EBioMedicine 7 (2016) 62–72

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

Pan-Cancer Analyses Reveal Long Intergenic Non-Coding RNAs Relevant to Tumor Diagnosis, Subtyping and Prognosis

Travers Ching [a,b], Karolina Peplowska [c], Sijia Huang [a,b], Xun Zhu [a,b], Yi Shen [b], Janos Molnar [c], Herbert Yu [b], Maarit Tiirikainen [c], Ben Fogelgren [d], Rong Fan [e], Lana X. Garmire [a,b,*]

[a] Molecular Biosciences and Bioengineering Graduate Program, University of Hawaii at Manoa, Honolulu, HI 96822, USA
[b] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA
[c] Genomics Shared Resource, University of Hawaii Cancer Center, Honolulu, HI, 96813, USA
[d] Department of Anatomy, Biochemistry and Physiology, John A. Burns School of Medicine, University of Hawaii, Honolulu, HI 96813, USA
[e] Department of Biomedical Engineering, Yale University, New Haven, CT 06520, USA

ARTICLE INFO

Article history:
Received 26 August 2015
Received in revised form 2 March 2016
Accepted 16 March 2016
Available online 19 March 2016

Keywords:
LincRNA
lncRNA
pan-cancer
RNASeq
biomarkers

ABSTRACT

Long intergenic noncoding RNAs (lincRNAs) are a relatively new class of non-coding RNAs that have the potential as cancer biomarkers. To seek a panel of lincRNAs as pan-cancer biomarkers, we have analyzed transcriptomes from over 3300 cancer samples with clinical information. Compared to mRNA, lincRNAs exhibit significantly higher tissue specificities that are then diminished in cancer tissues. Moreover, lincRNA clustering results accurately classify tumor subtypes. Using RNA-Seq data from thousands of paired tumor and adjacent normal samples in The Cancer Genome Atlas (TCGA), we identify six lincRNAs as potential pan-cancer diagnostic biomarkers (PCAN-1 to PCAN-6). These lincRNAs are robustly validated using cancer samples from four independent RNA-Seq data sets, and are verified by qPCR in both primary breast cancers and MCF-7 cell line. Interestingly, the expression levels of these six lincRNAs are also associated with prognosis in various cancers. We further experimentally explored the growth and migration dependence of breast and colon cancer cell lines on two of the identified lncRNAs. In summary, our study highlights the emerging role of lincRNAs as potentially powerful and biologically functional pan-cancer biomarkers and represents a significant leap forward in understanding the biological and clinical functions of lincRNAs in cancers.

1. Introduction

Advancement of high-throughput technologies such as RNA-Seq has recently allowed for the identification of tens of thousands of new lincRNAs in different tissues (Cabili et al., 2011; Ching et al., 2014; Garmire et al., 2011; Trapnell et al., 2010). The Encyclopedia of DNA Elements (ENCODE) project found that about 62% of the entire genome is transcribed to long (>200 base pairs) RNA sequences (Consortium, 2012). Given that 3% of the genome encodes protein-coding exons, the large majority of these transcripts are non-coding RNAs (lncRNAs). Among these lncRNAs, about one third come from intergenic regions (lincRNAs) (Consortium, 2012). Unlike small non-coding RNAs which may regulate target gene expression through simpler complementary recognition (Menor et al., 2014), the mechanisms of lincRNAs are complex and may depend on formation of RNA-protein complexes (Mchugh et al., 2014). Attempts have been made to extrapolate the functions of lincRNAs based on model lincRNAs, such as studies that predict lincRNAs binding to PRC2 or competing endogenous lincRNAs (microRNA "sponges") (Khalil et al., 2009; Liao et al., 2011; Liu et al., 2013; Salmena et al., 2011; Yuan et al., 2014). However, lincRNAs remain one of the most mysterious and least understood species of non-coding RNAs (Ching et al., 2014).

Regardless of the regulatory mechanisms, lincRNAs are becoming a relatively new class of cancer biomarker candidates. Several lincRNAs and overlapping lncRNAs have been relatively well-studied and indicated as potential biomarkers associated with tumor initiation, progression or prognosis, such as MALAT1(Ji et al., 2003; Tripathi et al., 2010; Ulitsky and Bartel, 2013), HOTAIR(Gupta et al., 2010; Rinn et al., 2007; Ulitsky and Bartel, 2013), XIST(Brockdorff et al., 1991; Penny et al., 1996; Weakley et al., 2011), PCAT1(Ge et al., 2013; Prensner et al., 2011; Ulitsky and Bartel, 2013) and CCAT2(Ling et al., 2013). However, most of the studies detect lincRNAs as candidate biomarkers of a specific cancer type. The pan-cancer biomarker-based design of clinical trials, on the other hand, can increase statistical power and greatly decreasing the size, expense, and duration of clinical trials (Cancer Genome Atlas Research et al., 2013). Towards this, we here propose a pan-cancer based lincRNA diagnostics biomarker study, which is aligned with the goal of The Cancer Genome Atlas (TCGA) analysis project that enables

* Corresponding author at: Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA.
E-mail address: lgarmire@cc.hawaii.edu (L.X. Garmire).

http://dx.doi.org/10.1016/j.ebiom.2016.03.023

the discovery of novel adaptive, biomarker-based strategies to be practiced across boundaries of different tumor types(Cancer Genome Atlas Research et al., 2013).

In this study, we have taken full advantage of the rich RNA-Seq data from the TCGA consortium, as well as thousands of RNA-Seq and microarray data from Gene Expression Omnibus (GEO) and our own collection of breast cancer samples. By combining data-mining and machine-learning methods with biological function validation experiments, we have highlighted lincRNAs as a new paradigm for actionable diagnostics in the pan-cancer setting. In addition, we have portrayed the comprehensive landscape of lincRNAs and their relationship to other omics data in pan-cancers. We found that the lincRNAs are more tissue-specific compared to protein-coding mRNAs, and they also convey complementary relevance to clinical information, including tumor molecular subtypes. Moreover, we have detected and thoroughly validated 6 lincRNAs as potential pan-cancer diagnostic biomarkers in over 3300 tissue samples. Lastly, we confirmed that the lincRNAs are biologically functional, by measuring the reduction of cell proliferation and migration in breast cancer cell lines with siRNA knockdown on two of the homologous lincRNAs.

2. Materials and Methods

2.1. RNA-Seq Datasets

2.1.1. TCGA Datasets

We used 12 cancer datasets from TCGA incorporating RNA-Seq data files from 1240 tissue samples (Supplementary Table I). RNA-Seq datasets were chosen from cancers in TCGA that have at least 25 pairs of primary tumor and paired adjacent normal tissue samples. These datasets include breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HSNC), kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD) and thyroid carcinoma (THCA). RNA-Seq BAM files were downloaded from UCSC Cancer Genomics Hub (https://cghub.ucsc.edu/) using the GeneTorrent program (Wilks et al., 2014). The TCGA alignment protocol used the Mapsplice alignment program (Wang et al., 2010) to align raw reads to the human genome, where loci with the same alignment score has equal probability to assign a read. Technical replicates were combined by merging the results from the BAM files. RefSeq genes and lincRNAs were quantified using featureCounts (Liao et al., 2013, 2014) from the Subread package (version 1.4.5-p1). RefSeq annotation was obtained from Illumina hg19 iGenomes and lincRNAs were obtained from Broad Institute Human Body Map project, so that we can directly compare the tissue specificity results between TCGA samples and those in Cabili et al.(Cabili et al., 2011). All alignments were conducted on the New Hampshire INBRE (IDeA Network of Biomedical Research Excellence) grid computing system. Batch effect was corrected, and DESeq2 (Love et al., 2013, 2014) (version 1.6.1) was used for calculating normalized count data and fragments per kilo bases of exons for per million mapped reads (FPKM) data. A combination of independent RNA-Seq and microarray datasets were used for verification, and the summary of the datasets is listed in Supplementary Table I.

2.1.2. GEO Datasets

A large-scale search of GEO RNA-Seq database was performed to find additional datasets for verification. Datasets with tumor and normal samples with good read quality (read mapping rate and low duplication rates) were selected. These included GSE25599 (liver cancer), GSE58135 (breast cancer) and GSE50760 (colon cancer). In addition, normal breast tissue samples were taken from GSE52194, GSE45326 and GSE30611 for comparison with our cancer samples. GEO datasets were aligned to the UCSC hg19 genome using Tophat2 with default parameters for either single-end or paired-end protocols. LincRNA count quantification and FPKM data were generated as above. Microarray datasets from GEO with tumor and normal samples were selected based on platforms that had probes mapping to the six lincRNAs of interest.

2.1.3. Our Own Dataset

Our primary breast cancer samples were extracted with RNeasy Mini Kit (Qiagen), followed by quality control with RNA 6000 chips (Agilent Bioanalyzer). RNA species with RIN values >7 were sent to the Genomics Core of Yale Stem Cell Centre. Ribo-depleted RNA-Seq was conducted with 100 bp read length. The read count quantification and FPKM data were generated as above. The RNA-Seq reads of our samples will be deposited to GEO upon publishing of this manuscript.

