



Check for updates

## SOFTWARE TOOL ARTICLE

# miND (miRNA NGS Discovery pipeline): a small RNA-seq analysis pipeline and report generator for microRNA biomarker discovery studies

[version 1; peer review: 2 approved with reservations]

Andreas Diendorfer , Kseniya Khamina , Marianne Pultar, Matthias Hackl

TAmiRNA GmbH, Vienna, Austria

**v1** First published: 24 Feb 2022, 11:233  
<https://doi.org/10.12688/f1000research.94159.1>

Latest published: 24 Feb 2022, 11:233  
<https://doi.org/10.12688/f1000research.94159.1>

## Abstract

In contrast to traditional methods like real-time polymerase chain reaction, next-generation sequencing (NGS), and especially small RNA-seq, enables the untargeted investigation of the whole small RNAome, including microRNAs (miRNAs) but also a multitude of other RNA species. With the promising application of small RNAs as biofluid-based biomarkers, small RNA-seq is the method of choice for an initial discovery study. However, the presentation of specific quality aspects of small RNA-seq data varies significantly between laboratories and is lacking a common (minimal) standard.

The miRNA NGS Discovery pipeline (miND) aims to bridge the gap between wet lab scientist and bioinformatics with an easy to setup configuration sheet and an automatically generated comprehensive report that contains all essential qualitative and quantitative results that should be reported. Besides the standard steps like preprocessing, mapping, visualization, and quantification of reads, the pipeline also incorporates differential expression analysis when given the appropriate information regarding sample groups.

Although miND has a focus on miRNAs, other RNA species like tRNAs, piRNA, snRNA, or snoRNA are included and mapping statistics are available for further analysis. miND has been developed and tested on a multitude of data sets with various RNA sources (tissue, plasma, extracellular vesicles, urine, etc.) and different species.

miND is a Snakemake based pipeline and thus incorporates all advantages using a flexible workflow management system. Reference databases are downloaded, prepared and built with an included (but separate) workflow and thus can easily be updated to the most recent version but also stored for reproducibility.

In conclusion, the miND pipeline aims to streamline the bioinformatics processing of small RNA-seq data by standardizing the processing from raw data to a final, comprehensive and reproducible report.

## Open Peer Review

### Approval Status ? ?

	1	2
<b>version 1</b>		
24 Feb 2022	<a href="#">view</a>	<a href="#">view</a>

1. **Kristian Almstrup** , Copenhagen  
 University Hospital - Rigshospitalet,  
 Copenhagen, Denmark

**Nina Mørup** , Copenhagen University  
 Hospital - Rigshospitalet, Copenhagen,  
 Denmark

**Ailsa Maria Main**, Copenhagen University  
 Hospital - Rigshospitalet, Copenhagen,  
 Denmark

2. **Francisco J. Enguita** , Universidade de  
 Lisboa, Lisboa, Portugal

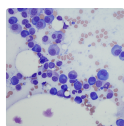
Any reports and responses or comments on the article can be found at the end of the article.

**Keywords**

microRNA, Next-Generation Sequencing, differential expression, smallRNA sequencing, biomarkers, spike-in, discovery study



This article is included in the **Bioinformatics** gateway.



This article is included in the **Cell & Molecular Biology** gateway.

**Corresponding author:** Matthias Hackl ([matthias.hackl@tamirna.com](mailto:matthias.hackl@tamirna.com))

**Author roles:** **Diendorfer A:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation; **Khamina K:** Conceptualization, Investigation, Writing – Review & Editing; **Pultar M:** Methodology, Software, Visualization, Writing – Review & Editing; **Hackl M:** Conceptualization, Data Curation, Funding Acquisition, Methodology, Project Administration, Resources, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** The TransBioLine project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 821283. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This communication reflects the authors' view and neither IMI nor the European Union or EFPIA are responsible for any use that may be made of the information contained therein.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2022 Diendorfer A *et al.* This is an open access article distributed under the terms of the **Creative Commons Attribution License**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Diendorfer A, Khamina K, Pultar M and Hackl M. **miND (miRNA NGS Discovery pipeline): a small RNA-seq analysis pipeline and report generator for microRNA biomarker discovery studies [version 1; peer review: 2 approved with reservations]** F1000Research 2022, 11:233 <https://doi.org/10.12688/f1000research.94159.1>

**First published:** 24 Feb 2022, 11:233 <https://doi.org/10.12688/f1000research.94159.1>

## Introduction

Small RNA-seq has been a well-established tool for the quantification of short RNA molecules like microRNAs (miRNAs) in various biofluids (Murillo *et al.*, 2019). Those short RNA molecules (17 to 25nt) play an important role in the cellular regulation of gene expression by interacting with specific complementary sites in targeted messenger RNAs (mRNAs). mRNAs that contain these target sites are then either down- or (rarely) up-regulated, resulting in a regulatory effect on the downstream translation of the mRNA (O'Brien *et al.*, 2018). In this context, miRNAs are part of a complex regulatory network where their expression does not only affect other mRNAs, but also the expression of miRNAs themselves is highly controlled (Lee & Ambros, 2001). Thus, the levels of miRNAs can be indicators of a cell's regulatory state and correlate with an organism's health status. For example the liver specific miR-122-5p was shown to be a suitable marker for liver injury when measured in serum or plasma (Llewellyn *et al.*, 2021) and as part of a miRNA expression signature can even be used to predict recovery after liver resection (Starlinger *et al.*, 2019).

This makes them interesting targets as biomarkers in liquid biopsy (Larrea *et al.*, 2016). But the search for miRNAs or miRNA signatures that are suitable as biomarkers requires a very specialized approach regarding computational biology. As next-generation sequencing (NGS) is often used in the discovery phase of studies (de Ronde *et al.*, 2018), a standardized and specialized analysis pipeline is highly important. Existing tools and pipelines for small RNA NGS analysis (like miRDeep2, miRExpress, miRNAkey, and sRNAbench) mostly focus on single steps, like the quantification of miRNAs or differential expression profiling, but either not provide additional analysis such as unsupervised analysis methods and data quality checks, or only in hard to interpret and inaccessible ways (Friedländer *et al.*, 2012; Wang *et al.*, 2009; Ronen *et al.*, 2010; Aparicio-Puerta *et al.*, 2019).

