Journal of the Endocrine Society, 2021, Vol. 5, No. 11, 1–23
https://doi.org/10.1210/jendso/bvab153
Research Article

Research Article

Characterization of the Testis-specific *LINC01016* Gene Reveals Isoform-specific Roles in Controlling Biological Processes

Enrique I. Ramos,[1] Barbara Yang,[1] Yasmin M. Vasquez,[2] Ken Y. Lin,[3] Ramesh Choudhari,[1] and Shrikanth S. Gadad[1,4,5]

[1]Center of Emphasis in Cancer, Department of Molecular and Translational Medicine, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, El Paso, Texas 79905, USA; [2]Department of Pathology, Baylor College of Medicine, Houston, Texas 77030, USA; [3]Department of Obstetrics & Gynecology and Women's Health, Division of Gynecologic Oncology, Montefiore Medical Center, Albert Einstein College of Medicine, Bronx, New York 10461, USA; [4]Graduate School of Biomedical Sciences, Texas Tech University Health Sciences Center, El Paso, Texas 79905, USA; and [5]Mays Cancer Center, UT Health San Antonio MD Anderson Cancer Center, San Antonio, Texas 78229, USA.

ORCiD number: 0000-0001-9921-7944 (E. I. Ramos); 0000-0001-7361-1652 (R. Choudhari); 0000-0003-0030-5472 (S. S. Gadad).

Abbreviations: cDNA, complementary DNA; CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; DFI, disease-free interval; FPKM, fragments per kilobase of transcript per million mapped reads; GDC, Genomic Data Commons database; GEO, Gene Expression Omnibus; GTEx, Genotype-Tissue Expression database; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; *LINC01016*, long intergenic non-protein–coding RNA 01016; lncRNA, long noncoding RNA; mRNA, messenger RNA; NGS, next-generation sequencing; NHPRTR, Nonhuman Primate Reference Transcriptome Resource; PANCAN, Pan-Cancer; PCA, principal component analysis; PCR, polymerase chain reaction; QC, quality control; RNA-seq, RNA sequencing; TCGA, The Cancer Genome Atlas; TGCT, testicular germ cell tumor; TTU HPCC, Texas Tech University High-Performance Computing Center; UCEC, uterine corpus endometrial carcinoma; UCSC, University of California, Santa Cruz

Received: 20 May 2021; Editorial Decision: 22 September 2021; First Published Online: 27 September 2021; Corrected and Typeset: 22 October 2021.

Abstract

Long noncoding RNAs (lncRNAs) have emerged as critical regulators of biological processes. However, the aberrant expression of an isoform from the same lncRNA gene could lead to RNA with altered functions due to changes in their conformations, leading to diseases. Here, we describe a detailed characterization of the gene that encodes long intergenic non–protein-coding RNA 01016 (*LINC01016*, also known as *LncRNA1195*) with a focus on its structure, exon usage, and expression in human and macaque tissues. In this study we show that it is among the highly expressed lncRNAs in the testis, exclusively conserved among nonhuman primates, suggesting its recent evolution and is processed into 12 distinct RNAs in testis, cervix, and uterus tissues. Further, we integrate de novo annotation of expressed *LINC01016* transcripts and isoform-dependent gene expression analyses to show that human *LINC01016* is a multiexon gene, processed through differential exon usage with isoform-specific roles. Furthermore, in cervical, testicular, and

ISSN 2472-1972

uterine cancers, *LINC01016* isoforms are differentially expressed, and their expression is predictive of survival in these cancers. This study has revealed an essential aspect of lncRNA biology, rarely associated with coding RNAs, that lncRNA genes are precisely processed to generate isoforms with distinct biological roles in specific tissues.

Key Words: genomics, gene expression, LINC01016, cervix, uterus, testis

Long noncoding RNAs (lncRNAs) play essential roles in normal physiological processes and aberrant conditions, such as cancer [1, 2]. Similar to protein-coding messenger RNAs (mRNAs), lncRNAs are transcribed by RNA polymerase II, spliced, polyadenylated, and capped. However, most lncRNAs lack protein-coding potential [1-3], with a few exceptions [4]. LncRNAs diversity is due to different modes of transcription, and they are expressed in a tissue-specific manner. And, the constant expansion of newly identified lncRNA genes requires that each one be comprehensively annotated to understand its molecular functions. Moreover, lncRNAs have been demonstrated to be excellent tumor biomarkers with therapeutic potential [2, 5, 6]; however, their therapeutic application has been limited by the lack of a mechanistic understanding of their actions in the cell. This knowledge gap is partly due to the limited or poor annotation of lncRNAs, which hinders functional studies [7]. Thus, it is imperative to determine the gene structure, exon usage, and expression pattern of developmental and disease-relevant lncRNAs to facilitate their functional characterization.

The annotation of lncRNAs is often more challenging than the annotation of protein-coding mRNA because of low copy numbers and extreme tissue specificity. Further, it is more difficult to determine the 5′/3′ ends of lncRNAs that are chromatin associated, transcribed from a divergent promoter, or transcribed in an antisense direction since they coexist with unprocessed transcripts [3, 7, 8]. These challenges have been partly addressed by recent advances in next-generation sequencing (NGS) technologies, which provide unparalleled prospects in elucidating gene structure, exon usage, expression, isoform-specific biological roles, and evolution of specific noncoding genes [3, 9-11]. This high-quality sequencing data across multiple species can be accessed through publicly available genome and gene expression databases [10, 12-14]. Notably, a recent study has identified 15 000 to 35 000 differentially expressed lncRNAs across different organs and species, emphasizing the cell-, evolution-, and developmental-specific transcriptional programs [15]. However, these studies lack information about isoform-specific functions and expression, which is critical in understanding lncRNA-regulated processes; unlike proteins, RNA plays a central role. The nucleotide length and composition are very crucial in defining lncRNA secondary structures that dictate its interaction with proteins, RNA, and DNA [16, 17]. Isoforms of varying length could fold into different structures, defining their specific functions, which warrants isoform-specific annotation and characterization of each lncRNA in target tissues [2].

In this study, we functionally characterize long intergenic non-protein–coding RNA 01016 (*LINC01016*), which is among the top 58 lncRNAs expressed in the testis. To define the *LINC01016* gene structure and identify isoform-specific expression, we have used data presented in the Nonhuman Primate Reference Transcriptome Resource (NHPRTR) [18] and RNA sequencing (RNA-seq) data from the human testis and cervix [19-21]. We have leveraged these data sets generated through NGS to annotate and study the expression and differential exon usage in human and nonhuman primates. Furthermore, we also cloned and expressed differentially processed transcripts to uncover isoform-specific roles of *LINC01016*. In addition, we have interrogated its association with clinical outcomes in cervical, testicular, and uterine cancer patients.

Results show that the *LINC01016* gene is transcribed using differential exon usage and alternative splicing and, in addition, different isoforms control specific biological processes. *LINC01016* is abundantly expressed in the testis in the late stages of tissue development, and its gene structure is conserved among nonhuman primate species. Intriguingly, interrogation of the publicly available resource containing cervical squamous cell carcinoma, testicular germ cell, and uterine corpus endometrial carcinoma cancer patients suggests that *LINC01016* isoforms are differentially expressed and that higher levels of the aggregate expression of *LINC01016* isoforms predict outcome in testicular cancer patients [22-24]. Interestingly, in testicular cancer, expression levels of different isoforms are associated with altered patient outcomes. Overall, these findings have revealed an essential aspect of lncRNA biology: Each lncRNA isoform may regulate a completely different set of biological processes that are rarely associated with coding RNAs.

Material and Methods

Database Searches and Analyses

Primate genomic databases were accessed from the Ensembl Genome Browser (www.ensembl.org). Searches were performed as described elsewhere [25]. Briefly, BlastN was employed under optimal parameters (maximum e-value of 10; mismatch scores: 1, -3; gap penalties: opening 5, extension, 2; filtered low complexity regions, and repeat sequences masked) by means of probes: human *LINC01016* DNA segments (*Homo sapiens* genome assembly GRCh38.p13). The following genome assemblies were examined: chimpanzee (*Pan troglodytes*, Pan_tro_3.0), gorilla (*Gorilla*, gorGor4), macaque (*Macaca mulatta*, Mmul_10), marmoset (*Callithrix jacchus*, ASM275486v1), and mouse lemur (*Microcebus murinus*, Mmur_3.0). The highest scoring results in all cases mapped to chromosome 6 in a region analogous to the location of human *LINC01016*, except for marmoset, in which there was no chromosome assignment, and in macaque mapped to chromosome 4. Human *LINC01016* complementary DNA (cDNA) sequences were obtained from the Ensembl genome browser database.

Data Sources

A summary of the samples, sources, and annotations used for this study is presented in Table 1.

LINC01016 Gene Models, Splice Variants, Conservation Tracks, and Expression

Collapsed exon structure and splice variants for human *LINC01016* were obtained from the Ensembl database annotation (release 101, August 2020; https://www.ensembl.org) with 12 splice site variants covering exons 1 to 4 in chromosome 6. Conservation tracks were obtained from the University of California, Santa Cruz (UCSC) Genome Browser database (https://genome.ucsc.edu) from the 100 vertebrate basewise conservation by PhyloP score and multiz alignment and conservation for 7 vertebrate species. Gene models were created in R using the package from Bioconductor Gviz [27]. Exon and exon-exon junction read counts were accessed from https://gtexportal.org. The data used for the analyses described in this manuscript were obtained from the Genotype-Tissue Expression database (GTEx) Portal April 11, 2020, or dbGaP accession number phs000424.v8.p2 April 11, 2020.

RNA Sequencing Expression Data Sets for Testes in Rhesus Macaque

NGS paired-end reads (101 bp) were downloaded from the NHPRTR (http://www.nhprtr.org/) [18]. The data sets are of Indian-origin rhesus macaque (University of Washington) from testes samples. FASTQ files were downloaded and analyzed in the subsequent RNA-seq–analysis pipeline. Briefly, FASTQ files were inspected for quality control (QC) using FastQC, alignment to the rhesus macaque genome (rheMac10; UCSC) was performed using TopHat2 [28] with the following options: a maximum number of multihits allowed (-g) - 2, annotation of spliced junctions (-j) and alignment to the entire rheMac10 genome. Alignments and data analyses were performed using the Texas Tech University High-Performance Computing Center (TTU HPCC) and TTUHSC El Paso NGS Computing Server. BAM files were sorted and indexed using Samtools [29], and alignment visualization was carried out using Integrative Genomic Viewer [30]. Generation of gene read counts was performed using the utility FeatureCounts from the Subread package using the options -t exon, -g gene_id or -g exon_id, and -a annotation.gtf. Normalization of read counts was conducted using the R package from Bioconductor DESeq2 [31]. Isoform expression and transcriptome assembly were performed using Cufflinks [32] with the options -g annotation.gtf, -b genome.fa and -u. Gene-level expression values (FPKM) were analyzed from the output files: genes.fpkm_tracking, isoforms.fpkm_tracking, and transcripts.gtf.

Amplification and Cloning of *LINC01016* Isoforms for Overexpression Experiments

Full length of the predicted isoforms was amplified from the human cervix (catalog No. CR559475 OriGene Technologies), uterus (catalog No. CR559475 OriGene Technologies), or testis RNA (catalog No. CR560016, OriGene Technologies) in polymerase chain reactions (PCRs) carried out with a Phusion DNA polymerase kit (catalog No. M0530, New England BioLabs). Primers were designed based on the sequence information on GTEx and Ensembl databases. PCR products were then analyzed on 1% SYBR-stained agarose gel, and bands at expected sizes were excised from agarose gel and purified with QIAquick Gel Extraction kit (catalog No. 28704, Qiagen). Critical observations included the following: (a) there is an additional exon toward the 5′ end on the 201-isoform amplified from the uterus (CU2), compared to the 201-isoform documented in the databases. There was no PCR product amplified from the cervix sample. (b) There are few mutations presented on the amplified 202 (T1) and 204 (T2) isoforms from testis, but most of them are documented single-nucleotide variations (formerly single-nucleotide polymorphisms). We chose the clones with the least mutations for the downstream experiments: 2 mutations on 202 and 3 on 204. (c) Isoform 205 shorter version shown on GTEx, was amplified as predicted (CU1) both from the cervix and uterus. However, for the longer version of the 205 isoform as noted in Ensembl, we obtained 205 ENU from the uterus that fully matched, but

Table 1. Summary of samples used in this study

Description	No. of samples	Reference study	Reference annotations	Independently processed	*LINC01016* isoforms
Nonhuman primates–comparative genomics	1	NHPRTR University of Washington [18]	Chimpanzee (*Pan troglodytes*, Pan_tro_3.0), gorilla (*Gorilla gorilla*, gorGor4), macaque (*Macaca mulatta*, Mmul_10), marmoset (*Callithrix jacchus*, ASM275486v1), mouse lemur (*Microcebus murinus*, Mmur_3.0). *Homo sapiens* (GRCh38.p13)	Yes Transcriptome	NA
Normal human testis	8	Djureinovic et al [19]	*H sapiens* (GRCh38.p13) Ensembl v101	Yes Transcriptome	12
Normal human cervix	4	Xu et al [20, 21]	*H sapiens* (GRCh38.p13) Ensembl v101	Yes Transcriptome	12
GTEx Cervical	18	GTEx Portal dbGaP Accession phs00424.v8.p2	GTEx v8 GENCODE v26	No	5
GTEx Testicular	345	GTEx Portal dbGaP Accession phs00424.v8.p2	GTEx v8 GENCODE v26	No	5
GTEx Uterine	134	GTEx Portal dbGaP Accession phs00424.v8.p2	GTEx v8 GENCODE v26	No	5
LINC01016 Cloned transcripts	6	Novel–GEO submission No. GSE171047	*H sapiens* (GRCh38.p13) Ensembl v101	Yes Transcriptome	12
TCGA TGCT	156	TCGA GDC v28	*H sapiens* (GRCh38) GENCODE v22	No	5
TCGA CESC	306	TCGA GDC v28	*H sapiens* (GRCh38) GENCODE v22	No	5
TCGA UCEC	552	TCGA GDC v28	*H sapiens* (GRCh38) GENCODE v22	No	5
TCGA PANCAN TOIL RSEM	10 535	UCSC Xena Browser [26]	*H sapiens* (GRCh38) TCGA GDC v18	No	5
GTEx TOIL RSEM	7862	UCSC Xena Browser [26]	*H sapiens* (GRCh38) TCGA GDC v18	No	5

Abbreviations: CESC, cervical squamous cell carcinoma and endocervical adenocarcinoma; GDC, Genomic Data Commons database; GEO, Gene Expression Omnibus; GTEx, Genotype-Tissue Expression database; *LINC01016*, long intergenic non-protein–coding RNA 01016; NA, not available; NHPRTR, Nonhuman Primate Reference Transcriptome Resource; PANCAN, Pan-Cancer; TCGA, The Cancer Genome Atlas; UCEC, uterine corpus endometrial carcinoma; UCSC, University of California, Santa Cruz; v, version.

