Pervasive translation of circular RNAs driven by short IRES-like elements

Xiaojuan Fan[1, 2, *], Yun Yang[1, *], Zefeng Wang[1, 2, 3#]

[1.] CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, [2] University of Chinese Academy of Sciences, [3] CAS Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China

* These authors contributed equally to this work

Corresponding to: wangzefeng@picb.an.cn

Running title: Pervasive translation of circRNAs

Abstract

Circular RNAs (circRNAs) are an abundant and conserved class of RNAs with ambiguous functions. Some circRNAs have been shown to be translated through IRES driven cap-independent translation, yet the scope of circRNA translation is unclear because endogenous IRESs are rare. To determine the prevalence and importance of circRNA translation, we screened a random 10-nt library using a cell-based reporter system and identified a large number of AU-rich motifs that are capable of initiating circRNA translation. These IRES-like elements are significantly enriched in circRNAs *vs.* linear mRNA and are sufficient to drive robust translation of circRNAs containing solely the coding sequences. Importantly, these elements are predicted to present in thousands of circRNAs with 67% of which encode a truncated isoform of host genes, and are found to bind *trans*-acting factors that promote circRNA translation. Our results indicate that extensive circRNA translation can be driven by short IRES-like elements, suggesting a general function of circRNAs in cytoplasm through translation.

Introduction

Circular RNAs (circRNAs) have recently been demonstrated as an abundant and conserved class of RNAs in animals and plants (for review, see (Barrett and Salzman, 2016; Chen, 2016)). Most circRNAs are produced from a special type of alternative splicing known as back-splicing of pre-mRNA, and are predominantly localized in cytoplasm (Jeck et al., 2013; Li et al., 2018; Salzman et al., 2012). CircRNAs are relatively stable due to their resistance to exonuclease (Jeck et al., 2013; Wang and Wang, 2015), and thus have been used as new type of molecular markers for human diseases such as cancers (Lyu and Huang, 2017; Qian et al., 2018). However, the general function of circRNA *in vivo* is still an open question. Several circRNAs have been reported to function as molecular sponges for miRNAs (Hansen et al., 2013; Memczak et al., 2013) or RNA binding proteins (RBPs) (Ashwal-Fluss et al., 2014) (i.e., as competitors of the linear mRNAs), whereas some nuclear circRNAs were reported to promote transcription of nearby genes (Li et al., 2015; Zhang et al., 2013). Since *in vitro* synthesized circRNAs can be translated in cap-independent fashion (Chen and Sarnow, 1995) and most circRNAs are localized in cytoplasm, it is highly possible that many circRNAs function as mRNAs to direct protein synthesis.

Recently we and other groups reported that some circRNAs are indeed translated *in vivo* via different internal ribosome entry sites (IRESs) (Legnini et al., 2017; Pamudurti et al., 2017; Wang and Wang, 2015; Yang et al., 2017), however the scope of circRNA translation is unclear. Because circRNAs lack a 5′ end, the translation of circRNAs has to be initiated through a cap-independent mechanism that requires the

IRES. However the endogenous IRESs are infrequent in eukaryotic transcriptomes, and their existence is sometimes under debate (Gilbert, 2010; Kozak, 2005, 2007). A recent study has identified hundreds of putative IRESs by systematically searching selected viral sequences and 5′-UTR of human mRNA (Weingarten-Gabbay et al., 2016), however in the set of ~100,000 known circRNAs (Glazar et al., 2014), only a small fraction (<1.5%) contain these newly identified IRESs.

Here we developed a cell-based reporter system to screen a random library for short elements that drive circRNA translation. Through a near-saturated screen and subsequent bioinformatics analyses, we identified 97 IRES-like hexamers that can be clustered into 11 groups with AU rich consensus motifs, and experimentally validated their IRES-like activities. Importantly, the IRES-like elements are significantly enriched in human circRNAs compared to all linear RNAs, suggesting that they are positively selected in circRNAs. Since these IRES-like hexamers account for ~2% of all hexamers (97/4096), any sequences longer than 50-nt should contain such an element by chance, implying that most circRNAs in human cells can potentially be translated. Consistently, we found that circRNAs containing only the coding sequences can indeed be translated in a rolling cycle fashion, presumably from internal IRES-like short elements in the coding region. We further explored the potential roles of circRNA-encoded proteins and identified the *trans*-acting factors that bind these IRES-like elements to promote circRNA translation. Collectively, our data indicate that short IRES-like elements can drive extensive circRNA translation, which may represent a general function of circRNAs in cytoplasm.

Result

Unbiased identification of short IRES-like elements

Previously we reported that GFP coding circRNAs can be translated from different viral or endogenous IRESs (Wang and Wang, 2015; Yang et al., 2017). Surprisingly, three of the four short poly-N sequences used as negative controls for known IRESs in that study were also found to promote GFP translation, with the only exception of poly-G (Fig. S1A). This observation indicates that certain short elements other than known IRESs are sufficient to initiate circRNA translation. To systematically identify additional sequences that drive circRNA translation, we adopted an unbiased screen approach originally developed to identify splicing regulatory *cis*-elements (Wang et al., 2012; Wang and Wang, 2013; Wang et al., 2013). Briefly, a library of random 10-nt sequences was inserted before the start codon of circRNA-encoded GFP (Yang and Wang, 2018), which was transfected into 293T cells to generate circRNAs that can be translated into intact GFP (see methods and table S1). To achieve full coverage of entire library, >100 million cells were transfected (Fig. 1A). We used fluorescence activated cell sorting (FACS) to recover the cells with green fluorescence resulted from circRNA translation, and then sequenced the inserted fragments to identify the IRES-like elements that drive circRNA translation.

We sequenced the inserted fragments from both dark and green cells using high-throughput sequencing, and compared the resulting 10-mers between these two cell populations to extract the hexamers enriched in the green cells *vs.* the dark cells

(Fig. 1B). We have identified 97 hexamers that are significantly enriched in the cells with GFP fluorescence (table S2). These enriched sequences are generally AU-rich despite the pre-sorting library has roughly even base composition (Fig. 1C and Fig. S1B). In addition, these hexamers have strong dinucleotide biases toward AT, AG and GA (Fig. 1D). Based on the sequence similarity, the 97 hexamers enriched in green cells were further clustered into 11 groups to produce consensus motifs (Fig. 1E, top). Consistently most consensus motifs are AU-rich motifs and enriched in 3′-UTR of linear mRNA (Fig. S1C). Several enriched hexamers contain the RRACH signature of the m6A modification site, however they were not prevalent enough to be clustered into a consensus motif.