With the need for a standardized report that contains all relevant data and an initial statistical analysis, we developed a small RNA-seq data processing pipeline that not only provides one centralized report with all relevant information, but also bridges the gap between biologists and bioinformaticians with very easy to prepare data submission files as input and a detailed and well documented and interactive report as output.

In this study, we developed a robust and portable analysis pipeline for NGS data with a focus on biomarkers in discovery studies. With this in mind, we targeted the following goals: (1) standardized data inputs, (2) reproducible analysis, and (3) ease of use for both bioinformaticians and study statisticians (including publication ready figures and a clear and intuitive representation of results).

The miND pipeline can be used on many operating systems and in various setups with the only requirement of being able to run Snakemake workflows (Köster & Rahmann, 2012). Wrapper scripts for startup of the pipeline on Linux based systems are provided which can be adapted for the use on different platforms.

## Methods

### Implementation

The pipeline is based on Snakemake (Köster & Rahmann, 2012), a scalable bioinformatics workflow engine which incorporates many features needed for reproducible computational analysis (Mölder *et al.*, 2021). This includes handling the installation and provisioning of software tools via conda ("Anaconda Software Distribution," 2020) and bioconda (Grüning *et al.*, 2018) and overall the orchestration of individual steps of the pipeline to optimize usage of limited resources like central processing unit (CPU) and memory. Configuration files in yml format are used and contain settings for multithreading to adapt the pipeline for various computing platforms (Diendorfer *et al.*, 2022).

### Use case

An example protocol demonstrating the analysis of a public data set is available at protocols.io under the name **miND pipeline AWS EC2 installation and setup V.2** and can be reproduced not only as a guide for following data analysis, but also to setup the pipeline and data repository. The protocol describes the setup in an Amazon Web Services EC2 (Amazon Web Services, Inc, 2015) instance but has also been developed and tested on other platforms and systems. Only operating system specific parts would have to be adapted (e.g., installation of tools like git or wget would be done via apt on Debian based Linux distributions). For scientists interested in running the miND pipeline themselves, it is highly recommended to follow the provided protocol with the example data before running analysis on their own data sets.

The generated miND report for this example data set is available on [GitHub](#).

### Operation

The miND pipeline was developed and tested on Debian Linux (v11.2) running Snakemake (v6.0.5) and conda (v.4.10.3). The hardware requirements depend on the size of analyzed datasets, but in general it is recommended to provide at least 4 CPU cores and 8GB of memory. The pipeline will scale according to the available resources.

### Data repository

The pipeline requires data from three reference data sets: (1) host genomes from ENSEMBL (Zerbino *et al.*, 2018), (2) RNA sequences from RNAcentral (Sweeney *et al.*, 2019), and (3) miRNA mature and precursor sequences from miRbase (Griffiths-Jones, 2004).

In order to download and prepare these datasets in the formats and structures required, miND provides separate workflows to build the data repository. These workflows can be executed with a shell script that will read configurations for each data source and then download, format and build the reference databases based on Snakemake workflows.

The data repository only has to be built once and will then provide the data needed for all future miND analysis runs. In case of updates of reference data sets, the repository can be rebuilt or extended by adding sources to the configuration files and running the build script again.

### NGS raw data and metadata file

The miND pipeline requires two types of data for each experiment: raw NGS data and a meta data file with additional sample information. Raw data can be supplied either in fastq, fastq.gz or BAM (without alignments) files. The given format will be detected based on the file extensions.

Experimental meta data and details about the samples is provided in a XLS file containing three sheets: (1) Project details sheet, with general information and data of the project. This includes project title and comments but also settings relevant for the processing of the data like the sample species, adapter sequences, and cutoff levels for significance and quality filtering. (2) Sample group matrix sheet, which lists all samples that are part of this experiment and links them to additional group information. Up to five grouping variables can be set with unlimited levels each. The last sheet contains the (3) Contrast selection and allows the selection of groups and group-combinations based on the data provided in the sample group matrix sheet. The contrasts selected here will be used for the differential expression analysis.

### Pipeline analysis steps

The overall flow of data through the pipeline is shown in Figure 1. This flow diagram outlines the most important steps of data processing in the miND pipeline, especially the quality control steps with FastQC (Andrews, 2010) and multiQC (Ewels *et al.*, 2016), followed by hierarchical mapping using bowtie1 (Langmead *et al.*, 2009) and miRDeep2 (Friedländer *et al.*, 2012), where either mapped or unmapped reads are further processed by the next step. The final “R scripts processing” step includes multiple scripts that preprocess and analyze that data (including mapping statistics, unsupervised analysis methods and differential expression analysis) to then generate an interactive HTML report based on R markdown.

