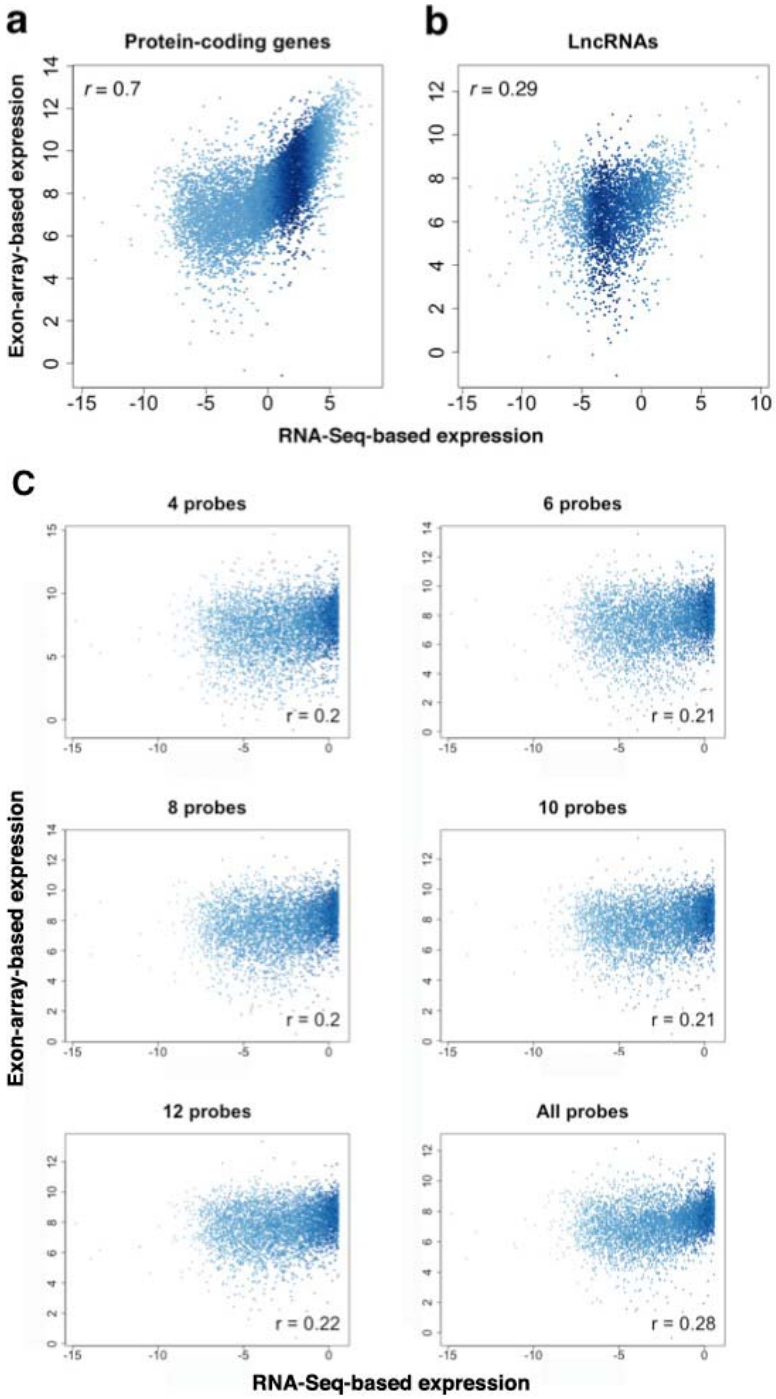
Supplementary Information

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

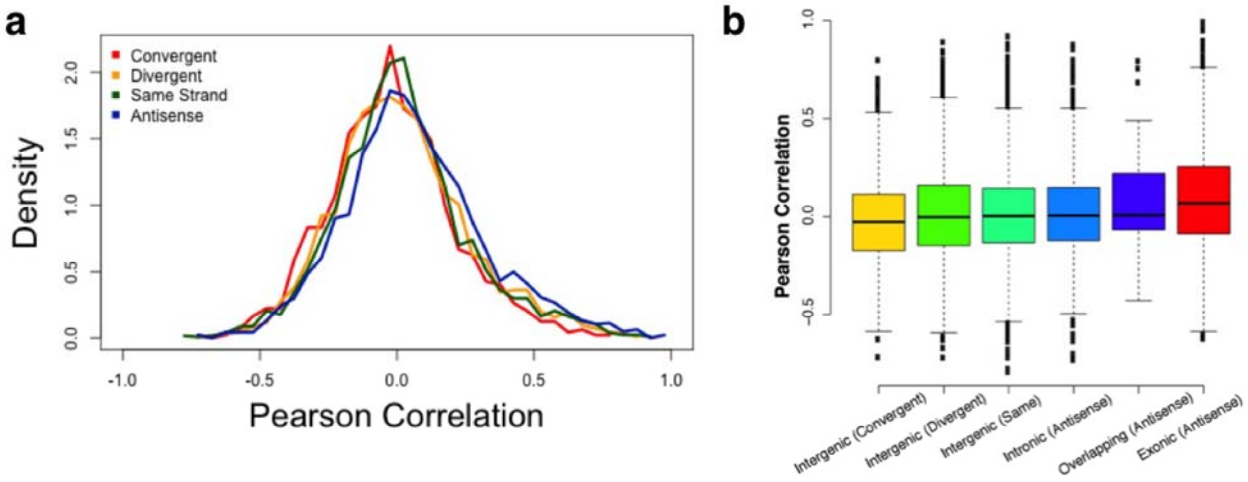
Zhou Du[1#], Teng Fei[2,3,4,5#], Roel G.W. Verhaak[6], Zhen Su[7], Yong Zhang[1], Myles Brown[2,3,4*], Yiwen Chen[5#,*], X. Shirley Liu[2,5*]

[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.

