Resource

Diverse alternative back-splicing and alternative splicing landscape of circular RNAs

Xiao-Ou Zhang,[1,2,5] Rui Dong,[1,2,5] Yang Zhang,[2,3,5] Jia-Lin Zhang,[1,2,3] Zheng Luo,[1,2] Jun Zhang,[3] Ling-Ling Chen,[2,3,4] and Li Yang[1,2,4]

[1]*Key Laboratory of Computational Biology, CAS Center for Excellence in Brain Science and Intelligence Technology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;* [2]*University of Chinese Academy of Sciences, Beijing 100049, China;* [3]*State Key Laboratory of Molecular Biology, CAS Center for Excellence in Molecular Cell Science, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China;* [4]*School of Life Science, ShanghaiTech University, Shanghai 20003, China*

Circular RNAs (circRNAs) derived from back-spliced exons have been widely identified as being co-expressed with their linear counterparts. A single gene locus can produce multiple circRNAs through alternative back-splice site selection and/or alternative splice site selection; however, a detailed map of alternative back-splicing/splicing in circRNAs is lacking. Here, with the upgraded CIRCexplorer2 pipeline, we systematically annotated different types of alternative back-splicing and alternative splicing events in circRNAs from various cell lines. Compared with their linear cognate RNAs, circRNAs exhibited distinct patterns of alternative back-splicing and alternative splicing. Alternative back-splice site selection was correlated with the competition of putative RNA pairs across introns that bracket alternative back-splice sites. In addition, all four basic types of alternative splicing that have been identified in the (linear) mRNA process were found within circRNAs, and many exons were predominantly spliced in circRNAs. Unexpectedly, thousands of previously unannotated exons were detected in circRNAs from the examined cell lines. Although these novel exons had similar splice site strength, they were much less conserved than known exons in sequences. Finally, both alternative back-splicing and circRNA-predominant alternative splicing were highly diverse among the examined cell lines. All of the identified alternative back-splicing and alternative splicing in circRNAs are available in the CIRCpedia database (http://www.picb.ac.cn/rnomics/circpedia). Collectively, the annotation of alternative back-splicing and alternative splicing in circRNAs provides a valuable resource for depicting the complexity of circRNA biogenesis and for studying the potential functions of circRNAs in different cells.

[Supplemental material is available for this article.]

Circular RNAs (circRNAs) formed by exon back-splicing (circularization) were originally identified in the 1990s (Nigro et al. 1991; Capel et al. 1993). Recently, circRNAs were rediscovered and shown to be the products of thousands of loci in eukaryotes, from fly and worm to mouse and human (Jeck et al. 2013; Memczak et al. 2013; Salzman et al. 2013; Guo et al. 2014; Westholm et al. 2014; Zhang et al. 2014; Ivanov et al. 2015). Recent research into circRNA biogenesis has shown that back-splicing is catalyzed, though inefficiently (Zhang et al. 2016), by the canonical spliceosomal machinery (Ashwal-Fluss et al. 2014; Starke et al. 2015; Wang and Wang 2015) and modulated by both *cis*-elements and *trans*-factors (Ashwal-Fluss et al. 2014; Zhang et al. 2014; Conn et al. 2015; Ivanov et al. 2015; Starke et al. 2015; for review, see Chen and Yang 2015; Chen 2016). Different from the canonical splicing that joins an upstream 5′ splice (donor) site with a downstream 3′ splice (acceptor) site in a sequential order to produce a linear RNA, back-splicing occurs in a reversed orientation that links a downstream 5′ splice (donor) site to an upstream 3′ splice (acceptor) site to yield a circRNA. Thus, the identification of back-splice junctions is crucial to annotate circRNAs (Jeck and Sharpless 2014).

With the intrinsic feature of being covalently closed without open ends, circRNAs are largely missed by polyadenylated transcriptome profiling but can be captured by RNA deep-sequencing (RNA-seq) from nonpolyadenylated RNAs (Yang et al. 2011; Jeck et al. 2013; Memczak et al. 2013; Salzman et al. 2013; Zhang et al. 2014). Two nonpolyadenylated RNA isolation strategies have been applied to RNA-seq to retrieve back-splice junction reads for circRNA annotation. On the one hand, nonpolyadenylated RNAs can be co-collected with polyadenylated transcripts after depleting ribosomal RNAs (ribo$^-$) (Memczak et al. 2013). On the other hand, the relatively purer nonpolyadenylated RNA fractionation can be enriched by depleting both ribosomal RNAs and polyadenylated RNAs (poly(A)$^-$/ribo$^-$, or p(A)$^-$ for simplicity) (Yang et al. 2011; Zhang et al. 2014). By counting back-splice junction reads, the expression of individual circRNA can be quantitatively evaluated (Memczak et al. 2013; Zhang et al. 2014).

Interestingly, a single gene locus can produce multiple circRNAs through alternative back-splicing (circularization) by a mechanism associated with the competition of putative RNA pairs across introns that bracket the circle-forming exons (Zhang et al.

[5]**These authors contributed equally to this work.**
Corresponding authors: linglingchen@sibcb.ac.cn, liyang@picb.ac.cn
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.202895.115. Freely available online through the *Genome Research* Open Access option.

2014). Theoretically, there are two types of alternative back-splicing (Fig. 1A). One type is alternative 5′ back-splicing, in which two or more 5′ downstream back-splice sites alternatively link to the same upstream 3′ back-splice site in a reversed orientation. The other one is alternative 3′ back-splicing, in which two or more upstream 3′ back-splice sites alternatively link to the same downstream 5′ back-splice site in a reversed orientation. Apparently, alternative back-splicing further expands the complexity of circRNA formation; however, the detailed annotation of alternative back-splicing is largely unknown.

Because the majority of annotated human circRNAs consist of multiple exons (Zhang et al. 2014), alternative splicing, in principle, should be yet another mechanism that can enlarge the diversity of circRNAs. It is well known that most multiexonic genes undergo alternative splicing to generate multiple (linear) mRNAs (Nilsen and Graveley 2010). Thus, alternative splicing can also expand the diversity of circRNAs. For example, some cassette exons are more favorably included in circRNAs than in linear mRNAs (Zhang et al. 2014), and some introns are retained in circRNAs (Salzman et al. 2013; Zhang et al. 2014; Li et al. 2015) through alternative splicing. Nevertheless, the precise annotation of alternative splicing within circRNAs is unclear, and the degree to which the difference in the alternative splicing pattern between circRNAs and their correlated linear counterparts remains to be investigated.

We hereby applied CIRCexplorer2, an upgraded computational pipeline, to identify both alternative back-splicing (Fig. 1A) and alternative splicing (Fig. 1B) events in circRNAs from various p(A)$^-$ RNA-seq data sets. Through a comparison with parallel poly(A)$^+$ (p(A)$^+$ for simplicity) RNA-seq data sets in the same cell lines, thousands of alternatively back-spliced and circRNA-predominant alternatively spliced exons, including previously unannotated ones, were identified in circRNAs with variable expression patterns from different p(A)$^-$ and p(A)$^-$/RNase R RNA-seq data sets. Together, the diverse landscape of alternative back-splicing and alternative splicing in circRNAs provides a valuable resource for depicting the complexity of circRNA formation.

Results

An upgraded computational pipeline for circRNA annotation

Multiple computational methods have been recently developed to detect back-splice junctions for circRNA annotation (Memczak et al. 2013; Hoffmann et al. 2014; Westholm et al. 2014; Zhang et al. 2014). Our previously reported pipeline, CIRCexplorer (Zhang et al. 2014), has been reported as one of the best pipelines for circRNA annotation by identifying back-splice junctions (Hansen et al. 2016). To systematically identify both alternative back-splice junctions (Fig. 1A) and alternative splice junctions (Fig. 1B) in circRNAs, we upgraded our pipeline to CIRCexplorer2 (Fig. 1C; Methods). Several major improvements have been implemented in the upgraded pipeline. First, according to requests from many users, we have incorporated additional aligners, such as STAR (Dobin et al. 2013), MapSplice (Jeck et al. 2013), and segemehl (Hoffmann et al. 2014), to fit the different requirements/preferences of RNA-seq mapping. All of these aligners in CIRCexplorer2 can be used to identify back-splicing with similar outcomes (Supplemental Fig. S1A,B). It is worth noting, however, that combining different aligners might provide a better prediction of back-splicing/circularization (Zhang et al. 2014; Hansen et al. 2016). Second, the p(A)$^-$ RNA-seq reads that mapped to the genome and the collinear exon-exon junctions were not simply discarded but were instead further de novo assembled for the identification of novel exons and novel splicing events. Finally, TopHat-unmapped but TopHat-Fusion-mapped reads were realigned to both known and de novo assembled annotations to determine back-splice junctions from either annotated and/or novel exons (Fig. 1C; Supplemental Methods). The false discovery rate of the upgraded CIRCexplorer2 pipeline remains at the low level (Supplemental Fig. S1C) when checked with the reported strategy (Hansen et al. 2016).

Detection of alternative back-splicing/splicing in circRNAs from p(A)$^-$ RNA-seq data sets

Next, we applied CIRCexplorer2 to identify both alternative back-splicing and alternative splicing in circRNAs from various p(A)$^-$ RNA-seq data sets of human cell lines, including the human embryonic stem cell (hESC) H9 line, human ovarian carcinoma PA1 cells, and 11 ENCODE cell lines (Methods). These data sets contain tens of thousands to hundreds of millions of RNA-seq reads (Supplemental Table S1). Parallel p(A)$^+$ RNA-seq data sets were used to discriminate circRNA-specific/-predominant alternative splicing from the linear RNA counterparts. The p(A)$^-$/RNase R RNA-seq data sets from both H9 and PA1 cells were generated in the laboratory to confirm that the detected alternative back-splicing and alternative splicing events were from circRNAs but not from their linear counterparts.

In total, more than 10,000 alternative back-splicing and alternative splicing events were identified in at least one of the examined cell lines (Fig. 1D; Supplemental Table S2). Strikingly, with an additional de novo assembly step in the updated CIRCexplorer2, thousands of novel exons were detected in circRNAs from examined cell lines (Fig. 1D). The identified alternative back-splicing and alternative splicing can be visualized at CIRCpedia (http://www.picb.ac.cn/rnomics/circpedia) (Supplemental Fig. S2).

