ciRS-7 exonic sequence is embedded in a long-noncoding RNA locus

Steven P. Barrett[1], Kevin R. Parker[2], Caroline Horn[1], Miguel Mata[3], Julia Salzman[1,4*]

[1]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA

[2]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, CA, USA

[3] Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA

[4] Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA

*correspondence to:
julia.salzman@stanford.edu

Abstract

Despite thorough analysis, the human transcriptome is incompletely annotated: some genes lack accurate transcriptional start sites and in some genes, splicing events have been missed. In this paper, we report a significant example of this incompleteness in both the promoter and splicing of ciRS-7, a highly expressed circRNA thought to be exceptional because it is transcribed from a locus lacking any mature linear RNA transcripts of the same sense. Using unbiased computational approaches, we have discovered that the human ciRS-7 exonic sequence is spliced into linear transcripts. Further, we use statistical approaches to discover that its promoter coincides with that of the long non-coding RNA, LINC00632. We validate this prediction using multiple experimental assays and show that the splicing of ciRS-7 into linear transcripts is conserved to mouse. Together, experimental and computational evidence argue that expression of ciRS-7 is primarily determined by epigenetic state of LINC00632 promoters and that transcription and splicing factors sufficient for ciRS-7 biogenesis are expressed in cells that lack detectable ciRS-7 expression. This unbiased joint analysis of RNA-seq and epigenetic data reveal the potential for discovering important biological regulation missed in current reference annotations. Here, we focus on the significant biological implications for the most intensely-studied circular RNA, ciRS-7.

Introduction

Until recently, the expression of circRNA was an almost completely uncharacterized component of eukaryotic gene expression, with a few important exceptions [1–4]. It is now appreciated that circRNAs are a ubiquitous feature of eukaryotic gene expression [1,3,5,6]. While many functions have been posited for circRNAs, few have been supported with experimental evidence.

ciRS-7, one of the most highly expressed and most intensely studied circRNAs, is an exception to this, where recent work has shown it functions as a miRNA sponge [6,7]. The sequence of ciRS-7 is highly repetitive with over 70 repeated miR-7 seed sequences in humans, most of which are conserved across eutherian mammals, and its expression is highly tissue-specific with highest expression in the brain [6–8]. ciRS-7 also exhibits increasing expression in

neuronal differentiation models *in vitro [9]*. When expressed in zebrafish, which do not have an endogenous copy of ciRS-7 but do express miR-7, expression of the ciRS-7 sequence results in a defect in midbrain development [6]. In spite of these functional findings, model for the biogenesis and regulation of ciRS-7 is lacking.

Biogenesis of circular RNAs has been found to be regulated by intronic sequence flanking the circularized exon [10,11]. However, this mechanism does not appear to control the biogenesis ciRS-7; inserting 1 kb of the endogenous sequence flanking the ciRS-7 exon into a plasmid driven by a CMV promoter was not sufficient to produce ciRS-7 [7], implying that additional sequence is necessary for circularization. Identifying this additional sequence is an especially difficult problem in the case of ciRS-7, as the intron upstream of the circularized exon has not been described due to the lack of an annotated promoter. Even educated guesses about this intron have not been possible because, unlike every other known human circular RNA, ciRS-7 is not included in a linear transcript, obscuring possible transcriptional start sites that would be shared with these isoforms.

Identifying the promoter a circular RNA such as ciRS-7 is non-trivial: unlike for linear RNAs where 5' RACE can determine the transcription start site (TSS), no such approach can be used for ciRS-7. Thus, we designed a method correlating epigenetic marks in the genomic region up- and downstream of the annotated ciRS-7 exon to its expression as an unbiased method for identifying its promoter. This is a general analytic framework that could be applied to any circRNA, but we chose to focus on ciRS-7 because of the biological significance described above and because it has stood out as the only case of a circRNA with no known linear counterpart.

Our analysis led us to discover that the promoters of a nearby uncharacterized locus long non-coding RNA (LINC00632) were responsible for driving ciRS-7 expression. We also discovered that exons from this LINC could splice to the ciRS-7 exon, resulting in unexpected novel linear transcripts that include this exon and whose localization vary depending on the presence of the ciRS-7 sequence.

Together, these results support post-transcriptional coupling between a long-noncoding RNA and ciRS-7, the most abundant circRNA in the human brain, and raise a host of important functional questions about this locus. Our results also represent the first steps toward pinpointing the mechanisms underlying the regulation and biogenesis of ciRS-7.

Results

Computational methods predict the ciRS-7 promoter

As a first step toward identifying the promoters and regions regulating the expression of ciRS-7, we tested for statistically significant correlations between six epigenetic marks in the genomic region surrounding the ciRS-7 locus (+/- 50 kb) and ciRS-7 expression (See Methods for details). Unexpectedly, the Pearson correlation between H3K4me3, H3K4me1, and H3K27ac marks, associated with *cis*-regulatory elements, and ciRS-7 was both high and statistically significant ($p<.001$) at the promoters of a nearby lincRNA, LINC00632, compared to an empirical null distribution composed of a handful of other genomic regions (Fig 1A). We also found a number of other activating marks in the region with signatures suggestive of potential enhancers. No significant signal at regions more proximal to the ciRS-7 exon or upstream of annotated LINC00632 isoforms were observed, providing strong correlative support that ciRS-7 expression is driven from these promoters.