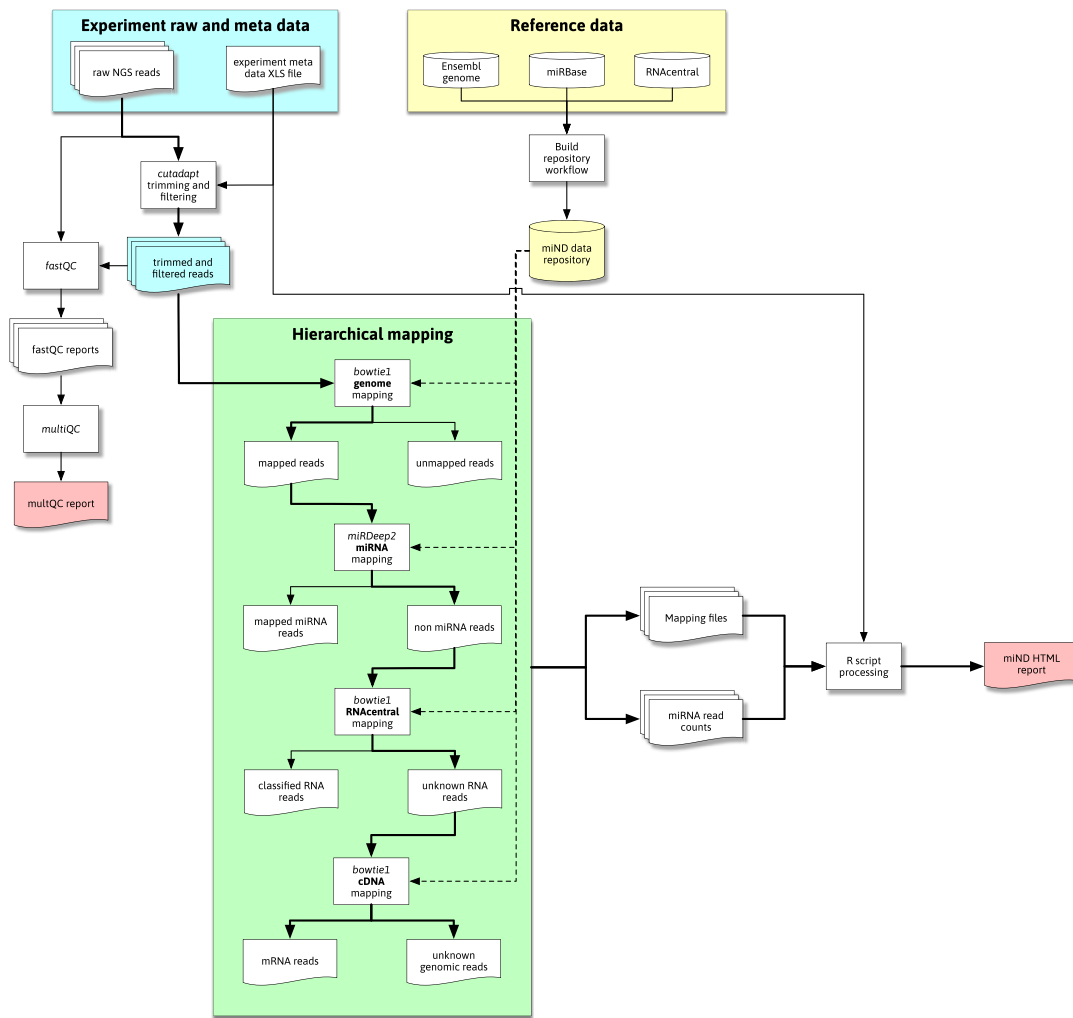
The hierarchical mapping uses genome datasets from the prepared data repository (generated once before the initial run as described in the “Data repository” subsection) in a first step to filter out reads that do not map to the host organism’s genome (bowtie1, allowing for two mismatches). The genome-mapped reads are further processed by miRDeep2 to accurately quantify miRNAs. To identify further remaining (genome mapping but non-miRNA) reads, bowtie1 is used to first map against the RNAcentral database and then complementary DNA sequences (to assign mRNA reads), both steps allowing for one mismatch. Reads that remain unmapped after these hierarchical clustering are classified as either “unknown genomic” (if they mapped against the host genome) or “unmapped” (in case of reads that did not map against the host genome and were thus filtered in the first mapping step). The generated mapping files are processed by R scripts to prepare mapping statistics for the different RNA species in each sample.

The mapping process focuses on miRNAs and prioritizes them by using the specialized mapping tool miRDeep2 directly after an initial genome mapping step. It utilizes bowtie1 for mapping of the reads but performs a more sophisticated assignment of miRNA IDs to the reads. This includes detailed information of isomiRs (mature miRNAs with highly similar sequences) that is prepared for further analysis steps.

For the identification of other RNA species RNAcentral is used. This comprehensive database contains non-coding RNA (ncRNA) sequences from a broad range of species. This step focuses on the classification of reads and uses bowtie1 (allowing for one mismatch) reporting the first (best) hit. This limits the use of the mapping data to the required classification, as reads could map to multiple references which are not reported mainly for performance reasons.

### Differential expression and independent filtering

miND pipeline uses the popular R package EdgeR (Robinson *et al.*, 2009) for differential expression analysis (DEA) with the quasi-likelihood negative binomial generalized log-linear model functions provided by the package.



**Figure 1. Flowchart representing the high-level steps of data processing through the pipeline.** Reference data is downloaded and processed by the repository build process (yellow area; top right) and then available for the miND pipeline in the repository/subfolder. Raw next-generations sequencing (NGS) data (blue area) is first adapter and quality trimmed and then handled by quality control (QC) tools and processed through hierarchical mapping steps (green area). These steps produce a set of mapping files that are then ingested and analyzed by R scripts, producing the miND report in the end.

A central role of NGS data processing and especially in DEA is the filtering of reads with low expression levels. Those reads would otherwise increase the noise level in the data and result in a high rate of false positives in the following DEA. Recommendations on fixed reads per million reads (RPM) based cutoff values (e.g., filtering all reads with less than 10 RPM) do not adequately account for variations in library size and miRNA reads ratio in the library and are thus arbitrary cutoffs. The DEA package DESeq2 (Love *et al.*, 2014) implements an independent filtering method that was adapted in miND to be used also with EdgeR. Assuming that most false-positives are caused by low abundant miRNAs, the algorithm removes quantiles of miRNAs from the low-abundance end and checks if the number of significant miRNAs increases after false-discovery rate (FDR) adjustment. This would be the case if mostly false positives have been removed because FDR adjustment would now be more sensitive and not remove as many true positives, increasing the overall number of significant results. This method works reliably if there are any true positive results. If the result set consist only of false positives, then even after removing the low abundant miRNAs, results would not increase the number of significant results (as there are no true positives to enrich). In this case our implementation of the algorithm has a fallback, where very lowly expressed miRNAs are pre-filtered prior to DEA and FDR adjustment: In a first step, miRNAs that are only expressed at very low levels are filtered, which is defined as RPM values that are lower than 10 divided by the smallest library size in at least half the number of samples of the smaller group. Those miRNAs carry no biological and statistical relevance (Chen *et al.*, 2016) as they have very low read counts in both groups.

An exemplary relation between a given quantile cut-off and the resulting number of differentially expressed miRNAs after FDR is shown in [Figure 2](#).

