Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology) 2023 24(1):15-31
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.B2200206

Check for updates

Spatial transcriptome analysis of long non-coding RNAs reveals tissue specificity and functional roles in cancer

Kang XU[1], Xiyun JIN[1], Ya LUO[1], Haozhe ZOU[1], Dezhong LV[1], Liping WANG[1], Limei FU[1], Yangyang CAI[1], Tingting SHAO[1], Yongsheng LI[2]✉, Juan XU[1]✉

[1]*College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China*
[2]*Key Laboratory of Tropical Translational Medicine of Ministry of Education, College of Biomedical Information and Engineering, Hainan Women and Children's Medical Center, Hainan Medical University, Haikou 571199, China*

Abstract: Long non-coding RNAs (lncRNAs) play a significant role in maintaining tissue morphology and functions, and their precise regulatory effectiveness is closely related to expression patterns. However, the spatial expression patterns of lncRNAs in humans are poorly characterized. Here, we constructed five comprehensive transcriptomic atlases of human lncRNAs covering thousands of major tissue samples in normal and disease states. The lncRNA transcriptomes exhibited high consistency within the same tissues across resources, and even higher complexity in specialized tissues. Tissue-elevated (TE) lncRNAs were identified in each resource and robust TE lncRNAs were refined by integrative analysis. We detected 1 to 4684 robust TE lncRNAs across tissues; the highest number was in testis tissue, followed by brain tissue. Functional analyses of TE lncRNAs indicated important roles in corresponding tissue-related pathways. Moreover, we found that the expression features of robust TE lncRNAs made them be effective biomarkers to distinguish tissues; TE lncRNAs also tended to be associated with cancer, and exhibited differential expression or were correlated with patient survival. In summary, spatial classification of lncRNAs is the starting point for elucidating the function of lncRNAs in both maintenance of tissue morphology and progress of tissue-constricted diseases.

Key words: Spatial expression; Transcriptome; Long non-coding RNA; Cancer; Biomarker

1 Introduction

Long non-coding RNAs (lncRNAs), a class of non-coding RNAs in the human genome, control various crucial biological functions (Xu et al., 2020). As a large group of RNA regulators, their precise regulatory effectiveness is closely related to the tissue context (Lv et al., 2020), as well as their expression patterns in tissues under both healthy and diseased states. Moreover, spatial characterization of lncRNA expression across tissues is important, because tissue-constricted lncRNA expression and pathway regulation underlie human physiology, and their dysfunction often results in disease. These factors combined to make the understanding of the tissue context of lncRNA functions, disease pathophysiology, and lncRNA–disease associations particularly challenging.

Identifying tissue-elevated (TE) lncRNAs across tissues is the first step in exploring their spatial expression patterns. Indeed, apart from a small number of genes only expressed in one tissue, more and more studies have observed an interesting group of protein-coding genes with TE expression in certain tissues or groups, which are either not expressed or expressed at lower levels in other tissues (She et al., 2009; Burgess, 2015; Uhlén et al., 2016; Aran et al., 2017; Wang et al., 2019). Thus, TE genes, including tissue-specific ones, can serve as biomarkers of specific biological processes or tissues in which they are expressed. To obtain a more comprehensive picture of the functions of lncRNAs, a multiple-tissue analysis is usually performed to look at their expression in a panel of tissues or organs. Indeed, it has been shown

✉ Juan XU, xujuanbiocc@ems.hrbmu.edu.cn
Yongsheng LI, liyongsheng@hainmc.edu.cn
Juan XU, https://orcid.org/0000-0002-3709-4165
Yongsheng LI, https://orcid.org/0000-0003-1914-0727

Received Apr. 8, 2022; Revision accepted July 13, 2022;
Crosschecked Jan. 11, 2023

© Zhejiang University Press 2023

that disease genes, including lncRNAs, generally tend to be expressed in a limited number of tissues (Li et al., 2020). However, the spatial expression pattern of lncRNAs is still not comprehensively characterized across human tissues and cancers.

Over the years, quantitative transcriptomic technologies such as RNA-sequencing (RNA-seq) and cap analysis of gene expression (CAGE) have been used to build genome-wide expression drafts of human genes, including both lncRNAs and protein-coding genes. The spatial patterns of protein-coding genes have been studied (Rawal et al., 2021). However, for lncRNAs, broader scale analyses had mostly focused on a single tissue or cell-type-resolved analysis. To the best of our knowledge, no broad-scale quantitative or integrative analysis of available lncRNA transcriptomes across healthy human tissues has been performed that would enable a comprehensive investigation of the spatial expression patterns of lncRNAs. Fortunately, several recent efforts to build large-scale expression profiles in samples representing most of the major human organs and tissues have been published. Well-accepted transcriptome studies on human tissues have been published by the Genotype-Tissue Expression (GTEx) consortium (Uhlén et al., 2015), the Human BodyMap 2 (HBM2) (Schroth, 2011), the Human Protein Atlas (HPA) consortium (Uhlen et al., 2010), the Functional Annotation of the Mammalian Genome (FANTOM) consortium (Hon et al., 2017), and The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research Network et al., 2013), collectively comprising thousands of samples from all major human tissues. The former three were profiled from the normal physiological state in humans; in particular, the GTEx generated normal transcriptome with the greatest number of samples pooled from 30 healthy human tissues. On the other hand, TCGA provides a comprehensive transcriptome for various cancers. Four (GTEx, HBM2, HPA, TCGA) of these five expression profiles were derived from RNA-seq assemblies, and the last one (FANTOM) was derived from CAGE. Genes with accurate 5'-ends were identified. These five expression profiles are notable for representing independent surveys of human genes by distinct methods (RNA-seq vs. CAGE), in different tissues under distinct conditions (normal vs. disease). Therefore, these five large-scale expression profiles provide an opportunity to integrate data sets and further refine the spatial classification of lncRNAs in association with tissue context expression, and not just for protein-coding gene analysis.

To improve our understanding of the spatial expression patterns of lncRNA across human tissues, we built a comprehensive spatial expression atlas of lncRNA in humans by re-analyzing a large-scale transcriptomic atlas and provided a comprehensive baseline map of lncRNA expression across the human body. Using integrative analysis of a lncRNA transcriptome covering 20 519 samples across 38 tissues, we firstly assessed the relative expression of lncRNAs in each tissue in all publications. Next, robust TE lncRNAs in normal tissues were identified by analyzing four transcriptomes of normal tissues. We extended this analysis to investigate the expression feature of TE lncRNA in cancer, and discovered candidate clinically associated lncRNA markers by comparing transcriptional profiles of pathological samples to the normal transcriptome. Our integrative analysis substantially obtained the robust TE lncRNAs across the datasets. The comprehensive atlas of TE lncRNAs across multiple tissues and cancers provides data for a better understanding of lncRNA function and disease development and progression.

2 Materials and methods

2.1 Construction of a lncRNA and mRNA transcriptome across tissues

We used patients' data acquired from publicly available datasets that were collected with patients' informed consent. To conduct a systematic analysis of the transcriptome across multiple tissues, we used five widely available lncRNA and messenger RNA (mRNA) expression datasets (Table S1). From the GTEx (Lonsdale et al., 2013), we obtained the expression profiles for 8555 samples across 30 tissues. Expression of lncRNAs and protein-coding genes was measured by reads per kilobase of transcript per million mapped reads (RPKM). We obtained the raw RNA-seq data from two other tissue banks, HBM2 (Schroth, 2011) and HPA (Rustici et al., 2013; Uhlén et al., 2015). The HBM2 includes 16 samples across 16 tissues, and the HPA includes 95 samples across 27 tissues. The sequence reads were mapped to the human genome 19 using TopHat (http://ccb.jhu.edu/

software/tophat/index.shtml) (Kim et al., 2013), and the expression levels were calculated using Cufflinks v2.1.1 based on fragments per kilobase of transcript per million mapped reads (FPKM) (Trapnell et al., 2012). In addition, we obtained gene and lncRNA expression profiles from 760 samples across 34 tissue types from the FANTOM project (https://fantom.gsc.riken.jp/5) (Noguchi et al., 2017). Expression of genes and lncRNAs was measured by maximum counts per million (CPM).

Moreover, we obtained the lncRNA and gene expression datasets generated by the TCGA Research Network (http://cancergenome.nih.gov). In total, 33 different TCGA projects representing different cancer types were analyzed. RNA-seq-based expression profile data were obtained via the R package "TCGAbiolinks" (Colaprico et al., 2016). We downloaded the FPKM-based gene expression for 33 types of cancer. Based on the annotations in GENCODE (https://www.gencodegenes.org) (Derrien et al., 2012), we divided the gene expression profiles into lncRNA and protein-coding gene expression for each cancer type.

2.2 Cluster analysis of tissues based on lncRNA expression

To investigate whether the expression of lncRNAs can distinguish different tissues, the top 1000 lncRNAs with higher variance were obtained from each data resource. The overlapping lncRNAs in four resources (GTEx, HBM2, HPA, FANTOM) were selected. We next performed uniform manifold approximation and projection (UMAP) dimension reduction with the R package "umap" (https://github.com/tkonopka/umap), using these shared top lncRNAs. Each sample was visualized in two dimensions and the samples were colored by tissue type and data resource.

2.3 Analysis of the complexity of the lncRNA transcriptome for each tissue

In each tissue, lncRNAs were first ranked in descending order based on their average expression levels in the tissue. For a given top number of lncRNAs (k), we defined the complexity (C) of the lncRNA transcriptome in tissue (i) as follows:

$$C(i,k)=\frac{\sum_{a=1}^{k}\exp(a,i)}{\sum_{a=1}^{n}\exp(a,i)},k=1,2,3,\cdots,n,$$

where n is the total number of lncRNAs in tissue i and exp (a, i) represents the average expression value of lncRNA a in tissue i.

2.4 Identification of robust TE lncRNAs by integration analysis

All the human TE lncRNAs were identified based on the spatial expression patterns of lncRNAs across different tissues, as has been done with coding genes in a previous study (Uhlén et al., 2016). Collectively, three sub-categories of TE lncRNAs were further stratified to reflect increasing degrees of elevated expression in a particular tissue, including "tissue-specific (TS)," "tissue-enriched (TER)," and "tissue-enhanced (TEH)": (1) TS lncRNAs were expressed only in a particular tissue; (2) TER lncRNAs had at least five-fold higher expression level in a particular tissue compared to other tissues; and (3) TEH lncRNAs had at least five-fold higher expression level in a particular tissue compared to the average expression levels in all other tissues. Similarly, we identified TE lncRNAs in each cancer type.