A set of orthogonal experiments validate that ciRS-7 shares a promoter with annotated lincRNAs

To determine whether specific activation upstream of the transcription starts of

LINC00632 isoforms we computationally predicted to drive expression of ciRS-7 were sufficient, we used the promoter-activating CRISPRa system in HeLa [12], which have little to no detectable ciRS-7 expression [7]. Based on this analysis, guides were designed to target the most distal and proximal annotated TSSs. Targeting of the CRISPRa system to either region resulted in induction of the intended ASINC.1 isoforms, and activating either of these promoters strongly induced expression of ciRS-7 (Fig 1B).

This experiment supports the model that the lack of LINC00632 and ciRS-7 expression in HeLa is due to epigenetic regulation, rather than the absence of necessary trans-acting factors. To test the hypothesis that HeLa is competent to express ciRS-7 absent any repressive epigenetic regulation, we transfected a Bacterial Artificial Chromosome (BAC O), containing a genomic fragment starting upstream of the proximal LINC00632 promoter and ending ~50kb downstream of ciRS-7, into HeLa cells (Fig 1C). We detected significant expression of ciRS-7 and LINC00632 from this BAC by both RT-PCR and Northern blot after one day of transfection, suggesting that any (post-)transcriptional machinery required for ciRS-7 expression is present in HeLa cells (Fig 1D, S1 Fig).

As an orthogonal test that these promoters direct ciRS-7 expression, we transfected three other BACs (A-C) each containing the exonic sequence of ciRS-7, but differing in their inclusion of the proximal endogenous promoter of LINC00632 (Fig 1D). Because BAC O transfects more efficiently than BACs A-C, likely due to its relatively small size, we excluded it in this comparison (S2 Fig). Expression of ASINC isoforms and ciRS-7 were correlated and highly dependent on inclusion of the LINC00632 promoter, further supporting the hypothesis that ciRS-7 and LINC00632 isoforms share the same promoters (Fig 1E, S3 Fig).

As a final test of our identification of the ciRS-7 promoter, we created a genomic deletion in HEK293T cells, known to express ciRS-7, that encompasses the approximate positions both putative promotes (based on our CRISPRa experiments). We obtained one homozygous clone and measured the expression levels of ciRS-7 in this strain, which decreased by approximately one thousand-fold (S3 Fig).

ciRS-7 is an exon embedded in mature human linear RNA transcripts

In parallel with our computational approach to identify the promoter of ciRS-7, we tested whether gaps in the reference annotation of human exons might have missed upstream and or downstream splice sites which can be paired, via canonical linear splicing, with splice sites in ciRS-7. Indeed, spliceosomal circRNAs contain both 5' and 3' splice sites flanking their exons and in almost all known cases, circRNA exons are alternatively contained in linear transcripts with upstream and downstream exons [1,3].

To do this, we developed an analysis of RNA-seq reads capable of detecting splicing outside of annotated exonic sequences, as most algorithms that do not use annotations have high false positive and negative rates [13]. We designed a conceptually straightforward approach, entailing breaking a single RNA-seq read into two pseudo paired-end reads and aligning to a reference genome (See Methods for details). Unlike most other approaches, this one permits discovery of splicing at un-annotated exonic boundaries.

We applied this algorithm to H1ESC RNA-seq data. It's only prediction for upstream sequence spliced into ciRS-7 was a cryptic donor splice site within an annotated LINC00632 exon. The same algorithm also predicted splicing between the 3' end of ciRS-7 to a cryptic downstream exon lacking any annotation (Fig 2A). To test these predictions, we performed RT-PCR using primers in in HEK293T (Fig 2B, S4 Fig), a cell line known to express ciRS-7 [8]. Direct Sanger sequencing of resulting products validated our predictions, and included two cryptic 5' splice sites in the final exon of LINC00632 paired with the annotated acceptor of ciRS-7 (S5 Fig). RT-PCR for the downstream exon yielded two splice isoforms, one of which was predicted computationally and the other using an acceptor ~1 kb upstream (Fig 2B (*right*), S5 Fig, S7 Fig).

qPCR for three variants: LINC00632, the LINC00632-ciRS-7 transcript, and ciRS-7 in HEK293T showed that ciRS-7 was ~250-fold more abundant than the LINC00632-ciRS-7 transcript and about ~50-fold more abundant than LINC00632 (S8 Fig). LINC00632-ciRS-7 and LINC00632 transcripts were also RNase R sensitive, demonstrating that they are indeed linear transcripts, while the resistance of ciRS-7 to RNase R treatment was consistent with it being circular (S6 Fig).

The splicing of ciRS-7 into mature linear RNA transcripts is conserved to mouse

In mouse, the same analytic approach above (Methods) predicted cryptic exons flanking ciRS-7. These variants were confirmed by RT-PCR in mouse brain (Fig 2C). qPCR for the junction between the novel upstream exon and ciRS-7 showed this isoform was RNase R sensitive, evidence of it being linear, and ~250 fold less abundant than ciRS-7 (S9 Fig). In this experiment, ciRS-7 was also strongly sensitive to RNase R, as has been reported by others [3]. While exonic sequences flanking ciRS-7 in linear transcripts have no detectable primary sequence homology between human and mouse, such conservation is not necessarily expected for long non-coding RNA [14].

ciRS-7 locus has complex alternative splicing

Exploratory PCR in human uncovered isoforms that splice directly from LINC00632 to cryptic exons downstream of ciRS-7, including skipping of the ciRS-7 sequence and direct splicing into the downstream internal exon of ciRS-7 (Fig 2D, S7 Fig). In mouse, RNA-seq analysis predicted a new circRNA resulting from back-splicing of a cryptic exon 15kb downstream of ciRS-7 to its annotated acceptor, which we validated by PCR and sequencing (Fig 2C). We attempted several PCRs not guided by the RNA-seq analysis described above; in general, these PCR reactions were negative, evidence against a model of pervasive noisy splicing in the locus and in support of the importance of unbiased RNA-seq based algorithms to identify splicing.

