Published in final edited form as:
Nat Struct Mol Biol. 2013 July ; 20(7): 908–913. doi:10.1038/nsmb.2591.

Integrative genomic analyses reveal clinically relevant long non-coding RNA in human cancer

Zhou Du[1,#], **Teng Fei**[2,3,4,5,#], **Roel G.W. Verhaak**[6], **Zhen Su**[7], **Yong Zhang**[1], **Myles Brown**[2,3,4,*], **Yiwen Chen**[5,#,*], and **X. Shirley Liu**[2,5,*]

[1]Department of Bioinformatics, School of Life Sciences and Technology, Tongji University, Shanghai, China

[2]Center for Functional Cancer Epigeneitcs, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[3]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[4]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[5]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts, USA

[6]Department of Bioinformatics and Computational Biology, Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

[7]College of Biological Sciences, China Agriculture University, Beijing, China

Abstract

Despite growing appreciations of the importance of long non-coding RNA (lncRNA) in normal physiology and disease, our knowledge of cancer-related lncRNA remains limited. By repurposing microarray probes, we constructed the expression profile of 10,207 lncRNA genes in approximately 1,300 tumors over four different cancer types. Through integrative analysis of the lncRNA expression profiles with clinical outcome and somatic copy number alteration (SCNA), we identified lncRNA that are associated with cancer subtypes and clinical prognosis, and predicted those that are potential drivers of cancer progression. We validated our predictions by experimentally confirming prostate cancer cell growth dependence on two novel lncRNA. Our analysis provided a resource of clinically relevant lncRNA for development of lncRNA biomarkers and identification of lncRNA therapeutic targets. It also demonstrated the power of integrating publically available genomic datasets and clinical information for discovering disease associated lncRNA.

Systematic efforts to catalogue long non-coding RNA (lncRNA) using traditional cDNA Sanger sequencing[1], histone mark ChIP-seq[2, 3], or RNA-seq[4, 5] data revealed that the human genome encodes over 10,000 lncRNA with little coding capacity. Growing evidences

*To whom correspondence should be addressed: X. Shirley Liu (xsliu@jimmy.harvard.edu) or Yiwen Chen (ywchen@jimmy.harvard.edu) or Myles Brown (myles_brown@dfci.harvard.edu).
#These authors contributed equally

Author Contributions

YC conceived the project. ZD and YC designed the algorithms and performed computational analyses. RVG contributed to the subtype analyses of ovarian cancer. TF performed all the experimental validation. ZD, TF, ZS, YZ, MB, YC and XSL participated in the discussions and contributed to the analysis of the intermediate results throughout the project. YC, MB and XSL supervised the project. YC and XSL wrote the manuscript with the help from other co-authors.

suggest that in cancer lncRNA, like protein-coding genes (PCGs), may mediate oncogenic or tumor suppressing effects and promise to be a new class of cancer therapeutic targets[6]. While a handful of lncRNA have been functionally characterized, little is known about the function of most lncRNA in normal physiology or disease[7]. LncRNA may also serve as cancer diagnostic or prognostic biomarkers that are independent of PCG. A well-known example of a potential cancer diagnostic biomarker is *PCA3*, a prostate-specific lncRNA gene that is significantly overexpressed in prostate cancer. Noninvasive monitoring of urinary *PCA3* transcript level is currently being developed for diagnostics in the clinic[8].

As lncRNA do not encode proteins, their functions are closely associated with their transcript abundance. RNA-seq is a comprehensive way to profile lncRNA expression. However, due to the higher cost associated with the adoption of this technique, publically available RNA-seq datasets of tumors are relatively limited compared with array-based expression profiles. In addition, RNA-seq datasets with low sequencing coverage or small sample numbers have only limited statistical power to discover clinically relevant lncRNA. In contrast, there are a large number of datasets that contain array-based gene expression profiles across hundreds of tumor samples. These array-based expression profiles are often accompanied with matched clinical annotation and/or somatic genomic alteration profiles such as somatic copy number alteration (SCNA). Although lncRNA are not the intended targets of measurement in the original array design, microarray probes can be re-annotated for interrogating lncRNA expression[9-14]. Compared with RNA-seq data of low sequencing coverage, array-based expression data may have lower technical variation and better detection sensitivity for low-abundance transcripts[15, 16], a prominent feature of lncRNA[5]. Moreover, array-based expression data contain strand information and allow for interrogating expression of anti-sense single-exon lncRNA, whereas most of current RNA-seq data in clinical applications do not have strand information and thus are unable to accurately quantify the expression of this class of lncRNA[17].

To repurpose the publically available array-based data to interrogate lncRNA expression in tumor samples, we developed a computational pipeline to re-annotate the probes that are uniquely mapped to lncRNA using the latest annotations of lncRNA and PCG. We further performed integrative genomic analyses of lncRNA expression profiles, clinical information and SCNA profiles of tumors in four different cancer types including 150 tumor samples of prostate cancer from the MSKCC Prostate Oncogenome Project[18] and 451 tumor samples of glioblastoma multiforme (GBM), 585 tumor samples of ovarian cancer (OvCa) and 113 tumor samples of lung squamous cell carcinoma (Lung SCC) from the Cancer Genome Atlas Research Network (TCGA) project[19]. We identified lncRNA that are significantly associated with cancer subtypes or cancer prognosis and predicted those that may play tumor promoting or suppressing function.

Results

Repurposing microarray data for probing lncRNA expression

Among the different gene expression microarray platforms, we focused on reannotating the probes from Affymetrix microarrays. These arrays not only have many more short probes that are likely to map to lncRNA genes, but have been the most widely used platforms for gene expression profiling of patient tumor samples. We designed a computational pipeline to re-annotate the probes from five Affymetrix array types (**Methods**, Fig. 1a), and kept annotated lncRNA and PCG transcripts with at least 4 probes uniquely mapped to them. Among the five Affymetrix array types, Affymetrix Human Exon 1.0 ST array has the most comprehensive coverage of the annotated human lncRNA (Supplementary Table 1). In total, 10,207 lncRNA genes have at least 4 probes covering their annotated exons (Fig. 1a), which constitute approximately 64% of all 15,857 lncRNA genes (with over 60% coverage in each

category[20] of lncRNA genes) collected in this study (**Methods**, Fig. 1b,c, Supplementary Table 2). We focused our studies on the Affymetrix exon-array-expression profiles because of its most comprehensive coverage of lncRNA.

We used a model-based method[21] (**Methods**) to derive the gene expression index of all the PCGs and lncRNA on exon arrays. To gauge the reliability of our approach, we examined the correlation of both lncRNA and PCG expression between exon array and RNA-seq data that were generated from two different laboratories using the same prostate cancer cell line LNCaP[18, 22]. We found that both PCG (r=0.70, p<2.2x10^{-16}) and lncRNA (r=0.29, p<2.2x10^{-16}) showed significant concordance of expression between exon array and RNA-seq data (Supplementary Fig. 1a,b). This observation is consistent with the previous finding that the correlation between microarrays and RNA-seq is lower in lowly-expressed genes[23], as lncRNA generally are expressed at lower levels than PCG[5]. As the level of probe coverage could also influence the accuracy of lncRNA expression derived from microarray, we further investigated how the correlation of expression between exon array and RNA-seq changes at different probe-coverage by examining those PCGs that have similar expression level to that of lncRNA (Supplementary Fig. 1c). We found that the correlation between exon-array and RNA-seq based expression showed a moderate increase when all probes (0.28) were used compared with when only 4 probes (0.20) were used (Supplementary Fig. 1c). The correlations were similar for PCGs (0.28) and lncRNA (0.29) when the expression level was controlled for. These results suggest that although the probe coverage may influence the array-based lncRNA expression estimation, the dominant factor that governs the observed difference in correlation between array and RNA-seq for PCGs versus lncRNA is their expression level. A recent study, in which a 60-mer custom oligonucleotide array was designed to investigate lncRNA expression, showed that the correlation of lncRNA expression between the custom array and RNA-seq data was between 0.24 and 0.31[20]. Therefore, although the concordance between exon arrays and RNA-seq is lower for lncRNA than for PCG expression, it may represent the typical performance when comparing lncRNA expression between an array-based platform and RNA-seq.

