Published online 1 September 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D1055–D1061***
https://doi.org/10.1093/nar/gkab733

DIANA-miTED: a microRNA tissue expression database

Ioannis Kavakiotis[1,*], Athanasios Alexiou[1,2], Spyros Tastsoglou[1,2], Ioannis S. Vlachos[3,4,5] and Artemis G. Hatzigeorgiou[1,2,*]

[1]DIANA-Lab, Dept. of Computer Science and Biomedical Informatics, Univ. of Thessaly, 35131 Lamia, Greece, [2]Hellenic Pasteur Institute, 11521 Athens, Greece, [3]Cancer Research Institute | Harvard Medical School Initiative for RNA Medicine, Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA, USA, [4]Harvard Medical School, Boston, MA, USA and [5]Broad Institute of MIT and Harvard, Cambridge, MA, USA

Received July 13, 2021; Editorial Decision August 6, 2021; Accepted August 27, 2021

ABSTRACT

microRNAs (miRNAs) are short (∼23nt) single-stranded non-coding RNAs that act as potent post-transcriptional gene expression regulators. Information about miRNA expression and distribution across cell types and tissues is crucial to the understanding of their function and for their translational use as biomarkers or therapeutic targets. DIANA-miTED is the most comprehensive and systematic collection of miRNA expression values derived from the analysis of 15 183 raw human small RNA-Seq (sRNA-Seq) datasets from the Sequence Read Archive (SRA) and The Cancer Genome Atlas (TCGA). Metadata quality maximizes the utility of expression atlases, therefore we manually curated SRA and TCGA-derived information to deliver a comprehensive and standardized set, incorporating in total 199 tissues, 82 anatomical sublocations, 267 cell lines and 261 diseases. miTED offers rich instant visualizations of the expression and sample distributions of requested data across variables, as well as study-wide diagrams and graphs enabling efficient content exploration. Queries also generate links towards state-of-the-art miRNA functional resources, deeming miTED an ideal starting point for expression retrieval, exploration, comparison, and downstream analysis, without requiring bioinformatics support or expertise. DIANA-miTED is freely available at http://www.microrna.gr/mited.

INTRODUCTION

microRNAs (miRNAs) are abundant regulatory RNAs that primarily guide the cleavage, degradation, and/or translational repression of their target transcripts (1). Suppression of gene expression at the post-transcriptional level is by and large attributed to miRNA function (2). To this end, the accurate cataloguing of their abundance across tissues and cell types is an indispensable tool towards the understanding of gene regulation and dysregulation in physiological and pathological conditions. Equally importantly, altered miRNA levels in tissues and biofluids between disease and healthy states, or within the course of a disease, can discriminate conditions or correlate with clinical phenotypes and outcomes, highlighting the diagnostic, prognostic, or predictive biomarker capabilities that specific miRNAs may carry (3,4).

Initially, limited numbers of known miRNAs were being quantified using low-yield methods, including Northern blotting and quantitative reverse transcriptase PCR (RT-qPCR), as well as miRNA-tailored microarrays (5–7). Advances in sequencing technologies (Next Generation Sequencing, NGS) greatly boosted data yields and provided a means to massively quantify miRNAs and discover novel ones (8). Currently, the state-of-the-art high throughput method to derive miRNA abundance estimates is small RNA sequencing (sRNA-Seq). Thousands of publicly available sRNA-Seq datasets have been deposited in raw or count format in repositories such as the Sequence Read Archive (SRA) (9) and Gene Expression Omnibus (GEO) (10), while large consortia such as The Cancer Genome Atlas (TCGA) have also generated vast datasets capturing miRNA abundance in thousands of patient samples (11).

Creating a comprehensive and consistent human atlas of miRNA expression is a challenging task, since (i) available datasets are provided for retrieval at various analysis levels (e.g. raw FASTQ files, count-level estimates, normalized expression values), they have been produced utilizing (ii) variable miRNA annotation sources and versions, as well as (iii) distinct quantification pipelines and algorithmic choices (11–14). Moreover, public repositories and individual submitters annotate sample- and study-specific meta-

*To whom correspondence should be addressed. Tel: +30 24210 74758; Fax: +30 24210 74997; Email: arhatzig@uth.gr
Correspondence may also be addressed to Ioannis Kavakiotis. Tel: +30 24210 74758; Fax: +30 24210 74997; Email: ikavakiotis@uth.gr

data in a minimally systematic manner, impeding the integration of datasets from various sources into one uniform database.

