miEAA 2.0: Integrating multi-species microRNA enrichment analysis and workflow management systems

Fabian Kern [1,†], Tobias Fehlmann [1,†], Jeffrey Solomon [1], Louisa Schwed [1], Christina Backes [1], Eckart Meese [2], and Andreas Keller [1,3,*]

[1] Chair for Clinical Bioinformatics, Saarland University, Saarbrücken, Germany

[2] Department of Human Genetics, Saarland University Hospital, Homburg, Germany

[3] School of Medicine Office, Stanford University, Stanford, CA, USA

Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA

[†] Both authors added equally to the work

[*] Correspondence: andreas.keller@ccb.uni-saarland.de; Tel.: +49-(0) 681-302-68611

Abstract

Gene set enrichment analysis has become one of the most frequently used applications in molecular biology research. Originally developed for gene sets, the same statistical principles are now available for all omics types. In 2016, we published the miRNA enrichment analysis and annotation tool (miEAA) for human precursor and mature miRNAs.

Here, we present miEAA 2.0, supporting miRNA input from *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*. To facilitate inclusion of miEAA in workflow systems, we implemented an Application Programming Interface (API). Users can perform miRNA set enrichment analysis using either the web-interface, a dedicated Python package, or custom remote clients. Moreover, the number of category sets was raised by an order of magnitude. We implemented novel categories like annotation confidence level or localisation in biological compartments. In combination with the miRBase miRNA-version and miRNA-to-precursor converters, miEAA supports research settings where older releases of miRBase are in use. The web server also offers novel comprehensive visualisations such as heatmaps and running sum curves with background distributions. Lastly, additional methods to correct for multiple hypothesis testing were implemented. We demonstrate the new features using case studies for human kidney cancer and mouse samples. The tool is freely accessible at: https://www.ccb.uni-saarland.de/mieaa2.

Keywords— Gene set enrichment analysis; MicroRNAs; Post-transcriptional gene regulation; Annotation;

Introduction

Transcriptomics designates an indispensable set of techniques to study gene expression, often in a genome-wide manner, as the backbone of modern molecular biology and clinical research. The innumerable amount of classical bulk-sequencing datasets is further augmented by the recent advancements in high-resolution single-cell approaches. Since gene expression is constituted by many biological factors, experimental focus has been enlarged to include the regulatory non-coding transcriptome (ncRNAs), i.e. to RNA classes that regulate messenger RNAs (mRNAs) either directly or indirectly. Among these, microRNAs (miRNAs) are small non-coding RNAs, typically 18-25 nucleotides in length, loaded by proteins of the AGO-family to build RNA-induced silencing complexes (RISC) [1]. Gene regulation through the RISC complex is facilitated by one or two mature ($-5p$; $-3p$) miRNA arms, arising from one or several transcribed precursors [2]. Besides other modes of action, activated complexes target preferentially $3'$-untranslated regions of mRNAs to induce either catalytic cleavage or translation repression. Hence, profiling miRNA expression contributes to the understanding of gene regulation and potentially portrays cellular states. To date, numerous studies highlight their informative role in disease detection, sub-type classification, or progression, such as for cancer [3], neurodegenerative [4], or metabolic disorders [5] with a variety of bio-specimens [6].

Considering that several thousands of miRNAs have already been discovered, many novel miRNA candidates have been additionally proposed [7], while the total number of human miRNAs is estimated to be 2,300 [8]. Finding differences in expression for miRNAs is similar to mRNAs and therefore non-trivial. Differential gene expression studies often lead to dozens, hundreds, or even thousands of de-regulated genes. Thus, large scale studies often make use of the functionality of gene set enrichment analysis (GSEA) [9]. GSEA can further reduce large amounts of information towards a significant set of molecular functions, biological properties, or pathways of genes. In principle, a user inputs either a set or ordered list of genes and the tool runs the required statistical algorithms and provides background datasets to compare against.
Similar functionality was also implemented for other omics types, including proteomics, metagenomics, or epigenomics. An in-depth review of gene set analysis methods for data other than mRNAs demonstrates the increasing interest and demand of the community in respective tools [10]. We previously developed an approach tailored for both miRNA precursor enrichment and mature miRNA enrichment analyses, the miRNA enrichment analysis and annotation tool (miEAA) [11]. Here, we present an update of this tool that includes more categories, supports more organisms, has new statistical functionality and offers a standardised Application Programming Interface (API) to facilitate the inclusion of miEEA in modern data analysis workflows [12].

Given the growing interest in miRNAs, other tools with similar functionality to miEAA exist. Among the most functional tools, the recent successor version of TAM [13] introduced 1,238 human miRNA set categories obtained from manual literature review of approximately 9,000 scientific manuscripts, along with new query and visualisation features. In addition to the over- and under-representation analysis, users can compare the correlation of two miRNA lists under different disease conditions. All kinds of enrichment tools rely on high quality sets of miRNA categories that were either obtained by curation of scientific literature or collected from specific databases. For instance, curated miRNA annotations can be obtained from miRBase or miRCarta [14], miRNA-target interactions from miRTarBase [15], miRNA-pathway associations from miRPathDB [16], tissue-specific miRNAs from the human TissueAtlas [17], or miRNA-disease associations from HMDD [18] or MNDR [19], many of which were updated in the last two years. Further specialised annotations like miRNA and transcription factor interactions provided by TransmiR [20], miRNA sub-cellular localisations collected in RNALocate [21], or extra-cellular circulating miRNAs contained in miRandola [22] provide target categories for integrated enrichment analysis.

