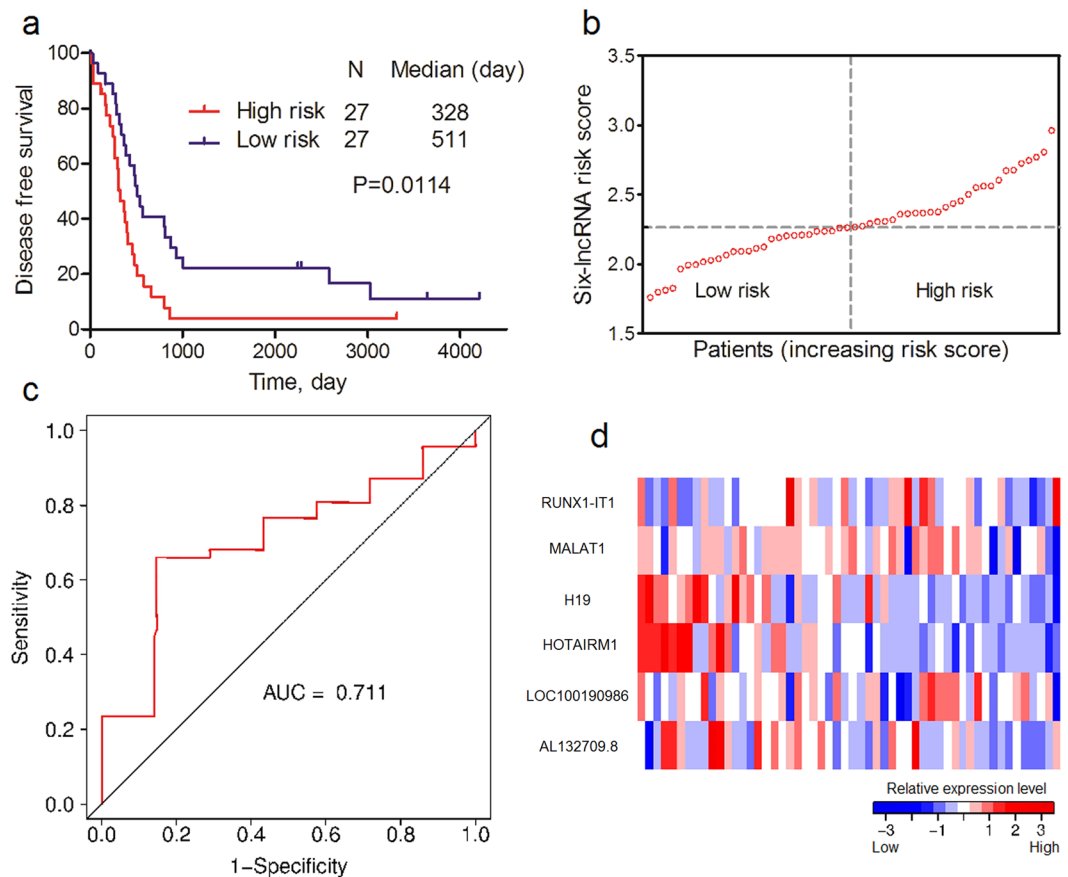


Figure 3. Association between six-lncRNA signature and DFS of OvCa patients in GSE30161 external validation cohort. (a) K-M curve of DFS between low- and high-risk patients. (b) Risk scores of each patient in GSE9891 internal validation cohort (sorted by risk score). (c) Time-dependent ROC curve analysis of the DFS prediction based on the risk score with three years as the time point. (d) Expression heat map of six lncRNAs in OvCa patients in the GSE30161 external validation cohort (sorted by risk score).

its credibility from a biological perspective. The signature also showed a great ability to stratify OvCa patients into low- and high-risk subgroup with significantly different overall survival.

