Heliyon 9 (2023) e22666

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

miRGediNET: A comprehensive examination of common genes in miRNA-Target interactions and disease associations: Insights from a grouping-scoring-modeling approach

Emma Qumsiyeh [a,*], Zaidoun Salah [b], Malik Yousef [c]

[a] Department of Computer Science and Information Technology, Al-Quds University, Palestine
[b] Molecular Genetics and Genetic Toxicology, Arab American University, Ramallah, Palestine
[c] Information Technology Engineering, Al-Quds University, Abu Dis, Palestine

ABSTRACT

In the broad and complex field of biological data analysis, researchers frequently gather information from a single source or database. Despite being a widespread practice, this has disadvantages. Relying exclusively on a single source can limit our comprehension as it may omit various perspectives that could be obtained by combining multiple knowledge bases. Acknowledging this shortcoming, we report on miRGediNET, a novel approach combining information from three biological databases. Our investigation focuses on microRNAs (miRNAs), small non-coding RNA molecules that regulate gene expression post-transcriptionally. We delve deeply into the knowledge of these miRNA's interactions with genes and the possible effects these interactions may have on different diseases. The scientific community has long recognized a direct correlation between the progression of specific diseases and miRNAs, as well as the genes they target. By using miRGediNET, we go beyond simply acknowledging this relationship. Rather, we actively look for the critical genes that could act as links between the actions of miRNAs and the mechanisms underlying disease. Our methodology, which carefully identifies and investigates these important genes, is supported by a strategic framework that may open up new possibilities for comprehending diseases and creating treatments. We have developed a tool on the Knime platform as a concrete application of our research. This tool serves as both a validation of our study and an invitation to the larger community to interact with, investigate, and build upon our findings. miRGediNET is publicly accessible on GitHub at https://github.com/malikyousef/miRGediNET, providing a collaborative environment for additional research and innovation for enthusiasts and fellow researchers.

1. Introduction

Small non-coding RNAs known as microRNAs (miRNAs) are a class of RNAs that are crucial for controlling gene expression. These molecules' involvement in a range of cellular processes, including cell growth, differentiation, and apoptosis, has piqued the interest of the scientific community [1]. Recent technological developments have made it possible to identify many miRNAs and their target genes, greatly enhancing our knowledge of how these molecules function in biological processes.

However, understanding the relationship between microRNAs and diseases is intricate and fraught with difficulties. According to previous research, miRNAs can function as oncogenes or tumor suppressors based on the cellular context [2]. Their dual function complicates our understanding of their role in disease mechanisms.

Furthermore, the current state of research is dispersed, with results frequently limited to particular diseases or miRNAs, which hinders our ability to create a comprehensive understanding of the interactions between miRNAs and diseases [3]. Additionally, there are insufficient computational tools to combine various biological data sources and present a complete picture of these interactions [4].

[*] Corresponding author.
E-mail address: emma.qumsiyeh@hotmail.com (E. Qumsiyeh).

https://doi.org/10.1016/j.heliyon.2023.e22666
Received 19 July 2023; Received in revised form 15 November 2023; Accepted 16 November 2023
Available online 22 November 2023

Given these difficulties, the importance of our work lies in our attempt to provide a single framework, called miRGediNET, that combines information from various sources to investigate the intricate connections between miRNAs, the genes they target, and the diseases they are linked to. Our method seeks to open up new perspectives on these complex interactions, which may lay the groundwork for further studies and treatment approaches in the future.

Machine learning algorithms have been increasingly applied in the analysis of miRNA-disease associations. One such application is the development of miRNA-based disease prediction models using machine learning algorithms. In [2], the authors developed a deep learning algorithm called DeepMiR that integrates multiple data sources, including miRNA expression data, disease phenotypes, and protein-protein interaction networks, to predict miRNA-disease associations. The DeepMiR algorithm uses a graph convolutional neural network (GCN) to extract features from the heterogeneous network and a multi-layer perceptron (MLP) to perform binary classification of miRNA-disease associations. The algorithm was trained on a dataset of known miRNA-disease associations and evaluated using cross-validation and independent testing datasets.

In [3], the authors introduced RBMMMDA, a method for predicting multiple types of disease-microRNA associations. RBMMMDA combines miRNA functional similarity, disease semantic similarity, and known miRNA-disease associations to create a weighted tensor representation. The algorithm then uses graph-regularized weighted tensor decomposition to discover hidden associations between miRNAs and diseases. RBMMMDA outperforms existing methodologies in predicting miRNA-disease associations, as demonstrated by a comprehensive evaluation.

In another study by Lan et al. [5], a computational framework named KBMF-MDI was introduced. This method predicts associations between miRNAs and diseases based on their similarities. The study employed both sequence and function information of miRNAs to measure similarity among miRNAs and used semantic and function information to measure similarity among diseases. By integrating these data sources, the kernelized Bayesian matrix factorization method was used to deduce potential miRNA-disease associations. When applied to 6084 known miRNA-disease associations using 5-fold cross-validation, the results showed that KBMF-MDI could effectively predict unknown miRNA-disease associations.

Graph Neural Networks (GNNs) have indeed made significant strides in analyzing the miRNA-gene-disease network, leveraging the inherent graph structure of biological interactions to model and predict associations. However, our suggested miRGediNET method presents several distinct advantages over traditional GNN-based approaches. While GNNs predominantly focus on the graph structure of data, miRGediNET integrates information from three distinct biological databases. This ensures a multi-faceted perspective, tapping into a broader spectrum of biological insights. miRGediNET has a unique framework that offers a systematic approach, allowing for structured categorization, prioritization, and predictive modeling of genes. In contrast, many GNNs might be more focused on the holistic network structure without delving deep into individual entity prioritization.

Many feature selection algorithms employed in gene expression data analysis rely on statistical and machine learning approaches. However, these methods often overlook the valuable biological knowledge embedded in the data that could significantly enhance the feature selection process. Bellazzi and Zupan have extensively explored the advancements in gene expression-based analysis techniques, with a particular focus on studies involving associations and classification and the implications of reverse-engineering gene-gene networks and resulting phenotypes [4]. Notably, incorporating biological knowledge into clustering algorithms poses a formidable challenge. In a separate study, Kustra and Zagdanski addressed this challenge in a separate study by incorporating Gene Ontology (GO) annotations into gene expression data, employing a correlation-based dissimilarity matrix to derive a GO-based dissimilarity matrix [6].

Traditional gene selection approaches suffer from various limitations. A prominent drawback is their inadequate ability to provide meaningful biological interpretations, consequently hindering the generation of novel biological knowledge [7,8]. Recognizing this limitation, researchers have recently focused on developing integrative gene selection approaches. These novel methodologies aim to incorporate domain knowledge derived from external biological resources (prior-biological knowledge) into the analysis of gene expression data. By leveraging such integrative approaches, scientists can harness the power of additional biological information to enhance their understanding and interpretation of gene expression patterns.

Integrative machine-learning approaches that incorporate prior-biological knowledge leverage existing knowledge about genes, pathways, and biological processes to improve the accuracy and interpretability of the machine-learning models [7,9]. These approaches aim to integrate prior knowledge with data-driven analysis to enhance the understanding of complex biological systems. One common strategy is incorporating prior knowledge as features or constraints in the machine learning model. For example, prior knowledge about gene-gene interactions or gene-pathway associations can be used as additional features to improve the model's predictive performance [10]. This integration allows the model to capture the underlying biological relationships and dependencies, leading to more accurate predictions and better biological interpretation. Another approach is to use prior knowledge to guide the model selection or regularization process. Prior knowledge can be utilized to impose constraints on the model parameters or structure, promoting solutions that are consistent with the known biological context. This helps to prevent overfitting and enhances the interpretability of the resulting model [11].

Raghu et al. [12] proposed an integrative gene selection approach combining KEGG, DisGeNET, and additional genetic meta-information. They calculate each gene's importance score and distance metrics for each gene, using a combination of gene-disease association scores from DisGeNET and gene expression levels. On the other hand, Perscheid et al. [8] suggested a flexible approach that combines traditional gene selection methods with multiple knowledge bases. They compared traditional gene selection approaches with integrative gene selection approaches and found that incorporating external data improves the effectiveness of simple traditional filter methods. Integrating external biological data makes these methods compatible with more advanced machine learning techniques, achieving similar classification accuracies while reducing computational running times. The integration process also makes the computation processes more transparent and interpretable.

Perscheid recently comprehensively examined integrative biomarker detection methods that incorporate prior knowledge using gene expression datasets [13]. In her survey article, she critically assessed the distinctive features of various integrative gene selection approaches and provided an overview of external knowledge bases employed in these strategies.

GediNET [14] adopts an integrative approach to discovering disease-disease associations (DDAs) using machine learning techniques. Unlike traditional methods focusing on individual genes, GediNET groups genes based on their disease associations and scores these gene groups to identify significant biomarkers. The tool begins by grouping genes based on their disease associations. These gene groups are then subjected to a scoring component that evaluates their classification significance, allowing the identification of top-performing groups. These high-ranked gene groups are subsequently used to train a machine-learning model. By leveraging this Grouping, Scoring, and Modeling process (G-S-M), GediNET can uncover other diseases with similar associations to the initial disease signature. An enhanced version of it called GediNETPro [15] that utilizes Cross-Validation (CV) information and clustering techniques, such as K-means, to detect patterns of disease group associations. A recent tool, miRdisNET [16], is an integrative computational model for predicting potential miRNA-disease associations in human complex diseases. The tool utilizes miRNA expression profiles and known disease-miRNA associations as input data, employing the G-S-M approach. miRdisNET can predict miRNAs for new diseases and identify disease-disease associations based on shared miRNA knowledge.

maTE [17] is the first tool based on G-S-M that integrates prior-biological knowledge of miRNA target gene and gene expression to identify miRNAs potentially contributing to the disease under study. Interestingly, there is growing evidence of a relationship between miRNAs, their target genes, and diseases. Some miRNAs target specific genes, while others are implicated in disease etiology. However, the intersection between these two sets of genes remains intriguing, as it represents potential key players that link miRNA regulatory networks to disease mechanisms.

