miRMaster 2.0: multi-species non-coding RNA sequencing analyses at scale

Tobias Fehlmann [1], **Fabian Kern** [1], **Omar Laham**[1], **Christina Backes** [1], **Jeffrey Solomon**[1], **Pascal Hirsch**[1], **Carsten Volz**[2], **Rolf Müller**[2] and **Andreas Keller** [1,3,*]

[1]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany, [2]Department of Microbial Natural Products, Helmholtz-Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Centre for Infection Research (HZI) and Department of Pharmacy, Saarland University, Campus E8 1, 66123 Saarbrücken, Germany and [3]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, CA, USA

ABSTRACT

Analyzing all features of small non-coding RNA sequencing data can be demanding and challenging. To facilitate this process, we developed miRMaster. After the analysis of over 125 000 human samples and 1.5 trillion human small RNA reads over 4 years, we present miRMaster 2 with a wide range of updates and new features. We extended our reference data sets so that miRMaster 2 now supports the analysis of eight species (e.g. human, mouse, chicken, dog, cow) and 10 non-coding RNA classes (e.g. microRNAs, piRNAs, tRNAs, rRNAs, circRNAs). We also incorporated new downstream analysis modules such as batch effect analysis or sample embeddings using UMAP, and updated annotation data bases included by default (miRBase, Ensembl, GtRNAdb). To accommodate the increasing popularity of single cell small-RNA sequencing data, we incorporated a module for unique molecular identifier (UMI) processing. Further, the output tables and graphics have been improved based on user feedback and new output formats that emerged in the community are now supported (e.g. miRGFF3). Finally, we integrated differential expression analysis with the miRNA enrichment analysis tool miEAA. miRMaster is freely available at https://www.ccb.uni-saarland.de/mirmaster2.

GRAPHICAL ABSTRACT
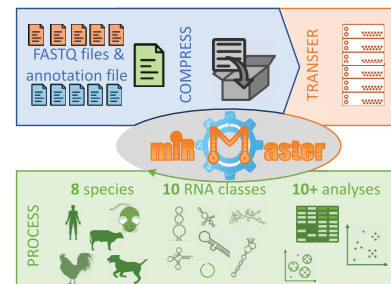
INTRODUCTION

The reliable analysis of small non-coding RNA (sncRNAs) sequencing data can be challenging, time consuming and varies in many aspects. This includes the primary processing of sequencing data for quality control but also downstream analyses. Besides tRNAs and snoRNAs, microRNAs are sncRNAs that are extensively studied already for over two decades. A comprehensive summary on the state-of-the art in microRNA biology has been published by Bartel in 2018 (1). In addition to the canonical biology of microRNAs, more and more non-canonical aspects of miRNA biology are becoming obvious (2), calling for a broad variety of analysis aspects. It is thus not surprising, that many miRNA analysis tools, online and stand-alone, are available. The current release of Aviator (https://www.ccb.uni-saarland.de/aviator), a tool that aims to provide accessibility statistics of all web servers and data bases in life sciences, lists 322 web-based resources for microRNAs, of which 235 are currently working. By developing miRMaster (3,4), we provide a tool with a strong focus on the analysis of all aspects of miRNAs described in a systematic manner. While the tool became stepwisely broader in its functionality aspect to cover other sncRNA classes, microRNAs are still its anchor point. Based on the original version of miRMaster, 1500 runs have been completed, over

---

125 000 human samples have been analyzed, and 1.5 trillion human small RNA reads were processed over four years. To make further use of the uploaded data, we ask miRMaster users whether we can re-analyze the aggregated sequencing reads as further feedback for the development of our tool and as comparison to our small RNA research projects.

