PLOS ONE

RESEARCH ARTICLE

Literature-based condition-specific miRNA-mRNA target prediction

Minsik Oh[1], Sungmin Rhee[1], Ji Hwan Moon[2], Heejoon Chae[3], Sunwon Lee[4], Jaewoo Kang[4], Sun Kim[1,2,5]*

1 Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea, **2** Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea, **3** Division of Computer Science, Sookmyung Women's University, Seoul, Republic of Korea, **4** Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea, **5** Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea

* sunkim.bioinfo@snu.ac.kr

Citation: Oh M, Rhee S, Moon JH, Chae H, Lee S, Kang J, et al. (2017) Literature-based condition-specific miRNA-mRNA target prediction. PLoS ONE 12(3): e0174999. https://doi.org/10.1371/journal.pone.0174999

Editor: Geraldo A Passos, University of São Paulo, BRAZIL

Received: November 27, 2016

Accepted: March 17, 2017

Published: March 31, 2017

Data Availability Statement: GSE21411, GSE40059 and GSE53482 are available from the GEO database(accession numbers GSE21411, GSE40059, GSE53482).

Funding: This work was supported by grant numbers 2012M3A9D1054622, 2014M3C9A3063541, and 2012M3C4A7033341, National Research Foundation of Korea (URL: http://www.nrf.re.kr/nrf_tot_cms/index.jsp?pmi-sso-return2=none). The authors who received the funding are: Minsik, Sungmin, Ji Hwan, Heejoon, Sunwon, Jaewoo, Sun. The funders had no role in

Abstract

miRNAs are small non-coding RNAs that regulate gene expression by binding to the 3′-UTR of genes. Many recent studies have reported that miRNAs play important biological roles by regulating specific mRNAs or genes. Many sequence-based target prediction algorithms have been developed to predict miRNA targets. However, these methods are not designed for condition-specific target predictions and produce many false positives; thus, expression-based target prediction algorithms have been developed for condition-specific target predictions. A typical strategy to utilize expression data is to leverage the negative control roles of miRNAs on genes. To control false positives, a stringent cutoff value is typically set, but in this case, these methods tend to reject many true target relationships, i.e., false negatives. To overcome these limitations, additional information should be utilized. The literature is probably the best resource that we can utilize. Recent literature mining systems compile millions of articles with experiments designed for specific biological questions, and the systems provide a function to search for specific information. To utilize the literature information, we used a literature mining system, BEST, that automatically extracts information from the literature in PubMed and that allows the user to perform searches of the literature with any English words. By integrating omics data analysis methods and BEST, we developed Context-MMIA, a miRNA-mRNA target prediction method that combines expression data analysis results and the literature information extracted based on the user-specified context. In the pathway enrichment analysis using genes included in the top 200 miRNA-targets, Context-MMIA outperformed the four existing target prediction methods that we tested. In another test on whether prediction methods can re-produce experimentally validated target relationships, Context-MMIA outperformed the four existing target prediction methods. In summary, Context-MMIA allows the user to specify a context of the experimental data to predict miRNA targets, and we believe that Context-MMIA is very useful for predicting condition-specific miRNA targets.

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

MicroRNAs (miRNAs) are small non-coding RNAs that are 19-24 nucleotides in length. These RNAs regulate gene expression at the post-transcriptional level by binding to the 3′-UTR of mRNAs [1, 2]; thus, miRNAs are functionally important. There are numerous scientific findings on the functional roles of miRNAs by regulating specific genes. For example, it is reported that miR-15 and miR-16-1 bind to BCL2 [3] and that apoptosis is induced. Another example is that miR-125b, miR-145, miR-21 and miR-155 are dysregulated in breast cancer cells, and different expression levels of these miRNAs have significant correlations with breast cancer phenotypes, such as tumor stages and status of estrogen and progesterone receptors [4]. Moreover, it is well known that miRNAs are related to proliferation, differentiation, and cell death [5].

The functional roles of miRNAs differ in different contexts. In other words, the relationship between miRNA and target genes is dynamic in different conditions. Thus, it is very important to identify which genes are targeted by miRNAs in a given context. There are more than 1000 miRNAs, and approximately 60% of protein-coding genes are regulated by miRNAs [6]. Since it is not possible to perform biological experiments for such a large number of miRNAs and genes, computational prediction is very important, and numerous computational methods have been developed for predicting targets of miRNAs. The first generation of computational tools leverage sequence complementary information and binding energy potentials. These prediction methods include TargetScan [7], PITA [8], mirSVR [9], miRanda [10] and PicTar [11]. These tools generally come with corresponding databases that compile miRNA-target information. In addition to sequence complementary information, there are different approaches used in each of these methods. miRanda estimates the energy on sequence matching of miRNA and mRNA pairs to predict targets [10]. PicTar first finds candidate 3′-UTR sites and uses a hidden Markov model (HMM) to filter out target sites [11]. TargetScan considers a conservation seed match and then considers regions outside seed matches [7]. The mirSVR algorithm uses a support vector regression method to compute scores on candidate target sites that are identified by miRanda [9]. PITA uses the accessibility of target sites as a main feature to predict targets [8].

Target prediction methods based on the sequence similarity score rely on the existence of target sites, and these methods are accompanied by target databases. However, such target information is not condition specific without considering which miRNAs and which genes are expressed; thus, there are many false positives even if the target information is accurate, which is not the case since many target databases do not agree on the miRNA-target relationship. To make the target information condition specific, many expression-based target prediction methods have been developed. These methods take miRNA-mRNA expression data and several sequence-based target databases as input data and filter out miRNA-mRNA targets using statistical significance or computational algorithms. We briefly summarize the previous expression-based algorithms. GenMiR++ used a Bayesian model and expectation maximization algorithm to predict the posterior probability of a miRNA target for mRNA [12]. MMIA employs a two-step method, where the first step is to select differentially expressed miRNA, and the second step is to select negatively correlated differentially expressed mRNA [13] only for the differentially expressed miRNAs. MMIA also supports sequence data analysis on a cloud environment, which enables the user to utilize both microarray data and NGS data [14]. MAGIA2 is a web-based tool that considers the correlation among miRNA and mRNA and transcription factor (TF) regulation [15]. CoSMic extracts the significant target mRNA cluster for each miRNA [16]. CoSMic employs methods similar to gene set enrichment analysis (GSEA) to identify miRNA targets [17]. miRNAmRNA is a target prediction algorithm based

on the global test of a linear regression model [18]. To extract condition-specific miRNA activity, identifying causal relationships using intervention calculus when the DAG is absent was proposed [19]. A recent tool, PlantMirnaT, was designed as a plant-specific miRNA-mRNA sequencing data analysis algorithm [20]. The unique feature of PlantMirnaT is using the expression quantity information from sequencing data and employing a split ratio model to identify the relationship of target pairs.

