Hamid Hamzeiy[1,a] / Rabia Suluyayla[2,a] / Christoph Brinkrolf[3] / Sebastian Jan Janowski[3] / Ralf Hofestaedt[3] / Jens Allmer[2]

Visualization and Analysis of MicroRNAs within KEGG Pathways using VANESA

[1] NDAL, Bogazici University, Faculty of Science, Department of Molecular Biology and Genetics, 34342 Istanbul, Turkey
[2] Izmir Institute of Technology, Faculty of Science, Department of Molecular Biology and Genetics, 35430 Urla, Izmir, Turkey, E-mail: jens@allmer.de
[3] Bielefeld University, Faculty of Technology, Department of Bioinformatics and Medical Informatics, D-33501 Bielefeld, Germany

Abstract:

MicroRNAs (miRNAs) are small RNA molecules which are known to take part in post-transcriptional regulation of gene expression. Here, VANESA, an existing platform for reconstructing, visualizing, and analysis of large biological networks, has been further expanded to include all experimentally validated human miRNAs available within miRBase, TarBase and miRTarBase. This is done by integrating a custom hybrid miRNA database to DAWIS-M.D., VANESA's main data source, enabling the visualization and analysis of miRNAs within large biological pathways such as those found within the Kyoto Encyclopedia of Genes and Genomes (KEGG). Interestingly, 99.15 % of human KEGG pathways either contain genes which are targeted by miRNAs or harbor them. This is mainly due to the high number of interaction partners that each miRNA could have (e.g.: hsa-miR-335-5p targets 2544 genes and 71 miRNAs target *NUFIP2*). We demonstrate the usability of our system by analyzing the measles virus KEGG pathway as a proof-of-principle model and further highlight the importance of integrating miRNAs (both experimentally validated and predicted) into biological networks for the elucidation of novel miRNA-mRNA interactions of biological importance.

Keywords: VANESA, microRNA, KEGG pathway, target database, measles

1   Introduction

Many endogenous RNA molecules have been shown to be influential in regulating protein expression, and, therefore, are important for the subsequent cellular events that involve those proteins. One such class of RNA molecules is the evolutionary conserved microRNA (miRNA) group which performs its function via binding to complementary sequences on their target messenger RNAs (mRNAs). Target sequences are usually found within the 3′UTR of the target transcript, however, some studies [1], [2], [3] have suggested that miRNAs may also target the 5′UTR and coding regions. This kind of interaction is believed to mainly result in the translational repression of the target mRNA by means of mRNA degradation or blocking of translation [4] but interestingly, it may also result in an increase in the expression levels of the target genes [3], [5], [6].

Currently, miRBase (miRBase 21) is the primary online database housing miRNA sequences and their annotation [7]. While miRBase contains primarily miRNA information, TarBase [8] mainly contains miRNA target information and interactions derived from different experimental and computational techniques which include gene specific and high-throughput techniques. Similar to TarBase, miRTarBase [9] also contains miRNA targets, but different from TarBase it only contains experimentally proven miRNA-target interactions. The Kyoto Encyclopedia of Genes and Genomes (KEGG) [10], on the other hand, contains a large body of information on gene regulatory pathways additionally including a visualization feature. This has made KEGG the primary information source for the modeling and simulation of biological systems and networks [11].

It is the aim of this work to integrate miRNA and gene regulatory networks. To achieve that end, a novel functionality has been added to the VANESA software by integrating a custom miRNA database to DAWIS-M.D., which allows for merging, overlay, and the intersection of miRNA and gene regulatory networks. This is especially important since a single miRNA can target many mRNAs, while conversely, a single mRNA may

be targeted by more than one miRNAs and also there may be more than one recognition site per mRNA for a given miRNA [12], [13], [14], [15], making it difficult to analyze miRNA regulatory effects efficiently. Our initial data analysis showed that most human KEGG pathways have miRNA interaction partners or genes that harbor miRNAs (99.15 %). We also used the human measles pathway within KEGG as a proof of principle model and tested the different functionalities of VANESA including the ability to add interactions within the network for pre-miRNAs to both target genes and sources. Considering the importance of network analysis for the better understanding of the impact of miRNAs on KEGG pathways, VANESA proves to be a powerful tool for this purpose as it features a wide range of network analysis, simulation, and editing tools allowing for complex yet visually supported analysis of gene and miRNA networks in tandem.

