Integrated Systems and Technologies

Cancer Research

# TANRIC: An Interactive Open Platform to Explore the Function of lncRNAs in Cancer

Jun Li[1], Leng Han[1], Paul Roebuck[1], Lixia Diao[1], Lingxiang Liu[1,2], Yuan Yuan[1], John N. Weinstein[1,3], and Han Liang[1,3,4,5]

## Abstract

Long noncoding RNAs (lncRNA) have emerged as essential players in cancer biology. Using recent large-scale RNA-seq datasets, especially those from The Cancer Genome Atlas (TCGA), we have developed "The Atlas of Noncoding RNAs in Cancer" (TANRIC; http://bioinformatics.mdanderson.org/main/TANRIC:Overview), a user-friendly, open-access web resource for interactive exploration of lncRNAs in cancer. It characterizes the expression profiles of lncRNAs in large patient cohorts of 20 cancer types, including TCGA and independent datasets (>8,000 samples overall). TANRIC enables researchers to rapidly and intuitively analyze lncRNAs of interest (annotated lncRNAs or any user-defined ones) in the context of clinical and other molecular data, both within and across tumor types. Using TANRIC, we have identified a large number of lncRNAs with potential biomedical significance, many of which show strong correlations with established therapeutic targets and biomarkers across tumor types or with drug sensitivity across cell lines. TANRIC represents a valuable tool for investigating the function and clinical relevance of lncRNAs in cancer, greatly facilitating lncRNA-related biologic discoveries and clinical applications. *Cancer Res; 75(18); 3728–37. ©2015 AACR.*

## Introduction

The human genome encodes approximately 20,000 protein-coding genes and also a large number of transcriptionally active, noncoding RNAs (~14,000 according to the ENCODE annotation; ref. 1). Among noncoding RNAs, long noncoding RNAs (lncRNA), typically >200 bp, have increasingly been recognized as playing essential roles in tumor biology, representing a new focus in cancer research (2–4). Emerging evidence has indicated that lncRNAs contribute to tumor initiation and progression through diverse mechanisms ranging from epigenetic regulation of key cancer genes (5, 6) and enhancer-associated activity (7) to post-transcriptional processing of mRNAs (8, 9). Therefore, central tasks in cancer research are the identification of lncRNA components involved in carcinogenesis and elucidation of their functions in specific tumor contexts. That inquiry is expected to lay the foundation for development of novel biomarkers and therapeutic agents.

Recent RNA-seq data over large cancer patient cohorts provide an unprecedented opportunity to pursue that inquiry in a systematic way. In particular, The Cancer Genome Atlas (TCGA) represents a unique resource as it generates multidimensional data at the DNA, RNA, and protein levels for a broad range of human tumor types (10). However, there are several computational challenges for biomedical researchers to make full use of these data and prioritize lncRNAs for further functional investigations. First, the number of expressed lncRNAs in human cancers is large. For example, a very recent pan-cancer analysis reveals approximately 8,000 tumor-specific or lineage-specific lncRNAs (11). In terms of prioritizing lncRNAs with potential clinical relevance and elucidating their mechanisms, it is very informative to perform the correlation analysis of lncRNA expression with clinical variables (e.g., patient survival) or with the molecular characteristics of driver genes or therapeutic targets (e.g., PTEN loss or HER2 status) over large patient cohorts. But because of high dimension and complexity of the data involved, such analyses are often daunting and time-consuming. Second, the annotation of lncRNAs in the human genome is rough, very incomplete and fast evolving, so it is important for researchers to be able to query the expression profiles of user-defined lncRNAs (based on genomic coordinates). This function is not available in current lncRNA-related bioinformatics resources as it requires the calculation directly from a huge amount of raw RNA-seq mapping files. Third, given lncRNA candidates of interest, it is critical to examine their profiles in a variety of cancer cell lines, which allows researchers to choose appropriate model systems for experimental studies. Unfortunately, efficient bioinformatics tools with the above functions are still missing, representing a major barrier for the cancer research community to a systems-level understanding of the function and underlying mechanisms of lncRNAs.

To fill the gap, we have developed The Atlas of Noncoding RNAs in Cancer (TANRIC), a user-friendly, open resource for interactive exploration of lncRNAs in the context of TCGA clinical and

[1]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas. [2]Department of Oncology, Nanjing Medical University, Nanjing, Jiangsu, China. [3]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas. [4]The University of Texas Graduate School of Biomedical Sciences, Houston, Texas. [5]Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, Houston, Texas.

**Note:** Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

J. Li and L. Han contributed equally to this article.

**Corresponding Author:** Han Liang, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030. Phone: 713-745-9815; Fax: 713-563-4242; E-mail: hliang1@mdanderson.org

**doi:** 10.1158/0008-5472.CAN-15-0273

genomic data. Using TANRIC, we have demonstrated that a large number of lncRNA species show differential expression among known tumor subtypes or in correlation with clinical variables; many lncRNAs show strong correlations with established therapeutic targets and biomarkers across tumor types or with drug sensitivity across cell lines; and the tumor subtypes defined by lncRNA-expression profiles show extensive concordance with established tumor subtypes and provide potential prognostic value.

## Materials and Methods

### Data resource

We downloaded RNA-seq BAM files of 6,309 patient samples (including 6,083 primary tumor samples and 226 metastasis samples) across 20 TCGA cancer types and their related 564 non-tumor tissue samples (if available; ref. 10) from the UCSC Cancer Genomics Hub (CGHub, https://cghub.ucsc.edu/). Included were bladder urothelial carcinoma (BLCA), brain lower-grade glioma (LGG), breast invasive carcinoma (BRCA), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), colon adenocarcinoma (COAD), skin cutaneous melanoma (SKCM), glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), rectum adenocarcinoma (READ), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), and uterine corpus endometrioid carcinoma (UCEC). We also downloaded 739 BAM files of Cancer Cell Line Encyclopedia (CCLE) cell lines (12) from CGHub. In addition, we obtained the RNA-seq files of 531 samples from another three independent studies, including LUAD (13), clear-cell renal cell carcinoma (14), and glioblastomas (15). In total, the current TANRIC release includes RNA-seq data from 8,143 samples (1,142 billion reads).