Motivation

There are approximately 1,500 known miRNAs in the human genome. The number of possible miRNA-gene pairs exceeds 30 million when more than 20,000 protein-coding genes are considered. Among these pairs, only a fraction of the relationships are significant in terms of biological functions, e.g., phenotypes or cancer subtypes. Computational methods for predicting the miRNA target employ various techniques to identify phenotype-specific miRNA targets. Because this is a typical prediction problem, the challenges can be summarized in terms of false positives and false negatives.

- **Target databases have high false positive rates**: Sequence-based target prediction algorithms, such as TargetScan, mirSVR, and PITA, and their corresponding databases generally produce high false positives. There are two major reasons for these high false positives. First, these databases contain all known targets; thus, the target information is not condition specific. For this reason, when transcriptome data measured in a specific condition are analyzed, many targets are false positives. Second, sequence-based prediction methods do not consider the regulatory role of miRNA, which generally results in a negative correlation between miRNA and the target gene. In addition, sequence-based prediction methods do not consider sample-specific sequence information. For example, sequence variations in the target regions can affect the target relationship, but the current algorithms do not consider minor but subtle sequence variations.
- **Expression-based methods may have false negative rates**: Expression-based methods utilize negative correlation information between miRNA and targets or similar approaches. For these methods, there is always an issue of establishing a cutoff threshold value, e.g., for a negative correlation. If the cutoff value is not stringent, then there are too many miRNA-target relationships. Thus, in general, it is a common practice to set a quite stringent cutoff value. In this case, many true miRNA-target relationships can be rejected, i.e., the false negative issue.

Addressing the false positive and false negative issues is a very challenging problem unless we fully understand how miRNAs regulate target genes. Using sequence pairing information and gene expression information is very useful because such methods have already produced many biologically meaningful results. However, one important information source, the literature, is not utilized in current methods. The scientific literature is currently growing exponentially. As shown in Fig 1, more than 100,000 papers related to 'cancer' are published every year. Thus, if we combine sequence pairing information and gene expression information with the literature information, we can certainly make a good improvement in predicting miRNA targets, reducing both false positives and false negatives. In particular, as with the use of gene expression information, the use of the literature information should be condition specific. The main issues are how to handle the vast amount of studies in the literature, how to allow the user to specify the experimental conditions, and finally, how to combine sequence pairing information, gene expression information and the literature information in a single computational framework.

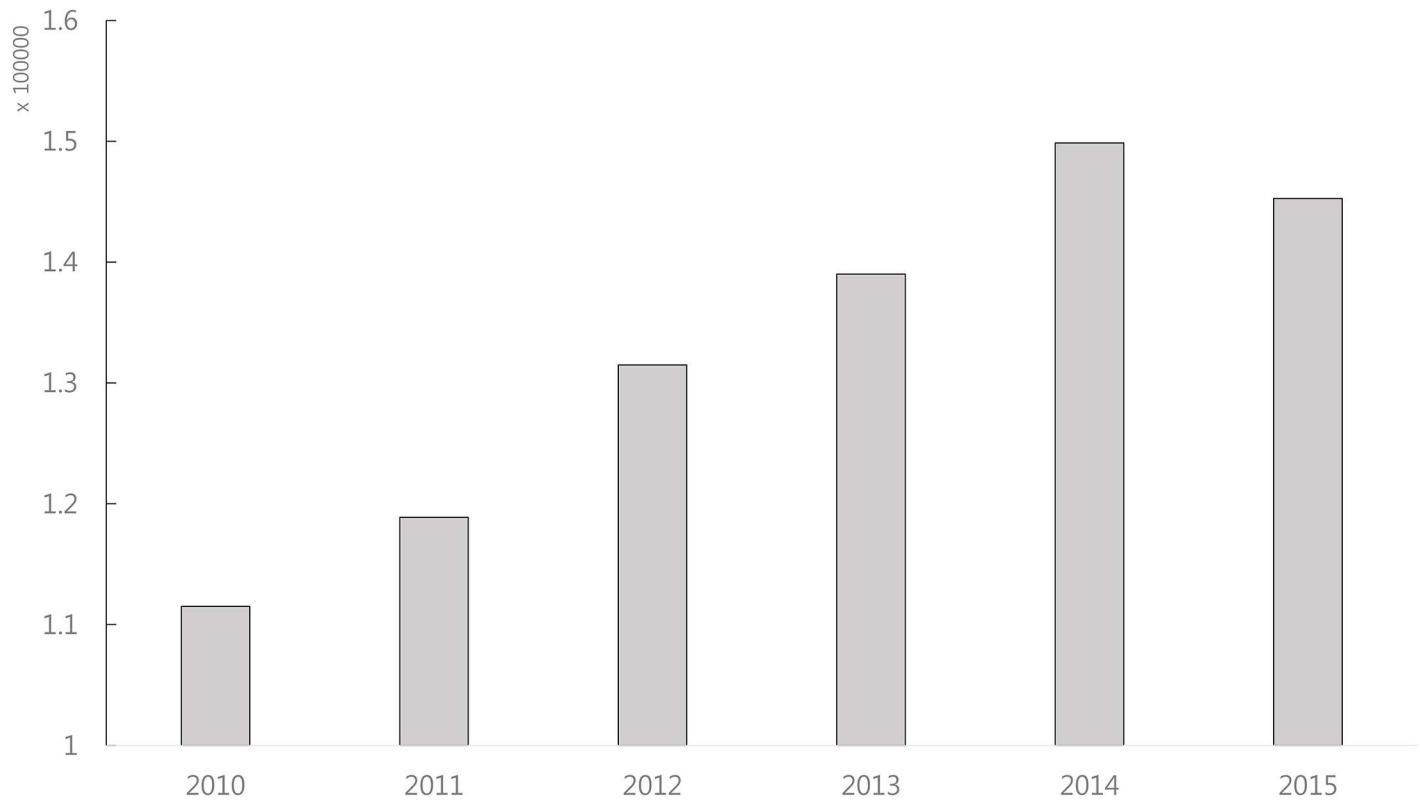

Fig 1. The number of published papers related to the keyword 'cancer' since 2010. More than 100,000 papers have been published every year.

https://doi.org/10.1371/journal.pone.0174999.g001

Toward this goal, two research groups are working together to design and implement a novel human-specific miRNA-target prediction method.

