Identification of lncRNA Signature Associated With Pan-Cancer Prognosis

Guoqing Bao , Ran Xu, Xiuying Wang , Jianxiong Ji, Linlin Wang , Wenjie Li , Qing Zhang, Bin Huang, Anjing Chen, Di Zhang, Beihua Kong, Qifeng Yang, Cunzhong Yuan, Xinyu Wang , Jian Wang , and Xingang Li

Abstract—Long noncoding RNAs (lncRNAs) have emerged as potential prognostic markers in various human cancers as they participate in many malignant behaviors. However, the value of lncRNAs as prognostic markers among diverse human cancers is still under investigation, and a systematic signature based on these transcripts that related to pan-cancer prognosis has yet to be reported. In this study, we proposed a framework to incorporate statistical power, biological rationale, and machine learning models for pan-cancer prognosis analysis. The framework identified a 5-lncRNA signature (*ENSG00000206567*, *PCAT29*, *ENSG00000257989*, *LOC388282*, and *LINC00339*) from TCGA training studies ($n = 1,878$). The identified lncRNAs are significantly associated (all $P \leq 1.48E\text{-}11$) with overall survival (OS) of the TCGA cohort ($n = 4,231$). The signature stratified the cohort into low- and high-risk groups with significantly distinct survival outcomes (median OS of 9.84 years versus 4.37 years, log-rank $P = 1.48E\text{-}38$) and achieved a time-dependent ROC/AUC of 0.66 at 5 years. After routine clinical factors involved, the signature demonstrated better performance for long-term prognostic estimation (AUC of 0.72). Moreover, the signature was further evaluated on two independent external cohorts (TARGET, $n = 1,122$; CPTAC, $n = 391$; National Cancer Institute) which yielded similar prognostic values (AUC of 0.60 and 0.75; log-rank $P = 8.6E\text{-}09$ and $P = 2.7E\text{-}06$). An indexing system was developed to map the 5-lncRNA signature to prognoses of pan-cancer patients. In *silico* functional analysis indicated that the lncRNAs are associated with common biological processes driving human cancers. The five lncRNAs, especially *ENSG00000206567*, *ENSG00000257989* and *LOC388282* that never reported before, may serve as viable molecular targets common among diverse cancers.

Index Terms—lncRNA, pan-cancer, prognosis, machine learning.

Guoqing Bao and Xiuying Wang are with the School of Computer Science, The University of Sydney, Camperdown, NSW 2006, Australia (e-mail: guoqing.bao@sydney.edu.au; xiu.wang@sydney.edu.au).

Ran Xu was with the Department of Neurosurgery, Qilu Hospital, Shandong University, Jinan 250012, China. He is now with the Department of Neurosurgery, Xuanwu Hospital, Capital Medical University, Beijing 100053, China (e-mail: xrqssq@126.com).

Jianxiong Ji, Wenjie Li, Qing Zhang, Bin Huang, Anjing Chen, Di Zhang, Xinyu Wang, and Xingang Li are with the Department of Neurosurgery, Qilu Hospital, Shandong University, Jinan 250012, China (e-mail: jijx_sdu@163.com; 402861939@qq.com; 467621765@qq.com; huangbin_sdu@qq.com; caj669@qq.com; dzhang1982@hotmail.com; wangxinyu@sdu.edu.cn; lixg@sdu.edu.cn).

Linlin Wang is with the Department of Radiation Oncology, Shandong Cancer Hospital, Jinan 250117, China (e-mail: wanglinlinatjn@163.com).

Beihua Kong, Qifeng Yang, and Cunzhong Yuan are with the Department of Breast Surgery and Department of Obstetrics and Gynecology, Qilu Hospital, Shandong University, Jinan 250012, China (e-mail: kongbeihua@sdu.edu.cn; qifengy_sdu@163.com; yuancunzhong@126.com).

Jian Wang is with the Department of Neurosurgery, Qilu Hospital, Shandong University, Jinan 250012, China, and also with the Department of Neurosurgery, Qilu Hospital, Shandong University, Jinan 250012, China (e-mail: jian.wang@uib.no).

Digital Object Identifier 10.1109/JBHI.2020.3027680

I. INTRODUCTION

IN THE precision medicine era targeted molecular therapy is the main strategy for the management of cancer patients. Recent basket and umbrella trials demonstrate that actionable mutations are considered as an important predictor of tumor response, and the same molecular alteration can be effectively controlled across different cancers [1]. Hence, a deep understanding of the molecular events which underlie various biological behaviors is increasingly needed.

LncRNAs are transcripts longer than 200 nucleotides encoded by the genome without protein translation potential [2]. More than 75% of transcripts in the genome are noncoding RNAs [2], [3]. Increasing evidence indicates that lncRNAs function as critical mediators for many aggressive biological behaviors in human cancers [4]–[8]. For example, HOX antisense intergenic RNA (HOTAIR) has been found to be critical in driving multiple malignant behaviors, such as proliferation, migration and invasion, suppression of drug response, and genomic instability [9]. Several studies have also identified the prognostic value of lncRNAs in human cancers [10], [11]. In particular, the lncRNA PCA3/DD3 is uniquely expressed in prostate cancer tissues compared to normal tissues and has already been tested

as a biomarker in clinical settings [12]. Such studies provide convincing evidence for the identification of a lncRNA-based signature that associated with the prognosis of multiple cancers. However, a single lncRNA is not sufficient to reflect the complexity of cancer biological behaviors. Therefore, a study focused on the expression patterns of several lncRNAs could be more accurate and informative.

The Cancer Genome Atlas (TCGA) consists of original sequencing data from various sources, provides the opportunity to perform integrated studies on the commonalities and differences between diverse cancers. This multi-cancer (pan-cancer) approach to the analysis of such datasets has accelerated the study of the disease and improved treatment efficacy for different types of human cancers [13]–[15]. Through analysis of large-scale TCGA datasets, multiple lncRNAs have been identified as critical factors in gene regulatory network perturbations [16]. Besides, hundreds of lncRNAs have been proposed to be involved in oncogenic genes and pathways in each tumor context [17] and mounting evidence from the study of individual cancers has also established lncRNAs as cancer-specific prognosis predictors [18]–[20]. Although lncRNAs are commonly seen as dysregulated in a tumor-specific manner, a recent study demonstrated that some of the lncRNAs involved in multiple tumor contexts [21]. Studies of the lncRNA-disease association also suggested that similar diseases tend to be associated with functionally similar lncRNAs [22], [23]. Recent studies provide further justification for the importance of cancer prognosis in the field of pan-cancer research [24]–[26]. Meanwhile, another research group established a seven-gene signature for the prediction of patient prognosis in 13 cancer types [27]. However, the potential lncRNAs as pan-cancer prognostic markers have not yet been rigorously tested. Such results might provide insight into the biological processes involved in cancer initiation and development shared by many different human cancers.

Therefore, the main objective of this study is to identify pan-cancer prognostic lncRNAs that related to common and critical biological processes driven diverse human cancers. To do that, we propose a pan-cancer prognosis analysis framework by incorporating statistical power, biological rationale, and machine learning model as a whole biomarker identifier. As a result, the framework identified five prognostic lncRNAs associated with overall survival (OS) of pan-cancer studies. An indexing system was developed to map lncRNA signature to the OS of different types of cancer studies. A 5-lncRNA risk score model was established which stratified patient studies into high- and low-risk groups with significantly distinct survival outcomes. The prognostic power of the five lncRNAs and the risk score model was further validated in the testing dataset and another two independent cohorts. Functional analysis revealed potential underlying functions of these lncRNAs associated with common and critical cellular processes that drive human cancers. Gene enrichment and qPCR analyses identified *ENSG00000206567* associated with gynecologic cancers.

II. MATERIALS AND METHODS

An overview of our pan-cancer prognosis analysis framework is illustrated in Fig. 1.

A. Data Acquisition and Study Design

There is a total of 731 lncRNAs reported by TCGA for pan-cancer analysis (4,266 cancer studies covered). The corresponding RNA-seq data (normalized by the FPKM method [28]) and clinical information were downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/). After the exclusion of 35 cases without complete survival information, a total of 4,231 patients with a broad range of cancer types ($n = 33$) were enrolled in our study (TCGA cohort), Supplementary Table S1. To identify the prognostic lncRNAs related to multiple cancers, a study cohort ($n = 2,210$), which stratified into 85% cases for training/cross-validation ($n = 1,878$) and 15% for testing ($n = 332$), was identified from the TCGA cohort based on prognosis distribution (cutoff of 3.5 y to stratify the better and poor prognosis patients, Supplementary Table S2). The training data was used for the identification of prognosis-related lncRNAs, the test data was utilized to estimate the prognostic effectiveness, and the overall TCGA cohort was served as internal validation. Besides, another two independent cohorts, i.e. Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and Clinical Proteomic Tumor Analysis Consortium (CPTAC), which provided by National Cancer Institute (NCI), contain 23 different types of cancer studies were further collected. Similarly, patient data without survival information were excluded and the remaining patient studies, i.e. TARGET ($n = 1,122$) and CPTAC ($n = 391$) cohorts, were used for external validation. In addition, three routine clinical parameters, including age at diagnosis, gender, and tumor stage, were included as co-variables.

