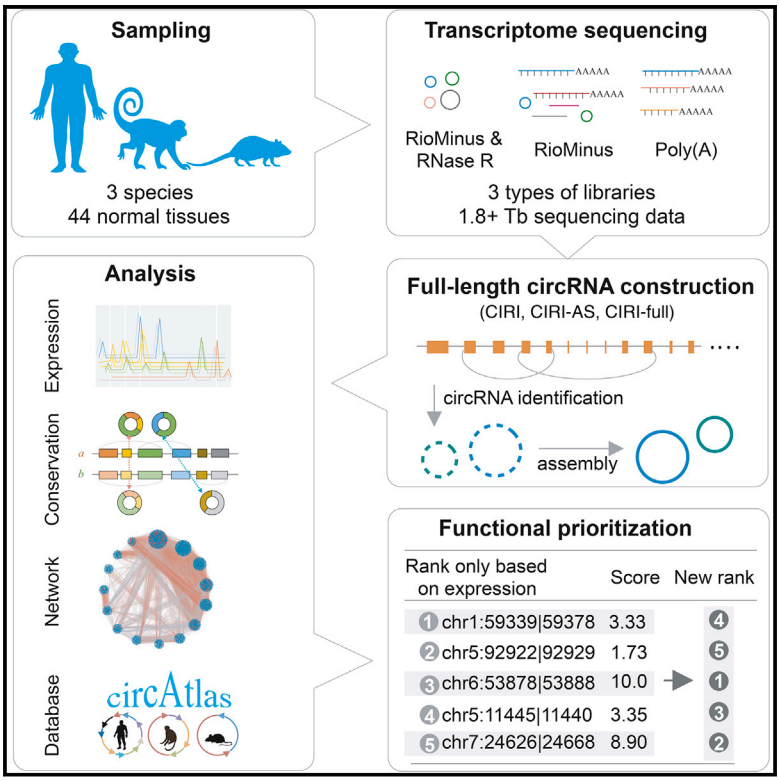
Cell Reports

Resource

Expanded Expression Landscape and Prioritization of Circular RNAs in Mammals

Graphical Abstract



Highlights

- RNA-seq libraries and data from 44 normal tissues of human, macaque, and mouse
- CircAtlas is the most comprehensive catalog of circRNAs from normal tissues
- 72.6% of circRNAs have been assembled into full-length circular transcripts
- Prioritized a new subset of circRNAs, overlapped orthologous circRNAs

Authors

Peifeng Ji, Wanying Wu, Shuai Chen, ..., Shaogeng Zhang, Penghui Yang, Fangqing Zhao

Correspondence

zhfq@biols.ac.cn

In Brief

Ji et al. present a large-scale study of circRNA repertoires from multiple tissues of human, macaque, and mouse and propose a new approach to annotate and prioritize functional circRNAs.

Ji et al., 2019, Cell Reports *26*, 3444–3460
March 19, 2019
https://doi.org/10.1016/j.celrep.2019.02.078

CellPress

OPEN ACCESS
CellPress

Cell Reports
Resource

Expanded Expression Landscape and Prioritization of Circular RNAs in Mammals

Peifeng Ji,[1,5] Wanying Wu,[1,2,5] Shuai Chen,[1,2,5] Yi Zheng,[1] Lin Zhou,[1] Jinyang Zhang,[1] Hao Cheng,[1] Jin Yan,[3] Shaogeng Zhang,[3] Penghui Yang,[3] and Fangqing Zhao[1,2,4,6,*]

[1]Computational Genomics Lab, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Beijing 302 Hospital, Beijing 100039, China
[4]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: zhfq@biols.ac.cn
https://doi.org/10.1016/j.celrep.2019.02.078

SUMMARY

Circular RNAs (circRNAs) are emerging as essential regulators of various biological and disease processes. To comprehensively understand the diversity of circRNAs and prioritize their importance, we present a large-scale study of circRNA repertoires from multiple tissues from human, macaque, and mouse. We discovered totals of 104,388, 96,675, and 82,321 circRNAs from the three species, respectively, with an average of 72.6% being successfully assembled into full-length transcripts for each species. Using these full-length circRNAs, we identified thousands of evolutionarily conserved circRNAs that were valuable for functional screening and prioritization. By constructing both species-specific and conserved gene co-expression networks, we inferred circRNA functions on a global scale and prioritized promising functional candidates. To illustrate how well-established prior knowledge facilitates to screen functional candidates, we used the circRNA co-expression networks to prioritize circRNAs that may be involved in liver tumorigenesis and experimentally validated their functions.

INTRODUCTION

Circular RNAs (circRNAs), RNA molecules with both ends covalently linked, have recently emerged as a large class of regulatory RNAs that are ubiquitous in animals. As a heterogeneous class, these circular transcripts may participate in different aspects of biological processes through yet unknown and diverse mechanisms. In addition to the well-studied function of circRNAs as microRNA sponges (Hansen et al., 2013; Memczak et al., 2013), extensive studies have revealed that circRNAs play important roles in gene regulation, development, and carcinogenesis (Chen et al., 2017; Kristensen et al., 2018a). However, our understanding of how circRNAs participate in biological processes is still preliminary.

Advances in deep sequencing are giving rise to the rapid accumulation of large circRNA datasets. For example, Rybak-Wolf et al. (2015) identified a large number of circRNAs in human and mouse brains. Meanwhile, circRNA databases have been established, such as circRNADB (Chen et al., 2016) and CSCD (Xia et al., 2018), to allow the scientific community to explore promising candidates. Nevertheless, existing investigations are far from comprehensive and systematic, because they focus predominantly on recognizing circRNAs from cell lines or a single or small collection of tissues, in particular from specific cancer samples. Therefore, considering that the vast majority of circRNAs are extremely cell type specific and usually transcribed at low levels, the discovery of novel circRNAs is an ongoing and improving process. Moreover, a substantial gap remains between the huge number of identified circRNAs and how to understand their functions. One of the major reasons accounting for this gap is the inability to prioritize functionally important candidates in a high-throughput manner. Many pioneering studies focused on characterizing circRNAs rely mainly on cherry-picking those circRNAs with the most distinct expression patterns in disease samples relative to normal samples. For example, Zheng et al. (2016) revealed that circHIPK3 regulates cell growth by sponging multiple miRNAs, and Hirsch et al. (2017) found that circNPM1 expression is highly associated with acute myeloid leukemia. However, these approaches are low throughput and limited to a small number of candidates or samples. Alternatively, evolutionary conservation is a valuable criterion for functional screening. Two key examples of highly conserved and expressed circRNAs that exhibit developmentally regulated expression are the circRNAs originating from the RIMS2 gene (Rybak-Wolf et al., 2015) and the CDR1 antisense locus (ciRS-7) (Venø et al., 2015). However, because of the lack of full-length circRNAs, sequence conservation comparison is restricted to flanking introns and coding DNA sequences as well as small-scale studies (Rybak-Wolf et al., 2015). Considering the prevalence of circRNA isoforms generated by combinations of internal components within back-splicing junctions (BSJs) (Gao et al., 2016) and the limited number of sequenced species, current cross-species conservation analyses based on partial sequences may not be as fruitful for understanding specific circRNA functions. Thus, the next major

challenge to overcome will lie in the elucidation of circRNA functions on a global scale.

To address these challenges, we used a compendium of 132 RNA sequencing (RNA-seq) libraries to generate comprehensive landscapes of the human, macaque, and mouse transcriptome and expression profiles across mammalian major primary tissues using three types of RNA-seq libraries. These datasets allowed us to discover and quantify circRNAs without bias and to assess their evolutionary conservation. We identified myriad circRNAs associated with the three species and systemically elucidated their diversity and expression patterns. We also identified thousands of evolutionarily conserved circRNAs and constructed co-expression networks to prioritize their functional importance. Furthermore, we illustrated the application of the networks by prioritizing differentially expressed circRNAs from liver cancer datasets and experimentally verified candidates that might be involved in liver tumorigenesis.

RESULTS

Expanded Landscapes of Human, Macaque, and Mouse Transcriptomes

We attempted to capture the spectrum of transcriptional diversity using RNA-seq of three species: human, macaque, and mouse. Seventeen human, 13 macaque, and 14 mouse tissue samples were used as sources to construct the RNA-seq libraries (Figure 1A). Each tissue was sequenced using RNA-seq with the following goals: (1) RiboMinus and RNase R treatment for the comprehensive identification of circRNAs, (2) poly(A) enrichment for accurate mRNA and long noncoding RNA (lncRNA) expression quantification, and (3) RiboMinus treatment for accurate circRNA and mRNA expression comparison. These three libraries were sequenced with paired-end 250 bp (PE250), paired-end 150 bp (PE150), and PE150 reads, respectively, and generated an average 10.9, 7.3, and 25.2 GB data per sample for the three types of libraries, respectively (Figure S1A). Linear transcripts at the threshold defined for expression quantitative trait loci analysis (reads per kilobase of transcript per million mapped reads [RPKM] > 0.1) (Melé et al., 2015) were detected using StringTie (Pertea et al., 2015). For circRNA prediction, four different tools were performed separately on all PE150 samples, including CIRI2 (Gao et al., 2018), DCC (Cheng et al., 2016), MapSplice (Wang et al., 2010), and CircExplorer2 (Zhang et al., 2016). Considering that other than CIRI2, these tools do not work for long reads (>200 bp), we trimmed the PE250 datasets to PE150 and then used these tools to detect circRNAs. circRNAs detected by at least two tools and supported by at least two independent BSJ reads were kept for downstream analyses (Figure S1B). Furthermore, full-length circRNAs were reconstructed on the basis of the PE250 datasets using CIRI-full, which is the only software to effectively assemble full-length circular transcripts (Zheng et al., 2019). We obtained 71,112, 77,812, and 56,769 full-length circRNAs for the three species, respectively. PCR-based validation for 20 randomly selected full-length circRNAs revealed a high level of accuracy of the reconstruction (Figure S1C; Table S1). By integrating these two sets of results, we obtained 104,388, 96,675, and 82,321 circRNAs for human, macaque, and mouse, respectively, outnumbering all existing studies. More important, this study showed, for the first time, high-quality full-length circular transcripts at high-throughput scales. These sequences will greatly facilitate downstream sequence-based analysis, such as conservation estimation and composition detection. We further used strategies to evaluate the robustness of our analyses, including RNase R resistance analysis (Figure S2A) and qPCR-based quantification (Figure S2B). Collectively, these results showed a high level of reliability of the identified circRNAs in this study. These datasets give us an opportunity to comprehensively compare the difference between circular and linear RNA transcripts.

We first explored circRNA alternative splicing (AS) events in the three species. By setting a cutoff of ten BSJ reads for circRNAs, we found that all four types of alternative splicing events could be observed within circRNAs in all three species (Figure 1B). Exon skipping (ES) was the most common alternative splicing type in circRNAs. Alternative 3′-splicing site (A3SS) and alternative 5′-splicing site (A5SS) were also major circular alternative splicing types, both of which occurred in >10% of the circRNAs, consistent with our previous study showing that alternative splicing events not only occur in mRNAs but are also prevalent in circRNAs (Gao et al., 2016). We also measured the alternative splicing events in long intervening noncoding RNAs (lincRNAs) and found that circRNAs exhibited a considerably greater proportion of alternative splicing events than did lincRNAs across all three species. However, lincRNAs are generally longer than circRNAs ($p < 0.01$, Wilcoxon test) (Figure S2C) when comparing full-length lincRNAs (Lizio et al., 2015; You et al., 2017) and circRNAs. After normalizing their exon numbers and lengths (Figure S2D), lincRNAs showed an increased number of isoforms compared with circRNAs (Figure S2E).

Then, the expression patterns of different RNA species were compared. To obtain an unbiased comparison, the expression levels of circRNAs, lincRNAs, and mRNAs using RNA-seq with RiboMinus treatment were normalized. The overall expression levels of circRNAs across all the studied tissues were surveyed. In general, circRNAs were transcribed at low levels, comparable with lincRNAs, but showed decreased expression levels relative to protein-coding transcripts (Figure 1C). Nevertheless, we found hundreds of circRNAs that were considerably more highly expressed than their cognate linear genes (Figure S2F), and certain loci showed a tendency to exclusively form circular transcripts. The low expression levels of circRNAs raise the question as to whether they are actively regulated or result from non-specific transcriptional noise. To test these possibilities, we evaluated the expression variations of expressed RNAs in three tissues (brain, liver, and stomach) across multiple human individuals. For each tissue, transcripts expressed in at least half of the samples were selected, and for each resulting transcript, the coefficient of variance, which represents the variance of expression, was computed. Strikingly, circRNAs exhibited remarkably decreased levels of expression variance relative to lincRNAs but comparable with mRNAs across all compared tissues (Figure 1D), indicating that their expression is actively regulated rather than stochastic. Considering that most circRNAs are derived from exons, we investigated whether the high expression variation of circRNAs originates from their