We identified TE lncRNAs individually from each resource and integrated the TE lncRNAs from different datasets. To integrate datasets and further refine the robust TE lncRNAs in normal tissues, we required that the robust TE lncRNAs be identified in the same tissue in at least two resources.

2.5 Assessment of the power of TE lncRNAs to predict tissue origin

We used TE lncRNAs as molecular markers to identify each tissue. First, we normalized the expression of TE lncRNAs in each resource by Z-score. Then, the normalized expression profiles were subjected to CIBERSORT (Newman et al., 2015) to predict the proportion in each tissue.

2.6 Enrichment analysis of cancer TE lncRNAs

To evaluate whether the TE lncRNAs identified in cancer patients were significantly enriched in the TE lncRNAs identified in corresponding normal tissues, we performed a hypergeometric test for each cancer type. The P-value was calculated by

$$P=1-F(x-1|N,M,n)=1-\sum_{i=0}^{x-1}\frac{\binom{n}{i}\binom{N-n}{M-i}}{\binom{N}{M}},$$

where N is the total number of lncRNAs, n and M are the numbers of TE lncRNAs identified in cancer patients and corresponding normal tissue, respectively, and x is the number of overlapping TE lncRNAs. We also calculated the odds ratio (OR) by $\mathrm{OR} = \frac{x/M}{n/N}$.

2.7 Identification of clinically relevant lncRNAs in cancer

To identify clinically relevant lncRNAs in cancer, we detected both differentially expressed and survival-related lncRNAs. First, a *t*-test was used to identify differentially expressed lncRNAs in each cancer type. We only considered cancer types with more than five normal samples, and lncRNAs with an expression level of 0 in less than 30% of samples were included for subsequent analysis. The lncRNAs with fold changes of >2 or <0.5 and adjusted *P*-value of <0.05 were identified as differentially expressed in each cancer (Li YS et al., 2018). To identify survival-related lncRNAs in cancer, we firstly ranked the cancer patients based on the expression of a specific lncRNA. The top-ranked and bottom-ranked 20% of patients were selected. The difference in the survival time of these two groups was analyzed with the log-rank test and lncRNAs with *P*-value of <0.05 were identified as survival-related lncRNAs in cancer.

2.8 Functional enrichment analysis of lncRNAs

To predict the function of TE lncRNAs, we detected their co-expressed protein-coding genes via Pearson's correlation coefficient (R>0.4 and P<0.05). The hypergeometric test was used for detecting enriched Gene Ontology (GO) categories. Then, the enriched significant *P*-values were adjusted with the Benjamini and Hochberg's methods, and finally, GO terms with adjusted *P*-value of <0.05 were considered.

3 Results

3.1 High consistency exhibited by lncRNA transcriptomes within tissues

Analysis of the spatial expression of the lncRNA transcriptome across different tissues and organs would greatly improve our understanding of human biology and disease. Five independent datasets that covered distinct methods and conditions were obtained, and provided the opportunity to perform an integrative analysis of the spatial features of lncRNA expression. First, we discovered that approximately 30 tissues were available in each resource, except for HBM2, which included 16 tissues (Figs. 1a and 1b, Table S1). There were totally 9426 normal samples across 38 normal tissues and 11 093 samples across 33 cancer types.

Next, we collected a comprehensive lncRNA annotation, including Ensembl (GRCH37) (Aken et al., 2017), GENCODE (Derrien et al., 2012), and an additional data of a previous study by Cabili et al. (2011). We compared the genomic coordinates of these lncRNAs to obtain non-redundant lncRNA annotation. If two lncRNA loci overlapped by more than 80%, we reserved these lncRNAs from GENCODE. Finally, 15 646 lncRNAs were analyzed. We calculated the average expression of lncRNAs and mRNAs across all tissues and compared the expression distribution of lncRNAs and coding genes. To explore lncRNA expression patterns across this heterogeneous pool of data among different datasets and tissue types, we performed unsupervised clustering based on the expression of lncRNAs with high variation.

We discovered that the sample groups of the same tissues from different data resources were clustered together (Fig. 1c). At the same time, these groups of samples were not clustered together in the same way as in the dataset (Fig. S1a). Moreover, the frequent high similarity within the same tissues between the HPA (fresh-frozen tissues) and GTEx (postmortem tissues) indicated negligible effects in the sampling procedures. These results suggested that lncRNA expression exhibited high consistency within the same tissue and could distinguish different tissues more reliably than other factors, including the laboratory of origin. Thus, we are able to present a comprehensive resource of lncRNA expression across multiple tissues and conditions obtained by different methods.

3.2 Higher complexity expressed by lncRNA within specialized tissues

The evidence is clear that lncRNA plays a significant role in maintaining the morphology and function of tissues (Cabili et al., 2011; Li et al., 2014). However, it is unclear how many human lncRNAs exist, where are they expressed, and in what quantities. First, we calculated the average expression of lncRNAs and mRNAs across all tissues and compared the expression

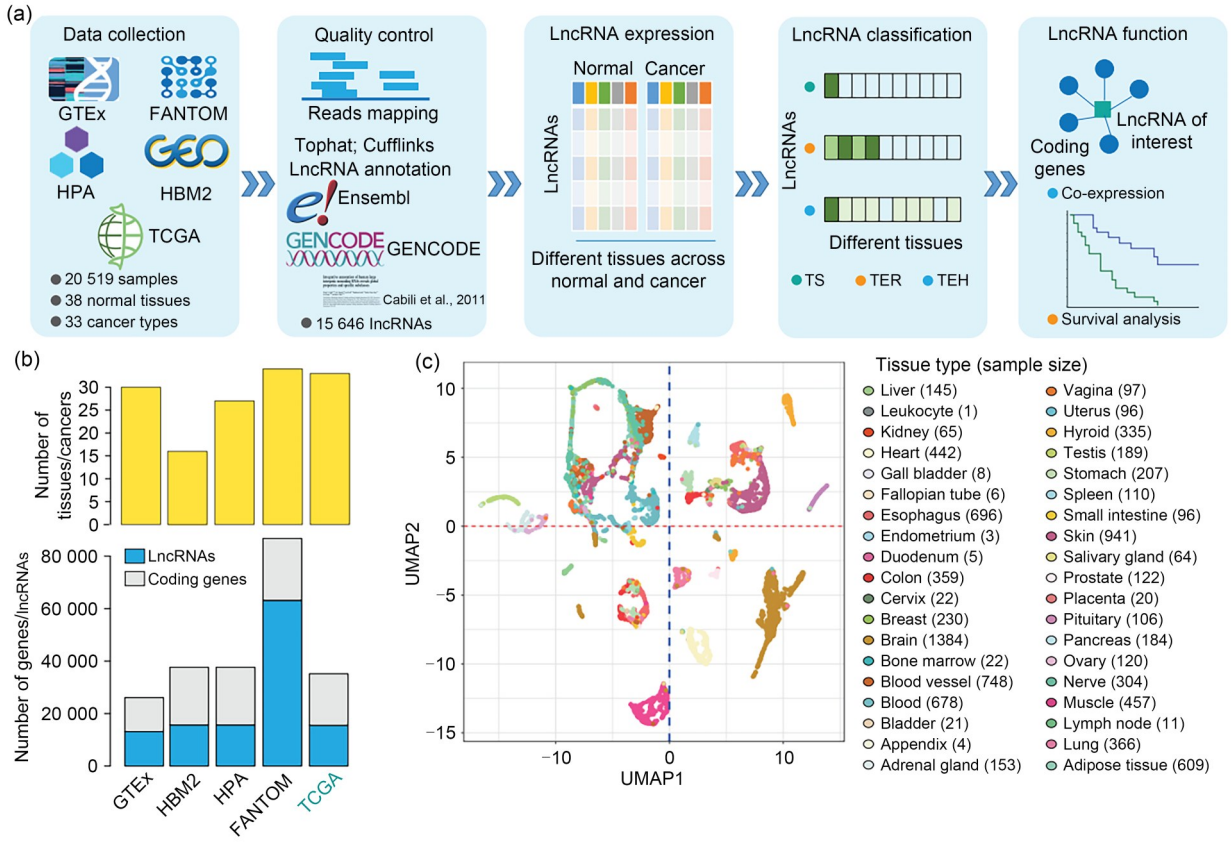

Fig. 1 Overview of the lncRNA spatial transcriptome. (a) Data collection and classification of lncRNAs across different resources. (b) Top bar plot shows the number of tissues or cancers obtained from each of the five resources. Bottom bar plot shows the number of lncRNAs and coding genes across resources. (c) UMAP visualization of all samples from four normal lncRNA transcriptome resources. Samples were colored by tissue type. LncRNA: long non-coding RNA; GTEx: Genotype-Tissue Expression; FANTOM: Functional Annotation of the Mammalian Genome; HPA: Human Protein Atlas; HBM2: Human BodyMap 2; TCGA: The Cancer Genome Atlas; TS: tissue-specific; TER: tissue-enriched; TEH: tissue-enhanced; UMAP: uniform manifold approximation and projection.

distribution of lncRNAs and coding genes. Consistent with previous studies (Eisenberg and Levanon, 2013; Lingadahalli et al., 2018), we found that lncRNAs showed lower expression levels in tissues obtained from all four projects (Fig. 2a). The numbers of expressed lncRNAs in each tissue were also far less than that in protein-coding genes (Fig. S2). In an attempt to explore the tissue complexity and composition of the human lncRNA transcriptome, the first question we asked was how many lncRNAs were expressed in a specific tissue. As a result, we found 11 035–13 023 lncRNAs (at the expression threshold of 0.1 in GTEx, HPA, and HBM2) expressed in at least one of the 38 studied tissues (Fig. 2b), suggesting that expression of approximately 80% of total lncRNAs seems to account for all basic and specialized functions in the studied tissues. The number of expressed lncRNAs in FANTOM was higher than that obtained from RNA-seq data at a similar proportion of expressed lncRNAs. Moreover, we found that the majority of lncRNAs were consistently expressed in all projects. When a relatively strict threshold of expression was set, we also found a similar proportion of expressed lncRNAs in the GTEx, HPA, and HBM2. Only approximately 1%–9% of lncRNAs were detected in individual projects. This may be due to the difference in the detection of different tissues in these projects. We also obtained similar results when we used a relatively strict expression threshold to define the expressed lncRNAs (Fig. S3).

