Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

Genome and Epigenome

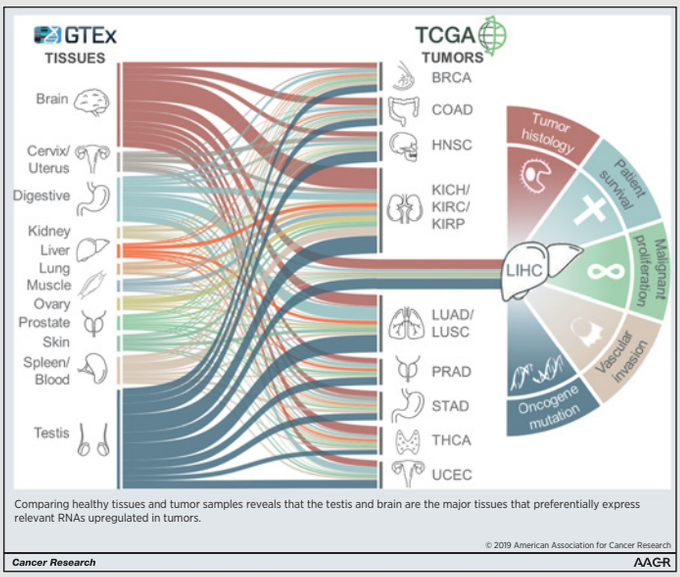
Cancer Research

Identification of Coding and Long Noncoding RNAs Differentially Expressed in Tumors and Preferentially Expressed in Healthy Tissues

Juan P. Unfried[1], Guillermo Serrano[2], Beatriz Suárez[1], Paloma Sangro[3,4,5], Valeria Ferretti[1], Celia Prior[1], Loreto Boix[5,6], Jordi Bruix[5,6], Bruno Sangro[3,4,5], Víctor Segura[2,4], and Puri Fortes[1,4]

Abstract

The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) datasets allow unprecedented gene expression analyses. Here, using these datasets, we performed pan-cancer and pan-tissue identification of coding and long noncoding RNA (lncRNA) transcripts differentially expressed in tumors and preferentially expressed in healthy tissues and/or tumors. Pan-cancer comparison of mRNAs and lncRNAs showed that lncRNAs were deregulated in a more tumor-specific manner. Given that lncRNAs are more tissue-specific than mRNAs, we identified healthy tissues that preferentially express lncRNAs upregulated in tumors and found that testis, brain, the digestive tract, and blood/spleen were the most prevalent. In addition, specific tumors also upregulate lncRNAs preferentially expressed in other tissues, generating a unique signature for each tumor type. Most tumors studied downregulated lncRNAs preferentially expressed in their tissue of origin, probably as a result of dedifferentiation. However, the same lncRNAs could be upregulated in other tumors, resulting in "bimorphic" transcripts. In hepatocellular carcinoma (HCC), the upregulated genes identified were expressed at higher levels in patients with worse prognosis. Some lncRNAs upregulated in HCC and preferentially expressed in healthy testis or brain were predicted to function as oncogenes and were significantly associated with higher tumor burden, and poor prognosis, suggesting their relevance in hepatocarcinogenesis and/or tumor evolution. Taken together, therapies targeting oncogenic lncRNAs should take into consideration the healthy tissue, where the lncRNAs are preferentially expressed, to predict and decrease unwanted secondary effects and increase potency.



Significance: Comprehensive analysis of coding and noncoding genes expressed in different tumors and normal tissues, which should be taken into account to predict side effects from potential coding and noncoding gene-targeting therapies.

Graphical Abstract: http://cancerres.aacrjournals.org/content/canres/79/20/5167/F1.large.jpg.

[1]University of Navarra (UNAV). Center for Applied Medical Research (CIMA). Program of Gene Therapy and Hepatology. Pamplona, Spain. [2]UNAV/CIMA. Bioinformatics Platform. Pamplona, Spain. [3]Clínica Universidad de Navarra. Liver Unit. Pamplona, Spain. [4]Navarra Institute for Health Research (IdiSNA), Pamplona, Spain. [5]Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBERehd), Spain. [6]BCLC Group, Hospital Clinic-IDIBAPS, Barcelona, Spain.

J.P. Unfrid and G. Serrano are co-first authors. B. Sangro, V. Segura, and P. Fortes are co-last authors of this article.

Corresponding Author: Puri Fortes, CIMA, IDISNA, Pio XII, 55, Pamplona 31008, Spain. Phone: 34-948-194700; Fax: 34-948-194717; E-mail: pfortes@unav.es

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

Cancer Res 2019;79:5167–80

doi: 10.1158/0008-5472.CAN-19-0400

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

Introduction

Traditionally, cancer has been classified following pathologic criteria and according to the tissue of origin. More recently, efforts carried out by different consortia, such as The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), or Functional Annotation of the Mammalian Genome (FANTOM), have allowed the characterization of molecular patterns in thousands of tumors in a high-throughput manner (1–3). The results from these large-scale analyses have changed the landscape of tumor evaluation and are expected to revolutionize tumor understanding. Molecular analyses of specific tumors have allowed subclassifications that correlate with clinical features, including patient survival (4–6). These analyses have helped to identify pathways that drive tumor growth and, therefore, may aid in the selection of effective therapies. When tumors from different tissues of origin are analyzed simultaneously, the study of molecular patterns fosters the discovery of common deregulated pathways that have an impact on cancer hallmarks, highlighting the relevance of pan-cancer analyses (7–10).

Several molecular parameters have been studied in specific tumors or in a pan-cancer manner to search for candidate driver genes and/or pathways, and to evaluate their impact on cancer therapy, prognosis, and classification. Most studies focus on DNA (mutations and copy number variations), epigenetic, and/or transcriptomic data (8, 9, 11, 12). Many gene expression analyses have studied deregulated coding transcripts and miRNAs, whereas some have focused on deregulated long noncoding RNAs (lncRNA; refs. 11, 13–16). LncRNAs are transcripts longer than 200 nucleotides with poor coding capacity and with the potential to function as RNA molecules (17). LncRNAs are similar to mRNAs as most are transcribed by polymerase II and many are processed by splicing and polyadenylation. However, compared with mRNAs, lncRNA splicing tends to be less efficient, and lncRNAs are poorly conserved, poorly expressed, preferentially nuclear, and more tissue-specific (18). Although lncRNA genes are numerous, very few have been well characterized (19). Available studies show that some lncRNAs act *in trans*, away from their site of transcription, whereas others act *in cis*, regulating the expression of neighboring genes. Most described lncRNAs carry out regulatory functions by interacting with proteins, DNA, or other RNAs. The function of some of them is relevant for cell homeostasis, growth, and differentiation. In fact, several lncRNAs such as HOTAIR or PVT1 behave as oncogenes (20, 21), whereas others such as MEG3 or PANDAR act as tumor suppressors (22, 23). LncRNAs involved in tumor growth have been found deregulated in transcriptome analyses of particular tumors or in pan-cancer approaches (14, 24). These studies show that most deregulated lncRNAs are tumor-specific, whereas a handful behave as "onco-lncRNAs," deregulated across many cancer types (14, 16).

We have compared the human coding and long noncoding transcriptome using RNA sequencing (RNA-seq) data from the TCGA and Genotype-Tissue Expression (GTEx) consortia. Our pan-cancer analyses of differential expression also show that deregulated lncRNAs are more tumor-specific. Some transcripts that are aberrantly upregulated in a tumor are preferentially expressed in specific healthy tissues, which could be defined as "donor-tissues". Interestingly, most lncRNAs upregulated in tumors are "donated" by testis, brain, the digestive tract, and blood/spleen, although different tumors have specific "donor-tissues" of tumor transcripts. Some lncRNAs upregulated in hepatocellular carcinoma (HCC) and "donated" by testis, brain, or other tissues associate significantly with important clinical parameters, and are predicted to enhance cell expression and growth, suggestive of their relevance in hepatocarcinogenesis and/or tumor progression. We propose that therapies aiming to target the expression of oncogenic lncRNAs should take into consideration the lncRNA "donor-tissue" to increase therapeutic efficacy and decrease unwanted secondary effects.

Materials and Methods

Human samples and cell lines

Public clinical and histologic data of patients with HCC from TCGA were used for analysis (Supplementary Table S1: Patient data; refs. 6, 25). Human liver cancer and peritumoral samples were obtained from 52 HCC patients treated with hepatic resection or liver transplantation from January 2011 to December 2017 in two hospitals, Clínica Universidad de Navarra in Pamplona (31 patients) and Hospital Clinic in Barcelona (21 patients; BCL-CUN cohort). The study was approved by the Institutional Ethics Committee of each hospital (reference number 121/2015), and all patients gave their written informed consent for medical research. Clinical and histologic characteristics as well as relevant outcomes were obtained from medical records.

HuH7 (Dr. Chisari's laboratory, Scripps Research Institute, La Jolla, CA), Hep3B, HepG2, PLC, and SK-Hep (ATCC) liver cells and 293T, A549, and nonimmortalized BJ fibroblasts (ATCC) were grown at 37°C in a humid atmosphere containing 5% CO_2 for less than 20 passages after thawing. Cell lines were cultured in DMEM supplemented with 2 mmol/L glutamine, 1% penicillin/streptomycin, and enriched with 10% FBS and were tested for *Mycoplasma* every other week with MycoAlert (Lonza).

RNA extraction, RT-PCR, and qPCR

RNA extraction from peritumoral and tumor tissue or cells was performed using the Maxwell 16 LEV simply RNA Purification Kit (Promega) following the manufacturer's recommendations. The RNA concentration was measured using a NanoDrop 1000 Spectrophotometer. For reverse transcription, 1 μg of RNA was incubated in M-MLV-RT buffer, 5 mmol/L DTT, 200 units M-MLV-RT enzyme (#28025013, Invitrogen), 0.5 mmol/L dNTPs (#10297018, Invitrogen), and 10 ng/mL random primers (#48190011, Invitrogen) in a final volume of 40 μL. The reaction was set at 37°C for 60 minutes and 95°C for 1 minute in the C1000 Touch Thermal Cycler from Bio-Rad and immediately placed at 4°C. Quantitative PCR was performed with 5 μL SYBR Green Supermix Reagent (#1708880, Bio-Rad), 0.27 μmol/L of each primer and 2 μL of the cDNA mix in a final volume of 11 μL in the CFX96 Real-Time system from Bio-Rad. The mixture was incubated at 95°C for 3 minutes, then at 95°C for 15 seconds, 60°C for 15 seconds, and 72°C for 25 seconds for 34 cycles, and finally, 1 minute at 95°C and 1 minute at 65°C. The results were analyzed with Bio-Rad CFX manager software. Gene expression was normalized to the *RPLP0* housekeeping gene, a known invariant control for HCC, and expressed as 2-ΔCT. The primers used were designed using the Primer3 online tool (http://primer3.ut.ee/; ref. 26) and are listed in Supplementary Table S1: Primer data.