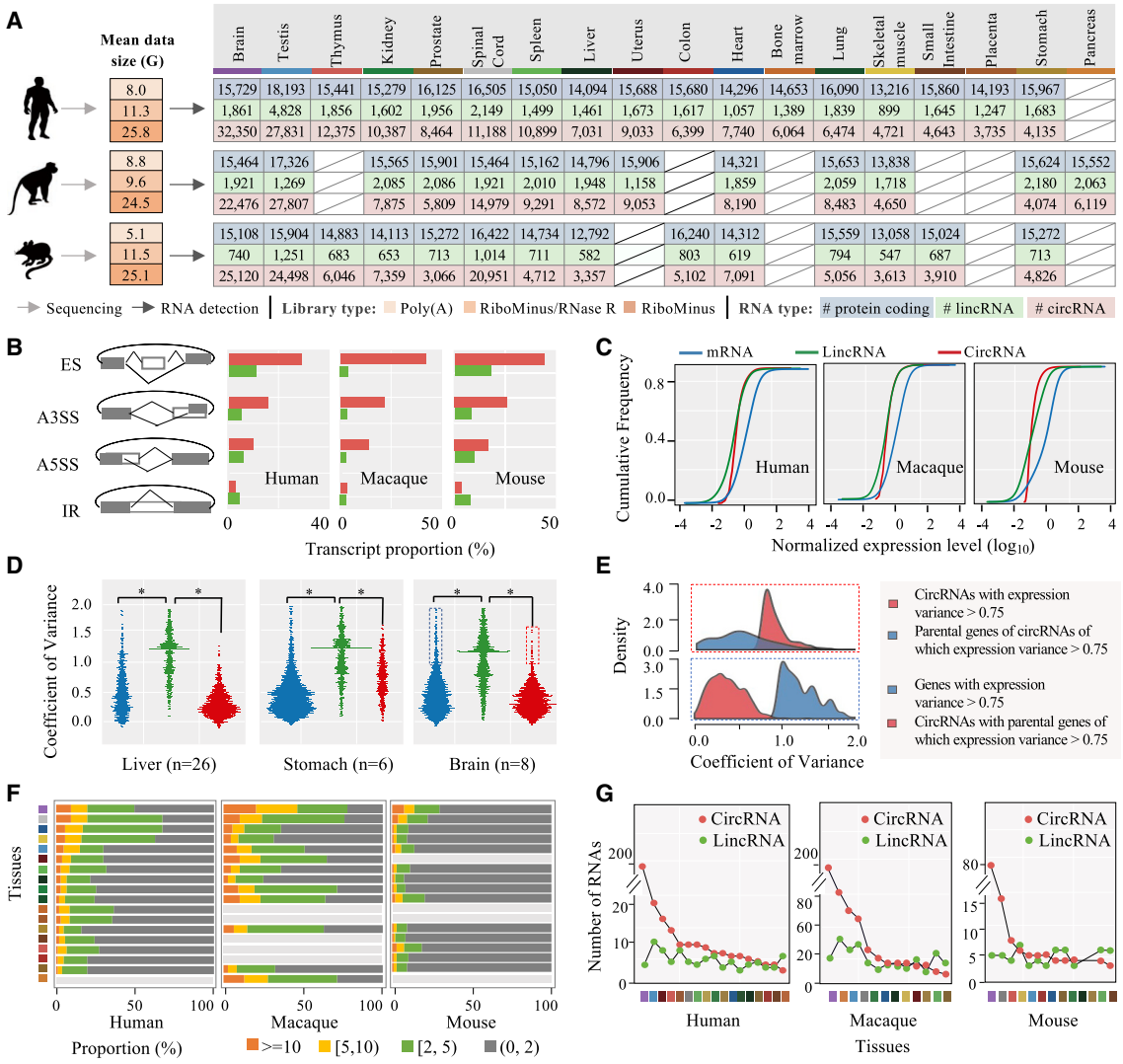


Figure 1. Extended Landscapes of circRNAs in Human, Macaque, and Mouse

(A) The mean sequencing data size (left) and the numbers of expressed genes, circRNAs, and lincRNAs (right) are listed in the table. The left table (in orange) lists the mean sequencing data size of each library type. The colored blocks above the right table represent different tissues. For each species, the first row represents the number of expressed protein-coding genes in a certain tissue, and the second and third rows represent the number of identified lincRNAs and circRNAs, respectively.

(B) Percentage of circRNAs (BSJ $\geq$ 10 reads) and lincRNAs containing four types of alternatively spliced exons in each species: ES (exon skipping), A3SS (alternative 3′-splicing site), A5SS (alternative 5′-splicing site), and IR (intron retention).

(C) Cumulative distributions of normalized expression levels for linear and circular RNAs (mRNA, lincRNA, and circRNA).

(D) Expression variance for the circRNAs (red), lincRNAs (green), and mRNAs (blue) in the three tissues for different human individuals. The "n" in the brackets represents the number of individuals used for the analysis. Each dot represents the expression variance of one transcript shared by multiple individuals (**$p < 0.01$, Wilcoxon test).

(E) Coefficient of expression variance for mRNAs and circRNAs. First, the circRNAs with high expression variance (within the red rectangle in Figure 2B) were extracted. The expression variance of these circRNAs and their parental genes were plotted (top, red rectangle). Second, the mRNAs with high expression variance (within the blue rectangle in D) were extracted. The expression variance of these mRNAs and the circRNAs derived from them were plotted (bottom, blue rectangle). The density plots colored in red and blue represent the circRNAs and mRNAs, respectively.

(F) Percentage of circRNAs is classified into the four groups according to their expression levels (normalized to the number of back-spliced reads per million mapped reads).

(G) Number of circRNAs (red) and lincRNAs (green) in the top 1,000 highest expressed transcripts in each tissue from the three species. Colored blocks correspond to the tissues in (Figure 1A).

See also Figures S1 and S2 and Table S1.

parental genes. We extracted circRNAs with highly variable expression levels across different human brain samples and compared their expression variance with their parental genes. Only a weak concordance between their variation distribution was observed (Figure 1E), indicating that the expression of circRNAs is largely independent of that of their parental transcripts.

Finally, we investigated the landscapes of circRNA expression using the RNA-seq libraries with RiboMinus treatment, which provides an unbiased quantification of both linear and circRNAs. By normalizing the expression of transcripts using the total sequencing depth, we classified circRNAs into four catalogs according to their abundance and subsequently compared the profiles of every catalog in each tissue. In principle, highly expressed circRNAs preferred neural tissues (e.g., brain and spinal cord) to other tissues, consistent with previous studies (Hansen et al., 2013; Memczak et al., 2013) (Figure 1F). Moreover, the proportion of highly expressed circRNAs from human and macaque was higher than in mouse. These observations are distinct from those for protein-coding genes and lincRNAs. Specifically, the proportion of abundant mRNAs was nearly constant across all tissues. However, abundant lincRNAs were enriched in the testis (Figure S2G). We further examined the most abundant RNAs in each tissue from the three species. The top 1,000 most abundant RNA transcripts in each tissue from these species were selected. As expected, the majority of the abundant RNAs were mRNAs. Regarding the remaining RNAs (Figure 1G), the number of circRNAs outnumbered that of lincRNAs in most tissues, especially in the brain, where circRNAs accounted for approximately 20% of the selected RNAs in human and macaque.

Most Protein Coding Genes Express One Dominant circRNA

Profiting from the obtained comprehensive landscapes of the transcriptomes, we investigated the number of genes expressing circRNAs in each species and found that more than half of the detected genes could express circRNAs (Figure 2A). Taking human as an example, as many as 61.0% of the genes were found to yield both linear and circular transcripts, and these genes tended to generate many more linear isoforms than those only expressing linear transcripts (Figure 2B). This difference may be because the genes that express only linear transcript have low exon number (on average 24.0) compared with the genes expressing both linear and circular transcripts (Liang et al., 2017) (on average 73.9). Moreover, circRNAs were far more prevalent than linear transcripts (Figures 2B and S3A). A striking example is the human BIRC6 gene, which could transcribe into 243 circular transcripts, whereas only 12 linear transcripts were recorded in the Ensembl database. To validate these findings, we further used the shared circRNAs between CIRI and DCC (Cheng et al., 2016) from macaque and mouse and obtained similar results (Figure S3B).

It has been proved that in a given condition, most protein-coding genes have one major transcript expressed at a significantly higher level than others (Gonzàlez-Porta et al., 2013). We therefore investigated whether the same scenario holds true for circRNAs. To this end, genes expressing at least two circRNAs were selected (67% of the total genes). The first two most abundant circular transcripts on a gene were termed the major and minor transcripts, respectively. If the major transcript was expressed at a remarkably (2-fold) higher level than the minor transcript, the major transcript was then referred to as the dominant transcript. We plotted the expression levels for the major and minor transcripts in the brain and observed the evident existence of dominant transcripts (Figure 2C). We then quantified the transcript dominance by calculating the ratio of the expression levels between the dominant circRNAs and the minor circRNAs. Overall, 23% of the genes exhibited a 2-fold dominant circRNA (i.e., expressed twice as high as that of the minor circRNA), and for 8% of the genes, the dominant circRNA was 5-fold dominant (Figure 2D).

We next sought to verify whether the existence of the dominant circRNA is a universal characteristic of the transcriptional signature. By the same measurement, we found that this was the case with the remaining tissues in human (Figure S3C). This finding further raised the question as to how often the same dominant circRNA can be detected for a gene across different tissues. We found that the dominant circRNAs were often the same major circRNAs for 18% of the genes that were expressed in at least two tissues. In addition to being highly expressed, the dominant circRNAs exhibited decreased levels of tissue specificity (Figure 2E) and increased junction ratio, which was defined as the ratio of BSJ reads and total number of reads aligned to the junction site (Figure 2F), suggesting that they may play more essential roles than the minor circRNAs. Finally, we examined whether the dominant circular transcripts were simply byproducts of the dominant linear transcripts. The patterns of overlap between dominant circular and linear transcripts could be classified into the following three categories according to shared exons: (1) overlapped: both of them use the same subset of exons; (2) partially overlapped: some but not all of the exons are shared; and (3) none overlapped: without shared exons (Figure 2G). Notably, the percentage of overlapped patterns was approximately 50%, while that of partially or none overlapped patterns was approximately 50% (Figure 2H), indicating that the formation of these dominant circular and linear transcripts is likely to be independently regulated. For further validation, we calculated the Pearson correlation coefficient between dominant circular and linear transcripts using their expression profiles. As shown in Figure 2I, the expression pattern correlation, ranging from −0.2 to 0.4, showed no relatedness between these two types of transcripts. These results collectively suggest that the dominant circular transcripts are not simply alternative splicing byproducts of the dominant linear transcripts but may have independent biological mechanisms accounting for their formation.

RNA Binding Proteins Are Key Regulators of Tissue-Preferential Expression Pattern in circRNAs

We studied the tissue specificity of circRNAs using human, macaque, and mouse tissues. As controls, both lincRNAs and mRNAs were also surveyed. The distribution of mRNA expression across tissues was U-shaped, whereas those of circRNAs and lincRNAs were typically L-shaped (Melé et al., 2015) (Figure 3A, top), suggesting increased levels of tissue specificity for noncoding transcripts compared with mRNAs. Indeed, when

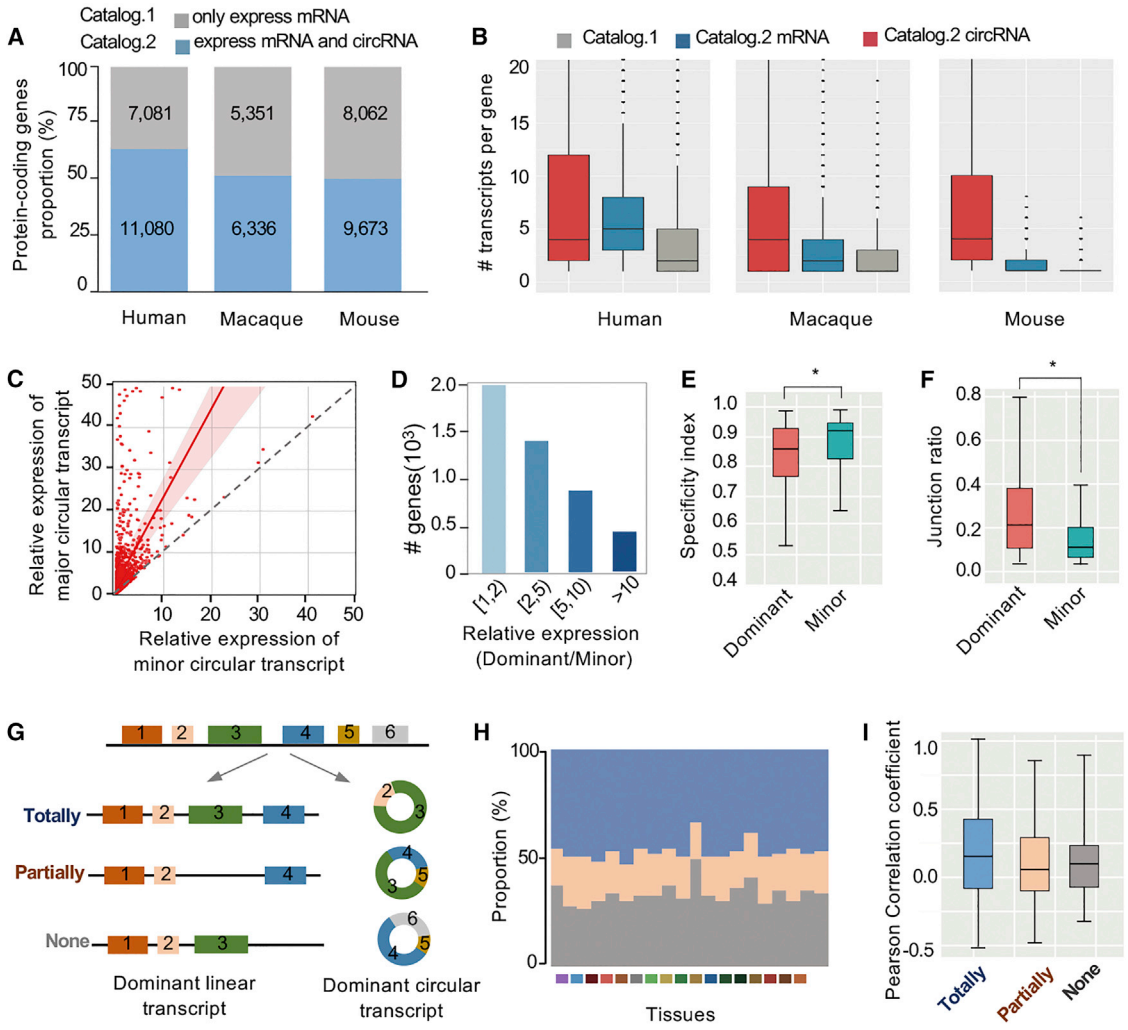


Figure 2. Characteristics of circRNA Expression Patterns

(A) Proportion of two catalogs of protein-coding genes. The protein-coding genes are classified into two catalogs, where the first catalog only expresses mRNA (gray) and the second one expresses both circular and linear transcripts (blue).

(B) Comparison of transcript number per gene between the two types of protein-coding genes.

(C) Relative expression levels of the major and minor circular transcripts. Specifically, the first two most abundant circular transcripts on a gene that can express more than two circRNAs are termed as the major and minor transcripts, respectively. Each dot represents a gene. The red solid line is the regression line that visualizes a linear relationship as determined through regression.

(D) Number of genes explained by each category for the expression ratio of the major to minor transcript.