205 ENC amplified from the cervix showed a different exon arrangement towards at the 5′ end, as seen in the amplified CU2.

Confirmed PCR products were cloned into a pcDNA3.0 vector. HEK-293 cells (catalog No. CRL-1573, ATCC) were transfected with these *LINC01016* isoforms using a Lipofectamine 3000 kit (catalog No. L3000015, Thermo Fisher Scientific) for 48 hours and were collected for total RNA extraction (catalog No. BS88136, Bio Basic). RNA quality was assessed by electrophoresis using the Agilent 2200 TapeStation system. Overexpression of the respective isoforms was confirmed by real-time PCR using PowerUp SYBR Green Master Mix (A25743, Thermo Fisher Scientific) with various sets of primers designed to recognize particular isoform(s). Real-time PCR reactions were carried out using the Applied Biosystems StepOne Real-Time PCR System. Directional mRNA library preparation (poly-A enrichment) and RNA-seq were performed at Novogene Corporation (Sacramento, California, USA) with 2 biological replicates of each overexpressed isoform along with empty vector control.

Primer list

CU1BamHI_FP: 5′CGCGGATCCACGAGGCAGCAAATCCGAAC3′

CU1XhoI_RP: 5′CCGCTCGAGAGTAGACCTTGGCTCCTCTC3′

CU1ENXhoI_RP: 5′CCGCTCGAGGCTTTCTATTTTAATGAATTTATACG3′

CU2BamHI_FP: 5′CGCGGATCCAGCAAATCCGAACAGGCAAAG3′

CU2XhoI_RP: 5′CCGCTCGAGTGAACTTGCAAATATACCATTTAATG3′
T1BamHI_FP: 5′CGCGGATCCACTCCACCAGCCGGCCC3′
T1XhoI_RP: 5′CCGCTCGAGTTACATTTTTAAAAAGGTCTTCCAC3′
T2BamHI_FP: 5′CGCGGATCCGGACGGAGGCAGCCGTTAG3′
T2XhoI_RP: 5′CCGCTCGAGTGGTTATAAGCTTTCTATTTTAATG3′

Overexpressed *LINC01016* Isoforms, RNA Sequencing Library Preparation, and Sequencing

RNA-seq libraries were prepared and sequenced at Novogene Corporation Inc. After sample preparation, sample QC was carried out by Nanodrop, agarose gel electrophoresis, and Agilent 2100. After the QC procedures, mRNA from eukaryotic organisms is enriched using oligo(dT) beads. First, the mRNA is fragmented randomly by adding fragmentation buffer. Then the cDNA is synthesized by using an mRNA template and random hexamers primer, after which a custom second-strand synthesis buffer (Illumina), dNTPs, RNase H, and DNA polymerase I are added to initiate the second-strand synthesis. Second, after a series of terminal repair, A ligation, and sequencing adaptor ligation, the double-stranded cDNA library is completed through size selection and PCR enrichment. QC of the library consists of 3 steps: Quibit 2.0, Agilent 2100, and Q-PCR. Finally, qualified libraries are fed into Illumina sequencers (NovaSeq 6000) after pooling according to its effective concentration and expected data volume. The quality of RNA-seq data is summarized in Table 2.

LINC01016 Assembly of Transcriptome Data and Differential Gene Expression

Paired-end reads (150 bp) were mapped to the human transcriptome (hg38) release 101 version (August 2020) from Ensembl using the HISAT2 aligner [33]. Samtools were used to convert SAM into BAM files and subsequently sort and index files. Read counts were generated from the alignments using the FeatureCounts program from the subread-2.0.1 package [34] with Ensembl annotation version 101 as the reference annotation. Normalization and differential gene expression analyses were performed using the DESeq2 R package [31]. Differentially expressed genes were extracted by applying a 1.5–fold-change cutoff.

Splice Event Prediction and Quantification From RNA Sequencing

Splice graph analysis based on de novo prediction was carried out using the R package SGSeq (Bioconductor) [35]. Input data are RNA-seq reads mapped to a reference genome in BAM format. Briefly, sample information was generated from the original BAM file with function getBamInfo(), genomic coordinates of *LINC01016* were stored in a GRanges object and used to filter to the appropriate coordinates, predictions were carried out by using the analyzeFeatures() function, annotation was performed using Ensembl version 101, and plots were generated using the plotFeatures(), and plotCoverage() functions.

Gene Ontology and Kyoto Encyclopedia of Genes and Genomes Pathway Analysis

Differentially expressed genes that passed the previously described cutoffs were used to obtain the biotype classification using Ensembl database REST API; composition analysis based on biotypes was generated in R. Gene Ontology (GO) analyses were performed using the fgsea (Bioconductor) R package for fast preranked gene set enrichment analysis [36]. Different gene data sets were obtained from the Molecular Signatures Database (https://www.gsea-msigdb.org) to query genes in the pathway: hallmark pathways, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, molecular functions, and biological processes. Heat maps were generated using Plotly (R package) using the normalized enrichment score.

Genotype-Tissue Expression and The Cancer Genome Atlas *LINC01016* Transcript Expression RNA Sequencing

Transcript expression from the GTEx and The Cancer Genome Atlas (TCGA) were obtained by using the UCSC XenaBrowser online exploration tool (https://xenabrowser.net). Data sets from both sources were downloaded using the TOIL RSEM isoform percentage resource [26].

Tumor Sample Analysis

The expression of *LINC01016* was determined using TCGA GDC Portal v28 in testicular germ cell tumor (TGCT), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), and uterine corpus endometrial carcinoma (UCEC) samples, as well as GTEx Portal version 8 for testis, cervix, and uterus samples.

Principal Component Analysis

We carried out dimension reduction statistical analysis for the expression of *LINC01016* (high/low), the clinical stage of the tumor (stage I, stage IA, stage IB, stage IS, stage II, stage IIA, stage IIB, stage IIC, stage IIIB), and the disease-free interval (DFI). The analysis was

Table 2. *LINC01016*-cloned transcripts RNA sequencing metrics

Sample	Average quality per read	Total raw reads (PE)	Uniquely aligned reads (PE)	Total aligned reads (PE)	FeatureCounts annotation *Homo_sapiens* GRCh38.101.gtf	Feature counts assignments
293 Parental rep. 1	36 Phred Score	22 262 581	19 956 962 (89.6%)	21 427 982 (96.25%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	16 729 869
293 Parental rep. 2	36 Phred Score	21 487 450	19 318 690 (89.91%)	20 710 869 (96.39%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	16 274 344
293 Empty vector rep. 1	36 Phred Score	21 665 711	19 139 792 (88.34%)	20 521 173 (94.72%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	16 037 761
293 Empty vector rep. 2	36 Phred Score	21 116 062	18 543 309 (87.82%)	19 924 475 (94.36%)	Features: 1 397v832 Meta-features: 60 671 Chrom/Contigs: 47	15 529 247
293 T1 rep. 1	36 Phred Score	23 108 070	20 516 178 (88.78%)	22 006 509 (95.23%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	17 255 935
293 T1 rep. 2	36 Phred Score	21 278 059	18 876 679 (88.71%)	20 229 572 (95.07%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	15 986 789
293 T2 rep. 1	36 Phred Score	21 407 269	19 096 155 (89.20%)	20 459 399 (95.57%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	16 254 700
293 T2 rep. 2	36 Phred Score	20 220 718	17 928 057 (88.66%)	19 242 228 (95.16%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	15 468 756
293 CU1 rep. 1	36 Phred Score	21 451 814	19 013 560 (88.63%)	20 407 671 (95.13%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	16 165 268
293 CU1 rep. 2	36 Phred Score	21 116 289	18 762 840 (88.85%)	20 117 960 (95.27%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	16 214 935
293 CU2 rep. 1	36 Phred Score	21 410 611	19 105 880 (89.23%)	20 477 150 (95.64%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	16 258 057
293 CU2 rep. 2	36 Phred Score	20 691 553	18 429 516 (89.07%)	19 730 592 (95.36%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	15 537 828
293 CU1ENC rep. 1	36 Phred Score	20 902 448	18 542 115 (88.71%)	19 893 366 (95.17%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	15 040 770
293 CU1ENC rep. 2	36 Phred Score	21 267 592	18 744 985 (88.14%)	20 142 742 (94.71%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	16 375 455
293 CU1ENU rep. 1	36 Phred Score	21 270 923	19 064 532 (89.63%)	20 363 019 (95.73%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	15 596 446
293 CU1ENU rep. 2	36 Phred Score	20 440 059	18 175 525 (88.92%)	19 474 267 (95.28%)	Features: 1 397 832 Meta-features: 60 671 Chrom/Contigs: 47	15 706 828

Each sample was sequenced on an Illumina NovaSeq 6000 platform. The sequence length of each read is 150 bp.
Abbreviations: *LINC01016*, long intergenic non-protein–coding RNA 01016; PE, paired ends.

conducted using R statistical software using the function prcomp() to generate the model along with the options center = TRUE, and scale = TRUE. The visual representation of the principal component analysis (PCA) was performed using the R package factoextra by using the function fviz_pca_ind() with settings to color the groups based on the variable in question and addElipses = TRUE.

Survival Analysis of Cervical Cancer, Testicular Germ Cell Tumor, and Uterine Corpus Endometrial Carcinoma Patients

To evaluate the prognostic value of the *LINC01016*, we explored its expression in samples of cervical squamous cell carcinoma, samples of TGCT, and UCEC cancer patient samples. Data for gene and isoform levels were obtained from the TCGA Pan-Cancer (PANCAN) via UCSC XenaBrowser online exploration tool (https://xenabrowser.net) [26]. Data sets from both sources were downloaded using the TOIL RSEM fpkm resource. The metric used for survival analysis was the DFI. The Kaplan-Meier plotter tool was utilized to plot a custom version of the data using KMplot [24] (http://kmplot.com).

Genomic Data Set Availability

The following new data sets generated for this study are available from the NCBI's Gene Expression Omnibus (GEO)database (http://www.ncbi.nlm.nih.gov/geo/) using accession number GSE171047.

Accession Nos.	
RNA sequencing	*GSE171047*
• 293 parental rep 1	GSM5217361
• 293 parental rep 2	GSM5217362
• 293 empty vector rep 1	GSM5217363
• 293 empty vector rep 2	GSM5217364
• Testis (GTEx) isoform 202 rep 1	GSM5217365
• Testis (GTEx) isoform 202 rep 2	GSM5217366
• Testis (GTEx) isoform 204 rep 1	GSM5217367
• Testis (GTEx) isoform 204 rep 2	GSM5217368
• Cervix/Uterus (GTEx) isoform 205 rep 1	GSM5217369
• Cervix/Uterus (GTEx) isoform 205 rep 2	GSM5217370
• Cervix/Uterus (GTEx) isoform 201 (4 exons) rep 1	GSM5217371
• Cervix/Uterus (GTEx) isoform 201 (4 exons) rep 2	GSM5217372
• Cervix/Uterus (Ensembl) isoform 205 (4 exons) rep 1	GSM5217373
• Cervix/Uterus (Ensembl) isoform 205 (4 exons) rep 2	GSM5217374
• Cervix/Uterus (Ensembl) isoform 205 (3 exons) rep 1	GSM5217375
• Cervix/Uterus (Ensembl) isoform 205 (3 exons) rep 2	GSM5217376

Results

Human *LINC01016* Gene Expression

Given the limited evidence of a functional role for the *LINC01016* gene in the literature, the relative abundance of the gene-specific transcripts could indicate its biological significance. Tissue-specific gene expression data were extracted from the GTEx database to determine the abundance of *LINC01016* transcript in healthy human tissues. *LINC01016* transcript levels varied slightly over a 40 transcripts-per-million range between tissues; the highest expression was seen in the testis, followed by the cervix and uterus (Fig. 1A). Further investigation in human testes during developmental stages suggests that *LINC01016* RNA starts appearing as early as the 13th year and as late as the 63rd year (Fig. 1B).

Analysis of transcriptomic data from "GTEx Portal dbGaP Accession phs00424.v8.p2" revealed that *LINC01016* is one among the highly expressed lncRNAs in the testis, and also among the top 58 lncRNAs expressed in testis with reads per kilobase of transcript per million mapped reads 10 (RPKM) (Fig. 1C).

The Human *LINC01016* Gene

The *LINC01016* gene is found on chromosome 6 of the human genome, located at the short arm of the chromosome (6p21.31). In human genome assembly, GRCh38.p13, Ensembl genome browser predicts 4 exons of *LINC01016*, spanning 29 408 base pairs between genomic coordinates 33 896 914 and 33 867 506 (Fig. 2A). According to Ensembl (version 101), the *LINC01016* gene is transcribed and processed into 12 different RNA species via a differential arrangement of exons and alternative splicing (Fig. 2B), and each transcript is composed of 1, 2, or 3 exons (see Fig. 2B). The sequences for *LINC01016* that are found in the NCBI nucleotide database (accession numbers NR_038989.1 and AK057709) do not contain the entire list of sequences found in Ensembl; however, they are similar to transcript 201 (see Fig. 2B). A summary of the samples, sources, and annotations used for this study is presented in Table 1.