Cataloguing miRNA expression estimates, especially in disease, is a high value scientific endeavor with numerous potential applications in basic and translational research. Currently available implementations vary in scope, breadth, and functionality. miRmine (15), HMED (16) and DASHR2 (17), aim to capture a wide variety of sample types and tissues, and comprise 304, 401 and 802 datasets, respectively. miRmine allows single/multiple miRNA queries and outputs sample-level RPM matrices for cell lines or tissues. DASHR2 focuses on genomic localization of small RNAs, provides information regarding their sequence and secondary structure (by RNAfold), and features a static expression table for each small RNA. YM500v3 (18) focuses only on cancer with samples solely derived from TCGA. SEAWeb (19) is a web application offering various analyses, such as differential expression and classification, allowing users to upload their results to compare against the database content. SEAWeb contains 4258 datasets from 10 organisms (3360 human datasets). Finally, DeepBase v3.0 (20) divides data into two classes, namely 'Cancer data' (from TCGA and ICGC) and 'Tissue data' (500 datasets from SRA/GEO, GTEx, ENCODE). Importantly, miRNA expression values collected in DeepBase refer to precursors or miRNA host genes, falling short on providing abundance estimates of the functional mature miRNA forms in every analysis type it provides. At the time of writing, HMED and YM500v3 were not in a functional state.

Until today and as mentioned above, most available databases either comprise a small number of datasets or focus solely or almost solely on TCGA and this is mostly due to practical reasons. TCGA is a rather uniform resource, while GEO/SRA studies are extremely diverse in terms of library preparations, adapters utilized, and sample quality. DIANA-miTED comes to bridge this gap and enable researchers investigate miRNA expression across the widest variety of libraries to date, as well as to perform simple or sophisticated analyses from a single resource. The capabilities and entries of all available databases are presented in Table 1, which captures the number of entries from each data source, the fold-change increase of miTED against that specific source, as well as the access to tools and downstream analyses provided from each resource. miTED not only is the largest such database to date but it provides direct interconnections to the DIANA-tools cosmos, enabling researchers to perform target prediction, prioritization, and functional investigations from a single user interface.

To generate miTED, we performed pre-processing and analysis of >15 000 sRNA-Seq datasets retrieved from TCGA and SRA, utilizing DIANA-mAP analysis workflow (21). DIANA-mAP ensures fully automated A-to-Z uniform analysis of sRNA-Seq datasets, from the raw file to the expression estimates. Our results are presented in DIANA-miTED (miRNA Tissue Expression Database), an online atlas of miRNA expression in healthy and disease states (http://www.microrna.gr/mited). In miTED, users can retrieve expression values for one or multiple miRNAs and/or tissues/cell lines, identify top expressed miRNAs, or top tissues/cell lines where a miRNA of interest is expressed the most (Figure 1).

METHODS AND RESULTS

Data collection and curation

Raw sRNA-Seq datasets originated from two resources, namely NCBI-SRA (22) and the TCGA project (23). NCBI-SRA data were identified by selecting all entries with Library Strategy and organism fields as 'miRNA-Seq' and '*Homo sapiens*', respectively, and retrieved locally. Only datasets comprising tissue or cell line information metadata were included in the analysis. TCGA sRNA-Seq datasets were retrieved using the GDC Data Transfer Tool, along with the respective transcriptomic metadata from the GDC data portal (24), while all available sample and patient metadata were retrieved. Only sRNA-seq TCGA datasets annotated with both patient and transcriptomic metadata were included in the study. The selected datasets were analyzed with the well-defined sRNA-Seq analysis workflow DIANA-mAP (21), following strict quality controls (described in Section 'High-throughput data analysis'), resulting in a collection of 4142 NCBI-SRA and 11 041 TCGA analyzed datasets.

We manually curated both NCBI-SRA and TCGA-derived metadata information to deliver to users a coherent and standardized set. The resulting collection offers the following metadata information. 'Sample_ID' contains the sample's identification code in SRA (e.g. SRR1774098) or TCGA (e.g. TCGA-P8-A5KD-11A-11R-A35M-13). 'Collection' holds the origin collection of the sample (SRA or TCGA). 'Project_ID' contains the sample's project identification code in NCBI-SRA (e.g. PRJDB2675) or TCGA (e.g. TCGA-PCPG). All entries in miTED mandatorily contain the aforementioned information. Moreover, each entry must either feature 'Tissue or organ of origin' or 'Cell line' information. Depending on the initial metadata, the following information is provided: 'Tissue subregion' refers to a more specific anatomical location of the 'Tissue or organ of origin'. 'Disease' contains information about the disease, if available. Notably, disease meta-information has been annotated even in healthy samples (control patients) from the same study or from control tissue regions derived from the same patient (e.g. matched healthy and neoplastic tissue from a cancer patient).

High-throughput data analysis

Data processing was performed using DIANA-mAP (v1.0) (21), which is a fully automated computational pipeline with an emphasis to pre-processing, that allows users to perform miRNA NGS data analysis from raw data up to quantification and differential expression in an easy, scalable, efficient and intuitive way. The tool was developed in R and performed the pre-processing, alignment and quantification steps on all samples included in miTED.