(E) Comparison of tissue specificity between dominant and minor transcripts.

(F) Comparison of the junction ratio between the dominant and minor transcripts.

(G) Examination of whether the major circular transcripts are derived from the same major linear transcripts. The patterns of overlap between dominant circular and linear transcripts are classified into the following three categories according to shared exons: (1) totally overlapped: both of them use the same subset of exons; (2) partially overlapped: some but not all of the exons are shared; and (3) none overlapped: without shared exons.

(H) Percentage of genes containing dominant circular and linear transcripts explained by each category noted above.

(I) Expression correlation between dominant linear transcripts and circular transcripts in each category.

See also Figure S3.

using the tissue specificity index (Yanai et al., 2005) to quantify how strongly the expression is dominated by a single tissue, the difference between noncoding RNAs and protein-coding transcripts was significant. This scenario held true when the BSJ read cutoff for circRNA was set to a value of 10 (Figure S4A). Both circRNAs and lincRNAs were expressed in a more tissue-specific manner than protein-coding genes ($p < 0.01$, Wilcoxon test) (Figure 3A, bottom). In particular, the vast majority of circRNAs (64%) were highly tissue specific (tissue specificity index = 1.0) relative to only ~15% of protein-coding genes. The discrepancy between this type of noncoding RNAs and protein-coding genes held true for the compared transcripts

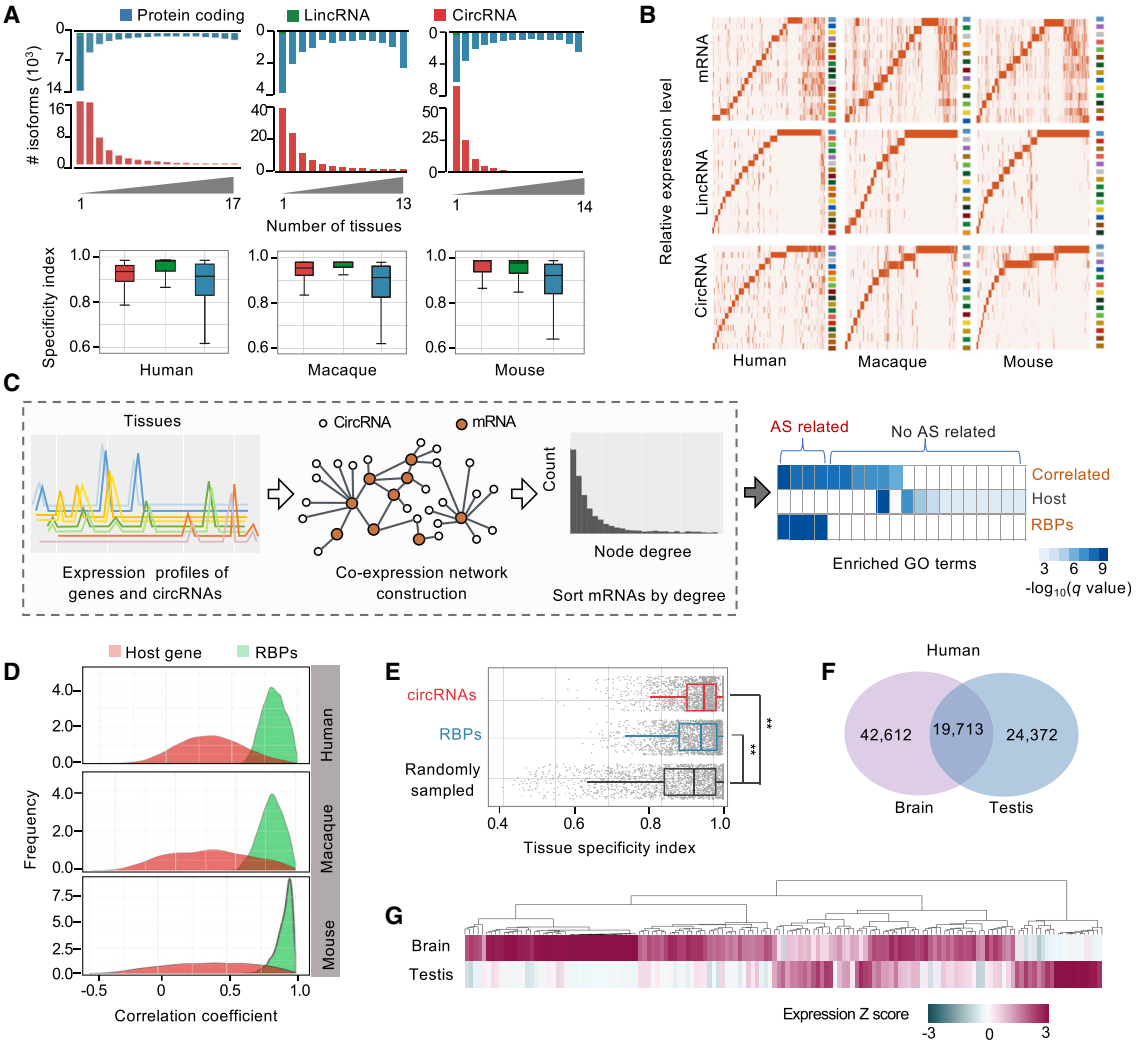


Figure 3. Tissue Specificity of circRNAs

(A) Number of tissues in which genes are expressed (top) and the tissue specificity index (bottom) of the three types of transcripts. Values close to 1 represent high tissue specificity.

(B) Abundance of the three types of transcripts (columns) across tissues (rows) for each species. Rows and columns are ordered on the basis of a k-means clustering algorithm. Color intensity represents fractional density across the columns of normalized sequencing depth.

(C) circRNA-mRNA co-expression network construction and functional enrichment. First, the Pearson correlation coefficient was computed for each circRNA-mRNA pair on the basis of the expression profiles. Then, the circRNA-mRNA pairs with Pearson correlation coefficients > 0.75 were screened out to construct the co-expression network, where each node denotes a circRNA or mRNA and the edges represent the correlation between the circRNAs and mRNAs. Finally, Gene Ontology (GO) functional enrichment analysis was performed using the top 5% of mRNAs with the highest degree (right). As controls, circRNA and RBP host genes were also simultaneously analyzed.

(D) Pearson correlation coefficient between circRNAs and RBPs and between circRNAs and their host genes.

(E) Tissue specificity of circRNAs and RBPs. Each node denotes one transcript. Both circRNAs and RBPs show significantly increased tissue specificity compared with other protein-coding genes (**$p < 0.01$, Wilcoxon test).

(F) Shared circRNAs between brain and testis.

(G) Expression profiles of different RBPs (columns) in brain and testis (rows).

See also Figure S4.

expressed at similar levels (Figure S4B). We next applied a focused analysis on the dominant tissues of circRNAs. As shown in Figure 3B, consistent with previous studies (Rybak-Wolf et al., 2015; You et al., 2015), the vast majority of circRNAs were exclusive to neural tissues, especially the brain. Both tissues showed well-conserved tissue specificity across species, suggesting that these circRNA expression patterns are not stochastic but instead are selectively maintained.

As the most striking characteristic of circRNAs, the extremely tissue-specific expression of circRNAs raises the following question: what is the key determinant of this tissue-preferential expression pattern? Considering that RNA binding proteins

(RBPs), such as QKI (Conn et al., 2015), ADAR (Ivanov et al., 2015), and DHX9 (Aktaş et al., 2017), have been cumulatively proved to regulate the formation of circRNAs and are involved in all aspects of post-transcriptional processes (Pereira et al., 2017), we speculated that these regulators may also be the primary factors that contribute to the tissue-specific expression of circRNAs. To test this possibility, we first examined whether circRNAs are more frequently associated with RBPs than other genes. A co-expression network between circRNAs and mRNAs was constructed using their expression profiles. The top 5% of the most frequently connected mRNAs (hub genes) were used for Gene Ontology (GO) enrichment analysis. Consistent with the GO enrichment result with the RBPs (Figure 3C), circRNAs were more frequently associated with mRNAs that were predominantly enriched for alternative splicing-related functions, such as RNA and mRNA splicing. Conversely, their host genes were found to be involved in other non-alternative splicing-related functions (Figure S4C). Moreover, the RBP binding sites were enriched in the flanking introns of circRNA compared with linear mRNA ($p = 9.217e-07$, Mann-Whitney U test) (Figure S3E). Additionally, the expression patterns of the circRNAs showed strong associations with those of RBPs, with an average Pearson correlation coefficient of 0.9. However, they were poorly correlated with their linear counterparts (Kristensen et al., 2018b) (Figure 3D), which was further recapitulated using time-series datasets (Rabani et al., 2014) (Figures S3E–S3G). RBPs exhibited a highly tissue-specific expression pattern, which was comparable with that of circRNAs (Figure 3E). Considering that the RBP proteins exhibited a similar expression profile compared with their genes, which had been verified using the data from the HUMAN PROTEOME MAP database (Kim et al., 2014), the RBP proteins should have strong associations with circRNAs in terms of expression profile (Figure S3H). Similar observations were also found when the same analysis was extended to the GTEx datasets (Figure S3I).

Taken together, these results suggest that RBPs are the key regulators of the tissue-preferential expression pattern of circRNAs. Taking the two most circRNA-enriched tissues (brain and testis) as examples, the RBPs in these two tissues were quite distinct from one another in terms of their expression profiles (Figure 3F), corresponding to only a small fraction of shared circRNAs between the two tissues (Figure 3G).

Orthologous circRNAs Exhibit Highly Conserved Expression and Splicing Patterns across Species

Evolutionary analysis is essential for insights into the genetic basis of phenotypes and for functional screening. For circRNAs, such analysis remains scarce despite growing attention to these circular transcripts. Thus, we sought to identify evolutionarily conserved circRNAs on the basis of orthologous genes. Specifically, orthologous genes were identified and the circRNAs on these orthologous genes were subsequently screened. The resulting circRNAs with the same BSJ sites between any two species were defined as overlapped orthologs (OO-type), with the remaining circRNAs termed non-overlapped orthologs (NO-type) (Figure 4A). We first estimated the presence of shared OO-type circRNAs and orthologous genes across species. We found that circRNAs evolve rapidly; approximately 19.1% of human circRNAs were also detected as expressed in macaque, and only approximately 4.4% were expressed in mouse, whereas more than 67% of conservation was observed in protein-coding genes. We also determined how well the OO-type circRNAs were conserved across tissues among species and found that approximate 4.8% of them were recurrently expressed in a certain tissue between any two different species (Figures S5A and S5B). Then, we compared the exon boundary conservation of the OO-type circular and linear transcripts between human and macaque. Considering that most of the cirexons in circRNAs are identical to those in linear mRNAs, only the intergenic/intronic circRNA fragments (ICFs) (Gao et al., 2015) that were exclusively present in circRNAs were used for the analysis. The boundary conservation level of ICFs was similar to that of protein-coding exons. Consequently, compared with lncRNAs, the OO-type circRNAs exhibited more constraint with respect to maintaining an exact splicing event position among orthologous pairs (Figure S5C). Next, the expression levels of conserved circRNAs were characterized. The OO-type circRNAs exhibited increased expression levels compared with species-specific circRNAs, whereas this scenario was not observed in the shared orthologous genes from which the circRNAs were derived. In summary, these findings suggest that there is a distinct evolutionary conservation pattern in orthologous circRNAs and that these shared circRNAs may undertake essential biological functions.

We next narrowed down to the OO-type circRNAs (2,772) that were conserved across the three species. We first characterized the expression patterns of the OO-type circRNAs and their linear counterparts. As shown in Figure 4B, the orthologous genes exhibited remarkably increased levels of concordant expression patterns among species compared with the OO-type circRNAs. That is, the expression pattern of ortholog genes across different tissues was more consistent than that of the OO-type circRNAs. The expression patterns of the OO-type circRNAs were more concordant in closely related species (e.g., human and macaque). Moreover, a number of OO-type circRNAs were expressed much higher than their host genes across all the studied tissues, suggesting their functional importance. Additionally, OO-type circRNAs exhibited remarkably increased expression levels (Figure 4C), reduced levels of tissue specificity (Figure 4D), decreased levels of expression divergence (Figures 4E and S5D), and an increased number of repetitive elements in the flanking introns (Figure S5E) compared with NO-type circRNAs across the surveyed species. We next investigated whether OO-type circRNAs in different species share similar alternative splicing patterns. To this end, we calculated the values of "percentage spliced in" (Ψ) for alternative splicing events in human OO-type circRNAs and compared these values with the Ψ values of their circular orthologs in macaque (Figure 4F). With few exceptions, the Ψ values for all the OO-type circRNA pairs were highly concordant between human and macaque ($p < 2.2e-16$; Figure 4G), indicating that the conservation of OO-type circRNAs lies not only in their expression levels but also in their internal splicing patterns.

