Comparative Analysis of Pan-cancer and Normal Tissues Reveals Cancer Tissue-enriched CircRNAs Associated with Cancer Mutations as Potential Exosomal Biomarkers

Xuezhu Wang  
Peking Union Medical College Hospital  https://orcid.org/0000-0002-9424-6540

Yucheng Dong  
Peking Union Medical College Hospital

Zilong Wu  
The First Affiliated Hospital of Hunan Normal University

Guanqun Wang  
Tsinghua University

Yue Shi  
Peking Union Medical College Hospital

Yongchang Zheng (✉ yong-chang_zheng@outlook.com)  
Peking Union Medical College Hospital

Research

Keywords: pan-cancer, circRNA, plasma exosomal biomarkers, cancer mutations, liquid biopsy

DOI: https://doi.org/10.21203/rs.3.rs-421398/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Abstract

**Background:** A growing body of evidence has shown that circular RNAs (circRNAs) are promising exosomal cancer biomarker candidates. However, global alterations of circRNAs in cancer and the driving force of circRNA biogenesis remain under investigation. Studies on these factors are needed to identify ideal circRNA biomarkers for cancer.

**Methods:** We comparatively analyzed the circRNA landscape in pan-cancer and normal tissues to investigate their biologically significant characteristics and identify circRNAs enriched in pan-cancer. We used co-expression analysis, LASSO regularization, and support vector machine to analyze 265 pan-cancer and 319 normal tissues to identify circRNAs with the highest ability to distinguish cancer tissues from normal tissues which showed high expression in plasma exosomes from patients with cancer (e.g., hepatocellular carcinoma, HCC) and were associated with cancer mutations.

**Results:** Expression of circRNAs was reduced in cancer tissues and plasma exosomes from patients with cancer than in normal tissues and exosomes from healthy controls. The circRNAs with the strongest ability to distinguish between cancer and normal tissues were among the top 10% of stably expressed circRNAs. Compared with normal-tissue-enriched circRNAs, cancer-tissue-enriched circRNAs exhibited a more prominent association with cancer mutations and higher levels in plasma exosomes from patients with HCC. Particularly, we identified dynein axonemal heavy chain 14 (DNAH14), which serves as the host gene of three cancer tissue-enriched circRNAs, as one of the top back-spliced genes exclusive to cancer tissues. Among these three circRNAs, chr1:224952669:224968874_+ was significantly elevated in plasma exosomes from patients with HCC and was associated with the cancer mutation chr1:224952669: G>A, a splice acceptor variant that is potentially a driving force of circRNA biogenesis.

**Conclusions:** Our bioinformatic analyses provide insights into the characteristics of the circRNA landscape in cancer and the potential of cancer mutation-associated and cancer tissue-enriched RNAs as plasma exosomal cancer biomarkers. Moreover, our results highlight DNAH14 as a host gene that should be further examined in studies of circRNA in cancer.

**Background**

Circular RNAs (circRNAs) are covalently closed circular and single-stranded non-coding RNAs that are universally generated by cancer and normal cells; they have been detected in plasma exosomes derived from these cells (1). CircRNAs are gaining increasing attention as promising cancer biomarkers using liquid biopsies and are associated with several types of cancers, such as gastric cancer, colon cancer, and hepatocellular carcinoma (HCC) (2). For example, circ-KIAA1244 was downregulated in gastric tissues and plasma samples in patients with gastric cancer, and this decrease was negatively correlated with the TNM stage, lymphatic metastasis, and overall survival of patients (3). In colon cancer, a scoring model involving four circRNAs effectively predicted the postoperative recurrence of stage II/III cancer (4). Zhang
et al. showed that elevation of circUHRF1 in HCC tissues and plasma exosomes correlated with poor prognosis and resistance to anti-PD1 immunotherapy (5).

In recent years, studies have revealed great variability in circRNA profiles between cancer tissues and normal tissues, for which many circRNA databases have been established, some of which are highly cited (6). The Cancer-Specific CircRNA Database (CSCD)/Interactional Database of Cancer-Specific CircRNAs (IDCSC) contains classifications of circRNAs that are “cancer-specific”, “normal-specific”, or “common” based on analysis of hundreds of cancer tissues and normal tissue samples (7). The MiOncoCirc database collects thousands of circRNA profiles in cancer tissues by performing exome capture RNA sequencing (8). The Circatlas database contains circRNA profiles from thousands of samples across 19 different normal tissues, which show that most circRNAs are expressed at low levels (which differs from long non-coding RNAs (lncRNAs), micro RNAs (miRNAs), and mRNAs) and are cell-type specific (9, 10). The exoRbase is a collection of exosomal circRNAs, IncRNAs, and mRNAs from patients with cancer and healthy controls (11).