As mentioned before, several other web servers and web-services for analyzing miRNA sequencing data exist, overlapping partially with miRMaster's functionality. We thus want to put our tool in the context of others and recent developments in the field. A broad tool meta review has been published by Chen *et al.*, which covers 95 review papers and about 1000 miRNA bioinformatics tools (5). The variety of tools reaches from rather specialized tools, e.g. for the detection of isomiRs or miRNA editing (6) to very broad analysis pipelines. Among the tools with broader analysis functionality with a particular focus on miRNAs, we want to mention CBS-miRSeq (7), miRquant (8), sRNAbench/sRNAtoolbox (9), Chimira (10), mirPRo (11), miRge (12) and CPSS2 (13). Among those, sRNAtoolbox was updated most recently with improvements made to e.g. the batch processing, library preparation protocols and differential expression analysis. Moreover, SPAR is a small RNA-seq portal for analysis of sequencing experiments (14). GLASSgo for example facilitates automated detection of sRNA homologs (15). Also, for viruses, sRNA analysis tools have been developed such as MISIS-2, a tool for the in-depth analysis of sRNAs and representation of consensus master genomes in viral quasispecies (16). With respect to non-canonical miRNA biology, such as miRNA sponges (17), also specialized software tools have been made available (18). Another aspect is the analysis of miRNA editing and chemical modifications. Likewise, for this task, specialized tools are available such as Prost! (19), DeAnniso (20) and others. Similarly, for tRNAs and other ncRNA classes, chemical modifications exist (21) that further complicate the analysis of ncRNA data sets.

The number of available tools, web-based and standalone, reflects the continuous interest of the research community in non-coding RNAs. At the same time, it pinpoints that a more integrative analysis of different ncRNA classes is required. Recently, a 'changing of the guards' model has been proposed, where microRNA levels decreased but small transfer RNA fragments increased in blood of patients (22).

Although a direct comparison in terms of scope, functionality, ease of use and other parameters is challenging and partially subjective, we aimed to provide at least an overview on a selection of commonly used broader analysis tools that are available as web-service. We thus evaluated the number of analyses, the number of supported organisms, the number of supported ncRNA classes and other parameters for 10 tools and present the result sorted by the publication date (Figure 1). The analysis reveals an expected pattern, the more recent tools have a broader scope of functionality as compared to the early tools. The recently updated sRNABench for example excels in basically all categories, e.g. offers more organisms as compared to miRMaster 2. miRMaster 2 offers in contrast more output options and covers more ncRNA classes. Figure 1 allows to compare the functionality of the 10 selected tools and supports users in their decision to select one tool.

MATERIALS AND METHODS

Reference data bases

miRMaster relies on public annotation data sets of several well-known and widely used reference data bases. These include for the different RNAs miRBase (version 22.1 (23)), Ensembl ncRNA (version 100 (24)), RNACentral (for piRNAs) (version 15 (25)), GtRNAdb (version 18.1 (26)), circBase (accessed 25.10.20 (27)), NONCODE (version 5 (28)) and NCBI RefSeq for the reference genomes as well as viruses and bacteria.

Supported sequencing protocols

miRMaster 2 directly supports the most common sequencing protocols. This includes Illumina Truseq, Bioo Scientific Nextflex, MGISeq and Diagenodes D-Plex and CATS technology.