First, we compute the **omics score** by utilizing sequence pairing information and gene expression information to produce candidate miRNA-target pairs. Then, we compute the literature-based **context score** to evaluate each candidate miRNA-target pair using the Biomedical Entity Search Tool (BEST) [21]. Using BEST, the user can specify the experimental condition using a set of any keywords, which will automatically be translated to a set of genes and related miRNAs. Subsequently, the two scores, the **omics score** and the **context score**, are combined into a single score in a conditional probabilistic form.

The remainder of this paper is organized as follow. In the Methods section, we explain how to compute the **omics score** based on the expression data and miRNA-gene relationship and the **context score** from the literature according to user-provided keywords. In the Results section, we show how our proposed method performs compared with four existing methods in experiments with omics datasets in the public domain.

Methods

In this section, we explain how our method, Context-MMIA, predicts human miRNA targets by combining the literature information and gene expression data. Context-MMIA takes two-class (control vs. treated) human miRNA-mRNA expression data as input. Then, with user-specified keywords as the context of the experiment, it computes the probabilities of miRNA-gene pairs relevant to the phenotype differences by combining gene/miRNA expression data and the literature data. Fig 2 illustrates the workflow of Context-MMIA. First, differentially

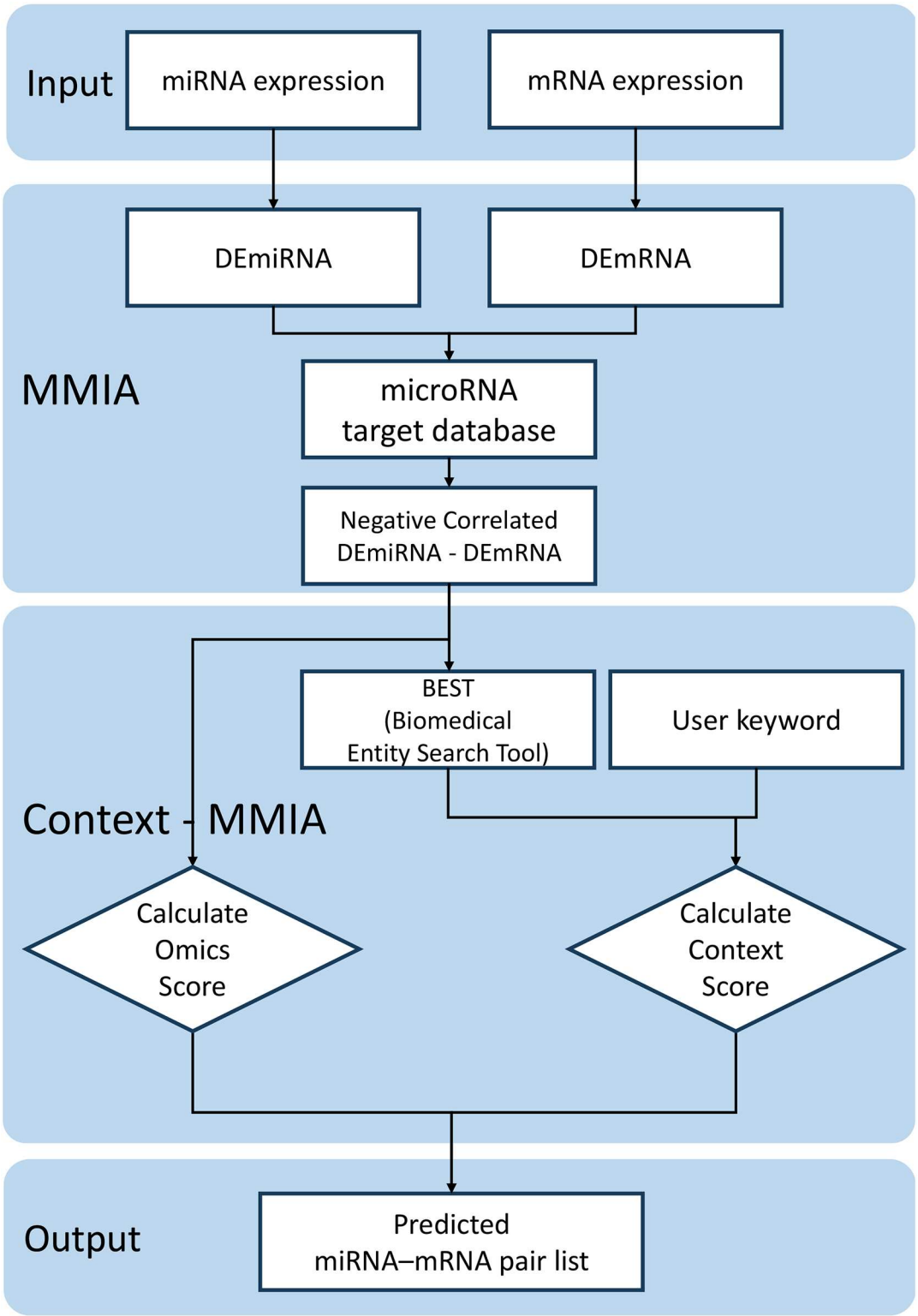

Fig 2. Schematic workflow for Context-MMIA. The system accepts expression information of miRNA and mRNA as inputs. In the MMIA step, DEmiRNAs and DEmRNAs are extracted based on their expression level difference, and their negative correlation is computed. In the Context-MMIA step, the system computes omics and context scores based on user-provided keywords by utilizing the BEST system. Finally, the system ranks miRNA-mRNA pairs using the scores.

https://doi.org/10.1371/journal.pone.0174999.g002

expressed miRNAs (DEmiRNAs) and differentially expressed mRNAs or genes (DEmRNAs) are determined with a cutoff value at the relaxed level such that most of the true positives can be retained in this step. Note that we use negative correlation information and the literature information to filter out and re-weight candidates for interaction pairs in the following steps. In the second step of processing omics data, human miRNA-mRNA pairs are predicted using miRNA target databases such as TargetScan, mirSVR, and PITA. These miRNA-mRNA pairs are further screened by negative correlation information between miRNA and mRNA. In the third step, for each pair of miRNA and mRNA, Context-MMIA calculates the **omics score** based on expression data and the **context score** based on the literature information compiled based on the user-provided keywords. Finally, target pairs are ranked by combining the **omics score** and **context score**. For each miRNA-mRNA pair, Context-MMIA computes alignments of human miRNA and the 3′-UTR of mRNA and generates the visualization of the miRNA-mRNA alignment on the website.