Moreover, the mechanisms of circRNA biogenesis are unclear, particularly those governing aberrant circRNA expression in cancer tissues. Studies in this field have described data supporting a back-splicing model, in which the double ends of a pre-mRNA fragment ligate to form a closed circular structure (12), although the driving force and machinery mediating back-splicing remain unclear. The alternative splicing factor Quaking has been implicated in circRNA regulation, as it has been reported to alter circRNA expression during the epithelial-mesenchymal transition, which is a critical process in cancer metastasis (13). CircRNA formation is also likely associated with H3K79me2 histone modifications (14), which have been shown to regulate co-transcriptional alternative splicing (15).

Despite the growing body of evidence and data in circRNA research, their global alterations in cancer, the principles of back-splicing, and the driving force of circRNA biogenesis remain under investigation. Most studies of circRNA in cancer have not addressed these issues but rather focused on cancer tissues and para-cancer tissues without examining the role of host genes or plasma exosomes from patients with cancer.

To this end, we performed a comparative analysis to determine the circRNA landscape in cancer and normal tissues and identify cancer tissue-enriched circRNAs. We observed an association between splicing sites in cancer-enriched circRNAs and cancer mutations, which is a potential driving force of circRNA biogenesis in cancer. We also examined the expression of cancer tissue-enriched circRNAs in plasma exosomes from patients with cancer (e.g., HCC) and healthy controls. Our study provides new insight into the landscape and biogenesis of circRNAs in cancer tissue and the potential of circRNAs as cancer biomarkers.

**Methods**

**CSCD (IDCSC): circRNA expression dataset comprising cancer and normal tissues**
We downloaded circRNA expression datasets from the IDCSC database (http://gb.whu.edu.cn/IDCSC/#), which is the successor of the highly cited CSCD database(7) (http://gb.whu.edu.cn/CSCD/#). We reorganized the original data format of cancer-specific, normal-specific, and common circRNA counts into circRNA profiles of individual samples. We selected the circRNA profile analyzed by the CIRCexplorer(16) circRNA prediction algorithm against the GRch38 human reference genome. We removed circRNA counts <2 and samples harboring total circRNA counts <10. After removing samples with ambiguous information regarding tissue types, 265 cancer tissues and 319 normal tissues were included for analysis (Additional file Table S1).

The circRNA nomenclature describes the chromosome, two back-splicing sites, and strandness, such as “chr1_224952669_224968874_+”.

**ExoRbase: plasma exosome circRNA profile datasets**

We downloaded raw RNA sequencing data of plasma exosomes collected by exoRbase(11) (http://www.exorbase.org/) from the Sequence Read Archive database (https://www.ncbi.nlm.nih.gov/sra/). All 21 patients with HCC and 32 healthy controls were included. We used the CIRCexplorer (16) circRNA prediction algorithm and GRch38 human reference genome to analyze the raw RNA sequencing data for circRNA profiles, removing any circRNA counts <2.

**Statistical analysis**

We used R software (Version 3.6.0) algorithms to conduct basic visualization and statistical analyses, including density plots, violin plots, line plots, Venn diagrams, bar plots, heatmaps, t-distributed stochastic neighbor embedding (t-SNE), and principal component analysis (PCA).

**Gene functional enrichment**

Metascape(17) (http://metascape.org/) is an online tool useful for functional enrichment analysis. We selected the Gene Prioritization by Evidence Counting algorithm and Reactome and Gene Ontology databases. The parameters for pathway and process enrichment were defined as follows: min overlap = 3, p-value (accumulative hypergeometric p-values) cutoff = 0.01, and min enrichments = 1.5. The parameters of protein-protein interaction enrichment were set as follows: min network size = 3, max network size = 500.

**LASSO regularization analysis**

We used the R package “glmnet” (18) to perform the least absolute shrinkage and selection operator (LASSO) regularization analysis, which is a statistical learning method. For the training set, we randomly selected 70% of cancer and normal tissue samples, with the other 30% comprising a validation set. For LASSO regularization, 50% of the training set was randomly sampled, and LASSO regression was applied for 50 repetitions. Five-fold cross-validation and Akaike information criterion (AIC) analyses were performed to estimate the expected generalization error and the selected optimal value of the “1-se”
lambda parameter. An adaptive general linear model to select for normal tissue-enriched circRNAs was constructed, with the random seeds set to 42 to ensure reproducibility of the results.

**Weighted gene co-expression network analysis**

We used the R package “WGCNA” (19) to perform co-expression analysis of the circRNAs and identify circRNA co-expression modules that were positively correlated with cancer tissues.

The parameters were selected as follows: (1) For the top 10% of stably expressed circRNAs: power = 4, max block size = 21,249, min module size = 5, reassign threshold = 0, merge cut height = 0.25, type of correlation = Pearson; (2) For the top 10%–20% of stably expressed circRNAs: power = 4, max block size = 23,709, min module size = 5, reassign threshold = 0, merge cut height = 0.25, type of correlation = Pearson; (3) For the top 20%–30% of stably expressed circRNAs: power = 6, max block size = 20,768, min module size = 5, reassign threshold = 0, merge cut height = 0.25, type of correlation = Pearson.