Although ribo$^-$ RNA-seq data sets were used for circRNA annotation, such data sets were not suitable for the analysis of circRNA-alternative splicing. Because both polyadenylated and nonpolyadenylated transcripts are included in the ribo$^-$ RNA population, it is impractical to discriminate whether the identified alternative splicing events are from linear (m)RNAs or circRNAs (Fig. 1E; Supplemental Fig. S3). Thus, in the current study, only p(A)$^-$ and p(A)$^-$/RNase R treated RNA-seq data sets were used to profile alternative back-splicing and alternative splicing in circRNAs.

Landscape of alternative back-splicing in circRNAs

With CIRCexplorer2, more than 10,000 circRNAs were identified, and their downstream 5′ back-splice sites and upstream 3′ back-splice sites were accordingly annotated (Supplemental Fig. S4A). Of the highly expressed circRNAs with mapped back-splice junction reads $\geq$ 0.1 RPM (mapped back-splice junction Reads Per Million mapped reads) (Zhang et al. 2014), up to 30% were alternatively back-spliced (Fig. 2A; Supplemental Fig. S4B; Supplemental Table S2), which suggests that alternative back-splicing is widely distributed in circRNAs. We further used PCU (Percent Circularized-site Usage) (Methods), the usage of each alternative back-splice site, to evaluate and compare the diverse alternative back-splicing events across samples (Fig. 2B).

In total, up to 70% of the alternative 5′/3′ back-splicing in highly expressed circRNAs was detected in multiple p(A)$^-$ RNA-seq data sets (Supplemental Fig. S4C), and the use of these alternative back-splice sites is significantly diverse among the examined

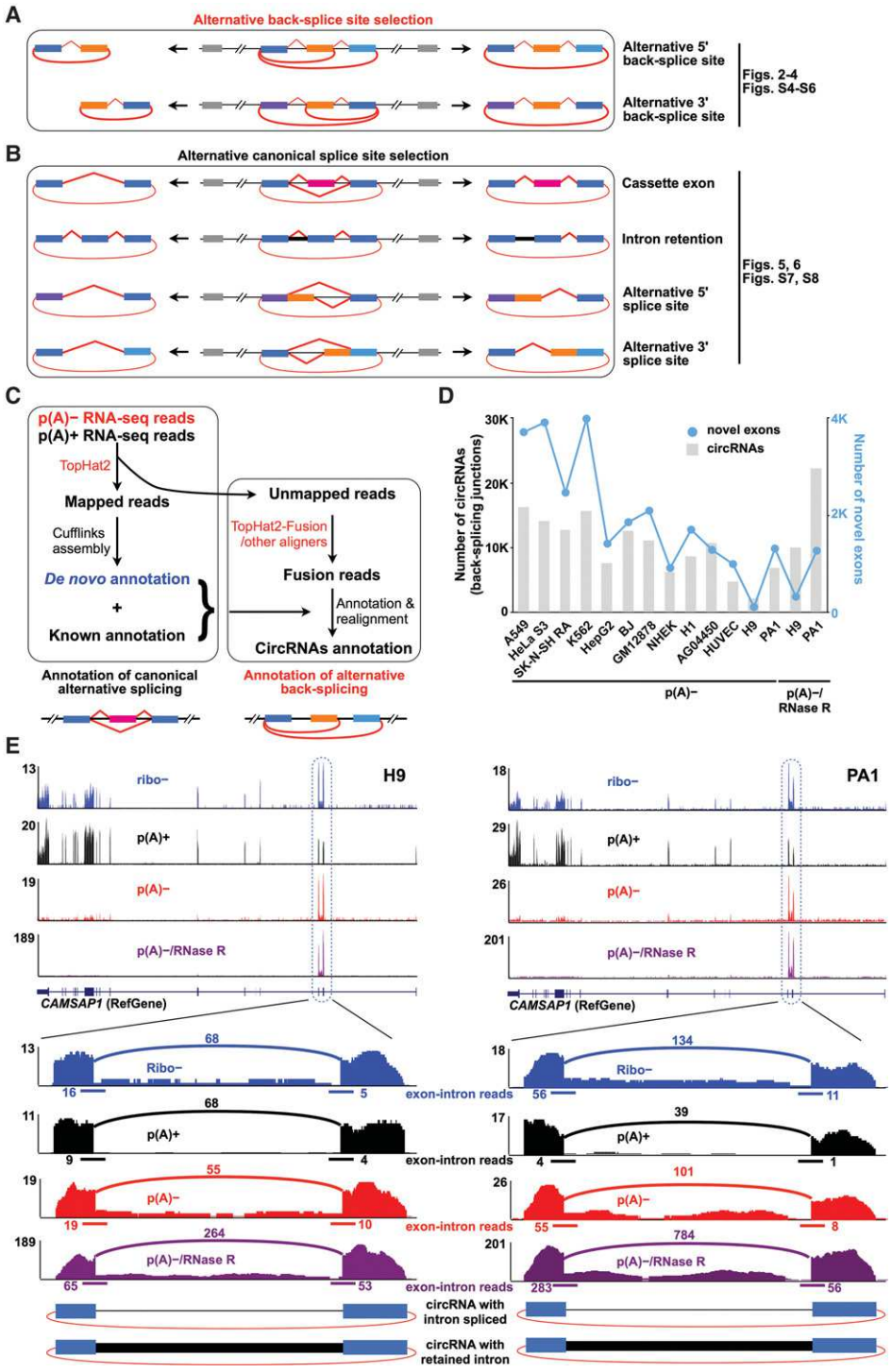

Figure 1. An upgraded computational pipeline (CIRCexplorer2) to systematically identify alternative (back-)splicing in back-spliced circular RNAs (circRNAs). (*A*) Schematic diagrams of two types of alternative back-splicing. Colored bars, exons. Black lines, introns. Red polylines, (canonical) collinear splicing. Red arc lines, back-splicing (circularization). (*B*) Schematic diagrams of four basic types of alternative splicing. Colored bars, exons. Black lines, introns. Red lines, splicing. Red arc lines, back-splicing (circularization). (*C*) The schematic diagram of CIRCexplorer2. The analysis was performed as described (Zhang et al. 2014) with modifications (Supplemental Methods). Alternative back-splicing and alternative splicing in circRNAs were determined with stringent criteria (Supplemental Methods). (*D*) Ten thousand circRNAs (gray bars) were detected by CIRCexplorer2. Thousands of novel exons (blue points) were identified in circRNAs in different human cell lines with de novo assembly (Supplemental Methods). (*E*) The identification and visualization of circRNAs in the *CAMSAP1* locus from H9 (*left* panel) or PA1 (*right* panel) cell lines. Different types of RNA-seq data sets from ribo$^-$, p(A)$^+$, p(A)$^-$ or p(A)$^-$/RNase R RNA populations were used for comparison. *CAMSAP1* circRNAs could be determined from ribo$^-$, p(A)$^-$, and p(A)$^-$/RNase R RNA-seq data sets by identifying back-splice junctions. Notably, ribo$^-$ RNA-seq is not suitable to study canonical splicing events (intron retention, in this case) that occur specifically within circRNAs, as ribo$^-$ RNAs contain both polyadenylated and nonpolyadenylated transcripts. Blues bars, exons. Black lines, introns. Black thick line, the retained intron. Red arc lines, back-splicing (circularization).

