biology

MDPI

Article

Identification of Colon Cancer-Related RNAs Based on Heterogeneous Networks and Random Walk

Bolin Chen [1], Teng Wang [1], Jinlei Zhang [1], Shengli Zhang [2] and Xuequn Shang [1,*]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China; blchen@nwpu.edu.cn (B.C.); tengwang@mail.nwpu.edu.cn (T.W.); jinleizhang@mail.nwpu.edu.cn (J.Z.)
[2] School of Information Technology, Minzu Normal University of Xingyi, Xingyi 562400, China; zhangshengli@xynun.edu.cn
* Correspondence: shang@nwpu.edu.cn

Simple Summary: Colon cancer is a complex disease with high incidence rates and mortality worldwide. Although some medical methods have been used for screening, prevention and treatment, its molecular mechanism is still unclear. Among all dysfunctional factors, the change of mutual regulation relationship between RNAs is an important factor affecting the development of cancer. Therefore, the purpose of this study is to find RNAs related to colon cancer that have not been verified. We used differential expression analysis to screen mRNAs, miRNAs and lncRNAs and further constructed a heterogeneous interaction network among these three kinds of RNAs. The network propagation algorithm RW-DIR was then developed to mine the biological information contained in the network and to identify RNAs closely related to colon cancer. The research results have provided some theoretical support for disease research and provide a basis for narrowing the research scope of medical experiments.

Abstract: Colon cancer is considered as a complex disease that consists of metastatic seeding in early stages. Such disease is not simply caused by the action of a single RNA, but is associated with disorders of many kinds of RNAs and their regulation relationships. Hence, it is of great significance to study the complex regulatory roles among mRNAs, miRNAs and lncRNAs for further understanding the pathogenic mechanism of colon cancer. In this study, we constructed a heterogeneous network consisting of differentially expressed mRNAs, miRNAs and lncRNAs. This contains three kinds of vertices and six types of edges. All RNAs were re-divided into three categories, which were "related", "irrelevant" and "unlabeled". They were processed by dynamic excitation restart random walk (RW-DIR) for identifying colon cancer-related RNAs. Ten RNAs were finally obtained related to colon cancer, which were hsa-miR-2682-5p, hsa-miR-1277-3p, ANGPTL1, SLC22A18AS, FENDRR, PHLPP2, hsa-miR-302a-5p, APCDD1, MEX3A and hsa-miR-509-3-5p. Numerical experiments have indicated that the proposed network construction framework and the following RW-DIR algorithm are effective for identifying colon cancer-related RNAs, and this kind of analysis framework can also be easily extended to other diseases, effectively narrowing the scope of biological experimental research.

Keywords: colon cancer; heterogeneous network; random walk; differential expression analysis

Citation: Chen, B.; Wang, T.; Zhang, J.; Zhang, S.; Shang, X. Identification of Colon Cancer-Related RNAs Based on Heterogeneous Networks and Random Walk. *Biology* **2022**, *11*, 1003. https://doi.org/10.3390/biology11071003

Academic Editor: Armando Varela-Ramirez

Received: 10 May 2022
Accepted: 28 June 2022
Published: 2 July 2022

1. Introduction

Due to the continuous improvement of medical standards, people's life expectancy has increased. Currently, people's diet has changed greatly and not only leads to higher incidences of cancer but also to younger individuals with higher incidence. Colon cancer is a common digestive tract malignant tumor occurring in the colon and about a tenth of all cancer cases, thus making it among the top three cancers in terms of incidence as well as mortality [1]. Metastatic seeding of colon cancer often occurs early when carcinoma is clinically undetectable and occurs years before diagnosis and surgery [2]. Although there are

many effective screening means, a further understanding of its occurrence mechanism will promote the further development of innovative screening methods, prognostic indicators and treatment. However, the molecular mechanism of colon cancer formation is still not completely and clearly elucidated.

There are many discussions about the relationship among mRNAs, miRNAs, lncRNAs and diseases, because more and more studies show that these RNAs play key roles in many important biological processes and diseases. The association of mRNAs with cancers has been widely studied, and evidence has been accumulated, with the exception of the relationship of mRNAs and other non-coding RNAs. The microRNAs (miRNAs) are a class of non-coding small RNA molecules encoded by endogenous genes with about 22 nucleotides in length. In animals and plants, it is mainly involved in the regulation of post-transcriptional gene expression [3]. Benefiting from the regulatory function of miRNAs, there are many studies using miRNAs in building networks for identifying disease-related miRNAs, such as the BNPMDA algorithm [4] and NTSMDA algorithm [5]. Long non-coding RNAs (lncRNAs) are defined as RNAs that are longer than 200 nucleotides and that are not translated into functional proteins. It has been found that lncRNAs are closely related to cell cycle, such as differentiation, development, reproduction, aging and many human diseases [6,7]. With the increasing understanding and attention to lncRNAs, the use of network modelling to predict their relationship with diseases has also increased in recent years, such as the GANLDA algorithm [8] and the BPLLDA algorithm [9].

At present, increasing attention has been paid to the data mining algorithms of graphs. Among them, random walk is a very classic algorithm for mining graph structures, which has widely been used. Random walk(RW) models have also been applied in various domains, such as locomotion and the foraging of animals, the dynamics of neuronal firing and decision-making in the brain, descriptions of financial markets, evolution of research interests ranking systems, dimension reduction and feature extraction from high-dimensional data and even sports statistics. RW theory can also help predict the arrival times of diseases spreading in networks [10].

Many current methods for analyzing RNA interaction networks ignore the heterogeneous characteristics of the network. They either only use the interactions between two types of RNAs, which ignore the interactions within the same type of RNA [11], or do not treat different types of RNAs (nodes) differently, which render the obtained results in a state of non-equilibrium. For instance, the label reasoning models often need to calculate entropy, but they cannot conduct the global random at the same time [12]. To overcome of this, this study proposes to combine the idea of maximum entropy with a tag inference by using random walk to identify key RNAs related to colon cancer by considering the overall property of mRNAs, miRNAs, and lncRNAs in the heterogeneous network. The results of different types of RNAs were balanced.

To be more specific, this study first proposed to construct a colon cancer-specific RNA interaction heterogeneous network. The traditional random walk algorithm was then improved to find and analyze the RNAs related to colon cancer. The details are provided as follows. Firstly, we constructed a heterogeneous biological network for colon cancer, in which mRNAs, miRNAs and lncRNAs are the vertices of the network, and the interactions between every two types of RNAs and within each RNA served as the edges. There were three types of vertices and six kinds of edges. Then, we designed a random walk transfer matrix for heterogeneous networks, and labelled all vertices as three categories, namely "related", "irrelevant" and "unlabeled", according to whether the vertices are related to colon cancer. Applying the idea of traditional random walk, different measures were taken for different category vertices encountered in the process of walking so as to achieve the purpose of classifying the "unlabeled" RNAs. Figure 1 indicates the processes of identifying colon cancer-related RNAs in this study, where part A is the process of building the heterogeneous network, and part B is the process of RW-DIR.