MATERIALS AND METHODS

In miEAA 2.0 we provide support for 3 species (2 new), 24 new category sets, and updates to our pre-existing datasets. To unify data preprocessing, we implemented an automated pipeline using Snakemake [23], Python 3.6, and the pandas [24] Python package facilitating data collection and filtering steps. For each species and their corresponding data sources our pipeline performs the same basic process, consisting of downloading the datasets, cleaning and updating the miRNA and precursor identifiers, transforming the results into a Gene Matrix Transposed (GMT) file, and creating background reference sets. Files were copied to the web server without further modification.

Data collection

Novel datasets were obtained to build our enrichment categories, consisting of Gene Ontology [25], miRTarBase 8.0 [15], KEGG [26], miRandola 2017 [22], miRPathDB 2.0 [16], TissueAtlas [17], MNDR v2.0 [19], NPInter 4.0 [27], RNALocate v2.0 [21], TAM 2.0 [13], and TransmiR v2.0 [20]. Other pre-existing datasets have been updated, including HMDD v3.0 [18] and miRBase v22.1 [28]. We retained the rest of our pre-existing datasets, namely miRWalk2.0 [29], our published age and gender dependent miRNAs and our distribution of miRNAs in immune cells [11]. All datasets contain miRNAs or precursors for *Homo sapiens*. When available, we also utilise the data for *Mus musculus* and *Rattus norvegicus*, allowing enrichment analysis on 39, 31, and 26 miRNA/precursor category sets, respectively. Raw datasets were obtained either through a direct download or via an API. In particular, the QuickGO and KEGG datasets are compiled by querying their corresponding REST APIs.

Category data preprocessing

First, data from QuickGO was mapped back to miRBase using RNAcentral [30]. NCBI Gene was used in conjunction with miRTarBase to produce the indirect annotations. With the aid of the miRBaseConverter R package [31], miRNA and precursor names were translated to the latest version of miRBase. For KEGG Pathways and GO Annotations (direct and indirect through target genes from miRTarBase) we only keep miRNAs for which functional MTI support is available. In the MNDR diseases category set, we exclude HMDD data as it is precursor based, and MNDR is for mature miRNAs.

Web server, statistics, and API implementation

The miEAA web server was built using a dockerized Django Web Framework v2.1, which exposes a web-API using the Django REST framework. The celery software was used as the job scheduler. Frontend libraries comprise Highcharts, dataTables, jquery, and Bootstrap. P-value correction methods were implemented using the R stats package. For the static GSEA running sum plots, a simulated background distribution is computed by randomly permuting the test set 100 times and traversing the running sum for each random permutation. Alongside our new API we provide a lightweight Python package, as well as a command line interface (CLI) tool, supporting Python 3.5 or higher. These are made freely available through the Python Package Index (pip) and through the *ccb-sb* conda channel.

Case studies

Raw and reads per million miRNA mapping (rpmmm) normalized miRBase v21 precursor counts and metadata of kidney renal clear cell carcinoma case and control samples were obtained from TCGA. Since multiple sequencing results might be associated with the same sample ID in TCGA, we kept only one result file for each sample by preferring files from H over R over T analytes and selecting the aliquot with the highest plate number and / or lexicographical sorting order. Subsequently, miRNAs with fewer than 5 raw reads in less than 50% of either case or control samples were discarded from the analysis. All remaining miRNA counts were log_2-scaled. Effect size was calculated using the implementation of Cohen's d from the R package effsize. Lists of precursor names, either selected by statistical significance or ordered by effect size, were converted from miRBase v21 to v22.1 using the online miRBase converter feature of miEAA. The list of all precursors from miRBase v21, converted to v22.1, were used as a reference set. The configured parameters included default precursor category sets without the *PubMed ID* and *TransMiR Tissues* sets, BH-FDR adjustment to a significance level of 0.05 with independently adjusted p-values per category set, and a minimum of 2 required hits per sub-category.

For the second case study, raw Agilent microarray data along with sample metadata was downloaded from NCBI's GEO using accession ID GSE117000. Array parsing and probe signal processing was performed identically to the description in the first publication of miEAA [11]. Subsequently, all counts were quantile-normalized and log_2-transformed. All further down-stream analyses were performed analogous to the first case-study described above.

RESULTS

Overview on miEAA 2.0

In the following, changes and novelties introduced by the second major release of miEAA are described. Since all annotations of miRNAs to categories and databases are with respect to the miRNA reference database, miRBase, we converted the datasets to match its latest public version 22.1. This also affects the miRBase-version and miRNA-to-precursor converters, the former of which was designed to be fully backwards compatible. Moreover, both ORA and GSEA algorithms accept lists of either precursors or miRNAs, from human, mouse, and rat species. In total, 123,655 categories from 15 published databases/resources are available to test against. A detailed breakdown of the counts by source and organism, on database and category set level, are available from Supplementary Table 1 and 2, respectively. For the precursor annotations, we curated family assignments, re-computed genomic clusters of miRNA genes, updated the chromosomal locations and source PubMed IDs for human, and added all similar categories for mouse and rat. All species are annotated with a new category containing high confidence precursors according to miRBase criteria. For human data, we transferred the disease annotations from HMDD to the new major release v3. We added associations from MNDR to allow disease comparisons against HMDD, and incorporated functional RNA interactions from NPInter. Lastly, novel categories such as the cellular localisation of miRNAs and regulatory interactions between miRNAs and transcription factors were incorporated from RNALocate and TransmiR, respectively. For the mature miRNAs, comparable changes apply as for the precursors in the cases of miRBase, MNDR, NPInter, and RNALocate-derived categorie sets. The gap between annotations of miRNA properties and their function is filled by categories on target genes taken from miRTarBase. To facilitate target-based enrichment of molecular pathways or biological function, we computed enrichments on target genes of miRNAs using Gene Ontology and KEGG. As an alternative for end-users, pre-computed significant enrichments of miRNAs associated with pathways provided by miRPathDB were made available for analysis. As the data from miRPathDB already involves a statistical pre-filtering, we implemented a new list of expert categories to highlight the underlying differences. Manually curated classifications from miRandola about known circular or extracellular miRNAs complete the final category dataset. Supposedly, the substantially enlarged number of categories might increase the average runtime of our algorithms, especially for the computationally intensive GSEA. Therefore, we profiled and improved our GeneTrail-based implementation to be three times faster, on average. [32].

