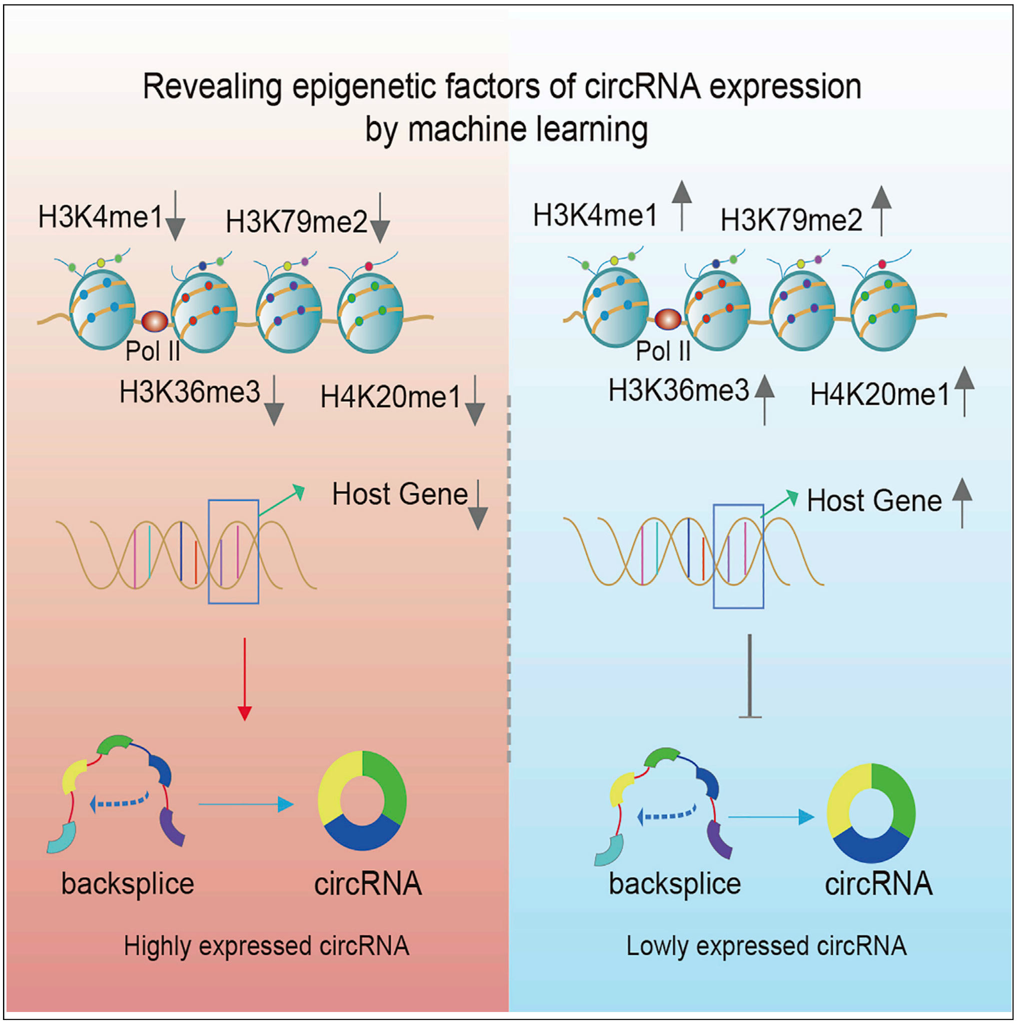
iScience

CellPress OPEN ACCESS

Article

Revealing Epigenetic Factors of circRNA Expression by Machine Learning in Various Cellular Contexts



Mengying Zhang, Kang Xu, Limei Fu, ..., Haozhe Zou, Yan Zhang, Yongsheng Li

zhangtyo@hit.edu.cn (Y.Z.)
liyongsheng@hainmc.edu.cn (Y.L.)

HIGHLIGHTS

CircRNAs exhibit specific expression in various cellular contexts

High and low expressed circRNAs exhibit different biological functions

Histone modifications are significantly correlated with circRNAs expression

Machine learning models were constructed for predicting circRNAs expression

Zhang et al., iScience *23*, 101842
December 18, 2020 © 2020 The Author(s).
https://doi.org/10.1016/j.isci.2020.101842

iScience

CellPress OPEN ACCESS

Article

Revealing Epigenetic Factors of circRNA Expression by Machine Learning in Various Cellular Contexts

Mengying Zhang,[1,4] Kang Xu,[1,4] Limei Fu,[1,2,4] Qi Wang,[1] Zhenghong Chang,[1] Haozhe Zou,[1,2] Yan Zhang,[3,*] and Yongsheng Li[1,2,5,*]

SUMMARY

Circular RNAs (circRNAs) have been identified as naturally occurring RNAs that are highly represented in the eukaryotic transcriptome. Although a large number of circRNAs have been reported, the underlying regulatory mechanism of circRNAs biogenesis remains largely unknown. Here, we integrated in-depth multi-omics data including epigenome, transcriptome, and non-coding RNA and identified candidate circRNAs in six cellular contexts. Next, circRNAs were divided into two classes (high versus low) with different expression levels. Machine learning models were constructed that predicted circRNA expression levels based on 11 different histone modifications and host gene expression. We found that the models achieve great accuracy in predicting high versus low expressed circRNAs. Furthermore, the expression levels of host genes of circRNAs, H3k36me3, H3k79me2, and H4k20me1 contributed greatly to the classification models in six cellular contexts. In summary, all these results suggest that epigenetic modifications, particularly histone modifications, can effectively predict expression levels of circRNAs.

INTRODUCTION

Circular RNA (circRNA) is a novel endogenous non-coding RNA that is common in the eukaryotic transcriptome (Glazar et al., 2014; Memczak et al., 2013; Salzman et al., 2013) and characterized by the presence of covalent bonds connecting the 3′ and 5′ ends (Jeck et al., 2013). Several circRNAs have been identified from a few transcriptional genes more than 30 years ago (Cocquerelle et al., 1993; Nigro et al., 1991; PG, 1996); the biogenesis is regulated by specific *cis*-acting elements and *trans*-acting factors (Kristensen et al., 2019). The cyclization of circRNAs is promoted by surrounding complementary sequences and regulated by specific RNA-binding proteins (Ashwal-Fluss et al., 2014; Conn et al., 2015; Ivanov et al., 2015; Liang and Wilusz, 2014; Zhang et al., 2014). In addition, both alternative splicing events within the same back-splice junction and alternative back-splice site selection can produce various circRNAs from the same gene locus (Gao et al., 2016). In the past few years, various studies have demonstrated that circRNAs are ubiquitous in animals and not the previously considered splicing by-product (Salzman et al., 2012).

Accumulating studies have shown that circRNAs play important roles in carcinogenesis and are expected to be therapeutic targets (Chen et al., 2019a; Han et al., 2018; Huang et al., 2020; Ju et al., 2019; Miranda et al., 2006; Qian et al., 2018). For example, a circRNA, CDR1 antisense RNA (CDR1as) (antisense to the cerebellar degeneration-related protein 1 transcript), was reported as miR-7 sponge and inhibited the function of miR-7 in colorectal cancer (Weng et al., 2017). *circMLL/AF9* that was derived from oncogenic fusion genes can contribute to tumor-promoting properties (Guarnerio et al., 2016). Moreover, the expression of circRNAs was extensively dysregulated in complex diseases. Cell cycle-related *circTP63* was up-regulated in lung squamous cell carcinoma (LUSC) tissues, and its up-regulation was directly correlated with larger tumor size and higher tumor node metastasis (TNM) stage in patients with LUSC (Cheng et al., 2019). In addition, *circ*ASAP1, a circRNA derived from exons 2 and 3 of the ASAP1 gene, was overexpressed in hepatocellular carcinoma (HCC) cell lines with high metastatic potential and in metastatic HCCs (Hu et al., 2019).

Recent studies have shown that RNA-binding protein (RBP) is a key regulator of the expression pattern of circRNAs. Such as QKI (Conn et al., 2015), and DExH-box helicase 9 (DHX9) (Aktas et al., 2017), they have

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

[2] Key Laboratory of Tropical Translational Medicine of Ministry of Education, Hainan Medical University, Haikou 571199, China

[3] School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, China

[4] These authors contribute equally to this work.

[5] Lead Contact

*Correspondence: zhangtyo@hit.edu.cn (Y.Z.), liyongsheng@hainmc.edu.cn (Y.L.)

https://doi.org/10.1016/j.isci.2020.101842

Cell Line	Number of circRNA	Number of Highly Expressed circRNA	Number of Lowly Expressed circRNA
A549	7,972	510	597
GM12878	10,767	657	662
H1-hESC	7,782	427	441
HepG2	8,278	319	322
Hela-S3	8,471	330	331
NHEK	3,131	96	97
Total	23,989	1,276	1,903

Table 1. Number of CircRNAs across Six Cell Lines

been accumulatively reported to regulate the formation of circRNAs and are key players in post-transcriptional events (Pereira et al., 2017). In addition, efforts in *D. melanogaster* suggest that the biogenesis of many circRNAs is influenced by a combination of *cis*-acting elements and *trans*-acting splicing factors, including heterogeneous nuclear ribonucleoproteins (hnRNPs) and SR proteins (that is, proteins containing a long repeat of serine and arginine amino acid residues) (Kramer et al., 2015). Furthermore, circRNAs may be specifically generated and regulated from a study using (estrogen-stimulated) MCF-7 cells, which show higher levels of H3K36me3 and a higher number of Ago-binding sites in circularizing exons (Tarrero et al., 2018).

Recent studies have shown that histone modifications affect the splicing mechanisms and splicing outputs by recruiting splicing regulators that affect chromatin-binding proteins (Luco et al., 2010). However, little is known about whether there exists specific regulatory mechanism between histone modification and circRNA expression. To fill this gap, we systematically analyzed global circRNAs expression in a large panel of cell lines from the Encyclopedia of DNA Elements (Consortium, 2004) (ENCODE). The origins and distribution of circRNAs on the chromosome were identified and analyzed in various cellular contexts. Next, epigenetic modification signals, mainly histone modifications, and expression of circRNA-related host genes were used to characterize high or low expression levels of circRNAs. We identified five important factors that can markedly distinguish the expression of circRNAs. Furthermore, we explored the relationship of epigenetic factors and circRNA expression in each cell line. In summary, we initially investigated the regulation relationship between circRNA expression and histone modifications, which provides an important reference for further study of circRNA biogenesis.

RESULTS

Identification of circRNAs in Various Cellular Contexts

We analyzed circRNA transcripts using RNA sequencing (RNA-seq) of ribosomal RNA-deplete RNA from six cell lines (A549, GM12878, H1-hEsc, HepG2, HeLa-S3, and NHEK). A detailed summary for the samples used in this study was provided in Table S1 and Table S2. The pipeline based on BWA-MEM alignment was used to identify circRNAs with gene annotations (Figure S1). In total, we identified 23,989 unique candidate circRNAs in all six cell lines. The numbers of circRNA identified in each cell line was provided in Table 1. Compared with previously identified circRNAs that were downloaded from circBase (Glazar et al., 2014), we found that there are 14,499 known circRNAs and 9,490 circRNAs identified in our current study (Table S3). Moreover, compared with circRNAs identified in circAtlas (Wu et al., 2020) and circRIC (Ruan et al., 2019), we also found that the majority of circRNAs has been identified in these databases (Table S4). Thus, all these results suggested that circRNA is the RNA family that plays an important role, not a by-product of splicing. Notably, the differences in sequencing depth, variable identification methods, and differences in the developmental stage of cells or tissues may all contribute to the final discovery of novel circRNAs.

Next, we annotated these circRNA candidates using the RefSeq database (Pruitt et al., 2012). Previous work has revealed the variant types of circRNAs in genomic regions (Memczak et al., 2013; Zheng et al., 2016). We

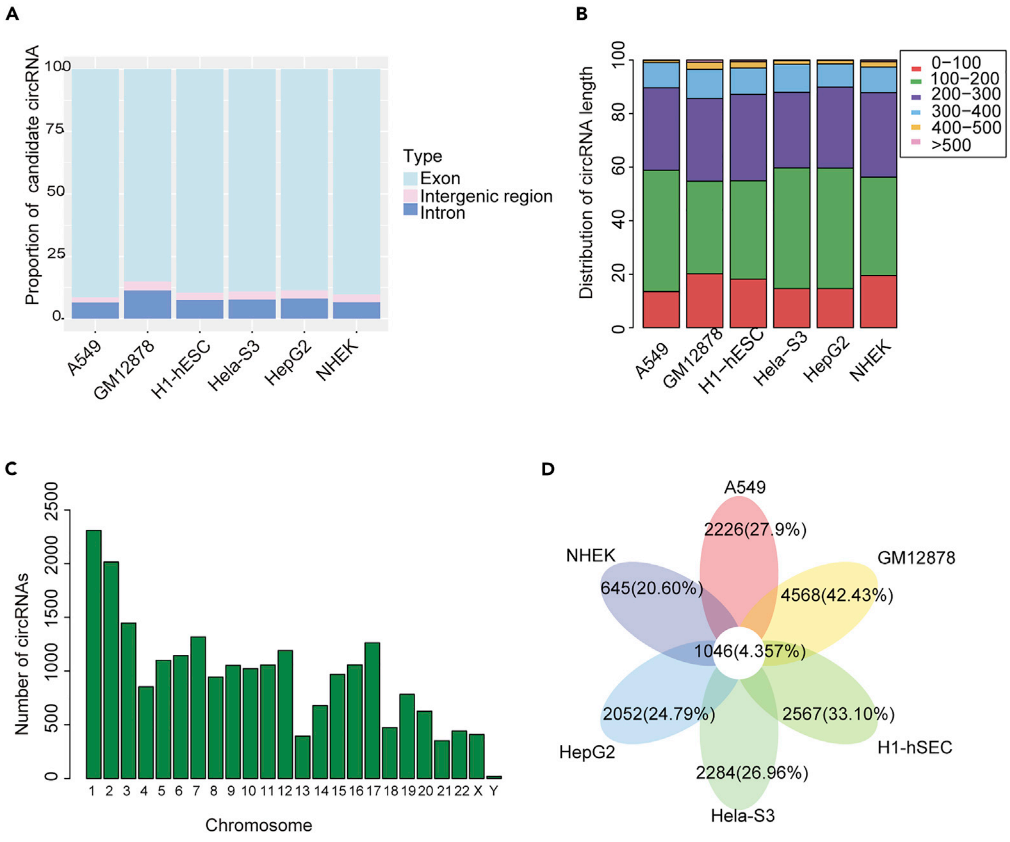


Figure 1. Profiling of Circular RNAs in Different Cell Lines

(A) Barplot showing the proportion of circRNAs with variant origins. Light blue for exon circRNAs, pink for intergenic circRNAs, and blue for the intron circRNAs.