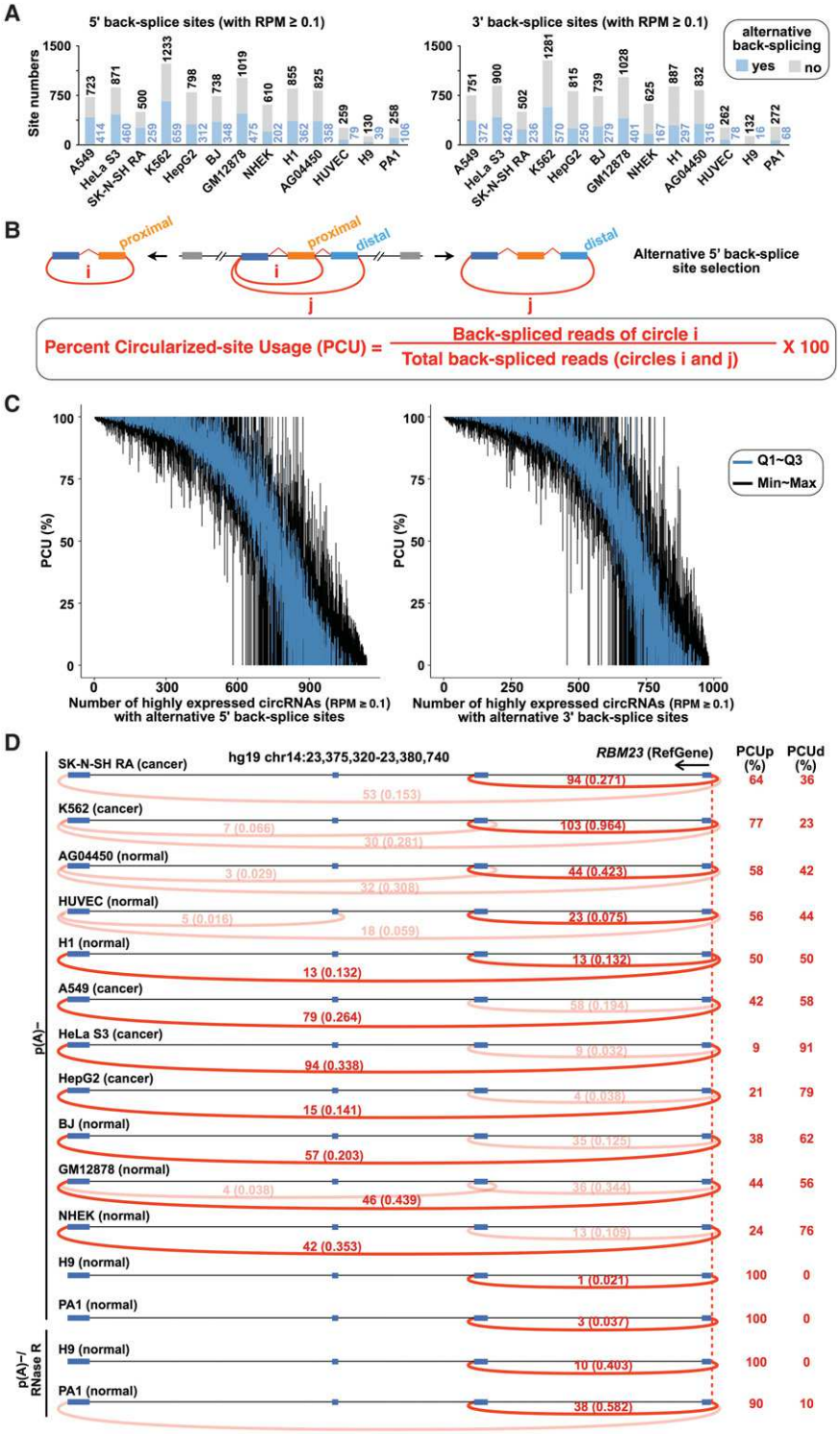

Figure 2. The diverse landscape of alternative back-splicing. (*A*) Approximately 12%–57% of back-splice sites are alternatively selected among high-confidence expressed circRNAs with RPM (mapped back-splice junction Reads Per Million mapped reads) ≥ 0.1. (*B*) A schematic diagram of alternative 5′ back-splicing and its quantification (*top* panel). The use of proximal and distal 5′ back-splice sites can be quantitated by the Percent Circularized-site Usage (PCU, *bottom* panel) with detected back-splice junction reads (i and j, respectively). (*C*) Diverse usage of alternative 5′ (*left*) and 3′ (*right*) back-splice sites among different cell lines. Each blue vertical line denotes PCU variation for one circRNA from the first quartile (Q1) to the third quartile (Q3) across cell lines, and each black vertical line denotes PCU variation from the minimum to the maximum. Note that only highly expressed circRNAs with RPM ≥ 0.1 in at least three cell lines were used for this analysis. (*D*) Visualization of alternative 5′ back-splice site usage in circRNAs produced from the *RBM23* locus across different cell lines. Predicted circRNAs in the *RBM23* locus were indicated by red arc lines with raw back-splice junction reads and normalized RPMs (numbers *above* each arc line, *left* panel). The use of proximal alternative 5′ back-splice site (PCUd) or distal alternative 5′ back-splice site (PCUp) was calculated accordingly (*right* panel).

cell lines (Fig. 2C). For instance, there are two choices for the alternative 5′ back-splice site selection, i.e., the proximal selection and the distal selection, in the human *RBM23* circRNAs (Fig. 2D). Both events were detected in 12 of the 13 examined cell lines with available p(A)$^-$ and/or p(A)$^-$/RNase R RNA-seq data sets. However, the use of these two alternative 5′ back-splice sites was largely different among cell lines. The PCU of the proximal alternative 5′ back-splice site (PCUp) varied from 9% to 100%, and, accordingly, the PCU of the distal alternative 5′ back-splice site (PCUd) varied from 91% to almost 0 (right panel, Fig. 2D). Notably, the diverse usages of alternative back-splice sites in highly expressed circRNAs (with RPM ≥ 0.1 in at least three cell lines) were less affected by sequence depths than by their variable expression in different cell lines (Supplemental Fig. S4D).

Since the complementary sequences in flanking introns can promote exon circularization (Liang and Wilusz 2014; Zhang et al. 2014), we evaluated whether the existence of multiple pairs of intronic complementary sequences would have an effect on alternative back-splice site selection. Theoretically, an across-intron RNA pair flanking the proximal back-splice sites would lead to proximal back-splice site selection (left panels of Fig. 3A,B, top). Similarly, an across-intron RNA pair flanking the distal back-splice sites could lead to distal back-splice site selection (left panels of Fig. 3A,B, bottom). Thus, the formation of a proximal RNA pair competes with a distal RNA pair, resulting in the alternative back-splice site selection in a single gene locus. A computational analysis of the orientation-opposite complementary sequences revealed that this scenario is indeed the case. More than 70% of the highly expressed circRNAs (junction reads ≥ 0.1 RPM) with alternative back-splice site selection contained the paired intronic complementary sequences flanking both the proximal and distal 5′/3′ back-splice sites (right panels of Fig. 3A,B); in comparison,

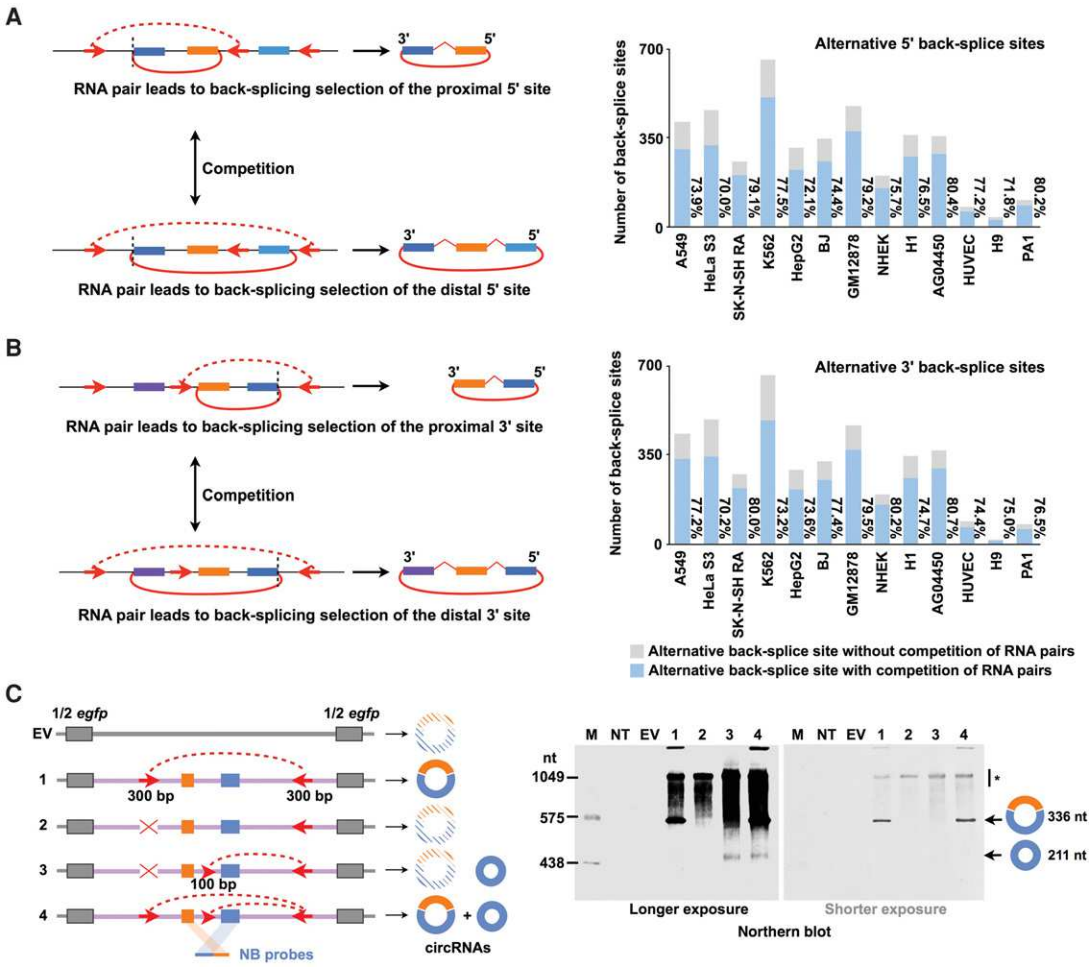

Figure 3. Competition of RNA pairs flanking proximal or distal back-splice sites leads to alternative back-splice site selection. (*A*,*B*) Potential RNA pairs (red dashed arc lines) produced by orientation-opposite complementary sequences (red arrows) flanking proximal (*left top* panels) or distal (*left bottom* panels) back-splice sites lead to alternative 5′ (*A*)/3′ (*B*) back-splice site selection, respectively (red arc lines). The competition of RNA pairs flanking proximal or distal back-splice sites leads to alternative back-splice site selection. More than 70% of the highly expressed circRNAs (RPM ≥ 0.1) with alternative back-splice site selection contain potential paired complementary sequences flanking both proximal and distal 5′/3′ back-splice sites (*right* panels). (*C*) Recapitulation of alternative back-splicing. (*Left*) A schematic drawing of *egfp* expression vectors with engineered complementary sequences for *POLR2A* circular RNA recapitulation. Half *egfp* sequences from the expression vector backbone are indicated as gray bars. *POLR2A* exonic and intronic sequences are indicated as colored bars and light purple lines, respectively. Nonrepetitive complementary sequences (red arrows) were inserted into multiple *POLR2A* intronic regions to form different RNA pairs (red dashed arc lines). Northern blot (NB) probes are indicated as colored bars. (*Right*) Validation of alternatively back-spliced *POLR2A* circRNAs by Northern blot on denaturing PAGE gel. Note that only partial complementary sequence (~100 bp) was inserted into the middle intron for smaller *POLR2A* circRNA. (*) Linear RNA background.

~20% of randomly selected nonalternative back-splicing events were flanking with paired intronic complementary sequences (Supplemental Fig. S5A,B). Clusters of proximal and distal RNA pairs across introns were seen in many gene loci (for example, the human *RBM23* locus) (Supplemental Fig. S5C,D), and this strong competition between these potential RNA pairs was correlated with the detected alternative back-splicing events (Fig. 2D). This analysis provides yet another line of evidence demonstrating that *cis*-elements can significantly affect the biogenesis of circRNAs.

We next used expression vectors to validate that the competition of paired intronic complementary sequences leads to alternative back-splice site selection. As previously reported (Zhang et al. 2014), paired complementary sequences engineered into the flanking introns could significantly increase the expression of *POLR2A* circular RNA that contains two exons (Fig. 3C, #1). When the paired structure was disrupted, the *POLR2A* circular RNA expression was dramatically reduced to undetectable levels (Fig. 3C, #2). When paired complementary sequences were engineered into the introns flanking only one exon, a smaller *POLR2A* circular RNA with that exon was induced (Fig. 3C, #3). Interestingly, when multiple complementary sequences were individually inserted into different introns that led to the competition of two RNA pairs, both the original *POLR2A* circular RNA with two exons and the smaller *POLR2A* circular RNA with only one exon could be expressed from the same expression vector (Fig. 3C, #4). Alternative back-splice sites of both recapitulated *POLR2A* circRNAs were further confirmed by Sanger sequencing after RT-PCR with divergent primers. However, it should be noted that the competition between intronic RNA pairs and their induced alternative back-splicing regulation could be more complicated under endogenous conditions (Supplemental Fig. S5C).