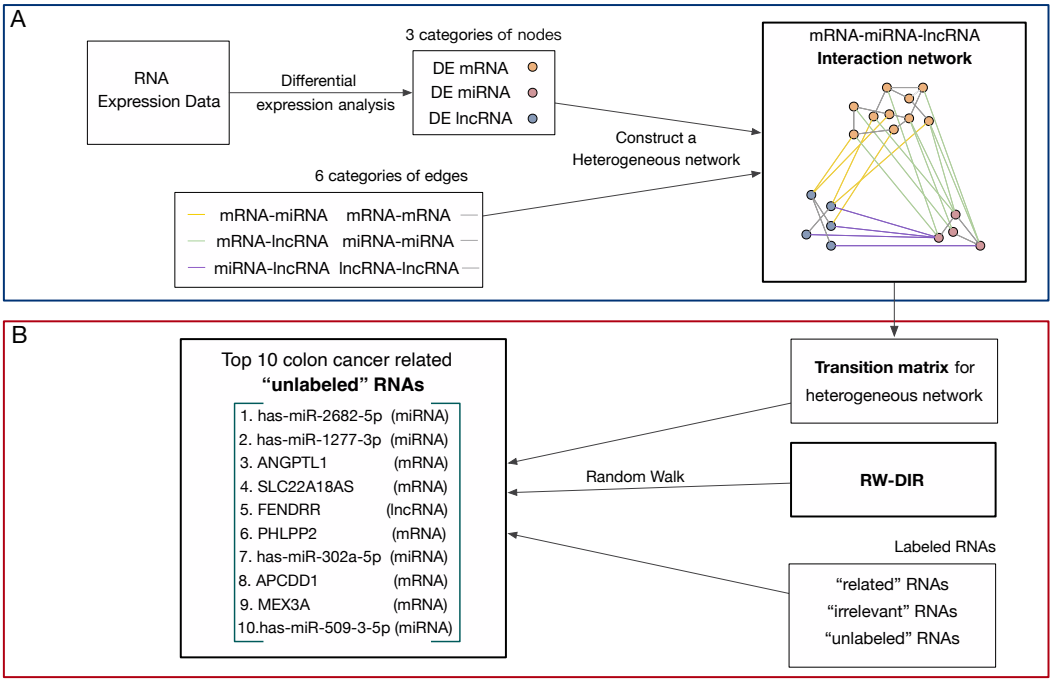

Figure 1. This is the workflow of this study. (**A**) Construct the mRNA-miRNA-lncRNA interaction network. (**B**) RW-DIR algorithm was applied to obtain the results of colon cancer-related RNAs.

2. Materials and Methods

2.1. RNA Expression Data

In this study, mRNA expression data, miRNA expression data, lncRNA expression data and clinical data were collected from the open-access dataset of The Cancer Genome Atlas (TCGA) database [13]. The project ID was "TCGA-COAD".

For mRNA and lncRNA expression data, the experimental strategy we downloaded was "RNA-Seq", the data type was "Gene Expression Quantification", and the data category was "transcriptome profiling". The data from the Ensemble database [14] annotated the type of all RNAs in the gene expression data, and it was downloaded from TCGA. In this study, we selected "protein-coding gene" and "lncRNA" as mRNA and lncRNA for subsequent analysis.

For miRNA, the data type we downloaded was "Isoform Expression Quantification", the workflow type was "BCGSC miRNA Profiling", and the data category was "Transcriptome Profiling".

The clinical data of colon cancer were also obtained from TCGA. The original clinical data contained a variety of clinical information items of the samples, and only the information about sample ID and cancer stage was selected. The sample ID was used to map the RNA expression data of the particular sample, and the information of the cancer stage was used to distinguish whether the samples were cancerous or paracancerous tissue; the latter will be used as normal samples.

2.2. The Relationship of RNAs

The connection in this study could divide into two categories. One was the connection between different kinds of RNA, and the other was the connection within the same type of RNA. The relationship and data source databases are shown in Table 1. The relationships between "miRNA-miRNA" and "lncRNA-lncRNA" are obtained by the Deepwalk algorithm [15], and their respective associations are related to their target genes.

Table 1. RNA association and interaction database.

Types of RNA Associations	Database
mRNA-miRNA	multiMiR [16]
mRNA-lncRNA	starbase V3.0 [17]
miRNA-lncRNA	LncBase V2.0 [18]
mRNA-mRNA	STRING [19]

2.3. Classification of RNA

The mRNAs, miRNAs, and lncRNAs that are related to colon cancer and are verified by experiments were obtained from the databases shown in Table 2. We selected known colon cancer-related RNAs as RNAs with "related" labels and randomly selected an equal amount of RNAs that related to other diseases and excluded colon cancer-related RNAs as RNAs with an "irrelevant" label. The remaining vertices were marked with the "unlabeled" label.

Table 2. Colon cancer-related RNAs and database.

Types of RNA	Database
mRNA	Comparative Toxicogenomics Database [20]
miRNA	miR2disease [21]
lncRNA	LncRNADisease [22]

2.4. Data Preprocessing

Since deeper sequencing always produces more sequence fragments, in differential expression analysis, the row counts were rarely used directly. In practice, the counts are usually normalized to eliminate sequencing differences due to sequencing depth. The log-CPM normalization method was used in this study.

The R package edgeR [23] was used for data preprocessing. In all datasets, there would be a mixture of expressed genes and non-expressed genes. Reducing these noises would not only significantly improve the accuracy of statistical inferences from RNA-seq but also allow mathematical models in the data to be more accurately estimated and reduce the amount of RNA analyzed downstream; thus, this study used the "filterByExpr" function in edgeR package to filter RNAs with low expression counts [24]. For each group of data, the TMM (Trimmed Mean of M-values) [25] algorithm was also considered to ensure that each sample has a similar distribution of expression data. After the data preprocessing, the number of three RNAs and the number of samples in their respective datasets are shown in Table 3.

Table 3. Edge Information in Heterogeneous Networks.