Because of the syntenic hosting of ciRS-7 in linear transcripts, and to simplify nomenclature, we have renamed the uncharacterized LINC00632 to **A**lternatively **S**pliced **IN**to **C**iRS-7 (ASINC) and the corresponding mouse locus, “Asinc”; we call the linear variants of LINC00632 lacking ciRS-7 sequence ASINC.1 and those containing it ASINC.2.

Linear and circular isoforms in the ASINC locus are differentially localized

To test for potential differential regulation of ciRS-7, ASINC.1, and ASINC.2, we profiled their localization by fractionating nuclear and cytoplasmic RNA from HEK293T. Using XIST and ACTB as controls for enrichment of nuclear and cytoplasmic fractions, respectively, qPCR demonstrated that, relative to ciRS-7, ASINC.2, and ASINC.1 were enriched in the nucleus with increasing degrees (~9 and 25-fold respectively vs. ciRS-7), suggesting that the ciRS-7 sequence impacts the steady-state localization of transcripts containing it (Fig 2E).

Discussion

Despite intensive study, the promoter and mechanisms regulating ciRS-7 expression have remained mysterious. A significant open question regarding the mechanism of its splicing has remained: is transcriptional regulation of ciRS-7 unique from other circRNAs in that it is not spliced into a linear transcript? We find that the answer to this question is ‘No.’ ciRS-7 is spliced into a linear transcript that is part of a lincRNA locus. This answer also has important implications for experiments that assign function to ciRS-7 sequence. We have re-named the locus ASINC in part because the sequence of ciRS-7 is contained in linear transcripts which could also sequester miR-7, the namesake of ciRS-7.

As a practical implication of this finding, experiments that study the function of the ciRS-7 sequence make what had been a reasonable simplifying assumption that the only transcripts containing the ciRS-7 sequence are circular. Given the results in this paper, it is possible that functions assigned ciRS-7 could in some cases be carried out by the linear ASINC.2 transcripts as both contain the exonic sequence of ciRS-7.

In addition, unbiased analysis and experimental confirmation have revealed the promoters for ciRS-7 expression, as well as other significant activating epigenetic marks that we predict function as enhancers. This work can also as a stepping stone for the field to begin to dissect transcriptional regulation and biogenesis of the ciRS-7 locus. Regulation of alternative splice variants in the ASINC locus, transcription factor binding patterns, and three-dimensional interactions with the promoters we identified can be analyzed across cell types and throughout development. Determining cis-sequence required for ciRS-7 circularization is now a tractable problem because the intronic sequence flanking the circularized exon has been identified.

This analysis could be generalized to other genes whose promoters are not well-annotated. As an example, another well-known circRNA with no annotated promoter is derived from the Sry gene in mouse, which is circularized in mature adult testes but expresses an unspliced mRNA in the developing genital ridge that governs sex determination [2]. It has been hypothesized, though remains untested, that the promoter for the Sry circRNA uses a separate promoter from the mRNA [15]. The analysis presented here could provide evidence for this model and, furthermore, could point to the unknown site of transcription initiation. Further, a handful of highly-expressed circRNAs in human are derived from the first exon of an annotated mRNA [16]. Given what we have found here, we predict that the promoter of these genes has been misidentified and that more upstream exons in the gene exist.

We also found that BACs containing a fragment of the ASINC locus that includes its promoter can produce ciRS-7 when introduced to HeLa cells which have little to no expression of ciRS-7. This suggests that the transcriptional and splicing machinery necessary for ciRS-7 expression is likely general and that expression of the locus may be primarily regulated at the level of epigenetic modification of the locus, either at the newly-discovered promoters or putative enhancers.

ASINC transcripts are differentially localized in the cell depending on their inclusion of the ciRS-7 sequence, suggesting active regulation of or by the transcripts potentially through factors that bind the sequence in ciRS-7. Specifically, our data support a model where the ciRS-7-containing linear ASINC.2 may have different functions in the nucleus than the ciRS-7 circRNAs in the cytoplasm. Indeed, the cytoplasmic ciRS-7 circles have been shown to function by sequestering mir-7 [6,7], a mechanism that unlikely to be employed by nuclear ASINC.2. Further study of transcripts in the entirety of the locus and their regulation may reveal new functions for ciRS-7 and this locus as a whole.

Methods

RNA-seq analysis for detection of novel splice isoforms

Raw RNA-seq reads from SRR5048080 (human) and SRR1785046 (mouse) were downloaded from the SRA. Reads were mapped and analyzed using KNIFE [17]. Reads failing to align ("unaligned") by KNIFE were used in a simple custom mapping approach to identify novel splicing events within a single read: single reads were split into pseudo-pairs (k-mers) by taking the first 20 mer in the unaligned read and the remaining k-mer defined by the start position O (30 for mouse and 50 for human because of differing input read lengths) and the minimum of O+20 and the remaining read length after trimming (see supplemental perl script). Only reads where each pseudo-pair > 18 nt were used for analysis. These pseudo-paired end reads, which actually came from the same single read were then realigned separately as pairs using bowtie2

to a custom index made by the following sequences:

>mm10_knownGene_uc012hid.1 range=chrX:61083246-61285558 5'pad=0 3'pad=0 strand=- repeatMasking=none
>hg38_knownGene_uc004fbf.2 range=chrX:140683176-140884660 5'pad=0 3'pad=0 strand=+ repeatMasking=none