2 Related Works

The emerging roles of miRNAs as important regulators of many cellular processes has led to an increasing number of studies [16], [17], [18] using the comprehensive biological pathways of KEGG to understand the possible functions of miRNAs. For example, DIANA miRPath v.3.0 [18] is an online tool which combines miRNAs with KEGG pathways to give a better understanding of the mechanisms by which biological pathways are regulated. Despite a good web user interface with links to other online resources, it lacks the necessary tools for network editing and simulation. DIANA-microT web server [19] is another tool which along with the limitations mentioned for DIANA miRPath v.3.0, lacks any visualization feature making it very limited for network analysis. Another similar tool to DIANA miRPath v.3.0 is the online service called miRTar [20]. It can visualize miRNAs within KEGG pathways using a similar representation as KEGG. Like many other tools, miRTar lacks features which would enable the user to create custom networks and analyze them separately. An alternative attempt to investigate the effects of miRNAs within KEGG pathways is miTALOS [17], [21]. It uses different target prediction tools and, besides KEGG, it also uses data from NCI PID [22]. On the other hand, there are tools that focus on network reconstruction, which is also very important for the understanding of the impact of miRNAs on biological systems. Tools such as CellMiner [23] work on integrating various biological databases. CellMiner includes a large number of genes as well as about 450 miRNAs along with chemical compounds and drugs. Relationships between miRNAs and genes can also be found in miRGen [24], which by using database integration focuses on the possible effects of the genomic organization of miRNAs on their function. CyTargetLinker [25], CluePedia [26] and CyTRANSFINDER [27] also aim to provide a visual interface for the analysis of miRNAs in regulatory biological networks within the Cytoscape framework [28] but lack the direct access to the many large biological databases.

In this study, we amended DAWIS-M.D. [29], which is an information system based on the data warehouse infrastructure BioDWH [30], with a novel resource containing miRNAs and their targets. BioDWH, DAWIS-M.D. as well as VANESA are open-source projects and free-of-charge for academic use. Several molecular databases, such as KEGG, MINT, IntAct and HPRD, are integrated and available via a web service. Depending on the ongoing project, new databases get integrated or already integrated databases are updated. VANESA [31] is a systems biology software capable of reconstructing, visualizing, and analysis of large biological networks. The rich array of tools available within VANESA facilitates the modeling and simulation of biological networks and biochemical processes. Via the implemented web service to DAWIS-M.D., VANESA has access to the integrated databases which can be used for modeling and reconstruction of biological networks. It also allows users to extract information of a biological component from different biological databases. This along with the ability to construct, combine, intersect, and merge networks can prove to be very important in elucidating the impact of miRNAs and compiling an overall picture of the different roles of miRNAs and their impact on subsequent biological processes and networks.

Any of the compared software solutions above lack either a visualization feature or comprehensive databases. Thus, VANESA and its connection to the data warehouse as well as the features of VANESA make it a suitable platform for the visualization and analysis of the impact of miRNA-mediated regulation within biological networks such as those found within KEGG.

3 Architecture and Implementation

This section describes the architecture and implementation of the miRNA database as well as the changes made in VANESA and the new opportunities which arise due to these changes. Figure 2 gives an overview of the basic features of VANESA.