Type of RNA	Number of RNA	Number of Normal Samples	Number of Tumor Samples
mRNA	116591	41	443
miRNA	302	8	444
lncRNA	1526	41	443

2.5. Differential Expression Analysis

Differential expression analysis [23] refers to obtaining normalized read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. There were two main parameters for using this method to screen differentially expressed RNAs: one is $|logFC|$ and the other one is the p-value.

In this study, the exact test method that is based on the negative binomial distribution in R package. EdgeR was used to identify differentially expressed RNAs. The threshold selection of the three differentially expressed RNAs was different. Specifically, the mRNA that had p-value < 0.05 and $|logFC| \geq 2$ could be chosen as the differentially expressed

(DE) mRNA; the miRNA that had p-value < 0.05 and $|logFC| \geq 2$ could be chosen as the DE miRNA; the lncRNA that had p-value < 0.05 and $|logFC| \geq 1$ could be chosen as the DE lncRNA. Finally, 1372 DE mRNAs, 175 DE miRNAs and 137 DE lncRNAs were obtained in this study.

2.6. Construct Heterogeneous Network

This study was based on the exploration of the complex regulatory relationship among mRNAs, miRNAs and lncRNAs. Therefore, a heterogeneous network was first constructed to express the relationship among the three in the form of a network for subsequent data mining. The heterogeneous network was shown in Figure 2. In this study, this network was defined as G(V,E), where V represents all vertices and E represents all edges. The adjacency matrix H can be obtained by assigning a value of 1 if there was an edge between two vertices, and it is 0 otherwise.

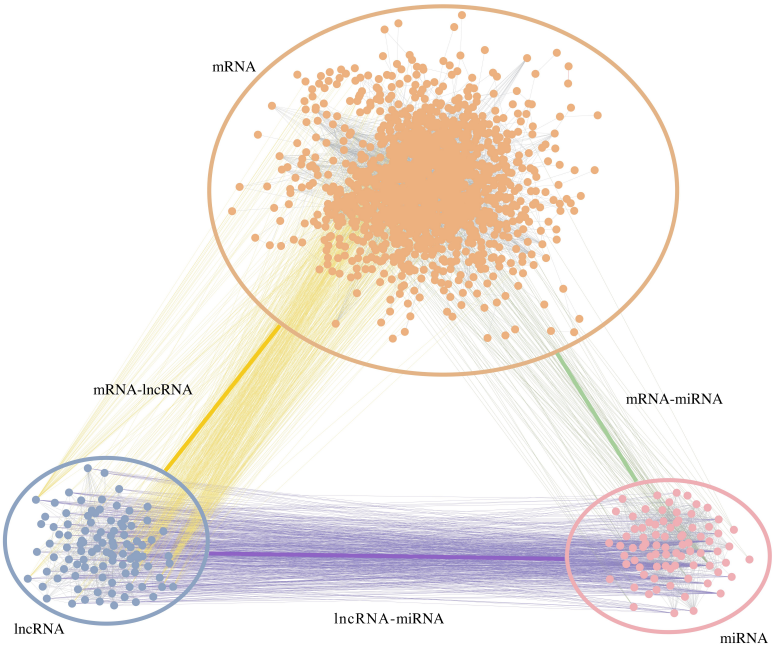

Figure 2. The heterogeneous network of RNA interactions. Orange vertices are mRNAs, pink vertices are miRNAs, and blue vertices are lncRNAs. The grey edges represent the connection within the same kind of RNAs, the yellow edges represent the relationship between mRNA and lncRNA, the green edges represent the relationship between mRNA and miRNA, and the purple edges represent the relationship between miRNA and lncRNA.

2.7. RW-DIR

Similarly to traditional random walk, RW-DIR required a transition matrix [26] for the subsequent walk in the network, but the transition matrix in this study was designed based on heterogeneous networks. A diagram of the calculation method is shown in Figure 3. In the figure, s represents the number of mRNAs, m represents the number of miRNAs, n represents the number of lncRNAs, and h represents the number of all RNAs. Obviously, h equals to the sum of s, m, and n. Three parameters, λ, δ and θ, are used to adjust the transition probability of different types of RNAs. The λ is the transfer parameter between mRNA and miRNA, the δ is the transfer parameter between mRNA and lncRNA, and the θ is the transfer parameter between miRNA and lncRNA. Specifically, we have the following.

$$\lambda = |mRNA - miRNA| / (|mRNA - mRNA| + |miRNA - miRNA|) \quad (1)$$

$$\delta = |mRNA - lncRNA| / (|mRNA - mRNA| + |lncRNA - lncRNA|) \quad (2)$$

$$\theta = |miRNA - lncRNA| / (|miRNA - miRNA| + |lncRNA - lncRNA|) \quad (3)$$

The method to calculate transition matrix W is written as Equation (4):

$$W(i,j) = X \cdot \frac{H(i,j)}{\sum_{k=a}^{b} H(i,k)}, \tag{4}$$

where $X = \{x \mid x \in \{A, (1-A-B)\}, A \neq B\}\}$, $\{A, B\} \in \{\lambda, \delta, \theta\}$, $\{i,j\} \in \{\{1,\ldots,s\}, \{s+1,\ldots,s+m\}, \{s+m+1,\ldots,h\}\}$, $a \in \{1, s+1, s+m+1\}$ and $b \in \{s, s+m, h\}$. The selection of parameter X, a and b could be more intuitive according to the Figure 3. Since H was comprised 9 sub-matrices, the sub-matrices should be calculated separately.

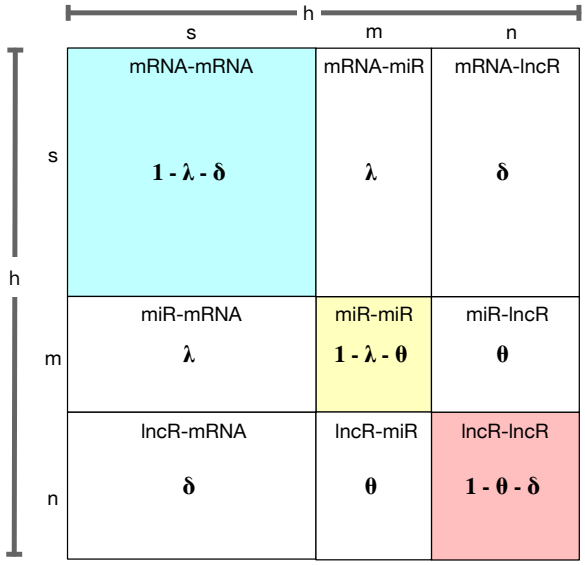

Figure 3. This is a diagram of the transition matrix calculation. Nine sub-matrices form a complete transition matrix. λ is the transfer parameter between mRNA and miRNA, δ is the transfer parameter between mRNA and lncRNA, and θ is the transfer parameter between miRNA and lncRNA. The transfer parameters in each small square correspond to the parameters that could be used to calculate the percentage of summarized weight corresponding to the sub-transition matrix.