Pre-processing utilized the external tools FastQC (v0.11.7) (25), DNApi (v1.1) (26), and Cutadapt (v1.16) (27). Raw datasets were quality checked allowing a minimum Phred score of 10, adapters were inferred when not provided and trimmed using an 18 bp minimum allowed

Table 1. Comparison of DIANA-miTED against existing resources cataloguing miRNA expression. Numbers of included datasets and their fold-difference compared to miTED, available expression units, version of miRBase and miRNA-relevant functionalities of each resource are presented. NA values correspond to cases in which accessing the resource or performing queries failed

Resource	Included datasets	Fold-increase in miTED	Expression units	miRBase version	miRNA-related functionalities	Links to tools and resources
miRmine	304	∼50	RPM	v21	Search by single/multiple miRNAs. Bulk data download.	miRBase, NCBI-SRA
HMED	401	∼38	RPM	v20	NA	NA
DASHR2	802	∼19	Counts, RPM	v19, v21	Search by single miRNA, coordinates, or sequence. Tissue specificity and read coverage information. Bulk data download.	UCSC Genome Browser
YM500v3	∼11 000	∼1.38	NA	v21	NA	NA
SEAWeb	4258	∼4.2	RPM	v21	Search by single/multiple miRNAs. Differential expression. Gene targets (miRTarBase). Disease associations. Result download.	miRBase, GeneCards, GEO, PubMed, Disease Ontology
DeepBase v3.0	∼11 500	∼1.31	RPM, $\log_2$(RPM)-mean	v22	Precursor-level miRNA expression (collected RPM values). Expression heatmaps. Query by dataset selection. Bulk data download.	UCSC Genome Browser, PubMed
DIANA-miTED	15 183	-	Counts, RPM, $\log_2$(RPM)	v22	Search by single/multiple miRNAs and/or single/multiple tissues or cell lines. Filter disease/healthy. Search for top expressed miRNAs. Search for top sites by miRNA abundance. Result download.	DIANA-tools (microT-CDS, TarBase, LncBase, miRPath), miRBase