LncRNA associated with cancer status, subtype and prognosis

To validate the utility of exon array data in combination with clinical annotation to identify cancer-related lncRNA, we examined the expression pattern of thirteen literature-curated cancer-related lncRNA[6] that have corresponding exon array probes in a prostate cancer dataset[18]. This dataset consists of 29 normal prostate samples, 131 primary and 19 metastatic prostate tumor samples with exon array data[18] (Fig. 2a). Interestingly, nine out of these thirteen known cancer-related lncRNA showed significantly differential expression between tumor and normal prostate samples (Mann-Whitney U test, p<0.05). Three out of these nine lncRNA were directly related to prostate cancer, including one known prostate cancer diagnostic biomarker *PCA3*[8], and two, *PCAT-1*[22] and *PCGEM1*[24], that have been functionally implicated in prostate cancer progression. *GAS5*, a tumor-suppressive lncRNA known to be down regulated in breast cancer[25], showed increased expression in prostate cancer (Table 1), suggesting complex and context-dependent functions of lncRNA in different cancer types. Interestingly, several lncRNA such as *NEAT1*[26], *DANCR*[27], *HOTTIP*[28], *PRINS*[29], and *EGOT*[30] that have established functions in forming nuclear speckles[26], in development[27] or in autoimmune disease[29], but were not previously known to be related to cancer, showed differential expression between tumor and normal prostate samples (Table 1), suggesting their potential function in prostate cancer.

We next sought to identify lncRNA that showed significant expression difference between tumors and normal prostate tissues, and found 109 up- and 104 down-regulated lncRNA (Mann-Whitney U test, false discovery rate (FDR) <0.05, fold-change≥1.5), respectively

(Fig. 2a). Notably, among the lncRNA with sufficient exon-array probe coverage, we re-discovered 7 out of 8 lncRNA which were reported to show higher expression in prostate cancer from an independent study based on RNA-seq data[22]. Furthermore, we identified an additional 102 lncRNA genes which were up-regulated in prostate cancer, but were missed by the other study[22], suggesting that arrays and RNA-seq may be complementary methods to identify clinically relevant lncRNA. When a lncRNA acts in *cis* and influences the expression of its neighboring PCG or a lncRNA and its neighboring PCG are under the same *cis*-regulation, they can show coordinated expressions. We compared the distribution of the correlation between lncRNA and its neighboring PCG from different lncRNA classes (Supplementary Table 2). Interestingly, antisense genic lncRNAs are slightly better correlated with their sense PCG than intergenic lncRNA (pvalue < 10^{-10}) (Supplementary Fig. 2a,b), suggesting a more co-coordinated expression between sense PCG and anti-sense lncRNA gene.

Cancer is a clinically heterogeneous disease and individual cancer types can be further divided into molecular subtypes, each with its specific biological and clinical behavior. Previous studies established 4 subtypes of GBM (proneural, neural, classical and mesenchymal subtype)[19], 4 subtypes of OvCa (immunoreactive, proliferative, mesenchymal and differentiated subtype)[31], 4 subtypes of Lung SCC (basal, classical, primitive and secretory subtype)[32] based on the expression profile of PCG, and 6 subtypes of prostate cancer based on the SCNA profiles[18].

LncRNA with subtype-specific expression may have important function in individual molecular subtypes. We compared the lncRNA expression across different subtypes (**Methods**) and identified hundreds of lncRNA showing subtype-specific expression patterns in GBM, OvCa and Lung SCC (Fig. 2b-e). The same approach did not yield any lncRNA that show significant subtype-specific expression in prostate cancer, which was a reminiscence of the lack of robust PCG-expression-based subtype of prostate cancer[18]. In addition, 628 lncRNA showed subtype-specific expression in more than one cancer type (Fig. 2b) and some of them have been functionally implicated in other physiological or pathological processes. For example, *MIAT*, a lncRNA which showed specific expression in mesenchymal subtype of OvCa and in proneural subtype of GBM, is known to confer risk of myocardial infarction[33] and regulate retinal cell fate specification[34]. Another example is *RMST*, a lncRNA known to be differentially expressed between rhabdomyosarcoma subtypes[35], also exhibited subtype-specific expression patterns in GBM, OvCa and Lung SCC.

A previous study of *HOTAIR*[36, 37] showed that patients with higher *HOTAIR* expression had poorer prognosis in colorectal cancer[38]. To identify the lncRNA, which are associated with clinical outcome in prostate cancer, GBM, OvCa and Lung SCC, we performed multivariate Cox regression analysis to evaluate the significance of correlation between individual lncRNA expression and overall- and progression-free survival in the presence of other confounding factors such as ethnicity, age and gender (**Methods**). We identified approximately 100 lncRNA each in prostate cancer, GBM, OvCa and Lung SCC, whose expression was significantly correlated with overall or progression-free survival ($p < 0.01$, Supplementary Table 3). Interestingly, nine lncRNA showed consistent positive or negative correlation between their expression and overall or progression-free survival in different cancer types, suggesting their potential as more general prognostic biomarkers. The lncRNA gene, with the Ensembl ID ENSG00000261582 is an example of a lncRNA that showed negative correlation between its expression and overall survival in both Lung SCC and OvCa (Fig. 3a). This lncRNA also showed subtype-specific expression in OvCa, but not in Lung SCC. Additionally, five lncRNA showed significant and consistent positive or

negative correlation between both overall and progression-free survival in OvCa and one such example (Ensembl ID ENSG00000225128) was shown (Fig. 3b).

Predicting lncRNA that are potential cancer drivers

An important form of somatic genetic alteration in cancer is SCNA, in which a genomic region is either amplified or deleted. Some of the genes within amplified (or deleted) regions show increased (or decreased) expression level, leading to altered activity in cancer cells. Studies suggest that the genes playing causal roles in oncogenesis are often located in the SCNA that are altered frequently across tumors[39, 40]. To reveal the lncRNA that may play tumor promoting or suppressing function, we identified hundreds of lncRNA that map to regions of recurrent SCNA across tumors (**Methods**) for prostate cancer, GBM, OvCa and Lung SCC (Fig. 3c). Some of these lncRNA also showed significant correlation between overall or progression-free survival (Supplementary Table 4). In addition, we identified lncRNA that were consistently located in the regions of SCNA across different cancers (Fig. 3c) and found a significant overlap of the lncRNA genes that are in the SCNA gain or loss regions between some of cancer types (Supplementary Table 5).