Implementation and graphical representation of the web service
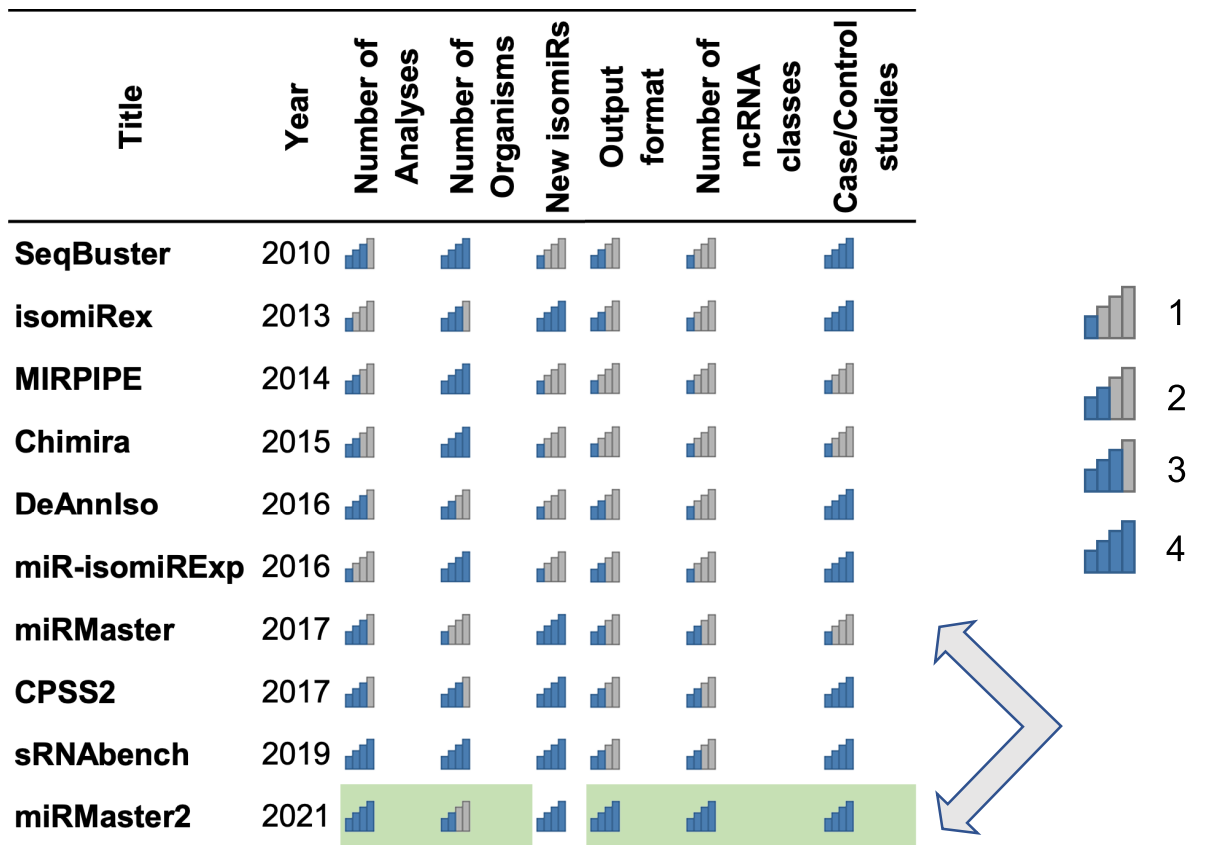
The miRMaster 2 web service was implemented using Python 3.7.6 with Django 2.2.10, Postgres 11.1 and Redis 5.0. All services are encapsulated in docker containers and bundled with docker compose. The job queue is handled with Celery 4.4.7 and Redis, and the jobs are executed via Snakemake 5.31.1 (29). The frontend was styled with Bootstrap 4.5.3 and the interactive features are based on the Angular JS 1.5.11 and jQuery 3.4.1 libraries. Plots are rendered with Highcharts 8.2.2 and Clustergrammer-GL 0.22.0 (30). RNA secondary structures are rendered with fornac 1.1.10 and interactive tables with DataTables 1.10.23.

Compression and quality control data

miRMaster accepts raw FASTQ and gzip compressed FASTQ files as input. At first and before the data is sent to our server, we perform three pre-processing steps encompassing adapter trimming, quality filtering and read collapsing via JavaScript on the upload page. The collapsed reads are transmitted to the servers as soon as possible but in chunks of ~16MB, thereby ensuring a low RAM consumption. The quality scores of the reads are collected on the user side and only aggregated metrics are sent to the server. Multiple annotations can be provided for the uploaded samples, which can then be highlighted in the resulting reports. In particular, a group annotation for differential expression downstream can be selected, in which case miRMaster will perform additional analyses.