2.2. Tissue Specificity

To analyze tissue specificity, Jensen-Shannon divergence score (JS score) was calculated from tumor and normal samples of each tissue, and the two distributions of JS scores were compared following the method of Cabili et al. (Cabili et al., 2011). Briefly, FPKM values were first calculated from the normalized count data from each sample. Then the mean FPKM for each tissue type was calculated and log transformed. The vector $\boldsymbol{e}$ that represents the distribution of expression is given by:

$$\boldsymbol{e}=\frac{log_2(FPKM+1)}{\sum_{i=1}^{n} log_2(FPKM_i+1)}$$

The JS_t score is the JS score for each tissue type t, calculated by the following:

$$JS_t(\boldsymbol{e},\boldsymbol{e^t}) = 1-\sqrt{H(\boldsymbol{e}+\boldsymbol{e^t})-\frac{H(e)+H(\boldsymbol{e^t})}{2}}$$

Where H is the Shannon entropy and $\boldsymbol{e^t}$ is the hypothetical distribution when a lincRNA is expressed in only one tissue type:

$$\boldsymbol{e^t} = \left(e^1,\ldots,e^i,\ldots,e^n\right), where\, e^i = \begin{cases} 1\, if\, i=t \\ 0\, if\, i\neq t \end{cases}$$

The JS score for a lincRNA is then defined as the maximum JS_t score across all tissue types.

2.3. Differential Expression

Each of the 12 TCGA cancer datasets was tested for differential expression (DE) using DESeq2 (Love et al., 2013, 2014). Statistically significant genes were selected with a FDR adjusted p-value threshold of 0.05 after Benjamini & Hochberg multiple hypothesis correction. As a result, six lincRNAs were discovered to be consistently upregulated or down-regulated in all twelve TCGA cancer datasets. These six lincRNAs were used subsequently for survival and pathway analysis.

2.4. Survival Analysis

These six lincRNAs with pan-cancer diagnostic potential were examined for their association with patient survival among four types of TCGA cancer types. Note that these lincRNAs were initially selected as diagnostic biomarkers, but not prognostic biomarkers. The survival data from the four types TCGA cancers were obtained in two approaches. LUAD, LUSC and OV have relapse free survival information directly available from the TCGA data repository. The fourth cancer type BRCA has overall survival data available, per the courtesy of Volinia and Croce (2013)). Patients who did not have an event (death or tumor relapse, depending on the data set) during the study were

considered as censored. The expression values of the six lincRNAs were used as predictors to fit a Cox-Proportional Hazards (Cox-PH) regression model, where the overall survival or disease free survival was the response variable. For each patient, a prognosis index (PI) score was generated from the Cox-PH model. The median PI score among all patients of the same cancer type was used as the threshold to dichotomize the patients into high vs. low risk groups, similar to others (Huang et al., 2014). The log-rank p-value was then calculated to assess the statistically significant difference between the Kaplan–Meier curves of the high vs. low risk groups.

2.5. Tumor Subtype Classification and Concordance Between Data Types Using Nmf

Non-negative matrix factorization (NMF) method was used to classify tumor subtypes with lincRNA expression values. The optimal number of clusters was selected using the maximum cophenetic correlation. The lincRNA clustering results were then compared to those of other data types, using the method similar to Han et al. (2014)). The other data types from the TCGA include mRNA-Seq, mature microRNA-Seq, methylation and reverse phase protein array (RPPA) for each cancer type (Liao et al., 2013), all obtained from the Broad institute Genomic Data Analysis Center (GDAC). The concordances from the chi-square tests between lincRNA and other data types were used to assess the correlations between clustering.

Additionally, lincRNA clustering was compared with another standard method, the PAM50 clustering (Cancer Genome Atlas, 2012), using the TCGA breast cancer samples. The correlation between these two clustering approaches was calculated using the concordance as mentioned above. Similarly, cluster correlation was computed for subtypes based on ER+/− information from the GSE58135 breast cancer dataset.

2.6. Lincrna Sequence Coding Potential and Homology Characterization

To predict the coding potential of the sequences, iSeeRNA (Sun et al., 2013) and Coding-Potential Assessment Tool (CPAT) (Wang et al., 2013) were used. The two programs are trained on long non-coding RNAs to assess the coding potential of transcripts. For iSeeRNA, the coordinates of lincRNA transcripts and exons were used as inputs in the form of GFF files. For CPAT, lincRNA sequences were used as inputs in the form of fasta files. To test for homology between transcripts, NCBI's command line BLAST+ suite (Camacho et al., 2009) was used. Pairwise BLAST was performed on all isoforms of the six differentially expressed lincRNAs. We calculated the percentages of homology by the number of matching base pairs divided by the total number of base pairs in the query sequence. Due to the high homology between three of the discovered lincRNAs (PCAN-2, PCAN-3 and PCAN-5), downloaded RNA-Seq reads may have slight ambiguity in counting these lincRNA expression, since they were generated by TCGA using the Mapsplice alignment program (Wang et al., 2010).

2.7. Quantitative RT-PCR (qRT-PCR) Analysis

Total RNA from MDA-MB-231 and MCF-7 cell lines was isolated using RNeasy Mini Kit (Qiagen). Pooled total RNA from five healthy normal breast cancer patients was ordered from Biochain (Total RNA - Human Adult Normal Tissue 5 Donor Pool: Breast, catalog# R1234086-P). To match these healthy controls, total RNA was isolated from five in-house breast cancer patient samples.

High Capacity cDNA Reverse Transcription kit (Life Technologies, Thermo Scientific) was used for random-primed first-strand complementary DNA synthesis. Real time quantitative PCR (qPCR) was performed with SYBR Green (Life Technologies) with primers against selected linc RNAs (primer sequences are listed in Supplementary Table VI). Amplification and real time measurement of PCR products was performed with 7900HT Fast Real-Time PCR System (Life Technologies). The comparative Ct method (Livak and Schmittgen, 2001) was used to quantify the expression levels of lincRNAs. Beta-glucuronidase (GUS) gene expression served as the internal control. GUS was selected as the internal control, as its expression level has been found to be comparable in range to the expression of linc RNAs and is stable in a wide variety of cancers (Habel et al., 2006; Rubie et al., 2005).

2.8. RNA Interference

The siRNA oligos were synthesized by GE Dharmacon. The target sequences are as follows: control siRNA: 5′-UGGUUUACAUGUCGACUAA-3′, 5′-UGGUUUACAUGUUGUGUGA-3′, 5′-UGGUUUACAUGUUUUCUGA-3′, 5′-UGGUUUACAUGUUUUCCUA-3′; lincRNA siRNA #1: 5′-UUCCUUUAGACCCAUUCUCUU-3′; lincRNA siRNA #2: 5′-GAACCCACCACUGCUUCUC-3′. This lincRNA siRNA targets PCAN-2 and PCAN-3 lincRNAs. Cells were transfected in a 6-well plate format with siRNA oligos at 40 nM (for cell proliferation assays) or 60 nM (for migration assays) concentration, using DharmaFECT 1 Transfection Reagent (Dharmacon). The knockdown efficiency was determined by qRT-PCR 24 h post transfection.

2.9. Cell Growth and Migration Assays

Cell proliferation analysis was done using CellTiter-Glo Luminescent Cell Viability Assay Kit (Promega). Briefly, MDA-MB-231 cells were transfected in biological triplicates with siRNA constructs (control siRNA and linc RNA siRNA). After 24 h, 400 cells of each condition were seeded in triplicates into 96-well plates and allowed to grow for another 48 h. Cells number estimation at different time points was based on the quantification of the present ATP using SpectraMax Gemini XPS microplate reader (Molecular Devices). Cell migration was analysed using well established wound-healing assay (Liang et al., 2007). Scratches in cell monolayer were made 30 h post siRNA transfection (3 scratches in each of the 3 biological replicates). Cell migration was analysed by time-lapse microscopy using IX81 Olympus microscope, with 10× objective (for MDA-MB-231 cells) and 4× objective with additional 1.6× magnification (for MCF-7 cells). Images were taken every 5 min over time period of 24 h. Migration rates and cell tracking were analysed using the Metamorph software.

3. Results

3.1. Overview of the Workflow

To detect genes differentially expressed between healthy and tumor tissues, we employed a two-factor (cancer/normal, and source of samples) experimental design in which patients with tumor samples and matched normal sample were selected. This approach allowed sufficient statistical power by reducing the variation of data (Ching et al., 2014). In total, we downloaded 1240 paired cancer and adjacent normal RNA-Seq samples in 12 different cancer types.

The 12 different cancer types include breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), 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), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD) and thyroid carcinoma (THCA). Details on the number of samples in each cancer type, sequencing strategies, total mappable reads, and detected lincRNAs are listed in Supplementary Table I. For lincRNA genomic coordinates, we used the UCSC genome browser's “lincRNA transcript track”, which is based on both the Broad Institute Human Body Map including the annotations of transcripts of uncertain coding potential (TUCP) (Cabili et al., 2011). We quantified lincRNA expression with normalized FPKM values.

Computationally, we have performed various analyses to study the biological and clinical relevance of lincRNAs to pan-cancer, including differential expression (DE), tissue specificity and molecular subtype analyses, as well as construction and verification of the diagnostic and survival models (Supplementary Fig. 1). Experimentally, we have verified the gene expression differences of a panel of 6 lincRNAs, which have pan-cancer diagnostic biomarker potential. Most importantly, we demonstrated the phenotypic changes of two of the over-expressed lincRNAs by siRNA knockdown experiments in two breast cancer cell lines MCF-7 and MDA-MB-231.