### Efficient algorithm for expression quantitation of user-defined lncRNAs

To calculate the expression of a user-defined lncRNA, TANRIC accepts the genomic coordinates of multiple segments as the input (e.g., given a lncRNA of 3 exons, the input could be "chr7:27135713–27136007;27138458–27138985;27139398–27139585"). The total exon length of a queried lncRNA should be shorter than 50 kb. To minimize the computation time for quantifying user-defined lncRNA expression, we preprocessed all raw BAM files through three steps: (i) extraction of sequence depth data from raw BAM files using SAMtools; (ii) division of genome-wide depth data into short segments (~3,000,000 bp); (iii) merged of the data into a single file, thereby minimizing the file input/output time when dealing with hundreds of samples for each cancer type; (iv) compression of the merged depth files using a block compression algorithm; and (v) generation of

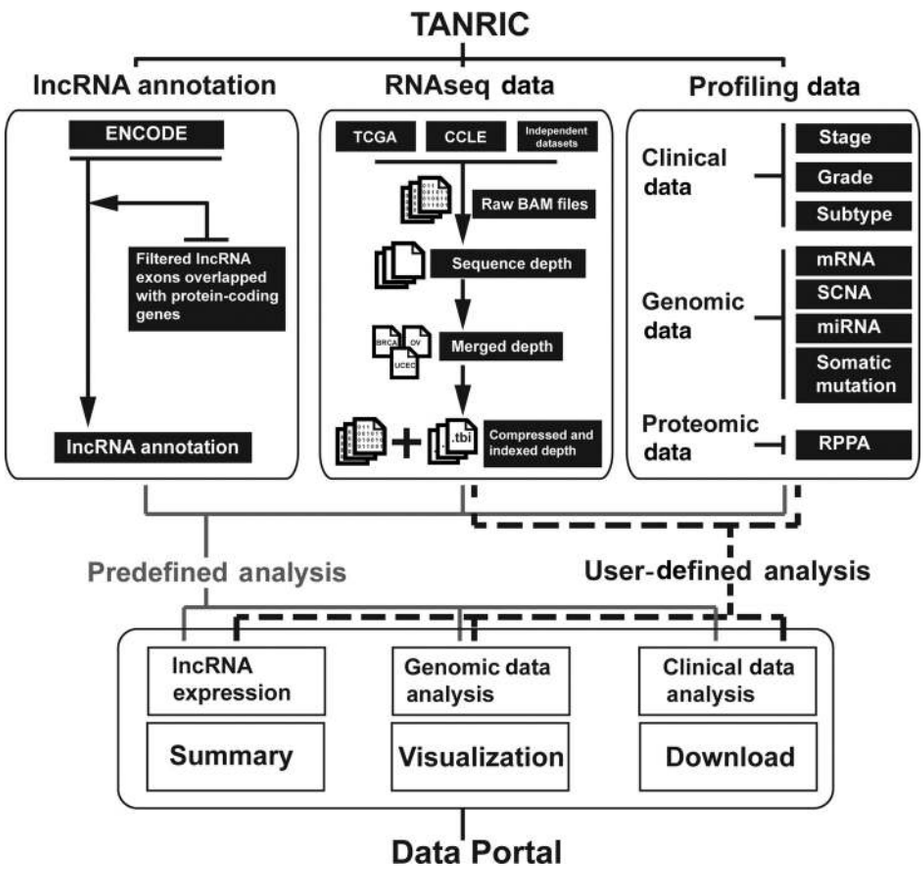


**Figure 1.**
Summary of TANRIC architecture.

**Table 1.** Summary of the data resources of the current TANRIC release

| Data source | Cancer type | #Normal samples | #Tumor samples | Sequencing strategy | Read length | #Expressed lncRNAs[a] |
|---|---|---|---|---|---|---|
| TCGA | Bladder urothelial carcinoma (BLCA) | 19 | 252 | Paired-end | 48 | 1,958 |
| TCGA | Brain lower grade glioma (LGG) | 0 | 486 | Paired-end | 48 | 2,301 |
| TCGA | Breast invasive carcinoma (BRCA) | 105 | 837 | Paired-end | 50 | 1,960 |
| TCGA | Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) | 3 | 196 | Paired-end | 48 | 1,846 |
| TCGA | Colon adenocarcinoma (COAD) | 0 | 157 | Single-end | 76 | 714 |
| TCGA | Skin cutaneous melanoma (SKCM) | 0 | 226 | Paired-end | 48 | 1,755 |
| TCGA | Glioblastoma multiforme (GBM) | 0 | 154 | Paired-end | 76 | 2,369 |
| TCGA | Head and neck squamous cell carcinoma (HNSC) | 42 | 426 | Paired-end | 48 | 1,357 |
| TCGA | Kidney chromophobe (KICH) | 25 | 66 | Paired-end | 48 | 1,971 |
| TCGA | Kidney renal clear cell carcinoma (KIRC) | 67 | 448 | Paired-end | 50 | 2,111 |
| TCGA | Kidney renal papillary cell carcinoma (KIRP) | 30 | 198 | Paired-end | 48 | 2,118 |
| TCGA | Liver hepatocellular carcinoma (LIHC) | 50 | 200 | Paired-end | 48 | 1,446 |
| TCGA | Lung adenocarcinoma (LUAD) | 58 | 488 | Paired-end | 48 | 2,031 |
| TCGA | Lung squamous cell carcinoma (LUSC) | 17 | 220 | Paired-end | 50 | 1,883 |
| TCGA | Ovarian serous cystadenocarcinoma (OV) | 0 | 412 | Paired-end | 75 | 1,866 |
| TCGA | Prostate adenocarcinoma (PRAD) | 52 | 374 | Paired-end | 48 | 2,010 |
| TCGA | Rectal adenocarcinoma (READ) | 0 | 71 | Single-end | 76 | 716 |
| TCGA | Stomach adenocarcinoma (STAD) | 33 | 285 | Paired-end | 75 | 1,328 |
| TCGA | Thyroid carcinoma (THCA) | 59 | 497 | Paired-end | 48 | 1,900 |
| TCGA | Uterine corpus endometrioid carcinoma (UCEC) | 4 | 316 | Single-end | 76 | 855 |
| CCLE | Tumor cell lines | 0 | 739 | Paired-end | 101 | 2,137 |
| Independent | Chinese_GBM | 0 | 274 | Paired-end | 101 | 2,419 |
| Independent | Japanese_KIRC | 0 | 97 | Paired-end | 100 | 2,308 |
| Independent | Korean_LUAD | 77 | 83 | Paired-end | 101 | 2,569 |

[a]Expressed lncRNAs defined as those with an average RPKM $\geq$ 0.3 across all samples in each cancer type.

corresponding index files for quick location and retrieval of queried data. We quantified the lncRNA expression as reads per kilobase per million mapped reads (RPKM; ref. 16) and generated the expression profile in a dynamic table. With data preprocessing, the time for calculating lncRNA expression was reduced by >100-fold compared with that of SAMtools. Currently, TANRIC, operating single threaded, can generate the expression profile for any user-defined lncRNA in a minute. That capability dramatically improves performance, enabling rapid analysis of specified lncRNAs through a web interface.

