


Sequence analysis

# Biogenesis mechanisms of circular RNA can be categorized through feature extraction of a machine learning model

Chengyu Liu <sup>1,†</sup>, Yu-Chen Liu<sup>2,3,4,†</sup>, Hsien-Da Huang<sup>5,6,7</sup> and Wei Wang<sup>1,\*</sup>

<sup>1</sup>Department of Chemistry and Biochemistry and <sup>2</sup>Institute of Engineering in Medicine, University of California, San Diego, La Jolla, CA, USA, <sup>3</sup>Institute of Bioinformatics and Systems Biology, National Chiao Tung University, HsinChu, Taiwan, <sup>4</sup>Institute of Pharmacology, National Yang-Ming University, Taipei, Taiwan and <sup>5</sup>School of Life and Health Sciences, <sup>6</sup>Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, Longgang District, Shenzhen, Guangdong Province, 518172, China and <sup>7</sup>Shenzhen Bay Laboratory, Nanshan District, Shenzhen, Guangdong Province, 518172, China

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on December 13, 2018; revised on August 21, 2019; editorial decision on September 8, 2019; accepted on September 12, 2019

## Abstract

**Motivation:** In recent years, multiple circular RNAs (circRNA) biogenesis mechanisms have been discovered. Although each reported mechanism has been experimentally verified in different circRNAs, no single biogenesis mechanism has been proposed that can universally explain the biogenesis of all tens of thousands of discovered circRNAs. Under the hypothesis that human circRNAs can be categorized according to different biogenesis mechanisms, we designed a contextual regression model trained to predict the formation of circular RNA from a random genomic locus on human genome, with potential biogenesis factors of circular RNA as the features of the training data.

**Results:** After achieving high prediction accuracy, we found through the feature extraction technique that the examined human circRNAs can be categorized into seven subgroups, according to the presence of the following sequence features: RNA editing sites, simple repeat sequences, self-chains, RNA binding protein binding sites and CpG islands within the flanking regions of the circular RNA back-spliced junction sites. These results support all of the previously reported biogenesis mechanisms of circRNA and solidify the idea that multiple biogenesis mechanisms co-exist for different subset of human circRNAs. Furthermore, we uncover a potential new links between circRNA biogenesis and flanking CpG island. We have also identified RNA binding proteins putatively correlated with circRNA biogenesis.

**Availability and implementation:** Scripts and tutorial are available at <http://wanglab.ucsd.edu/star/circRNA>. This program is under GNU General Public License v3.0.

**Contact:** wei-wang@ucsd.edu

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Circular RNAs (circRNAs) represent an emerging type of regulatory RNA with 3' and 5' ends covalently join together as 'back-spliced junctions'. circ RNAs have been reported to have multiple functions. Beside serving as miRNA sponges, many circRNAs are important for brain function, synaptic plasticity (Westholm *et al.*, 2014; Rybak-Wolf *et al.*, 2015; You *et al.*, 2015) and fetal development (Szabo *et al.*, 2015). Cell free circRNAs are found stable in saliva (Bahn *et al.*, 2015) as well as exosomes (Li *et al.*, 2015), which

makes circRNA a promising diagnosis biomarker. Most circRNAs are originated from circularization of coding gene exons, which leads to the hypothesis that circRNA biogenesis competes with pre-mRNA splicing (Ashwal-Fluss *et al.*, 2014). Literature evidences suggest that the biogenesis of each circRNA subset may likely be regulated by different mechanisms (Chen and Yang, 2015; Conn *et al.*, 2015; Ivanov *et al.*, 2015; Jeck *et al.*, 2013; Li *et al.*, 2017; Liang and Wilusz, 2014; Zhang *et al.*, 2013, 2014, 2016), which supports the existence of multiple subclasses of circRNAs and each with specific roles.