To compare the expression levels of OO-type circRNAs and their corresponding mRNAs, we used RiboMinus transcriptomic data from brain samples. Compared with the remaining

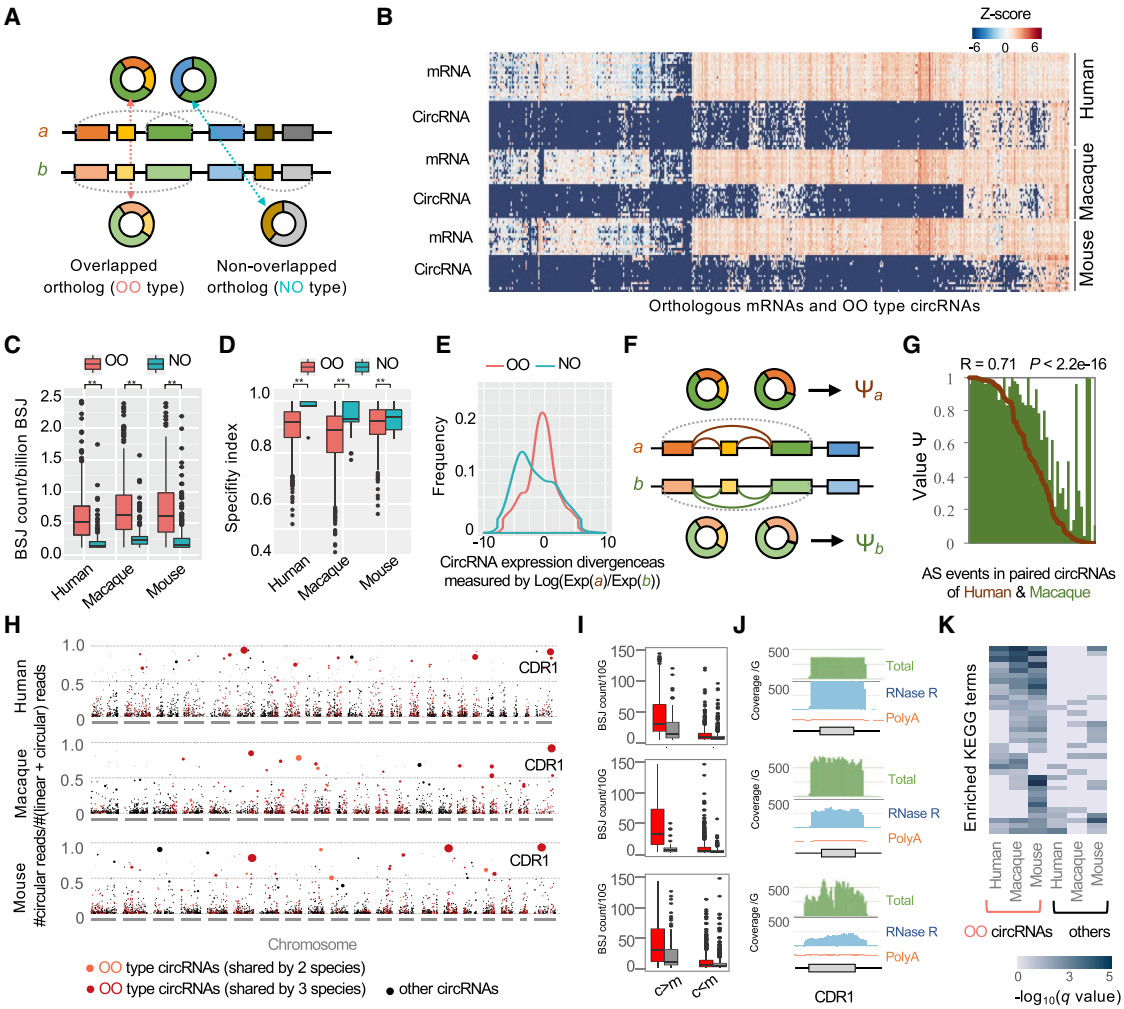


Figure 4. Cross-Species Conservation of Overlapped Orthologous circRNAs

(A) Overlapped orthologous (OO-type) circRNAs are derived from orthologous genes between species a and b and share the same BSJs. Non-overlapped orthologous (NO-type) circRNAs are derived from orthologous genes from species a and b but do not share the same BSJs.

(B) Expression of orthologous genes and OO-type circRNAs that are shared by the three species across tissues. Rows denote tissues, and columns represent orthologous genes as well as the OO-type circRNAs derived from them.

(C) Comparison of the expression levels of highly expressed OO-type circRNAs (top 10% of the most abundant circRNAs) and highly expressed NO-type circRNAs (top 10% of the most abundant circRNAs). Expression levels are normalized to the total BSJ read count.

(D) Tissue specificity of OO- and NO-type circRNAs in the three species.

(E) Expression divergence between OO- and NO-type circRNAs in related species. Expression divergence is measured as $\log_2$(expression level of species a)/expression level of species b). Expression levels are normalized to the total BSJ read count.

(F) Detection of splicing events in OO-type circRNAs between a pair of species (a and b).

(G) Comparison of Ψ values in OO-type circRNAs between human and macaque.

(H) Relative expression levels of circRNAs in brain transcriptomes of the three species. Each circle denotes a circRNA, with the circle size representing its normalized expression level. Red and black circles represent OO-type and other circRNAs, respectively. The dashed lines along the x axis represent the chromosomes from each species. The y axis represents the relative expression level of a circRNA and its corresponding mRNA as measured by the number of circular reads divided by the total read count in the same BSJ. The circles above the dotted line at Y = 0.5 indicate circRNAs whose expression levels are higher than those of their corresponding mRNAs.

(I) Comparison of the expression levels of OO-type circRNAs (red) and other circRNAs (gray). "c > m" represents circRNA loci at which the circRNA expression level is higher than the mRNA expression level, whereas "c < m" represents circRNA loci at which the circRNA expression level is lower than the mRNA expression level. In both categories, OO-type circRNAs exhibited elevated expression levels.

(J) Sequencing depth of the ciRS-7 locus in the CDR1 gene as measured in three independent RNA-seq datasets: RiboMinus, RiboMinus+RNase R, and PolyA-selected mRNA-seq.

(K) Functional enrichment analysis of OO-type circRNAs and other circRNAs in the brain transcriptomes of the three species.

See also Figure S5.

circRNAs, OO-type circRNAs exhibited significantly elevated expression levels ($p < 0.01$, Wilcoxon test; Figure 4H). We classified these circRNAs into the following two classes, $c > m$ (circRNA expression level higher than mRNA expression level) and $c < m$ (circRNA expression level lower than mRNA expression level), according to the relative abundance to their corresponding linear transcripts. As shown in Figure 4I, for both classes, the OO-type circRNAs exhibited substantially elevated expression levels relative to other circRNAs. For example, ciRS-7 was one of the most abundant circRNA in all the brain samples analyzed (Figure 4J). However, consistent with a previous study (Piwecka et al., 2017), its parental gene, CDR1, did not produce linear transcripts at this locus, as revealed by poly(A)-selected mRNA sequencing (Figure 4J). Finally, we performed functional enrichment analysis of the OO-type and other circRNAs in each species. As shown in Figure 4K, the enriched functions of the OO-type circRNAs were highly overlapped. Most of these circRNAs appear to be involved in neural processes and functions, including the MAPK signaling pathway and glutamatergic synapse and morphine addiction activities (Figures 4K and S5F). In contrast, there was no significant functional enrichment for other circRNAs. Collectively, we conclude that OO-type circRNAs are highly conserved across species with respect to both expression and splicing patterns. Finally, using CIRI-full, we identified a total of 37,591 (55.4% of 67,907) full-length OO-type circRNAs from the three species, providing the most comprehensive catalog of evolutionarily conserved circRNAs thus far, which will greatly facilitate the functional analysis of circRNAs.

Inference of circRNA Function Using a Co-expression Network

With a few exceptions, the function of most circRNAs remains elusive. A co-expression network presents a valuable platform for functional annotation and screening. To predict functional circRNAs and their potential regulatory mechanisms, we inferred co-expression networks independently for each of the three species, where nodes represented genes (circRNAs or mRNAs) and edges represented a co-expression relationship between gene pairs. This network consisted of 33,512,072 co-expression relationships from 38,532 nodes. As shown in Figure 5A, the correlations of nodes recapitulated tissue types (i.e., the nodes and edges on the network exhibited tissue-specific distributions). For example, the brain- and spinal cord-specific genes were clustered together, while the heart-specific genes were tightly connected.

We first aimed to verify the topological signature of the network in terms of the node degree, where a higher degree indicates that the node is likely to be a hub and participate in more interactions, and the betweenness centrality (BC), where a node with a higher BC indicates that it acts as a bridge to connect different components and control communication. An examination of the node degree on the network fit a power-law distribution ($p = 0.00049$) (Figure S6A), indicating that the network was similar to many other biological and scale-free networks, where high connectivity genes are few in number and most genes have low connectivity. Moreover, circRNAs showed significantly increased levels of BC compared with mRNAs ($p = 0.00029$, Wilcoxon test) (Figure S6B). For example, 42 of the top 50 largest BC nodes were circRNAs and exhibited increased levels of specific topological characteristics relative to mRNAs. To further validate the predictions provided by the network, we took advantage of gene knockdown experiments. Briefly, we knocked down three RBPs—TRA2B, MBNL1, and PTBP1—in HeLa cells. The CIRI2 pipeline was applied to the resulting knockdown transcriptomic datasets, and the expression levels of circRNAs were normalized. We used the proportion of shared circRNAs between the knockdown experiments and the network prediction to assess whether the direction of the alternations in the circRNAs after gene knockdown experiments could be predicted by the network (Figure 5B). To increase the reliability of the analyses, two publicly available gene knockdown datasets, DHX9 and QKI, were also analyzed using the same approach. We found high levels of concordance between the correlation predictions and the real alterations (Figure 5B; Tables S2–S6). Closer inspection revealed that only a few concordant circRNAs were shared among the RBPs (Figure 5C), suggesting that circRNAs tend to be exclusively regulated by specific RBPs. Intriguingly, these RBPs also exhibited distinct signatures in terms of the regulated circRNAs (Figure 5D). For example, the circRNAs regulated by TRA2B showed the highest level of BCs and expression and the lowest level of tissue specificity compared with other RBPs. It should be noted that the function of RBPs was reflected consistently by the direction of the expression changes after knockdown (Figure 5B, pie plots). DHX9, an abundant nuclear RNA helicase that prevents the formation of circRNAs (Aktaş et al., 2017), served as an example, and the expression levels of nearly all regulated RNAs (>99%) increased after its knockdown. In contrast, with the knockdown of QKI, an alternative splicing factor that promotes the formation of circRNAs (Conn et al., 2015), decreased expression levels were observed for nearly all the correlated circRNAs. Therefore, we further explored the entire connection signature of the network. As shown in Figure 5E, the vast majority (an average of 93%) of the network connections were positive correlations, and they occurred more frequently between protein-coding genes or between circRNAs. Connections between circRNAs and protein-coding genes, however, tended to be negative (14%). This may be because many protein-coding genes, such as RBPs that connect with circRNAs more frequently on the network, function as repressors of circRNA formation. Indeed, we found that 32% of the RBPs showed a negative correlation tendency. DHX9 and ADAR served as examples, consistent with previous studies, and an overwhelming majority of negative connections were observed. Additionally, the vast majority (>90%) of connections between circRNAs and HNRNPL, QKI and FUS, which can promote the circularization of exons by binding to flanking intronic sequences, were positive, which is in agreement with previous reports (Conn et al., 2015; Fei et al., 2017), revealing the reconstruction specificity of the network.

To further explore the detailed interplay between circRNAs and mRNAs, we emphasized the recently characterized m6A modification process (Zhou et al., 2017) (Figure 5F). Seventy percent of the experimentally identified m6A circRNAs were detected from circRNAs that are directly connected with the above

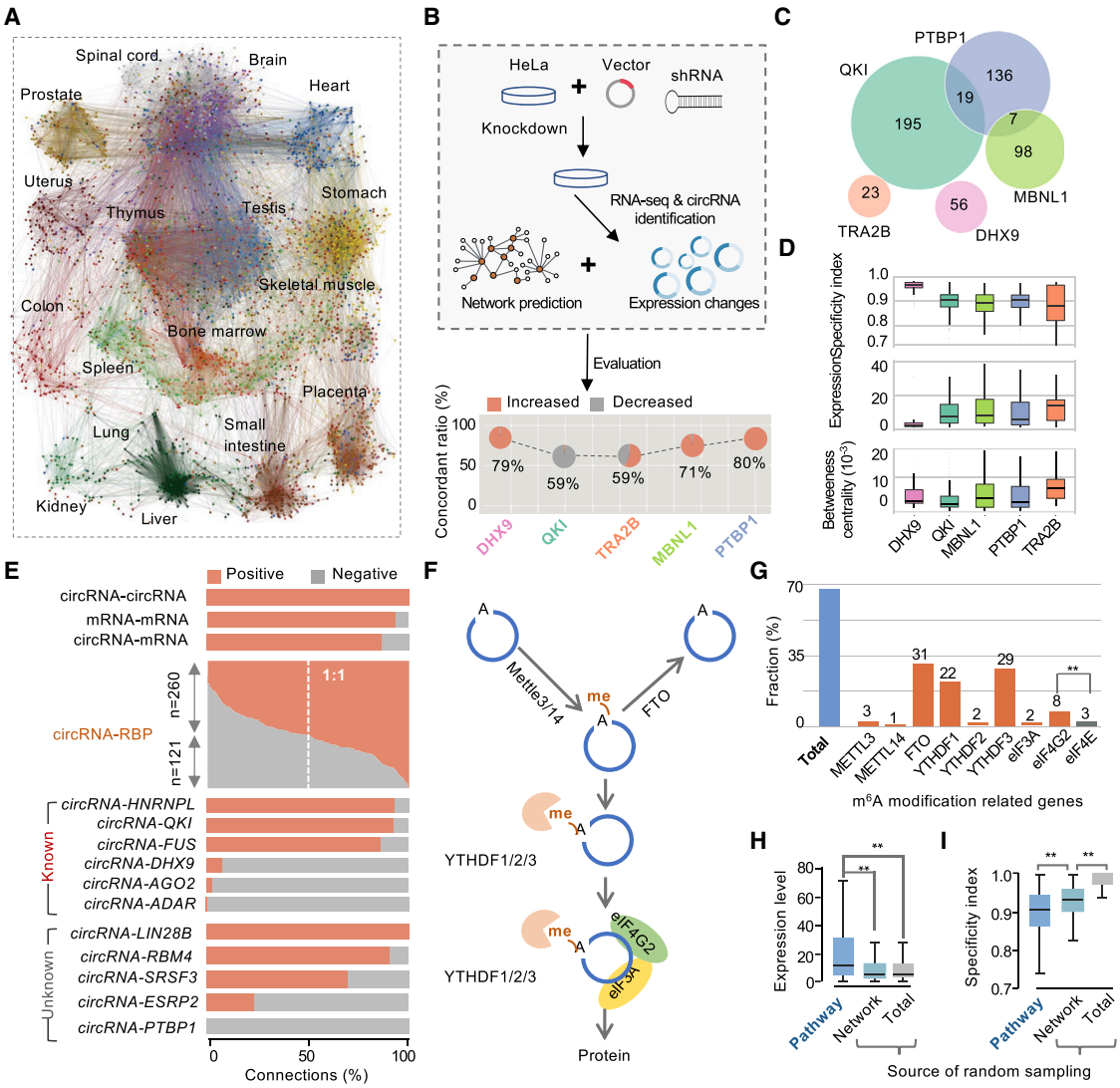


Figure 5. Co-expression Network of Human circRNAs and mRNAs

(A) Visualization of the Pearson correlation network of circRNAs and mRNAs from human tissues. Each node (dot) in the network represents a circRNA or mRNA, and the edges (lines) correspond to correlations between individual measurements above the defined threshold. Co-expressed nodes form closely connected complex clusters within the graph. Nodes are colored according to their specific tissues.