Random walk is a discrete-time Markov process where a walker located at vertex i at a certain moment will jump to adjacent vertex j at the next moment with probability $W(i,j)$. This jump is independent to the past. Before exhibiting random walk with dynamic incentive restart (RW-DIR), we introduced the random walk with restart (RWR) method [27] first. The difference between RWR and traditional random walk is that there is a certain probability of returning to the starting point after each step of walking. In RWR, based on the transition matrix W and the hopping process $P(t+1) = \alpha W P(t) + (1-\alpha)P(0)$, where $P(t)$ is a vector that represents the probability of walker at all vertex at time t, and $P(t)$ will converge, which means RWR will reach a stationary distribution.

Next, we would introduce RW-DIR. The vertices in the studied heterogeneous network have certain prior knowledge about colon cancer; thus, we considered that in the process of walking. Different measures should be taken for vertices with different labels such that the reliable colon cancer-related vertices can be finally obtained. The specific algorithm process is defined in Algorithm 1. In the initial round of the algorithm, we assigned the value of 1 in P_0 to the colon cancer-related RNA with the largest degree, and the rest was 0. In each round of the algorithm, judging which vertices the current round walks to was based on whether there were any changes in $P(t)$ compared with $P(t-1)$. Considering that the random walk process could spread the information of labelled RNAs, we used known knowledge for random walk and information dissemination. We inflated the effect of the "related" vertex, which was gradually added to P_0. It enhanced its global

influence in the process of repeatedly restarting. We shrunk the effect of the "irrelevant" vertex, which included adding a penalty value to the "irrelevant" vertex in each iteration to make it smaller; this would reduce their impact of the global process. For the "unlabeled" vertex, we referred to the idea of maximum entropy by calculating the entropy value of each vertex after each iteration, and the inference result of the vertex with the largest entropy value represented the highest uncertainty. The transition probability matrix W is recalculated according to the entropy value in each round iteration. The larger the entropy value, the larger transition probability matrix value of the vertex. It should be noted that the calculation of entropy needs to be started after walking to the global process; that is, the calculation of the transfer matrix also needs to wait until the process of walking to all vertices. It would help the vertex with the "unlabeled" result in collecting more information for further inferences.

Algorithm 1 Random Walk with Dynamic Incentive Restart (RW-DIR)

Input: transition matrix W, initial P_0, "related" label vertices set V_r, "irrelevant" label vertices set V_i, unlabeled vertex set V_u.
Output: The label weight set $m(v_i)$ of vertex $v_i \in V_u$

repeat
  Comparing $P_{(t)}$ with $P_{(t-1)}$, the vertex whose value has changed is the vertex that the t-th round traveled to, denoted as $V_{neighbor}$;
  $V_{rn} = \{v_i \mid v_i \in V_r \cap V_{neighbor}\}$, $V_{in} = \{v_i \mid v_i \in V_i \cap V_{neighbor}\}$, $V_{un} = \{v_i \mid v_i \in V_u \cap V_{neighbor}\}$.
  for $v_i \in V_{rn}$ **do**
    $P_0(v_i) \leftarrow$ initial value 0.1 ; `// v_i is a vertex with "related" label`
    $P_{t+1} \leftarrow \alpha W P_t$ + (1-α)P_0 ;
  end
  for $v_i \in V_{in}$ **do**
    $P'_{t+1} \leftarrow \alpha W P_t$ + (1-α)P_0 ; `// v_i is a vertex with "irrelevant" label`
    $P'_{t+1}[\text{index}(v_i)] \leftarrow P'_{t+1}[\text{index}(v_i)] \times \frac{1}{degree(v_i)}$;
    $P_{t+1} \leftarrow P'_{t+1}$;
  end
  for $v_i \in V_{un}$ **do**
    calculate $E(v_i)$; `// v_i is an unlabeled node`
    for $v_i \in V_{un}$ **do**
      update W ; `// update transition matrix by entropy`
      calculate $m(v_i, v_j)$
    end
    calculate $m(v_i)$; `// score two types of labels for vertex` v_i
  end
until $P_{(t)}$ *Converge*;

The entropy value $E(v_j)$ of the vertex v_j can be calculated by Equation (5), where $m_k(v_i)$ represents the possibility that vertex v_i belonged to label k. In this study, we set $k = 1$ represent the "related" vertex, and $k = 2$ represent the "irrelevant" vertex. When the $E(v_i)$ value of vertex v_i is calculated for the first time (the tth round walk), where v_i was the "unlabeled" vertex, we obtain initial $m_1(v_i) = \frac{h-ascending\ rank\ of\ P_t[index(v_i)]}{h}$, and $m_2(v_i) = 1-m_1(v_i)$. For "related" vertex v_r, we set $m_1(v_r) = 1$ and $m_2(v_r) = 0$. For "irrelevant" vertex v_l, we set $m_2(v_l) = 1$ and $m_1(v_l) = 0$.

$$E(v_i) = -\sum_{k=1}^{2} m_k(v_i) log_2 m_k(v_i), \tag{5}$$

When the random walk process had covered the entire network (in the tth round walk), transition matrix W needed to be updated according to the entropy value of E by Equation (6), where v_j represented all neighbor vertices of v_i, including v_i.

$$W(v_i, v_j) = \frac{E(v_i)}{\sum_{v_j \in N^+(v_i)} E(v_j)}, \tag{6}$$

After that, it was necessary to update and calculate the probability that each vertex belonged to each label. As shown in Equation (7), $m_k(v_i, v_j)$ represented the probability that label k propagated from vertex v_i to vertex v_j, which could be used to further calculate the probability of vertex v_i with label k, i.e., $m_k(v_i)$, in Equation (8).

$$m_k(v_i, v_j) = m_k(v_i) \times W_{v_i, v_j}, \tag{7}$$

$$m_k(v_i) = \frac{\sum_{v_j \in N^-(v_i)} m_k(v_i, v_j)}{\sum_{k=1}^{2} \sum_{v_j \in N^-(v_i)} m_k(v_i, v_j)}, \tag{8}$$

In summary, at the beginning of the algorithm, we started the RW-DIR algorithm with one known colon cancer-related vertex. In the subsequent iterative process of the walker, compared to the previous round, we classified the type of vertices that had been "walked". If it was a "related" vertex that has never been reached before, it would be added to P0 and assigned a value of 0.1. If it was an "irrelevant" label, add a penalty value to P_t; that is, we divide it by its degree. If the walker "walked" to vertices with no label, its entropy value would be calculated, and the transition matrix should be updated by the entropy value. Finally, the algorithm would stop after P_t convergence. The m_1 sorting result of the unlabeled vertex will be used as an indicator for screening colon cancer-related RNAs.