3.2. The High Tissue Specificities of Lincrnas are Diminished in Cancers

To investigate the expression patterns of the lincRNA transcripts among different tissue types, we conducted principal component analysis (PCA) for lincRNA expression on adjacent normal and cancer samples separately from 12 TCGA datasets (Fig. 1). As expected, the normal samples are clearly clustered by tissue type based on lincRNA expression (Fig. 1a). However, the cancer samples become less separable by tissue type (Fig. 1b). The less precise distinction of cancer samples in the PCA plot reflects a degree of de-differentiation of tumor cells. The possibility of confounding due to heterogeneity of tumors of the same type can be excluded, since the latter would lead to more spreading, rather than less spreading observed on the PCA plot. We therefore reason it as the loss of tissue specificity in cancers. Supporting this observation, the first three principal components of PCA account for less variance in cancer samples compared to those in the adjacent normal tissues, suggesting deregulation of lincRNAs in cancers (Fig. 1). We replicated the same analysis for protein-coding genes between tumor and adjacent normal tissues, and found the same trend of losing tissue specificity in the tumor samples (Supplementary Fig. 2).

To further analyze the tissue specificity of lincRNAs, we calculated the tissue specificity scores (JS scores) as defined in Cabili et al.[1], where a higher JS score indicates more tissue specificity. We compared the distributions of these JS scores in tumor and adjacent normal tissue, for both lincRNAs and RefSeq protein coding genes (Fig. 2). Consistent with the PCA plots, lincRNAs in cancer tissues are significantly less tissue specific than those in adjacent normal tissues (t-test, $p < 2.2e\text{-}16$) (Fig. 2a, c and d). Moreover, in comparison with RefSeq protein coding genes (Fig. 2b, e and f), lincRNAs have a much higher average JS score (t-test, $p < 2.2e\text{-}16$). Subsequently, we defined a subset of lincRNAs that are highly tissue specific with JS score greater than 0.75 and are expressed in at least 5% (32 out of 640) of the total normal samples (Supplementary Table II). To confirm that the tissue-specific lincRNAs defined by TCGA pan-cancer analysis are accurate, we then compared the tissue type assigned to lincRNAs by Cabili et al.[1] to the tissue types assigned to the same lincRNAs based on the TCGA data. We observed statistically significant correlations (χ^2-test, all $p < 0.0001$) between the two studies in all tissue categories (Supplementary Fig. 3). In addition, we plot the tissue specific JS score for each tissue type (JS_t score) and plotted their distributions (Supplementary Fig. 4). As expected, significant amounts of lincRNA have zero JS scores, as many lincRNAs are not expressed in certain tissues.

3.3. LincRNA Clustering Accurately Predicts Molecular Subtypes of Tumors

Given the tissue specificity of lincRNAs, we hypothesized that lincRNAs can accurately separate tumors by molecular subtype. To identify a representative cancer type, we first used consensus non-negative matrix factorization (CNMF) to cluster the patient samples from each of the 12 types of cancer. We then calculated the correlations between the clustering result based on lincRNAs and those based on four other high-throughput data types: mRNA expression, micro-RNA expression, DNA methylation and reverse phase protein array (RPPA) obtained from the Broad Institute Genomic Data Analysis Center (GDAC) (BROAD, 2014). The majority of lincRNA and GDAC clustering results are statistically significantly correlated (Fig. 3a). As expected, lincRNA and mRNA expression are the most highly correlated among all four high-throughput data types. Among the 12 cancer datasets, the BRCA dataset has the best agreements between lincRNAs and the other data types. We therefore focused on the correlation between lincRNA and molecular subtypes in breast invasive carcinoma.

We first applied CNMF to the TCGA BRCA dataset and used cophenetic correlation (Liao et al., 2013) to determine the optimal cluster number to be 5, the same number of clusters as in PAM50 based classification. We then compared the result of CNMF clustering to PAM50 based subtypes, which include basal-like, HER2-enriched, luminal A, luminal B and normal-like subtypes (Cancer Genome Atlas, 2012) (Fig. 3c). The concordance score based on the χ^2-test is highly significant ($p < 2.2e\text{-}16$), and the overall accuracy to clinical types is 71.6%, as measured by rand measure, a metric for the percentage of agreement on a pair of samples belonging to the same group. Interestingly, the first

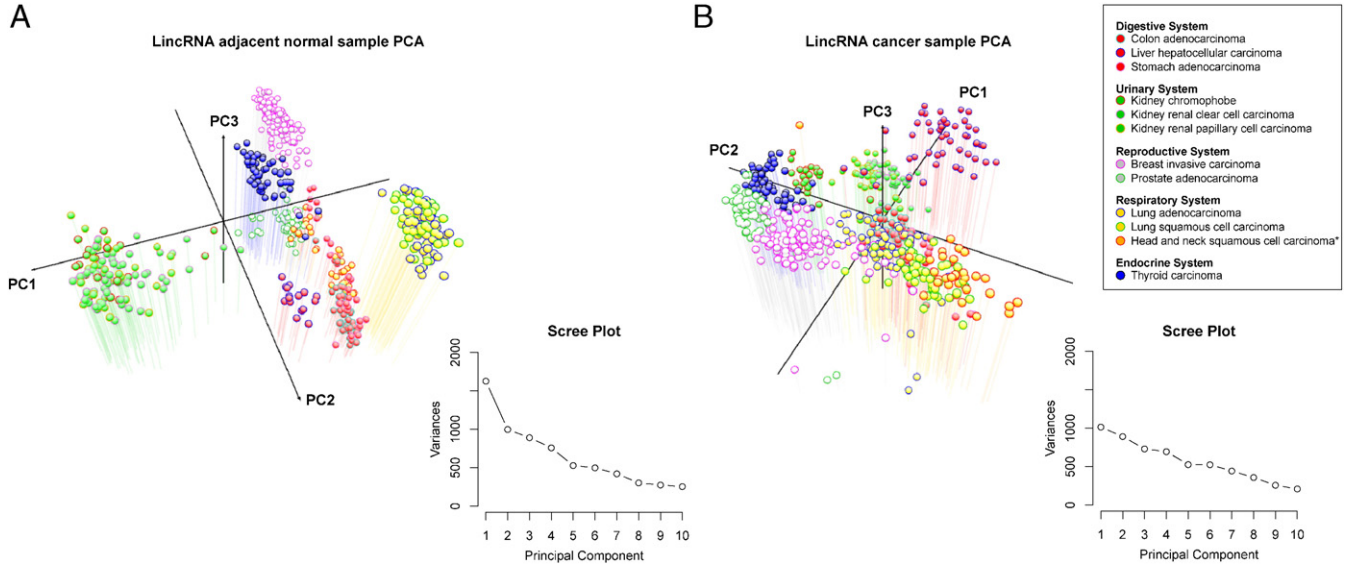

Fig. 1. Principal component analysis of lincRNA expression in 12 TCGA datasets. The first three principal components (PCs) were plotted using the log FPKM values of lincRNA expression in (a) normal adjacent tissue and (b) cancer samples. The variances associated with each of the first 10 principal components are plotted alongside each graph (Scree Plot).