Data pre-processing

While most of the analysis parameters in the pre-processing view are conveniently set to reasonable default values, the expert mode allows full control and maximal flexibility. In the pre-processing view, the 5′ barcode length, an adapter barcode length and the length of the unique molecular identifier (UMI) can be determined. Further, after activation of the expert mode, minimum read length and maximum adapter edit distance can be modified as well as the minimum read/adapter overlap. Finally, leading and trailing

Title	Year	Number of Analyses	Number of Organisms	New isomiRs	Output format	Number of ncRNA classes	Case/Control studies
SeqBuster	2010						
isomiRex	2013						
MIRPIPE	2014						
Chimira	2015						
DeAnnIso	2016						
miR-isomiRExp	2016						
miRMaster	2017						
CPSS2	2017						
sRNAbench	2019						
miRMaster2	2021						

Legend: 1, 2, 3, 4

	1	2	3	4
Number of Analyses	one analysis	up to three analyses	up to 6 analyses	over 6 analyses
Number of Organisms	one organism	up to 10 organisms	up to 100 organisms	over 100 organisms
New isomiRs	no	-	-	yes
Output format	tables	tables and images	-	tables/interactive images
Number of ncRNA classes	no	up to 8 other classes	-	over 8 other classes
Case/Control studies	no	-	-	yes

Figure 1. Tool comparison. Comparison of features provided by tools analyzing sncRNA-seq data. Improvements of miRMaster 2 in comparison to its original release are marked in green.

N's can be trimmed, or reads containing N's can be fully omitted. For quality trimming, the sliding window size and quality can be selected. To maximize the processing capabilities of the user's CPU, the number of threads that can be used on the client side for the pre-processing can be configured.

Mapping

The read mapping process in miRMaster 2 is carried out using Bowtie 1.2.3 (31) as the standard option. For the up-

date we added STAR 2.7.5a (32) as alternative mapper. In the standard mode, up to five hits in the reference genome are allowed for each read, where the mapping seed length is set to 18 nt and no mismatches in the seed region are allowed. In the expert mode, the user is free to change these parameters. For quantifying miRNAs, reads are mapped against miRNA precursors while allowing per default one mismatch. The resulting mappings are then filtered to count only reads mapping to the annotated miRNAs with at most two nucleotides differences at the 5′ end and five nucleotides at the 3′ end. For isomiR quantification, the mapping is

performed with one additional mismatch and then subsequently filtered, such that non-templated nucleotide additions are not counted as mismatches. Other ncRNAs are per default quantified without any mismatches and all multimapping reads with the lowest number of mismatches are considered. To generate the miRGFF3 isoform format, mirtop 0.4.23 (33) is run on the aligned files and subsequently filtered to accommodate the user selected number of allowed mismatches.

Detection filtering

Before applying analysis methods to the expression matrices of the different ncRNA types, these matrices are filtered per default, such that only those RNAs that are expressed in at least 50% of all the samples are kept, or in case a differential expression annotation variable is provided (e.g. Diagnosis), in at least 50% of the samples of one of the variable levels (e.g. Dementia or Control). For this filtering procedure, only RNAs that are expressed with at least three reads are considered detected. In addition to normalizing the reads by the sequencing depth (RPM) we $\log_2$ transform the expression data and add a pseudo count of one.

Embedding

miRMaster 2 is equipped with two common dimension reduction approaches, namely Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). The user can first select the embedding from a drop down followed by the response variable. This response variable is extracted from the annotation file and the data points are colored according to the respective grouping. As representation, 2D-scatter plots are provided to the user.

Clustering

Hierarchical clustering with Euclidean distance and complete linkage is performed on the sample Spearman correlation, which is determined based on the reads per million (RPM) normalized expression matrix for each ncRNA type separately, as well as for all sncRNAs. In addition, hierarchical clustering is also performed on the RPM $\log_2$ normalized expression matrix for each ncRNA type, as well as on subsets of the top RNAs with the largest variance.