The IRES-like activities of each cluster were further validated by inserting the representative hexamers in each cluster (or the control hexamers depleted in green cells) into the circRNA reporter and measuring the translation products of circRNAs with western blot (Fig. 1E, bottom). All the reporters inserted with the enriched hexamers showed robust GFP translation from circRNAs, whereas the GFP productions from the reporters inserted with control hexamers were barely detectable (Fig. 1F). The circRNA levels were similar in all reporters containing different hexamers as judged by RT-PCR and by northern blot (Fig. 1E and Fig. S2), suggesting that the differences in GFP production are probably due to distinct activities of these hexamers in driving translation (rather than by differences in back-splicing efficiency). Collectively, these results indicated that our screen can reliably identify short sequences that drive circRNA translation, and thus we refer this set of 97 hexamers as

IRES-like hexamers.

IRES-like hexamers are enriched in endogenous circRNAs to drive translation

We further examined the frequency of each hexamer in linear mRNAs and circRNAs and compared the average frequency of the IRES-like hexamers *vs.* the control set of hexamer sets in different types of RNAs. In the linear mRNAs (all annotated mRNAs from RefSeq), the average frequencies of the IRES-like hexamers are similar to the random hexamers or the hexamers depleted in green cells (Fig. 2A, left panel). Surprisingly, the average frequencies of IRES-like hexamers were significantly higher in all tested circRNA datasets (Gao et al., 2016; Jeck et al., 2013; Memczak et al., 2013; Rybak-Wolf et al., 2015) compared to the control hexamer sets (Fig. 2A), indicating that endogenous circRNAs are enriched with short IRES-like elements. Since these IRES-like hexamers were independently identified from unbiased screen of random sequences, such enrichment strongly suggests that circRNAs may be positively selected for their ability to be translated.

The 97 IRES-like hexamers account for ~2% of the entire hexamer population (4^6=4096), indicating that there will be an IRES-like hexamer in a 50-nt sequence by chance. Since >99% circRNAs are longer than 100-nt (Glazar et al., 2014) (Fig. S3A), most circRNAs should contain internal IRES-like short elements by chance. Therefore almost all open reading frames (ORFs) in circRNAs could potentially be translated using IRES-like short elements. To directly test this unconventional conclusion, we generated a series of GFP-coding circRNA reporters without stop

codon and/or known IRES sequences to measure GFP production using western blot (Fig. 2B). Consistent with our previous report, the control reporter containing the m6A modification sites at upstream of the start codon was reliably translated (Yang et al., 2017), and deletion of stop codon in this reporter led to production of GFP concatemers from a rolling circle translation of circRNA (Fig. 2B). When deleting all UTR sequences between the start and stop codon in the circRNA, the GFP translation is abolished, presumably because there is no room for any sequence to function as the IRES. However, when deleting the stop codon, the circRNA containing only the GFP coding sequence can also be translated in a rolling circle fashion to produce GFP concatemers, presumably through an internal sequence function as an IRES (lanes 4 and 5, Fig. 2B). As a control, we also confirmed the efficient circRNA expression in all the RNA samples using northern blot with the optional RNase R treatment (Fig. S3B). Interestingly, the rolling cycle translation can produce some huge GFP concatemers that are retained in loading wells (lanes 3 and 5, Fig. 2B, the huge proteins could not be efficiently transferred to the membrane and thus being underestimated). To eliminate the possible artifacts from rolling cycle transcription from circular plasmid, these reporter plasmids were stably inserted into genome using Flp-In system, and the similar rolling cycle translations were observed in these stably transfected cells (Fig. S3C).

In addition, the rolling circle translation initiated from internal coding sequence is not limited to the circRNA encoding GFP gene, as the circRNAs containing only the ORF sequences of different luciferase genes can also be translated from internal

coding sequence (Fig. 2C, Fig. S3D-F). Interestingly, the rolling cycle translation from circRNAs apparently produces more proteins than the circRNAs with stop codons (Fig. 2B), suggesting that initiation of circRNA translation is slow and may be a rate-limiting step.

Potential function of circRNA encoded proteins

Since most circRNAs can potentially be translated, we further examined molecular characteristic of the putative circRNA-encoded proteins. In accordance with previous observation (Pamudurti et al., 2017), we found that the exons in the 5′ end of a pre-mRNA are more likely to be included in the circRNAs (Fig. 3A), suggesting that many circRNAs could potentially encode N-terminal truncated protein isoforms of their host genes. Consistently we found that 79% of circRNA exons are from coding region, while 17% of circRNA exons spanning 5′-UTR and coding region and 4% spanning coding and 3′-UTR region (Fig. 3A inset).

Using the dataset of full length human circRNAs (Gao et al., 2016), we examined the putative circRNA-encoded ORFs that are longer than 20 amino acids (equivalent to 60 nt). We found that a large fraction of endogenous circRNAs (67%) can encode for proteins overlapping with their host genes (i.e., translated in the same reading frame as the host gene), including 14% of endogenous circRNAs that can be translated in a rolling cycle fashion (named as overlapped cORF and rcORF respectively, purple pie slices in the left panel of Fig. 3B, also see table S3 for all circRNA-encoded isoforms). In addition, 16% of human circRNAs can encode for

proteins that are different from host genes but are homologous to other known proteins (homologous cORF and rcORF, brown pie slices in Fig. 3B, left panel), whereas 10% circRNAs encode for proteins that are not homologous to any known proteins on earth. In fact, only 7% of circRNAs have no ORF longer than 20 aa. As a comparison, 44% or 21% of circRNAs from two different controls (reversed sequence or shuffled sequence controls) do not contain any ORF (Fig. 3B, right two panels), and the remaining control circRNAs are more likely to code for proteins that are not homologous to any known proteins, suggesting that the endogenous circRNAs are more likely to code for functional proteins.

We further examined the protein domains in these circRNA-coded ORFs, and found that 55% of circRNAs with overlapped ORFs contain some known protein domains (Fig. 3C). However only 9% of circRNAs with homologous ORFs include the known protein domains (Fig. 3C). These data suggest that circRNAs with overlapped ORF are more likely to produce truncated proteins with known functional domains (thus may play the similar or competitive roles to host genes) compared to the circRNAs with homologous ORF. We further examined the putative functions for these domains, and found that these circRNA-encode isoforms are enriched for P-loop NTP hydrolase domains and protein-kinase domains, the domains for protein-protein interaction or protein complex formation (e.g., armadillo-type fold, WD-40 repeats containing domain, ankyrin repeat domain), protein-RNA interacting domains, and signal transduction domains (PH-domain like superfamily) (Fig. 3C).