Pairs of pseudo-reads mapping on the same strand were identified and used for further analysis. This approach differs from other algorithms because it (a) does not use a seed and extend approach; and (b) reads are aligned to a 100kb radius of ciRS-7 rather than the reference genome The complete analysis, the R script used along with output is provided in Supplemental File 4. Although our analysis is likely to include other novel splicing in the locus, we focused our attention on queries for reads that supported a splice from an un-annotated location upstream of the acceptor in ciRS-7 and from the donor in ciRS-7 to an un-annotated downstream exon using criteria on the position of discordant pseudo-paired end mappings (see R script). An example of the reads that supported these events were (in human: SRR5048129.92905492_2774860 (upstream exon); SRR5048080.59411858_2626289 (downstream exon) and in mouse: SRR1785046.9826290 (upstream exon); @SRR1785046.13497132 HWI-ST1148:158:C3UJCACXX:2:2112:5228:93046 length=50 (downstream exon backsplicing to ciRS-7 acceptor).

ENCODE data analysis

For ChIP-seq analysis of histone modifications, processed narrowPeak files aligned to hg19 were downloaded from the ENCODE portal. All samples for ChIP-seq were selected with the following filtering criteria, based on annotations in the metadata annotation file downloaded from https://www.encodeproject.org/metadata/type=Experiment&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens/metadata.tsv: “Assay” == “ChIP-seq”, “File format” == “bed narrowPeak”, “Output type” == “peaks”. Only samples with availability of H3K4me1, H3K4me3, H3K9me3, H3K27ac, H3K27me3, and H3K36me3 were selected. For RNA-seq analysis, raw reads were similarly downloaded from the ENCODE portal. Total RNA-seq experimental data were filtered based on the following criteria: “File format” == “fastq”, “Output type” == “reads”, “Biosample treatment” == null, “Library depleted in” == “rRNA”, “Biosample subcellular fraction term name” == null. The lists of cell types for which satisfactory ChIP-seq and RNA-seq data were cross-referenced to identify the list of 34 cell types for which data was analyzed.

ChIP-seq narrowPeak file were processed according to the following pipeline: the enrichment scores of any peak calls in genomic 500bp bins spanning 50kb up- and down-stream of the genomic locus were identified using the bedtools merge command. If multiple peaks were called in a given bin, the peak with the highest enrichment score was reported. The absence of a peak was reported as a 0, which could indicate either complete lack of signal or insufficient signal over input to call a significant peak in the ENCODE pipeline. Fold enrichment values for replicate ChIP-seq experiments performed on the same cell type were averaged for a given genomic bin.

RNA-seq reads were quantified using Kallisto quant against a custom Kallisto index consisting of RefSeq cDNA transcripts, exclusive of any covering the ASINC/CiRS-7 locus, plus sequences corresponding to individual ASINC/ciRS-7 exons transcribed from both the plus and minus strands [18]. For single-ended data, the quant --[fr/rf]-stranded -b 0 -t 2 -l 200 -s 20 --

single command was used, using either --fr-stranded or --rf-stranded depending on the order of the input files. For paired-ended data, the quant --[fr/rf]-stranded -b 0 -t 2 command was used, again using either --fr-stranded or --rf-stranded depending on the input read files.
Expression data was quantified as 1000*tpm/transcript length. Expression values (RPKM) were averaged across replicates for a given cell type for a given transcript.

Decoy ChIP marks on chromosomes 7 and 12, corresponding to regions +/- 50 kb upstream and downstream of the annotated TSS and transcription stop sites respectively in the ACTB and HOTAIR loci, were used as an empirical null for determining the FDR for correlations between ChIP enrichment in putative promoters and enhancers (S3 File). We computed a conservative statistical significance value as follows. For each mark, we computed the correlation between ciRS-7, measured as exonic TPM for bins on chromosomes 7 and 12, and generated the empirical distribution of these correlations. The FDR was estimated using the empirical null distribution of correlations between ciRS-7 and ASINC transcripts with of marks on chromosomes 7 and 12 using a one-sided KS-test for the distribution of correlations between marks in the ciRS-7 locus exceeding from the distribution of null correlations, a test that would identify positively correlated marks. We report the output of the KS-test statistic as well as the binomial p value for testing whether more bins with the top correlations with ciRS-7 are found in the ciRS-7 locus then by chance, as well as their identity in S5 File. We model-checked our assumptions for the empirical null distribution by showing that there was no significant correlation or anticorrelation between ciRS-7 and ACTB (Pearson r = 0.27, p-value = 0.12) or ciRS-7 and HOTAIR (Pearson r = -0.09, p-value = 0.61), as such effects could distort our null model. We also note that because we exclude bins with no signal, our estimate of the FDR is conservative.

Heatmaps were generated by averaging ChIP-seq peak enrichment for a given 500bp genomic bin across all sample replicates, and computing the Pearson correlation coefficient of ChIP enrichment against RNA-seq expression for a given cell type.

Bacterial Artificial Chromosomes (BACs) and plasmid vectors

BACs were purchased from Thermo Fisher Scientific (Waltham, MA) in the case of BAC CTD-2166E9, and from the BACPAC Resources Center (Children's Hospital Oakland Research Institute, CA) in the case of all other BACs. BACs were purified from E. coli using the Nucleobond Xtra BAC Maxi Kit (Macherey-Nagel, Duren, Germany). SP-dCAS9-VPR (Addgene ID: 63798) was provided by the Qi lab [12], and the guide encoding plasmid, pMCB306 (Addgene ID: 89360), was a gift from Michael Bassik [19]. Guides were cloned into pMCB306 cloned into the BlpI/BstXI site using annealed oligos with the appropriate sticky ends (S1 File).