The second question we asked was how the various tissues and the lncRNA transcriptome differed in composition and complexity. We calculated tissue complexity by measuring the proportion of the

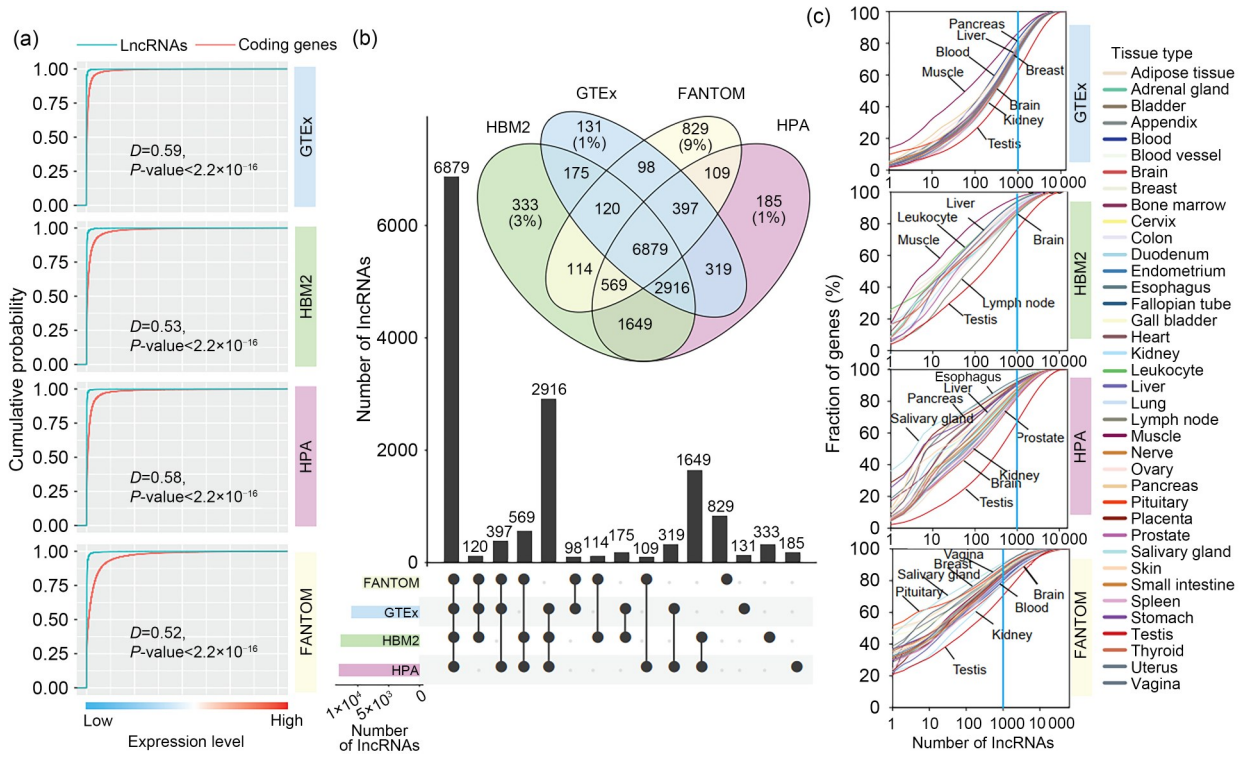

Fig. 2 Spatial complexity of lncRNAs from four transcriptome resources. (a) Empirical cumulative distribution of lncRNA and protein-coding genes across four transcriptome resources (FANTOM, HPA, HBM2, GTEx) with *P*-value<0.05 by Kolmogorov-Smirnov test; *D* is the maximum vertical difference value. (b) Overlap of tissue-elevated (TE) lncRNAs in four transcriptome resources in loose thresholds. (c) Spatial complexity of lncRNAs from four transcriptome resources. LncRNAs: long non-coding RNAs; GTEx: Genotype-Tissue Expression; FANTOM: Functional Annotation of the Mammalian Genome; HPA: Human Protein Atlas; HBM2: Human BodyMap 2.

transcriptome contributed by the most abundant lncRNAs at different gradients (Fig. 2c). Notably, the testis, brain, and kidney showed consistently high transcriptome complexity in all resources. These tissues expressed large percentages of lncRNAs in the genome, with a small fraction of the lncRNA pool contributing to the most highly expressed genes. On the other hand, the liver and muscle had a less complex transcriptome, expressing fewer genes in the genome. A large fraction of the transcriptome in these tissues contributed to the most highly expressed genes. For example, the top thousand most highly expressed genes contributed to 96.4% of the lncRNA population in muscle tissue in the HBM2, whereas they contributed less than 62.4% of the lncRNA in testis tissue in the GTEx. Similar trends in transcriptome complexity were reported from previous studies on protein-coding genes in mammals (Jongeneel et al., 2005; Ramsköld et al., 2009; Zhu et al., 2016), suggesting conservation of the tissue-controlled expression patterns. We proposed that this complexity in the expression pattern of lncRNAs might be explained in terms of not only the degree of specialization, but also the types of cells in each tissue. For example, the brain has a variety of cells specialized for equally important but different functions. As different cell types express different cell-controlled lncRNAs, tissue as a whole has a large collection of equally important TE genes expressed at comparable rates.

3.3 Spatial classification of lncRNAs based on tissue profiling

As described above, lncRNAs exhibit notably higher variation of expression among tissues when compared to protein-coding genes. However, knowledge about their expression patterns across tissues is still limited. We found that more than 80.61% of the lncRNAs were expressed in two or more tissues with a minimum expression value of 0.1. Moreover, 65.95% were expressed in five or more tissues. Thus, the comprehensive collection of a genome-wide lncRNA transcriptome would be an attractive way to establish a

refined classification of lncRNAs with regard to their spatial expression patterns; the expression pattern of TE lncRNAs can be explained in terms of tissue complexity (Salem et al., 2015). In this study, TE lncRNAs were detected only if their expression in one tissue was five-fold or higher compared to the expression values of the other tissues. Collectively, three sub-categories of TE lncRNAs were further stratified to reflect increasing degrees of elevated expression in a particular tissue, including TS, TER, and TEH (Fig. 3a).

Based on the definition, TE lncRNAs were identified based on analysis of the four resources of normal tissues (Figs. 3b–3e), and accounted for 56.47%–75.38% of the lncRNAs. Notably, there were a higher number of TE lncRNAs in testis and brain tissues. In addition, about one-third of TE lncRNAs were the TS subtype in the GTEx project, and the majority of lncRNAs were TEH across other tissues (Fig. 3b). For the other resources, there were more TEH lncRNAs in the HPA and FANTOM projects (Figs. 3c–3e). As in the GTEx project, the testis was observed to contain the largest number of TE lncRNAs again, followed by brain, blood, and skin tissues. We also found that the proportion distributions of these three TE lncRNA subtypes in the four resources were highly similar (Fig. 3f). For instance, approximately 44%–71% of the TS lncRNAs were shared among different resources. There were higher similarity scores between the GTEx and HPA projects for all three types of TE lncRNAs. Moreover, we obtained similar results when using the higher threshold of expression (Fig. S4). Overall, there was significant overlap in the spatial classification of lncRNAs based on these four independent datasets.

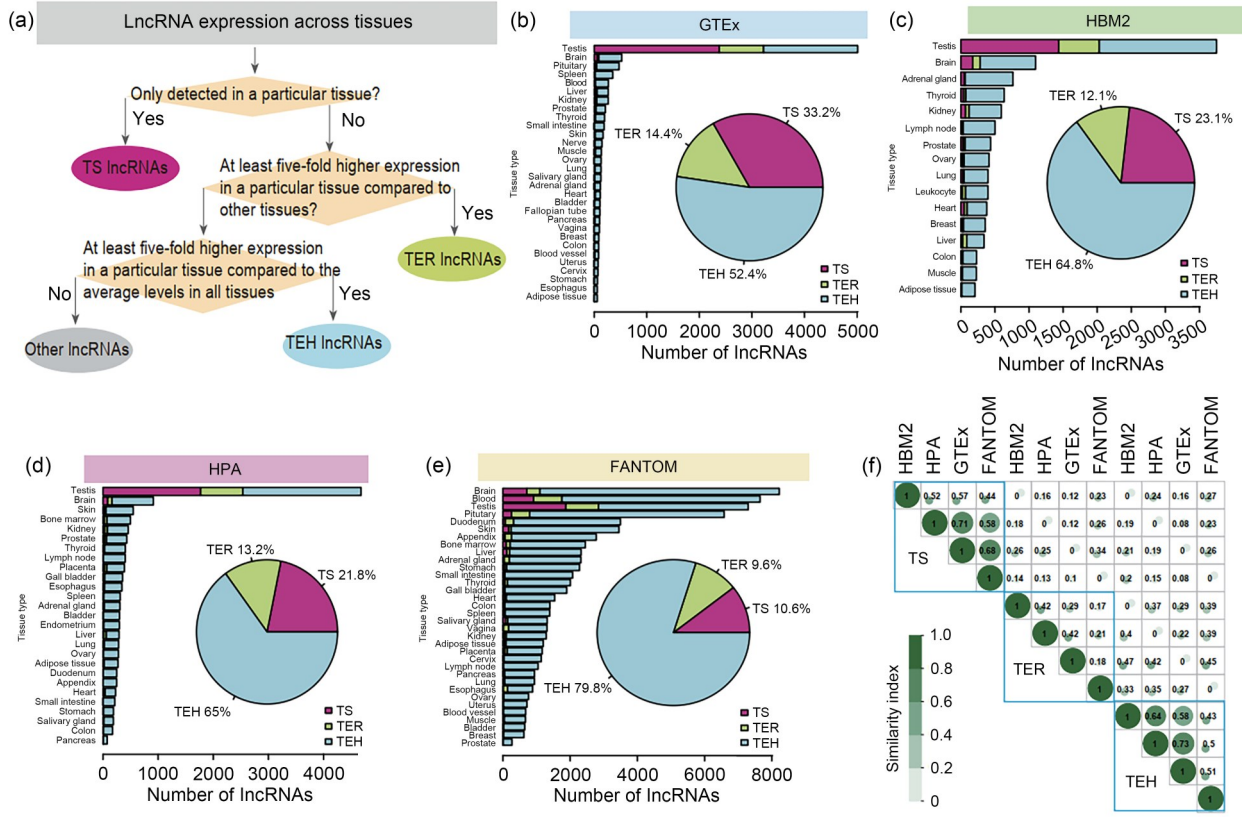

Fig. 3 Classification of lncRNAs and comparison of tissue-elevated (TE) lncRNAs among four transcriptome resources. (a) The workflow for lncRNA classification. (b–e) Bar plots showing the number of TE lncRNAs in each tissue across GTEx (b), HBM2 (c), HPA (d), and FANTOM (e). Purple represents TS, green represents TER, and light blue represents TEH. Pie charts showing the proportion of each type of TE lncRNAs in four transcriptome resources. (f) The similarity of different lncRNA categories among different transcriptome resources. Rows and columns indicate TE lncRNA categories, and the number is the similarity index between two types of lncRNA. LncRNAs: long non-coding RNAs; GTEx: Genotype-Tissue Expression; FANTOM: Functional Annotation of the Mammalian Genome; HPA: Human Protein Atlas; HBM2: Human BodyMap 2; TS: tissue-specific; TER: tissue-enriched; TEH: tissue-enhanced.