B. Screening of Prognostic LncRNAs Associated With Pan-Cancer Prognosis

As shown in Fig. 1(a), statistical power is incorporated with machine learning models for identification and evaluation of independent prognostic lncRNAs, which contains three major steps:

1) Machine learning-Driven Stepwise Feature Selection (MLSFS): A stepwise feature selection method, namely Recursive feature elimination with cross-validation (RFECV) [29], [30], was used to identify the candidate lncRNAs for prognosis prediction. RFECV is a self-contained algorithm equipped with cross-validation and can work independently without human intervention once configured properly, thus reduce the inter-observer variability and improve reproducibility and stability. The method is composed of two layers with Recursive Feature Elimination (RFE) as an inner layer being embedded in Cross-Validation (CV) as an outer layer. The outer CV layer consists of n folds of cross-validations. In each fold, the input data is first stratified into train/validation pairs, and then for each iteration in the inner RFE layer, Random Forest (RF) [31] assigns importance for lncRNAs (features) according to its internal decision trees. The least important feature is eliminated based on the feature importance. The trained RF is estimated on the validation data to calculate the performance score of the remaining features. Based on the performance scores of CV folds, the feature combination that achieves the maximum average accuracy score is chosen as optimal features. Finally,
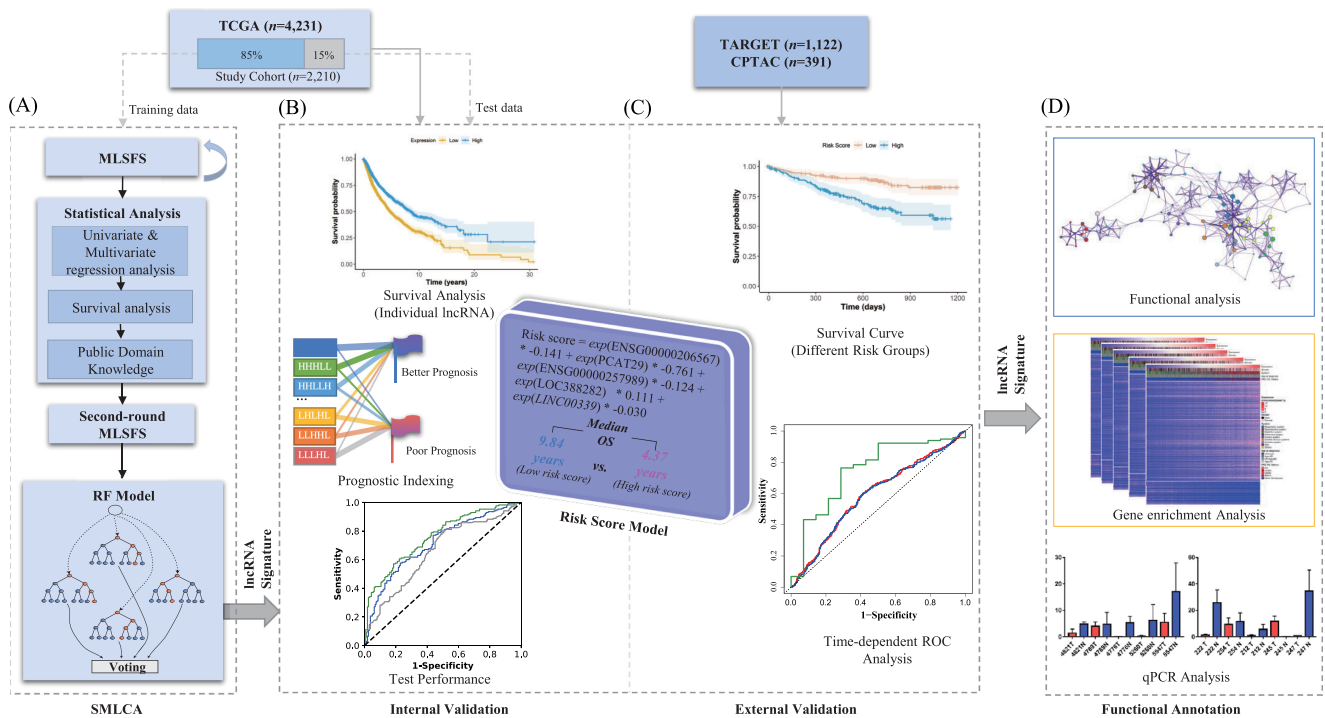
Fig. 1. Flowchart of the pan-cancer prognosis analysis framework. (A) Independent prognostic lncRNAs were identified through statistical and machine learning (ML) combined analysis (SMLCA) from TCGA training data. MLSFS: ML-based stepwise feature selection. (B) Prognostic value of lncRNAs was evaluated on the test dataset and overall TCGA cohort. (C) 5-lncRNA risk score model was developed and prognostic power of lncRNA signature was further validated on another two independent external cohorts. (D) Functional analysis identified the lncRNA signature associated with common and critical cellular processes that drive human cancers; gene enrichment analysis found one of the lncRNAs associated with gynecologic cancers, and qPCR analysis suggested that it was differentially expressed in pan-gynecologic tumor and normal tissue pairs.

the previous output served as the data source for the next-round of RFECV until the number of features was invariant, which yielded the candidate lncRNAs for further analysis.

2) Selection of Independent Prognostic LncRNAs With Statistical Analyses: We used *coxph*, a function in R survival package, to compute the Cox proportional hazards regression models, which measures the association between the survival time of cancer studies and one (Univariate) or more (Multivariate) predictor lncRNAs. Univariate Cox regression analyses were first performed for each of the candidate lncRNAs. Statistically significant lncRNAs ($P < 0.05$) in Univariate analyses were further evaluated with Multivariate Cox regression analyses where gender, age at diagnosis, and tumor stage were included as covariables. Prognostic lncRNAs identified by significant Hazard Ratio (HR) ($|HR - 1| \geq 0.1$) were finally subjected to Kaplan-Meier survival analyses to compute the survival probability over time, from which patient studies were separated into high and low expression groups based on their median values. The lncRNAs that exhibited significant survival difference (log-rank test $P < 0.001$) for the two expression groups were selected. In this work, HR measures the relative risk ratio of candidate lncRNAs. $HR > 1$: risk factor, $HR < 1$: protective factor, and $HR = 1$: not a valuable factor.

3) Evaluation With the Machine Learning Model: RF was introduced as the prognostic performance estimator. The supervised machine learning model leveraged multiple decision trees for ensemble voting and reduced the classification variance

through training on different parts of the training data [32]. A feature bagging technique was integrated to generate higher model performance by selecting a random subset of the features in each iteration [33]. The classifier was implemented using the scikit-learn library (https://scikit-learn.org/) in the present study. To tune the optimal hyperparameters, 5-fold nested cross-validation was adopted which partitioned the training/cross-validation data ($n = 1,878$) into 4:1 held-out dataset pairs. In the inner loop, each iteration (fold) further used additional cross-validation to search hyperparameters and fit the model. The following hyperparameters were tuned due to their superior performance under the nested cross-validation: a total of 1000 trees in the forest, with the maximum depth of the tree = 8, where samples to split an internal node are ≥ 25 and samples at a leaf node are ≥ 4.

The predictive performances of identified lncRNAs, three clinical factors, or their combination were further evaluated and compared on the test set ($n = 332$) after the model was trained on full training/cross-validation data using the identified hyperparameters.

C. Mapping of LncRNA Signature To Patient Prognosis

The study cohort was stratified into high- (H) and low-expression (L) groups based on the median expression values of the corresponding lncRNAs (Supplementary Table S3). There are 2^n possible expression permutations for the signature

(which composed of *n* lncRNAs). The relationship among the pan-cancer studies, permutations of the lncRNA signature, and corresponding patient survival outcomes were visualized in a Sankey diagram. The diagram was constructed using the Python plotly library.[1]

D. Construction of Prognostic Signatures and Performance Validation

The overall TCGA cohort was stratified into low- and high-expression subgroups based on the median expression value of each identified lncRNA. The prognostic value of each lncRNA was first evaluated by conducting Kaplan-Meier analysis and log-rank tests on different subgroups.

Two risk score systems were then derived to estimate patient mortalities based on identified lncRNA signature and hybrid signature (composed of lncRNA and clinical factors) respectively, as follow:

$$RiskScore = \sum_{k=1}^{n} Coef(factor_k) * Val(factor_k) \qquad (1)$$

where *Coef* is the coefficient of risk factor *k* (lncRNA or clinical factor) measured by multivariate Cox regression analyses; *Val* is the value of the risk factor *k*; *n* is the number of factors.

To evaluate the prognostic significance of the signature, the overall TCGA cohort was stratified into low- and high-risk subgroups based on the median risk score. The prognostic values of the signature were measured by Kaplan-Meier analysis and the log-rank test. Time-dependent ROC/AUCs at three, five, and ten years were then calculated. Similar measurements were also conducted on another two external datasets to further validate the effectiveness of identified signature on different platforms.

E. Analysis of LncRNA Functions and Their Correlation With Clinical Factors

Spearman correlation coefficients were computed to measure the correlation between each identified lncRNA and genome-wide RNA-Seq profiles (the other 60,482 genes). Genes that correlated with at least one of the lncRNAs in the TCGA cohort (Spearman correlation coefficients > 0.40 or < -0.30) were identified as co-expressed genes and enrolled in the functional enrichment analyses. The coefficient thresholds were determined under consideration of numbers reported in the literature as well as the characteristics of the datasets [34], [35]. The analyses were performed using an online gene annotation and analysis platform, named Metascape.[2]

The correlation between lncRNA and other clinical factors, including gender, cancer tissue of origin, age at diagnosis and tumor stage, was evaluated using annotated heatmaps based on RNA-Seq data of the TCGA cohort, Fig. 1(d). The co-expressed genes included in this analysis were selected using the same standard as in the functional analysis mentioned above. Patient molecular data and their corresponding clinical parameters were

accordingly sorted, and thereby, the correlation between the target lncRNA and the clinical factors can be observed.

Differential lncRNA(s) was evaluated by quantitative real-time polymerase chain reaction (qRT-PCR) using additional tissue sections (provided and approved by the Medical Ethics Committee of Qilu Hospital, Shandong University; consent was received from all involved subjects).

F. Statistical Analysis

Univariate and multivariate Cox proportional hazard regression analyses were performed to measure the association between lncRNAs (and/or clinical parameters) and patient overall survival. HR was calculated to measure the prognostic impact of various factors and a 95% confidence interval (CI) was used to indicate the precision of the estimated HR. Kaplan-Meier survival analyses were conducted to measure the overall survival difference between risk groups. The log-rank test was performed to evaluate the statistical significance. Time-dependent ROC analysis [36], which is used to assess the predictive power of diagnostic markers for time-dependent disease outcomes, was used to measure the predictive performance of the identified lncRNAs.

III. RESULTS

A. Screening of Independent Prognostic LncRNAs

The MLSFS method yielded 26 candidate lncRNAs that associated with OS in TCGA training studies under no clinical variables considered. After subjected to statistical analyses, *ZNF883*, *ENSG00000277476*, and *PVT1* were first filtered as they were not significantly associated with OS in univariate Cox regression analyses (light gray, Table I). Next, clinical covariables were enrolled and 11 lncRNAs were further removed based on HR ($|HR - 1| < 0.1$) calculated in multivariate Cox regression analysis (dark gray, Table II). Kaplan-Meier survival analysis identified *ENSG00000226380*, *FOXM1*, and *LINC02637* as insignificant factors to stratify different risk groups (light gray, Table II). In the second-round MLSFS, *MEG8*, *MIR3142HG*, and *MIR4435-2HG* were excluded because they did not emerge as significant factors in ML models once clinical features were involved. Although MEG3 exhibited prognostic significance, the lncRNA was not considered since it was a known prognostic biomarker in multiple cancers reported in the public literature (public domain knowledge). Thus, the signature associated with pan-cancer prognosis identified in this study is composed of five lncRNAs, i.e. *ENSG00000206567*, *PCAT29*, *ENSG00000257989*, *LOC388282*, and *LOC388282*.