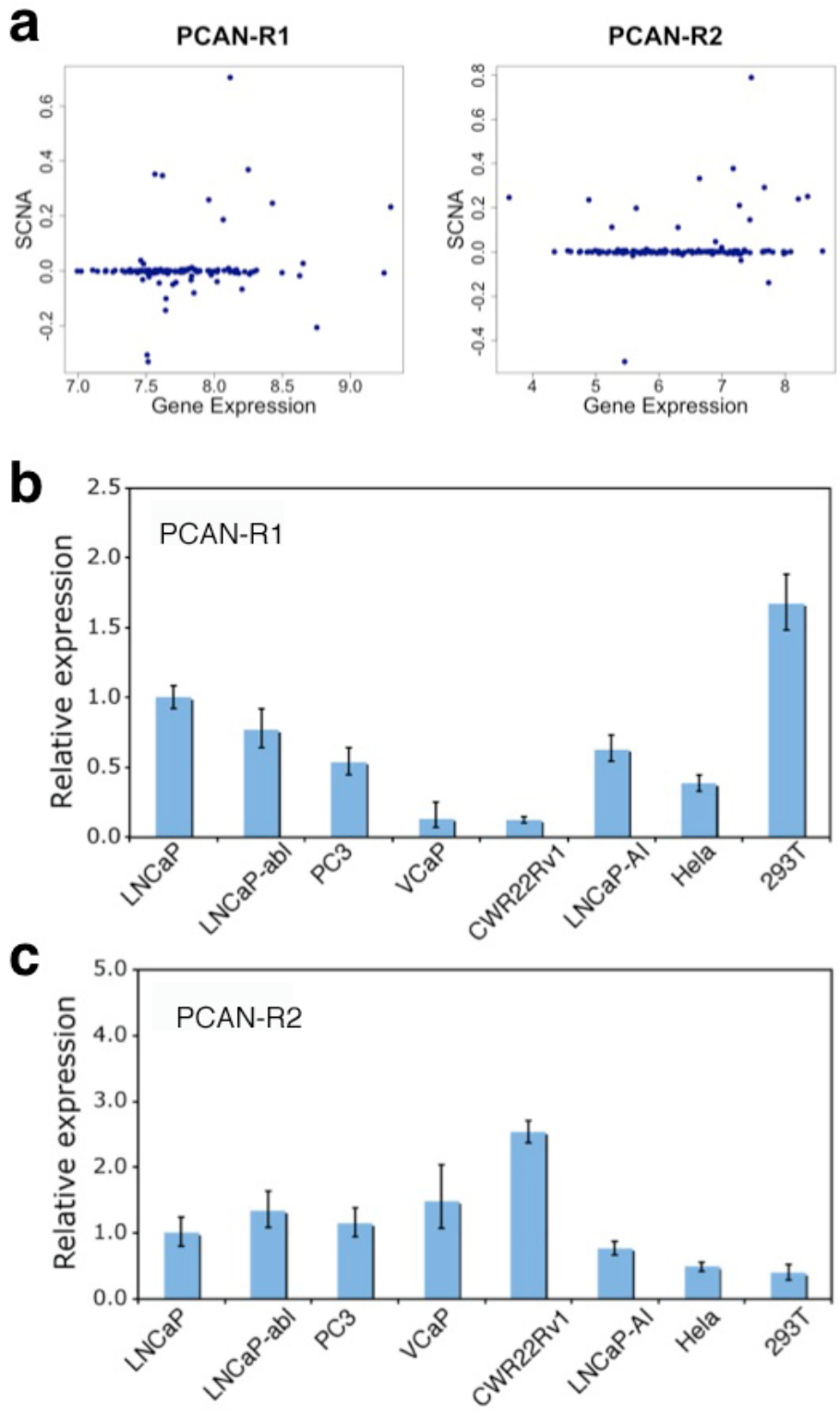


Supplementary Figure 1. A comparison between RNA-seq and exon-array based expression. The scatter-plot of gene expression measured by exon array and RNA-seq from LNCaP cell line was shown for (a) protein coding genes and (b) lncRNA. Only the genes with non-zero RNA-seq count were shown. The Pearson's correlation coefficient was used to quantify the strength of the correlation between RNA-seq and exon-array-based expression. (c) The scatter-plot of protein coding gene expression measured by exon array using randomly selected

4, 6, 8, 10, 12 and all probes and RNA-seq from LNCaP cell line was shown. The protein-coding genes, the expression of which is below the 95% quantile value of lncRNA expression based on RNA-seq data were selected. The Pearson's correlation coefficient was used to quantify the strength of the correlation between RNA-seq and exon-array-based expression using different number of probes.



Supplementary Figure 2. The Pearson Correlation Coefficient (PCC) of expression between lncRNA in different categories and their neighboring protein coding genes (PCGs). (a) The distribution of PCC was shown for the “convergent”, “divergent” and “same strand” intergenic lncRNA, and for all antisense genic lncRNA, including “exonic”, intronic” and “overlapping” classes. (b) The box plot of PCC was shown for the “convergent”, “divergent” and “same strand” intergenic lncRNA and for “exonic”, “intronic”, and “overlapping” antisense genic lncRNA. The sense genic lncRNA were not included because those lncRNA and their neighboring PCGs may be non-independent transcripts.