3.1 Human miRNA Database

Human miRNA data from miRBase, miRTarBase, and TarBase were collected and integrated to form a customized database which was incorporated into DAWIS-M.D. Data from each of the supporting databases were downloaded and filtered for the desired information such as the hairpin (pre-miRNA) sequence and its accession number. 2588 miRNAs were extracted from miRBase along with 10,966 targets from TarBase and 34,777 targets from miRTarBase which resulted in the formation of our database (Figure 1) needed for the integration of miRNAs with KEGG pathways. 3022 genes and 346 miRNAs were found to be common in both TarBase and miRTarBase. In total, the 2588 mature miRNAs extracted from miRBase have 45,344 mRNA targets within TarBase and miRTarBase, targeting 12,347 distinct genes.
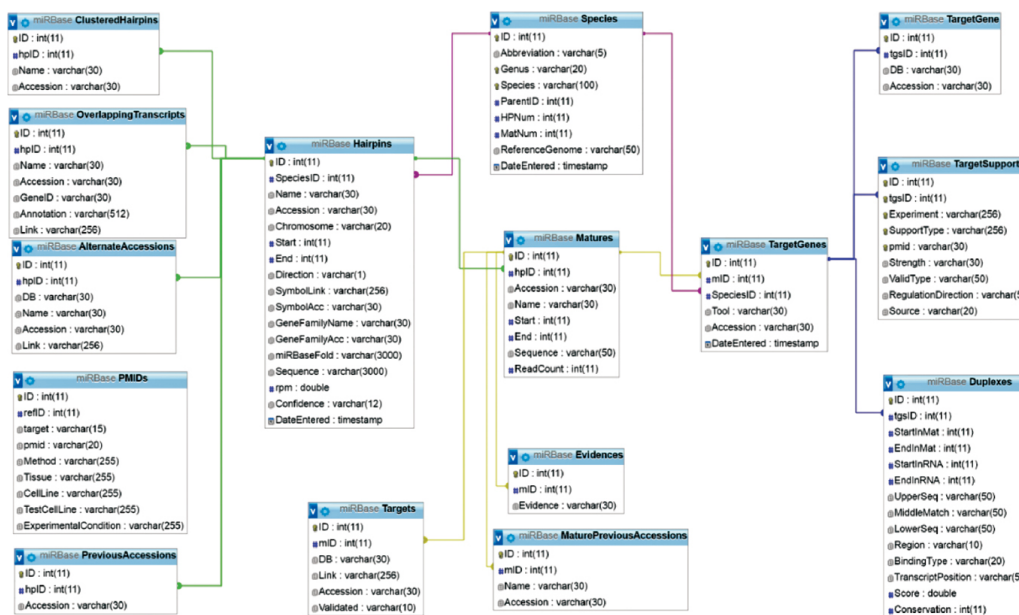

Figure 1: Schema of the integrated miRNA database for VANESA. Our miRNA database was curated from miRBase, Tar-Base, and miRTarBase data. Figure shows tables, fields and relationships within the database schema.

Data from miRBase (release 21) were downloaded in the form of the provided SQL dump and were transformed and integrated into our system using in-house scripts. TarBase V6.0 (http://diana.cslab.ece.ntua.gr/-tarbase/tarbase_download.php) was used for obtaining information regarding the target sequences of experimentally proven human miRNAs. Data were downloaded in Microsoft Excel file format. Data from MiRTarBase 6.1 were also downloaded (http://mirtarbase.mbc.nctu.edu.tw/php/download.php). All data were modeled and imported into our MySQL database. Finally, the database schema and data were imported into DAWIS-M.D.

MySQL Workbench 5.2 CE was used for creating our hand-curated miRNA database. Data from miRBase, TarBase, and miRTarBase were integrated into the database schema (Figure 1) using in-house scripts.