Batch effect analysis

To detect the influence of technical batches or attribute variance to biologically relevant parameters, a Principal Variance Component Analysis (PVCA) is performed. PVCA combines the strengths of two data analysis techniques, principal component analysis (PCA), which reduces the feature dimensions while maintaining the largest fraction of the variability in the data, and variance components analysis (VCA), which fits a mixed linear model using factors of interest as random effects. In more detail, all variables provided in the annotation file are fit as random effects including two-way interaction terms in the mixed model. Thereby, principal components obtained from the original data expression matrix are selected. As a result, the proportion of variance that is attributed to the variables from the annotation file is reported.

Differential gene expression analysis

The groupings for differential expression analysis are extracted from the annotation file provided by the user. If more than two groups are given, all pair-wise differential expression analyses are performed automatically. miRMaster first computes whether the features are normally distributed by applying the Shapiro Wilk test. As hypothesis test for assessing the degree of differential expression, *t*-test (for normally distributed data) and Wilcoxon Mann–Whitney test otherwise, are calculated. For multiple groups, also analysis of variance (ANOVA) as well as the nonparametric Kruskal–Wallis test are computed. The *P*-values are adjusted for multiple testing by controlling the false discovery rate using the Benjamini–Hochberg procedure. Further measures representing effect sizes are fold changes, the area under the receiver operator characteristics curve (AUC value) and Cohen's *d*. In addition to tabular output, volcano plots ($\log_2$ fold change versus negative decade logarithm of *P*-values) are displayed and boxplots per RNA are shown.

miRNA prediction

The prediction of new miRNAs follows the same principles as described in miRDeep (34) and our previous publications (3,4). Similarly, the expert mode allows maximal flexibility. As a first step, after the reads have been mapped to the genome, miRNA precursor candidates are determined. To this end local maximum read stacks, which are assumed to stem from potential miRNAs, are searched in downstream windows of per default 70 nucleotides and two precursors are excised from each stack. For this step, the required minimum read stack can be increased in order to improve the specificity of miRNA precursor predictions. Also, minimal and maximal length of the mature miRNA(s) can be set. Moreover, the mapping fraction consistent with Dicer processing can be increased or decreased. miRMaster groups new miRNA candidates into several categories, depending on their overlap with known miRNAs or known miRNA precursors. Precursor miRNA candidates belonging to the 'novel' category are not overlapping any known miRNAs and none of their annotated miRNAs have a similarity to any known miRNA of the same species. Details on the different categories are provided in the software tutorial page. To further rank and prioritize the predicted precursors, NovoMiRank (35) is subsequently applied and the scores are presented in the results table.

API to miEAA

Following our ambition to develop a fully integrated knowledge base on miRNAs (36), we started to integrate our tools such as miEAA (37,38) and miRSwitch (39) with APIs. For miRMaster we continued this process in the reverse direction. miRNAs can be sorted by their expression level, or if case–control studies are evaluated, also with respect to their differential expression. From the miRMaster results page, the miEAA APIs for over-/underrepresentation and

miRNA set enrichment analyses can be queried such that functional pathway enrichment analyses are feasible within minutes. Moreover, miRNAs are linked to target genes and target gene networks using miRTargetLink 2 (40), when applicable.

RESULTS

In this section, we aim for a complete and self-containing description of miRMaster 2, partially sketching features that have already been available in the original version such as the miRNA prediction module. The improvements are highlighted, summarized in the conclusion and also marked in green in Figure 1.

Data input

miRMaster relies on FASTQ files (optionally gzip compressed) as main input. Additionally, an annotation file can be provided by users to facilitate downstream analyses. While the annotation file is typically small, represented by few kilobytes of data, the FASTQ files can encompass several gigabytes per sample. Therefore, the data transfer for large studies could become a time-consuming task. We thus implemented an algorithm, exploiting the fact that miRNA-seq libraries are often of lower read complexity, that compresses the data typically by over 90%. This happens at the client side and only the compressed data are then transferred. Similarly, selected quality statistics are computed at the client side and only the relevant compressed information is transmitted. The server-side processing of the data starts immediately, i.e., while the data are transferred the processing is already initiated. We optimized these procedures in a way that studies with several hundred samples can be processed by miRMaster. The time-consuming analyses carried out at the server side afterwards are mostly implemented in efficient C++, such that even large-scale studies exceeding hundreds of samples are processed within a few hours.