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

Public datasets and data normalization for bioinformatic analyses

Tissue specificity and preferential expression studies were performed using the GTEx dataset (https://gtexportal.org/) with RNA-seq data from more than 7,000 samples from 53 different tissues (27). The matrix of raw counts obtained using STAR aligner with hg19 assembly and annotated with Gencode version 19 was downloaded from the web portal of the project (GTEx release V6). Tumor-specific, tumor preferential, and differential expression studies were performed using the RNA-seq data from cancer samples from TCGA (http://cancergenome.nih.gov/), which has analyzed tumors and matched tumor and peritumoral tissues from 11,000 patients to study 34 cancer types and subtypes (2). The matrix of raw counts obtained using STAR aligner with hg38 assembly and annotated with Gencode version 22 was downloaded from the GDC Data Portal (https://portal.gdc.cancer.gov/).

The analysis of differentially expressed genes and gene profiles of both protein coding and long noncoding genes in normal tissues (GTEx) and cancer samples (TCGA) was carried out following the bioinformatics workflow provided by limma to allow the analysis of RNA-seq experiments using linear models. In this pipeline, the datasets were first normalized with edgeR using the TMM (trimmed mean of M-values) normalization, then the $\log_2$CPM values were calculated using voom from limma package and the normalized expression matrix was used for the statistical analysis with limma (28). Before normalization, genes with less than 5 counts in all the samples (nonexpressed genes) were removed from the analysis.

Differential expression analyses

Only those cancer types from the TCGA with more than 20 paired tumor–peritumor samples were included in the analysis of differential gene expression, resulting in the study of 13 of the 34 tumor types analyzed by the TCGA. For the 13 tumor types, differentially expressed genes were selected using a sampling-based strategy, due to the low number of control samples available for some cancers. Using this approach, we identified those genes with higher probability of being deregulated in cancer. For each of the 13 TCGA tumor types studied, we performed the DEG analysis with the limma package workflow for RNA-seq data analysis from R/Bioconductor (29) using as input, data from all the peritumoral samples and a random sampling of the same number of tumoral samples with 200 iterations. We selected the set of genes differentially expressed for each cancer type using the criteria of FDR < 1% in all the iterations. As an example, in the case of liver hepatocellular carcinoma (LIHC), which has 50 peritumoral samples analyzed, we performed the DE analysis with data from the 50 peritumoral samples and 50 tumor samples taken randomly from all the 374 HCCs studied. This analysis was performed 200 times and the tumor samples chosen were different each time. The representative value of the fold change for each gene was calculated using all the samples for each cancer type.

Specific and preferential expression analysis

For each normalized dataset, GTEx and TCGA, two gene profiling studies were carried out (one for protein coding and another one for noncoding genes) using SOM neural networks (30) implemented in the Kohonen package of R of manually optimized dimension 15×15 (225 clusters) and 2,000 iterations. We confirmed the convergence of the algorithm following the error rate and recalculated the centroids of each cluster using the mean value to generate a representative expression profile for each of them. Afterwards, a manual curation of all the clusters was performed to define the set of expression profiles and assign the corresponding preferential tissue or cancer type to the cluster. This was done for those clusters whose first quartile of expression in the highest-expressing tissue or cancer-type was over the second quartile of expression in the second highest-expressing tissue or cancer type. The grouping of cancer and tissues was carried out manually after visual inspection of the centroid expression patterns.

In addition, gene specificity was evaluated for all genes for which the sum of counts in all the samples (GTEx or TCGA) was lower than 5 counts. We quantified the similarity between the expression profiles of those genes and a set of representative patterns in which extreme cases of expression in only one type of cancer or tissue were considered (31). Pearson correlation was used to measure similarity and only those clusters with $P < 0.05$ were considered for further analysis. Only those genes with a first similarity value to a representative expression profile higher than a predefined threshold of correlation >0.25 and a difference of correlation of 0.15 between the first and the second value were considered cancer or tissue specific and were assigned a specific type.

Data integration using a systems biology approach

We generated independent network analyses of cancer deregulated and cancer and tissue specific or preferentially expressed protein coding and noncoding genes. Tissues and tumors with less than 30 specific or preferentially expressed coding or long noncoding genes were excluded from the analysis. The networks were used to study the statistical significance of the intersections: (i) between deregulated genes in the 13 different tumor types included in the analysis; (ii) between the genes with specific/preferential expression in the tumors and the genes with specific/preferential expression in the tissues represented in the GTEx dataset; and (iii) between the genes upregulated in the studied cancer types and the genes with specific/preferential expression in the tissues.

The statistical analysis to evaluate the significance of the intersections was based on the hypergeometric distribution (32). For each pair of gene sets, we defined the reference set as the number of genes expressed in the TCGA, the category of interest as the number of genes of one of the gene sets, the selection as the number of genes in the other gene set and we evaluated the statistical significance of the intersection between them. The P values were corrected for multiple hypotheses testing using Bonferroni method. We calculated a score from the P value obtained using the expression $-\log_{10}$ (P value). Because of the large size of the graphs obtained, we filtered these networks removing all the genes with a degree $1 \leq d < 10$. The genes with degree 1 are the most specific ones and the genes with degree higher or equal to 10 are the hubs of the network. As expected, in the analysis of the intersection between genes with specific/preferential expression in the tumors or genes with specific/preferential expression in the tissues, the resulting score is 0.

Functional analyses

Functional enrichment analysis of Gene Ontology (GO) categories was carried out using standard hypergeometric test. The biological knowledge extraction was complemented using Ingenuity Pathway Analysis (IPA; Ingenuity Systems, www.ingenuity.

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

com). Ingenuity database includes manually curated and fully traceable data derived from literature sources. To predict lncRNA function, we performed a guilt-by-association analysis (GBA) considering the samples of the TCGA from the cancer types included in this study. The Pearson correlation value was calculated between each lncRNA of interest and the remaining set of protein coding genes. The correlation matrix obtained was used as input for giTools (33), where an enrichment analysis of GO categories was performed for each lncRNA using a method based on the *Z*-score to calculate the enrichment *P* values and corrected using the FDR (34). In addition, the ranked list of correlated genes for each lncRNA was used in the nonparametric Kolmogorov–Smirnoff rank test as implemented in the Gene Set Enrichment Analysis (GSEA) software (35). Gene categories were selected from MsigDB database using the CP (gene sets from pathway databases) collection of gene sets. The *P* values for each gene set were computed on the basis of 2,000 permutation iterations. We selected those categories enriched with FDR = 0.

Statistical analysis

All the enrichment analyses using the hypergeometric distribution and the association studies were performed using the statistical environment R/Bioconductor except the GBA study, which was performed using the giTools software (36). To evaluate prognosis potential, we used a classifier based on logistic regression and ROC analysis to measure the performance of the classifier using neoplasm histologic grade (G3 vs. G1; refs. 37, 38). Fisher exact test using higher/lower than median groups was used to find significant associations. Expression data are shown as means ± SD and statistical analyses were performed using Prism 5 (GraphPad Software). A descriptive analysis was carried out to analyze the distribution of the samples with D'Agostino and Pearson normality test. Nonparametric tests were used after normality failure. Differences between two groups were analyzed using two-tailed Student *t* test or U-Mann–Whitney, whereas differences between three groups were analyzed using the Kruskal–Wallis ANOVA test followed by Dunn multiple comparisons test. TCGA's paired samples were evaluated with paired *t* tests, while paired samples from the BCL-CUN cohort were analyzed with Wilcoxon matched pairs signed rank tests. Differences were deemed significant for a real alpha of 0.05 ($P < 0.05$). Statistical significance is indicated by *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P \leq 0.0001$. ns indicates nonsignificant differences.

Results

Pan-cancer analysis of coding and long noncoding deregulated genes

We identified recurrently deregulated transcripts using RNA-seq data from the TCGA for tumor/peritumor samples across 13 different cancer types in which at least 20 peritumoral samples had been studied. These tumors are breast 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 HCC (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), prostate adenocarcinoma (PRAD), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), and uterine corpus endometrial carcinoma (UCEC). For each tumor type, we defined recurrently deregulated transcripts as those with an FDR < 1% after performing 200 comparisons between RNA-seq data from all peritumoral samples and the same number of tumor samples chosen randomly (see Materials and Methods for details). As a control for the analysis, enrichment of functional categories using GO annotations and IPA was evaluated for protein-coding genes deregulated in each cancer type. The results showed an overrepresentation of cancer and cancer-related pathways ($P < 1e{-}27$). Although the human transcriptome has more long noncoding genes (including pseudogenes) than coding genes, in all tumors evaluated, we identified a higher number of deregulated coding transcripts versus long noncoding transcripts (Fig. 1A; Supplementary Table S2). Interestingly, 22% to 36% of deregulated lncRNA genes were located close (<5 Kb) to deregulated coding genes, suggesting that these neighboring genes are co-deregulated. Recurrently deregulated transcripts were preferentially downregulated in most tumors (Fig. 1B; Fisher exact test $P < 2.2e{-}16$ and Supplementary Fig. S1A). In addition, the degree of downregulation (absolute fold change) was significantly higher than the degree of upregulation for most tumors. Supplementary Fig. S1B shows the result of the analysis with the 100 upregulated or downregulated genes with the highest absolute fold change for each tumor type, but similar results have been obtained when all deregulated genes have been considered. Interestingly, the average fold change of the top coding and long noncoding deregulated transcripts showed a similar trend in most tumors (Supplementary Fig. S1C). In summary, this TCGA data analysis indicates that most tumors show more downregulated than upregulated transcripts and more coding than noncoding deregulated genes.