(B) The splice length distribution of circRNAs in six cell lines.

(C) Bar graph showing the distribution of circRNA on different chromosomes.

(D) Venn diagram showing candidate circRNAs across six cell lines. See also Figures S1 and S2 and Tables S1 and S2, Tables S3 and S4.

found that more than 85% of circRNA candidates were derived from exon, whereas smaller fractions aligned with introns, intergenic region (Figure 1A). We analyzed the length of exon circRNAs and found that the length in most of the six cell lines was less than 500 bp (Figure 1B). We also reconstructed the full-length sequences of circRNAs by CIRI-full (Zheng et al., 2019). We found that the average length distribution of circRNA was similar, and the majority of circRNAs was around 200 bp in length (Figure S2). Moreover, no significant correlation was observed between the distribution of candidate circRNAs and chromosome length and number of genes (Figure 1C). Next, we analyzed the number of circRNAs shared by six cell lines. Interestingly, we rarely found that the number of common circRNAs in six cell lines, which accounted for approximately 4.36% of all circRNAs. However, in a single cell line, such as GM12878, cell-type-specific circRNAs accounted for 42.43% of the total number (Figure 1D and Table S5). These findings were consistent with previous studies that a vast majority of circRNAs show tissue and cell type specificity (Salzman et al., 2013; Szabo et al., 2016). In addition, we focused on the differences between shared and specific circRNAs, including the expression status and the type of circRNAs. We found that shared circRNAs had a higher proportion of exon circRNAs and significantly higher length and expression than specific circRNAs (Figure S3).

Quantitative Analysis of circRNA Expression

Next, we examined the expression abundance of circRNAs in cell lines. We first divided circRNAs into two classes with different expression patterns, according to the expression values. We defined circRNAs with top 20% expression value as high expression circRNAs, the bottom 20% expression value as low expression

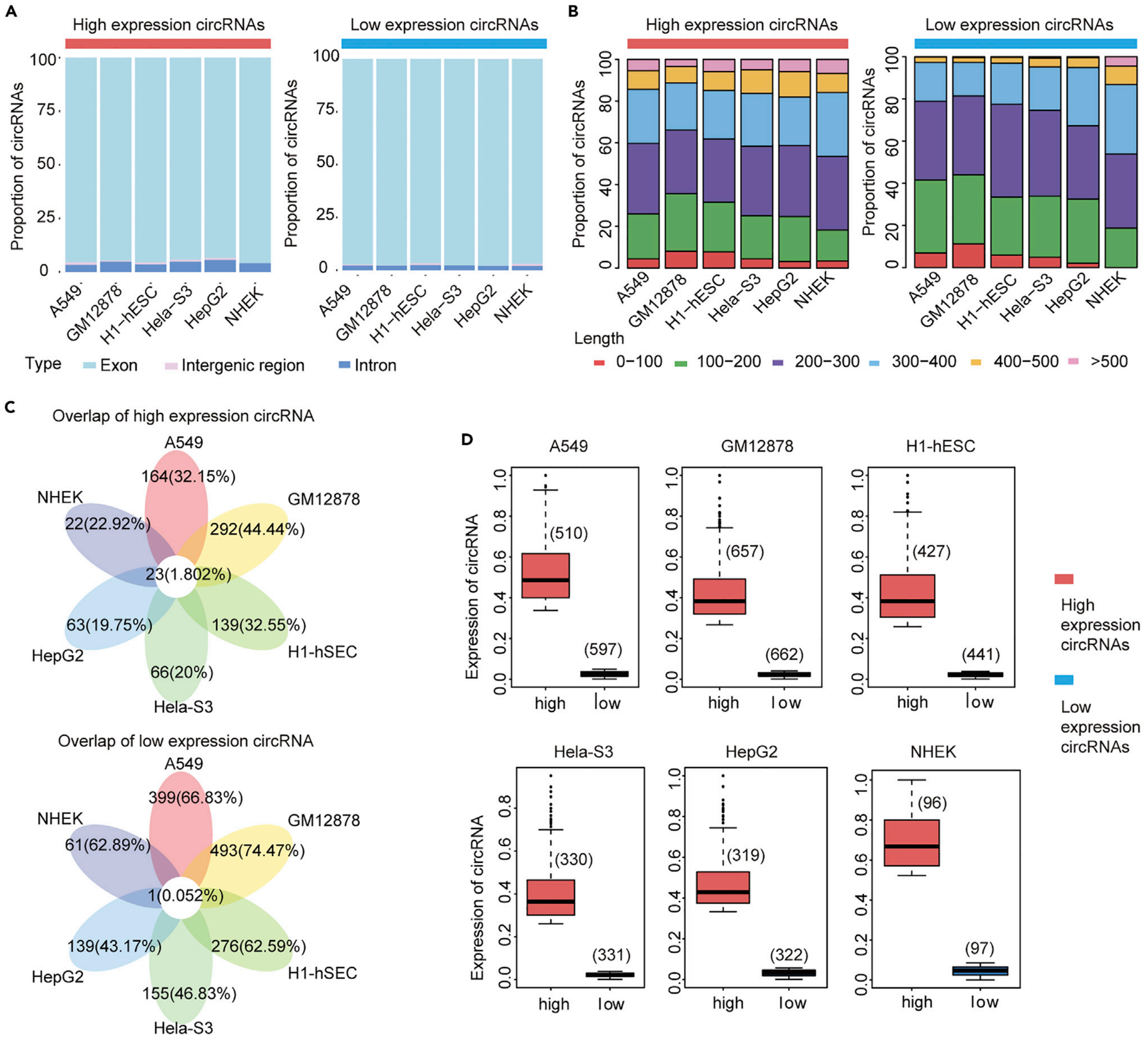


Figure 2. Profiling of High and Low Expression circRNAs in Six Cell Lines

(A) A bar graph showing the proportion of different types of high- and low-expression circRNAs. The left part is highly expressed circRNA, and the right part is low expression circRNA.

(B) The splice length distribution of circRNAs with different expression pattern in six cell lines. The left part is a high expression circRNA, and the right part is a low expression circRNA.

(C) The Venn diagram of high and low expression circRNA intersection in six cell lines.

(D) Box plot showing expression value and number of high and low expression circRNAs. See also Figures S3–S5 and Table S5.

circRNAs (Figure S4). The numbers of circRNA with different expression patterns identified in the six cell lines were listed in Table 1. We next explored the differences in expression patterns of circRNAs across six cell lines. Consistent with previous studies, exon circRNAs were the main type of circRNAs (Figure 2A). The length of most exon-circRNAs was <500 bp, and the median length was about 200 bp (Figure 2B). Moreover, the number of shared circRNAs across cell lines was quite small. This result revealed that numerous circRNAs seem to be specifically expressed across various cells. Notably, the circRNAs shared by cell lines in the low expression pattern was less than those in high expression pattern. This indicates that expression specificity of circRNA is higher in the low expression pattern (Figure 2C). A box plot of circRNA expression was shown in Figure 2D.

Although studies have shown that splice length of circRNA in the genome is relatively short, the number and length of exon-forming circRNAs are obviously different. Next, we explored whether the expression of circRNA is affected by the genome length of circRNA. Spearman correlation was performed for length and expression of all identified circRNAs. The analysis revealed that there was no significant correlation between length and expression of circRNA (Figure S5). This result indicated that expression of circRNAs is not actually affected by length. The expression levels of circRNA in cells varied greatly, suggesting that they play different functional roles in different cellular contexts.

Next, Gene Ontology (GO) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were performed to infer potential functions of circRNA. We identified terms that were over-represented or under-represented in the two expression patterns. Our analysis found that high expression circRNAs tend to be enriched in biological processes, such as receptor's metabolic process, negative regulation of mRNA 3-end processing, histone H3-K27 methylation, regulation of small GTPase-mediated signal transduction, and cell differentiation (Figure 3A). Low expression circRNAs were functionally enriched in positive regulation of gene silencing by miRNA, negative regulation of histone H3-K27 acetylation, histone H4-K20 methylation, cAMP metabolic process, and regulation of substrate adhesion-dependent cell spreading. Moreover, KEGG pathway results suggested that high expression circRNAs may be involved in mismatch repair pathway and basal transcription factors (Figure 3B). Notably, circRNAs in low expression groups were enriched in pathways in cancer, which was consistent with the previous conclusion that circRNAs in cancers tend to be low expression (Vo et al., 2019). Molecular function and cell components enrichment also showed differences between circRNAs in two groups (Figures S6 and S7). Taken together, these results suggest that there may be some relationship between the expression patterns of circRNA and histone modifications. It also shows that two different patterns of circRNA perform different biological functions in the life of an organism.

Epigenetic Features Are Predictive of circRNA Expression across Cellular Contexts

As an important epigenetic regulator, histone modification has been widely studied in recent years. Emerging evidence has shown important roles of histone modifications in circRNA biogenesis (Tarrero et al., 2018). To further explore the relationship between circRNA expression and histone modifications, we downloaded histone modification signals shared by six cell lines from ENCODE project. The host gene expression was obtained from RNA-seq using Cufflinks (Trapnell et al., 2010) (see Transparent Methods). We hypothesized that the binding peaks of histone modification obtained from ENCODE would be different between high expression circRNAs and low expression circRNAs. To test this hypothesis, we applied 11 histone modification peaks and expression of host gene to characterize circRNA with different expression patterns. For the robustness of classification model, the datasets were divided into training sets and testing sets. The models trained in training sets were applied to the testing sets to evaluate the robustness.

It is well known that the A549 cell line is human lung adenocarcinoma cell line. In recent years, circRNA expression in A549 has been extensively studied (Dai et al., 2018; Sun et al., 2019). We first constructed five frequently used machine learning classification algorithms (decision tree, logistic regression, SVM, naive Bayes, and random forest) to predict circRNA expression levels in A549 cell line. Indeed, those algorithms have different classification power on circRNA expression prediction. We evaluated the power of the classification models combining with three indicators, including the precision, accuracy, and area under the ROC curve (AUC) of the classifier (Figure S8). All these results suggest that classification model of random forest was the best one. Therefore, random forest was used to integrate the histone modification peaks and host gene expression for circRNA prediction.

