Complementary Sequence-Mediated Exon Circularization

Xiao-Ou Zhang,1,4 Hai-Bin Wang,2,3,4 Yang Zhang,2 Xuhua Lu,3 Ling-Ling Chen,2,* and Li Yang1,*
1Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
2State Key Laboratory of Molecular Biology, Shanghai Key Laboratory of Molecular Andrology, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
3Department of Orthopedic Surgery, Changzheng Hospital, Second Military Medical University, Shanghai 200003, China
4Co-first authors
*Correspondence: linglingchen@sibcb.ac.cn (L.-L.C.), liyang@picb.ac.cn (L.Y.)
http://dx.doi.org/10.1016/j.cell.2014.09.001

SUMMARY

Exon circularization has been identified from many loci in mammals, but the detailed mechanism of its biogenesis has remained elusive. By using genome-wide approaches and circular RNA recapitulation, we demonstrate that exon circularization is dependent on flanking intronic complementary sequences. Such sequences and their distribution exhibit rapid evolutionary changes, showing that exon circularization is evolutionarily dynamic. Strikingly, exon circularization efficiency can be regulated by competition between RNA pairing across flanking introns or within individual introns. Importantly, alternative formation of inverted repeated Alu pairs and the competition between them can lead to alternative circularization, resulting in multiple circular RNA transcripts produced from a single gene. Collectively, exon circularization mediated by complementary sequences in human introns and the potential to generate alternative circularization products extend the complexity of mammalian posttranscriptional regulation.

INTRODUCTION

Covalently closed circular RNA molecules were originally found to naturally exist in plant viroids (Sanger et al., 1976) and hepatitis delta virus (Kos et al., 1986). Later, endogenous circular RNAs processed from pre-mRNAs were identified in both human Ets-1 (Cocquereille et al., 1992; Nigro et al., 1991) and mouse Sry (Capel et al., 1993) genes. The presence of inverted sequences in flanking introns was suggested to be crucial for mouse Sry circularization (Dubin et al., 1995), especially with longer exon circularization (Pasman et al., 1996). However, only a few cases of pre-mRNA processed circular RNAs and their expression levels were reported to be very low, suggesting that they were byproducts of splicing errors (Cocquereille et al., 1993) and thus likely lacking biological function.

Recently, circular RNAs from human INK4a/ARF (Burd et al., 2010) and CDR1 (Hansen et al., 2013; Hansen et al., 2011) loci were identified and suggested to affect human atherosclerosis risk or regulate mRNA expression, thus shedding new light on physiological roles of circular RNAs. With the advent of high-throughput sequencing from nonpolyadenylated RNA transcripts, thousands of circular RNAs from back-spliced exons were successfully identified in multiple human cell lines (Jeck et al., 2013; Memczak et al., 2013; Salzman et al., 2012, 2013), with suggested function as miRNA sponges (Hansen et al., 2013; Memczak et al., 2013). However, only a few of such circular RNAs contain multiple binding sites to trap one particular miRNA (Jeck and Sharpless, 2014; Guo et al., 2014).

Most circular RNAs from back-spliced exons are stable and cytoplasmic (Jeck et al., 2013; Memczak et al., 2013), possibly due to their resistance to the cellular linear RNA decay machineries. Nevertheless, circular RNA transcripts were generally expressed at low levels compared with linear RNAs (Salzman et al., 2012, 2013). Bioinformatic analyses revealed that back-spliced exons were generally flankied by longer introns, and the existence of Alu elements in flanking introns was computationally predicted to be highly associated with the formation of human circular RNAs (Jeck et al., 2013). However, direct experimental evidence and detailed mechanism(s) supporting this model were still lacking.

Here, we take advantage of nonpolyadenylated and RNase R-treated RNA-seq from H9 human embryonic stem cells (hESCs) with a newly developed pipeline to predict back-spliced junctions and systematically characterize circular RNAs. Importantly, we have recapitulated circular RNA formation from a unique expression vector and offer multiple lines of evidence to support the conclusion that circular RNA formation is dependent on flanking complementary sequences, including either repetitive or nonrepetitive elements. Strikingly, such sequences exhibit rapid evolutionary changes among mammals, showing that exon circularization is evolutionarily dynamic. Furthermore, we show that the exon circularization efficiency is regulated by the competition of RNA pairing by complementary sequences within individual introns or across flanking introns. Alternative formation of inverted repeated Alu pairs (IRAlus) and the competition between them lead to alternative circularization, resulting
Figure 1. A Customized Algorithm to Determine Exon Circularization and Alternative Circularization

(A) A computational pipeline for back-spliced junction read calling to accurately annotate circular RNAs. The TopHat-Fusion algorithm (Kim and Salzberg, 2011) was first applied to call junction reads (green thin bars, step 1). Due to sequence similarity at the ends of certain exons (blue bars), mapped junction reads were split (dash arc lines) at incorrect sites, leading to mapping to incorrect locations (red thin bar). Such mapping errors were adjusted with a customized algorithm (step 2). Predicted circular RNAs from back-spliced exons (blue bars) were summed with predicted junction reads (step 3) in Table S1. Exons that do not back splice are indicated as gray bars.

(B) RNase R treatment significantly enriches circular RNA identification. Top, 2,119 or 9,639 (Table S1) events were identified with at least one back-spliced junction read in poly(A)– RNA-seq (green) or poly(A)–/RNase R treated RNA-seq (blue), respectively. Among them, about 1,061 circular RNAs were (legend continued on next page)
in multiple circular RNA transcripts produced from a single gene.

**RESULTS**

**Computational Pipeline to Precisely Identify Junction Reads from Circularized Exons**

Exonic circular RNA is produced from a splice donor site at a downstream exon joining to a splice acceptor site at an upstream exon (back splice). Due to the rearranged exon ordering, specific algorithms are required to annotate these back-spliced exon events for circular RNA prediction. Here, we developed a combined strategy (Figure 1A, CIRCexplorer, Experimental Procedures) to identify junction reads from back-spliced exons. In brief, RNA-seq reads were multiply aligned to the human hg19 reference genome using the TopHat algorithm (Kim et al., 2013) for RNA-seq, and unmapped reads were then uniquely mapped to the human hg19 reference genome with TopHat-Fusion (Kim and Salzberg, 2011) (Figure 1A, step 1). These reads, unmapped with TopHat but mapped with TopHat-Fusion on the same chromosome in a noncolinear ordering (back-spliced ordering), were extracted as candidate back-spliced junction reads. These junction reads were further realigned against existing gene annotation to determine the precise positions of downstream donor and upstream acceptor splice sites, respectively. Due to the sequence similarity at the ends of certain exons, some RNA-seq reads were split at incorrect sites, resulting in mapping to incorrect genomic locations (Figure 1A, red thin bar), usually with a 1–2 nucleotide shift (Figure S1A available online). Such mapping errors were then adjusted with a customized algorithm according to RefSeq exon annotation (Figure 1A, step 2). Finally, exonic circular RNAs were annotated with the support of identified junction reads (Figure 1A, step 3).