(B) Network validation using shRNA knockdown experiments. Briefly, shRNAs targeting TRA2B, MBNL1, and PTBP1 were cloned into vectors. The shRNA-expressing plasmids were co-transfected into 293T cells. Viral supernatants were collected and then used to transduce HeLa cells. After knocking down each RBP, we calculated the fold change of each circRNA and extracted the circRNAs with fold-change > 2. On the basis of the resulting circRNAs, we found that percentage of circRNAs that the expression alternation can be correctly predicted by the co-expression network. The pie plot shows the ratio of concordant predictions to experimental observations. The orange color in the pie indicates the circRNA expression levels increased after gene knockdown, whereas the gray color denotes that those of the circRNA decreased.

(C) The Venn plot shows the number of shared circRNAs with expression fold change > 2 after knocking down the RBPs.

(D) The tissue specificity, expression levels in the brain and betweenness centrality of RBP-regulated circRNAs.

(E) Percentage of positive connections for the entire network, for all human RBPs, for functionally known RBPs, and for functionally unknown RBPs. circRNA-circRNA, connections between two circRNAs; mRNA-mRNA, connections between two mRNAs; circRNA-mRNA, connection between a circRNA and an mRNA. circRNA-RBPs, connection between a circRNA and an RBP, where "n" denotes the number of RBPs (rows), and the white line on this panel indicates that the ratio of positive connections to negative connections equals 1. "n = 260" denotes the number of RBPs containing more positive connections (ratio > 1:1). "n = 121" represents the number of RBPs with fewer positive connections (ratio < 1:1).

(F) A schematic diagram of circRNA translation driven by m6A.

(G) Proportion of m6A circRNAs in previous studies that were identified in the network (blue bar) and that were connected by m6A modification-related genes (orange bar). Comparison of the percentage of correlated circRNAs between eIF4G2 and eIF4E (**$p < 0.01$, binomial test).

(H) Comparison of expression levels between m6A-related circRNAs and randomly sampled circRNAs from the network or total circRNAs (**$p < 0.01$, Wilcoxon test).

(I) Comparisons of tissue specificity between m6A-related circRNAs and randomly sampled circRNAs from the network or total circRNAs (**$p < 0.01$, Wilcoxon test).

See also Figure S6 and Tables S2–S6.

genes on the network. Moreover, eIF4G2 (Yang et al., 2017b), a cap-independent translation factor that leads to translation initiation in the absence of eIF4E, connected with more m^6A circRNAs than eIF4E (Figure 5G). The circRNAs correlated with these genes also displayed expression patterns distinct from those of other randomly sampled circRNAs. Specifically, these m6A circRNAs showed increased expression levels (Figure 5H) and decreased tissue specificity (Figure 5I) compared with other circRNAs.

Evolutionarily Conserved Co-expression Network

Evolutionary conservation is a powerful criterion to identify functionally essential genes from a set of co-regulated genes. The co-regulation of gene pairs over a large evolutionary distance implies that the co-regulation confers a selective advantage, most likely because the genes are functionally related (Stuart et al., 2003). Consequently, we reconstructed an evolutionarily conserved co-expression network for circRNAs and protein-coding genes across the three species. Briefly, for each species and for each pair of orthologs (circRNAs or mRNAs), we computed the Pearson correlation coefficients of their expression patterns (Figure 6A). Subsequently, the combination of each orthologous gene pair was ranked according to its correlation score. We defined an orthologous gene cluster as a set of orthologs (circRNAs or mRNAs) across the three organisms and assigned each gene to at most a single orthologous gene cluster. For example, a cluster referred to the human gene STK14B, the macaque gene p70S6Kb and the mouse gene Rps6kb2, all of which encode a ribosomal protein S6 kinase (Gout et al., 1998). Given two orthologous gene clusters, if the combination of correlation coefficients measured in each species was significantly higher or lower than expected by chance, they were considered evolutionarily relevant. Hence, these two orthologous gene clusters were considered co-expressed. On the basis of this approach, we combined all the links between pairs of co-expressed orthologous gene clusters to build an evolutionarily conserved co-expression network. In total, we calculated the co-expression relationships for 13,356 orthologous genes and 2,772 OO-type circRNAs that were conserved across the three species. The resulting co-expression relationships formed a network with 12,119 nodes (2,537 circRNAs and 9,582 protein-coding genes) and 435,025 edges. Network connectivity relies mainly on expression levels, as more connections are detected for highly expressed nodes (Necsulea et al., 2014). We therefore investigated whether this holds true for this network. Indeed, the connections of both circRNAs and mRNAs increased by increasing the maximum expression level (Figures 6C and S6C). Furthermore, circRNAs generally had higher connectivity (mean degree 98.8) than protein-coding genes (mean degree 64.6) ($p = 0.00033$, Wilcoxon test). The highly connected circRNAs may represent interesting candidates for further studies on gene expression regulation.

We used the constructed co-expression network to infer potential functions for circRNAs by using a "guilt-by-association" analysis. We identified 72 tightly intra-connected clusters with at least 20 nodes, with an average circRNA proportion of 18%. The largest 15 clusters are illustrated in Figure 6A. GO enrichment was performed for each cluster using the biological process terms. As shown in Figure 6B, the clusters were enriched for tissue-specific functions, such as modulation of synaptic transmission (cluster 1) and muscle cell development (cluster 3). Cluster 1 had the highest circRNA proportion and was enriched for synaptic transmission, consistent with the predominant circRNA neural specificity. GO enrichment analyses for individual clusters suggested that circRNAs may be involved in diverse biological functions, such as adaptive immunity, blood coagulation, and neurotransmitter transport. Using this approach, we provided functional implications for thousands of circRNAs that previously had no meaningful annotation.

On the basis of the annotations, we checked whether the annotation of the OO-type circRNAs was the same as that of their host genes. Strikingly, we found that most circRNAs (65%) and their host genes did not have the same GO annotations (Figure 6D). The network edges not only produce functional clusters but also facilitate a reliable capture of intra-cluster functional differences (Figure 6E). For example, all three circRNAs derived from the human RIMS2 gene were annotated with GO term GO:0050804 (modulation of synaptic transmission). We then performed functional enrichment for the uniquely connected mRNAs of these three circRNAs and discovered that they were enriched for different GO terms, including GO:0050890 (cognition), GO:0006813 (potassium ion transport), and GO:0001505 (regulation of neurotransmitter levels). The increased resolution by adding more transcriptomic data will facilitate more accurate circRNA annotations and extend the applications for the conserved co-expression network.

To this end, we integrated the identified circRNAs with their expression patterns, genomic features, conservations, and functional annotations into a web server called the circRNA Atlas (circatlas.biols.ac.cn) (Figure 6F). Users can browse, search, retrieve, visualize, and prioritize circRNAs and their related information. We believe that this resource will help the circRNA community to annotate and prioritize circRNAs.

Prioritization of Liver Cancer-Related circRNAs in the Context of the Network

Although RNA-seq has emerged as the choice for interrogating the transcriptome, most analysis methods do not consider prior knowledge of biological networks to detect differentially expressed genes, which may result in many irrelevant significant genes (Lei et al., 2017). We investigated the circRNA expression profiles in the 40 RNA-seq datasets of hepatocellular carcinoma (HCC) tumor tissues and their adjacent normal tissues (Yang et al., 2017a). First, expressed circRNAs in these samples were detected and ranked by their differential expression significance (Figure 7A, leftmost column). To prioritize these candidates in the context of co-expression network, a random walk algorithm (Köhler et al., 2008) was used to assign a score to each candidate circRNA, which measures the relative network distance of the circRNA to all known liver cancer associated genes. The candidate circRNAs were then re-ranked according to their assigned scores and conservation among species (Figure 7A). Next, we experimentally validate these circRNAs using cell proliferation experiments by overexpressing them in the LO2 cell line. Evidently, the new ranks of the circRNAs tended to include more functional candidates after the re-ranking process, highlighting the necessity of considering differentially expressed circRNAs in the

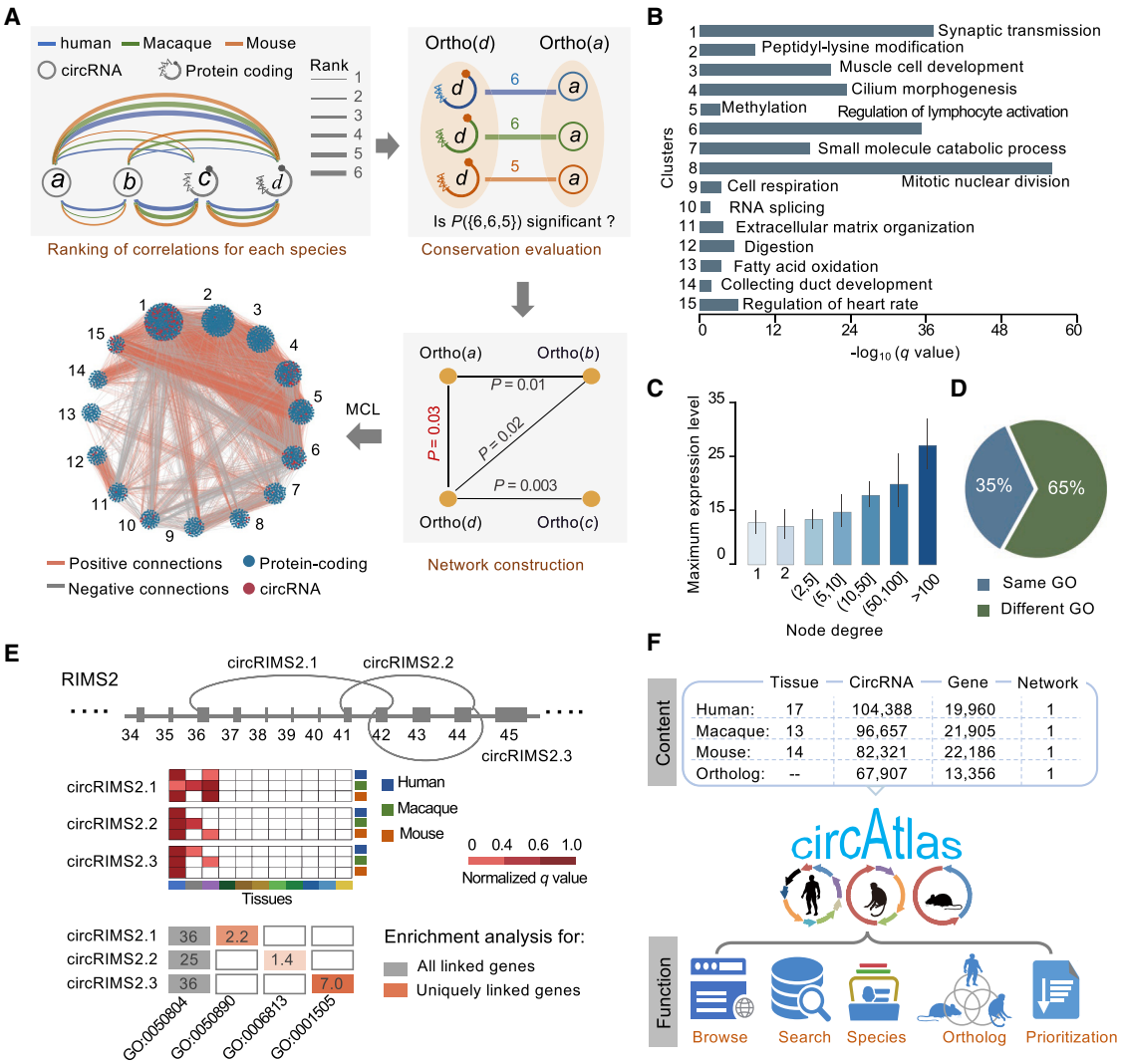


Figure 6. Evolutionarily Conserved Co-expression Network for mRNAs and circRNAs across the Three Species

(A) Construction of evolutionarily conserved co-expression network in human, macaque, and mouse. For each species and for each pair of orthologous circRNAs or mRNAs, the Pearson correlation coefficients of their expression patterns were computed. Subsequently, the combination of each orthologous gene pair was ranked according to its Pearson correlation coefficient. Then, an orthologous gene cluster was defined as a set of orthologous genes (circRNA or protein coding) across the three organisms, and each gene was assigned to at most a single orthologous gene cluster. Given two orthologous gene clusters, if the combination of the correlation coefficients measured in each species was significantly higher or lower than expected by chance, they were considered as co-expressed. Subsequently, all of the links between pairs of co-expressed orthologous gene clusters were combined to build the conserved co-expression network. Finally, the MCL algorithm was applied to cluster the constructed network, and the 15 largest MCL clusters in the co-expression network are shown.

(B) GO enrichment for the top 15 largest MCL clusters; only the most significant GO category is displayed.

(C) Gene expression levels for nodes with different degrees of network connectivity. Highly connected circRNAs tend to have higher expression levels.

(D) Percentage of circRNAs annotated with the same GO term as their host genes.

(E) Functional divergence of circRNAs derived from RIMS2. Top: illustration of the genomic region of three circRNAs derived from the human RIMS2 gene. Curved gray lines indicate the BSJs of circRNAs. Middle: expression profiles of the three circRNAs across the tissues (columns) in human, macaque, and mouse (rows). Bottom: functional annotation of the three circRNAs. On the basis of the conserved co-expression network, these three circRNAs are grouped together and annotated with GO term GO:0050804 (gray blocks). However, they are enriched in different functions (red blocks) when annotating their uniquely connected mRNAs.

(F) All the identified circRNAs and their annotations are integrated into the online platform circAtlas.

See also Figure S6C.

context of the co-expression network. Finally, we focused on the circRNA with the highest weighted score, which was formed by circularization of the second exon of the CDYL2 gene.