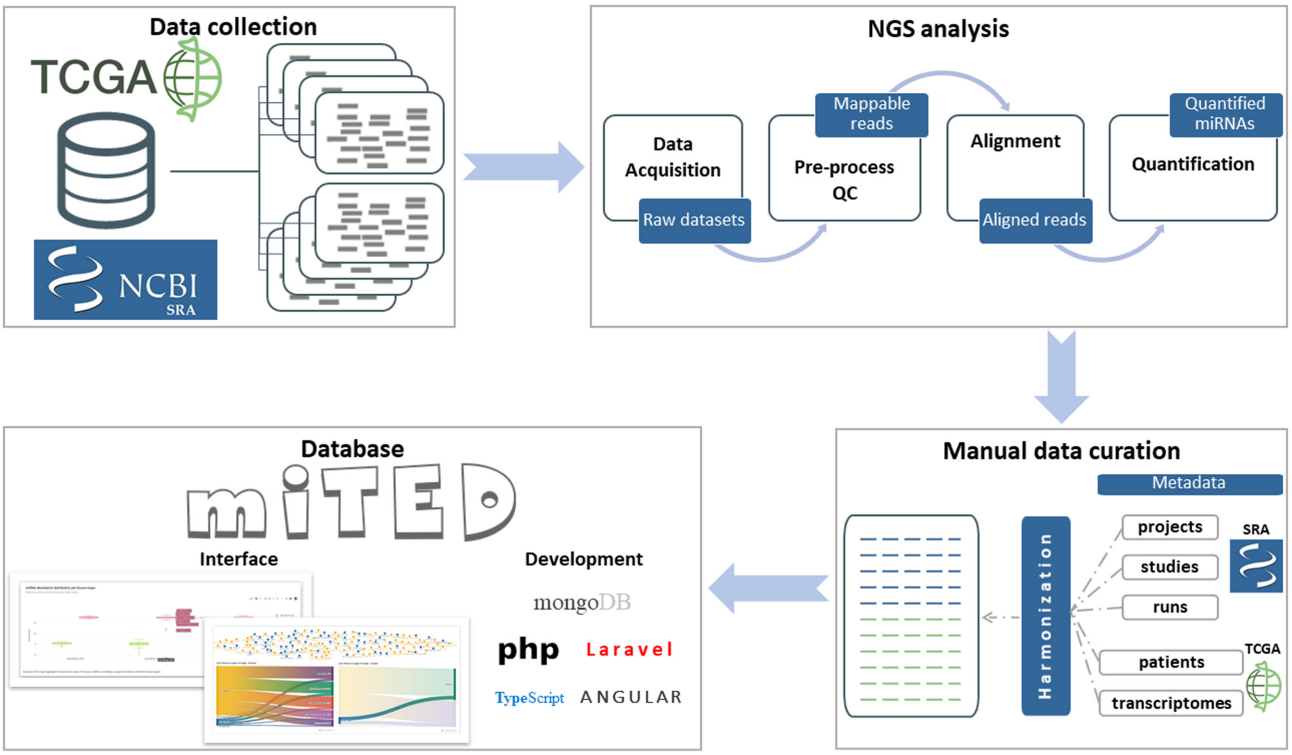

Figure 1. DIANA-miTED development workflow. Initially, human raw sRNA-Seq datasets were retrieved from NCBI-SRA and TCGA (alignment files obtained from TCGA and converted back to FASTQ format). Raw datasets were uniformly subjected to pre-processing and quality control, alignment and quantification. Read count, RPM and $\log_2$(RPM) values were calculated for miRBase miRNAs. Metadata from both resources were curated manually to create a comprehensive, standardized set of metadata annotations for the analyzed datasets. DIANA-miTED resource was developed utilizing MongoDB (noSQL database), PHP/Laravel for data access layer development, and Typescript/Angular for the application layer development. miTED features extensive query, filtering and visualization options and supports local retrieval of requested data.

read length after trimming. Reads were subsequently aligned to the GRCh38 human genome assembly using Bowtie (v1.1.1) (28), allowing up to five multi-maps per read. Finally, the aligned reads were quantified through miRDeep2 (v0.0.8) (12), using the miRBase v22 hairpin and mature human miRNA forms (29), and allowing one mismatch on the precursor during alignment. The remaining DIANA-mAP and external tools' parameters, were set to their default values.

NCBI-SRA datasets were analyzed utilizing adapter information derived from metadata when available, although such information was very scarce. sRNA-Seq datasets at times exhibit especially low mapping rates and even less than half of the reads can be assigned to miRNAs, due to sample contamination or some other biological explanation, yet quantification results can still be considered robust (30). In our case, thousands of samples were analyzed in bulk and adapter information was missing in numerous cases. In order to avoid the inclusion of erroneous results due to incorrect adapter inference, we opted to conservatively only include samples which had at least 50% of their pre-processed reads assigned to known miRNAs. The same parameters were used for the analysis of the TCGA datasets, resulting in an average 79.2% [76.6–81.9, 95% confidence interval] rate or pre-processed reads assigned to known miRNAs. Read counts and read per million (RPM) units were directly retrieved from DIANA-mAP results, while $\log_2$(RPM + 1) values were calculated specifically for visualization purposes. The addition of 1 across all RPM values during log-transformation was performed to avoid undefined values resulting from log-transforming zeros.

Database architecture and implementation

DIANA-miTED is a NoSQL database built using the MVC architecture and hosted on Apache HTTP server 2.4. The data access layer (back-end) consists of MongoDB (https://www.mongodb.com/) and the PHP framework Laravel 8 (https://laravel.com/) (PHP 7.2), while the presentation layer (front-end) is designed using Angular 9.1 (https://angular.io/) and the Angular Material UI library (https://material.angular.io/). miTED's data are stored in NoSQL database collections and Laravel handles the connection to them for storing and/or retrieval. Finally, on the presentation layer, database statistics are presented using Chart JS (https://www.chartjs.org/) and Plotly JavaScript Open Source Graphing Library (https://plotly.com/javascript/), while Flourish (https://flourish.studio/) is utilized for the more complex visualizations.

Database content

The analysis of >120 billion reads from 15 183 sRNA-seq datasets, 4142 (27.3%) of which were retrieved from NCBI-SRA and 11 041 (72.7%) from TCGA, yielded >120 million expression values (120 978 144), as read counts, reads per million (RPM), and $\log_2$(RPM). The database comprises 12 400 datasets from disease samples from 261 human pathologies, as well as 1386 datasets from healthy samples. 14 146 samples are derived from 199 tissues/organs and 1037 samples from 267 cell lines. Sex ratio is balanced within miTED, with 6751 female-, 6123 male-derived samples (2309 unannotated entries).

miTED INTERFACE

Querying miTED. A friendly online graphical user interface was implemented to enable the users search, browse, but also meta-analyze this extensive collection without requiring bioinformatics support or expertise. DIANA-miTED offers three main query pages, namely *Multi-query*, *Top-miRNAs* and *Top-sites* through the *Querying DB* top menu. The *Multi-query* page offers the ability to explore and compare the expression of specific miRNAs in tissues or cell lines. The *Top-miRNAs* page returns the top expressed miRNAs in a specific tissue or cell line. Finally, the *Top-sites* page provides the tissues or cell lines where a specific query miRNA is the most expressed. All generated results and plots can be downloaded without the need for login, application, or verification procedures, through the dedicated download buttons (both for data tables and plots).