B. Mapping of 5-lncRNA Signature to Patient Prognosis

The signature composed of five distinct lncRNAs and thus has 2^5 possible permutations if each lncRNA was stratified into high (H) and low (L) expression levels using the median value as cutoff (Supplementary Table S3). A Sankey diagram was developed to visualize the relationship between lncRNA signature (represented as 32 expression permutations or 5-lncRNA phenotypes) and patient prognosis in the study cohort, as shown in Fig. 2.

---

[1][Online]. Available: https://plot.ly/
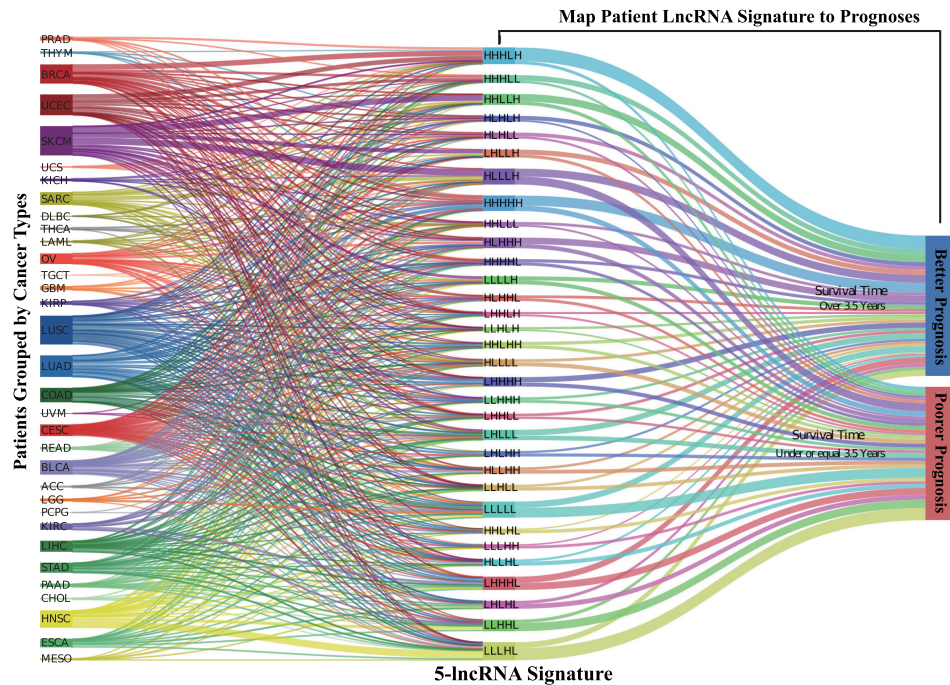[2][Online]. Available: http://metascape.org

Fig. 2. Map 5-lncRNA signature to patient prognosis. The left half diagram maps cancer studies to their lncRNA signatures (5-lncRNA phenotypes). The right half diagram maps lncRNA signatures to survival outcomes. H, high expression. L, low expression. Each permutation of the signature represents the expression levels of *ENSG00000206567*, *PCAT29*, *ENSG00000257989*, *LOC388282*, and *LINC003393* (from the left to the right), middle panel.

TABLE I
RESULTS OF UNIVARIATE COX REGRESSION ANALYSIS

Genes	HR	HR (95% CI)	*P*-value
LOC388282	1.484	1.321-1.667	2.56E-11
LINC00339	0.675	0.602-0.758	3.00E-11
PCAT29	0.678	0.604-0.761	4.35E-11
ENSG00000206567	0.681	0.607-0.764	6.70E-11
EGFR	1.448	1.290-1.626	3.61E-10
LINC00909	0.716	0.638-0.803	1.22E-08
PWAR6	0.717	0.639-0.805	1.53E-08
MEG8	1.378	1.228-1.546	4.96E-08
EWSAT1	0.731	0.651-0.820	9.03E-08
ENSG00000257989	0.741	0.660-0.831	3.73E-07
OGFRP1	1.329	1.184-1.491	1.30E-06
MIR3142HG	0.756	0.673-0.848	1.97E-06
LINC01547	0.766	0.683-0.859	5.45E-06
MEG3	1.278	1.139-1.434	2.98E-05
ENSG00000251615	0.807	0.720-0.906	2.76E-04
LINC00242	1.237	1.102-1.388	3.04E-04
LINC01537	1.221	1.089-1.370	6.42E-04
GASAL1	1.22	1.088-1.368	6.78E-04
MIR4435-2HG	1.215	1.083-1.362	9.14E-04
FOXM1	1.197	1.067-1.343	2.13E-03
ENSG00000226380	1.183	1.055-1.327	4.09E-03
LINC02637	1.173	1.046-1.316	6.30E-03
LINC02551	1.148	1.024-1.287	1.84E-02
ZNF883	1.093	0.975-1.226	1.29E-01
ENSG00000277476	1.069	0.953-1.199	2.54E-01
PVT1	1.046	0.932-1.173	4.46E-01

Gray color: lncRNAs filtered.

TABLE II
RESULTS OF MULTIVARIATE COX REGRESSION AND KAPLAN-MEIER SURVIVAL ANALYSIS

Factors	HR	HR (95%CI)	*P* (Cox)	*P* (KM)
Gender	1.125	0.991-1.276	6.85E-02	-
Age at diagnosis	1.02	1.015-1.025	8.36E-17	-
Tumor stage	1.303	1.197-1.418	1.08E-09	-
LOC388282	1.306	1.145-1.488	6.54E-05	1.94E-11
MEG8	1.206	1.055-1.378	6.14E-03	4.38E-08
MIR3142HG	0.804	0.706-0.915	9.83E-04	1.83E-06
ENSG00000226380	1.189	1.047-1.350	7.46E-03	4.03E-03
LINC00339	0.818	0.718-0.931	2.43E-03	2.28E-11
PCAT29	0.823	0.726-0.933	2.37E-03	3.30E-11
ENSG00000257989	0.837	0.736-0.952	6.58E-03	3.35E-07
FOXM1	1.158	1.013-1.324	3.16E-02	2.09E-03
MIR4435-2HG	1.144	1.010-1.295	3.36E-02	8.92E-04
MEG3	1.137	0.994-1.301	6.21E-02	2.84E-05
LINC02637	1.123	0.997-1.265	5.68E-02	6.23E-03
ENSG00000206567	0.879	0.767-1.009	6.65E-02	5.17E-11
OGFRP1	1.094	0.951-1.258	2.09E-01	-
LINC01537	1.081	0.956-1.223	2.16E-01	-
LINC00242	1.078	0.953-1.219	2.33E-01	-
LINC02551	1.049	0.929-1.185	4.42E-01	-
EWSAT1	0.965	0.850-1.097	5.89E-01	-
EGFR	0.971	0.846-1.114	6.76E-01	-
LINC00909	0.976	0.854-1.115	7.21E-01	-
ENSG00000251615	1.022	0.888-1.176	7.64E-01	-
PWAR6	0.979	0.845-1.135	7.78E-01	-
GASAL1	0.995	0.880-1.125	9.41E-01	-
LINC01547	0.999	0.878-1.136	9.84E-01	-

Dark gray: lncRNAs filtered by Multivariate Cox regression; Light gray: lncRNAs filtered by Kaplan-Meier analyses

The right part of the diagram can be consulted independently, for example, the majority of patients with lncRNA phenotypes HHHLH, HHHLL, HHLLH (favorable phenotypes) related to better prognosis, whereas, patients with LLHHL and LLLHL have poorer prognosis. The most valuable phenotypes located at the outer edges of the diagram. Observed from the left part of the diagram, there is more percentage of patients with BRCA, UCEC or SKCM mapped to favorable lncRNA phenotypes

compared to patients with HNSC, which represents patients with the first three cancer types tended to have better overall prognosis than patients with HNSC. This result is consistent with the literature [37]–[39], i.e. BRCA, UCEC, and SKCM patients have a much higher five-year survival rate (>90%, 90%, and 92% respectively) compared to HNSC (50%). Besides, we found patients with LUSC and LUAD, which unlike BRCA, UCEC, SKCM, and HNSC, etc., are not densely mapped to any specific lncRNA phenotypes, which indicates the prognoses of LUSC and LUAD patients are more heterogeneous, this finding is supported by recent research studies [40], [41]. The prognostic value gradually declined with the phenotype arrangements approaching the center of the diagram.

C. Prognostic Performance of The 5 LncRNAs on Test Data

The RF model with the 5-lncRNA signature demonstrated a prognostic AUC performance of 0.714 on the test data, which is significantly superior to the predictive performance of routine clinical factors (AUC of 0.656), Fig. 3(a). After the combination of the 5-lncRNA signature with clinical parameters, the model yielded a classification AUC of 0.758 for differentiation of better vs. poor prognosis in test studies, Fig. 3(a). These results indicated that the signature exhibited good discrimination and calibration.

D. Construction of 5-lncRNA Risk Score Model and Validation on Overall TCGA Cohort

The five lncRNAs were found all significantly associated with OS (log-rank test P $\leq$ 1.48E-11) of the TCGA cohort ($n = 4,231$), Fig. 3 B-F. ENSG00000206567, PCAT29, ENSG00000257989, and LINC00339 tended to be the protective factors since their high expressions were related to longer survival (Fig. 3 B-D and F), whereas LOC388282 might be a risk factor because its high expression was associated with poor prognosis (Fig. 3(e)).