**Support vector machine**

We used the R package “caret” and “e1071” (18) to construct a support vector machine, which is a type of machine learning model. For the training set, we randomly selected 70% of cancer and normal tissue samples, with the other 30% comprising a validation set. We used the training set to train a support vector machine model to perform binary classification of cancer and normal tissues, and we used the validation set (which was not used for feature selection in the LASSO regularization or support vector machine training) to evaluate the predictive performance of the model. During model training, performance was improved by using the support vector machine tuning function, which optimally determines the “gamma” and “cost” parameters by five-fold cross-validation. The performance was then evaluated quantitatively and represented by a receiver operating characteristics curve, which reflected the accuracy of the circRNAs involved in the model to classify cancer and normal tissues. The random seeds were set to 42 to ensure the reproducibility of the results.

**IntOGen: cancer mutation database**

The IntOGen (20) (https://intogen.org/) database is a compendium of mutational cancer drivers. We used the “Search” function to find potential cancer-associated mutations at the two splice sites of cancer-specific, cancer tissue-enriched, and normal tissue-enriched circRNAs. The human reference genome used for this analysis was GRCh38.

**Results**

**CircRNA is less abundant and less stably expressed in cancer tissues**

In total, 265 cancer samples across 15 different tissue types and 319 normal tissues from 38 anatomical sites were included in this study (Fig. 1a, Additional file Table S1). The expression of circRNAs in individual cancer tissue samples was significantly lower than that in the normal tissue samples, with the
normal tissues showing a greater range of expression (Fig. 1b). Most cancer tissues harbored extremely low levels of circRNAs, whereas some normal tissues expressed very high circRNA levels. The total types of circRNAs did not increase with the increase in the total counts of circRNAs (Additional file Fig. S1a), suggesting that the nature of tumorigenesis, rather than the sequencing depth, was the underlying mechanism.

Most circRNAs were expressed at low levels in the analyzed tissues. Of the combined samples (584 in total), the top 10% of stably expressed circRNAs occurred in $\geq 20$ samples, top 20% in $\geq 7$ samples, top 30% in $\geq 4$ samples, and top 40% in $\geq 2$ samples. Approximately 50% of circRNAs occurred in only one of the 584 samples. This sparsity of circRNA expression was more prominent in cancer tissues than in normal tissues (Fig. 1c).

Based on the hypothesis that more commonly expressed circRNAs have a higher potential to serve as biomarkers, the 210,784 differentially expressed circRNAs were divided into four groups: the top 10%, top 10%–20%, and top 20%–30% stably expressed and other less stably expressed circRNAs. t-SNE embedding of the four groups of top stably expressed circRNA profiles demonstrated that samples from the same tissue type tended to be neighbors. t-SNE embedding of the top 10% stably expressed circRNAs showed the most distinct separation of the different sample types, regardless of whether PCA was performed (Fig. 1d; Additional file Fig. S1b). These results support previous observations that circRNA expression profiles exhibit high tissue type-specificity (10). Downstream analyses were therefore employed separately for the different expression groups (top 10%, 10%–20%, and 20%–30% of stably expressed circRNAs).

**Cancer-specific circRNAs and cancer-specific host genes are associated with differentiation, apoptosis, cell growth, and cell cycle**

The CSCD (IDCSC) database defines cancer-specific circRNAs as those observed exclusively in cancer tissues. Our analysis indicated that 82.16% of the circRNAs present in cancer tissues were also observed in normal tissues. The number of cancer-specific circRNAs was 7.7% of the total number of normal-specific circRNAs (Fig. 2a). Most of the 11,343 cancer-specific circRNAs were not stably expressed, and only 74 circRNAs were stably expressed in $\geq 4$ tissues, among which 62 circRNA were derived from protein-coding host genes (Fig. 2b, Additional file Table S2). Interestingly, the host genes of these 74 circRNAs displayed functional enrichment in myeloid cell differentiation, regulation of lymphocyte apoptotic process, and regulation of growth, which are likely related to oncogenesis (Fig. 2c).

Similarly, 97.65% of circRNA host genes observed in the cancer tissues were also detected in normal tissues, and only 229 host genes were cancer-specific (Fig. 2d), of which 10 genes were part of the largest MCODE module in the protein-protein-interaction network and are related to the process of the cell cycle (Fig. 2e).