Next, random forest was applied to train the model in six cell lines. First, our datasets were divided into training sets and testing sets (see Transparent Methods). We performed 10-fold cross-validation on our datasets to verify that the correlation was not specific to a subset of data. Finally, the AUC indicated that both the training sets and the testing sets model are robust in the six cell lines (Figure 4). For example, in A549 cell line, the AUC reached 0.89 in the training sets, whereas the AUC of the model classification was as high as 0.95 in the testing sets. In the GM12878 cell line, the area under the ROC curve was 0.911 in the training sets, and this value reached 0.845 in the testing sets. The accuracy of the model in the training sets of the H1-hESC cell line was 0.897 and it was 0.78 in testing sets. Similarly, in the Hela-s3, HepG2, and NHEK cell-lines, the AUC predicted by the training sets and the testing sets model all reached 0.8 or more. All these

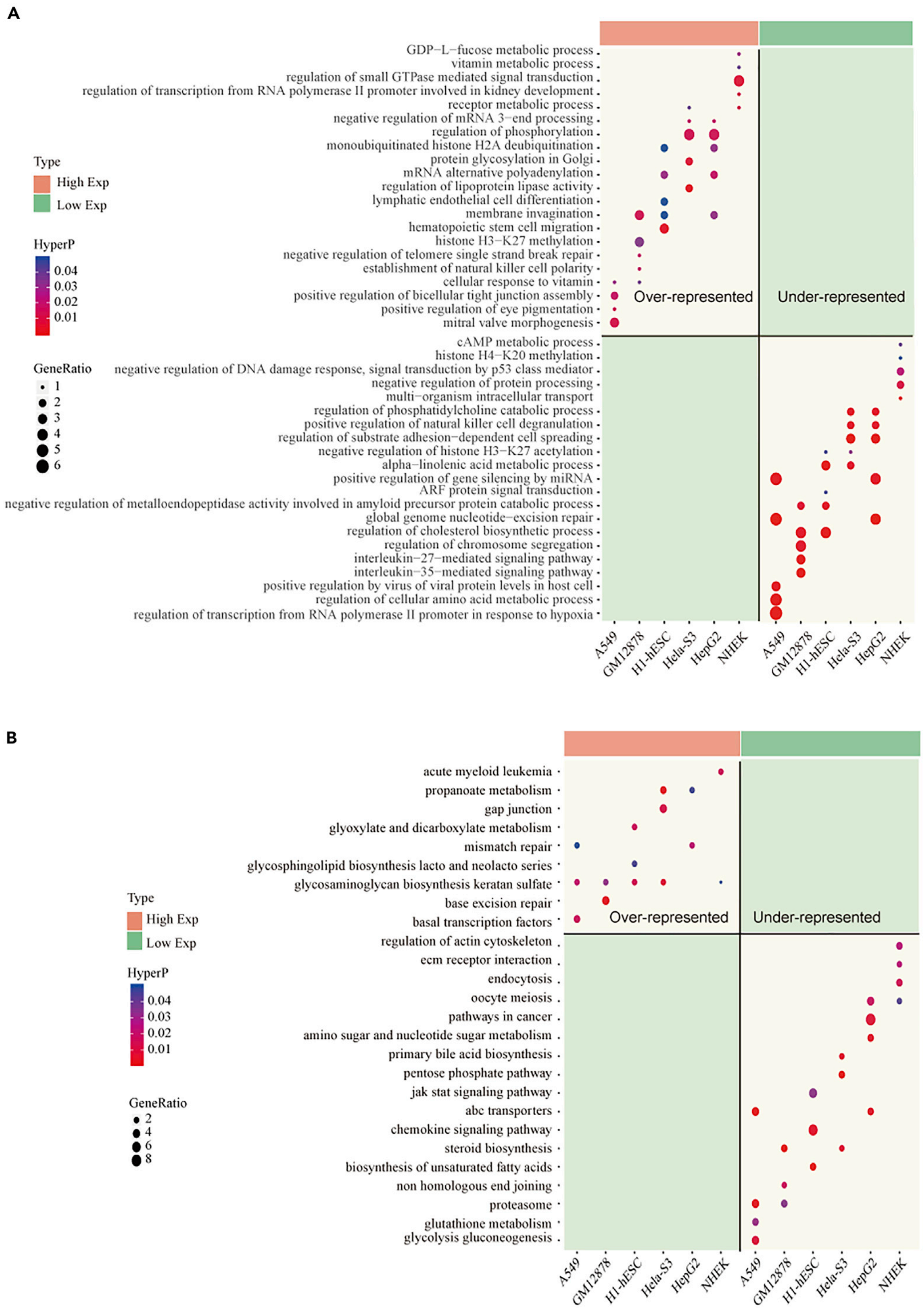


Figure 3. Over-Represented and Under-Represented Functions of circRNA with High and Low Expression Classes

(A) Enriched bubble diagrams of high and low expression patterns in biological processes. The portion covered by the red band represents a high expression circRNA, and the portion covered by the green band represents a low expression circRNA. The color of the bubble represents the p value; bubble size represents the number of circRNA host genes that present in one term.

(B) Bubble diagrams of high and low expression patterns in the KEGG pathway. See also Figures S6 and S7.

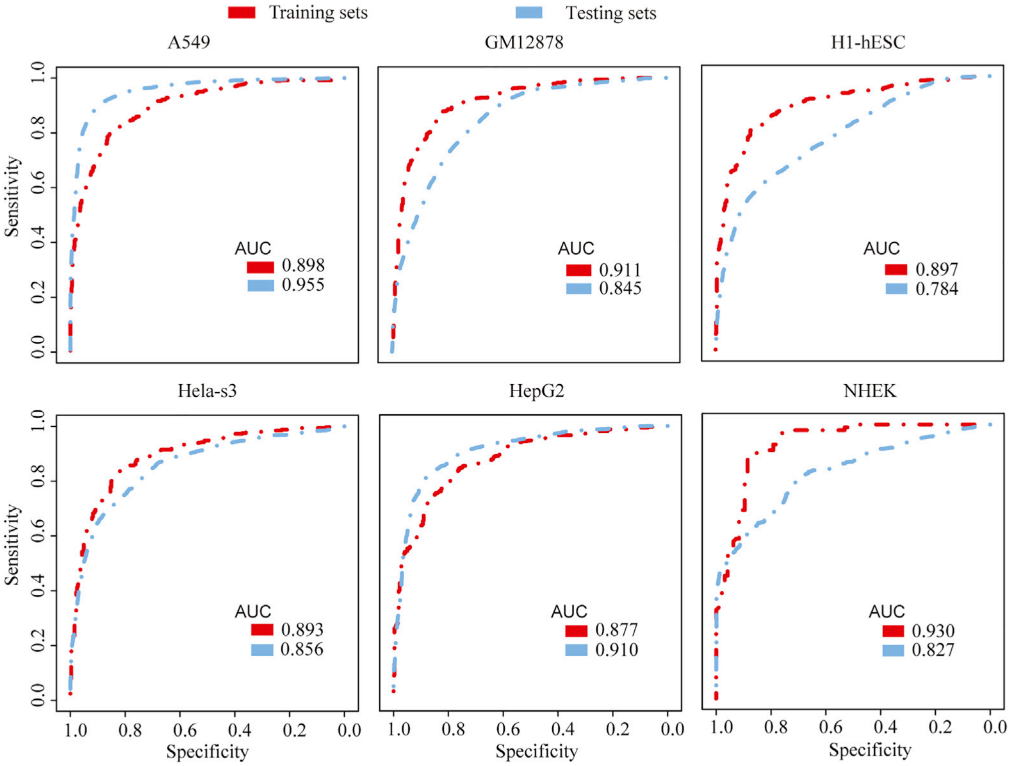


Figure 4. The Area under the ROC Curve of Random Forest Classification for Each Cell Line

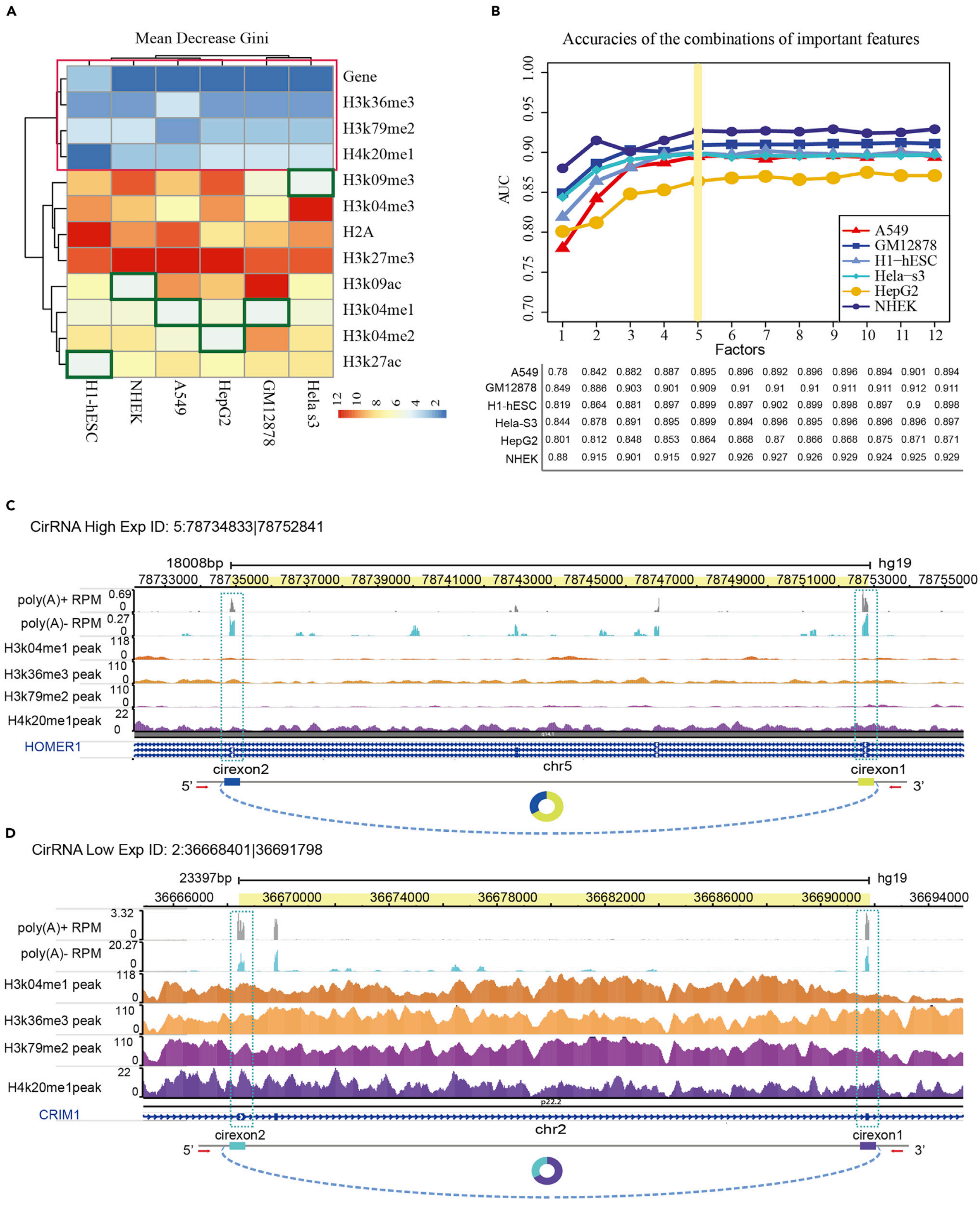
The predictive ability of 12 factors to characterize the different expression patterns of circRNA. The red line represents the AUC value in the training sets, and the blue line represents the AUC value in the testing sets. See also Figure S8.

results demonstrated that histone modifications were indicative of genomic circRNA expression in different cell types.

Epigenetic Determinants of circRNA Expression

The mean decrease Gini (MDG) was used to evaluate the importance of each feature to provide a relative ranking of the investigated features. The larger MDG indicates the increasing importance of the corresponding feature for prediction of circRNA expression patterns. Several important factors were revealed according to MDG in the random forest classifier (Figure 5A). The colors changed from blue to red and represented the decreasing importance of the factor. The host genes' expression, H3K79me2, H3K36me3, and H4K20me1, all contributed greatly in the classification of each cell line; these four factors showed a higher MDG and were identified as shared potential effectors (SPEs) (the factor in the red box in Figure 5A). Next, to avoid factor redundancy events in the model construction, we selected the top n (n = 1, 2, 3, 4 … 12) important factors in each cell line to construct the classifier to evaluate the prediction ability of the model (Figure 5B). Although the AUC varied in each cell line, when the factors of the constructed model reached five, the AUC value of the model tended to be stable. Therefore, we identified five important factors in each cell line (red box in Figure 5A and the factor represented by the green box) as maker factors for different patterns of circRNA expression.

Moreover, we confirmed our result in another independent validation sets (see Transparent Methods). We collected the available RNA-seq datasets from the Gene Expression Omnibus database (GEO) (Edgar et al., 2002) and downloaded the peaks of 11 histone modifications from The NIH Roadmap Epigenomics Mapping Consortium (http://www.roadmapepigenomics.org/). We obtained a total of five public data resources of each cell line (A549, GM12878, H1-hESC, HeLa-S3, HepG2). We next selected top five factors in each cell line for model construction. The results showed that the AUC was above 0.85 (Figure S9). These results further validated the reliability of our models and also showed that the top five factors may affect the expression pattern of circRNAs.



A549	0.78	0.842	0.882	0.887	0.895	0.896	0.892	0.896	0.896	0.894	0.901	0.894
GM12878	0.849	0.886	0.903	0.901	0.909	0.91	0.91	0.91	0.911	0.911	0.912	0.911
H1-hESC	0.819	0.864	0.881	0.897	0.899	0.897	0.902	0.899	0.898	0.897	0.9	0.898
Hela-S3	0.844	0.878	0.891	0.895	0.899	0.894	0.896	0.895	0.896	0.896	0.896	0.897
HepG2	0.801	0.812	0.848	0.853	0.864	0.868	0.87	0.866	0.868	0.875	0.871	0.871
NHEK	0.88	0.915	0.901	0.915	0.927	0.926	0.927	0.926	0.929	0.924	0.925	0.929

Figure 5. Model Evaluation Capabilities of Five Classifiers and Contribution of Factors to the Model

(A) MDG of each factor, the color from blue to red represents the importance of high to low, the column represents six cell lines, and the row represents 12 factors. The red box factors are SPE of each cell line, and the green box represents the fifth most important factor.

(B) Model predictive ability assessment of top n factors combination, the yellow line marks the fifth contribution factor, and the area under the ROC curve of the factor has stabilized.

(C and D) The circRNA visualization of four key histone modification signals, the origin of exons, their expression levels from the poly(A) RNA-seq (blue wiggle tracks), and the expression of their cognate mRNAs from the poly(A)+ RNA-seq (gray wiggle tracks) in A549 cells. See also Figure S9.

To investigate the differences in the modification signal of the five factors in circRNA, we selected high and low expression circRNA in A549 cells related to lung cancer in the circBase database. For example, high expression circRNA *5:78734833|78752841* and low expression circRNA *2:36668401|36691798* (Figures 5C and 5D). circRNA *5:78734833|78752841* is a known *circ0006916*, which is derived from the homer scaffold protein 1 (HOMER1) gene and is highly expressed in the A549 cell line (Dai et al., 2018). The circRNA *2:36668401|36691798* is known as *circ0007386.* These two circRNAs are formed by back splice of two exons on chromosome 5 and chromosome 2. Visualization shows that poly(A)+ RNA expression is higher in low expression pattern. These results demonstrated that the expression of circRNA and linear isoforms were independent of each other. The variations of H3K4me1, H3K36me3, H3K79me2, and H4K20me1 signals were strongly correlated with the circRNA expression dynamics. The signal of histone modification tended to be at a lower level in highly expressed circular RNA.