In this study, we aim to decipher the relationship between the sequence features and circRNA subclasses. To this end, we have developed a computational model to predict whether a genomic locus would generate circRNA based on the presence of the following sequence features within the flanking regions of the circular RNA back-spliced junction sites: CpG islands, enhancers, RNA binding protein (RBP) binding sites, simple repeats, RNA editing sites and DNA self-chains. These features were selected based on the hypothesized circRNA biogenesis mechanisms (Conn et al., 2015; Ivanov et al., 2015; Jeck et al., 2013; Li et al., 2017; Liang and Wilusz, 2014; Zhang et al., 2013, 2014, 2016) as well as sequence features that can potentially participate in biogenesis of non-coding RNAs including CpG islands (Lai and Shiekhattar, 2014) and enhancer regions (Chen et al., 2017; Lam et al., 2014). We have designed a contextual regression model (Liu and Wang, 2017) that successfully distinguish the circRNA back-spliced junction sites defined by transcriptome sequencing from randomly selected human genome loci in an average 72.6% accuracy. Using the feature extraction technique in the contextual regression model (Khalid et al., 2014), we found that the examined 21 427 circRNAs can be categorized into 7 groups based on the biogenesis contributing factors. Our analysis supports that multiple biogenesis mechanisms co-exist for different subset of human circRNAs. In particular, we found 79 RBPs were identified to be significantly correlated to circRNA biogenesis. Interestingly, we uncovered a potential new link between circRNA biogenesis and flanking CpG islands, which suggests the potential correlation between DNA methylation and circRNA biogenesis.

## 2 Materials and methods

The data analysis process of this research is summarized in [Supplementary Figure S1](#). First, 55 689 human circRNAs back-spliced junction sites were collected from the database CircNet (Liu et al., 2016) as positive training data, and equal amount of randomly selected locus on HG19 human genome as negative training data. These junction sites and randomly selected locus were then divided into training and testing sets (in a ratio of 7:3) for the contextual regression network model designed to predict whether a randomly selected locus from human genome would generate circRNA. The features of the training set include whether CpG islands, enhancer regions, RBP binding sites, simple repeats, A-to-I RNA editing sites (RNA editing sites for short in the later text) and DNA self-chains present in the upstream and downstream region of the selected locus. After reaching optimum average accuracy, through the application of feature extraction techniques (Khalid et al., 2014), we successfully found that the examined 21 427 circRNAs can be categorized into 7 groups based on the presumed biogenesis contributing factors.

The back-spliced junction sites in the positive training set were selected from the database CircNet (Liu et al., 2016). The data include reported human back-spliced junction sites were collected from 22 recent studies (Alhasan et al., 2016; Bachmayr-Heyda et al., 2015; Bahn et al., 2015; Boeckel et al., 2015; Cheng et al., 2016; Conn et al., 2015; Dang et al., 2016; Gao et al., 2015; Guo et al., 2014; Jeck et al., 2013; Kelly et al., 2015; Memczak et al., 2013; Rybak-Wolf et al., 2015; Salzman et al., 2012, 2013; Song et al., 2016; Zhang et al., 2013, 2014, 2016; Zheng et al., 2016) and 465 human transcriptome sequencing datasets were collected from NCBI Sequence Read Archive (Leinonen et al., 2011). The back-spliced junction sites in each RNA-seq sample were identified using a circRNA discovery pipeline referred as find\_circ (Glazar et al., 2014; Hansen et al., 2016; Memczak et al., 2013). The criteria defined in the pipeline hence the detected junction sites met same standards as those in the previous reports, as described in the Memczak et al.'s (2013) study was applied. Adhering to the suggestion of recent year comparison study (Hansen et al., 2016), we selected the back-spliced junction sites based on the number of previous peer review reports in which the back-spliced junction sites were reported and the number of samples among the 465 collected samples in which the back-spliced junction sites were found meeting the criteria defined in find\_circ (Glazar et al., 2014; Hansen et al.,

2016; Memczak et al., 2015). Only the circRNAs with the sum of these two numbers  $> 3$  were considered as positive training data in this study. Locus of the CpG islands, enhancer regions, simple repeats, RNA editing sites and DNA self-chains on HG19 human genome was collected from UCSC Genome Browser (Casper et al., 2017). Although the locus of the RBP binding sites was collected from the Ray et al. (2013) study.

## 3 Results

Using the feature extraction technique in the contextual regression model (Liu and Wang, 2017), we found that the examined 21 427 circRNAs can be categorized into seven groups based on the biogenesis contributing factors. Our analysis supports that multiple biogenesis mechanisms co-exist for different subset of human circRNAs. In particular, we found 79 RBPs were identified to be significantly correlated to circRNA biogenesis. Interestingly, we uncovered a potential new link between circRNA biogenesis and flanking CpG islands, which suggests the potential correlation between DNA methylation and circRNA biogenesis.