**Cancer and normal tissues share a large proportion of top actively spliced host genes and demonstrate differences in functional enrichment**
We also found that some genes were more actively back-spliced, therefore serving as host genes of a greater number of differentially expressed circRNAs. The top 30 actively back-spliced host genes in cancer and normal tissues showed prominent overlap (Fig. 2f), despite the difference in the ranking. Dynein axonemal heavy chain 14 (DNAH14) was the third most actively back-spliced gene that was exclusively expressed in cancer tissues. Titin (TTN) was the top actively back-spliced gene to be exclusively expressed in normal tissues. TTN was recently reported to serve as a host gene for regulatory circRNAs with important roles in the splicing of muscle genes in the human heart (21). Moreover, functional enrichment analysis of the top 30 highly spliced host genes showed an overlap in function, but those in cancer tissues were more enriched in the ubiquitin-dependent protein catabolic process, cell cycle, and negative regulation of the catabolic process, whereas those in normal tissues were more enriched in “MET activates PTK2” signaling, response to muscle stretch, heart development, cell-matrix adhesion, and cellular response to organonitrogen compounds (Fig. 2g). The overlap between the top actively back-spliced host genes in cancer and normal tissues increased steadily as the quantile of ranking increased, whereas the least actively back-spliced host genes (quantile of ranking < 0.3) also showed significantly increased overlap (Fig. 2h).

**Normal tissue-enriched and top 10% stably expressed circRNAs are associated with essential biological processes**

The normal tissue-enriched circRNAs were directly selected by LASSO regularization analysis of the top 10%, top 10%–20%, and top 20%–30% stably expressed circRNAs (Additional files Fig. S2a, S2b). The 14 circRNAs among the top 10% of stably expressed circRNAs exhibited the most distinct enrichment in normal tissues (Fig. 3a), and the strongest ability to classify cancer and normal tissues via the support vector machine (Fig. 3b). All of these normal tissue-enriched circRNAs were derived from protein-coding host genes (Table 1). The functions of the genes hosting the normal tissue-enriched circRNAs were enriched in essential biological processes, including endosomal transport and the phosphate metabolic process (Fig. 3c).
Table 1
Normal tissue-enriched circRNAs associated with protein-coding genes

| circRNA                        | Host gene name (Protein-coding) |
|-------------------------------|----------------------------------|
| chr1_155438326_155459898_-    | ASH1L                            |
| chr1_1804418_1817875_-         | GNB1                             |
| chr10_1072115_1096246_+        | WDR37                            |
| chr11_36227084_36227430_+      | LDLRAD3                          |
| chr12_123498543_123499536_-    | RILPL1                           |
| chr14_24266429_24268619_-      | RABGGTA                          |
| chr16_11020192_11126146_+      | CLEC16A                          |
| chr17_31940285_31940486_+      | SUZ12                            |
| chr3_114350273_114351878_-     | ZBTB20                           |
| chr3_172247532_172251541_+     | FNDC3B                           |
| chr4_40934455_40945070_-       | APBB2                            |
| chr5_38523418_38530666_-       | LIFR                             |
| chr6_138943512_138944622_-     | REPS1                            |
| chr7_98190727_98194572_+       | LMTK2                            |

Cancer tissue-enriched and top 10% stably expressed circRNAs were predominantly back-spliced from oncogenes

Because of the more active and stable expression of circRNAs in normal tissues, LASSO regularization was not adequate to identify cancer tissue-enriched circRNAs. Therefore, we performed co-expression analysis (Additional file Fig. S4a). Co-expression modules positively correlated with cancer were the firebrick, orange-red, and salmon modules in the top 10% of stably expressed circRNA group; white-smoke, sienna, and dark-olive-green modules in the top 10%–20% of stably expressed circRNA group; and coral and deep-pink modules in the top 20%–30% of stably expressed circRNA group (Additional files Fig. S3a, S3b). Enrichment of these circRNAs in cancer tissues was observed, although with obvious variations between different cancer types. The cancer-enriched circRNAs were most highly conserved in HCC and T-cell acute lymphoblastic leukemia bone marrow, whereas they were sparsely expressed in pancreatic and kidney cancers (Fig. 4a). Similar to the normal tissue-enriched circRNAs, cancer tissue-enriched circRNAs within the top 10% stably expressed circRNAs showed the strongest ability to
distinguish cancer from normal tissues, whereas those from the top 10%–20% or 20%–30% of stably expressed circRNAs groups could not classify cancer and normal tissues (Fig. 4b, Additional file Fig. S4a). The 22 circRNAs with the strongest ability to classify the tissues were selected by LASSO regularization and examined by support vector machine (Fig. 4c, Additional files Fig. S4b, S4c), among which 18 circRNAs were derived from protein-coding host genes (Table 2). Interestingly, the function of cancer tissue-enriched circRNAs in the top 10% stably expressed circRNA group that was most significantly enriched was “oncogene-induced senescence”, which implicated the potential tendency of oncogenes to be circRNA host genes in cancer (Fig. 4d).