BAC Fingerprinting

To ensure BACs had the proper insert, 3 µg of each BAC were digested with 12 units of Ban I (NEB, Ipswich, MA) in the manufacturer's buffer for 1.5 hours at 37°C. The digests were then heated to 65°C for 20 min. The digestion fragments were separated by loading 750 ng per lane on a 1% LE Agarose (GeneMate) gel with 0.5X TAE running buffer. The DNA was visualized with ethidium bromide staining. Simulated BAC fingerprints were created using SnapGene software (from GSL Biotech) (S10 Fig).

Transfections

All transfections were performed in 6-well plates using 7.5 µL of Lipofectamine 3000 and 2.5 µg of total DNA per well according to the manufacturer's protocol. Unless noted otherwise, cells

were harvested 24 hours after transfection.
For CRISPRa experiments, we introduced 1.25 µg of the SP-dCAS9-VPR plasmid and 1.25 µg of combined sgRNA plasmids (four for each promoter tested). Cells were harvested 48 hours after transfection.

Nuclear/Cytoplasmic Fractionation

1-2x10^6 293T cells were fractionated for nuclear/cytoplasmic RNA using the PARIS kit (Thermo Fisher Scientific) according to the manufacturer's instructions. 0.25-0.5 µg of RNA of each fractionated sample was used in the RT prior to qPCR, using an equal RNA mass for both the nuclear and cytoplasmic fractions in each experiment.

RNA Purification (primary tissue)

Snap-frozen total brain tissue from a 12-week old C57BL/6 pregnant female mouse was homogenized in TRIZOL and the aqueous phase was purified on Purelink RNA column with on-column DNase treatment.

RNA Purification (cell lines)

Cells were lysed directly in tissue culture plates by the direct addition of TRIzol reagent (Thermo Fisher Scientific). The manufacturer's protocol was followed with the following modifications: after isolation of the aqueous phase, 1 volume of 100% EtOH was added to the sample, and then the entire volume was applied to and spun through a RNA Clean & Concentrator-5 column (Zymo, Irvine, CA). The column protocol was performed as per the manual's instructions starting from the application of the RNA Prep Buffer.

RNase R treatment

1 µg of RNA was treated with 5 U RNase R (Epicentre, Madison, WI) (or no enzyme in the case of the mock) in 10 µL total reaction volume at 37 C for 30 min. 1 µL 1 mM EDTA, 10 mM dNTPs, and 25 uM random hexamers were then added to the sample and the sample was then heated to 65 C for 5 min to denature RNA structures. The RNA was then reverse transcribed without purification with the addition of 4 µL 5x supplement buffer (250 mM Tris pH 8, 125 mM KCl, 15 mM MgCl2) and 2 µL of 0.1 M DTT to provide the necessary conditions for the RT reaction.

RT-PCR and qPCR

Total RNA was reverse transcribed with random hexamers using 100 U Maxima Reverse Transcriptase (Thermo Fisher Scientific) according to the manufacturer's instructions.
Endpoint PCRs were performed using DreamTaq DNA Polymerase (Thermo Fisher Scientific, Waltham, MA). For all PCR reactions, 1.5 µl of the unpurified RT-reaction was used per 50 µl reaction volume. All RT-PCR reactions were performed using the recommended cycling protocol for 35 cycles.
qPCR reactions were assembled as 10 µl reactions using AccuPower 2X GreenStar qPCR Master Mix (Bioneer, Daejeon, Korea) with 0.3 µl of template used per reaction. qPCRs were performed on an ABI 7900HT using following cycling protocol: 50℃ for 20 min, 95℃ for 10 min, (95℃ for 15 s and 60˚C for 60 s) × 45 cycles, followed by a dissociation stage.

Northern Blot

Northern Blots were performed using 5 µg of total RNA per well using the NorthernMax kit (Thermo Fisher Scientific) according to the manufacturer's recommendations. Single-stranded DNA oligos were used as probes and were purchased from IDT (Coralville, IA). Probe sequences can be found in S1 File.

Acknowledgments

We thank Peter Wang, Peter Sarnow, Zoe Davis and Robert Bierman for discussion and critical comments that improved our work; the Krasnow and Herschlag Labs for sharing reagents and space; the Qi and Bassik labs for plasmids. Inga Jarmoskaite for her help with radiation safety training.

References

1. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO (2012) Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS ONE 7: e30733. doi:10.1371/journal.pone.0030733.

2. Capel B, Swain A, Nicolis S, Hacker A, Walter M, et al. (1993) Circular transcripts of the testis-determining gene Sry in adult mouse testis. Cell 73: 1019–1030. doi:10.1016/0092-8674(93)90279-Y.

3. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, et al. (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA 19: 141–157. doi:10.1261/rna.035667.112.

4. Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, et al. (1991) Scrambled exons. Cell 64: 607–613.

5. Wang PL, Bao Y, Yee M-C, Barrett SP, Hogan GJ, et al. (2014) Circular RNA is expressed across the eukaryotic tree of life. PLoS ONE 9: e90859. doi:10.1371/journal.pone.0090859.

6. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, et al. (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 495: 333–338. doi:10.1038/nature11928.

7. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, et al. (2013) Natural RNA circles function as efficient microRNA sponges. Nature 495: 384–388. doi:10.1038/nature11993.