### Implementation of the TANRIC data portal

The expression data on annotated lncRNAs and the precalculated correlations with clinical and genomic data are stored in CouchDB. Correlation, differential analyses, and survival analyses were performed in R. The Web interface was implemented in JavaScript; tables were visualized by DataTables; the embedded plots were based on HighCharts; and heat maps were generated using the Next-Generation Clustered Heat Map tool (Broom, Weinstein and colleagues; unpublished data).

### Expression quantitation of annotated lncRNAs

To perform a comprehensive survey of human lncRNAs, we obtained the genomic coordinates of 13,870 human lncRNAs from the GENCODE Resource (version 19; ref. 1). We further filtered those lncRNA exons that overlapped with any known coding genes based on the gene annotations of GENCODE (1) and RefGene. As a result, the analysis focused on the remaining 12,727 lncRNAs. On the basis of the BAM files, we quantified the expression levels of lncRNAs as RPKM, and the lncRNAs with detectable expression were defined as those with an average RPKM $\geq$ 0.3 across all samples in each cancer type, as defined in the literature (17).

### Analysis of expressed lncRNAs for biomedical significance

We obtained the clinical information associated with tumor samples, including the patient's overall survival time, tumor stage, and tumor grade from Synapse TCGA Pan-Cancer data portal (https://www.synapse.org/), with ID syn300013. We also obtained known tumor subtype information from TCGA marker articles (if available). To identify lncRNAs differentially expressed between tumors and matched normal samples, we used the paired Student *t* test to assess the statistical difference between the two groups. To identify lncRNAs differentially expressed among established tumor subtypes or tumor stages, we used ANOVA to assess the statistical difference. Groups with fewer than five samples were excluded from the analysis.

### Analysis of lncRNA expression related to potential clinical applications

We obtained a list of 121 actionable target genes from Van Allen and colleagues (18) and added two genes that are well-established targets in immune therapy. We downloaded TCGA molecular profiling data of these target genes, including somatic mutations, mRNA expression, miRNA expression, and somatic copy-number alteration (SCNA) data from Synapse TCGA Pan-Cancer data portal. Student *t* tests were used to assess the statistical difference in lncRNA expression between mutated and wild-type samples given a gene of interest, and Spearman rank correlations were used to assess relationships between lncRNA expression and SCNA or mRNA, with a coefficient (absolute value) cutoff of 0.6. Multiple comparisons correction was performed using the Benjamini–Hochberg method with a corrected false discovery rate (FDR) cutoff of 0.05, and a 2-fold change between at least two groups was also required. To assess the effects of lncRNA expression on drug sensitivity, we downloaded the drug screening data from CCLE (http://www.broadinstitute.org/ccle/home), and calculated the correlations between the expression levels of approximately 1,290 expressed

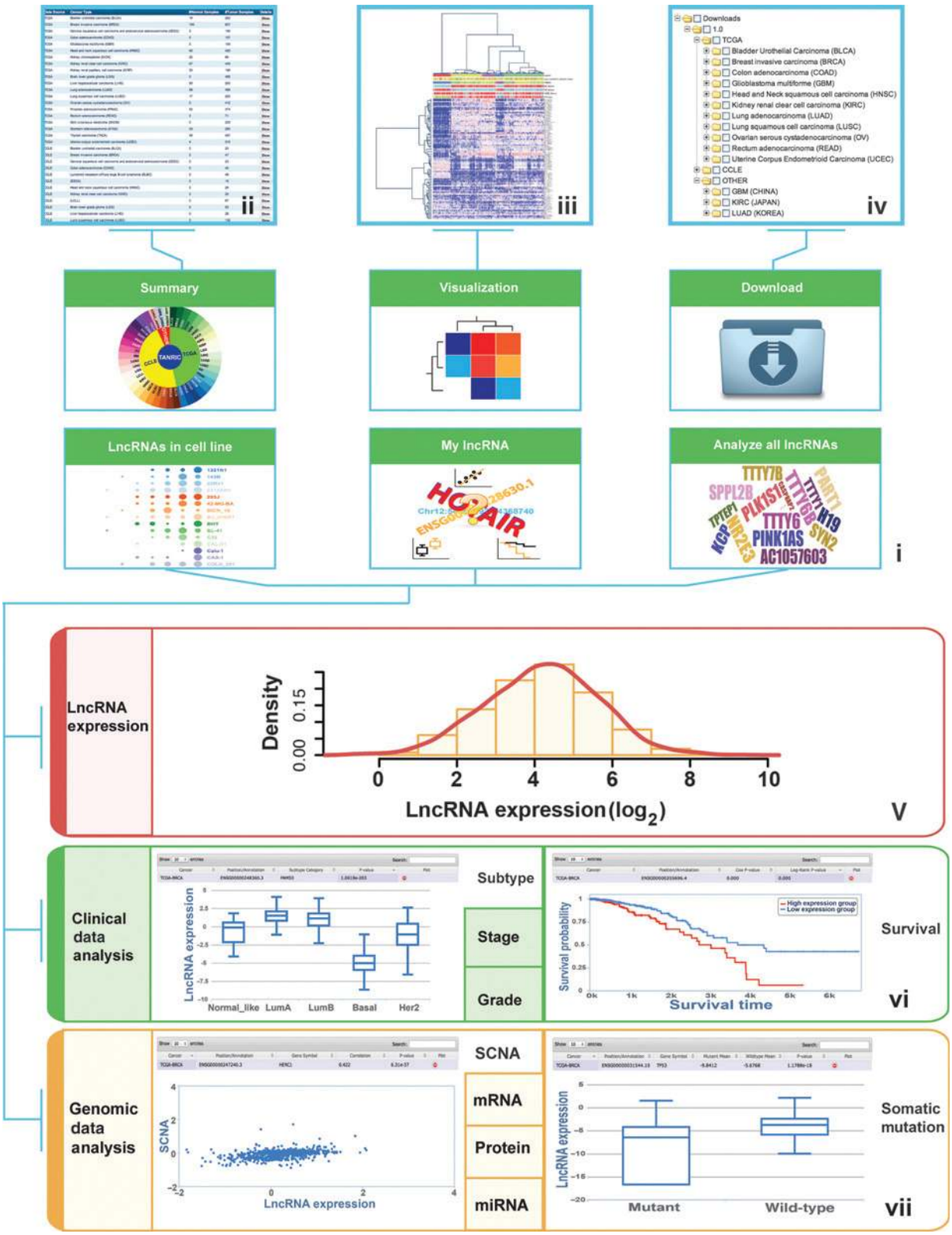


**Figure 2.**
Overview of TANRIC data portal. The panel of six modules (i); the "Summary" module (ii); the "next-generation clustered heat map" view in the "Visualization" module (iii); the "Download" module (iv); the three analysis modules provide raw expression data on a lncRNA of interest (v); the analysis modules provide clinical data analysis of lncRNAs (including differential analysis among tumor subtypes, stages, and grades) and analysis of correlation with patient survival (vi); and the analysis modules provide genomic data analysis of lncRNAs, including differential analysis between mutated and wild-type samples for a protein-coding gene of interest and analysis of correlations with SCNA, miRNA, mRNA, and protein expression (vii).