| circRNA                        | Host gene name (Protein-coding) |
|-------------------------------|---------------------------------|
| chr1_224952669_224968874_+    | DNAH14                          |
| chr1_224952669_224974153_+    | DNAH14                          |
| chr1_225231072_225266769_+    | DNAH14                          |
| chr10_110964124_110965061_+   | SHOC2                           |
| chr10_42631928_42636972_-     | ZNF33B                          |
| chr2_45546731_45553730_-      | SRBD1                           |
| chr2_55025085_55028220_-      | RTN4                            |
| chr20_13483226_13587370_-     | TASP1                           |
| chr20_13559007_13580981_-     | TASP1                           |
| chr20_13559007_13587370_-     | TASP1                           |
| chr20_13569506_13587370_-     | TASP1                           |
| chr4_118105017_118114960_+    | NDST3                           |
| chr4_177353307_177360677_+    | NEIL3                           |
| chr5_109781395_109789527_+    | MAN2A1                          |
| chr5_36207169_36227565_-      | NADK2                           |
| chr7_116110707_116112038_-    | TFEC                            |
| chr8_96879797_96880005_+      | CPQ                             |
| chrX_6150994_61517771_-       | NLGN4X                          |
Increased level of cancer tissue-enriched circRNAs related to cancer mutations are present in plasma exosomes from patients with HCC

We next examined the potential of the cancer-specific and cancer tissue-enriched circRNAs as exosomal biomarkers, using normal tissue-enriched circRNAs for comparison. We used HCC as an example, given that the expression of these cancer tissue-enriched circRNAs was relatively conserved in HCC (Fig. 4a), such as the circRNA chr1_224952669_224968874_+ (Additional file Fig. S4d). Additionally, the plasma exosomes from patients with HCC comprised the largest pool of such data in the exoRbase database.

All 21 samples from patients with HCC and 32 samples from healthy controls were included in our analyses. PCA demonstrated that circRNA expression was less variable in the plasma exosomes from patients with HCC than those from healthy controls (Fig. 5a). The circRNA was less abundant in the plasma exosome from patients with HCC than in those from healthy controls (Fig. 5b), similar to cancer and normal tissue samples (Fig. 1b), which has not been reported previously. The sparsity of circRNA expression was also observed (Fig. 5c) but was similar in patients with HCC and in healthy controls, different from that in cancer and normal tissues (Fig. 1c).

Thereafter, we determined the expression of cancer-specific, cancer tissue-enriched, and normal tissue-enriched circRNAs in the plasma exosomes. Only 10 of the 74 cancer-specific circRNAs were captured in exosomes, most of which were expressed at very low levels in the plasma exosomes of both patients with HCC and healthy controls (Fig. 5g). Therefore, these cancer-specific circRNAs were less likely to be ideal exosomal biomarkers. In contrast, cancer tissue-enriched circRNAs were more abundant in the exosomes, with some displaying elevated expression in HCC exosomes. Specifically, chr1_224952669_224968874_+ (circ-DNAH14) was a cancer tissue-enriched circRNA with significantly elevated levels in plasma exosomes from patients with HCC (Fig. 5e, Fig. 5g). In comparison, most normal tissue-enriched circRNAs were more abundant and showed higher levels in the exosomes of healthy controls (Fig. 5g).

We also investigated the associations between circRNA back-splicing sites and cancer mutations indexed by the IntOGen database. chr3:111598119:G > C of CD96 and chr10:6430809:C > A of protein kinase C theta (PRKCQ) were splice variants related to cancer-specific circRNAs chr3_111598119_111606792_+ and chr10_6430809_6442081_ (Fig. 5d). chr1:224952669:G > A is a splice acceptor variant of DNAH14 that is associated with two cancer tissue-enriched circRNAs, chr1_224952669_224968874_+ and chr1_224952669_224974153_+. Furthermore, chr1_224952669_224968874_+ was steadily expressed in exosomes from patients with HCC but was rarely expressed in healthy controls, suggesting its potential value as a biomarker. chr10:110964124:G-> is a splice acceptor variant of SHOC2 leucine-rich repeat scaffold protein (SHOC2), although the corresponding cancer tissue-enriched circRNA chr10_110964124_110965061_+ was more abundant in the exosomes from healthy controls.

chr2:45546731: C > A is a splice donor variant of S1 RNA binding domain 1 (SRBD1), and the associated circRNA chr2_45546731_45553730_- was elevated in a specific subgroup of patients with HCC (Fig. 5e). However, no splice donor or splice acceptor variants were observed to be associated with the normal tissue-enriched circRNAs, despite a splice region variant chr5:38530666:C > G of LIF receptor subunit...
alpha (LIFR), which was related to circRNA chr5:38523418_38530666_. The splice region variant was much less significantly associated with back-splicing in circRNA formation than the splice donor variant or splice acceptor variant (Fig. 5f).

**Discussion**

To our knowledge, there has been no comparative analysis of the circRNA profiles in pan-cancer and normal tissues, or reports regarding associations between cancer mutations and circRNAs. We integrated the comparative analysis of circRNA profiles in pan-cancer and normal tissues with our analysis of the plasma exosomal circRNA landscape to account for the fact that circRNAs in exosomes are secreted by a wide variety of normal tissues. This was important because differential analysis of cancer and para-cancer tissue was not adequate to establish which of the highly expressed circRNAs are potential plasma exosome biomarkers.