3.2 VANESA

In VANESA, the visual elements for the search in the new miRNA database were added, as well as the related queries. Data can be queried by miRNA name, gene name, accession number or sequence. Retrieved data is shown as a biological network. Additional queries for enriching KEGG pathways with available miRNA data were implemented. For this project, the highlighted databases (KEGG and miRNA) in Figure 2 were added or updated.
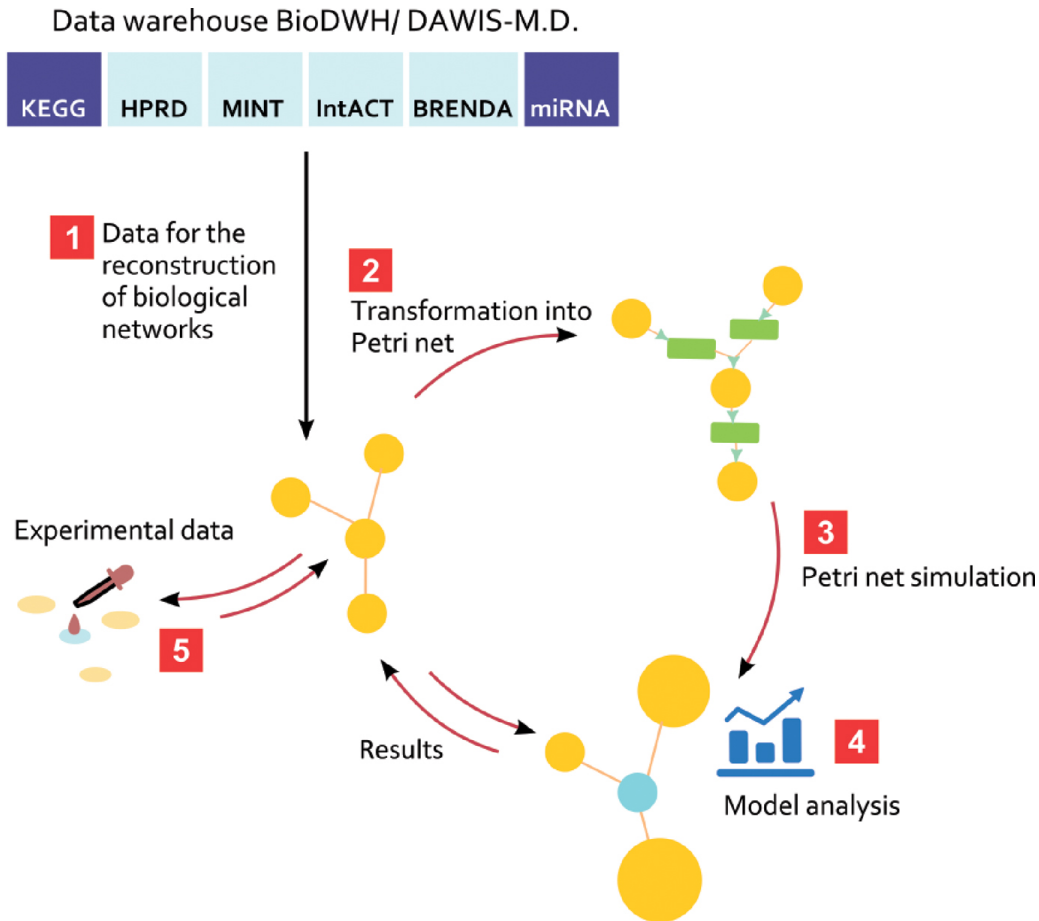
Figure 2: Workflow of VANESA. The highlighted databases KEGG and miRNA were updated or added, respectively.

The addition of our custom miRNA database to DAWIS-M.D. enables the analysis and visualization of miR-NAs within a rich array of biological pathways and further enriches these pathways with regulatory information allowing for more comprehensive analyses. Experimentally validated miRNAs and predicted pre-miRNAs can be added to gene pathways, if they target specific genes within the pathway, or in cases where they are co-expressed with genes therein, which is usually the case when a miRNA is found within an intron of a gene or other regions of a transcript. The same applies for miRNAs originating from intergenic regions, but this is only useful if they target at least one gene within the pathway. This further enhances the possible applications of VANESA and allows for a variety of different kinds of analyses on networks; more than other similar tools described in Section 2. Advantages of VANESA include network modeling techniques such as Petri net modeling and graph analysis. Furthermore, the user can use a large array of databases available within DAWIS-M.D to further enrich each biological pathway directly from VANESA such as KEGG [11], BRENDA [32], and HPRD [33] for further analysis of the impact of miRNAs and to investigate the general network dynamics. This further allows for drill down analysis of the functions of a certain miRNA or a family of miRNAs from a parent network to other related or targeted networks.

4 Application

As a proof of principle, we apply the system allowing the merging of gene and microRNA regulatory pathways to the measeles virus. First, we will provide information about the miRNA database content and then detail how the information allowed a new perspective on measeles infection.

4.1 Connection of the miRNA Database to KEGG

For human, 234 (99.15 %) KEGG pathways out of 236 KEGG pathways in total have miRNA sources and targets; emphasizing the vast impact of miRNA-mediated regulation within biological pathways. The 234 KEGG pathways which are regulated by miRNAs have 65,473 miRNA gene interactions. Keeping in mind that not

every validated miRNA has a validated target, these numbers show how versatile and dynamic miRNAs are in regulating different mechanisms within the cell. Some miRNAs, such as has-miR-335-5p, may have higher potential activity by targeting 2544 genes uniquely whereas the median is 100 targets according to the compiled data within our database, but this is likely to change as more miRNA targets are being validated. This pattern is also seen at the gene level where some genes, such as *Nfat5* and *NUFIP2*, are targeted by multiple miRNAs, 13 and 71, respectively.