Supplementary Figure 3. **The somatic copy number alteration (SCNA) and expression of PCAN-R1 and PCAN-R2 in prostate cancer patients and their expression in different cell lines**. (a) The scatter-plot of PCAN-R1/-R2 expression and their corresponding SCNA across different individuals was shown. Expression level of PCAN-R1 (b) and PCAN-R2 (c) across different cell lines. Prostate cancer cell lines (LNCaP, LNCaP-abl, PC3, VCaP, CWR22Rv1 and LNCaP-AI) as well as Hela and 293T cells were employed to determine the expression level of PCAN-R1 and PCAN-R2. These two lncRNA exhibit reasonably high expression levels in LNCaP cells, in which we determined lncRNA identity and carried out functional validation.

(a)

PCAN-R1-A

AGAGGCCGGACCTGGGCAACCCCAGCCTGGAGGTGCCGGGGCCGGAGCTCCCAGAGGGCTGGGTGCGAGGCCTAGGCG
GGGTCAGGTTCGGGTTCCTAGGCCATAGCGGAGCTGCAGCCCAGGCGGCCGGAGCGGAACCCAGCCCCGCTCCGAGTG
CCACGTCTCCAGGAACCCCCTCCTTACTCTTGGACAACACTCCGCCCCCGCCCGGGCCTCCGTCCCCCAAACCGCCCT
CATTTGTGCAGCGCCAGATCCTTCGGACACATCCCTAGGTGTCTCCATCCTCATTCGGTCCATCCAACTCCCAGACCT
CACGTCAACCGGCTGCACCCCACTTTCCAGCCTGCGCCCCAGATCTGCAGCCTTCGCCCCTAGATACACCCGCCTGGT
GATGAGGCGCTCCTCGCGTTCCTTCCGTCTCCAGTGCACCGGCTTTCCCTCGTCCTCTGCGCAGTCCATCTCAGCTCA
TCTCTCCAATTCAATGCCATCATCTCTCCTCACCATCTCTCGGTGCCCTGGAATGTTTGCTGTCAGATGTCCCCTGTG
AAACCCACAAACGCTTGCGATTTGGCCTCCTTGTTTTATTTTGTGTAGTCCTACAACGTCTTGTTACTACCCCCTATT
ACAACACTTATAACTCA-polyA