Some recurrently deregulated transcripts have been identified only in one of the 13 tumor types studied. For those identified in 2 or more tumors, the transcript could be upregulated, downregulated, or bimorphic, that is, upregulated in some tumors and downregulated in others. Under arbitrary conditions, the number of bimorphic transcripts should increase with the number of tumors in which recurrently deregulated transcripts are found, as the possibility that a transcript is upregulated in some tumors and downregulated in others increases. However, we observed that the number of bimorphic transcripts gradually decreases when transcripts are deregulated in more than 5 tumors (Fig. 1C). This suggests a purifying selection that favors the relevance of those candidates that are either upregulated or downregulated in a high number of tumors. We found 128 lncRNAs to be significantly upregulated and 320 lncRNAs downregulated in more than 5 tumors (Supplementary Table S2: >5 tumors). They include previously described oncogenes such as PVT1, GAS5, or HAGLROS (20, 39, 40) and tumor suppressors such as FENDRR or ADAMTS9-AS2 (41, 42). Unlike noncoding genes, most coding genes deregulated in more than 10 tumor types are upregulated (Fig. 1C).

Comparing the deregulated coding transcriptome among the 13 evaluated tumors brings together lung cancers (LUAD and LUSC), gynecologic cancers (BRCA and UCEC), gastrointestinal tumors (COAD, STAD), and kidney tumors (KIRC and KIRP; Fig. 2A). Evaluation of the deregulated long noncoding transcriptome shows expected associations (lung to lung or kidney to kidney tumors) but also some unexpected and highly significant associations such as breast to kidney tumors. Interestingly, network analysis shows that coding gene deregulation is very promiscuous, while deregulated lncRNAs are more tumor-

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

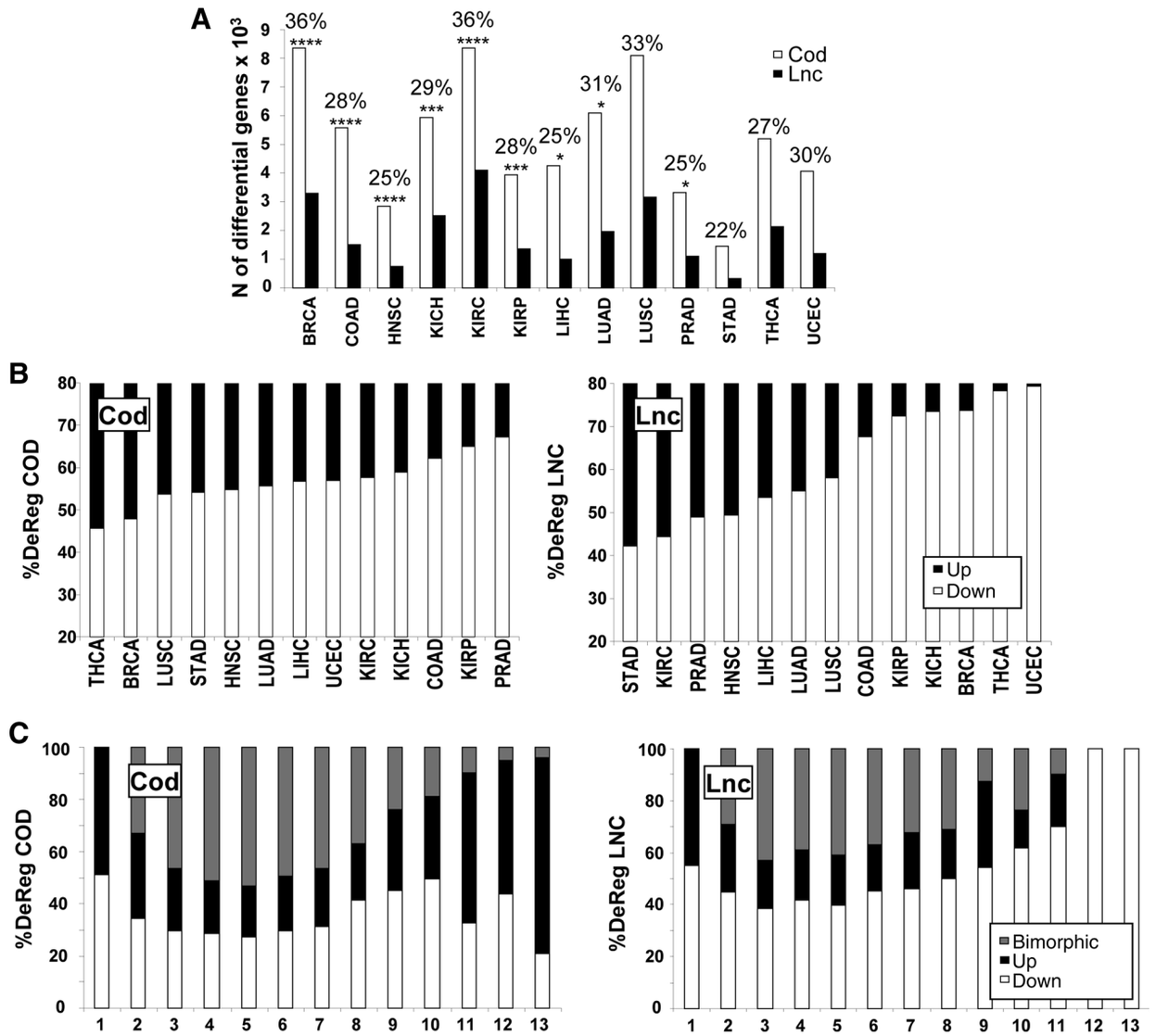


Figure 1.

Analysis of mRNAs and lncRNAs deregulated in individual tumors from the TCGA. **A,** Number of recurrently deregulated coding and long noncoding genes in the indicated tumors. For each tumor, the percentage of deregulated long noncoding genes located close to deregulated coding genes versus total deregulated long noncoding genes is indicated at the top of the bar. Statistical significance of the co-deregulation of neighboring genes is also indicated. **B,** Percentage of upregulated and downregulated coding (left) and long noncoding (right) transcripts versus the total number of coding and long noncoding transcripts, respectively, deregulated in each tumor. **C,** Percentage of upregulated, downregulated, or bimorphic coding (left) and long noncoding (right) transcripts versus the total number of transcripts deregulated in one or more tumors, up to the 13 tumors examined.

specific (Fig. 2B). In fact, most coding genes deregulated in a particular tumor are also deregulated in other 3 to 5 tumors, whereas this number is reduced to 0 to 3 tumors when lncRNAs are evaluated (Fig. 2C).

Pan-cancer and pan-tissue analyses of specific and preferential expression

Cancer-deregulated lncRNAs are more tumor-specific and the tissue specificity of lncRNAs has been well described (43). Hence, we searched for healthy tissues where cancer-upregulated lncRNAs are specifically or preferentially expressed. Coding transcripts were evaluated in parallel for comparison. First, we identified coding and long non-coding transcripts that are specifically expressed in a tissue or in a tumor using GTEx and TCGA data, respectively (see methods for details). As previously described, evaluation of GTEx data shows that testis and brain express many specific lncRNAs and that there are more tissue-specific lncRNAs than coding transcripts (super-specific genes in Fig. 3A; ref. 43). Similarly, TCGA analysis shows that several tumors express specific lncRNAs and that there are more tumor-specific lncRNAs than coding transcripts. Overall, the number of specific transcripts was low. To identify a collection of transcripts preferentially expressed in a particular tissue or tumor, we used neural networks (see Materials and Methods for details; Supplementary

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

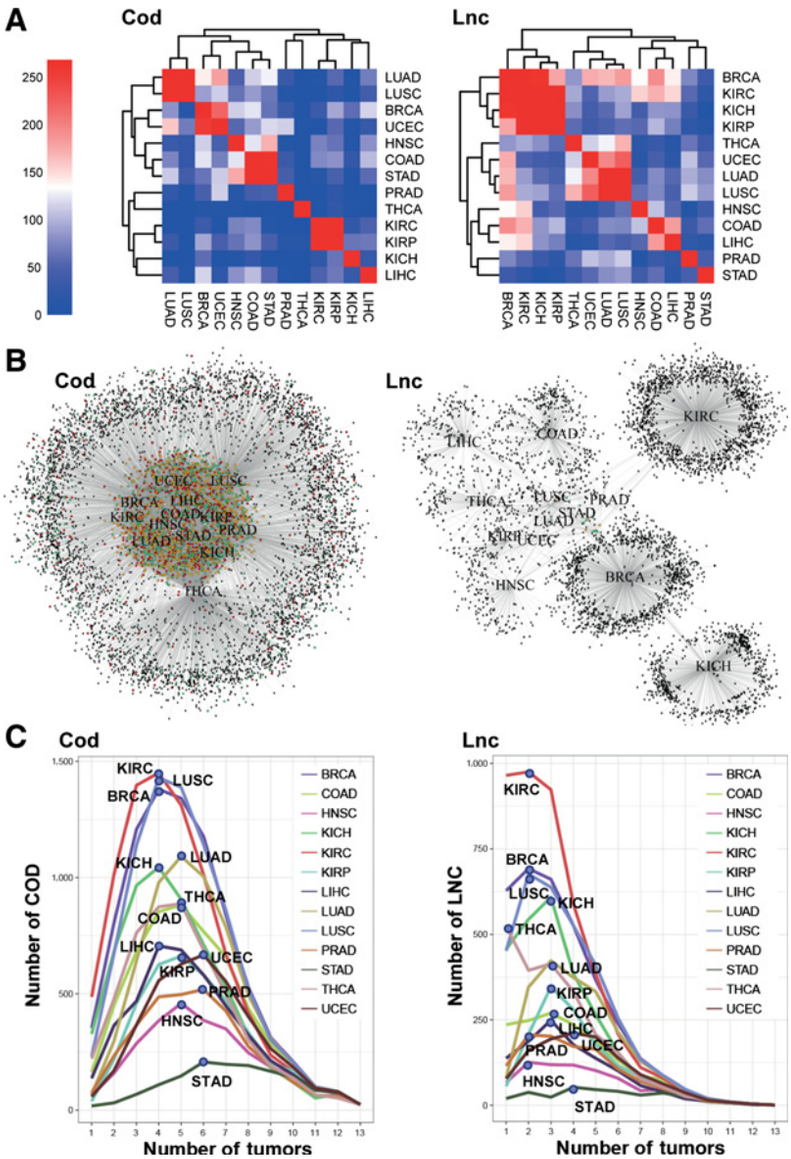


Figure 2.