In the *Multi-query* page, users can perform queries, retrieve, and compare the expression of one or more (even all) miRNAs in tissues or cell lines. The dedicated search boxes allow free-text search and selection of specific tissues or cell lines and miRNAs (Figure 2A). The *Multi-query* form gives the opportunity to restrict a search to specific diseases, include only results from SRA or TCGA data collections, retrieve data depending on health status (i.e. 'Healthy' or 'Disease'), and choose the appropriate expression unit among read counts, RPM and $\log_2$(RPM). Results are organized into three distinct sections. (A) The first section is dedicated to visualizing the retrieved results. Grouped boxplots enable the comparison of miRNA abundance in specific tissues/diseases (Figure 2C). Moreover, sample distributions are explored through a Sankey diagram depicting Tissue–Disease relationships and pie charts for gender, collection and health status (Figure 2D). (B) The second section caters the interconnection of miTED results with related DIANA resources (tools and databases) for each miRNA. miTED provides for each input miRNA hyperlinks towards DIANA-microT-CDS (31), a web server of predicted miRNA targets, DIANA-Tarbase v.8 (32) a reference database of experimentally supported miRNA targets, DIANA-LncBase v.3 (33), a reference repository with experimentally supported miRNA targets on long non-coding RNAs, and DIANA-miRPath v.3 (34), a web server dedicated to the assessment of miRNA regulatory roles and the identification of controlled pathways. (C) Finally, in the third section, a data table with sample metadata as well as the expression of the user requested miRNAs, as described in Materials and Methods section, is provided (Figure 2B).

The *Top-miRNAs* page is the second query page in miTED resource. Through this page users can search for the top expressed miRNAs in a specific Tissue or Cell line. The displayed results include a data table showing the expression of all miRNAs in descending order and a bar chart depicting the top expressed miRNAs in the desired tissue or cell line.

Top-sites page is dedicated to retrieving Tissues or Cell lines where a specific miRNA is the most abundant. Similar to the *Top-miRNAs* page, results include a table containing