PCAN-R1-B

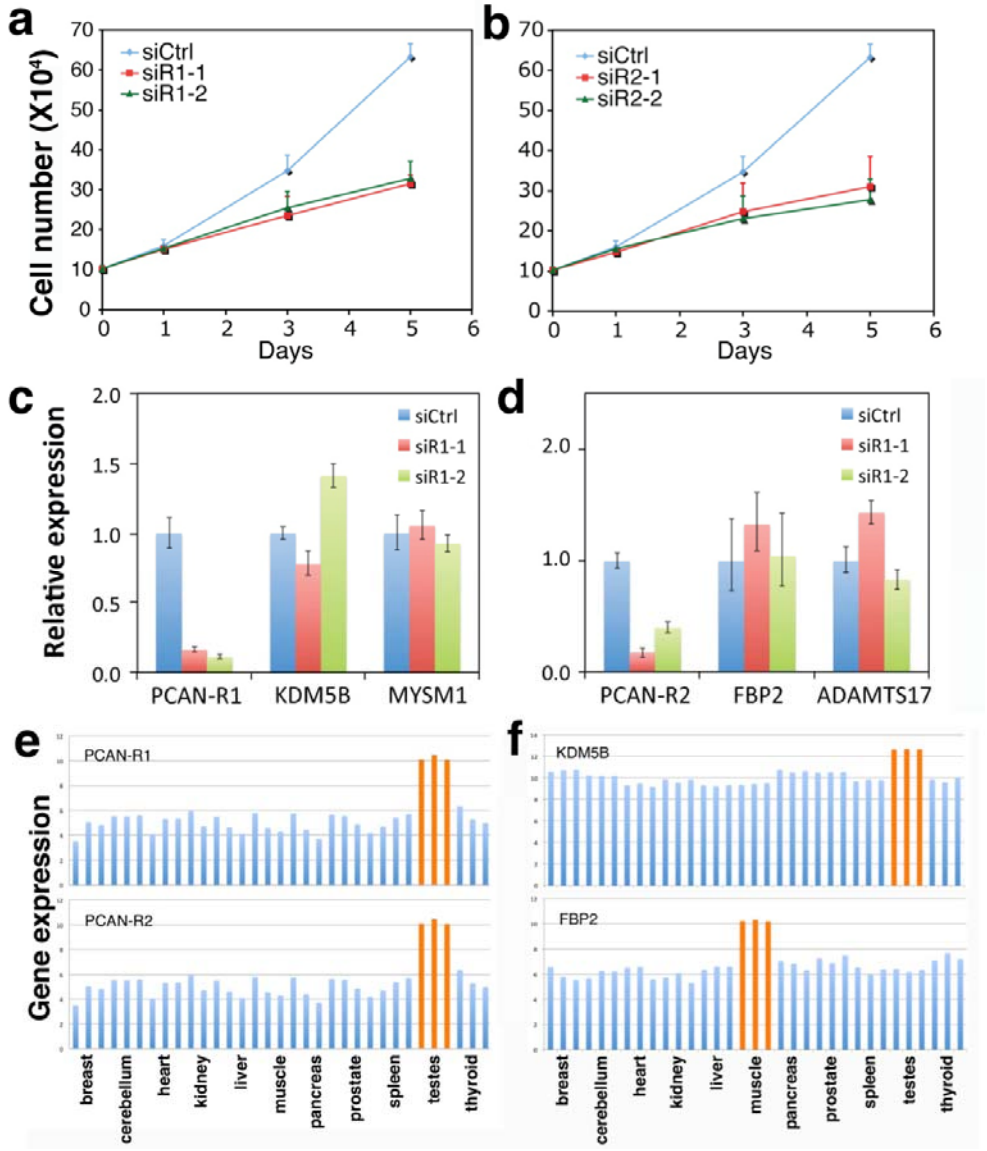
AGAGGCCGGACCTGGGCAACCCCAGCCTGGAGGTGCCGGGGCCGGAGCTCCCAGAGGGCTGGGTGCGAGGCCTAGGCG
GGGTCAGGTTCGGGTTCCTAGGCCATAGCGGAGCTGCAGCCCAGGCGGCCGGAGCGGAACCCAGCCCCGCTCCGAGTG
CCACGTCTCCAGGAACCCCCTCCTTACTCTTGGACAACACTCCGCCCCCGCCCGGGCCTCCGTCCCCCAAACCGCCCT
CATTTGTGCAGCGCCAGATCCTTCGGACACATCCCTAGGTGTCTCCATCCTCATTCGGTCCATCCAACTCCCAGACCT
CACGTCAACCGGCTGCACCCCACTTTCCAGCCTGCGCCCCAGATCTGCAGCCTTCGCCCCTAGATACACCCGCCTGGT
GATGAGGCGCTCCTCGCGTTCCTTCCGGTAACCGCGTTGCGAAGACCACGCTGCCGGGTTGCAAACTTGGGGGGACTT
CCTCCCTCCCCTCCCCCTGGGCGCCGTGCAACTGCCCTGGGACCGGGTTCTGGGATGAGGGGGGCAGACCGGGCTCCC
CAGCGGCCGGCGCAGCACGTAGCGCACGTGTAGGGTCCGCTCCCCACCCCCTCGCCGCCTCTGACAACTTTTCAGGGC
TCCAGGTGTCCGTGAGCCTCCCTTCCGCCCTGGCCTCCGGTCTCTGCCTTGCTCGTGCTTCTACCACCACCCTTCCCC
TCCCAACCCGGTGGATCCTCTCGTCTCCCCCAGTCTCCAGTGCACCGGCTTTCCCTCGTCCTCTGCGCAGTCCATCTC
AGCTCATCTCTCCAATTCAATGCCATCATCTCTCCTCACCATCTCTCGGTGCCCTGGAATGTTTGCTGTCAGATGTCC
CCTGTGAAACCCACAAACGCTTGCGATTTGGCCTCCTTGTTTTATTTTGTGTAGTCCTACAACGTCTTGTTACTACCC
CCTATTACAACACTTATAACTCA-polyA