8. Hansen TB, Wiklund ED, Bramsen JB, Villadsen SB, Statham AL, et al. (2011) miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. EMBO J 30: 4414–4422. doi:10.1038/emboj.2011.359.

9. Rybak-Wolf A, Stottmeister C, Glažar P, Jens M, Pino N, et al. (2015) Circular rnas in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol Cell 58: 870–885. doi:10.1016/j.molcel.2015.03.027.

10. Liang D, Wilusz JE (2014) Short intronic repeat sequences facilitate circular RNA production. Genes Dev 28: 2233–2247. doi:10.1101/gad.251926.114.

11. Kramer MC, Liang D, Tatomer DC, Gold B, March ZM, et al. (2015) Combinatorial control of Drosophila circular RNA expression by intronic repeats, hnRNPs, and SR proteins. Genes Dev 29: 2168–2182. doi:10.1101/gad.270421.115.

12. Chavez A, Scheiman J, Vora S, Pruitt BW, Tuttle M, et al. (2015) Highly efficient Cas9-mediated transcriptional programming. Nat Methods 12: 326–328. doi:10.1038/nmeth.3312.

13. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, et al. (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10: 1185–1191. doi:10.1038/nmeth.2722.

14. Johnsson P, Lipovich L, Grandér D, Morris KV (2014) Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochim Biophys Acta 1840: 1063–1071. doi:10.1016/j.bbagen.2013.10.035.

15. Hacker A, Capel B, Goodfellow P, Lovell-Badge R (1995) Expression of Sry, the mouse sex determining gene. Development 121: 1603–1614.

16. Zhang X-O, Wang H-B, Zhang Y, Lu X, Chen L-L, et al. (2014) Complementary sequence-mediated exon circularization. Cell 159: 134–147. doi:10.1016/j.cell.2014.09.001.

17. Szabo L, Morey R, Palpant NJ, Wang PL, Afari N, et al. (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biol 16: 126. doi:10.1186/s13059-015-0690-5.

18. Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34: 525–527. doi:10.1038/nbt.3519.

19. Han K, Jeng EE, Hess GT, Morgens DW, Li A, et al. (2017) Synergistic drug combinations for cancer identified in a CRISPR screen for pairwise genetic interactions. Nat Biotechnol 35: 463–474. doi:10.1038/nbt.3834.

Figure Captions

Fig 1. ciRS-7 shares a promoter with annotated LINC isoforms. **(A)** Heatmap correlation (Pearson r) between strand-specific ciRS-7 expression and enrichment of histone marks across the ciRS-7 locus and surrounding genomic region (50 kb up- and downstream of annotated elements). Correlations are plotted in 500 nucleotide bins, and a depiction of annotated genes is shown below. **(B)** RT-PCR for ciRS-7 and LINC00632 transcripts in HeLa (+/- VPR CRISPR activation) with guides targeting distal and proximal promoters identified in (A). **(C)** Schematic of BAC inserts with respect to the ciRS-7 and LINC00632 genes. **(D)** RT-PCR *(Left)* and Northern blot *(Right)* for ciRS-7 and LINC00632 transcripts generated after BAC O transfection. **(E)** RT-PCR for ciRS-7 and LINC00632 transcripts from HeLa cells transfected with BACs A-C.

Fig 2. ciRS-7 exonic sequence is included in linear transcripts. **(A)** Schematic of the locus including PCR primers used in this study. Dotted lines indicate approximate positions of newly discovered introns. Figure is not drawn to scale. **(B)** RT-PCR of circular and linear ciRS-7 splice products from HEK293T. *(Left)* (1,A): control PCR for ciRS-7; other lanes: LINC00632 exons spliced to the ciRS-7 exonic sequence;*(Right)* PCR of spliced products that include cryptic exons downstream of ciRS-7. (‡) represents rolling circle ciRS-7 PCR products with and without intron retention. (*) Other products were also identified (see S7 Fig). **(C)** RT-PCR of circular and

linear ciRS-7 splice products from mouse brain RNA. **(D)** Examples of novel splicing observed in the human and mouse ASINC loci. Curved line in mouse indicates a backsplice. **(E)** qPCR quantification of nuclear-cytoplasmic fractionated RNA from HEK293T. Error bars represent the standard deviation of biological replicates.

Supporting Information

S1 Fig. RNase R sensitivity of transcripts generated from BAC O vs HEK293T. LINC00632 isoform T3 was measured in both cases. Error bars represent the standard deviation of biological replicates.

S2 Fig. Transfection efficiency of the BACs relative to BAC O quantified by DNA-qPCR of BAC backbone DNA from HeLa cells transfected with the BAC. Error bars represent the standard deviation of biological replicates (with error propagated from BAC O).

S3 Fig. qPCR quantification of ciRS-7 and LINC00632 isoform T3 in HeLa transfected with BACs and HEK mutants. (A) RNA expression in HeLa transfected with BACs A, B, and C. All values have been normalized to those for BAC A, and error bars represent the standard deviation of biological replicates (with error propagated from BAC A). (B) RNA expression of isoforms in wild-type HEK293T and in a cloned strain of HEK293T in which the putative ciRS-7 promoters have been deleted.

S4 Fig. Additional PCRs to determine connectivity between exons of LINC00632 and ciRS-7.

S5 Fig. Sanger sequencing traces for novel linear ciRS-7 junctions in human.

S6 Fig. RNase R sensitivity of ciRS-7, LINC00632, and LINC00632-CDR1AS isoforms in HEK293T.