Pan-cancer analysis of deregulated transcripts. **A,** Heatmap and hierarchical clustering of the pairwise comparisons of coding (left) and long noncoding (right) transcripts deregulated in each of the evaluated tumors. The color code indicates the value of the enrichment score calculated as $-\log_{10}$ (*P* value). The color gradient is indicated for each heatmap. **B,** Visualization of the filtered network of interactions between coding (left) and long noncoding (right) genes and tumor types. Only genes with a degree 1 or >10 are displayed. **C,** Number of recurrently deregulated coding (left) and long noncoding (right) genes in a single tumor that are only deregulated in one or more tumors, up to the 13 tumors examined. The peak of the curve has been highlighted for each tumor.

Fig. S2A). When we visualized the output of the analysis, we noted that in most clusters (over 50%), transcripts preferentially expressed in heart were also preferentially expressed in muscle tissues. This was also observed for blood and spleen and for all brain tissues, pituitary, and peripheral nerve (see a representative cluster in Supplementary Fig. S2B). Therefore, we combined these related tissues (Fig. 3A; Supplementary Table S3). On the cancer counterpart, we combined a number of tumors for similar reasons including those originated in the kidney (KICH, KIRC, KIRP), gastrointestinal tract (COAD, READ for rectum adenocarcinoma, STAD), hematologic tissues (LAML for acute myeloid leukemia, DLBC for diffuse large B-cell lymphoma, THYM for thymoma), neurologic tissues (LGG for lower grade glioma, GBM for glioblastoma multiforme, PCPG for pheochromocytoma and paraganglioma), and squamous cells of the lung and head and neck (LUSC, HNSC; Fig. 3A; Supplementary Table S3). We verified that statistical significance of the results was higher when the samples were grouped than when the original labels were used. More coding genes than long noncoding genes were defined as preferentially expressed in cancer. This may be due to the higher expression levels of coding genes and the restrictions imposed for the selection of preferentially expressed transcripts. The percentage of preferential lncRNA genes located close (<5 Kb) to preferential coding genes was very variable (0%–27%; Fig. 3A).

Cross-comparison of transcripts preferentially expressed in tissues and tumors showed that tumors were significantly similar to their tissue of origin (Fig. 3B and C). Nervous system tumors (LGG, GBM, and PCPG) were similar to brain, hematologic tumors (LAML, DLBC, and THYM) to spleen/blood, squamous cancers (LUSC and HNSC) to skin, and HCC to liver. These significant similarities were observed for preferentially expressed

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

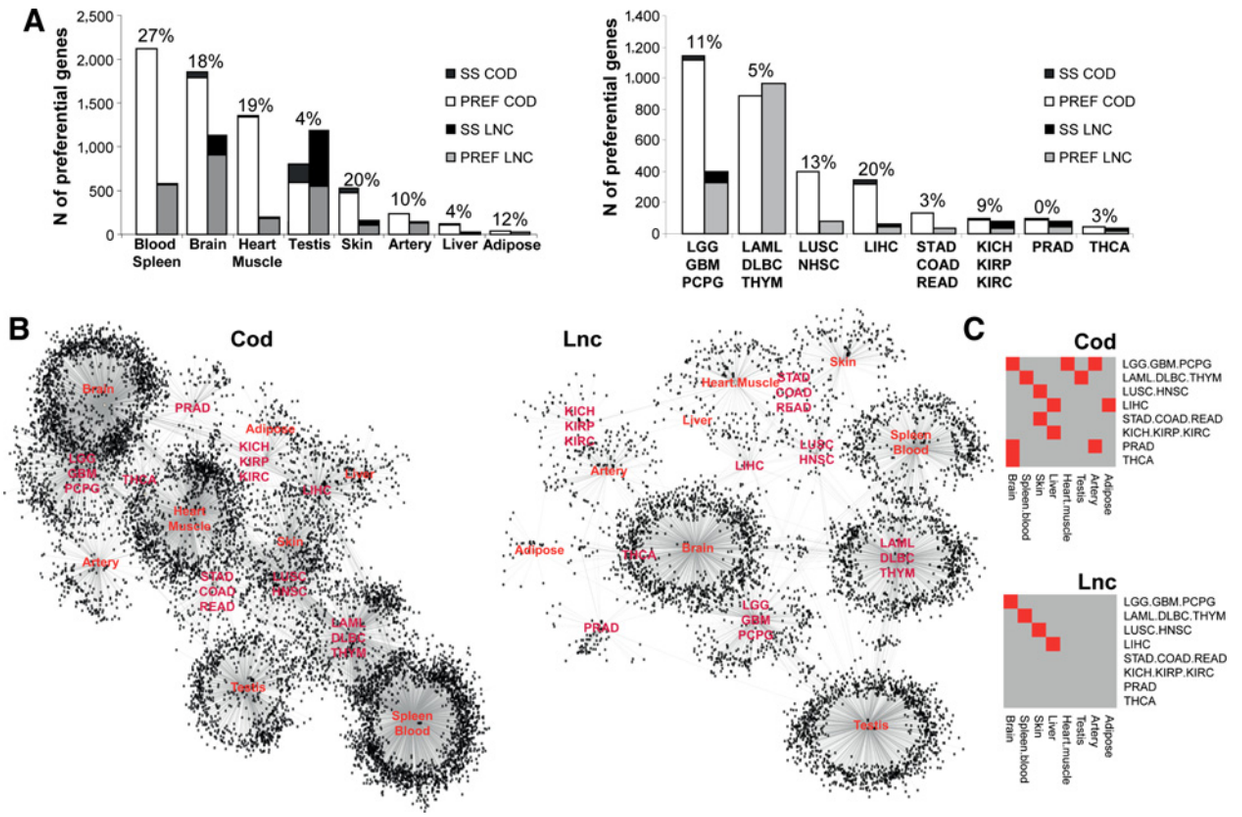


Figure 3.

Pan-cancer and pan-tissue analyses of specific and preferential expression. **A,** Number of super-specific (SS) or preferentially expressed (PREF) coding and noncoding genes in the indicated tissues (left) and tumors (right). For each condition, the percentage of preferentially expressed lncRNAs located close to the preferentially expressed coding genes versus the total number of preferentially expressed lncRNAs is indicated at the top of the bar. **B,** Network of interaction of preferentially expressed coding (left) and noncoding (right) genes in all tumors and tissues examined. **C,** Heatmap of the pairwise comparisons of coding (upregulation) and noncoding (downregulation) transcripts preferentially expressed in each of the evaluated tissues versus tumors. The color code indicates statistical significance of the enrichment analysis using the hypergeometric distribution. Red, corrected $P < 0.05$; gray, nonsignificant associations.

coding and long noncoding transcripts. However, preferential coding transcripts showed increased connectivity in the interaction network compared with lncRNAs and also significant associations between tumors and tissues of different origin.

Analysis of tissues that preferentially express upregulated tumor transcripts

All studies reported reveal that lncRNAs are more specific than coding transcripts. Therefore, we used lncRNAs to address whether some tissues preferentially express the transcripts that are upregulated in a particular tumor. To this aim, we combined the list of lncRNAs preferentially expressed in each tissue with the list of lncRNAs upregulated in each tumor. The hypergeometric enrichment analysis shows that most tumors significantly upregulate lncRNAs that are preferentially expressed in brain, spleen/blood, and testis (Fig. 4A). Given that neural networks do not allow the identification of all lncRNAs preferentially expressed in a given tissue, we repeated the analysis in a more stringent manner. We selected the 100 lncRNAs with the highest fold change in each of the 13 tumor types examined, and we interrogated GTEx to extract their tissue expression pattern. Then, we classified the 1,300 lncRNAs as tissue specific (when expression in other tissues was nondetected) or preferentially expressed (following the criteria described in Materials and Methods), promiscuous, nonannotated, or nondetected (lower than 0.2 RPKM according to GTEx data).

Similar to the previous analysis, related tissues were grouped for the analysis when they shared preferential expression in more than 50% of the cases. Expected associations were found around muscle (heart and peripheral muscle), digestive tract (colon, rectum, stomach, and esophagus), adipose (breast and adipose), female gynecologic organs (uterus, cervix, and fallopian), and nervous system (all brain tissues, pituitary, and peripheral nerve). Unexpected associations (in less than 10% of the cases) were found between testis and brain and between esophagus, skin, intestine, and vagina. Similar results were obtained when these unexpected associations were not considered in the analysis (selecting only one tissue of preferential expression) or when they were included (by selecting the top two tissues of preferential expression).

We observed that most lncRNAs highly upregulated in the combination of all tumors (TCGA) were preferentially expressed in testis, brain, spleen/blood, and the digestive tract (Fig. 4B; Supplementary Table S3: donor tissues). The results were similar when the preferential expression was analyzed in the top 100 or in the top 130 lncRNAs upregulated in each of the 13 tumors.