The relative contributions (coefficients) of the five prognostic lncRNAs to OS were derived from multivariate Cox regression measured in the study cohort. A 5-lncRNA risk score model was established and applied on entire TCGA cohort through integrating the obtained coefficients and gene expression values, as follow: $Risk\,score = exp(ENSG00000206567) * -0.141 + exp(PCAT29) * -0.761 + exp(ENSG00000257989) * -0.124 + exp(LOC388282) * 0.111 + exp(LINC00339) * -0.030$. As a continuous variable, 5-lncRNA risk score was found significantly associated with OS of TCGA patients (HR = 2.732, 95% CI = 2.332-3.202, P = 1.63E-35). The overall TCGA studies were stratified into two risk groups using the median risk score as cutoff and Kaplan-Meier analysis demonstrated patients' OS in the low-risk group is significantly better than the high-risk group (median OS of 9.84 versus 4.37 years, log-rank test P = 1.48E-38), Fig. 4(a). Time-dependent ROC analysis showed that the score model achieved AUC of 0.66 at five years (95% CI: 0.632-0.684), which is relatively higher than three (AUC of 0.64) and ten years (AUC of 0.63),
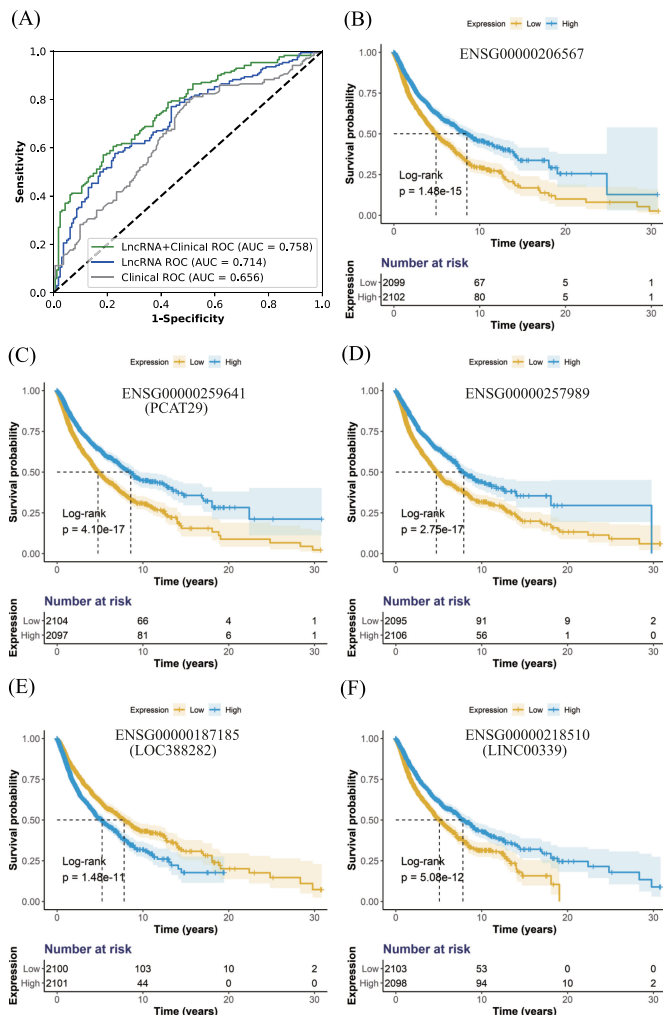

Fig. 3. Prognostic performance of 5 lncRNAs on TCGA cohort. (A) ROC/AUC performance of RF model for differentiation of better and poor prognosis on TCGA test data. (B-F) Kaplan-Meier analysis on overall TCGA studies ($n = 4,231$) based on expression of ENSG00000206567 (B), PCAT29 (C), ENSG00000257989 (D), LOC388282 (E), and LINC00339 (F). The high- and low-expression were assigned based on their median expression values (Supplementary Table S3). P values were calculated by two-sided log-rank tests.
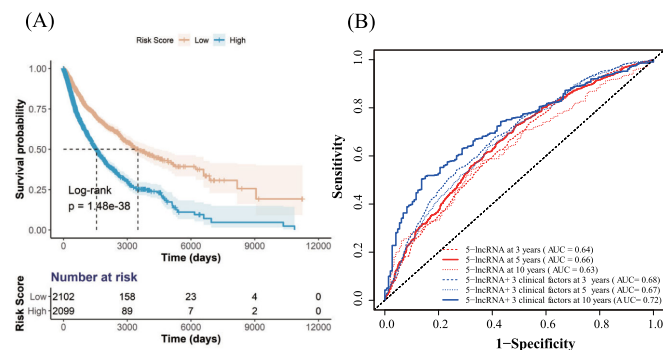

Fig. 4. Predictive performance of 5-lncRNA risk score model on TCGA cohort. (A) Kaplan-Meier analysis for patient OS with high- and low-risk groups. (B) Time-dependent ROC analysis based on 5-lncRNA signature (red line) and 5-lncRNA + 3 clinical factors (blue line).
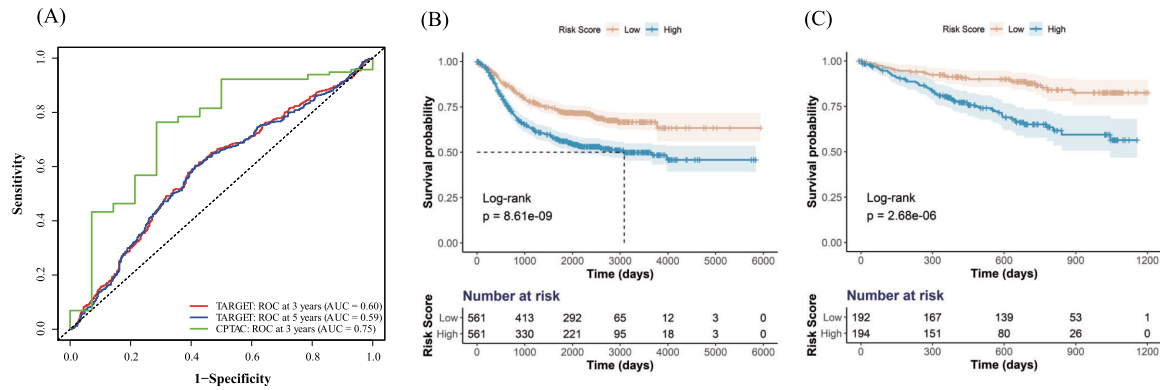
Fig. 5. Predictive performance of 5-lncRNA risk score model on external cohorts. (A) Time-dependent ROC analysis on TARGET and CPTAC cohorts based on 5-lncRNA signature. (B, C) Kaplan-meier analysis for patient OS with high- and low-risk groups in TARGET (B) and CPTAC (C) cohort studies.

Fig. 4(b). The signature also demonstrated good prognostic performance on individual cancers (Supplementary Table S4).

Combined with gender, age at diagnosis and optional tumor stage, the hybrid risk score model achieved a higher time-dependent ROC/AUC of 0.718 (95% CI: 0.677-0.760) for long-term (ten years) risk estimation but no significant boost in three (AUC of 0.68) and five years (AUC of 0.67), Fig. 4(b). Similarly, the hybrid model stratified patient studies into two risk groups with distinct survival outcomes (median OS of 11.86 years in the low-risk group compared with 4.22 years in the high-risk group, log-rank test $P = 4.86\text{E-}51$).

E. Validation of 5-lncRNA Risk Score Model on Another Two Independent External Cohorts

The risk score model was further applied on TARGET ($n = 1,122$) and CPTAC ($n = 391$) cohorts using the coefficients derived from the TCGA study cohort. The lncRNA risk score achieved HR of 2.393 (95% CI: 1.611-3.553, $P = 1.53\text{E-}05$) on TARGET and 1.922 (95% CI: 1.436-2.572, $P = 1.121\text{E-}05$) on CPTAC. The two external cohorts were stratified into high- and low-risk groups respectively. Time-dependent ROC analysis showed that the lncRNA model achieved AUC of 0.60 at three years and 0.59 at five years respectively on TARGET, and AUC of 0.75 at three years on CPTAC cohort (5-year AUC could not be calculated due to uneven survival distribution), Fig. 5(a). Kaplan-Meier analyses showed the two risk groups in CPTAC and TARGET demonstrated significantly different survival outcomes respectively (log-rank test $P = 8.61\text{E-}09$ and $P = 2.68\text{E-}06$), Fig. 5 B-C. The performance of the hybrid model (lncRNA with clinical factors) was not presented here since no sufficient clinical parameters, especially age at diagnosis and tumor stage, provided in the two datasets.

F. Identification of Biological Processes Associated With the Lncrna Signature

By computing Spearman correlation coefficients between lncRNA signature and genome-wide RNA-Seq data in the TCGA cohort, a total of 1,299 positively or negatively genes were found correlated with at least one of the five lncRNAs in

our signature. Functional and pathway analyses showed those co-expressed genes with prognostic lncRNAs were significantly enriched in 215 GO terms and 18 KEGG pathways (Log10(P) $< \text{-}2$ or $P < 0.01$), which mainly involved in cilium movement and organization, cell projection, lymphocyte activation, T cell activation and differentiation, cellular response, primary immunodeficiency, and metabolic processes of xenobiotic, drug, flavonoid, hormone, retinoic acid and so on (Table III).

Next, we thoroughly investigated each top enriched GO term or KEGG pathway in relation to cancer from public literature. We found some of the functions including cilium movement and organization [42]–[45], lymphocyte activation and differentiation [46]–[50], T-cell activation and differentiation [51]–[55], flavone metabolic process [56], [57], cytokine production [58]–[61], terpenoid metabolic process [62], [63], and regulation of leukocyte cell-cell adhesion [64], [65] are fundamental to cancer initialization and progression and related to pan-cancer, whereas others like microtubule formation and movement [66]–[68], inner dynein arm assembly [69], xenobiotic glucuronidation [70], [71], regulation of cell-cell adhesion [72]–[75], mammary gland branching [76], hormone metabolic process [77]–[79], leukocyte differentiation [80], [81] and retinoic acid metabolic process [82]–[84] are associated with specific cancer types (Table III). In contrast, functions like xenobiotic metabolic process [85], [86] and positive T-cell selection [87] are suspected to be related to cancer whereas cell projection assembly, flavonoid glucuronidation, and metabolic process, cellular glucuronidation and cellular response to lipid are yet to be explored in-depth.