Identifying genes and miRNAs based on the user-provided context

Context-MMIA takes a set of keywords from the user to specify the context of the experiment. Currently, the most widely used biomedical literature database, PubMed, contains over 26 millions records. When we perform a search with the keyword 'cancer', over 3 million records are retrieved. Thus, we believe that this literature database contains enough articles to rank miRNA-gene pairs in terms of the user-provided context. However, there are two major issues in ranking miRNA-gene pairs: given the keywords, relevant papers should be identified and relevant gene names and miRNA names should also be identified. Since not all papers contain the user-provided keywords, it is necessary to infer the relevance of the words to extract genes and miRNAs in the relevant articles. To address this issue, we use BEST to identify relevant words and genes/miRNAs [21]. BEST has predefined biomedical entities for each category, such as drug, pathway, gene, and disease, and then it identifies relevant entities extracted from PubMed articles from the user query. For example, it returns entities such as 'ERBB2', 'wnt signaling pathway', and 'tamoxifen' with the keyword 'breast cancer' as an input. BEST has its own scoring system for entities, which is very useful in ranking gene-miRNA pairs with respect to the user-provided keywords. For example, there are keywords 'breast cancer' and entities 'cell cycle', 'mir-200c', 'BRCA1', and 'ESR'. At the beginning, BEST compiles PubMed articles containing 'breast cancer' and the four entities in the abstract. Then, it measures the score and the rank for each entity and lists entities ordered by score. After compiling articles containing 'BRCA1' and 'breast cancer', BEST calculates a document score for each article and sums the score to measure the entity score, which is denoted as *BEST*(*BreastCancer*, *BRCA*1). In this paper, we use BEST to measure the relevance of each miRNA and mRNA for a given user query.

Omics score

The **omics score** (OS) is the probability of a gene-miRNA contributing to the class difference when expression data are analyzed. The OS is based on the general principle that differentially expressed miRNA targets genes differentially, resulting in negative correlations between genes and miRNA; then, differentially expressed gene explains the phenotype differences. Context-MMIA computes the **omics score** based on a strategy similar to MMIA. It measures miRNA differential scores, mRNA differential scores, and then correlation scores. The DEmiRNAs and DEmRNAs can be determined by MMIA. After the DEmRNAs and DEmiRNAs are determined, the probability of miRNA-mRNA contributing to the class difference is calculated. Let the p-values of miRNA and mRNA be p_{m_i} and p_{g_j}, respectively. For miRNA m_i, m_i's differential

score $diff(m_i)$ is defined by Eq 1, and its normalization $diff_n(m_i)$ is defined by Eq 2.

$$diff(m_i) = -log_2(p_{m_i}) \tag{1}$$

$$diff_n(m_i) = \frac{diff(m_i) - \min(diff)}{\max(diff) - \min(diff)} \tag{2}$$

The calculation of $diff_n$ for mRNA is similar to that of miRNA. The range of $diff_n$ is between 0 and 1 by Eq 2. If miRNA is significantly differentially expressed in a given condition, then the value of $diff_n$ will be close to 1.

Correlation score is defined by measuring the Pearson's correlation coefficient of the miRNA-mRNA pair's logarithmic expression as in [22]. Context-MMIA considers only negatively correlated miRNA-mRNA pairs; thus, a negative value of the coefficient is defined as the correlation score as in Eq 3.

$$corr(m_i, g_j) = -pearson\_correlation(m_i, g_j) \tag{3}$$

The **omics score** of miRNA-mRNA $OS(m_i, g_j)$ is defined in Eq 4.

$$OS(m_i, g_j) = diff_n(m_i) * corr(m_i, g_j) * diff_n(g_j) \tag{4}$$

By definition, $OS(m_i, g_j) \in [0, 1]$; thus, a value of OS close to 1 means that the miRNA and mRNA are both significantly differentially expressed and anticorrelated. Thus, we predict that the pair is related to the phenotype difference with a high confidence in terms of expression data.

Context score

We defined the context score (CS) to measure the probability of a miRNA-mRNA pair contributing to the phenotype difference in terms of the literature information. As described in the previous section, BEST estimates a score between predefined entities and keywords. We denoted the user-input keyword as k, which is context specified by the user (e.g., disease, gene, pathway, and so forth). As shown in Eq 5, $CS(m_i, g_j|k)$ measures the significance of the m_i-g_j pair for k in terms of the literature information.

$$CS(m_i, g_j|k) = P(m_i|k) * P(g_j|k) \tag{5}$$

To compute $P(m_i|k)$, we used Bayes' rule and transformed $P(m_i|k)$ into Eq 6 because BEST only measures the score for predefined entities and does not support undefined keywords (e.g., broad keyword, new drug or pathway, and so on) [23].

$$P(m_i|k) = \frac{P_n(k|m_i) * P_n(m_i)}{\sum_{l=1}^{p} P_n(k|m_l) * P_n(m_l)} \tag{6}$$

By converting $P(m_i|k)$ using Bayes' rule, our method provides the user with a freeform keyword environment, which allows the user to easily utilize our system even when the user is not familiar with biological terms.

$$P(k|m_i) = log_2(BEST(k, m_i) + 1) \tag{7}$$

The literature significance of miRNA (m_i) for a given keyword k, $P(k|m_i)$, is computed as shown in Eq 7. $BEST(k, m_i)$ is the score of m_i for k computed by BEST, and we converted the scale of the score by taking the logarithm of the BEST score. For example, assume that the

keyword 'immune system' and the miRNA 'miR-155' are used in an analysis. If the relation between 'miR-155' and 'immune system' is well studied, then $P(immune\ system \mid miR155)$ and $BEST(immune\ system, miR155)$ will have a high score.

$$P(m_i) = log_2(BEST(m_i, m_i) + 1) \quad (8)$$

Eq 8 describes how to compute $P(m_i)$, which denotes how much literature information exists for m_i; the more that papers report m_i, the higher the value it will have. After computing $P(m_i)$ and $P(k|m_i)$, normalization terms $P_n(m_i)$ and $P_n(k|m_i)$ are defined by the min-max normalization.