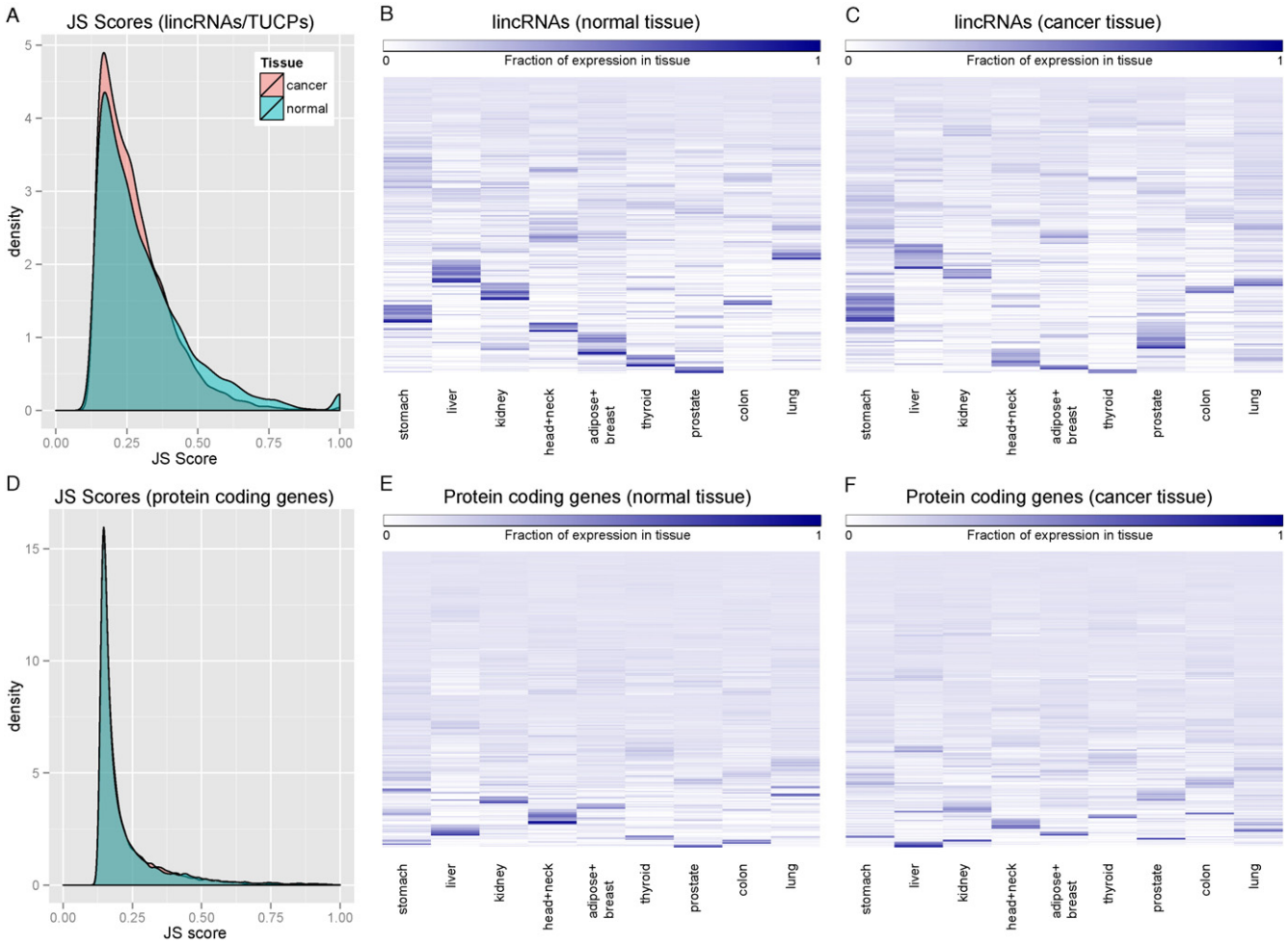

Fig. 2. Tissue specificities of lincRNAs and protein coding genes. (a-b) Maximum JS scores for were used to measure tissue specificity in primary tumors and adjacent normal samples, based on either lincRNAs (a) or protein coding genes (b). A value of 1 indicates that the lincRNA is expressed in only one tissue. (c-f) Fractional expression of lincRNAs or protein coding genes in each tissue was plotted in the adjacent normal or cancer samples.

CNMF cluster has the strongest correlation with the basal-like subtype among all molecular subtypes, with an accuracy of 95% based on rand measure. Additionally, we examined the GSE58135 breast cancer dataset that has primary tumor samples in ER+/HER2- and triple negative subtypes (Fig. 3b). The unsupervised CNMF clustering on these cancer samples yields highly accurate separation between ER+/HER2- and triple negative samples (χ^2-test p < 2.2e-16, and rand measure 84.5%). These results show that lincRNAs are well correlated with the molecular subtypes of tumors.

3.4. Transcriptome Analysis Reveals a Pan-Cancer Panel of Six LincRNAs

To seek a panel of lincRNAs as pan-cancer diagnostic biomarkers, we performed differential expression analysis on the above 12 TCGA datasets and detected thousands of differentially expressed lincRNAs in each TCGA dataset (Supplementary Fig. 5). Among them, six lincRNAs are consistently and significantly altered in all 12 cancers, with five of them being up-regulated and one down-regulated (Fig. 4a, Supplementary Fig. 6 and Supplementary Table III). In contrast, when we applied the same selection criteria to protein coding genes, we identified 47 mRNAs. The much larger number of mRNAs is presumably due to the less tissue specificity of mRNAs and more annotated mRNAs compared to lincRNAs at the time of investigation.

Several other lincRNAs, such as PCAT1, MALAT1, HOTAIR, have previously been reported to associate with a variety of cancers (Ge et al., 2013; Ji et al., 2003; Prensner et al., 2011). We re-analyzed their expression in our pan-cancer data set (Supplementary Fig. 7). These three lincRNAs are not pan-cancer lincRNAs, but the TCGA results confirmed the previous findings based on several cancer types. PCAT1 was discovered in prostate cancer (Prensner et al., 2011), and is indeed extremely significant in the TCGA PRAD data. MALAT1 is known to be primarily associated with liver cancer, lung cancer and kidney cancer (Ji et al., 2003), and it is recapitulated in the TCGA data. HOTAIR is also known to be highly upregulated in many different TCGA cancer types.

To confirm that the six lincRNAs are indeed associated with pan-cancers, we processed additional 833 samples from a wide range of resources including three public RNA-Seq datasets and eleven microarray datasets (Supplementary Table I). All three public RNA-Seq datasets (GSE58135 breast cancer, GSE50760 colon cancer, and GSE25599 liver cancer) show consistent directions of fold change for all six lincRNAs (Fig. 4b). Although the microarray platforms are not designed to detect lincRNAs, some probes are nevertheless overlapped with non-coding RNAs as shown by others (Du et al., 2013), and thus they can be another source of empirical verification. Among the various microarray platforms examined, 24 of the 29 microarray probe sets have the same overall directions of fold changes as those in the RNA-Seq datasets (Supplementary Fig. 8). Moreover, the expression levels of the six lincRNAs in 28 breast cancer cell lines from the GSE58135 dataset and 5 breast cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) are all comparable with

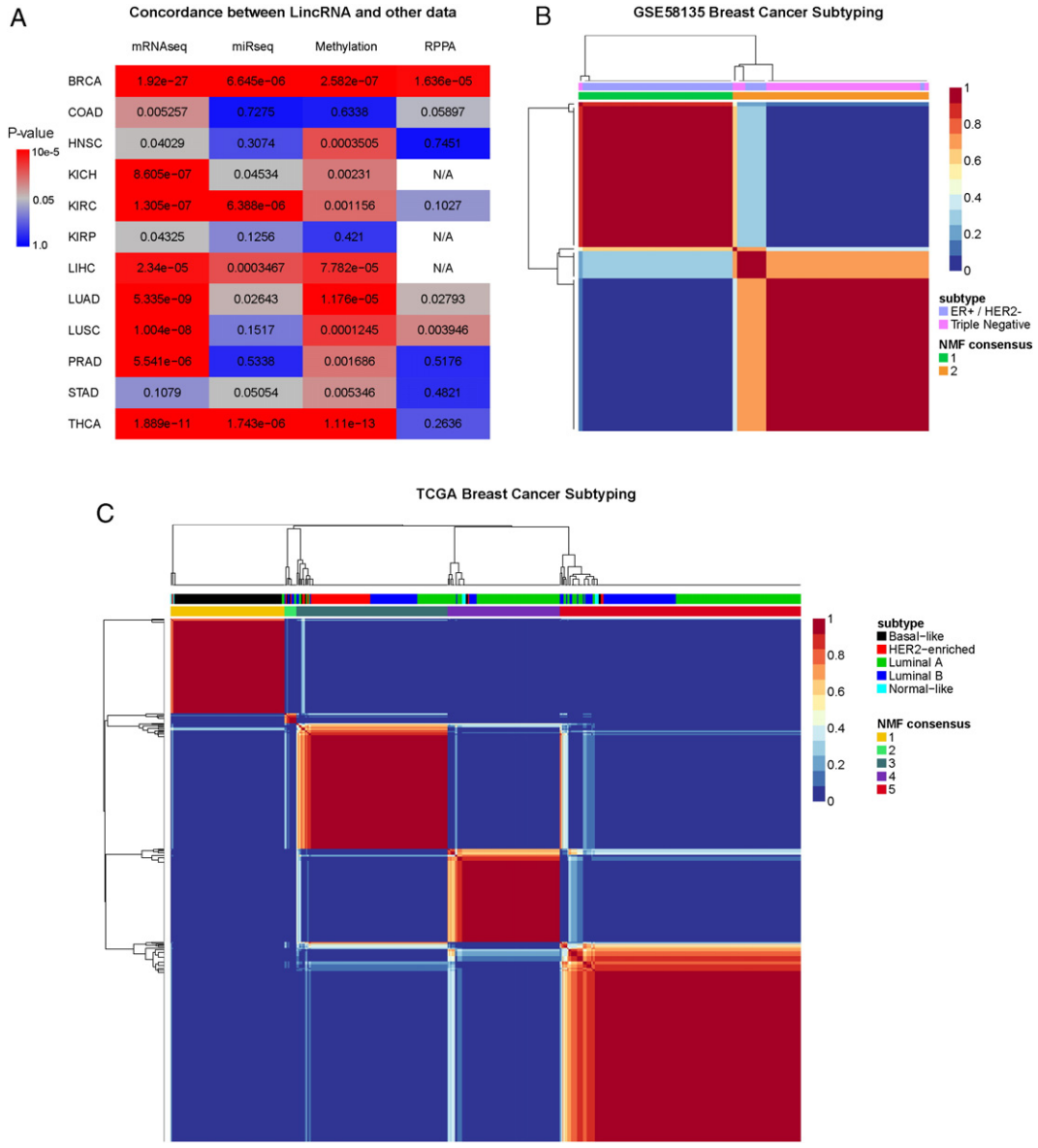

Fig. 3. Correlation of lincRNAs with other data types and cancer subtypes. (a) The concordance between clustering results of lincRNAs and other high throughput data types in TCGA based on χ-square statistical test. (b) CNMF was used to determine the clustering of lincRNAs in the GSE58135 Breast Cancer dataset. The concordance of the clustering with the tumor subtypes in the dataset is significant (chi-square, $p < 2.2e\text{-}16$). (c) CNMF was used to determine the clustering of lincRNA in the TCGA BRCA dataset. The concordance of the CNMF clustering with the tumor subtypes in the dataset is significant (chi-square, $p < 2.2e\text{-}16$).

those from the TCGA BRCA samples (Supplementary Fig. 9), further supporting the robustness of these lincRNAs as potential pan-cancer biomarkers.