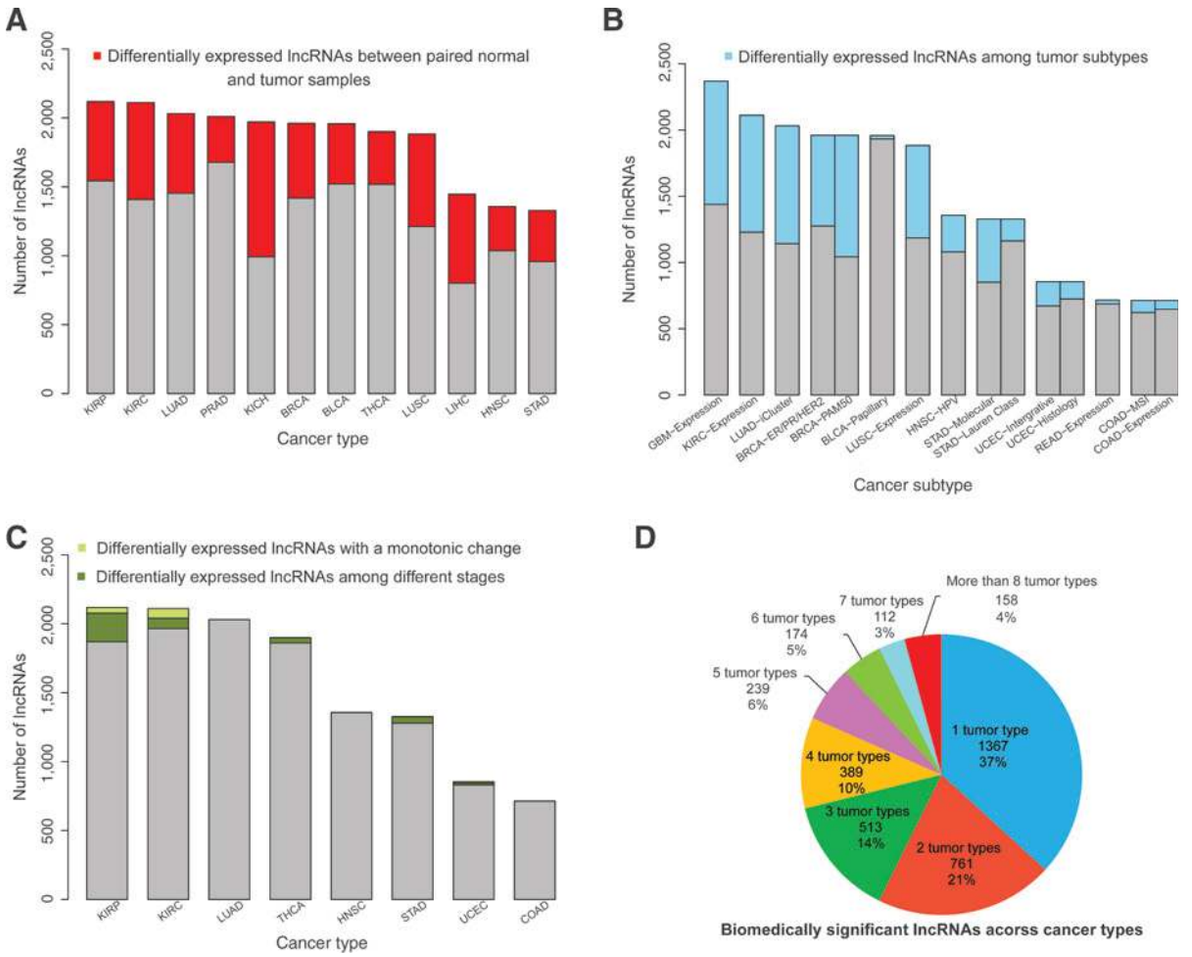


**Figure 3.**
A large number of lncRNAs with potential biomedical significance in various cancer types. A, the total bars represent the numbers of expressed lncRNAs; red represents the numbers of lncRNAs differentially expressed between tumor and matched normal samples across tumor types. B, the total bars represent the numbers of expressed lncRNAs; blue represents the numbers of differentially expressed lncRNAs among known tumor subtypes. C, the total bars represent the numbers of expressed lncRNAs; green represents the numbers of differentially expressed lncRNAs among clinical stages, among which the light green parts represent those with a pattern of consistent increase or decrease across stages. D, the pie chart showing the numbers of lncRNAs with biomedical significance across tumor types.

lncRNAs and the $IC_{50}$ values of 24 drugs across approximately 330 cell lines. Spearman rank correlations were used to detect significant correlations with a coefficient (absolute value) cutoff of 0.3.

### Analysis of tumor subtypes revealed by lncRNA expression

To classify tumor subtypes based on lncRNA expression, for each cancer type, we selected the 500 lncRNAs with the most variable expression pattern and used ConsensusClusterPlus (19) to classify the tumor samples into sample clusters (subtypes). We then used the $\chi^2$ test to determine concordance between lncRNA-expression subtypes and known subtypes, and the log-rank test to examine whether lncRNA-expression subtypes significantly correlated with the overall patient survival times. To understand the molecular mechanisms associated with lncRNA subtypes, we downloaded reverse-phase protein array (RPPA) expression data from The Cancer Proteome Atlas (20). Pathway analysis was conducted as previously described (21). Briefly, the members of each pathway were predefined on the basis of a literature search. RPPA data were median-centered and normalized by SD across all samples for each component to obtain relative protein levels. The pathway score was then taken as the sum of the relative protein levels of all positive regulatory components minus the equivalent sum for the negative regulatory components in a particular pathway. Antibodies targeting different phosphorylated forms of the same protein with Pearson correlation coefficient >0.85 were averaged. We used a Student *t* test or ANOVA analysis to assess statistical differences in pathway score among groups, using the Benjamini–Hochberg correction, with FDR cutoff of 0.05.

