International Journal of
Molecular Sciences

MDPI

Article

Systematical Identification of Breast Cancer-Related Circular RNA Modules for Deciphering circRNA Functions Based on the Non-Negative Matrix Factorization Algorithm

Shuyuan Wang [1,†], Peng Xia [1,†], Li Zhang [1,†], Lei Yu [1,†], Hui Liu [1], Qianqian Meng [1], Siyao Liu [1], Jie Li [1], Qian Song [1], Jie Wu [1], Weida Wang [1], Lei Yang [1,*], Yun Xiao [1,2,*] and Chaohan Xu [1,*]

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China; bioccwsy@gmail.com (S.W.); xiapeng231515@outlook.com (P.X.); Bio_MiniZhang@outlook.com (L.Z.); biomathnmghebyulei@outlook.com (L.Y.); liuhui870320@gmail.com (H.L.); mqq1992hmu@outlook.com (Q.M.); liusiyao29@outlook.com (S.L.); jacklee2THU@outlook.com (J.L.); songqian@hrbmu.edu.cn (Q.S.); wujie_bio@outlook.com (J.W.); zjhzwang@outlook.com (W.W.)

[2] Key Laboratory of Cardiovascular Medicine Research, Ministry of Education, College of Pharmacy, Harbin Medical University, Harbin 150081, China

* Correspondence: leiyang@hrbmu.edu.cn (L.Y.); xiaoyun@ems.hrbmu.edu.cn (Y.X.); chaohanxu@hrbmu.edu.cn (C.X.); Tel.: +86-13100883633 (C.X.)

† These authors contributed equally to this work.

Received: 21 November 2018; Accepted: 12 February 2019; Published: 20 February 2019

Abstract: Circular RNA (circRNA), a kind of special endogenous RNA, has been shown to be implicated in crucial biological processes of multiple cancers as a gene regulator. However, the functional roles of circRNAs in breast cancer (BC) remain to be poorly explored, and relatively incomplete knowledge of circRNAs handles the identification and prediction of BC-related circRNAs. Towards this end, we developed a systematic approach to identify circRNA modules in the BC context through integrating circRNA, mRNA, miRNA, and pathway data based on a non-negative matrix factorization (NMF) algorithm. Thirteen circRNA modules were uncovered by our approach, containing 4164 nodes (80 circRNAs, 2703 genes, 63 miRNAs and 1318 pathways) and 67,959 edges in total. GO (Gene Ontology) function screening identified nine circRNA functional modules with 44 circRNAs. Within them, 31 circRNAs in eight modules having direct relationships with known BC-related genes, miRNAs or disease-related pathways were selected as BC candidate circRNAs. Functional enrichment results showed that they were closely related with BC-associated pathways, such as 'KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAYS IN CANCER', 'REACTOME IMMUNE SYSTEM' and 'KEGG MAPK SIGNALING PATHWAY', 'KEGG P53 SIGNALING PATHWAY' or 'KEGG WNT SIGNALING PATHWAY', and could sever as potential circRNA biomarkers in BC. Comparison results showed that our approach could identify more BC-related functional circRNA modules in performance. In summary, we proposed a novel systematic approach dependent on the known disease information of mRNA, miRNA and pathway to identify BC-related circRNA modules, which could help identify BC-related circRNAs and benefits treatment and prognosis for BC patients.

Keywords: Circular RNA (circRNA); breast cancer; non-negative matrix factorization (NMF) algorithm

1. Introduction

Breast cancer (BC) is the most frequent malignancy in women, affecting more than 10% of women in western countries [1]. To improve the BC diagnosis and therapy with efficiency, it is

imperative to explore the molecular mechanisms of BC pathogenesis [2,3]. Therefore, some biological biomarkers involved in the development of BC, including mRNAs, lncRNAs and miRNAs, have been detected [4–6]. Further, several studies have shown that their corresponding molecular modules play important roles in BC [7–9]. How to efficiently identify these molecular modules that could be potentially used as diagnostic markers and therapeutic targets has been a big challenge.

With the enormous development in the field of high-throughput RNA sequencing technology, a novel class of endogenous RNA, circular RNA (circRNA), has been extensively studied [10,11]. Some researches demonstrated that circRNAs could be involved in many biological processes, including regulation of transcription [12], neuronal development [13], cell cycle control [14], and tumorigenesis [15,16]. For instance, Simon J. Conn et al. suggested that circRNAs biogenesis could be modified during the human epithelial-mesenchymal transition (EMT), and more than 30% productions of circRNAs were dynamically controlled by the alternative splicing [17]. In addition, Guarnerio J. et al. discovered that the generation of fusion circRNAs from chromosomal translocations displayed remarkable ability in promoting cellular transformation in vitro and initiate tumors [15]. Liang et al. found that circRNA, hsa_circ_0008717 (namely circ-ABCB10), was significantly upregulated in BC tissue, and knockdown of this circRNA could restrain the proliferation of BC cells [18]. Fang et al. disclosed that the delivery of a circRNA circ-Ccnb1 suppressed the effect of p53 mutations and enhance tumor progression in BC patients [19]. The identification of circRNA biomarkers largely benefited the in-depth exploration and investigation of the developmental mechanisms of BC and provided more promises for BC patients' diagnosis and therapy [20–23].

Although some advance in biological protocols has been made, it is time-consuming and expensive to identify BC-circRNAs only by using experimental technologies. Thus, some systematical approaches have been developed and proposed to identify BC-related circRNAs. Lu et al. identified 1155 differentially expressed circRNAs in BC tissues through analysis of a genome-wide circRNA profile data and found the expression levels of six circRNAs were related to BC which participated in cancer-related pathways [24]. Chen et al. identified the functional roles of circEPSTI1 on proliferation, clonal formation, and apoptosis in three triple-negative breast cancer (TNBC) cell lines by knocking down experiments. They confirmed that circEPSTI1 binds to miR-4753 and miR-6809 as a miRNA sponge to affect TNBC proliferation and apoptosis [25]. Many BC-related circRNAs had been identified and promoted the development of circRNA research. However, the disease circRNA list of BC is relatively incomplete and the systematical researches for their relevant functions remain poor, which handles the BC diagnosis and therapy.