In this study, we investigate the common genes associated with miRNAs and diseases. By identifying and examining these shared genes, we aim to shed light on their biological significance, potential functional roles, and relevance in disease pathogenesis. Through an integrated analysis of miRNA-target gene interactions and disease-associated gene lists, we seek to unravel the underlying molecular mechanisms that drive cellular processes and contribute to the development and progression of diseases. The findings from this study will not only enhance our understanding of the complex interplay between miRNAs, their target genes, and diseases and potentially identify novel therapeutic targets or strategies for intervention. Furthermore, by elucidating the biological information of these common genes, we can gain insights into the regulatory networks and pathways that are disrupted in diseases, paving the way for future investigations and advancements in precision medicine.

2. Method

2.1. Grouping-scoring-modeling approach

miRGediNET is a novel approach based on the G-S-M approach.

The Grouping-Scoring-Modeling (G-S-M) approach is a method that combines machine learning and prior biological knowledge to detect the most significant groups of features/genes. It involves grouping and scoring features/genes based on their association with a binary-labeled target, such as control or disease. The unique aspect of this approach is the simultaneous utilization of computational and domain knowledge.

The G-S-M approach employs embedded feature selection, which involves repeatedly using machine learning algorithms to identify the most discriminative groups of features. By incorporating prior domain knowledge, such as miRNA regulation, KEGG pathways, and disease databases, the G-S-M approach aims to gain a more comprehensive understanding of the underlying mechanisms of a biological system, potentially leading to new insights and discoveries.

The primary objective of the G-S-M approach is to generalize its application to any form of prior knowledge that can group measured features. It uses a two-class classification approach, requiring data from two classes (e.g., control and disease) and prior knowledge (e.g., genes associated with a disease) as inputs. Grouping is performed based on the provided prior knowledge, and a scoring process is applied to each group using internal cross-validation and statistical analysis.

Previous implementations of the G-S-M approach, such as CogNet [18], maTE [17], mirCornet [19], miRModuleNet [20], SVM-RCE-R [21], PriPath [22], miRdisNET [16] and GediNET [14], GeNetOntology [23], have served as inspiration for further development. These implementations have considered specific prior knowledge of group genes, such as miRNA-target interactions, disease-gene associations, and KEGG pathways.

2.2. The integration of three pillars of biological knowledge

A revolutionary approach in molecular biology and genetics is signaled by the integration of biological knowledge, with a particular focus on miRNAs, their target genes, and disease correlation. By simultaneously examining miRNAs, target genes, and disease correlations, researchers can comprehensively understand the complex network of interactions and associations. This comprehensive perspective can reveal patterns or connections that might be missed when studying these entities separately. Frequently, diseases result from a series of molecular events involving multiple genes and regulatory elements. Integrating miRNAs with target genes and disease correlation can elucidate these complex mechanisms, providing information on the onset, progression, and potential therapeutic targets.

These three dimensions of biological knowledge can be combined to improve the accuracy and dependability of predictive models. These integrative models are more capable of predicting the course of a disease, the roles of miRNAs, or putative gene targets.

2.2.1. *In this study, we have integrated three pillars of prior-biological knowledge as the following*

1. microRNA and its target genes. The most commonly used database for microRNA and its target gene interactions is miRTarBase [24]. miRTarBase is a comprehensive and manually curated database that provides experimentally validated miRNA-target interactions. It collects data from published articles, and each entry in miRTarBase includes detailed information about the miRNA, target gene, experimental validation method, and the associated literature reference.
2. Disease gene associations: We have used DisGeNET [25]. DisGeNET is a database that focuses on gene-disease associations. It integrates information from multiple sources, including scientific literature, public databases, and expert-curated resources. DisGeNET provides a scoring system to evaluate the strength of the gene-disease associations.
3. miRNA-disease associations: We have used the Human microRNA Disease Database (HMDD), a curated database specifically focusing on miRNA-disease associations [26]. It integrates information from literature mining and manual curation information and provides comprehensive annotations and functional data for miRNA-disease relationships.

miRGediNET aims to explore the role of common or shared genes associated with microRNAs and diseases in the disease understudy. The fact that the common genes are associated with microRNAs and diseases suggests their potential importance in biological processes and disease mechanisms.

The hypotheses that lead us to consider those common genes are: Firstly, microRNAs play a crucial regulatory role by binding to messenger RNA (mRNA) molecules and modulating gene expression. When a microRNA targets a gene, it indicates its involvement in specific cellular processes or disease pathways [27]. The association of these common genes with diseases suggests their potential relevance in disease development, progression, or treatment response. These genes may contribute to disease-associated pathways or manifest the disease phenotype [28]. Thirdly, the consistency of evidence strengthens the functional relevance of these genes. When a gene is independently associated with both microRNAs and diseases, it increases confidence in its importance. The convergence of multiple lines of evidence supports its significance in microRNA regulation and disease biology [29]. Fourthly, genes associated with microRNAs and diseases may be potential therapeutic targets. Modulating their activity or interactions with microRNAs could influence disease-related pathways or restore normal gene expression patterns [30]. Lastly, the research and literature significance of genes extensively studied in the context of microRNAs and diseases contribute to their importance. Experimental evidence, functional

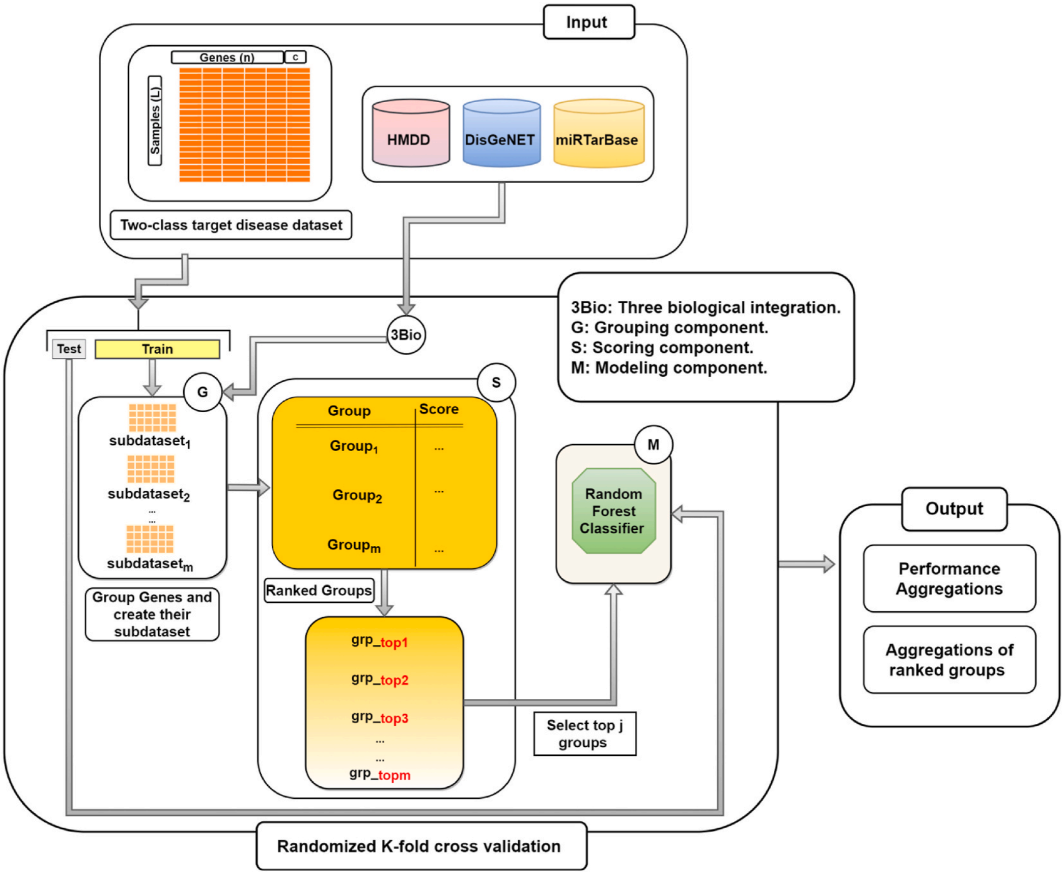

Fig. 1. miRGediNET workflow. The main workflow of 3Bio-G-S-M integrates three pre-existing biological knowledge sources for grouping genes.

studies, and clinical observations support their effectiveness [31]. It is important to note that further investigation, such as experimental validation or functional studies, is necessary to fully understand these common genes' precise roles and significance of these common genes in specific microRNA-mediated regulatory networks and disease contexts. Nonetheless, the association of these genes with both microRNAs and diseases provides valuable insights and suggests their potential importance in biological processes and disease pathogenesis.

In this study, we have used the miRTarBase database for the microRNA target association. Moreover, several computational methods, such as TargetScan [32], and others, developed over the years, each with its own unique approach to predicting miRNA targets.

Among these tools, miRTRS, developed by Jiang et al. [33] stands out due to its innovative approach. Employing a recommendation algorithm, miRTRS can predict targets for new miRNAs using sequence similarity without the typical requirement of selecting negative samples. When benchmarked against other recent methods, miRTRS showcased superior performance, highlighting its potential as a leading tool in the realm of miRNA target prediction.

We utilized the HMDD for microRNA disease associations. However, one could also employ computational tools like DNRLMF-MDA [34] to predict microRNA-disease associations.

2.3. miRGediNET

We have developed a new integrative tool named miRGediNET that integrates prior-biological knowledge from three resources. The main workflow of miRGediNET is illustrated in Fig. 1.

The miRGediNET comprises four main components, namely G, S, M, and 3Bio. The G, S, and M components have been inherited from the previous tool GediNET, while the 3Bio component is a newly introduced addition.

The following is the miRGediNET methodology overview:

1. Data Preparation and Splitting:

The initial step involves partitioning the dataset into two segments: training and testing. This separation ensures that the model is both trained and evaluated on different subsets of data, guaranteeing a robust assessment of its performance.