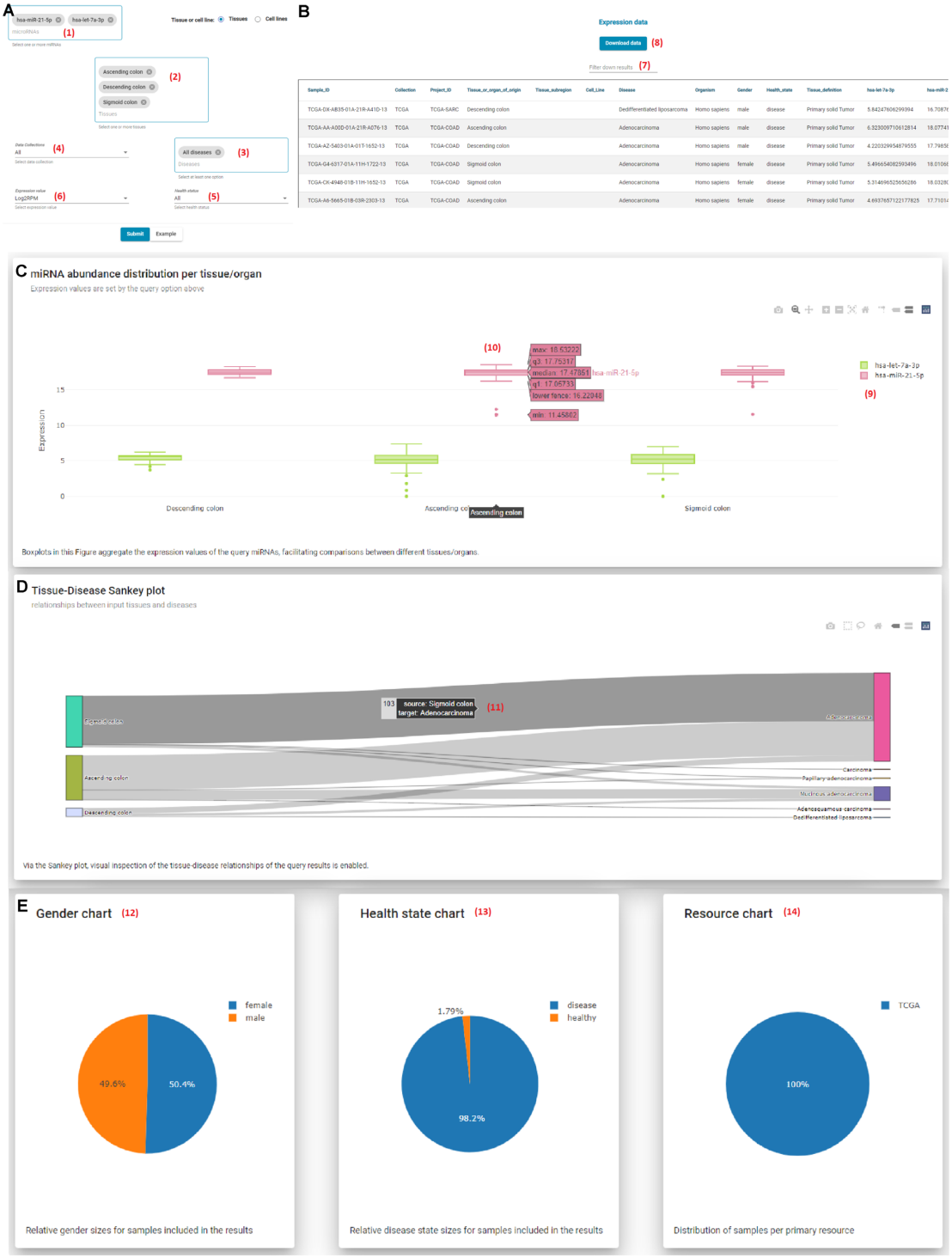

Figure 2. Multi-query page interface. (**A**) Submission form. Users may search for one or more of 2656 miRNAs (1) and Tissues (2). Both miRNAs and Tissue query boxes support free text search. Through filtering options, users may restrict their query to specific Diseases (3), Collections (4) and Health status (5). Via the Expression value drop-down menu (6), users can choose the desired expression unit that will be returned (read counts, RPM, $\log_2$(RPM)). (**B**) Results table. All entries compliant to the applied criteria are returned, along with their metadata and the expression values of the selected miRNAs. The results list can be customized to show 20, 50, 100, 150 and 200 items per page. A useful word-based filter, Filter-down results (7), has been implemented to narrow-down the returned entries and focus on these that contain a very specific term of interest. Users can retrieve the results of their query in tab-delimited format by clicking on the Download data button (8), without the need for any sign-up, application, or verification procedure. (**C**) Interactive boxplot showing the miRNA abundance distribution per tissue/organ. Users can select (9) which miRNAs are visible in the diagram offering direct comparison among miRNAs. On hover, (10) boxplots reveal the corresponding boxplot statistics (minimum, maximum, median, lower fence, first quartile and third quartile). (**D**) Interactive Sankey diagrams enable visual inspection of the tissue-disease relationships of the query results. On hover, (11) users may explore in more detail the distribution of samples. (**E**) Pie charts offer visual representation of the distribution of samples across Gender (12), Health state (13) and Collection (14) variables.