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

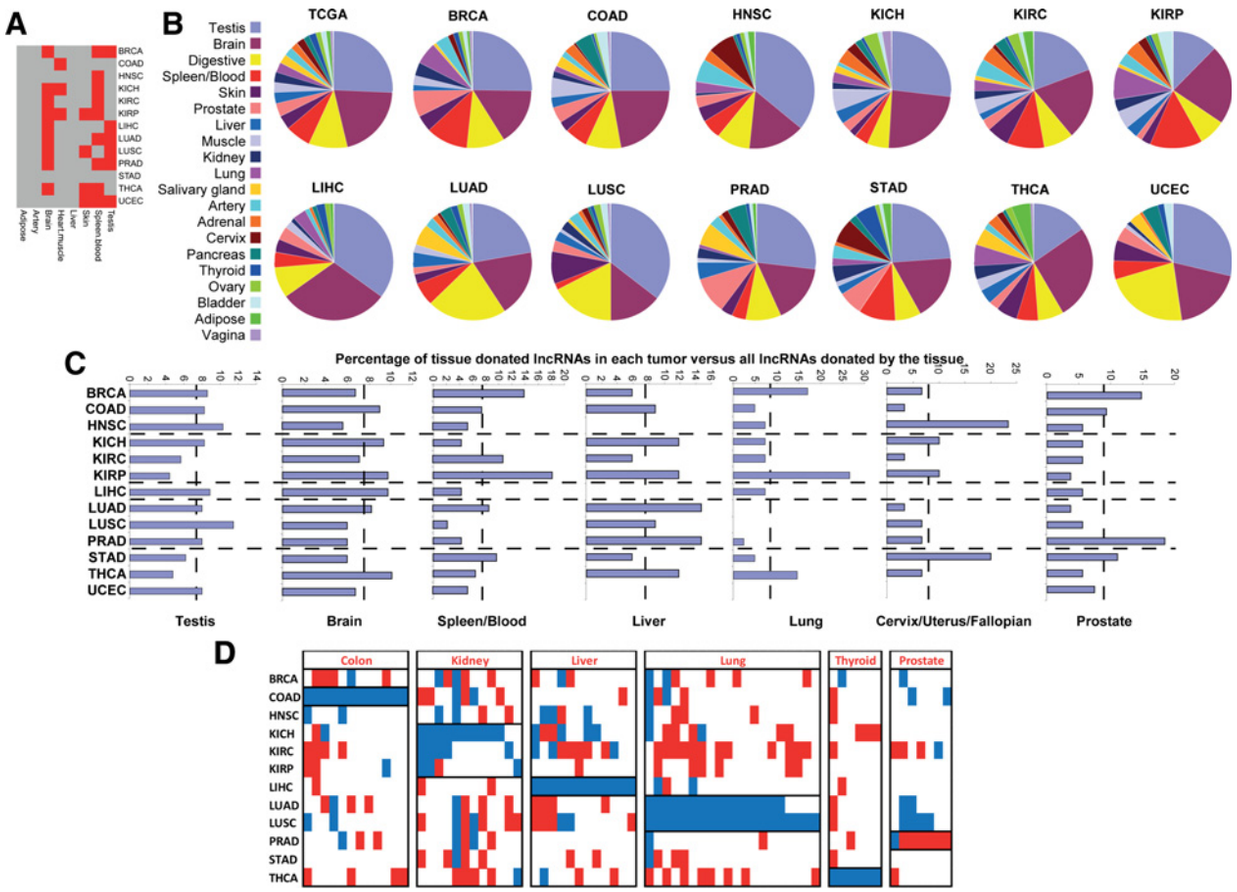


Figure 4.

Pan-cancer analysis of tissues that preferentially express upregulated tumor lncRNAs. **A,** Heatmap of the pairwise comparisons of lncRNAs upregulated in each of the evaluated tumors versus lncRNAs preferentially expressed in each of the selected tissues. The color code indicates statistical significance of the enrichment analysis using the hypergeometric distribution. Red, corrected $P < 0.05$; gray, nonsignificant associations. **B,** Pie chart showing the "lncRNA-donor-tissue" of the top 100 upregulated lncRNAs in all (TCGA) or in each of the tumors examined. **C,** Graphs showing the percentage of "tissue-donated lncRNAs" in each tumor versus all lncRNAs "donated" by the tissue. The average value is indicated by a discontinuous line. **D,** Pan-cancer deregulation analysis of lncRNAs preferentially expressed in the indicated tissue that are downregulated (COAD, pan-kidney tumors, LIHC, LUAD, LUSC, and THCA) or upregulated (only PRAD) in the tumor originated from that tissue. Each rectangle represents significant upregulation (red) or downregulation (blue) of a particular lncRNA in the indicated tumors.

Interestingly, the contribution of each "lncRNA-donor-tissue" is different for each tumor. Although the contribution of testis, brain, and the digestive tract is similar for all tumors, it represents almost 75% of the total in LIHC and less than half in KIRP or KIRC. Furthermore, lung "donated" more lncRNAs to KIRP, and cervix "donated" more to STAD and HSNC than they did to other tumors (Fig. 4B and C).

Overall, tumors downregulate lncRNAs preferentially expressed in their tissue of origin. LIHC, LUAD/LUSC, UCEC, or THCA do not upregulate liver, lung, cervix, or thyroid lncRNAs, respectively (Fig. 4C). This may be indicative of tumor dedifferentiation. The exception to this rule is PRAD, which upregulates a high proportion of transcripts preferentially expressed in prostate. Interestingly, lncRNAs that are preferentially expressed in prostate and upregulated in PRAD can be downregulated in other tumors (Fig. 4D). Similarly, lncRNAs that are downregulated in COAD, kidney tumors, LIHC, lung tumors, or THCA and preferentially expressed in their respective tissue of origin, can be upregulated in other tumors. All these candidates fall into the category of bimorphic lncRNAs.

Similar analyses with coding genes resulted in a different pattern of "transcript-donor tissue" for each tumor (Supplementary Fig. S3; Supplementary Table S3: donor tissues). Much like lncRNAs, most of the coding transcripts highly upregulated in the combination of the tumors (TCGA) are preferentially expressed in testis, brain, and the digestive tract (Supplementary Fig. S3A). Furthermore, in our analysis, we observed that tumors such as LIHC, LUAD/LUSC, THCA, or UCEC, do not upregulate mRNAs preferentially expressed in their tissue of origin (Supplementary Fig. S3B). IPA of the coding genes upregulated in a tumor and expressed preferentially in one tissue indicates that these genes associate highly significantly with different tumors (Supplementary Table S3: IPA). Interestingly, the collection of genes preferentially expressed in testis and spleen/blood is enriched in genes required for cell division, immune and inflammatory response, and cell migration (Supplementary Table S3: GO). Genes

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

preferentially expressed in brain, cervix, digestive tract, liver, and skin are enriched in genes related to development and differentiation, cell signaling, secretion, and metabolism.

LncRNAs upregulated in LIHC and preferentially expressed in testis or brain are associated with clinically relevant parameters

A very high proportion of lncRNAs found upregulated after analysis of LIHC data are preferentially expressed in testis or brain. We studied some of these candidates in more detail to determine whether they are associated with parameters related to hepatocarcinogenesis, tumor progression, or prognosis. We also studied other candidates to validate our list of deregulated lncRNAs. Among lncRNAs upregulated in LIHC, we selected two that were preferentially expressed in testis, two in brain, one in both, one in other tissue, one in several tissues, and two in none (Fig. 5). We also selected lncRNAs downregulated in LIHC that were and were not preferentially expressed in the liver. Of all 13 TCGA tumors examined, some lncRNAs were only deregulated in LIHC, others were upregulated or downregulated in other tumors, and some were bimorphic.

We studied whether the selected lncRNAs were expressed in 5 cell lines derived from HCC, adenocarcinoma and hepatoblastoma, one kidney tumor cell line (293T), one lung tumor cell line (A549), and fibroblasts (BJ). All the upregulated lncRNAs were highly expressed in one or more liver cancer cell lines (Supplementary Fig. S4A). Interestingly, some lncRNAs were poorly expressed in cell lines derived from lung and kidney tumors or in fibroblasts. With the exception of FAM99A, the downregulated lncRNAs studied were poorly expressed in all cell lines tested. Validation was also performed in patient samples. First, using TCGA data, we verified that all LIHC-upregulated and downregulated lncRNAs selected showed significantly increased and decreased expression, respectively, in all and in paired tumor samples compared to peritumoral tissue (Fig. 6A; Supplementary Fig. S4B). LIHC patients from the TCGA showed only rarely (25%) an underlying chronic liver disease (Supplementary Table S1). This is a strong bias because most HCCs develop in livers with chronic inflammation (44). Therefore, we evaluated the levels of the selected lncRNAs in HCCs and paired peritumoral tissue samples from a much more representative cohort (BCL-CUN cohort). In this cohort, 50 of 52 patients had steatosis, hepatitis, or cirrhosis. Healthy livers were used as controls. The expression of most deregulated lncRNA candidates was also significantly altered in this cohort (Fig. 5, Fig. 6B). The only exception was RP11-242J7.1, whose levels were not significantly elevated in tumor compared with peritumor tissue.

To further determine the importance of the selected lncRNAs in HCC, we studied their clinical associations. Using TCGA data and neoplasm histologic grade to classify, we observed that several lncRNAs are good prognostic markers (Fig. 7A). In addition, several of the selected lncRNAs associate with molecular classifications (Hoshida's and iCluster) and clinical parameters [vascular invasion, the macrotrabecular massive (MTM)

		Strat		Clinical					Genetic					
Name	Validation	1	2	1	2	3	4	5	1	2	3	Other TCGA	GTEX	Coexpression
LINC01419	0.0016											LIHC specific	Testis	Testis genes
HAGLR	<0.0001											COAD·HNSC·KICH·KIRP·LUSC·THCA	Brain and testis	AS HOXD1
LINC00176	0.0004											BRCA·HNSC·KIRC·LUAD·LUSC	Brain	ND
RP11-242J7.1	ns											LIHC specific	None	Tandem CASP3
RP5-1011O1.2	0.0009											COAD·HNSC·LUSC	Uterus	Testis genes
RP11-284F21.10	0.0018											HNSC·KICH·KIRC·KIRP·LUAD·LUSC	Several	AS BCAN
MAFG-AS1	0.0005											BRCA·COAD·KICH·KIRC·LUAD·LUSC	Testis	Chr17.q25.3
LINC00665	0.0002											BRCA·LUAD·PRAD	Brain	ZFP at Chr19.p12; q13.12 and q13.41
RP11-739N20.2	0.0430											HNSC·KICH·LUAD	None	Kinesins and centromer proteins
LINC01093	<0.0001											BRCA·LUSC	Liver	ACSL1 neighbour and liver genes
PLGLA	0.0111											KICH·KIRC·KIRP	Liver	PLG parental and liver genes
RP11-421P23.1	0.0071											LIHC specific	Liver and testis	ND
FAM99A	0.0051											LIHC specific	Liver	Liver genes
RP4-564F22.6	0.0065											LIHC specific	NA	Tandem LBP
RP11-353N14.5	0.0001											COAD·LUAD	Brain and testis	ND

0.04 0.03 0.02 0.01 0

Figure 5.