3. Results

3.1. The Heterogeneous Network

The relationships of "miRNA-miRNA" and "lncRNA-lncRNA" were obtained by constructing the interaction network of their respective target genes and then applying the Deepwalk algorithm. After applied the Deepwalk algorithm in miRNA-target genes interaction networks, we obtained 273,431 edges and 740 miRNA vertices. Moreover, the miRNA-target gene database was miRtarbase. The DE miRNAs obtained in this study were screened out; finally, the miRNA functional similarity network had 81 vertices and 389 edges, which was shown in the pink circle part in Figure 2. We also applied the Deepwalk algorithm in the lncRNA-target genes interaction network. The lncRNA-target gene database was starBase, and the relationship of "lncRNA-target" was screened with a threshold greater than 0.5; we obtained 63,546 edges and 357 lncRNA vertices. In this study, we set the edge weight of the threshold value to be greater than 0.9 in the network and selected DE lncRNAs. Finally, we obtained 65 lncRNA vertices and 604 edges, which are shown in the blue circle in Figure 2.

The RNA interaction heterogeneous network constructed in this study had 1521 vertices and 9651 edges, as shown in Figure 2. Among them, the number of mRNA vertices was 1340, the number of miRNA vertices was 80, and the number of lncRNA vertices was 101. The edge's information is shown in Table 4.

Table 4. Edge Information in Heterogeneous Networks.

Type of Edge	Number of Node	Number of Edge
mRNA-mRNA	1300 (mRNA)	7408
miRNA- miRNA	81 (miRNA)	389
lncRNA-lncRNA	389 (lncRNA)	604
mRNA-miRNA	56 (mRNA) - 326(miRNA)	569
mRNA-lncRNA	33 (mRNA) - 70(lncRNA)	94
miRNA-lncRNA	99 (miRNA) - 57(lncRNA)	587

3.2. The Result of RW-DIR

We selected the top 10 RNAs in descending order with respect to the m_1 value as candidate colon cancer-related RNAs, where m_1 represented the probability that the vertex was classified as colon cancer-related RNAs. These RNAs were hsa-miR-2682-5p, hsa-miR-1277-3p, ANGPTL1, SLC22A18AS, FENDRR, PHLPP2, hsa-miR-302a-5p, APCDD1, MEX3A and hsa-miR-509-3-5p. Among them, FENDRR is lncRNA; hsa-miR-2682-5p, hsa-miR-1277-3p, hsa-miR-302a-5p and hsa-miR-509-3-5p are miRNAs; ANGPTL1, SLC22A18AS, PHLPP2, APCDD1 and MEX3A are mRNAs.

3.2.1. Colon Cancer Related mRNAs

ANGPTL is a family of proteins similar to angiopoietins. They affect angiogenesis, inflammation, metabolic disturbances, hematopoiesis, and cancer development. Studies have shown that ANGPTL1 can act as an anti-angiogenic factor and a tumor suppressor [28]. ANGPTL1 has been reported to suppress migration and invasion in lung, breast and colorectal cancer, acting as a novel tumor suppressor candidate [29]. For SLC22A18AS, high expression levels are significantly associated with worsening disease progression. In addition, low levels of SLC22A18AS are also correlated with better overall survival for lung adenocarcinoma patients [30]. For PHLPP2, maintaining balanced PHLPP2 expression levels is critical for disease prevention, as changes in steady-state levels of PHLPP2 are associated with many diseases, including diabetes, hepatic steatosis, and cancer. Recently, many studies have shown that the expression of PHLPP2 is universally absent in a variety of cancers and plays a key role in a wide range of biological processes, including cancer cell proliferation, metastasis, autophagy and apoptosis [31]. For APCDD1, there is a study that suggested that APCDD1 regulated breast cancer progression by targeting canonical WNT signaling and modulating breast cancer cell invasion [32]. For MEX3A, it may promote glioma development by regulating cell proliferation, cell apoptosis, cell cycle and cell migration, and MEX3A has been identified as a potential tumor promoter in glioma development and therapeutic target in glioma treatment [33]. Taken together, these mRNAs are all related to the survival process of cells and play important roles in some cancers.

3.2.2. Colon Cancer Related miRNAs

Hsa-miR-2682-5p and hsa-miR-1277-3p are the top two results. The neighbor vertices of hsa-miR-2682-5p and hsa-miR-1277-3p in the heterogeneous network, which were constructed in this study, were all "related" vertices. For hsa-miR-2682-5p, the study has suggested that miR-2682-5p promotes cell proliferation and migration in oral squamous cell carcinoma, and its target mRNA and lncRNA in this study were all known colon cancer-related RNAs [34]. For hsa-miR-302a-5p, some studies showed that the miR-302 family, which includes miR-302b, miR-302c, and miR-302d, exerts antitumor effects in several cancers, such as endometrial carcinoma, glioma and breast cancer. MiR-302a has been shown to function as a tumor suppressor by regulating diverse cellular functions [35]. For example, HMGA2 has been implicated as a driver of tumor metastasis; however, hsa-miR-302a-5p is the powerful post-transcriptional regulator of HMGA2 [36]. For hsa-miR-509-3-5p, the decreased expression of miR-509-3-5P promoted the colony, migration and invasion abili-

ties of gastric cancer cells in vitro as well as tumorigenesis and lymph vertex metastasis in vivo [37]. In summary, the regulation of miRNAs on cancer is generally reflected in the regulation of their target genes. The four candidate colon cancer-related miRNAs obtained in this study were basically closely related to the occurrence of some common cancers, and their relationship with colon cancer deserves further study.

3.2.3. Colon Cancer Related lncRNAs

Table 5 has summarized the top 10 RNAs and related diseases. We can see from the table that FENDRR is the only lncRNA among the top 10 candidate colon cancer-related RNAs. Studies have shown that the low expression of the FENDRR occurs in gastric cancer and is associated with poor prognosis; thus, FENDRR plays an important role in the progression and metastasis of gastric cancer [38]. FENDRR is expressed in a variety of cancers and is significantly associated with different clinical features. Furthermore, FENDRR has shown potential as a biomarker for cancer diagnosis, prognosis and treatment. Therefore, FENDRR is a potential candidate lncRNA for studying colon cancer-related RNAs [39].