(b)

PCAN-R2

GTGGTGGCGGTGGACATTGCAGCGCGGCTGGAGGGGGTCCTTAGACAAGGTGCAAGACAAACAGAAGAGGGCATGTGG
GGTCAAACTCGCTAGCTGCCTGCCTGATTTTTCTGCACACAGGACAAATTCACCAAGAGCCTAGCAACATGAAGAGAG
ATGCCAGGAAGAAGAGAGAAGCCAGGAAACAAGCCAACCGCACAATCCCCACATCAGAGCAGGAGAAGATGGGGGCCT
GCTGGCAGAGCTGGGGCTTGGCTGTGGTCACTCTGAACCTGCTCTTTGGTGTTTTCATGAGTGGTGGGAAGAATAGGG
ACCATATGGAGCCCACACAGGAAGCTCTAGCAGTAACACAGCAAGCAGGAAGACAATTCTAAGGAAGCAGCCCATAGT
CTTCTTTCTTTTCCTGTGCATCTTCCACTGTCAGTGAGGCTCCTCATTTATGGTGAACCCAACTGTGTGTATCTCCCA
AGTTCTCACCCGCAGATTAATGTTTTCAGGAAGATAGGCCATCAACAGTGAGAGGAAGAAGTTACATTGTCGTATGAG
GGATGCATTTTAACCATTAATTTGTGGTACAGGCTGGGCGCAGTGGCTTATGCATGTAATCCCAGCACTTTGGGAGGC
CGAGGTGGGTGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACATGGTGAAACCCCGTCTTTACTAAATATA
CAAAAAATTGGCCGGGCGTGGTGGTGGGCACCTGTAGTCCCAGCTACTCGGGGGGCTGAGGCAGGAGAATGGTGTGAA
CCCGGGAGGCAGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGATGACAGAGCAAGACTCCATCTC
A-polyA

Supplementary Figure 4. **The sequences of PCAN-R1 and PCAN-R2 identified by 3'- and 5'-RACE in LNCaP cells**. PCAN-R1 has two transcripts: the short two-exon PCAN-R1-A transcript (641bp without counting polyA tail) and long single-exon PCAN-R1-B transcript (959bp without counting polyA tail); while PCAN-R2 has one single strong transcript with three exons (859bp without counting polyA tail). Different colors denote different exons.