We further analyzed the protein-protein interaction (PPI) network of the host genes to infer potential functions of these circRNA-encoded proteins. Based on the PPI evidences in public domain, these genes formed a densely connected interaction network, which can be further clustered into different groups with distinct functions (e.g. neurotrophin signaling pathway, DNA replication, ubiquitin mediated proteolysis, and clathrin coat assembly et al.) (Fig. 3D, Fig. S4A). Consistent with the enriched domains in circRNA-encoded proteins, the major molecular function of these host genes is protein-protein interaction (Fig. S4B). Along with the observation that these overlapped circORFs are enriched with protein binding and protein kinase domains, these data suggest circRNA-encoded proteins may function as regulators or competitors for the canonical isoforms that bind to their targets to regulate key biological pathways like DNA replication, signal transduction and protein degradation.

Trans-acting factors that bind to IRES-like short elements to promote circRNA translation

Next, we seek to uncover the molecular mechanisms by which these newly IRES-like short elements initiate cap-independent translation of circRNA. An earlier report showed that short sequences may function as IRES by pairing with certain regions of 18S rRNA (i.e. “active region”) (Weingarten-Gabbay et al., 2016). However we found little correlation between our newly identified IRES-like elements

to these “active 18S rRNA regions” (Fig. S5A), suggesting that these newly identified elements may not function by paring with 18S rRNA.

Previously, we found that m6A reader protein YTHDF3 can recognize N6-methyladenosine in circRNA to directly recruit translation initiation factors (Yang et al., 2017). By analogy the newly identified IRES-like short elements may also function as regulatory *cis*-elements to recruit *trans*-acting factors that promote translation (Harvey et al., 2018). To identify such *trans*-acting factors, we used the consensus sequences of IRES-like elements as bait for affinity purification of their specific binding proteins (Fig. 4A). Briefly, ~20 nt RNA sequences containing three copies of IRES-like hexamers were synthesized with a 5′ end biotin modification and a $(CH_2)_{18}$ spacer, and the resulting RNA baits were incubated with HeLa cell extracts, subsequently the RNA-protein complexes were purified with streptavidin beads (Wang et al., 2012; Wang and Wang, 2013; Wang et al., 2013) (Fig. 4A). We found that, all the five probes consist of IRES-like elements showed robust binding of several proteins, whereas the negative sequence (ACCGCG) had weak background of non-specific RBP binding (Fig. 4B). The specific protein bands in each lane were cut for identification with mass-spectrometry (LC-MS/MS), and the top candidates in each band were shown (Fig. 4B).

In total 58 protein candidates were identified with high confidence, with many overlapping proteins in different baits (table S4). The majority of these proteins are known to bind RNAs and can be clustered into two major groups based on the protein-protein interaction (PPI) network: the proteins involved in RNA processing

(e.g., hnRNPs) and the proteins involved in mRNA translation (e.g., the ribosomal proteins) (Fig. 4C). The identified RBPs are enriched for regulatory function in RNA processing, splicing, transport and stabilization as judged by gene ontology analysis (Fig. S5B).

We further tested several proteins by fusing the candidate RBPs to a customized Puf domain, a programmable RNA binding domain that can be designed to bind any 8-nt RNA (Wei and Wang, 2015). We found that the specific tethering of PABPC1 and hnRNP U to the circRNA reporter can promote translation, whereas the ELAVL1 (HuR) and hnRNP A1 do not affect translation when tethered to the same site in the circRNA (Fig. 4D). Such translation promoting activity required the specific binding of *trans*-acting factors to circRNAs, as disrupting the Puf-RNA interactions had abolished the regulatory effect and restoring the specific interaction can rescue the translation-promoting activity (Fig. 4D).

PABPC1 is an abundant protein that bind to poly-A or AU-rich sequences (Wigington et al., 2014). We further test the role of PABPC1 in regulating circRNA using reporters contain short poy-A elements at upstream of start codon in circRNA (Fig. 4F). The results showed that the over-expression of PABPC1 can indeed promote translation of circRNA encoded GFP. As controls, the circRNA translation was not affected by another poly-A binding protein PABPC4, or by proteins previously reported to enhance IRES activity (Fig. 4F). Taking together, our results showed that certain RBPs (e.g., PABPC1) are capable to recognize these IRES-like

elements to promote cap-independent translation of circRNAs, exemplifying a new model for circRNA translation initiation independent of canonical IRESs.

Discussion

Several recent studies indicated that some circRNAs function as template to direct protein synthesis, however the nature of circRNA translation is still under debate because other studies failed to detect the significant association of circRNA with polysome (Guo et al., 2014; Jeck et al., 2013; You et al., 2015). In addition, the mechanism of circRNA translation is not clear. While the cap-independent translation of circRNA generally requires a viral or endogenous IRES, we have demonstrated that a short *cis*-elements containing m6A modification site is sufficient to drive circRNA translation by directly recruiting initiation factors (Yang et al., 2017). Here we surprisingly found that the requirement of IRESs can be easily fulfilled, and many additional short sequences (~2% of all hexamers) are capable of driving cap-independent translation in circRNAs. In fact, any circRNAs longer than 50 nt should contain an IRES-like hexamer by chance, and more importantly endogenous circRNAs are enriched with such IRES-like elements compared to linear mRNAs. As a result, thousands of cytoplasmic circRNAs can be potentially translated, the majority of which produce protein isoforms overlapping with canonical gene products.

Since many circRNAs contain the natural start codon of host gene, this study raises an interesting open question regarding the biological function of the circRNA-coded protein isoforms. Because these circRNA encoded isoforms have large overlap with the canonical host gene, these circRNA-encoded isoforms may

function as competitive regulator of canonical isoforms or play the similar function in different subcellular location.

Compared to linear mRNA, a relatively small number of circRNAs are reported to be associated with polysomes (Guo et al., 2014; You et al., 2015), suggesting an inefficient translation of circRNAs. In addition, the minimal requirement for cap-independent translation of circRNAs and the difficulties to detect circRNA-encoded products presented an interesting paradox. While there is no clear explanation, several factors may help to reconcile such contradiction. First, although most circRNAs have the capacity to be translated, it does not necessarily mean that they are indeed translated efficiently *in vivo* to produce stable proteins. It is possible that some products from pervasive translation of circRNAs may not be folded correctly and thus be rapidly degraded. This scenario is conceptually similar to the transcriptional control, where the pervasive transcription happen in many genomic regions on both directions but most of the transcribed products are degraded and only a small fraction of products are stable and functional (Jensen et al., 2013; Wade and Grainger, 2014). In another word, the protein folding and stability may function as the quality control step for circRNA translation.