Further analysis, including sub-group analysis and adjustment of clinical factors, indicated that the six-lncRNA signature could still predict recurrence in most sub-groups and were independent of other clinical factors, including age, tumor stage, tumor grade and histology type. It is well known that the prognosis of late-stage OvCa patients is worse than early-stage OvCa patients (Fig. 4a,b). In our study, the six-lncRNA signature could identify all high-risk patients with late-stage OvCa in different cohorts (Figs 3a and 4b). This could be of great use to improve the prognosis of these severe cancer patients.

We also explored the possible functions of these genes. Although the exact mechanism of most of the six lncRNAs was unclear, the lncRNAs were still closely related with cancer, as described in the literature, which strengthened the reliability and possibility of our six-lncRNA signature as a predictor of OvCa recurrence. In addition to the reports in the literature, our functional enrichment analysis of the six lncRNA-related genes may also shed new light on the possible functions of these lncRNAs in OvCa.

RUNX1-IT1, MALAT1, H19, and HOTAIRM1 have been widely associated with cancer in recent years. RUNX1-IT1 is an oncogenic lncRNA that can promote tumor progression and metastasis²⁰. RUNX1-IT1 was overexpressed in non-responder chronic myeloid leukemia²¹. MALAT1 was found to be upregulated in a variety of human cancers, such as lung cancer, breast cancer, prostate cancer, colon cancer, and liver cancer^{22–24}. Some studies showed this lncRNA was involved in the regulation of cell mobility²⁵, which was consistent with our findings about cell migration (Fig. 5). H19 has been widely linked to oncogenesis, although the exact mechanism remains unclear²⁶. H19 is a precursor of mir-675, which down-regulates tumor suppressor genes in cancer²⁷. H19 is also up-regulated in basal cell cancer compared with normal skin specimens²⁸, and is associated with poor prognosis in breast cancer²⁹. However, it should be noted that H19 is down-regulated in high-risk patients in our study, which is inconsistent with these studies. HOTAIRM1 is a kind of lncRNA that plays an important role in the development of immune cells³⁰. HOTAIRM1 is also reported to be a tumor suppressor by affecting a series of genes related to cell proliferation in colon cancer³¹.

There were still relatively few studies describing LOC100190986 and AL132709.8 in cancer research. LOC100190986 is associated with HCV genotype 1b transfection in the HepG2 cell line³². AL132709.8 was up-regulated in neural precursor cells from patients of lethal congenital contracture syndrome³³.