3.4 Closely correlation between robust TE lncRNAs and tissue-related functions

The above three types of lncRNA all exhibited high expression in specific tissues. Thus, we combined them and identified robust TE lncRNAs (Section 2.4). To integrate the TE lncRNAs across different resources, we next focused on the 34 tissues that were investigated in at least two projects. In total, we identified 1–4684 TE lncRNAs across these tissues (Fig. 4a). Similar to our results above, the male reproductive tissue and the testis had by far the highest number of TE genes, followed by the brain. Moreover, the majority (40.7%) of TE lncRNAs in testis tissue were identified in more than two resources. Approximately 12.0% of the TE lncRNAs in testis tissue were observed in all resources. This indicated their robust spatial expression in the testis. Interestingly, these two tissues have been found to have the greatest number of TE protein-coding genes in both the HPA and GTEx (Uhlén et al., 2016). However, the number of TE lncRNAs is greater than the number of protein-coding genes, indicating that lncRNAs have a more important role in these tissues. Functional enrichment analysis of the TER protein-coding genes in the HPA has been performed and the results showed links to the function of each tissue (Uhlén et al., 2015). Similarly, we found that the TE lncRNA testis developmental related gene 1 (*TDRG1*) was expressed in normal testis tissue in the four resources, and the expression level was higher than that in other tissues (Fig. 4b) (Jiang et al., 2011). Functional analysis of their co-expressed genes by the Genomic Regions Enrichment of Annotations Tool (GREAT) (http://great.stanford.edu/

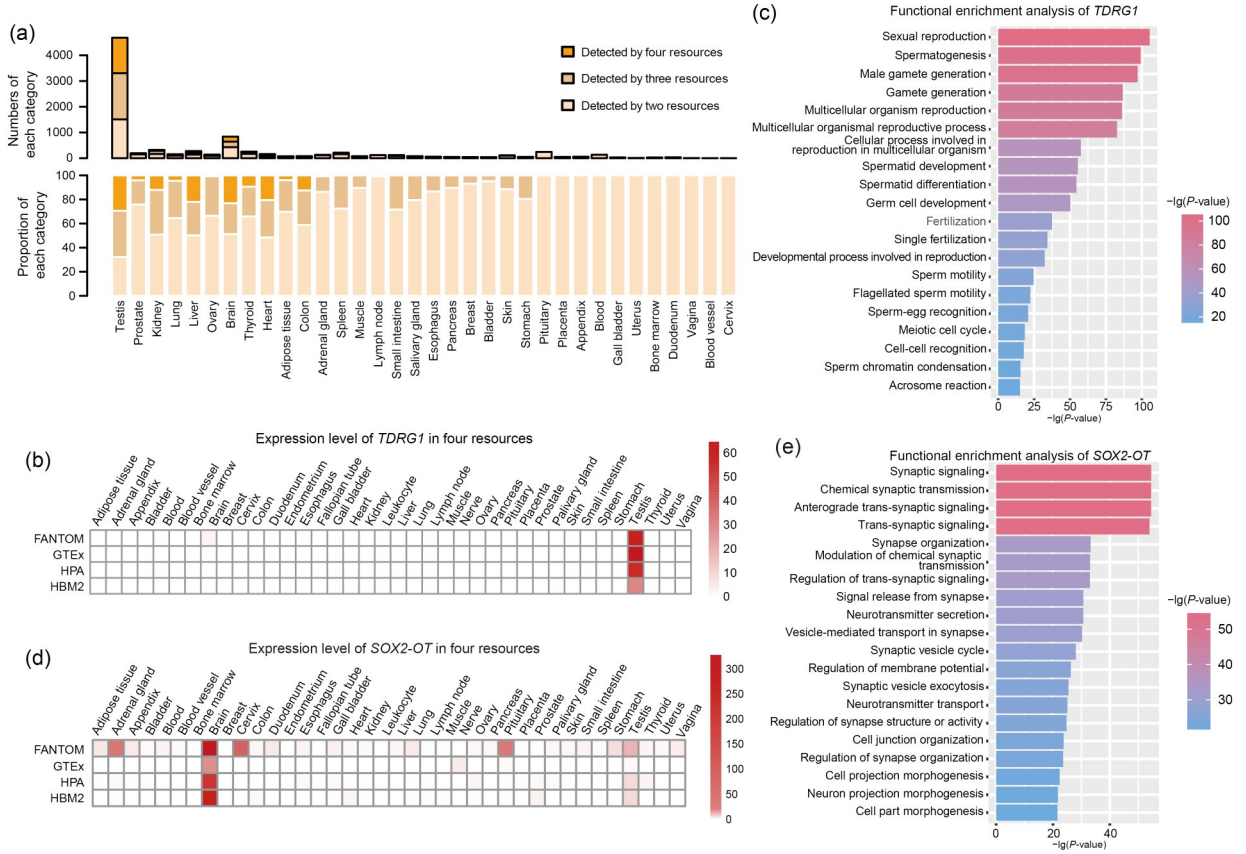

Fig. 4 Statistics of robust tissue-elevated (TE) lncRNAs in each tissue. (a) Number of robust TE lncRNAs in each tissue and number of TE lncRNAs identified by at least two, three, or four resources (top); proportion of integrated TE lncRNAs in each tissue and proportion of TE lncRNAs identified by at least two, three, or four resources (bottom). (b) Expression value of *TDRG1* in the four transcriptome sources. (c) Functional enrichment analysis of *TDRG1* correlated-coding genes. (d) Expression value of *SOX2-OT* in the four transcriptome sources. (e) Functional enrichment analysis of *SOX2-OT* correlated-coding genes. LncRNAs: long non-coding RNAs; GTEx: Genotype-Tissue Expression; FANTOM: Functional Annotation of the Mammalian Genome; HPA: Human Protein Atlas; HBM2: Human BodyMap 2; *SOX2-OT*: SRY-box transcription factor 2 overlapping transcript; *TDRG1*: testis developmental related gene 1.

public/html) revealed that they regulated testis-related biological functions, including spermatogenesis, male gamete generation, and spermatid differentiation (Fig. 4c) (Gan et al., 2016; Wang et al., 2016; Wei et al., 2018). Previous research has also shown that TDRG1 promoted development and migration of sperm and that over-expression might induce tumor development (Chen et al., 2015).

In addition, we found that the numbers of TE lncRNAs in normal brain tissues ranked third; this rank had priority over that based on TE protein-coding genes. An example is SRY-box transcription factor 2 overlapping transcript (*SOX2-OT*), which had enriched expression in the brain (Fig. 4d). Functional analysis suggested that *SOX2-OT* was involved in chemical synaptic transmission and regulation, regulation of membrane potential, and neurotransmitter secretion (Fig. 4e), and was associated with brain function. It is evident that the expression patterns of lncRNAs were likely to be associated with the phenotype consequences. Conversely, the cervix expressed the lowest number of TE lncRNAs, followed by blood vessel, vagina, duodenum, bone marrow, and then uterus, in increasing order. The uterus, cervix, and ovary have been found to have the least number of TE protein-coding genes; however, the blood vessel has not yet been investigated.

What is the contribution of the TE lncRNAs to the transcriptome in different tissues? Although TE lncRNAs were relatively highly expressed in testis, brain, and liver tissues, these lncRNAs significantly contributed to the total cellular lncRNA pool (28.8%, 5.1%, and 1.7% of total lncRNAs, respectively). Conversely, in the cervix, blood vessel, and vagina, which expressed a small number of TE lncRNAs, these genes contributed only 0.0061%, 0.0061%, and 0.0246% of total cellular lncRNA, respectively. These findings indicated the wide variation in the number of genes and regulation of lncRNA expression that determine tissue variation.

As described above, these TE lncRNAs showed high consistency across different projects. To determine whether expression profiles of the identified TE lncRNAs could distinguish different normal tissue types, we used the CIBERSORT tool to perform linear support vector regression (Newman et al., 2015). The expression profiles of TE lncRNAs were used as the feature matrix. We found that the majority of tissue types were recalled based on the expression matrix of TE lncRNAs (Fig. S5). These results suggested that the TE lncRNAs we identified could have critical tissue-specific functions.

3.5 Trend of evalated expression of TE lncRNAs within the same tissues in cancers

It has been widely accepted that the identification and appropriate use of TE lncRNAs could provide important insights into disease mechanisms and tissue-specific therapeutic targets. Besides elevated expression in normal tissues, high disease-type specificity has been discovered (Deras et al., 2008; Gupta et al., 2010; Hung et al., 2011; Fu et al., 2019; Qu et al., 2022), and genes with disease-restricted expression are highly desirable both as markers and pharmacologic targets because selective expression imparts the specificity required for successful disease-context targeting approaches. To better understand the TE features of lncRNAs for each cancer type and to advance the discovery of clinically related lncRNAs, we also identified lncRNAs with especially elevated expression in the corresponding cancer samples that were absent or expressed weakly in other cancers.