Table 5. Top 10 RNAs and Related Diseases.

Top 10 RNAs	Number of RNA	Related Diseases
hsa-miR-2682-5p	miRNA	Oral squamous cell carcinoma
hsa-miR-1277-3p	miRNA	/
ANGPTL1	mRNA	Lung cancer, breast cancer, colorectal cancer
SLC22A18AS	mRNA	Lung adenocarcinoma
FENDRR	lncRNA	Gastric cancer, lung cancer, hepatocellular carcinoma (HCC), gastric cancer
PHLPP2	mRNA	Diabetes, hepatic steatosis, and cancer
hsa-miR-302a-5p	miRNA	Endometrial carcinoma, glioma and breast cancer
APCDD1	mRNA	Breast cancer
MEX3A	mRNA	Glioma
hsa-miR-509-3-5p	miRNA	Gastric cancer

3.3. Performance of RW-DIR

In this study, the ROC curve was used to visually display the classification performance of the algorithm, and the AUC value was used to measure the classification ability of the algorithm [40]. In this study, we used LOO-CV(Leave-One-Out Cross-Validation) to test the performance of RW-DIR. Specifically, we placed one "related" vertex into "unlabeled" vertices each time and tested its m_1 value. Equal amounts of RNAs were randomly selected from the candidate RNAs related to other cancers, with the exception of colon cancer as negative samples, and checked their m_1 value. After the results were obtained, the ROC diagram was made, as shown in Figure 4, in which the color of the curve is red, and AUC is 0.8212.

We also evaluated the performance of the traditional restart random walk (TRWR) algorithm [27] on heterogeneous networks with a traditional transition matrix and the performance of the RW-DIR algorithm without using entropy. In detail, the process of using TRWR algorithm on the RNA interaction heterogeneous network was as follows: first, in order to obtained the relationship between "unlabeled" RNAs and "related" RNAs, the "related" vertices in P0 all were assigned a value of 0.1, and the transition matrix was calculated according to the degree of the vertex that needed to meet the standardization rules of the transition matrix; finally, walker can walk according to Equation (6) until

convergence. The leave-one-out method was used during testing and the ROC diagram is shown in Figure 4, and the color of its curve is blue.

$$P_{t+1} = \alpha W P_t + (1 - \alpha) P_0 \, . \tag{9}$$

The yellow curve in Figure 4 referred to the method that lacked the part of entropy in RW-DIR (lack of processing for "unlabeled" nodes). Specifically, the idea of not considering maximum entropy was to only consider RNAs that are known to be associated with colon cancer, and we only expand or shrink these vertices at this time, and perform nothing else for the "unlabeled" vertex. We also performed the leave-one-out to test the performance of the method.

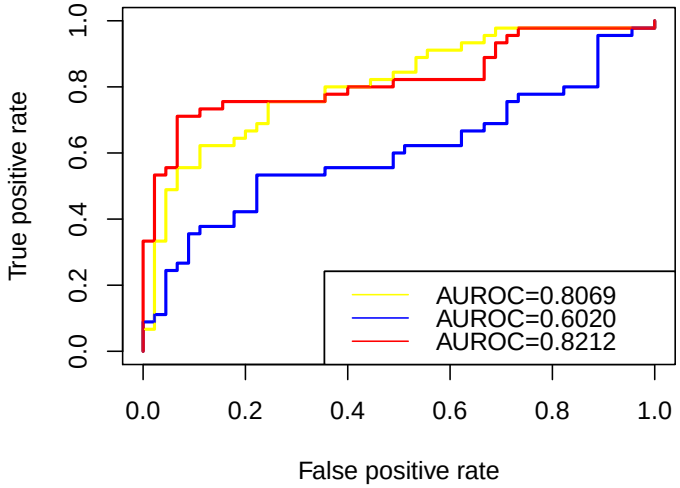

Figure 4. The prediction performance for prioritizing colon cancer causal RNAs. The ROC curves illustrating the performance in distinguishing "related" label RNAs from "irrelevant" label RNAs. Red curve represented RW-DIR; the blue curve represented TRWR, which was directly used on the basis of the heterogeneous network that is constructed in this study; the yellow curve represented RW-DIR without entropy.

It can be seen from the results that designing a transfer matrix for heterogeneous networks is very necessary for the network's propagation algorithm, which can try to avoid the deviation of the results caused by the difference in the number of different types of vertices or edges. For the three types of RNAs, the number of mRNAs was far greater than that of miRNAs and lncRNAs, and the number of edges in different subnetworks was also very different. On the other hand, it was reasonable to use the idea of maximum entropy for final RNA screening. If only the walk probability (P_t) ranking was used as the final result, the aggregation or dispersion of vertices will be ignored, and the results were average, while in this study, the m_1 value was used as the parameter for the final comparison, which was obtained by aggregating the information of the global vertices and not the score that transferred from the rank of P_t. Comparing the AUC values of the three methods, it could be seen that RW-DIR performed the best, and RWR performed poorly.

4. Discussion

Colon cancer has become the third most common cancer in the world. In recent years, the characteristics of its younger patient population, urbanization and easy metastasis in the early stage have attracted our attention. At present, its molecular mechanism is still unclear. Cancer-related RNAs are the key to targeted therapy. Therefore, this study is committed to find mRNAs, miRNAs and lncRNAs that are closely related to colon cancer. Based on RNA expression data, we analyzed and mined it at the data level and topology level in order to obtain relevant results and applied them to medical experiments, provide data evidence and narrow the research scope of medical experiments.

In this study, we started with RNA expression data and conducted differential expression analysis to obtain DEmRNA, DEmiRNA and DElncRNA. Combined with the RNA interaction database and the graph-embedding method, the heterogeneous network of mRNA-miRNA-lncRNA interaction was constructed. On this basis, we designed an innovative network propagation and data mining algorithm. The main idea is to treat the vertices with different types and labels differently, and finally, we obtained the relevant RNAs that are most related to colon cancer but not confirmed by research. We obtained the top ten unproven RNAs associated with colon cancer. They are hsa-miR-2682-5p, hsa-miR-1277-3p, ANGPTL1, SLC22A18AS, FENDRR, PHLPP2, hsa-miR-302a-5p, APCDD1, MEX3A and hsa-miR-509-3-5p. Moreover, most of them have a certain inhibitory effect on the development of other types of cancer, and some can even be used as biomarkers.