The other possibility is that the initiation of cap-independent translation of circRNAs is less efficient than linear mRNAs. However the translation elongation rate should be comparable between circular *vs.* linear RNAs once the active ribosomes are assembled onto mRNAs. As a result, the translation of most circRNAs may be carried out by monoribosomes rather than polysomes, which could

explain the lack of circRNAs in polysome-associated RNAs. This is supported by the observation that rolling cycle translation of the reporter circRNA has produced much more products than the same circRNA containing stop codon (Fig. 2B and 2C), the latter of which requires re-initiation at each round of translation. Therefore the circRNAs with rolling cycle open reading frame should be an important subject for future study, because their products are usually misfolded to form protein aggregates due to concatemeric repeats, which may have pathogenic properties. On the other hand, the misfolded protein products often induce unfolded protein response that leads to protein degradation and apoptosis (Yang et al, unpublished data).

It is well accepted that the translation through cap-independent pathways are less efficient under normal physiological condition. However under certain cellular stress condition (like heat stress) or in certain cell types (like cancer cells), the canonical cap-dependent translation is inhibited and the cap-independent translation may become more predominant (Leprivier et al., 2015; Liu and Qian, 2014). Consistently we found that the translation of GFP from circRNA is promoted in heat shock conditions (Yang et al., 2017), suggesting that circRNA encoded proteins may be induced under such conditions (or in certain cells where canonical translation is suppressed), implying potential roles for the circRNA-encoded proteins in stress response and cancer cell progression.

Methods Summary (see supplementary material for detailed methods)

Plasmid library construction and screening

In order to screen short elements for the initiation of circRNA translation, the previously described pcircGFP reporter (Wang and Wang, 2015) was modified and inserted with a random 10mer sequence library. We obtained sufficient numbers of *E. coli* clones to achieve ~2-fold coverage of all possible decamers. The resulting library was transfected into 293T cells (20 μg/ per 15 cm dish) and the green cells were sorted 48 hours after transfection. In total 122 million cells were sorted, we collected 4 million cells without GFP fluorescence (negative controls), 13 million cells with low GFP fluorescence, 5 million cells with medium GFP fluorescence and 0.5 million cells with high GFP. Then RNAs were extracted, and sequencing library was generated by using RT-PCR. RNA-seq was performed with Hiseq 2500.

Identification of enriched motifs in IRES-like elements

We used a statistic enrichment analysis to extract enriched motifs from decamer sequences recovered in green cells. Briefly, each inserted 10-mer was extended into a 14-mer by appending 2 nt of the vector sequence at each end to allow for cases in which IRES activity derived from sequences overlapping the vector. The resulting 14-nt sequences were broken into overlapping hexamers, and .the enrichment score of each 6-mer between different datasets was calculated using Z-test. Hexamers with score > 7 are defined as IRES-like elements, while the hexamers with score < -7 are defined as depleted elements.

circRNA datasets

The circRNA datasets were obtained by analyzing previously published RNA-seq data from 293 cells treated with RNase R (BioProject database of Genbank, accession number PRJNA266072). The sequencing data were further processed to obtain full length circRNA data, see supplementary methods for detail.

Identification of *trans*-factors with RNA affinity purification

The RNA affinity purification method was adopted from the previously described protocol (Wang and Wang, 2013). Each biotin labeled RNA sample was incubated with cell extracts of 2.5×10^8 HeLa cells for 2 hrs at 4 °C in a 5 ml mixture. Next, 50 μl Streptavidin-agarose beads were added into the mixture and incubated for 2 hrs at 4 °C with slow rotation. The beads were washed 3 times and eluded, and the total proteins were then separated with a 4-20% PAGE Gel for mass spectrometry analysis.

circORF prediction and domain prediction

All potential ORFs encoded by the circRNAs in three reading frames of the sense strand were predicted. We set a threshold of ORF length at >20aa. If one ORF doesn't contain stop codon, it was defined as rolling cycle translation (i.e. without UTR). The circRNAs that do not contain any IRES-like element in the putative UTR region were removed from future analyses.

We used the blastp 2.6.0+ to detect the homologues of circRNA encoded ORFs in

non-redundant protein database. If no homologous proteins were detected for the circORFs, we labeled the circRNA as “blast null”. If circORF is partially overlapped with host gene ORF (at least 7aa are the same), we named it “overlapped circORF”, if the circORF do not overlap with the host protein but are homologous to other known proteins, we named it “homologous circORF”. The domains of circORFs were predicted using InterPro package and SUPERFAMILY analyses (Gough et al., 2001). All predicted domains with e value less than 0.0001 in SUPERFAMILY database were used for detailed analyses. As controls, we reversed circRNA sequence (as “reversed control circRNA”), or shuffled circRNA sequence three times (as “shuffled control circRNA”) with the mean values from three shuffled control datasets being shown in the figure 3B (the standard deviation between shuffled sets is less than 0.005).

Acknowledgment

The authors want to thank Drs. Albrecht Bindereif, Reinhard Luehrmann and Xiaoling Li for useful discussions and comments, and Yue Hu for help in analyzing RBP binding data. This work is supported by National Natural Science Foundation of China to Z.W. (31570823, 31661143031, and 31730110) and Y.Y. (91753135, 31870814). Z.W. is also supported by the type A CAS Pioneer 100-Talent program.

Author Contributions

Conceptualization, Z.W. and Y.Y; Methodology, X. F., Y. Y., and Z.W.; Software, X.F.; Experiments, X.F., Y.Y.; Writing, X.F., Y.Y., X. L. and Z.W.; Funding Acquisition, Y.Y., and Z.W.

Declaration of Interests

Z. W. and Y. Y. has co-founded a company, CirCode Biotech, to commercialize the application of circular RNA as template of protein production/expression. The other authors declare no competing financial interests.

Reference:

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

Barrett, S.P., and Salzman, J. (2016). Circular RNAs: analysis, expression and potential functions. Development *143*, 1838-1847.

Chen, C.Y., and Sarnow, P. (1995). Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs. Science *268*, 415-417.

Chen, L.L. (2016). The biogenesis and emerging roles of circular RNAs. Nat Rev Mol Cell Biol *17*, 205-211.

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.

Gilbert, W.V. (2010). Alternative ways to think about cellular internal ribosome entry. J Biol Chem *285*, 29033-29038.

Glazar, P., Papavasileiou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. RNA *20*, 1666-1670.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol *313*, 903-919.

Guo, J.U., Agarwal, V., Guo, H., and Bartel, D.P. (2014). Expanded identification and characterization of mammalian circular RNAs. Genome biology *15*, 409.

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.

Harvey, R.F., Smith, T.S., Mulroney, T., Queiroz, R.M.L., Pizzinga, M., Dezi, V., Villenueva, E., Ramakrishna, M., Lilley, K.S., and Willis, A.E. (2018). Trans-acting translational regulatory RNA binding proteins. Wiley Interdiscip Rev RNA *9*, e1465.

Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., and Sharpless, N.E. (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA *19*, 141-157.

Jensen, T.H., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. Mol Cell *52*, 473-484.

Kozak, M. (2005). A second look at cellular mRNA sequences said to function as internal ribosome entry sites. Nucleic Acids Res *33*, 6593-6602.