The *LINC01016* Gene in Nonhuman Primates

In Ensembl, the *LINC01016* gene has not been annotated in any known nonhuman primate species, which encouraged us to investigate its presence and location in the target nonhuman primate genomes. Through examining relevant genomic databases, the *LINC01016* sequence was discovered in 5 different primate species (Fig. 3A and 3B and Table 3), in a region analogous to the location of human

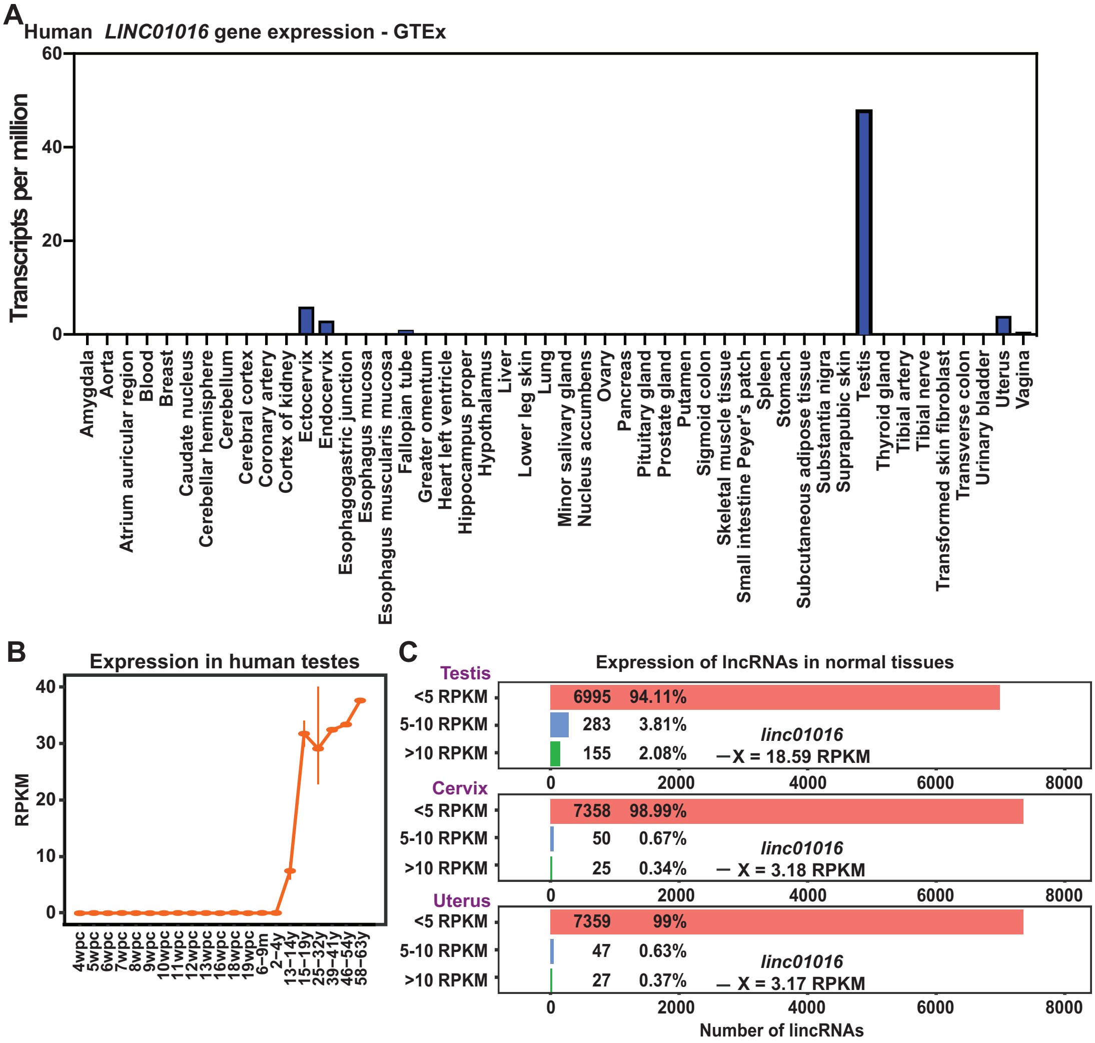

Figure 1. Expression of *LINC01016* in different human tissues. A, *LINC01016* gene expression in all the normal human tissues, as accessed from GTEx. Results are graphed as transcripts per million. B, *LINC01016* gene expression across developmental stages of the human testis, represented as RPKM. The expression data were accessed using the https://apps.kaessmannlab.org/lncRNA_app/ [15]. C, Expression of all lncRNAs in testis, cervix, and uterus. LncRNAs are detected using GTEx [9, 10]. GTEx, Genotype-Tissue Expression database; *LINC01016*, long intergenic non-protein–coding RNA 01016; lncRNA, long noncoding RNA; m, months; RPKM, reads per kilobase of per million mapped reads; wpc, weeks post conception; y, years.

LINC01016, except for in marmoset, where the chromosome annotation was absent, and in macaque, where it was found on chromosome 4 (see Fig. 3B). The data showed the *LINC01016* of interest is composed of 4 exons (see Fig. 3B and Table 3), similarly to human *LINC01016* (see Fig. 3A). There was also a similarity in the predicted *LINC01016* gene structure in these primates, and significant DNA conservation was shown for all putative exons, especially in chimpanzee and gorilla (99.37%-99.19%, respectively) (see Fig. 3B and Table 3). However, mouse lemur and marmoset were exceptions, as segments of exon 3 deviated from the other species (see Fig. 3B and Table 3). Moreover, RNA-seq analyses showed that *LINC01016* RNAs were present in the testis of macaque species, and an isoform with 4 exons was predominantly expressed (Fig. 3C). The overall organization of the *LINC01016* gene structure is similar to other reproductive tissue-specific genes, such as *MIR503HG* and other recently identified lncRNAs [5, 8].

LINC01016 Gene Structure in Different Human Tissues

We sought to understand the extent to which the human *LINC01016* gene had been annotated in Ensembl; analyses were performed by interrogating human gene expression data available from GTEx. Intriguingly, exon-exon junction expression differs across testicular, cervical, and uterine tissues (Fig. 4A and 4B). In testis, exon-exon junction 5 (see Fig. 4A and 4B) is predominantly expressed, leading to the formation of a single continuous exon (Fig. 4C and 4D), which is similar to isoform 209 and 212 described in Ensembl (see Fig. 2B). In contrast, both in the cervix and uterus, junctions

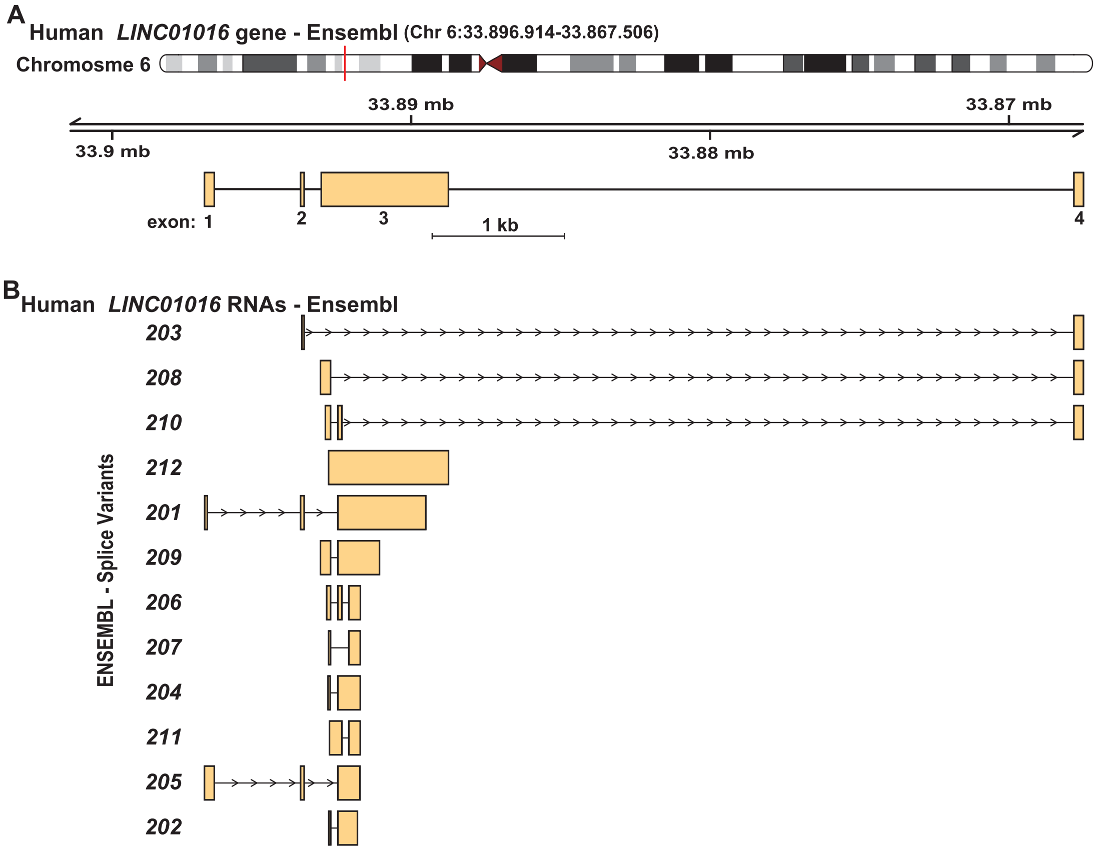
A
Human *LINC01016* gene - Ensembl (Chr 6:33.896.914-33.867.506)
Chromosme 6
33.89 mb
33.87 mb
33.9 mb
33.88 mb
exon: 1
2
3
1 kb
4
B
Human *LINC01016* RNAs - Ensembl
ENSEMBL - Splice Variants
203
208
210
212
201
209
206
207
204
211
205
202

Figure 2. The human long intergenic non-protein–coding RNA 01016 (*LINC01016*) gene in the Ensembl genome database. A, Diagram of the human *LINC01016* gene on chromosome 6 from Ensembl. Exons are depicted as boxes and introns as lines. Chromosomal coordinates and a scale bar are shown. B, Diagrams of human *LINC01016* RNAs from Ensembl. The exons or portions of exons found in each transcript are illustrated and are aligned with their location within the gene. The scale is the same as in A.

2 and 3 are expressed (see Fig. 4A and 4B), giving rise to exons 1, 2, and 4 (see Fig. 4C), corresponding precisely to what was defined in the genome database for isoform 201 (see Fig. 2B). However, in the cervix, junction 6 is also expressed but does not affect the exon arrangement (see Fig. 4B and 4C). Although these data vary from the information presented in Ensembl, they concur firmly with findings recently reported by Sarropoulos et al [15] (https://apps.kaessmannlab.org/lncRNA_app/). The aforementioned observations define a human *LINC01016* gene of 2 exons in the testis and 3 exons both in the cervix and uterus (Fig. 4D).

De novo Prediction of Splice Events, Expression, and Isoform Identification in Testis and Cervix

To independently evaluate and define the *LINC01016* gene structure, RNA-seq data sets [19-21] were obtained, and splice events were determined based on de novo predictions for *LINC01016* in the testis (Fig. 5A) and cervix (Fig. 5B). Testis-specific splice graph analysis (see Fig. 5A) shows 3 predicted exons and 2 junctions for 8 normal testis samples. The heat map shows differential exon usage and expression (FPKM) based on each sample. On the other hand, cervix-specific splice graph analysis (see Fig. 5B) shows 4 predicted exons and 3 junctions for 4 normal cervix samples. The heat map shows differential exon usage and expression (FPKM) based on each sample, with the lowest expression for junction 1.

Gene models for the *LINC01016* isoforms detected on the RNA-seq datasets are shown for normal testis (Fig. 6A) and cervix (Fig. 6B). The top panel shows isoform 201 (NCBI Refseq) and the predicted collapse structure for each tissue. The bottom panel depicts the different RNA-seq detected isoforms; in the case of the testis, a total of 6 different isoforms were detected with an expression range of 0.19 to 12.50 FPKM, while cervix tissue analysis identified a total of 4 isoforms with an expression range of 0.12 to 0.36 FPKM. The gene-specific mature transcript analyses reiterate differential exon usage and isoform expression based on tissue-specificity for normal samples.