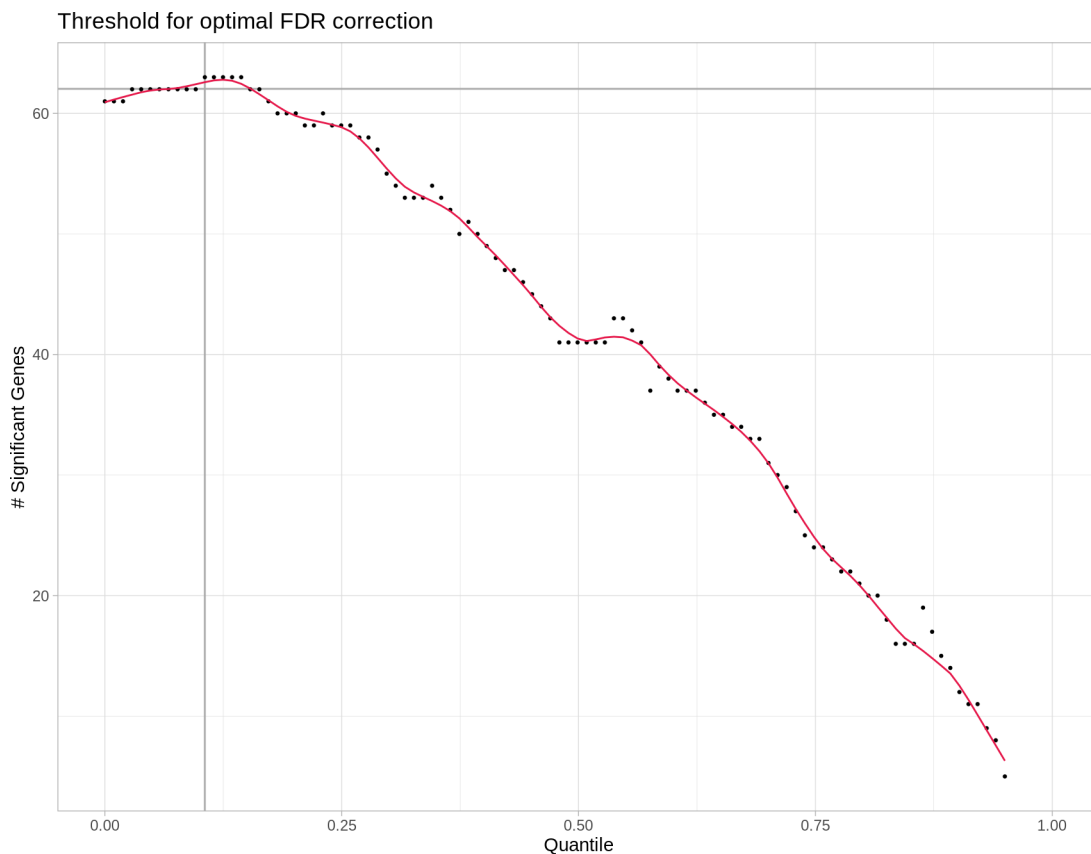
For differential expression the contrasts of interest can be selected in the experiment meta data XLS file (last sheet of the SampleContrastSheet.xlsx). Either groups or combinations of groups can be selected based on the group information provided for the samples. Each selected contrast will be part of the final interactive HTML report. In addition, a blocking factor can be selected if applicable. This blocking factor will be included in the model for the differential expression as additive factor and thus can be used e.g., for a paired experimental design or to account for batch effects.

### Interactive HTML report and statistical analysis

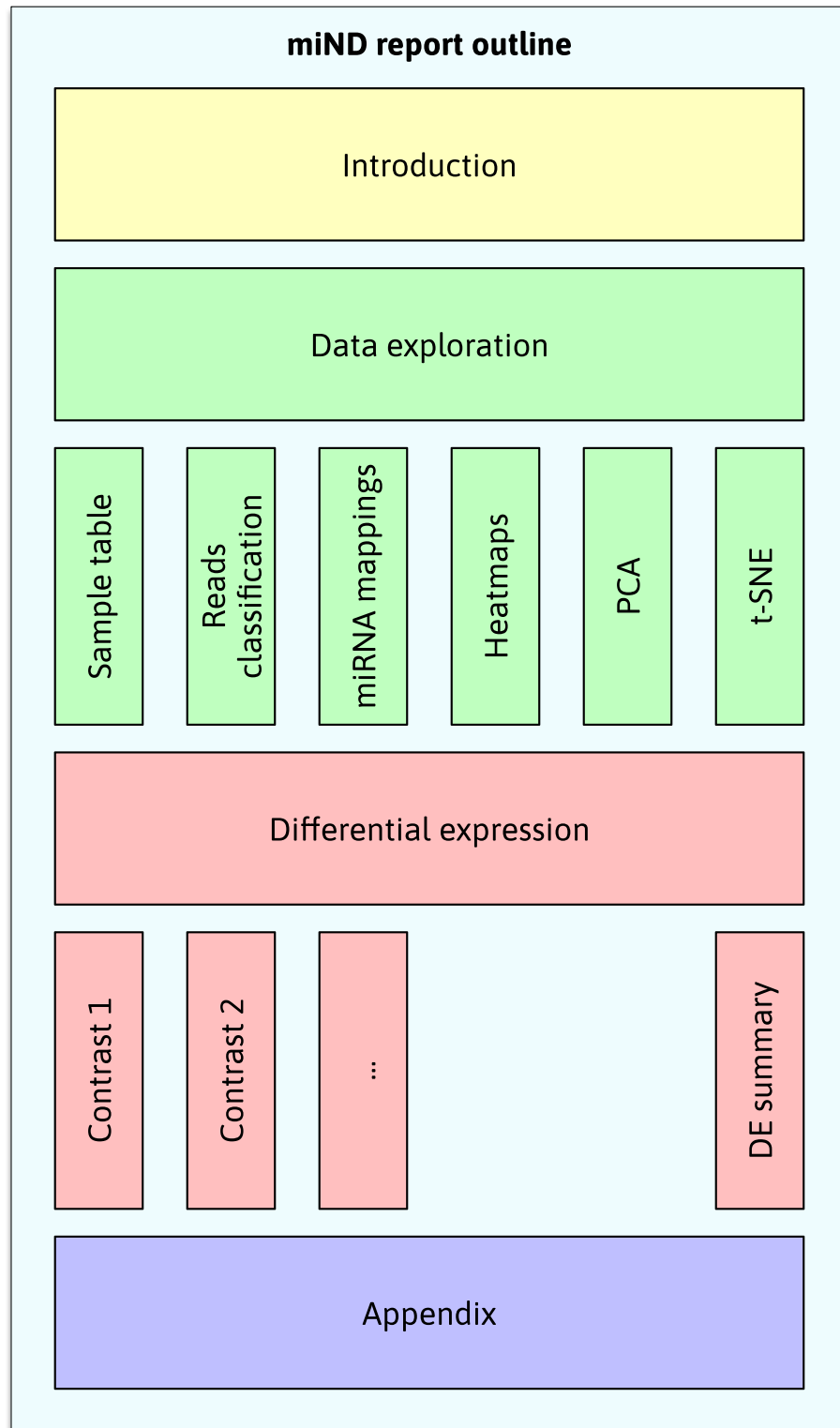
Although DEA is a central point of biomarker discovery studies, other statistical methods are needed to put this analysis into context and ensure valid results. The miND pipeline report contains a series of additional graphs and tables to present the data in a way that is interactive and easy to browse. The main sections (see [Figure 3](#)) are (1) introduction, (2) data exploration (including a sample table, reads classification plots, miRNA mapping tables, heatmaps, principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE plots)), (3) differential expression results, and (4) an appendix (references and run information).