Among the many genes located within regions of SCNA, only a fraction of them are likely to be drivers of cancer. To further distinguish driver from passenger lncRNA in the regions of SCNA, we integrated SCNA and expression profiles of lncRNA in tumors. We reasoned that driver lncRNA with SCNAs should result in corresponding gene expression changes[40, 41], as only those SCNAs that cause the change of transcript abundance could possibly alter lncRNA activity. Therefore, we selected lncRNA whose SCNAs showed positive correlation with expression level change as the candidate drivers (**Methods**, Supplementary Table 3) for prostate cancer, GBM, OvCa and Lung SCC.

Experimental validation of two novel lncRNA

As it is prohibitive to validate all the candidate driver lncRNA in four cancer types, we focused our experimental validation on candidate lncRNA that may have tumor promoting function in prostate cancer (i.e. those in recurrent SCNA (gain) regions, which showed positive correlation between their SCNAs and expression level). Among all the candidate driver lncRNA that showed increasing expression from normal to primary to metastatic prostate cancer, we chose the two that showed most significant expression difference between tumor and normal prostate tissue (i.e. the smallest *p*- value from Mann-Whitney U test) for experimental validation. We named these two lncRNA as Prostate Cancer Associated Non-coding RNA 1 and 2, abbreviated as PCAN-R1 (Ensembl ID ENSG00000228288) and PCAN-R2 (Ensembl ID ENSG00000231806), respectively. Both lncRNA showed positive correlation between gene expression and SCNA (Fig. 4a and Supplementary Fig. 3a). The criteria of increasing expression from normal to primary to metastatic prostate cancer aimed to uncover lncRNA that may be important therapeutic targets for both primary and metastatic cancers (Fig. 4b). Coding potential analysis confirmed the non-coding nature of these two LncRNA (**Methods**). We chose the prostate cancer cell line LNCaP for experimental validation in which both lncRNA have moderate or higher expression level compared with other prostate cancer or non-prostate cancer cell lines (Supplementary Fig. 3b,c). Using 5′ and 3′ RACE, we found that for PCAN-R1, while one isoform PCAN-R1-A was almost identical to the Ensembl annotated transcript ENST00000425295 (Fig. 4c, Supplementary Fig. 4a), the other isoform PCAN-R1-B was a spliced variant of PCAN-R1-A with an intron retention (Fig. 4c). Interestingly, for PCAN-R2, the major isoform had an extra exon in the 5′ end, and the remaining two exons also had different lengths from the Ensembl annotation (Fig. 4c, Supplementary Fig. 4b). The new 5′ exon of PCAN-R2 was more consistent with the profile of H3K4me3, a histone mark of active promoter and the profile of DNase I hypersensitive regions (i.e. the regions with an

open chromatin state) in LNCaP cells. We further confirmed the transcript structure of PCAN-R1 and PCAN-R2 by Northern blot (Fig. 5a, **Methods**).

Based on the determined lncRNA transcript structures, we designed siRNAs that targeted the common exon of each lncRNA gene. Notably, knockdown of either PCAN-R1 or PCAN-R2 using two different siRNA (Fig. 5b) resulted in substantial decrease in both cell growth (Fig. 5c) and soft-agar colony formation in the androgen-dependent prostate cancer cell line LNCaP (Fig. 5d). We further confirmed this growth inhibition upon lncRNA knockdown in the androgen-independent prostate cancer cell line LNCaP-abl (Supplementary Fig. 5a,b). To rule out the possibility that the observed phenotypes were from siRNA off-target effect on PCG expression, we searched the homologous sequences of the designed siRNA sequences in all protein-coding transcripts. We found no hit with the perfect match or one mismatch, and only found five transcripts from five genes for all the siRNAs when two mismatches were allowed. Among these, two PCGs *MYSM1* (potentially targeted by siR1-1) and *ADAMTS17* (potentially targeted by siR2-2) showed elevated expression in prostate tumors than in normal samples, which resembled PCAN-R1 and PCAN-R2 in terms of expression pattern and accordingly indicated their potentials of functionality. The expression of these two genes were unaffected upon corresponding siRNA treatment, suggesting that the observed cellular phenotype upon siRNA knockdown of the selected lncRNA was unlikely to be from off-target effect on PCGs (Supplementary Fig. 5c,d).

As a lncRNA may act in *cis* and influence the expression of its neighboring PCG, we investigated whether the expression of the neighboring PCG was regulated by PCAN-R1 or PCAN-R2. The siRNA knockdown of PCAN-R1 or PCAN-R2 had no effect on the expression of their neighboring PCG *KDM5B* and *FBP2* (Supplementary Fig. 5c,d) respectively, suggesting that the functional mechanism of PCAN-R1/-R2 are not directly through their neighboring PCG. Interestingly, in normal tissues, PCAN-R1 and its neighboring PCG *KDM5B* showed the highest expression in testis (Supplementary Fig. 5e,f). In contrast, while PCAN-R2 showed a similar expression across different tissues, its neighboring PCG *FBP2* exhibited muscle-specific expression pattern (Supplementary Fig. 5e,f), suggesting that the expression of PCAN-R2 and *FBP2* may be differently regulated.

Discussion

Our analyses demonstrated that repurposing microarray probes to construct lncRNA expression profile in patient sample is a cost-effective approach, given the large number of such datasets available in public repositories. The constructed gene expression profiles of both lncRNA and PCGs from our analyses is a valuable resource for understanding the similarity and difference of transcriptional (e.g. antisense RNA[42]) regulation of PCGs by lncRNA across different cancer types. In combination of matched SCNA profile and clinical information, these gene expression profiles also allow for inferring network models[43, 44] that will help to advance the understanding of lncRNA function in cancer etiology.

More importantly, the experimental validation of two lncRNA without previous implication in cancer suggested the effectiveness of our integrative analyses in finding functionally important lncRNA in cancer. Our analyses predicted about 80 to 300 candidate driver lncRNA that may have tumor promoting function in each of four cancers, respectively. An intersection of such list of candidate driver lncRNA with the list of lncRNA generated from orthogonal functional genomic datasets such as ribonucleoprotein immunoprecipitation (RIP) followed by sequencing (RIP-seq)[45] data (a genomic technique for identifying lncRNA physically associated with the protein of interest), would greatly help to prioritize

their functional valuation in different biological context including epigenetic regulation and facilitate the discovery of lncRNA therapeutic targets.

In current study, we only utilized SCNA and expression data in combination with clinical information for our integrative analysis. It is conceivable that other types of genomic data such as SNP array[46] and genome sequencing data[47] can be further integrated to reveal the multifaceted relationship between mutation spectrum and expression of lncRNA, disease status, and clinical outcome.

In summary, our study represents a proof-of-principle study for identifying clinically relevant lncRNA through integrative analyses of orthogonal genomic datasets and clinical information. It opens new avenues for leveraging publicly available genomic data to study the functions and mechanisms of lncRNA in human diseases.