2. G Component (Grouping):

This phase is about organizing genes into specific groups based on prior-biological knowledge. In miRGediNET, the G component benefits from the 3Bio component, which offers a set of groups connected to common genes linked with microRNAs and disease associations.

The G component extracts specific sub-datasets of genes from the training data, corresponding to each group. The inherent class labels of the samples (positive or negative) are preserved.

3. S Component (Scoring):

The S component steps in to assign scores to the groups. The scores reflect the difference in expression between the two classes (positive and negative) within the sub-datasets. A machine learning approach, particularly the Random Forest algorithm, is applied, incorporating a randomized k-fold cross-validation technique. While accuracy serves as the primary scoring metric in miRGediNET, other metrics like sensitivity, specificity, or AUC can also be applied, depending on the user's preference.

4. M Component (Machine Learning Model):

It considers the top-ranked j groups. Genes from these groups amalgamate to form a set of top-ranked associated genes. Using these genes, a machine learning model, like the Random Forest classifier, is trained. The model's prowess is evaluated using performance statistics, gauging metrics like accuracy, precision, recall, or AUC.

5. Iterative Evaluation:

The entire process, from data splitting to model evaluation, operates within a randomized k-fold cross-validation loop, set to 100 iterations in miRGediNET. This ensures comprehensive and robust evaluations of the tool's performance, enhancing its reliability.

The following describes the G, S, and M components. For more details, we refer to the GediNET study [35]. The first step in the workflow is to split the dataset into two parts, one for training and the second for testing. The G, or grouping component, is responsible for grouping genes based on prior-biological knowledge. In miRGediNET, the G component utilizes the knowledge from the 3Bio component (see below for details of the 3Bio component), which provides a list of groups that are associated with the common genes associated with microRNAs and disease associations. The process within the G component involves extracting a sub-dataset of genes from the training part of the dataset under study associated with each group. This is achieved by selecting the gene columns corresponding to the specific group. The original class labels for the samples represented by the rows (positive or negative) are also retained. The sub-datasets generated by the G component serve as inputs for the following S (Scoring) component, where further group scoring takes place.

The S component, also known as the scoring component, assigns scores to the groups created in the G component. These scores measure the extent to which each group exhibits differential expression between the two classes (positive and negative) within the two-class sub-datasets. The S component utilizes an internal randomized k-fold cross-validation technique [36] based on a machine learning algorithm (such as Random Forest). The average of the accuracy or AUC is computed as the group's score. We have set k to be 5. It is worth noting that while miRGediNET employs accuracy as the scoring measurement, other metrics such as sensitivity, specificity, or area under the curve (AUC) can also be utilized. The choice of measurement can be adjusted based on specific requirements or preferences.

The M component uses the top-ranked j groups. The genes from these top-ranked groups are combined to form a set of top-ranked associated genes.

Once the set of top-ranked associated genes is obtained, a sub-dataset is extracted from the training part (90 % training, 10 % testing). A machine learning model, such as a Random Forest (RF) classifier, is trained. Finally, the trained RF model is evaluated in the testing part. The model's performance statistics are recorded, such as accuracy, precision, recall, or the area under the curve. This evaluation is typically performed for different values of j, assessing the model's performance when considering other numbers of top-ranked associated genes.

The miRGediNET evaluates the performance of the model by splitting the data into training and testing subsets. The workflow uses the randomized k-fold cross-validation loop. We have set k to 100. In each iteration, the input dataset is randomly partitioned into 90 % for training and 10 % for testing.

2.4. *The 3Bio component*

We have considered three biological resources that represent the three biological prior-knowledge summarized in Table 1. The first data is the miRTarBase [24] miRNA-Genes targets. The second is DisGeNET [25] of Disease-Genes associations, while the third is HMDD [26] of miRNA-disease associations.

Table 1 summarizes data from three databases related to miRNA-gene targets, Disease-Genes associations, and miRNA-disease associations. In miRTarBase, there are 2599 unique miRNAs and 15,064 genes, with a total of 502,652 associations between them. DisGeNET focuses on disease-gene associations, containing information on 3929 unique diseases and 15,991 genes, with a total of 329,936 associations between them. HMDD provides data on miRNA-disease associations, including 1207 unique miRNAs and 894 unique diseases, with 35,547 associations between them.

3Bio component uses the three prior-biological knowledge that represent the relationships that define the groups' content. The main relationship that is used for creating the groups is the miRNA-Disease association, as illustrated in Fig. 2. For the miRNA, we retrieve its list of target genes, and for the disease, we retrieve its list of associated genes. The intersection of those two lists is the group that we define. Mathematically, we might express it as:

1. First relationship: Given a specific $miRNA_x$ that is associated with a specific $Disease_y$. We refer to it as a pair of ($miRNA_x$, $Disease_y$).
2. The second relationship: $miRNA_x$ is associated with target genes T_x. Let's say the target genes of $miRNA_x$ are $T_x = [gene_{1x}, gene_{2x}, gene_{3x}, \ldots]$.
3. Third relationship: $Disease_y$ is associated with genes T_y. Let's say the genes associated with $Disease_y$ are $T_y = [gene_{1y}, gene_{2y}, gene_{3y}, \ldots]$.

Step 1. to 3 have created 427 pairs of ($miRNA_x$, $Disease_y$) with T_x genes and T_y genes. We will refer to each group as: pair($miRNA_x$,

Table 1
Summary information about the 3 biological prior-knowledge used in miRGediNET.

miRNA-Genes Targets miRTarBase	#of items
miRNA	2599
Genes	15064
miRNA-Genes	502652
Disease-Genes Associations DisGENET	
Disease	3929
Genes	15991
Disease-Genes	329936
miRNA-Disease Associations HMDD	
miRNA	1207
Disease	894
miRNA-Disease	35547

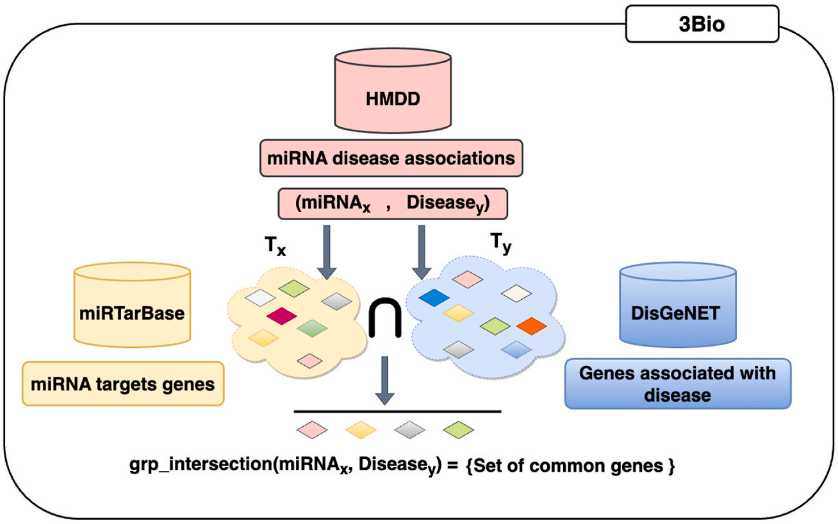
3Bio
HMDD
miRNA disease associations
(miRNA$_x$, Disease$_y$)
T$_x$
T$_y$
miRTarBase
miRNA targets genes
∩
DisGeNET
Genes associated with disease
grp_intersection(miRNA$_x$, Disease$_y$) = {Set of common genes }

Fig. 2. The 3Bio component.

$Disease_y) = \{T_x, T_y\}$

Finally each group is computed as the intersection of T_x and T_y,

grp_intersection($miRNA_x$, $Disease_y$) = $T_x \cap T_y = [gene_{1x}, gene_{2x}, gene_{3x}, \ldots] \cap [gene_{1y}, gene_{2y}, gene_{3y}, \ldots]$

The resulting groups, will contain the genes that are common to both T_x and T_y.

In our study, the 3Bio component serves as a foundational layer responsible for the integration of three distinct biological databases. The primary function of 3Bio is to generate coherent groups based on shared features or associations across these resources. These groups are not just random collections but are carefully curated sets of miRNAs, genes, and diseases that share a contextual relationship.

Once these groups are formed, they are fed into the G component of our methodology.

Table 2 is a summary of the relationship between the three biological resources. Table 2 consists of two columns for the names of miRNAs and diseases. The third column presents the set of target genes corresponding to the miRNA, while the fourth column displays the set of associated genes related to the diseases. The *'Common Genes'* column indicates the intersection between these two sets. Subsequently, three columns provide the respective counts or numbers of genes within each set. In addition, the last column presents the Jaccard similarity score [37] between the target and disease gene sets. The Jaccard similarity coefficient measures the size of the intersection of two sets divided by the size of their union.

Table 2 shows in terms of the number of target genes of miRNAs, a minimum of 32 genes and a maximum of 891 genes. On average, the miRNAs in our study targeted approximately 282 genes. This wide range suggests that different miRNAs may have varying degrees of influence on gene regulation, with some targeting a few specific genes while others have a broader impact.

Regarding the number of diseased genes, we found a minimum of 142 genes and a maximum of 3087 genes. The average count of associated genes of disease was approximately 2000 genes. This variation highlights the complexity and heterogeneity of diseases, with some conditions involving a relatively small number of genes while others exhibit extensive genetic perturbations.

Regarding the number of common genes between miRNA targets and disease-associated genes, we observed a minimum of 4 genes and a maximum of 77 genes. On average, there were around 26 common genes identified. Common genes suggest potential regulatory relationships between miRNAs and disease-associated genes, indicating their potential roles in disease pathogenesis and progression.

Finally, Jaccard similarity scores range from 0.001 to 0.024. The average similarity score across the common genes was approximately 0.009. These scores represent the degree of overlap between the target gene set and the disease gene set, with higher values indicating a greater similarity. The variations in similarity scores reflect the diverse relationships between miRNA targets and disease-associated genes, emphasizing the unique regulatory mechanisms underlying different diseases.