To validate the circular transcript of CDYL2, we designed primers (Table S7) that specifically amplified the canonical or back-spliced isoforms of CDYL2 and confirmed the amplified

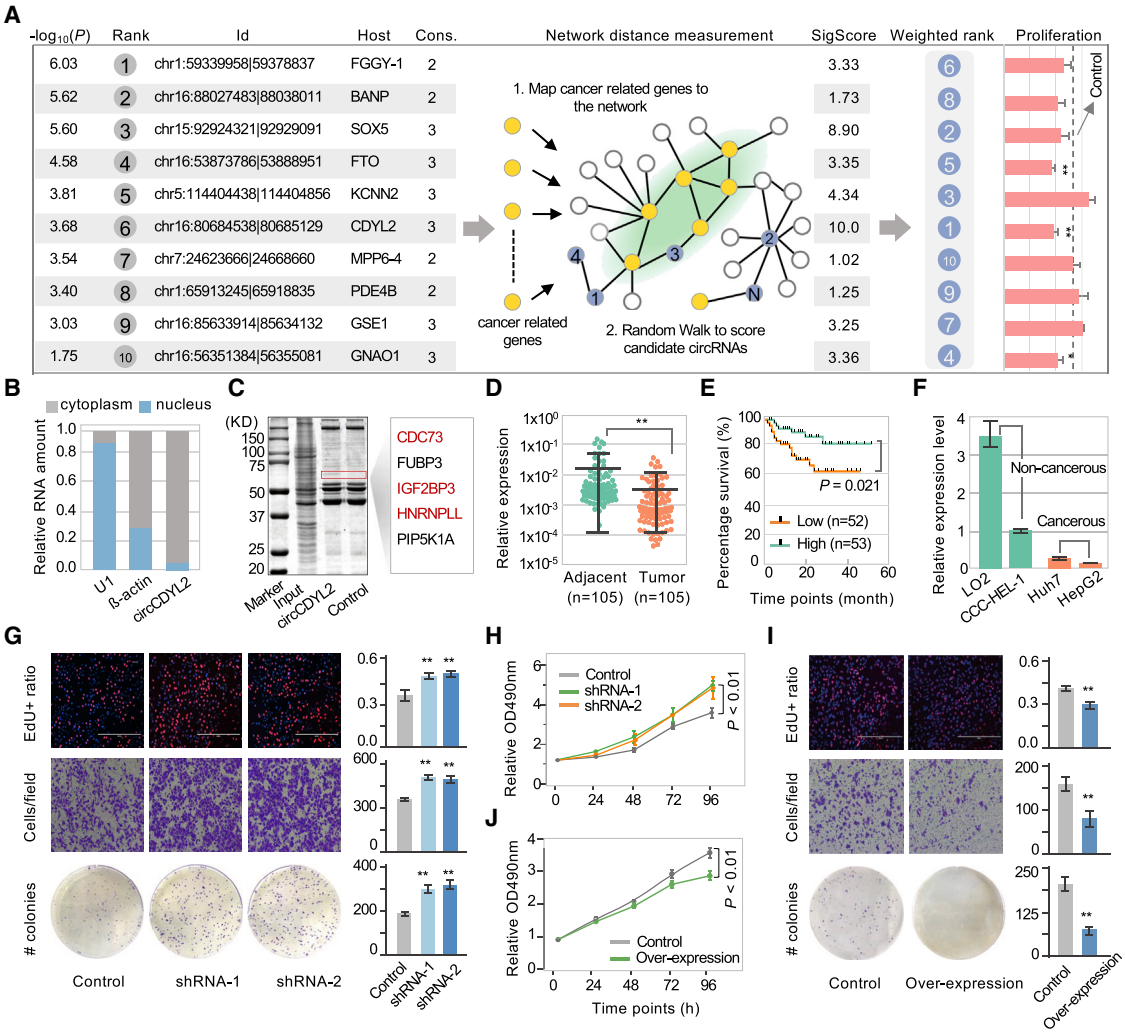


Figure 7. Prioritizing Candidate Liver Cancer Related circRNAs in the Context of the Co-expression Network

(A) Ranking candidate differentially expressed circRNAs in the context of co-expression network. circRNAs were first ranked by the differential expression significance value (leftmost column). A random walk algorithm (see STAR Methods) was used to assign a score to each candidate circRNA. The candidates were then re-ranked according to their assigned scores and conservation to define a priority list of circRNA candidates. Next, we experimentally validated functional circRNAs by overexpressing these circRNAs in the LO2 cell line (*p < 0.05 and **p < 0.01, Wilcoxon test). Cons., conservation; SigScore, significance score. In the network, yellow nodes represent known liver cancer-related genes and blue nodes represent circRNAs strongly associated with these genes.

(B) circCDYL2 was detected in different cell fractions. Nuclear and cytoplasmic RNA was extracted, and junction primers were used for circCDYL2 detection. U1 and β-actin were used as the control for nuclear and cytoplasmic RNA, respectively.

(C) RNA pull-down assay to detect interacted proteins of circCDYL2. The proteins in the red rectangle were extracted and subjected to mass spectrometry analysis. As a result, five proteins were identified and three (in red color) of them were also included in the co-expression network.

(D) Expression levels of circCDYL2 in 105 randomly selected paired tumor and adjacent liver tissues (**p < 0.01). Data are presented as mean value ± SD. The two-tailed unpaired Student's t test was used to determine statistical significance between the adjacent and tumor groups.

(E) Kaplan-Meier survival curve indicating that the low expression of circCDYL2 was correlated with low survival rates. Patients in the cohort were divided into two groups according to the median value of circCDYL2 expression level. The p value between the circCDYL2-high and circCDYL2-low groups is 0.021 (Gehan-Breslow-Wilcoxon test).

(F) Real-time qPCR showing that the expression levels of circCDYL2 in HCC cell lines were lower than in non-cancerous cell lines.

(G) Proliferation, migration, and colony formation assay of the control LO2 cells or circCDYL2 stably silenced LO2 cells. Top: cell proliferations were assessed using EdU immunofluorescence staining method. Middle: cell migration assays were performed using transwell chamber. Bottom: colony formation assay. Bar plots show the quantification of left panels, respectively. The results showed that silencing circCDYL2 significantly enhanced the proliferation, migration, and colony formation capabilities in LO2 cells.

(H) The growth curves of LO2 cells were measured after silencing circCDYL2 using MTS assays.

(legend continued on next page)

sequence using Sanger sequencing and RNase R resistance analysis (Figure S7A). Detection of circCDYL2 on nuclear and cytoplasmic RNA showed that this circRNA was located predominantly in the cytoplasm (Figure 7B). Additionally, knockdown analysis on the LO2 cell line revealed that the expression profile of circCDYL2 was independent of the corresponding parental gene (Figure S7B). We next confirmed the interactions of circCDYL2 on the network using pull-down experiments. Using mass spectrometry (Figure S7C), we successfully identified five proteins interacting with circCDYL2, three of which were confirmed by the co-expression network (Figure 7C), demonstrating a high level of accuracy of the network. One of the identified proteins was parafibromin, a tumor suppressor that inhibits cancer cell growth (Zhang et al., 2006). The connection of circCDYL2 with this protein suggests that this circular transcript may be involved in the biological process of suppressing cancer. Indeed, by quantifying circCDYL2 expression levels using RT-PCR in 105 pairs of cancer and normal liver samples, we found that the levels of circCDYL2 in tumor specimen were significantly lower than in the adjacent benign tissue, which is consistent with the RNA-seq analysis (Figure 7C). The Kaplan-Meier survival curve revealed that downregulated circCDYL2 level was associated with poor survival in liver cancer patients (Figure 7D). Furthermore, our results showed that levels of circCDYL2 were downregulated in cancerous cell lines compared with normal cell lines (Figure 7E), suggesting circCDYL2 may correlate with liver cancer progression. To further understand the biological function of circCDYL2, we performed knockdown and overexpression experiments on the LO2 and Huh7 cell lines, respectively. Subsequently, the effects of circCDYL2 on cell proliferation and migration were examined. Downregulation of circCDYL2 by two short hairpin RNAs (shRNAs) targeting the junction sites of circCDYL2 significantly enhanced cell proliferation and migratory capabilities (Figures 7G and 7H), whereas overexpression of this circular transcript in the cancerous cell line (Huh7) exhibited the opposite trend (Figures 7I and 7J). Taken together, these results suggest that circCDYL2 may play a role in maintaining normal cell functions and a dysregulated level may be involved in liver tumorigenesis. Although further experiments are needed to elucidate its function, this study illustrates how the prior knowledge of co-expression networks, enabled by the evolutionary perspective of this study, can prioritize functional circRNAs and stimulate further investigations.

DISCUSSION

In this study, we explored and characterized expanded transcription landscapes from human, macaque, and mouse. We discovered totals of 104,388, 96,675, and 82,321 highly confident circRNAs from the three species, and 71,112, 77,812, and 56,769 were successfully assembled into full-length circRNAs, respectively. On the basis of these datasets, we comprehensively investigated the expression pattern and evolutionary conservation of circRNAs and identified 70,186 evolutionarily conserved circRNAs and elaborated their importance. We also constructed networks to assign functional annotations and prioritize promising functional circRNA candidates using liver cancer datasets.

Advances in transcriptome sequencing are leading to a deeper understanding of the intricate nature of transcription by identifying a vast number of noncoding RNAs. Recent efforts in the discovery and characterization of circRNAs have resulted in a wealth of data and extended our knowledge of circRNAs, including their composition, expression, and modification (Gao et al., 2016; Yang et al., 2017b). However, current knowledge on circRNAs is far from adequate. High-throughput characterization efforts have thus far been confined to limited tissues from a few species, cell lines, and cancer types. Considering that circRNAs are highly species and tissue specific (e.g., as many as 64% of circRNAs in our study are present in only one tissue), existing identification efforts may not necessarily be considered comprehensive. Therefore, it is difficult to discover and recognize functional circRNAs without knowing their expression landscape in a particular organism or tissue. Here, we generated 132 RNA-seq libraries from 44 normal tissues from human, macaque, and mouse comprising more than 1.8 TB of sequencing data. This huge dataset provides the broadest collection of normal tissues for circRNA studies thus far and is at least one order of magnitude larger than those used in previous studies. Furthermore, this dataset involves three types of RNA-seq libraries—RiboMinus/RNase R, poly(A), and RiboMinus—which outnumber previous circRNA-related studies, of which most only contained RiboMinus/RNase R libraries. The types of libraries used in this study will not only facilitate the sensitive detection of circRNAs but will also enable accurate quantification between circRNAs and their parental genes. For example, we identified an average of 117,035 circRNAs for each species and discovered 10,698 human circRNAs that are more abundant than their parental genes in at least one tissue. More important, we leveraged these datasets to identify approximately 61,956 full-length circRNAs for each species. These full-length circular transcripts provide, for the first time, a comprehensive view of circRNAs in mammals and enable an improved understanding of the composition of these circular transcripts.

The large-scale transcription characterization provides an opportunity to compare the properties of circular and linear noncoding transcripts. In this study, we found that both circRNAs and lincRNAs are transcribed at low levels, are highly tissue specific, are preferentially expressed, and evolve rapidly, which is in agreement with previous investigations (Necsulea and Kaessmann, 2014; Necsulea et al., 2014). However, circRNAs differ from lincRNAs in several aspects. First, there are far more circRNAs than lncRNAs. For example, we identified approximately 139,000 circRNAs in human, which is twice as many as

(I) Proliferation, migration, and colony formation assay of the control Huh7 cells or circCDYL2 stably overexpressed Huh7 cells. The subgraphs have the same order as those in the Figure 7G.

(J) The growth curves of Huh7 cells were measured after overexpressing circCDYL2 using MTS assays.

Data in (F)–(J) are presented as mean value ± SD of three independent experiments. The two-tailed unpaired Student's t test was used to determine statistical significance in and between indicated groups in (F)–(J). See also Figure S7 and Table S7.

in all currently known lncRNAs reported in the MiTranscriptome assembly project (58,648) (Iyer et al., 2015). Furthermore, ES, A3SS, A5SS, and intron retention (IR) events occur more frequently in circRNAs. In human, for example, these four types of alternative splicing events account for approximately 57% of all circRNA alternative splicing events, whereas the number is only 26% in lincRNAs. Second, although both circRNAs and lincRNAs are transcribed at low levels, several circRNAs have extraordinarily high expression levels, such as the circRNAs generated from CDR1 and RIMS1. Among the top 1,000 most abundant RNAs, the number of circRNAs is greater than that of lincRNAs in most tissues. More important, circRNAs exhibit remarkably decreased levels of expression variance across different human individuals relative to lincRNAs but comparable levels to mRNAs, indicating that circRNA expression is under tight regulation rather than being stochastic. Finally, compared with lincRNAs, OO-type circRNAs exhibit increased levels of exon boundary conservation and conserved splicing patterns. Collectively, these findings indicate that the circRNA transcriptome is far more complex than the lincRNA transcriptome and highlight its potential roles in various biological processes.

Although high-throughput sequencing greatly facilitates the accumulation of identified circRNAs, there is a substantial gap between the growing number of circRNAs and our ability to characterize their functions. Indeed, among the 351,105 circRNAs identified in this study, fewer than 0.1% have been individually characterized. Therefore, high-throughput screening of functionally important circRNAs is an essential step toward understanding their functions. Evolutionary conservation analysis has often been used to prioritize noncoding RNAs for functional studies, but this analysis is hampered by an inability to determine circRNAs' complete sequences. Current methods are only able to obtain the BSJ sites of circRNAs rather than their full-length sequences. In this study, we constructed long-insert (400–800 bp) RNA-seq libraries combined with longer reads (2 × PE250) and high sequencing depth (10 + 25 GB per sample) for all 34 tissues, which greatly facilitated the reconstruction of their full sequences. On the basis of the full-length circRNAs, we identified a subset of highly conserved OO-type circRNAs that exhibit a higher level of conservation than other orthologous circRNAs in terms of expression level, junction ratio, tissue specificity, and functional enrichment. We also incorporated two types of networks to maximize the prioritization power of these datasets by assigning putative functions to circRNAs on the basis of their connected genes and screening promising candidates in light of their topological characteristics. The gene knockdown experiments showed high levels of concordance between changes in expression and network prediction. By using these networks, we provide functional annotations for thousands of circRNAs, representing a promising resource for large-scale functional studies of circRNAs. Moreover, the well-established prior knowledge of the networks greatly facilitates screening true functional candidates. By testing on the datasets from liver cancer samples, we demonstrated that compared with the methods that simply choose top differentially expressed candidates, improved accuracy was achieved when prioritizing candidate circRNAs in the context of the networks.