In line with this approach, we collected the circRNA profiles from the CSCD (IDCSC) database, as it contains the most balanced number of cancer and normal tissues. First, we revisited the concept of cancer-specific circRNAs (circRNAs expressed in cancer tissues, but not in normal tissues), as proposed by the developer of the CSCD database (7). Overall, cancer-specific was not an ideal criterion for screening potential circRNA biomarkers. Most cancer-specific circRNAs were expressed at low levels in cancer tissues, and even the most stably expressed cancer-specific circRNAs were present at very low levels in plasma exosomes. In contrast, cancer tissue-enriched circRNAs were more stably expressed in cancer, with their host genes enriched in the category of “oncogene-induced senescence”. Oncogene-induced senescence is a cellular system responsive to oncogenic signaling, which is reported to be a "double-edged sword" that can either induce or inhibit oncogenesis (22). The normal tissue-enriched circRNAs were stably expressed in a variety of normal tissues but rarely expressed in the cancer tissues, suggesting that expression of these circRNAs was lost during the transition from normal to cancer tissues.

We observed that the total circRNA count was less abundant in cancer tissues than in normal tissues, which was also the case for plasma exosomes. However, the total types of circRNAs did not evidently elevate with the increase in the total counts of circRNAs, suggesting that the sequencing depth was not the reason for this difference. Because circRNA are long-lived RNA molecules, the rapid proliferation of cancer cells may lead to a decreased circRNA abundance, as observed in colorectal and ovarian cancer (23). Furthermore, changes in the level of splicing factors involved in circRNA biogenesis may contribute to decreased circRNA levels (24). Notably, the mechanism underlying the global reduction of circRNAs in plasma exosomes from patients with HCC remains to be investigated.

In this study, the potential role of DNAH14 as an important circRNA host gene in cancer was highlighted. DNAH14 was the third-highest back-spliced host gene in pan-cancer tissues but was not among the top back-spliced host genes in the normal tissues, although the overlap between the top back-spliced genes in cancer and normal tissues was considerable. DNAH14 is the host gene of three cancer-enriched
circRNAs, of which chr1_224952669_224968874_+ and chr1_224952669_224974153_+ were associated with cancer splice acceptor variant chr1:224952669:G > A. Particularly, chr1_224952669_224968874_+ was significantly elevated in plasma exosomes from patients with HCC compared to in healthy controls. Thus, we hypothesize that circRNA chr1_224952669_224968874_+ is a biomarker of HCC or even pan-cancer, and is likely reflective of the cancer mutation chr1:224952669:G > A, which may be associated with dynein expression and therefore centrosome abnormalities present in cancer cells.

**DNAH14** encodes a heavy chain of axonemal dynein, a microtubule-associated motor protein that participates in maintaining the integrity of the centrosome, which is often numerically, positionally, or structurally dysregulated in cancer (25). In fact, dynein encoding genes (DNAH family) are among the most frequently mutated genes in cancer (20). In recent studies, somatic mutations in DNAH genes have been associated with a higher chemotherapy response rate in patients with gastric cancer (26). These findings and the literature highlight DNAH14 as a host gene that should be further examined in research of circRNA in cancer.

There were several limitations to our study. First, we did not analyze a pan-cancer plasma exosome circRNA profile, as the resources of RNA sequencing data of plasma exosomes from patients with cancer are limited. Second, the cancer mutations were not inferred from cancer tissue samples involved in the analysis of cancer-enriched circRNAs, as mutation data was not provided by the circRNA database. Therefore, to validate our hypothesis, efforts are needed to collect the circRNA profile and mutation data for cancer tissues, together with the plasma exosome circRNA profile, in a large cohort of patients with cancer.

**Conclusion**

Our bioinformatic analyses provide insights into the characteristics of the circRNA landscape in cancer and the potential of cancer mutation-associated and cancer tissue-enriched RNAs as plasma exosomal cancer biomarkers. Furthermore, our results highlight **DNAH14** as a host gene that should be further examined in studies of circRNA in cancer.

**List Of Abbreviations**

circRNA: circular RNA; HCC: hepatocellular carcinoma; TNM: tumor, node, metastasis; DNAH14: CSCD (IDCSC): cancer-specific circRNA database (Interactional Database of Cancer-Specific CircRNAs); t-SNE: t-distributed stochastic neighbor embedding; PCA: principal component analysis; LASSO: least absolute shrinkage and selection operator; MET: MET proto-oncogene, receptor tyrosine kinase; PTK2: protein tyrosine kinase 2

**Declarations**

**Ethics approval and consent to participate**
Consent for publication

Not applicable.

Availability of data and material

This study has not generated new data or materials. The data and codes are available in the GitHub repository (https://github.com/Selecton98/CircRNA_pan-cancer).