2.5. Data

We extracted from the GEO database [38] ten human gene expression datasets related to diverse and complex diseases. Each dataset contains its GEO accession, the name of the disease, the number of samples, and the availability of both positive and negative samples. Table 3 provides a comprehensive description of the ten datasets listed.

2.6. Evaluation

The miRGediNET was evaluated on the ten GEO datasets. The evaluation procedure involved the application of a random split to

Table 2
Intersecting Genes between miRNA Targets and Disease-Associated Genes: Common Genes and Counts.

miRNA	Disease	Targeted Genes of miRNA	Associated Genes of the Disease	Common_genes	# Taget_genes	#Diseased_genes	#Common_genes	Similarity Score
hsa-mir-575	leukemia	ccdc69,ifnlr1,ppp1r3b, wipi2,klf6,znf677,	dpb1,egln1,h4c15,mir375, mllt11,gsn,xpr1,	xiap,usp1,ccna2,plcg2,dusp2,rad51, hdac7,zc3h12a,rbfox2,tsg101,mef2d, bcl2l1	130	2111	12	0.00538358
hsa-mir-107	melanoma	ppil1,sun2,ei24,add2,ago3, cdc42se2	ifnl1,unc5c,sammson, lncr1, kctd12,stk3,siglec9,	hmga2,arnt,jak1,mdm4,rad51,fbxw7, xpc,oprm1,itga2,slc2a3,	300	3087	77	0.02326284
hsa-mir-217	melanoma	snrnp27,stk40,kras, trappc2b,flvcr1,dnal1, rab11b,	,tas2r60,rad51,f9,tlr7,mir373, opn1lw,	hif1a,tmsb4x,sirt1,smad7,tp53inp1, ezh2,ubxn11,kras,	80	3087	22	0.006995231
hsa-mir-621	melanoma	pabpc1l2a,dad1,klhl15, togaram2,oip5,slc25a46,	a6,irf3,cdh2,raly,mir199a2, anxa6,thrb, tyro3,sipa1,	foxo3,pawr, rora,cdk6	32	3087	4	0.001284109
hsa-mir-646	melanoma	dynll2,dmtf1,cdc42se2, golga8a,	sox5,mak16,gata3,meis1, ywhazp5,itpr3,phf20,inpp4b,	chek1,hspa8,dctn5,wee1,crk,slit3,tlk1, cul3,sparc, lamtor1,decr1,	251	3087	49	0.014898145
hsa-mir-638	leukemia	cnnm4,cnga2,pld1,dctn5, nr4a3,ldha,	ddx4,bcl10,ldha, mrtfa,trim28, prss1,ikbke, ctnnb1,tox,gpt, epc2	tmed2,bub1,nr4a3,ldha,sox2,pten, oscp1,hoxb6,sod2,cdk2,tp53	53	2111	11	0.00510915
hsa-mir-484	early-stage breast carcinoma	,trim65,kiaa0895,acta2, pgd,mgat5,brca1,puf60, hars2,	thop1,esr1,tacc3,itgal, hpse, il17rb,cntn3,ncoa3,	thop1,brca1,mdm2,slc6a8,col18a1, cdc25a,ezh2,fasn, fkbp4	891	142	9	0.008789063

Table 3
Description of the 10 datasets used in the study. Each entry has the GEO accession, the name of the disease, the number of samples, and the data classes.

GEO accession	Title	Disease	#Samples Classes
GDS1962	Glioma-derived stem cell factor effect on angiogenesis in the brain	Glioma	180 negative = 23 positive = 157
GDS2545	Metastatic prostate cancer (HG-U95A)	Prostate cancer	171 negative = 81 positive = 90
GDS2771	Large airway epithelial cells from cigarette smokers with suspect lung cancer	Lung cancer	192 negative = 90 positive = 102
GDS3257	Cigarette smoking effect on lung adenocarcinoma	Lung adenocarcinoma	107 negative = 49 positive = 58
GDS4206	Pediatric acute leukemia patients with early relapse: white blood cells	Leukemia	197 negative = 157 positive = 40
GDS5499	Pulmonary hypertension: PBMCs	Pulmonary hypertension	140 negative = 41 positive = 99
GDS3837	Non-small cell lung carcinoma in female nonsmokers	Lung cancer	120 negative = 60 positive = 60
GDS4516_4718	Colorectal cancer: laser microdissected tumor tissues	Colorectal cancer	148 negative = 44 positive = 104
GDS2547	Metastatic prostate cancer (HG-U95C)	Prostate cancer	164 negative = 75 positive = 89
GDS3268	Colon epithelial biopsies of ulcerative colitis patients	Colitis	202 negative = 73 positive = 129

each dataset, dividing the data into training and testing sets. The split ratio was 90 % for training and 10 % for testing. This process was repeated 100 iterations to ensure robustness and obtain reliable results. This kind of cross-validations is named the randomized k-fold cross-validation technique [36].

By accumulating the results from the randomized k-fold cross-validation, we calculated the mean and standard deviation for all the following performance metrics. These metrics include Accuracy (Acc), Sensitivity (Sen), Specificity (Spe), Precision (Prec), and F-measure (F-m) [39]. The mean values provide an estimate of the central tendency of the performance metrics, while the standard deviation indicates the variability across the repeated evaluations. After generating lists of miRNA disease groups and their associated genes in each iteration, there was a need to prioritize them due to slight variations. To address this, we employed a rank aggregation approach, explicitly utilizing the RobustRankAggreg R package developed by Kolde et al. [40]. The RobustRankAggreg package was embedded into the miRGediNET workflow, enabling the prioritization of the aggregated lists. This approach assigns a P-value to each element in the aggregated list, indicating how well each element/entity was ranked compared to the expected value. The P-value provides a measure of the significance or reliability of the ranking for each element, aiding in identifying the most relevant genes associated with the disease groups.

3. Results

Table 4 displays the performance of miRGediNET on 10 datasets, explicitly focusing on the top 2 groups. The values presented in Table 4 are the average results obtained from 100 iterations using a randomized 100-fold cross-validation loop. The performance metrics are reported in terms of the area under the curve (AUC), accuracy (ACC), sensitivity (SEN), specificity (SPE), and F-measure (F-m). The first column of the table represents the GEO accession, while the second column indicates the number of groups. The column labeled "#Genes" provides the number of genes.

Notably, the dataset GDS4206 showed only one unsuccessful result. However, it is worth mentioning that this particular dataset yielded similar observations when other tools, such as CogNet [18], matE [17], PriPath [22], and GediNET [14], were applied to it.

Table 4
Summary of performance metrics for gene expression datasets in disease classification: Accuracy (Acc), sensitivity (Sen), specificity (Spe), precision (Prec), and F-measure (F-m)".

GEO accession	#G	#Genes	AUC	Acc	Sen	Spe	Prec	F-m
GDS1962	2	7.89	0.966	0.906	0.93	0.845	0.945	0.932
GDS2545	2	20.05	0.785	0.707	0.677	0.74	0.761	0.706
GDS2547	2	10.12	0.767	0.695	0.721	0.668	0.697	0.699
GDS2771	2	16.95	0.712	0.669	0.703	0.632	0.69	0.688
GDS3257	2	13.25	0.998	0.979	0.994	0.964	0.971	0.981
GDS3268	2	8.55	0.661	0.607	0.604	0.61	0.642	0.619
GDS3837	2	8.58	0.948	0.884	0.907	0.862	0.882	0.886
GDS4206	2	5.92	0.52	0.608	0.245	0.77	0.325	0.362
GDS4516	2	8.07	0.998	0.987	0.986	0.99	0.995	0.99
GDS5499	2	6.63	0.897	0.858	0.912	0.735	0.892	0.899

This suggests that the challenges or limitations encountered with GDS4206 may not be specific to miRGediNET alone but extend to other analysis tools.

Table 5 shows an example of the miRGediNET performance table. The values in the table are the average measurements performance over randomized 100-fold cross-validation, for the aggregated top-ranked 10 groups in the GDS3257 dataset. The first row represents the performance of the first top-ranked group (#Groups = 1), which achieved a remarkable AUC of 98 % with an average of 7.73 genes. The subsequent row (#Groups = 2) presents the performance metrics when combining the genes from the first and second top-ranked groups.

Table 6 is an example of the top-ranked groups of miRGediNET, along with their assigned P-values obtained through robust rank aggregation. This table plays a crucial role in the biological interpretation section, allowing for an in-depth analysis of the relationship between miRNA-disease associations and their target genes. Specifically, Table 6 focuses on the top-ranked groups, such as "HSA-MIR-1297_COLORECTAL CARCINOMA" and "HSA-MIR-498_COLORECTAL CARCINOMA," among others. It raises an important biological question regarding the association between these top-ranked groups and the disease being investigated in the dataset, which in this case is Lung Cancer within the dataset GDS3837. These results provide valuable insights into the potential involvement of these miRNAs in the development or progression of Lung Cancer, thereby enabling further exploration and analysis of their associated target genes.

miRGediNET also generates a list of significant genes aggregated using the Robust Rank Aggregation tool. During the scoring process for each group, the genes associated with the group receive the same score as the group itself. This aggregated list combines the scores of the genes to create a comprehensive list of significant genes. Table 7 is an example of significant genes for the GDS3837 dataset.

3.1. Comparison with other similar tools

In our study, we compared three computational tools, each based on the G-S-M model, that are designed to analyze the relationships among miRNAs, genes, and diseases. These tools are miRGediNET (miRNA and genes), maTE (miRNA and genes), and GediNET (disease and genes).

All three tools were assessed using ten GEO datasets (as detailed in Table 3). To ensure robustness and reliable results, we employed the randomized k-fold cross-validation technique [28]. Specifically, each dataset was divided into a training set (90 %) and a testing set (10 %). This partitioning was iteratively repeated 100 times. Based on the outcomes of these iterations, we calculated the mean for several performance metrics, including accuracy, Sensitivity, specificity, precision, F-measure, and AUC [39].

The summarized outcomes for the three tools over the ten datasets, specifically for the top 2 groups, are presented in Table 8.

As observed from Table 8, GediNET boasts the highest mean AUC at 0.85 for the top 2 groups. miRGediNET and maTE have closely matched mean AUC values at 0.83 and 0.82, respectively, when considering the top 2 groups.