Modeling the circRNA Expression Based on Epigenetic Features

We used random forest to screen five important factors based on MDG to characterize circRNA. Four of the important factors are common to each cell line. We next explored the existing connection of important factors and circRNA expressions. We subdivided the expression of circRNA and observed the trends of the five signal factors at different expression levels (Figures 6A–6E). In the A549 cell line, all five factors showed negative correlation with circRNA expression, indicating that the expression of host gene and four epigenetic factors has a negative regulatory relationship to circRNA expression (Figure 6). The host gene, H3k36me3, H3k79me2, H4k20me1 in the GM12878 cell line showed the negative correlation with circRNA expression, whereas H3K4me1 peak and circRNA expression revealed weak promotion (Figure S10). In the H1-hESC cell line, host gene, H3k36me3, H3k79me2, H4k20me1, H3K4me1, and circRNA expression showed a negative correlation trend (Figure S11). In the Hela-s3 cell line, the expression of five factors with circRNA was consistent with H1-hESC, and both factors were negatively regulated (Figure S12). The relationship between each factor with the expression of circRNA in HepG2 cell line was similar to that of GM12878 cell line. The expression of host gene, H3k36me3, H3k79me2, H4k20me1, and circRNA showed a negative correlation, whereas the expression of the fifth important factor H3k04me2 in HepG2 cell line displayed positive promotion with circRNA expression (Figure S13). In the NHEK cell line, the top four shared factors, including host genes expression, H3k36me3, H3k79me2, and H4k20me1, were negatively correlated with circRNA expression, whereas H3k9ac and circRNA expression show a positive promotion (Figure S14).

In total, we identified four important factors shared by the cell lines, including host gene expression, H3k36me3, H3k79me2, and H4k20me1, were negatively correlated with the expression of circRNA (Figure 6F). We speculated that the expression of circRNA may be negatively affected by these four factors. The fifth important factor was the cell-specific factor, and the expression of circRNA showed an inconsistent trend. It suggested that this feature may have a cell line-specific relationship with circRNAs.

DISCUSSION

In this study, we identified and analyzed the global landscape of circRNAs in six different cell lines and found large and different genomic locations of circRNAs that are specifically expressed in tissues or cell lines as in previous studies (Guo et al., 2014; Salzman et al., 2013). To measure the expression of circRNA, we used the junction ratio as measure of the relative expression value of circRNA by CIRI (Gao et al., 2015). Consistently, we found that most of the identified circRNA expressions were of less abundance. This may be one of the reasons why circRNA has long been considered a by-product of the pre-mRNA splice process (Li et al., 2017). Importantly, to ensure the accuracy of analysis, we also performed CIRIquant (Zhang et al., 2020) for circRNA quantification and analyzed the Spearman correlation between the junction ratio in CIRI and CIRIquant. Compared with CIRI, the junction ratio in CIRIquant was higher, but overall, there was strong correlation between the two methods (Figure S15).

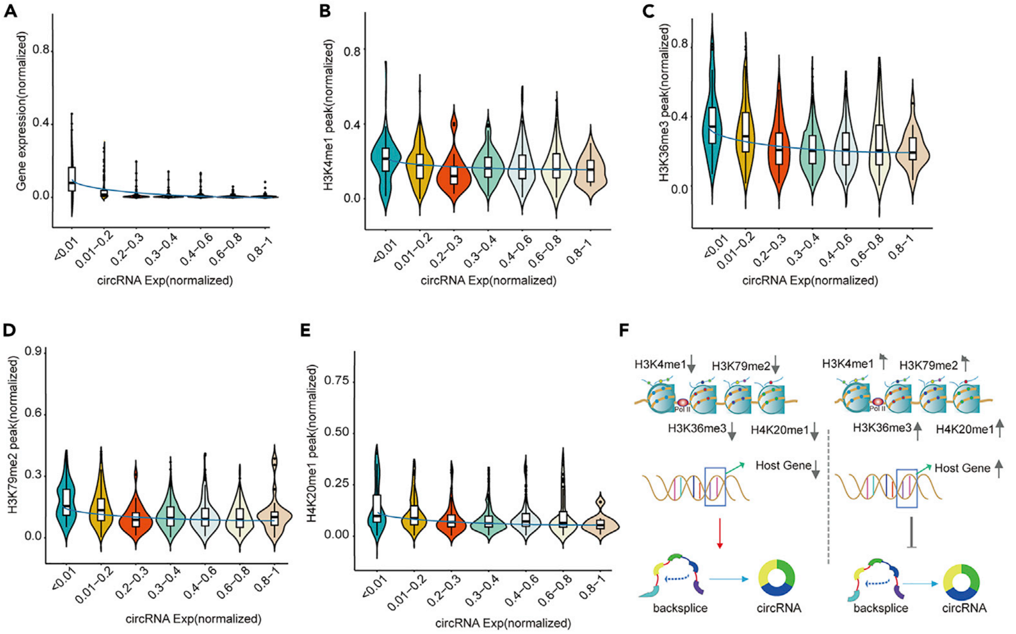


Figure 6. The Relationship between circRNA Expression and Five Important Signals in the A549 Cell Line

(A) Violin plot of host gene and circRNA expression.

(B) Violin plot of histone-modified H3K4me1 and circRNA expression.

(C) Violin plot of histone modification of H3K36me3 and circRNA expression.

(D) Violin plot of histone modification of H3K79me2 and circRNA expression.

(E) Violin plot of histone-modified H4K20me1 and circRNA expression.

(F) Schematic diagram of the relationship between five signals and circRNA in the A549 cell line. The data used in the figure was normalized. See also Figures S10–S15.

The current study found that the formation of circular RNA is regulated by a number of factors, of which Lariat-driven circularization and intron-pairing-driven circularization are the two most typical models (Chen and Yang, 2015). Moreover, it was discovered that the inverted repeat Alu (IRAlu) plays an important role in exon splicing and formation of circular RNA (Liang and Wilusz, 2014; Petkovic and Müller, 2015; Zhang et al., 2014). In addition, RNA-binding proteins (RBPs) also regulate the formation of circular RNA (Conn et al., 2015; Ruan et al., 2019). A novel computational method called iCircRBP-DHN was also proposed for discriminating circRNA-RBP bingeing sites (Yang et al., 2020). However, there are still many unknowns about the mechanism of the formation and function of circRNA. Although little research has been done on circRNA and epigenetics, there is increasing evidence that circRNA is associated with epigenetics. A recent study led to further explicit support for the possible specific production and regulation of circRNAs (Tarrero et al., 2018), which showed higher H3K36me3 levels (post-transcriptional histone modifications) in circularized exons.

In this work, we have quantified the relative contribution of host gene and histone modifications to circRNA expression regulation. We identified the important factors affecting the different expression patterns of circRNA by random forest. Moreover, the AUC of the random forest reached 0.78 or more, regardless of the training set or the testing set. These results showed that the factors that we identified show good effect on distinguishing the expression of circRNA. Next, we selected the factor features of the top five contributions to construct classifiers and estimate the area under the curve, thus eliminating the occurrence of factor redundancy in model construction. Although some histone modifications were redundant for predicting the overall circRNA expression, there might be different circRNA groups that are regulated by histone modifications in different ways. It should be noted that the redundancy only exists with regard to circRNA expression prediction. Essentially, distinct histone modification types play very different roles during transcriptional regulation (Xu et al., 2018). For example, both H3K4me3 and H3K36me3 act as marks for active genes; the former mainly occurs in the promoter regions facilitating the initiation of transcription, whereas the latter functions mainly in the transcribed regions involved in transcriptional elongation (Cheng and Gerstein, 2012).

Finally, we screened five important factors as features to characterize the different expression of circRNA. We observed that the four shared factors, host gene expression, H3k36me3, H3k79me2, and H4k20me1, showed negative effects on the expression of circRNA in each cell line, whereas the fifth important factor was cell lineage specificity. According to our analysis, the higher expression of the linear subtype of the host gene resulted in a decrease in the proportion of circulation of circRNA. Although several studies have found that circRNAs were generally poorly correlated with the expression of host genes, some circRNAs still show high correlation with host genes (Ruan et al., 2019). This is not a contradiction, as our conclusion is that host gene expressions contribute to the classification of high and low expressed circRNAs. We analyzed the contribution of each factor to the classifier using the Gini index. The mean decrease Gini (MDG) was used as the importance score of each feature to provide a relative rank of investigated features. The larger MDG indicates the increasing importance of the corresponding feature for predicting of circRNA expression patterns. In addition, we also verified the contribution of the top five factors to the accuracy of the model in the independent verification sets.

As we know, histone modification of H3k36me3 is related to the transcriptional region of the gene (Wagner and Carpenter, 2012; Zuo et al., 2018), and down-regulation of H3k36me3 and H3k79me2 expression promotes the back splice of exons. Histone modification of H4k20me1 is catalyzed by *PR-Set7* and is associated with the cell cycle (Beck et al., 2012) and down-regulated H4k20me1 promotes up-regulated expression of circRNA. In total, we explore the molecular mechanism of circRNA from the perspective of epigenetics. In addition, recent breakthrough researches have demonstrated that N6-methyladenosine (m6A) modification occurs in circRNAs and promotes protein translation through recruitment of the initiation factor *eIF4G2* and the m6A reader *YTHDF3* (Li et al., 2019; Paramasivam and Vijayashree Priyadharsini, 2020; Zhang et al., 2019). Chen et al. (2019b) reported that m6A modifications of human endogenous circRNAs exerted an important function of suppressing innate immune responses by inhibiting RIG-I activation. These may become a new direction for research in the field of circRNA and epigenetics.

In summary, we have shown that histone modification and host gene expression signals can predict circRNA expression in various cell lines. In addition, we selected a small number of modifications, which together can explain a large part of the difference in circRNA expression. The level of these modifications can be used to infer the expression of circRNA, thereby providing some information about the transcription process, which provides the possibility for many unknown mechanisms and functions of circRNA.

Limitations of the Study

In this study, we provide a new perspective on the regulation of circRNA expression and use a machine learning method to model the high and low expression levels of circRNA and histone modification status to facilitate the understanding and discovery of new circRNA regulatory mechanisms. This work still has some limitations that deserve attention and further study. First, the histone modification data obtained from public databases for establishing regulatory models are not complete, although to our knowledge, these databases are already of high quality and relatively comprehensive. More complete data will be more conducive to comprehensive modeling of the relationship between biomolecules. For the types of histone modifications that do not exist in the model, we do not know whether there are regulatory relationships between histone modifications and circRNA. More complete data will be more conducive to the comprehensive modeling of regulatory relationships between biomolecules. Second, we selected a small number of modifications, which together can explain a large part of the difference in circRNA expression. Whether these modifications play a crucial role in the transcription process, or whether they represent equally important modification groups, must be clarified through further experimental studies.

Resource Availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Yongsheng Li (liyongsheng@hainmc.edu.cn).

Materials Availability

This study did not generate new materials.

Data and Code Availability

This study did not generate datasets/code.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101842.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China (2018YFC2000100); the National Natural Science Foundation of China (32060152, 31970646, 31871338 and 61873075); Natural Science Foundation for Distinguished Young Scholars of Heilongjiang Province (JQ2019C004); and Heilongjiang Touyan Innovation Team Program.

AUTHOR CONTRIBUTIONS

M.Z. and L.F. conducted experiments and collected data; M.Z., Q.W., and Z.C. analyzed results; M.Z. and K.X. created figures; M.Z. and H.Z. performed statistical analysis; M.Z. prepared first draft; M.Z., K.X., L.F., Q.W., Z.C., H.Z., Y.Z., and Y.L. corrected and proofread manuscript; Y.Z. and Y.L. supervised the study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 3, 2020
Revised: November 11, 2020
Accepted: November 18, 2020
Published: December 18, 2020

REFERENCES

Aktas, T., Avsar Ilik, I., 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.

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. Cel. *56*, 55–66.

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

Beck, D.B., Oda, H., Shen, S.S., and Reinberg, D. (2012). PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. Genes Dev. *26*, 325–337.

Chen, L.L., and Yang, L. (2015). Regulation of circRNA biogenesis. RNA Biol. *12*, 381–388.

Chen, X., Yang, T., Wang, W., Xi, W., Zhang, T., Li, Q., Yang, A., and Wang, T. (2019a). Circular RNAs in immune responses and immune diseases. Theranostics *9*, 588–607.

Chen, Y.G., Chen, R., Ahmad, S., Verma, R., Kasturi, S.P., Amaya, L., Broughton, J.P., Kim, J., Cadena, C., Pulendran, B., et al. (2019b). N6-methyladenosine modification controls circular RNA immunity. Mol. Cel. *76*, 96–109.e9.

Cheng, C., and Gerstein, M. (2012). Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. Nucleic Acids Res. *40*, 553–568.

Cheng, Z., Yu, C., Cui, S., Wang, H., Jin, H., Wang, C., Li, B., Qin, M., Yang, C., He, J., et al. (2019). circTP63 functions as a ceRNA to promote lung squamous cell carcinoma progression by upregulating FOXM1. Nat. Commun. *10*, 3200.