Competing interests

The authors declare that they have no competing interests.

Funding

This study was supported by grants from WBE Liver Fibrosis Foundation (CFHPC 2020021), Beijing Dongcheng District outstanding talent funding project and the Beijing Undergraduate Training Programs for Innovation and Entrepreneurship (202010023046).

Authors’ contributions

XW conceptualized the study, analyzed the data and drafted the work; YD and GW collected the data and revised the work; ZW and YS collected the data; YZ designed the project and revised the work. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Meng S, Zhou H, Feng Z, Xu Z, Tang Y, Li P, et al. CircRNA: functions and properties of a novel potential biomarker for cancer. Molecular cancer. 2017;16(1):94.
2. Wang S, Zhang K, Tan S, Xin J, Yuan Q, Xu H, et al. Circular RNAs in body fluids as cancer biomarkers: the new frontier of liquid biopsies. Molecular cancer. 2021;20(1):13.
3. Tang W, Fu K, Sun H, Rong D, Wang H, Cao H. CircRNA microarray profiling identifies a novel circulating biomarker for detection of gastric cancer. Molecular cancer. 2018;17(1):137.
4. Ju H-Q, Zhao Q, Wang F, Lan P, Wang Z, Zuo Z-X, et al. A circRNA signature predicts postoperative recurrence in stage II/III colon cancer. EMBO Mol Med. 2019;11(10):e10168-e.
5. Zhang PF, Gao C, Huang XY, Lu JC, Guo XJ, Shi GM, et al. Cancer cell-derived exosomal circUHRF1 induces natural killer cell exhaustion and may cause resistance to anti-PD1 therapy in hepatocellular
cancer. Molecular cancer. 2020;19(1):110.

6. Vromman M, Vandesompele J, Volders PJ. Closing the circle: current state and perspectives of circular RNA databases. Briefings in bioinformatics. 2021;22(1):288-97.

7. Xia S, Feng J, Chen K, Ma Y, Gong J, Cai F, et al. CSCD: a database for cancer-specific circular RNAs. Nucleic acids research. 2018;46(D1):D925-d9.

8. Vo JN, Cieslik M, Zhang Y, Shukla S, Xiao L, Zhang Y, et al. The Landscape of Circular RNA in Cancer. Cell. 2019;176(4):869-81.e13.

9. Ji P, Wu W, Chen S, Zheng Y, Zhou L, Zhang J, et al. Expanded Expression Landscape and Prioritization of Circular RNAs in Mammals. Cell reports. 2019;26(12):3444-60.e5.

10. Wu W, Ji P, Zhao F. CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. Genome biology. 2020;21(1):101.

11. Li S, Li Y, Chen B, Zhao J, Yu S, Tang Y, et al. exoRBase: a database of circRNA, IncRNA and mRNA in human blood exosomes. Nucleic acids research. 2018;46(D1):D106-d12.

12. Chen LL. The expanding regulatory mechanisms and cellular functions of circular RNAs. Nature reviews Molecular cell biology. 2020.

13. Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, et al. The RNA binding protein quaking regulates formation of circRNAs. Cell. 2015;160(6):1125-34.

14. Chen JB, Dong SS, Yao S, Duan YY, Hu WX, Chen H, et al. Modeling circRNA expression pattern with integrated sequence and epigenetic features demonstrates the potential involvement of H3K79me2 in circRNA expression. Bioinformatics (Oxford, England). 2020.

15. Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, et al. MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic acids research. 2018;46(D1):D303-d7.

16. Zhang X-O, Wang H-B, Zhang Y, Lu X, Chen L-L, Yang L. Complementary Sequence-Mediated Exon Circularization. Cell. 2014;159(1):134-47.

17. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature communications. 2019;10(1):1523.

18. Yu S, Li Y, Liao Z, Wang Z, Wang Z, Li Y, et al. Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. Gut. 2020;69(3):540-50.

19. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008;9:559.

20. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nature reviews Cancer. 2020;20(10):555-72.

21. Tijsen AJ, Cócer Ortega L, Reckman YJ, Zhang X, van der Made I, Aufiero S, et al. TTN Circular RNAs Create A Backsplice Motif Essential for SRSF10 Splicing. Circulation. 2021.
22. Liu X-l, Ding J, Meng L-h. Oncogene-induced senescence: a double edged sword in cancer. Acta Pharmacologica Sinica. 2018;39(10):1553-8.

23. Bachmayr-Heyda A, Reiner AT, Auer K, Sukhbaatar N, Aust S, Bachleitner-Hofmann T, et al. Correlation of circular RNA abundance with proliferation–exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis, and normal human tissues. Scientific reports. 2015;5:8057.

24. Goodall GJ, Wickramasinghe VO. RNA in cancer. Nature Reviews Cancer. 2021;21(1):22-36.

25. Priyanga J, Guha G, Bhakta-Guha D. Microtubule motors in centrosome homeostasis: A target for cancer therapy? Biochimica et biophysica acta Reviews on cancer. 2021;1875(2):188524.