## Results

### A user-friendly, interactive, open-access platform for exploring the function of lncRNAs in cancer

To provide a comprehensive lncRNA resource to the cancer research community, we have collected large-scale RNA-seq

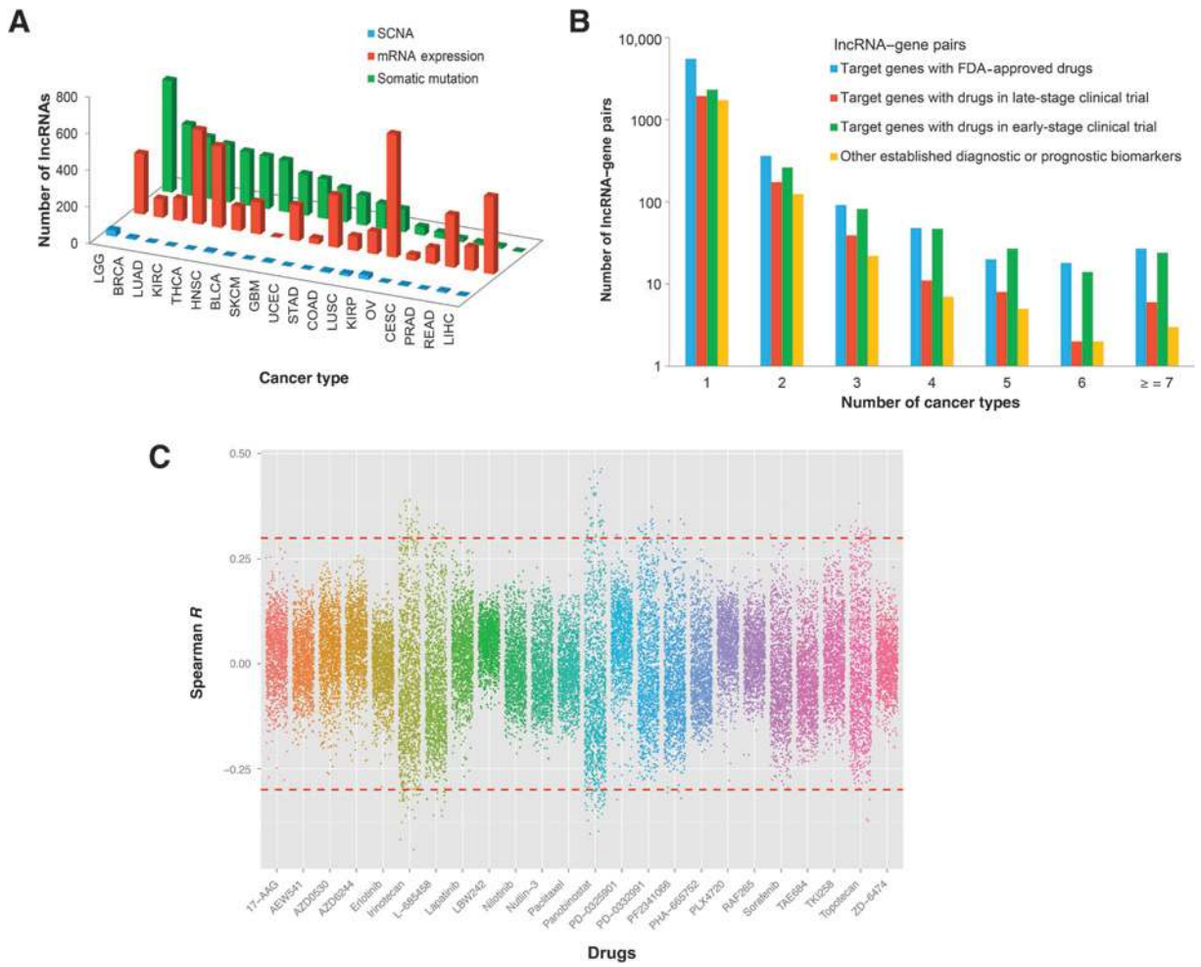


**Figure 4.**
Associations of lncRNAs with clinically actionable genes or drug sensitivity. A, numbers of lncRNAs for which the expressed levels are associated with an SCNA, mRNA expression, or somatic mutation of clinically actionable genes in each cancer type. B, numbers of lncRNA–gene pairs across multiple cancer types. The color bars represent the frequencies according to the clinical utility of actionable genes. C, a Manhattan plot showing the correlations of lncRNA expression and drug $IC_{50}$ across CCLE cell lines. Each dot represents one lncRNA–drug correlation and correlations for different drugs are shown in different colors.

datasets from TCGA and other, independent studies and have made processed lncRNA expression data plus multiple analysis and visualization modules available through TANRIC (http://bioinformatics.mdanderson.org/main/TANRIC:Overview; Fig. 1). This data release, which covers 8,143 samples, has three parts (Table 1). (i) Part one consists of TCGA tissue sample sets: 6,309 tumor samples from 20 cancer types and 564 normal samples from 11 tissues. Other TCGA cancer sets will be added in the coming months. (ii) Part two consists of independent tumor tissue sample sets: one GBM set (274 samples; ref. 15), one KIRC set (97 samples; ref. 14), and one LUAD set (83 samples; ref. 13). Other independent sample sets will be added when available. (iii) Part three consists of tumor cell lines: 739 cell line samples from CCLE (12). To our knowledge, this represents the largest publicly available collection of lncRNA data with parallel multidimensional cancer genomic data.

TANRIC integrates lncRNA expression data with clinical and genomic data (Fig. 1) and provides a user-friendly interface consisting of six modules: Summary, Visualization, Download, My lncRNA, Analyze all lncRNAs and lncRNAs in cell lines (Fig. 2,i). The "Summary" module shows an overview of RNA-seq datasets in TANRIC with a detailed description of each set (e.g., source, read length, sequencing platform, and sequencing strategy; Fig. 2,ii). The "Visualization" module offers an innovative way to examine the global patterns of lncRNA expression in a specific sample set through "next-generation clustered heat maps" (Fig. 2,iii). The interactive heat maps allow users to zoom, navigate, and drill down on clustering patterns (subtypes) of samples or lncRNAs and link to relevant biologic information sources. The "Download" module allows users to obtain the expression data of approximately 13,000 annotated lncRNAs for analysis (Fig. 2, iv; Materials and Methods).