Methods

Repurposing data from Affymetrix Human Exon 1.0 ST array and Affymetrix 3′ IVT arrays to interrogate lncRNA expression

We collected long non-coding RNA (lncRNA) annotation from two sources: the catalogue of lncRNA from Ensembl database[48] (Homo sapiens GRCh37, release 67) and the catalog of lncRNA generated based on transcriptome assembly from RNA-seq data[5]. For those lncRNA transcripts that have overlap on the same strand between these two sources, we only kept the Ensembl annotation (Fig. 1a) to avoid redundancy. This resulted in a total of 15,857 lncRNA genes. We re-annotated probe sets of Human Exon array for lncRNA by mapping all probes to the human genome (hg19) using SeqMap[49]. We kept those probes that were uniquely mapped to the genome with no mismatch. We then removed all probes that were mapped to protein-coding transcripts (183,252) and pseudogene transcripts (15,789) based on the annotations from Ensembl[48] (http://www.ensembl.org) and UCSC[50] (http://www.genome.ucsc.edu) database. By matching the rest probes to lncRNA sequences, we obtained 202,449 probes and 10,207 corresponding lncRNA genes with at least 4 probes. The same strategy was applied to generate the probes that correspond to lncRNA transcripts for other 3′ IVT Affymetrix array platforms. The raw intensity of exon array probes was corrected using a probe-sequence-specific background model and the expression level of a lncRNA gene was calculated by summarizing the background-corrected intensity of all probes corresponding to this gene[21]. The lncRNA expression was quantile-normalized across different biological samples. The gene expression calculation was implemented using Jetta[51]. When the batch information is available, Combat[52], an empirical Bayes method was used to remove potential batch effect.

LncRNA classification

The classification scheme was adopted from Derrien et al[20]. The lncRNA were categorized into intergenic and genic ones. The neighboring protein-coding genes of lncRNA were selected based on (1) The nearest distance of the lncRNA or (2) the longest overlapped regions. The intergenic lncRNA were sub-classified as “same strand”, “convergent” and “divergent” according to their relative orientation with the neighboring protein-coding genes. The genic lncRNA were classified as being exonic, intronic and overlapping and sense and antisense according to their relation with neighboring protein-coding genes. The classification of all lncRNA genes that have at least 4 exon-array probes was listed as a supplementary dataset (supplementary-file1.xlsx)

Comparative analysis of exon array and RNA-seq data

We obtained RNA-seq[22] and exon array[18] data of LNCaP cell line from two different studies[18, 22]. The RNA-seq-based gene expression was calculated using Cufflinks1.0.2[53](default parameter and -G option) and the exon-array-based gene expression was calculated using the same procedure as was described in the last section. The Pearson correlation coefficient was used to quantify the strength of the association between exon-array-based and RNA-seq-based expression.

Exon array data, clinical annotation and SCNA data of different cancers

For prostate cancer, we obtained exon array data, clinical annotation and SCNA data generated by the MSKCC Prostate Oncogenome Project[18] from Gene expression Omnibus (GEO) (GSE21034). This dataset included 29 normal adjacent, 131 primary and 19 metastatic tissue specimens as well as 4 prostate cell lines with exon-array data. The exon array data, clinical annotation and SCNA data of 451 GBM[19], 585 OvCa[31] and 113 Lung SCC[32] primary tumors were downloaded from The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov). We further obtained exon array data of 11 human normal tissues from Affymetrix (http://www.affymetrix.com/).

Identifying lncRNA associated with overall- and progression-free survival or cancer subtype

We performed multivariate Cox proportional hazard (Cox regression) analysis to assess the association between different covariates including lncRNA expression, ethnicity, age and gender with overall or progression-free survival. In addition to lncRNA expression, we only included the clinical outcome and covariate data that were available in individual dataset for analysis. For GBM and Lung SCC, we included ethnicity, age and gender, whereas for prostate cancer and OvCa, we only included ethnicity and age as additional covariates. The Cox regression analyses were performed for overall survival in GBM and Lung SCC, for progression-free survival in prostate cancer, and for both overall and progression-free survival in OvCa. The molecular subtype information of 220, 487, 89 and 150 tumor samples from GBM, OvCa, Lung SCC and prostate cancer were obtained from previous studies[18,19,31,32]. One tailed Mann-Whitney U test was performed to compare the lncRNA expression in each subtype with other subtypes in the same cancer. The lncRNA that showed statistically higher expression (FDR < 0.05) in only one subtype were considered as subtype-specific ones.

Identifying candidate driver lncRNA by integrating SCNA and expression data

The recurrent somatic copy number alteration (SCNA) regions of prostate cancer, GBM, HGS-OvCa and Lung SCC were identified using GISTIC[54,55] or RAE[56] algorithm in previous studies[19,31,32]. For prostate cancer, the SCNA regions were determined as the union of SCNA regions from two different studies[18,39]. The magnitude of SCNAs was estimated as log2 ratios of segmented copy number between cancer and control DNAs. Among the lncRNA in the SCNA regions, we selected those that showed significant and concordant expression change (one tailed Mann-whitney U test, $p<0.05$) in tumor samples with corresponding somatic copy number gain (log2 ratio > 0.2) or loss (log2 ratio <-0.2), in comparison with the other samples (Supplementary Table 2).

Coding potential analysis

To confirm that the two lncRNA PCAN-R1 and PCAN-R2 are non-coding, we used two different methods, txCdsPredict from UCSC and phyloCSF[57] to calculate their coding potential. For coding potential calculation with phyloCSF, we used the multiple sequence alignment of 29 mammalian genomes[58]. We chose the thresholds used previously

(txCdsPredict = 800[22] and phyloCSF = 100[5]), below which the transcripts were considered non-coding. We found that the scores of all possible opening reading frames (ORFs) from PCAN-R1 and PCAN-R2 transcripts were well below the thresholds (txCdsPredict score: PCAN-R1, 470 and PCAN-R2, 359; phyloCSF score: PCAN-R1, -123.1434 and PCAN-R2, -148.5448), supporting that these two lncRNA genes are non-coding.

Cell culture

LNCaP, CWR22Rv1 and PC3 cells were cultured in PRMI 1640 medium supplemented with 10% fetal bovine serum (FBS). LNCaP-abl and LNCaP-AI cells were maintained in phenol red-free RPMI 1640 medium with 10% charcoal/dextran-treated FBS. VCaP, Hela and 293T cells were grown in DMEM with 10% FBS.

Quantitative RT-PCR (qRT-PCR) analysis

RNA was isolated using RNeasy Mini Kit (Qiagen). Reverse transcriptase (Invitrogen) was employed for random-primed first-strand complementary DNA (cDNA) synthesis. Real-time PCR was carried out on ABI Prism 7300 detection system using SYBR Green PCR master mix. The ΔΔCt method was used to comparatively quantify the amount of mRNA level. RPS28 gene expression served as the internal control. Primer sequences are listed below: RPS28, 5′-CGATCCATCATCCGCAATG-3′(forward) and 5′-AGCCAAGCTCAGCGCAAC-3′(reverse); PCAN-R1, 5′-CAGGAACCCCCTCCTTACTC-3′(forward) and 5′-CTAGGGATGTGTCCGAAGGA-3′ (reverse); PCAN-R2, 5′-CTTGGCTGTGGTCACTCTGA-3′ (forward) and 5′-ACACACAGTTGGGTTCACCA-3′ (reverse); KDM5B, 5′-ATTGCCTCAAAGGAATTTGGCAGTG-3′(forward) and 5′-CATCACTGGCATGTTGTTCAAATTC-3′(reverse); MYSM1, 5′-CAAATCAGAAGACCGGCCATAA-3′(forward) and 5′-GCACGTCCCCTTAAACATGATG-3′(reverse); FBP2, 5′-GGGTCAAGCATGAAGAGGTC-3′ (forward) and 5′-CAGAGGATGAGCCTTCTGAAA-3′ (reverse); ADAMTS17, 5′-ACGACAACGTCCCGCTAAG-3′ (forward) and 5′-TCCTCCATACTCCTCGTTCTG-3′ (reverse);