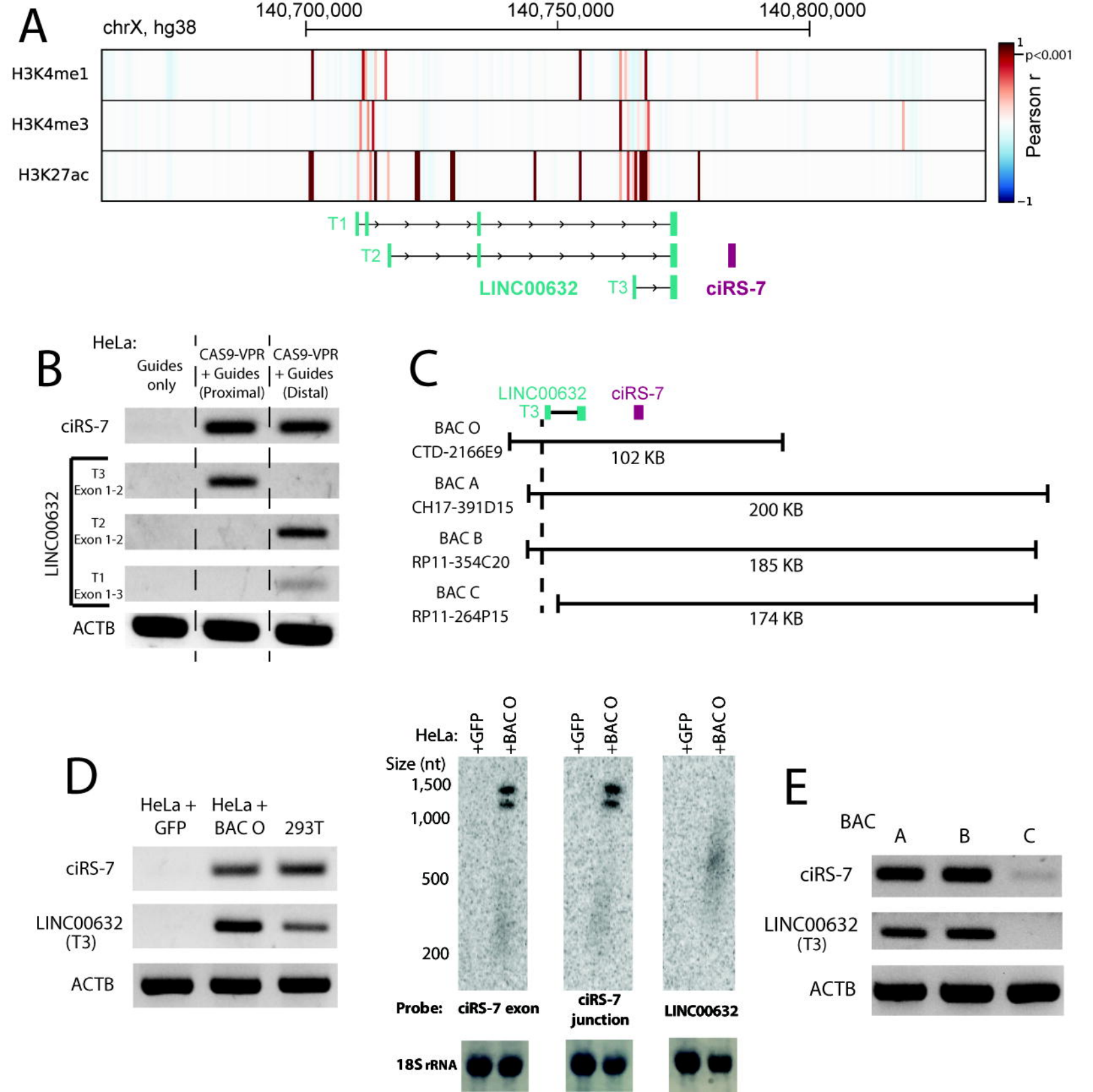
S7 Fig. Products from TOPO cloning of bands in Fig 2B--*left* (lanes 2 and 3).