To verify this lincRNA panel experimentally, we performed additional RNA-Seq and qPCR experiments on our own breast cancer samples. First, we sequenced fresh frozen primary tumor samples from 10 individual patients using the ribosomal depletion RNA-Seq method. We then compared them to normal breast tissue RNA-Seq data from GEO (GSE52194, GSE45326 and GSE30611). All six lincRNAs have the same trends of changes as in the other GEO RNA-Seq datasets (Fig. 4c) and five of them are significantly differentially expressed. We followed up with the qPCR validation and designed seven PCR primer pairs for selected transcripts in the lincRNA panel (supplementary Table IV). The qPCR results in pooled breast tumor samples ($n = 5$), pooled normal breast samples ($n = 5$) and MCF-7 cell lines are shown in Fig. 4d. In all cases, the expression levels show statistically significant differential expression in the same directions as the RNA-Seq data, both between primary tumor and normal sample pools and between normal and MCF-7 cancer cell lines.

3.5. Sequence Features among the Six-Lincrna Biomarkers

To confirm the non-coding nature of the lincRNA transcripts, we used the iSeeRNA (Sun et al., 2013) and Coding-Potential Assessment Tool (CPAT) (Wang et al., 2013). Both programs are specifically trained on long non-coding RNAs to assess the non-coding potential of RNA transcripts. Out of the 52 isoforms from the lincRNA panel, iSeeRNA predicted 49 to be non-coding. For the three transcripts that are ambiguous, we used a second tool, CPAT, to obtain further evidence for the coding or non-coding nature of these transcripts. CPAT classifies all three of them as non-coding RNAs. In contrast, both CPAT and iSeeRNA correctly classified all isoforms of house-keeping genes GUS and GAPDH as protein coding. Overall, both programs provide strong evidence for the non-coding nature of the six lincRNAs (Supplementary Table V).

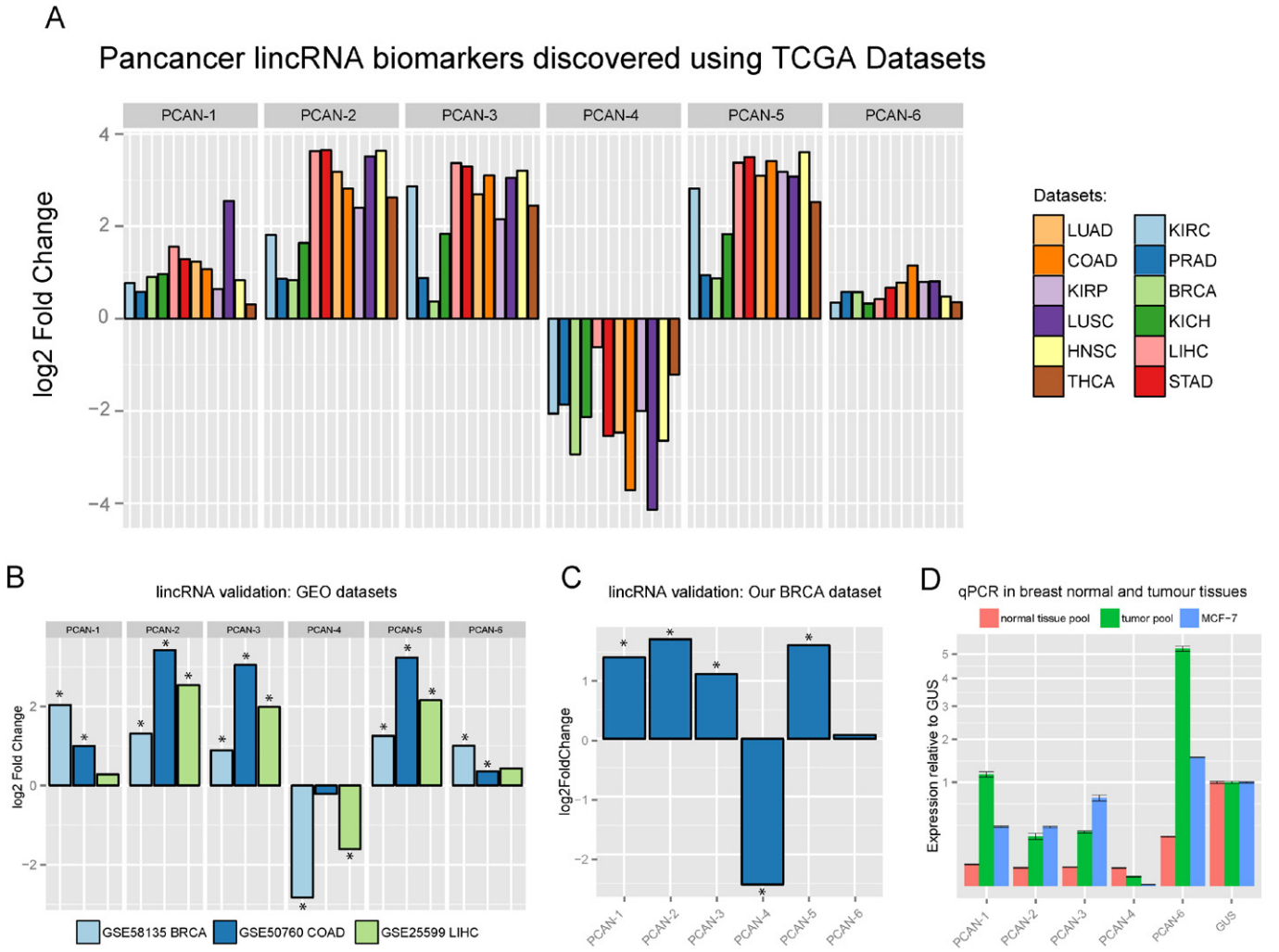

Fig. 4. Differentially expressed pan-cancer lincRNAs. (a) Six lincRNAs are consistently differentially expressed in 12 TCGA datasets. Each of the six lincRNAs shown is either significantly upregulated or significantly downregulated across the various cancers. The six lincRNAs in three independent RNA-Seq datasets from GEO (b), our own breast cancer dataset (c) and qPCR of pooled 5 normal tissues, pooled 5 tumors and the MCF-7 cell line (d).

To examine the relationship between the six lincRNAs, we first checked the correlations of their expression values in all TCGA samples. Three of the lincRNAs, PCAN-2, PCAN-3 and PCAN-5, are highly correlated with spearman correlation coefficients of approximately 0.92 between them (Supplementary Fig. 10). The high correlations among expression prompted us to check if sequence similarities exist. Thus, we tested the pairwise homology among all transcripts of the six lincRNAs, using NCBI's BLAST+ suite (Camacho et al., 2009) (Supplementary Fig. 11). Indeed, the three lincRNAs mentioned above are highly homologous, and some of the annotated transcripts are 99% identical. Two of the lincRNAs, PCAN-2 and PCAN-3, are in the tandem locations on chromosome 14 and the third lincRNA PCAN-5 is located on chromosome 22, suggesting potential gene duplication events from a common origin.

3.6. The LincRNA Biomarker Panel Robustly and Accurately Predicts Pan Cancers

To quantitatively assess the value of the six lincRNAs as pan-cancer diagnostic biomarkers, we built a classification model upon them (Fig. 5a). First, we split the TCGA pan-cancer data into 80% training and 20% holdout testing sets. Given that some lincRNAs are highly correlated (Supplementary Fig. 10) and thus potentially redundant as biomarker predictors, we used correlation feature selection (CFS) method to select the most relevant and least redundant subset of lincRNAs among them. As a result, five of the lincRNAs were chosen: PCAN-1, PCAN-2, PCAN-3, PCAN-4, and PCAN-6.

We then compared the classification results on the training dataset using four widely used machine-learning algorithms: Random Forest (RF), Linear Support Vector Machines (LSVM), Gaussian Support Vector Machines (GSVM) and Logistic Regression with L2 regularization (L2-LR). As shown by the receiver operator characteristics (ROC) curves on the TCGA training data set, RF has the best AUC of 0.947 (95% confidence interval, or CI: 0.9343–0.9603) on the training data among the four methods (Supplementary Fig. 12). We thus selected the RF model to test the classification performance on additional 496 samples from the hold out test set. As expected, the trained RF model has very similar prediction result on the TCGA hold-out testing set, with an AUC = 0.947, sensitivity = 0.817 and specificity = 0.970 (Fig. 5d).