Our study explored the landscape of circRNAs in human, macaque, and mouse; systematically elucidated their diversities in various tissues; and greatly expanded our knowledge of circRNAs on a genome-wide scale. Moreover, our work has generated a wealth of datasets that provide an essential resource for future functional studies. To allow the scientific community to explore these circRNAs, we developed an online portal, circAtlas, which will provide a foundation for circRNA studies and serve as a powerful starting point to investigate their biological importance.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Sample collection
  - Cell lines
- METHOD DETAILS
  - RNA extraction, library construction and sequencing
  - RNA isolation, RNase R resistance analysis, RT-PCR and comparisons with other tools
  - RNA overexpression and knockdown assay
  - Reverse transcription PCR and quantitative real-time qPCR
  - Cell transfection and proliferation, colony formation and migration assays
  - RNA pull-down assay
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - CircRNA detection and full-length circRNA assembly
  - Reference genome, gene expression analysis and functional enrichment
  - Identification of orthologous genes that express orthologous circRNAs
  - Overlapped orthologous circRNA identification
  - Species-specific co-expression network construction and analysis
  - Conserved co-expression network construction
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found with this article online at https://doi.org/10.1016/j.celrep.2019.02.078.

ACKNOWLEDGMENTS

This work was supported by grants from the Beijing Natural Science Foundation (JQ18020), the National Natural Science Foundation of China (31722031, 31701148, 91640117, 31671364, and 91531306), the National Key R&D Program (2018YFC0910400), and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB13000000).

AUTHOR CONTRIBUTIONS

F.Z. conceived the project and designed the approach. P.J., W.W., H.C., and Y.Z. analyzed the data. S.C. sequenced the samples and performed the experiments. L.Z. and W.W. designed the web portal. J.Y., S.Z., and P.Y.

contributed liver cancer samples for survival analysis. F.Z. and P.J. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: October 2, 2018
Revised: January 18, 2019
Accepted: February 20, 2019
Published: March 19, 2019

REFERENCES

Aktaş, T., Avşar Ilık, İ., Maticzka, D., Bhardwaj, V., Pessoa Rodrigues, C., Mittler, G., Manke, T., Backofen, R., and Akhtar, A. (2017). DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. Nature *544*, 115–119.

Chen, X., Han, P., Zhou, T., Guo, X., Song, X., and Li, Y. (2016). circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations. Sci. Rep. *6*, 34985.

Chen, Y.G., Satpathy, A.T., and Chang, H.Y. (2017). Gene regulation in the immune system by long noncoding RNAs. Nat. Immunol. *18*, 962–972.

Cheng, J., Metge, F., and Dieterich, C. (2016). Specific identification and quantification of circular RNAs from sequencing data. Bioinformatics *32*, 1094–1096.

Conn, S.J., Pillman, K.A., Toubia, J., Conn, V.M., Salmanidis, M., Phillips, C.A., Roslan, S., Schreiber, A.W., Gregory, P.A., and Goodall, G.J. (2015). The RNA binding protein quaking regulates formation of circRNAs. Cell *160*, 1125–1134.

Fei, T., Chen, Y., Xiao, T., Li, W., Cato, L., Zhang, P., Cotter, M.B., Bowden, M., Lis, R.T., Zhao, S.G., et al. (2017). Genome-wide CRISPR screen identifies HNRNPL as a prostate cancer dependency regulating RNA splicing. Proc. Natl. Acad. Sci. U S A *114*, E5207–E5215.

Gao, Y., and Zhao, F. (2018). Computational strategies for exploring circular RNAs. Trends Genet. *34*, 389–400.

Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biol. *16*, 4.

Gao, Y., Wang, J., Zheng, Y., Zhang, J., Chen, S., and Zhao, F. (2016). Comprehensive identification of internal structure and alternative splicing events in circular RNAs. Nat. Commun. *7*, 12060.

Gao, Y., Zhang, J., and Zhao, F. (2018). Circular RNA identification based on multiple seed matching. Brief. Bioinform. *19*, 803–810.

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. (2013). Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. Genome Biol. *14*, R70.

Gout, I., Minami, T., Hara, K., Tsujishita, Y., Filonenko, V., Waterfield, M.D., and Yonezawa, K. (1998). Molecular cloning and characterization of a novel p70 S6 kinase, p70 S6 kinase beta containing a proline-rich region. J. Biol. Chem. *273*, 30061–30064.

Hansen, T.B., Jensen, T.I., Clausen, B.H., Bramsen, J.B., Finsen, B., Damgaard, C.K., and Kjems, J. (2013). Natural RNA circles function as efficient microRNA sponges. Nature *495*, 384–388.

Hirsch, S., Blätte, T.J., Grasedieck, S., Cocciardi, S., Rouhi, A., Jongen-Lavrencic, M., Paschka, P., Krönke, J., Gaidzik, V.I., Döhner, H., et al. (2017). Circular RNAs of the nucleophosmin (NPM1) gene in acute myeloid leukemia. Haematologica *102*, 2039–2047.

Ivanov, A., Memczak, S., Wyler, E., Torti, F., Porath, H.T., Orejuela, M.R., Piechotta, M., Levanon, E.Y., Landthaler, M., Dieterich, C., and Rajewsky, N. (2015). Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. Cell Rep. *10*, 170–177.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. *47*, 199–208.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Köhler, S., Bauer, S., Horn, D., and Robinson, P.N. (2008). Walking the interactome for prioritization of candidate disease genes. Am. J. Hum. Genet. *82*, 949–958.

Kristensen, L.S., Hansen, T.B., Venø, M.T., and Kjems, J. (2018a). Circular RNAs in cancer: opportunities and challenges in the field. Oncogene *37*, 555–565.

Kristensen, L.S., Okholm, T.L.H., Venø, M.T., and Kjems, J. (2018b). Circular RNAs are abundantly expressed and upregulated during human epidermal stem cell differentiation. RNA Biol. *15*, 280–291.

Lei, M., Xu, J., Huang, L.C., Wang, L., and Li, J. (2017). Network module-based model in the differential expression analysis for RNA-seq. Bioinformatics *33*, 2699–2705.

Liang, D., Tatomer, D.C., Luo, Z., Wu, H., Yang, L., Chen, L.L., Cherry, S., and Wilusz, J.E. (2017). The output of protein-coding genes shifts to circular RNAs when the pre-mRNA processing machinery is limiting. Mol. Cell *68*, 940–954.e943.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al.; FANTOM consortium (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biol. *16*, 22.

Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., et al.; GTEx Consortium (2015). Human genomics. The human transcriptome across tissues and individuals. Science *348*, 660–665.

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

Necsulea, A., and Kaessmann, H. (2014). Evolutionary dynamics of coding and non-coding transcriptomes. Nat. Rev. Genet. *15*, 734–748.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grützner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature *505*, 635–640.

Pereira, B., Billaud, M., and Almeida, R. (2017). RNA-binding proteins in cancer: old players and new actors. Trends Cancer *3*, 506–528.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Piwecka, M., Glažar, P., Hernandez-Miranda, L.R., Memczak, S., Wolf, S.A., Rybak-Wolf, A., Filipchyk, A., Klironomos, F., Cerda Jara, C.A., Fenske, P., et al. (2017). Loss of a mammalian circular RNA locus causes miRNA deregulation and affects brain function. Science *357*, eaam8526.

Rabani, M., Raychowdhury, R., Jovanovic, M., Rooney, M., Stumpo, D.J., Pauli, A., Hacohen, N., Schier, A.F., Blackshear, P.J., Friedman, N., et al. (2014). High-resolution sequencing and modeling identifies distinct dynamic RNA regulatory strategies. Cell *159*, 1698–1710.

Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., et al. (2015). Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. Mol. Cell *58*, 870–885.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249–255.

Venø, M.T., Hansen, T.B., Venø, S.T., Clausen, B.H., Grebing, M., Finsen, B., Holm, I.E., and Kjems, J. (2015). Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development. Genome Biol. *16*, 245.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178.

Xia, S., Feng, J., Chen, K., Ma, Y., Gong, J., Cai, F., Jin, Y., Gao, Y., Xia, L., Chang, H., et al. (2018). CSCD: a database for cancer-specific circular RNAs. Nucleic Acids Res. *46*, D925–D929.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics *21*, 650–659.

Yang, Y., Chen, L., Gu, J., Zhang, H., Yuan, J., Lian, Q., Lv, G., Wang, S., Wu, Y., Yang, Y.T., et al. (2017a). Recurrently deregulated lncRNAs in hepatocellular carcinoma. Nat. Commun. *8*, 14421.

Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., Jin, Y., Yang, Y., Chen, L.L., Wang, Y., et al. (2017b). Extensive translation of circular RNAs driven by N^6-methyladenosine. Cell Res. *27*, 626–641.

You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., Akbalik, G., Wang, M., Glock, C., Quedenau, C., et al. (2015). Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. Nat. Neurosci. *18*, 603–610.

You, B.H., Yoon, S.H., and Nam, J.W. (2017). High-confidence coding and noncoding transcriptome maps. Genome Res. *27*, 1050–1062.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287.

Zhang, C., Kong, D., Tan, M.H., Pappas, D.L., Jr., Wang, P.F., Chen, J., Farber, L., Zhang, N., Koo, H.M., Weinreich, M., et al. (2006). Parafibromin inhibits cancer cell growth and causes G1 phase arrest. Biochem. Biophys. Res. Commun. *350*, 17–24.

Zhang, X.O., Dong, R., Zhang, Y., Zhang, J.L., Luo, Z., Zhang, J., Chen, L.L., and Yang, L. (2016). Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res. *26*, 1277–1287.

Zheng, Y., and Zhao, F. (2018). Detection and reconstruction of circular RNAs from transcriptomic data. Methods Mol. Biol. *1724*, 1–8.

Zheng, Q., Bao, C., Guo, W., Li, S., Chen, J., Chen, B., Luo, Y., Lyu, D., Li, Y., Shi, G., et al. (2016). Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. Nat. Commun. *7*, 11215.

Zheng, Y., Ji, P., Chen, S., Hou, L., and Zhao, F. (2019). Reconstruction of full-length circular RNAs enables isoform-level quantification. Genome Med. *11*, 2.

Zhou, C., Molinie, B., Daneshvar, K., Pondick, J.V., Wang, J., Van Wittenberghe, N., Xing, Y., Giallourakis, C.C., and Mullen, A.C. (2017). Genome-wide maps of m6A circRNAs identify widespread and cell-type-specific methylation patterns that are distinct from mRNAs. Cell Rep. *20*, 2262–2276.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Human Total RNA Master Panel	Clontech, Palo Alto, CA	Cat# 636643
Deposited Data		
Human liver transcriptomic data	Raw sequencing datasets from the paper	GEO: GSE77509
Human stomach transcriptomic data	Raw sequencing datasets from the paper	GEO: GSE69360
13 time-series rRNA-depleted and 4-thiouridine (4sU)-labeled RNA-seq datasets of mouse immune dendritic cells during the LPS response	Raw sequencing datasets from the paper	GEO: GSE56977
Mouse DHX9 gene kockdown transcriptomic datasets	Raw sequencing datasets from the paper	GEO: GSE85164
Mouse QKI gene kockdown transcriptomic datasets	Raw sequencing datasets from the paper	SRA: ERS640281
RiboMinus treated RNA-seq datasets of tumor and NAT tissue from 20 HCC patients	Raw sequencing datasets from the paper	SRA: from SRX1558026 to SRX1558064
RNA-seq datasets of multiple tissue samples from human, macaque and mouse.	This study	BIGD ID: PRJCA000751
Experimental Models: Cell Lines		
HEK293T	ATCC	Cat#CRL-3216
CCC-HEL-1	ATCC	Cat#BJ-X0027
LO2	ATCC	Cat#CRL-2706
Huh7	Yu Bo Biotech	Cat#YB-H1900
HepG2	ATCC	Cat#HB-8065
Experimental Models: organisms and strains		
Male adult macaque	Kunming Primate Research Center at the Chinese Academy of Sciences (KPRC)	N/A
Female adult macaque	Kunming Primate Research Center at the Chinese Academy of Sciences (KPRC)	N/A
Male adult mouse	Beijing Institutes of Life Science	N/A
Oligonucleotides		
For oligonucleotide sequence information, see Table S1	This study	N/A
Software and Algorithms		
CIRI-full Version 2.0	Zheng et al., 2019	https://sourceforge.net/projects/ciri-full
CIRI Version 2.0	Gao et al., 2018	https://sourceforge.net/projects/ciri
CIRI-AS Version 1.2	Gao et al., 2016	https://sourceforge.net/projects/ciri
DCC version 0.4.6	Cheng et al., 2016	https://github.com/dieterich-lab/DCC
CIRCexplorer version 2.3.0	Zhang et al., 2016	https://github.com/YangLab/CIRCexplorer
MapSplice version 2.1.8	Wang et al., 2010	http://www.netlab.uky.edu/p/bioinfo/MapSplice2
HISAT2 version 2.1.0	Kim, et al., 2015	https://ccb.jhu.edu/software/hisat2/
StringTie version 1.3.3b	Pertea et al., 2015	https://ccb.jhu.edu/software/stringtie/
ClusterProfiler	Yu, et al., 2012	https://github.com/GuangchuangYu/clusterProfiler
CircAtlas	This study	http://circatlas.biols.ac.cn/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by Prof. Fangqing Zhao (zhfq@biols.ac.cn) and Dr. Peifeng Ji (jipeifeng@biols.ac.cn).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Sample collection

Seventeen human samples (Clontech, Palo Alto, CA) were used in this study, including brain, testis, thymus, spleen, spinal cord, kidney, uterus, prostate, heart, lung, liver, colon, bone marrow, small intestine, skeletal muscle, stomach and placenta. Macaque tissue samples were systematically collected from well-characterized rhesus monkeys born and raised at the Kunming Primate Research Center at the Chinese Academy of Sciences (KPRC) in outdoor, 6-acre enclosures that provide a naturalistic setting and a normal social environment. Eleven specimens from one male adult macaque (4 years of age) were collected, including brain, testis, spleen, spinal cord, kidney, prostate, heart, lung, liver, stomach and pancreas. Moreover, uterus samples were collected from one female adult macaque at 13 years of age. Mouse tissue samples were collected from one male adult mouse at the Beijing Institutes of Life Science. Fourteen samples were collected, including brain, testis, thymus, spleen, spinal cord, kidney, prostate, heart, lung, liver, colon, small intestine, skeletal muscle and stomach. All animal procedures were in strict accordance with the guidelines from the National Care and Use of Animals approved by the National Animal Research Authority (P.R. China) and the Institutional Animal Care and Use Committee (IACUC) of the Kunming Institute of Zoology of Chinese Academy of Sciences. The non-human primate care and experimental protocols were approved by the Ethics Committee of Kunming Institute of Zoology and the Kunming Primate Research Center, Chinese Academy of Sciences (AAALAC accredited), and the methods were performed in accordance with the approved guidelines.