4.2 Analyzing Predicted Pre-miRNAs Implicated in Measles within VANESA

In order to test our system, we chose to analyze the human measles pathway within KEGG (Figure 3). Measles pre-miRNAs were predicted using a machine learning approach [34] based on human miRNA features since only measles miRNAs that are similar to human miRNAs can influence the human host. For genes in the measles KEGG pathway, we predicted 22 putative measles pre-miRNAs but pre-miRNAs predicted from *TAB2* and *BCC3* stood out as their confidence was 0.99 according to our model and their mature miRNA sequences had perfect BLAST matches with already known, experimentally validated miRNAs (Figure 3).
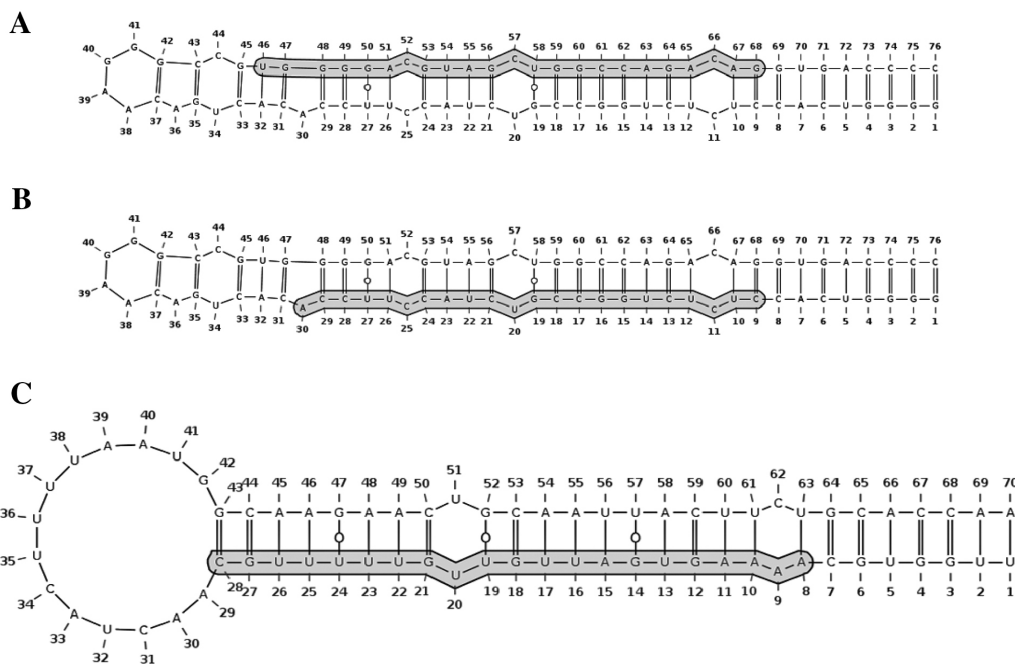

Figure 3: The pre-miRNAs predicted from *BCC3* and *TAB2*. The mature miRNAs similar to known human miRNAs are highlighted with a grey background. A) One hairpin predicted from the *BCC3* similar to the mature miRNA hsa-miR-3191-3p. B) One hairpin predicted from *BCC3* gene similar to the mature miRNA hsa-miR-3191-3p. C) One hairpin predicted from *TAB2* gene similar to the hsa-miR-548au-5p miRNA.