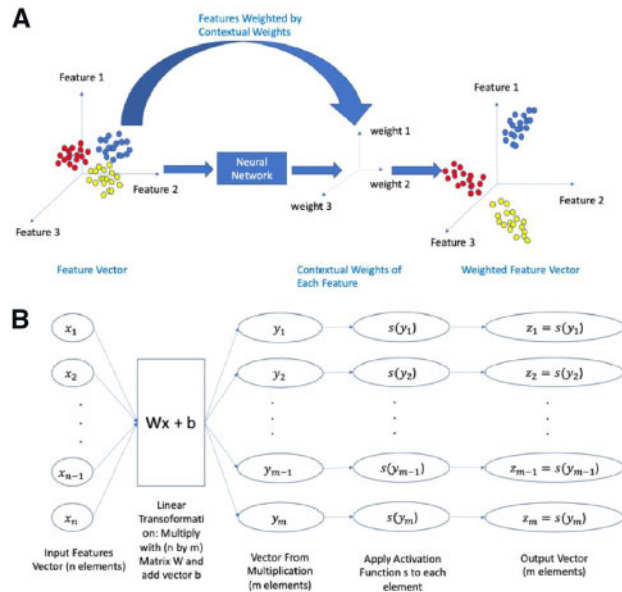
### 3.1 An interpretable neural network model to predict circRNAs

We implemented a contextual regression model to predict circRNA using these features. Instead of letting the neural network learn a function that maps features to target values, our method let the neural network learn a function that maps features to local linear models that best predict the target value from the features, thus generating a model that can both achieve state-of-the-art accuracy like a deep neural network while giving human interpretable quantification of feature contribution ([Fig. 1A](#)). As summarized in [Supplementary Figure S1C](#), in this contextual regression model, we used a residual neural network (He et al., 2016) that is composed of three layers of FNN (feed forward neural network, [Fig. 1B](#)) as the embedding function to generate the linear models. Batch normalization (Ioffe and Szegedy, 2015) is applied in the input layer to reduce the variance between input batches and make the training process more stable. The output of the embedding function is then dot-producted with the features and fed into a logistic function to output the prediction result. The FNN model is implemented as an operation that maps a vector  $x$  to  $s(Ax+b)$  where  $A$  is an matrix,  $b$  is an vector and  $s$  is the activation function. Both  $A$  and  $b$  are first initialized and then trained with tensorflow AdamOptimizer (Kingma and Ba, 2014). As for the parameter setting, all three layers of the FNN have 10 hidden units and tanh as their activation function, batch size is set to 50, max gradient norm to 10, learning rate to 0.0001. The weight matrix of each neural network is initialized with the tensorflow tf.truncated\_normal function with standard deviation of 0.05 to prevent vanishing gradient problem. The bias term in each layer is initialized all to value 0. We used cross-entropy as the loss function during training. To make the feature contribution easily interpretable, we applied a Lasso penalty in the form of L1 regularization on the context weight with penalty coefficient 0.0001.

### 3.2 Prediction and the feature selection results

As summarized in [Supplementary Table S1](#), the designed contextual regression model successfully distinguish circRNA back-spliced junction sites defined by transcriptome sequencing data from randomly selected human genome loci, in an average 72.6% accuracy and the area under curve of ROC curve 0.801 ([Supplementary Fig. S2](#)).

To extract the informative features, we selected the run with the highest accuracy (Run no. 5) from 10 runs. The difference between the accuracy on the training and testing sets were negligible (72.5 versus 72.8%), which suggested no overfitting. Therefore, we pooled together the training and testing sets in the feature analysis. We selected the most confidently predicted 21 427 circRNAs that have confident scores  $> 0.7$  to evaluate the contribution of each feature. The weighted feature contribution (WFC) was then obtained