Kozak, M. (2007). Lessons (not) learned from mistakes about translation. Gene *403*, 194-203.

Legnini, I., Di Timoteo, G., Rossi, F., Morlando, M., Briganti, F., Sthandier, O., Fatica, A., Santini, T., Andronache, A., Wade, M.*, et al.* (2017). Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. Mol Cell *66*, 22-37 e29.

Leprivier, G., Rotblat, B., Khan, D., Jan, E., and Sorensen, P.H. (2015). Stress-mediated translational control in cancer cells. Biochimica et biophysica acta *1849*, 845-860.

Li, X., Yang, L., and Chen, L.L. (2018). The Biogenesis, Functions, and Challenges of Circular RNAs. Mol Cell *71*, 428-442.

Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., Zhong, G., Yu, B., Hu, W., Dai, L., *et al.* (2015). Exon-intron circular RNAs regulate transcription in the nucleus. Nat Struct Mol Biol *22*, 256-264.
Liu, B., and Qian, S.B. (2014). Translational reprogramming in cellular stress response. Wiley Interdiscip Rev RNA *5*, 301-315.
Lyu, D., and Huang, S. (2017). The emerging role and clinical implication of human exonic circular RNA. RNA Biol *14*, 1000-1006.
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.
Pamudurti, N.R., Bartok, O., Jens, M., Ashwal-Fluss, R., Stottmeister, C., Ruhe, L., Hanan, M., Wyler, E., Perez-Hernandez, D., Ramberger, E., *et al.* (2017). Translation of CircRNAs. Mol Cell *66*, 9-21 e27.
Qian, L., Yu, S., Chen, Z., Meng, Z., Huang, S., and Wang, P. (2018). The emerging role of circRNAs and their clinical significance in human cancers. Biochim Biophys Acta Rev Cancer *1870*, 247-260.
Rybak-Wolf, A., Stottmeister, C., Glazar, 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.
Salzman, J., Gawad, C., Wang, P.L., Lacayo, N., and Brown, P.O. (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS One *7*, e30733.
Wade, J.T., and Grainger, D.C. (2014). Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. Nat Rev Microbiol *12*, 647-653.
Wang, Y., Ma, M., Xiao, X., and Wang, Z. (2012). Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. Nat Struct Mol Biol *19*, 1044-1052.
Wang, Y., and Wang, Z. (2013). Systematical identification of splicing regulatory cis-elements and cognate trans-factors. Methods.
Wang, Y., and Wang, Z. (2015). Efficient backsplicing produces translatable circular mRNAs. RNA *21*, 172-179.
Wang, Y., Xiao, X., Zhang, J., Choudhury, R., Robertson, A., Li, K., Ma, M., Burge, C.B., and Wang, Z. (2013). A complex network of factors with overlapping affinities represses splicing through intronic elements. Nat Struct Mol Biol *20*, 36-45.
Wei, H., and Wang, Z. (2015). Engineering RNA-binding proteins with diverse activities. Wiley Interdiscip Rev RNA *6*, 597-613.
Weingarten-Gabbay, S., Elias-Kirma, S., Nir, R., Gritsenko, A.A., Stern-Ginossar, N., Yakhini, Z., Weinberger, A., and Segal, E. (2016). Systematic discovery of cap-independent translation sequences in human and viral genomes. Science *351*.
Wigington, C.P., Williams, K.R., Meers, M.P., Bassell, G.J., and Corbett, A.H. (2014). Poly(A) RNA-binding proteins and polyadenosine RNA: new members and novel functions. Wiley Interdiscip Rev RNA *5*, 601-622.
Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., Jin, Y., Yang, Y., Chen, L.L., Wang, Y., *et al.* (2017). Extensive translation of circular RNAs driven by N(6)-methyladenosine. Cell Res *27*, 626-641.

Yang, Y., and Wang, Z. (2018). Constructing GFP-Based Reporter to Study Back Splicing and Translation of Circular RNA. Methods Mol Biol *1724*, 107-118.
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. Nature neuroscience *18*, 603-610.
Zhang, Y., Zhang, X.O., Chen, T., Xiang, J.F., Yin, Q.F., Xing, Y.H., Zhu, S., Yang, L., and Chen, L.L. (2013). Circular intronic long noncoding RNAs. Mol Cell *51*, 792-806.

Figure Legend:

Figure 1. Extensive IRES-like elements can drive circRNA translation.

(A) Schematics diagram for the screen of short IRES-like elements. Random decamers were inserted into pcircGFP-BsmBI reporter, and the resulting library was transfected into 293T cells and sorted by FACS. The cells with different GFP fluorescence signals were collected and the inserted sequences were amplified and sequenced using high-throughput sequencing. The hexamers enriched in green cells were identified by computational analysis (also see method).

(B) Flow-cytometry analysis of cells transfected with circRNA reporter containing the random 10-mer library. The cells were classified into four groups based on their GFP fluorescence (GFP negative cells and cells with low, medium or high GFP signals). The cells with medium and high fluorescence were sorted out as "green cells".

(C) Single nucleotide frequency in the starting library (top) and the sequences enriched in green cells (bottom).

(D) Dinucleotide frequency and enrichment of the sequences enriched in green cells.

(E) The 97 enriched hexamers (i.e., IRES-like elements, z score >7) were clustered into 11 groups and the consensus motifs were shown as pictogram (top). The representative hexamers in each cluster were inserted into the circRNA reporter that was transiently transfected into 293T cells, the translation of GFP was assayed by western blot 48 hours after transfection (bottom).

(F) The 122 depleted hexamers (i.e., negative control, z score < -7) were clustered into 11 groups and the consensus motifs were shown as pictogram (top). The

activities of representative hexamers in all clusters were tested using the same condition as panel E. An IRES-like element, AAAAAA, was included as the positive marker.

Figure 2. circRNAs contain many IRES-like elements that initiate translation.

(A) IRES-like elements are significantly enriched in circRNAs. Average frequency of different types of hexamers (all hexamers, IES-like hexamers and depleted hexamers) in linear mRNA *vs.* circRNAs. N: the number of RNA sequences used in each dataset. **: p-value < 0.001 with KS test.

(B) Translation of circRNAs can be initiated by internal coding sequence. Left panel, schematic diagrams of four reporter circRNAs that encode GFP. From top to bottom: circRNA with an m6A site at upstream of the start codon; circRNA with an upstream m6A site and no stop codon; circRNA with start codon immediately following the stop codon; circRNA containing only the coding sequence (no stop codon or UTR). Red line indicates stop codon. Right panel, western blot analysis to detect translation from circRNAs. The circRNA plasmids were transfected into 293T cells, and samples were analyzed by western blot 2 days after transfection. Both rcGFP with m6A and rcGFP can generate several large protein concatemers of GFP through rolling cycle translation. The level of GAPDH was measured as a loading control.