Cell lines

HeLa (female, *Homo sapiens* cervical cancer), HEK293T (male, *Homo sapiens* kidney), CCC-HEL-1 (unknown sex, *Homo sapiens* embryonic liver cells), LO2 (male, *Homo sapiens* hepatic cell), Huh7 (male, *Homo sapiens* liver cancer) and HepG2 (male, *Homo sapiens* liver cancer) cells were cultured at 37°C in an incubator containing 5% CO_2 in Dulbecco's modified Eagle's medium (GIBCO, USA) supplemented with 10% fetal bovine serum (GIBCO, USA) and 1% penicillin-streptomycin (GIBCO, USA). Cells were transfected with plasmids using HieffTrans Liposomal Transfection Reagent (YEASEN, China) according to manufacturer's protocol. All cells were cultured at 37°C in 5% CO_2 and tested routinely for Mycoplasma contamination.

METHOD DETAILS

RNA extraction, library construction and sequencing

For each human, macaque and mouse tissue, total RNA was isolated using TRIZOL (Invitrogen, Carlsbad, CA). The RNA concentration and quality were determined with NanoDrop, Qubit and Agilent 2100 instruments. Total RNA was then divided into three replicates containing equal amounts of RNA to construct three types of cDNA libraries. Specifically, a RiboMinus kit (KAPA, USA) was used to deplete ribosomal RNA in these replicates, one of which was further incubated at 37°C with 10 U μg^{-1} RNase R (Epicenter, Madison, WI). rRNA- and rRNA-/RNAase R-treated samples were used as templates for separate cDNA libraries following the TruSeq protocol (Illumina, San Diego, CA), while another total RNA replicate was used to prepare a poly(A)-selected library following the TruSeq v2 guide. The first two libraries were then sequenced on the Illumina HiSeq 2500 platform at the Research Facility Center at Beijing Institutes of Life Science, CAS, with PE250 kits. The poly(A)-selected libraries were sequenced using PE150 kits. The sequence data were submitted to public databases and will be released upon the acceptance of this manuscript.

RNA isolation, RNase R resistance analysis, RT-PCR and comparisons with other tools

Total RNA of mouse brain sample was isolated using TRIZOL. RNA concentration and quality were determined by NanoDrop, Qubit and Agilent 2100. Then, cDNA was synthesized using a SuperScript III first-strand kit (Invitrogen) with random hexamers as primers for all three samples. Outward-facing primer sets (Table S1) were designed for circRNA candidates identified by PCR reactions were performed for the three cDNA samples using 35 cycles. PCR products were directly sequenced to validate circularity.

We applied a similar criterion that was used in CIRI2 paper to evaluate candidate circRNAs with BSJ-read count > 3 in the dataset without RNase R treatment. Specifically, circRNAs were labeled as enriched if a 3-fold increase of BSJ-read count was observed after RNase R treatment. In contrast, candidate circRNAs not detected after RNase R treatment were labeled as depleted. Otherwise, they were classified as unaffected. We also detected circRNAs from the tissue samples of the three species using DCC (v0.4.6), CIRCexplorer (v2.3.0) and MapSplice (v2.1.8). Furthermore, we deposited the predicted circRNAs to the circAtlas database for users to access our data.

RNA overexpression and knockdown assay

To overexpress circCDYL2 in liver cell lines, the genomic region of circCDYL2 with two flanking introns (see Table S7) were amplified by PCR from HEK293T cells. Resulting products were cloned into BamHI and NotI sites of pCDNA3.1+ vector. Huh7 cells were transfected with recombinant plasmid and selected with G418 (300 μg/ml). cDNA oligonucleotides suppressing gene expression were synthesized (sequences are available in Table S7), annealed and inserted into BamHI and EcoRI sites of pSIHI-H1-puro vector (System Biosciences, Mountain View, CA, USA). To produce lentivirus expressing shRNAs, HEK293T cells were co-transfected with the recombinant vector described above, pCMV-VSV-G and pCMV-dR8.2 dvpr vectors. After 48 hours, supernatants containing lentivirus were harvested and filtered through 0.45 μm filters. HeLa and LO2 cells were infected by lentivirus for 48 hours and selected with puromycin (1.5 μg/ml).

Reverse transcription PCR and quantitative real-time qPCR

First-strand cDNA was synthesized with random primer using the FastKing RT kit (Tiangen, China) according to the manufacturer's instruction. Primers for circ-CDYL2, TRA2B, PTBP1, MBNL1 are available in Table S7. Real-time qPCR was performed in the StepOne Plus Real-time PCR system (Applied Biosystem, USA) Using Hieff qPCR SYBR Green Master Mix (YEASEN, China) according to the manufacturer's instruction. The relative expression of RNA was calculated using comparative Ct method.

Cell transfection and proliferation, colony formation and migration assays

Cells were transfected with plasmids using HieffTrans Liposomal Transfection Reagent (YEASEN, China) according to manufacturer's protocol. For cell proliferation assays, 2500 cells were seeded into 96-well flat-bottomed plates. After 12h of culture, cell viability was measured using CellTiter 96 Aqueous One Solution Cell Proliferation Assay (Promega, USA). EdU immunofluorescence staining was performed using Cell-Light EdU Apollo567 *In Vitro* Kit (RibioBio, China) according to the manufacturer's instruction. For colony formation assay, 350 cells were seeded in the 6-well plates and cultured with complete growth medium for 10 days. Clones were fixed with 4% paraformaldehyde, stained with 0.5% crystal violet and counted. Migration assay was performed in Transwell chambers with 8-μm polycarbonate membrane. 5×10^4 cells with serum-free medium were seeded into the upper chambers, complete medium was added to the lower chambers. After culturing for 24h, cells that migrated through membrane were fixed with 4% paraformaldehyde, stained with 0.5% crystal violet and counted.

RNA pull-down assay

A total of 10^7 cells were washed by phosphate-buffered saline, lysed in 500 μL cell lysis buffer for western and IP (Beyotime, China) and centrifuged at 4°C, 12000rpm for 10min. The supernatants were incubated with 3 μg biotin labeled DNA probe (see Table S7) against circ-CDYL2 or control probe at room temperature for 1h. A total of 30 μL washed Streptavidin C1 magnetic beads (Invirtogen, USA) were added to each reaction and incubated at room temperature for 1h. beads were washed by cell lysis buffer for five times. 30 μL elution buffer were added, samples were incubated in boiled water for 10min, and the supernatants were collected.

QUANTIFICATION AND STATISTICAL ANALYSIS

For cell line experiments, data were shown as mean ± SD, ''n'' represents the number of samples used. The type of test method used for statistical analysis was specified in the text where the results were described and details for the test can be found in the relevant figure legend and method section. All tests were two-sided unless otherwise specified.

CircRNA detection and full-length circRNA assembly

BSJs in the RNA-seq reads were detected using CIRI2, DCC, CIRCexplorer2 and MapSplice with default parameters. Since DCC, CIRCexplorer2 and MapSplice are specifically designed for short reads (data not shown), sequencing reads for each sample were trimmed to a length of 150 bp. The single-splice events within these BSJs were inferred with CIRI-AS (v1.2) (parameter -d yes). Within each BSJ, all cirexons inferred from the single splice events were collected, sorted and recorded. Orthologous alternative splicing events in related species were determined using the Liftover tool in the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgLiftOver), which converts genome coordinates from one species to another. Ψ values were calculated by CIRI-AS. The full-length circRNA construction was performed using the CIRI-full pipeline (https://sourceforge.net/projects/ciri-full/), which employs reverse overlapping information for the amplified circular transcripts, where the 5′- and 3′ ends of paired reads were reversely overlapped with one another (Gao and Zhao, 2018; Zheng and Zhao, 2018). Expression variation of a given transcript across multiple individuals was determined using the coefficient of variance, which represents the variance of expression. This score was defined as the ratio of the standard deviation to the mean expression level.

Reference genome, gene expression analysis and functional enrichment

Analyses in this study were performed on genome version GRCh38 downloaded from GENCODE (https://www.gencodegenes.org/) for human, GRCm38 downloaded from GENCODE for mouse and Mmul 8.0.1 for macaque downloaded from Ensembl (http://www.ensembl.org/info/data/ftp/index.html). RiboMinus treated and Poly(A)-enriched RNA-seq datasets were mapped onto the genomes

using HISAT2 (v2.1.0) (Kim et al., 2015), and gene expression was quantified using StringTie (v1.3.3b) (Pertea et al., 2015) with default parameters. KEGG pathway enrichment analysis of the genes from which the circRNAs were derived was performed using the ClusterProfiler (Yu et al., 2012) package.

Identification of orthologous genes that express orthologous circRNAs

circRNAs were first annotated using GTF files to determine their genes of origin. We identified orthologous gene pairs in which both members can express circRNAs using a pairwise orthologous gene list (a list of one-to-one orthologous gene pairs) downloaded from the OMA orthology database (http://omabrowser.org). To identify circRNAs that were conserved at the BSJ level between two species, 50-bp fragments on both sides of the BSJ were extracted from the reference orthologous sequence and used to represent the BSJ sequence. Then, all circRNA BSJ sequences in one species were aligned to those of the other species using BLAT with default parameters. The reciprocal best hit strategy was used to find orthologous circRNAs; then, the orthologous circRNAs pairs with alignment scores < 150 were filtered.

Overlapped orthologous circRNA identification

The complete sequences of the circRNAs in each species were obtained using CIRI-full. circRNA sequences from orthologous genes from multiple species were aligned with each other using BLAT with default parameters. The reciprocal best hit strategy was then employed to process the alignment results and determine the orthologous circRNA pairs at the full-length level. If the overlapping region was larger than 90% of the length of the orthologous circRNAs, they were treated as overlapped orthologous circRNAs (OO-type). OO-type circRNA pairs were defined as circRNAs that are expressed from orthologous genes and share the same BSJs in different species. Non-overlapped orthologous (NO-type) circRNAs were also expressed from orthologous genes, but their BSJ positions differed among species. The expression level of a circRNA was calculated by counting the number of BSJ reads and normalization based on the total number of BSJ reads.

Species-specific co-expression network construction and analysis

We constructed co-expression networks for circRNAs and protein-coding genes in each species. Similarities between individual gene expression patterns were determined by computing a Pearson correlation coefficient matrix for gene-to-gene comparisons and filtering weak correlations, where $r < 0.75$ for protein-coding gene pairs and $r < 0.5$ for gene pairs containing circRNAs. To avoid false positives in the co-expression network analysis, the nodes on the network were restricted to mRNAs or circRNAs expressed in at least 3 tissues.

To determine the reliability of the network, we performed gene knockdown experiments. We knocked down three RBPs—TRA2B, MBNL1 and PTBP1—in HeLa cells. We included two publically available gene knockdown datasets, QKI and DHX9, for analysis. The CIRI pipeline was applied to the knockdown transcriptomic datasets, and the expression levels of circRNAs were normalized. We used the shared circRNAs between knockdown experiments and the co-expression network to assess whether the direction of the alternations in the circRNAs after the gene knockdown experiments could be predicted by the network. Specifically, the changes in circRNA were calculated. Circular transcripts with an expression fold change > 2 that were shared with the network were extracted for verification. If a given circRNA is positively correlated with a given gene in the network, after knocking down this gene, the circRNA's expression level should show a decreasing tendency and vice versa.

To prioritize disease related circRNAs, we developed a method which ranked candidate circRNAs by considering both circAtlas networks and circRNA conservation. In detail, we performed random walk algorithm following a previous method (Köhler et al., 2008), which is defined as the transition of an iterative walker from its current node on the circAtlas network to a randomly selected neighbor starting at a given source node. After assigning scores to the circRNAs on the network, the candidate circRNAs were then re-ranked according to their assigned scores and conservation across species according to the formula: SigScore = Conservation*2-100/$\log_2$(assigned score). At last, the scores of candidates were normalized between 1 and 10.

Conserved co-expression network construction

We built an evolutionarily conserved co-expression network for circRNAs and protein-coding genes following a previous method [36,38], which was used to construct co-expression networks for lncRNAs and protein-coding genes. For each species and for each pair of genes (circRNA or protein-coding gene), we calculated the Pearson correlation coefficients based on their expression profiles. Given two genes, we tested whether the combination of the correlation coefficients computed for each species was significantly higher or lower than expected by chance ($p < 0.05$). To test the significance, we compared the observed ranks of the correlation coefficients with random n-dimensional order statistics as described previously [38]. Note that in the original study presenting the evolutionarily conserved co-expression network reconstruction, negative correlations were discarded. However, negative correlations of expression profiles are expected for repressor regulatory factors and their downregulated targets. Therefore, the correlation was considered relevant if the combination of the correlation ranks across all three species was significantly lower than expected by chance ($p < 0.05$). We used the FDR method for multiple testing corrections to ensure that the co-expression relationships obtained were truly biologically relevant. Specifically, for co-expression between pairs of protein-coding genes, we used an FDR threshold of 0.01, while for pairs of genes including circRNAs, a more relaxed threshold of 0.05 was used to increase the sensitivity.

DATA AND SOFTWARE AVAILABILITY

The sequence data generated in this study have been deposited to BIGD with the following accession number: BIGD ID: PRJCA000751. Data may be accessed at the following site website: http://bigd.big.ac.cn/bioproject/browse/PRJCA000751. Source codes can be downloaded at http://circatlas.biols.ac.cn.