Along with improving the data, we raised the available number of statistical parameter settings as well. First, users can request unadjusted or adjusted p-values using six published techniques to account for multiple hypothesis testing on the same dataset. In addition to the classical Bonferroni and Benjamini-Hochberg False discovery rate

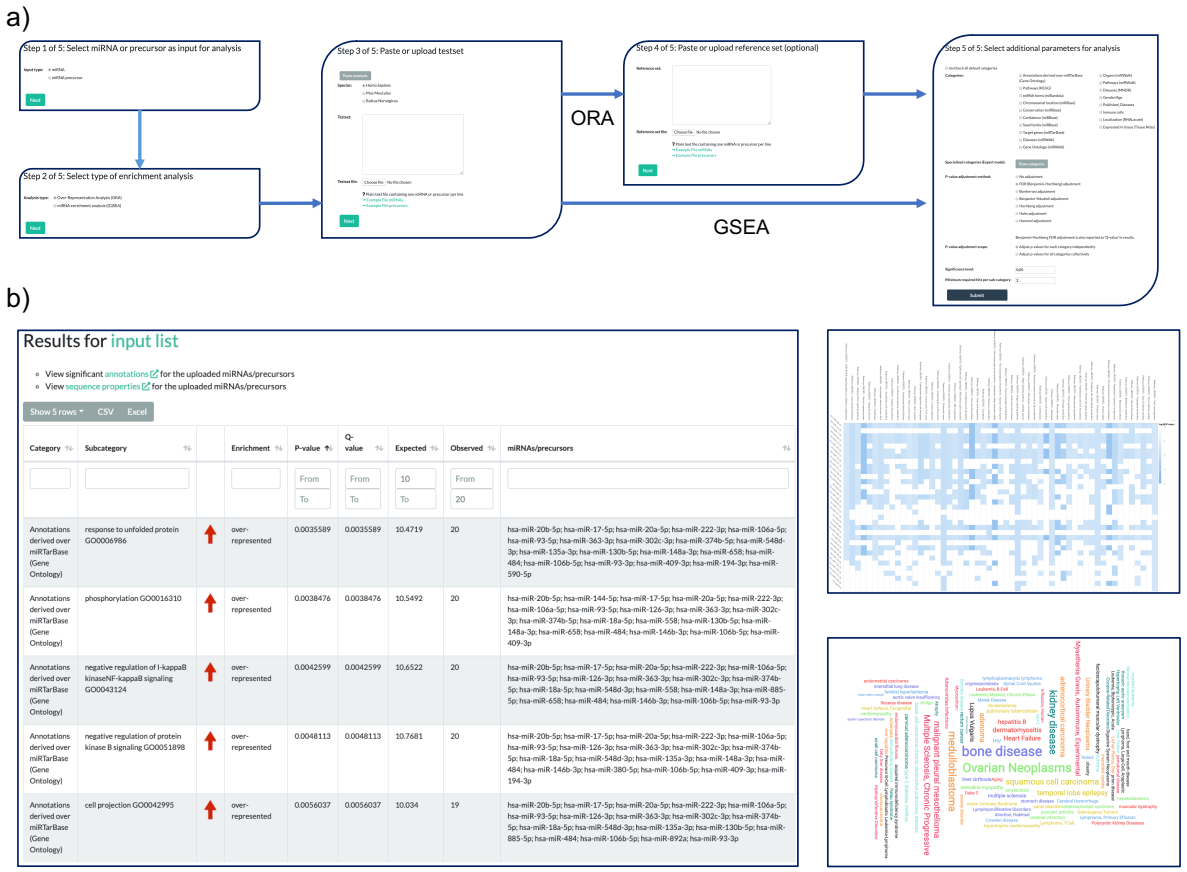

Figure 1: miEAA workflow and exemplary results. **(a)** Each miRNA / precursor enrichment analysis consists of at most five steps. First, users should select whether they want to perform enrichment on precursors or miRNAs. Second, the enrichment algorithm, i.e. either ORA or GSEA must be selected. Next, the desired test set can be defined either through a textbox or a file upload. The fourth step only appears for ORAs where custom background reference sets can be inserted or uploaded. This is optional since miEAA provides pre-computed reference sets for all categories. Lastly, the set of categories and databases as well as statistical parameters should be selected. **(b)** Typical result view for an ORA. Users can sort, select, filter, and export the obtained enrichment results interactively. Moreover, several visualisations of the results are provided for each run, such as the precursor / miRNA to category heatmap and the category wordcloud.

(BH-FDR) procedures, the adjustments proposed by Benjamini-Yekuteli, Hochberg, Holm, and Hommel can be selected. Moreover, the default behaviour of miEAA to correct p-values database / category set-wise was extended by a p-value pooling approach. In summary, the well-established alternatives for p-value correction can support highly customised research setups where alternate levels of stringency are required [33].