For miRNAs in the results, there was increasing evidence that indicated that hsa-miR-2682-5p acted as a tumor suppressor in various cancers, such as non-small cell lung cancer (NSCLC) and Pancreatic cancer (PC) [41,42]; hsa-miR-302a-5p also suppresses proliferation and invasion in NSCLC [35], and hsa-miR-509-3-5p can suppress lung cancer by inhibiting the proliferation and migration of lung cancer cells [37]. The first two are regulated by targeting mRNA, while the last one regulates cancer cells through the relationship with lncRNA, which also showed that the competitive and cooperative relationship between different RNAs was close and further strengthens the possibility that the experimental results are likely to be related to colon cancer. For the mRNA results obtained in the study, we found that the mRNA of the top four has been experimentally verified, and when it is highly expressed in other types of cancer, it has positive significance for the development and prognosis of cancer. The top four mRNAs are ANGPTL1, SLC22A18AS, PHLPP2 and APCDD1. They have proved that they could inhibit the proliferation and metastasis of cancer cells in many other cancers, such as lung cancer, breast cancer and colon cancer. The last ranked mRNA, MEX3A [33], is a promoter for glioma and a therapeutic target in the treatment of glioma.

The model constructed in this study needs to be supported by a large number of databases. For some diseases, the amount of data may not be large enough, resulting in inaccurate results in data mining. However, at present, the cancer-related databases are relatively complete, and the information about RNA is relatively perfect. Therefore, most common cancers can find relevant mRNA, miRNA and lncRNA by this method. Moreover, the current RNA interaction network was built on the basis of differentially expressed RNA, and some cancer-related RNAs had not been screened by differential expression analysis. In addition, there are still some problems in using machine learning and other computing methods to identify cancer-related RNAs, such as little data quantity, unbalanced sample data, and difficult modeling. Moreover, it still requires follow-up biological experiments for further verification. Therefore, it is necessary to find a better method to screen the vertices of the interaction network in the future.

The network topology model and global heterogeneous network analysis algorithm proposed in this study provided new inspiration and ideas for finding RNA related to colon cancer and other diseases. Although there were some limitations in the data, they still did not affect the reliability of the final result.

5. Conclusions

Colon cancer is a complex disease with a high incidence rate and high mortality. Although there are certain medical methods for its prevention and treatment, its molecular mechanism has not been clear. The complex regulation between RNAs is an important cause of cancer. Therefore, the purpose of this study is to find RNA related to colon cancer that has not been verified. We used the regulatory relationship between mRNA, miRNA and lncRNA screened by differential expression analysis to construct a heterogeneous network, and then we analyzed its topological characteristics and used the RW-DIR method to find RNAs that are closely related to colon cancer. The results can provide some theoretical support for disease research and provide a basis for medical experiments.

Author Contributions: Conceptualization, B.C. and T.W.; methodology, B.C. and T.W.; software, T.W. and J.Z.; validation, B.C., T.W. and X.S.; formal analysis, T.W.; investigation, B.C.; resources, T.W.; data curation, T.W. and J.Z.; original draft preparation, T.W.; review and editing, B.C. and S.Z.; visualization, T.W.; supervision, B.C. and X.S.; project administration, B.C. and X.S.; funding acquisition, B.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant No. 61972320, the National Key RD Program of China (No. 2021YFA1000402).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data are available upon request from the corresponding author.

Acknowledgments: We would like to thank the reviewers for their valuable comments and suggestions.

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

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2020**, *71*, 209–249. [CrossRef] [PubMed]
2. Hu, Z.; Ding, J.; Ma, Z.; Sun, R.; Seoane, J.A.; Scott Shaffer, J.; Suarez, C.J.; Berghoff, A.S.; Cremolini, C.; Falcone, A.; et al. Quantitative evidence for early metastatic seeding in colorectal cancer. *Nat. Genet.* **2019**, *51*, 1113–1122. [CrossRef]
3. Krol, J.; Loedige, I.; Filipowicz, W. The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.* **2010**, *11*, 597–610. [CrossRef] [PubMed]
4. 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] [PubMed]
5. Sun, D.; Li, A.; Feng, H.; Wang, M. NTSMDA: Prediction of miRNA–disease associations by integrating network topological similarity. *Mol. Biosyst.* **2016**, *12*, 2224–2232. [CrossRef]
6. Serviss, J.T.; Johnsson, P.; Grandér, D. An emerging role for long non-coding RNAs in cancer metastasis. *Front. Genet.* **2014**, *5*, 234. [CrossRef]
7. Wilusz, J.E.; Sunwoo, H.; Spector, D.L. Long noncoding RNAs: Functional surprises from the RNA world. *Genes Dev.* **2009**, *23*, 1494–1504. [CrossRef]
8. Lan, W.; Wu, X.; Chen, Q.; Peng, W.; Wang, J.; Chen, Y.P. GANLDA: Graph attention network for lncRNA-disease associations prediction. *Neurocomputing* **2022**, *469*, 384–393. [CrossRef]
9. Xiao, X.; Zhu, W.; Liao, B.; Xu, J.; Gu, C.; Ji, B.; Yang, J. BPLLDA: Predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network. *Front. Genet.* **2018**, *9*, 411. [CrossRef]
10. Masuda, N.; Porter, M.A.; Lambiotte, R. Random walks and diffusion on networks. *Phys. Rep.* **2017**, *716*, 1–58. [CrossRef]
11. Zhou, R.S.; Zhang, E.X.; Sun, Q.F.; Ye, Z.J.; Liu, J.W.; Zhou, D.H.; Tang, Y. Integrated analysis of lncRNA-miRNA-mRNA ceRNA network in squamous cell carcinoma of tongue. *BMC Cancer* **2019**, *19*, 779. [CrossRef] [PubMed]
12. Pan, J.; Yang, Y.; Hu, Q.; Shi, H. A label inference method based on maximal entropy random walk over graphs. In Proceedings of the Asia-Pacific Web Conference, Suzhou, China, 23–25 September 2016; pp. 506–518.
13. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **2015**, *19*, A68. [CrossRef] [PubMed]
14. Yates, A.D.; Achuthan, P.; Akanni, W.; Allen, J.; Allen, J.; Alvarez-Jarreta, J.; Flicek, P. Ensembl 2020. *Nucleic Acids Res.* **2020**, *48*, D682–D688. [CrossRef]
15. Perozzi, B.; Al-Rfou, R.; Skiena, S. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 24–27 August 2014; pp. 701–710.
16. Ru, Y.; Kechris, K.J.; Tabakoff, B.; Hoffman, P.; Radcliffe, R.A.; Bowler, R.; Theodorescu, D. The multiMiR R package and database: Integration of microRNA–target interactions along with their disease and drug associations. *Nucleic Acids Res.* **2014**, *42*, e133. [CrossRef] [PubMed]
17. 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]
18. Karagkouni, D.; Paraskevopoulou, M.D.; Tastsoglou, S.; Skoufos, G.; Karavangeli, A.; Pierros, V.; Hatzigeorgiou, A.G. DIANA-LncBase v3: Indexing experimentally supported miRNA targets on non-coding transcripts. *Nucleic Acids Res.* **2020**, *48*, D101–D110. [CrossRef] [PubMed]
19. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Mering, C.V. STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [CrossRef]