Top enriched pathways are also related to diverse human cancers and are critical in cancer progression and treatment, for example, drug metabolism affects multidrug resistance and chemotherapy in cancer [88], chemical carcinogenesis is a major reagent in the etiology of cancer [89], [90], ECM-receptor interaction is involved in six critical cancer hallmarks [91], [92] and patients with primary immunodeficiency are at increased risk to develop certain cancers [93], [94]; ascorbate and aldarate metabolism and cytokine-cytokine receptor interaction are related to clinical outcomes of colorectal cancer and recurrence of childhood acute lymphoblastic leukemia [95]–[97]; steroid hormone biosynthesis was identified as critical targets for breast

TABLE III
TOP ENRICHED GO TERMS AND KEGG PATHWAYS RELATED TO LNCRNA SIGNATURE

GO Terms & KEGG Pathways	Description	Cancer relation	Enrichment.	NO. of Genes	Log10(P)
GO Terms					
GO:0003341	cilium movement	Pan-cancer [42-45]	6.6	16	-8.9
GO:0044782	cilium organization	Similar to GO:0003341	2.8	41	-8.5
GO:0030031	cell projection assembly	-	2.4	52	-8.4
GO:0120031	plasma membrane bounded cell projection assembly	-	2.4	51	-8.3
GO:0046649	lymphocyte activation	Breast, cervical, lung, gastric and colorectal cancers [46-48]	2.2	58	-7.9
GO:0035082	axoneme assembly	Similar to GO:0003341	6.4	14	-7.6
GO:0060271	cilium assembly	Similar to GO:0003341	2.7	38	-7.5
GO:0042110	T cell activation	Pan-cancer [51-53]	2.5	42	-7.4
GO:0001578	microtubule bundle formation	Prostate, breast, ovarian and melanoma cancers [66-68]	5	16	-6.9
GO:0007018	microtubule-based movement	Similar to GO:0001578	3	30	-6.9
GO:0052696	flavonoid glucuronidation	-	19	6	-6.8
GO:0009812	flavonoid metabolic process	-	13	7	-6.5
GO:0036159	inner dynein arm assembly	Lung cancer, mediastinal tumor [69]	13	7	-6.5
GO:0052697	xenobiotic glucuronidation	Colon cancer [70,71]	15	6	-6.1
GO:0070286	axonemal dynein complex assembly	-	7.9	9	-5.9
GO:0030098	lymphocyte differentiation	Pan-cancer [49,50]	2.6	32	-5.9
GO:0006805	xenobiotic metabolic process	Under exploration [85,86]	3.8	17	-5.7
GO:0071466	cellular response to xenobiotic stimulus	-	3.1	20	-5.2
GO:0022407	regulation of cell-cell adhesion	Lung and colon cancer [72-75]	2.3	33	-5
GO:0030217	T cell differentiation	Pan-cancer [55]	2.7	23	-4.8
GO:0043368	positive T cell selection	Under exploration [87]	6.6	8	-4.7
GO:0060745	mammary gland branching involved in pregnancy	Breast cancer [76]	19	4	-4.7
GO:0051552	flavone metabolic process	Pan-cancer [56,57]	19	4	-4.7
GO:0042445	hormone metabolic process	Breast, prostate and uterine cancers [77-79]	2.7	22	-4.5
GO:0052695	cellular glucuronidation	-	8.9	6	-4.4
GO:0002534	cytokine production involved in inflammatory response	Similar to GO:0001816	5.2	9	-4.3
GO:0031077	post-embryonic camera-type eye development	-	16	4	-4.3
GO:0001539	cilium or flagellum-dependent cell motility	-	5.8	8	-4.3
GO:0060285	cilium-dependent cell motility	-	5.8	8	-4.3
GO:0001816	cytokine production	Pan-cancer [58-61]	1.8	50	-4.2
GO:0071396	cellular response to lipid	-	1.9	41	-4.1
GO:0006721	terpenoid metabolic process	Pan-cancer [62,63]	3.3	14	-4
GO:1903037	regulation of leukocyte cell-cell adhesion	Pan-cancer [64,65]	2.3	25	-4
GO:0042073	intraciliary transport	-	4.8	9	-4
GO:0001817	regulation of cytokine production	Similar to GO:0001816	1.8	46	-4
GO:0002521	leukocyte differentiation	Breast cancer, leukemia, lymphoma [80,81]	2	36	-4
GO:0007017	microtubule-based process	Similar to GO:0001578	1.8	48	-4
GO:0042573	retinoic acid metabolic process	Prostate cancer [82-84]	6.2	7	-4
KEGG Pathways					
hsa00983	Drug metabolism - other enzymes	Pan-cancer [88]	6.1	10	-5.4
hsa00601	Glycosphingolipid biosynthesis	-	7.3	7	-4.5
hsa00860	Porphyrin and chlorophyll metabolism	-	5.4	8	-4
hsa00830	Retinol metabolism	-	4.3	10	-4
hsa00040	Pentose and glucuronate interconversions	-	5.8	7	-3.8
hsa00014	Glucuronate pathway (uronate pathway)	-	6.3	6	-3.5
hsa00053	Ascorbate and aldarate metabolism	Colorectal cancer [95]	6.3	6	-3.5
hsa04640	Hematopoietic cell lineage	Under exploration [100,101]	3.2	11	-3.2
hsa04060	Cytokine-cytokine receptor interaction	colorectal cancer, leukemia [96,97]	2.2	21	-3.2
hsa00140	Steroid hormone biosynthesis	Breast and prostate cancer [98,99]	3.9	8	-3
hsa05142	Chagas disease	Drug repurpose [118]	3	11	-3
hsa00980	Metabolism of xenobiotics by cytochrome P450	Similar to GO:0006805	3.4	9	-2.9
hsa05204	Chemical carcinogenesis	Pan-cancer [89,90]	3.1	9	-2.6
hsa00982	Drug metabolism - cytochrome P450	Similar to hsa00983	3.2	8	-2.5
hsa04512	ECM-receptor interaction	Pan-cancer [91,92]	2.8	8	-2.1
hsa05144	Malaria	Drug repurpose [116,117,119]	3.5	6	-2.1
hsa05340	Primary immunodeficiency	Pan-cancer [93,94]	3.8	5	-2

and prostate cancer therapy [98], [99]; hematopoietic stem cell approach was recently used in cancer immunotherapy [100], [101]. More interestingly, the co-expressed genes are also enriched in Malaria and Chagas disease pathways (Table III).

G. Association Between ENSG00000206567 and Gynecologic Cancers

Expression patterns of the five lncRNAs were investigated through the identification of co-expressed genes of each lncRNA from the entire TCGA cohort and ranking the studies according to corresponding lncRNA expression value. Such gene enrichment analysis was used to determine whether gene expression patterns plotted according to specific lncRNA were associated with any clinical factors, including age at diagnosis, gender, cancer type, and so on. Genes positively correlated with *ENSG00000206567* displayed relatively higher expression in gynecologic cancers (also known as female reproductive system cancers), including ovarian (OV), uterine corpus endometrial cancer (UCEC), cervical squamous cell carcinoma (CESC), and breast cancer (BRCA), rather than in other human cancers (heatmap in Fig. 6A). However, no such significant correlation was observed for the other four lncRNAs (data not shown here).
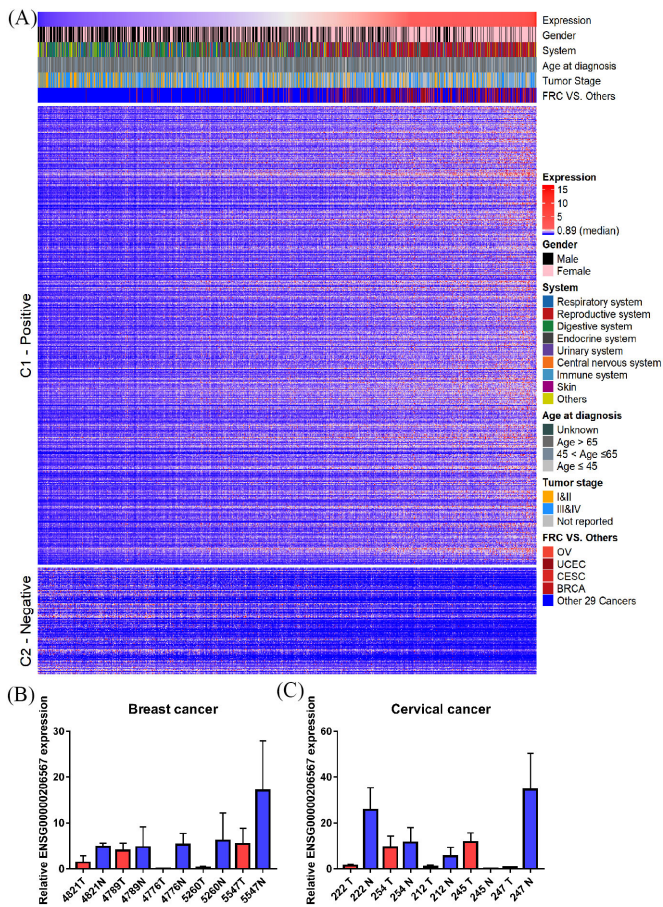
Fig. 6. ENSG00000206567 is associated with gynecologic cancers. (A) Correlation between the co-expressed genes of *ENSG00000206567* and clinical factors. Rows represent co-expressed genes of *ENSG00000206567*, which include 560 positively correlated genes (upper half) and 130 negatively correlated genes (lower half); whereas columns denote 4,231 tumor studies covered 33 cancer types in overall TCGA cohort. Patient cases are ranked according to their expression of *ENSG00000206567* (from low to high or from left to right). The upper second to fifth rows of the heatmap are corresponding to clinical factors. FRC, female reproductive cancers (or gynecologic cancers). (B-C) *ENSG00000206567* is deferentially expressed in breast (B) and cervical (C) cancer and normal tissue pairs.

To further validate this finding, qRT-PCR was conducted using RNAs isolated from breast and cervical tumors and normal tissue pairs. Total RNA was isolated with TRIzol Reagent (Thermo Fisher Scientific; Waltham, MA, USA) following the manufacturer's instructions. cDNA was synthesized from total RNA (2 μg) using a reverse transcription kit (Toyobo Life Science; Osaka, Japan). qPCR was performed in triplicate using 1 μL of cDNA in a standard SYBR Premix Ex Taq (Roche; Pleasanton, CA, USA) on a Real-Time PCR Detection System (480II, Roche). As a result, *ENSG00000206567* was found deferentially expressed in tumor and normal tissue pairs (Fig. 6(B)–C)).

IV. DISCUSSION AND CONCLUSION

Along with the improvement of sequencing techniques, more meaningful noncoding transcripts have been identified as underlying drivers in human cancers [102]–[104]. LncRNAs, the "dark matter" of the non-coding family, have been gradually identified as cancer prognostic predictors and proposed to be therapeutic targets in recent years [11], [105]. In our present work, statistical and machine learning combined analysis enabled us to identify and validate lncRNA biomarkers from pan-cancer studies. The 5-lncRNA risk score model stratified cancer studies into high- and low-risk groups with significantly distinct survival outcomes. An indexing system was developed to illustrate the relationship between 5-lncRNA signature and OS of different types of cancer patients. The lncRNA signature demonstrated superior predictive performance than clinical factors and achieved more accurate longer-term (10 years) mortality prediction after combined with gender, age at diagnosis and tumor stage.