To further verify the robustness of the five-lincRNA panel, we tested the TCGA data based RF model on four independent RNA-Seq datasets: GSE58135 breast cancer, GSE50760 colon cancer, GSE25599 liver cancer and our breast cancer dataset (Fig. 5b, c and d). Impressively, this model predicts the other four independent data sets very well, with AUCs of 0.972 (95% CI: 0.95–0.9946), 0.841(95% CI: 0.6875–0.9946), 0.970 (95% CI: 0.9108–1) and 0.950 (95% CI: 0.867–1) for GSE58135, GSE50760, GSE25599 and our dataset, respectively (Fig. 5c and d). Other model evaluation metrics including Sensitivity, Specificity, Precision, Matthew's Correlation Coefficient, F-score and Accuracy in the validation datasets further demonstrate the excellent performance of the

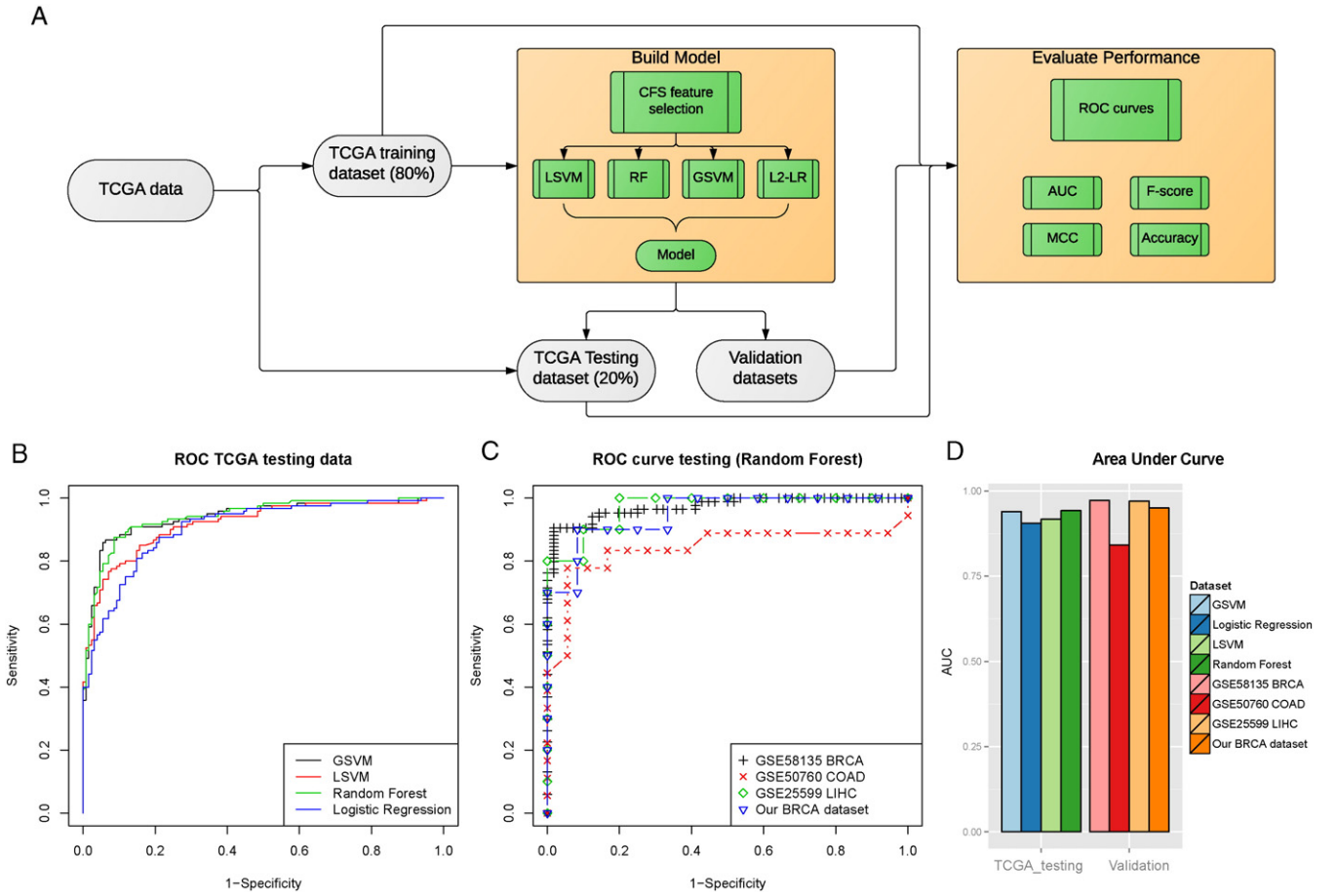

Fig. 5. The pan-cancer diagnostic model for the lincRNA panel. (a) The classification of the lincRNA panel was based on a computational RNA-Seq pipeline. The TCGA data were split into 80% training and 20% testing subsets. Five out of the six lincRNAs were selected as predictive features using Correlation Feature Selection (CFS). Pan-cancer diagnostic models were constructed using four standard classification machine learning methods: Random Forest (RF), Linear Support Vector Machines (LSVM), Gaussian Support Vector Machines (GSVM) and Logistic Regression (L2-LR). The best model was chosen based on various metrics of the Receiver operating characteristic (ROC) curves, including Area Under the Curve (AUC), F-score, Matthew's correlation coefficient (MCC) and Accuracy. (b) The performance of the classifier was analysed with the ROC curves on the TCGA hold-out testing data, based on the four classification methods mentioned above and (c) ROC curves of the top Random Forest model on four independent RNA-Seq validation datasets. (d) AUCs were calculated on the TCGA hold-out testing data in and the four validation datasets.

model (Supplementary Table VI). We therefore conclude that the panel of six lincRNAs are potential biomarkers for pan-cancer diagnosis.

3.7. The LincRNA Panel is Associated with Prognosis in Cancer Patients

Although the six lincRNAs were detected as potential diagnosis markers for pan-cancer, we were curious if they might be associated with the prognosis of cancer patients as well. Thus we performed survival analysis on 1201 samples from four TCGA datasets: namely BRCA, LUAD, LUSC datasets, and additionally the TCGA ovarian cancer (OV) dataset which was not used in the lincRNA signature discovery phase due to lack of normal samples (Supplementary Fig. 13). Since only overall survival information is available in TCGA in BRCA and OV datasets, we fit the overall survival with Cox-PH regression models and categorized the patient risks by prognosis index (PI) (Huang et al., 2014). The resulting Kaplan–Meier survival curves show that the lincRNA panel is able to separate patients into higher and lower risk groups by median PI, with log-rank tests p-values of 0.012 and 0.010 for BRCA and grade 3 OV, respectively (Supplementary Fig. 13a and b). On the other hand, the more preferable relapse free survival (RFS) in LUAD and LUSC datasets are available, thus we fit RFS with Cox-PH models, and obtained significant p-values of 0.0416 and 0.013 for differential survivals of LUAD and LUSC samples, respectively (Supplementary Fig. 13c and d). In summary, although the lincRNA panel was not purposely discovered as prognosis markers but rather diagnostic markers, their expression values are associated with the prognosis outcomes in various types of cancers.

3.8. Biological Relevance of LincRNAs Explored By Cell Culture Experiments

To explore the relationship between the lincRNAs panel and tumorigenic phenotypes, we conducted experiments using two breast cancer and colon cancer cell lines as examples. Given the extremely high homology between PCAN-2 and PCAN-3, we specifically designed siRNAs that target both of them so as to observe phenotypes. In non-aggressive MCF-7 and highly metastatic MDA-MB-231 cell lines, we efficiently knocked down two lincRNAs PCAN-2 and PCAN-3 (Fig. 6a). Transient knockdown allowed us to analyse cell proliferation and cell migration rate. Interestingly, the growth rate of fast proliferating MDA-MB-231 cells significantly decreased upon transfection with lincRNAs siRNA (Fig. 6b). To assess cell migration rates we employed the well-established wound-healing assay and followed the cell movement with time-lapse microscopy over the time of 24 h. As expected, the migration rate was significantly inhibited upon lincRNAs knockdown (Fig. 6c, d). The effect of lincRNA down-regulation on cell migration was more pronounced in a highly aggressive MDA-MB-231 cell line (0.349 versus 0.059 mm over 24 h for control and lncRNA siRNA, respectively) but it was also observed in much slower migrating MCF-7 cells (0.127 versus 0.096 mm over 24 h for control and lncRNA siRNA, respectively). We repeated the cell migration experiment on MDA-MB-231