Cocquerelle, C., Mascrez, B., Hetuin, D., and Bailleul, B. (1993). Mis-splicing yields circular RNA molecules. FASEB J. *7*, 155–160.

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.

Consortium, E.P. (2004). The ENCODE (ENCyclopedia of DNA Elements) project. Science *306*, 636–640.

Dai, X., Zhang, N., Cheng, Y., Yang, T., Chen, Y., Liu, Z., Wang, Z., Yang, C., and Jiang, Y. (2018). RNA-binding protein trinucleotide repeat-containing 6A regulates the formation of circular RNA 0006916, with important functions in lung cancer cells. Carcinogenesis *39*, 981–992.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

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.

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

Guarnerio, J., Bezzi, M., Jeong, J.C., Paffenholz, S.V., Berry, K., Naldini, M.M., Lo-Coco, F., Tay, Y., Beck, A.H., and Pandolfi, P.P. (2016). Oncogenic role of fusion-circRNAs derived from cancer-associated chromosomal translocations. Cell *165*, 289–302.

Han, B., Chao, J., and Yao, H. (2018). Circular RNA and its mechanisms in disease: from the bench to the clinic. Pharmacol. Ther. *187*, 31–44.

Hu, Z.Q., Zhou, S.L., Li, J., Zhou, Z.J., Wang, P.C., Xin, H.Y., Mao, L., Luo, C.B., Yu, S.Y., Huang, X.W., et al. (2019). Circular RNA sequencing identifies

CircASAP1 as a key regulator in hepatocellular carcinoma metastasis. Hepatology. https://doi.org/10.1002/hep.31068.

Huang, W., Yang, Y., Wu, J., Niu, Y., Yao, Y., Zhang, J., Huang, X., Liang, S., Chen, R., Chen, S., et al. (2020). Circular RNA cESRP1 sensitises small cell lung cancer cells to chemotherapy by sponging miR-93-5p to inhibit TGF-beta signalling. Cell Death Differ. *27*, 1709–1727.

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

Jeck, 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.

Ju, H.Q., Zhao, Q., Wang, F., Lan, P., Wang, Z., Zuo, Z.X., Wu, Q.N., Fan, X.J., Mo, H.Y., Chen, L., et al. (2019). A circRNA signature predicts postoperative recurrence in stage II/III colon cancer. EMBO Mol. Med. *11*, e10168.

Kramer, M.C., Liang, D., Tatomer, D.C., Gold, B., March, Z.M., Cherry, S., and Wilusz, J.E. (2015). Combinatorial control of Drosophila circular RNA expression by intronic repeats, hnRNPs, and SR proteins. Genes Dev. *29*, 2168–2182.

Kristensen, L.S., Andersen, M.S., Stagsted, L.V.W., Ebbesen, K.K., Hansen, T.B., and Kjems, J. (2019). The biogenesis, biology and characterization of circular RNAs. Nat. Rev. Genet. *20*, 675–691.

Li, Y., Zhang, J., Huo, C., Ding, N., Li, J., Xiao, J., Lin, X., Cai, B., Zhang, Y., and Xu, J. (2017). Dynamic organization of lncRNA and circular RNA regulators collectively controlled cardiac differentiation in humans. EBioMedicine *24*, 137–146.

Li, Y., Xiao, J., Bai, J., Tian, Y., Qu, Y., Chen, X., Wang, Q., Li, X., Zhang, Y., and Xu, J. (2019). Molecular characterization and clinical relevance of m(6)A regulators across 33 cancer types. Mol. Cancer *18*, 137.

Liang, D., and Wilusz, J.E. (2014). Short intronic repeat sequences facilitate circular RNA production. Genes Dev. *28*, 2233–2247.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science *327*, 996–1000.

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.

Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell *126*, 1203–1217.

Nigro, J.M., Cho, K.R., Fearon, E.R., Scott, E., Kern, J.M.R., Oliner, J.D., Kinzler, K.W., and Vogelstein, B. (1991). Scrambled exon. Cell *64*, 607–613.

Paramasivam, A., and Vijayashree Priyadharsini, J. (2020). Novel insights into m6A modification in circular RNA and implications for immunity. Cell. Mol. Immunol. *17*, 668–669.

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

Petkovic, S., and Müller, S. (2015). RNA circularization strategies in vivo and in vitro. Nucleic Acids Res. *43*, 2454–2465.

PG, Z. (1996). Circular RNAs from transcripts of the rat cytochrome P450 2C24 gene: correlation with exon skipping. Proc. Natl. Acad. Sci. U. S. A. *93*, 6536–6541.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. *40*, D130–D135.

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. Bioch. Biophy. Acta Rev. Cancer *1870*, 247–260.

Ruan, H., Xiang, Y., Ko, J., Li, S., Jing, Y., Zhu, X., Ye, Y., Zhang, Z., Mills, T., Feng, J., et al. (2019). Comprehensive characterization of circular RNAs in ∼ 1000 human cancer cell lines. Genome Med. *11*, 55.

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.

Salzman, J., Chen, R.E., Olsen, M.N., Wang, P.L., and Brown, P.O. (2013). Cell-type specific features of circular RNA expression. PLoS Genet. *9*, e1003777.

Sun, L.F., Zhang, B., Chen, X.J., Wang, X.Y., Zhang, B.W., Ji, Y.Y., Wu, K.C., Wu, J., and Jin, Z.B. (2019). Circular RNAs in human and vertebrate neural retinas. RNA Biol. *16*, 821–829.

Szabo, L., Morey, R., Palpant, N.J., Wang, P.L., Afari, N., Jiang, C., Parast, M.M., Murry, C.E., Laurent, L.C., and Salzman, J. (2016). Erratum to: statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biol. *17*, 263.

Tarrero, L.C., Ferrero, G., Miano, V., De Intinis, C., Ricci, L., Arigoni, M., Federica Riccardo, Annaratone, L., Castellano, I., Calogero, R.A., et al. (2018). Luminal breast cancer-specific circular RNAs uncovered by a novel tool for data analysis. Oncotarget *9*, 14580–14596.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515.

Vo, J.N., Cieslik, M., Zhang, Y., Shukla, S., Xiao, L., Zhang, Y., Wu, Y.M., Dhanasekaran, S.M., Engelke, C.G., Cao, X., et al. (2019). The landscape of circular RNA in cancer. Cell *176*, 869–881 e813.

Wagner, E.J., and Carpenter, P.B. (2012). Understanding the language of Lys36 methylation at histone H3. Nat. Rev. Mol. Cell Biol. *13*, 115–126.

Weng, W., Wei, Q., Toden, S., Yoshida, K., Nagasaka, T., Fujiwara, T., Cai, S., Qin, H., Ma, Y., and Goel, A. (2017). Circular RNA ciRS-7-A promising prognostic biomarker and a potential therapeutic target in colorectal cancer. Clin. Cancer Res. *23*, 3918–3928.

Wu, W., Ji, P., and Zhao, F. (2020). CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. Genome Biol. *21*, 101.

Xu, J., Wang, Z., Li, S., Chen, J., Zhang, J., Jiang, C., Zhao, Z., Li, J., Li, Y., and Li, X. (2018). Combinatorial epigenetic regulation of non-coding RNAs has profound effects on oncogenic pathways in breast cancer subtypes. Brief. Bioinform. *19*, 52–64.

Zhang, X.O., Wang, H.B., Zhang, Y., Lu, X., Chen, L.L., and Yang, L. (2014). Complementary sequence-mediated exon circularization. Cell *159*, 134–147.

Zhang, C., Fu, J., and Zhou, Y. (2019). A review in research progress concerning m6A methylation and immunoregulation. Front. Immunol. *10*, 922.

Yang, Y., Hou, Z., Ma, Z., Li, X., and Wong, K.C. (2020). iCircRBP-DHN: identification of circRNA-RBP interaction sites using deep hierarchical network. Brief. Bioinform. bbaa274, https://doi.org/10.1093/bib/bbaa274.

Zhang, J., Chen, S., Yang, J., and Zhao, F. (2020). Accurate quantification of circular RNAs identifies extensive circular isoform switching events. Nat. Commun. *11*, 90.

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.

Zuo, X., Rong, B., Li, L., Lv, R., Lan, F., and Tong, M.H. (2018). The histone methyltransferase SETD2 is required for expression of acrosin-binding protein 1 and protamines and essential for spermiogenesis in mice. J. Biol. Chem. *293*, 9188–9197.

iScience, Volume *23*

Supplemental Information

Revealing Epigenetic Factors of circRNA Expression by Machine Learning in Various Cellular Contexts

Mengying Zhang, Kang Xu, Limei Fu, Qi Wang, Zhenghong Chang, Haozhe Zou, Yan Zhang, and Yongsheng Li

Supplemental Figures

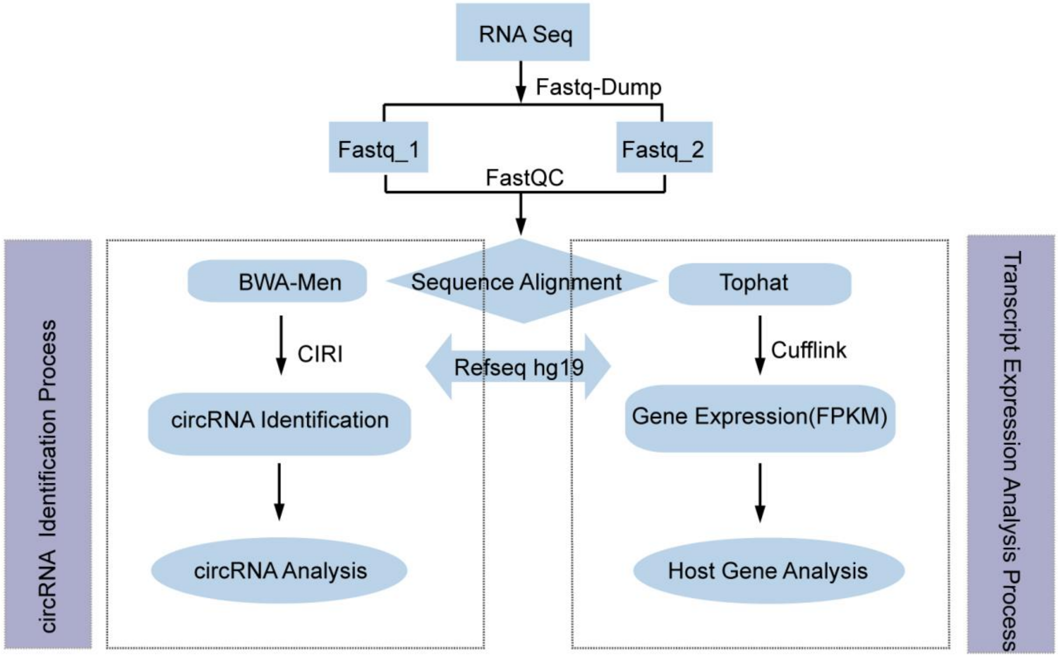


Figure S1. Flowchart for identification of circRNAs and host genes. Related to Figure 1.

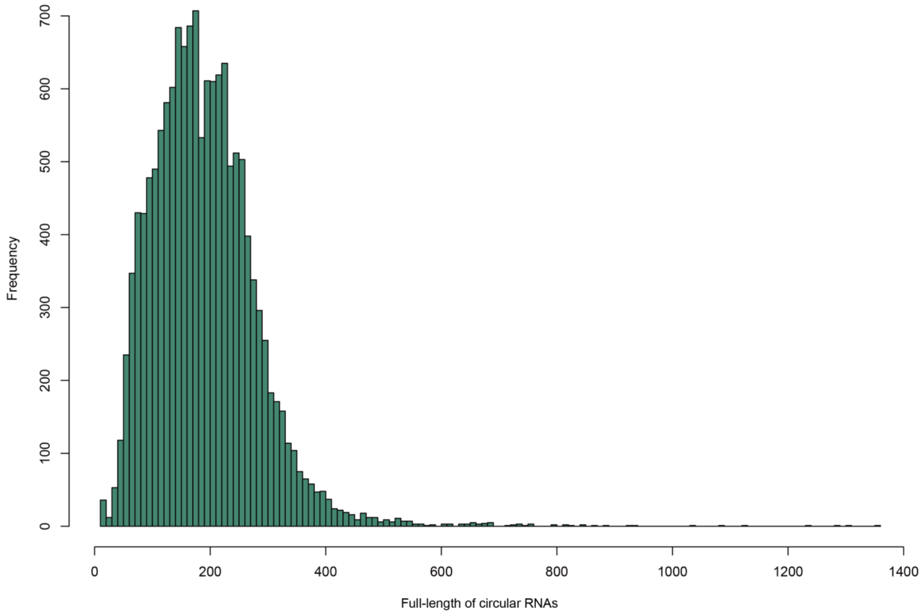


Figure S2. Density plot showing full length of circRNAs in six cell lines. Related to Figure 1.