(C) Translation of the circRNA encoding Renilla luciferase (Rluc) using internal IRES-like elements. Top: schematic diagram of two Rluc circRNAs. cRLuc contains a sequence coding a T2A peptide which can cleave large repeat protein into small

pieces between start and stop codon of Rluc. rcRluc do not have stop codon. Bottom: dual luciferase assay of cRluc and rcRluc. Control and circular Rluc plasmid were co-transfected with Fluc (firefly luciferase) reference reporter into 293T cells. 48 hours after transfection, cells were lysed and luminescence measured by luminescence reader. (n = 3 independent experiments; error bars represent the mean ± SD).

Figure 3. Functional analysis of putative circRNA encoded proteins.

(A) Position distribution for the first exon of circRNAs in their host genes. Previously published full length circRNA sequences were retrieved from BioProject database, and the histogram was plotted according to the positions of the first circRNA exon (i.e., the exon number) in the host genes. The inserted pie chart presents the percent of circRNAs overlapping with different regions of mRNA.

(B) Survey of the potential coding products of circRNAs. Left panel, percent of endogenous circRNAs that code for an ORF longer than 20aa. Purple pie slices: circRNAs translated in a regular fashion (cORF, dark purple) or a rolling cycle fashion (rcORF, light purple) into proteins that are partially overlapped with their host genes; brown pie slices: circRNAs translated in a regular fashion (cORF, dark brown) or a rolling cycle fashion (rcORF, light brown) into proteins that are not overlapped with their host genes but are homologous to other known proteins; blue pie slices: circRNAs translated into proteins that are not homologous to any known protein (cORF and rcORF combined); grey pie slices: circRNAs do not contain any potential ORF longer than 20 aa. Right panels: the same analysis of putative coding

product from two control circRNA sets, reversed sequences of endogenous circRNA and random shuffled sequences of endogenous circRNA.

(C) Potential functions of circRNA encoded proteins. The circular RNA encoded ORFs with overlap to host genes (left bar plot) and without overlap to host genes (right bar plot) were analyzed for the protein domains using InterPro package and SUPERFAMILY database. All predicted domains with e value less than 0.0001 in SUPERFAMILY database were plotted according to their number of occurrence. The pie chart in middle panel is same as that in Fig. 3B, and the percentages of circORFs with or without known domains were shown.

(D) Protein-protein interaction network for the putative circRNA-encoded protein isoforms. The host genes with a putative circRNA-encoded isoform (highly expressed, see details in method) were used as query to search the STRING-database for protein-protein interaction using default setting (with medium confidence, considering of all interaction evidences and discarding disconnected nodes), and the resulting network was clustered using MCODE tool.

Figure 4. Systematically identification of *trans*-factors that recognize IRES-like elements.

(A) Schematic diagram of RNA affinity purification. Biotin-labeled RNAs containing consensus motifs of IRES-like elements were incubated with HeLa cell lysate, and RNA-protein complexes were purified by streptavidin beads. The proteins were further identified by Mass spectrometry (see method for details).

(B) Identification of *trans*-factors bound by each RNA probe. The probes presenting five consensus motifs of IRES-like elements (red) and a control probe (blue) were used (see table S2 for full sequence). The total proteins eluded from each RNA probe were separated with SDS-PAGE, and each band was cut and analyzed by Mass spectrometry. The top three identified proteins in each band were labeled at right in red.

(C) Protein-protein interaction network of identified *trans*-factors. Top proteins bound by all RNA probes (i.e., IRES-like elements) were analyzed by STRING and clustered into two main groups by MCODE tool.

(D) Test the activity of *trans*-factors by tethering the candidate factors to circRNA translation reporter with specific Puf-RNA interaction. The circRNA reporter inserted with an 8-nt sequence (with weak IRES-activity by itself) were co-transfected with different Puf-fusion proteins into 293T cells. The resulting cells were collected 2 day after transfection to analyze the protein productions and RNA levels by western blot and RT-PCR. Different pairs of Puf proteins and 8-nt targets were used as specificity control. Puf-N2 can specifically bind AGUGUCAG, whereas Puf-N8 can specifically bind to GCGUCUGC.

(E) Validation of PABPC1 activity. The expression vector of PABPC1 and various control RBPs were co-transfected with circRNA translation reporter containing $(A)_{10}$ sequence before the starting codon, and the protein production were assayed 48 hours after transfection.