Comparison of *LINC01016* Gene Models Between GTEx and RNA Sequencing Sample Predictions

Human *LINC01016* shows diverse isoform expression depending on the tissue. However, there is variability in

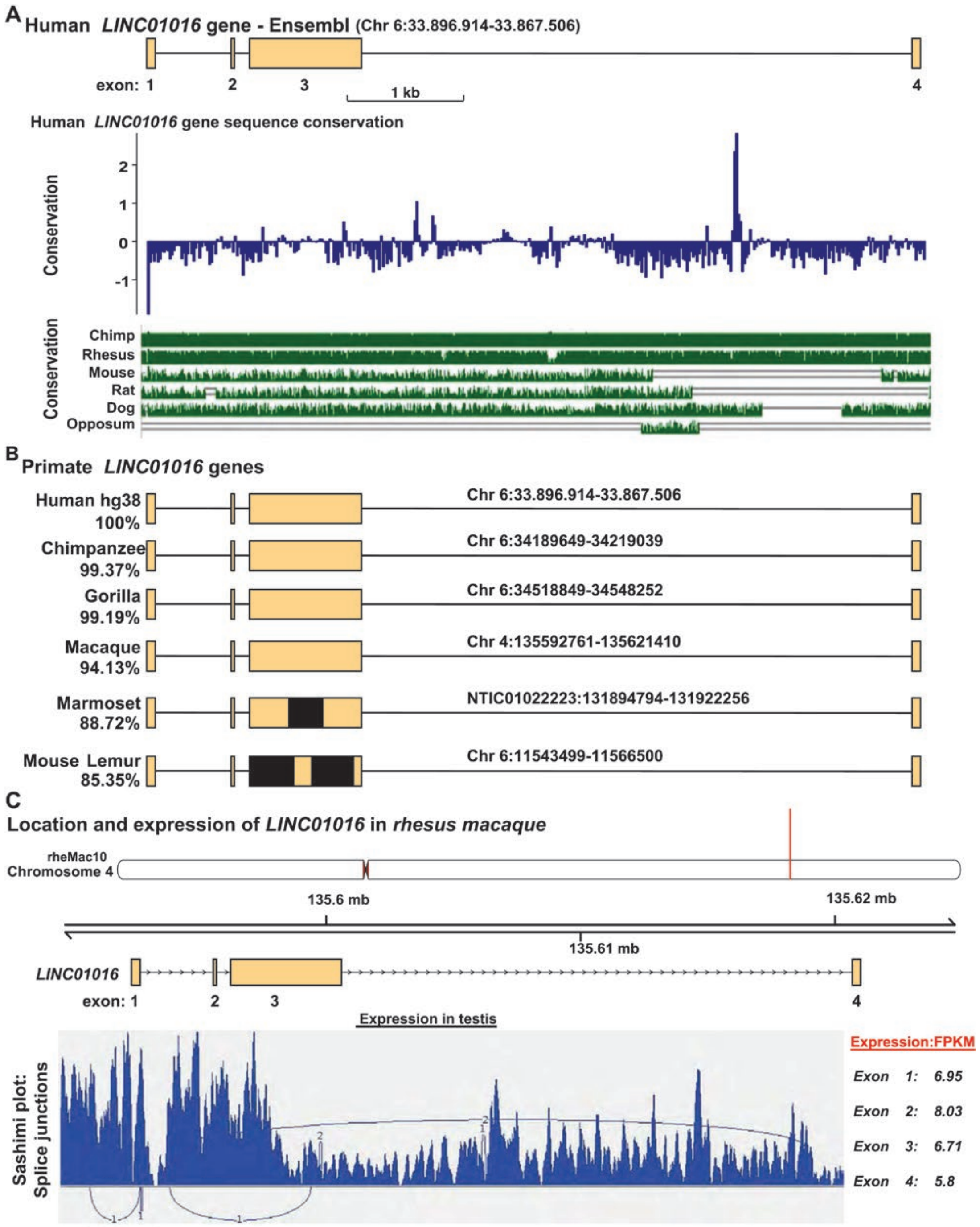

Figure 3. Long intergenic non-protein–coding RNA 01016 (*LINC01016*) gene structure and expression in primates. A, Human *LINC01016* gene conservation between different vertebrates. Top panel (blue histogram track) shows 100 vertebrate Basewise conservation by PhyloP score. The bottom panel (green alignment track) shows Multiz Alignment & Conservation for 7 vertebrate species. B, Conserved identity between different primates and *LINC01016*. Depicted are the predicted genomic structures with conserved exons, and in black are genomic regions missing from BLAST alignments to their respective human sequence. Percentage identity from BLAST search results is provided for each of the species. C, The top panel shows the genomic location of the predicted structure for *LINC01016* in rhesus macaque with 4 conserved exons, similar to human RNA sequencing experiments. Data from the Nonhuman Primate Reference Transcriptome Resource project are shown for expression in testis. The expression of different exons is shown as fragments per kilobase of transcript per million mapped reads (FPKM).

the expression of the isoforms of *LINC01016* depending on the data source and method used to analyze the samples. GTEx database (release v8) shows only 2 isoforms for the cervix/uterus (Fig. 7A), while our RNA-seq de novo analysis predicts a total of 4 detected isoforms (Fig. 7B). Moreover, for testis, GTEx (release v8) shows 3 isoforms (see Fig. 7A) while our RNA-seq de novo analysis predicts 6 different isoforms for the samples (Fig. 7C). Importantly, our RNA-seq data analysis uses the latest release of the gene annotations for *LINC01016*, which includes a higher

Table 3. Nucleotide identity with human *LINC01016* exons

Species	Exon 1 (329)	Exon 2 (122)	Exon 3 (4284)	Exon 4 (321)
Chimpanzee	99.7	100	98.72	99.07
Gorilla	99.08	100	98.93	98.75
Macaque	94.9	94.21	93.33	94.08
Marmoset	90.62	91.89	86.91	85.45
Mouse lemur	88	86.59	84.39[c]	82.42[e]
Dog	92.86[a]	96.15[b]	84.26[d]	88.89[f]

Abbreviation: *LINC01016*, long intergenic non-protein–coding RNA 01016.
[a] 28 bp aligned.
[b] 26 bp aligned.
[c] 300 bp aligned.
[d] 269 bp aligned.
[e] 273 bp aligned.
[f] 36 bp aligned.

number of isoforms. Nevertheless, the intricate and differential expression profile of *LINC01016* transcripts indicates that each isoform is plausibly associated with distinct biological processes.

Functional Analysis of *LINC01016* Isoforms

To unravel and define isoform-specific roles, 6 different transcripts of the *LINC01016* human gene were cloned in a mammalian expression vector and sequenced (Fig. 8A). Each transcript was expressed in a heterologous cell line (HEK293 cells, *LINC01016* is not transcribed). An advantage of this strategy is that endogenous lncRNA-driven compensatory mechanisms, which could mask the isoform-specific analyses, can be avoided. Fig. 8B shows the gene model diagrams of the cloned transcripts with the chromosome location, sequencing coverage, junction plots, and RNA-seq transcript expression (FPKM) for each independent replicate. The heat map depicts the differential expression fold change for each cloned transcript (Fig. 9A). Results show significant variability in gene expression depending on the transcript. Intersections of differentially expressed genes were analyzed between samples in different groups: all regulated genes (see Fig. 9B), upregulated genes (Fig. 9C), and downregulated genes (Fig. 9D). These results clearly show that although there are a common set of genes regulated by all the transcripts, the nonoverlapping isoform-dependent gene sets are more extensive and mutually exclusive.

Transcript-specific Molecular Signatures of Regulated Biological Pathways

To study *LINC01016* transcripts' potential roles in biological pathways, we performed a GO analysis. Different molecular signature databases were queried to obtain more insights into the potential transcript-specific biological functions of *LINC01016*. Heat maps showing normalized enrichment scores were generated for the different datasets: Hallmark gene sets (Fig. 10A), KEGG pathways (Fig. 10B), molecular functions (Fig. 10C), and biological processes (Fig. 10D). All molecular signature databases showed significantly different regulated pathways depending on the selected transcript, which correlates well with the identified gene sets specific to each isoform.

Expression and Possible Prognostic Value of *LINC01016* in Cervical Squamous Cell Carcinoma, Testicular Germ Cell Tumor, and Endometrial Carcinoma Cancer

Differentially expressed lncRNAs in reproductive tissues have been thought to be potential tumor biomarkers [5], prompting us to explore the expression of *LINC01016* in malignant tissue types of the cervix, testis, and uterus.

We analyzed different *LINC01016* transcripts expression using gene expression profiles from normal (GTEx) and PANCAN samples. Surprisingly, all the isoforms showed varied expression levels and appeared to be differentially regulated in normal and tumor tissues (cervical, testicular, and uterine), suggesting an isoform-specific functional role, which could reveal the regulatory mechanisms specific to lncRNA biology (Fig. 11). These results strongly suggest that *LINC01016* plausibly plays a critical role in reproductive biology.

To assess the clinical relevance of biomarkers, we conducted a multivariate statistical analysis comparing the expression of *LINC01016* (high/low), the clinical stage of the tumor, and DFI. To this end, we carried out PCA (Fig. 12). The expression of *LINC01016* on a PCA plot shows 2 distinctive clusters for low and high expression separated by PC2, which accounts for 35.3% of the variability (Fig.

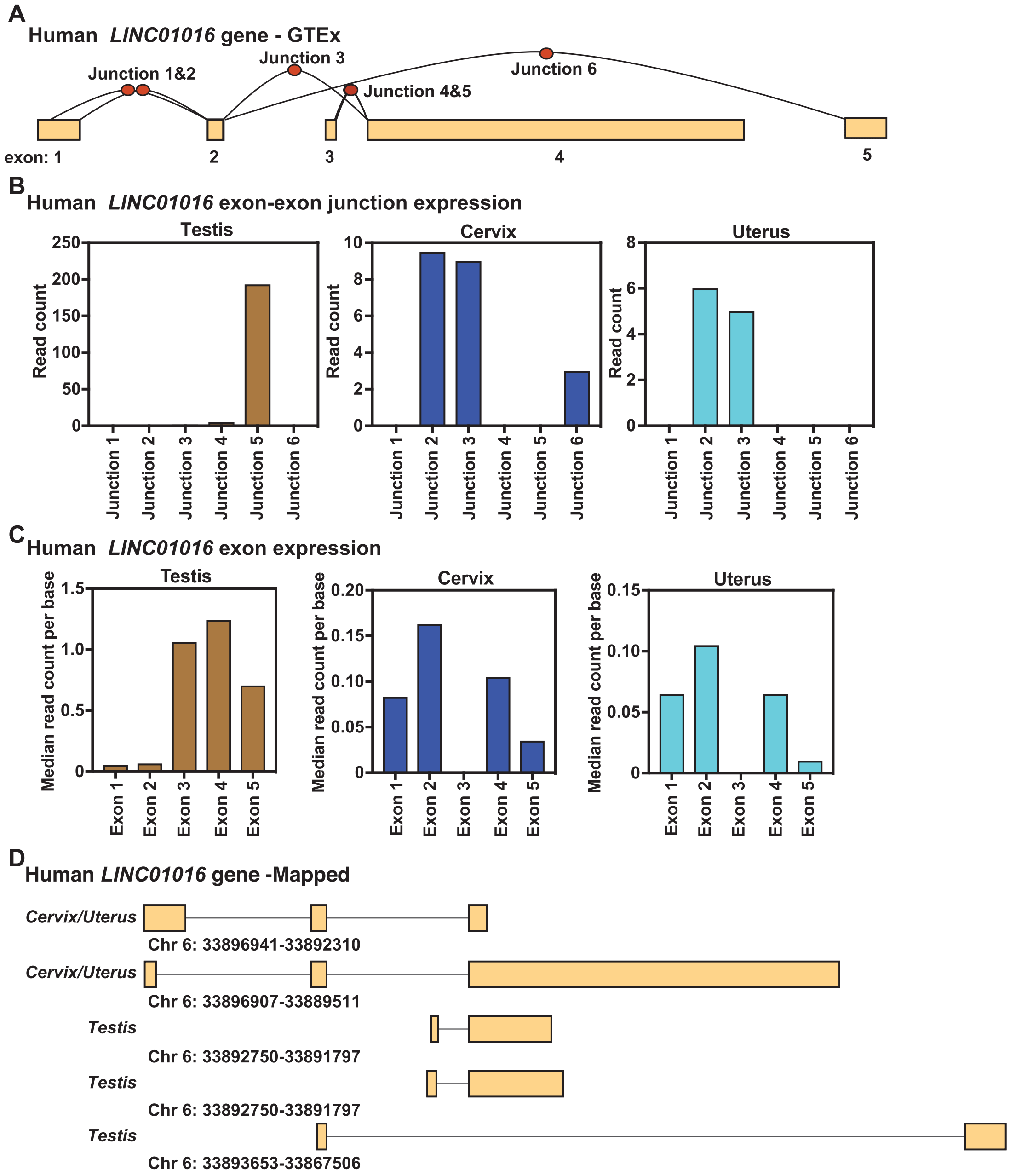

Figure 4. Characterization of the structure of the human long intergenic non-protein–coding RNA 01016 (*LINC01016*) gene by analysis of RNA sequencing data. A, Collapsed view of human *LINC01016* gene exon and exon-exon junction structure from the Genotype-Tissue Expression database (GTEx) [9, 10]. B, Exon-exon junction levels in testes, cervix, and uterus expressed as read count, accessed from GTEx. C, Exon-specific expression in testes, cervix, and uterus obtained from GTEx. D, Diagram of the human *LINC01016* isoforms as analyzed in parts A to C. Exons are depicted as boxes and introns as lines.

12A). Additionally, the analysis for the clinical-stage data shows 3 distinct clusters depending on the clinical stage (stage I, stage II, and stage III) separated by PC1, which accounts for 38.6% variability (Fig. 12B).

Moreover, by interrogating the TCGA database, we also found that aggregate expression of *LINC01016* isoforms is downregulated in CESC, TGCT, and UCEC tissues [6, 22, 23] (Fig. 13A). To determine the clinical value of its altered expression patterns across these cancers, we generated Kaplan-Meier plots using a cancer resource containing samples of cervical squamous cell carcinoma, TGCT, or UCEC [24, 37]. We observed elevated levels of *LINC01016* RNA at the gene level predicted DFI in cervical squamous cell carcinoma and TGCTs, but not in UCEC tumors (see Fig. 13). More important, we performed Kaplan-Meier analysis for isoforms of *LINC01016* (201, 202, 203, 204, and 205) across samples of TGCT, cervical squamous cell carcinoma, or UCEC using the DFI as our metric [36] (Fig.