Thus, we integrated circRNA-mRNA, miRNA-mRNA, and pathway-mRNA data to identify BC-related circRNA modules based on a non-negative matrix factorization (NMF) algorithm [26], and deciphered relevant circRNA functions. NMF has been demonstrated to be a powerful tool for detecting modules in heterogeneous multi-omics data, and biological entities and mechanisms can be naturally described in the biological contexts [27]. Our approach could systematically identify BC-related circRNA modules relying on relatively complete disease information of mRNAs, miRNAs and pathways, which may provide more confident knowledge for the further identification of BC candidate circRNAs. Through integration analysis, thirteen circRNA modules, containing 4164 nodes (2703 genes, 80 circRNAs, 63 miRNAs and 1318 pathways) and 67,959 edges, were identified by our approach. Among them, we found 31 circRNAs in eight modules that closely related to known BC-related genes, miRNAs or pathways, which might be associated with the development and progression of BC. These identified circRNA modules could provide more insights into the investigation of their functional mechanisms, and will benefit the illumination of circRNA functions for clinical applications for BC patients in the future. To our knowledge, we first applied the NMF method to identify BC-related circRNA modules through the integration of multiple omics data from mRNAs, miRNAs and pathways, which largely facilitated the efficient prioritization and identification of BC candidate circRNAs and provided the potential circRNA biomarkers for the clinical diagnosis and treatment in BC patients.

2. Results

2.1. Identification of Differentially Expressed circRNAs and miRNAs in BC

In total, there are 8953 human circRNAs recorded in the circBase database (http://www.circbase.org/) were identified from the RNA-seq data of BC patients by using the UROBORUS tool (Figure 1A). Then, 854 circRNAs expressed in at least 50% of patients were retained. Since several studies have shown that differentially expressed circRNAs or miRNAs would like to be associated with the disease with high probabilities [28,29], 80 differential expression circRNAs (four upregulated and 76 downregulated) with fold change (FC) values >2 or <0.5 were gathered (Supplementary Table S1). And 63 miRNAs that differently expressed (FC values >2 or <0.5 and Wilcoxon signed rank $p < 0.005$) in at least 50% of BC patient samples were retained for the following analysis.

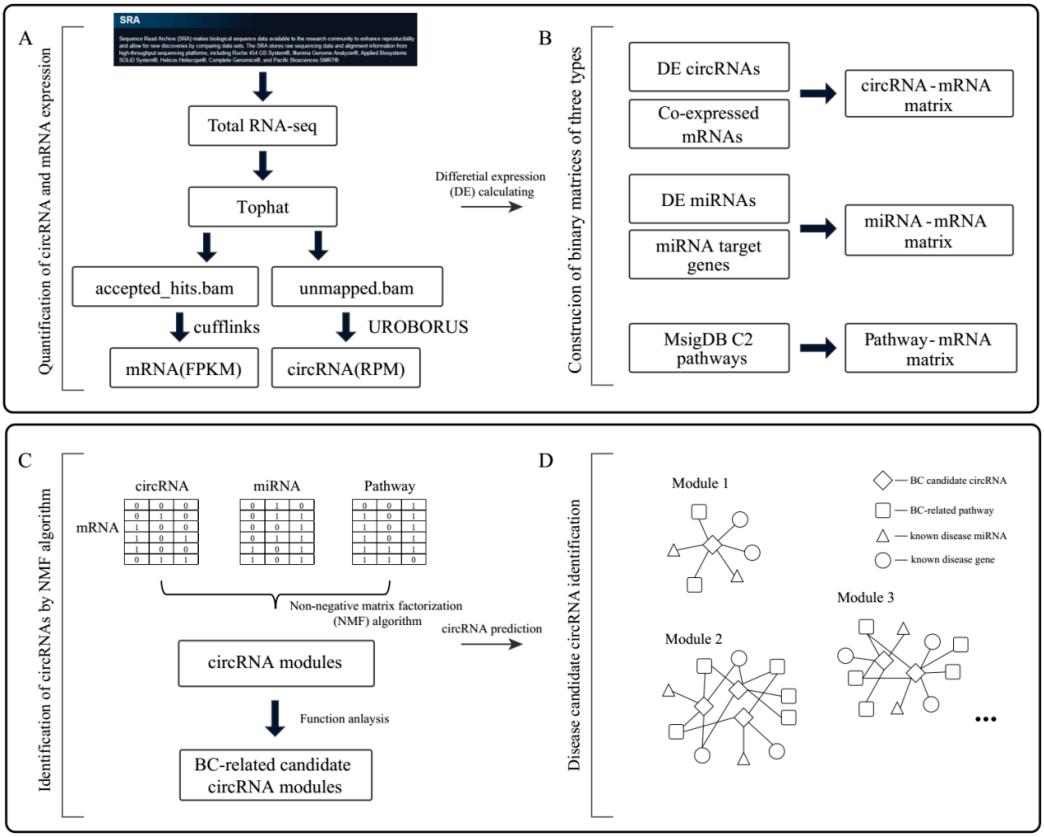

Figure 1. The flowchart of identification of breast cancer (BC)-related circRNA modules. The flowchart depicted a summary of the most important steps of the analysis workflow.

Differentially expressed circRNAs were used to construct the circRNA-mRNA co-expression relations by calculating the Pearson correlation coefficient (PCC) values (Figure 1B). In total, 80 circRNAs and 17,519 mRNAs associated with 124,486 co-expressed pairs (PCC > 0.4 and $p < 0.05$) were obtained. Further, 80 circRNAs and 13,251 mRNAs with degrees more than three (119,528 co-expressed circRNA-mRNA pairs) were selected and were used to characterize the circRNA-mRNA binary matrix. MiRNA and mRNA relationships were integrated by the miRNA-target gene data, which were collected from starBase, miRTarBase, and PITA. 63 miRNAs and 8385 mRNAs with more than three partners were retained and were used to construct the miRNA-mRNA binary matrix. The pathway-mRNA relations were integrated from pathway data obtained from the Molecular Signatures Database (http://software.broadinstitute.org/gsea/msigdb), which contained a large number of functional annotation information that was curated from BioCart, Kyoto Encyclopedia of Genes and Genomes (KEGG), the NCI Pathway Interaction Database (PID), and Reactome. Finally,

1329 pathways and relevant 8904 mRNAs were used to characterize mRNA-pathway binary matrix. Three characterized binary matrices in total contained 2703 common mRNAs, 63 miRNAs, 80 circRNAs and 1318 pathways (Table 1).