$P(m_i|k)$ is computed using Bayes' rule and specifies the significance of m_i given the literature domain k, and the value of $P(m_i|k)$ has a correlation with the amount of studies, i.e., the number of papers about m_i in domain k. For mRNA g_j, $P(g_j|k)$ is computed in a similar way, and we measured the significance of the m_i-g_j pair in k by computing $CS(m_i, g_j|k)$ using $P(m_i|k)$ and $P(g_j|k)$.

Pair score

The pair score of m_i, g_j and k is denoted as $Score(m_i, g_j, k)$, which is a confidence value of target prediction in terms of both expression and literature data.

$$Score(m_i, g_j, k) = OS(m_i, g_j) \ * \ CS(m_i, g_j \mid k) \quad (9)$$

Eq 9 can be interpreted as a weighted **omics score**, where the weight is determined by a probability of a m_i, g_j pair being true in terms of the user-provided context given keywords k.

Results

To evaluate Context-MMIA, we performed three experiments in comparison with four existing tools: MMIA, MAGIA2, CoSMic and GenMiR++.

The three experiments were pathway analysis, reproducibility of validated miRNA targets in human, and sensitivity tests when different keywords were used for specifying the experimental context. We used 2-class microarray datasets containing miRNA and mRNA expression profiles in humans. GSE21411 [24], GSE40059 [25], and GSE53482 [26] from human disease studies were used. Each study reports experimentally validated miRNA and the correlated target mRNA pair, which was used to evaluate the miRNA target prediction methods in this section. A detailed description of each dataset is listed in Table 1.

Table 1 summarizes the validated target pair and the domain of the experimental design in each dataset. In the interstitial lung diseases (ILD) study, it was reported that ZEB-1 affects the persistence of disease in ILD through suppression of NEDD4L by miR-23a. In the GSE40059 breast cancer study, the authors investigated differences between aggressive breast cancer cell lines and less-aggressive cell lines and reported that CFL2 was up-regulated by miR-200c. The authors also reported that CFL2 expression was correlated with tumor grade. In the primary myelofibrosis (PMF) study, the authors revealed that overexpressed miR-155-5p regulates JARID2, and they suggested that regulated JARID2 may be related to MK hyperplasia in PMF. Disease information was used to test performances when different contexts are specified for Context-MMIA. It is necessary to choose keywords to specify contexts. 'Interstitial lung disease' and 'primary myelofibrosis' are too specific to use literature data; thus, we used the more general words 'lung disease' and 'myelofibrosis' as the keywords for Context-MMIA.

Table 1. Dataset summary. Each GEO study comes with an experimentally validated miRNA-mRNA target (the second column) to affect their disease domain (the third column). Disease information was used to test performances when different contexts are specified.

Data	Experimentally validated target	Disease
GSE21411	hsa-miR-23a—NEDD4L	Interstitial Lung Diseases
GSE40059	hsa-miR-200c—CFL2	Breast Cancer
GSE53482	hsa-miR-155—JARID2	Primary Myelofibrosis

https://doi.org/10.1371/journal.pone.0174999.t001

Pathway analysis

To evaluate the effectiveness of the approach used in Context-MMIA, we compared it with four expression-based methods: MMIA, MAGIA2, GenMiR++, and CoSMic. GenMiR++ computes probabilities for target pairs using an EM algorithm. MMIA extracts DEmiRNA to reduce the search space by a user-defined cutoff and finds negatively expressed target DEmRNAs. MAGIA2 provides several methods for the integrated analysis, and we chose Pearson's correlation method from among these methods. After measuring the correlation, MAGIA2 calculates the false discovery rate (FDR) for each target. CoSMic extracts an mRNA cluster for each miRNA and computes the significance of a cluster using permutation tests. Likewise, each algorithm uses a different strategy to predict the miRNA target and to reduce the search space. We used these four algorithms to compare performances in terms of the predictive power. The methods compute confidence values for the predicted miRNA and mRNA targets, typically probability or p-value. We ranked the prediction results in terms of the confidence values. In the experiments, we used a p-value cutoff of 0.1 for Context-MMIA. For MMIA, a p-value of 0.05 was used for both DEmiRNA and DEmRNA selection.

For the performance evaluation, we used the top 200 predicted miRNA-mRNA pairs predicted by each method. Then, we mapped genes included in the interacting pairs to human pathways using DAVID [27, 28] to determine which pathways were significantly enriched. Among these pathways, we carefully selected pathways that are most likely related to the disease through the literature study as shown in Table 1. We set evaluation criteria as how these literature-guided pathways were predicted by each method. Table 2 shows the ratios of the number of genes that are mapped to significantly enriched pathways to the number of genes included in the top 200 miRNA-target edges. The number of genes is less than 200 because the same gene was multiply targeted, e.g., miR-200c-BRCA1 and miR-23a-BRCA1.

As shown in Table 2, the number of genes mapped to the significantly enriched pathways is quite different for each method even though the number of genes does not considerably differ for each method. In terms of the ratio of mapped genes to predicted genes, Context-MMIA outperforms the existing methods 2 to 4 times. A gene set in a pathway means that genes have similar biological functions in terms of regulating molecular processes. Thus, the ratios in Table 2 indicate that Context-MMIA produces more functionally coherent gene sets.

Table 2. The ratio of the mapped genes and the number of the genes in the top 200 miRNA-target pairs. From each method, we extracted the top 200 target pairs using each method and performed pathway analysis using DAVID. The numerator is the number of genes mapped to the enriched pathways, and the denominator is the genes in the top 200 edges. The ratio of Context-MMIA is the largest for each dataset.

Methods	GSE21411	GSE40059	GSE53482
Context-MMIA	37 / 79	45 / 157	42 / 127
MMIA	12 / 157	20 / 179	11 / 124
GenMiR++	0 / 194	18 / 197	26 / 200
MAGIA2	18 / 182	12 / 191	19 / 193
CoSMic	24 / 196	9 / 195	X

https://doi.org/10.1371/journal.pone.0174999.t002

Table 3. Enriched pathway analysis on GSE40059 breast cancer data. Breast-cancer-related pathways are selected by the literature search. A circle in a cell means that the pathway is enriched by the gene set predicted by each method (A: Context-MMIA, B: MMIA, C: GenMiR++, D: MAGIA2, and E: CoSMic). More pathways are enriched by the gene set in the Context-MMIA result.