The signature was found associated with characteristics and processes fundamental to human cancers. To a large extent, this relationship supports the signature as a pan-cancer prognostic biomarker. For example, extracellular matrix (ECM) receptor interaction [106], cilium organization [107] and T cell activation [108] were related to positively co-expressed genes of our signature, whereas xenobiotic metabolism [109], [110], inflammatory response [111], and cell adhesion regulation [112] correlated to negatively co-expressed genes. ECM and its receptors have been demonstrated to be involved in all six hallmarks of human cancers [91], including resistance to cell death, induction of angiogenesis, replicative immortality, invasion and metastasis, loss of growth suppression and dysregulated proliferation [113]. The 5-lncRNA signature might, therefore, be a molecular reference for understanding the interplay between ECM, and cancer and stromal cells which might be critical for cancer prevention [91]. The correlation with primary immunodeficiency, T cell activation, inflammatory response, and drug metabolism also implicates a role in the immune response to cancer [114]. The prognosis of cancer patients has been linked to the activation status of innate T cells which is the front line of host defense against cancer [115]. Malaria and Chagas disease were found associated with cancer [116]–[119]. The relation to Malaria and Chagas disease pathways indicates the lncRNA signature identified from the pan-cancer scenario captures shared characteristics that drive diverse human diseases. This may facilitate drug repurposes, for example, Weyerhäuser *et al.* repurposed antimalarial chloroquine for treatment of glioblastoma [119] and Kraus *et al.* used anti-cancer drugs to treat Chagas disease [118]. Functional and tissue studies are needed, however, to determine their exact role in these cellular processes.

There are considerable difficulties for the identification of prognostic biomarkers in the pan-cancer scenario due to cancer-specific context and cancer subtype in the clinic, however, the interdisciplinary research collaborations among computer scientists, biologists, and clinicians empower the capacities and opportunities in searching prognostic predictors common effective among multiple cancers. With the help of artificial intelligence, bioinformatics analysis, and multi-center genomic databases, recent achievements in this field further ascertain the significance and viability of research on lncRNAs with pan-cancer prognostic significance. For instance, a recent study investigated

the unique molecular features common to pan-gynecologic cancers and related them with the prognosis of 2,579 patient studies [120]. Another research investigated lncRNAs with cancer patient prognosis by identification of survival-related lncRNAs in multiple-cancer context [121]. This study demonstrates the great value of pan-cancer research for capturing shared characteristics among cancer-specific prognostic markers. Rather than conflicting with cancer-specific prognosis, the investigation of the pan-cancer prognosis opens up a new perspective for a better and more complete understanding of the cancer mechanisms.

In conclusion, we identified prognostic lncRNAs that significantly associated with the OS of pan-cancer studies by innovatively combining the machine learning model and traditional statistical analysis as a whole biomarker identifier. Although it requires intensive computing power, the proposed framework is a new, streamlined, and intuitive pan-cancer analysis approach, which can be easily understood and implemented. Besides, the lncRNA signature identified with the framework is robust, independent of clinical factors, and demonstrated significant prognostic value on different pan-cancer cohorts and cancer types. With the framework, as far as we know, it is the first time to map lncRNA expression profiles to patient prognosis across multiple cancer types. The lncRNA indexing system and the risk score model may provide insights to develop a "cancer prognosis screening" tool. Meanwhile, the identified lncRNAs are found associated with common biological processes fundamental for cancer initialization and progression, and the majority of them have never reported before which may serve as novel molecular targets for therapy and show efficacy across a broad spectrum of human cancers.

V. DATA AVAILABILITY AND EXPERIMENT REPRODUCIBILITY

All the data used in our study can be downloaded from the TCGA data portal[3] using the official GDC Data Transfer Tool. The computer code used in this study and supplementary material have been released at: https://github.com/guoqingbao/PanCancerLncRNA.

REFERENCES

[1] H. J. West, "Novel precision medicine trial designs: Umbrellas and baskets," *JAMA Oncol.*, vol. 3, no. 3, p. 423, Mar. 2017.

[2] A. Bhan, M. Soleimani, and S. S. Mandal, "Long noncoding RNA and cancer: A new paradigm," *Cancer Res.*, vol. 77, no. 15, pp. 3965–3981, Aug. 2017.

[3] P. G. Sanfilippo and A. W. Hewitt, "Translating the ENCyclopedia of DNA elements project findings to the clinic: ENCODE's implications for eye disease," *Clin. Exp. Ophthalmol*, vol. 42, no. 1, pp. 78–83, Jan./Feb. 2014.

[4] M. E. Forrest and A. M. Khalil, "Review: Regulation of the cancer epigenome by long non-coding RNAs," *Cancer Lett.*, vol. 407, pp. 106–112, Oct. 2017.

[5] Y. Tang *et al.*, "The regulatory role of long noncoding RNAs in cancer," *Cancer Lett.*, vol. 391, pp. 12–19, Apr. 2017.

[6] Y. Tang *et al.*, "Linking long non-coding RNAs and SWI/SNF complexes to chromatin remodeling in cancer," *Mol. Cancer*, vol. 16, no. 1, p. 42, Feb. 2017.

[7] X. Chen, C. C. Yan, X. Zhang, and Z.-H. You, "Long non-coding RNAs and complex diseases: From experimental results to computational models," *Brief. Bioinf.*, vol. 18, no. 4, pp. 558–576, 2017.

[8] X. Chen *et al.*, "Computational models for lncRNA function prediction and functional similarity calculation," *Brief. Functional Genomics*, vol. 18, no. 1, pp. 58–82, Feb. 2019.

[9] Q. Tang and S. S. Hann, "HOTAIR: An oncogenic long non-coding RNA in human cancer," *Cell Physiol. Biochem.*, vol. 47, no. 3, pp. 893–913, 2018.

[10] A. J. Gooding, B. Zhang, L. Gunawardane, A. Beard, S. Valadkhan, and W. P. Schiemann, "The lncRNA BORG facilitates the survival and chemoresistance of triple-negative breast cancers," *Oncogene.*, vol. 38, no. 12, pp. 2020–2041, Nov. 2018.

[11] B. J. Reon, B. Takao Real Karia, M. Kiran, and A. Dutta, "LINC00152 promotes invasion through a 3'-hairpin structure and associates with prognosis in glioblastoma," *Mol. Cancer Res.*, vol. 16, no. 10, pp. 1470–1482, Oct. 2018.

[12] R. Spizzo, M. I. Almeida, A. Colombatti, and G. A. Calin, "Long non-coding RNAs and cancer: A new frontier of translational research?," *Oncogene.*, vol. 31, no. 43, pp. 4577–87, Oct. 2012.

[13] K. A. Hoadley *et al.*, "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer," *Cell*, vol. 173, no. 2, pp. 291–304.e6, 2018.

[14] F. Sanchez-Vega *et al.*, "Oncogenic signaling pathways in the cancer genome atlas," *Cell*, vol. 173, no. 2, pp. 321–337. e10, 2018.

[15] Y. Deng *et al.*, "A pan–cancer atlas of cancer hallmark–associated candidate driver lnc RNA s," *Mol. Oncol.*, vol. 12, no. 11, pp. 1980–2005, 2018.

[16] Y. Li *et al.*, "LncMAP: Pan-cancer atlas of long noncoding RNA-mediated transcriptional network perturbations," *Nucleic Acids Res.*, vol. 46, no. 3, pp. 1113–1123, 2018.

[17] H. S. Chiu *et al.*, "Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context," *Cell Rep.*, vol. 23, no. 1, pp. 297–312 e12, Apr. 2018.

[18] M. Zhou, Z. Zhang, H. Zhao, S. Bao, and J. Sun, "A novel lncRNA-focus expression signature for survival prediction in endometrial carcinoma," *BMC Cancer*, vol. 18, no. 1, 2018, Art. no. 39.

[19] Y. Mao *et al.*, "A seven-lncRNA signature predicts overall survival in esophageal squamous cell carcinoma," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 8823.

[20] M. Zhou, Z. Zhang, H. Zhao, S. Bao, L. Cheng, and J. Sun, "An immune-related six-lncRNA signature to improve prognosis prediction of glioblastoma multiforme," *Mol. Neurobiol.*, vol. 55, no. 5, pp. 3684–3697, 2018.

[21] H.-S. Chiu *et al.*, "Pan-cancer analysis of lncRNA regulation supports their targeting of cancer genes in each tumor context," *Cell Rep.*, vol. 23, no. 1, pp. 297–312. e12, 2018.

[22] X. Chen and G.-Y. Yan, "Novel human lncRNA-disease association inference based on lncRNA expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, Oct. 2013.

[23] X. Chen, "KATZLDA: KATZ measure for the lncRNA-disease association prediction," *Sci. Rep.*, vol. 5, 2015, Art, no. 16840. doi: 10.1038/srep16840

[24] S. Bin Lim, M. L. K. Chua, J. P. S. Yeong, S. J. Tan, W.-T. Lim, and C. T. Lim, "Pan-cancer analysis connects tumor matrisome to immune response," *NPJ Precis. Oncol.*, vol. 3, no. 1, May 2019, Art. no. 15.

[25] M. M. Ali *et al.*, "PAN-cancer analysis of S-phase enriched lncRNAs identifies oncogenic drivers and biomarkers," *Nat. Commun.*, vol. 9, no. 1, Feb. 2018, Art. no. 883.

[26] M. Ung, E. Schaafsma, D. Mattox, G. L. Wang, and C. Cheng, "Pan-cancer systematic identification of lncRNAs associated with cancer prognosis," *Peer J.*, vol. 8, pp. e8797–e8797, 2020.

[27] R. Kandimalla *et al.*, "RNAMethyPro: A biologically conserved signature of N6-methyladenosine regulators for predicting survival at pan-cancer level," *NPJ Precis. Oncol.*, vol. 3, no. 1, May 2019, Art. no. 13.

[28] S. Anders, P. T. Pyl, and W. Huber, "HTSeq-a Python framework to work with high-throughput sequencing data," *Bioinformatics*, vol. 31, no. 2, pp. 166–169, 2015.

[29] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products," (in English), *Chemometrics Intell. Lab. Syst.*, vol. 83, no. 2, pp. 83–90, Sep. 2006.

[30] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," (in English), *Mach. Learn.*, vol. 46, no. 1-3, pp. 389–422, 2002.

[31] T. K. Ho, "The random subspace method for constructing decision forests," (in English), *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, Aug. 1998.

[3][Online]. Available: https://portal.gdc.cancer.gov/

[32] J. Friedman, T. Hastie, and R. Tibshirani, "Random forests," *The Elements Of Statistical Learning* series in statistics New York, NY, USA: Springer, 2001.

[33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[34] M. Gry *et al.*, "Correlations between RNA and protein expression profiles in 23 human cell lines," *BMC Genomics*, vol. 10, no. 1, Aug. 2009, Art. no. 365.

[35] H. Wang, Q. Wang, X. Li, B. Shen, M. Ding, and Z. Shen, "Towards patterns tree of gene coexpression in eukaryotic species," *Bioinformatics*, vol. 24, no. 11, pp. 1367–1373, 2008.