Table 1. Summary information of three characterized binary matrixes.

Association Matrix	#(circRNA/miRNA/patnway)	#(mRNA)	Dimensions
circRNA-mRNA matrix	80	2703	80 × 2703
miRNA-mRNA matrix	63	2703	63 × 2703
pathway-mRNA matrix	1318	2703	1318 × 2703

2.2. Identification of circRNA Modules Based on a Non-Negative Matrix Factorization (NMF) Algorithm

The NMF algorithm was previously shown to be a useful decomposition method for multivariate data, in which the existing features can be transformed into a lower dimensional space. This algorithm can be applied to many practical problems in bioinformatics and computational biology such as integration analysis of different data. Therefore, based on three binary matrices of circRNA-mRNA, miRNA-mRNA and pathway-mRNA, we used the NMF algorithm to identify modules that were more representative and associated with BC-related functions. When K (the default parameter ranges from 5 to 20) equals to 13, the value of objective function F reached the minimum Euclidean error and the corresponding 13 circRNA modules were generated, including 4164 nodes (80 circRNAs, 2703 genes, 63 miRNAs and 1318 pathways) and 67,959 edges. Subsequently, 9 circRNA modules (Table 2) having more than 10 GO biological process (BP) functional categories were retained as BC-related circRNA modules (see details in Methods and Materials), including 1174 mRNAs, 44 circRNAs, 30 miRNAs and 325 pathways.

Table 2. Summary of 9 circRNA modules, including 2703 genes, 80 circRNAs, 63 miRNAs and 1318 pathways.

Modules	Nodes	CircRNAs	mRNAs	miRNAs	Pathways	Edges
1	222	8	136	6	72	1069
2	415	8	271	7	129	3299
3	172	8	129	6	29	864
4	233	5	163	4	61	1375
5	382	8	271	7	96	2708
6	141	8	83	6	44	665
7	171	8	130	7	26	827
8	216	8	152	7	49	1237
9	331	7	217	6	101	2054

Modules 1 to 9 contained 222, 415, 172, 233, 382, 141, 171, 216 and 331 nodes (circRNAs, mRNAs, miRNAs or pathways), and 1069, 3299, 864, 1375, 2708, 665, 827, 1237 and 2054 edges, respectively (Figure 2 and Table 2). Within them, hsa_circ_0006528 in module 1 and module 3 has been validated to be related to BC [30]. There was a common gene named DDX3X in five modules including module 1, 3, 6, 7 and 8. DDX3X was abnormally expressed in breast epithelial cancer cells in which its expression was activated by HIF1A during hypoxia. Meanwhile, eight known BC-related genes—AKT1, CHEK2, ERBB2, PIK3CA, PPM1D, PTEN, SMAD4 and TSG101—were found in the nine modules (Supplementary Table S2). Twenty known BC-related miRNAs were also found: hsa-mir-7, hsa-let-7f, hsa-mir-103a, hsa-mir-130b, hsa-mir-135a, hsa-mir-144, hsa-mir-146a, hsa-mir-182, hsa-mir-185, hsa-mir-190a, hsa-mir-200a, hsa-mir-204, hsa-mir-216b, hsa-mir-224, hsa-mir-26a, hsa-mir-34b, hsa-mir-374b, hsa-mir-378a, hsa-mir-449a and hsa-mir-625 were also found in these modules (Supplementary Table S2).