Figure 4 indicates that when confronted with a fully enriched KEGG pathway, not much information can be directly gleaned from the result. Therefore, it is essential to enable filtering and selection processes leading to more accessible views (Figure 5).
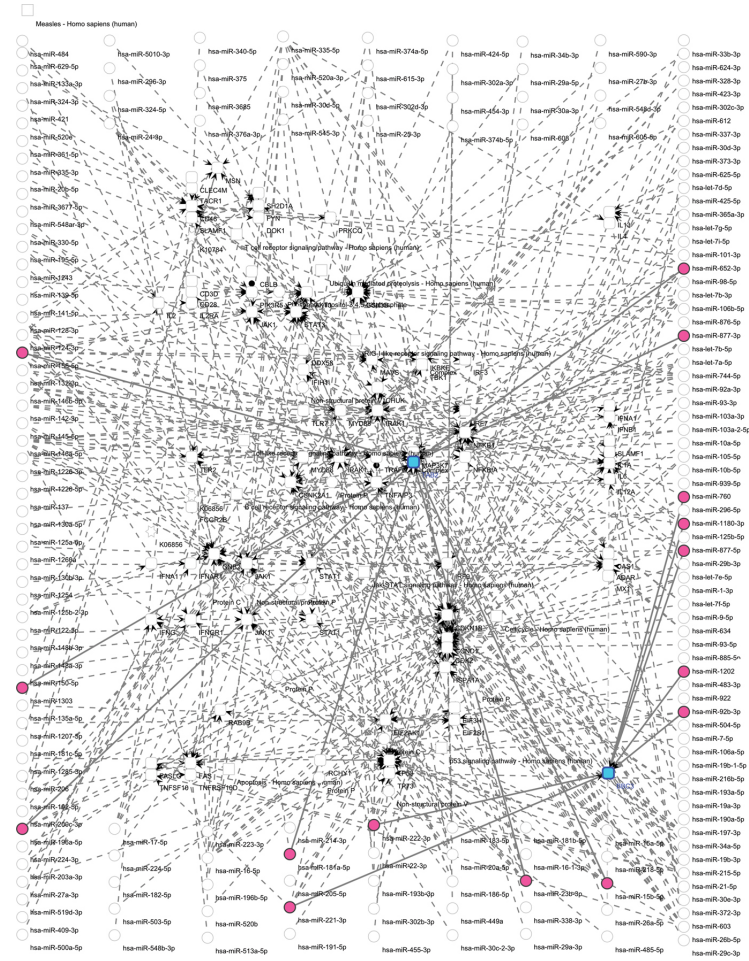
Figure 4: The miRNA-enriched human KEGG measles pathway. The blue squares represent *TAB2* and *BCC3* while miR-NAs targeting these two genes are marked pink. All miRNAs mapped to the pathway have been arranged along the sides not to disturb the view of the pathway.
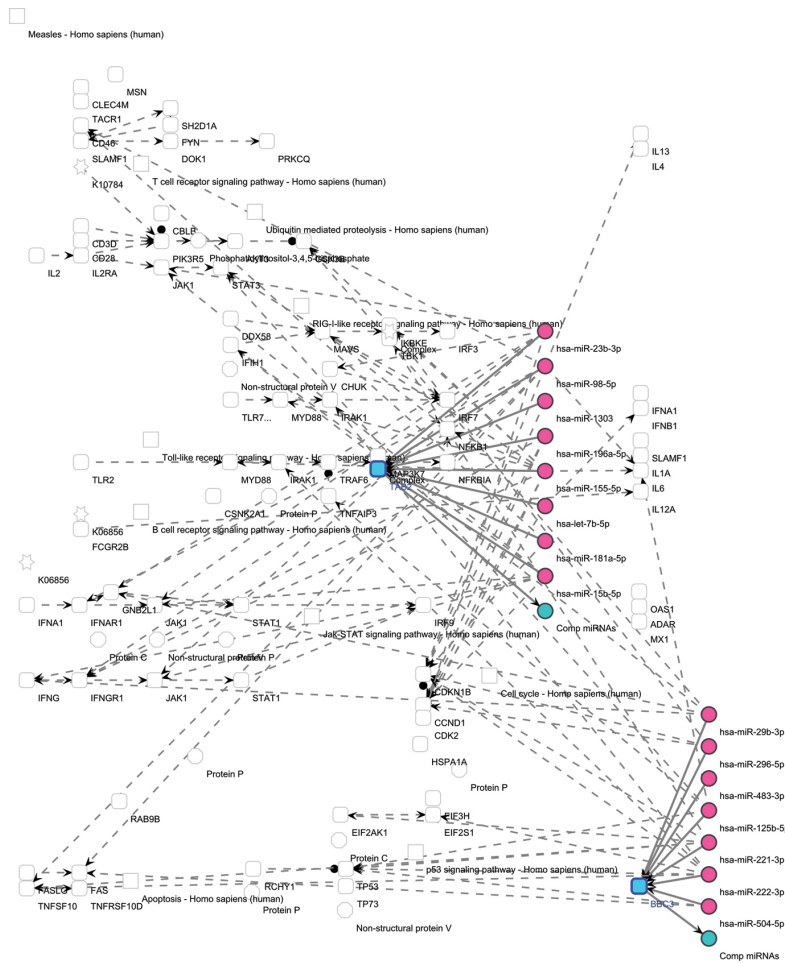
Figure 5: The blue squares represent *TAB2* and *BCC3* while the two miRNAs predicted from these two genes are marked green. The network has been reduced to only include the most relevant nodes.

Therefore, Figure 5 shows only the miRNAs targeting *TAB2* and *BCC3* along with their predicted miRNAs and their subsequent partner nodes within the human KEGG Measles pathways. We employed the same approach for the prediction of novel human pre-miRNAs that may be co-expressed with genes involved in the KEGG human measles pathway. Using VANESA we were able to integrate all these predicted pre-miRNAs and experimentally validated miRNA into the KEGG human measles pathway which gave a more comprehensive view of the possible role of miRNAs in virus and host interactions.