[36] P. J. Heagerty, T. Lumley, and M. S. Pepe, "Time-dependent ROC curves for censored survival data and a diagnostic marker," *Biometrics*, vol. 56, no. 2, pp. 337–344, 2000.

[37] L. Chen, H. M. Linden, B. O. Anderson, and C. I. Li, "Trends in 5-year survival rates among breast cancer patients by hormone receptor status and stage," *Breast Cancer Res. Treat.*, vol. 147, no. 3, pp. 609–616, 2014.

[38] R. Zou, M. Zheng, M. Tan, H. Xu, N. Luan, and L. Zhu, "Decreased PTGDS expression predicting poor survival of endometrial cancer by integrating weighted gene co-expression network analysis and immuno-histochemical validation," *Cancer Manage. Res.*, vol. 12, 2020, Art. no. 5057.

[39] L. Cui, S. Cheng, X. Liu, D. Messadi, Y. Yang, and S. Hu, "Syntenin-1 is a promoter and prognostic marker of head and neck squamous cell carcinoma invasion and metastasis," *Oncotarget*, vol. 7, no. 50, 2016, Art. no. 82634.

[40] M. Lucchetta, I. da Piedade, M. Mounir, M. Vabistsevits, T. Terkelsen, and E. Papaleo, "Distinct signatures of lung cancer types: Aberrant mucin O-glycosylation and compromised immune response," *BMC Cancer*, vol. 19, no. 1, 2019, Art. no. 824.

[41] Z. Liu, M. Deng, L. Wu, and S. Zhang, "An integrative investigation on significant mutations and their down-stream pathways in lung squamous cell carcinoma reveals CUL3/KEAP1/NRF2 relevant subtypes," *Mol. Med.*, vol. 26, no. 1, pp. 48–48, 2020.

[42] I. Sánchez and B. D. Dynlacht, "Cilium assembly and disassembly," *Nat. Cell. Biol.*, vol. 18, no. 7, pp. 711–717, 2016.

[43] J. J. Malicki and C. A. Johnson, "The cilium: Cellular antenna and central processing unit," *Trends Cell Biol.*, vol. 27, no. 2, pp. 126–140, 2017.

[44] M. Mirvis, T. Stearns, and W. James Nelson, "Cilium structure, assembly, and disassembly regulated by the cytoskeleton," *Biochem. J.*, vol. 475, no. 14, pp. 2329–2353, 2018.

[45] L. Fabbri, F. Bost, and N. M. Mazure, "Primary cilium in cancer hall-marks," *Int. J. Mol. Sci.*, vol. 20, no. 6, 2019, Art. no. 1336.

[46] N. A. Hutnick *et al.*, "Exercise and lymphocyte activation following chemotherapy for breast cancer," *Med. Sci. Sports Exerc.*, vol. 37, no. 11, 2005, Art. no. 1827.

[47] F. Triebel, K. Hacene, and M.-F. Pichon, "A soluble lymphocyte activation gene-3 (sLAG-3) protein as a prognostic factor in human breast cancer expressing estrogen or progesterone receptors," *Cancer Lett.*, vol. 235, no. 1, pp. 147–153, Apr. 2006.

[48] Y. He *et al.*, "Lymphocyte–activation gene–3, an important immune checkpoint in cancer," *Cancer Sci.*, vol. 107, no. 9, pp. 1193–1197, 2016.

[49] B. Ruffell, D. G. DeNardo, N. I. Affara, and L. M. Coussens, "Lympho-cytes in cancer development: Polarization towards pro-tumor immunity," *Cytokine Growth Factor Rev.*, vol. 21, no. 1, pp. 3–10, 2010.

[50] B. J. Altman and C. V. Dang, "Normal and cancer cell metabolism: lymphocytes and lymphoma," *FEBS J.*, vol. 279, no. 15, pp. 2598–2609, 2012.

[51] B. J. Miles *et al.*, "Prostate-specific antigen response and systemic T cell activation after in situ gene therapy in prostate cancer patients failing radiotherapy," *Hum. Gene Ther.*, vol. 12, no. 16, pp. 1955–1967, 2001.

[52] S. J. Cronin and J. M. Penninger, "From T–cell activation signals to signaling control of anti-cancer immunity," *Immunol. Rev.*, vol. 220, no. 1, pp. 151–168, 2007.

[53] T. Pentcheva–Hoang, E. Corse, and J. P. Allison, "Negative regulators of T–cell activation: Potential targets for therapeutic intervention in cancer, autoimmune disease, and persistent infections," *Immunol. Rev.*, vol. 229, no. 1, pp. 67–87, 2009.

[54] P.-Y. Pan *et al.*, "Immune stimulatory receptor CD40 is required for T-cell suppression and T regulatory cell activation mediated by myeloid-derived suppressor cells in cancer," *Cancer Res.*, vol. 70, no. 1, pp. 99–108, 2010.

[55] D. E. Speiser, D. T. Utzschneider, S. G. Oberle, C. Münz, P. Romero, and D. Zehn, "T cell differentiation in chronic infection and cancer: functional adaptation or exhaustion?," *Nat. Rev. Immunol.*, vol. 14, no. 11, pp. 768–774, 2014.

[56] M. Abotaleb *et al.*, "Flavonoids in cancer and apoptosis," *Cancers*, vol. 11, no. 1, 2019, Art. no. 28.

[57] P. Batra and A. K. Sharma, "Anti-cancer potential of flavonoids: Recent trends and future perspectives," *3 Biotech.*, vol. 3, no. 6, pp. 439–459, 2013.

[58] T. Fukuyama *et al.*, "Cytokine production of lung cancer cell lines: Cor-relation between their production and the inflammatory/immunological responses both in vivo and in vitro," *Cancer Sci.*, vol. 98, no. 7, pp. 1048–1054, 2007.

[59] D. W. Edwardson, J. Boudreau, J. Mapletoft, C. Lanner, A. T. Kovala, and A. M. Parissenti, "Inflammatory cytokine production in tumor cells upon chemotherapy drug exposure or upon selection for drug resistance," *PLoS One*, vol. 12, no. 9, 2017, Art. no. e0183662.

[60] G. Landskron, M. De la Fuente, P. Thuwajit, C. Thuwajit, and M. A. Hermoso, "Chronic inflammation and cytokines in the tumor microenvironment," *J. Immunol. Res.*, vol. 2014, 2014, Art. no. 149185.

[61] R. J. Dunlop and C. W. Campbell, "Cytokines and advanced cancer," *J. Pain Symptom Manag.*, vol. 20, no. 3, pp. 214–232, 2000.

[62] A. Salminen, M. Lehtonen, T. Suuronen, K. Kaarniranta, and J. Hu-uskonen, "Terpenoids: Natural inhibitors of NF-kB signaling with anti-inflammatory and anticancer potential," *Cell. Mol. Life Sci.*, vol. 65, no. 19, pp. 2979–2999, 2008.

[63] M. Huang, J.-J. Lu, M.-Q. Huang, J.-L. Bao, X.-P. Chen, and Y.-T. Wang, "Terpenoids: Natural products for cancer therapy," *Expert Opin. Investig. Drugs*, vol. 21, no. 12, pp. 1801–1818, 2012.

[64] P. Friedl, Y. Hegerfeldt, and M. Tusch, "Collective cell migration in mor-phogenesis and cancer," *Int. J. Dev. Biol.*, vol. 48, no. 5-6, pp. 441–449, 2004.

[65] G. W. Swart, "Activated leukocyte cell adhesion molecule (CD166/ALCAM): Developmental and mechanistic aspects of cell clustering and cell migration," *Eur. J. Cell Biol.*, vol. 81, no. 6, pp. 313–322, 2002.

[66] M. A. Jordan and L. Wilson, "Microtubules and actin filaments: Dynamic targets for cancer chemotherapy," *Curr. Opin. Cell Biol.*, vol. 10, no. 1, pp. 123–130, 1998.

[67] L. Sepp-Lorenzino *et al.*, "The microtubule-stabilizing agents epothilones A and B and their desoxy-derivatives induce mitotic arrest and apoptosis in human prostate cancer cells," *Prostate Cancer Prostatic Dis.*, vol. 2, no. 1, pp. 41–52, 1999.

[68] A. I. Einzig, P. H. Wiernik, and E. L. Schwartz, "Taxol: a new agent active in melanoma and ovarian cancer," *New Drugs, Concepts Results in Cancer Chemotherapy* Berlin, Germany: Springer, 1992, pp. 89–100.

[69] M. Horie, H. Arai, S. Noguchi, M. Suzuki, Y. Sakamoto, and T. Oka, "Kartagener syndrome with lung cancer and mediastinal tumor," *J. Japanese Respir. Soc.*, vol. 48, no. 5, pp. 375–378, 2010.

[70] J. Cummings *et al.*, "Glucuronidation as a mechanism of intrinsic drug resistance in human colon cancer: Reversal of resistance by food additives," *Cancer Res.*, vol. 63, no. 23, pp. 8443–8450, 2003.

[71] H. A. Zahreddine and K. L. Borden, "Molecular pathways: Gli1-induced drug glucuronidation in resistant cancer cells," *Clin. Cancer Res.*, vol. 21, no. 10, pp. 2207–2210, 2015.

[72] S. Hirohashi, "Inactivation of the E-cadherin-mediated cell adhesion system in human cancers," *Amer. J. Pathol.*, vol. 153, no. 2, pp. 333–339, 1998.

[73] B. Wijnhoven, W. Dinjens, and M. Pignatelli, "E–cadherin–catenin cell–cell adhesion complex and human cancer," *Brit. J. Surg.*, vol. 87, no. 8, pp. 992–1005, 2000.

[74] A. S. Jaiswal, B. P. Marlow, N. Gupta, and S. Narayan, "β-catenin-mediated transactivation and cell–cell adhesion pathways are important in curcumin (diferuylmethane)-induced growth arrest and apoptosis in colon cancer cells," *Oncogene*, vol. 21, no. 55, pp. 8414–8427, 2002.

[75] R. M. Bremnes, R. Veve, F. R. Hirsch, and W. A. Franklin, "The E-cadherin cell–cell adhesion complex and lung cancer invasion, metastasis, and prognosis," *Lung Cancer*, vol. 36, no. 2, pp. 115–124, May 2002.

[76] M. D. Sternlicht, "Key stages in mammary gland development: The cues that regulate ductal branching morphogenesis," *Breast Cancer Res.*, vol. 8, no. 1, 2005, Art. no. 201.