We applied this computational strategy to annotate back-spliced junction reads in either poly(A)–/Ribo– (poly(A)–) or RNase R-treated poly(A)–/RNase R RNA-seq (Figure S1B) from H9 hESCs. This computational pipeline is highly efficient (Figure S1C), memory economy (Figure S1D), easily accessed, and user friendly due to the application of TopHat (Kim et al., 2013) and TopHat-Fusion (Kim and Salzberg, 2011). It allowed us to detect 2,119 and 9,639 (Figure 1B) exonic circular RNA candidates with at least one back-spliced junction read in poly(A)– (GEO: GSE24399 and GSE60467) or poly(A)+/RNase R (GEO: GSE48003) RNA-seq, respectively (Table S1) for a full list of circular RNAs with at least one junction read). Among them, 1,061 circular RNAs were identified from both samples. On average, their junction reads were about 10-fold enriched after RNase R treatment (Figure 1B, bottom), consistent with previous reports that RNase R treatment is able to enrich circular RNAs from both exonic (Jeck et al., 2013; Salzman et al., 2013; Salzman et al., 2012) or intronic (Zhang et al., 2013) sequences.

To characterize intrinsic features of exon circularization, we next focused in great detail on highly expressed circular RNAs with at least five junction reads from RNase R-treated or untreated poly(A)– RNA-seq samples. Among 1,662 such circular RNAs (Table S2), about 88% of them could also be identified from at least one other cell line (Jeck et al., 2013; Salzman et al., 2013; Memczak et al., 2013), and 12% were predicted only in H9 cells (Figure 1C). Importantly, many of these circular RNAs were confirmed to be processed from back-spliced exons by divergent PCR (Figure S1E and data not shown). In addition, randomly selected circular RNAs (Figures 1D and 1E, left) were further detected by northern blots (NB) on either denaturing PAGE or native agarose gels (Figures 1D and 1E, right). While circular RNAs were resolved on native agarose gels with the same predicted sizes as their linear isoforms, they migrated much more slowly in denaturing PAGE gels, as expected. These RNAs remained stable after RNase R treatment (Suzuki et al., 2006) in both NB (Figures 1D and 1E) and divergent PCR (Figure S1C), further confirming them to be circular.

**Alternative Circularization**

We identified a number of multiple exon circularization events that were produced from single gene loci. For instance, one region of the human CAMSAP1 gene produced at least seven distinct exonic circular RNAs from H9 cells, as indicated in Figure 1F, although the majority of these circular RNA isoforms were expressed at low levels. We named this phenomenon alternative circularization (AC). Strikingly, all these AC events and CAMSAP1 circular RNAs from H9 cells were also predicted from poly(A)–/Ribo– RNA-seq data sets from other ENCODE cell lines (data not shown). In addition, the alternative circularization that leads to multiple circular RNA formation with different expression was observed in many gene loci with different expression levels (Table S1). Together, the existence of alternative circularization determined before and after RNase R treatment. (Bottom) Box plots of relative expression of predicted circular RNAs from different groups. RNase R treatment significantly enriched circular RNA identification by about 10-fold (*p < 2.2 × 10⁻¹⁶, Wilcoxon rank-sum test), RPM, reads per million mapped reads. (C) High-confidence circular RNAs from H9 cells (1,662, Table S2) were determined with at least five junction reads from either poly(A)–/RNase R or poly(A)– RNA-seq sample. About 88% (1,461 out of 1,662) of these high-confidence circular RNAs were also identified from six ENCODE cell lines (Salzman et al., 2013). About 12% (201 out of 1,662) were only identified in H9 cells using these accessible data sets. (D and E) Visualization and validation of two circular RNAs from CAMSAP1 (D) and CRKL (E) loci. (Left) Visualization of predicted circular RNAs. Deep-sequencing signals from poly(A)–/RNase R (purple), poly(A)– (red) or poly(A)+ (black) RNAs were shown. Predicted circular RNAs were indicated with mapped junction reads from either poly(A)– or poly(A)+/RNase R RNA-seq. Northern blot (NB) probes for the circularized exons were highlighted as blue bars. Black arrow indicates the direction of transcription. (Right) RNase R-treated or untreated H9 total RNAs were loaded on a 5% denaturing PAGE gel or on a 1% native agarose gel. NB was performed with probes indicated in left panels. NB bands for circular RNAs are indicated with schematic circles (in blue). Circular RNAs remained stable after RNase R treatment, whereas linear RNAs (asterisk) were removed. Note that a longer CAMSAP1 circular RNA with an intervening intron between circularized exons was also detected by NB; however, it was unstable by applied RNase R treatment with an unknown mechanism. (F) A schematic diagram to show alternative circularization in one region of human CAMSAP1 locus. Circular RNAs with their back-spliced junction reads (numbers) from H9 poly(A)–/RNase R RNA-seq were indicated by arc lines. Different shades of black colors indicated different number of junction reads. See also Figure S1 and Tables S1 and S2.
circularization suggests yet another layer of gene expression regulation of circular RNA formation.

**Circularized Exons Are Preferentially Flanked by Orientation-Opposite Alu Elements to Form IRAlus Pairs**

Next, we investigated genomic features of highly expressed circular RNAs from Figure 1C. First, the vast majority of highly expressed exonic circular RNAs identified were processed from exons located in the middle of RefSeq genes, with only a few, including the first or the last exons (Figure 2A), suggesting that the circular RNA formation is generally coupled to RNA splicing.

In addition, although 168 out of 1,662 highly expressed circular RNAs contained only one annotated exon, most circular RNAs contain multiple exons, most commonly two to three exons (Figure 2A), suggesting that the average Alu density in such long flanking introns is comparable to that in other control introns (Figure 2D). This observation thus suggests that longer flanking introns are not necessary for circular RNA formation, but the extended length could introduce more Alu elements that, in turn, promote exon circularization.

On average, there were about three Alu elements in both upstream and downstream flanking introns of circularized exons (Figures 2E and 2F). Interestingly, Alu elements that could form IRAlus, either convergent (Figure 2G) or divergent (Figure 2H), are juxtaposed across flanking introns in a parallel way with similar distances to adjacent exons. In contrast, randomly selected long control intron pairs (at least 10,000 nt in length with no circularized exons identified in H9 cells) have comparable numbers of Alus (Figure S2C); however, the juxtaposition of convergent and divergent IRAlus across them is negligibly correlated (Figure S2D). This thus suggests that IRAlus pairing might bring a splice donor site at a downstream exon and a splice acceptor site at an upstream exon close to each other and promote back splicing. Finally, we also observed that other nonrepetitive complementary sequences were involved in exon circularization, as discussed below.