Back-splicing with novel exons

The de novo assembly of the unmapped reads from the p(A)$^-$ and/or p(A)$^-$/RNase R RNA-seq data sets by the upgraded CIRCexplorer2 pipeline revealed that many alternative 5′/3′ back-splice sites from previously unannotated exons (non-RefSeq, non-UCSC Known Genes, or non-Ensembl) were predominantly detected in circRNAs (Supplemental Table S3). For instance, in the human *MED13L* locus, at least four previously unannotated exons were identified in p(A)$^-$ and/or p(A)$^-$/RNase R RNA-seq data sets in PA1 and/or other cell lines (Fig. 4A; Supplemental Fig. S6A), and three of the four were alternatively back-spliced, as shown by the identified back-splice junction reads (red arc lines in Fig. 4A). The existence of these novel alternative 5′ back-spliced sites was further confirmed by Sanger sequencing (Fig. 4A, bottom panel) after RT-PCR amplification (Supplemental Fig. S6B) and by Northern blot analysis (Fig. 4B). In contrast, the inclusion of these novel exons was rarely detected in the linear *MED13L* mRNA (Fig. 4A, p(A)$^+$ RNA-seq; Supplemental Fig. S6B, RT-PCR). Although with a similar sequence feature (Supplemental Fig. S6C), hundreds of 5′/3′ back-splice sites from previously unannotated exons were identified in circRNAs among different cell lines (Fig. 4C; Supplemental Table S3) with at least one back-splice junction read from available p(A)$^-$ RNA-seq data sets but were largely missed in linear RNAs (Supplemental Fig. S6D).

These novel back-splice sites have a slightly lower splicing strength than the randomly selected annotated exons (Fig. 4D). In addition, such novel back-splice sites in general contain similar sequence signatures to those of annotated exons (Fig. 4E). Interestingly, our analysis further revealed that these novel exons are less conserved in sequence than annotated exons (Fig. 4F). As it has been suggested that back-splicing is unfavorably processed by the spliceosome (Jeck and Sharpless 2014; Chen and Yang 2015; Starke et al. 2015; Zhang et al. 2016), it is unclear how the spliceosome could specifically recognize these exons during back-splicing but not during canonical splicing. In this case, novel mechanisms that are associated with back-splicing await discovery.

The complexity of alternative splicing within circRNAs

The majority of the annotated human circRNAs consist of multiple exons (Zhang et al. 2014), indicating that potential alternative splicing events could occur during circRNA formation. With the upgraded CIRCexplorer2 pipeline and available p(A)$^-$ and p(A)$^-$/RNase R RNA-seq data sets, all four basic types of alternative splicing were identified within the circRNAs from the examined cell lines (Supplemental Fig. S7). We further quantitated the extent of different types of alternative splicing by PSI (Percent Spliced In) for cassette exon selection, PIR (Percent Intron Retention) for intron retention, and PSU (Percent Splice-site Usage) for alternative 5′/3′ splice site selection. All of the splicing events with more selection in circRNAs than those in their linear cognates were counted (Supplemental Table S4; Supplemental Methods). These analyses revealed that 20%–30% of the circRNA-specific/-predominant alternative splicing events could be detected in multiple examined cell lines (Supplemental Fig. S7).

In the current study, we focused on the analysis of alternative cassette exon inclusion/exclusion in circRNAs. A positive correlation was observed between the number of circRNAs and the number of circRNA-predominant cassette exons (Supplemental Fig. S8A). High-confidence circRNA-predominant cassette exons were further identified with a stringent pipeline (Fig 5A,B; Supplemental Fig. S8B). Genomic feature analysis revealed that the splice site strength and the density of different splicing regulators were not much different among these high-confidence circRNA-predominant cassette exons, constitutive exons, and the cassette exons that were identified in linear mRNAs (Fig. 5C,D; Supplemental Fig. S8C,D). However, circRNA-predominant cassette exons were generally less conserved than constitutive exons and cassette exons that were identified in linear mRNAs (Fig. 5E). It is unclear how these cassette exons could be predominantly spliced in circRNAs but not in their linear cognate RNAs.

Novel circRNA-predominant cassette exons

Strikingly, we have identified hundreds to thousands of previously uncharacterized circRNA-predominant cassette exons (Fig. 6A; Supplemental Table S5), representing up to 25% of the highly expressed circRNAs in different cell lines. These circRNA-predominant novel cassette exons are much less conserved than are the annotated circRNA-predominant exons or other cassette exons that were present in linear RNAs (Fig. 6B). Examples of randomly selected circRNA-predominant cassette (novel) exons (Fig. 6C, bottom), are shown in both p(A)$^-$ and p(A)$^-$/RNase R RNA-seq data sets in PA1 cells (Fig. 6C) and validated by RT-PCR and/or by Northern blot in PA1 and H9 cells (Fig. 6C,D). Finally, circRNA-predominant cassette exons could be detected in multiple cell lines with different inclusion rates, as exemplified by the new circRNA-predominant cassette exon in the human *PIP5K1C* gene (Fig. 6E).

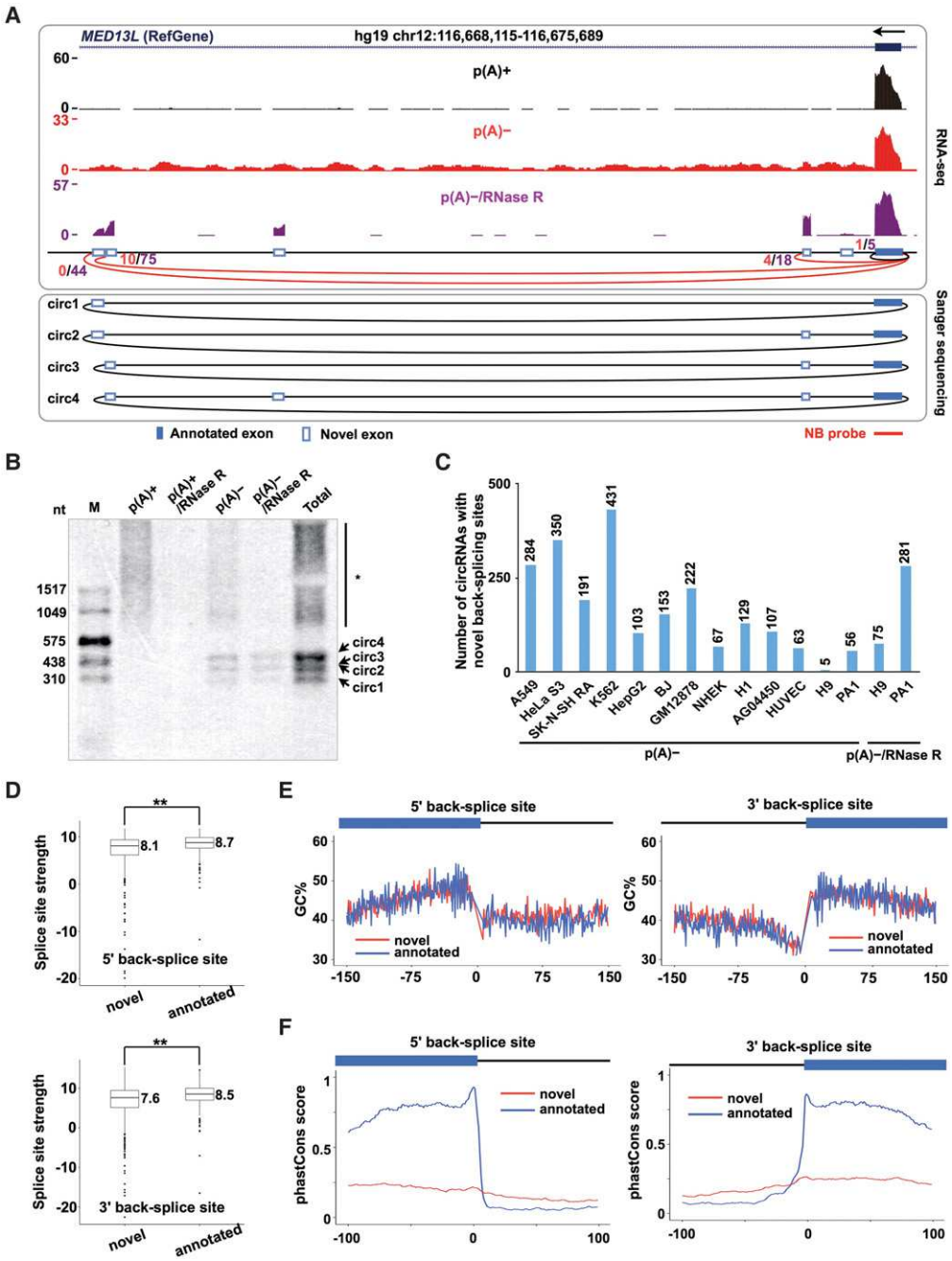

Figure 4. Unannotated exons produced from alternative back-splicing. (*A*) At least four novel exons (white bars) in the human *MED13L* locus were identified in PA1 p(A)$^-$ and/or p(A)$^-$/RNase R RNA-seq data sets. The predicted circRNAs in the *MED13L* locus were indicated by red arc lines with raw back-splice junction reads from p(A)$^-$ (red) and/or p(A)$^-$/RNase R (purple) RNA-seq data sets. Alternative back-splice sites were determined by both RNA-seq (*top* panel) and Sanger sequencing (*bottom* panel). Note that these new exons were barely detected in linear counterparts from parallel p(A)$^+$ RNA-seq (the wiggle track in black). (*B*) Validation of multiple *MED13L* circRNAs with previously unannotated exons by Northern blot on native agarose gel. Note that the validation of these *MED13L* circRNAs with novel exons was consistent with RT-PCR (Supplemental Fig. S6B). (*) Linear RNA background. (*C*) Hundreds to thousands of circRNAs were identified with novel back-splice sites across different cell lines. (*D*) Splicing strength of novel back-splice sites is comparable to that of annotated back-splice sites. (**) *P* value < 0.01, Wilcoxon rank-sum test. (*E*) Novel (red) and annotated back-spliced exons (blue) have similar GC contents. (*F*) Novel back-spliced exons (red) are less conserved in sequences than are annotated back-spliced exons (blue).