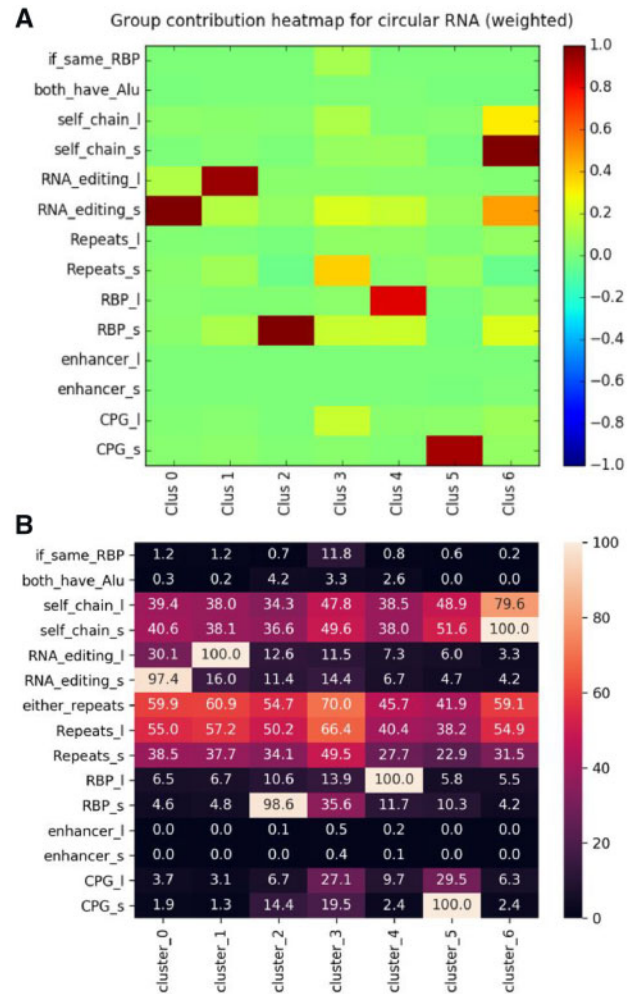


**Fig. 1.** Illustration of contextual regression. (A) Graphic illustration of the mechanism of contextual regression: the features are inputted into a neural network that generates a contextual weight for each feature which represents the importance of the features. Then, the features are then weighted by the corresponding weights to makes an easier separation of samples. In classification or regression tasks, the weighted features are then summed to yield the prediction. (B) A graphic demonstration of the FNN parts in the contextual regression model

from the model output. The features for each circRNA genesis mechanism (except ‘flanking short ALU repeats’ (both\_have\_AlU) and ‘binding sites of single RBP’ (if\_same\_RBP) since they only have 1 value for the whole region) were pooled into short (within 1000 bp radius from the circRNA) and long range (1000–2000 bp from the circRNA) by summing the WFC in each category for better data visualization. This resulted in 14 total feature contributions for each circular RNA data point (CpG short range, CpG long range, enhancer short range, enhancer long range, RBP short range, RBP long range, repeats short range, repeats long range, RNA editing short range, RNA editing long range, self-chain short range, self-chain long range, if head and tail both have Alu, if head and tail have the same RBP) that could be treated as a vector with 14 elements. Then, each of these vectors was normalized to the unit length and PCA was applied to the whole data point collection. The top 10 principal components were extracted which explained 98.9% of the total variance. Then a K-mean clustering was applied to separate the processed data into subclasses. Multiple values of  $k$  were experimented and  $k = 7$  was chosen for being the largest  $k$  that ensured all cosine similarities between cluster centers are  $< 0.5$ . The significantly contributing features were plotted in the heat map format and confirmed in the original feature data.

As a result, we found that these circRNAs could be clustered into seven different categories according to their biogenesis factors (Fig. 2A). To validate that the features with high contribution scores are enriched in each cluster, we calculated the percent enrichment of each feature in each cluster (Fig. 2B). The enrichment plot supported our clustering result.