Breast-Cancer-Related Pathway	A	B	C	D	E
Purine metabolism [29]		O			
Pyrimidine metabolism [30]		O			
ABC transporters [31]			O		
MAPK signaling pathway [32]	O				
Cytokine-cytokine receptor interaction [33]	O				
Neuroactive ligand-receptor interaction [34]			O		
p53 signaling pathway [35]	O	O			
Apoptosis [36]		O			
Notch signaling pathway [37]				O	
TGF-beta signaling pathway [38]	O				
Axon guidance [39]				O	
Focal adhesion [40]	O	O		O	
ECM-receptor interaction [41]		O			
Cell adhesion molecules (CAMs) [42]	O		O		
Adherens junction [43]	O				
Regulation of actin cytoskeleton [44]	O				
Glioma [45]	O				
Melanoma [46]	O				

https://doi.org/10.1371/journal.pone.0174999.t003

Table 3 lists pathways related to 'breast cancer' and enriched pathways predicted by each method for the GSE40059 dataset. The enriched pathway analysis for the data from all three experiments is presented in S1 File. The circles in Table 3 mean an enriched pathway when DAVID pathway analysis was performed by using genes in the top 200 edges. For example, if the ECM-receptor interaction is enriched in the Context-MMIA and GenMiR++ results, circles are marked in the context column and the second column for the corresponding tools. As shown in Table 3, more pathways related to 'breast cancer' were enriched in the gene sets produced by Context-MMIA than in the gene sets produced by the competing methods. In addition, several important pathways were enriched only in Context-MMIA. For example, it is well known that approximately half of breast tumors have stronger MAP kinase activity than the surrounding benign tissues [32]. Inflammation plays a pivotal role in tumor initiation, promotion, angiogenesis and metastasis. Cytokines are important in all the phenomena, and it has been reported that cytokines participate in regulating both induction and protection in breast cancer [33]. In addition, many studies have reported that TGF-beta signaling is critically important in the regulation of breast cancer [38]. High focal adhesion kinase expression is known to be related to aggressive breast cancer phenotypes [47]. Furthermore, cell adhesion molecules (CAMs) have a strong relationship with the process of metastasis, which is an important feature in predicting breast cancer prognosis [42]. Moreover, a study revealed that activated leukocyte cell adhesion molecule (ALCAM) expression has a correlation with clinical outcomes such as grade, TNM stage, and NPI [48].

Reproducibility of validated targets in humans

Table 4 shows the rankings of experimentally validated targets among the targets predicted by each method. Because Context-MMIA computes the context score using the literature data for given keywords, there is a possibility that the original papers of the datasets can affect the

Table 4. Reproducibility of validated targets. This table contains the rankings of validated target pairs in three datasets. The validated targets are listed in the second column of Table I. Context-MMIA outperformed existing tools in predicting the validated targets. MAGIA2 and CoSMic failed to reproduce the validated targets.

Data	GSE21411	GSE40059	GSE53482
Context-MMIA	**481**	**338**	**21**
MMIA	1411	387	1465
GenMiR++	8625	1673	95492
MAGIA2	X	X	X
CoSMic	X	X	X (Not Work)

https://doi.org/10.1371/journal.pone.0174999.t004

context score. Thus, we penalized the validated targets to compute $P(k|m_i)$ by excluding each paper when the BEST tool measures a score $BEST(k, m_i)$.

As shown in Table 4, Context-MMIA outperformed the other expression-based methods even though the penalized score is used. MMIA took the second place in reproducing the validated targets, but it ranked validated targets much lower than Context-MMIA. Although not rejecting the validated targets, GenmiR++ ranked validated targets very low. This result shows that GenmiR++ produced too many false positives for the three datasets. MAGIA2 failed to identify the validated targets as positive target pairs in any datasets because none of the validated target pairs satisfied the statistical cutoff. CoSMic also failed to identify the validated target pairs for two datasets, GSE21411 and GSE40059. In addition, CoSMic did not run successfully for dataset GSE53482 due to an input error issue. Many tools were not successful in reproducing validated targets, which can be an indication of false negatives.

To further confirm the reproducibility of our algorithm, we investigated how many experimentally verified targets in humans are detected in the top 200 miRNA-mRNA pairs by each of the methods. Experimentally validated human miRNA-mRNA pairs were extracted from miRTarBase [49], which curated experimentally validated miRNA-target interactions (MTI) by reporter assay, western blot, microarray, and next-generation sequencing experiments. We used human functional MTIs with strong evidence for functionality in humans as true interacting pairs. Table 5 summarizes the number of validated targets in the top 200 miRNA-mRNA pairs predicted by each method.

As shown in 5, Context-MMIA predicted two to five times more validated targets compared to the existing methods. Context-MMIA predicted more than 10% of the experimentally validated MTIs in humans, with is a considerably higher prediction accuracy than existing methods; thus, we believe that Context-MMIA suggests good candidates for further experimental validation.

Sensitivity tests when different keywords are used

The performance of Context-MMIA depends on how the keywords to specify context are related to the goal of the experiment. In addition to disease-related keywords, we performed

Table 5. Detection of human-specific validated targets. This table contains the number of validated target pairs in three datasets. The validated targets are extracted from miRTarBase target pairs filtered by human functional miRNA target interaction (MTI).

Data	GSE21411	GSE40059	GSE53482
Context-MMIA	**27**	**38**	**24**
MMIA	5	4	12
GenMiR++	3	4	3
MAGIA2	0	0	0
CoSMic	7	0	X (Not Work)

https://doi.org/10.1371/journal.pone.0174999.t005

Table 6. Sensitivity tests when different keywords are used. Rankings of validated targets are shown when different keywords are used. The validated targets had high ranks when disease-related keywords were used.

Keyword	GSE21411	GSE40059	GSE53482
Correct keyword	**481**	**338**	**21**
Insulin resistance	12479	2036	4250
Influenzas	6826	1169	1623
HIV	5865	4002	3238
Hepatocellular carcinoma	5278	3265	7180

https://doi.org/10.1371/journal.pone.0174999.t006

experiments using less-relevant keywords such as insulin resistance, influenzas, HIV and hepatocellular carcinoma. The results of Context-MMIA using less-relevant keywords are presented in Table 6. The relevant keywords for the three datasets are listed in the third column of Table 1. As shown in Table 6, the rankings of the validated pairs were considerably higher when the keywords that reflect experimental designs were used. This result indicates that our method is able to reflect the degree of relevance to the experimental design and capture the different miRNA-mRNA pairs when different keywords were used. In summary, the experiments with irrelevant keywords showed that our method can capture the miRNA-mRNA pairs, reflecting the user-specified biological context.