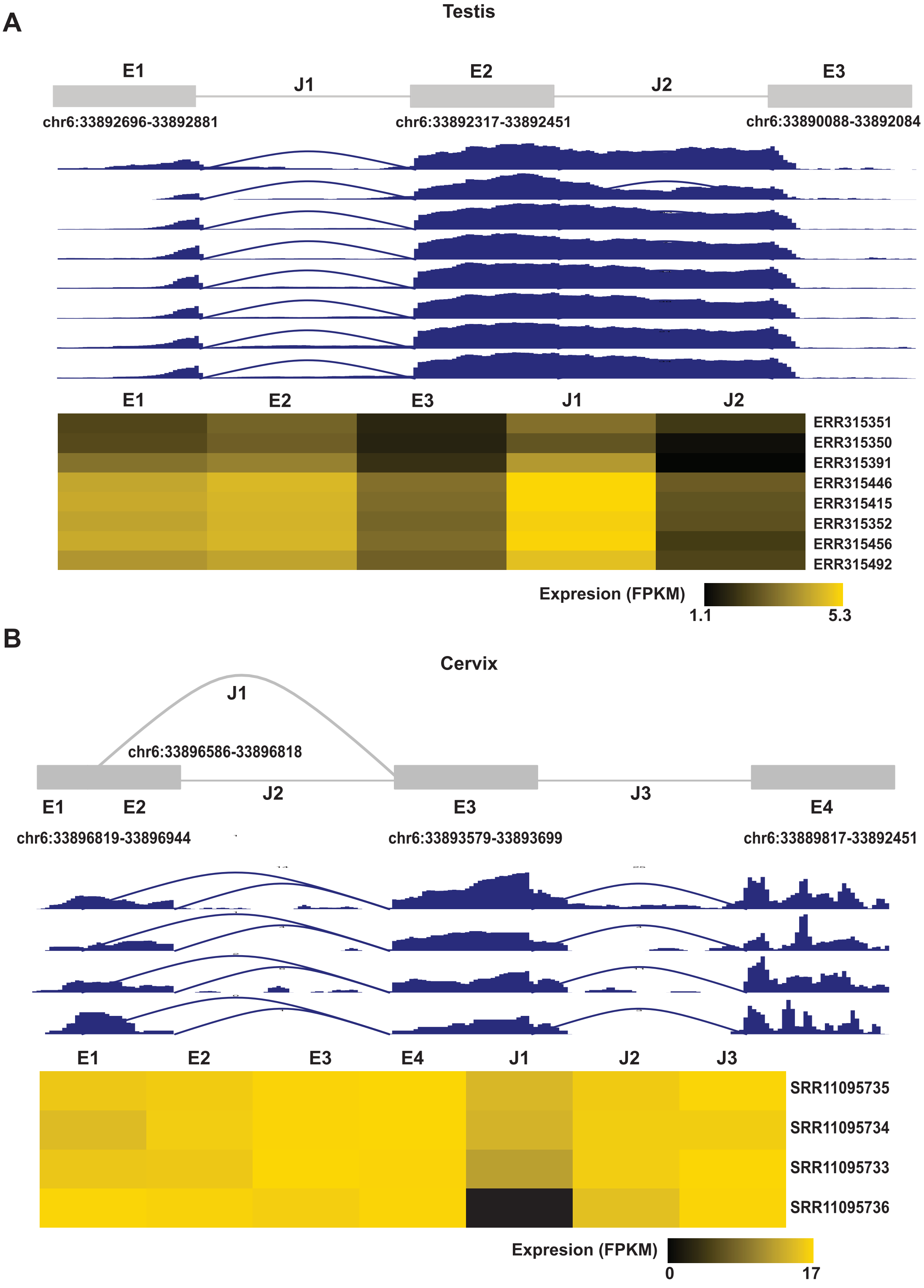

Figure 5. RNA-sequencing predicted splice events, quantification expression, and isoform identification. Splice graph analysis based on de novo prediction for long intergenic non-protein–coding RNA 01016 (*LINC01016*) coverage and expression quantification from RNA sequencing normal data sets of the normal A, testis, and B, cervix.

14). Interestingly, we identified distinct isoform-specific patterns of survival to be dependent on low/high expression of *LINC01016*. Our data show that in TGCTs, isoforms 203 and 204 correlate with better outcomes for DFI with increased expression, whereas elevated expression of isoform 205 showed poor DFI outcome (Fig. 14A).

Moreover, cervical squamous cell carcinoma showed a generally poor DFI prognosis with high expression levels

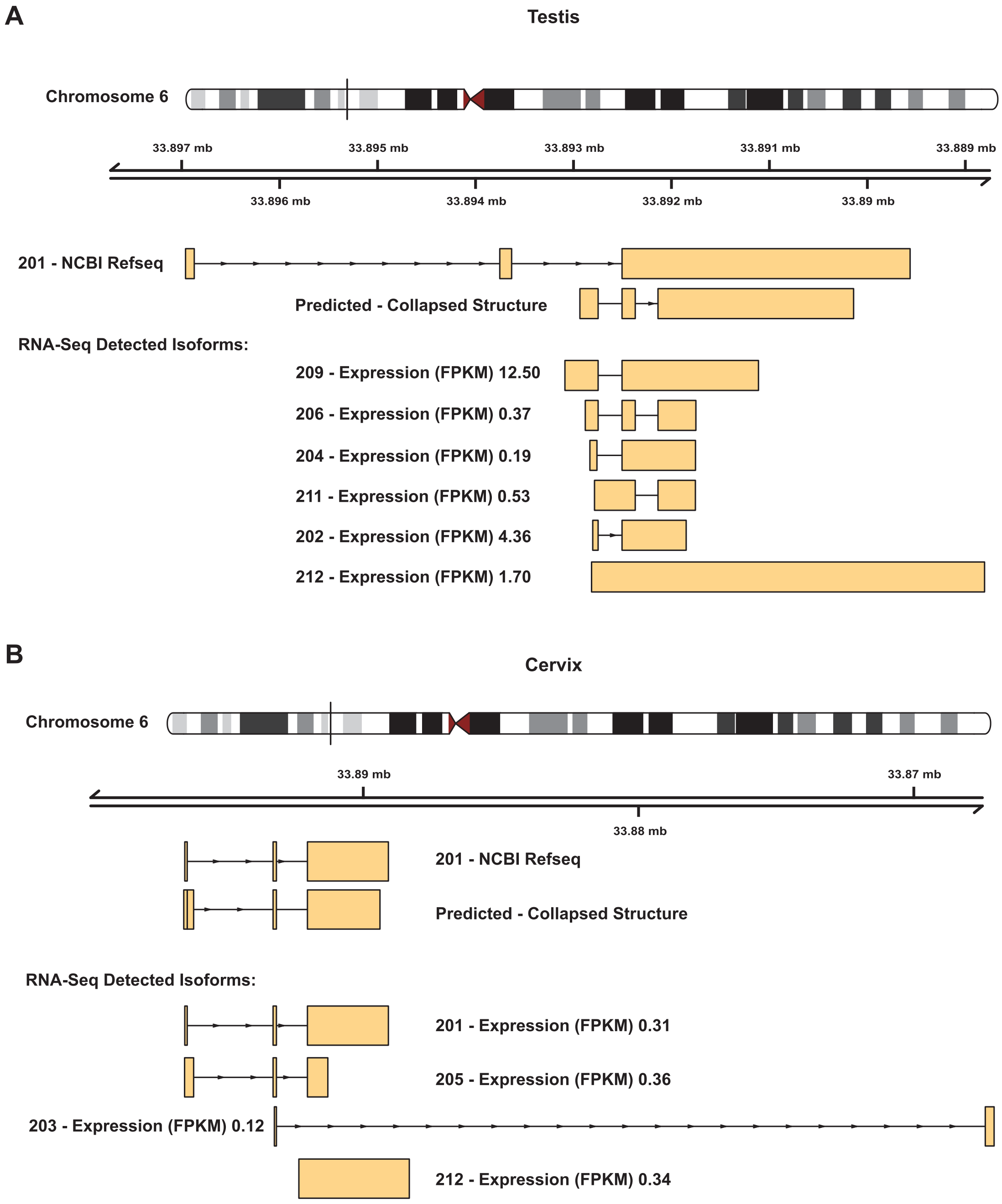
A
Testis
Chromosome 6
33.897 mb
33.895 mb
33.893 mb
33.891 mb
33.889 mb
33.896 mb
33.894 mb
33.892 mb
33.89 mb
201 - NCBI Refseq
Predicted - Collapsed Structure
RNA-Seq Detected Isoforms:
209 - Expression (FPKM) 12.50
206 - Expression (FPKM) 0.37
204 - Expression (FPKM) 0.19
211 - Expression (FPKM) 0.53
202 - Expression (FPKM) 4.36
212 - Expression (FPKM) 1.70
B
Cervix
Chromosome 6
33.89 mb
33.87 mb
33.88 mb
201 - NCBI Refseq
Predicted - Collapsed Structure
RNA-Seq Detected Isoforms:
201 - Expression (FPKM) 0.31
205 - Expression (FPKM) 0.36
203 - Expression (FPKM) 0.12
212 - Expression (FPKM) 0.34

Figure 6. RNA sequencing (RNA-seq)–predicted gene models. A, Gene models for long intergenic non-protein–coding RNA 01016 (*LINC01016*) isoforms expression detected from RNA-seq data sets of the normal A, testis, and B, cervix. Included is isoform 201, which is NCBI Refseq, and the predicted collapsed structure.

across the different isoforms (Fig. 14B). Last, UCEC does not show a significant difference between *LINC01016* isoforms (Fig. 14C). Altogether, these results suggest that respective *LINC01016* isoforms are associated with prognostic value in a subset of cancer types.

Discussion

LncRNAs are found to be aberrantly expressed in tumors and can be excellent diagnostic markers with therapeutic value [1, 38]. Interestingly, a subset of these lncRNAs, including isoforms of the same lncRNA gene, are differentially expressed spatiotemporally across reproductive tissue types such as ovary, placenta, and testis [8, 38-40]. While few have been investigated for their function, the lack of proper annotation presents a formidable challenge in understanding the molecular role of newly identified lncRNAs [7]. Genome-wide transcriptomic analyses help define molecular features of lncRNAs, including ends of transcripts, exon usage,

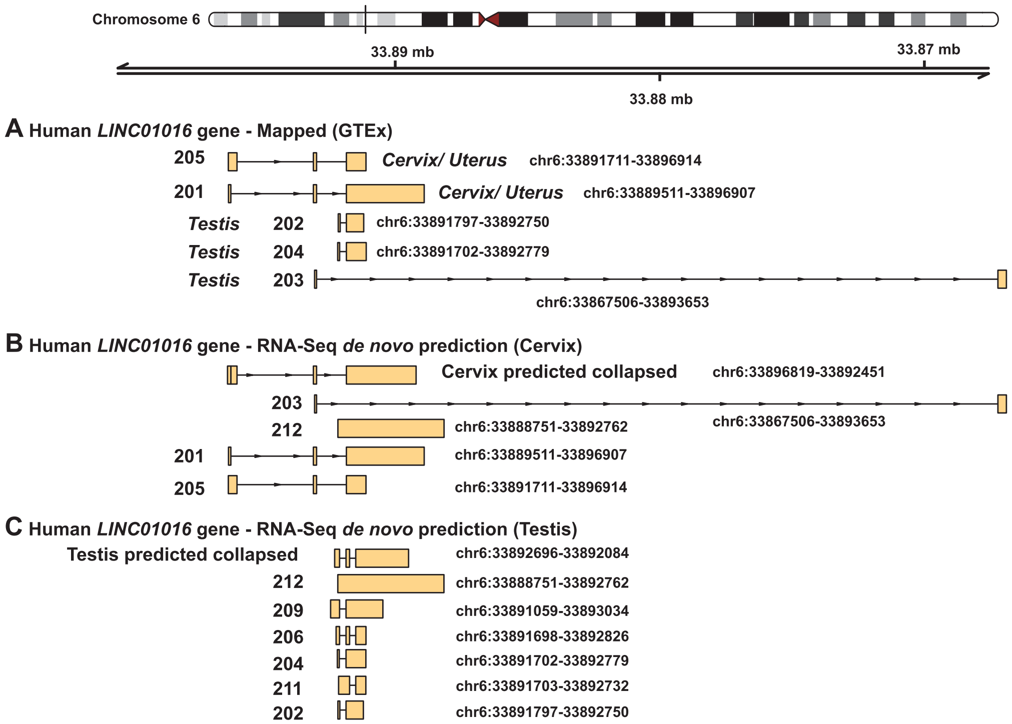
Chromosome 6
33.89 mb
33.87 mb
33.88 mb
A Human LINC01016 gene - Mapped (GTEx)
205 Cervix/ Uterus chr6:33891711-33896914
201 Cervix/ Uterus chr6:33889511-33896907
Testis 202 chr6:33891797-33892750
Testis 204 chr6:33891702-33892779
Testis 203 chr6:33867506-33893653
B Human LINC01016 gene - RNA-Seq de novo prediction (Cervix)
Cervix predicted collapsed chr6:33896819-33892451
203 chr6:33867506-33893653
212 chr6:33888751-33892762
201 chr6:33889511-33896907
205 chr6:33891711-33896914
C Human LINC01016 gene - RNA-Seq de novo prediction (Testis)
Testis predicted collapsed chr6:33892696-33892084
212 chr6:33888751-33892762
209 chr6:33891059-33893034
206 chr6:33891698-33892826
204 chr6:33891702-33892779
211 chr6:33891703-33892732
202 chr6:33891797-33892750

Figure 7. Comparison of gene models for long intergenic non-protein–coding RNA 01016 (*LINC01016*) between normal the Genotype-Tissue Expression database (GTEx), RNA-sequencing (RNA-seq)–predicted normal testis, and RNA-seq–predicted normal cervix. A. Isoforms detected by GTEx database v8 (dbGaP Accession phs000424.v8.p2) for normal cervix/uterus and testis samples. Gene models for *LINC01016*, de novo predicted isoforms using RNA-seq in the B, normal cervix, and C, testis samples.