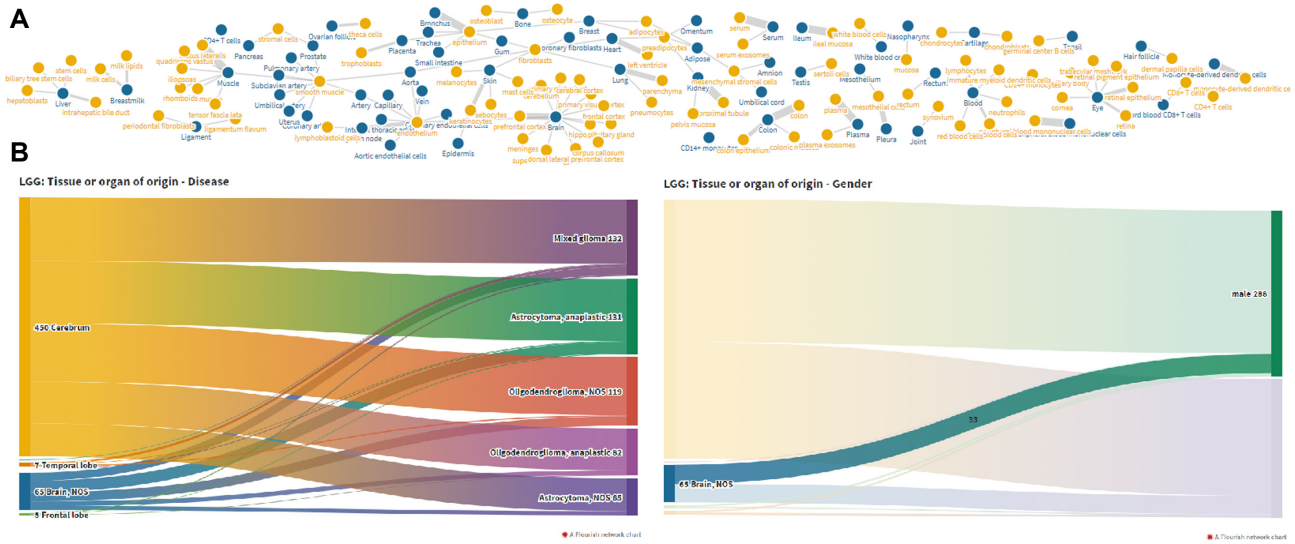

Figure 3. Visualizations. (**A**) Interactive graph network relating tissues—organs and tissue subregions in samples included in miTED. Users may explore the graph and highlight nodes of interest, revealing most/least populated tissues and organs. (**B**) Interactive Sankey diagrams depicting relationships between 'Tissue or organ of origin' and Disease and 'Tissue or organ of origin' and Gender. On hover, users can explore the distribution of samples per category.

expression values of Tissues/Cell lines in descending order and a bar chart depicting the top tissues or cell lines where the input is expressed the most.

Visualization options in miTED

DIANA-miTED also provides three visualization pages via its *Visualizations* menu. The first page, 'Tissue - Subregions | Graph Network', provides a graph network depicting the relations between Tissue or organ of origin and Tissue subregions (Figure 3A). It is an interactive graph offering the ability to highlight and move nodes, in order to explore degrees of interconnection between them. 'TCGA Projects Exploration' page contains Sankey diagrams for exploring relationships Tissue-Disease and Tissue-Gender of the TCGA datasets separately (Figure 3B). Sankey diagrams are also interactive, enabling the exploration of sample distributions in each category. Finally, 'DB statistics' page provides supplemental graphs depicting the overall database content.

FUTURE WORK

DIANA-miTED is the first version of an effort to provide a standardized set of all the available sRNA-Seq datasets analyzed though a well-defined pipeline. Future updates of the database will ensure it remains timely and relevant, constantly integrating and providing new available datasets analyzed and annotated in a uniform and accurate manner. We believe that miTED will prove to be particularly useful to users who are not non-coding RNA bioinformatics specialists, enabling them to easily retrieve compare and explore expression metrics and proceed with downstream in silico functional analyses.

ACKNOWLEDGEMENTS

The results published here are in part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov. We acknowledge use of controlled-access data from dbGaP Study Accession phs000178 – General Research Use – under dbGaP project #23441.

FUNDING

Personal postdoctoral fellowship to Ioannis Kavakiotis that was carried out under the call 'call for interest for postdoctoral researchers, scholarship for postdoctoral research' of University of Thessaly, that is implemented by University of Thessaly and funded by the 'Stavros Niarchos Foundation'. Funding for open access charge: Stavros Niarchos Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
2. O'Brien,J., Hayder,H., Zayed,Y. and Peng,C. (2018) Overview of MicroRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol.*, **9**, 402.
3. Condrat,C.E., Thompson,D.C., Barbu,M.G., Bugnar,O.L., Boboc,A., Cretoiu,D., Suciu,N., Cretoiu,S.M. and Voinea,S.C. (2020) miRNAs as biomarkers in disease: Latest findings regarding their role in diagnosis and prognosis. *Cells*, **9**, 276.
4. Wang,J., Chen,J. and Sen,S. (2016) MicroRNA as biomarkers and diagnostics. *J. Cell. Physiol.*, **231**, 25–30.
5. Koscianska,E., Starega-Roslan,J., Sznajder,L.J., Olejniczak,M., Galka-Marciniak,P. and Krzyzosiak,W.J. (2011) Northern blotting analysis of microRNAs, their precursors and RNA interference triggers. *BMC Mol. Biol.*, **12**, 14.
6. Liu,C.G., Calin,G.A., Volinia,S. and Croce,C.M. (2008) MicroRNA expression profiling using microarrays. *Nat. Protoc.*, **3**, 563–578