We also evaluated new visualisation features for the output of enrichment analyses to provide a simple overview and to improve comprehension. As a result, we made existing graphs interactive and implemented enrichment graphs with simulated background distributions for GSEA as well as automatic word cloud and heatmap plots for all enrichment algorithms. Word clouds display the names of obtained categories while scaling the size of the terms relatively to the number of hits that occurred and allow us to qualitatively compare the categories. On top of that, category to miRNA heatmaps depict log-transformed p-values at the combinations where hits occurred. This feature permits a simple way to compare the similarity of enriched / depleted categories with respect to associated miRNAs or precursors. The workflow of miEAA and example visualisations are displayed in Figure 1. Finally, we enhanced the general accessibility of miEAA through the implementation of a public API and a Python package, for which more details are provided below.

Case study 1: Human kidney renal clear cell carcinoma

As the first case-study of miEAA 2.0, we acquired 591 human miRNA-seq samples from the kidney renal clear cell carcinoma (KIRC) project of TCGA, which can be divided into 520 Primary tumor (PT) and 71 Solid tissue normal (STN) samples. Sample information can be found in Supplementary Table 3. Of the 1,881 precursors

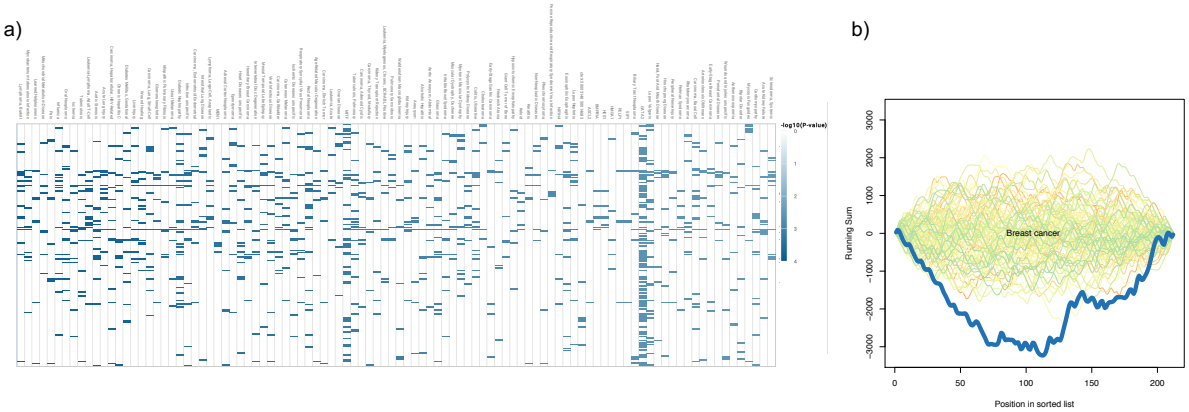

Figure 2: Web server visualisation of case study results. **(a)** Category to miRNA heatmap with $-log10$-scaled enrichment p-values from the first case study. **(b)** GSEA plot with simulated background distributions (green to orange lines) and actual depletion observed for breast cancer (dark blue line) during evaluation of the second case study.

from miRBase v21, 321 are consistently detected in at least 50% of the samples for each biogroup. Among these, 282 were differentially expressed between PT and STN according to the FDR-adjusted wilcoxon test p-values ($p < 0.01$). Over-representation analysis of the precursors resulted in 541 significantly enriched and 7 significantly depleted (FDR-adjusted; $p < 0.05$) categories. As shown in Figure 2(b), a subset of miRNAs is ubiquitously present in significant categories, while others seem to be more specific. The top 10 categories sorted by increasing p-value are associated with cancer, including renal cell carcinoma. Also, the observed over expected ratio (123/48.6) indicates a strong enrichment ($p = 2.80 \times 10^{-38}$) of the de-regulated precursors with kidney and other types of cancer. A miRNA set enrichment analysis, using the list of detected precursors and sorted by effect size, revealed 253 enriched and 40 depleted categories. Here, the miRNA-precursor cluster $147,189,704 : 147,284,728$ on the X chromosome is the most depleted category ($p = 8.64 \times 10^{-10}$), an observation that is in line with the depletion of precursor family hsa-mir-506. Interestingly, the list of highly enriched terms contains many transcription factors, the top 5 being *HEY1*, *WDR5*, *ELF1*, *BRD4*, and *FLI1*.