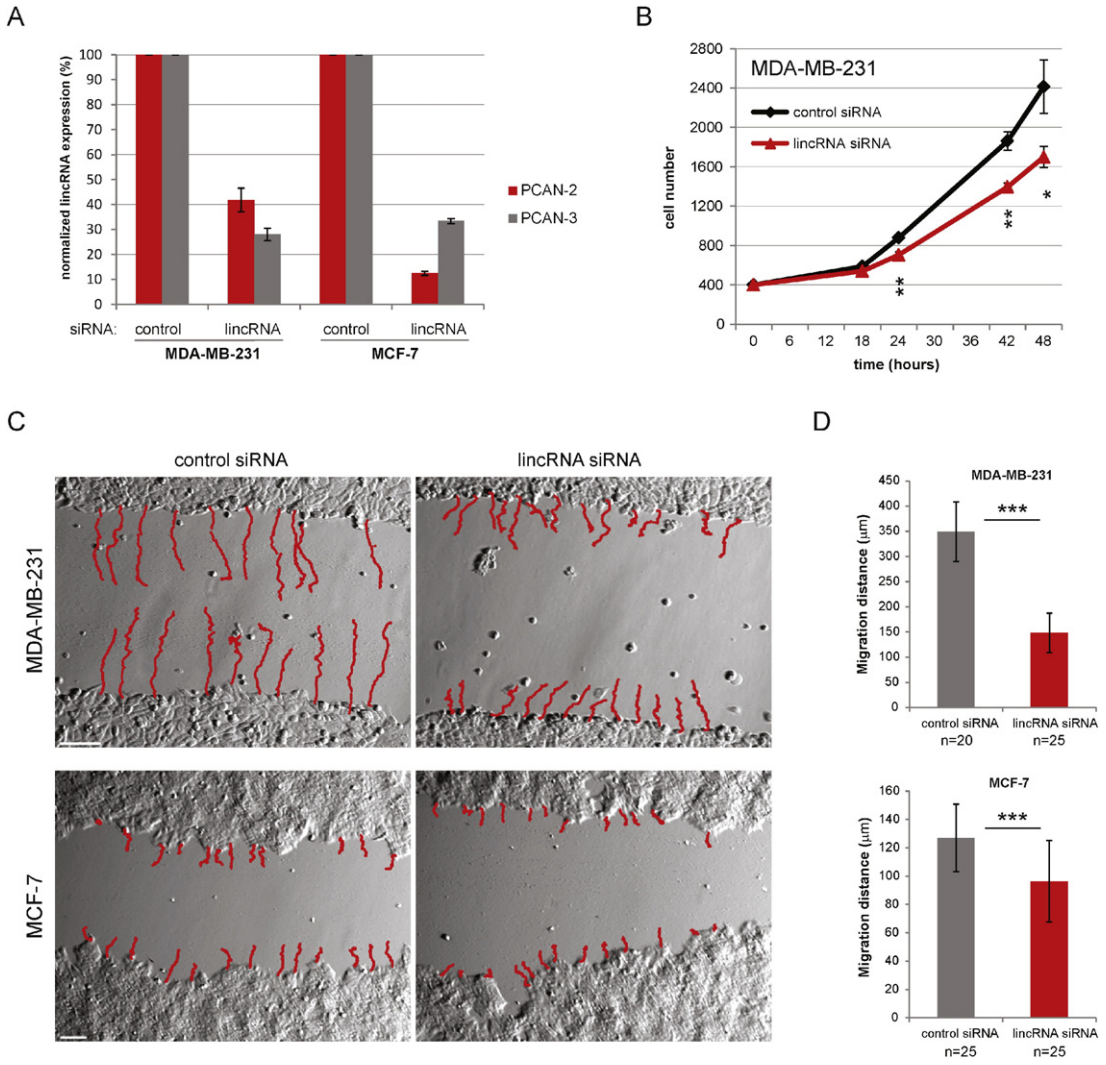

Fig. 6. The effect of lincRNAs downregulation on cell proliferation and migration. (a) PCAN-2 and PCAN-3 lincRNAs can be efficiently knocked down in MDA-MB-231 and MCF-7 cell lines. Bars represent RT-qPCR results of PCAN-2 and PCAN-3 expression. siRNA lincRNA bars show mean expression (n = 3) with S.D. normalized to the control condition. (b) Transient knockdown of PCAN-2 and PCAN-3 inhibits the growth rate of MDA-MB-231 cells. 30 h after transfection (time point "0") 400 cells were seeded in 96-well plates and processed for luminescent cell viability assay at indicated time points. Data points represent mean value (n = 3), error bars, S.D. *, $P < 0.05$, **$P < 0.01$. (c) lincRNA knockdown inhibits migration of MDA-MB-231 and MCF-7 cells in wound-healing assay. Cells were transiently transfected 30 h before making scratches in the cell monolayer. Cell migration rate was analysed with time-lapse microscopy. Red lines – cells tracks analysed over 24 h. Size bars – 100 μm. (d) Quantification of MDA-MB-231 and MCF-7 cells migration distance over 24-h time period. Bars – value of mean migration distance, error bars – S.D. (n = 20–25 analysed cells), ***$P < 0.001$.

with another less effective siRNA, and observed similar significantly slower ($P < 0.0001$) migrating rate (Supplementary Fig. 14).

Furthermore, we repeated these experiments in another HCT116 colon cancer cell line with the more efficient siRNA (Supplementary Fig. 15). Using the same experimental procedures, we observed significant differences in both cell proliferation ($p < 0.0001$) and migration ($p = 0.036$), between the lncRNA knockdown and the siRNA scrambled control. These results suggest that down-regulation of cancer cell abundant PCAN-2 and PCAN-3 lincRNAs weakens the typical cancer phenotypic features, such as proliferation and migration.

4. Discussion

Since 2012, a community effort has launched towards TCGA pan-cancer analysis across many different tumor types (Han et al., 2014; Weinstein et al., 2013), where the main focus has been the mutational landscape (Kandoth et al., 2013). Pan-Cancer Initiative aims to enable the discovery of novel intervention strategies that can be tested clinically, including developing novel adaptive biomarker-based clinical trials that cross boundaries between tumor types (Cancer Genome Atlas Research et al., 2013). One can expect that in the future, a pan-cancer screening biomarker panel from blood or other body fluids could become a useful, routine, and economical screening tool (Cancer Genome Atlas Research et al., 2013) applied before the patients have typical cancer symptoms that indicate late-stage character of the disease. Once an individual is identified as high-risk in the test, he or she can be followed up with more confirmative tests, such as imaging scanning. In the field of cancer biomarkers, although many lincRNAs and other lncRNAs have recently been implicated in cancer initiation and progression (Han et al., 2014; Vitiello et al., 2014; Iyer et al., 2015), the clinical potential of lincRNAs remains under-explored across different tumor types. In this study, our goals were to (1) depict the landscape of lincRNAs in pan-cancers, (2) demonstrate their relevance to clinical outcomes, such as tumor subtype, diagnosis and patient survival; and (3) explore the utilities of lincRNAs as pan-cancer diagnostic biomarkers.

Towards these goals, we have performed a new dimension of pan-cancer analysis using the lincRNA transcriptome. In total, we analyzed 3354 patient RNA-Seq samples from 12 types of cancers in TCGA (13 including OV in survival analysis) as well as an additional 15 independent

datasets (three RNA-Seq datasets from GEO, one in-house RNA-Seq breast cancer dataset and 11 microarray datasets from GEO). By systematically analyzing 12 types of RNA-Seq datasets in TCGA, we show that lincRNAs are more tissue specific than protein-coding genes. The loss of tissue specificity due to cancer is greater for lincRNAs compared to protein-coding genes. This suggests that lincRNAs can potentially be more sensitive biomarkers than protein coding genes. In addition, unsupervised clustering results of lincRNAs demonstrate significant correlations with molecular subtypes. CNMF clustering based on lincRNAs almost perfectly divided the Triple Negative and ER+/Her2- breast cancers into distinct groups in GSE58135 data set. Furthermore, CNMF clustering of TCGA BRCA samples detected 5 distinct clusters that highly correspond to the five widely used molecular subtypes based on the PAM50 signatures.

Although others have suggested that lincRNAs have potential as biomarkers (Prensner et al., 2011; Sun et al., 2013), we pinpoint a promising six-lincRNA pan-cancer diagnostics panel quantitatively, rigorously and robustly. Despite all the potential issues including population heterogeneity and sample size limitation in high throughput datasets (Berrar et al., 2006), the six-lincRNA biomarker model performs well overall with AUCs ranging from 0.972 to 0.841. Moreover, we verify the alteration of these lincRNAs with eleven additional microarray gene expression data sets. Our most unexpected finding is that the six lincRNA diagnostic signature is also associated with the survival prognosis of cancer patients, based on the TCGA datasets (BRCA, OV, LUAD and LUSC). Furthermore, we have demonstrated that the lincRNAs have biological functions, by knocking-down experiments on two of them, PCAN-2 and PCAN-3. Our preliminary results indicate that down-regulation of only two out of six panel lincRNAs is sufficient to partially revert some of the typical physiological hallmarks of cancer cells including fast proliferation and more importantly, migration.

Developing a pan-cancer biomarker model based on the lincRNA signatures could be very significant clinically, providing complementary values to protein-coding gene based biomarker panels. We plan to continue our translational investigations in this direction. Yet our next challenge is to understand how each of the identified lincRNA biomarkers function in tumorigenesis and progression. Although lincRNAs do not encode proteins, it's clear that they play important roles in cellular biology. Currently, multiple hypotheses exist on how lincRNAs regulate cellular functions (Ching et al., 2014), which include functioning as scaffold structure (Kowalczyk et al., 2012; Ling et al., 2013), sponge of small regulatory RNAs (Liu et al., 2013; Salmena et al., 2011) or direct interaction with proteins to modulate localization and activity(Ma et al., 2012). To better understand the phenotypic effects of the six lincRNAs, we will proceed with experiments that address the physiological functions of these lincRNAs as well as molecular mechanisms by which they promote tumorigenesis and/or malignancy. We are aware that the repertoire of lincRNAs is evolving and thus we may miss some newly identified lincRNAs, such as reported recently (Iyer et al., 2015). However, given the fact that the six lincRNAs in this report have reached very high and robust accuracy in pan-cancer data, the addition of other new lincRNAs is expected to add very small effect on diagnosis at most.

In summary, our initial pan-cancer analysis has demonstrated that lincRNAs accurately classify cancer subtypes through supervised as well as unsupervised methods. The panel of six lincRNAs is a highly accurate diagnostic biomarker signature with additional prognostic value. These results highlight lincRNAs as a new paradigm for actionable pan-cancer diagnosis and prognosis.