Based on lncRNA expression profiles of more than 10 000 cancer patients across 33 cancer types from the TCGA project, we identified the TE lncRNAs in each cancer type. We found that 57.69% of lncRNAs were TE lncRNAs, 4.80% were TS lncRNAs, and there were high proportions of both TER (18.00%) and TEH (77.20%) lncRNAs, unlike the distribution in the normal tissue states (Fig. S6a). In addition, we identified a higher number of TE lncRNAs in acute myeloid leukemia (LAML) than in other solid tumors (Fig. S6b). In solid tumors, there were more TE lncRNAs in brain lower grade glioma (LGG), testicular germ cell tumor (TGCT), and liver hepatocellular carcinoma (LIHC). Interestingly, the corresponding normal tissues (brain, testis, and liver) also had many TE lncRNAs. Next, we further explored the association between TE lncRNAs within the same tissue under normal or cancer conditions. Significantly higher overlaps than in random conditions were discovered through both the OR index and hypergeometric test method across all tissues considered, which is a notable finding (Fig. 5a). In particular, the highest number of TE lncRNAs was found in the brain vs. LGG and glioblastoma multiforme (GBM), followed by testis vs. TGCT (Table S2). This suggested that

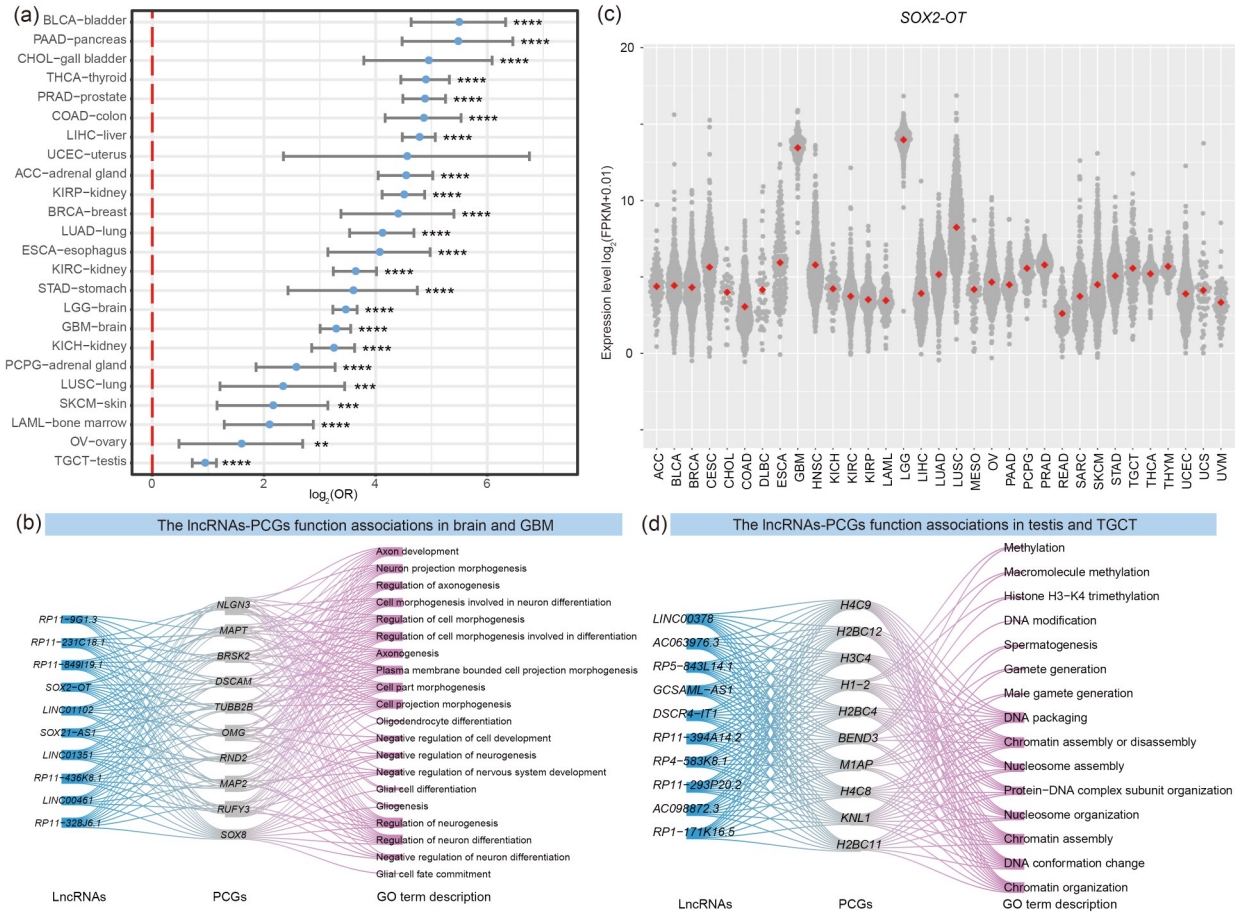

Fig. 5 Comparison of lncRNA transcriptome in cancer and corresponding tissue. (a) The odds ratio (OR) of normal tissue and cancer. ** *P*<0.01, *** *P*<0.001, **** *P*<0.0001. (b) The coding genes of GBM, which are co-expressed with the overlapping TE lncRNAs of the brain (at least in two transcriptome sources) and GBM. (c) Scatter plot showing the expression level distribution of *SOX2-OT* in samples of 33 cancer types, with red dots for the average expression of *SOX2-OT* in different cancer types. (d) The coding genes of TGCT, which are co-expressed with the overlapping TE lncRNAs of the testis (at least in two transcriptome sources) and TGCT. PCGs: protein-coding genes; FPKM: fragments per kilobase of transcript per million mapped reads; *SOX2-OT*: SRY-box transcription factor 2 overlapping transcript; TGCT: testicular germ cell tumor; GO: Gene Ontology; GBM: glioblastoma multiforme; ACC: adrenocortical carcinoma; BLCA: bladder urothelial carcinoma; BRCA: breast invasive cancer; CESC: cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL: cholangiocarcinoma; COAD: colon adenocarcinoma; DLBC: lymphoid neoplasm diffuse large B-cell lymphoma; ESCA: esophageal cancer; HNSC: head and neck squamous cell carcinoma; KICH: kidney chromophobe; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LAML: acute myeloid leukemia; LGG: lower grade glioma; LIHC: liver hepatocellular carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; MESO: mesothelioma; OV: ovarian serous cystadenocarcinoma; PAAD: pancreatic adenocarcinoma; PCPG: pheochromocytoma/paraganglioma; PRAD: prostate adenocarcinoma; READ: rectum adenocarcinoma; SARC: sarcoma; SKCM: skin cutaneous melanoma; STAD: stomach adenocarcinoma; THCA: thyroid carcinoma; THYM: thymoma; UCEC: uterine corpus endometrial carcinoma; UCS: uterine carcinosarcoma; UVM: uveal melanoma.

these TE lncRNAs in normal tissue were likely to exhibit high expression in a cancer state.

By functional enrichment analysis of their co-expressed genes, we found that TE lncRNAs in brain tissue under both normal and cancer conditions were involved in axon development, neuron projection morphogenesis, and regulation of axonogenesis (Fig. 5b). For example, *SOX2-OT* has been demonstrated to be up-regulated during central nervous system development, and is associated with Parkinson's disease and a number of cancers (Guo et al., 2021). *SOX2-OT* is elevated more than 47-fold in normal brain tissue and also has >32 times higher expression than that in other cancers (Fig. 5c). Similarly, TE lncRNAs in testis

tissue under both normal and cancer conditions tend to regulate spermatogenesis and functions associated with DNA epigenetic modification (Fig. 5d). Thus, we conjecture that the strong bias for the high expression of cancer lncRNAs might be related to more active functions during the progression of cancer.

3.6 Strong clinical associations exhibited by TE lncRNAs across cancers

Given the significant association of TE lncRNAs within tissues under normal and cancer conditions, a fundamental question is what fraction of TE lncRNAs in tumors is functionally tumorigenic or clinically valuable. To address this question, we focused on cancer TE lncRNAs that showed differential expression or correlations with patient survival. First, we identified the differentially expressed TE (DTE) lncRNAs in each cancer type with more than five normal tissues. A DTE lncRNA is defined by at least two-fold changes and *P*-value of <0.05 (*t*-test followed by multiple test correction using Benjamini-Hochberg's method). All 15 of the analyzed cancer types contained DTE lncRNAs, ranging from 21 in ESCA to 256 in LIHC (Table S3). Interestingly, we found that cancer DTE lncRNAs were significantly enriched at the genome-wide level (Fig. 6a, Table S3), indicating that these TE lncRNAs are likely to be associated with the development of cancer.

Moreover, we found that a large number of DTE lncRNAs in most cancer types were likely to exhibit up-regulated expression in cancer tissues (Figs. 6a and 6b). For example, 92.97% (119 out of 128) DTE lncRNAs were up-regulated in prostate adenocarcinoma (PRAD). Most of the up-regulated DTE lncRNAs were cancer-constricted and up to 92.01% and 92.06% of DTE lncRNAs were up-regulated and down-regulated in only one cancer type, respectively (Fig. 6c). Among these DTE lncRNAs, only 11.13% are reported to be related with cancers (Fig. 6d), indicating that most DTE lncRNAs in cancers have not yet been described. For example, prostate cancer associated transcript 18 (*PCAT18*) is elevated more than 16-fold in normal prostate tissue and also has more than 31 times higher expression in PRAD than in other cancers. Interestingly, it has been shown to be more highly expressed in blood in the case of prostate cancer. Tyrosine aminotransferase (TAT) antisense RNA 1 (*TAT-AS1*) is another example; the expression level is elevated more than 33-fold in normal liver tissue and also more than 11 times higher in LIHC than in other cancers, but its association with disease has not been reported. Thus, these DTE lncRNAs are good candidates for cancer diagnosis as well as investigation of their function in cancer.