Case study 2: Mouse model for breast cancer progression

To showcase the novel support for model organisms in miEAA, we selected a dataset from GEO where circulating miRNAs from a breast-cancer mouse model were measured with microarrays [34]. The dataset comprises in total 36 samples from mutation-carrier (NeuT+) and age-matched wildtype (NeuT-) mice that were collected at the premalignant, preinvasive, and invasive stages of the disease. In this particular study, agilent microarrays probed with miRNAs from miRBase v19 were used on mice's plasma extracted RNA samples. Sample information can be found in Supplementary Table 4. Following a detection threshold procedure similar to our first case study, 212 miRNAs remained for differential expression analysis. Of these, mmu-miR-6243 had to be discarded as a result of mapping the identifiers from miRBase v19 to v22.1, which we performed with the miEAA miRBase version converter. Subsequently, we applied GSEA on the list of miRNAs sorted by decreasing effect size between the premalignant and the invasive stage, for NeuT+ and NeuT- samples separately. Strikingly, the former run returned 311 significant categories, while the latter returned none. Overall, many more categories seemed to be depleted ($N = 301$) than enriched ($N = 9$), suggesting a wide-spread up-regulation of molecular pathways by miRNAs being down-regulated in NeuT+. For example, we found Macrophage differentiation ($p = 2.54 \times 10^{-5}$), Vasculature development ($p = 1.60 \times 10^{-4}$), and VEGF signaling pathway ($p = 0.0016$) to be depleted, which might be a signal for the increased tumor burden of NeuT+ mice at the invasive breast cancer stage. Moreover, we evaluated GSEAs for the comparison of NeuT+ and NeuT- at all three stages. While the first two setups returned a rather unspecific set of categories with all p-values located close to the significance boundary, the last comparison yielded many interesting results. First, observations were in line with the group-wise comparison along the age dimension, because all categories are depleted, i.e. no enrichments. Further, the results show that several dozen conserved miRNAs ($p = 4.53 \times 10^{-5}$) are down-regulated in the NeuT+ model at the invasive stage. More significant categories we found like exosome ($p = 2.31 \times 10^{-5}$) and circulating ($p = 0.0086$) miRNAs, breast cancer ($p = 0.0094$, Figure 2(b)), microRNAs in cancer ($p = 0.028$), and PI3K-Akt signaling pathway ($p = 0.028$) can be associated with this exemplifying study.

New data export and browsable API

All data, results, and interactive plots shown on the web server are exportable to common data formats. Also, we were seeking to support the trend towards the development of reproducible and automated data analysis pipelines

[12]. To this end, miEAA hosts a public, browsable API offering the same functionality as the web site, allowing one to access the miRNA converters and statistical algorithms remotely. This functionality is further augmented by a full-feature Python package with API library code and a command-line interface (CLI). For example, a regular workflow as performed on the website can be accomplished with three sequential calls to the web API or one call to the CLI. We provide code examples in the common data science programming languages Python and R to demonstrate this use-case. We also implemented the interface to solve two recurring problems in biological data analysis. First, reproducibility of statistical experiments can be improved, because usage of the versioned API in the context of a workflow manager such as Snakemake [23] or Nextflow fosters self-documenting research setups [35]. Second, oftentimes the analysis of miRNA high-throughput data involves the comparison of multiple biogroups, timepoints or other annotation variables. With the aid of our API and the package, multiple runs of miEAA can be performed at ease while minimising the time spent for set up and results aggregation.

DISCUSSION

Statistical tools for biological enrichment analysis are a key to understanding data from high-throughput omics assays. However, the performance primarily depends on the quality of the underlying annotations and the statistical soundness. We show that new developments in the miRNA research field yielded an unprecedented set of biological categories, covering most aspects of miRNA properties and function, with cross-species analysis becoming increasingly important. On the other side, as with every statistical framework applied on biological data, assumptions are not always met and findings should be assessed critically in the light of further validation experiments. The novel release of miEAA attempts to cover these aspects by enhancing the set of available categories both quantitatively and qualitatively as well as through offering more (stringent) approaches for p-value correction. Also, a major limitation of some datasets concerns the availability of mature miRNA identifiers, as only precursor names were available from source databases. However, especially in the context of diseases, mature miRNA resolution is preferable to match the biological selectivity for one major miRNA arm being expressed. Datasets incorporated in miEAA were compiled either automatically or manually. TAM, another miRNA enrichment tool with functionality similar to miEAA, uses a fewer number of high-quality annotations, which come exclusively from manual curation [13]. A detailed comparison with respect to 22 criteria between our tool and TAM is shown in Supplementary Table 5.

We have demonstrated the capability of miEAA to yield novel biological results in cancer research. For the kidney renal clear cell carcinoma case study, we found a depletion of the mir-506 precursor family, which has been observed before in other types of cancers [36, 37]. Many interactions to transcription factors were also found for the up-regulated miRNAs, suggesting an increased regulatory burden due to the exceedingly transcriptional up-regulation observed in cancer. For example, HEY1, which is a transcriptional repressor has been characterised to be up-regulated in renal cell carcinomas [38]. For the mouse breast cancer progression study, we illustrated the backwards compatibility of miEAA with respect to miRBase. The overall observed depletion of pathways in mice agrees with our first case study. Moreover, the significant categories like vasculature development that are associated with morphogenesis, resemble an increased tumor burden of NeuT+ mice, which was previously confirmed with a large human RNA-seq dataset on breast cancer [39]. In both case studies we observed many associations with other types of cancers or diseases. While this may speak for a molecular and biological similarity, a certain publication bias, e.g. for cancer, is a confounding factor that skews the statistics [13].

Finally, we sought to improve accessibility of miEAA and develop a web-API in combination with a Python package and code examples. These features can also enhance its usability in other applications for miRNA research, for example to annotate functional sub-graphs in regulatory network analysis [40]. In conclusion, miEAA 2.0 is a flexible, comprehensive, and highly accessible tool for high-throughput miRNA annotation and enrichment analysis.

DATA AVAILABILITY

miEAA 2.0 is freely available at https://www.ccb.uni-saarland.de/mieaa2. No login is required. Example code for API-usage and a pre-compiled Python package are freely available from https://github.com/Xethic/miEAA-API.

SUPPLEMENTARY DATA

Supplementary Data are available online.

ACKNOWLEDGEMENTS

We thank the authors of the utilised GEO dataset for providing their microarray samples to the general public. The results shown here are in whole or part based upon data generated by the TCGA Research Network:

https://www.cancer.gov/tcga. We would like to express gratitude towards all specimen donors and research groups involved in the sample acquisition.

FUNDING

None declared.