**Circular RNA Formation Is Promoted by IRAlus or by Other Nonrepetitive but Complementary Sequences**

Although computational analyses have led to the speculation that IRAlus could promote exon circularization, experimental evidence supporting this model has been lacking. As we have computed that the vast majority of highly expressed circularized exon(s) were found in middle exons of annotated genes (Figure 2A), we cloned the circularized exons along with their full-length flanking introns into the middle of the split egfp gene in pZW1 (Wang et al., 2004). The application of such vectors allowed us to visualize the processed back-spliced exons by NB and linear RNA isoforms by semiquantitative RT-PCR. We selected a representative circular RNA for detailed characterization. This circular RNA contains two exons residing in the human POLR2A locus, flanked by relatively short introns with one reverse Alu in the upstream intron and two forward Alus in the downstream intron (Figures 3A and 3B). Note that there are two potential pairs of IRAlus formed across flanking introns (red arcs, Figure 3B) with similar predicted free energies (Figure 3B). The endogenous POLR2A circular RNA was validated by divergent PCR (Figure S1C) and NB on denaturing PAGE gels in both H9 and HeLa lines (Figure S3A).

For recapitulation, wild-type (Figure 3C, #1) or a series of deletions (Figure 3C, #2–#7) constructs for POLR2A circular RNA were individually cloned into pZW1 (Wang et al., 2004). NB validated that recapitulated POLR2A circular RNA, with the same migration as endogenous one (Figure S3A), could be produced in clones when flanking IRAlus were formed, including wild-type (Figure 3C, #1 and Alu deletions that retained a pair of IRAlus formed across flanking introns (Figure 3C, #3 and #4). In contrast, POLR2A circular RNA could not be detected in clones in which deletions eliminated IRAlus pairing across flanking introns (Figure 3C, #2 and #5–#7). It should be noted that the expression of linearized hybrid egfp mRNA (with POLR2A exon included) was much higher than that of linearized egfp mRNA (with POLR2A exon skipped), as revealed by RT-PCR (Figure 3C), suggesting that exon circularization efficiency is low.

Interestingly, no mouse POLR2A circular RNA was identified from RNA-seq (Figure S3B) or divergent PCRs (data not shown).

---

**Figure 2. Genomic Feature Analyses of Exon Circularization Revealed Juxtaposition of Orientation-Opposite Alu in Flanking Introns**

(A) Genomic distribution of back-spliced exons. Most back-spliced exons are located in the middle of genes, and very few of them are the first or the last exons.

(B) Number and length distribution of back-spliced exons. (Left) Most (about 90%) of circular RNAs contain multiple back-spliced exons. (Right) Box plots indicate the distribution of exon length (y axis) from circular RNAs containing different numbers of back-spliced exons (x axis). *p < 2.2 x 10^-16, Wilcoxon rank-sum test.

(C) Length distribution of flanking introns. Both upstream (blue line) and downstream (yellow line) flanking introns are much longer than control introns (gray line). p < 2.2 x 10^-16, Wilcoxon rank-sum test.

(D) The density of Alu elements in flanking introns. The average Alu density in flanking introns (upstream in blue line and downstream in yellow line) is similar to that in randomly selected control introns (gray line).

(E and F) On average, about three Alu elements, either convergent (E) or divergent (F), were identified in flanking introns.

(G and H) Juxtaposition of orientation-opposite Alu elements in flanking introns. Convergent (G) or divergent (H) Alu elements in flanking introns (top) were divided according to their relative positions as first (25%), median (50%) or third quantile (75%). Their relative positions (distances to splice sites) were clustered respectively to create isodenses (curves on which points have the same density) to estimate the overall distribution of Alus in both upstream and downstream flanking introns. Note that IRAlus juxtapose in a parallel manner in introns flanking circular RNA exons with positive correlation (Spearman correlation coefficient > 0.5) (Mukaka, 2012).

See also Figure S2.
in R1 mouse embryonic stem cells (mESCs). Correspondingly, there are no inverted repeats in the related flanking introns in mouse, providing an explanation for the lack of POLR2A circular RNA in mouse. This result indicates the important correlation of inverted repeated complementary sequences (IRAlus) in human with the expression of circular RNAs. Interestingly, human POLR2A circular RNA could be also recapitulated when transfected into NIH 3T3 cells in a way that complementary sequences still could promote circular RNA formation (Figure S3C). These results suggest that the machinery for circular RNA formation is evolutionarily conserved and that the different organization of inverted repeated complementary sequences between species could lead to different expression levels of circular RNAs.

In addition, POLR2A circular RNA could be recapitulated when only a half Alu element still remained (Figure 3D, #2 and #3), suggesting that even partially complementary Alus are sufficient to promote RNA pairing (Figure S4A) and exon circularization. Moreover, POLR2A circular RNA could also be detected in the reconstituted clone (Figure 3D, #4) in which a pair of IRAIs from the 3′ UTR of the human NICTN1 gene (Chen et al., 2008) (Figure S4B) was inserted into flanking introns to replace the wild-type POLR2A IRAlus. Although circular RNAs were expressed at low levels in certain deletions (Figures 3C, #3 and 3D, #3) and replacement (Figure 3D, #4), these results suggest that the formation of this circular RNA is not dependent on specific Alu species.

We next hypothesized that any complementary pairing across circularized exons should be able to induce exon circularization. To examine this hypothesis, we further inserted a nonrepetitive fragment from another POLR2A intron into both upstream and downstream flanking introns in an orientation-opposite pattern to obtain a new expression vector (Figure 3D, #5). Such insertion allowed the formation of a strong and completely complementary pairing across circularized exons (Figure S4C). Strikingly, POLR2A circular RNA expression was much higher in this clone (Figure 3D, #5) than all of the others that we have examined. Correspondingly, in this case, the production of linearized egfp mRNA (with POLR2A exon skipping for circularization) was comparable with the linearized hybrid egfp mRNA (with POLR2A exon inclusion), as revealed by RT-PCR (Figure 3D, #5). When this completely complementary structure was disrupted, circular RNA was barely detectable (Figure 3D, #6). Taken together, our results demonstrate that the circular RNA formation is dependent on the pairing capacity of complementary sequences, whether or not they are from repetitive elements.

**Naturally Existing Nonrepetitive Complementary Sequences in Flanking Introns Promote Exon Circularization and Undergo Rapid Evolutionary Changes**

As we have shown that artificial nonrepetitive complementary sequences can promote circular RNA formation (Figure 3D), it became of interest to identify whether naturally existing nonrepetitive complementary sequences play a role in circular RNA processing. As indicated in Figure 3E, a pair of nonrepetitive complementary sequences (about 180 nt) that naturally exist in the human GCN1L1 locus with naturally existing nonrepetitive complementary sequences. (Left) A schematic diagram of the human GCN1L1 locus for one predicted circular RNA with three back-spliced exons (blue bars). Repetitive elements are indicated with colored arrows in human or mouse. Nonrepetitive complementary sequences were identified in the human, but not in the mouse genome, and are highlighted with black lines. (Right) Junction reads for GCN1L1 circular RNA were only identified from human H9 cells, but not from mouse R1 cells. The existence of GCN1L1 circular RNA was confirmed by RT-PCR with divergent primers (blue arrowheads) in human H9, but not in mouse R1 cells.