When considering the number of genes, GediNET considers a substantially larger number of genes (77.51 on average) in its analyses for the top 2 groups. In contrast, miRGediNET and maTE focus on a more concise gene set with averages of 10.60 and 7.61 genes, respectively, for the top 2 groups.

3.2. Biological validation of miRGediNET

For this section, we have used the Cancer Genome Atlas - Breast Invasive Carcinoma (TCGA-BRCA) [41] dataset on the Genomic Data Commons hosted by the National Cancer Institute. miRGediNET was applied, and results were obtained. We focused on gene expression (mRNA) datasets with reads mapped to GRCh38, which were downloaded from Xena Public Data Hubs for their analysis [42].

The molecular intrinsic subtype classes in the BRCA dataset were determined using the PAM50 assay (Prediction Analysis of Microarray 50). This assay is based on 50 gene signatures and helps classify samples into different molecular subtypes [43].

The downloaded tumor samples were filtered based on molecular subtypes, specifically Luminal A, Luminal B, Her2-enriched, and Basal-like. The samples were divided into two categories: 248 samples with ER+/PR+/− (Luminal label, comprising Luminal A and

Table 5
An example of Cumulative Performance of miRGediNET for the Top 10 Gene Groups in the GDS3257 Dataset - Averaged Over 100 Iterations.

#Groups	#Genes (Mean)	Sensitivity (Mean)	Specifity (Mean)	Precision (Mean)	Accuracy (Mean)	Area Under Curve (Mean)	F-measure (Mean)
1	7.73	0.978	0.958	0.966	0.968	0.989	0.969
2	13.25	0.994	0.964	0.971	0.979	0.998	0.981
3	17.6	0.998	0.962	0.969	0.98	0.999	0.982
4	22.14	0.992	0.96	0.968	0.976	0.999	0.978
5	26.07	0.994	0.96	0.968	0.977	0.999	0.979
6	29.44	0.996	0.956	0.965	0.976	0.999	0.978
7	32.74	1	0.962	0.969	0.981	0.998	0.983
8	36.07	1	0.966	0.972	0.983	0.999	0.985
9	39.68	1	0.966	0.972	0.983	0.999	0.985
10	42.64	1	0.962	0.971	0.981	0.999	0.983

Table 6
An Example of an output of the RobustRankAggreg tool for top significant groups for GDS3837 dataset.

Groups	p-value	List of genes	#Genes
HSA-MIR-1297_COLORECTAL CARCINOMA	7.62747E-71	E2F7, MAD2L1, C3, RGS17 …	7
HSA-MIR-498_COLORECTAL CARCINOMA	1.24786E-59	CDK6, GDE1(, EPHA4, CENPI …	9
HSA-MIR-320A_COLORECTAL CARCINOMA	3.5021E-57	MCM4, CLDN12, ZIC2 …	18
HSA-MIR-1258_COLORECTAL CARCINOMA	5.61299E-53	TRIM2, E2F8, CPM, IPP.	4
HSA-MIR-3142_COLORECTAL CARCINOMA	6.42081E-53	GJB2, GSTO2	2
HSA-MIR-4478_COLORECTAL CARCINOMA	1.68725E-52	BRIP1, PDPN, GDE1, KNL1 …	8
HSA-MIR-603_COLORECTAL CARCINOMA	1.90478E-49	CDK6, GREM1, TBX4 …	12
HSA-MIR-429_COLORECTAL CARCINOMA	1.08304E-47	DLC1, ZEB2 SHCBP1 …	6
HSA-MIR-375_COLORECTAL CARCINOMA	8.40599E-46	COL12A1, ELAVL2, CDKN2B …	16
HSA-MIR-429_GLIOBLASTOMA	1.86825E-45	DLC1, ZEB2, RBBP4 …	6
HSA-MIR-320A_MELANOMA	7.9096E-44	CDK6, AQP4, EPHA4 …	11

Table 7
Displays the top 10 significant genes aggregated using the RobustRankAggreg tool for the GDS3837 dataset.

Gene	The Group	Score
E2F7	HSA-MIR-1297_COLORECTAL CARCINOMA	3.26757E-46
MAD2L1	HSA-MIR-1297_COLORECTAL CARCINOMA	5.39518E-45
C3	HSA-MIR-1297_COLORECTAL CARCINOMA	7.61581E-45
RGS17	HSA-MIR-1297_COLORECTAL CARCINOMA	3.53472E-44
GPT2	HSA-MIR-320A_COLORECTAL CARCINOMA	2.07603E-36
CENPI	HSA-MIR-498_COLORECTAL CARCINOMA	1.4539E-35
ZIC2	HSA-MIR-320A_COLORECTAL CARCINOMA	3.20091E-35
RGS17	HSA-MIR-498_COLORECTAL CARCINOMA	4.00696E-35
CLDN12	HSA-MIR-320A_COLORECTAL CARCINOMA	6.61018E-34

Table 8
The average results of all 10 datasets over the top 2 groups.

	miRGediNET	maTE	GediNET
AUC	0.83	0.82	0.85
#Genes	10.60	7.61	77.51

Luminal B subtypes) and 124 samples with ER-/PR- (ER-negative label, comprising Her2-enriched and Basal-like subtypes), excluding the normal-like subtype. However, the focus of further analysis in the study was on the molecular subtype pair BRCA LumAB_Her2-Basal, with Luminal A and Luminal B samples considered positive (pos: LumAB) and Her2-enriched and Basal-like samples considered negative (neg: Her2Basal). The necessary data preprocessing steps were completed, where the raw TCGA gene expression counts were downloaded and normalized using edgeR with the trimmed mean of M-values (TMM) method [44].

Table 9 shows the performance results obtained by miRGediNET.

Table 9 displays the aggregated performance of the top 10 ranked groups in the LumAB_Her2Basal dataset using miRGediNET. The performance table represents the average of randomized 100-fold cross-validation. In the first row, the performance of the highest-ranked group is presented, showing an AUC of 96 % with an average of 3.32 genes. The second row shows the performance metrics for the top 2 groups, where the genes from both the highest-ranked group and the second-highest-scoring group are combined. In other words, miRGediNET provides cumulative performance results for the top 10 groups.

Table 9
The averages of randomized 100-fold cross-validation performance table of miRGediNET for top-ranked 10 groups for Breast Cancer LumAB_-Her2Basal dataset cumulatively.

#Groups	#Features (Mean)	Sensitivity (Mean)	Specificity (Mean)	Precision (Mean)	Accuracy (Mean)	Area Under Curve (Mean)	F-measure (Mean)
1	3.32	0.982	0.93	0.965	0.964	0.975	0.973
2	4.47	0.986	0.932	0.966	0.968	0.978	0.976
3	5.41	0.988	0.936	0.968	0.97	0.979	0.978
4	5.8	0.988	0.937	0.969	0.97	0.982	0.978
5	6.3	0.988	0.937	0.969	0.97	0.982	0.978
6	6.74	0.99	0.936	0.968	0.972	0.983	0.979
7	7.33	0.991	0.935	0.968	0.972	0.983	0.979
8	8.15	0.992	0.937	0.969	0.973	0.984	0.98
9	8.53	0.992	0.938	0.969	0.974	0.984	0.98
10	9.57	0.994	0.938	0.97	0.975	0.985	0.981

Our analysis aims to show that integrating prior biological knowledge might provide valuable information about the pathogenesis of a specific disease or group of related diseases and identify new biomarkers with diagnostic, prognostic, or therapeutic value. To show this, we focus on miRNA-206, which has been ranked first among the significant miRNAs involved in the pathogenesis of different types of cancer, as shown in Table 10. MiRNA-206 is a tumor suppressor miRNA. It's well established that this miRNA suppresses the cancer progression of different types of tumors by targeting many different genes [45]. Interestingly, our miRNA-target interactions and disease associations proved the association between different miRNA-206 target genes previously known to play roles in breast cancer tumorigenesis. This, of course, proves the validity of our approach to studying disease associations with gene expression patterns. Examples of these genes include but are limited to, ESR1, SFRP1, VEGFA, STC2, CREB3L2, MET, PAX3, ANP32B, G6PD, and TBX3. Moreover, our association's study identified new potential breast cancer genes (like RPS7 and GALNT4) not functionally shown before to play a role in breast tumorigenesis. Notably, these new genes were shown to play roles in other cancer types but not breast cancer [46]. miRNA-206 targets genes involved in breast cancer progression to more advanced stages [45]. In fact, our data support this notion by revealing genes that are known to support breast cancer progression. Our finding further supports that miRNA-206 expression is more down-regulated in more advanced breast cancer subtypes than less advanced luminal A and B subtypes, as shown in Table 10.

Examples of other miRNAs revealed by our tool miRGediNET and thus predicted to be important in cancer biology are miRNA-3666 and miRNA-520B. MiRNA-3666 was shown before to inhibit the progression of different types of tumors, including; lung, colorectal, cervical, thyroid, breast, and other types of cancer [47–50].

Also, miRNA-520B was shown before to play different roles in different types of cancer, including colorectal, head and neck, liver, breast, and other cancers [51], [[52] p. 4], [53,54]. Our data confirm that our model can discover new genes important in tumor biology. This was partly proved by confirming the role of genes that are previously known to have important roles in carcinogenesis.

3.3. Enrichment analysis

Table 10 exhibits one of the outputs of miRGediNET, presenting a collection of top significant groups that were scored using the S component and then ranked using RobustRankAggreg, based on their P-values. The LumAB_Her2Basal dataset was considered in this analysis, and the top 10 groups and their respective sets of genes were selected for enrichment analysis. Among these top 10 groups, a total of 60 genes were identified. Notably, there are 20 distinct genes, meaning that some genes are shared among multiple groups.

To apply the enrichment analysis, we uploaded the 20 distinct genes on the Enrichr-KG [55] application. It is an advanced web-server application and knowledge graph database that enhances gene set enrichment analysis capabilities offered by Enrichr [56]. By integrating specific gene set libraries from Enrichr, the tool facilitates integrative enrichment analysis and visualization. The tool employs a bipartite graph representation, connecting genes to their corresponding annotation terms, bridging results across multiple gene set libraries and presenting an integrated network of enriched terms associated with overlapping genes.