Analysis of lncRNAs deregulated in LIHC. The name of the lncRNA and the statistical analysis (*P* value) of the validation of the lncRNA expression on the BCL-CUN cohort of HCC samples are provided (validation). Significant associations with clinical data are also indicated. Patient stratification: 1, Hoshida; 2, i-cluster. Clinical data: 1, vascular invasion; 2, MTM; 3, α-fetoprotein levels; 4, pathologic stage; 5, neoplasm histologic grade. Genetic associations: 1, *TERT* expression; 2, *TP53* mutations; 3, *CTNNB1* mutations. Red gradient from lowest to highest indicates significance. "Other TCGA" describes tumors different than LIHC, in which the expression of the lncRNA is significantly deregulated. Upregulation, red; downregulation, blue. "LIHC specific" denotes that significant deregulation in other tumors have not been found. GTEx describes the tissue of preferential expression for each lncRNA. "Several" denotes that more than two tissues express the lncRNA to similar levels. "Brain and testis" or "liver and testis" indicate that these tissues express the lncRNA to higher levels than the rest. "None" indicates that there is no significant expression in any of the tissues evaluated by GTEx. "NA" indicates that the lncRNA has not been analyzed in the GTEx database. Coexpression shows characteristics of genes highly coexpressed with each lncRNA. Testis or liver genes are genes with preferential expression in the indicated tissues. AS and tandem denotes that the lncRNA is coexpressed with the neighboring gene located in antisense orientation or in tandem, respectively. ND, none determined. *MAFG-AS1* is located at chr17 q25.3 and its expression correlates with that of several genes located in the same region. *LINC00665* is located at chr19 q13.12 and its expression correlates with that of several zinc finger genes located in the same region or in chr19 p12 or q13.41. *PLGLA* is a pseudogene and correlates with the expression of the *PLG* parental gene.

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

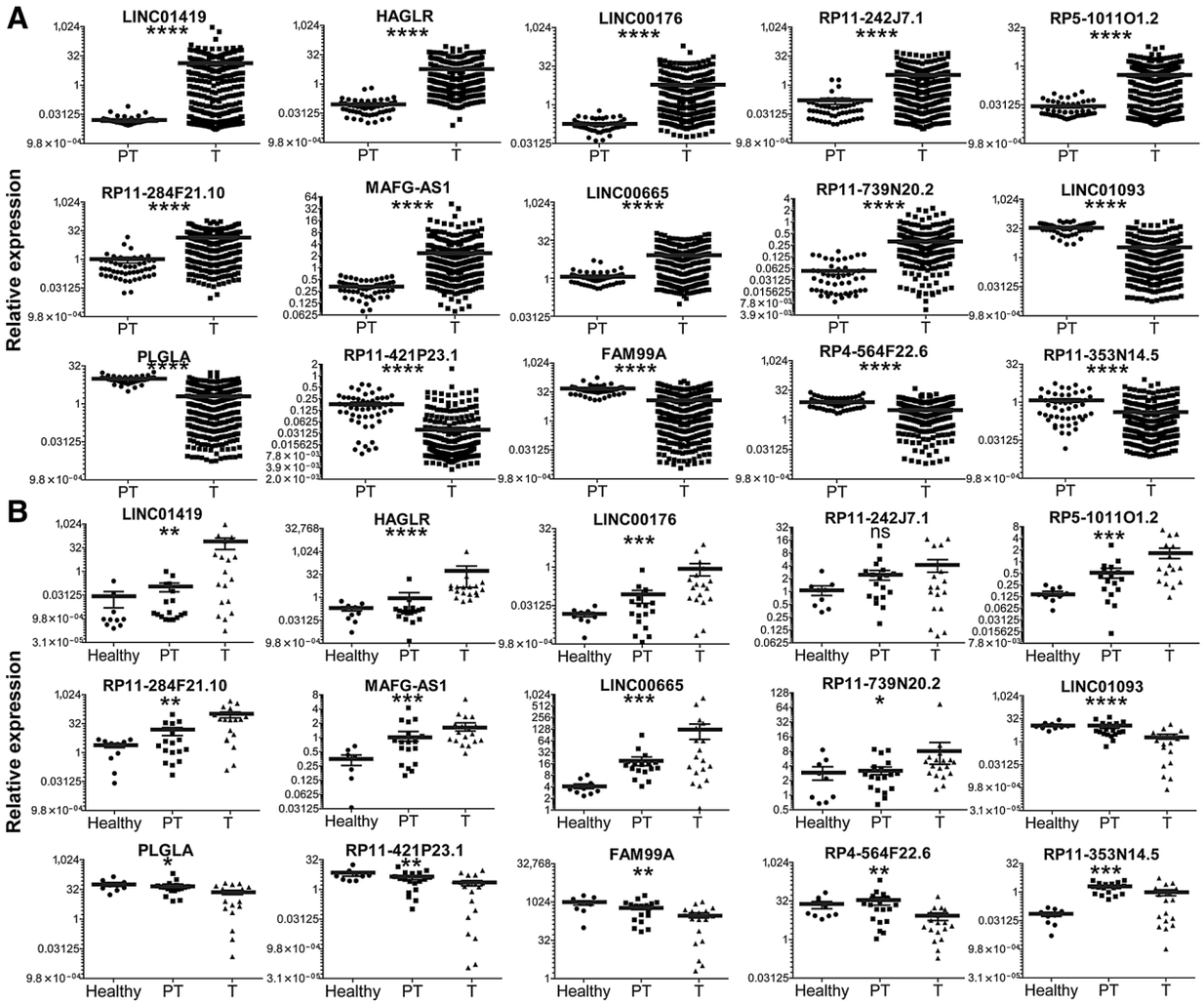


Figure 6.

Relative expression of several lncRNAs deregulated in LIHC and liver tumors. **A,** The relative expression of each lncRNA was evaluated in all peritumoral (PT) and tumor (T) samples from the TCGA RNA-seq data. **B,** RNA was isolated from healthy livers and paired peritumoral and tumor samples from a cohort of HCC patients and the expression of the indicated lncRNAs was quantified by qRT-PCR. The levels of RPLP0 mRNA were also evaluated and used as a reference to calculate the relative expression. The average is indicated as a line. The result of the statistical analysis (two-tailed unpaired *t* test in **A** and Kruskal–Wallis ANOVA test in **B**) is indicated in each graph. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P \leq 0.0001$; ns, nonsignificant differences.

histologic subtype, α-fetoprotein levels, pathologic stage, and histologic grade; Fig. 5; Supplementary Table S2: clinical LIHC; refs. 6, 25, 45). Similarly, significant associations were found with *TERT* expression and mutations in *TP53* and *CTNNB1*, well-known drivers of cancer progression. Interestingly, when our BCL-CUN cohort of patients was classified according to prognosis, patients with the best prognosis showed significantly lower levels of RP11-242J7.1 in the tumor versus peritumor tissue, whereas the opposite scenario was observed in patients with worse prognosis (Supplementary Fig. S4C). This indicates that this lncRNA could be a relevant target even if its levels were not significantly altered when all tumors were evaluated as a single group. Further analysis of the BCL-CUN cohort showed that the expression of some lncRNAs is significantly different in patients with different tumor differentiation and tumor burden (size or number of nodules; Supplementary Fig. S4D). Similar results were obtained with the TCGA cohort. The expression of some lncRNAs is significantly different in patients with different Hoshida classification, clinical parameters (vascular invasion, MTM), *TERT* expression, and with or without mutations in *TP53* or *CTNNB1* (Fig. 7B). In these analyses, we observed that the selected lncRNAs upregulated in HCC were significantly more expressed in samples from patients with a more aggressive clinical phenotype (larger tumors with poor differentiation, MTM subtype, vascular invasion, and with mutations in p53). Conversely, selected lncRNAs downregulated in HCC were significantly more expressed in patients with a less aggressive phenotype and better prognosis. To determine whether this is a general feature, we searched for significant differences in the expression of all the lncRNAs identified as deregulated in LIHC in patients with different variables related to prognosis (vascular invasion, high α-fetoprotein levels, MTM, worse

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

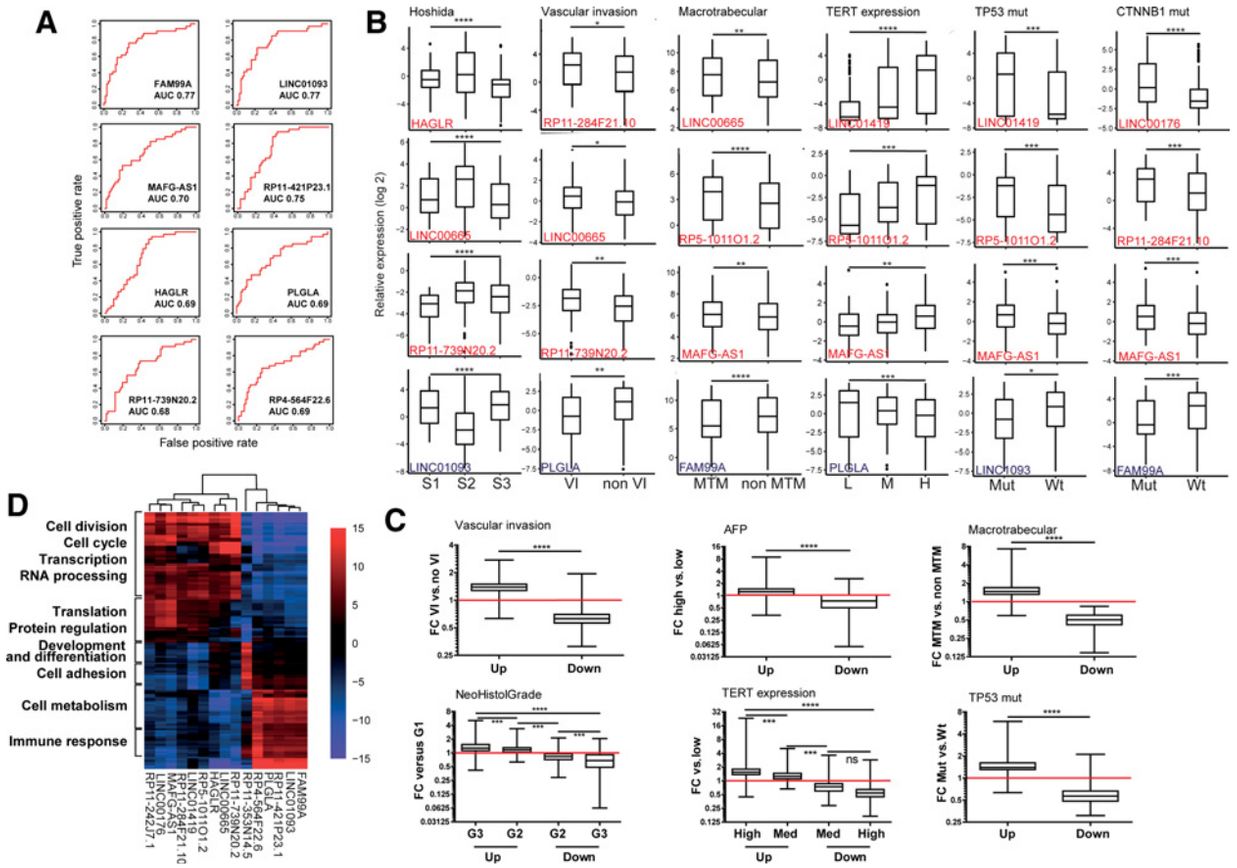


Figure 7.