Discussion

Different from canonical splicing that joins an upstream 5′ splice (donor) site with a downstream 3′ splice (acceptor) site, splicing also occurs in a reversed orientation (back-splicing), by which a downstream 5′ splice (donor) site links an upstream 3′ splice (acceptor) site, resulting in the production of circRNAs from back-spliced exons from pre-RNAs (Chen 2016). Interestingly, a single gene locus can produce multiple circRNAs through alternative 5′/3′ back-splice site selections that are uniquely identified in

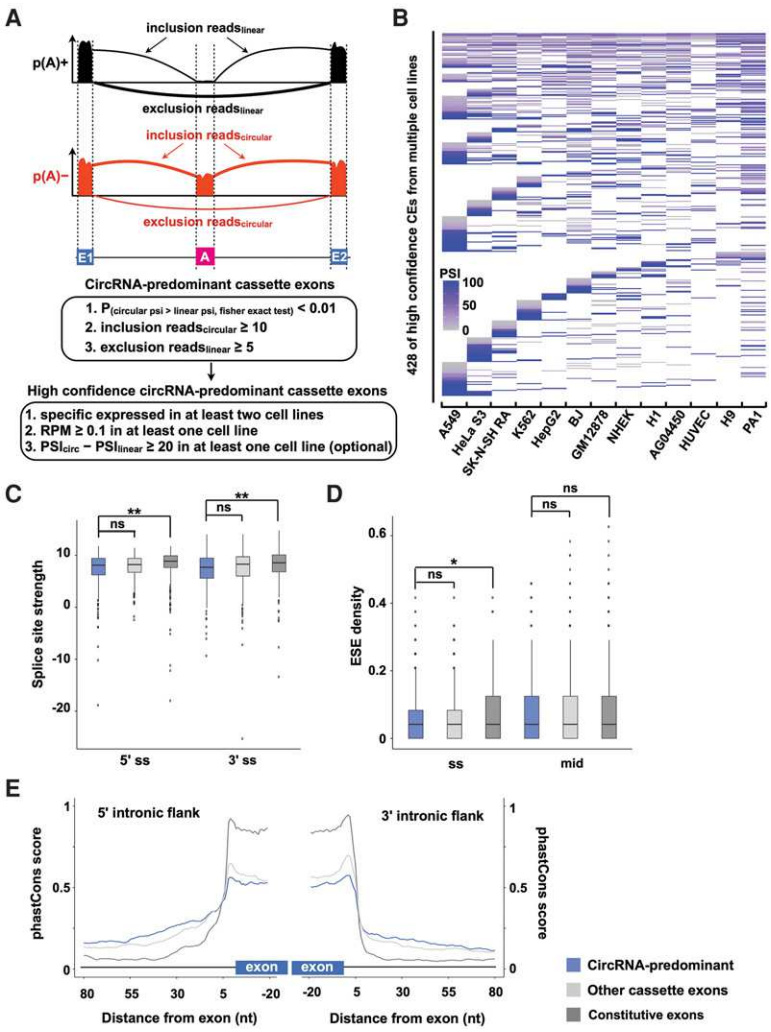

Figure 5. Characterization of circRNA-predominant alternative cassette exons. (*A*) A strategic pipeline to identify high-confidence circRNA-predominant alternative cassette exons. By comparing the alternative cassette exon selection between $p(A)^+$ and $p(A)^-$ RNA-seq data sets, circRNA-predominant alternative cassette exons were selected using stringent criteria. (*B*) The high-confidence circRNA-predominant cassette exons were determined with expression in at least two cell lines. (*C*) The strength of the 5′/3′ splice sites of high-confidence circRNA-predominant cassette exons (blue) was comparable with those of cassette exons identified in linear RNAs (light gray) and constitutive exons (dark gray). (ns) Not significant, (**) *P* value <0.01, Wilcoxon rank-sum test. (*D*) Similar densities of exonic splicing enhancers (ESE) were identified between high-confidence circRNA-predominant cassette exons (blue) and cassette exons identified in linear RNAs (light gray) and constitutive exons (dark gray). (ns) Not significant, (*) *P* value < 0.05, Wilcoxon rank-sum test. (*E*) The high-confidence circRNA-predominant cassette exons (blue) were slightly less conserved than were cassette exons identified in linear RNAs (light gray) and constitutive exons (dark gray).

circRNAs (Fig. 1A; Zhang et al. 2014). In addition, it is well known that alternative splicing significantly contributes to expanding the complexity and diversity of transcriptomes. We concluded from the current study that this scenario is also the case for circRNA. Alternative splicing greatly expands the circRNA complexity (Fig. 1B), as the majority of human circRNAs consist of multiple exons.

With CIRCexplorer2, thousands of instances of alternative back-splicing and alternative splicing were found in circRNAs (Figs. 2, 5). Importantly, thousands of new circRNA-related exons and new cassette exons were identified (Figs. 4, 6). Interestingly, the alternative use of back-splice sites and canonical splice sites was strikingly diverse among different cell lines (Figs. 2, 5), which suggests that circRNA-related alternative back-splicing/splicing events are under regulation.

What factors could contribute to alternative back-splicing regulation? It has been reported that both *cis*-elements and *trans*-factors can facilitate back-splicing, presumably by bridging two back-splice sites close together to overcome the unfavorable catalysis by the spliceosome (Ashwal-Fluss et al. 2014; Liang and Wilusz 2014; Zhang et al. 2014; Chen and Yang 2015; Conn et al. 2015). Indeed, the majority of circRNAs that undergo alternative back-splicing contain paired complementary sequences in introns flanking both proximal and distal back-splice sites, and we further confirmed that the competition of RNA pairing results in alternative back-splicing by taking advantage of engineered expression vectors. Although the paired complementary sequences are presumably identical among all tested human cell lines, diverse alternative back-splicing landscapes were observed, suggesting that the regulation of alternative back-splicing is more complicated than the current depiction. Other factors, such as additional RNA binding proteins (RBPs) that are differentially expressed in various cell lines, may contribute to the selection of alternative back-splicing, resulting in the diverse regulation of alternative back-splicing among different cell lines. In fact, as RBP-mediated regulation of alternative splicing is prevalent (Nilsen and Graveley 2010), we suspect that back-splicing could be regulated by similar mechanisms.

It is also interesting to find all four basic types of alternative splicing patterns in circRNA production (Supplemental Fig. S7; Supplemental Methods). Strikingly, many alternatively spliced cassette exons appeared to be circRNA-predominant. Although the detailed mechanism is unclear, it is possible that such events could occur post-transcriptionally during circRNA biogenesis (Kramer et al. 2015; Zhang et al. 2016). Finally, the biological significance of circRNA-predominant alternative splicing awaits further investigation.

Since most circRNAs are expressed at low levels, RNase R, an enzyme that digests linear RNAs but preserves circRNAs (Suzuki et al. 2006), has been used to enrich circRNAs and to further verify the existence of circRNA-specific alternative splicing from both PA1 and H9 samples that were prepared in the lab (Fig. 6). However, $p(A)^-$/RNaseR RNA-seq data sets were largely absent in publicly available ENCODE samples. Although we have applied the same stringent criteria (Supplemental Fig. S8B; Supplemental Methods) to annotate high-confidence circRNA-predominant