TANRIC provides three analysis modules that enable users to examine the function and underlying mechanisms of lncRNAs in a flexible, interactive way. The "My lncRNA" module provides detailed information about one lncRNA of interest in a user-specified patient sample set. With the module, users can obtain the expression data for any annotated lncRNA (Fig. 2,v) and examine whether the lncRNA shows differential expression between tumor and normal samples or among tumor subgroups

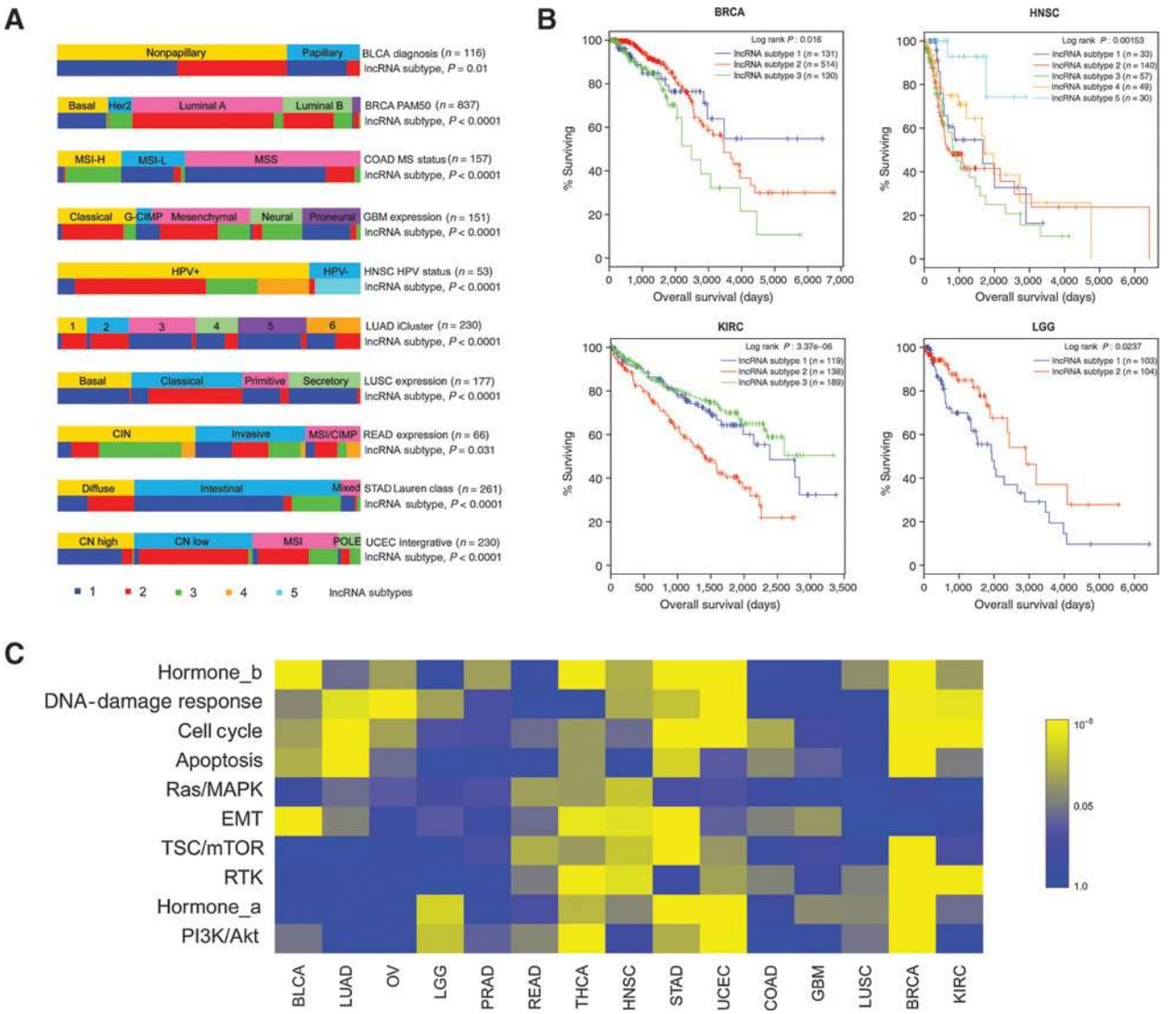


**Figure 5.**
lncRNA expression reveals clinically and biologically relevant tumor subtypes. A, lncRNA-expression subtypes show extensive, strong concordance with established tumor subtypes. B, lncRNA-expression subtypes appear to be correlated with overall patient survival times in BRCA, HNSC, KIRC, and LGG. C, key signaling pathways are differentially expressed among tumor subtypes defined by lncRNA expression. The colors in the heat map represent the statistical significance (FDR) of the associations between lncRNA-expression tumor subtypes and the protein-expression pathway scores.

(as visualized through the box plots, Fig. 2,vi) or whether it correlates with patient survival time (based on *P* values from the univariate Cox proportional hazards model and log-rank test and visualization through a Kaplan–Meier plot, Fig. 2,vi). This module also enables users to examine the correlations of the lncRNA with various molecular data for protein-coding and miRNA genes. The data types include SCNAs, mRNA expression, miRNA expression, protein expression (as visualized through the scatter plots in Fig. 2,vii) and somatic mutations (as visualized through the box plots, Fig. 2,vii). For example, elevated *BCAR4* expression has been shown to significantly correlate with shorter survival time of breast cancer patients (22); and *HOTAIR*, a well-studied lncRNA, is known to be coexpressed with *HOXC* genes (23). Through this module, these findings can be easily confirmed on the basis of TCGA cohorts. Because the annotation of lncRNAs is still rough and incomplete, this module also allows for the query of any user-defined lncRNA or its isoform (based on genomic coordinates) and returns the analysis results. The "Analyze all lncRNAs" module allows users to analyze approximately 13,000 ENCODE-annotated RNAs in a user-specified patient sample set. With this module, users can easily identify the most differentially expressed lncRNAs among tumor subtypes or those with the strongest correlations with patient survival times (Fig. 2,vi). Given a known coding/miRNA gene of interest, this module helps identify those lncRNAs with the strongest associations for various types of molecular data (Fig. 2,vi). The results are presented in a table, and users can search the results by lncRNA name, rank the correlations, and visually examine the details. The "lncRNAs in cell lines" module provides analyses similar to those in "My lncRNA," but in sets based on cell lines. It can help users identify

appropriate cell line models for functional experiments. Through the TANRIC portal, users can perform extensive analyses on lncRNAs, both within and across tumor types and obtain publication-quality figures in a convenient way.