Author contributions

LXG and TC envisioned the project and designed the work. TC conducted the data analysis, with assistance from SH and XZ. TC and LXG wrote the manuscript. SY and HY provided the UHCC RNA samples, and RF helped to sequence the UHCC breast cancer samples. MT, TC and LXG designed the qPCR primers and KP, TC, JM, MT and BF collaborated on cell culture and qPCR validations. All authors have read, revised and approved the final manuscript.

Competing Financial Interests

The author(s) declare no competing financial interests.

Acknowledgements

This research was supported by grants K01ES025434 awarded by NIEHS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), P20 COBRE GM103457 awarded by NIH/NIGMS, and Hawaii Community Foundation Medical Research Grant 14ADVC-64566 to L.X. Garmire. B. Fogelgren is supported by awards from NIH (1K01DK087852, R03DK100738, and P20GM103456-06 A1-8293); the March of Dimes (#5-FY14-56); Hawaii Community Foundation (12ADVC-51347); University of Alabama at Birmingham HepatoRenal Fibrocystic Disease Core Center (5P30DK074038), and RCMI-BRIDGES at the University of Hawaii (5G12MD007601). The UHCC GSR is supported by the NCI P-30 grant CA071789-15. We would like to thank Dr. Joe Ramos and Paul Anastasiadis for providing breast cancer cell lines, Dr. Peiwen Fei for the colon cancer cell line, and the UHCC Microscopy and Imaging Shared Resource for using the facility. We would also like to thank Dr. Jason Moore for providing access to the GPU system at Dartmouth College.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.ebiom.2016.03.023.

References

Berrar, D., Bradbury, I., Bubitzky, W., 2006. Avoiding model selection bias in small-sample genomic datasets. Bioinformatics 22, 1245–1250.
BROAD, 2014. Broad Institute TCGA Genome Data Analysis Center (2014): Analysis Overview for 15 July 2014. Broad Institute of MIT and Harvard.
Brockdorff, N., Ashworth, A., Kay, G.F., Cooper, P., Smith, S., Mccabe, V.M., Norris, D.P., Penny, G.D., Patel, D., Rastan, S., 1991. Conservation of position and exclusive expression of mouse xist from the inactive X chromosome. Nature 351, 329–331.
Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J.L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 25, 1915–1927.
Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinf. 10, 421.
Cancer Genome Atlas, N., 2012. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70.
Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., 2013. The cancer genome atlas Pan-cancer analysis project. Nat. Genet. 45, 1113–1120.
Ching, T., Huang, S., Garmire, L.X., 2014. Power analysis and sample size estimation for RNA-seq differential expression. RNA 20, 1684–1696.
Consortium, E.P., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.
Du, Z., Fei, T., Verhaak, R.G., Su, Z., Zhang, Y., Brown, M., Chen, Y., Liu, X.S., 2013. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nat. Struct. Mol. Biol. 20, 908–913.
Garmire, L.X., Garmire, D.G., Huang, W., Yao, J., Glass, C.K., Subramaniam, S., 2011. A global clustering algorithm to identify long intergenic non-coding RNA-with applications in mouse macrophages. PLoS One 6, e24051.
Ge, X., Chen, Y., Liao, X., Liu, D., Li, F., Ruan, H., Jia, W., 2013. Overexpression of long noncoding RNA PCAT-1 is a novel biomarker of poor prognosis in patients with colorectal cancer. Med. Oncol. 30, 1–6.
Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J.L., 2010. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464, 1071–1076.
Habel, L.A., Shak, S., Jacobs, M.K., Capra, A., Alexander, C., Pho, M., Baker, J., Walker, M., Watson, D., Hackett, J., 2006. A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. Breast Cancer Res. 8, R25.
Han, L., Yuan, Y., Zheng, S., Yang, Y., Li, J., Edgerton, M.E., Diao, L., Xu, Y., Verhaak, R.G., Liang, H., 2014. The Pan-cancer analysis of pseudogene expression reveals biologically and clinically relevant tumour subtypes. Nat. Commun. 5.
Huang, S., Yee, C., Ching, T., Yu, H., Garmire, L.X., 2014. A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. PLoS Comput. Biol. 10, e1003851.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P.M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., 2003. MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. Oncogene 22, 8031–8041.
Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., 2015. The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. 47 (3), 199–208.
Kandoth, C., Mclellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., Mcmichael, J.F., Wyczalkowski, M.A., Leiserson, M.D., Miller, C.A., Welch, J.S., Walter, M.J., Wendl, M.C., Ley, T.J., Wilson, R.K., Raphael, B.J., Ding, L., 2013. Mutational landscape and significance across 12 major cancer types. Nature 502, 333–339.
Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Morales, D.R., Thomas, K., Presser, A., Bernstein, B.E., Van Oudenaarden, A., 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc. Natl. Acad. Sci. 106, 11667–11672.
Kowalczyk, M.S., Higgs, D.R., Gingeras, T.R., 2012. Molecular biology: RNA discrimination. Nature 482, 310–311.
Liang, C.C., Park, A.Y., Guan, J.L., 2007. In vitro scratch assay: a convenient and inexpensive method for analysis of cell migration in vitro. Nat. Protoc. 2, 329–333.
Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhao, G., Luo, H., Bu, D., Zhao, H., 2011. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. Nucleic Acids Res. 39, 3864–3878.
Liao, Y., Smyth, G., Shi, W., 2013. featureCounts: an efficient general-purpose read summarization program. (*arXiv*, 1305, 16).
Liao, Y., Smyth, G.K., Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930.
Ling, H., Spizzo, R., Atlasi, Y., Nicoloso, M., Shimizu, M., Redis, R.S., Nishida, N., Gafâ, R., Song, J., Guo, Z., 2013. CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. Genome Res. 23, 1446–1461.
Liu, K., Yan, Z., Li, Y., Sun, Z., 2013. Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis. Bioinformatics 29, 2221–2222.
Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. Methods 25, 402–408.
Love, M., Anders, S., Huber, W., 2013. Differential Analysis of RNA-Seq Data at the Gene Level Using the DESeq2 Package.
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. (*bioRxiv*).
Ma, H., Hao, Y., Dong, X., Gong, Q., Chen, J., Zhang, J., Tian, W., 2012. Molecular mechanisms and function prediction of long noncoding RNA. Sci. World J. 2012.
Mchugh, C.A., Russell, P., Guttman, M., 2014. Methods for comprehensive experimental identification of RNA–protein interactions. Genome Biol. 15, 203.
Menor, M., Ching, T., Zhu, X., Garmire, D., Garmire, L.X., 2014. mirMark: a site-level and UTR-level classifier for miRNA target prediction. Genome Biol. 15, 500.
Penny, G.D., Kay, G.F., Sheardown, S.A., Rastan, S., Brockdorff, N., 1996. Requirement for xist in X chromosome inactivation. Nature 379, 131–137.
Prensner, J.R., Iyer, M.K., Balbin, O.A., Dhanasekaran, S.M., Cao, Q., Brenner, J.C., Laxman, B., Asangani, I.A., Grasso, C.S., Kominsky, H.D., 2011. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat. Biotechnol. 29, 742–749.
Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., 2007. Functional demarcation of active and silent chromatin domains in human < i > HOX</i > loci by noncoding RNAs. Cell 129, 1311–1323.
Rubie, C., Kempf, K., Hans, J., Su, T., Tilton, B., Georg, T., Brittner, B., Ludwig, B., Schilling, M., 2005. Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. Mol. Cell. Probes 19, 101–109.
Salmena, L., Poliseno, L., Tay, Y., Kats, L., Pandolfi, P.P., 2011. A < i > ceRNA</i > hypothesis: the Rosetta Stone of a hidden RNA Language? Cell 146, 353–358.
Sun, K., Chen, X., Jiang, P., Song, X., Wang, H., Sun, H., 2013. iSeeRNA: identification of long intergenic non-coding RNA transcripts from transcriptome sequencing data. BMC Genomics 14, S7.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515.
Tripathi, V., Ellis, J.D., Shen, Z., Song, D.Y., Pan, Q., Watt, A.T., Freier, S.M., Bennett, C.F., Sharma, A., Bubulya, P.A., 2010. The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation. Mol. Cell 39, 925–938.
Ulitsky, I., Bartel, D.P., 2013. lincRNAs: genomics, evolution, and mechanisms. Cell 154, 26–46.
Vitiello, M., Tuccoli, A., Poliseno, L., 2014. Long non-coding RNAs in cancer: implications for personalized therapy. Cell. Oncol. 1–12.
Volinia, S., Croce, C.M., 2013. Prognostic microRNA/mRNA signature from the integrated analysis of patients with invasive breast cancer. Proc. Natl. Acad. Sci. U. S. A. 110, 7413–7417.
Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., 2010. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. gkq622.
Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.-P., Li, W., 2013. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. Nucleic Acids Res. 41, e74–e74.
Weakley, S.M., Wang, H., Yao, Q., Chen, C., 2011. Expression and function of a large non-coding RNA Gene XIST in human cancer. World J. Surg. 35, 1751–1756.
Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J.M., Network, C.G.A.R., 2013. The cancer genome atlas pan-cancer analysis project. Nat. Genet. 45, 1113–1120.
Wilks, C., Cline, M.S., Weiler, E., Diehkans, M., Craft, B., Martin, C., Murphy, D., Pierce, H., Black, J., Nelson, D., 2014. The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. Database 2014, bau093.
Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., Chen, R., 2014. NPInter v2. 0: an updated database of ncRNA interactions. Nucleic Acids Res. 42, D104–D108.