Conclusion

We presented Context-MMIA, a human-specific miRNA-mRNA target pair prediction system that utilizes both expression profiles and the literature information from the user-specified experimental design goals. A major contribution of our system is that we handled the false positives and false negatives, which are an inherent issue in expression-based prediction tools, by incorporating the user-specified context information from the literature. Analyses on three independent human datasets showed that Context-MMIA can capture the true positive miRNA-mRNA target pairs that are specific to a biological context. Context-MMIA outperformed existing tools in a series of experiments, such as pathway analysis, validated target ranking, and irrelevant keyword experiments.

We emphasize that computational predictions of miRNA-mRNA target pairs should be further validated in biological experiments and that our system is intended to provide good candidates for experimental validation. Context-MMIA is available at http://biohealth.snu.ac.kr/software/contextMMIA

Supporting information

S1 File. Pathway analysis results. S1 File contains pathway results for the other two datasets. (PDF)

Author Contributions

Conceptualization: SK JK.

Data curation: MO.

Formal analysis: MO.

Funding acquisition: SK JK.

Investigation: SK MO.

Methodology: MO SR.

Project administration: SK JK.

Software: MO HC SL.

Supervision: SK.

Validation: SK MO JHM SR.

Writing – original draft: SK MO.

Writing – review & editing: SK MO SR JHM HC.

References

1. Ambros V. The functions of animal microRNAs. Nature. 2004; 431(7006):350–355. https://doi.org/10.1038/nature02871 PMID: 15372042
2. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. cell. 2004; 116(2):281–297. https://doi.org/10.1016/S0092-8674(04)00045-5 PMID: 14744438
3. Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, Shimizu M, et al. miR-15 and miR-16 induce apoptosis by targeting BCL2. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102(39):13944–13949. https://doi.org/10.1073/pnas.0506654102 PMID: 16166262
4. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer research. 2005; 65(16):7065–7070. https://doi.org/10.1158/0008-5472.CAN-05-1783 PMID: 16103053
5. Hwang H, Mendell J. MicroRNAs in cell proliferation, cell death, and tumorigenesis. British journal of cancer. 2006; 94(6):776–780. https://doi.org/10.1038/sj.bjc.6603023 PMID: 16495913
6. Friedman RC, Farh KKH, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome research. 2009; 19(1):92–105. https://doi.org/10.1101/gr.082701.108 PMID: 18955434
7. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. cell. 2005; 120(1):15–20. https://doi.org/10.1016/j.cell.2004.12.035 PMID: 15652477
8. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nature genetics. 2007; 39(10):1278–1284. https://doi.org/10.1038/ng2135 PMID: 17893677
9. Betel D, Koppal A, Agius P, Sander C, Leslie C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome biology. 2010; 11(8):R90. https://doi.org/10.1186/gb-2010-11-8-r90 PMID: 20799968
10. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human microRNA targets. PLoS Biol. 2004; 2(11):e363. https://doi.org/10.1371/journal.pbio.0020363 PMID: 15502875
11. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nature genetics. 2005; 37(5):495–500. https://doi.org/10.1038/ng1536 PMID: 15806104
12. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, et al. Using expression profiling data to identify human microRNA targets. Nature methods. 2007; 4(12):1045–1049. https://doi.org/10.1038/nmeth1130 PMID: 18026111
13. Nam S, Li M, Choi K, Balch C, Kim S, Nephew KP. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. Nucleic acids research. 2009; 37 (suppl 2):W356–W362. https://doi.org/10.1093/nar/gkp294 PMID: 19420067
14. Chae H, Rhee S, Nephew KP, Kim S. BioVLAB-MMIA-NGS: microRNA–mRNA integrated analysis using high-throughput sequencing data. Bioinformatics. 2014; p. btu614. PMID: 25270639
15. Bisognin A, Sales G, Coppe A, Bortoluzzi S, Romualdi C. MAGIA2: from miRNA and genes expression data integrative analysis to microRNA–transcription factor mixed regulatory circuits (2012 update). Nucleic acids research. 2012; p. gks460. https://doi.org/10.1093/nar/gks460 PMID: 22618880
16. Ben-Moshe NB, Avraham R, Kedmi M, Zeisel A, Yitzhaky A, Yarden Y, et al. Context-specific microRNA analysis: identification of functional microRNAs and their mRNA targets. Nucleic acids research. 2012; 40(21):10614–10627. https://doi.org/10.1093/nar/gks841
17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings

of the National Academy of Sciences. 2005; 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102