For our analysis, we carefully selected four distinct libraries: DisGeNET [25], Reactome 2022 [57], WikiPathway 2021 Human [58], and KEGG 2021 Human [59,60]. Each library contains a curated collection of gene sets. To ensure focused and relevant results, we included the top five terms from each library. Moreover, we applied filtering criteria to enhance the quality of the analysis. Specifically, we filtered based on the minimum number of libraries per gene, a minimum number of links per gene, and a minimum number of links per term, setting all three thresholds to two. This filtering process helps refine the enrichment analysis and ensures that only significant and well-supported associations are considered.

The subnetwork shows the following associations: From KEGG: The gene products IRF1, ESR1, and ESR2 are members of the KEGG pathway Prolactin signaling pathway. The gene products ESR1, and ESR2 are members of the KEGG pathway Estrogen signaling pathway. The gene products VEGFA, ESR1, and ESR2 are members of the KEGG pathway Chemical carcinogenesis. The gene products DAPK1, and VEGFA are members of the KEGG pathway Bladder cancer. The gene products MET, DAPK1, VEGFA, ESR1, and ESR2 are members of the KEGG pathway Pathways in cancer.

From Reactome: The genes VEGFA, and ESR1 are members of the Reactome pathway TFAP2 (AP-2) Family Regulates Transcription Of Growth Factors And Their Receptors R-HSA-8866910. The genes ESR1, and ESR2 are members of the Reactome pathway Extranuclear Estrogen Signaling R-HSA-9009391. The genes ESR1, and ESR2 are members of the Reactome pathway Constitutive

Table 10
The output of the RobustRankAggreg tool for top significant groups for the Breast Cancer LumAB_Her2Basal dataset.

Group	p-value	Unique Genes	#Genes
HSA-MIR-206_GLIOBLASTOMA	2.6E-121	ESR1, SFRP1, VEGFA, STC2, CREB3L2, MET, PAX3, ANP32B	8
HSA-MIR-206_ASTROCYTOMA	6.8E-119	ESR1, SFRP1, VEGFA, MET, PAX3	5
HSA-MIR-206_COLORECTAL CARCINOMA	7.2E-118	ESR1, SFRP1, VEGFA), STC2, G6PD, MET, PAX3, ANP32B, TBX3	9
HSA-MIR-3666_COLORECTAL CARCINOMA	2.5E-116	ACSL4, ESR1, ZIC5, NFIB, FOXQ1, IRF1, MASTL, DAPK1, GALNT4	9
HSA-MIR-206_MEDULLOBLASTOMA	2.3E-114	ESR1, SFRP1, VEGFA, MET, PAX3	5
HSA-MIR-206_OSTEOSARCOMA	2.8E-114	SR1, SFRP1, VEGFA, MET, PAX3	5
HSA-MIR-206_MELANOMA	1.2E-113	ESR1, SFRP1, VEGFA, G6PD, MET, PAX3, TBX3	7
HSA-MIR-206_CHONDROSARCOMA	4E-112	ESR1, VEGFA, MET, TBX3	4
HSA-MIR-520B_PITUITARY ADENOMA	5.5E-110	ESR1, ESR2	2
HSA-MIR-206_RHABDOMYOSARCOMA	3.3E-108	ESR1, VEGFA, G6PD, MET, PAX3, TBX3	6
HSA-MIR-206_PROLACTINOMA	3.1E-105	ESR1, VEGFA	2

Signaling By Aberrant PI3K In Cancer R-HSA-2219530. The genes ESR1, and ESR2 are members of the Reactome pathway Nuclear Receptor Transcription Pathway R-HSA-383280. The genes VEGFA, and ESR1 are members of the Reactome pathway Transcriptional Regulation By AP-2 (TFAP2) Family Of Transcription Factors R-HSA-8864260.

From WikiPathways: The genes ESR1, and ESR2 are members of the WikiPathway Nuclear receptors WP170. The ESR1 and ESR2 genes are members of the WikiPathway Mammary gland development pathway - Pregnancy and lactation (Stage 3 of 4) WP2817. The genes VEGFA and ESR1 are WikiPathway Aryl Hydrocarbon Receptor Netpath WP2586 members. The genes DAPK1, and VEGFA are members of the WikiPathway Bladder cancer WP2828. The genes PAX3 and SFRP1 are WikiPathway Endoderm differentiation WP2853 members.

From DisGeNET: The disease Medulloblastoma is associated with the following genes: MET, PAX3, IRF1, DAPK1, VEGFA, ESR1, ESR2, and SFRP1. The disease Solid Neoplasm is associated with the following genes: MET, PAX3, IRF1, VEGFA, ESR1, and SFRP1. The disease Triple Negative Breast Neoplasms is associated with the following genes: MET, VEGFA, ESR1, ESR2, and SFRP1. The disease of Stomach Carcinoma is associated with the following genes: MET, PAX3, IRF1, DAPK1, VEGFA, ESR1, ESR2, and SFRP1. The disease Malignant neoplasm of the stomach is associated with the following genes: MET, PAX3, IRF1, DAPK1, VEGFA, ESR1, ESR2, and SFRP1.

To obtain more focused results on disease-genes associations, we have selected only the DisGeNET library and set the threshold to include the top 10 diseases. The output subnetwork is presented in Fig. 4.

The subnetwork shows that those genes are significantly associated with different diseases; however, our interest is in diseases related to breast cancer. The tool has identified two breast cancer diseases, namely Triple Negative Breast Neoplasms, and invasive Carcinoma of the breast, as being associated with the discovered genes. Interestingly, these two breast cancer diseases are among the ten diseases associated with the identified genes.

4. Discussion and conclusion

This study introduced a novel approach named miRGediNET, which integrates three prior biological knowledge-based machine-learning techniques. We aim to explore the intersection of genes associated with microRNAs and diseases, as this intersection represents potential key players in linking miRNA regulatory networks to disease mechanisms. To achieve this, we utilized three important resources: miRTarBase, DisGeNET, and the Human microRNA Disease Database (HMDD). miRTarBase provided information about the association between microRNAs and their target genes. At the same time, DisGeNET and HMDD supplied us with data on disease-gene associations and miRNA-disease associations, respectively. By integrating these three sources of biological knowledge, we could identify and examine the shared genes associated with miRNAs and diseases. Our analysis focused on the biological significance, potential functional roles, and relevance of these shared genes in disease pathogenesis.

By carefully grouping genes based on their miRNA-disease associations, we scored these groups in terms of their classification significance. This scoring allowed us to train our machine-learning model. The significance of our approach lies in the fact that the identified common genes are associated with both microRNAs and diseases. This implies their potential importance in biological processes and disease mechanisms. Our study supports the growing evidence supporting the relationship between miRNAs, their target genes, and diseases.

Moreover, our approach goes beyond the traditional method of searching for biomarker genes. Instead, we considered prior gene knowledge, leveraging the associations between miRNAs, target genes, and diseases. This integrative strategy allows us to uncover hidden connections and provides a more comprehensive understanding of the complex interactions occurring within biological systems.

While several tools and methods have previously explored miRNA-target interactions and their association with diseases, miRGediNET distinguishes itself in several key aspects. Unlike many existing tools that focus on specific subsets of miRNA-gene interactions or disease associations, miRGediNET provides a comprehensive analysis, capturing a broader spectrum of interactions and associations. This unique approach allows for a more structured and detailed analysis. By grouping relevant genes and scoring them based on specific criteria, it ensures a more accurate representation of interactions. The modeling aspect further refines this data, offering predictive insights that can guide future research. Many tools are limited by their reliance on a single database. In contrast, miRGediNET integrates data from three distinct biological knowledge bases, ensuring a richer and more holistic dataset for analysis.

With its implementation in Knime and availability on GitHub, miRGediNET is not just a research tool but a collaborative platform. This open-access nature encourages community engagement and collective advancement in the field.

In essence, the innovation of miRGediNET lies not just in its methodological approach but in its comprehensive, integrative, and collaborative nature, pushing the boundaries of current research on miRNA-disease associations.

In conclusion, miRGediNET represents a valuable tool for exploring the significance of integrating more than one prior-biological knowledge associated with microRNAs and diseases. By incorporating prior biological knowledge, we have demonstrated the potential of this approach to unravel key players in disease mechanisms and shed light on the underlying molecular processes. Our findings contribute to the field of miRNA research and may pave the way for the development of targeted therapies and improved disease management strategies.

Data availability statements

All data supporting the findings of this study are publicly available. The datasets were obtained from The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). Specific datasets and accessions used in this study can be accessed directly from the respective repositories.