### Reads classification plots

The reads classification plots (see [Figure 4](#)) present the amounts of reads mapped to different RNA species (miRNAs, tRNAs, piRNA, rRNA, lncRNA, etc.) based on the hierarchical mapping done by the miND pipeline. This is plotted as

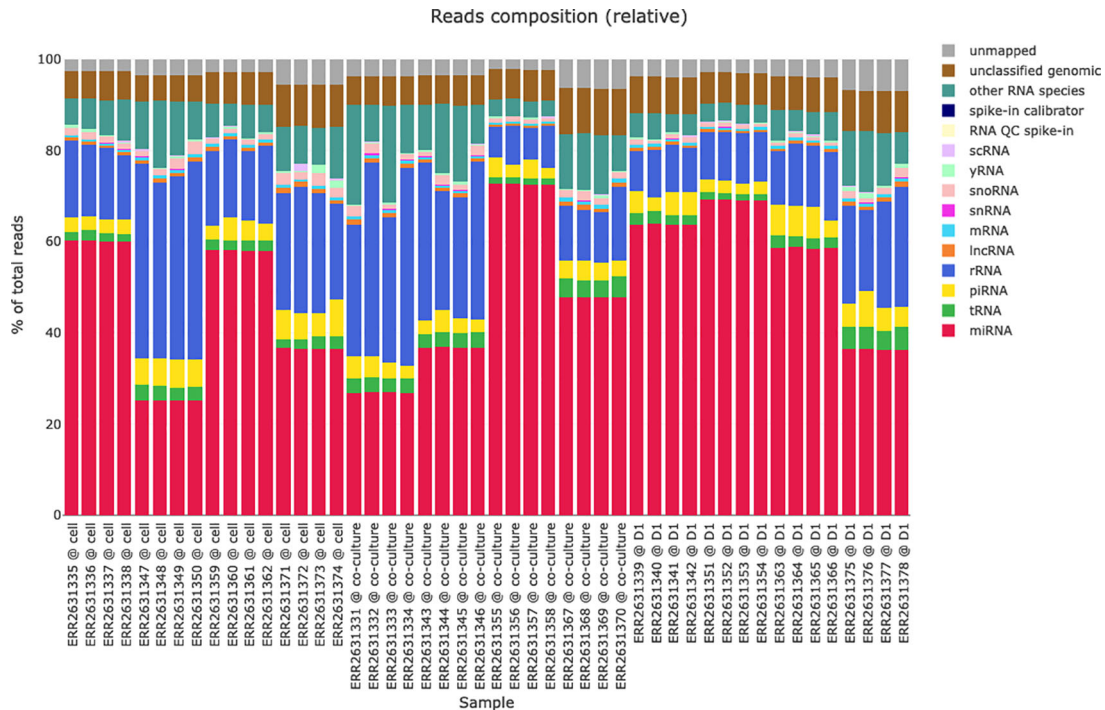


**Figure 2. DESeq2's false discovery rate (FDR) based independent filtering method.** Each point represents the number of differentially expressed micro ribonucleic acids (miRNAs) after false discovery rate (FDR) adjustment and done in steps of increasingly stringent quantile-based reads filtering. With more and more low read count miRNAs removed from the differential expression analysis, the number of significant (FDR) differentially expressed (DE) miRNAs increases to the point where more and more true positives get removed, thus decreasing the total amount of DE miRNAs. This is shown in the graph as the maximum of the red line. The optimal quantile cutoff value is then determined by finding this maximum.



**Figure 3. Outline of the interactive miND report.** The main sections (1) introduction, (2) data exploration, (3) differential expression, and (4) appendix each contain multiple subsections. The standardized structure of the report allows for the quick assessment and comparison of experiment results. t-distributed stochastic neighbor embedding (t-SNE), micro ribonucleic acids (miRNAs), differentially expressed (DE).





**Figure 4. Reads classification of all samples scaled to 100% of total reads.** Each bar represents an individual sample, while colors of the bar charts give insights in the mapped ribonucleic acid (RNA) species. This representation helps with a quick identification of library prep or sequencing issues if the ratios or total number of reads are not as expected.

absolute reads but also as relative ratios (percent) to get a quick impression of the RNA classes that are present in the data set. Especially for samples with low numbers of miRNAs present (e.g. extracellular vesicles) these two graphs give important information about the success of library preparation and sequencing.

The data on which the reads classification plots are based on is also included in the HTML report and can be either browsed directly in the HTML file or (as all tables and figures) or exported in various data formats (CSV or XLS for tabular data and PNG for graphics) for further analysis or publications.