RACE and northern blot analysis

5′ and 3′ rapid amplification of cDNA ends (RACE) were performed using the RLM-RACE Kit (Ambion) following the manufacturer's manual. Northern blot were performed using DIG Northern Starter Kit (Roche). Digoxigenin (DIG) labeled RNA probes were generated by *in vitro* transcription with T7 RNA polymerase from PCR products of corresponding regions to detect specific lncRNA transcripts in poly(A) enriched mRNAs of LNCaP cells. The PCR primers used to amplify specific regions for northern probes are listed below: PCAN-R1, 5′-GACCTGGGCAACCCCAGCCTG-3′(sense) and 5′-GATCACTAATACGACTCACTATAGGGCGCGAGGAGCGCCTCATCACC-3′ (antisense, including T7 promoter sequence); PCAN-R2, 5′-GACAAATTCACCAAGAGCCTAG-3′(sense) and 5′-GATCACTAATACGACTCACTATAGGGGAAGACTATGGGCTGCTTCCTT-3′ (antisense, including T7 promoter sequence).

RNA interference

The siRNA oligos were synthesized by Dharmacon, Inc. The target sequences are as follows: siControl, 5′-GCGACCAACGCCTTGATTG-3′; siR1-1, 5′-GGTGTCTCCATCCTCATTC-3′; siR1-2, 5′-CTCCCAGACCTCACGTCAA-3′; siR2-1, 5′-ACAGGAAGCTCTAGCAGTA-3′; siR2-2, 5′-CCATCAACAGTGAGAGGAA-3′.

Cells were transfected with 20nM siRNA oligos by RNAiMax reagent (Invitrogen) in 24-well plate. The knockdown efficiency was determined by qRT-PCR at 48-72 hours post transfection.

Cell growth and soft agar assay

For cell growth assay, cells were plated in 24-well plates, transfected with indicated siRNA oligos in triplicate and allowed to grow for another 5 days. Cells were counted every other day by a hemocytometer. Anchorage-independent cell growth in soft agar was performed in triplicate with 10,000 LNCaP cells per well suspended in 1.5 ml of medium containing 0.35% agar spread on top of 1.5 ml of 0.7% solidified agar in six-well plates. Colonies were stained with crystal violet and counted after four weeks plating. Data were shown as Mean +S.D. n=3.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially funded by National Natural Science Foundation of China [31028011] (X.S.L.), National Basic Research (973) Program of China [2010CB944904] (Y.Z.), and National Institutes of Health of US GM099409 (X.S.L.).

References

1. Ota T, et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nat Genet. 2004; 36:40–45. [PubMed: 14702039]
2. Guttman M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009; 458:223–227. [PubMed: 19182780]
3. Khalil AM, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A. 2009; 106:11667–11672. [PubMed: 19571010]
4. Guttman M, et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol. 2010; 28:503–510. [PubMed: 20436462]
5. Cabili MN, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25:1915–1927. [PubMed: 21890647]
6. Prensner JR, Chinnaiyan AM. The emergence of lncRNAs in cancer biology. Cancer Discov. 2011; 1:391–407. [PubMed: 22096659]
7. Wapinski O, Chang HY. Long noncoding RNAs and human disease. Trends Cell Biol. 2011; 21:354–361. [PubMed: 21550244]
8. Lee GL, Dobi A, Srivastava S. Prostate cancer: diagnostic performance of the PCA3 urine test. Nat Rev Urol. 2011; 8:123–124. [PubMed: 21394175]
9. Liao Q, et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res. 2011; 39:3864–3878. [PubMed: 21247874]
10. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A. 2008; 105:716–721. [PubMed: 18184812]
11. Michelhaugh SK, et al. Mining Affymetrix microarray data for long non-coding RNAs: altered expression in the nucleus accumbens of heroin abusers. J Neurochem. 2010; 116:459–466. [PubMed: 21128942]
12. Gellert P, Ponomareva Y, Braun T, Uchida S. Noncoder: a web interface for exon array-based detection of long non-coding RNAs. Nucleic Acids Res. 2013; 41:e20. [PubMed: 23012263]

13. Johnson R. Long non-coding RNAs in Huntington's disease neurodegeneration. Neurobiol Dis. 2012; 46:245–254. [PubMed: 22202438]
14. Zhang X, et al. Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiol Dis. 2012; 48:1–8. [PubMed: 22709987]
15. Raghavachari N, et al. A systematic comparison and evaluation of high density exon arrays and RNA-seq technology used to unravel the peripheral blood transcriptome of sickle cell disease. BMC Med Genomics. 2012; 5:28. [PubMed: 22747986]
16. Xu W, et al. Human transcriptome array for high-throughput clinical studies. Proc Natl Acad Sci U S A. 2011; 108:3707–3712. [PubMed: 21317363]
17. Levin JZ, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. 2010; 7:709–715. [PubMed: 20711195]
18. Taylor BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell. 2010; 18:11–22. [PubMed: 20579941]
19. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008; 455:1061–1068. [PubMed: 18772890]
20. Derrien T, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 2012; 22:1775–1789. [PubMed: 22955988]
21. Kapur K, Xing Y, Ouyang Z, Wong WH. Exon arrays provide accurate assessments of gene expression. Genome Biol. 2007; 8:R82. [PubMed: 17504534]
22. Prensner JR, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol. 2011; 29:742–749. [PubMed: 21804560]
23. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10:57–63. [PubMed: 19015660]
24. Petrovics G, et al. Elevated expression of PCGEM1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. Oncogene. 2004; 23:605–611. [PubMed: 14724589]
25. 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. [PubMed: 18836484]
26. Clemson CM, et al. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol Cell. 2009; 33:717–726. [PubMed: 19217333]
27. Kretz M, et al. Suppression of progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev. 2012; 26:338–343. [PubMed: 22302877]
28. Wang KC, et al. A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature. 2011; 472:120–124. [PubMed: 21423168]
29. Szegedi K, et al. The anti-apoptotic protein G1P3 is overexpressed in psoriasis and regulated by the non-coding RNA, PRINS. Exp Dermatol. 2010; 19:269–278. [PubMed: 20377629]
30. Wagner LA, et al. EGO, a novel, noncoding RNA gene, regulates eosinophil granule protein transcript expression. Blood. 2007; 109:5191–5198. [PubMed: 17351112]
31. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615. [PubMed: 21720365]
32. Hammerman PS, et al. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489:519–525. [PubMed: 22960745]
33. Ishii N, et al. Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. Journal of Human Genetics. 2006; 51:1087–1099. [PubMed: 17066261]
34. Rapicavoli NA, Poth EM, Blackshaw S. The long noncoding RNA RNCR2 directs mouse retinal cell specification. BMC Dev Biol. 2010; 10:49. [PubMed: 20459797]
35. Chan AS, Thorner PS, Squire JA, Zielenska M. Identification of a novel gene NCRMS on chromosome 12q21 with differential expression between rhabdomyosarcoma subtypes. Oncogene. 2002; 21:3029–3037. [PubMed: 12082533]