7. Chen,C., Tan,R., Wong,L., Fekete,R. and Halsey,J. (2011) Quantitation of microRNAs by real-time RT-qPCR. In: *PCR Protocols*. Humana Press, pp. 113–134.
8. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–D73.
9. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database Collaboration (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
10. Clough,E. and Barrett,T. (2016) The Gene Expression Omnibus Database. *Methods Mol. Biol.*, **1418**, 93–110.
11. Chu,A., Robertson,G., Brooks,D., Mungall,A.J., Birol,I., Coope,R., Ma,Y., Jones,S. and Marra,M.A. (2016) Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Res.*, **44**, e3.
12. Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
13. Aparicio-Puerta,E., Lebrón,R., Rueda,A., Gómez-Martín,C., Giannoukakos,S., Jaspez,D., Medina,J.M., Zubkovic,A., Jurak,I., Fromm,B. *et al.* (2019) sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression. *Nucleic Acids Res.*, **47**, W530–W535.
14. Rozowsky,J., Kitchen,R.R., Park,J.J., Galeev,T.R., Diao,J., Warrell,J., Thistlethwaite,W., Subramanian,S.L., Milosavljevic,A. and Gerstein,M. (2019) exceRpt: A comprehensive analytic platform for extracellular RNA profiling. *Cell Syst.*, **8**, 352–357.
15. Panwar,B., Omenn,G.S. and Guan,Y. (2017) miRmine: a database of human miRNA expression profiles. *Bioinformatics*, **33**, 1554–1560.
16. Gong,J., Wu,Y., Zhang,X., Liao,Y., Sibanda,V.L., Liu,W. and Guo,A.Y. (2014) Comprehensive analysis of human small RNA sequencing data provides insights into expression profiles and miRNA editing. *RNA biology*, **11**, 1375–1385.
17. Kuksa,P.P., Amlie-Wolf,A., Katanić,Ž., Valladares,O., Wang,L.S. and Leung,Y.Y. (2019) DASHR 2.0: integrated database of human small non-coding RNA genes and mature products. *Bioinformatics*, **35**, 1033–1039.
18. Chung,I.F., Chang,S.J., Chen,C.Y., Liu,S.H., Li,C.Y., Chan,C.H., Shih,C.C. and Cheng,W.C. (2017) YM500v3: a database for small RNA sequencing in human cancer research. *Nucleic Acids Res.*, **45**, D925–D931.
19. Rahman,R.U., Liebhoff,A.M., Bansal,V., Fiosins,M., Rajput,A., Sattar,A., Magruder,D.S., Madan,S., Sun,T., Gautam,A. *et al.* (2020) SEAweb: the small RNA Expression Atlas web application. *Nucleic Acids Res.*, **48**, D204–D219.
20. Xie,F., Liu,S., Wang,J., Xuan,J., Zhang,X., Qu,L., Zheng,L. and Yang,J. (2021) deepBase v3.0: expression atlas and interactive analysis of ncRNAs from thousands of deep-sequencing data. *Nucleic Acids Res.*, **49**, D877–D883.
21. Alexiou,A., Zisis,D., Kavakiotis,I., Miliotis,M., Koussounadis,A., Karagkouni,D. and Hatzigeorgiou,A.G. (2020) DIANA-mAP: analyzing miRNA from raw NGS data to quantification. *Genes*, **12**, 46.
22. Leinonen,R., Sugawara,H., Shumway,M. and International Nucleotide Sequence Database Collaboration (2011) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
23. Cancer Genome Atlas Research Network, Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
24. Grossman,R.L., Heath,A.P., Ferretti,V., Varmus,H.E., Lowy,D.R., Kibbe,W.A. and Staudt,L.M. (2016) Toward a shared vision for cancer genomic data. *N. Engl. J. Med.*, **375**, 1109–1112.
25. Andrews,S. (2010) In: *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Babraham Institute, Cambridge, UK.
26. Tsuji,J. and Weng,Z. (2016) DNApi: a de novo adapter prediction algorithm for small RNA sequencing data. *PLoS One*, **11**, e0164228.
27. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, **17**, 10–12
28. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
29. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
30. Tarallo,S., Ferrero,G., Gallo,G., Francavilla,A., Clerico,G., Realis Luc,A., Manghi,P., Thomas,A.M., Vineis,P., Segata,N. *et al.* (2019) Altered fecal small RNA profiles in colorectal cancer reflect gut microbiome composition in stool samples. *mSystems*, **4**, e00289-19.
31. Paraskevopoulou,M.D., Georgakilas,G., Kostoulas,N., Vlachos,I.S., Vergoulis,T., Reczko,M., Filippidis,C., Dalamagas,T. and Hatzigeorgiou,A.G. (2013) DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.*, **41**, W169–W173.
32. Karagkouni,D., Paraskevopoulou,M.D., Chatzopoulos,S., Vlachos,I.S., Tastsoglou,S., Kanellos,I., Papadimitriou,D., Kavakiotis,I., Maniou,S., Skoufos,G. *et al.* (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.*, **46**, D239–D245.
33. Karagkouni,D., Paraskevopoulou,M.D., Tastsoglou,S., Skoufos,G., Karavangeli,A., Pierros,V., Zacharopoulou,E. and Hatzigeorgiou,A.G. (2020) DIANA-LncBase v3: indexing experimentally supported miRNA targets on non-coding transcripts. *Nucleic Acids Res.*, **48**, D101–D110.
34. Vlachos,I.S., Zagganas,K., Paraskevopoulou,M.D., Georgakilas,G., Karagkouni,D., Vergoulis,T., Dalamagas,T. and Hatzigeorgiou,A.G. (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.*, **43**, W460–W466.