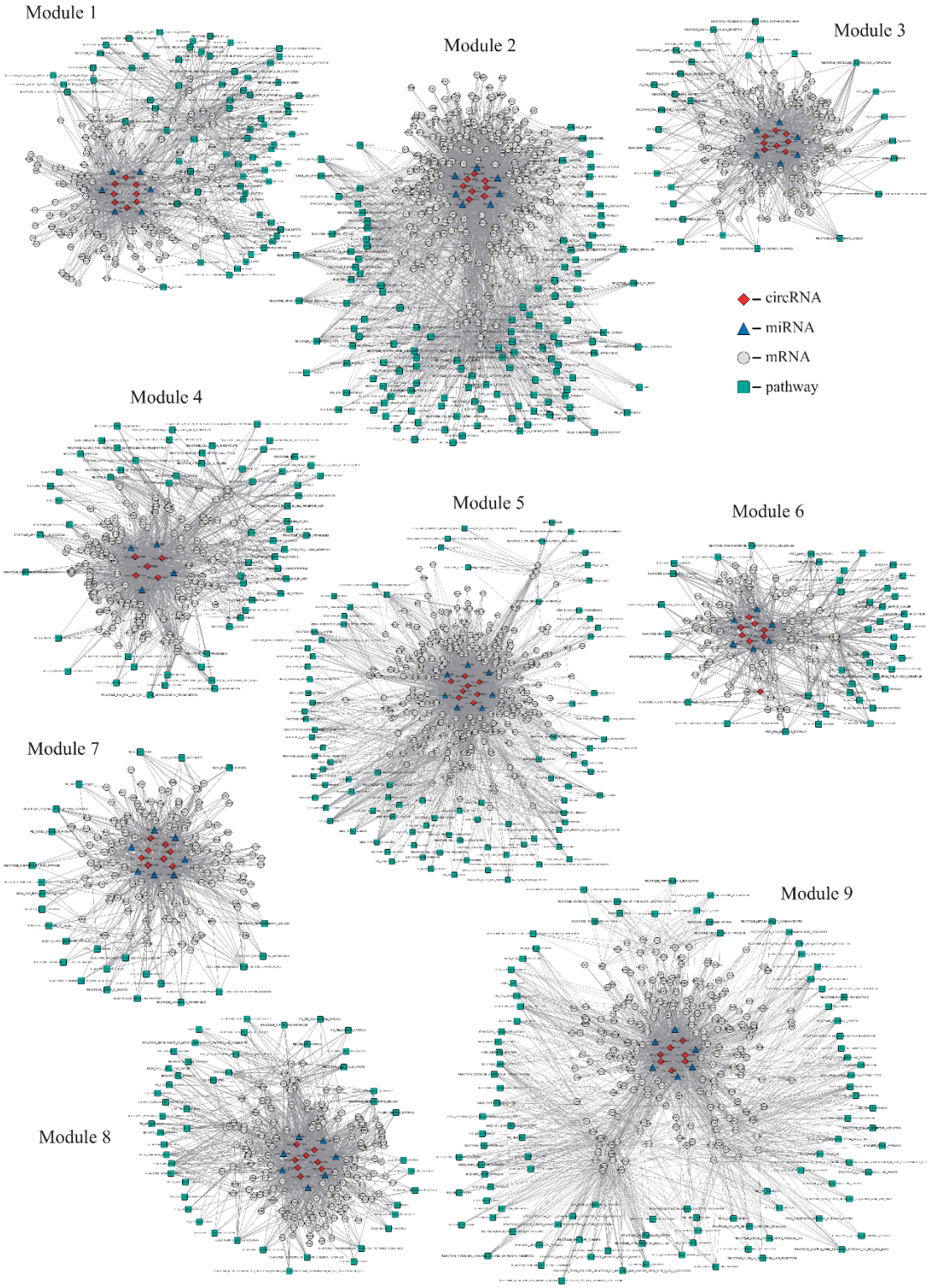

Figure 2. The overview of nine circRNA functional modules, including 1174 mRNAs, 44 circRNAs, 30 miRNAs and 325 pathways.

To better characterize the relationships between circRNAs and pathways or miRNAs in the nine modules, the normalized term overlap (NTO) scores [31] were calculated for each candidate circRNA-pathway pair and circRNA-miRNA pair (see details in Methods and Materials). Then, 44 circRNAs and 20 BC-related miRNA with high similar relations (NTO $\geq$ 0.5) were obtained

(Supplementary Table S3). Also, 44 circRNAs and 14 KEGG pathways with high similar relations (NTO $\geq$ 0.5) were obtained (Supplementary Table S4), including 'KEGG PATHWAYS IN CANCER', 'REACTOME IMMUNE SYSTEM', 'KEGG MAPK SIGNALING PATHWAY', 'KEGG CALCIUM SIGNALING PATHWAY', 'KEGG PROSTATE CANCER', 'PID ERBB1 DOWNSTREAM PATHWAY', 'PID P53 DOWNSTREAM PATHWAY', 'KEGG GNRH SIGNALING PATHWAY', 'KEGG P53 SIGNALING PATHWAY', 'KEGG SMALL CELL LUNG CANCER', 'KEGG WNT SIGNALING PATHWAY', 'PID CXCR4 PATHWAY', 'PID NOTCH PATHWAY' and 'REACTOME ACTIVATED TLR4 SIGNALLING'. Most of these pathways were cancer-related, showing that these circRNA-related modules in the nine modules may play important functional roles during the BC development and progression.

2.3. Prediction of Disease Candidate circRNAs in BC

To further identify candidate circRNAs that may be potentially associated with BC patients, circRNAs having more than four direct interaction partners (known BC genes, miRNAs or pathways, NTO score $\geq$ 0.5) were extracted (Supplementary Tables S2–S4, see details in Methods and Materials). Then, 31 unique candidate BC circRNAs were identified from module 1 to 9 (no circRNA in module 7), including 1, 8, 2, 5, 7, 4, 1 and 5 BC candidate circRNAs, respectively. Most of these circRNAs were associated with BC known disease genes (AKT1, PIK3CA, PPM1D, SMAD4 and TSG101) or miRNAs (hsa-let-7f, hsa-mir-7, hsa-mir-103a, hsa-mir-135a, hsa-mir-144, hsa-mir-146a, hsa-mir-182, hsa-mir-185, hsa-mir-190a, hsa-mir-204, hsa-mir-216b, hsa-mir-26a, hsa-mir-374b, hsa-mir-378a, hsa-mir-449a and hsa-mir-625). Especially, 30 out 31 BC candidate circRNAs: hsa_circ_0001222, hsa_circ_0002886, hsa_circ_0004458, hsa_circ_0004575, hsa_circ_0004910, hsa_circ_0007895, hsa_circ_0027842, hsa_circ_0079753 in module 2, hsa_circ_0002138, hsa_circ_0003614, hsa_circ_0003638, hsa_circ_0007766 in module 4, hsa_circ_0001725, hsa_circ_0007843, hsa_circ_0008362, hsa_circ_0017242, hsa_circ_0069244, hsa_circ_0073901, hsa_circ_0086375 in module 5, hsa_circ_0001558, hsa_circ_0017924, hsa_circ_0044177, hsa_circ_0069492 in module 6, hsa_circ_0001119, hsa_circ_0004513, hsa_circ_0007785, hsa_circ_0020399, hsa_circ_0037130 in module 9, hsa_circ_0003759 in module 3 and 4, hsa_circ_0001447 in module 3 and 8 were related to 'KEGG PATHWAYS IN CANCER', 'REACTOME IMMUNE SYSTEM' and 'KEGG MAPK SIGNALING PATHWAY', 'KEGG P53 SIGNALING PATHWAY' or 'KEGG WNT SIGNALING PATHWAY', which were closely involved with BC.

Interestingly, several circRNAs' corresponding parental genes, such as has_circ_0007766, has_circ_0017242, has_circ_0037130, has_circ_0003759, has_circ_0007843, has_circ_0086375 and has_circ_0003614, were respectively recorded as ERBB2, AKT3, NPRL3, LPP, ARHGAP32, NFIB and ASPH in the circBase database. All these genes were remarked as 'Cancer-related genes' or 'Disease related genes' in the Human Protein Atlas (https://www.proteinatlas.org/), which also suggested that these circRNAs may be potentially served as disease biomarkers for disease diagnosis or therapy during disease development or progression.