### A large number of lncRNAs with potential biomedical significance across cancer types

Using the data and analysis modules available at TANRIC, we performed a comprehensive survey to assess the potential biomedical significance of lncRNAs. First, for 12 TCGA cancer types with available non-tumor samples, we found large numbers of lncRNAs with significant differential expression between tumor and matched normal samples (paired *t* test; FDR < 0.05; fold change ≥2; Fig. 3A). As an independent validation, 81% of the differentially expressed lncRNAs identified in the TCGA LUAD set were confirmed in a Korean sample set (13). Second, for 11 TCGA cancer types with established biologic or molecular subtypes (Supplementary Table S1), we found considerable numbers of differentially expressed lncRNAs among the known tumor subtypes (*t* test or ANOVA; FDR < 0.05; fold change≥ 2 in at least two groups; Fig. 3B), and those lncRNAs may play a role in defining tumor heterogeneity within a cancer type. Third, for eight TCGA cancer types with sufficient samples available across different disease stages (tumor stages I–IV), we identified some lncRNAs for which the expression patterns correlated with disease stage. Some showed a monotonic change [e.g., 71 and 41 in kidney clear cell cancer (KIRC) and KIRP, ANOVA analysis, FDR < 0.05, fold change ≥2 in at least two groups; Fig. 3C]. Those lncRNAs may be involved in tumor progression. Across the above three analyses, we demonstrate an abundance of lncRNAs with potential biomedical relevance, and many of the lncRNAs show significance in more than one cancer type (Fig. 3D).

To examine the potential impact of lncRNAs on clinical practice, we focused on 123 clinically actionable genes (18). According to their clinical utility, we classified the genes into four groups: (i) therapeutic targets with FDA drugs approved for cancer treatment; (ii) therapeutic targets with drugs in late-stage clinical trials; (iii) therapeutic targets with drugs in early-stage clinical trials; and (iv) other established diagnostic and prognostic biomarkers (Supplementary Table S2). We then examined the correlations between the expressed lncRNAs and the actionable genes, and found considerable numbers of lncRNAs strongly correlated with one or more targets in terms of (i) differential expression between samples with wild-type and mutated genes (*t* test, FDR < 0.05, and fold change ≥2); (ii) in correlation with SCNAs (Spearman rank correlation $|Rs| > 0.6$); and (iii) in correlation with mRNA expression (Spearman rank correlation $|Rs| > 0.6$; Fig. 4A). Focusing on strongly correlated lncRNA-target pairs, we found that many of the pairs are consistently identified in multiple TCGA cancer types (Fig. 4B). These results highlight the potential of lncRNAs as regulators of key therapeutic targets for clinical practice.

To explore the potential effects of lncRNAs on drug sensitivity, we identified the expressed lncRNAs in the CCLE cell lines (12) and examined their correlations with the sensitivity data ($IC_{50}$) of 24 drugs available. Interestingly, we found 202 lncRNA–drug pairs with significant correlations (Spearman rank correlation $|Rs| > 0.3$ and $FDR < 0.01$; Fig. 4C). These results suggest a critical role of some lncRNAs in affecting the response of cancer therapies.

### Biologic and clinical relevance of tumor subtypes revealed by lncRNA expression

Finally, we examined the clinical relevance of tumor subtypes revealed by TCGA lncRNA expression profiles. On the basis of the top 500 lncRNAs with the most variable expression, we defined sample subtypes (sample clusters) by ConsensusClusterPlus (Materials and Methods; ref. 19). For each of the TCGA cancer types we studied, lncRNA-expression subtypes show extensive, strong concordance with established subtypes ($\chi^2$ test; $P < 0.05$; FDR < 0.05; Fig. 5A). For example, lncRNA subtype 1 in breast cancer (BRCA) almost exclusively corresponds to the basal subtype; lncRNA subtype 5 in HNSC primarily corresponds to HPV-negative tumors; and lncRNA subtype 1 in endometrial cancer (UCEC) mainly represents the high copy-number molecular subtype (24). We next assessed the prognostic value of lncRNA-

**Table 2.** Comparison of TANRIC with other available lncRNA-focused bioinformatics resources

| | Data resource | | | Analysis | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Allow user input? | | Clinical data analysis | Molecular data analysis | | |
| Name[a] | Total sample size | Cancer cell line | Non-TCGA large patient cohorts | User-defined lncRNAs | User-defined sample set | Survival, grade, stage, subtype | Genomic data | Proteomic data | lncRNA-expression subtype |
| TANRIC | 8,143 | 739 | √ | √ | √ | √ | mutation SCNA mRNA miRNA | √ | √ |
| MiTranscriptome | 6,938 | ∼200 | √ | O | O | O | O | O | O |
| lncRNAtor | 4,995 | NA | O | O | O | O | mRNA | O | O |
| lncRNABase | ∼6,000 | NA | O | O | O | O | O | O | O |
| ChIPBase | NA | ∼80 | O | O | O | O | O | O | O |
| LNCipedia | NA | NA | O | O | O | O | O | O | O |
| lncRNAdb | NA | NA | O | O | O | O | O | O | O |
| NONCODE | NA | NA | O | O | O | O | O | O | O |
| lncRNome | NA | NA | O | O | O | O | O | O | O |
| NRED | NA | NA | O | O | O | O | O | O | O |
| DIANA-LncBase | NA | NA | O | O | O | O | O | O | O |
| lncRNADisease | NA | NA | O | O | O | O | O | O | O |

Abbreviation: NA, not available.

[a]The resources with a primary focus on cancer lnRNAs are shaded.

expression subtypes. For BRCA, HNSC, KIRC, and brain LGG, the lncRNA-expression subtypes show distinct patient survival profiles (log-rank test, $P < 0.05$, Fig. 5B). As an independent validation, the three lncRNA-expression subtypes in another independent KIRC cohort (14) also show a significant correlation with the overall patient survival times (Supplementary Fig. S1). Furthermore, given clinical variables (i.e., disease stage and tumor grade), the lncRNA subtypes confer additional prognostic power in BRCA and KIRC (multivariate Cox proportional hazards model, $P < 0.05$).

To explore molecular mechanisms associated with the tumor subtypes defined by lncRNA expression, we examined whether some biologic pathways showed some differential expression among the tumor subtypes based on pathway scores calculated from TCGA protein expression data (21). We found that the tumor subtypes defined by lncRNA expression (Fig. 5A) are associated with activation or inhibition of some pathways (Fig. 5C). These results suggest that lncRNA expression represents one meaningful dimension; therefore, integrating lncRNA expression with other molecular data may help characterize the molecular basis of human cancer more fully.