(F) The naturally existing nonrepetitive complementary sequences undergo rapid evolutionary changes across species. Sequence conservation analysis of GCN1L1 circular RNA region in different species, visualized with the VISTA browser. (Top) RefSeq exons, including back-spliced (blue bars) and non-backspliced (gray bars) exons, are indicated. A nonrepetitive complementary sequence (black arrow) was individually identified in both upstream and downstream flanking introns, forming a potential RNA pair highlighted with a black arc line. (Bottom) Rapid evolutionary changes of complementary sequences among primates. Y axis, species selected for comparison (left) and levels of conservation based on genomic sequence similarity (right). Colors of conserved regions are labeled by VISTA according to RefSeq annotations (exons in blue shadow and introns in red shadow).

In (C) and (D), equivalent amounts of RNA from cells transfected with each indicated plasmid were resolved on a 5% denaturing PAGE gel for NB to detect circular RNAs. Semiquantitative RT-PCRs further confirmed two isoforms of linearized RNAs. Three biological replicates have been done, and only one set of representative results was shown. The relative abundance of circular RNA and linear RNA with skipped exons was determined by using Quantity One (Bio-Rad). NT, not transfected; EV, empty vector. See also Figures S3, S4, and S6.
Please cite this article in press as: Zhang et al., Complementary Sequence-Mediated Exon Circularization, Cell (2014), http://dx.doi.org/10.1016/j.cell.2014.09.001
the human GCN1L1 locus was identified with only 3 nt difference (Figures 3E and S5A). As expected, the formation of GCN1L1 circular RNA was detected by both RNA-seq and RT-PCR in H9 cells (Figure 3E, right).

Intriguingly, these nonrepetitive complementary sequences were largely missing in the mouse GCN1L1 locus, and circular RNA was undetectable by either RNA-seq or RT-PCR in R1 mESCs (Figure 3E, right). In fact, the endogenous complementary sequences in flanking introns are not only nonconserved between human and mouse, but also undergo rapid changes even among primates (Figures 3F and S5B). These observations suggest that exon circularization may undergo rapid evolutionary changes, shown by GCN1L1 locus reported here as well as some other gene loci (data not shown).

**The Regulation of Circular RNA Formation: Competition of IRAUs Formation across or within Individual Flanking Introns**

We have offered new lines of evidence by both computational and experimental analyses to confirm that complementary sequences (mainly IRAUs in the human context) promote circular RNA formation (Figures 2 and 3). However, the existence of IRAUs does not always lead to exon circularization. As indicated in Figure 4A, although multiple Alu and IRAUs across introns could be identified in the human ZWILCH locus (left), circular RNA from this locus was undetectable from RNA-seq (middle) or divergent PCR (right) in H9 hESCs. Interestingly, circular RNA in the mouse ZWILCH locus was clearly identified from RNA-seq (Figure 4A, middle) and divergent PCR (Figure 4A, right) in R1 mESCs. Thus, what is the mechanism leading to the different expression of ZWILCH circular RNA in mouse and human?

Three potential IRAUs within the downstream intron (IRAUs\textsubscript{within}) and another three potential IRAUs across flanking introns (IRAUs\textsubscript{across}) were predicted in the human ZWILCH locus (Figure 4B, top). In contrast, the mouse ZWILCH region contains three potential inverted complementary SINEs (two pairs of B1 elements and one pair of B2 element) within the downstream intron but six potential inverted repeats (B1 elements) across flanking introns (Figure 4B, bottom). Theoretically, the competition of RNA pairing by complementary sequences across or within individual flanking introns could significantly affect splicing selection and thus exon circularization. With the strong competition for IRAUs\textsubscript{within} formation, for instance, in the human ZWILCH region (Figure 4B, top), splicing occurs at two adjacent exons as a normal constitutive splicing event, which results in a linearized RNA transcript with exon inclusion (Figure S6A), but not exon circularization (Figure 4C, left). In contrast, the predominant formation of IRAUs\textsubscript{across}, such as in the mouse ZWILCH region (Figure 4B, bottom), could bring two exons together for back splicing, leading to a linearized RNA transcript with exon skipping (Figure S6B). Such skipped exons undergo circularization to process a circular RNA transcript (Figure 4C, right).

To test this model, we took advantage of the expression construct that expresses POLR2A circular RNA at a very high level (#5, Figure 3D) and inserted an additional copy of complementary sequence to introduce potential RNA pairing within the downstream flanking intron (Figure 4D). As expected, because of the competition of RNA pairing within an individual intron (Figure 4D, left, arc lines), splicing was prone to occur at two adjacent exons as a normal constitutive splicing event, with significantly reduced exon skipping and circularization (#2 and #3, Figure 4D). The reduction of exon skipping was further monitored by RT-PCR (Figure 4D, lower-right). In addition, we found that the expression of recapitulated POLR2A circular RNA is positively correlated with that of linear egfp mRNA with POLR2A exon skipping (Figure 4E). However, to what extent circularization correlates with exon-skipped splicing in endogenous...
conditions remains to be defined. Taken together, our results strongly indicate that exon circularization can be altered by the competition of RNA pairing within individual introns or across flanking introns. This alteration can be species specific due to the different distribution of complementary sequences across species (for instance, in ZWILCH locus, Figure 4).

The Generality of Exon Circularization with the Competition of IR\textit{Alu}s

We suggest a model that the competition between IR\textit{Alu}s\textsubscript{within} and IR\textit{Alu}s\textsubscript{across} alters exon circularization (Figure 4C). To identify the generality of circular RNA formation with the competition of IR\textit{Alu}s\textsubscript{within} and IR\textit{Alu}s\textsubscript{across} (Figure 5A), we performed genome-wide comparison of the complementary \textit{Alu} element distribution between the highly expressed circular RNA flanking introns (Table S2) and randomly selected control introns (Figure 5B). Strikingly, the percentage of circular RNA flanking introns with across-intron IR\textit{Alu}s is 3.49-fold higher than that of control introns (compared 92.8% with 26.6%). The percentage of circular RNA flanking introns with (IR\textit{Alu}s\textsubscript{across} – IR\textit{Alu}s\textsubscript{within})\textsuperscript{R1} is 4.74-fold higher than that of control introns (compared 46.5% with 9.8%).

Together, these genome-wide analyses suggest that the circular RNA formation is generally associated with the competition of IR\textit{Alu}s, but not just IR\textit{Alu}s themselves. The competition of RNA pairing can be very complicated in endogenous conditions. In addition to the number of potential IR\textit{Alu}s formed across or within flanking introns, the distance between each potential pair of \textit{Alu}s may also affect the pairing capacity. It has been reported that the capacity to form complementary pairs is dramatically decreased when it exceeds several kilobases (Athanasiadis et al., 2004). We found that the distance between the nearest pair of convergent (D) and divergent (E) \textit{Alu}s is much longer in long intron pairs than that in circular RNA flanking introns, \( \text{**p} < 2.2 \times 10^{-16} \), Wilcoxon rank-sum test.