2.4. Comparison with Other circRNA Prioritization Approaches

To further evaluate the performance of our approach in the identification of BC-related circRNA modules, we compared the circRNA modules generated by our approach with those yielded by the MCL algorithm. MCL is a traditional cluster method for networks, which has been widely used for clustering of genes, proteins or other biomarkers according to their expression profile or other experimentally detected data [32,33]. According to the same comprehensive network, the MCL algorithm was performed by the Cytoscape plugin clusterMaker (the minimum number of nodes in each module was set to 20). Then, 8 modules (Supplementary Table S5 and Supplementary Figure S1) were generated, of which module 1 was the biggest one with 1678 nodes, and module 2 to 7 contained 60, 40, 36, 25, 22 and 20 nodes, respectively. Due to the different sizes of circRNA modules and the different nodes in each modules generated by MCL and our approach, it was impossible to directly compare the results of these two approaches in the identification of BC-related circRNAs. Thus, we

indirectly compared circRNA modules obtained from these two approaches by statistic proportion of nodes with known BC information. We found BC known disease circRNAs in our module 1 and 3 account for 0.45% and 0.58%, and in MCL module 1 account for 0.12%. Other nodes (include genes, miRNAs and pathways) with known BC information in circRNA modules of our approach account for 5.405%, 2.651%, 4.651%, 3.433%, 3.141%, 8.511%, 3.509%, 4.167% and 2.720% while 1.728%, 0, 0, 0, 0, 0, and 5% in MCL circRNA modules, respectively (Supplementary Table S5). The comparison results suggested that our approach identified circRNA modules with more BC-related information, which could be able to capture more characterization about BC. Furthermore, GO enrichment analysis suggested that more functional GO terms were enriched in modules generated by our approach (R-package 'clusterProfiler', Benjamini–Hochberg correction, FDR < 0.05), which demonstrated that our circRNA modules were more closely related to the development and progression of BC (Supplementary Table S5).

3. Discussion

Increasing numbers of disease-related molecular biomarkers, including gene, protein, miRNA, lncRNA, and circRNA could provide more promises to improve the diagnosis and treatment for BC patients [34–36]. However, relatively incomplete disease information about circRNA brings a challenge to biological researchers to uncover their functional mechanisms and roles. Towards this end, we developed a computational pipeline with the goal of identifying BC-related circRNA modules by integration of circRNA, mRNA, miRNA, and pathway data based on an NMF algorithm in this work. Our approach integrated known disease information and omics data, whereby we could identify BC candidate circRNAs and infer their functional roles.

Employing the systemic pipeline in 33 BC RNA-seq data with tumor and normal samples, we identified 13 circRNA modules in BC, containing 80 circRNAs, 2703 genes, 63 miRNAs and 1318 pathways with 6,795,944 interactions. After screening by functional enrichment analysis, nine circRNA modules potentially associated with BC were obtained. Within them, one circRNA hsa_circ_0006528 had been recognized as known disease circRNA. Simultaneously, eight genes and twenty miRNAs in circRNA modules have been validated as known BC biomarkers. Functional enrichment results showed that other 31 circRNAs were closely related with known disease miRNA or BC associated pathways. The circRNA prioritization result of our approach suggested that known disease information curated by circRNA direct partners, including genes, miRNAs and pathways, could give more chances to recognize disease related circRNAs. Comparison with other module identification methods like MCL, NMF algorithm identified more BC informative circRNA modules and could comprehensively characterize BC from circRNA perspective.

There are some limitations to our approach. Relative small numbers of BC candidate circRNAs were included and analyzed in our approach, which limited the ability of our approach in the prediction of circRNA's functions in human BC context. In addition, relatively strict screening criteria were adopted in functional modules identification, which tended to remove some meaningful circRNA modules. Thus, there are also some other proposed methods for the identification of disease-related factors and modules which could be used for reference [37–39]. For example, Chen X et al. developed decision tree learning-based model (EGBMMDA) for predicting miRNA-disease associations, by integrating the miRNA functional similarity, the disease semantic similarity, and known miRNA–disease associations [40]. BC-related circRNAs would be efficiently identified by this method if more functional categories are explicit for circRNAs.

In summary, we proposed a systematic approach to identify BC-related circRNA modules through the NMF algorithm. These identified circRNA modules provide novel insights into the potentially BC-associated circRNAs, which will benefit the clinical applications of circRNA biomarkers for BC diagnosis, treatment and prognosis in the future.

4. Materials and Methods

4.1. Data Acquisitions

Paired-end RNA-seq data of SRP062132, which detected by "Illumina Genome Analyzer II" were downloaded from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP062132&go=go). This dataset included 15 disease samples and 18 normal samples (Figure 1A). The miRNA expression profile data, GSE83270, which detected by 'Exiqon miRCURY LNA microRNA array, 7th generation' (GPL22003) for 12 BC patients, including 12 BC patients, was downloaded from the GEO database. The corresponding miRNA target genes were obtained from starBase [41] (http://starbase.sysu.edu.cn/), miRTarBase [42] (http://mirtarbase.mbc.nctu.edu.tw/php/index.php) and PITA (https://genie.weizmann.ac.il/pubs/mir07/mir07_data.html). Pathway data were downloaded from the Molecular Signatures Database [43] (http://software.broadinstitute.org/gsea/msigdb) database, which contained a large number of annotated functional genes collected from existing public databases, such as BioCart, KEGG, PID, and Reactome. We selected pathway data from the curated gene sets (c2) in MsigDB V6.1, which contained a total of 1329 metabolic and signaling pathways.

4.2. Quantification of circRNA-mRNA, miRNA-mRNA, and Pathway-mRNA Binary Matrices

For each sample of BC patients in SRP062132, the RNA-seq reads were first mapped to the human reference genome (GRCh37/hg19, UCSC Genome Browser [44]) by the TopHat2 [45] tool, which was capable of detecting the canonical splicing event (Figure 1A). In addition, the unmapped reads were then used to identify the circRNAs by the pipeline proposed by UROBORUS [46]. During the process of quantification of human circRNAs, the unmapped reads were extracted to 20-bp anchors from head ends and tail ends. The short 20-bp paired-end seed sequence reads were aligned to the human reference genome (hg19) with a maximum of 2 bp mismatches using TopHat2 with a default parameter value. Balanced mapped junction (BMJ) reads and unbalanced mapped junction (UMJ) reads were generated as two sets spanning the spliced site. BMJ or UMJ reads were represented as reads aligned to the joining region of two back spliced exons with minimum 20 bp of overhang at any an end or with less than 20 bp of overhang at one end, respectively. To evaluate the relative expression of circRNAs in different disease and normal tissues, we normalized the number of circular reads to per kilobase per million reads sequenced (RPKM) values. To quantify the expression levels of mRNAs, we used Cufflinks [47] software to process the accepted hits.bam file in the TopHat2 results, which contained all reads mapped to the human reference genome. We also used RPKM value to identify the relative expression of each mRNA.