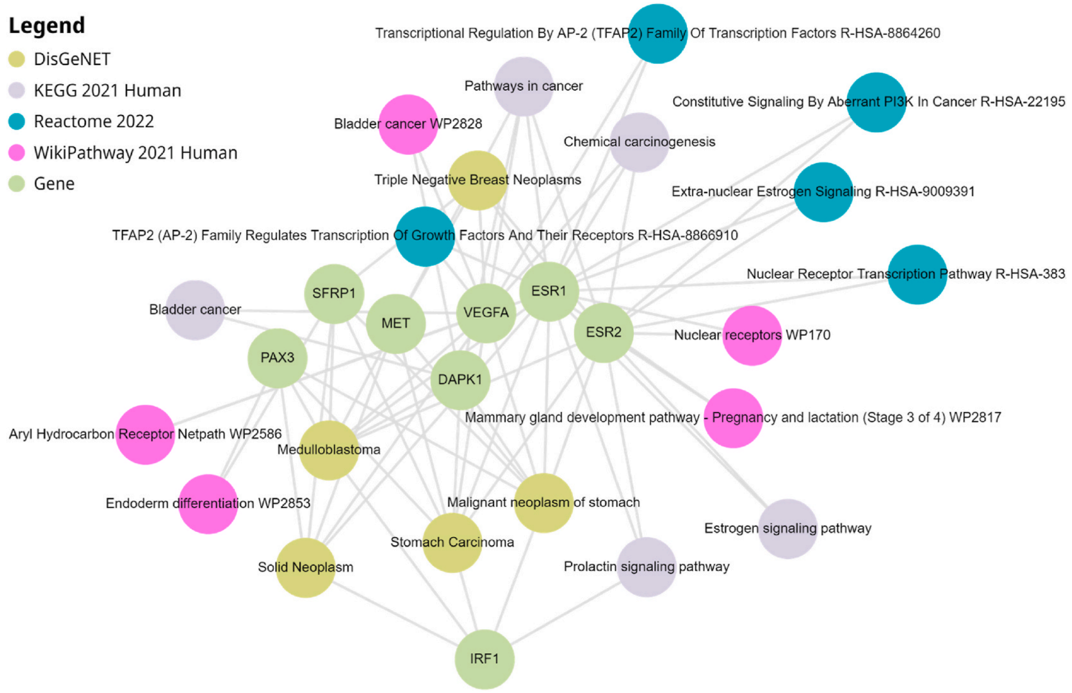
Legend
DisGeNET
KEGG 2021 Human
Reactome 2022
WikiPathway 2021 Human
Gene
Transcriptional Regulation By AP-2 (TFAP2) Family Of Transcription Factors R-HSA-8864260
Pathways in cancer
Constitutive Signaling By Aberrant PI3K In Cancer R-HSA-22195
Bladder cancer WP2828
Chemical carcinogenesis
Triple Negative Breast Neoplasms
Extra-nuclear Estrogen Signaling R-HSA-9009391
TFAP2 (AP-2) Family Regulates Transcription Of Growth Factors And Their Receptors R-HSA-8866910
Nuclear Receptor Transcription Pathway R-HSA-383
SFRP1
ESR1
Bladder cancer
MET
VEGFA
ESR2
Nuclear receptors WP170
PAX3
DAPK1
Mammary gland development pathway - Pregnancy and lactation (Stage 3 of 4) WP2817
Aryl Hydrocarbon Receptor Netpath WP2586
Medulloblastoma
Endoderm differentiation WP2853
Malignant neoplasm of stomach
Estrogen signaling pathway
Stomach Carcinoma
Solid Neoplasm
Prolactin signaling pathway
IRF1

Fig. 3. The subnetwork is associated with 4 resources.

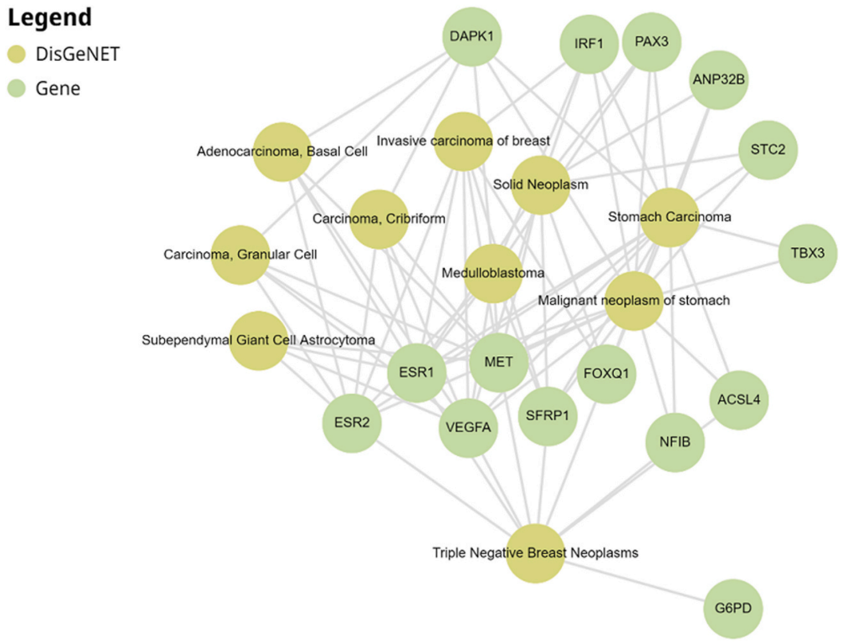

Fig. 4. The subnetwork of only the Disease-Genes associations.

CRediT authorship contribution statement

Emma Qumsiyeh: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Zaidoun Salah:** Formal analysis. **Malik Yousef:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

List of abbreviations

Abbreviations	*Full Form / Explanation*
ACC	Accuracy
AUC	Area under the curve
CV	Cross Validation
GEO	Gene Expression Omnibus
HMDD	Human microRNA Disease Database
RF	Random Forest
SEN	Sensitivity
SPE	Specificity
TCGA	The Cancer Genome Atlas
TMM	Trimmed Mean of M-values
BRCA	Breast Invasive Carcinoma
SVM	Support Vector Machines
GNN	Graph Neural Networks
F-m	F-measure
Prec	Precision

References

[1] C.E. Condrat, et al., miRNAs as biomarkers in disease: latest findings regarding their role in diagnosis and prognosis, Cells 9 (2) (Jan. 2020) 276, https://doi.org/10.3390/cells9020276.

[2] T. Rashid, et al., DEEPMIR: a deep neural network for differential detection of cerebral microbleeds and iron deposits in MRI, Sci. Rep. 11 (1) (Jul. 2021), 14124, https://doi.org/10.1038/s41598-021-93427-x.

[3] D. Ouyang, et al., Predicting multiple types of associations between miRNAs and diseases based on graph regularized weighted tensor decomposition, Front. Bioeng. Biotechnol. 10 (Jul. 2022), 911769, https://doi.org/10.3389/fbioe.2022.911769.

[4] R. Bellazzi, B. Zupan, Towards knowledge-based gene expression data mining, J. Biomed. Inf. 40 (6) (Dec. 2007) 787–802, https://doi.org/10.1016/j.jbi.2007.06.005.

[5] W. Lan, J. Wang, M. Li, J. Liu, F.-X. Wu, Y. Pan, Predicting MicroRNA-disease associations based on improved MicroRNA and disease similarities, IEEE ACM Trans. Comput. Biol. Bioinf 15 (6) (Nov. 2018) 1774–1782, https://doi.org/10.1109/TCBB.2016.2586190.

[6] R. Kustra, A. Zagdanski, Incorporating Gene Ontology in Clustering Gene Expression Data, vol. 2006, 2006, p. 563, https://doi.org/10.1109/CBMS.2006.100.

[7] M. Yousef, A. Kumar, B. Bakir-Gungor, Application of biological domain knowledge based feature selection on gene expression data, Entropy Basel Switz 23 (1) (Dec. 2020) E2, https://doi.org/10.3390/e23010002.

[8] C. Perscheid, B. Grasnick, M. Uflacker, Integrative gene selection on gene expression data: providing biological context to traditional approaches, J. Integr. Bioinforma. 16 (1) (Feb. 2019), https://doi.org/10.1515/jib-2018-0064.

[9] C. Kuzudisli, B. Bakir-Gungor, N. Bulut, B. Qaqish, M. Yousef, Review of Feature Selection Approaches Based on Grouping of Features, PeerJ, 2023.

[10] M. Yousef, F. Ozdemir, A. Jaber, J. Allmer, B. Bakir-Gungor, PriPath: identifying dysregulated pathways from differential gene expression via grouping, scoring, and modeling with an embedded feature selection approach, BMC Bioinf. 24 (1) (Feb. 2023) 60, https://doi.org/10.1186/s12859-023-05187-2.

[11] V. Gligorijević, N. Pržulj, Methods for biological data integration: perspectives and challenges, J. R. Soc. Interface 12 (112) (Nov. 2015), 20150571, https://doi.org/10.1098/rsif.2015.0571.

[12] V.K. Raghu, X. Ge, P.K. Chrysanthis, P.V. Benos, Integrated theory-and data-driven feature selection in gene expression data analysis, in: 2017 IEEE 33rd International Conference on Data Engineering (ICDE), IEEE, San Diego, CA, USA, Apr. 2017, pp. 1525–1532, https://doi.org/10.1109/ICDE.2017.223.

[13] C. Perscheid, Integrative biomarker detection on high-dimensional gene expression data sets: a survey on prior knowledge approaches, Briefings Bioinf. 22 (3) (May 2021) bbaa151, https://doi.org/10.1093/bib/bbaa151.

[14] E. Qumsiyeh, L. Showe, M. Yousef, GediNET for discovering gene associations across diseases using knowledge based machine learning approach, Sci. Rep. 12 (1) (Nov. 2022), https://doi.org/10.1038/s41598-022-24421-0. Art. no. 1.

[15] E. Qumsiyeh, M. Yazıcı, M. Yousef, GediNETPro: discovering patterns of disease groups, in: Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies - BIOINFORMATICS, SciTePress, 2023, pp. 195–203, https://doi.org/10.5220/0011690800003414.

[16] A. Jabeer, M. Temiz, B. Bakir-Gungor, M. Yousef, miRdisNET: discovering microRNA biomarkers that are associated with diseases utilizing biological knowledge-based machine learning, Front. Genet. 13 (Jan. 2023), 1076554, https://doi.org/10.3389/fgene.2022.1076554.

[17] M. Yousef, L. Abdallah, J. Allmer, maTE: discovering expressed interactions between microRNAs and their targets, Bioinformatics 35 (20) (Oct. 2019) 4020–4028, https://doi.org/10.1093/bioinformatics/btz204.

[18] M. Yousef, E. Ülgen, O. Uğur Sezerman, CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis, PeerJ Comput. Sci. 7 (2021) e336, https://doi.org/10.7717/peerj-cs.336.

[19] M. Yousef, G. Goy, R. Mitra, C.M. Eischen, A. Jabeer, B. Bakir-Gungor, miRcorrNet: machine learning-based integration of miRNA and mRNA expression profiles, combined with feature grouping and ranking, PeerJ 9 (2021), e11458, https://doi.org/10.7717/peerj.11458.