On the other hand, lncRNA expression has been demonstrated to be associated with cancer patients' survival (Du et al., 2013; Zhou et al., 2016; Li YS et al., 2018). We next explored the extent to which the TE lncRNAs were associated with the survival of patients in each cancer type. A survival-related TE (STE) lncRNA was defined by a *P*-value in survival analysis of <0.05 (log-rank test). STE lncRNAs were found in all 33 of the cancers analyzed (Figs. 6e–6g, Table S3). Interestingly, kidney renal clear cell carcinoma (KIRC) had 155 STE lncRNAs and also had the most DTE lncRNAs. Moreover, unlike DTE lncRNAs, the number of STE lncRNAs in six cancers was relatively small; for instance, no more than ten STE lncRNAs were found. These findings indicated that TE lncRNA had relatively high power as a diagnostic marker than a prognostic one. We also focused on risky STE lncRNAs, high expression of which is associated with poor survival, and found just 476 of them across cancer types. Among these risk STE lncRNAs, just 9.21% are reported to be related to survival, and these include rhabdomyosarcoma 2-associated transcript (*RMST*), MIR155 host gene (*MIR155HG*), DiGeorge syndrome critical region gene 9 (*DGCR9*), and *SOX2-OT*. We found that high expression of surfactant associated 1, pseudogene (*SFTA1P*) and SLIT-ROBO Rho GTPase-activating protein 3 antisense RNA 2 (*SRGAP3-AS2*) was associated with better survival of patients in lung adenocarcinoma (LUAD), and high expression of *RP11-63A11.1* and *RP11-536I6.2* was associated with better survival of patients in kidney renal papillary cell carcinoma (KIRP) (Fig. S7). These STE lncRNAs provided good candidates for cancer prognostic analysis.

In summary, the association of clinically relevant TE lncRNAs with cancer points to a strategy of systematically exploring the relationship between pathology and the spatial expression patterns of lncRNAs across various human tissues.

4 Discussion

Gaining a more complete understanding of the relationship between the human genome and phenotypes

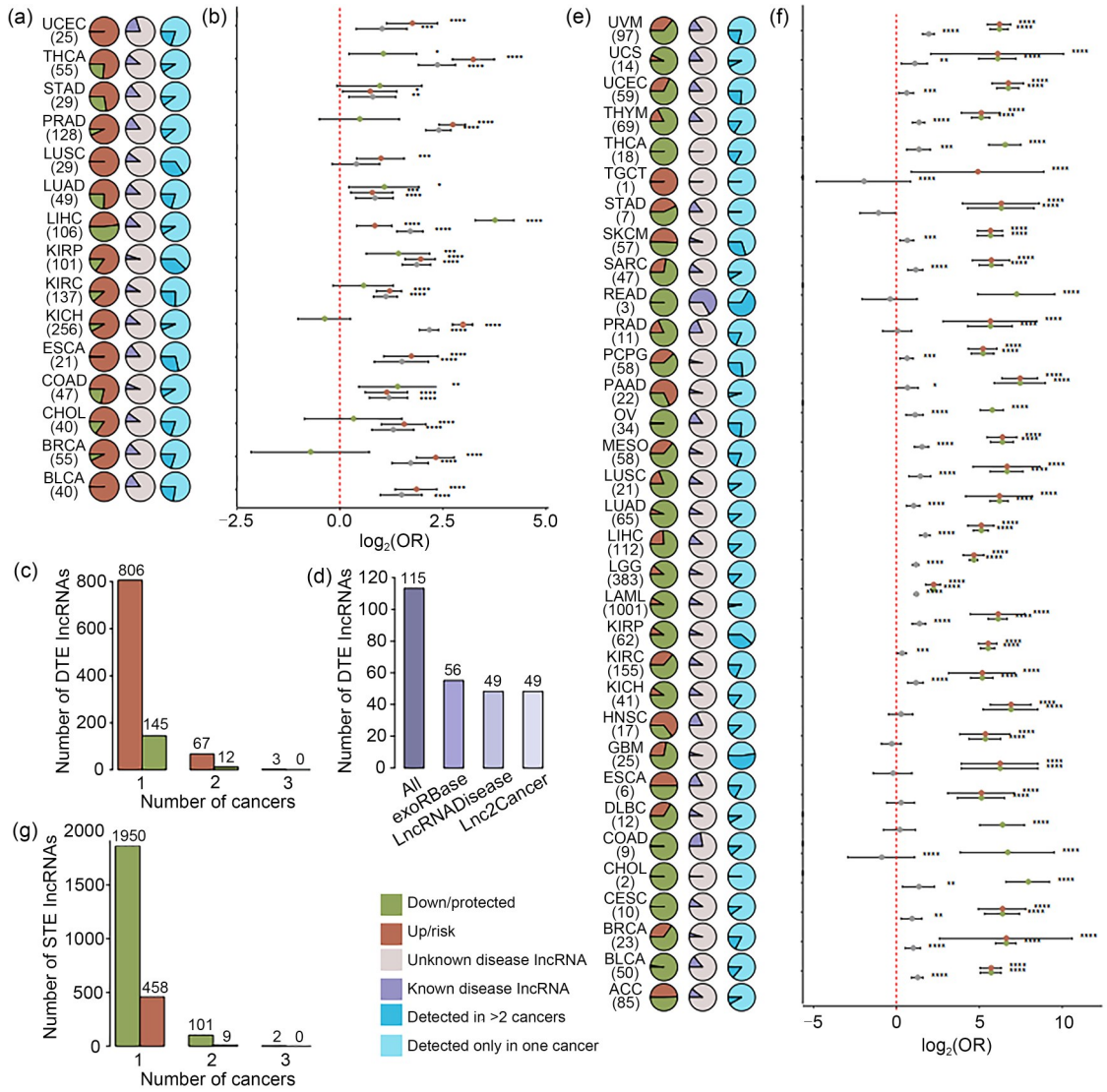

Fig. 6 Common TE lncRNAs of tissues and cancers associated with differentially expressed lncRNAs. * ***P*<0.1,** ** ***P*<0.01,** *** ***P*<0.001,** **** ***P*<0.0001. (a) Proportion of differentially expressed TE (DTE) lncRNAs in fifteen cancer types. (b) Enrichment analysis of TE lncRNAs and DTE lncRNAs in cancer. Grey dots represent the OR of all DTE lncRNAs in each tissue type; red dots represent the OR of up-regulated DTE lncRNAs in each tissue type; green dots represent the OR of down-regulated DTE lncRNAs in each tissue type. (c) Numbers of differentially up-regulated and down-regulated TE lncRNAs shared by cancers. (d) Numbers of DTE lncRNAs confirmed to be associated with specific diseases in three public databases. All represents the union dataset of three public databases. (e) Proportion of survival-related TE (STE) lncRNA in 33 cancer types. (f) Enrichment analysis between the TE lncRNAs of cancer- and survival-related lncRNAs. Grey dots represent the OR of all STE lncRNAs in each tissue type; red dots represent the OR of risk STE lncRNAs in each tissue type; green dots represent the OR of protected STE lncRNAs in each tissue type. (g) The numbers of STE lncRNAs shared by cancers. LncRNAs: long non-coding RNAs; TE: tissue-elevated; OR: odds ratio; ACC: adrenocortical carcinoma; BLCA: bladder urothelial carcinoma; BRCA: breast invasive cancer; CESC: cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL: cholangiocarcinoma; COAD: colon adenocarcinoma; DLBC: lymphoid neoplasm diffuse large B-cell lymphoma; ESCA: esophageal cancer; GBM: glioblastoma multiforme; HNSC: head and neck squamous cell carcinoma; KICH: kidney chromophobe; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LAML: acute myeloid leukemia; LGG: lower grade glioma; LIHC: liver hepatocellular carcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; MESO: mesothelioma; OV: ovarian serous cystadenocarcinoma; PAAD: pancreatic adenocarcinoma; PCPG: pheochromocytoma/paraganglioma; PRAD: prostate adenocarcinoma; READ: rectum adenocarcinoma; SARC: sarcoma; SKCM: skin cutaneous melanoma; STAD: stomach adenocarcinoma; TGCT: testicular germ cell tumors; THCA: thyroid carcinoma; THYM: thymoma; UCEC: uterine corpus endometrial carcinoma; UCS: uterine carcinosarcoma; UVM: uveal melanoma.

will require an improved understanding of the nature of their specific or abnormal changes across tissues from health to disease. However, our ability is hindered by the lack of a complete portrayal of lncRNA with spatial expression patterns. In this study, we collected four large-scale expression profiles of lncRNA compiled for all the major normal tissues of the human body, which allowed a comprehensive identification of TE lncRNAs. The tissue-to-tissue correlation of lncRNA expression across these resources revealed high similarity within the same tissue types and dominant distinction among different tissues, rather than similarity linked to other factors such as laboratory of origin, sampling procedures, or detection technologies. Based on expression profiles of protein-coding genes, previous studies had reported high similarity within the same tissues after analyzing the HPA and GTEx (Danielsson et al., 2015; Yu et al., 2015). Thus, the commonly used transcriptomes of normal tissues are valuable resources for lncRNAs, and not just for protein-coding gene analyses. Notably, analysis of the transcriptome by RNA-seq and CAGE is complementary, and it would be an attractive endeavor to integrate data obtained by these two approaches to refine gene expression models. Thus, our integrated analysis of these data resources provides the most comprehensive transcriptome of lncRNAs currently available, and allows the construction of a robust TE lncRNA map to explore the spatial expression patterns of lncRNAs.

The precise actions of lncRNAs are frequently dependent on their spatial expression patterns. TE lncRNAs were consistently detected in a genome-wide manner from four independent datasets. More than half of human lncRNAs display TE expression patterns across the human body. Functional analysis of these TE lncRNAs is well in line with their function in the respective tissue or organ. Moreover, about 80.61% of TE lncRNAs are found in at least two tissues, which is compatible with their function pleiotropy across different tissues based on their regulation of different biological functions. Indeed, calculation of the degree of shared TE lncRNAs based on the Jaccard index makes it clear that cancer types originating from similar tissues tend to share TE lncRNAs (Fig. S8). These results showed that TE lncRNAs might be interesting starting points for further in-depth studies to gain a better molecular understanding of the cellular phenotypes that define the function of each respective tissue and organ.