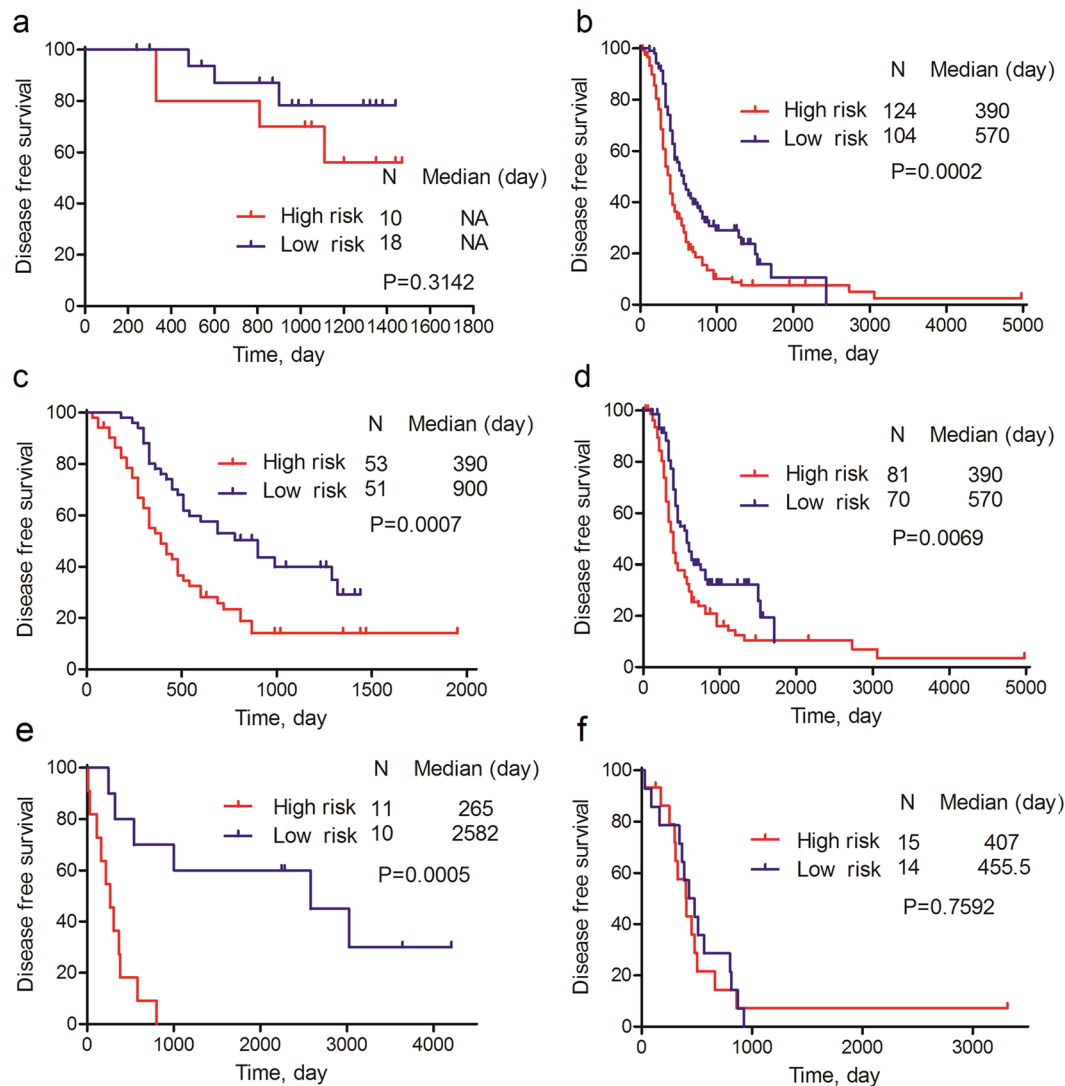


Figure 4. Sub-group analysis of association between six-lncRNA signature and DFS of OvCa patients. **(a)** K-M curve of early-stage OvCa patients in GSE9891. **(b)** K-M curve of late-stage OvCa patients in GSE9891. **(c)** K-M curve of low-grade OvCa patients in GSE9891. **(d)** K-M curve of high-grade OvCa patients in GSE9891. **(e)** K-M curve of low-grade OvCa patients in GSE30161. **(f)** K-M curve of high-grade OvCa patients in GSE30161.

The selected six lncRNAs were highly reliable from the perspective of both statistics and biology, since we conducted rigorous internal validation, external validation, and biological explanation. However, our study has some limitations. First, we used different cutoff values for GSE9891 and GSE30161. The overall expression levels of lncRNAs in GSE30161 were lower than that of GSE9891. The main reason for this phenomenon was the different experimental conditions in different labs. We can solve this problem by using a rigorous experimental procedure and adopting a unique cut-off value in the future. Second, the differences of K-M survival curves between low- and high-risk patients were not statistically significant for early-stage patients in GSE9891 (Fig. 4a) and high-grade patients in GSE30161 (Fig. 4f), although patients in the low-risk group still tended to have better DFS. This may be because the sample sizes of these sub-groups were too small (28 cases and 29 cases, respectively). Hence further studies are needed to validate the signature in these groups.

We applied the LASSO method to high-dimensional lncRNA expression data and identified a six-lncRNA signature that was highly associated with OvCa recurrence. This signature was validated in internal and external validation cohorts and was independent of other clinical factors. Furthermore, from a literature review and our functional analysis, we found that the six lncRNAs were closely related with cancer. Thus, this six-lncRNA signature may be a promising method to personalize OvCa therapy and improve patient quality of life according to patients' condition in the future.