#### *miRNA mappings table*

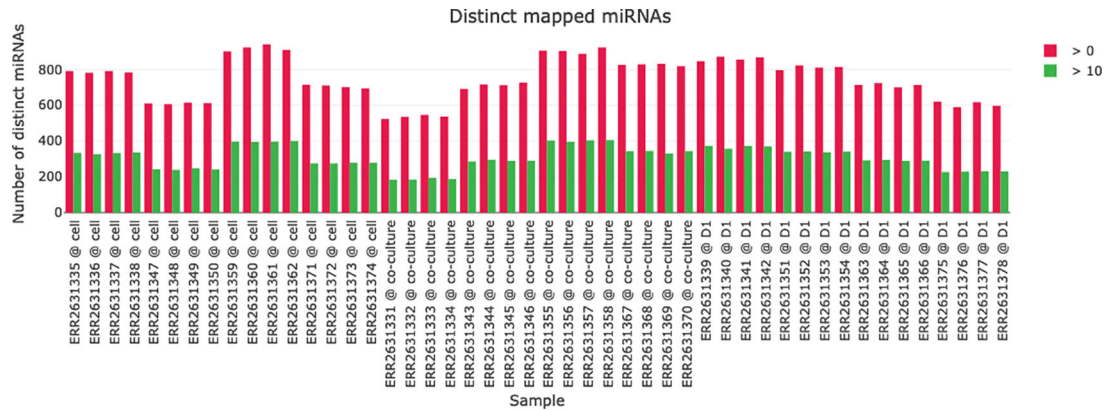
The miRNA mappings table contains read counts for each miRNA that was found in at least one of the samples. The table is available with raw read counts but also as RPM (normalized to the total number of miRNAs mapped in each sample). Group information is included in this table, if provided by the experiment metadata XLS file.

A visualization of the miRNA mapping statistics helps in comparing the number of identified miRNAs in the samples (see Figure 5). For each sample the number of distinct miRNAs with a read count above 0 and above 10 is plotted to give an impression about the abundance of distinct miRNAs and their read counts in the samples.

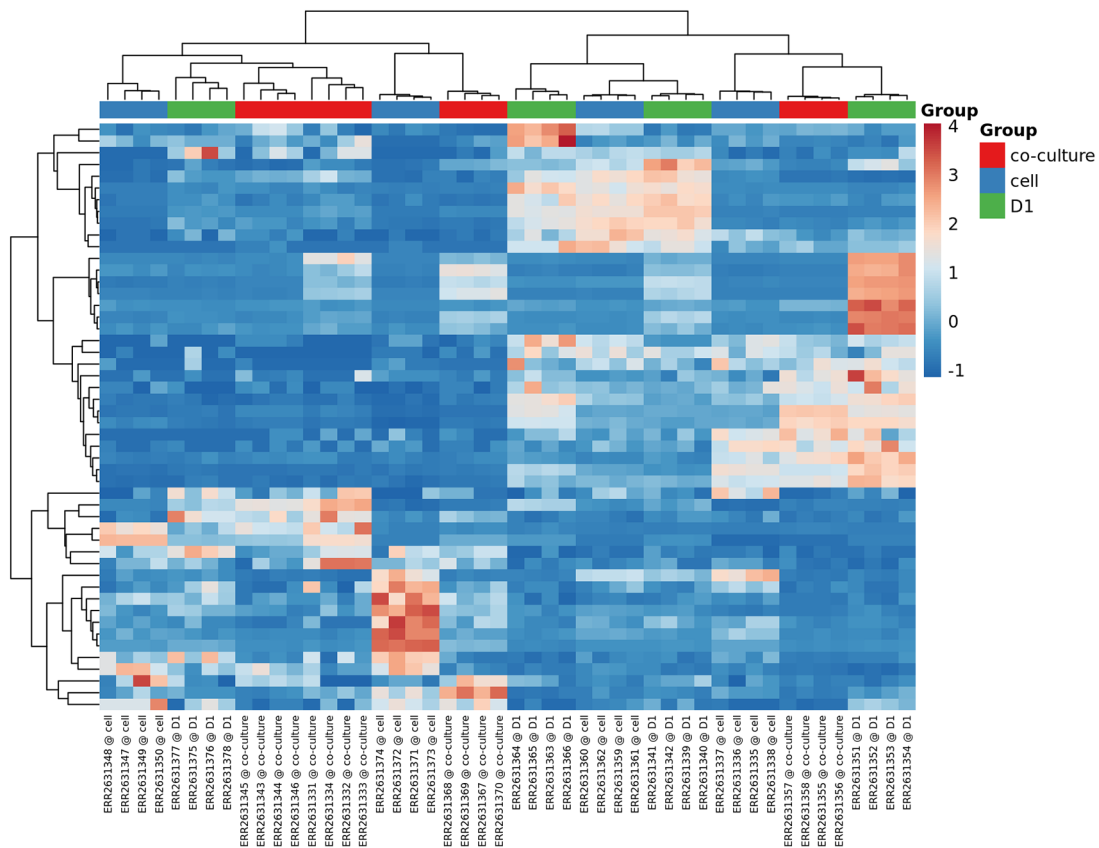
#### *Heatmaps, PCA and t-SNE plots*

The heatmaps, PCA and t-SNE plots are part of the unsupervised clustering methods that are applied by the miND pipeline and included in the report. For better understanding of underlying group relationships, any grouping information available in the meta data file will be included in the graphs. Two heatmaps are generated in the interactive HTML report. The first includes only the top 50 miRNAs based on the coefficient of variation (see Figure 6) while the second one contains all miRNAs that were detected in all samples. Both heatmaps are based on RPM normalized reads and scaled using the unit variance method for visualization. Clustering is based on complete clusters using Euclidean distances as these methods are applicable for most experimental setups. The group association of each sample is shown in the heatmaps with colored bars at the top to visualize clustering of samples based on the provided grouping information. Multiple groups are supported for heatmaps (no groups limit) and PCA/t-SNE (maximum of two groups are shown by colors and shapes).





**Figure 5. Distinct mapped micro ribonucleic acids (miRNAs) for each sample.** The number of identified miRNAs with either a read count above 0 (red) or 10 (green) is plotted for each sample.



**Figure 6. Heatmap of top 50 miRNAs.** Group information provided with the experiment meta data XLS file is included if available.