Figure 5 highlights the two newly predicted miRNAs, but only for hsa-miR-548au-5p targeting information was available in our miRNA database. Unfortunately, none of the targets were located in the KEGG pathway, which would have enabled us to add another internal link to the existing pathway. Therefore, we used its known targets to perform Reactome gene enrichment (Table 1). Among the top ten enriched pathways, three of them refer directly to the immune system. Additionally, other pathways like transcription are enriched, which makes sense in the context of infection. After using VANESA to reconstruct and visualize the human measles pathway, we moved further to predict novel miRNAs from both, the measles genome and the respective human genes of the human measles KEGG pathway. This was done by fragmenting the genomic sequences into overlapping sequences and calculating their secondary structure using RNAfold. Hairpin structures were extracted and features were calculated for miRNA prediction using a random forest classifier. This was done with KNIME, a data analytics platform [35].

Table 1: Excerpt of the enrichment report created using Reactome based on the miRNA's targets of hsa-miR-548au-5p.

Pathway identifier	Pathway name	#Entities found	#Entities total	Entities *p*-Value
R-HSA-74160	Gene expression	14	2085	0.238269722
R-HSA-168256	Immune system	11	2059	0.580989218
R-HSA-212436	Generic transcription pathway	9	1032	0.10732199
R-HSA-3700989	Transcriptional regulation by TP53	8	476	0.004308557

R-HSA-162582	Signal transduction	8	2837	0.993325881
R-HSA-1280218	Adaptive immune system	6	1144	0.601587964
R-HSA-983169	Class I MHC mediated antigen processing & presentation	5	478	0.119575644
R-HSA-1640170	Cell cycle	5	638	0.268044602
R-HSA-168249	Innate immune system	5	1079	0.71156603
R-HSA-1430728	Metabolism	5	3351	0.999990651

5 Discussion

Currently, a large amount of biological data are available and this pool is growing at an increasing pace due to the rapid generation of large amounts of data generated via high-throughput technologies. This has led to a surge in the number of biological databases during the past decade aiming to store and share data within the scientific community. However, this increase in number leads to the problem of data fragmentation, emphasizing the need to perform data integration. An example of such an effort is DAWIS-M.D. which includes data from major biological databases. We enabled the further enrichment of the comprehensive biological pathways within KEGG and also provide a suitable platform for analyzing the impact of miRNAs-mediated regulation within biological pathways. This was achieved by including our custom miRNA database, created using data from miRBase, TarBase, and miRTarBase, to DAWIS-M.D. and using VANESA and its new miRNA query extensions. We were able to test our system using the human Measles pathway as a proof of principle model. The ability to visualize, edit, and simulate biological networks is an important aspect in trying to shed light on the numerous events that work in concert within a biological system. VANESA provides these functions as different networks can be compared and the importance of each miRNA can be weighted, which can prove helpful in experiment design and for visualizing the interactomics of the cell in greater detail. Keeping in mind that miRNAs pose a great potential for being used as future therapeutics, the visualization of their regulatory pathways will be helpful in the development of novel medications, emphasizing the importance of data fusion and data integration within DAWIS-M.D. Using such a system allows for performing drill down analyses of the different elements found in biological pathways. We applied these analyses to the Measles pathway and were able to find two novel miRNAs which are similar to already known miRNAs. The targets of one of the miRNAs further showed that it is likely to be implicated in infection thus highlighting the efficiency of the overall approach.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

References

[1] Place RF, Li L-C, Pookot D, Noonan EJ, Dahiya R. MicroRNA-373 induces expression of genes with complementary promoter sequences. Proc Natl Acad Sci USA. 2008;105:1608–13.
[2] Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos l. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. Nature. 2008;455:1124–8.

[3] Ørom UA, Nielsen FC, Lund AH. MicroRNA-10a binds the 5′UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell. 2008;30:460–71.

[4] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136:215–33.

[5] Hu LL, Huang Y, Wang QC, Zou Q, Jiang Y. Benchmark comparison of ab initio microRNA identification methods and software. Genet Mol Res. 2012;11:4525–38.

[6] Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. Science. 2007;318:1931–4.

[7] Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucl Acids Res. 2011;39:D152–7.