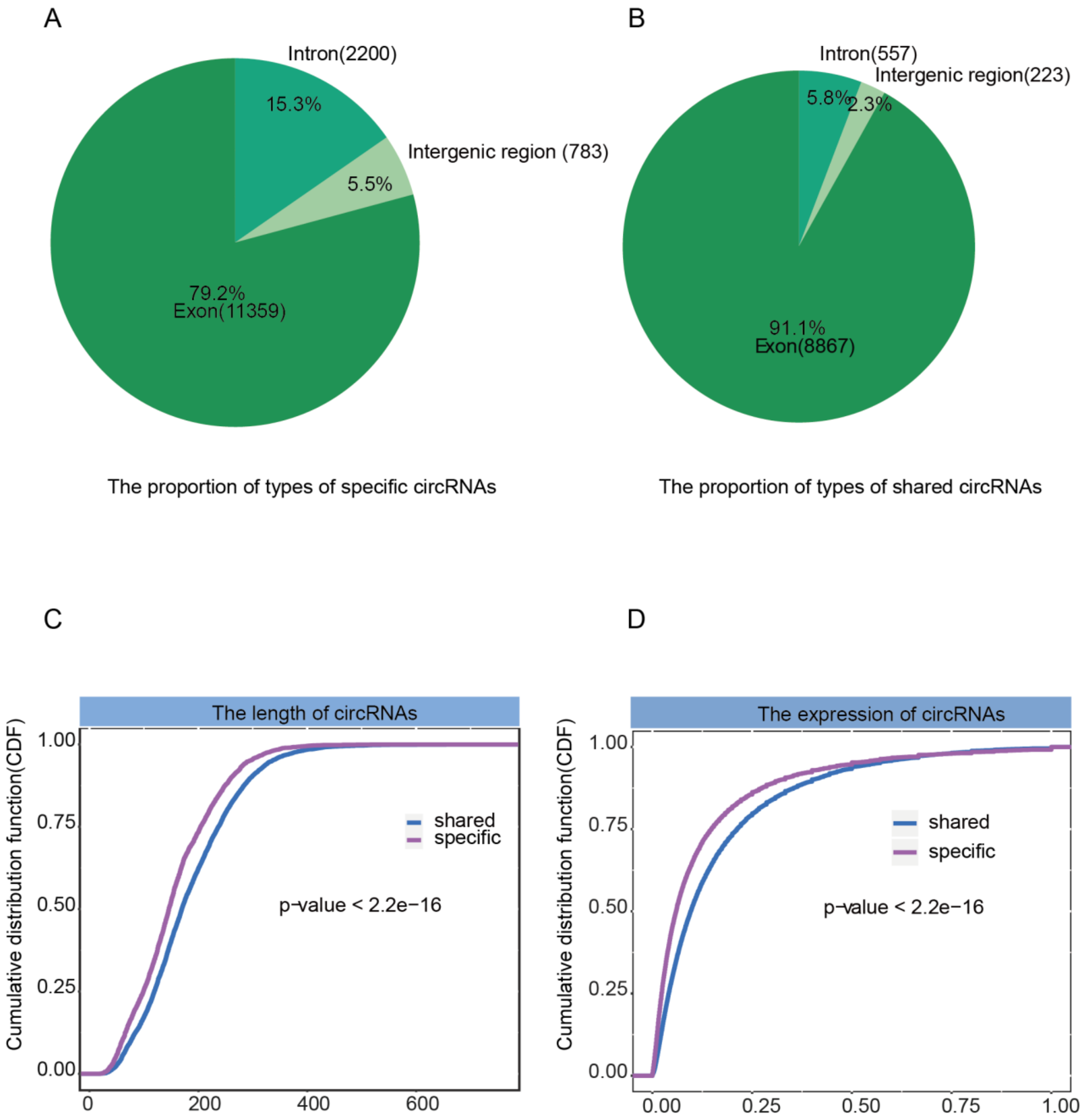


Figure S3. The differences between shared and specific circRNAs. (A-B) The proportion of types of specific and shared circRNAs. (C-D) The cumulative distribution of length and expression between specific and shared circRNAs. Related to Figure 2.

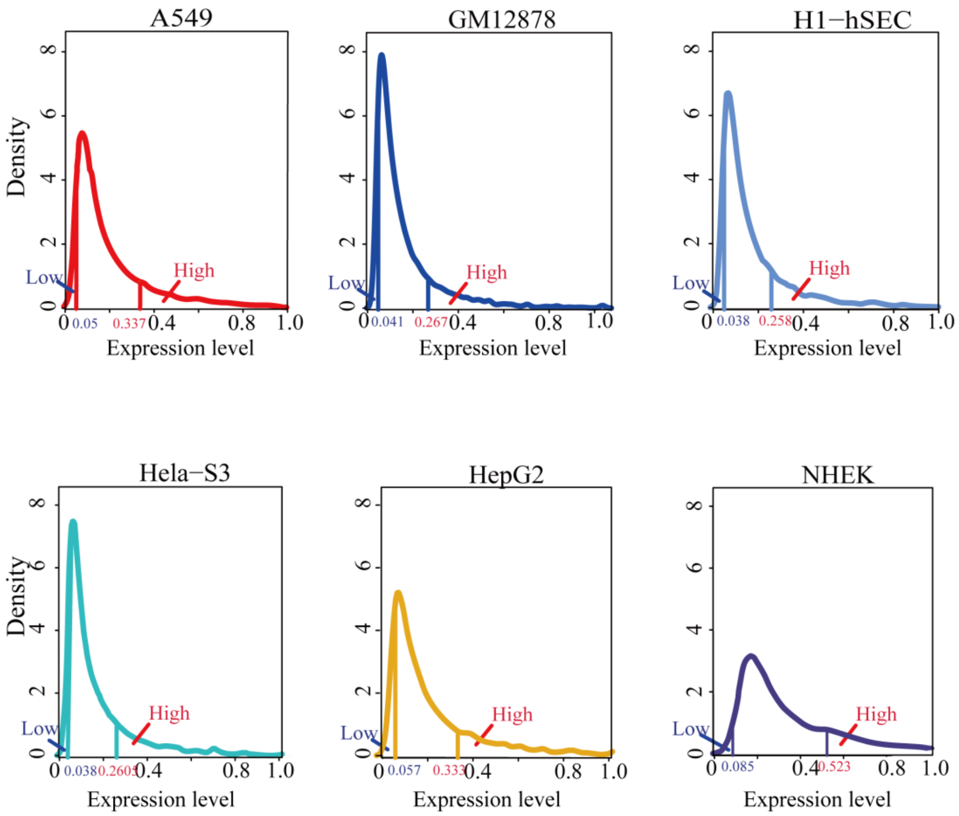


Figure S4. Density plots showing expression values distribution in six cell lines. A density map of expression values in six cell lines and a division of expression levels of high and low expression in each cell line. Related to Figure 2.

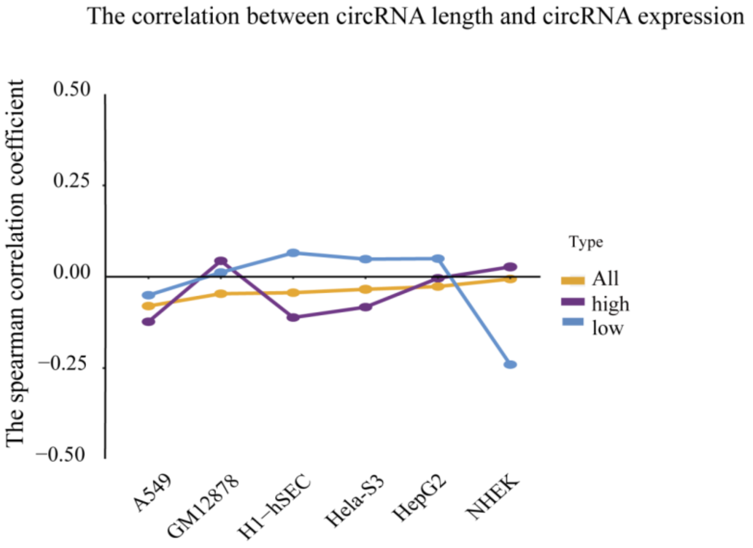


Figure S5. The correlation between circRNA length and circRNA expression. The spearman correlation of circRNA expression and circRNA length in all circRNAs, high expression of circRNA, low expression of circRNA in 6 cell lines. Related to Figure 2.

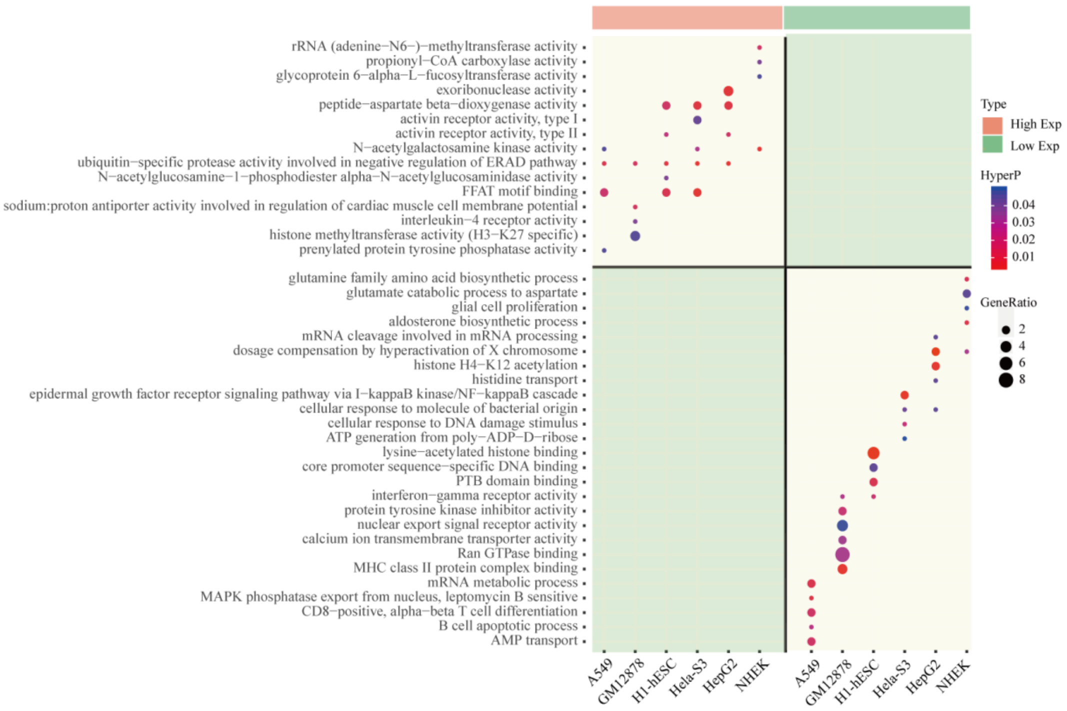


Figure S6. Enrichment and resistance events occur in circRNA high and low expression patterns. Enriched bubble diagrams of high and low expression patterns in molecular function. The portion covered by the red band represents a high expression circRNA, and the portion covered by the green band represents a low expression circRNA. The color of the bubble represents the p-value, bubble size represents the number of circRNA host genes which present in one term. Related to Figure 3.

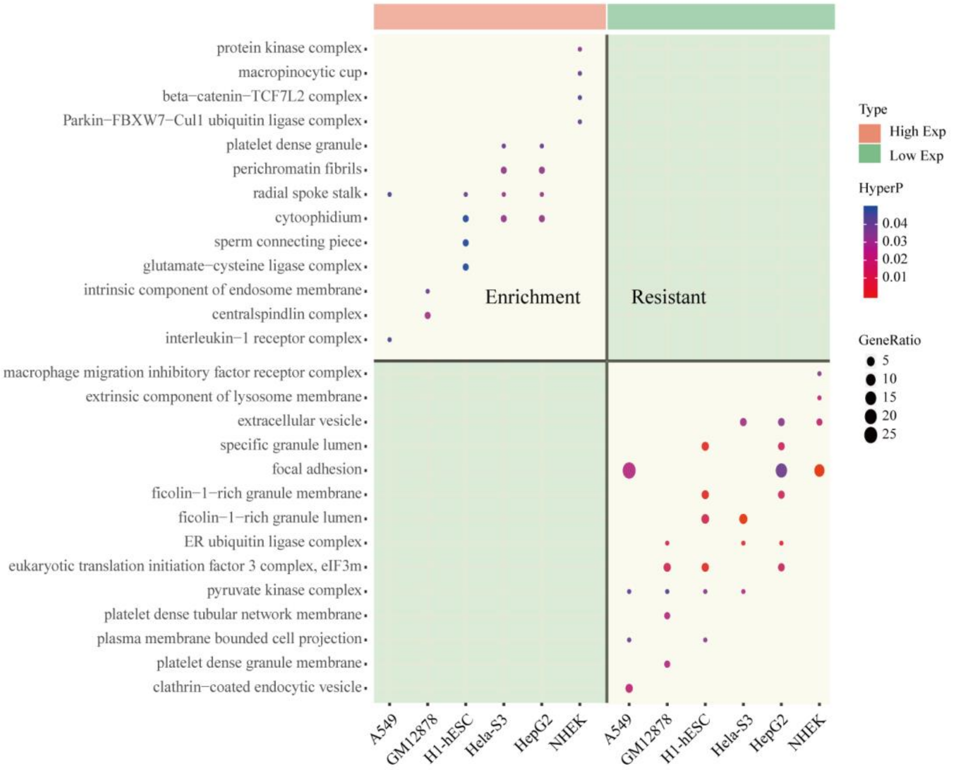


Figure S7. Enrichment and resistance events occur in circRNA high and low expression patterns. Enriched bubble diagrams of high and low expression patterns in cellular component. The portion covered by the red band represents a high expression circRNA, and the portion covered by the green band represents a low expression circRNA. The color of the bubble represents the p-value, bubble size represents the number of circRNA host genes which present in one term. Related to Figure 3.