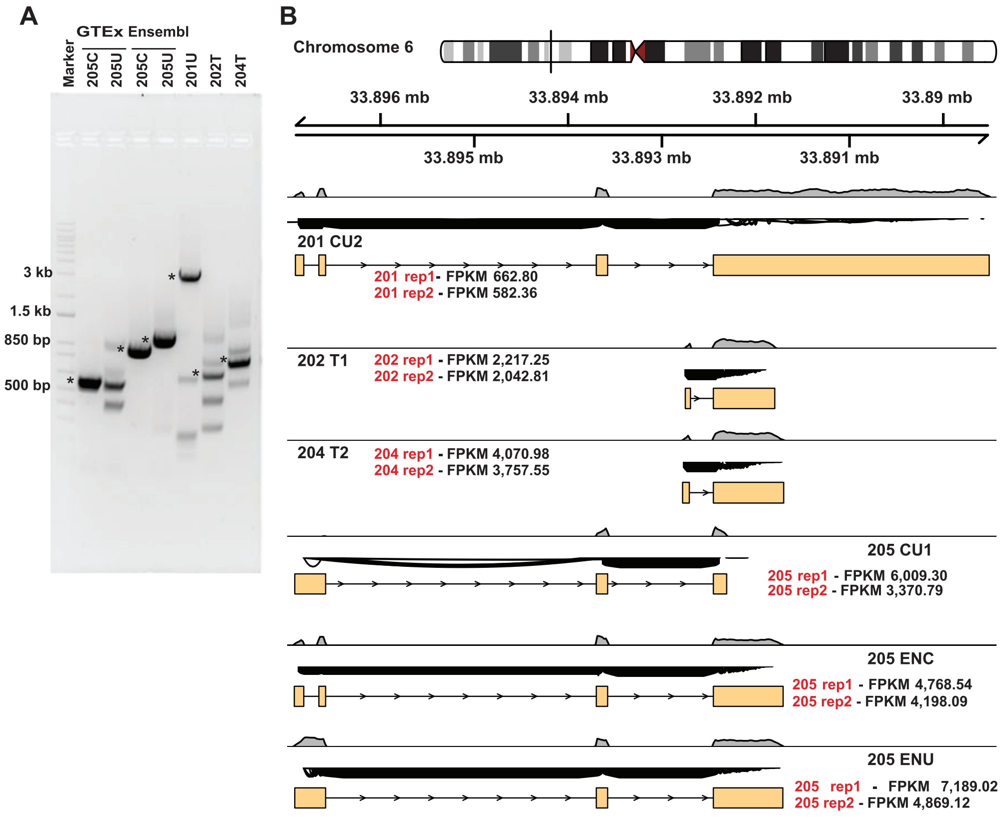

Figure 8. Gene models, coverage, junctions, and expression for long intergenic non-protein–coding RNA 01016 (*LINC01016*) cloned transcripts on experimental RNA sequencing (RNA-seq) data. A, Polymerase chain reaction (PCR) analysis of selected *LINC01016* isoforms. Agarose gel profile of full-length reverse transcription–PCR of *LINC01016* isoforms, cloned in PCDNA3.0 mammalian expression vector. B, The diagram shows the chromosome location, sequencing coverage, junction plots, gene models, and RNA-seq transcript expression (fragments per kilobase of transcript per million mapped reads; FPKM) for each cloned transcript: 201 CU2, 202T1, 204T2, 205 CU1, 205 ENC, and 205 ENU.

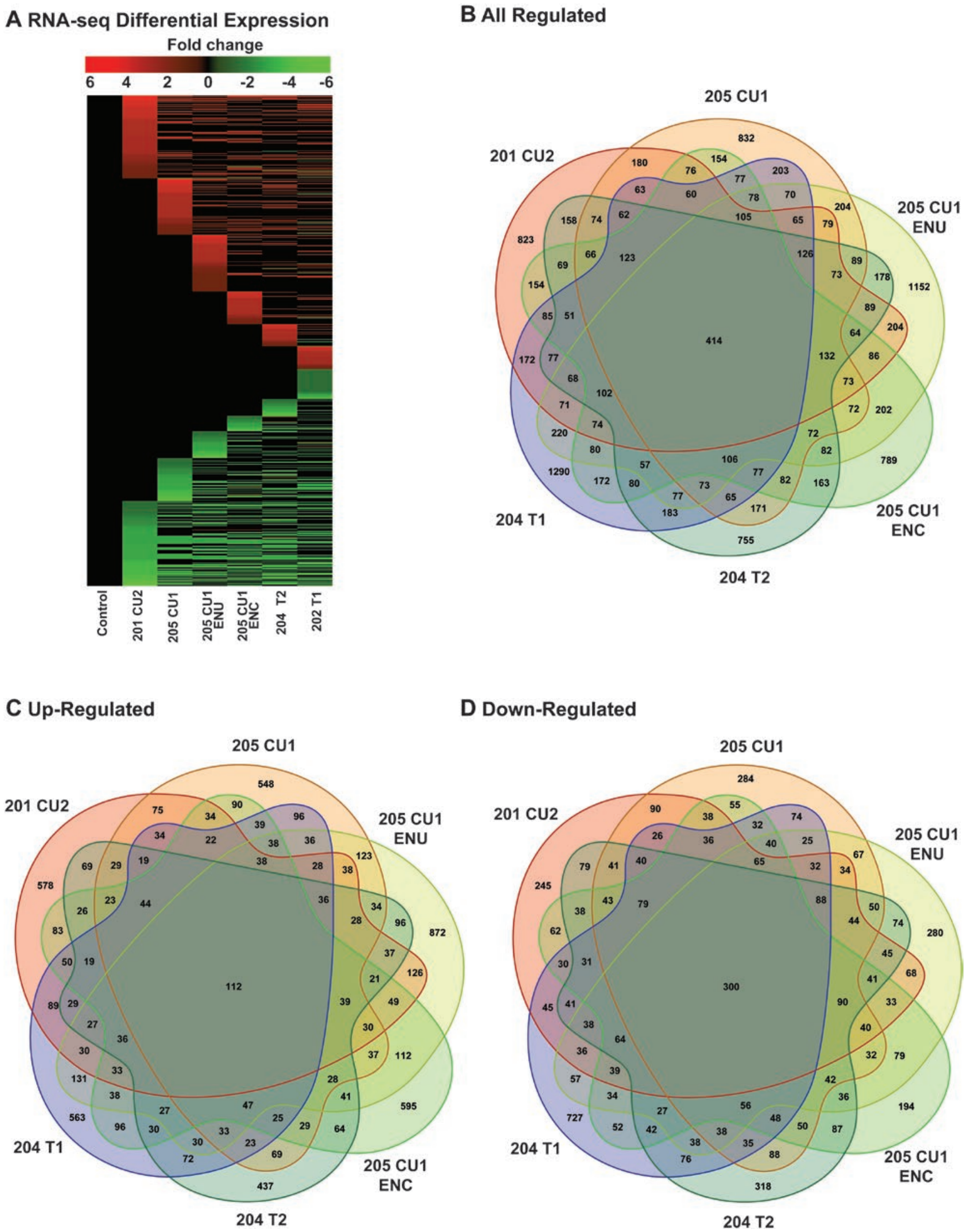

Figure 9. Long intergenic non-protein–coding RNA 01016 (*LINC01016*) cloned transcripts RNA sequencing (RNA-seq) differential expression of genes analyses. A, Heat map showing the log2(fold change) differential expression on cloned *LINC01016* transcripts. B to D, Venn diagrams for all regulated genes from overlapping intersections between the different categories B, all regulated; C, upregulated; and D, downregulated.

and relative isoform abundance [41]; however, detailed gene-specific analysis remains the gold standard to accurately determine the gene structure and quantify the isoform-specific expression of relatively less abundant lncRNAs. In this regard, we have used novel and curated data sets to annotate and quantify a testis-specific long intergenic noncoding RNA, *LINC01016*, in humans and nonhuman primates. Our results indicate that the Ensembl annotation predicts isoforms with 1 to 4 exons (see Fig. 2). In contrast, exon and exon-exon expression analyses show that *LINC01016* with 3 exons is the predominantly expressed transcript in cervix and uterus tissue. The signal-, space-, and time-dependent variability in the expression of isoforms is a widely observed feature

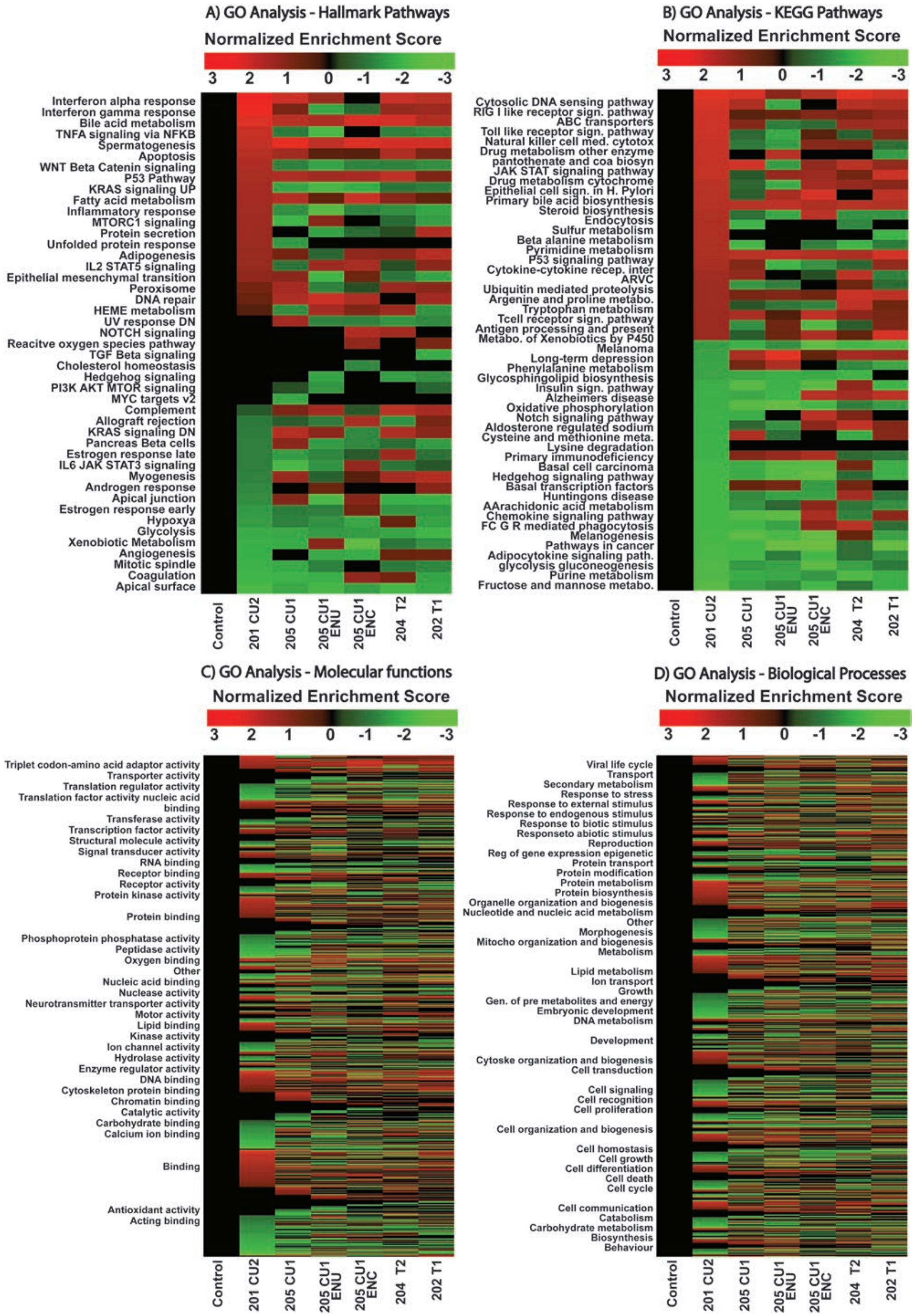

Figure 10. Long intergenic non-protein–coding RNA 01016 (*LINC01016*) cloned transcripts RNA sequencing (RNA-seq) Gene Ontology (GO) analysis. A, GO analysis for cancer hallmark pathways; B, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways heat map for top 25 upregulated/downregulated pathways on the RNA-seq data set; C, molecular functions signatures (molecular functions are grouped by parental term); D, and biological processes signatures (biological processes are grouped by parental term) using normalized enrichment score (NES).

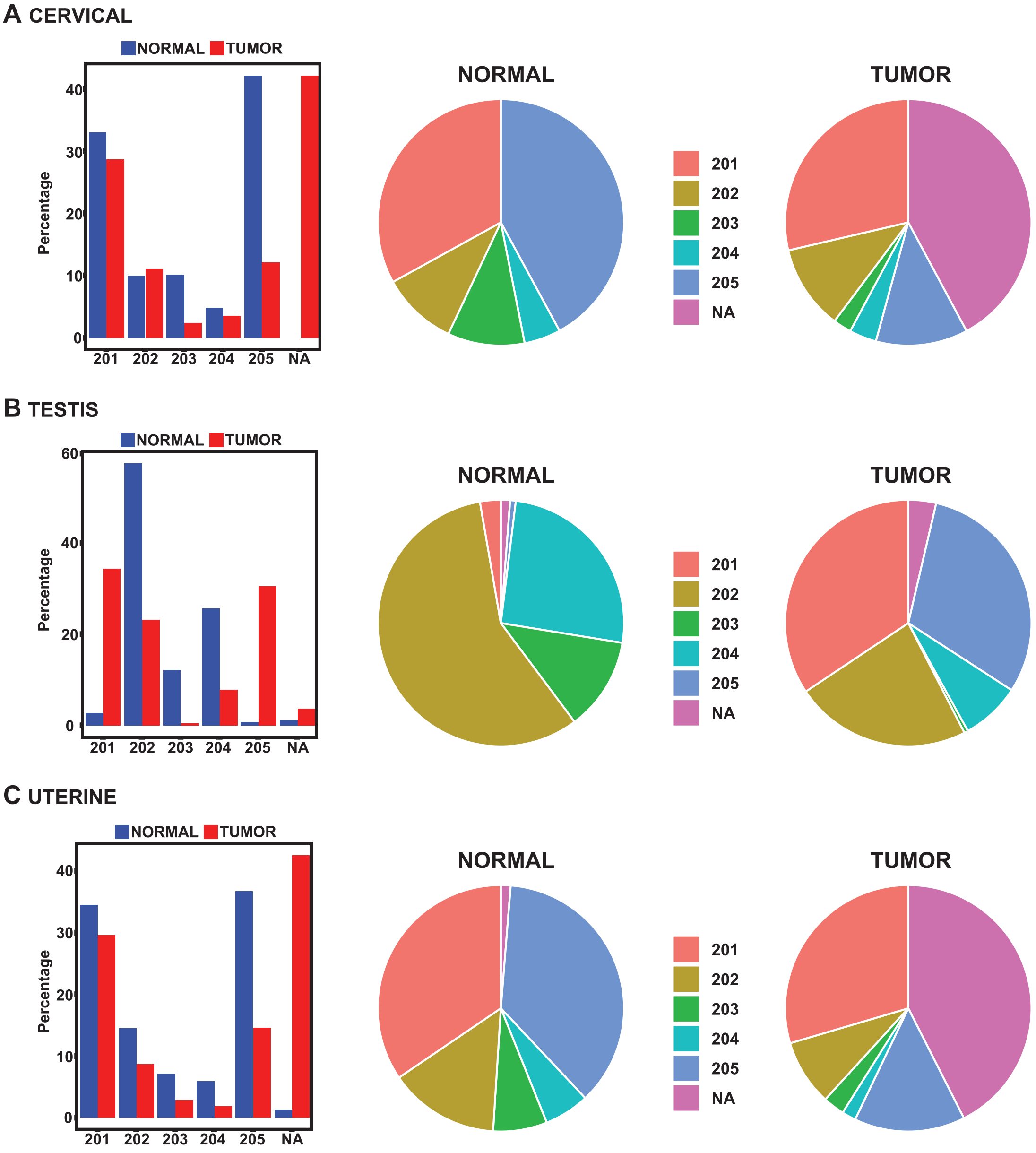

Figure 11. Genotype-Tissue Expression database (GTEx) and The Cancer Genome Atlas (TCGA) long intergenic non-protein–coding RNA 01016 (*LINC01016*) transcript expression from available RNA sequencing data sets on different tissues. *LINC01016* isoforms detected by GTEx [9, 10] for normal tissue and TCGA Pan-Cancer (PANCAN) tumor samples for A, cervical; B, testis; and C, uterine tissue samples. Isoform annotation is based on GENCODE v22.

of human transcripts [42]. Overall, these results suggest an intricate organization of the *LINC01016* locus. This information is very intriguing since variable isoform expression through differential exon usage has been linked to several cancers [43, 44]. The aberrant expression of an isoform from the same gene could lead to protein or RNA with altered functions due to changes in their secondary structures or conformations [45]. Hence, it is imperative to study individual isoforms of a lncRNA gene since each one of them could be functioning independently.