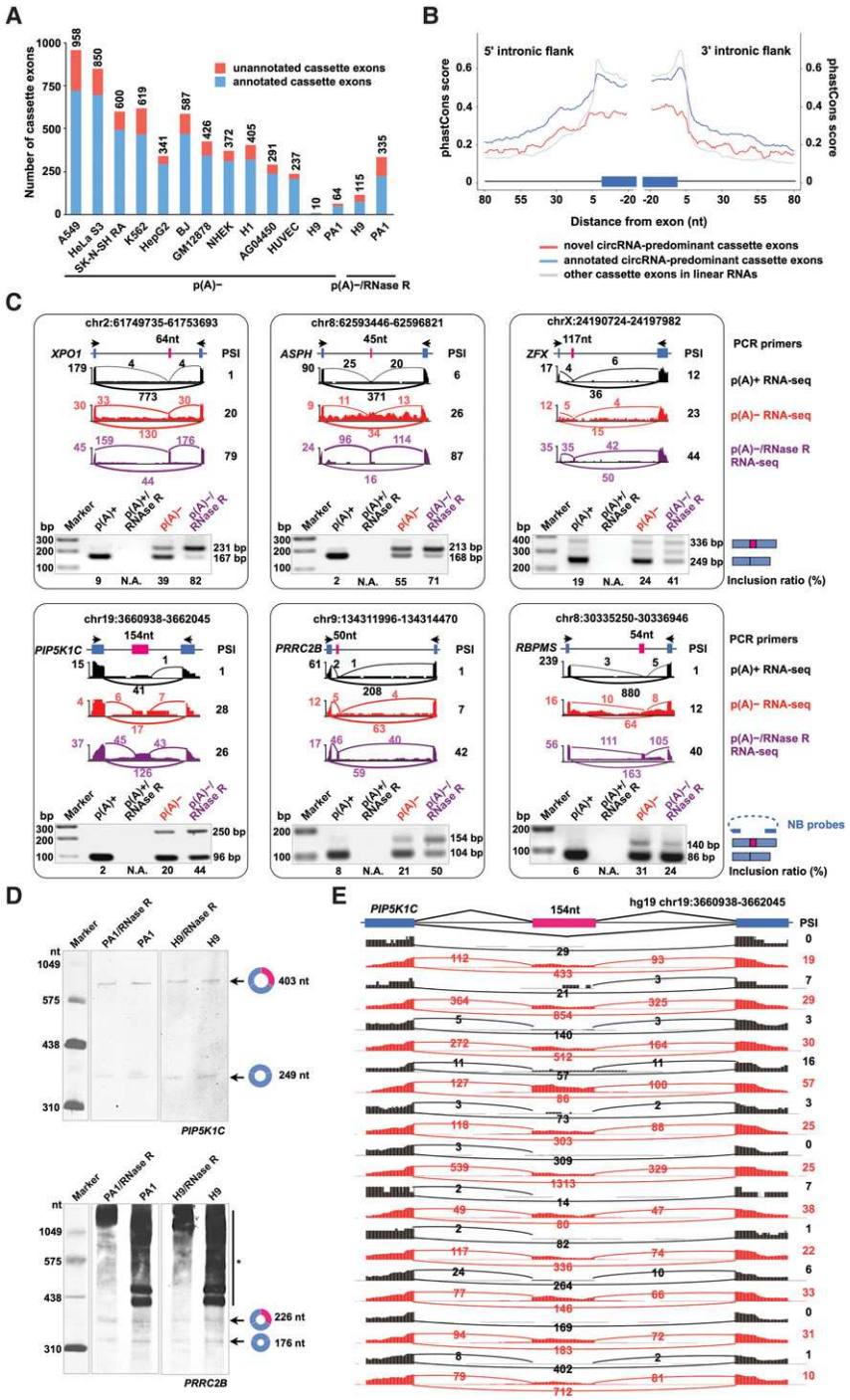

Figure 6. Unannotated circRNA-predominant alternative cassette exons. (*A*) Hundreds of previously unannotated circRNA-predominant cassette exons (red) were identified in circRNAs from individual cell lines. (*B*) Previously unannotated circRNA-predominant cassette exons (red) were much less conserved than the annotated circRNA-predominant cassette exons (blue) or cassette exons in linear RNAs (gray). (*C*) Validation of circRNA-predominant cassette exons in PA1 cells by RT-PCR. Similar to the RNA-seq results (PSI ratio), semiquantitative RT-PCR showed the detection of six circRNA-predominant cassette exons from the $p(A)^-$ and $p(A)^-$/RNase R RNA population but barely any from $p(A)^+$ RNAs. The inclusion ratio of circRNA-predominant cassette exons from RT-PCR was determined by Quantity One (Bio-Rad). Note that circRNA-predominant cassette exons in *PIP5K1C*, *PRRC2B*, and *RBPMS* loci (*bottom*) were not previously annotated by RefGene, UCSC Known Genes, or Ensembl. Magenta bars, circRNA-predominant cassette exons. Blue bars, known exons. Divergent PCR primers were indicated as black arrows. (*D*) Validation of circRNA-predominant cassette exon inclusion from different cell lines by Northern blot on denaturing PAGE gels. The circRNAs with alternative cassette exon inclusion/exclusion in both *PIP5K1C* and *PRRC2B* loci were detected in PA1 and H9 cell lines. Note that the circRNA-predominant cassette exons in *PIP5K1C* or *PRRC2B* loci are previously unannotated. Magenta in the circles, circRNA-predominant cassette exons. Blue in the circles, known exons. (*) Linear RNA background. (*E*) Visualization of circRNA-predominant cassette exon inclusion in *PIP5K1C* locus from different cell lines. The inclusion ratio of the circRNA-predominant cassette exons from RNA-seq was indicated by PSI. Note that the circRNA-predominant cassette exon in the *PIP5K1C* locus is a newly identified exon in this study. Magenta bar, circRNA-predominant cassette exon. Blue bars, other exons.

alternative splicing, we cannot rule out higher rates of false-positives in these ENCODE samples due to the lack of RNase R treatment. Nevertheless, we assume that more circRNA-specific alternative (back-)splicing events could be further revealed with the addition of extra RNase R samples and the application of stringent methods used in this study.

Collectively, we concluded that alternative back-splicing and different types of alternative splicing in circRNAs are prevalent in human transcriptomes. The alternative back-splicing and circRNA-predominant alternative splicing events are highly diverse among different cell lines, indicating that additional *cis*-elements and *trans*-factors involved in circRNA biogenesis are yet to be identified. Finally, the involvement of thousands of previously uncharacterized exons during the alternative back-splicing/splicing of circRNAs suggests an even more complex landscape of RNA (back-)splicing and its regulation in human transcriptomes.

Methods

Upgraded CIRCexplorer2 pipeline

We have upgraded the previously reported computational pipeline CIRCexplorer (Zhang et al. 2014) to a new version (CIRCexplorer2) (Supplemental Methods; Supplemental Material) to comprehensively decipher the alternative back-splicing/splicing pattern of circRNAs in multiple cell lines (GEO: GSE26284, GSE24399, GSE60467, GSE48003, and GSE75733) (Djebali et al. 2012; Zhang et al. 2014).

Characterization of alternative back-splicing

To systematically evaluate each alternative 5′/3′ back-splicing event, PCU was defined as in Figure 2B. High-confidence alternative 5′/3′ back-splicing events were further selected for more detailed characterization (Supplemental Table S2). Briefly, after clustering all circRNAs based on their 5′/3′ back-splice sites, circRNAs with highly expressed 5′/3′ back-splice sites (at least one circRNA with RPM $\geq$ 0.1) were selected for further analysis (Fig. 2A; Supplemental Fig. S4B).

Characterization of alternative splicing in circRNAs

All four basic types of alternative splicing events were detected and quantitated in highly expressed circRNAs (RPM $\geq$ 0.1) from $p(A)^-$ and/or $p(A)^-$/RNase R RNA-seq data sets with relevant metrics (Supplemental Methods). At the same time, parallel $p(A)^+$ RNA-seq data sets were aligned to the GRCh37/hg19 human reference genome by TopHat2 (Kim et al. 2013), and all relevant alternative splicing events in the linear RNAs were identified accordingly. By comparing alternative splicing between highly expressed circRNAs (RPM $\geq$ 0.1) and their linear cognates, all types of circRNA-specific/-predominant alternative splicing were determined with stringent criteria described in the Supplemental Methods.

Cell culture, total RNA isolation, polyadenylated/nonpolyadenylated RNA separation, RNase R treatment, and RNA-seq

PA1 cells were cultured using standard protocol provided by ATCC. PA1 cells were grown in MEMα (Gibco) with 10% FBS and 1× GlutaMax (Gibco). H9 cells were maintained as described (Zhang et al. 2013). Total RNA isolation, polyadenylated/nonpolyadenylated RNA separation, RNase R treatment, and RNA-seq were performed as described (Yang et al. 2011; Zhang et al. 2013; Yin et al. 2015).

RT-PCR, Northern blot, and Sanger sequencing

RT-PCR and Northern blots were used to evaluate the relative abundance of circRNAs as described (Zhang et al. 2013, 2014). PCR bands with novel circRNA-predominant exons were further subjected to Sanger sequencing. PCR primers and Northern blot probes are listed in Supplemental Table S6.

Data access

Raw and processed RNA-seq data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE75733. Source code of CIRCexplorer2 is included in the Supplemental Material. Sanger trace files are available at NCBI Trace Archives (http://www.ncbi.nlm.nih.gov/Traces/home/index.cgi) with TI numbers from 2343264014 to 2343264040.

Acknowledgments

We thank Hua-Hong Fang for technical support and deep-sequencing library preparation. H9 cells were obtained from the WiCell Research Institute. RNA-seq was performed at the CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China. This work was supported by grants 2014CB910600 and 2014CB964800 from the Ministry of Science and Technology (MoST) and grants 91540115, 91440202, 31471241, and 31322018 from the National Natural Science Foundation of China (NSFC).

Author contributions: L.Y. and L.-L.C. conceived, designed, and supervised the project. X.-O.Z. and R.D. performed the bioinformatics analysis with the assistance of Z.L. Y.Z. performed experiments with assistance from J.-L.Z. and J.Z. L.Y., L.-L.C., X.-O.Z., and R.D. analyzed the data. L.Y. and L.-L.C. wrote the paper with contributions from co-authors.

References

Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S. 2014. circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell* **56:** 55–66.

Capel B, Swain A, Nicolis S, Hacker A, Walter M, Koopman P, Goodfellow P, Lovell-Badge R. 1993. Circular transcripts of the testis-determining gene *Sry* in adult mouse testis. *Cell* **73:** 1019–1030.

Chen LL. 2016. The biogenesis and emerging roles of circular RNAs. *Nat Rev Mol Cell* **17:** 205–211.

Chen LL, Yang L. 2015. Regulation of circRNA biogenesis. *RNA Biol* **12:** 381–388.

Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, Roslan S, Schreiber AW, Gregory PA, Goodall GJ. 2015. The RNA binding protein quaking regulates formation of circRNAs. *Cell* **160:** 1125–1134.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. 2012. Landscape of transcription in human cells. *Nature* **489:** 101–108.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21.

Guo JU, Agarwal V, Guo H, Bartel DP. 2014. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol* **15:** 409.

Hansen TB, Venø MT, Damgaard CK, Kjems J. 2016. Comparison of circular RNA prediction tools. *Nucleic Acids Res* **44:** e58.

Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt LM, Teupser D, Hackermuller J, et al. 2014. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol* **15:** R34.

Ivanov A, Memczak S, Wyler E, Torti F, Porath HT, Orejuela MR, Piechotta M, Levanon EY, Landthaler M, Dieterich C, et al. 2015. Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep* **10:** 170–177.

Jeck WR, Sharpless NE. 2014. Detecting and characterizing circular RNAs. *Nat Biotechnol* **32:** 453–461.

Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. 2013. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19:** 141–157.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14:** R36.

Kramer MC, Liang D, Tatomer DC, Gold B, March ZM, Cherry S, Wilusz JE. 2015. Combinatorial control of *Drosophila* circular RNA expression by intronic repeats, hnRNPs, and SR proteins. *Genes Dev* **29:** 2168–2182.

Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, et al. 2015. Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* **22:** 256–264.

Liang D, Wilusz JE. 2014. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev* **28:** 2233–2247.

Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, et al. 2013. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495:** 333–338.

Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, Oliner JD, Kinzler KW, Vogelstein B. 1991. Scrambled exons. *Cell* **64:** 607–613.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463:** 457–463.

Salzman J, Chen RE, Olsen MN, Wang PL, Brown PO. 2013. Cell-type specific features of circular RNA expression. *PLoS Genet* **9:** e1003777.

Starke S, Jost I, Rossbach O, Schneider T, Schreiner S, Hung LH, Bindereif A. 2015. Exon circularization requires canonical splice signals. *Cell Rep* **10:** 103–111.

Suzuki H, Zuo Y, Wang J, Zhang MQ, Malhotra A, Mayeda A. 2006. Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing. *Nucleic Acids Res* **34:** e63.

Wang Y, Wang Z. 2015. Efficient backsplicing produces translatable circular mRNAs. *RNA* **21:** 172–179.

Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, Celniker SE, Graveley BR, Lai EC. 2014. Genome-wide analysis of *Drosophila* circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep* **9:** 1966–1980.

Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL. 2011. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol* **12:** R16.

Yin QF, Chen LL, Yang L. 2015. Fractionation of non-polyadenylated and ribosomal-free RNAs from mammalian cells. *Methods Mol Biol* **1206:** 69–80.

Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH, Zhu S, Yang L, Chen LL. 2013. Circular intronic long noncoding RNAs. *Mol Cell* **51:** 792–806.

Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL, Yang L. 2014. Complementary sequence-mediated exon circularization. *Cell* **159:** 134–147.

Zhang Y, Wei X, Li X, Zhang J, Zhang JL, Yang L, Chen LL. 2016. The biogenesis of nascent circular RNAs. *Cell Rep* **15:** 611–624.

Received December 8, 2015; accepted in revised form June 28, 2016.

Diverse alternative back-splicing and alternative splicing landscape of circular RNAs

Xiao-Ou Zhang, Rui Dong, Yang Zhang, et al.

Genome Res. 2016 26: 1277-1287 originally published online June 30, 2016
Access the most recent version at doi:10.1101/gr.202895.115

Supplemental Methods

Upgraded CIRCexplorer2 pipeline

To comprehensively decipher the alternative back-splicing/splicing pattern of circRNAs, we upgraded our previously reported computational pipeline CIRCexplorer (Zhang et al. 2014) to a new version (CIRCexplorer2). Several major improvements have been implemented in the upgraded pipeline. First, we have incorporated other aligners, such as STAR (Dobin et al. 2013), MapSplice (Jeck et al. 2013) and segemehl (Hoffmann et al. 2014), to fit the different requirements/preferences of RNA-seq mapping from different users. Second, poly(A)− RNA-seq reads that mapped to the genome and collinear exon-exon junctions were not simply discarded but were instead further *de novo* assembled to find novel exons and thus novel splicing events. Finally, TopHat-unmapped but TopHat-Fusion-mapped reads were realigned to both known and *de novo* assembled annotations to determine back-splice junctions from either annotated and/or novel exons (Fig. 1C). The application of *de novo* assembly leads to the discovery of a large number of novel exons. The source codes of CIRCexplorer2 can be accessed from https://github.com/YangLab/CIRCexplorer2, and the step-by-step usage is described as below.

Step 1: RNA-seq reads mapping with multiple aligners

CIRCexplorer2 makes full use of TopHat2/TopHat-Fusion for RNA-seq read mapping (Zhang et al. 2014). In addition, other aligners, such as STAR (Dobin et al. 2013), MapSplice (Jeck et al. 2013) and segemehl (Hoffmann et al. 2014), are also integrated into CIRCexplorer2 to meet different requirements for RNA-seq read mapping and data mining.

For the TopHat2/TopHat-Fusion pipeline, a two-step mapping strategy was exploited as previously described (Zhang et al. 2014), with the modifications suggested as below. Briefly, reads of human embryonic stem cell H9 with/without RNase R-treated poly(A)–/ribo– (poly(A)– for simplicity) RNA-seq (GEO:GSE48003, GEO:GSE24399 and GEO:GSE60467), human ovarian carcinoma PA1 with/without RNase R-treated poly(A)–/ribo– RNA-seq (GEO: GSE75733), or poly(A)–/ribo– RNA-seq of 11 ENCODE cell lines (GEO:GSE26284) (Supplemental Table S1) were first mapped using TopHat2 (Kim et al. 2013) (TopHat 2.0.9 with parameters: -g 1 --microexon-search -m 2) against the GRCh37/hg19 human reference genome with the UCSC Genes annotation (hg19 knownGene.txt updated at 2013/6/30). Unmapped reads were then extracted and aligned onto the GRCh37/hg19 human reference genome with TopHat-Fusion (Kim et al. 2013) (TopHat 2.0.9 with parameters: --fusion-search --keep-fasta-order --bowtie1 --no-coverage-search).

For other aligners, the same RNA-seq datasets were aligned using different aligners with respective parameters (STAR 2.4.0j with the following parameters: --chimSegmentMin 10; segemehl 0.2.0-418 with the following parameters: -S -M 1; and MapSplice 2.1.9 with the following parameters: -k 1 --non-canonical --fusion-non-canonical --min-fusion-distance 200).

Step 2: *de novo* assembly to annotate novel RNA transcripts/exons

The Cufflinks reference annotation based transcript (RABT) assembly method (Roberts et al. 2011) was used to identify new transcripts for circRNAs. In brief, TopHat2-mapped reads of poly(A)– RNA-seq were assembled using filtered gene annotations (gene annotations were collected from hg19 knownGene.txt updated at

2013/6/30, refFlat.txt updated at 2013/10/13 and ensGene.txt updated at 2014/4/6 and then filtered with at least two junction reads supporting all of the isoform junctions) using Cufflinks 2.2.1 with the following parameters: -u -F 0 -j 0. The assembled novel transcripts were combined with existing gene annotations (knownGene.txt, refFlat.txt and ensGene.txt) for later use.

Step 3: circRNA annotation with junction read re-alignments

Candidate back-splice junction reads (aligned on the same chromosome but in the non-collinear order) were extracted from the fusion alignment and re-aligned against combined (annotated and new) gene annotations to determine the precise positions of back-splice sites as previously described (Zhang et al. 2014). Note that all linear RNAs, including nascent linear RNAs, randomly degraded/deadenylated linear RNAs and spliced-out intermediates, were filtered out due to the lack of back-splicing junctions.

Similar strategy was employed as previously described (Hansen et al. 2016) to evaluate the false discovery rate of CIRCexplorer2. In brief, circRNAs identified in p(A)− RNA-seq datasets with at least three back-splicing junction reads were checked in corresponding p(A)−/RNase R RNA-seq datasets. For one specific circRNA, if the RPM of back-splicing junction reads in p(A)−/RNase R RNA-seq is higher than that in p(A)− RNA-seq, this circRNA is defined as enriched by RNase R. Otherwise, it is considered as depleted by RNase R. The false discovery rate of upgraded CIRCexplorer2 pipeline remains as low (Supplemental Fig. S1C) as that of CIRCexplorer (Hansen et al. 2016). It is worth noting that the false discovery rate analysis might also depend on many other factors, such as different sequencing depths and variable sequencing quality.

Characterization of alternative splicing in circRNAs

All four basic types of alternative splicing events, including alternative cassette exon selection, intron retention, alternative 5' splice site selection and alternative 3' splice site selection (Fig. 1B), were widely detected and quantitated in highly expressed circRNAs (RPM ≥ 0.1) from poly(A)– and/or poly(A)–/RNase R RNA-seq datasets with relevant metrics (alternative cassette exons: Percent Spliced In (PSI, Supplemental Fig. S7A) (Han et al. 2013; Irimia et al. 2014); intron retention: Percent Intron Retention (PIR, Supplemental Fig. S7B) (Braunschweig et al. 2014; Irimia et al. 2014); and alternative 5'/3' splice site selections: Percent Splice-site Usage (PSU, Supplemental Figs. S7C and S7D) (Irimia et al. 2014)). Because circRNAs lack poly(A) tails, we presumed that the splicing pattern in poly(A)– RNA-seq could represent the splicing landscape of circular RNAs and used the splicing pattern in poly(A)+ RNA-seq to evaluate the alternative splicing of linear RNAs. At the same time, parallel poly(A)+ RNA-seq datasets of relevant cell lines were aligned to the GRCh37/hg19 human reference genome by TopHat2 (Kim et al. 2013), and all of the relevant alternative splicing events in the linear RNAs were accordingly identified. By comparing alternative splicing between circRNAs and their linear cognates, all types of circRNA-specific/-predominant alternative splicing were determined based on the following criteria:

1. alternative cassette exons (Fig. 5A)

$P_{(\text{circular psi} > \text{linear psi, fisher exact test})} < 0.01$

$\text{Inclusion reads}_{\text{circular}} \geq 10$

$\text{Exclusion reads}_{\text{linear}} \geq 5$

High-confidence circRNA-predominant cassette exons were filtered as follows: 1) detected in at least two cell lines and 2) RPM ≥ 0.1 in at least one cell line in the current study. With these stringent cutoffs, about 90% of high-confidence circRNA-predominant cassette exons identified in PA1 p(A)− RNA-seq could be enriched by RNase R treatment in related p(A)−/RNase R RNA-seq (Supplemental Fig. S8B), suggesting a low false discovery rate (~10%) in our analyses. An optional filter (PSIcirc − PSIlinear ≥ 20% in at least one cell line) could be further applied to reduce the false discovery rate.

2. intron retention

Introns have no overlap with the annotated exons of any annotated genes.

Introns are covered by *de-novo*-assembled transcripts.

$PIR_{circular} > PIR_{linear}$

$P_{(exon\text{-}intron\ reads\ !=\ intron\ reads,\ binomial\ test)} \geq 0.05$

$E1I_{circular} + IE2_{circular} \geq 10$

$E1E2_{linear} \geq 5$

3. alternative 3' splice site selections

$PSU_{circular} > PSU_{linear}$

$0 < PSU_{circular} < 100\%$

Total splice site junction reads ≥ 5

4. alternative 5' splice site selections

$PSU_{circular} > PSU_{linear}$

$0 < PSU_{circular} < 100\%$

Total splice site junction reads ≥ 5

It is worth noting that intron retentions in circRNAs were frequently filtered out with RNase R treatment in PA1 and H9 cells (Supplemental Fig. S7B), which is very different to all the other three alternative splicing events that were enriched by RNase R (Supplemental Figs. S7A, S7C and S7D). It is possible that this type of circRNAs with retained intron were not stable with *in vitro* RNase R treatment, as previously reported (Zhang et al. 2014). In addition, the subgroup of circRNAs with retained introns have been recently reported to be mostly located in the nucleus (Li et al. 2015b), which is distinct from most other circRNAs that are located in the cytoplasm. It was suspected that the binding with protein cofactors in nucleus might stabilize these intron-exon circRNAs; however, the naked intron-exon circRNAs after purification is quite unstable with RNase R treatment.

CIRCpedia: an integrative database of circRNAs with detected alternative back-splicing and alternative splicing

All of the identified alternative back-splicing and alternative splicing events in circRNAs, together with newly identified exons, are available in the CIRCpedia database (http://www.picb.ac.cn/rnomics/circpedia). In this online database, multiple circRNAs produced from any individual gene locus in different cell lines can be easily searched, browsed and downloaded (Supplemental Fig. S2A). Currently, the database contains circRNA back-splicing and alternative splicing from 13 human cell lines, and information on a wider spectrum of cell-line, tissue and species samples will be constructed when additional high-quality RNA-seq datasets are available.