Analysis of clinical associations and functional predictions of lncRNAs deregulated in LIHC. **A,** ROC curves for lncRNAs that are good prognostic indicators. Neoplasm histologic grade was used to classify. **B,** Box plots of lncRNAs whose levels are significantly different in patients with Hoshida's classification S1 to S3; detectable or undetectable vascular invasion (VI) or MTM pattern; low (L, <2 units), medium (M), or high (H, >20 units) *TERT* expression and wild-type or mutant *TP53* or *CTNNB1* according to TCGA data. **C,** Upregulated lncRNAs are expressed to higher levels in patients with worse prognosis. Fold change of the average expression in patients with worse prognosis versus better prognosis was calculated for each lncRNA whose levels were significantly different in patients with the indicated features. Up, upregulated; down, downregulated. Fold change 1 (no difference) is highlighted with a red line. AFP, α-fetoprotein. Neoplasm histologic grade (NeoHistGrade). The results of the statistical analyses (*t* test for pairwise comparisons or ANOVA, followed by multiple comparison test) are indicated. **D,** GBA analysis of several lncRNAs deregulated in LIHC. Red, positive *Z*-score; blue, negative *Z*-score. Only functions with *Z*-scores higher than 10 were evaluated. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P \leq 0.0001$.

histologic grade, *TERT* expression, and mutations in *TP53*). Then, the lncRNAs identified were divided into upregulated or downregulated and the fold change was calculated as the average expression of each lncRNA in all patients with worse prognosis versus those with better prognosis. Statistical comparisons between upregulated and downregulated lncRNAs show that most upregulated lncRNAs express higher levels (fold change higher than 1) in patients with worse prognosis while most downregulated lncRNAs express lower levels in those patients (Fig. 7C). Similar results were found for deregulated coding genes, suggesting that the deregulated genes identified in our study could be excellent therapeutic targets.

Finally, to predict the function of the chosen candidates, we performed a GBA analysis. The results indicate that those candidates upregulated in LIHC correlated positively with cell cycle, cell division, and gene expression and negatively with development and differentiation, adhesion, metabolism, and immune response (Fig. 7D). Downregulated candidates showed the opposite pattern. GSEA showed similar results (Supplementary Fig. S5A and S5B). Interestingly, expression of testis or liver lncRNAs such as LINC01419, LINC01093, PLGLA, or FAM99A, correlates with expression of genes preferentially or specifically expressed in testis or liver, respectively (Fig. 5). As expected, the expression of several lncRNAs highly correlated with the levels of neighbor genes located antisense, in tandem or in the same chromosomal region. The expression of RP11-739N20.2 highly correlated with the expression of several centromere proteins, whereas the levels of LINC00665 highly correlated with those of zinc finger proteins located in three clusters in chromosome 19 (Supplementary Fig. S5C).

Discussion

The analyses of the massive amount of data generated by the TCGA are helping to better shape our understanding of how cancer arises, develops, and evolves in the patient, with

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

the hope of making an impact in the clinical setting in the near future (10). To achieve this goal, further analyses using multifocus and multidisciplinary approaches are needed. Most published studies have focused on the analysis of paired samples (14). This is far from ideal because transcriptomes of peritumoral tissue and healthy tissue are different (46). In addition, most tissues are complex mixtures of different kinds of cells, while tumors arise from one cell type within that mixture. However, analyses of paired samples have successfully identified key oncogenes and tumor suppressors. It should be noted that in the analyses of paired samples, the data obtained from many tumors that do not have a matching peritumoral tissue is not considered. To avoid this problem, we hypothesized that because peritumoral tissues tend to be homogeneous among patients with similar diagnosis (46), data from peritumoral samples could be used to compare matching and nonmatching tumor samples. Then, we identified those transcripts recurrently deregulated after 200 comparisons between a randomized pool of tumor transcriptomes and all peritumoral samples. As expected, the identified deregulated coding genes were enriched for those involved in cancer hallmarks. Among lncRNAs, we identified several with known oncogenic and tumor suppressor functions (20, 39–42, 47). In addition, deregulation of all randomly selected lncRNAs deregulated in HCC was validated in another cohort (Fig. 6). Furthermore, several of these candidates associate significantly with clinical and genetic parameters indicating their prognostic value (Fig. 5; Fig. 7, Supplementary Fig S4D; Supplementary Table S2). Finally, we observe with high significance that those genes upregulated in HCC are expressed to higher levels in patients with worse prognosis (Fig. 5; Fig. 7; Supplementary Fig S4D; Supplementary Table S2). All these results suggest that the deregulated lncRNAs identified could play a role in HCC development or progression.

In the TCGA analyses, we and others have observed that most deregulated transcripts are downregulated and that the absolute fold change of downregulated genes is generally higher than that of upregulated genes (Fig. 1B; Supplementary Fig. S1; refs. 11, 14). We were surprised to observe the opposite results when we attempted to validate some of the HCC deregulated genes obtained from the LIHC samples from the TCGA in an independent cohort of HCC patients (Fig. 6B). Similarly, when CAGE data from RIKEN are used to search for cancer-deregulated genes, a lower proportion of genes are downregulated than when TCGA data are analyzed (11). Further analysis of TCGA and other datasets will be required to clarify these discrepancies. This is important because upregulated genes, susceptible for antisense drug targeting, may have higher therapeutic potential than downregulated genes.

Pan-cancer analysis of TCGA data results in cancer classification according to the cell of origin (48, 49). It has been reasoned that tumor transcriptomes are very similar to those of their tissue of origin because deregulated tumor transcripts are only a few among the thousands of transcripts that define a specific tissue (48). Interestingly, we also observed that tumors aggregate according to their cell of origin when deregulated genes are evaluated (Fig. 2A). This would suggest that similar genes are deregulated in any process of tumorigenesis from a particular tissue (from kidney to KIRC/KIRP or from lung to LUAD/LUSC) or from two related tissues (from colon and stomach to COAD/STAD). In addition, genes specifically or preferentially expressed in a tumor are significantly similar to those expressed in their tissue of origin (Fig. 3B and C), suggesting that these genes could serve as a signature that it is not substantially lost with tumorigenesis.

Our pan-cancer analysis has also helped to compare general features of deregulated coding and noncoding transcripts. Compared with lncRNAs, recurrently deregulated coding transcripts are more numerous (Fig. 1A), more promiscuous (Fig. 2B and C), and those deregulated in most tumors are upregulated (Fig. 1C). In turn, lncRNAs are more tumor-specific and cell-type specific (Fig. 2B and C; Fig. 3). These facts have been described previously (14, 50, 51) and support the idea that many cancers have distinct lncRNA signatures (19). Then, we have addressed what is the tissue of preferential expression of lncRNAs upregulated in cancer using GTEx data, which have been previously analyzed jointly with TCGA data with a different aim (46). Although the tissues analyzed by GTEx cannot be considered, strictly speaking, as healthy, they are nontumoral tissues and they are a magnificent source of data. As expected, we found that most tumors do not upregulate transcripts that are preferentially expressed in their tissue of origin (Fig. 4C and D). Interestingly, some transcripts that are preferentially expressed in colon, kidney, liver, lung, and thyroid are downregulated in tumors derived from those tissues, whereas upregulated in other tumors (Fig. 4D). Tumor dedifferentiation may be behind this differential deregulation and could be one of the reasons why some transcripts show bimorphic features, being upregulated in some tumors and downregulated in others (Fig. 1C). Unexpectedly, PRAD upregulates lncRNAs preferentially expressed in prostate. The reason for this is unclear and we cannot exclude an artifact. However, most GTEx data were obtained from males older than 50 years (68%) and prostate hyperplasia increases with age. Therefore, prostate transcriptome from GTEx may be enriched in growth-related transcripts that increase in the process of tumorigenesis.

The best-represented cancer lncRNA "donor-tissues" are testis, brain, and spleen/blood. Our hypergeometric tests show that this reaches statistical significance even when the organs that have more preferentially expressed genes are the best "donors". In addition, the digestive tract is a good "donor" that does not show much preferential expression in our neural network analysis (Fig. 3) and the best cancer "donor-tissues" are similar when coding transcripts and lncRNAs are evaluated. Therefore, there is not a perfect correlation between the number of mRNAs preferentially expressed in a tissue and the number of "donated" mRNAs. We hypothesize that tumors may have their preferred "donor-tissues" in those who express genes that benefit tumor growth. In fact, the "donated" coding genes are enriched in those involved in cell division, immune, and inflammatory response, cell migration, signaling and secretion, development, and differentiation (Supplementary Table S3: GO). All these are features that could contribute to cancer growth. We speculate that "donated" lncRNAs could have similar functions and potentiate tumor development. In fact, individual lncRNAs "donated" from testis or brain to HCC are associated with relevant clinical features and their putative functions are related to cell growth and gene expression (Figs. 5 and 7). We were truly surprised to find that the brain, considered a paradigm of quiescent tissue, could "donate" features of proliferation. However, it has been described that there is a neural specific expression of pan-cancer–

promoting genes as cancer cells and nondifferentiated neuronal cells share regulatory networks mediating tumorigenesis and neural development (52).