In Cluster 0, RNA editing sites occurrence within 1000 nucleotide upstream or downstream of the circRNA locus was considered to be the most important factor for the biogenesis of 4576 circRNAs. For the other 4585 circRNAs in Cluster 1, appearance of RNA editing site within range of 1000–2000 nucleotides upstream or downstream of circRNA was considered as the most important biogenesis factor. Similarly, in Cluster 2, existence of RBP binding sites within short, which means within 1000 nucleotides upstream or downstream of the circRNA locus, was considered as the most important biogenesis factor for the 7503 circRNAs. For the 2342 circRNAs in the Cluster 4, occurrence in the long range, which



**Fig. 2.** Prediction and feature collection result. The result of the feature collection is summarized in this figure. (A) Through the result we found that these circRNAs can be into seven different categories according to their biogenesis factors. The long range features are marked with ‘\_l’ and the short range ones are marked with ‘\_s’. (B) Percentage of members in each cluster that contains each of the features

means within range of 1000–2000 nucleotides upstream or downstream of the circRNA locus, of RBP binding sites was suggested as the main biogenesis factor. Short repeat sequence flanking circRNA locus was clustered as the main factor for the 1344 circRNAs in Cluster 3. Biogenesis of 620 circRNAs was linked to flanking CpG islands, whereas 457 circRNAs’ biogenesis mechanism appeared to relate to the flanking DNA self-chains. The amount of these seven clusters is summarized in Supplementary Figure S3, while a complete list of the circRNAs is available in Supplementary Additional File S1. Through examining the input data that forms Clusters 2 and 4, which were associated with RBP, we identified 79 RBPs presumably participate in the biogenesis of these 9845 circRNAs. A complete list of the circRNA locus and associated RBP is available in Supplementary Additional File S2. These 79 RBPs (available in Supplementary Additional File S4) appear both in the short range (within 1000 nucleotides) in Cluster 2 and long range (from 1000 nucleotides to 2000 nucleotides) in Cluster 4 consistently. Among these RBPs, binding motifs of SRSF1, PUM1, SF3B4, ELAVL1, HNRNPA1, CSDA, SNRPA, RBFOX1 and PABPC1 were found appear in upstream or downstream of thousands of circRNAs in both clusters, as summarized in Supplementary Table S2.

Hence based on the result of this study, circRNAs can be categorized into multiple different subclasses, each with specific different function and biogenesis mechanism. Result of this research support all of previous discoveries and solidify the idea that multiple

biogenesis mechanisms co-exist for different subset of human circRNAs. Result of this study can also contribute to the advance of circRNA detection algorithm. Current mainstream research methods adopted to discover circRNAs are based on detection of back-spliced junction sites spanning reads within transcriptome sequencing data. However, this kind of approaches tends to have high false positive rate. Sensitivity of the junction site detection is also limited by sequencing depth. Had a clear concept of circRNA biogenesis mechanism is available, improved algorithm ruling out sequence base false positive might be able to be developed.

## Acknowledgements

We would like to thank Dr Shu Chien for his helpful comments and discussions.

## Funding

This work has been supported by the National Institutes of Health R01 [grant HG009626].

*Conflict of Interest:* none declared.