A simple search is available from the search page of CIRCpedia (Supplemental Fig. S2B). Users can easily query circRNA information in different cell lines and different types of alternative splicing or back-splicing. CIRCpedia provides query support by gene symbols and genomic locations. A specific gene symbol (/genomic location) will retrieve all of the circRNAs that have been identified in this gene locus (/genomic location), together with relevant alternative back-splicing and alternative splicing. In addition, users can also restrict their query to a specific type of alternative back-splicing/splicing or specific cell lines by different setting options. After the query, an informative table with genomic locations, circRNA ids, host gene names, relative expression, alternative (back)-splicing of relevant circRNAs and exon identity will be available to check online or download for further analysis. Useful links are also available to access more information or gene descriptions in GeneCard websites.

The detected alternative back-splicing and alternative splicing in circRNAs, together with the available gene annotation can be visualized in the website-embedded JBrowse (Skinner et al. 2009) (Supplemental Fig. S2C). Different types of tracks, such as gene annotation (including knownGene, refGene and ensGene), different RNA-seq tracks, alternative back-splicing/splicing of circRNAs, exon identity, and novel exons, are available to be visualized in the JBrowse.

Finally, tables for alternative back-splicing, alternative splicing and novel exons from each cell lines can be accessed from the download page (Supplemental Fig. S2D).

Splice site strength analysis

Splice site strength analysis was employed on novel back-splice sites using MaxEntScan (Yeo and Burge 2004). To better characterize novel back-splice sites, 500 annotated back-splice sites were randomly selected from annotated circRNAs as controls (Fig. 4D). A similar splice site strength analysis was performed with high-confidence circRNA-predominant cassette exons, 500 randomly selected cassette exons only in linear RNAs and 500 randomly selected constitutive exons (Fig. 5C).

GC content analysis

In sequences of approximately 150 nt upstream and downstream, 5'/3' novel and annotated back-splice sites were individually fetched and tiled up to compare the GC content (Fig. 4E).

Conservation analysis

Sequence conservation analysis was carried out by PhastCons metrics (Siepel et al. 2005). The PhastCons scores for multiple alignments of placental genomes were downloaded from UCSC, and the corresponding PhastCons scores of relevant regions were summarized to assess the conservation levels (Figs. 4F, 5E and 6B). The sequence conservation of novel/annotated back-splice sites (Fig. 4F), circRNA-predominant cassette exons (Fig. 5E) or novel/annotated circRNA-predominant cassette exons (Fig. 6B) was analyzed. A total of 500 randomly selected cassette exons only in linear RNAs and/or 500 randomly selected constitutive exons were used as controls (Figs. 5E and 6B).

Splicing regulatory element analysis

Sequences of previously established splicing regulatory elements (including ESE (Fairbrother et al. 2004), ESS (Wang et al. 2004), ISE (Wang et al. 2012) and ISS (Wang et al. 2013)) were examined from circRNA-predominant cassette exons, 500 randomly selected cassette exons only in linear RNAs and 500 randomly selected constitutive exons (Fig. 5D and Supplemental Figs. S8C and S8D). Briefly, splicing elements in the central 24 nt and 12 nt of the exonic sequences adjacent to the 5' and 3' splice sites (24 nt in total) (for ESE and ESS) or the flanking 200 nt intronic region around the 5' and 3' splice sites (for ISE and ISS) were individually analyzed and compared, as previously reported (Li et al. 2015a).

Complementary sequence analysis

For each highly expressed alternative 5'/3' back-splicing event (at least one circRNA with RPM ≥ 0.1, Fig. 2A), BLASTn (parameters: -word_size 11 -gapopen 5 -gapextend 2 -penalty -3 -reward 2) was used to detect complementary sequences (with a requirement of ≥ 50 nt) flanking the most proximal/distal back-splice sites. If both the most proximal circRNA and the most distal circRNA could be flanked by corresponding complementary sequences, this cluster of 5'/3' back-splicing events was defined as containing the feature of “competition of RNA pairs” (Figs. 3A and 3B).

To check whether the competition of RNA pairs flanking alternative 5'/3' back-splice sites is more frequent than expected by chance, control intron pairs were selected as described below. For each alternative 5' back-splicing event, two downstream non-circular RNA flanking introns were randomly selected to be individually

paired with the common upstream circRNA flanking intron. For each alternative 3' back-splicing event, two upstream non-circular RNA flanking introns were randomly selected to be individually paired with the common downstream circRNA flanking intron. This random selection of control intron RNA pairs was performed up to five times for each alternative 5'/3' back-splicing event. BLASTn (parameters: -word_size 11 -gapopen 5 -gapextend 2 -penalty -3 -reward 2) was then used to detect complementary sequences (with a requirement of ≥ 50 nt) within the control intron pairs.

Cell culture, plasmid construction, cell transfection, total RNA isolation, polyadenylated/non-polyadenylated RNA separation and RNase R treatment

Human ovarian carcinoma PA1 cells were cultured using standard protocol provided by ATCC. PA1 cells were grown in MEMα (Gibco) with 10% FBS and 1× GlutaMax (Gibco). Human embryonic stem cell line H9 cells were maintained as described previously (Zhang et al. 2013). Stem cell cultures were regularly evaluated for *POU5F1* expression every 3-4 weeks and cells were passaged every 6-7 days.

POLR2A circular RNA expression vectors with engineered complementary sequencing in two circRNA-flanking introns or one side of circRNA-flanking introns were obtained as described previously (Zhang et al. 2014). About 100 bp long complementary sequences were inserted into the middle intron between circRNA-residing exons by using BbvCI site with ClonExpress™ II One Step Cloning Kit (Vazyme) (Fig. 3C). Primers for plasmid construction were listed in Supplemental Table S6. All the expression vectors were individually transfected into human HeLa-J cells

with X-tremeGENE 9 (Roche) according to the manufacturer's protocol. Total RNAs were extracted 24 hr after transfection.

Cultured cell lines with different treatments were harvested in Trizol (Pufei) and RNAs were extracted with Trizol Reagent (Pufei) according to the manufacturer's protocol, followed by DNase I treatment at 37 °C for 30 mins (DNA-freeTM kit, Ambion). Polyadenylated and non-polyadenylated RNA separation was carried out as described previously (Yang et al. 2011; Yin et al. 2015). RNase R treatment was carried out as described previously (Zhang et al. 2013). Briefly, purified RNAs were incubated with 40 U of RNase R (Epicentre) for 3 h at 37 °C and then were subjected to purification with Trizol.

RT-PCR, Sanger sequencing, Northern blot and RNA-seq

Each from 5 μg total RNAs, p(A)+, p(A)+/RNase R, p(A)–, or p(A)–/RNase R RNA sample was used for RT-PCR and/or Northern blot analyses as described previously (Zhang et al. 2013; Zhang et al. 2014). The first strand cDNA was transcribed with SuperScript III (Invitrogen) with random hexamers in a total volume of 20 μl, and 0.5 μl first-strand cDNA products were further amplified for 30 cycles (94 °C 30 sec, 55 °C 30 sec, 72 °C 20 sec) with 2×Taq Plus Master Mix (Vazyme) according to the manufacturer's protocol. PCR bands of individual novel circRNAs were further subjected to Sanger sequencing. Northern blots were carried out according to the manufacturer's protocol (DIG Northern Starter Kit, Roche). Briefly, RNAs were loaded on native Agarose or denaturing PAGE gels. Digoxigenin (Dig) labeled antisense and sense probes were made using T7 RNA polymerase by in vitro transcription with the

RiboMAX^{TM} Large Scale RNA Production Systems (Promega). PCR primers and Northern blot probes were listed in Supplemental Table S6.

RNA-seq libraries were prepared by using Illumina TruSeq Total RNA LT Sample Prep Kit (P/N 15026495) and subjected to deep sequencing with Illumina HiSeq 2000 at the CAS-MPG Partner Institute for Computational Biology Omics Core, Shanghai, China.

References

Braunschweig U, Barbosa-Morais NL, Pan Q, Nachman EN, Alipanahi B, Gonatopoulos-Pournatzis T, Frey B, Irimia M, Blencowe BJ. 2014. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res* 24(11): 1774-1786.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1): 15-21.

Fairbrother WG, Yeo GW, Yeh R, Goldstein P, Mawson M, Sharp PA, Burge CB. 2004. RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res* **32**: W187-190.

Han H, Irimia M, Ross PJ, Sung HK, Alipanahi B, David L, Golipour A, Gabut M, Michael IP, Nachman EN et al. 2013. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature* 498(7453): 241-245.

Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt LM, Teupser D, Hackermuller J et al. 2014. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol* 15(2): R34.

Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallieres M, Tapial J, Raj B, O'Hanlon D et al. 2014. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* 159(7): 1511-1523.

Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. 2013. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19(2): 141-157.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4): R36.

Li YI, Sanchez-Pulido L, Haerty W, Ponting CP. 2015a. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res* **25**(1): 1-13.

Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L et al. 2015b. Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* 22(3): 256-264.

Roberts A, Pimentel H, Trapnell C, Pachter L. 2011. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27(17): 2325-2329.

Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. 2009. JBrowse: a next-generation genome browser. *Genome Res* **19**(9): 1630-1638.

Wang Y, Ma M, Xiao X, Wang Z. 2012. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat Struct Mol Biol* **19**(10): 1044-1052.

Wang Y, Xiao X, Zhang J, Choudhury R, Robertson A, Li K, Ma M, Burge CB, Wang Z. 2013. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat Struct Mol Biol* **20**(1): 36-45.

Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. 2004. Systematic identification and analysis of exonic splicing silencers. *Cell* **119**(6): 831-845.

Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL. 2011. Genomewide characterization of non-polyadenylated RNAs. *Genome Biol* 12(2): R16.

Yeo G, Burge CB. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* **11**(2-3): 377-394.

Yin QF, Chen LL, Yang L. 2015. Fractionation of non-polyadenylated and ribosomal-free RNAs from mammalian cells. *Methods Mol Biol* 1206: 69-80.

Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL, Yang L. 2014. Complementary sequence-mediated exon circularization. *Cell* 159(1): 134-147.

Zhang Y, Zhang XO, Chen T, Xiang JF, Yin QF, Xing YH, Zhu S, Yang L, Chen LL. 2013. Circular intronic long noncoding RNAs. *Mol Cell* 51(6): 792-806.