Methods

OvCa patient dataset and clinical information. Microarray data for GSE9891 and GSE30161 were downloaded from the Gene Expression Omnibus (GEO; <https://www.ncbi.nlm.nih.gov/geo/>)^{34, 35}. Clinical

Variable	Univariate analysis				Multivariate analysis			
	C	P	HR	95% CI of HR	C	P	HR	95% CI of HR
GSE9891 training cohort (N = 100)								
Risk score	1.667	<0.0001	5.296	2.925–9.589	1.6849	<0.0001	5.392	2.84–10.237
Age	−0.0062	0.6223	0.994	0.970–1.019	−0.0015	0.9129	0.999	0.973–1.025
Stage	1.7276	0.0170	5.627	1.362–23.251	1.4465	0.0554	4.248	0.967–18.656
Grade	−0.0052	0.9828	0.995	0.620–1.595	0.1633	0.5308	1.177	0.707–1.962
Histology subtype	0.5031	0.2811	1.654	0.663–4.128	−0.1193	0.8088	0.888	0.338–2.332
GSE9891 internal validation cohort (N = 157)								
Risk score	0.79	0.0016	2.203	1.351–3.594	0.5924	0.0255	1.808	1.075–3.041
Age	0.0183	0.0699	1.018	0.999–1.039	0.0152	0.1542	1.015	0.994–1.037
Stage	1.7095	0.0002	5.526	2.246–13.597	1.5578	0.0009	4.748	1.899–11.871
Grade	0.2027	0.3086	1.225	0.829–1.809	−0.1177	0.5633	0.889	0.596–1.325
Histology subtype	1.5048	0.0103	4.503	1.426–14.22	0.846	0.1606	2.33	0.715–7.597
GSE9891 entire cohort (N = 257)								
Risk score	1.0096	<0.0001	2.745	1.902–3.962	0.9028	<0.0001	2.466	1.689–3.601
Age	0.0087	0.2618	1.009	0.994–1.024	0.006	0.4587	1.006	0.99–1.022
Stage	1.7306	<0.0001	5.644	2.642–12.058	1.6227	<0.0001	5.067	2.337–10.985
Grade	0.1245	0.4151	1.133	0.839–1.528	−0.0524	0.7384	0.949	0.698–1.291
Histology subtype	0.9768	0.007	2.656	1.305–5.404	0.3133	0.3998	1.368	0.66–2.837
GSE30161 entire cohort (N = 54)								
Risk score	1.1528	0.0309	3.167	1.112–9.022	1.5104	0.0114	4.528	1.405–14.597
Age	0.0037	0.8029	1.004	0.975–1.034	0.0064	0.7046	1.006	0.974–1.041
Stage	—	—	—	—	—	—	—	—
Grade	0.4346	0.1765	1.544	0.822–2.9	0.7009	0.0429	2.015	1.023–3.972
Histology subtype	0.3131	0.4456	1.368	0.612–3.057	1.2788	0.0283	3.592	1.146–11.26

Table 2. Univariate and multivariate cox regression analyses of DFS in GSE9891 and GSE30161. Abbreviations: C Coefficient, P P value, HR Hazard Ratio, CI Confidence Interval.

Rank	Biological Process	FDR	KEGG pathway	FDR
1	Cell adhesion	1.10E-24	ECM-receptor interaction	7.30E-10
2	Biological adhesion	1.40E-24	Focal adhesion	4.00E-09
3	Response to wounding	2.80E-15	Cell adhesion molecules (CAMs)	7.50E-04
4	Vasculature development	2.10E-13	Leukocyte transendothelial migration	2.60E-03
5	Blood vessel development	2.30E-12	Cytokine-cytokine receptor interaction	2.70E-03
6	Inflammatory response	5.10E-10	Chemokine signaling pathway	4.50E-02
7	Immune response	6.30E-10		
8	Extracellular matrix organization	1.40E-09		
9	Blood vessel morphogenesis	3.00E-09		
10	Regulation of response to external stimulus	3.30E-09		