Combined with clinical studies, knowledge of TE lncRNAs provides a new means to generate hypotheses of the molecular basis of human disease. TE lncRNAs are expected to underlie many human diseases. Indeed, it has been shown that disease genes generally tend to be expressed in a limited number of tissues. Disentangling TE genes represented in large data compendia, along with their functions in specific tissues, offers means to address these challenges. To better understand the TE features of lncRNAs in disease states, we also identified TE lncRNAs in each cancer type. Interestingly, TE lncRNAs exhibited elevated expression in the corresponding cancer tissues compared to other cancer types, indicating that their higher expression might be a risk factor for disease. We used two criteria to prioritize the clinically related cancer lncRNAs (DTE lncRNAs and STE lncRNAs). The DTE lncRNAs tend to be up-regulated with cancer-constricted patterns, indicating their potential as diagnostic biomarkers. Similarly, TE lncRNAs tend to be related to patient survival in cancer-constricted patterns. Most of these DTE and STE lncRNAs have not yet been described, and could be good candidates for use in cancer diagnosis and prognostic analysis. Here, we provided an atlas of lncRNA marker candidates with elevated expression under both normal and disease conditions, and their clinical correlations (Fig. 7).

In addition, an awareness of the tissues that express elevated lncRNAs could allow a lncRNA found in whole blood or serum to be used as a circulating biomarker for a specific disease. For example, expression levels of TE lncRNAs prostate cancer antigen 3 (*PCA3*) and *PCAT18* are elevated in testis tissue under both normal and cancer conditions, and are much higher in tumor tissue. Expression of *PCA3* and *PCAT18* was also found to be up-regulated in blood samples of prostate cancer patients (Ren et al., 2013; Crea et al., 2014; Merola et al., 2015). The TE lncRNA *HULC* in liver tissue has also been proposed as a circulating biomarker for hepatocellular carcinoma, because it is up-regulated, and can be detected in blood (Panzitt et al., 2007). Furthermore, TE lncRNAs are susceptible to drug interference, which may provide entry points for the design of novel and specific therapeutics. For convenience, we also added TE lncRNAs that are known to act as circulating markers to our online

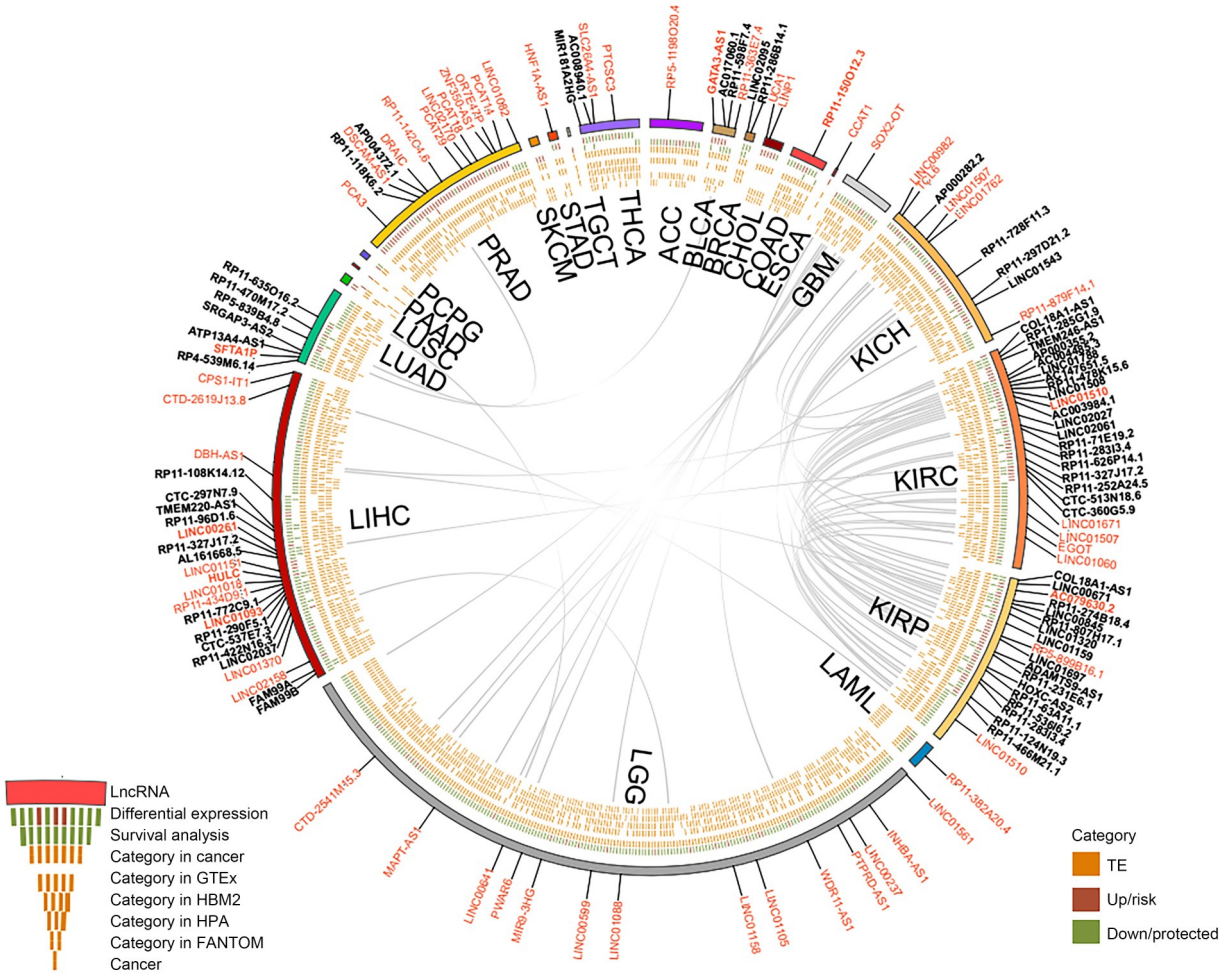

Fig. 7 TE lncRNAs that are elevated in both normal and cancer tissues and then are considered as differentially expressed TE (DTE) or survival-related TE (STE). Bold text indicates both DTE and STE lncRNAs; pink represents lncRNAs that are confirmed to be associated with specific diseases in three public databases. ACC: adrenocortical carcinoma; BLCA: bladder cancer; BRCA: breast invasive cancer; CHOL: cholangiocarcinoma; COAD: colon adenocarcinoma; ESCA: esophageal cancer; FANTOM: Functional Annotation of the Mammalian Genome; GBM: glioblastoma multiforme; GTEx: Genotype-Tissue Expression; HBM2: Human BodyMap 2; HPA: Human Protein Atlas; KICH: kidney chromophobe; KIRC: kidney renal clear cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LAML: acute myeloid leukemia; LIHC: liver hepatocellular carcinoma; LGG: lower grade glioma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; LncRNAs: long non-coding RNAs; PAAD: pancreatic adenocarcinoma; PCPG: pheochromocytoma/paraganglioma; PRAD: prostate adenocarcinoma; SKCM: skin cutaneous melanoma; STAD: stomach adenocarcinoma; TE: tissue-elevated; TGCT: testicular germ cell tumors; THCA: thyroid carcinoma.

resource LncSpA (Lv et al., 2020) using information from published databases, including Lnc2Cancer (Gao et al., 2019), LncRNADisease (Chen et al., 2013), and exoRBase (Li SL et al., 2018).

5 Conclusions

In summary, we used a framework to identify lncRNAs with spatial expression patterns based on an integrative transcriptome atlas of lncRNA across human tissues, and this framework can readily incorporate future lncRNA datasets to increase its utility. Spatial classification of lncRNAs is a powerful approach in the analysis of lncRNA function, and can be used as a starting point for elucidating the role of lncRNA in both the maintenance of tissue morphology and the progress of tissue-constricted diseases, crossing the gap between spatial expression patterns of lncRNAs and tissue physiology or pathology.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 31970646, 32060152, 32070673, and 32170676), the Hainan Province Science and Technology Special Fund (No. ZDYF2021SHFZ051), the Harbin Medical University Marshal Initiative Funding (No. HMUMIF-21024), the Marshal Initiative Funding of Hainan Medical University (No. JBGS202103), and the Heilongjiang Touyan Innovation Team Program.

Author contributions

Juan XU and Yongsheng LI designed the study. Kang XU, Xiyun JIN, and Ya LUO analyzed and interpreted the data. Kang XU, Haozhe ZOU, Dezhong LV, Liping WANG, and Yangyang CAI performed the cancer analysis. Kang XU, Limei FU, Tingting SHAO, and Yangyang CAI designed the figures. Kang XU, Juan XU, Yongsheng LI, and Tingting SHAO wrote and edited manuscript. All authors have read and approved the final manuscript, and therefore, have full access to all the data in the study and take responsibility for the integrity and security of the data.

Compliance with ethics guidelines

Kang XU, Xiyun JIN, Ya LUO, Haozhe ZOU, Dezhong LV, Liping WANG, Limei FU, Yangyang CAI, Tingting SHAO, Yongsheng LI, and Juan XU declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Aken BL, Achuthan P, Akanni W, et al., 2017. Ensembl 2017. *Nucleic Acids Res*, 45(D1):D635-D642. https://doi.org/10.1093/nar/gkw1104

Aran D, Camarda R, Odegaard J, et al., 2017. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun*, 8:1077. https://doi.org/10.1038/s41467-017-01027-z

Burgess DJ, 2015. Spatial characterization of proteomes. *Nat Rev Genet*, 16(3):129. https://doi.org/10.1038/nrg3910

Cabili MN, Trapnell C, Goff L, et al., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25(18):1915-1927. https://doi.org/10.1101/gad.17446611

Chen G, Wang ZY, Wang DQ, et al., 2013. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res*, 41(D1):D983-D986. https://doi.org/10.1093/nar/gks1099

Chen HY, Sun J, He YQ, et al., 2015. Expression and localization of testis developmental related gene 1 (TDRG1) in human spermatozoa. *Tohoku J Exp Med*, 235(2):103-109. https://doi.org/10.1620/tjem.235.103

Colaprico A, Silva TC, Olsen C, et al., 2016. *TCGAbiolinks*: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*, 44(8):e71. https://doi.org/10.1093/nar/gkv1507

Crea F, Watahiki A, Quagliata L, et al., 2014. Identification of a long non-coding RNA as a novel biomarker and potential therapeutic target for metastatic prostate cancer. *Oncotarget*, 5(3):764-774. https://doi.org/10.18632/oncotarget.1769