## Discussion

We have developed TANRIC, a user-friendly, interactive, open-access web resource for exploring the functions and mechanisms of lncRNAs in cancer. Compared with other available lncRNA-focused bioinformatics resources (11, 25–34), TANRIC has several unique features (Table 2): (i) It provides extensive, intuitive, and interactive analyses on lncRNAs of interest for their interactions with other TCGA genomic/proteomic/epigenomic and clinical data types, both within a tumor type and across tumor types; (ii) it enables users to query expression profiles of user-defined lncRNAs quickly; (iii) it includes RNA-seq data from well-characterized cell lines and other large, non-TCGA patient cohorts, thereby allowing users to validate a pattern of interest or identify model cell lines for experimental characterization. With the efficient analytic modules, TANRIC substantially lowers the barriers between cancer researchers and complex cancer transcriptomic data (>60TB and 1,142 billion reads in the current release). Going forward, we will constantly incorporate newly available large-scale cancer RNA-seq data into TANRIC.

We have further demonstrated the utility of TANRIC through a comprehensive pan-cancer analysis of expressed lncRNAs. Consistent with previous studies (4, 11, 35, 36), our analysis revealed a large number of tumor-associated lncRNAs. More importantly, we report that some lncRNAs show strong correlations with established therapeutic targets across tumor types or with drug sensitivity across cell lines. Although the correlations do not necessarily indicate direct cause–effect relationships, they highlight the potential of lncRNAs as a novel class of biomarkers or therapeutic targets. TANRIC, thus, represents a starting point for exploration of particular lncRNA species and for generation of testable hypotheses for further experimental investigation.

### Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

### Authors' Contributions

**Conception and design:** J. Li, L. Han, H. Liang
**Development of methodology:** J. Li, L. Han, H. Liang
**Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.):** J. Li, L. Han, L. Liu, H. Liang
**Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis):** J. Li, L. Han, P. Roebuck, L. Diao, L. Liu, Y. Yuan, H. Liang
**Writing, review, and/or revision of the manuscript:** J. Li, L. Han, P. Roebuck, J.N. Weinstein, H. Liang
**Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases):** J. Li, P. Roebuck, L. Liu, J.N. Weinstein
**Study supervision:** H. Liang

### Acknowledgments

The authors gratefully acknowledge contributions from the TCGA Research Network and its TCGA Pan-Cancer Analysis Working Group. The authors thank the MD Anderson high-performance computing core facility for computing resources and LeeAnn Chastain for editorial assistance.

### Grant Support

This study was supported by the NIH (CA143883 to J.N. Weinstein; CA175486 to H. Liang; and the MD Anderson Cancer Center Support Grant P30 CA016672 to J.N. Weinstein and H. Liang); the R. Lee Clark Fellow Award from The Jeanne F. Shelby Scholarship Fund (H. Liang); a grant from the Cancer Prevention and Research Institute of Texas (RP140462 to H. Liang); the Mary K. Chapman Foundation and the Lorraine Dell Program in Bioinformatics for Personalization of Cancer Medicine to J.N. Weinstein.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 27, 2015; revised April 27, 2015; accepted June 11, 2015; published OnlineFirst July 24, 2015.

### References

1. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature 2012;489:101–8.
2. Wapinski O, Chang HY. Long noncoding RNAs and human disease. Trends Cell Biol 2011;21:354–61.
3. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov 2011;1:391–407.
4. Du Z, Fei T, Verhaak RG, Su Z, Zhang Y, Brown M, et al. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat Struct Mol Biol 2013;20:908–13.
5. Yap KL, Li S, Munoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, et al. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell 2010;38:662–74.
6. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 2007;129:1311–23.
7. Wang D, Garcia-Bassets I, Benner C, Li W, Su X, Zhou Y, et al. Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature 2011;474:390–4.
8. Bernard D, Prasanth KV, Tripathi V, Colasse S, Nakamura T, Xuan Z, et al. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. EMBO J 2010;29:3082–93.
9. Tripathi V, Ellis JD, Shen Z, Song DY, Pan Q, Watt AT, et al. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol Cell 2010;39:925–38.
10. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013;45:1113–20.
11. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet 2015;47:199–208.

12. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
13. Seo JS, Ju YS, Lee WC, Shin JY, Lee JK, Bleazard T, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res 2012;22:2109–19.
14. Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, et al. Integrated molecular analysis of clear-cell renal cell carcinoma. Nat Genet 2013;45:860–7.
15. Bao ZS, Chen HM, Yang MY, Zhang CB, Yu K, Ye WL, et al. RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. Genome Res 2014;24:1765–73.
16. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8.
17. Han L, Yuan Y, Zheng S, Yang Y, Li J, Edgerton ME, et al. The Pan-Cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nat Commun 2014;5:3963.
18. Van Allen EM, Wagle N, Stojanov P, Perrin DL, Cibulskis K, Marlow S, et al. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. Nat Med 2014;20:682–8.
19. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.
20. Li J, Lu Y, Akbani R, Ju Z, Roebuck PL, Liu W, et al. TCPA: a resource for cancer functional proteomics data. Nat Methods 2013;10:1046–7.
21. Akbani R, Ng PK, Werner HM, Shahmoradgoli M, Zhang F, Ju Z, et al. A pan-cancer proteomic perspective on The Cancer Genome Atlas. Nat Commun 2014;5:3887.
22. Xing Z, Lin A, Li C, Liang K, Wang S, Liu Y, et al. lncRNA directs cooperative epigenetic regulation downstream of chemokine signals. Cell 2014;159:1110–25.
23. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 2010;464:1071–6.
24. The Cancer Genome Research Network. Integrated genomic characterization of endometrial carcinoma. Nature 2013;497:67–73.
25. Bhartiya D, Pal K, Ghosh S, Kapoor S, Jalali S, Panwar B, et al. lncRNome: a comprehensive knowledgebase of human long noncoding RNAs. Database 2013;2013:bat034.
26. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic Acids Res 2013;41:D983–6.
27. Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS. NRED: a database of long noncoding RNA expression. Nucleic Acids Res 2009;37:D122–6.
28. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res 2014;42:D92–7.
29. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Reczko M, Maragkakis M, Dalamagas TM, et al. DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. Nucleic Acids Res 2013;41:D239–45.
30. Park C, Yu N, Choi I, Kim W, Lee S. lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. Bioinformatics 2014;30:2480–5.
31. Quek XC, Thomson DW, Maag JL, Bartonicek N, Signal B, Clark MB, et al. lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res 2015;43:D168–73.
32. Volders PJ, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, et al. An update on LNCipedia: a database for annotated human lncRNA sequences. Nucleic Acids Res 2015;43:D174–80.
33. Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, et al. NONCODEv4: exploring the world of long non-coding RNA genes. Nucleic Acids Res 2014;42:D98–103.
34. Yang JH, Li JH, Jiang S, Zhou H, Qu LH. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res 2013;41:D177–87.
35. Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011;29:742–9.
36. White NM, Cabanski CR, Silva-Fisher JM, Dang HX, Govindan R, Maher CA. Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. Genome Biol 2014;15:429.