Frequent Alternative Circularization by Widely Distributed \textit{Alu}s in Human Introns

From our computational predictions, we have observed multiple exon circulations from one gene locus (Figure 1E). In addition to human CAMSAP1 circular RNAs (Figure 1E), multiple DNMT3B (Figure 6A) and XPO1 (Figure 6B) circular RNAs were also predicted in H9 cells. Importantly, some highly expressed, alternatively circularized DNMT3B and XPO1 RNAs could be separated from NB on denaturing PAGE gels (Figures 6A and 6B), further confirming the existence of alternative circularization as predicted from our computational approach. There are multiple juxtaposed \textit{Alu} elements in DNMT3B introns (Figure 6A) and
XPO1 introns (Figure 6B). Computational analyses revealed that these Alu elements could form a variety of IRAIs across orientation-opposite patterns (data not shown), which flank alternatively circularized exons (black/gray arcs), suggesting a role of Alu pairing and the competition between them in the formation of alternative circularization. We further confirmed that alternative circularization could be identified from >50% gene loci that produce circular RNAs in examined cell lines (Figures 6C and 6D and Table S1), suggesting the generality of alternative circularization. Frequent alternative circularization might be associated with widely distributed Alus in human introns (Figure 6E). This finding further suggests a previously underappreciated role of intronic Alu sequences in the formation of alternative circularization.

**DISCUSSION**

Due to intrinsic structure features such as lack of 3’ poly(A) tails and no 5’ end caps, genome-wide identification of circular RNAs has been missed in transcriptome profiling from polyadenylated RNAs. Deep sequencing from fractionated nonpolyadenylated RNA populations indicated accumulated signals from certain exons (called excised exons) (Yang et al., 2011), which were further confirmed as circular RNAs (Salzman et al., 2012). These observations were rapidly followed by others who reported numbers of circular RNA transcripts from back-spliced exons from multiple human cell lines after nonpolyadenylated RNA and/or circular RNA enrichment (Jeck et al., 2013; Memczak et al., 2013; Salzman et al., 2012, 2013). However, the mechanism for circular RNA biogenesis and its possible regulation have remained unclear, despite a noted association with Alu elements (Jeck et al., 2013).

Here, we developed a computational pipeline to precisely identify back-spliced junction reads for circular RNAs (Figure 1). We characterized circular RNA formation by both bioinformatic (Figure 2) and biochemical (Figure 3) lines of evidence, demonstrating that flanking complementary sequences, including both repetitive and nonrepetitive sequences, play important roles in exon circularization. Interestingly, recapitulation with a pair of perfectly matched complementary sequences led to efficient expression of circular RNAs (Figure 3D), suggesting that the pairing capacity across the circularized exons regulates the productivity of back-spliced circular RNA formation. Strikingly, our data also revealed that naturally existing nonrepetitive complementary sequences are highly associated with exon circularization and undergo rapid evolutionary changes (Figures 3E and 3F). It thus will be of great interest to compare genome-wide the formation of circular RNAs and relevant complementary sequences among different species and to further investigate the functional implications of these molecules during evolution. However, we could not exclude the possibility that lack of conservation of some circular RNAs indicates lack of function.

The existence of complementary sequences (either repetitive or nonrepetitive) is necessary, but not sufficient, for exon circularization. The competition of IRAIs formation within individual introns or across flanking introns significantly affects splicing selection and exon circularization (Figure 4), leading to the hypothesis that exon circularization efficiency is altered by the competition of RNA pairing by complementary sequences within individual introns or across flanking introns (Figure 4). It should be noted that the competition of RNA pairing by different distributions of repetitive elements in introns, as well as the circular RNA formation, is evolutionarily dynamic (Figure 4B).

We found that multiple circularized exons could be produced from a single gene locus (Figures 1E, 6A, and 6B), and this phenomenon has also been indicated in some other genes (Burd et al., 2010; Jeck et al., 2013; Salzman et al., 2012). This process is referred to as alternative circularization here to indicate multiple exon circularization selection from a single gene. Because IRAIs forming across flanking introns are strongly associated with exon circularization, we reasoned that the widely distributed Alu elements in human introns (Figure 6E) could allow multiple RNA pairing and competitive IRAIs formation to introduce alternative circularization. Accordingly, alternative circularization events were identified in thousands of genomic loci in different cell lines (Figures 6C and 6D). Our findings thus indicate that widespread Alu elements in introns and the competition of IRAIs could be actively involved in the selection of alternative circularization in human.

Alternative circularization generates a variety of circular RNAs. In addition to exons only (described here and Jeck et al., 2013; Memczak et al., 2013; Salzman et al., 2012, 2013) or introns only (ciRNAs, Zhang et al., 2013), we found that circular RNAs could be produced from both exons and introns (Figure 1D and Salzman et al., 2012) or from previously unannotated exons (exon x, Figure 6B). Thus, the endogenous conditions for circular RNA formation are very complex. For instance, both the number and the distance between complementary sequences across circularized exons can affect the pairing capability and their competition (Figure 5). Two other mechanisms have also been proposed to indicate circular RNA formation from back-spliced exons, named “direct backsplicing” and “lariat intermediate” (Salzman et al., 2012; Jeck et al., 2013; Jeck and Sharpless, 2014). However, what exactly happens within the circularization-related spliceosomes is largely unknown. In addition, the coordination of RNA transcription and exon circularization events also remains elusive. The observation that some circular RNAs are formed by long-range pairing (Figure 6B) suggests that splicing/circularization could occur after transcription despite the cotranscriptional assembly of the spliceosome (Black, 2003). Nevertheless, besides complementary sequences and their competition, it would be interesting to identify other cis elements and trans factors that are involved in alternative circularization, as has been done in the study of alternative splicing (Fairbrother et al., 2002; Lovci et al., 2013; Nilsen and Graveley, 2010).

The identification of alternative circularization further expands our understanding of gene expression regulation. Through alternative splicing, multiple functional mRNAs (and proteins) could be produced from a single gene (Nilsen and Graveley, 2010). These multiple functional mRNAs are generally thought to exist only as linearized molecules. Our work shows that alternative circularization coupled with alternative splicing can produce a variety of additional circular RNAs from one gene. Taken together, these lines of evidence imply a new level of complexity in transcriptomes and their regulation.
Figure 6. The Diversity of Inverted Repeated Alu Pairing and the Competition among Them Lead to Alternative Circularization

(A and B) Alternative circularization leads to a variety of circular RNA transcripts in human DNMT3B (A) and XPO1 (B) loci. (Top) Schematic diagrams to show alternative circularization in one region of examined human loci. Alu elements are indicated by red arrows in introns. Circular RNAs with their back-spliced junction reads (numbers) from H9 poly(A)–RNase R RNA-seq are indicated by arc lines. Different shades of black colors indicated different numbers of junction reads. Note that orientation-opposite Alu elements could form potential RNA pairs that flank all of the predicted circular RNAs in DNMT3B (A) or XPO1 (B). (Lower-left)
EXPERIMENTAL PROCEDURES

Additional details can be found in the Extended Experimental Procedures.