Danielsson F, James T, Gomez-Cabrero D, et al., 2015. Assessing the consistency of public human tissue RNA-seq data sets. *Brief Bioinform*, 16(6):941-949. https://doi.org/10.1093/bib/bbv017

Deras IL, Aubin SMJ, Blase A, et al., 2008. PCA3: a molecular urine assay for predicting prostate biopsy outcome. *J Urol*, 179(4):1587-1592. https://doi.org/10.1016/j.juro.2007.11.038

Derrien T, Johnson R, Bussotti G, et al., 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res*, 22(9):1775-1789. https://doi.org/10.1101/gr.132159.111

Du Z, Fei T, Verhaak RGW, et al., 2013. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol*, 20(7):908-913. https://doi.org/10.1038/nsmb.2591

Eisenberg E, Levanon EY, 2013. Human housekeeping genes, revisited. *Trends Genet*, 29(10):569-574. https://doi.org/10.1016/j.tig.2013.05.010

Fu PF, Zheng X, Fan X, et al., 2019. Role of cytoplasmic lncRNAs in regulating cancer signaling pathways. *J Zhejiang Univ-Sci B (Biomed & Biotechnol)*, 20(1):1-8. https://doi.org/10.1631/jzus.B1800254

Gan Y, Wang Y, Tan ZY, et al., 2016. TDRG1 regulates chemosensitivity of seminoma TCam-2 cells to cisplatin via PI3K/Akt/mTOR signaling pathway and mitochondria-mediated apoptotic pathway. *Cancer Biol Ther*, 17(7):741-750. https://doi.org/10.1080/15384047.2016.1178425

Gao Y, Wang P, Wang YX, et al., 2019. Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers. *Nucleic Acids Res*, 47(D1):D1028-D1033. https://doi.org/10.1093/nar/gky1096

Guo YB, Liu YY, Wang H, et al., 2021. Long noncoding RNA SRY-box transcription factor 2 overlapping transcript participates in Parkinson's disease by regulating the microRNA-942-5p/nuclear apoptosis-inducing factor 1 axis. *Bioengineered*, 12(1):8570-8582. https://doi.org/10.1080/21655979.2021.1987126

Gupta RA, Shah N, Wang KC, et al., 2010. Long non-coding RNA *HOTAIR* reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071-1076. https://doi.org/10.1038/nature08975

Hon CC, Ramilowski JA, Harshbarger J, et al., 2017. An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, 543(7644):199-204. https://doi.org/10.1038/nature21374

Hung T, Wang YL, Lin MF, et al., 2011. Extensive and coordinated transcription of noncoding RNAs within cell-cycle

promoters. *Nat Genet*, 43(7):621-629. https://doi.org/10.1038/ng.848

Jiang XZ, Li DJ, Yang JF, et al., 2011. Characterization of a novel human testis-specific gene: testis developmental related gene 1 (TDRG1). *Tohoku J Exp Med*, 225(4):311-318. https://doi.org/10.1620/tjem.225.311

Jongeneel CV, Delorenzi M, Iseli C, et al., 2005. An atlas of human gene expression from massively parallel signature sequencing (MPSS). *Genome Res*, 15(7):1007-1014. https://doi.org/10.1101/gr.4041005

Kim D, Pertea G, Trapnell C, et al., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, 14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36

Klijn C, Durinck S, Stawiski EW, et al., 2015. A comprehensive transcriptional portrait of human cancer cell lines. *Nat Biotechnol*, 33(3):306-312. https://doi.org/10.1038/nbt.3080

Li L, Eichten SR, Shimizu R, et al., 2014. Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol*, 15(2):R40. https://doi.org/10.1186/gb-2014-15-2-r40

Li SL, Li YC, Chen B, et al., 2018. exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res*, 46(D1):D106-D112. https://doi.org/10.1093/nar/gkx891

Li YS, Li LL, Wang ZS, et al., 2018. LncMAP: pan-cancer atlas of long noncoding RNA-mediated transcriptional network perturbations. *Nucleic Acids Res*, 46(3):1113-1123. https://doi.org/10.1093/nar/gkx1311

Li YS, Jiang TTF, Zhou WW, et al., 2020. Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers. *Nat Commun*, 11:1000. https://doi.org/10.1038/s41467-020-14802-2

Lingadahalli S, Jadhao S, Sung YY, et al., 2018. Novel lncRNA *LINC00844* regulates prostate cancer cell migration and invasion through AR signaling. *Mol Cancer Res*, 16(12):1865-1878. https://doi.org/10.1158/1541-7786.MCR-18-0087

Lonsdale J, Thomas J, Salvatore M, et al., 2013. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*, 45(6):580-585. https://doi.org/10.1038/ng.2653

Lv DZ, Xu K, Jin XY, et al., 2020. LncSpA: lncRNA spatial atlas of expression across normal and cancer tissues. *Cancer Res*, 80(10):2067-2071. https://doi.org/10.1158/0008-5472.CAN-19-2687

Merola R, Tomao L, Antenucci A, et al., 2015. PCA3 in prostate cancer and tumor aggressiveness detection on 407 high-risk patients: a National Cancer Institute experience. *J Exp Clin Cancer Res*, 34:15. https://doi.org/10.1186/s13046-015-0127-8

Newman AM, Liu CL, Green MR, et al., 2015. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*, 12(5):453-457. https://doi.org/10.1038/nmeth.3337

Noguchi S, Arakawa T, Fukuda S, et al., 2017. FANTOM5 CAGE profiles of human and mouse samples. *Sci Data*, 4:170112. https://doi.org/10.1038/sdata.2017.112

Panzitt K, Tschernatsch MMO, Guelly C, et al., 2007. Characterization of HULC, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding RNA. *Gastroenterology*, 132(1):330-342. https://doi.org/10.1053/j.gastro.2006.08.026

Qu L, He XY, Tang Q, et al., 2022. Iron metabolism, ferroptosis, and lncRNA in cancer: knowns and unknowns. *J Zhejiang Univ-Sci B (Biomed & Biotechnol)*, 23(10):844-862. https://doi.org/10.1631/jzus.B2200194

Ramsköld D, Wang ET, Burge CB, et al., 2009. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol*, 5(12):e1000598. https://doi.org/10.1371/journal.pcbi.1000598

Rawal HC, Angadi U, Mondal TK, 2021. *TEnGExA*: an R package based tool for tissue enrichment and gene expression analysis. *Brief Bioinform*, 22(3):bbaa221. https://doi.org/10.1093/bib/bbaa221

Ren SC, Wang FB, Shen J, et al., 2013. Long non-coding RNA metastasis associated in lung adenocarcinoma transcript 1 derived miniRNA as a novel plasma-based biomarker for diagnosing prostate cancer. *Eur J Cancer*, 49(13):2949-2959. https://doi.org/10.1016/j.ejca.2013.04.026

Rustici G, Kolesnikov N, Brandizi M, et al., 2013. ArrayExpress update-trends in database growth and links to data analysis tools. *Nucleic Acids Res*, 41(D1):D987-D990. https://doi.org/10.1093/nar/gks1174

Salem M, Paneru B, Al-Tobasei R, et al., 2015. Transcriptome assembly, gene annotation and tissue gene expression atlas of the rainbow trout. *PLoS ONE*, 10(3):e0121778. https://doi.org/10.1371/journal.pone.0121778

Schroth G, 2011. RNA-Seq of human individual tissues and mixture of 16 tissues (Illumina Body Map). BioStudies, E-MTAB-513. Retrieved from https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-513

She XW, Rohl CA, Castle JC, et al., 2009. Definition, conservation and epigenetics of housekeeping and tissue-enriched genes. *BMC Genomics*, 10:269. https://doi.org/10.1186/1471-2164-10-269

The Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al., 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, 45(10):1113-1120. https://doi.org/10.1038/ng.2764

Trapnell C, Roberts A, Goff L, et al., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, 7(3):562-578. https://doi.org/10.1038/nprot.2012.016

Uhlen M, Oksvold P, Fagerberg L, et al., 2010. Towards a knowledge-based Human Protein Atlas. *Nat Biotechnol*, 28(12):1248-1250. https://doi.org/10.1038/nbt1210-1248

Uhlén M, Fagerberg L, Hallström BM, et al., 2015. Tissue-based map of the human proteome. *Science*, 347(6220):1260419.

https://doi.org/10.1126/science.1260419

Uhlén M, Hallström BM, Lindskog C, et al., 2016. Transcriptomics resources of human tissues and organs. *Mol Syst Biol*, 12(4):862. https://doi.org/10.15252/msb.20155865

Wang DX, Eraslan B, Wieland T, et al., 2019. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol Syst Biol*, 15(2):e8503. https://doi.org/10.15252/msb.20188503

Wang Y, Gan Y, Tan Z, et al., 2016. *TDRG1* functions in testicular seminoma are dependent on the PI3K/Akt/mTOR signaling pathway. *OncoTargets Ther*, 9:409-420. https://doi.org/10.2147/OTT.S97294

Wei JC, Gan Y, Peng DY, et al., 2018. Long non-coding RNA H19 promotes TDRG1 expression and cisplatin resistance by sequestering miRNA-106b-5p in seminoma. *Cancer Med*, 7(12):6247-6257. https://doi.org/10.1002/cam4.1871

Xu J, Shao TT, Song MX, et al., 2020. MIR22HG acts as a tumor suppressor via TGFβ/SMAD signaling and facilitates immunotherapy in colorectal cancer. *Mol Cancer*, 19:51. https://doi.org/10.1186/s12943-020-01174-w

Yu NYL, Hallström BM, Fagerberg L, et al., 2015. Complementing tissue characterization by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium. *Nucleic Acids Res*, 43(14):6787-6798. https://doi.org/10.1093/nar/gkv608

Zhou CC, Yang F, Yuan SX, et al., 2016. Systemic genome screening identifies the outcome associated focal loss of long noncoding RNA PRAL in hepatocellular carcinoma. *Hepatology*, 63(3):850-863. https://doi.org/10.1002/hep.28393

Zhu JH, Chen G, Zhu SB, et al., 2016. Identification of tissue-specific protein-coding and noncoding transcripts across 14 human tissues using RNA-seq. *Sci Rep*, 6:28400. https://doi.org/10.1038/srep28400

Supplementary information

Tables S1–S3; Figs. S1–S8