Table 3. Top 10 significant GO BP terms and KEGG pathways enriched with DEGs in OvCa.

information for OvCa patients in these data were extracted from the R curatedOvarianData Bioconductor package³⁶. The microarray data were measured by the Affymetrix Human Genome U133A Plus 2.0 Array microarray platform. Borderline tumor patients and patients without days to tumor recurrence were excluded from this study. As a result, 257 patients in GSE9891 and 54 patients in GSE30161 were enrolled in this study. The OvCa patients in GSE9891 were randomly divided into a training cohort (n = 100) and an internal validation cohort (n = 157). Additionally, the OvCa patients in GSE30161 were analyzed as an external validation cohort (n = 54).

Microarray data preprocessing and lncRNA acquisition. The downloaded microarray data for GSE9891 and GSE30161 were normalized using the robust multi-array average (RMA) method³⁷. The probe set IDs of lncRNA were acquired from the study of Zhang *et al.*³⁸. Briefly, the probe set IDs were mapped to the NetAffx Annotation Files. Based on the Refseq transcripts ID and/or Ensembl gene ID in the annotation files, non-coding RNAs were retained and other types of non-coding RNA except lncRNA were then filtered. Finally, 2446 lncRNAs with corresponding probe set IDs were generated in our study.

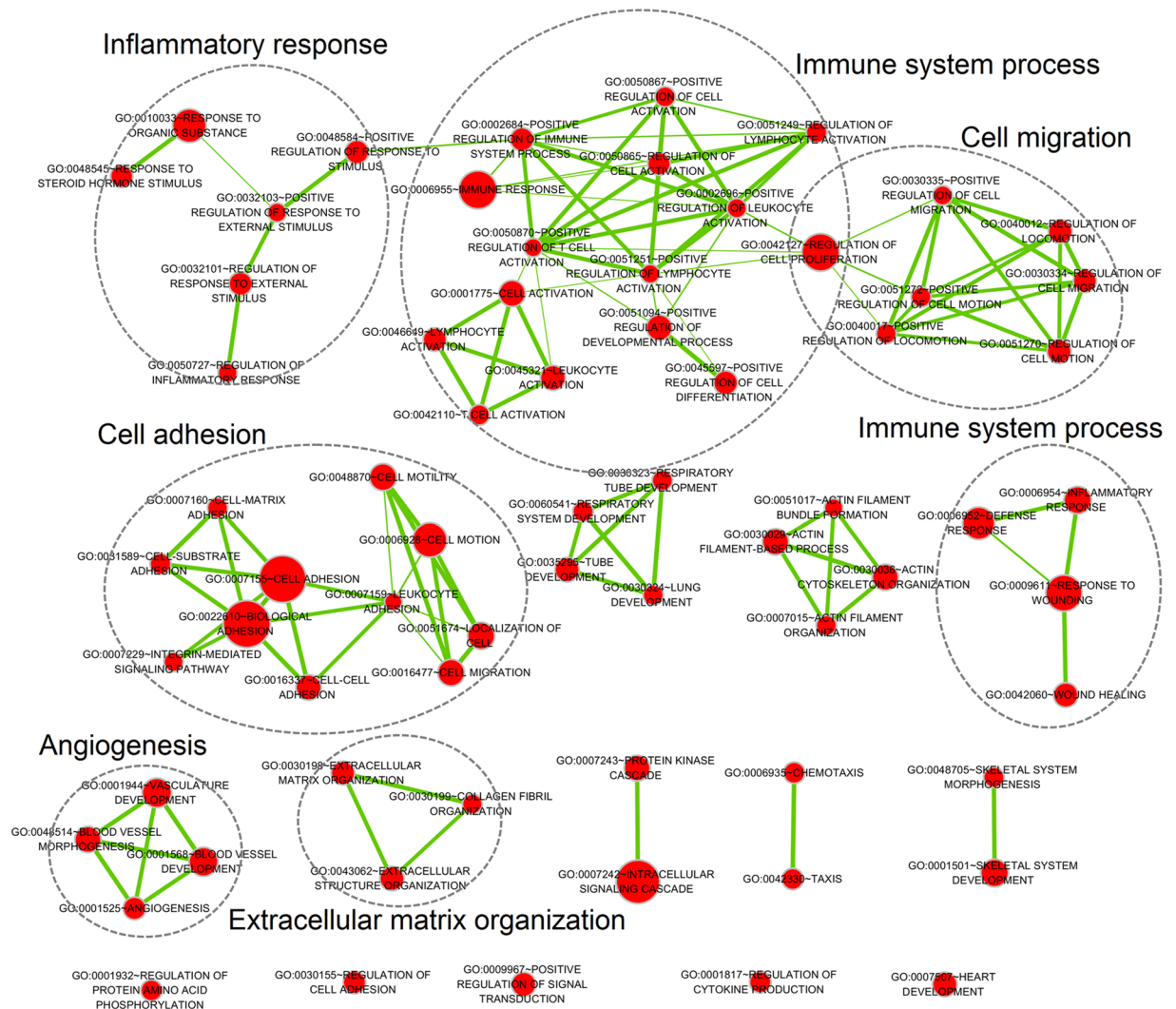


Figure 5. Interaction network of significant GO biological process terms with similar functions associated with six lncRNAs. Red nodes represent GO biological process terms. Node size is proportional to the total number of DEGs in that term. Thickness of green lines is proportional to the shared DEGs of two connected terms.

Signature generation and statistical analysis. LASSO penalized regression was conducted to select the lncRNAs associated with OvCa recurrence³⁹. The optimal tuning parameter lambda1 was chosen after 100 times of 10-fold cross validation. The function for the selection of lambda1 was “optL1” (fold = 10), and the function for LASSO penalized regression was “penalized” (lambda1 = lambda1). Other parameters of the functions were set to default values. A risk score was generated using the sum of lncRNA expression values weighted by the coefficients from LASSO penalized regression⁴⁰. The OvCa patients were then divided into low- and high-risk groups according to the median risk score.