After recognized by the UROBORUS, circRNAs recorded in the circBase database and expressed in more than 50% patient samples were retained (Figure 1A). The differentially expressed circRNAs with FC values >2 or <0.5 were identified. In GSE83270, miRNAs with FC >2 or <0.5 and Wilcoxon signed rank test $p < 0.005$ were considered as differentially expressed miRNAs. The PCC was then used to measure the co-expression relationships between differentially expressed circRNAs and mRNAs. CircRNA-mRNA pairs with PCC > 0.4 and p-value < 0.05 were used to construct the binary matrix. As for the miRNAs and mRNAs, the corresponding miRNA target mRNAs relations were used to build the miRNA-mRNA binary matrix. If a pair of miRNA and mRNA was recorded in any one of the three database starBase, miRTarBase and PITA, the miRNA and mRNA was denoted as "1" in the miRNA-mRNA binary matrix. The pathway-mRNA binary matrix was similarly constructed based on pathway information from the Molecular Signatures Database.

4.3. Construction of circRNA Modules Basing on Non-Negative Matrix Factorization (NMF) Algorithm

We identified different numbers K of BC-related functional modules from these three matrices by using non-negative matrix factorization (NMF). The objective function F for NMF was defined as:

$$F(W, H) = \sum_{I=1}^{3} \|X_I - WH_I\|^2 \quad (1)$$

where X_I $(I \in (1, 2, 3))$ represented the characterized binary circRNA-mRNA, miRNA-mRNA, and pathway-mRNA matrix, respectively. The same penalization parameters for characterization of binary circRNA-mRNA, miRNA-mRNA, and pathway-mRNA matrices were assigned as described in Liu's NMF approaches [26], and the penalization parameters were set as default zero. W and H were both non-negative matrices. W was an $M \times K$ (M was the number of common mRNAs in three matrices) matrix representing the basis vector. H_I $(I \in (1, 2, 3))$ was a $K \times N$ (N is the numbers of circRNAs, miRNAs, and pathways) matrix, representing the coefficient vector in the dimension reduction process (Figure 1C). We selected different K (from 5 to 20) numbers and calculated the Euclidean errors between the input matrices, and the model reconstructed data. The Euclidean error measured the distance between the input matrices and the model reconstructed data. By comparing the Euclidean errors, we selected the smallest one to build the functional modules. W and H were updated at each iteration step by using the generalized multiplicative update rules as follows:

$$W_{ij} = W_{ij} \frac{(X_1 {H_1}^T + X_2 {H_2}^T + X_3 {H_3}^T)_{ij}}{(W(H_1 {H_1}^T + H_2 {H_2}^T + H_3 {H_3}^T))_{ij}} \quad (2)$$

$$(H_I)_{ij} = (H_I)_{ij} \frac{(W^T X_I)_{ij}}{(W^T W H_I)_{ij}}, I = 1, 2, 3. \quad (3)$$

It was worth noting that when we used the randomly generated initial matrices W and H_I ($I \in$ (1,2,3)) to minimize the Euclidean distance function, a local minimum solution occasionally appeared. We thus repeated the optimization procedure 100 times with random initial solution matrices to address this limitation. The lowest object function value was selected as the final factorization solution, and the selected value K meant that we finally got K modules. Then the obtained decomposing matrices W and H_I ($I \in$ (1,2,3)) were normalized through Z-score normalization by the following formula:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \quad (4)$$

where μ_i represented the mean value of elements in the *i-th* column of W matrix or in the *i-th* row of H_I ($I \in$ (1,2,3)) matrix, and σ_i was the corresponding standard deviation. The obtained Z-score values were used to determine each module members (including mRNAs, miRNAs, circRNAs and pathways) according to a published method [48]. For each column of matrix W (corresponding to an identified module), we separately retrieved the top 1% to top 10% ranked mRNAs according to the Z-score values to perform the GO BP enrichment analysis, by using R-package 'clusterProfiler' (Benjamini–Hochberg correction, FDR < 0.05). Then, when the top k% mRNAs enriched the most BP GO terms, we assigned the top k% genes to this module. Similarly, we assigned the top k% miRNAs, circRNAs or pathways in the corresponding row of matrix H_I ($I \in$ (1,2,3)) to the same module (Figure 1C). Subsequently, disease candidate circRNAs were identified according to the relationships between circRNAs and known BC genes, miRNAs or BC-related pathways in these functional modules (Figure 1D). Those circRNAs having more than four direct interaction partners (known BC genes, miRNAs or pathways) were identified as BC relative circRNAs. The known BC genes and miRNAs were obtained from GeneCards (https://www.genecards.org/) and HMDD (http://www.cuilab.cn/hmdd) respectively.

In addition, we calculated the normalized term overlap (NTO) score to further determine the relationships between circRNAs and pathways or miRNAs. The NTO score was calculated by the formula as follows:

$$NTO = \frac{|E_G \cap E_T|}{\min(|E_G|, |E_T|)} \quad (5)$$

where E_G represented the number of associated mRNAs for a specific circRNA, E_T represented the number of mRNAs associated with a pathway or miRNA, $|E_G \cap E_T|$ represents the number of common mRNAs for circRNAs and pathways or miRNAs, and $\min(|E_G|, |E_T|)$ represented the minimum numbers of mRNAs of circRNAs and pathways or miRNAs. The above processing was implemented using the R software environment.

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/20/4/919/s1.

Author Contributions: C.X., S.W. and P.X. conceived and designed the study; S.W. and P.X. processed the next generation sequencing data; L.Z. and L.Y. implemented the NMF algorithm; P.X., W.W., L.Z., and L.Y. analyzed the result data; Q.M., S.L., J.L., Q.S., and J.W. made the tables; C.X. wrote the paper; H.L., L.Y., Y.X. and S.W. revised the manuscript for important intellectual content; all authors have read and approved the final version prior to publication.