## Conclusions

The miND pipeline was developed and optimized for a multitude of small RNA-seq studies. While other available tools focus on specific aspects of the analysis (e.g., miRDeep2 on quantification of miRNAs and annotation of possible novel miRNAs and sRNAbench on differential expression), miND generates an extensive and standardized report suitable for the discovery phase of biomarker studies. The prepared HTML report provides a solid basis for further research and communicates the most important results in a structured and accessible way. Especially parameters relevant to quality control of the whole sequencing experiment (from library preparation to the in-silico analysis) are reported in standardized formats, to allow for a reliable and quick analysis of the overall quality of the experiment.

Besides the results, the generated HTML report includes descriptions, hints, and details about the methods used. This ensures that the results can be interpreted and understood easily by non-statisticians or bioinformaticians. In addition, it ensures that the final HTML report contains all information needed for reproducibility and documentation of the analysis.

Data input and experimental setup of the miND pipeline can be adjusted with the given meta data file, making it possible to use the miND pipeline for various species, sample matrices and library preparation protocols.

With the availability of the source code of the pipeline under the GNU General Public License, additional analysis steps can be integrated into the R markdown report if needed, allowing the pipeline to be tailored to other specialized applications.

While the miND pipeline includes an extensive set of analysis often needed in early phases of biomarker discovery studies, it is important to highlight the fact that no standardized pipeline will be sufficient and flexible enough to be used exclusively for every study. The results generated are meant to be a starting point for further analysis and optimizations, as parameters. For example, differential expression or heatmaps are chosen to give good results in most use cases but might not be the optimal for an individual project.

The miND pipeline was developed as part of the Translational Safety Biomarker Pipeline (TransBioLine) project from the IMI2 consortium. This project focuses on the discovery of miRNAs as novel biomarkers in the context of drug safety. In this case, the miND pipeline provides a standardized but still extensive first analysis of NGS data. In addition, the miND pipeline includes an extra module for the implementation of miND spike-ins for absolute quantification of microRNAs as recently published by [Khamina \*et al.\* \(2022\)](#).

In another recently published article by [Gutmann \*et al.\* \(2021\)](#) the pipeline was used in the discovery phase of the study to identify miRNAs that are associated with COVID-19 severity and mortality. The miRNAs reported by the miND pipeline were later manually selected and evaluated based on the HTML report for further confirmation with RT-qPCR, where the confirmation showed a high level of reproducibility from the NGS data.

We will continue working on the pipeline and release updates to the public version if needed. Especially in regard to the miND spike-ins that allow for the absolute quantification of miRNA in biofluids we expect to release an updated version soon.

## Data availability

### Source data

Mature and hairpin sequences of miRBase are available at: <https://www.mirbase.org/ftp/22.1>

Genome sequences (DNA and cDNA) is available at Ensembl (for human): [http://ftp.ensembl.org/pub/release-105/fasta/homo\\_sapiens](http://ftp.ensembl.org/pub/release-105/fasta/homo_sapiens)

Non-coding RNA sequences are available at RNAcentral: [http://ftp.ebi.ac.uk/pub/databases/RNAcentral/current\\_release](http://ftp.ebi.ac.uk/pub/databases/RNAcentral/current_release)

Data associated with the example use case are not owned by the authors. Requirements to access these datasets is given in the protocol (<https://dx.doi.org/10.17504/protocols.io.b3f6qjre>).

## Software availability

Source code available from: <https://github.com/tamirna/miND>

Archived source code available from: <https://doi.org/10.5281/zenodo.6080470> (Diendorfer *et al.*, 2022)

License: [GNU GPL 3.0](#)

## References

Anaconda Software Distribution: *Anaconda Documentation*. Anaconda Inc.; 2020.  
Reference Source

Andrews S: *FastQC: A quality control tool for high throughput sequence data*. 2010  
Reference Source

- Aparicio-Puerta E, Lebrón R, Rueda A, *et al.*: **sRNAbench and sRNAtoolbox 2019: intuitive fast small RNA profiling and differential expression.** *Nucleic Acids Res.* 2019; **47**(W1): W530–W535.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Chen Y, Lun ATL, Smyth GK: **From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; referees: 5 approved].** *F1000Res.* 2016; **5**: 1–49.  
[Publisher Full Text](#)
- de Ronde MWJ, Ruijter JM, Moerland PD, *et al.*: **Study Design and qPCR Data Analysis Guidelines for Reliable Circulating miRNA Biomarker Experiments: A Review.** *Clin. Chem.* 2018; **64**(9): 1308–1318.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Diendorfer A, Khamina K, Pultar M, *et al.*: **miND (miRNA NGS Discovery pipeline): a small RNA-seq analysis pipeline and report generator for microRNA biomarker discovery studies (v1.2RC2).** *Zenodo.* 2022.  
[Publisher Full Text](#)
- Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: Summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Friedländer MR, MacKowiak SD, Li N, *et al.*: **MIRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades.** *Nucleic Acids Res.* 2012; **40**(1): 37–52.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Griffiths-Jones S: **The microRNA registry.** *Nucleic Acids Res.* 2004; **32** (Database issue): 109D–1111D.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat. Methods.* 2018; **15**(7): 475–476.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gutmann C, Khamina K, Theofilatos K, *et al.*: **Association of cardiometabolic microRNAs with COVID-19 severity and mortality.** *Cardiovasc. Res.* 2021; **118**: 461–474.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Khamina K, Diendorfer AB, Skalicky S, *et al.*: **A MicroRNA Next-Generation-Sequencing Discovery Assay (miND) for Genome-Scale Analysis and Absolute Quantitation of Circulating MicroRNA Biomarkers.** *Int. J. Mol. Sci.* 2022; **23**(3): 1226.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Köster J, Rahmann S: **Snakemake-a scalable bioinformatics workflow engine.** *Bioinformatics.* 2012; **28**(19): 2520–2522.  
[Publisher Full Text](#)
- Langmead B, Trapnell C, Pop M, *et al.*: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol.* 2009; **10**(3): R25.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Larrea E, Sole C, Manterola L, *et al.*: **New Concepts in Cancer Biomarkers: Circulating miRNAs in Liquid Biopsies.** *Int. J. Mol. Sci.* 2016; **17**(5): 627.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lee RC, Ambros V: **An Extensive Class of Small RNAs in *Caenorhabditis elegans*.** *Science.* 2001; **294**(5543): 862–864.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Llewellyn HP, Vaidya VS, Wang Z, *et al.*: **Evaluating the Sensitivity and Specificity of Promising Circulating Biomarkers to Diagnose Liver Injury in Humans.** *Toxicol. Sci.* 2021; **181**(1): 23–34.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 521–550.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mölder F, Jablonski KP, Letcher B, *et al.*: **Sustainable data analysis with Snakemake.** *F1000Res.* 2021; **10**: 33.  
[Publisher Full Text](#)
- Murillo OD, Thistlethwaite W, Rozowsky J, *et al.*: **exRNA Atlas Analysis Reveals Distinct Extracellular RNA Cargo Types and Their Carriers Present across Human Biofluids.** *Cell.* 2019; **177**(2): 463–477.e15.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- O'Brien J, Hayder H, Zayed Y, *et al.*: **Overview of microRNA biogenesis, mechanisms of actions, and circulation.** *Front. Endocrinol.* 2018; **9**(AUG): 1–12.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Robinson MD, McCarthy DJ, Smyth GK: **edgeR: A Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2009; **26**(1): 139–140.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Ronen R, Gan I, Modai S, *et al.*: **miRNAkey: a software for microRNA deep sequencing analysis.** *Bioinformatics.* 2010; **26**(20): 2615–2616.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Starlinger P, Hackl H, Pereyra D, *et al.*: **Predicting Postoperative Liver Dysfunction Based on Blood-Derived MicroRNA Signatures.** *Hepatology.* 2019; **69**(6): 2636–2651.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Sweeney BA, Petrov AI, Burkov B, *et al.*: **RNAcentral: A hub of information for non-coding RNA sequences.** *Nucleic Acids Res.* 2019; **47** (D1): D221–D229.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Wang W-C, Lin F-M, Chang W-C, *et al.*: **miExpress: Analyzing high-throughput sequencing data for profiling microRNA expression.** *BMC Bioinformatics.* 2009; **10**(1): 328.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Zerbino DR, Achuthan P, Akanni W, *et al.*: **Ensembl 2018.** *Nucleic Acids Res.* 2018; **46**(D1): D754–D761.  
[PubMed Abstract](#) | [Publisher Full Text](#)

# Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 05 October 2022

<https://doi.org/10.5256/f1000research.101122.r151316>

© 2022 Enguita F. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Francisco J. Enguita**

Faculdade de Medicina, Instituto de Medicina Molecular João Lobo Antunes, Universidade de Lisboa, Lisboa, Portugal

The manuscript by Diendorfer and coworkers describes a pipeline for NGS data processing specially devoted to the analysis of small non-coding RNAs, mainly focused on miRNAs.

The manuscript is well written, but the authors would need to give further details in order to compare their pipeline with the already existing ones. It is not clear for the reader what are the main advantages of miND pipeline in comparison with the already available ones.

I would advise to perform a small benchmarking study using a test dataset that could be any one existing in public databases. The authors would need to answer the questions: what are the advantages of miND, its weaknesses and why the user should give a try to this new pipeline.

**Is the rationale for developing the new software tool clearly explained?**

Partly

**Is the description of the software tool technically sound?**

Partly

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**

Partly

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**

Partly

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**

No

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** non-coding RNAs; miRNAs; lncRNAs; circRNAs; structural biology

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Reviewer Report 14 April 2022

<https://doi.org/10.5256/f1000research.101122.r129126>

© 2022 Almstrup K et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Kristian Almstrup**

Department of Growth and Reproduction, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

**Nina Mørup**

Department of Growth and Reproduction, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

**Ailsa Maria Main**

Department of Growth and Reproduction, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

In the manuscript by Diendorfer et al., a bioinformatic pipeline for analysis of data from small RNA sequencing is presented. The pipeline, named miND, allows identification and annotation of small RNA reads as well as differential expression analysis.

I have the following major concerns about the study as it is:

Several other small RNA sequencing pipelines, like Oasis2.0 (<https://oasis.dzne.de/index.php>), sRNAWorkbench (<https://sourceforge.net/projects/srnaworkbench/>), sRNAPipe (<https://github.com/GReD-Clermont/sRNAPipe>), miRge3.0 (<https://sourceforge.net/projects/mirge3/>) already exist and some support both identification and differential expression analysis. It is hence unclear what novelty miND brings compared to other pipelines. To allow the reader to make an informed choice about which pipeline to choose for analysis, miND should be benchmarked against some of the already existing pipelines. What are the differences when the same dataset (PRJEB27261/E-MTAB-6885) is analysed with e.g. Oasis2.0 (Rahman *et al.*, 2018)?

The authors argue that miND “bridges the gap between biologists and bioinformaticians”, and this

is also evident from the easy-to-use Excel files. However, the pipeline is based on Snakemake workflows and a conda install and hence require *a priori* knowledge of conda, which would not be common knowledge to biologists. I encourage the authors to make miND available as a standalone app or web portal (as is the case for the similar pipeline Oasis2.0). Since, at least, parts of the miND pipeline are based on R-scripts it might be easy to make a Shiny app or similar.

The pipeline focuses on miRNAs. This reviewer encourages the authors also to include analysis of other small RNA species as these are likely to be equally important as biomarkers in a liquid biopsy. Furthermore, on a whole, the authors do not discuss in detail the possible problems and downsides of their pipeline. A more in-depth and critical discussion of strengths and limitations is warranted.

The readme file on github should contain instructions on how to install miND in a language that biologists can understand.

Finally, in some places the authors should consider shortening sentences/simplifying statements so, again, it is easier for the readers to follow.

## References

1. Rahman R, Gautam A, Bethune J, Sattar A, et al.: Oasis 2: improved online analysis of small RNA-seq data. *BMC Bioinformatics*. 2018; **19** (1). [Publisher Full Text](#)

## Is the rationale for developing the new software tool clearly explained?

Partly

## Is the description of the software tool technically sound?

Yes

## Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

## Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

## Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Endocrinology, small RNAs, genetics

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have**

**significant reservations, as outlined above.**

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**