Conflict of interest statement.

None declared.

References

[1] David P. Bartel. Metazoan MicroRNAs. *Cell*, 173(1):20–51, 2018. [PubMed:29570994] [PubMed Central:PMC6091663] [doi:10.1016/j.cell.2018.03.006].

[2] Fabian Kern, Christina Backes, Pascal Hirsch, Tobias Fehlmann, Martin Hart, Eckart Meese, and Andreas Keller. What's the target: understanding two decades of in silico microRNA-target prediction. *Briefings in Bioinformatics*, dec 2019. [PubMed:31792536] [doi:10.1093/bib/bbz111].

[3] Laura Cantini, Gloria Bertoli, Claudia Cava, Thierry Dubois, Andrei Zinovyev, Michele Caselle, Isabella Castiglioni, Emmanuel Barillot, and Loredana Martignetti. Identification of microRNA clusters cooperatively acting on epithelial to mesenchymal transition in triple negative breast cancer. *Nucleic Acids Research*, 47(5):2205–2215, jan 2019. [PubMed:30657980] [PubMed Central:PMC6412120] [doi:10.1093/nar/gkz016].

[4] Nicole Ludwig, Tobias Fehlmann, Fabian Kern, Manfred Gogol, Walter Maetzler, Stephanie Deutscher, Simone Gurlit, Claudia Schulte, Anna Katharina von Thaler, Christian Deuschle, Florian Metzger, Daniela Berg, Ulrike Suenkel, Verena Keller, Christina Backes, Hans Peter Lenhof, Eckart Meese, and Andreas Keller. Machine Learning to Detect Alzheimer's Disease from Circulating Non-coding RNAs. *Genomics, Proteomics and Bioinformatics*, 17(4):430–440, 2019. [PubMed:31809862] [PubMed Central:PMC6943763] [doi:10.1016/j.gpb.2019.09.004].

[5] Thomas Thomou, Marcelo A Mori, Jonathan M Dreyfuss, Masahiro Konishi, Masaji Sakaguchi, Christian Wolfrum, Tata Nageswara Rao, Jonathon N Winnay, Ruben Garcia-Martin, Steven K Grinspoon, Phillip Gorden, and C Ronald Kahn. Adipose-derived circulating miRNAs regulate gene expression in other tissues. *Nature*, 542(7642):450–455, 2017. [PubMed:28199304] [PubMed Central:PMC5330251] [doi:10.1038/nature21365].

[6] Christina Backes, Eckart Meese, and Andreas Keller. Specific miRNA Disease Biomarkers in Blood, Serum and Plasma: Challenges and Prospects. *Molecular Diagnosis and Therapy*, 2016. [PubMed:27378479] [doi:10.1007/s40291-016-0221-4].

[7] Tobias Fehlmann, Christina Backes, Julia Alles, Ulrike Fischer, Martin Hart, Fabian Kern, Hilde Langseth, Trine Rounge, Sinan Ugur Umu, Mustafa Kahraman, Thomas Laufer, Jan Haas, Cord Staehler, Nicole Ludwig, Matthias Hübenthal, Benjamin Meder, Andre Franke, Hans-Peter Lenhof, Eckart Meese, and Andreas Keller. A high-resolution map of the human small non-coding transcriptome. *Bioinformatics (Oxford, England)*, 34(10):1621–1628, may 2018. [PubMed:29281000] [doi:10.1093/bioinformatics/btx814].

[8] Julia Alles, Tobias Fehlmann, Ulrike Fischer, Christina Backes, Valentina Galata, Marie Minet, Martin Hart, Masood Abu-Halima, Friedrich A. Grässer, Hans Peter Lenhof, Andreas Keller, and Eckart Meese. An estimate of the total number of true human miRNAs. *Nucleic acids research*, 47(7):3353–3364, mar 2019. [PubMed:30820533] [PubMed Central:PMC6468295] [doi:10.1093/nar/gkz097].

[9] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 2005. [PubMed:16199517] [PubMed Central:PMC1239896] [doi:10.1073/pnas.0506580102].

[10] Antonio Mora. Gene set analysis methods for the functional interpretation of non-mRNA data—Genomic range and ncRNA data. *Briefings in Bioinformatics*, oct 2019. [PubMed:31612220] [doi:10.1093/bib/bbz090].

[11] Christina Backes, Qurratulain T. Khaleeq, Eckart Meese, and Andreas Keller. MiEAA: MicroRNA enrichment analysis and annotation. *Nucleic Acids Research*, 2016. [PubMed:27131362] [PubMed Central:PMC4987907] [doi:10.1093/nar/gkw345].

[12] Jeffrey M Perkel. Workflow systems turn raw data into scientific knowledge. *Nature*, 573(7772):149–150, sep 2019. [PubMed:31477884] [doi:10.1038/d41586-019-02619-z].

[13] Jianwei Li, Xiaofen Han, Yanping Wan, Shan Zhang, Yingshu Zhao, Rui Fan, Qinghua Cui, and Yuan Zhou. TAM 2.0: Tool for MicroRNA set analysis. *Nucleic Acids Research*, 46(W1):W180–W185, jun 2018. [PubMed:29878154] [PubMed Central:PMC6031048] [doi:10.1093/nar/gky509].

[14] Christina Backes, Tobias Fehlmann, Fabian Kern, Tim Kehl, Hans Peter Lenhof, Eckart Meese, and Andreas Keller. MiRCarta: A central repository for collecting miRNA candidates. *Nucleic Acids Research*, 46(D1):D160–D167, 2018. [PubMed:29036653] [PubMed Central:PMC5753177] [doi:10.1093/nar/gkx851].