Figure 1. Extensive IRES-like elements can drive circRNA translation

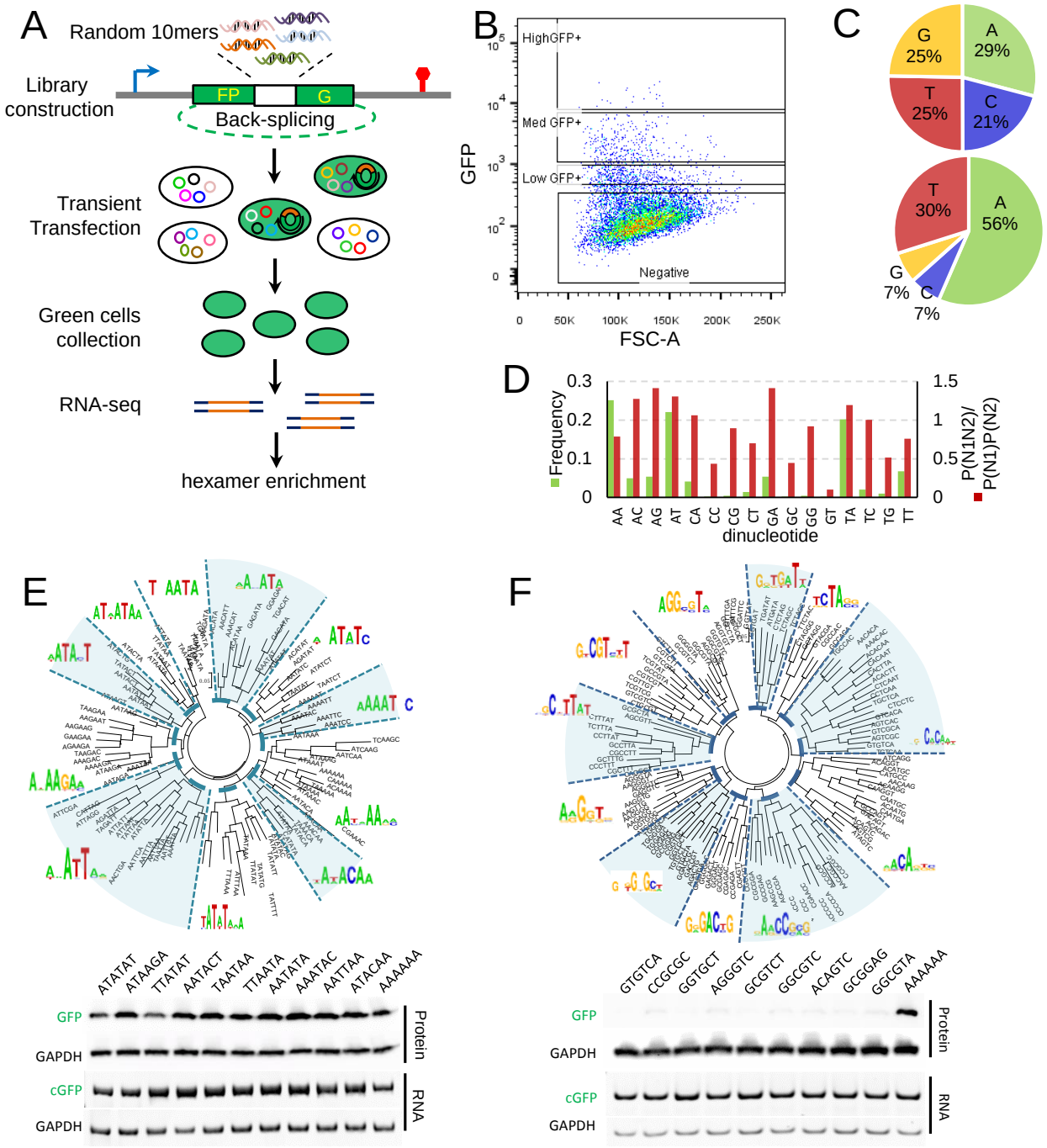

Figure 2. circRNAs contain many IRES-like elements that initiate translation

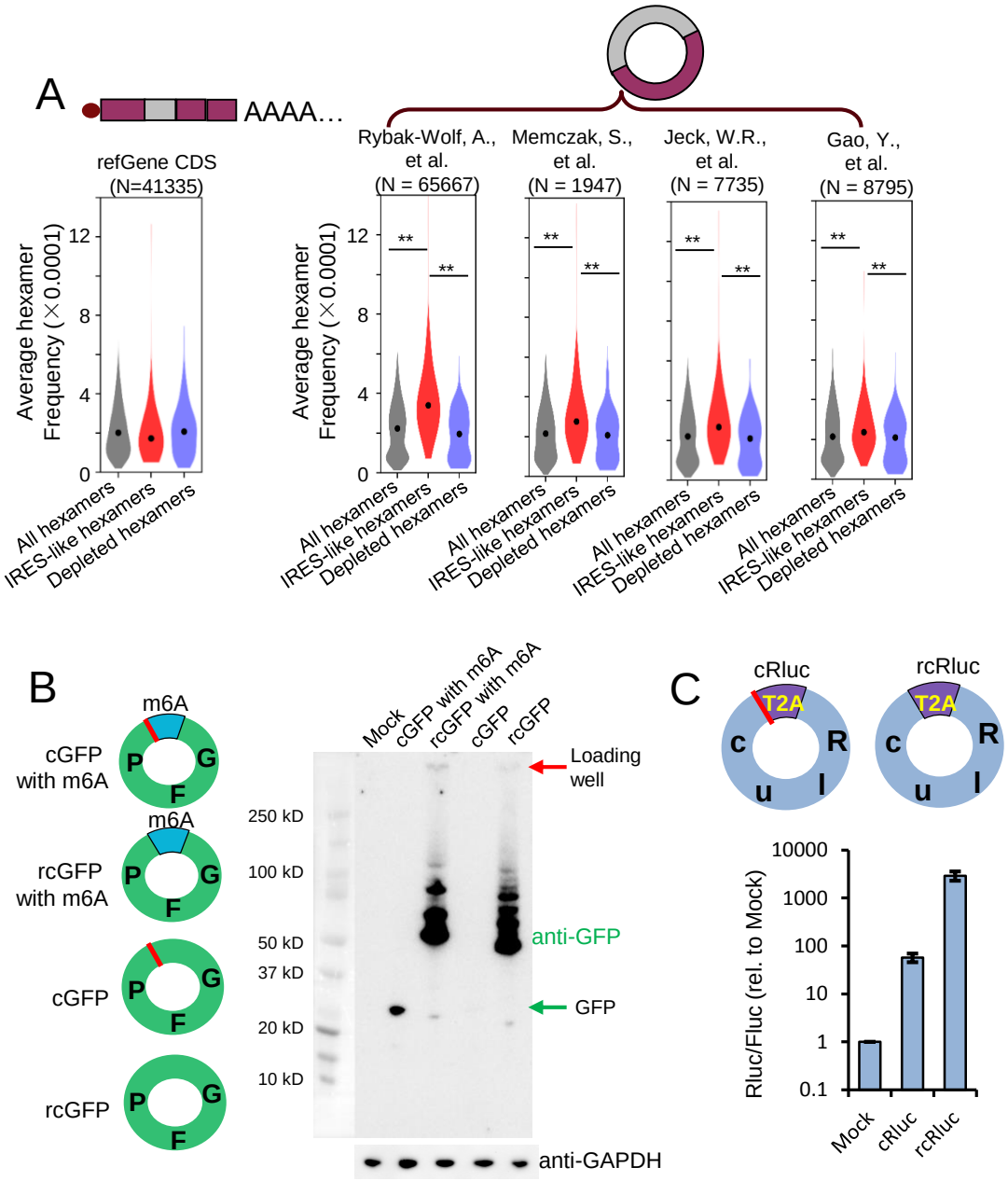