Computational Pipeline to Precisely Predict Back-Spliced Junctions for Circular RNAs

We developed a new computational pipeline (CIRCexplorer) by combining TopHat (Kim et al., 2013) and TopHat-Fusion (Kim and Salzberg, 2011) to obtain back-spliced junction reads for circular RNA prediction. See the Extended Experimental Procedures and Figure 1 for details.

Characterization of Back-Spliced Circular RNAs

Information of circularized exons and their flanking introns was extracted from existing gene annotations. Wilcoxon rank-sum test was used for length comparison between exons or introns. Information of repetitive elements was extracted from UCSC RepeatMasker database (rmsk.txt updated at 2009/4/27). Spearman’s rank correlation coefficient (Mukaka, 2012) was employed to evaluate the correlation of the juxtaposition of Alu elements for divergent or convergent IR-ALUs. See the Extended Experimental Procedures for details.

Plasmid Construction

For recapitulation of circular RNA, genomic region for POLR2A circular RNA with its wild-type flanking introns was amplified and cloned into Nhel/Mulu-digested pZW1 vector (Zhang et al., 2013). A series of deletions or insertions was further obtained. See the Extended Experimental Procedures for details.

Total RNA Isolation, Polyadenylated/Nonpolyadenylated RNA Separation, RNase R Treatment and RNA-Seq

Total RNA preparation and polyadenylated and nonpolyadenylated RNA separations were carried out as described previously (Yang et al., 2011). RNase R treatment was carried out at 37°C for 3 hr as described previously (Zhang et al., 2013). RNA-seq libraries were prepared by using Illumina TruSeq RNA Sample Prep Kit V2 and were subjected to deep sequencing with Illumina HiSeq 2000 at CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China. See the Extended Experimental Procedures for details.

ACCESSION NUMBERS

Additional raw sequencing data sets and bigwig track files of poly(A)+ and poly(A)−/Ribo− RNA-seq from hESCs H9 and mESCs R1 are available for download from NCBI Gene Expression Omnibus under accession number GSE60467.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.09.001.

AUTHOR CONTRIBUTIONS

L.Y. and L.-L.C. conceived, designed, and supervised the project. X.-O.Z. and L.-L.C. performed the bioinformatics analysis. H.-B.W., Y.Z., and X.L. performed experiments. L.Y. and L.-L.C. analyzed the data and wrote the paper with input from authors.

ACKNOWLEDGMENTS

We are grateful to Gordon Carmichael for critical reading of the manuscript and Qing-Fei Yin and Hua-Hong Fang for technical support. H9 cells were obtained from the WiCell Research Institute. RNA-seq was performed at CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China. This work was supported by grants XDA01010206 and 2012041TP8 from CAS; 2014CB964800, 2014CB910600, and 2011CB910105 from MOST; 31322018 and 31271376 from NSFC; and 2012SSSTP01 from SiBS.

Received: July 1, 2014
Revised: August 15, 2014
Accepted: August 27, 2014
Published: September 18, 2014

REFERENCES


Ago2-mediated cleavage of a circular antisense RNA. EMBO J. 30, 4414–4422.


Supplemental Information

EXTENDED EXPERIMENTAL PROCEDURES

Computational Pipeline to Precisely Predict Back-Spliced Junctions for Circular RNAs
To comprehensively map back-spliced junction reads with high confidence, two-step mapping strategy (CIRCexplorer) was exploited from hESCs H9 poly(A)/Ribo- RNA-seq (GEO:GSE24399 and GEO:GSE60467), hESCs H9 RNase R treated poly(A)/Ribo- RNA-seq (GEO:GSE48003), or mouse ESCs R1 poly(A)/Ribo- RNA-seq (GEO:GSE53942 and GEO:GSE60467) data sets, respectively. Sequence reads from each sample were first multiply mapped using TopHat (Kim et al., 2013) (TopHat 2.0.9 (parameters: -a 6 --microexon-search -m 2) against the GRCh37/hg19 human reference genome or NCBI37/mm9 mouse reference genome with the UCSC Genes annotation (Human: hg19 knownGene.txt updated at 2013/6/30, Mouse: mm9 knownGene.txt updated at 2011/5/30). Unmapped reads were then extracted and mapped onto the relevant reference genome using TopHat-Fusion (Kim and Salzberg, 2011)(TopHat 2.0.9, parameters: --fusion-search --keep-fasta-order --bowtie1 --no-coverage-search). Reads that split and aligned on the same chromosome but in non-colinear ordering (indicated with special XF tags in output BAM files) were extracted as candidate back-spliced junction reads. Back-spliced junction reads were further realigned against existing gene annotations (Human: hg19 knownGene.txt updated at 2013/6/30 and refFlat.txt updated at 2013/10/13, Mouse: mm9 knownGene.txt updated at 2011/5/30 and refFlat.txt updated at 2013/10/13) to determine the precise positions of donor or acceptor splice site for each back-spliced event. Junction reads with shifted alignments against canonical splice sites (with the consensus sequence flanking the donor splice site and the acceptor splice site) were adjusted to the correct positions with a customized script, and reads with alignments on different genes or non-canonical splice sites were largely from trans-splicing or PCR errors, thus were discarded. Finally, remaining back-spliced junction reads were combined and scaled to RPM ( Reads Per Million mapped reads, including TopHat mapping and TopHat-Fusion mapping) to quantify every back-spliced event. Wilcoxon rank-sum test was used for the comparison between different samples. Transcriptome de novo assembly was performed using Cufflinks v2.0.2 (Trapnell et al., 2013) as described previously (Zhang et al., 2014).

Comparison with Other Aligners for Back-Spliced Junction Prediction
Two available pipelines were chosen for comparison with ours. Segemehl (Hoffmann et al., 2014) and MapSplice (Jeck et al., 2013). Data sets from H9 poly(A)/Ribo– RNA-seq (GEO:GSE24399 and GEO:GSE60467) and H9 RNase R treated poly(A)/Ribo– RNA-seq (GEO:GSE48003), were chosen for testing with Segemehl 0.1.7-407 (parameters: -S -M 1) or MapSplice 2.1.3 (parameters: -k 1 --non-canonical --fusion-non-canonical), respectively. Predicted circular RNAs with defined junctions reads were further used for comparison.