Our results suggest that some transcripts that are preferentially expressed in certain healthy tissues can be upregulated in certain tumors and they may be interesting therapeutic targets. Therapies developed to target cancer-upregulated genes with oncogenic potential should consider the tissue of preferential expression of these targets to anticipate unwanted on-target off-organ side effects. In line with this, we propose that testis-specific oncogenic lncRNAs would be ideal targets for therapeutic intervention.

Disclosure of Potential Conflicts of Interest

J. Bruix has received speakers bureau honoraria from Bayer and Ipsen, and is a consultant/advisory board member for Bayer, BMS, KOWA, Gilead, Abbvie, ASTRA, Sirtex, Nerviano, Lilly, Sanofi, Incyte, Adaptimmune, MSD, Terumo, BTG, Basilea, Novartis, Roche, EISAI, and BIO-Alliance. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: L. Boix, J. Bruix, B. Sangro, V. Segura, P. Fortes
Development of methodology: G. Serrano, B. Suárez, V. Ferretti, C. Prior, V. Segura, P. Fortes
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): J.P. Unfried, P. Sangro, V. Ferretti, L. Boix, J. Bruix, B. Sangro
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J.P. Unfried, G. Serrano, P. Sangro, C. Prior, L. Boix, J. Bruix, V. Segura, P. Fortes
Writing, review, and/or revision of the manuscript: J.P. Unfried, G. Serrano, B. Suárez, L. Boix, J. Bruix, B. Sangro, V. Segura, P. Fortes
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): G. Serrano, B. Suárez, V. Ferretti, C. Prior, P. Fortes
Study supervision: V. Segura, P. Fortes

Acknowledgments

We particularly acknowledge the patients for their participation and the Biobank of the University of Navarra for its collaboration. We thank Nerea Razquin for excellent technical assistance. The results shown here are in part based upon data generated by the TCGA Research Network, http://cancergenome.nih.gov/, and by the GTEx Project, https://gtexportal.org/. This work was supported by European FEDER funding and grants from the Ministry of Economy (SAF2015-70971-R, DPI2015-68982-R); MCIU/AEI/FEDER (UE/RTI2018-101759-B-I00), Gobierno de Navarra (DIANA: 0011-1411-2017-000029, 33/2015); Fundación Echevano; Fundación Unicaja; Scientific Foundation of the Spanish Association Against Cancer (AECC PI044031); the Worldwide Cancer Research Foundation (16-0026); and Fondo de Investigación Sanitaria (PI16/01845, PI14/00962, PI18/00763), financed by the Instituto de Salud Carlos III. J.P. Unfried is a recipient of a University of Navarra's Asociación de Amigos fellowship and BSu of a Gobierno de Navarra's 0011-1408-2017-000008 fellowship. The Bioinformatics unit of CIMA is member of the ProteoRed-ISCIII platform. CIBERehd is funded by the Instituto de Salud Carlos III. The GTEx project was supported by the NIH and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS.

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 31, 2019; revised June 6, 2019; accepted August 2, 2019; published first August 6, 2019.

References

1. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature 2010;464:993–8.
2. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013;45:1113–20.
3. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, et al. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol 2015;16:1–14.
4. Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. Integrated genomic characterization of endometrial carcinoma. Nature 2013;497:67–73.
5. Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. Cell 2015;163:1011–25.
6. Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell 2017;169:1327–41.
7. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic signaling pathways in the cancer genome atlas. Cell 2018;173:321–37.
8. Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, et al. Pan-cancer patterns of somatic copy number alteration. Nat Genet 2013;45:1134–40.
9. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013;502:333–9.
10. Rubio-Perez C, Tamborero D, Schroeder MP, Antolín AA, Deu-Pons J, Perez-Llamas C, et al. In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. Cancer Cell 2015;27:382–96.
11. Kaczkowski B, Tanaka Y, Kawaji H, Sandelin A, Andersson R, Itoh M, et al. Transcriptome analysis of recurrently deregulated genes across multiple cancers identifies new pan-cancer biomarkers. Cancer Res 2016;76:216–26.
12. Witte T, Plass C, Gerhauser C. Pan-cancer patterns of DNA methylation. Genome Med 2014;6:66.
13. Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I. Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types. Nucleic Acids Res 2017;45:2973–85.
14. Cabanski CR, White NM, Dang HX, Silva-Fisher JM, Rauck CE, Cicka D, et al. Pan-cancer transcriptome analysis reveals long noncoding RNAs with conserved function. RNA Biol 2015;12:628–42.
15. Wang Z, Yang B, Zhang M, Guo W, Wu Z, Wang Y, et al. lncRNA epigenetic landscape analysis identifies EPIC1 as an oncogenic lncRNA that interacts with MYC and promotes cell-cycle progression in cancer. Cancer Cell 2018;33:706–720.
16. Chiu HS, Somvanshi S, Patel E, Chen TW, Singh VP, Zorman B, et al. Pan-cancer analysis of lncRNA Regulation supports their targeting of cancer genes in each tumor context. Cell Rep 2018;23:297–312.
17. 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.
18. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet 2015;17:47–62.
19. 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.
20. Tseng YY, Moriarity BS, Gong W, Akiyama R, Tiwari A, Kawakami H, et al. PVT1 dependence in cancer with MYC copy-number increase. Nature 2014;512:82–6.
21. 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.

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

22. Braconi C, Kogure T, Valeri N, Huang N, Nuovo G, Costinean S, et al. microRNA-29 can regulate expression of the long non-coding RNA gene MEG3 in hepatocellular cancer. Oncogene 2011;30:4750–6.
23. Han L, Zhang E, Yin D, Kong R, Xu T, Chen W, et al. Low expression of long noncoding RNA PANDAR predicts a poor prognosis of non-small cell lung cancer and affects cell apoptosis by regulating Bcl-2. Cell Death Dis 2015;6: e1665.
24. 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.
25. Calderaro J, Couchy G, Imbeaud S, Amaddeo G, Letouzé E, Blanc JF, et al. Histological subtypes of hepatocellular carcinoma are related to gene mutations and molecular tumour classification. J Hepatol 2017;67: 727–38.
26. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3–new capabilities and interfaces. Nucleic Acids Res 2012;40:e115.
27. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet 2013;45:580–5.
28. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012;40:4288–97.
29. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 2015;43:e47.
30. Kohonen T. Self-organized formation of topologically correct feature maps. Biol Cybern 1982;43:59–69.
31. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011;25:1915–27.
32. Draghici S, Lenhart S, Safer H, Maini P, Etheridge A, Gross L. Data analysis tools for DNA microarrays. 1st ed. New York, NY: Chapman and Hall/CRC.; 2003.
33. Perez-Llamas C, Lopez-Bigas N. Gitools: analysis and visualisation of genomic data using interactive heat-maps. Aerts S, editor. PLoS One 2011;6:e19541.
34. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 1995;57: 289–300.
35. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102:15545–50.
36. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5:R80.
37. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. New York, NY: Springer; 2009.
38. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: Visualizing classifier performance in R. Bioinformatics 2005;21:3940–1.
39. Mourtada-Maarabouni M, Pickard MR, Hedge VL, Farzaneh F, Williams GT. GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. Oncogene 2009;28:195–208.
40. Chen JF, Wu P, Xia R, Yang J, Huo XY, Gu DY, et al. STAT3-induced lncRNA HAGLROS overexpression contributes to the malignant progression of gastric cancer cells via mTOR signal-mediated inhibition of autophagy. Mol Cancer 2018;17:6.
41. Grote P, Wittler L, Hendrix D, Koch F, Währisch S, Beisaw A, et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Dev Cell 2013;24:206–14.
42. Yao J, Zhou B, Zhang J, Geng P, Liu K, Zhu Y, et al. A new tumor suppressor LncRNA ADAMTS9-AS2 is regulated by DNMT1 and inhibits migration of glioma cells. Tumour Biol 2014;35:7935–44.
43. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012;22:1775–89.
44. Zucman-Rossi J, Villanueva A, Nault J-C, Llovet JM. Genetic landscape and biomarkers of hepatocellular carcinoma. Gastroenterology 2015;149: 1226–39.
45. Chiang DY, Villanueva A, Hoshida Y, Peix J, Newell P, Minguez B, et al. Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. Cancer Res 2008;68:6779–88.
46. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun 2017;8:1077.
47. Shen Y, Katsaros D, Loo LWM, Hernandez BY, Chong C, Canuto EM, et al. Prognostic and predictive values of long non-coding RNA LINC00472 in breast cancer. Oncotarget 2015;6:8579–92.
48. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 2014;158:929–44.
49. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 2018;173:291–304.
50. Werner MS, Sullivan MA, Shah RN, Nadadur RD, Grzybowski AT, Galat V, et al. Chromatin-enriched lncRNAs can act as cell-type specific activators of proximal gene transcription. Nat Struct Mol Biol 2017; 24:596–603.
51. Casero D, Sandoval S, Seet CS, Scholes J, Zhu Y, Ha VL, et al. Long noncoding RNA profiling of human lymphoid progenitor cells reveals transcriptional divergence of B cell and T cell lineages. Nat Immunol 2015;16: 1282–91.
52. Zhang Z, Lei A, Xu L, Chen L, Chen Y, Zhang X, et al. Similarity in gene-regulatory networks suggests that cancer cells share characteristics of embryonic neural cells. J Biol Chem 2017;292:12842–59.

Published OnlineFirst August 6, 2019; DOI: 10.1158/0008-5472.CAN-19-0400

Cancer Research
The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

AACR American Association for Cancer Research

Identification of Coding and Long Noncoding RNAs Differentially Expressed in Tumors and Preferentially Expressed in Healthy Tissues

Juan P. Unfried, Guillermo Serrano, Beatriz Suárez, et al.

Cancer Res 2019;79:5167-5180. Published OnlineFirst August 6, 2019.