Figure 3. Functional analysis of putative circRNA encoded proteins

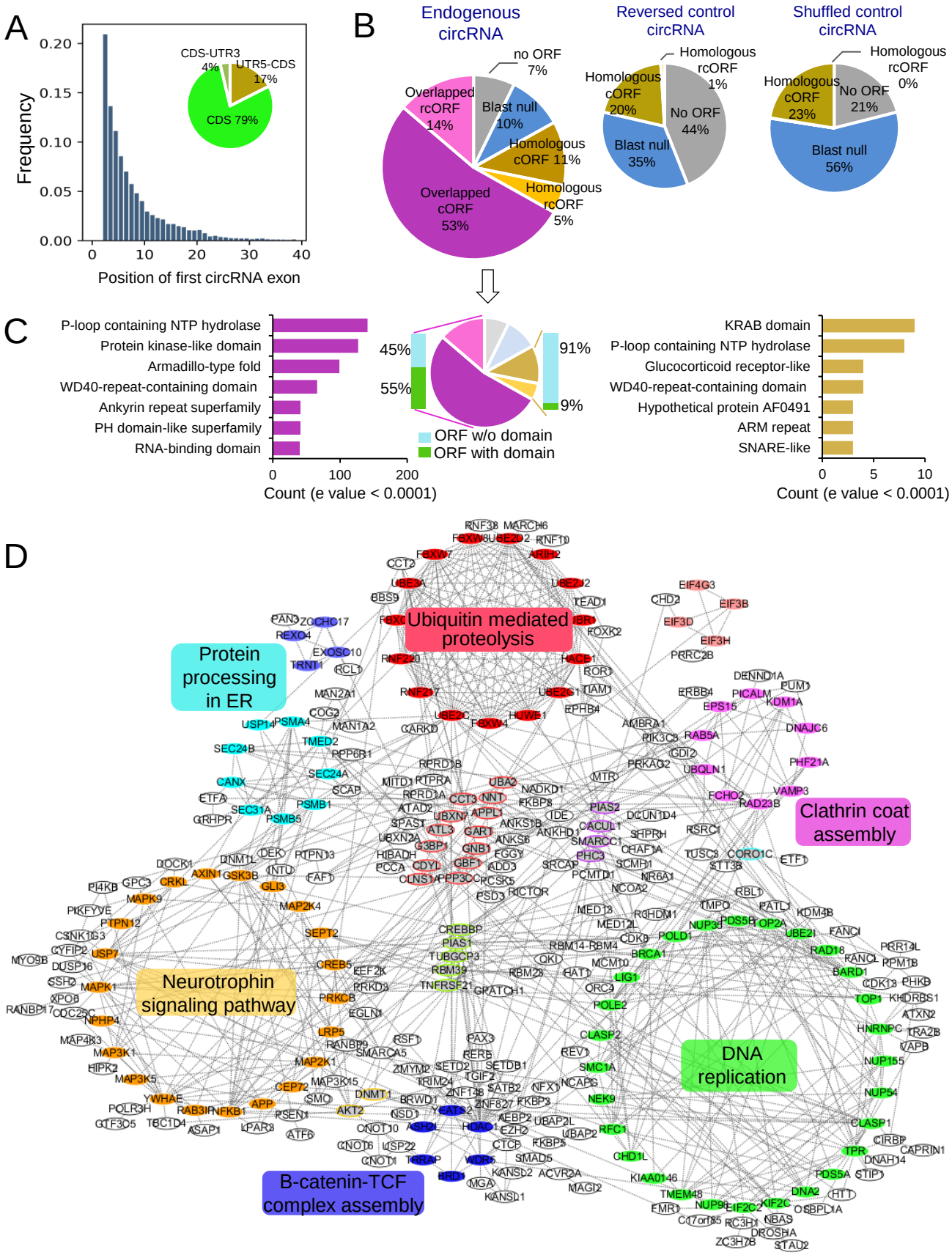
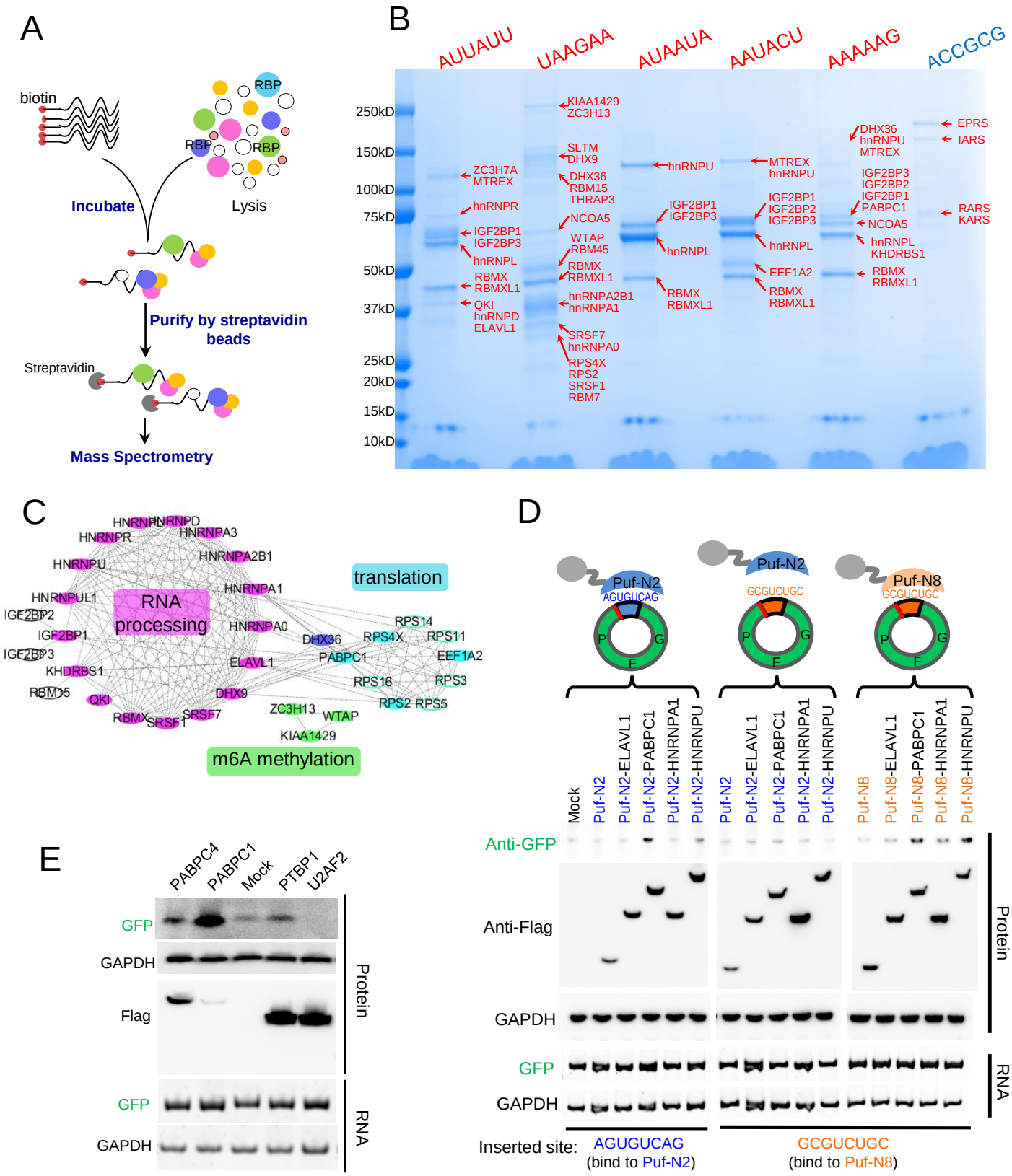
Figure 4. Systematically identification of *trans*-factors that recognize IRES-like elements