[8] Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucl Acids Res. 2012;40:D222–9.

[9] Hsu S-D, Lin F-M, Wu W-Y, Liang C, Huang WC, Chan WL, et al. miRTarBase: a database curates experimentally validated microRNA-target interactions. Nucl Acids Res. 2011;39:D163–9.

[10] Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucl Acids Res. 2012;40:D109–14.

[11] Aoki KF, Kanehisa M. Using the KEGG database resource. Curr Protoc Bioinformatics. 2005;Chapter 1:Unit 1.12.

[12] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116:281–97.

[13] Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet. 2005;37:495–500.

[14] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120:15–20.

[15] Wu S, Huang S, Ding J, Zhao Y, Liang L, Liu T, et al. Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3′ untranslated region. Oncogene. 2010;29:2302–8.

[16] Shen E, Diao X, Wang X, Chen R, Hu B. MicroRNAs involved in the mitogen-activated protein kinase cascades pathway during glucose-induced cardiomyocyte hypertrophy. Am J Pathol. 2011;179:639–50.

[17] Kowarsch A, Preusse M, Marr C, Theis FJ. miTALOS: Analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. RNA. 2011;17:809–19.

[18] Vlachos IS, Zagganas K, Paraskevopoulou MD, Georgakilas G, Karagkouni D, Vergoulis T, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. Nucl Acids Res. 2015;43:W460–6.

[19] Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, et al. DIANA-microT web server: elucidating microRNA functions through target prediction. Nucl Acids Res. 2009;37:W273–6.

[20] Hsu JB, Chiu C-M, Hsu S-D, Huang WY, Chien CH, Lee TY, et al. miRTar: an integrated system for identifying miRNA-target interactions in human. BMC Bioinf. 2011;12:300.

[21] Preusse M, Theis FJ, Mueller NS. miTALOS v2: Analyzing Tissue Specific microRNA Function. PLoS One. 2016;11:e0151771.

[22] Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, et al. PID: the pathway interaction database. Nucl Acids Res. 2009;37:D674–9.

[23] Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. CellMiner: a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. Cancer Res. 2012;72:3499–511.

[24] Megraw M, Sethupathy P, Corda B, Hatzigeorgiou AG. miRGen: a database for the study of animal microRNA genomic organization and function. Nucl Acids Res. 2007;35:D149–55.

[25] Kutmon M, Kelder T, Mandaviya P, Evelo CT, Coort SL. CyTargetLinker: a cytoscape app to integrate regulatory interactions in network analysis. PLoS One. 2013;8:e82160.

[26] Bindea G, Galon J, Mlecnik B. CluePedia cytoscape plugin: pathway insights using integrated experimental and in silico data. Bioinformatics. 2013;29:661–3.

[27] Politano G, Orso F, Raimo M, Benso A, Savino A, Taverna D, et al. CyTRANSFINDER: a Cytoscape 3.3 plugin for three-component (TF, gene, miRNA) signal transduction pathway construction. BMC Bioinf. 2016;17:157.

[28] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software Environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.

[29] Hippe K, Kormeier B, Janowski SJ, Töpel T, Hofestädt R. DAWIS-M.D. 2.0 – A data warehouse information system for metabolic data. In: Fähnrich K.-P., Franczyk B., editors. In Informatik 2010: Service Science – Neue Perspektiven für die Informatik, Beiträge der 40. Leipzig: Jahrestagung der Gesellschaft für Informatik e.V. (GI), 2010:720–5.

[30] Töpel T, Kormeier B, Klassen A, Hofestädt R. BioDWH: a data warehouse kit for life science data integration. J Integr Bioinform. 2008;5. DOI:10.2390/biecoll-jib-2008-93.

[31] Brinkrolf C, Janowski SJ, Kormeier B, Lewinski M, Hippe K, Borck D, et al. VANESA – a software application for the visualization and analysis of networks in system biology applications. J Integr Bioinform. 2014;11:239.

[32] Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, et al. BRENDA, the enzyme information system in 2011. Nucl Acids Res. 2011;39:D670–6.

[33] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference database – 2009 update. Nucl Acids Res. 2009;37:D767–72.

[34] Saçar Demirci MD, Bağcı C, Allmer J. Non-coding RNAs and inter-kingdom Communication, 1st ed Cham: Springer International Publishing, 2016.

[35] Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: the konstanz information miner. Berlin Heidelberg: Springer, 2008:319–26.