The association of risk score, clinical factors, disease-free survival (DFS) and overall survival (OS) were assessed by univariate and multivariate cox regression. Kaplan-Meier (K-M) survival curves were used to estimate DFS and OS for patients in the low- and high-risk groups, and the DFS and OS differences between the two groups were assessed using the log-rank test. Time-dependent receiver operating characteristic (ROC) curve analysis with three years as the time point was used to compare the sensitivity and specificity of the DFS prediction based on the risk score⁴¹. A heat map was used to present the relative expression levels of lncRNAs in this study.

Satterthwaite t-test was performed to determine the significance of each gene, and the corresponding false discovery rate (FDR) value was estimated for correcting multiple comparisons. Differentially expressed genes (DEGs) were selected as FDR < 0.05. Functional annotation of DEGs for Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways was performed using the Database for Annotation, Visualization and Integrated Discovery (version 6.7, DAVID, <https://david.ncicrf.gov/>) tool⁴². Significant GO biological process terms with similar function were visualized as interaction networks using the Enrichment Map plugin in Cytoscape^{43, 44}.

LASSO penalized regression, time-dependent ROC curve analysis and heat map analyses were conducted on penalized³⁹, survivalROC⁴¹ and gplots⁴⁵ packages, respectively, in the R platform. Univariate and multivariate cox regression and log-rank test were performed using SAS (version 9.3, SAS Institute, Cary, NC, USA). K-M survival curves and scatterplots of risk score were performed in GraphPad Prism (version 5.0, Graphpad Software, San Diego, CA, USA)⁴⁶. All other statistical analyses were performed in the R platform (version 3.3.2). All statistical tests were two-sided and a P value of less than 0.05 was considered statistically significant.