36. Gupta RA, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071–1076. [PubMed: 20393566]
37. Rinn JL, et al. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell. 2007; 129:1311–1323. [PubMed: 17604720]
38. Kogo R, et al. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. Cancer Res. 2011; 71:6320–6326. [PubMed: 21862635]
39. Beroukhim R, et al. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463:899–905. [PubMed: 20164920]
40. Garraway LA, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature. 2005; 436:117–122. [PubMed: 16001072]
41. Akavia UD, et al. An integrated approach to uncover drivers of cancer. Cell. 2010; 143:1005–1017. [PubMed: 21129771]
42. Tran VG, et al. H19 antisense RNA can up-regulate Igf2 transcription by activation of a novel promoter in mouse myoblasts. PLoS One. 2012; 7:e37923. [PubMed: 22662250]
43. Califano A, Butte AJ, Friend S, Ideker T, Schadt E. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. Nat Genet. 2012; 44:841–847. [PubMed: 22836096]
44. Pe'er D, Hacohen N. Principles and strategies for developing network models in cancer. Cell. 2011; 144:864–873. [PubMed: 21414479]
45. Zhao J, et al. Genome-wide identification of polycomb-associated RNAs by RIP-seq. Mol Cell. 2010; 40:939–953. [PubMed: 21172659]
46. Syvanen AC. Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet. 2001; 2:930–942. [PubMed: 11733746]
47. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet. 2010; 11:685–696. [PubMed: 20847746]
48. Flicek P, et al. Ensembl 2012. Nucleic Acids Res. 2012; 40:D84–90. [PubMed: 22086963]
49. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008; 24:2395–2396. [PubMed: 18697769]
50. Kuhn RM, Haussler D, Kent WJ. The UCSC genome browser and associated tools. Brief Bioinform. 2012
51. Seok J, Xu W, Gao H, Davis RW, Xiao W. JETTA: junction and exon toolkits for transcriptome analysis. Bioinformatics. 2012; 28:1274–1275. [PubMed: 22433281]
52. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118–127. [PubMed: 16632515]
53. Trapnell C, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28:511–515. [PubMed: 20436464]
54. Beroukhim R, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci U S A. 2007; 104:2000720012.
55. Mermel CH, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011; 12:R41. [PubMed: 21527027]
56. Taylor BS, et al. Functional copy-number alterations in cancer. PLoS One. 2008; 3:e3179. [PubMed: 18784837]
57. Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics. 2011; 27:275–282. [PubMed: 21075743]
58. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]

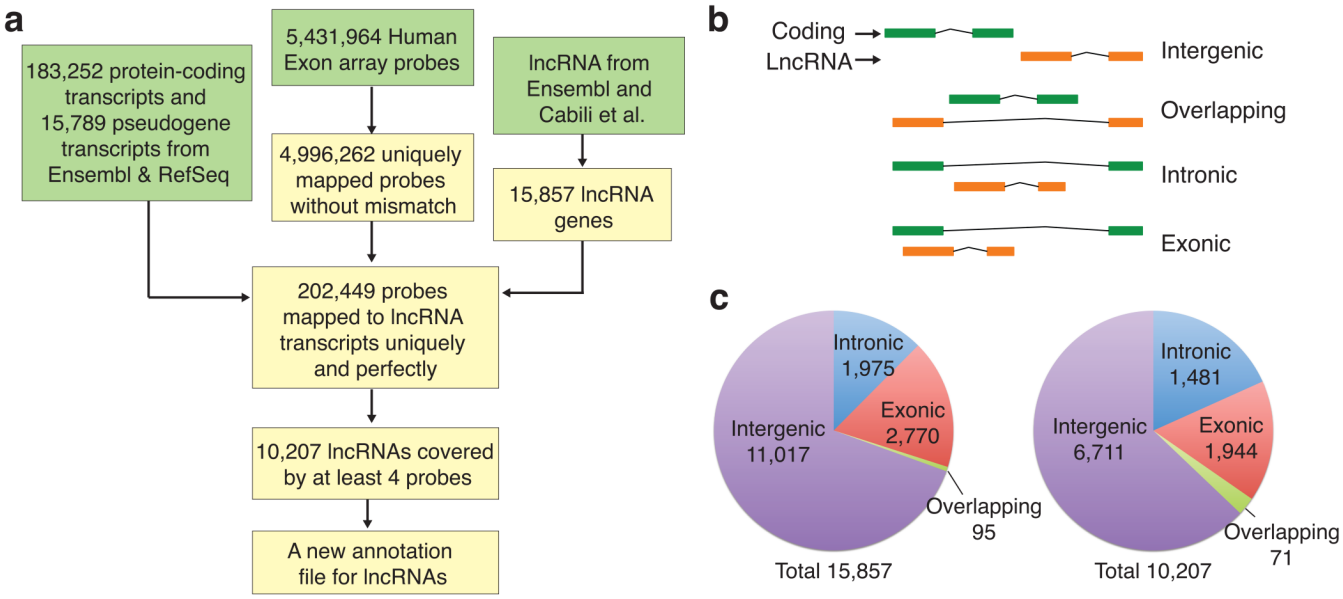

Figure 1. Human Exon array re-annotation and lncRNA classification

Affymetrix Human Exon array probe re-annotation pipeline for lncRNA was shown in (a). (b) Adopting the classification scheme from a previous study (Ref. 20), lncRNA were classified into four categories: intergenic, overlapping, intronic and exonic on the basis of their relationship with protein-coding genes. (c) Pie charts showed the number of lncRNA in each category for all collected lncRNA and for those with at least 4 uniquely mapped exon array probes.

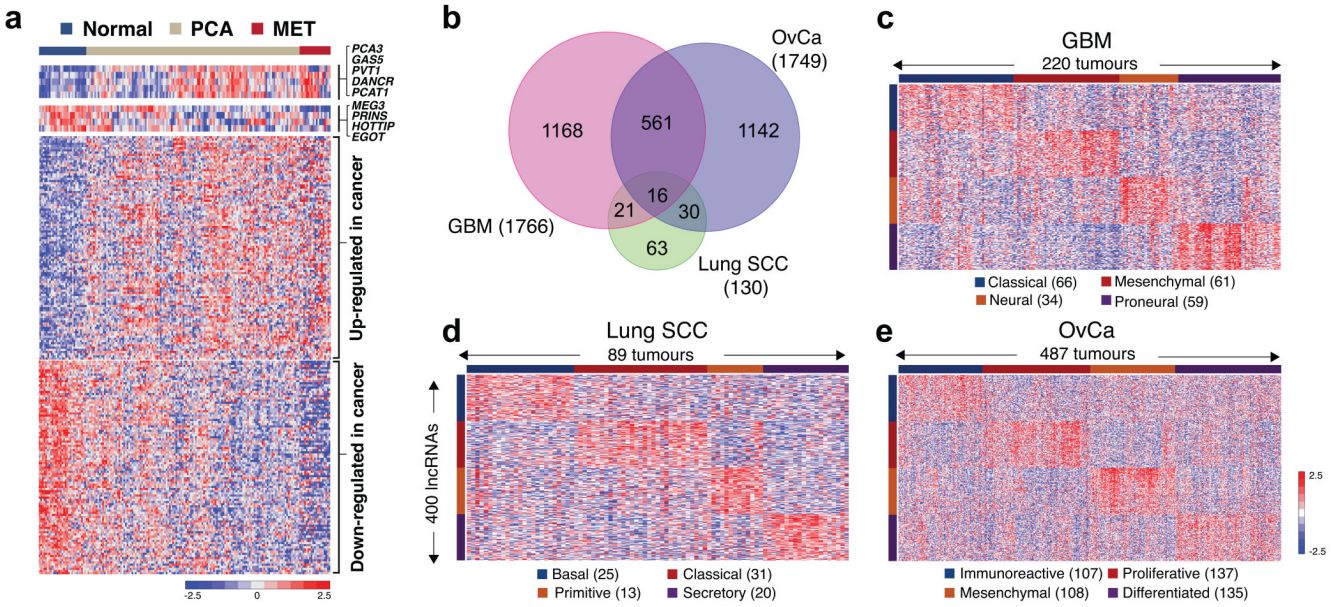

Figure 2. The number and the expression profile of lncRNA that have disease-specific or subtype-specific expression in prostate cancer, GBM, OvCa and Lung SCC