[20] M. Yousef, G. Goy, B. Bakir-Gungor, miRModuleNet: detecting miRNA-mRNA regulatory modules, Front. Genet. 13 (2022), 767455, https://doi.org/10.3389/fgene.2022.767455.

[21] M. Yousef, B. Bakir-Gungor, A. Jabeer, G. Goy, R. Qureshi, L.C. Showe, Recursive cluster elimination based rank function (SVM-RCE-R) implemented in KNIME, F1000Research 9 (Jan. 2021) 1255, https://doi.org/10.12688/f1000research.26880.2.

[22] M. Yousef, F. Ozdemir, A. Jaber, J. Allmer, B. Bakir-Gungor, PriPath: identifying dysregulated pathways from differential gene expression via grouping, scoring, and modeling with an embedded feature selection approach, BMC Bioinform. 24 (1) (2023).

[23] N.S. Ersoz, B. Bakir-Gungor, M. Yousef, GeNetOntology: identifying affected gene ontology terms via grouping, scoring, and modeling of gene expression data utilizing biological knowledge-based machine learning, Front. Genet. 14 (2023).

[24] miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database | Nucleic Acids Research | Oxford Academic [Online]. Available: https://academic.oup.com/nar/article/44/D1/D239/2503072. (Accessed 30 November 2021).

[25] J. Piñero, et al., DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Res. 45 (D1) (Jan. 2017) D833–D839, https://doi.org/10.1093/nar/gkw943.

[26] Z. Huang, et al., HMDD v3.0: a database for experimentally supported human microRNA-disease associations, Nucleic Acids Res. 47 (D1) (Jan. 2019) D1013–D1017, https://doi.org/10.1093/nar/gky1010.

[27] D.P. Bartel, MicroRNAs: target recognition and regulatory functions, Cell 136 (2) (Jan. 2009) 215–233, https://doi.org/10.1016/j.cell.2009.01.002.

[28] X. Chen, et al., Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases, Cell Res. 18 (10) (Oct. 2008) 997–1006, https://doi.org/10.1038/cr.2008.282.

[29] R.C. Friedman, K.K.-H. Farh, C.B. Burge, D.P. Bartel, Most mammalian mRNAs are conserved targets of microRNAs, Genome Res. 19 (1) (Jan. 2009) 92–105, https://doi.org/10.1101/gr.082701.108.

[30] L. He, G.J. Hannon, MicroRNAs: small RNAs with a big role in gene regulation, Nat. Rev. Genet. 5 (7) (Jul. 2004), https://doi.org/10.1038/nrg1379. Art. no. 7.

[31] M.Y. Shah, G.A. Calin, MicroRNAs as therapeutic targets in human cancers, Wiley Interdiscip. Rev. RNA 5 (4) (2014) 537–548, https://doi.org/10.1002/wrna.1229.

[32] S.E. McGeary, et al., The biochemical basis of microRNA targeting efficacy, Science 366 (6472) (Dec. 2019) eaav1741, https://doi.org/10.1126/science.aav1741.

[33] H. Jiang, J. Wang, M. Li, W. Lan, F.-X. Wu, Y. Pan, miRTRS: a recommendation algorithm for predicting miRNA targets, IEEE ACM Trans. Comput. Biol. Bioinf 17 (3) (May 2020) 1032–1041, https://doi.org/10.1109/TCBB.2018.2873299.

[34] C. Yan, J. Wang, P. Ni, W. Lan, F.-X. Wu, Y. Pan, DNRLMF-MDA:Predicting microRNA-disease associations based on similarities of microRNAs and diseases, IEEE ACM Trans. Comput. Biol. Bioinf 16 (1) (Jan. 2019) 233–243, https://doi.org/10.1109/TCBB.2017.2776101.

[35] E. Qumsiyeh, L. Showe, M. Yousef, GediNET for discovering gene associations across diseases using knowledge based machine learning approach, Sci. Rep. 12 (1) (Nov. 2022), 19955, https://doi.org/10.1038/s41598-022-24421-0.

[36] Y. Jung, J. Hu, A K-fold averaging cross-validation procedure, J. Nonparametric Statistics 27 (2) (2015) 167–179, https://doi.org/10.1080/10485252.2015.1010532.

[37] J. Hancock, Jaccard Distance (Jaccard Index, Jaccard Similarity Coefficient) (2004), https://doi.org/10.1002/9780471650126.dob0956.

[38] T. Barrett, et al., NCBI GEO: archive for functional genomics data sets—update, Nucleic Acids Res. 41 (2013), https://doi.org/10.1093/nar/gks1193.

[39] R. Trevethan, Sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice, Front. Public Health 5 (Nov. 2017) 307, https://doi.org/10.3389/fpubh.2017.00307.

[40] R. Kolde, S. Laur, P. Adler, J. Vilo, Robust rank aggregation for gene list integration and meta-analysis, Bioinformatics 28 (4) (Feb. 2012) 573–580, https://doi.org/10.1093/bioinformatics/btr709.

[41] K. Tomczak, P. Czerwińska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, Contemp. Oncol. 19 (1A) (2015), https://doi.org/10.5114/wo.2014.47136. A68–A77.

[42] M.J. Goldman, et al., Visualizing and interpreting cancer genomics data via the Xena platform, Nat. Biotechnol. 38 (6) (Jun. 2020) 675–678, https://doi.org/10.1038/s41587-020-0546-8.

[43] J.S. Parker, et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, J. Clin. Oncol. 27 (8) (Mar. 2009) 1160–1167, https://doi.org/10.1200/JCO.2008.18.1370.

[44] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinforma. Oxf. Engl. 26 (1) (Jan. 2010) 139–140, https://doi.org/10.1093/bioinformatics/btp616.

[45] Emerging roles and mechanisms of miR-206 in human disorders: a comprehensive review | Cancer Cell International | Full Text [Online]. Available: https://cancerci.biomedcentral.com/articles/10.1186/s12935-022-02833-2. (Accessed 3 July 2023).

[46] Y. Liu, et al., Loss of N-Acetylgalactosaminyltransferase-4 orchestrates oncogenic MicroRNA-9 in hepatocellular carcinoma* *this work was supported by national key projects for infectious diseases of China grants 2012zx10002012-007 and 2016zx10002018-008; national natural science foundation of China grants 31100629, 31270863, 81401988, 81471621, 81472227, 31570803, and 81572352; and program for new century excellent talents in university grant NCET-13-0146. The authors declare that they have no conflicts of interest with the contents of this article, J. Biol. Chem. 292 (8) (Feb. 2017) 3186–3200, https://doi.org/10.1074/jbc.M116.751685.

[47] H. Shi, Y. Ji, D. Zhang, Y. Liu, P. Fang, MicroRNA-3666-induced suppression of SIRT7 inhibits the growth of non-small cell lung cancer cells, Oncol. Rep. 36 (5) (Nov. 2016) 3051–3057, https://doi.org/10.3892/or.2016.5063.

[48] G. Wang, C. Cai, L. Chen, MicroRNA-3666 regulates thyroid carcinoma cell proliferation via MET, Cell. Physiol. Biochem. 38 (3) (2016) 1030–1039, https://doi.org/10.1159/000443054.

[49] D. Yang, R. Li, J. Xia, W. Li, H. Zhou, miR-3666 suppresses cellular proliferation and invasion in colorectal cancer by targeting SATB2, Mol. Med. Rep. (Oct. 2018), https://doi.org/10.3892/mmr.2018.9540.

[50] D. Li, L. Li, MicroRNA-3666 inhibits breast cancer cell proliferation by targeting sirtuin 7, Mol. Med. Rep. 16 (6) (Dec. 2017) 8493–8500, https://doi.org/10.3892/mmr.2017.7603.

[51] J. Xiao, et al., MicroRNA-520b functions as a tumor suppressor in colorectal cancer by inhibiting defective in cullin neddylation 1 domain containing 1 (DCUN1D1), Oncol. Res. 26 (4) (May 2018) 593–604, https://doi.org/10.3727/096504017X14920318811712.

[52] Y.-C. Lu, et al., MiR-520b as a novel molecular target for suppressing stemness phenotype of head-neck cancer by inhibiting CD44, Sci. Rep. 7 (1) (May 2017) 2042, https://doi.org/10.1038/s41598-017-02058-8.

[53] F. Zhang, et al., MLK3 is a newly identified microRNA-520b target that regulates liver cancer cell migration, PLoS One 15 (3) (Mar. 2020), e0230716, https://doi.org/10.1371/journal.pone.0230716.

[54] W. Cui, et al., miRNA-520b and miR-520e sensitize breast cancer cells to complement attack via directly targeting 3′UTR of CD46, Cancer Biol. Ther. 10 (3) (Aug. 2010) 232–241, https://doi.org/10.4161/cbt.10.3.12277.

[55] J.E. Evangelista, Z. Xie, G.B. Marino, N. Nguyen, D.J.B. Clarke, A. Ma'ayan, Enrichr-KG: bridging enrichment analysis across multiple libraries, Nucleic Acids Res. (May 2023) gkad393, https://doi.org/10.1093/nar/gkad393.

[56] M.V. Kuleshov, et al., Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, Nucleic Acids Res. 44 (W1) (Jul. 2016), https://doi.org/10.1093/nar/gkw377. W90–W97.

[57] M. Gillespie, et al., The reactome pathway knowledgebase 2022, Nucleic Acids Res. 50 (D1) (Jan. 2022) D687–D692, https://doi.org/10.1093/nar/gkab1028.

[58] M. Martens, et al., WikiPathways: connecting communities, Nucleic Acids Res. 49 (D1) (Jan. 2021) D613–D621, https://doi.org/10.1093/nar/gkaa1024.

[59] M. Kanehisa, Toward understanding the origin and evolution of cellular organisms, Protein Sci. 28 (11) (2019) 1947–1951, https://doi.org/10.1002/pro.3715.

[60] KEGG: kyoto encyclopedia of genes and genomes | nucleic acids research | oxford academic [Online]. Available: https://academic.oup.com/nar/article/28/1/27/2384332?login=false. (Accessed 4 July 2023) https://academic.oup.com/nar/article/28/1/27/2384332?login=false.