In addition to gene structure, *LINC01016* expression varies across tissues and organs (see Fig. 1). This tissue-specific expression can be leveraged as a potential diagnostic or prognostic marker, a feature common to several lncRNAs [6, 8, 46]. We also observed that in a manner similar to several reproductive tissue-specific genes [8], *LINC01016* is expressed during developmental stages of testis tissue, indicating a plausible role in germ cell development [47]. Given the tissue specificity of *LINC01016*, it will be interesting to study its role in the testis, which could lead to potential functioning in reversibly arresting sperm production. For this purpose, the majority of genes currently studied are those encoding proteins, leaving the gold mine of unstudied lncRNA genes unexplored, which is

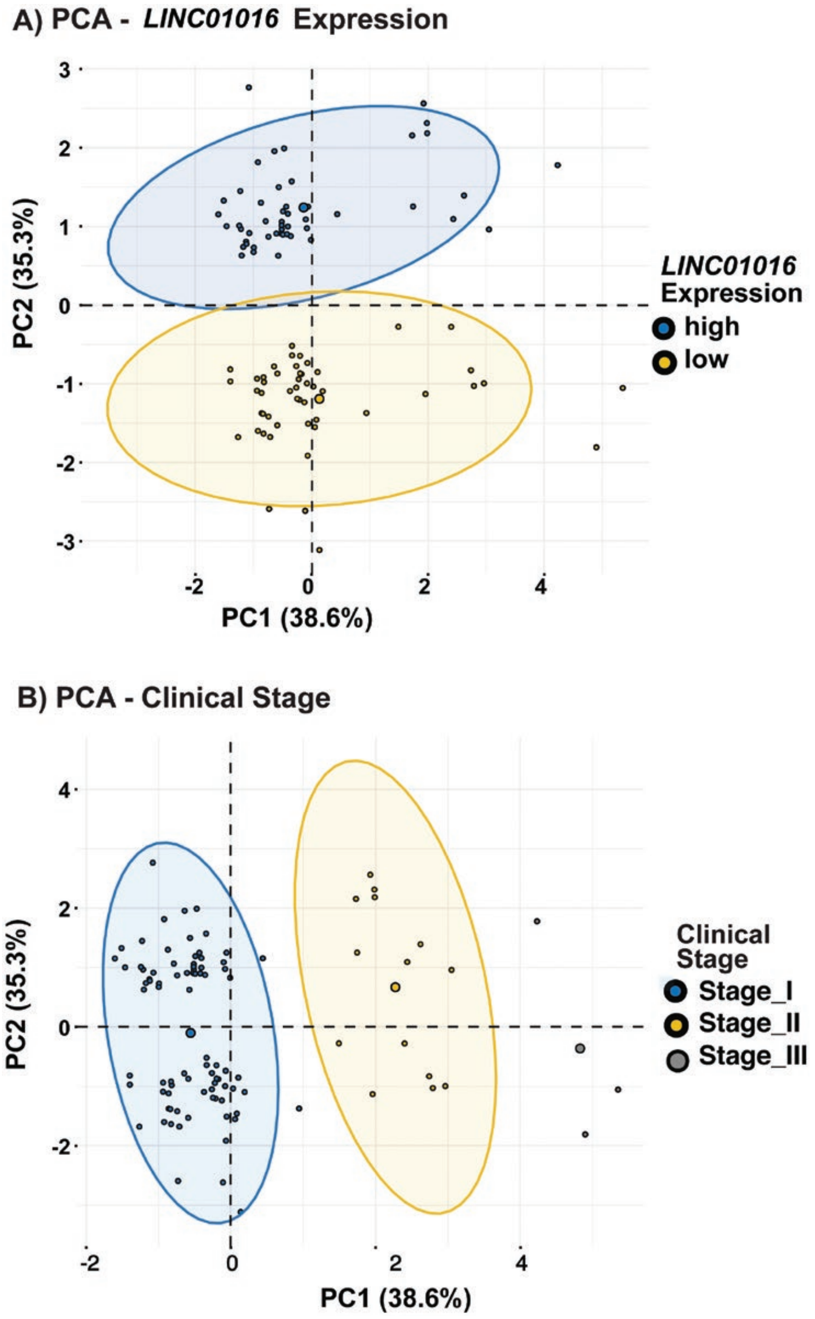

Figure 12. Principal component analysis (PCA) of long intergenic non-protein–coding RNA 01016 (*LINC01016*) expression and clinical staging. A, PCA on *LINC01016* expression shows 2 distinct clusters based on high/low levels of expression separated by principal component 2 (PC2), which accounts for 35.3% variability. B, PCA on the clinical stages of the tumor shows 3 distinct clusters based on stage I/stage II/stage III separated by PC1, which accounts for 38.6% variability.

necessary and valuable for their potential to uncover unexplained phenomena in a medically appropriate way.

Our data show *LINC01016* has significant differential exon usage (see Fig. 3-7), as well as isoform-specific roles in diverse biological processes (see Fig. 10). This is important because this characteristic contributes to the variability in detected isoforms across different samples and to plausible functions regulating various biological processes depending on isoform, tissue-specificity, and present conditions at a specific time point. Differential gene expression of *LINC01016*-regulated genes from RNA-seq using diverse transcripts is a clear representation of gene regulation variability (see Fig. 9). Intersectional Venn diagrams from differentially expressed genes show some overlap of regulation between different isoforms of *LINC01016*, but for the most part, each transcript has a regulation of specific gene networks. Similarly, GO analysis (see Fig. 10) shows that regulation of biological pathways is isoform dependent.

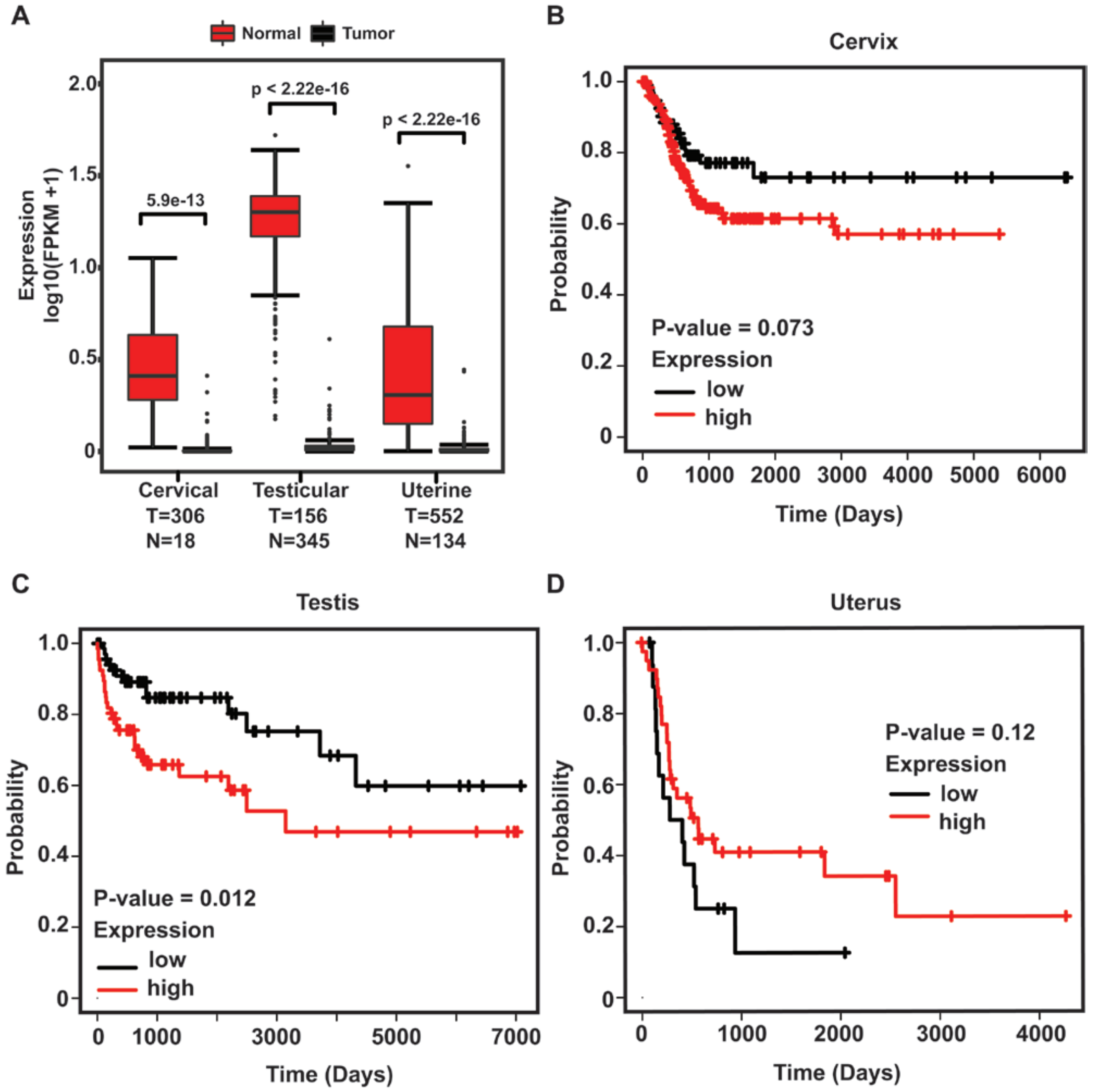

Figure 13. Expression of long intergenic non-protein–coding RNA 01016 (*LINC01016*) in reproductive cancers. A, Box plot representation of *LINC01016* in cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC); testicular germ cell tumors (TGCT); and uterine corpus endometrial carcinoma (UCEC). *LINC01016* RNA levels are lower in tumors (T, black box) than in normal tissues (N, red box). The expression data was accessed from the Genotype-Tissue Expression database (GTEx) portal (version 8) and The Cancer Genome Atlas (TCGA) Genomic Data Commons databases (v28). Kaplan-Meier survival analyses of patients expressing higher (red line) or lower (black line) levels of *LINC01016* RNA in B, cervical squamous cell carcinoma; C, testicular germ cell tumors; and D, uterine corpus endometrial carcinoma. The cancer outcome–linked gene expression data (disease-free interval) was accessed using TCGA Pan-Cancer (PANCAN) and graphed a custom version using kmplot.com [24, 37].

It is critical to point out the latest releases of widely used genomic databases such as TCGA, and GTEx shows only a portion of the identified *LINC01016* isoforms: 5 (see Figs. 7 and 11) because their analysis is based on previous gene annotations and therefore might not reflect the full extent of known isoforms. Newer annotations from Ensembl (version 101) and GENCODE (version 37) show 12 additional isoforms of *LINC01016*, matching predictions from RNA-seq experimental sample data sets. Therefore, it is crucial to compare versions and parameters of genomic databases when designing a study.

Many reproductive tissue-specific lncRNAs are differentially expressed in various cancers due to epigenetically dysregulated promoters and enhancers [2]. Notably, *LINC01016* is downregulated in cervical squamous cell carcinoma, TGCTs, and UCEC samples (see Fig. 13). Survival analysis in TGCTs and cervical squamous cell carcinoma suggests higher *LINC01016* expression to be a predictor of a clinical outcome in patients. Interestingly, testicular cancer shows that isoform-specificity plays an important role in predicting clinical outcomes for patients (see Fig. 14). Previously, *LINC01016* has also been implicated in endometrial and breast cancer [48, 49]. Collectively, these results imply that *LINC01016* could be a promising tumor biomarker.