References

- Torre, L. A. *et al.* Global cancer statistics, 2012. *CA Cancer J Clin* **65**, 87–108 (2015).
- Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics. *CA Cancer J Clin* **66**, 7–30 (2016).
- Jayson, G. C., Kohn, E. C., Kitchener, H. C. & Ledermann, J. A. Ovarian cancer. *The Lancet* **384**, 1376–1388 (2014).
- Jelovac, D. & Armstrong, D. K. Recent progress in the diagnosis and treatment of ovarian cancer. *CA Cancer J Clin* **61**, 183–203 (2011).
- Hangauer, M. J., Vaughn, I. W. & Mcmanus, M. T. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS genetics* **9**, 388–393 (2013).
- Zhu, X. *et al.* A long non-coding RNA signature to improve prognosis prediction of gastric cancer. *Molecular cancer* **15**, 60 (2016).
- Tu, Z. *et al.* An eight-long non-coding RNA signature as a candidate prognostic biomarker for lung cancer. *Oncology reports* **36**, 215–222 (2016).
- Jin, M., Ping, L., Zhang, Q., Yang, Z. & Shen, F. A four-long non-coding RNA signature in predicting breast cancer survival. *Journal of Experimental & Clinical Cancer Research* **33**, 1–10 (2014).
- Hu, Y. *et al.* A long non-coding RNA signature to improve prognosis prediction of colorectal cancer. *Oncotarget* **5**, 2230–2242 (2014).
- Li, J. *et al.* LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut* **63**, 1700–1710 (2014).
- Meryet-Figüiere, M. *et al.* An overview of long non-coding RNAs in ovarian cancers. *Oncotarget* **7**, 44719–44734 (2016).
- Huang, S., Qing, C., Huang, Z. & Zhu, Y. The long non-coding RNA CCAT2 is up-regulated in ovarian cancer and associated with poor prognosis. *Diagn Pathol* **11**, 49 (2016).
- Qiu, J. J. *et al.* Overexpression of long non-coding RNA HOTAIR predicts poor patient prognosis and promotes tumor metastasis in epithelial ovarian cancer. *Gynecol Oncol* **134**, 121–128 (2014).
- Cheng, Z. *et al.* A long noncoding RNA AB073614 promotes tumorigenesis and predicts poor prognosis in ovarian cancer. *Oncotarget* **6**, 25381–25389 (2015).
- Qiu, J. J. *et al.* Long non-coding RNA ANRIL predicts poor prognosis and promotes invasion/metastasis in serous ovarian cancer. *Int J Oncol* **46**, 2497–2505 (2015).
- Zhou, M. *et al.* Comprehensive analysis of lncRNA expression profiles reveals a novel lncRNA signature to discriminate nonequivalent outcomes in patients with ovarian cancer. *Oncotarget* **7**, 32433–32448 (2014).
- Petrillo, M. *et al.* Ovarian cancer patients with localized relapse: clinical outcome and prognostic factors. *Gynecol Oncol* **131**, 36–41 (2013).
- Kim, A., Ueda, Y., Naka, T. & Enomoto, T. Therapeutic strategies in epithelial ovarian cancer. *Journal of experimental & clinical cancer research* **31**, 14 (2012).
- Jiang, C., Li, X., Zhao, H. & Liu, H. Long non-coding RNAs: potential new biomarkers for predicting tumor invasion and metastasis. *Molecular cancer* **15**, 62 (2016).
- Yang, Z. *et al.* Long Noncoding RNA C21orf96 Promotes the Migration, Invasion and Lymph Node Metastasis in Gastric Cancer. *Anti-cancer agents in medicinal chemistry* **16**, 1101–1108 (2016).
- de Lavallade, H. *et al.* A gene expression signature of primary resistance to imatinib in chronic myeloid leukemia. *Leuk Res* **34**, 254–257 (2010).
- Hauptman, N. & Glavac, D. Long non-coding RNA in cancer. *International journal of molecular sciences* **14**, 4655–4669 (2013).
- Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031–8041 (2003).
- Massimo, M., Palmiro, P. & Fernando, D. U. O. Non-protein coding RNA biomarkers and differential expression in cancers: a review. *Journal of Experimental & Clinical Cancer Research* **27**, 19 (2008).