Supplementary Figure 5. The impact of siRNA-mediated knowdown of PCAN-R1 and PCAN-R2 on cell growth and the expression other PCGs, and the expression profile of these two lncRNA and their neighboring PCGs across different types of normal tissues. The growth curves of LNCaP-abl cell with or without targeted siRNA-mediated knockdown of (a) PCAN-R1 and (b) PCAN-R2 were shown. The growth curves of control siRNA-treated cells and the growth curves of two different targeted siRNA-treated cells were plotted in blue, red, and green, respectively. The relative expression level of PCAN-R1 (c) and PCAN-R2 (d), their neighboring PCGs KDM5B and FBP2, as well as PCGs MYSM1 and ADAMTS17 with homologue sequence (2 mismatches) to respective siRNA sequences was shown. The expression upon knockdown by two different siRNA (purple and orange) to either lncRNA and upon control siRNA was shown (green). (e) The expression pattern of PCAN-R1 and PCAN-R2 across 11 human normal tissues was shown. The exon array data of 11 human normal tissues were obtained from Affymetrix (http://www.affymetrix.com/). (f) The expression pattern of PCAN-R1 neighboring PCG, KDM5B and PCAN-R2 neighboring PCG, FBP2 across 11 human normal tissues were shown and each tissue has three replicates.

Supplementary Table 1. The number of Affymetrix microarray probes corresponding to lncRNA

	Number of probes corresponding to lncRNA	Number of lncRNA with at least 4 probes
Affymetrix Human Exon array	202449	10207
Affymetrix U95Av array	1865	76
Affymetrix U133 plus 2.0 array	43752	2561
Affymetrix U133B array	21880	1181
Affymetrix U133A array	2830	143

*The number of lncRNA with at least 4 probes coverage in five major Affymetrix array platforms were listed.

Supplementary Table 2. The number of lncRNA genes and those lncRNA genes with at least 4 Affymetrix Human Exon array probes for each category of lncRNA

lncRNA genes (15857/10207)										
	Intergenic (11017/6711)				**Genic (4840/3496)**					
	Same Strand	**Convergent**	**Divergent**	**Contig**	**Exonic (2770/1944)**		**Intronic (1975/1481)**		**Overlapping (95/71)**	
all lncRNA	4928	1847	3800	442	S	AS	S	AS	S	AS
					584	2186	247	1728	83	12
lncRNA with at least 4 probes	3048	1110	2550	3	S	AS	S	AS	S	AS
					255	1689	143	1338	61	10

*S = Sense, AS = Antisense

Supplementary Table 3. The number of lncRNA associated with prognosis or in the SCNA regions.

	GBM	Lung SCC	OvCa	PC
Overall survival	133	124	211	-
Progression-free survival	-	-	85	120
SCNA (gain)	86	112	271	111
SCNA (loss)	55	279	571	192

*The number of lncRNA associated with overall or progression-free survival and the number of lncRNA that were in the SCNA (gain)/SCNA (loss) regions and showed positive correlation between the SCNA (gain) /SCNA (loss) and expression level change in GBM, Lung SCC, OvCa and prostate cancer (PC) were shown.

Supplementary Table 4. The number of lncRNA associated with overall- or progression-free survival and located in the recurrent somatic copy number gain SCNA (gain) or SCNA (loss) regions

	GBM (O)	Lung SCC (O)	OvCa (O)	OvCa (P)	PC (P)
SCNA (gain)	16	35	70	44	49
SCNA (loss)	8	152	162	65	179

*O = overall survival, P = progression-free survival

Supplementary Table 5. The significance of overlap of lncRNA in the SCNA gain or loss regions among different cancer types.

	GBM	Lung SCC	OvCa	Prostate Cancer
GBM		**4.24E-08**	**0.4675162**	**0.1594936**
Lung SCC	**3.41E-07**		**9.69E-70**	**4.65E-69**
OvCa	**9.58E-08**	**4.39E-141**		**0.8674728**
Prostate Cancer	**0.04995362**	**7.06E-82**	**7.92E-33**	

*The fisher's exact test was used for *p*-value calculation. (SCNA gain: red; SCNA loss: blue)