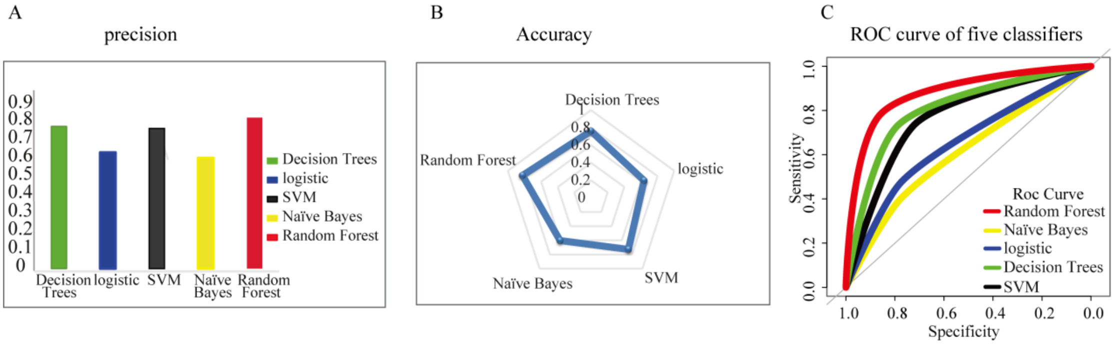


Figure S8. Evaluation of the effects of five classifiers in the A549 cell line. (A) Bar graph of the precision of the five classifiers. (B) Radar plots of the accuracy of the five classifiers. (C) The ROC curve of the five classifiers. Related to Figure 4.

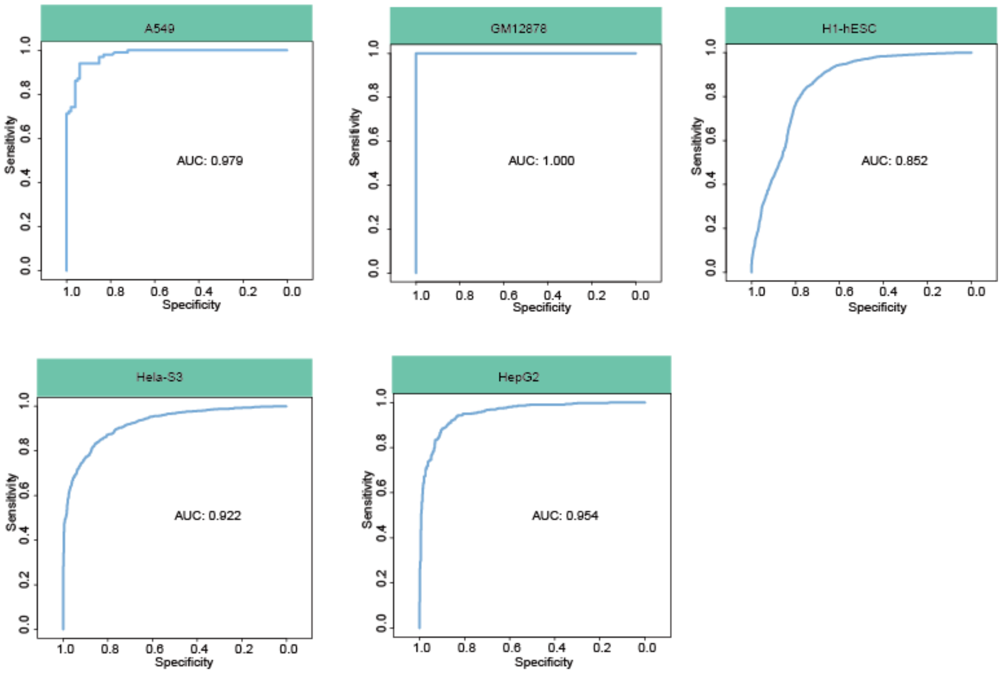


Figure S9. The area under the ROC curve of independent validation sets for each cell line. The predictive ability of top 5 factors to characterize the different expression patterns of circRNA. Related to Figure 5.

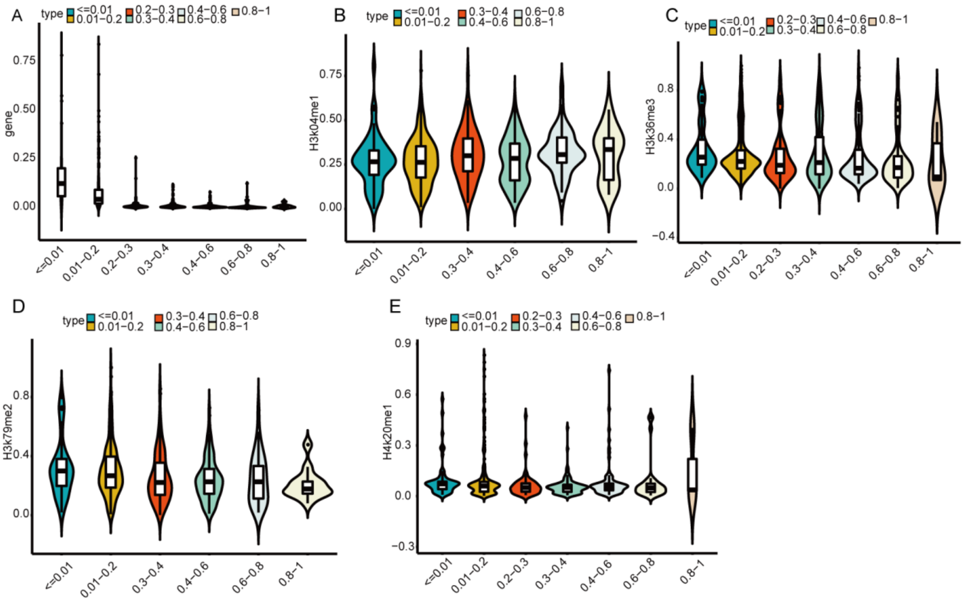


Figure S10. The relationship between circRNA expression and five important signals in the GM12878 cell line. (A) Violin map of host gene and circRNA expression. (B) Violin map of histone-modified H3K4me1 and circRNA expression. (C) Violin plot of histone modification of H3K36me3 and circRNA expression. (D) Violin plot of histone modification of H3K79me2 and circRNA expression. (E) Violin plot of histone-modified H4K20me1 and circRNA expression. Related to Figure 6.

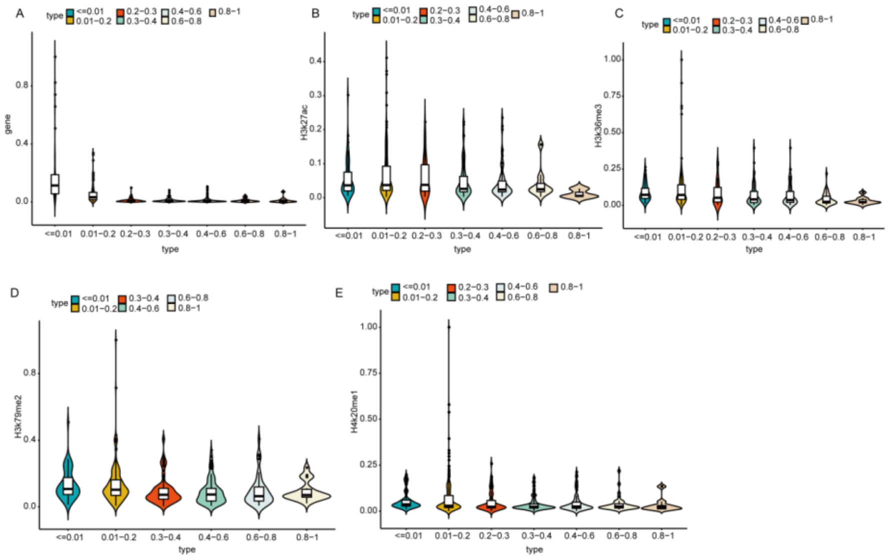


Figure S11. The relationship between circRNA expression and five important signals in the H1-hESC cell line. (A) Violin map of host gene and circRNA expression. (B) Violin map of histone-modified H3K27ac and circRNA expression. (C) Violin plot of histone modification of H3K36me3 and circRNA expression. (D) Violin plot of histone modification of H3K79me2 and circRNA expression. (E) Violin plot of histone-modified H4K20me1 and circRNA expression. Related to Figure 6.

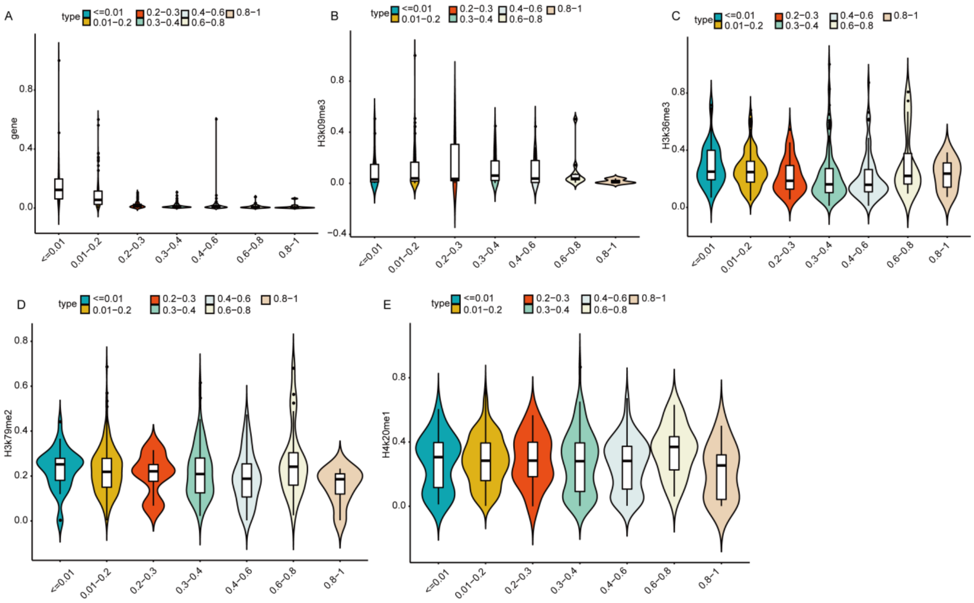


Figure S12. The relationship between circRNA expression and five important signals in the Hela-S3 cell line. (A) Violin map of host gene and circRNA expression. (B) Violin map of histone-modified H3K9me3 and circRNA expression. (C) Violin plot of histone modification of H3K36me3 and circRNA expression. (D) Violin plot of histone modification of H3K79me2 and circRNA expression. (E) Violin plot of histone-modified H4K20me1 and circRNA expression. Related to Figure 6.

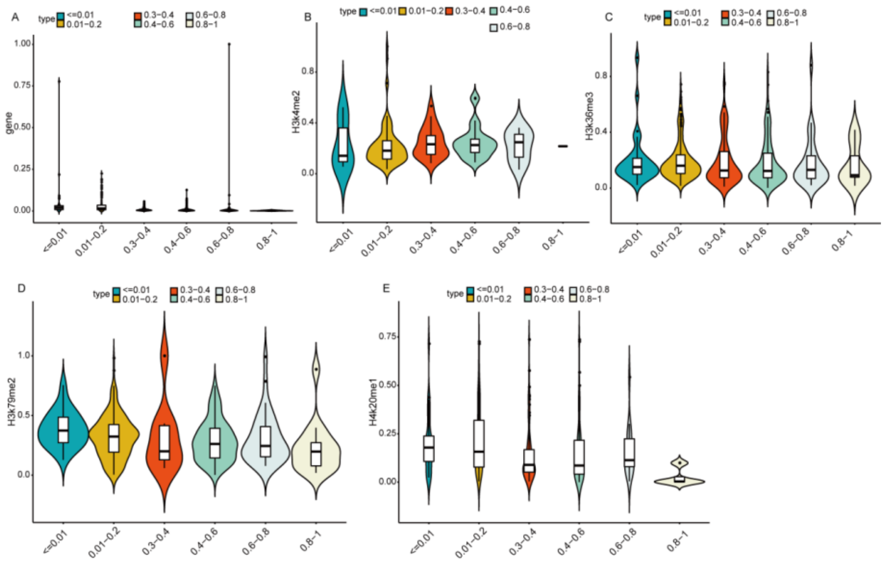


Figure S13. The relationship between circRNA expression and five important signals in the HepG2 cell line. (A) Violin map of host gene and circRNA expression. (B) Violin map of histone-modified H3K4me2 and circRNA expression. (C) Violin plot of histone modification of H3K36me3 and circRNA expression. (D) Violin plot of histone modification of H3K79me2 and circRNA expression. (E) Violin plot of histone-modified H4K20me1 and circRNA expression. Related to Figure 6.