(a) The expression level of lncRNA that showed significantly differential expression between cancer and normal prostate tissues were shown in heatmap across 29 normal prostate samples, 131 primary and 19 metastatic prostate tumor samples. Several known cancer-related lncRNA or lncRNA with established function in non-cancer context were highlighted. (b) Venn diagram represented the number of subtype-specific lncRNA in three cancers. The expression profile of top 100 lncRNA that exhibited significantly higher expression in one subtype than the others for (c) GBM, (d) OvCa and (e) Lung SCC were shown in heatmap (Note: the rank was based on the ascending order of the *p*-value). Tumor samples were hierarchically clustered within each subtype.

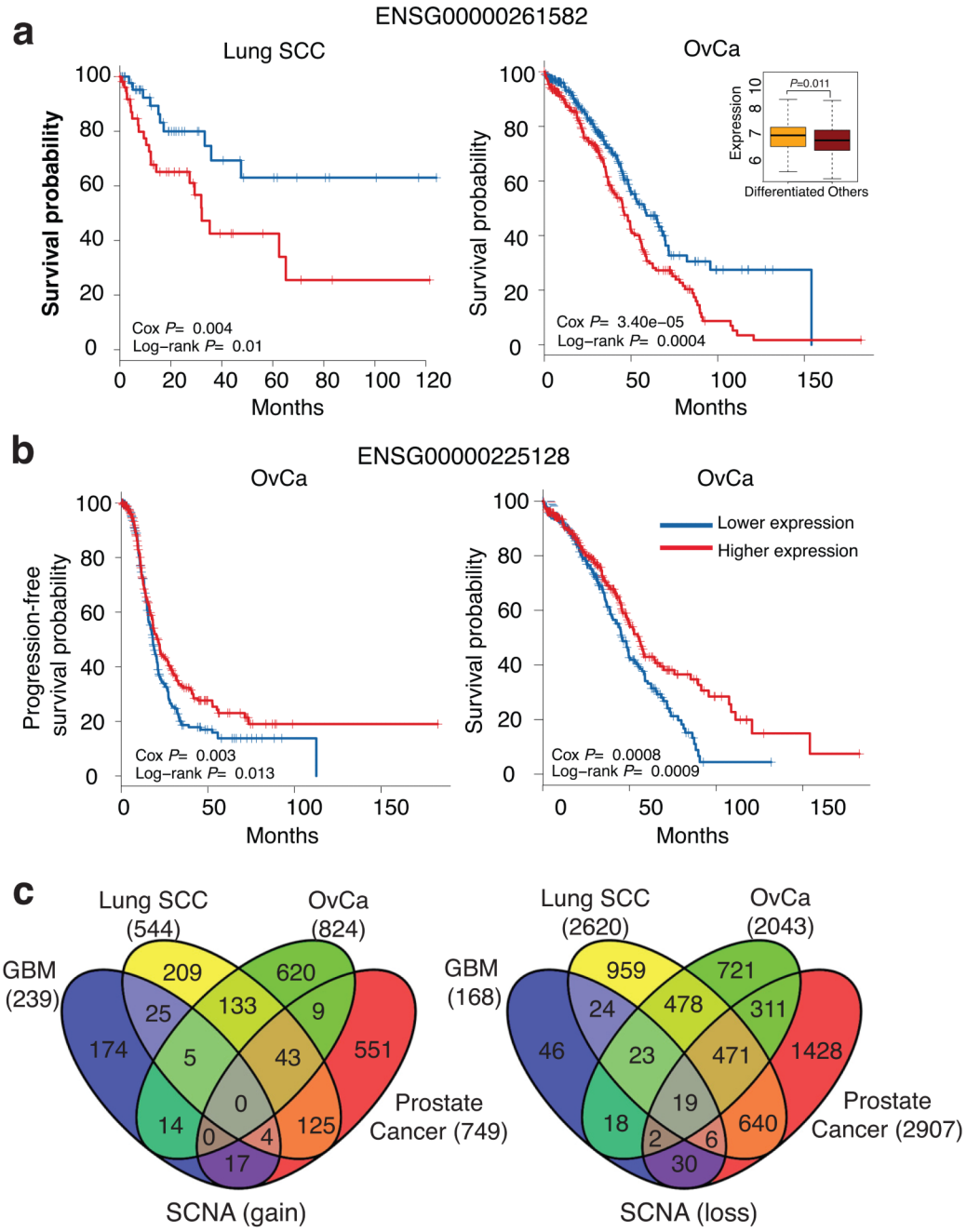

Figure 3. LncRNA associated with prognosis or in the genomic regions of SCNA

(a) Kaplan-Meier curve of two patient groups with higher (top 50%, n = 64) and lower expression (bottom 50%, n = 64) of ENSG00000261582 in Lung SCC and OvCa (Red: higher expression, blue: lower expression) was shown. The boxplot demonstrated that ENSG00000261582 expressed higher in the “differentiated” subtype of OvCa than the other subtypes. Both the *p*-value of multivariate Cox model for lncRNA expression and the *p*-value of log-rank test were shown. (b) Kaplan-Meier curve for overall and progression-free survival of two patient groups with higher (top 50%) and lower expression (bottom 50%) of ENSG00000225128 in OvCa was shown. (c) The number of lncRNA located in the SCNA (gain) and SCNA (loss) regions in different cancers was shown as Venn diagrams.

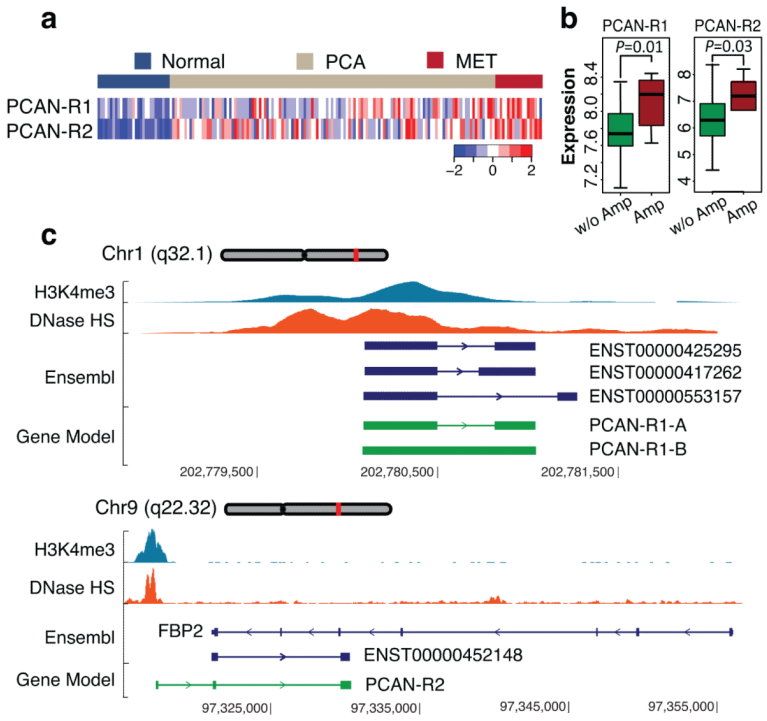

Figure 4. The genetic alteration and the expression profile of PCAN-R1 and PCAN-R2 in normal prostate tissues or prostate tumors and their transcript structure in cell line