## References

- Alhasan, A.A. *et al.* (2016) Circular RNA enrichment in platelets is a signature of transcriptome degradation. *Blood*, **127**, e1–e11.
- Ashwal-Fluss, R. *et al.* (2014) circRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell*, **56**, 55–66.
- Bachmayr-Heyda, A. *et al.* (2015) Correlation of circular RNA abundance with proliferation-exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. *Sci. Rep.*, **5**, 8057.
- Bahn, J.H. *et al.* (2015) The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. *Clin. Chem.*, **61**, 221–230.
- Boeckel, J.-N. *et al.* (2015) Identification and characterization of hypoxia-regulated endothelial circular RNA. *Circ. Res.*, **117**, 884–890.
- Casper, J. *et al.* (2017) The UCSC genome browser database: 2018 update. *Nucleic Acids Res.*, **46**, D762–69.
- Chen, H. *et al.* (2017) Non-coding transcripts from enhancers: new insights into enhancer activity and gene expression regulation. *Genomics Proteomics Bioinformatics*, **15**, 201–207.
- Chen, L.-L. and Yang, L. (2015) Regulation of circRNA biogenesis. *RNA Biol.*, **12**, 381–388.
- Cheng, J. *et al.* (2016) Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics*, **32**, 1094–1096.
- Conn, S.J. *et al.* (2015) The RNA binding protein quaking regulates formation of circRNAs. *Cell*, **160**, 1125–1134.
- Dang, Y. *et al.* (2016) Tracing the expression of circular RNAs in human pre-implantation embryos. *Genome Biol.*, **17**, 1.
- Gao, Y. *et al.* (2015) CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.*, **16**, 4.
- Glažar, P. *et al.* (2014) circBase: a database for circular RNAs. *RNA*, **20**, 1666–1670.
- Guo, J.U. *et al.* (2014) Expanded identification and characterization of mammalian circular RNAs. *Genome Biol.*, **15**, 409.
- Hansen, T.B. *et al.* (2016) Comparison of circular RNA prediction tools. *Nucleic Acids Res.*, **44**, e58.
- He, K. *et al.* (2016) *Deep Residual Learning for Image Recognition*. arXiv Preprint arXiv: 1512.03385.
- Ioffe, S. and Szegedy, C. (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv Preprint arXiv: 1502.03167.
- Ivanov, A. *et al.* (2015) Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep.*, **10**, 170–177.
- Jeck, W.R. *et al.* (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, **19**, 141–157.
- Kelly, S. *et al.* (2015) Exon skipping is correlated with exon circularization. *J. Mol. Biol.*, **427**, 2414–2417.
- Khan, M.A. *et al.* (2016) RBM20 regulates circular RNA production from the Titin gene. *Circ. Res.*, **119**, 996–1003.
- Khalid, S. *et al.* (2014) In: *A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning*. IEEE.
- Kingma, D.P. and Ba, J. (2014) Adam: a method for stochastic optimization. *arXiv Preprint arXiv: 1412.6980*.
- Lai, F. and Shiekhata, R. (2014) Where long noncoding RNAs meet DNA methylation. *Cell Res.*, **24**, 263–264.
- Lam, M.T.Y. *et al.* (2014) Enhancer RNAs and regulated transcriptional programs. *Trends Biochem. Sci.*, **39**, 170–182.
- Leinonen, R. *et al.* (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
- Li, B. *et al.* (2017) Discovering the interactions between circular RNAs and RNA-binding proteins from CLIP-seq data using circScan. *bioRxiv*, 115980, doi: 10.1101/115980.
- Li, Y. *et al.* (2015) Circular RNA Is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res.*, **25**, 981–984.
- Liang, D. and Wilusz, J.E. (2014) Short intronic repeat sequences facilitate circular RNA production. *Genes Dev.*, **28**, 2233–2247.
- Liu, C. and Wang, W. (2017) Contextual regression: an accurate and conveniently interpretable nonlinear model for mining discovery from scientific data. *arXiv Preprint arXiv: 1710.10728*.
- Liu, Y.-C. *et al.* (2016) CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res.*, **44**, D209–D115.
- Memczak, S. *et al.* (2013) Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, **495**, 333–338.
- Memczak, S. *et al.* (2015) Identification and characterization of circular RNAs as a new class of putative biomarkers in human blood. *PLoS One*, **10**, e0141214.
- Ray, D. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172.
- Rybak-Wolf, A. *et al.* (2015) Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Mol. Cell*, **58**, 870–885.
- Salzman, J. *et al.* (2012) Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One*, **7**, e30733.
- Salzman, J. *et al.* (2013) Cell-type specific features of circular RNA expression. *PLoS Genet.*, **9**, e1003777.
- Shi, L. *et al.* (2017) Circular RNA expression is suppressed by androgen receptor (AR)-regulated adenosine deaminase that acts on RNA (ADAR1) in human hepatocellular carcinoma. *Cell Death Dis.*, **8**, e3171.
- Song, X. *et al.* (2016) Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res.*, **44**, e87.
- Suzuki, H. *et al.* (2006) Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from Pre-mRNA splicing. *Nucleic Acids Res.*, **34**, e63.
- Szabo, L. *et al.* (2015) Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.*, **16**, 126.
- Westholm, J.O. *et al.* (2014) Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Rep.*, **9**, 1966–1980.
- You, X. *et al.* (2015) Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nat. Neurosci.*, **18**, 603–610.
- Zhang, X.-O. *et al.* (2014) Complementary sequence-mediated exon circularization. *Cell*, **159**, 134–147.
- Zhang, X.O. *et al.* (2016) Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Res.*, **26**, 1277–1287.
- Zhang, Y. *et al.* (2013) Circular intronic long noncoding RNAs. *Mol. Cell*, **51**, 792–806.
- Zheng, Q. *et al.* (2016) Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nat. Commun.*, **7**, 11215.