20. Davis, A.P.; Grondin, C.J.; Johnson, R.J.; Sciaky, D.; King, B.L.; McMorran, R.; Mattingly, C.J. The comparative toxicogenomics database: Update 2017. *Nucleic Acids Res.* **2017**, *45*, D972–D978. [CrossRef]
21. Jiang, Q.; Wang, Y.; Hao, Y.; Juan, L.; Teng, M.; Zhang, X.; Liu, Y. miR2Disease: A manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* **2009**, *37*, D98–D104. [CrossRef]
22. Bao, Z.; Yang, Z.; Huang, Z.; Zhou, Y.; Cui, Q.; Dong, D. LncRNADisease 2.0: An updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* **2019**, *47*, D1034–D1037. [CrossRef]
23. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140. [CrossRef] [PubMed]
24. Law, C.W.; Alhamdoosh, M.; Su, S.; Dong, X.; Ritchie, M.E. Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger. *F1000Research* **2016**, *5*, 1408. [CrossRef]
25. Robinson, M.D.; Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **2010**, *11*, R25. [CrossRef] [PubMed]
26. Lovász, L. Random walks on graphs. In *Combinatorics, Paul Erdos is Eighty*; Yale University: New Haven, CT, USA, 1993.
27. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*; Technical Report; Stanford InfoLab: Stanford Digital Libraries Working Paper: Stanford, CA, USA, 1998.
28. Chen, H.; Xiao, Q.; Hu, Y.; Chen, L.; Jiang, K.; Tang, Y.; Ding, K. ANGPTL1 attenuates colorectal cancer metastasis by up-regulating microRNA-138. *J. Exp. Clin. Cancer Res.* **2017**, *36*, 78. [CrossRef]
29. Sun, R.; Yang, L.; Hu, Y.; Wang, Y.; Zhang, Q.; Zhang, Y.; Zhao, D. ANGPTL1 is a potential biomarker for differentiated thyroid cancer diagnosis and recurrence. *Oncol. Lett.* **2020**, *20*, 240. [CrossRef]
30. Noguera-Uclés, J.F.; Boyero, L.; Salinas, A.; Cordero Varela, J.A.; Benedetti, J.C.; Bernabé-Caro, R.; Sánchez-Gastaldo, A.; Alonso, M.; Paz-Ares, L.; Molina-Pinelo, S. The Roles of Imprinted SLC22A18 and SLC22A18AS Gene Overexpression Caused by Promoter CpG Island Hypomethylation as Diagnostic and Prognostic Biomarkers for Non-Small Cell Lung Cancer Patients. *Cancers* **2020**, *12*, 2075. [CrossRef]
31. Wang, H.; Gu, R.; Tian, F.; Liu, Y.; Fan, W.; Xue, G.; Cai, L.; Xing, Y. PHLPP2 as a novel metastatic and prognostic biomarker in non-small cell lung cancer patients. *Thorac. Cancer* **2019**, *10*, 2124–2132. [CrossRef]
32. Cho, S.G. APC downregulated 1 inhibits breast cancer cell invasion by inhibiting the canonical WNT signaling pathway. *Oncol. Lett.* **2017**, *14*, 4845–4852. [CrossRef]
33. Yang, C.; Zhan, H.; Zhao, Y.; Wu, Y.; Li, L.; Wang, H. MEX3A contributes to development and progression of glioma through regulating cell proliferation and cell migration and targeting CCL2. *Cell Death Dis.* **2021**, *12*, 14. [CrossRef]
34. Lu, N.; Yin, Y.; Yao, Y.; Zhang, P. SNHG3/miR-2682-5p/HOXB8 promotes cell proliferation and migration in oral squamous cell carcinoma. *Oral Dis.* **2021**, *27*, 1161–1170. [CrossRef]
35. Chen, W.; Zhuang, X.; Qi, R.; Qiao, T. MiR-302a-5p suppresses cell proliferation and invasion in non-small cell lung carcinoma by targeting ITGA6. *Am. J. Transl. Res.* **2019**, *11*, 4348–4357. [PubMed]
36. Ma, J.; Li, D.; Kong, F.F.; Yang, D.; Yang, H.; Ma, X.X. miR-302a-5p/367-3p-HMGA2 axis regulates malignant processes during endometrial cancer development. *J. Exp. Clin. Cancer Res.* **2018**, *37*, 19. [CrossRef] [PubMed]
37. Liang, J.J.; Wang, J.Y.; Zhang, T.J.; An, G.S.; Ni, J.H.; Li, S.Y.; Jia, H.T. MiR-509-3-5p-NONHSAT112228. 2 Axis Regulates p21 and Suppresses Proliferation and Migration of Lung Cancer Cells. *Curr. Top. Med. Chem.* **2020**, *20*, 835–846. [CrossRef] [PubMed]
38. Xu, T.P.; Huang, M.D.; Xia, R.; Liu, X.X.; Sun, M.; Yin, L.; Shu, Y.Q. Decreased expression of the long non-coding RNA FENDRR is associated with poor prognosis in gastric cancer and FENDRR regulates gastric cancer cell metastasis by affecting fibronectin1 expression. *J. Hematol. Oncol.* **2014**, *7*, 63. [CrossRef]
39. Zheng, Q.; Zhang, Q.; Yu, X.; He, Y.; Guo, W. FENDRR: A pivotal, cancer-related, long non-coding RNA. *Biomed. Pharmacother.* **2021**, *137*, 111390. [CrossRef]
40. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
41. Mao, G.; Mu, Z.; Wu, D. Exosome-derived miR-2682-5p suppresses cell viability and migration by HDAC1-silence-mediated upregulation of ADH1A in non-small cell lung cancer. *Hum. Exp. Toxicol.* **2021**, *40*, S318–S330. [CrossRef]
42. Zhang, L.; Wang, Y.; Zhang, L.; You, G.; Li, C.; Meng, B.; Zhang, M. LINC01006 promotes cell proliferation and metastasis in pancreatic cancer via miR-2682-5p/HOXB8 axis. *Cancer Cell Int.* **2019**, *19*, 320. [CrossRef]