18. van Iterson M, Bervoets S, de Meijer EJ, Buermans HP, AC't Hoen P, Menezes RX, et al. Integrated analysis of microRNA and mRNA expression: adding biological significance to microRNA target predictions. Nucleic acids research. 2013; 41(15):e146–e146. https://doi.org/10.1093/nar/gkt525 PMID: 23771142
19. Zhang J, Le TD, Liu L, Liu B, He J, Goodall GJ, et al. Inferring condition-specific miRNA activity from matched miRNA and mRNA expression data. Bioinformatics. 2014; p. btu489. https://doi.org/10.1093/bioinformatics/btu489
20. Rhee S, Chae H, Kim S. PlantMirnaT: miRNA and mRNA integrated analysis fully utilizing characteristics of plant sequencing data. Methods. 2015; 83:80–87. https://doi.org/10.1016/j.ymeth.2015.04.003 PMID: 25863133
21. Lee S, Kim D, Lee K, Choi J, Kim S, Jeon M, et al. BEST: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. PloS one. 2016; 11(10):e0164680. https://doi.org/10.1371/journal.pone.0164680 PMID: 27760149
22. Mukherji S, Ebert MS, Zheng GX, Tsang JS, Sharp PA, van Oudenaarden A. MicroRNAs can generate thresholds in target gene expression. Nature genetics. 2011; 43(9):854–859. https://doi.org/10.1038/ng.905 PMID: 21857679
23. Lee J, Jo K, Lee S, Kang J, Kim S. Prioritizing biological pathways by recognizing context in time-series gene expression data. BMC Bioinformatics. 2016; 17(17):17.
24. Cho JH, Gelinas R, Wang K, Etheridge A, Piper MG, Batte K, et al. Systems biology of interstitial lung diseases: integration of mRNA and microRNA expression changes. BMC medical genomics. 2011; 4(1):1. https://doi.org/10.1186/1755-8794-4-8
25. Luo D, Wilson JM, Harvel N, Liu J, Pei L, Huang S, et al. A systematic evaluation of miRNA: mRNA interactions involved in the migration and invasion of breast cancer cells. Journal of translational medicine. 2013; 11(1):1. https://doi.org/10.1186/1479-5876-11-57
26. Norfo R, Zini R, Pennucci V, Bianchi E, Salati S, Guglielmelli P, et al. miRNA-mRNA integrative analysis in primary myelofibrosis CD34+ cells: role of miR-155/JARID2 axis in abnormal megakaryopoiesis. Blood. 2014; 124(13):e21–e32. https://doi.org/10.1182/blood-2013-12-544197 PMID: 25097177
27. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2009; 37(1):1–13. https://doi.org/10.1093/nar/gkn923
28. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2009; 4(1):44–57. https://doi.org/10.1038/nprot.2008.211
29. Schramm G, Surmann EM, Wiesberg S, Oswald M, Reinelt G, Eils R, et al. Analyzing the regulation of metabolic pathways in human breast cancer. BMC medical genomics. 2010; 3(1):1. https://doi.org/10.1186/1755-8794-3-39
30. Sigoillot FD, Sigoillot SM, Guy HI. Breakdown of the regulatory control of pyrimidine biosynthesis in human breast cancer cells. International journal of cancer. 2004; 109(4):491–498. https://doi.org/10.1002/ijc.11717 PMID: 14991569
31. Fletcher JI, Haber M, Henderson MJ, Norris MD. ABC transporters in cancer: more than just drug efflux pumps. Nature Reviews Cancer. 2010; 10(2):147–156. https://doi.org/10.1038/nrc2789 PMID: 20075923
32. Santen RJ, Song RX, McPherson R, Kumar R, Adam L, Jeng MH, et al. The role of mitogen-activated protein (MAP) kinase in breast cancer. The Journal of steroid biochemistry and molecular biology. 2002; 80(2):239–256. https://doi.org/10.1016/S0960-0760(01)00189-3 PMID: 11897507
33. Esquivel-Velázquez M, Ostoa-Saloma P, Palacios-Arreola MI, Nava-Castro KE, Castro JI, Morales-Montor J. The role of cytokines in breast cancer development and progression. Journal of Interferon & Cytokine Research. 2015; 35(1):1–16. https://doi.org/10.1089/jir.2014.0026
34. Finak G, Bertos N, Pepin F, Sadekova S, Souleimanova M, Zhao H, et al. Stromal gene expression predicts clinical outcome in breast cancer. Nature medicine. 2008; 14(5):518–527. https://doi.org/10.1038/nm1764 PMID: 18438415
35. Gasco M, Shami S, Crook T. The p53 pathway in breast cancer. Breast Cancer Research. 2002; 4(2):70. https://doi.org/10.1186/bcr426 PMID: 11879567
36. Lipponen P. Apoptosis in breast cancer: relationship with other pathological parameters. Endocrine-related cancer. 1999; 6(1):13–16. https://doi.org/10.1677/erc.0.0060013 PMID: 10732780
37. Reedijk M. Notch signaling and breast cancer. In: Notch Signaling in Embryology and Cancer. Springer; 2012. p. 241–257.

38. Moses H, Barcellos-Hoff MH. TGF-β biology in mammary development and breast cancer. Cold Spring Harbor perspectives in biology. 2011; 3(1):a003277. https://doi.org/10.1101/cshperspect.a003277 PMID: 20810549

39. Mehlen P, Delloye-Bourgeois C, Chédotal A. Novel roles for Slits and netrins: axon guidance cues as anticancer targets? Nature reviews Cancer. 2011; 11(3):188–197. https://doi.org/10.1038/nrc3005 PMID: 21326323

40. McLean GW, Carragher NO, Avizienyte E, Evans J, Brunton VG, Frame MC. The role of focal-adhesion kinase in cancer-a new therapeutic opportunity. Nature Reviews Cancer. 2005; 5(7):505–515. https://doi.org/10.1038/nrc1647 PMID: 16069815

41. Lu P, Weaver VM, Werb Z. The extracellular matrix: a dynamic niche in cancer progression. The Journal of cell biology. 2012; 196(4):395–406. https://doi.org/10.1083/jcb.201102147 PMID: 22351925

42. Saadatmand S, De Kruijf E, Sajet A, Dekker-Ensink N, van Nes J, Putter H, et al. Expression of cell adhesion molecules and prognosis in breast cancer. British Journal of Surgery. 2013; 100(2):252–260. https://doi.org/10.1002/bjs.8980 PMID: 23175431

43. Haidari M, Zhang W, Wakame K. Disruption of endothelial adherens junction by invasive breast cancer cells is mediated by reactive oxygen species and is attenuated by AHCC. Life sciences. 2013; 93 (25):994–1003. https://doi.org/10.1016/j.lfs.2013.10.027 PMID: 24211779

44. Flamini M, Sanchez A, Goglia L, Tosi V, Genazzani A, Simoncini T. Differential actions of estrogen and SERMs in regulation of the actin cytoskeleton of endometrial cells. Molecular human reproduction. 2009; 15(10):675–685. https://doi.org/10.1093/molehr/gap045 PMID: 19541800

45. Piccirilli M, Salvati M, Bistazzoni S, Frati A, Brogna C, Giangaspero F, et al. Glioblastoma multiforme and breast cancer: report on 11 cases and clinico-pathological remarks. Tumori. 2005; 91(3):256. PMID: 16206651

46. Goggins W, Gao W, Tsao H. Association between female breast cancer and cutaneous melanoma. International journal of cancer. 2004; 111(5):792–794. https://doi.org/10.1002/ijc.20322 PMID: 15252852

47. Lark AL, Livasy CA, Dressler L, Moore DT, Millikan RC, Geradts J, et al. High focal adhesion kinase expression in invasive breast carcinomas is associated with an aggressive phenotype. Modern Pathology. 2005; 18(10):1289–1294. https://doi.org/10.1038/modpathol.3800424 PMID: 15861214

48. King JA, Ofori-Acquah SF, Stevens T, Al-Mehdi AB, Fodstad O, Jiang WG. Activated leukocyte cell adhesion molecule in breast cancer: prognostic indicator. Breast Cancer Research. 2004; 6(5):1. https://doi.org/10.1186/bcr815

49. Hsu Sheng-Da and Lin Feng-Mao and Wu Wei-Yun and Liang Chao and Huang Wei-Chih and Chan, et al. miRTarBase: a database curates experimentally validated microRNA–target interactions. Nucleic acids research. 2010; p. gkq1107.