Funding: This work was supported in part by the National Natural Science Foundation of China [Grant Nos. 31871336, 61573122, 31501078, 61801150, 61803130]. Wu lien-teh youth science fund project of Harbin medical university [Grant Nos. WLD-QN1407]. The Health Department Science Foundation of Heilongjiang Province (Grant Nos. 2013128), the Education Department Science Foundation of Heilongjiang Province (Grant Nos. 12541415), the Postdoctoral project of Heilongjiang Province (Grant Nos. LBH-Z14130).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. DeSantis, C.E.; Ma, J.; Goding Sauer, A.; Newman, L.A.; Jemal, A. Breast cancer statistics, 2017, racial disparity in mortality by state. *CA: Cancer J. Clin.* **2017**, *67*, 439–448. [CrossRef]
2. Tokunaga, E.; Hisamatsu, Y.; Tanaka, K.; Yamashita, N.; Saeki, H.; Oki, E.; Kitao, H.; Maehara, Y. Molecular mechanisms regulating the hormone sensitivity of breast cancer. *Cancer Sci.* **2014**, *105*, 1377–1383. [CrossRef]
3. Kozlowski, J.; Kozlowska, A.; Kocki, J. Breast cancer metastasis—Insight into selected molecular mechanisms of the phenomenon. *Postęp. Hig. Med. Dośw.* **2015**, *69*, 447–451. [CrossRef]
4. Matsumoto, A.; Jinno, H.; Ando, T.; Fujii, T.; Nakamura, T.; Saito, J.; Takahashi, M.; Hayashida, T.; Kitagawa, Y. Biological markers of invasive breast cancer. *Jpn. J. Clin. Oncol.* **2016**, *46*, 99–105. [CrossRef]
5. Wang, W.; Luo, Y.P. MicroRNAs in breast cancer: oncogene and tumor suppressors with clinical potential. *J. Zhejiang Univ. Sci. B* **2015**, *16*, 18–31. [CrossRef] [PubMed]
6. Soudyab, M.; Iranpour, M.; Ghafouri-Fard, S. The Role of Long Non-Coding RNAs in Breast Cancer. *Arch. Iran. Med.* **2016**, *19*, 508–517. [PubMed]
7. Li, S.; Li, B.; Zheng, Y.; Li, M.; Shi, L.; Pu, X. Exploring functions of long noncoding RNAs across multiple cancers through co-expression network. *Sci. Rep.* **2017**, *7*, 754. [CrossRef]
8. Nogales-Cadenas, R.; Cai, Y.; Lin, J.R.; Zhang, Q.; Zhang, W.; Montagna, C.; Zhang, Z.D. MicroRNA expression and gene regulation drive breast cancer progression and metastasis in PyMT mice. *Breast Cancer Res. BCR* **2016**, *18*, 75. [CrossRef] [PubMed]
9. Yu, K.; Ganesan, K.; Miller, L.D.; Tan, P. A modular analysis of breast cancer reveals a novel low-grade molecular signature in estrogen receptor-positive tumors. *Clin. Cancer Res Off. J. Am. Assoc. Cancer Res.* **2006**, *12*, 3288–3296. [CrossRef] [PubMed]
10. Salzman, J.; Gawad, C.; Wang, P.L.; Lacayo, N.; Brown, P.O. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* **2012**, *7*, e30733. [CrossRef] [PubMed]
11. Danan, M.; Schwartz, S.; Edelheit, S.; Sorek, R. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res.* **2012**, *40*, 3131–3142. [CrossRef] [PubMed]
12. Zhang, Y.; Zhang, X.O.; Chen, T.; Xiang, J.F.; Yin, Q.F.; Xing, Y.H.; Zhu, S.; Yang, L.; Chen, L.L. Circular intronic long noncoding RNAs. *Mol. Cell* **2013**, *51*, 792–806. [CrossRef] [PubMed]