- Tano, K. *et al.* MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. *FEBS Letters* **584**, 4575–4580 (2010).
- Gabory, A., Jammes, H. & Dandolo, L. The H19 locus: role of an imprinted non-coding RNA in growth and development. *BioEssays* **32**, 473–480 (2010).
- Tsang, W. P. *et al.* Oncofetal H19-derived miR-675 regulates tumor suppressor RB in human colorectal cancer. *Carcinogenesis* **31**, 350–358 (2010).
- O'Driscoll, L. *et al.* Investigation of the molecular profile of basal cell carcinoma using whole genome microarrays. *Molecular cancer* **5**, 74 (2006).
- Yoshimura, K. *et al.* Prognostic value of matrix Gla protein in breast cancer. *Molecular medicine reports* **2**, 549–553 (2009).
- Atianand, M. K. & Fitzgerald, K. A. Long non-coding RNAs and control of gene expression in the immune system. *Trends Mol Med* **20**, 623–631 (2014).
- Wan, L. *et al.* HOTAIRM1 as a potential biomarker for diagnosis of colorectal cancer functions the role in the tumour suppressor. *J Cell Mol Med* **20**, 2036–2044 (2016).
- Yan, X. B., Chen, Z. & Brechot, C. The differences in gene expression profile induced by genotype 1b hepatitis C virus core isolated from liver tumor and adjacent non-tumoral tissue. *Hepatitis Monthly* **11**, 255–262 (2011).
- Pakkasjarvi, N. *et al.* Neural precursor cells from a fatal human motoneuron disease differentiate despite aberrant gene expression. *Dev Neurobiol* **67**, 270–284 (2007).
- Tohill, R. W. *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clinical Cancer Research An Official Journal of the American Association for Cancer Research* **14**, 5198–5208 (2008).
- Ferriss, J. S. *et al.* Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: predicting platinum resistance. *PloS one* **7**, e30550 (2012).
- Ganzfried, B. F. *et al.* curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. *Database (Oxford)* **2013**, bat013 (2013).
- Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).
- Zhang, X. *et al.* Long non-coding RNA expression profiles predict clinical phenotypes in glioma. *Neurobiology of Disease* **48**, 1–8 (2012).
- Goeman, J. J. L. 1 Penalized Estimation in the Cox Proportional Hazards Model. *Biometrical Journal* **52**, 70–84 (2010).

40. Hayes, J. *et al.* Prediction of clinical outcome in glioblastoma using a biologically relevant nine-microRNA signature. *Mol Oncol* **9**, 704–714 (2015).
41. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker. *Biometrics* **56**, 337–344 (2000).
42. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1–13 (2009).
43. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS one* **5**, e13984 (2010).
44. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498–2504 (2003).
45. Warnes, G. R. *et al.* gplots: various R programming tools for plotting data (2005).
46. Motulsky, H. Analyzing Data with GraphPad Prism (1999).

Acknowledgements

This work was funded by National Natural Science Foundation of China (project number: 81573256, 81473072, 81302511).

Author Contributions

K.Y. and Y.H. wrote the manuscript. K.Y., A.L. and Z.L. downloaded and analysed the data. W.W., H.X. and Z.R. prepared all the figures and tables. G.L. and K.L. designed the entire study and edited the manuscript. All authors reviewed and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:[10.1038/s41598-017-00763-y](https://doi.org/10.1038/s41598-017-00763-y)

Competing Interests: The authors declare that they have no competing interests.

Change History: A correction to this article has been published and is linked from the HTML version of this paper. The error has been fixed in the paper.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017