[15] Hsi-Yuan Huang, Yang-Chi-Dung Lin, Jing Li, Kai-Yao Huang, Sirjana Shrestha, Hsiao-Chin Hong, Yun Tang, Yi-Gang Chen, Chen-Nan Jin, Yuan Yu, Jia-Tong Xu, Yue-Ming Li, Xiao-Xuan Cai, Zhen-Yu Zhou, Xiao-Hang Chen, Yuan-Yuan Pei, Liang Hu, Jin-Jiang Su, Shi-Dong Cui, Fei Wang, Yue-Yang Xie, Si-Yuan Ding, Meng-Fan Luo, Chih-Hung Chou, Nai-Wen Chang, Kai-Wen Chen, Yu-Hsiang Cheng, Xin-Hong Wan, Wen-Lian Hsu, Tzong-Yi Lee, Feng-Xiang Wei, and Hsien-Da Huang. miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Research*, 48(D1):D148–D154, oct 2019. [PubMed:31647101] [doi:10.1093/nar/gkz896].

[16] Tim Kehl, Fabian Kern, Christina Backes, Tobias Fehlmann, Daniel Stöckel, Eckart Meese, Hans Peter Lenhof, and Andreas Keller. miRPathDB 2.0: a novel release of the miRNA Pathway Dictionary Database. *Nucleic acids research*, 48(D1):D142–D147, nov 2020. [PubMed:31691816] [doi:10.1093/nar/gkz1022].

[17] Nicole Ludwig, Petra Leidinger, Kurt Becker, Christina Backes, Tobias Fehlmann, Christian Pallasch, Steffi Rheinheimer, Benjamin Meder, Cord Stähler, Eckart Meese, and Andreas Keller. Distribution of miRNA expression across human tissues. *Nucleic Acids Research*, 44(8):3865–3877, 2016. [PubMed:26921406] [PubMed Central:PMC4856985] [doi:10.1093/nar/gkw116].

[18] Zhou Huang, Jiangcheng Shi, Yuanxu Gao, Chunmei Cui, Shan Zhang, Jianwei Li, Yuan Zhou, and Qinghua Cui. HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Research*, 47(D1):D1013–D1017, oct 2018. [PubMed:30364956] [PubMed Central:PMC6323994] [doi:10.1093/nar/gky1010].

[19] Tianyu Cui, Lin Zhang, Yan Huang, Ying Yi, Puwen Tan, Yue Zhao, Yongfei Hu, Liyan Xu, Enmin Li, and Dong Wang. MNDR v2.0: an updated resource of ncRNA–disease associations in mammals. *Nucleic Acids Research*, 46(D1):D371–D374, nov 2017. [PubMed:29106639] [PubMed Central:PMC5753235] [doi:10.1093/nar/gkx1025].

[20] Zhan Tong, Qinghua Cui, Juan Wang, and Yuan Zhou. TransmiR v2.0: an updated transcription factor-microRNA regulation database. *Nucleic Acids Research*, 47(D1):D253–D258, oct 2018. [PubMed:30371815] [PubMed Central:PMC6323981] [doi:10.1093/nar/gky1023].

[21] Ting Zhang, Puwen Tan, Liqiang Wang, Nana Jin, Yana Li, Lin Zhang, Huan Yang, Zhenyu Hu, Lining Zhang, Chunyu Hu, Chunhua Li, Kun Qian, Changjian Zhang, Yan Huang, Kongning Li, Hao Lin, and Dong Wang. RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Research*, 45(D1):D135–D138, aug 2016. [PubMed:27543076] [PubMed Central:PMC5210605] [doi:10.1093/nar/gkw728].

[22] Francesco Russo, Sebastiano Di Bella, Federica Vannini, Gabriele Berti, Flavia Scoyni, Helen V Cook, Alberto Santos, Giovanni Nigita, Vincenzo Bonnici, Alessandro Laganà, Filippo Geraci, Alfredo Pulvirenti, Rosalba Giugno, Federico De Masi, Kirstine Belling, Lars J Jensen, Søren Brunak, Marco Pellegrini, and Alfredo Ferro. miRandola 2017: a curated knowledge base of non-invasive biomarkers. *Nucleic Acids Research*, 46(D1):D354–D359, sep 2017. [PubMed:29036351] [PubMed Central:PMC5753291] [doi:10.1093/nar/gkx854].

[23] Johannes Köster and Sven Rahmann. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012. [PubMed:22908215] [doi:10.1093/bioinformatics/bts480].

[24] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J van der Walt, Matthew Brett, Joshua Wilson, K Jarrod Millman, Nikolay Mayorov, Andrew R.˜J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, \.Ilhan Polat, Yu Feng, Eric W Moore, Jake Vand erPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.˜A. Quintero, Charles R Harris, Anne M Archibald, Antônio H Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1. 0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. [PubMed:32015543] [doi:10.1038/s41592-019-0686-2].

[25] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, 47(D1):D330–D338, nov 2018. [PubMed:30395331] [PubMed Central:PMC6323945] [doi:10.1093/nar/gky1055].

[26] Minoru Kanehisa, Yoko Sato, Miho Furumichi, Kanae Morishima, and Mao Tanabe. New approach for understanding genome variations in KEGG. *Nucleic Acids Research*, 47(D1):D590–D595, oct 2018. [PubMed:30321428] [PubMed Central:PMC6324070] [doi:10.1093/nar/gky962].