13. You, X.; Vlatkovic, I.; Babic, A.; Will, T.; Epstein, I.; Tushev, G.; Akbalik, G.; Wang, M.; Glock, C.; Quedenau, C.; et al. Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nat. Neurosci.* **2015**, *18*, 603–610. [CrossRef] [PubMed]
14. Du, W.W.; Yang, W.; Liu, E.; Yang, Z.; Dhaliwal, P.; Yang, B.B. Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. *Nucleic Acids Res.* **2016**, *44*, 2846–2858. [CrossRef] [PubMed]
15. Guarnerio, J.; Bezzi, M.; Jeong, J.C.; Paffenholz, S.V.; Berry, K.; Naldini, M.M.; Lo-Coco, F.; Tay, Y.; Beck, A.H.; Pandolfi, P.P. Oncogenic Role of Fusion-circRNAs Derived from Cancer-Associated Chromosomal Translocations. *Cell* **2016**, *166*, 1055–1056. [CrossRef] [PubMed]
16. Yang, W.; Du, W.W.; Li, X.; Yee, A.J.; Yang, B.B. Foxo3 activity promoted by non-coding effects of circular RNA and Foxo3 pseudogene in the inhibition of tumor growth and angiogenesis. *Oncogene* **2016**, *35*, 3919–3931. [CrossRef] [PubMed]
17. Conn, S.J.; Pillman, K.A.; Toubia, J.; Conn, V.M.; Salmanidis, M.; Phillips, C.A.; Roslan, S.; Schreiber, A.W.; Gregory, P.A.; Goodall, G.J. The RNA binding protein quaking regulates formation of circRNAs. *Cell* **2015**, *160*, 1125–1134. [CrossRef] [PubMed]
18. Liang, H.F.; Zhang, X.Z.; Liu, B.G.; Jia, G.T.; Li, W.L. Circular RNA circ-ABCB10 promotes breast cancer proliferation and progression through sponging miR-1271. *Am. J. Cancer Res.* **2017**, *7*, 1566–1576. [PubMed]
19. Fang, L.; Du, W.W.; Lyu, J.; Dong, J.; Zhang, C.; Yang, W.; He, A.; Kwok, Y.S.S.; Ma, J.; Wu, N.; et al. Enhanced breast cancer progression by mutant p53 is inhibited by the circular RNA circ-Ccnb1. *Cell Death Differ.* **2018**, *25*, 2195–2208. [CrossRef] [PubMed]
20. Han, Y.N.; Xia, S.Q.; Zhang, Y.Y.; Zheng, J.H.; Li, W. Circular RNAs: A novel type of biomarker and genetic tools in cancer. *Oncotarget* **2017**, *8*, 64551–64563. [CrossRef] [PubMed]
21. Zhang, H.D.; Jiang, L.H.; Sun, D.W.; Hou, J.C.; Ji, Z.L. CircRNA: A novel type of biomarker for cancer. *Breast Cancer* **2018**, *25*, 1–7. [CrossRef] [PubMed]
22. Meng, S.; Zhou, H.; Feng, Z.; Xu, Z.; Tang, Y.; Li, P.; Wu, M. CircRNA: Functions and properties of a novel potential biomarker for cancer. *Mol. Cancer* **2017**, *16*, 94. [CrossRef] [PubMed]
23. Zhang, Z.; Yang, T.; Xiao, J. Circular RNAs: Promising Biomarkers for Human Diseases. *EBioMedicine* **2018**, *34*, 267–274. [CrossRef]
24. Lu, L.; Sun, J.; Shi, P.; Kong, W.; Xu, K.; He, B.; Zhang, S.; Wang, J. Identification of circular RNAs as a promising new class of diagnostic biomarkers for human breast cancer. *Oncotarget* **2017**, *8*, 44096–44107.
25. Chen, B.; Wei, W.; Huang, X.; Xie, X.; Kong, Y.; Dai, D.; Yang, L.; Wang, J.; Tang, H.; Xie, X. circEPSTI1 as a Prognostic Marker and Mediator of Triple-Negative Breast Cancer Progression. *Theranostics* **2018**, *8*, 4003–4015. [CrossRef] [PubMed]
26. Liu, K.; Beck, D.; Thoms, J.A.I.; Liu, L.; Zhao, W.; Pimanda, J.E.; Zhou, X. Annotating function to differentially expressed LincRNAs in myelodysplastic syndrome using a network-based method. *Bioinformatics* **2017**, *33*, 2622–2630. [CrossRef] [PubMed]
27. Yang, Z.; Michailidis, G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* **2016**, *32*, 1–8. [CrossRef]
28. Abdellatif, M. Differential expression of microRNAs in different disease states. *Circ. Res.* **2012**, *110*, 638–650. [CrossRef]
29. Greene, J.; Baird, A.M.; Brady, L.; Lim, M.; Gray, S.G.; McDermott, R.; Finn, S.P. Circular RNAs: Biogenesis, Function and Role in Human Diseases. *Front. Mol. Biosci.* **2017**, *4*, 38. [CrossRef]
30. Gao, D.; Zhang, X.; Liu, B.; Meng, D.; Fang, K.; Guo, Z.; Li, L. Screening circular RNA related to chemotherapeutic resistance in breast cancer. *Epigenomics* **2017**, *9*, 1175–1188. [CrossRef]
31. Mistry, M.; Pavlidis, P. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinform.* **2008**, *9*, 327. [CrossRef] [PubMed]
32. Azad, A.; Pavlopoulos, G.A.; Ouzounis, C.A.; Kyrpides, N.C.; Buluc, A. HipMCL: A high-performance parallel implementation of the Markov clustering algorithm for large-scale networks. *Nucleic Acids Res.* **2018**, *46*, e33. [CrossRef] [PubMed]
33. Krejci, A.; Hupp, T.R.; Lexa, M.; Vojtesek, B.; Muller, P. Hammock: A hidden Markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets. *Bioinformatics* **2016**, *32*, 9–16. [CrossRef] [PubMed]

34. Duffy, M.J.; Walsh, S.; McDermott, E.W.; Crown, J. Biomarkers in Breast Cancer: Where Are We and Where Are We Going? *Adv. Clin. Chem.* **2015**, *71*, 1–23. [PubMed]
35. Banin Hirata, B.K.; Oda, J.M.; Losi Guembarovski, R.; Ariza, C.B.; de Oliveira, C.E.; Watanabe, M.A. Molecular markers for breast cancer: prediction on tumor behavior. *Dis. Markers* **2014**, *2014*, 513158. [CrossRef] [PubMed]
36. Cole, K.D.; He, H.J.; Wang, L. Breast cancer biomarker measurements and standards. *Proteomics. Clin. Appl.* **2013**, *7*, 17–29. [CrossRef] [PubMed]
37. Chen, X.; Yin, J.; Qu, J.; Huang, L. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput. Biol.* **2018**, *14*, e1006418. [CrossRef]
38. Chen, X.; Huang, L. LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. *PLoS Comput. Biol.* **2017**, *13*, e1005912. [CrossRef]
39. Chen, X.; Xie, D.; Wang, L.; Zhao, Q.; You, Z.H.; Liu, H. BNPMDA: Bipartite Network Projection for MiRNA-Disease Association prediction. *Bioinformatics* **2018**, *34*, 3178–3186. [CrossRef]
40. Chen, X.; Huang, L.; Xie, D.; Zhao, Q. EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction. *Cell Death Dis.* **2018**, *9*, 3. [CrossRef]
41. Li, J.H.; Liu, S.; Zhou, H.; Qu, L.H.; Yang, J.H. starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **2014**, *42*, D92–D97. [CrossRef] [PubMed]
42. Chou, C.H.; Shrestha, S.; Yang, C.D.; Chang, N.W.; Lin, Y.L.; Liao, K.W.; Huang, W.C.; Sun, T.H.; Tu, S.J.; Lee, W.H.; et al. miRTarBase update 2018: A resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* **2018**, *46*, D296–D302. [CrossRef] [PubMed]
43. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdottir, H.; Tamayo, P.; Mesirov, J.P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011**, *27*, 1739–1740. [CrossRef] [PubMed]
44. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. *Genome Res.* **2002**, *12*, 996–1006. [CrossRef] [PubMed]
45. Kim, D.; Pertea, G.; Trapnell, C.; Pimentel, H.; Kelley, R.; Salzberg, S.L. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **2013**, *14*, R36. [CrossRef]
46. Song, X.; Zhang, N.; Han, P.; Moon, B.S.; Lai, R.K.; Wang, K.; Lu, W. Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Res.* **2016**, *44*, e87. [CrossRef]
47. Trapnell, C.; Roberts, A.; Goff, L.; Pertea, G.; Kim, D.; Kelley, D.R.; Pimentel, H.; Salzberg, S.L.; Rinn, J.L.; Pachter, L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **2012**, *7*, 562–578. [CrossRef]
48. Wang, L.; Wang, Y.; Hu, Q.; Li, S. Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks. *CPT: Pharmacomet. Syst. Pharmacol.* **2014**, *3*, e146. [CrossRef]