Characterization of Back-Spliced Circular RNAs
Information of back-spliced exons and flanking introns of circular RNAs was extracted from existing gene annotations (knownGene.txt updated at 2013/6/30 and refFlat.txt updated at 2013/10/13). Information of repetitive elements was extracted from the UCSC RepeatMasker database (rmsk.txt updated at 2009/4/27). Two control sets were randomly selected from existing gene annotations. One control (control introns) consists of 5,000 randomly selected unique intron pairs, and another control (long control introns) contains 1,000 randomly selected unique long (>10,000nt) intron pairs with no circular RNA predicted from H9 gene annotations. One control (control introns) consists of 5,000 randomly selected unique intron pairs, and another control (long control introns) contains 1,000 randomly selected unique long (>10,000nt) intron pairs with no circular RNA predicted from H9 gene annotations. Two control sets were randomly selected from existing gene annotations (knownGene.txt updated at 2013/6/30 and refFlat.txt updated at 2013/10/13) to determine the precise positions of donor or acceptor splice site for each back-spliced event. Junction reads with shifted alignments against canonical splice sites (with the consensus sequence flanking the donor splice site and the acceptor splice site) were adjusted to the correct positions with a customized script, and reads with alignments on different genes or non-canonical splice sites were largely from trans-splicing or PCR errors, thus were discarded. Finally, remaining back-spliced junction reads were combined and scaled to RPM ( Reads Per Million mapped reads, including TopHat mapping and TopHat-Fusion mapping) to quantify every back-spliced event. Wilcoxon rank-sum test was used for the comparison between different samples. Transcriptome de novo assembly was performed using Cufflinks v2.0.2 (Trapnell et al., 2013) as described previously (Zhang et al., 2014).

Inverted Sequence Pairing Analysis
Sequences from flanking introns were extracted and inverted complementary sequences were selected to represent the distribution of inverted sequences by employing blastn (parameters: -word_size 11 -gapopen 5 -gapextend 2 -penalty -3 -reward 2). Secondary
structure prediction of inverted complementary sequences and their MRE (Minimum Free Energy) evaluation were preformed with RNAfold 2.1.5 (Lorenz et al., 2011).

Sequence Conservation Analysis

VISTA Browser was employed to inspect the conservation landscape for a given region from different genomes. Multiple sequence alignments for given species and given regions were extracted from multiple alignments of 100 vertebrate species (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/maf/, updated at 2014/2/7).

Cell Culture and Transfection

Human HeLa and mouse NIH 3T3 cell lines were cultured using standard protocols provided by ATCC. Human ES H9 and mouse ES R1 cell lines were maintained as described previously (Zhang et al., 2013). Stem cell cultures were regularly evaluated for Oct3/4 expression every 3-4 weeks and cells were passaged every 6-7 days. The pZW1 series constructs were individually transfected into human HeLa or mouse NIH 3T3 cell lines with either X-tremeGENE 9 (Roche) or Lipofectamine™ 2000 reagent (Invitrogen) according to the manufacturer's instructions. Total RNAs were extracted 24 hr after transfection.

Plasmid Construction

Genomic region for POLR2A circular RNA with its wild-type flanking introns was amplified from H9 genomic DNAs. PCR product was digested with NheI/MluI and cloned into NheI/MluI digested pZW1 vector (a gift from Zefeng Wang, UNC School of Medicine) by using the method described previously (Zhang et al., 2013) to obtain a construct for the wild-type expression of POLR2A circular RNA (#1, Figure 3C). A series of deletions at Alus, insertion with a pair of IRAlus from 3’ UTR of human NICN1 gene (Chen et al., 2008), or insertion with a nonrepetitive fragment from another POLR2A intron were further obtained with PCR amplification and individually inserted into the same pZW1 vector (Figures 3C, 3D and 4D) by using ClonExpress™ II One Step Cloning Kit (Vazyme). All primers used for constructs are listed in Table S3.

Total RNA Isolation, Polyadenylated/Nonpolyadenylated RNA Separation, RNase R Treatment and RNA-Seq

Cultured cell lines or cells with different treatments/transfections were harvested in Trizol (Invitrogen) and RNAs were extracted with Trizol Reagent (Invitrogen) according to the manufacturer's protocol, followed by DNase I treatment at 37°C for 30 min, twice (DNA-free kit, Ambion). Polyadenylated and nonpolyadenylated RNA separation was carried out as described previously (Yang et al., 2011). RNase R treatment was carried out as described previously (Zhang et al., 2013). Briefly, purified RNAs were incubated with 40 U of RNase R (Epicenter) for 3 hr at 37°C and then were subjected to purification with Trizol. RNA-seq libraries were prepared by using Illumina TruSeq RNA Sample Prep Kit V2 and subjected to deep sequencing with Illumina HiSeq 2000 at CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China.

RT-PCR and Northern Blot

Similar amount of RNAs were used for RT-PCR validation. Semi-quantitative RT-PCR from similar amount of RNAs before and after RNase R treatment was used to evaluate the relative abundance of RNA transcripts as described previously (Zhang et al., 2013). The first strand cDNA was transcribed with SuperScript III (Invitrogen) with random hexamers, and PCR amplification was performed by using 2× Taq Plus Master Mix (Vazyme) according to the manufacturer's protocol. Primers for PCR amplification were listed in Table S3. Convergent primer sets were applied to detect the relative abundance of circular RNAs by regular RT-PCR (Zhang et al., 2013). Primer sequences were listed in Table S3. Northern blots were carried out according to the manufacturer's protocol (DIG Northern Starter Kit, Roche). RNAs were loaded on native Agrose or denaturing PAGE gels. Digoxigenin (Dig) labeled antisense and sense probes were made using either SP6 or T7 RNA polymerase by in vitro transcription with the AmpliScribe SP6 and T7 High Yield Transcription Kits (Epicenter).

SUPPLEMENTAL REFERENCES


Recalibrating Back-Spliced Junction Reads with a Customized Algorithm and RT-PCR Validation of Predicted Circular RNAs, Related to Figure 1

(A) Junction reads with shifted alignments near to canonical splice sites. Due to sequence similarity at the ends of certain exons (blue bars), mapped back-spliced junction reads were split (dash arc lines) at incorrect sites, leading to mapping to incorrect locations (red bars). Normally spliced exons were labeled with gray bars. Mis-mapped back-spliced junction reads were corrected with a customized algorithm (Figure 1 and Experimental Procedures).

(B) Summary of RNA-seq data sets from H9 hESCs. Polyadenylated RNAs (poly(A)+ for short), poly(A)–/Ribo– RNAs (poly(A)– for short) and RNase R-treated poly(A)– RNAs were individually prepared for RNA-seq according to previous studies (Yang et al., 2011; Zhang et al., 2013). Back-spliced junction reads were mainly determined from poly(A)– and RNase R-treated poly(A)– samples.

(C) Comparison of our pipeline with MapSplice in circular RNA prediction. More circular RNAs (at least one junction read) could be predicted by using our pipeline (blue pie) than by those MapSplice (Jeck et al., 2013) (pink pie) from H9 poly(A)–/RNase R RNA-seq data set.

(D) Comparison of our pipeline with Segemehl in circular RNA prediction. Top, the number of predicted circular RNAs are comparable between our pipeline (green bars) and Segemehl (Hoffmann et al., 2014) (gray bars) from H9 poly(A)–/RNase R RNA-seq data set. Bottom, our pipeline base on TopHat (Kim et al., 2013) and TopHat-Fusion (Kim and Salzberg, 2011) requires much less memory than Segemehl does.