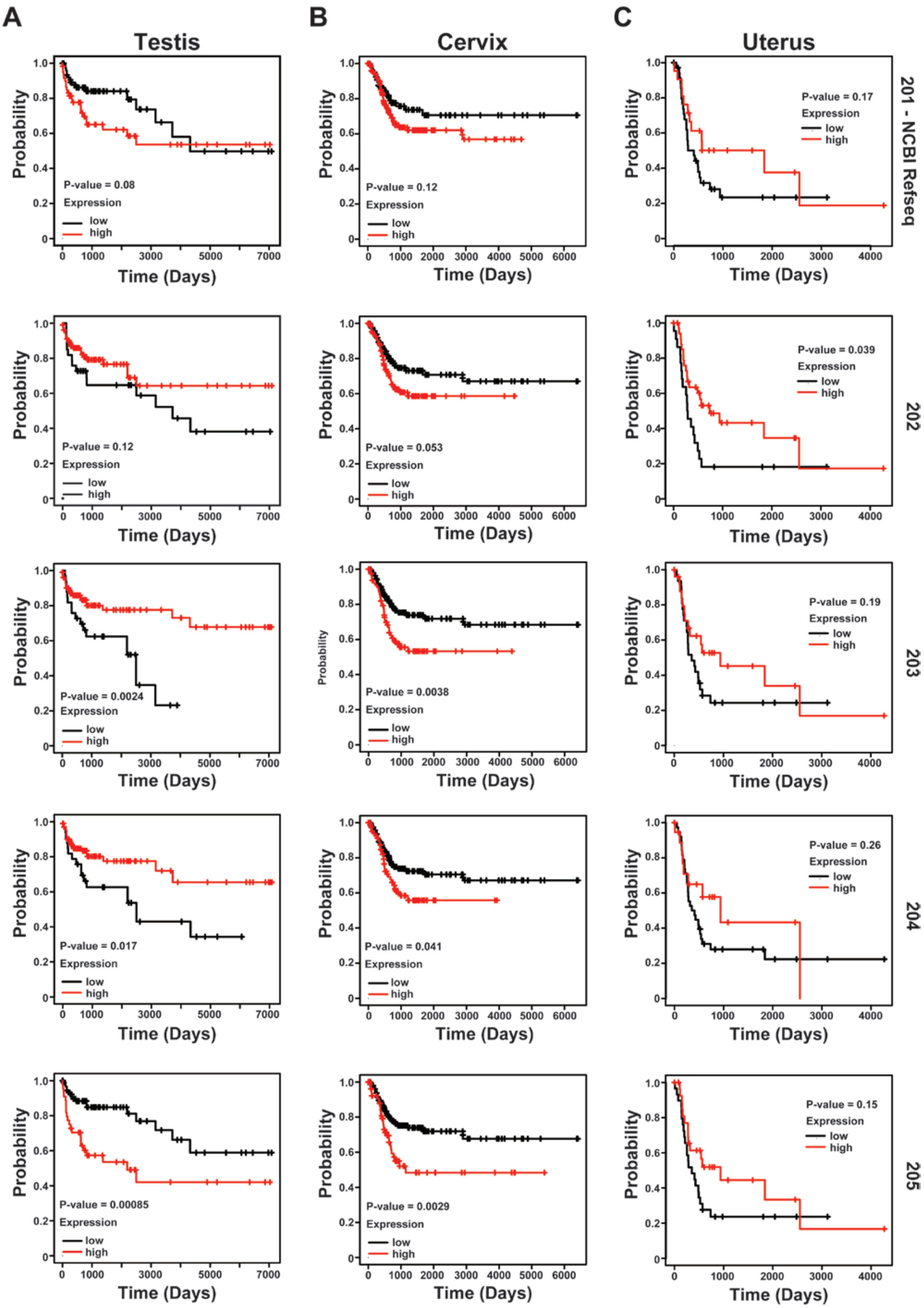

Figure 14. Expression of long intergenic non-protein–coding RNA 01016 (*LINC01016*) in reproductive cancers. Kaplan-Meier survival analyses of patients expressing higher (red line) or lower (black line) levels of *LINC01016* RNA based on different isoforms (201, 202, 203, 204, and 205) in B, cervical squamous cell carcinoma; A, testicular germ cell tumors; and C, uterine corpus endometrial carcinoma. The cancer outcome–linked gene expression data (disease-free interval) was accessed using The Cancer Genome Atlas Pan-Cancer (TCGA PANCAN) and graphed a custom version using kmplot.com [24, 37].

As described earlier, the isoform-specific gene expression analysis showed that the nonoverlapping isoform-dependent gene sets are more extensive and mutually exclusive (see Fig. 9). Therefore, the mechanistic investigation of *LINC01016* expression in target normal and cancer tissues could uncover details such as its interacting protein or RNA partners and its plausible role in chromatin-dependent processes. These details may further open potential avenues for research related to the diagnostic or therapeutic potential of *LINC01016*.

Our study also provides a unique perspective on using previously untapped information embedded in deep-sequencing results from species that are otherwise intractable experimental animal models [50, 51]. Because of the poor conservation in mice, the most widely used mammalian system, studying primate-specific lncRNA genes is challenging because of limited comparable models in which application of lncRNA findings could be recapitulated [2]. Thus, our results provide resources/experimental models as sources of information to bridge this gap in understanding, potentially aiding in developing an experimentally testable hypothesis.

In summary, through our study, we show a) that complexity of gene organization at the DNA and transcript level can be understood by integrating sequencing data with targeted experimental approaches, b) differential exon usage, c) evolutionary significance of *LINC01016*, d) relative expression of *LINC01016* in tissues with a plausible role in reproductive physiology, e) isoform-dependent differences in regulation of gene networks and biological pathways, and f) *LINC01016* as a potential tumor biomarker.

Acknowledgments

We thank Alana L. Harrison and members of the Gadad laboratory for their helpful comments. We thank Texas Tech University High-Performance Computing Center (Lubbock) and TTUHSC El Paso NGS Computing Server (El Paso).

Financial Support: This work was supported by a first-time faculty recruitment award from the Cancer Prevention and Research Institute of Texas (CPRIT; No. RR170020). S.S.G. is a CPRIT Scholar in Cancer Research.

Author Contributions: E.R.: data curation, methodology, and formal analysis; B.Y.: data curation, investigation, writing, review, and editing; Y.M.V.: investigation, writing, review, and editing; R.C.: data curation; K.Y.L.: writing, investigation, review, and editing; and S.S.G.: conceptualization, methodology, investigation, funding acquisition, project administration, writing of original draft, writing, review, and editing.

Additional Information

Correspondence: Shrikanth S. Gadad, PhD, Center of Emphasis in Cancer, Department of Molecular and Translational Medicine, Paul L. Foster School of Medicine, Texas Tech University Health Sciences Center, 5001 El Paso Dr, El Paso, TX 79905, USA. Email: shrikanth.gadad@ttuhsc.edu.

Disclosures: The authors have nothing to disclose.

Data Availability: The new RNA-seq data sets generated for this study are available from the NCBI's Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/) using accession number GSE171047.

References

1. Camacho CV, Choudhari R, Gadad SS. Long noncoding RNAs and cancer, an overview. *Steroids*. 2018;**133**:93-95.
2. Choudhari R, Sedano MJ, Harrison AL, et al. Long noncoding RNAs in cancer: from discovery to therapeutic targets. *Adv Clin Chem*. 2020;**95**:105-147.
3. Sun M, Gadad SS, Kim DS, Kraus WL. Discovery, annotation, and functional analysis of long noncoding RNAs controlling cell-cycle gene expression and proliferation in breast cancer cells. *Mol Cell*. 2015;**59**(4):698-711.
4. Wu P, Mo Y, Peng M, et al. Emerging role of tumor-related functional peptides encoded by lncRNA and circRNA. *Mol Cancer*. 2020;**19**(1):22.
5. Hosono Y, Niknafs YS, Prensner JR, et al. Oncogenic role of THOR, a conserved cancer/testis long non-coding RNA. *Cell*. 2017;**171**(7):1559-1572.e20.
6. Vasquez YM, Nandu TS, Kelleher AM, Ramos EI, Gadad SS, Kraus WL. Genome-wide analysis and functional prediction of the estrogen-regulated transcriptional response in the mouse uterus. *Biol Reprod*. 2020;**102**(2):327-338.
7. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat Rev Genet*. 2018;**19**(9):535-548.
8. Choudhari R, Yang B, Rotwein P, Gadad SS. Structure and expression of the long noncoding RNA gene MIR503 in humans and non-human primates. *Mol Cell Endocrinol*. 2020;**510**:110819.
9. Stranger BE, Brigham LE, Hasz R, et al. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. The eGTEx Project. *Nat Genet*. 2017;**49**(12):1664-1670.
10. GTEx Consortium; Laboratory Data Analysis & Coordinating Center (LDACC)–Analysis Working Group; Statistical Methods groups–Analysis Working Group; et al. Genetic effects on gene expression across human tissues. *Nature*. 2017;**550**(7675):204-213.
11. Quintana-Murci L. Understanding rare and common diseases in the context of human evolution. *Genome Biol*. 2016;**17**(1):225.
12. Manolio TA, Fowler DM, Starita LM, et al. Bedside back to bench: building bridges between basic and clinical genomic research. *Cell*. 2017;**169**(1):6-12.
13. Soumillon M, Cacchiarelli D, Semrau S, van Oudenaarden A, Mikkelsen TS. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *bioRxiv*. March 5, 2014, preprint: not peer reviewed.
14. Vera M, Biswas J, Senecal A, Singer RH, Park HY. Single-cell and single-molecule analysis of gene expression regulation. *Annu Rev Genet*. 2016;**50**:267-291.
15. Sarropoulos I, Marin R, Cardoso-Moreira M, Kaessmann H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature*. 2019;**571**(7766):510-514.

16. Zampetaki A, Albrecht A, Steinhofel K. Long non-coding RNA structure and function: is there a link? *Front Physiol*. 2018;**9**:1201.
17. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. *Nature*. 2012;**482**(7385):339-346.
18. Peng X, Thierry-Mieg J, Thierry-Mieg D, et al. Tissue-specific transcriptome sequencing analysis expands the non-human primate reference transcriptome resource (NHPRTR). *Nucleic Acids Res*. 2015;**43**(Database issue):D737-D742.
19. Djureinovic D, Fagerberg L, Hallström B, et al. The human testis-specific proteome defined by transcriptomics and antibody-based profiling. *Mol Hum Reprod*. 2014;**20**(6):476-488.
20. Xu J, Zou J, Wu L, Lu W. Transcriptome analysis uncovers the diagnostic value of miR-192-5p/HNF1A-AS1/VIL1 panel in cervical adenocarcinoma. *Sci Rep*. 2020;**10**(1):16584.
21. Xu J, Zhang Y, Huang Y, et al. circEYA1 functions as a sponge of miR-582-3p to suppress cervical adenocarcinoma tumorigenesis via upregulating CXCL14. *Mol Ther Nucleic Acids*. 2020;**22**:1176-1190.
22. Weinstein JN, Collisson EA, Mills GB, et al; Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat Genet*. 2013;**45**(10):1113-1120.
23. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res*. 2017;**45**(W1):W98-W102.
24. Nagy Á, Lánczky A, Menyhárt O, Győrffy B. Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets. *Sci Rep*. 2018;**8**(1):9227.
25. Choudhari R, Yang B, Rotwein P, Gadad SS. Structure and expression of the long noncoding RNA gene MIR503 in humans and non-human primates. *Mol Cell Endocrinol*. 2020;**510**:110819.
26. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol*. 2020;**38**(6):675-678.
27. Hahne F, Ivanek R. Visualizing genomic data using Gviz and bioconductor. *Methods Mol Biol*. 2016;**1418**:335-351.
28. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;**14**(4):R36.
29. Li H, Handsaker B, Wysoker A, et al; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;**25**(16):2078-2079.
30. Robinson JT, Thorvaldsdóttir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol*. 2011;**29**(1):24-26.
31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;**15**(12):550.
32. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012;**7**(3):562-578.
33. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*. 2015;**12**(4):357-360.
34. Liao Y, Smyth GK, Shi W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;**30**(7):923-930.
35. Goldstein LD, Cao Y, Pau G, et al. Prediction and quantification of splice events from RNA-Seq data. *PLoS One*. 2016;**11**(5):e0156132.
36. Sergushichev AA. An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*. June 20, 2016, preprint: not peer reviewed.
37. Lánczky A, Győrffy B. Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. *J Med Internet Res*. 2021;**23**(7):e27633.
38. Vallone C, Rigon G, Gulia C, et al. Non-coding RNAs and endometrial cancer. *Genes (Basel)*. 2018;**9**(4):187.
39. Taylor DH, Chu ET, Spektor R, Soloway PD. Long non-coding RNA regulation of reproduction and development. *Mol Reprod Dev*. 2015;**82**(12):932-956.
40. Wang Q, Wang N, Cai R, et al. Genome-wide analysis and functional prediction of long non-coding RNAs in mouse uterus during the implantation window. *Oncotarget*. 2017;**8**(48):84360-84372.
41. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*. 2010;**28**(5):511-515.
42. Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res*. 2018;**46**(2):582-592.
43. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant review with the Integrative Genomics Viewer. *Cancer Res*. 2017;**77**(21):e31-e34.
44. Zhang Z, Pal S, Bi Y, Tchou J, Davuluri RV. Isoform level expression profiles provide better cancer signatures than gene level expression profiles. *Genome Med*. 2013;**5**(4):33.
45. Tian N, Li J, Shi J, Sui G. From general aberrant alternative splicing in cancers and its therapeutic application to the discovery of an oncogenic DMTF1 isoform. *Int J Mol Sci*. 2017;**18**(3):191.
46. Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*. 2011;**25**(18):1915-1927.
47. Deng X, Berletch JB, Nguyen DK, Disteche CM. X chromosome regulation: diverse patterns in development, tissues and disease. *Nat Rev Genet*. 2014;**15**(6):367-378.
48. Jonsson P, Coarfa C, Mesmar F, et al. Single-molecule sequencing reveals estrogen-regulated clinically relevant lncRNAs in breast cancer. *Mol Endocrinol*. 2015;**29**(11):1634-1645.
49. Pan X, Li D, Huo J, Kong F, Yang H, Ma X. LINC01016 promotes the malignant phenotype of endometrial cancer cells by regulating the miR-302a-3p/miR-3130-3p/NFYA/SATB1 axis. *Cell Death Dis*. 2018;**9**(3):303.
50. Rotwein P. Quantifying promoter-specific insulin-like growth factor 1 gene expression by interrogating public databases. *Physiol Rep*. 2019;**7**(1):e13970.
51. Rotwein P. The insulin-like growth factor 2 gene and locus in nonmammalian vertebrates: organizational simplicity with duplication but limited divergence in fish. *J Biol Chem*. 2018;**293**(41):15912-15932.