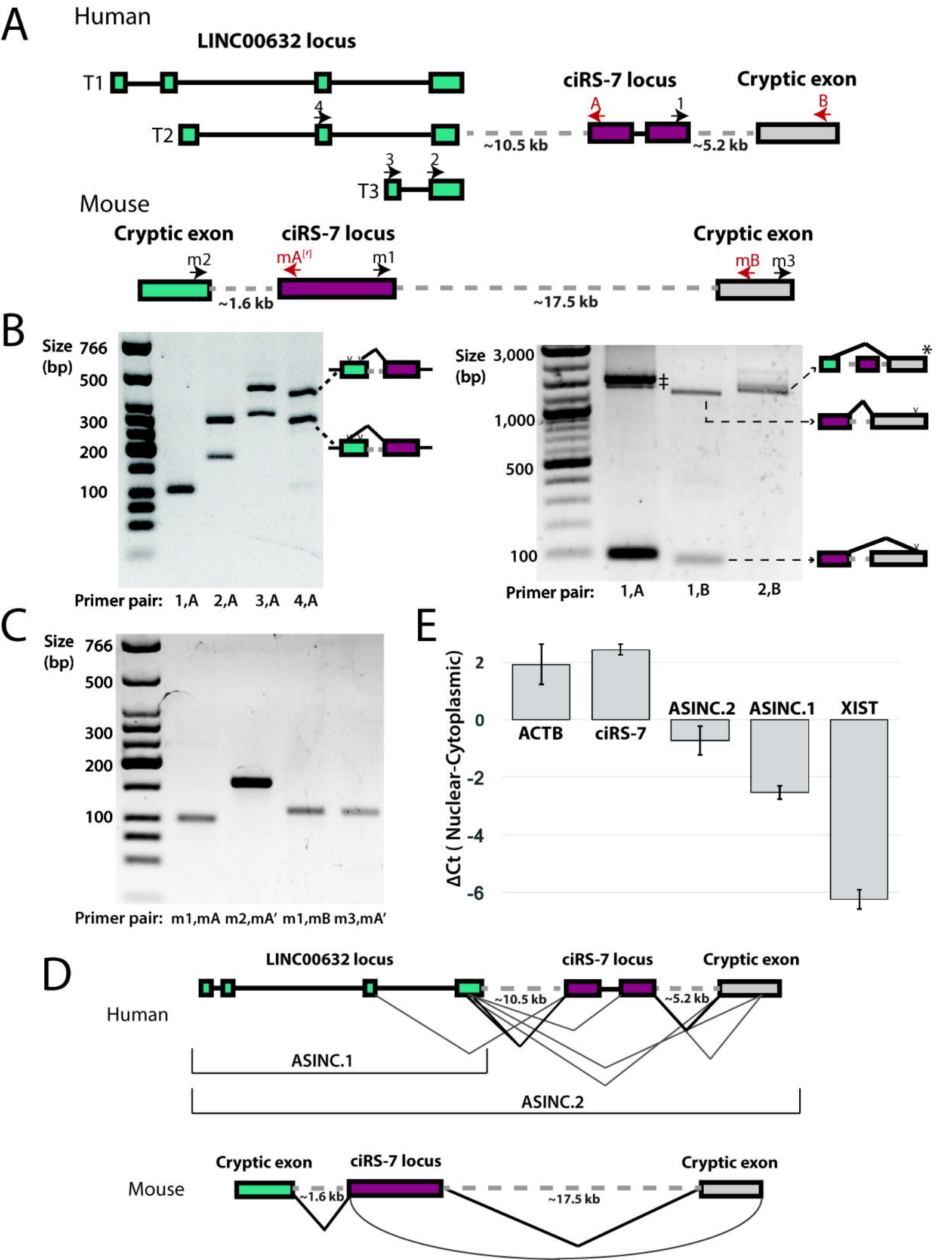
S8 Fig. qPCR ΔCt of human isoforms. Higher values indicate lower expression. Error bars represent standard deviation of biological replicates.

S9 Fig. qPCR ΔCt of mouse isoforms *(Left*). RNase R sensitivity of transcripts in mouse (*Right*). Error bars represent standard deviation of technical replicates.

S10 Fig. BAC quality checks. (A) Simulated and experimental BanI digest of the four BACs used in this study. The agreement of these footprints supports that BAC inserts are as reported and have not been significantly altered by bacterial recombination. (B) Sanger sequencing of the 5' ends of BAC genomic inserts. The vector sequence is in lowercase; the genomic insert sequence is in uppercase.

S1 File. PCR primers, Northern probes, and sgRNA sequences.
S2 File. RNA-seq scripts and output for novel isoform discovery.
S3 File. Table of ChIP-seq peak enrichment and RNA expression levels.
S4 File. Quantification of RNA expression in ENCODE data.
S5 File. Calculation of empirical FDR for ChIP-seq and RNA-seq correlations.


A
chrX, hg38
140,700,000
140,750,000
140,800,000
H3K4me1
H3K4me3
H3K27ac
1
p<0.001
Pearson r
−1
T1
T2
LINC00632
T3
ciRS-7
B
HeLa:
Guides only
CAS9-VPR + Guides (Proximal)
CAS9-VPR + Guides (Distal)
ciRS-7
LINC00632
T3 Exon 1-2
T2 Exon 1-2
T1 Exon 1-3
ACTB
C
LINC00632
ciRS-7
T3
BAC O
CTD-2166E9
102 KB
BAC A
CH17-391D15
200 KB
BAC B
RP11-354C20
185 KB
BAC C
RP11-264P15
174 KB
D
HeLa + GFP
HeLa + BAC O
293T
ciRS-7
LINC00632 (T3)
ACTB
HeLa:
+GFP
+BAC O
Size (nt)
1,500
1,000
500
200
Probe:
ciRS-7 exon
ciRS-7 junction
LINC00632
18S rRNA
E
BAC
A
B
C
ciRS-7
LINC00632 (T3)
ACTB


A
Human
LINC00632 locus
T1
T2
T3
4
3
2
ciRS-7 locus
A
1
Cryptic exon
B
~10.5 kb
~5.2 kb
Mouse
Cryptic exon
ciRS-7 locus
Cryptic exon
m2
mA[r]
m1
mB
m3
~1.6 kb
~17.5 kb
B
Size (bp)
766
500
300
200
100
Primer pair: 1,A 2,A 3,A 4,A
Size (bp)
3,000
1,000
500
100
Primer pair: 1,A 1,B 2,B
C
Size (bp)
766
500
300
200
100
Primer pair: m1,mA m2,mA' m1,mB m3,mA'
E
ΔCt (Nuclear-Cytoplasmic)
2
0
-2
-4
-6
ACTB
ciRS-7
ASINC.2
ASINC.1
XIST
D
Human
LINC00632 locus
ciRS-7 locus
Cryptic exon
~10.5 kb
~5.2 kb
ASINC.1
ASINC.2
Mouse
Cryptic exon
ciRS-7 locus
Cryptic exon
~1.6 kb
~17.5 kb