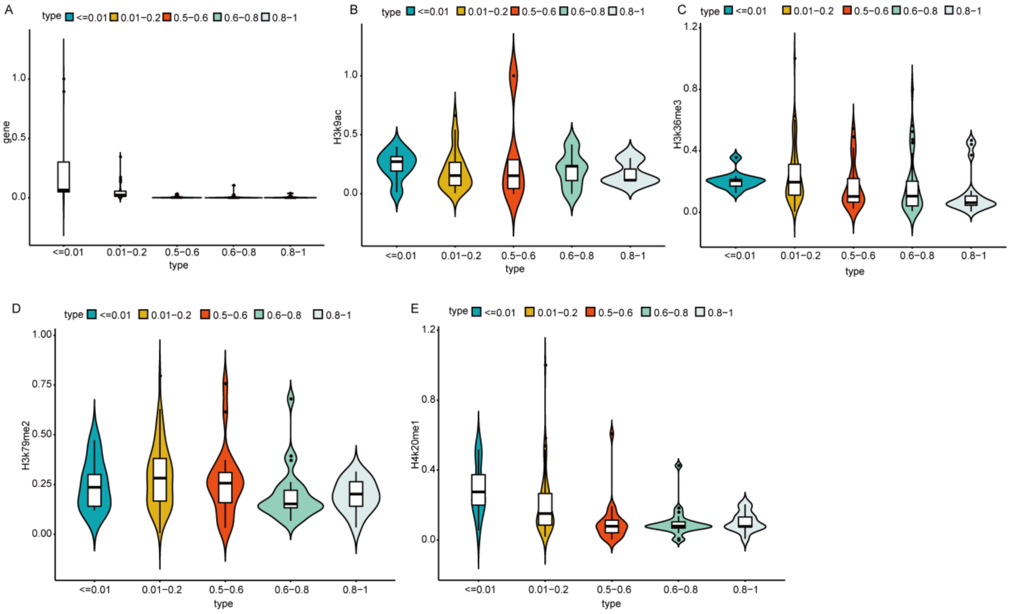


Figure S14. The relationship between circRNA expression and five important signals in the NHEK cell line. (A) Violin map of host gene and circRNA expression. (B) Violin map of histone-modified H3K9ac and circRNA expression. (C) Violin plot of histone modification of H3K36me3 and circRNA expression. (D) Violin plot of histone modification of H3K79me2 and circRNA expression. (E) Violin plot of histone-modified H4K20me1 and circRNA expression. Related to Figure 6.

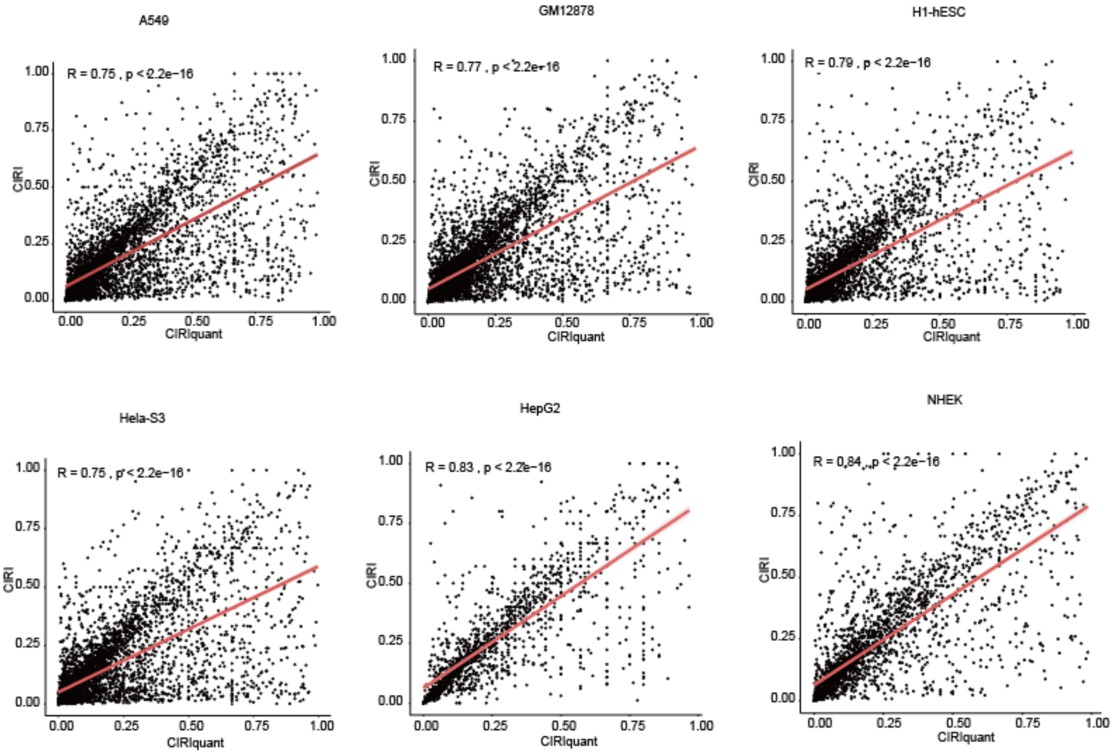


Figure S15. Scatter plot showing the correlation between the junction ratio in CIRI and CIRIquant, related to Figure 6.

Table S1. List of raw RNA-seq sequencing data for each cell line, related to Figure 1.

Cell line	Data
A549	SRR5048109.sra
	SRR5048110.sra
GM12878	SRR5048071.sra
	SRR5048072.sra
	SRR5048111.sra
	SRR5048112.sra
	SRR5048115.sra
	SRR5048116.sra
	SRR5048187.sra
Hela-s3	SRR5048117.sra
	SRR5048118.sra
	SRR5048119.sra
	SRR5048120.sra
	SRR5048132.sra
HepG2	SRR5048083.sra
	SRR5048084.sra
	SRR5048121.sra
	SRR5048122.sra
	SRR5048135.sra
	SRR5048136.sra
	SRR3192612.sra
NHEK	SRR3192482.sra
	SRR3192483.sra
	SRR3192484.sra
	SRR3192504.sra
	SRR3192505.sra
	SRR3192506.sra
H1-hESC	SRR5048069.sra
	SRR5048070.sra
	SRR5048129.sra
	SRR5048130.sra

Table S2. List of raw RNA-seq sequencing data for independent validation, related to Figure 4.

Cell line	Data
A549	SRR8371688.sra
	SRR8371689.sra
	SRR8371690.sra
GM12878	SRR3103887.sra
H1-hESC	SRR7230396.sra
	SRR7230403.sra
	SRR7230410.sra
Hela-s3	SRR8449597.sra
	SRR8449598.sra
	SRR8449599.sra
	SRR8449600.sra
HepG2	SRR4422330.sra
	SRR4422331.sra

Transparent Methods

RNA-Seq data in various cell lines. Raw sequencing data of RNA-Seq for six cell lines (including A549, GM12878, H1-hESC, HepG2, HeLa-S3 and NHEK) were downloaded from The Encyclopedia of DNA Elements (ENCODE) project (Consortium, 2004; Davis et al., 2018). Detailed samples information for cell lines was listed in Table S1. We next used fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to evaluate the quality of the raw reads. Trimmomatic (Bolger et al., 2014) was used to filter low quality bases and adaptor sequences at both ends of the reads. Finally, clean reads were aligned to human reference genome (hg19) by TopHat v2.1.1 (Trapnell et al., 2009).

Identification of circRNAs and protein-coding genes in various cell lines. We applied CIRI (Gao et al., 2015) which is an efficiency and accuracy method to detect the back-splice junction sites of circRNAs. In brief, we used FastQC for quality control of RNA-Seq data. BWA-MEM software (Li, 2013) was used to map reads to the reference genome hg19. We next used CIRI script to identify circRNAs in six cell lines. The default parameters were used for identifying circRNAs. The identified circRNAs were overlapped with circRNAs in circBase (Glazar et al., 2014), circAtlas (Wu et al., 2020) and circRIC (Ruan et al., 2019). The expressions of circRNAs were evaluated by relative counts of backsplice-junction reads and non-junction reads. We analyzed the genomic features of circRNAs, including circRNA types, circRNA exon length, and chromosome distribution. In addition, the expressions of protein-coding genes were calculated as Fragments Per Kilobase of exon per Megabase of library size (FPKM) by cufflinks v2.2.1 in each cell line (Trapnell et al., 2010).

Epigenetic modification of circRNAs and protein-coding genes. We utilized the co-localization approach to match histone modification signal of the circRNAs. In brief, 11 histone modification signals were downloaded in the University of California Santa Cruz (UCSC) Genome Browser database(W. James Kent et al., 2002) and matched to circRNAs according to the chromosomal location. If multiple histone modification peaks were aligned to the same circRNA, we used mean value as the circRNA-related histone modification signal. Moreover, circRNA-associated host gene expression levels were also defined by co-localization approach. For replicates of RNA-seq, we selected circRNAs that were shared in replicates for each cell line, and circRNA expression value was calculated as the mean value of all replicates.

Association analysis of circRNA expression. We used CIRIquant (Zhang et al., 2020) to quantify the expression of circRNA with the following parameter settings: CIRIquant -t 4 -1 ./test_1.fq.gz -2 ./test_2.fq.gz --config ./chr.yml --no-gene- o ./test -p test. Next, the spearman correlation between the junction ratio in CIRI and the junction ratio in CIRIquant was calculated.

Construction of classifiers. We first analyzed the distribution of circRNA expression in six cell lines. We found that the majority of circRNAs were low level expression and the distribution exhibited a skewed distribution. The normalized expressions of circRNAs were around 0.1. To better distinguish the difference of high and low expressed circRNAs, we removed the circRNAs with moderate expression levels. Thus, we first ranked the circRNAs based on their expression levels and defined the top 20% and bottom 20% circRNAs as high and low expressed groups. CircRNA-related host genes and 11 histone modifications were selected to characterize circRNAs with different expression patterns. Next, 12 features were normalized to eliminate the dimension effect. The normalization formula is as follows:

$$z_i = \frac{x_i}{\sqrt{\sum_{i=0}^{n} x_i^2}} * 100 (\text{Eq. } 1)$$

Finally, five classifiers including decision tree, logistic regression, SVM, naïve bayes, and random forest was constructed to predict circRNA expression patterns with the host gene and histone modifications as features.

Cross-validation of classifiers. We performed 10-fold cross validation for sampled datasets to evaluate the accuracy of the models. The process was repeated ten times and the average area under the receiver operating characteristic curve (AUC) was calculated as the major indicator of prediction accuracy .For duplicate samples of each cell line, the test set in each cell line was set up to assess the robustness of the model according to the ratio n-1:1 (n represents the experimental number of replicates of the sample).

Independent validation set. The RNA-Seq datasets for independent validation were downloaded from the Gene Expression Omnibus database(GEO) (Ron Edgar et al., 2002). 11 histone modification signals were downloaded in The NIH Roadmap Epigenomics Mapping Consortium (http://www.roadmap epigenomics.org/). Detailed samples information for cell lines was listed in Table S2.

Function analysis of circRNAs. To explore the differences in the function of circRNA-related host genes with different expression classes, Gene Ontology (GO) terms and term gene sets in gene2Go were downloaded from National Center for Biotechnology Information (NCBI, https://www.ncbi.nlm.nih.gov/). KEGG pathways and pathway gene sets were obtained in The Molecular Signatures Database (Liberzon et al., 2011). Furthermore, we applied hyper-geometric distribution statistical theory to calculate p-values for enrichment analysis. We defined the over-represented and under-represented of host gene in the high and low expression patterns of circRNA across six cell lines. A function term presents over-represented when formula is satisfied and the p-value is under 0.05.

$$\mathrm{p}(\mathrm{X} > \mathrm{x}) = \sum_{\mathrm{x}}^{\mathrm{n}} \frac{\mathrm{C}_{\mathrm{M}}^{\mathrm{x}} \mathrm{C}_{\mathrm{N-M}}^{\mathrm{n-x}}}{\mathrm{C}_{\mathrm{N}}^{\mathrm{n}}} \text{(Eq. 2)}$$

A functional term was under-represented when formula is satisfied and the p-value is under 0.05.

$$\mathrm{p}(\mathrm{X} \leq \mathrm{x}) = 1 - \sum_{\mathrm{x}}^{\mathrm{n}} \frac{\mathrm{C}_{\mathrm{M}}^{\mathrm{x}} \mathrm{C}_{\mathrm{N-M}}^{\mathrm{n-x}}}{\mathrm{C}_{\mathrm{N}}^{\mathrm{n}}} \text{(Eq. 3)}$$

Where x is the number of circRNA host genes annotated to a certain GO term, and N is the background gene sets. We selected the hg19 reference genome as the background gene sets. M is the number of all genes in a certain functional term, and n is the number of host genes of circRNAs.

Genome visualization of expression and histone modification. To discover the differences in modification signal of the important factors in circRNA, we further examined A549 cell line that is related to lung cancer. We obtained BigWig files of polyA RNA-seq, ployA deplept RNA-seq and histone modification peaks in A549 cell line from ENCODE (http://genome.ucsc.edu/ENCODE/downloads.html). The tool bigWigToBedGraph, which was acquired from UCSC, was used to convert BW files to BedGraph format. The WashU Epigenome Browser (https://epigenomegateway.wustl.edu/) was used to visualize high- and low-expression circRNA-modified signals.

Supplemental References：

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114-2120.

Consortium, E.P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636-640.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K.*, et al.* (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic acids research *46*, D794-D801.

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

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

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv *00*, 1–3.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739-1740.

Ron Edgar, Michael Domrachev, and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research *30*, 207–210.

Ruan, H., Xiang, Y., Ko, J., Li, S., Jing, Y., Zhu, X., Ye, Y., Zhang, Z., Mills, T., Feng, J.*, et al.* (2019). Comprehensive characterization of circular RNAs in ~ 1000 human cancer cell lines. Genome Med *11*, 55.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology *28*, 511-515.

W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and Haussler, D. (2002). The human genome browser at UCSC Genome research *12*, 996–1006.

Wu, W., Ji, P., and Zhao, F. (2020). CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. Genome biology *21*.

Zhang, J., Chen, S., Yang, J., and Zhao, F. (2020). Accurate quantification of circular RNAs identifies extensive circular isoform switching events. Nature communications *11*, 90.