(E) Validation of predicted circular RNAs. Top, schematic diagram to show the genomic structure and divergent PCR primers for circular RNAs. Blur bars, back-spliced exons. Grey bars, normally spliced exons. Dash arc lines, back-spliced circularization. Arrows, divergent PCR primers. Bottom, RT-PCR validation of circular RNAs with divergent PCR primers. Note that examined circular RNAs were stable before and after RNase R treatment.
(A and B) Genomic features of back-spliced exons in HeLa S3 poly(A)/-Ribo- RNA-seq. Highly expressed (junction read RPM ≥ 0.1) circular RNAs in HeLa S3 poly(A)/-Ribo- RNA-seq (GEO: GSE26284) were predicted with our pipeline (A) or Salzman’s methods (Salzman et al., 2013) (B). Left, most circularized exons are located in the middle of genes, and much less are from the first or the last exons. Middle, most (about 90%) of circular RNAs contain multiple circularized exons. Right, circular RNAs with only one back-spliced exon were much longer than those containing multiple back-spliced exons (**, P-value < 2.2 × 10^{-15} (A), P-value = 1.995 × 10^{-13} (B), Wilcoxon rank-sum test).

(C) On average, more than four Alu elements, either convergent (left) or divergent (right), were identified in 1,000 of randomly selected long control intron pairs (at least 10,000 nt in length).

(D) No juxtaposition of orientation-opposite Alu elements from long control intron pairs. Convergent (G) or divergent (H) Alu elements in flanking introns (top) were divided according to their relative positions as 1st (25%), median (50%) or 3rd quantile (75%). Their relative positions (distances to splice sites) were clustered respectively to create isodenses (curves on which points have the same density) to estimate the overall distribution of Alus in both upstream and downstream flanking introns. Spearman correlation coefficient < 0.3.
Figure S3. Recapitulation of Circular RNA Formation with Repetitive or Nonrepetitive Complementary Sequences, Related to Figure 3

(A) The same migration of recapitulated and endogenous POLR2A circular RNA. Total RNAs from HeLa and H9 cells were used as NB controls. Recapitulated POLR2A circular RNAs were transfected with three vectors used in (C). Total RNAs were resolved on a 5% denaturing PAGE gel for NB with the same probe used in Figure 3. NB images were shown with short (left) and long (right) exposure time. Red arrow, endogenous POLR2A circular RNA in untransfected H9 and HeLa cells. *, linear RNAs.

(B) Visualization of the mouse POLR2A locus from UCSC genome browser. Only one annotated SINE element and a low complexity RNA element are located in the mouse POLR2A locus, compared with pairs of Alu elements in the related human POLR2A locus. Note that no mouse POLR2A circular RNA was identified in this region from mouse RNA-seq data sets.

(C) Recapitulation of POLR2A circular RNA formation in mouse NIH 3T3 cell lines. Left, a schematic drawing of recapitulated vectors (Figures 3C and 3D) with a variety of genomic sequences for POLR2A circular RNA recapitulation (#1 to #6). NB probes for the circularized exons are highlighted as blue bars with dash lines. PCR primers for spliced RNAs are indicated as gray arrows. Right, validation of recapitulated circular RNAs by NB and linearized RNAs by RT-PCR. Note that the expression pattern of either the circular RNA (NB) or linearized RNAs (RT-PCR) in each expression vector in mouse NIH 3T3 cell lines is similar to those expressed in human HeLa cell lines (Figure 3C and 3D). NT, not transfected; EV, empty vector. *, linear RNAs.
Figure S4. Validation of POLR2A Circular RNA Formation and Replacement Strategy for Recapitulated POLR2A Circular RNA, Related to Figure 3

(A) Pairs formed with a 1/2 Alu and a full length Alu element. These pairs of partial IRAUs were predicted from plasmids used in Figure 3D (#2 and #3, respectively). Their minimum free energies ($\Delta G$) are also indicated. Note that $\Delta G$s of partial IRAUs are much lower than those of full length IRAUs (Figure 3B).

(B) Strategy for IRAU replacement with another pair of IRAUs from other gene. A pair of IRAUs from 3’ UTR of human NCI1 gene (Chen et al., 2008) was inserted into flanking introns to replace the wild-type POLR2A IRAUs to obtain a new expression vector for POLR2A circular RNA (#4, Figure 3D).

(C) Strategy for IRAU replacement with a pair of nonrepetitive complementary sequences. A nonrepetitive fragment from another POLR2A intron was cloned into both upstream and downstream POLR2A flanking introns with an orientation-opposite pattern to obtain a new expression vector for POLR2A circular RNA (#5, Figure 3D).
Figure S5. Complementary Sequences Identified in the Human GCN1L1 Locus and Conservation Analysis across Species, Related to Figure 3

(A) Complementary sequences in the examined human GCN1L1 locus (Figure 3E). Note that this pair of complementary sequences is almost perfectly matched with only three nucleotide differences.

(B) Evolutionary dynamics of complementary sequences in the examined human GCN1L1 locus. Sequence conservation analysis was visualized with the VISTA browser. Top, RefSeq exons, including back-spliced (blue bars) or non-back-spliced (gray bars) exons are indicated. The nonrepetitive complementary sequences (black lines) were predicted to form a perfect RNA pairing (black arc line). Middle, rapid evolutionary changes of complementary sequences among primates. y axis, species selected for comparison (left panel) and levels of conservation based on the genomic sequence similarity (right panel). Colors of conserved regions are labeled by VISTA according to RefSeq annotations (exons in blue shadow and introns in red shadow). Bottom, genomic DNA sequences from human, non-human primates to mouse. Note that the identified nonrepetitive complementary sequences in the human GCN1L1 locus undergo rapid evolutionary changes across species and even among primates (such as in rhesus monkey, labeled in red) than other intronic sequences.
Figure S6. De Novo Cufflinks Assembly Reveals Species-Specific Alternative Splicing in ZWILCH Locus, Related to Figure 4

(A) No alternative splicing in the examined human ZWILCH locus. Because of the competition of RNA pairing within the individual intron, splicing happened in a constitutive way to generate linearized RNAs without exon skipping. No alternatively spliced transcript could be determined in the examined human ZWILCH locus with de novo Cufflinks v2.0.2 (Trapnell et al., 2013) assembly from poly(A)+ RNA-seq. Note that no circular RNA was identified in the examined human ZWILCH locus (top panel, Figure 4B).

(B) Alternative splicing in the examined mouse ZWILCH locus. More potential RNA pairing could be formed across flanking introns than within the individual intron, thus splicing could occur with exon skipping (highlighted with red arrow on the right), which could be determined in the examined mouse ZWILCH locus with de novo Cufflinks v2.0.2 (Trapnell et al., 2013) assembly from poly(A)+ RNA-seq. Note that skipped exons were further processed by back-spliced circularization to form a circular RNA, identified in the related mouse ZWILCH locus (bottom panel, Figure 4B).

Cell 159, 1–14, September 25, 2014 ©2014 Elsevier Inc.