[27] Xueyi Teng, Xiaomin Chen, Hua Xue, Yiheng Tang, Peng Zhang, Quan Kang, Yajing Hao, Runsheng Chen, Yi Zhao, and Shunmin He. NPInter v4.0: an integrated database of ncRNA interactions. *Nucleic Acids Research*, 48(D1):D160–D165, oct 2019. [PubMed:31670377] [doi:10.1093/nar/gkz969].

[28] Ana Kozomara, Maria Birgaoanu, and Sam Griffiths-Jones. miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1):D155–D162, 2018. [PubMed:30423142] [PubMed Central:PMC6323917] [doi:10.1093/nar/gky1141].

[29] Harsh Dweep and Norbert Gretz. MiRWalk2.0: A comprehensive atlas of microRNA-target interactions. *Nature Methods*, 12(8):697, 2015. [PubMed:26226356] [doi:10.1038/nmeth.3485].

[30] The RNAcentral Consortium. RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Research*, 47(D1):D221–D229, nov 2018. [PubMed:30395267] [PubMed Central:PMC6324050] [doi:10.1093/nar/gky1034].

[31] Taosheng Xu, Ning Su, Lin Liu, Junpeng Zhang, Hongqiang Wang, Weijia Zhang, Jie Gui, Kui Yu, Jiuyong Li, and Thuc Duy Le. miRBaseConverter: an R/Bioconductor package for converting and retrieving miRNA name, accession, sequence and family information in different versions of miRBase. *BMC Bioinformatics*, 19(19):514, 2018. [PubMed:30598108] [PubMed Central:PMC6311916] [doi:10.1186/s12859-018-2531-5].

[32] Daniel Stöckel, Tim Kehl, Patrick Trampert, Lara Schneider, Christina Backes, Nicole Ludwig, Andreas Gerasch, Michael Kaufmann, Manfred Gessler, Norbert Graf, Eckart Meese, Andreas Keller, and Hans-Peter Lenhof. Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, 32(10):1502–1508, 2016. [PubMed:26787660] [doi:10.1093/bioinformatics/btv770].

[33] Keegan Korthauer, Patrick K Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, Eric J Alm, and Stephanie C Hicks. A practical guide to methods controlling false discoveries in computational biology. *Genome Biology*, 20(1):118, 2019. [PubMed:31164141] [PubMed Central:PMC6547503] [doi:10.1186/s13059-019-1716-1].

[34] Claudia Chiodoni, Valeria Cancila, Tiziana A Renzi, Milena Perrone, Andrea M Tomirotti, Sabina Sangaletti, Laura Botti, Matteo Dugo, Matteo Milani, Lucia Bongiovanni, Maurizio Marrale, Claudio Tripodo, and Mario P Colombo. Transcriptional Profiles and Stromal Changes Reveal Bone Marrow Adaptation to Early Breast Cancer in Association with Deregulated Circulating microRNAs. *Cancer Research*, 80(3):484 LP – 498, feb 2020. [PubMed:31776132] [doi:10.1158/0008-5472.CAN-19-1425].

[35] Paolo DI Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017. [PubMed:28398311] [doi:10.1038/nbt.3820].

[36] J Li, H Wu, W Li, L Yin, S Guo, X Xu, Y Ouyang, Z Zhao, S Liu, Y Tian, Z Tian, J Ju, B Ni, and H Wang. Downregulated miR-506 expression facilitates pancreatic cancer progression and chemoresistance via SPHK1/Akt/NF-κB signaling. *Oncogene*, 35(42):5501–5514, oct 2016. [PubMed:27065335] [PubMed Central:PMC5078861] [doi:10.1038/onc.2016.90].

[37] Linfei Zhang, Huadong Zhou, and Gang Wei. miR-506 regulates cell proliferation and apoptosis by affecting RhoA/ROCK signaling pathway in hepatocellular carcinoma cells. *International journal of clinical and experimental pathology*, 12(4):1163–1173, apr 2019. [PubMed:31933931] [PubMed Central:PMC6947048].

[38] Sajjad Karim, Jaudah A Al-Maghrabi, Hasan M A Farsi, Ahmad J Al-Sayyad, Hans-Juergen Schulten, Abdelbaset Buhmeida, Zeenat Mirza, Alaa A Al-boogmi, Fai T Ashgan, Manal M Shabaad, Hend F NourEldin, Khalid B M Al-Ghamdi, Adel Abuzenadah, Adeel G A Chaudhary, and Mohammed H Al-Qahtani. Cyclin D1 as a therapeutic target of renal cell carcinoma- a combined transcriptomics, tissue microarray and molecular docking study from the Kingdom of Saudi Arabia. *BMC Cancer*, 16(2):741, 2016. [PubMed:27766950] [PubMed Central:PMC5073805] [doi:10.1186/s12885-016-2775-2].

[39] Diana Tapia-Carrillo, Hugo Tovar, Tadeo Enrique Velazquez-Caldelas, and Enrique Hernandez-Lemus. Master Regulators of Signaling Pathways: An Application to the Analysis of Gene Regulation in Breast Cancer. *Frontiers in genetics*, 10:1180, dec 2019. [PubMed:31850059] [PubMed Central:PMC6902642] [doi:10.3389/fgene.2019.01180].

[40] Yannan Fan, Keith Siklenka, Simran K Arora, Paula Ribeiro, Sarah Kimmins, and Jianguo Xia. miRNet - dissecting miRNA-target interactions and functional associations through network-based visual analysis. *Nucleic Acids Research*, 44(W1):W135–W141, 2016. [PubMed:27105848] [PubMed Central:PMC4987881] [doi:10.1093/nar/gkw288].