(a) The heatmap showed the expression of PCAN-R1 and PCAN-R2 in normal prostate tissue, primary and metastatic prostate cancer. (b) The boxplot of PCAN-R1 and PCAN-R2 expression in tumors with genomic amplification (n = 7 and n = 9) and in the tumors without genomic amplification (n = 121 and n = 119) were compared. The boxplot showed the expression distribution of PCAN-R1 and PCAN-R2 in two groups and mann-Whitney U test was performed for the comparison. (c) The transcript structures of PCAN-R1 and PCAN-R2 from Ensembl annotation and determined by 5′ and 3′ RACE experiments in LNCaP cell were shown. In addition, the H3K4me3, and DNase I hypersensitive region profile in the same cell line were shown.

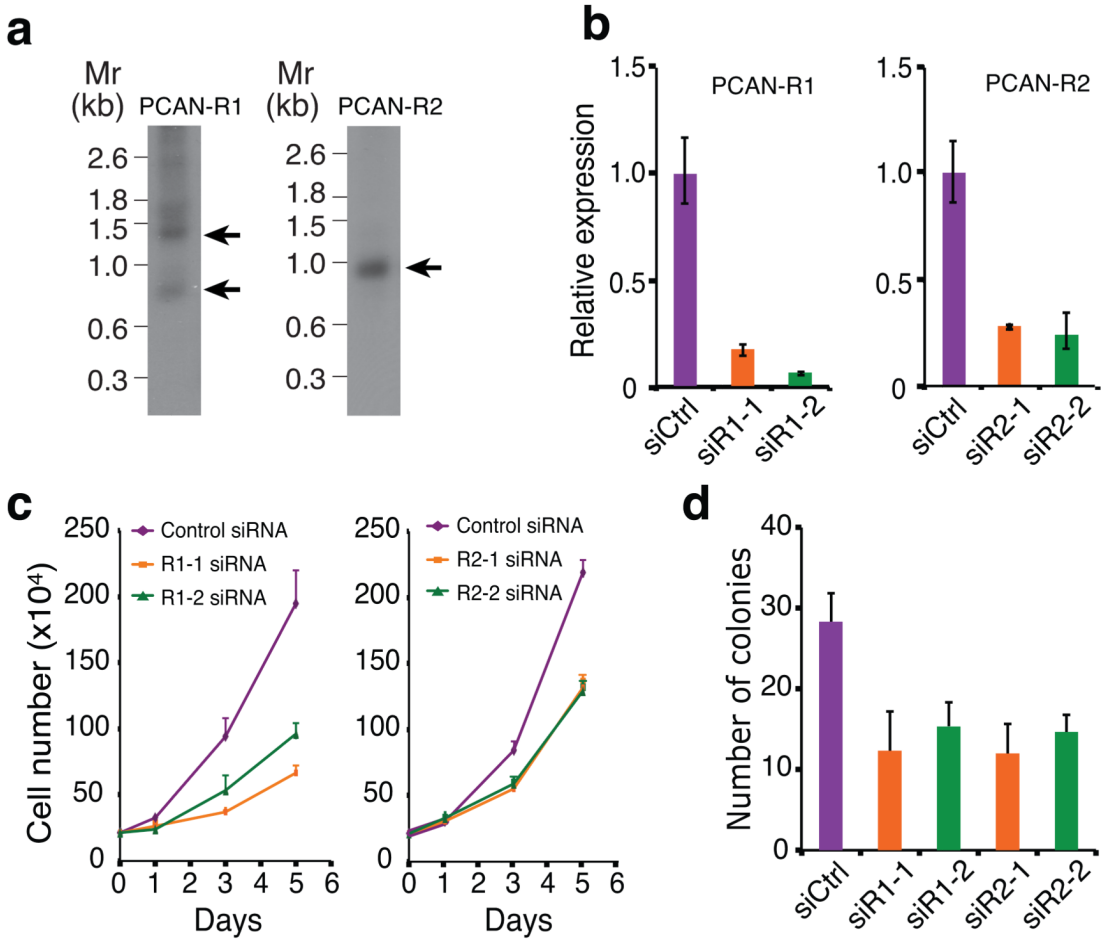

Figure 5. Functional validation of PCAN-R1 and PCAN-R2

(a) The Northern blot of PCAN-R1 and PCAN-R2 transcripts was shown (Mr: RNA marker). (b) The relative expression level of PCAN-R1 and PCAN-R2 upon knockdown by two different siRNA (orange and green) and upon control siRNA treatment (purple) was shown. (c) The growth curves of LNCaP cell with or without targeted siRNA-mediated knockdown of PCAN-R1 or PCAN-R2 were shown. The growth curves of control siRNA-treated cells and the growth curves of two targeted siRNA-treated cells were plotted in purple, orange, and green, respectively. Data were shown as Mean+S.D. n=3. (d) The number of soft-agar colony formation of LNCaP cell with or without targeted siRNA-mediated knockdown of PCAN-R1 or PCAN-R2 was shown.

Table 1

A summary table of literature-curated lncRNA

Ensembl ID	Gene Name	MW-U test *p*-value	Cancer vs Normal	Function annotation
ENSG00000225937	*PCA3*	9.50E-12	Up	Prostate cancer
ENSG00000234741	*GAS5*	1.77E-06	Up	Breast cancer
ENSG00000249859	*PVT1*	4.93E-11	Up	Multiple cancers
ENSG00000226950	*DANCR*	3.03E-08	Up	Development
ENSG00000253438	*PCAT1*	1.12E-05	Up	Prostate cancer
ENSG00000227418	*PCGEM1*	4.49E-04	Up	Prostate cancer
ENSG00000245532	*NEAT1*	0.00642	Up	Nuclear speckle
ENSG00000258492	*KCNQ1OT1*	0.0103	Up	Colon cancer
ENSG00000251164	*HULC*	0.0311	Up	Multiple cancers
ENSG00000251562	*MALAT1*	0.285	-	Multiple cancers
ENSG00000214548	*MEG3*	3.92E-08	Down	Multiple cancers
ENSG00000238115	*PRINS*	1.37E-07	Down	Autoimmune disease
ENSG00000243766	*HOTTIP*	1.95E-06	Down	Development
ENSG00000235947	*EGOT*	2.48E-05	Down	Development
ENSG00000214049	*UCA1*	2.11E-02	Down	Bladder cancer
ENSG00000228630	*HOTAIR*	0.0573	-	Multiple cancers
ENSG00000130600	*H19*	0.0842	-	Multiple cancers
ENSG00000240498	*ANRIL*	0.699	-	Prostate cancer

*Known cancer-related lncRNA or lncRNA with established function in non-cancer context, and their regulation in cancer compared with normal prostate tissue were listed. The statistical significance of their expression difference between cancer and normal prostate tissue was evaluated by Mann–Whitney U test (MW-U test).