[77] J. E. Mortimer, F. Dehdashti, B. A. Siegel, K. Trinkaus, J. A. Katzenel-lenbogen, and M. J. Welch, "Metabolic flare: Indicator of hormone responsiveness in advanced breast cancer," *J. Clin. Oncol.*, vol. 19, no. 11, pp. 2797–2803, 2001.

[78] H. P. Yang *et al.*, "Common genetic variation in the sex hormone metabolic pathway and endometrial cancer risk: Pathway-based eval-uation of candidate genes," *Carcinogenesis*, vol. 31, no. 5, pp. 827–833, 2010.

[79] N. M. Makridakis and J. K. V. Reichardt, "Molecular epidemiology of hormone-metabolic loci in prostate cancer," *Epidemiol. Rev.*, vol. 23, no. 1, pp. 24–29, 2001.

[80] T. Lança and B. Silva-Santos, "The split nature of tumor-infiltrating leukocytes: Implications for cancer surveillance and immunotherapy," *Oncoimmunology*, vol. 1, no. 5, pp. 717–725, 2012.

[81] Y.-g. Man, "Aberrant leukocyte infiltration: a direct trigger for breast tumor invasion and metastasis," *Int. J. Biol. Sci.*, vol. 6, no. 2, pp. 129–132, 2010.

[82] W. H. Miller Jr, "The emerging role of retinoids and retinoic acid metabolism blocking agents in the treatment of cancer," *Cancer: Interdisciplinary Int. J. Amer. Cancer Soc.*, vol. 83, no. 8, pp. 1471–1482, 1998.

[83] X.-H. Tang and L. J. Gudas, "Retinoids, retinoic acid receptors, and cancer," *Annu. Rev. Pathol.: Mechanisms Dis.*, vol. 6, pp. 345–364, 2011.

[84] C. Huynh, A. Brodie, and V. Njar, "Inhibitory effects of retinoic acid metabolism blocking agents (RAMBAs) on the growth of human prostate cancer cells and LNCaP prostate tumour xenografts in SCID mice," *Brit. J. Cancer*, vol. 94, no. 4, pp. 513–523, 2006.

[85] R. H. Dashwood, "Xenobiotic metabolism relevance to cancer," *J. Nutrition*, vol. 136, no. 10, pp. 2681S–2682S, 2006.

[86] V. Tamási, K. Monostory, R. A. Prough, and A. Falus, "Role of xenobiotic metabolism in cancer: Involvement of transcriptional and miRNA regulation of P450s," *Cell. Mol. Life Sci.*, vol. 68, no. 7, pp. 1131–1146, 2011.

[87] T. K. Starr, S. C. Jameson, and K. A. Hogquist, "Positive and negative selection of T cells," *Annu. Rev. Immunol.*, vol. 21, no. 1, pp. 139–176, 2003.

[88] D. W. Nebert, "Role of genetics and drug metabolism in human cancer risk," *Mutation Res./Fundam. Mol. Mechanisms Mutagenesis*, vol. 247, no. 2, pp. 267–281, Apr. 1991.

[89] G. N. Wogan, S. S. Hecht, J. S. Felton, A. H. Conney, and L. A. Loeb, "Environmental and chemical carcinogenesis," *Seminars in Cancer Biol.*, vol. 14, no. 6, pp. 473–486, 2004.

[90] L. A. Loeb and C. C. Harris, "Advances in chemical carcinogenesis: A historical review and prospective," *Cancer Res.*, vol. 68, no. 17, pp. 6863–6872, 2008.

[91] M. W. Pickup, J. K. Mouw, and V. M. Weaver, "The extracellular matrix modulates the hallmarks of cancer," *EMBO Rep.*, vol. 15, no. 12, pp. 1243–1253, 2014.

[92] C. Walker, E. Mojares, and A. del Rìo Hernández, "Role of extracellular matrix in development and cancer progression," *Int. J. Mol. Sci.*, vol. 19, no. 10, 2018, Art. no. 3028.

[93] N. F. de Miranda, A. Björkman, and Q. Pan–Hammarström, "DNA repair: The link between primary immunodeficiency and cancer," *Ann. New York Acad. Sci.*, vol. 1246, no. 1, pp. 50–63, 2011.

[94] P. C. Mayor *et al.*, "Cancer in primary immunodeficiency diseases: Cancer incidence in the united states immune deficiency network registry," *J. Allergy Clin. Immunol.*, vol. 141, no. 3, pp. 1028–1035, 2018.

[95] C. Tao, J. Sun, W. J. Zheng, J. Chen, and H. Xu, "Colorectal cancer drug target prediction using ontology-based inference and network analysis," *Database*, vol. 2015, 2015, Art. no. bav015.

[96] S. Wu, R. Geß ner, A. von Stackelberg, R. Kirchner, G. Henze, and K. Seeger, "Cytokine/cytokine receptor gene expression in childhood acute lymphoblastic leukemia: Correlation of expression and clinical outcome at first disease recurrence," *Cancer: Interdisciplinary Int. J. Amer. Cancer Soc.*, vol. 103, no. 5, pp. 1054–1063, 2005.

[97] C. Dong *et al.*, "Identification of a cytokine–cytokine receptor interaction gene signature for predicting clinical outcomes in patients with colorectal cancer," *Int. J. Clin. Exp. Med*, vol. 10, pp. 9009–9018, 2017.

[98] E. A. Mostaghel, "Steroid hormone synthetic pathways in prostate cancer," *Transl. Androl. Urol.*, vol. 2, no. 3, pp. 212–227, 2013.

[99] D. Africander and K.-H. Storbeck, "Steroid metabolism in breast cancer: Where are we and what are we missing?," *Mol. Cell. Endocrinol.*, vol. 466, pp. 86–97, May 2018.

[100] J. E. Adair, S. P. Kubek, and H.-P. Kiem, "Hematopoietic stem cell approaches to cancer," *Hematol./Oncol. Clin.*, vol. 31, no. 5, pp. 897–912, 2017.

[101] E. Gschweng, S. De Oliveira, and D. B. Kohn, "Hematopoietic stem cells for cancer immunotherapy," *Immunol. Rev.*, vol. 257, no. 1, pp. 237–249, 2014.

[102] G. Ciriello, M. L. Miller, B. A. Aksoy, Y. Senbabaoglu, N. Schultz, and C. Sander, "Emerging landscape of oncogenic signatures across human cancers," *Nat. Genet*, vol. 45, no. 10, pp. 1127–33, Oct. 2013.

[103] C. Kandoth *et al.*, "Mutational landscape and significance across 12 major cancer types," *Nature*, vol. 502, no. 7471, pp. 333–339, Oct. 2013.

[104] T. I. Zack *et al.*, "Pan-cancer patterns of somatic copy number alteration," *Nat. Genet*, vol. 45, no. 10, pp. 1134–40, Oct. 2013.

[105] H. Wang *et al.*, "Long noncoding RNA miR503HG, a prognostic indicator, inhibits tumor metastasis by regulating the HNRNPA2B1/NF-kappaB pathway in hepatocellular carcinoma," *Theranostics*, vol. 8, no. 10, pp. 2814–2829, 2018.

[106] P. Lu, V. M. Weaver, and Z. Werb, "The extracellular matrix: A dynamic niche in cancer progression," *J. Cell Biol.*, vol. 196, no. 4, pp. 395–406, Feb. 2012.

[107] A. M. Fry, M. J. Leaper, and R. Bayliss, "The primary cilium: Guardian of organ development and homeostasis," *Organogenesis*, vol. 10, no. 1, pp. 62–8, Jan. 2014.

[108] D. E. Speiser, P. C. Ho, and G. Verdeil, "Regulatory circuits of T cell function in cancer," *Nat. Rev. Immunol.*, vol. 16, no. 10, pp. 599–611, 2016.

[109] V. P. Androutsopoulos, A. M. Tsatsakis, and D. A. Spandidos, "Cytochrome P450 CYP1A1: Wider roles in cancer progression and prevention," *BMC Cancer*, vol. 9, Jun. 2009, Art. no. 187.

[110] K. F. Windmill, R. A. McKinnon, X. Zhu, A. Gaedigk, D. M. Grant, and M. E. McManus, "The role of xenobiotic metabolizing enzymes in arylamine toxicity and carcinogenesis: Functional and localization studies," *Mutat. Res.*, vol. 376, no. 1-2, pp. 153–60, May 1997.

[111] M. Moossavi, N. Parsamanesh, A. Bahrami, S. L. Atkin, and A. Sahebkar, "Role of the NLRP3 inflammasome in cancer," *Mol. Cancer*, vol. 17, no. 1, 2018, Art. no. 158.

[112] A. M. Mendonsa, T. Y. Na, and B. M. Gumbiner, "E-cadherin in contact inhibition and cancer," *Oncogene*, vol. 37, no. 35, pp. 4769–4780, 2018.

[113] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, no. 5, pp. 646–74, Mar. 2011.

[114] E. Dugnani, V. Pasquale, C. Bordignon, A. Canu, L. Piemonti, and P. Monti, "Integrating T cell metabolism in cancer immunotherapy," *Cancer Lett.*, vol. 411, pp. 12–18, Dec. 2017.

[115] D. Stolk, H. J. van der Vliet, T. D. de Gruijl, Y. van Kooyk, and M. A. Exley, "Positive & negative roles of innate effector cells in controlling cancer progression," *Front. Immunol.*, vol. 9, 2018, Art. no. 1990.

[116] D. F. Robbiani *et al.*, "Plasmodium infection promotes genomic instability and AID-dependent B cell lymphoma," *Cell*, vol. 162, no. 4, pp. 727–737, 2015.

[117] A. V. Nordor, D. Bellet, and G. H. Siwo, "Cancer-malaria: Hidden connections," *Open Biol.*, vol. 8, no. 10, 2018, Art. no. 180127.

[118] J. M. Kraus, C. L. Verlinde, M. Karimi, G. I. Lepesheva, M. H. Gelb, and F. S. Buckner, "Rational modification of a candidate cancer drug for use against Chagas disease," *J. Med. Chem.*, vol. 52, no. 6, pp. 1639–1647, 2009.

[119] P. Weyerhäuser, S. R. Kantelhardt, and E. L. Kim, "Re-purposing chloroquine for glioblastoma: Potential merits and confounding variables," *Front. Oncol. Mini Rev.*, vol. 8, 2018, Art. no. 335.

[120] A. C. Berger *et al.*, "A comprehensive pan-cancer molecular study of gynecologic and breast cancers," *Cancer Cell*, vol. 33, no. 4, pp. 690–705. e9, 2018.

[121] S. V. Vasaikar, P. Straub, J. Wang, and B. Zhang, "LinkedOmics: Analyzing multi-omics data within and across 32 cancer types," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D956–D963, 2017.