26. Zhu C, Yang Q, Xu J, Zhao W, Zhang Z, Xu D, et al. Somatic mutation of DNAH genes implicated higher chemotherapy response rate in gastric adenocarcinoma patients. Journal of translational medicine. 2019;17(1):109.

Figures
Figure 1

Characteristics of circRNA landscape in cancer and normal tissues. a. Composition of the 265 cancer tissue samples and 319 normal tissue samples. b. Total count of circRNAs in each cancer and normal tissue sample. c. Total types of circRNAs harbored by different proportions of cancer and normal tissue samples. d. t-SNE embedding of circRNA profiles in 584 tissue samples, after PCA. The top 10%, top
10%–20%, top 20%–30% of stably expressed circRNAs and the others were analyzed separately. Red: cancer tissues; blue: normal tissues. Rainbow: tissue types.

**Figure 2**

Cancer-specific circRNAs, cancer-specific host genes, and top actively back-spliced host genes in cancer and normal tissues. a. Venn diagram showing the overlap between the circRNA profiles of cancer and normal tissues. b. Types of cancer-specific circRNAs harbored by different counts of samples. c.
Functional enrichment of the host genes bearing cancer-specific circRNAs. 

d. Venn diagram showing overlap between the host genes of circRNAs in cancer and normal tissues.

e. Most significant MCODE module in the protein-protein interaction network involving host genes of cancer-specific circRNAs.

f. Top 30 actively back-spliced host genes in cancer and normal tissues, respectively. The proportion of total types of circRNAs related to the same host gene represents the back-splicing activity of the host gene.

g. Functional enrichment of the top 30 actively back-spliced host genes in the cancer tissues and normal tissues.

h. Overlap between the top actively back-spliced circRNAs in cancer and normal tissues.

Fig. 3
Figure 3

Normal tissue-enriched circRNAs identified using machine learning methods.  

a. Heatmap illustrating the expression of normal tissue-enriched circRNAs selected by LASSO regularization among the top 10%, top 10%–20%, and top 20%–30% of stably expressed circRNAs. Turquoise: top 10%; yellow: top 10%–20%; gray: top 20%–30%. Red: cancer tissues; blue: normal tissues. Rainbow: tissue types.  

b. Receiver operating characteristic curve showing the sensitivity of normal tissue-enriched circRNAs to classify cancer and normal tissues via the support vector machine. The circRNAs among the top 10%, top 10%–20%, and top 20%–30% of stably expressed circRNAs were computed separately. Blue: training set; red: validation set.  

c. Functional enrichment of host genes of normal tissue-enriched circRNAs among the top 10% of stably expressed circRNAs,
Figure 4

Cancer tissue-enriched circRNAs identified by co-expression analysis. a. Heatmap illustrating the expression of cancer tissue-enriched circRNA modules identified by co-expression analysis. Firebrick, orange-red, salmon, white-smoke, sienna, dark-olive-green, coral, and deep-pink represent co-expression modules. Red: cancer tissues; blue: normal tissues. Rainbow: tissue types. b. Receiver operating characteristic curve showing the sensitivity of cancer tissue-enriched circRNAs to classify cancer and...
normal tissues via the support vector machine. Different circRNA co-expression modules were computed. Blue: training set; red: validation set. c. Receiver operating characteristic curve showing the sensitivity of cancer tissue-enriched circRNAs to classify cancer and normal tissues via the support vector machine. These cancer tissue-enriched circRNAs were selected from the firebrick, orange-red, and salmon co-expression modules. Blue: training set; red: validation set. d. Functional enrichment of host genes bearing cancer tissue-enriched circRNAs among the top 10% stably expressed circRNAs, which are the firebrick, orange-red, and salmon co-expression modules.
Figure 5

CircRNA expression in exosomes from patients with HCC and associated cancer mutations. a. Plasma exosomal circRNA profiles from patients with HCC (HCC) and healthy controls (HC). b. Total circRNA count in each plasma exosome sample from patients with HCC and healthy controls. c. Total types of circRNAs harbored by different proportions of plasma exosomes from patients with HCC and healthy controls. d. Expression of cancer-specific circRNAs in plasma exosomes of patients with HCC and healthy controls, and associated cancer mutations. e. Expression of cancer tissue-enriched circRNAs in plasma exosomes of patients with HCC and healthy controls, and associated cancer mutations. f. Expression of normal tissue-enriched circRNAs in plasma exosomes of patients with HCC and healthy controls. g. Expression of cancer-specific circRNAs, cancer tissue-enriched circRNAs and normal tissue-enriched circRNA in plasma exosomes of patients with HCC and healthy controls.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureAbstract.jpg
- MSpancancercirc2AdditionalFile.docx
- MSpancancercirc2AdditionalFile2.xlsx