Supported organisms and RNA classes

While the original version of miRMaster was centered around the analysis of human microRNA data, supporting only a few other small RNA data analyses, we now provide full support for several other common organisms and more ncRNA classes. Importantly, the analysis of the RNA classes is not only restricted to small non-coding RNAs but also to longer RNAs such as circRNAs. With respect to the organisms, miRMaster 2 supports, besides *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Monodelphis domestica*, *Macaca mulatta*, *Gallus gallus*, *Bos taurus* and *Canis familiaris*.

Pre-processing functionality and mapping

In the pre-processing step, adapters are trimmed and only sequences exceeding the selected minimum length are extracted. Furthermore the *GC* content is calculated. The results of this step are available as table and displayed as boxplots (Figure 2A). If the user provided several groups, as for instance cases and controls, or grades of severity for a disease, the information is displayed per group. In addition to

the aggregated statistics, detailed per-sample statistics also are available. All tables can be downloaded in excel and csv format. All graphics are available as jpg, pdf, png and svg files. The underlying data for each graphic panel can also be downloaded in case users want to make figures on their own.

The second analysis step is genome and non-coding RNA mapping. First, the mapping to the reference genome is performed using the user specified input parameters and mapper. Mapping to the following 10 RNA classes is also performed: microRNA (Figure 2B), tRNA, piRNA, rRNA, scaRNA, lncRNA, snoRNA, snRNA, miscRNA and circRNA (Figure 2C). Additionally, mapping against viruses and bacteria from RefSeq is carried out for each sample, based on reads that did not map against the reference genome. As for the adapter trimming, all available information can be downloaded in excel and csv format and for each sample detailed mapping statistics are presented.

Sample embedding, clustering, batch effect analysis

After completing the pre-processing and mapping, different aggregated analyses on the sample and RNA class level are carried out. First, an embedding using UMAP or alternatively PCA is available. The embedded graphics as 2D scatter plot can be colored with respect to arbitrary input variables extracted from the annotation file. For each RNA class, a distinct embedding is available (miRNA Figure 2D, tRNA Figure 2E). This allows users for example to visually compare whether for one RNA class a better clustering with respect to a disease phenotype is observed as compared to another.

Next, a sample correlation analysis is performed. Here, the correlation of the RNA expression between all pairs of samples is computed and shown as heatmap. The ordering of rows and columns can be modified and colored representations on top of the heatmaps show for each provided variable a color code (Figure 2F). Per default, hierarchical clustering with Euclidean distance and complete linkage is shown. In addition to the sample-to-sample correlation clustering, also the clustering of the expression values for each of the 10 RNA classes is computed, with the possibility to focus on the subset of RNAs with the largest variance. For both, rows (features) and columns (samples), clusters are defined and automatically adjusted. If a cluster is selected, the distribution of representatives within the cluster is displayed and *P*-values for enrichments are provided.

As last consideration of this functionality aspect, miRMaster 2 estimates the proportion of variance that can be attributed to variables provided in the annotation file. This can be biologically relevant metadata (again, case or control, or different severity grade of a disease, sex and many others) or it can be technical batches (e.g. the information which samples have been sequenced together or come from the same site in a multi-centric study). To this end, miRMaster 2 performs a PVCA, that combines aspects of principal component analysis as well as variance components analysis. The results are provided as bar charts, again for each of the RNA classes separately (Figure 2G for miRNA and Figure 2H for tRNAs). While miRMaster 2 highlights potential experimental batches it currently does not provide function-
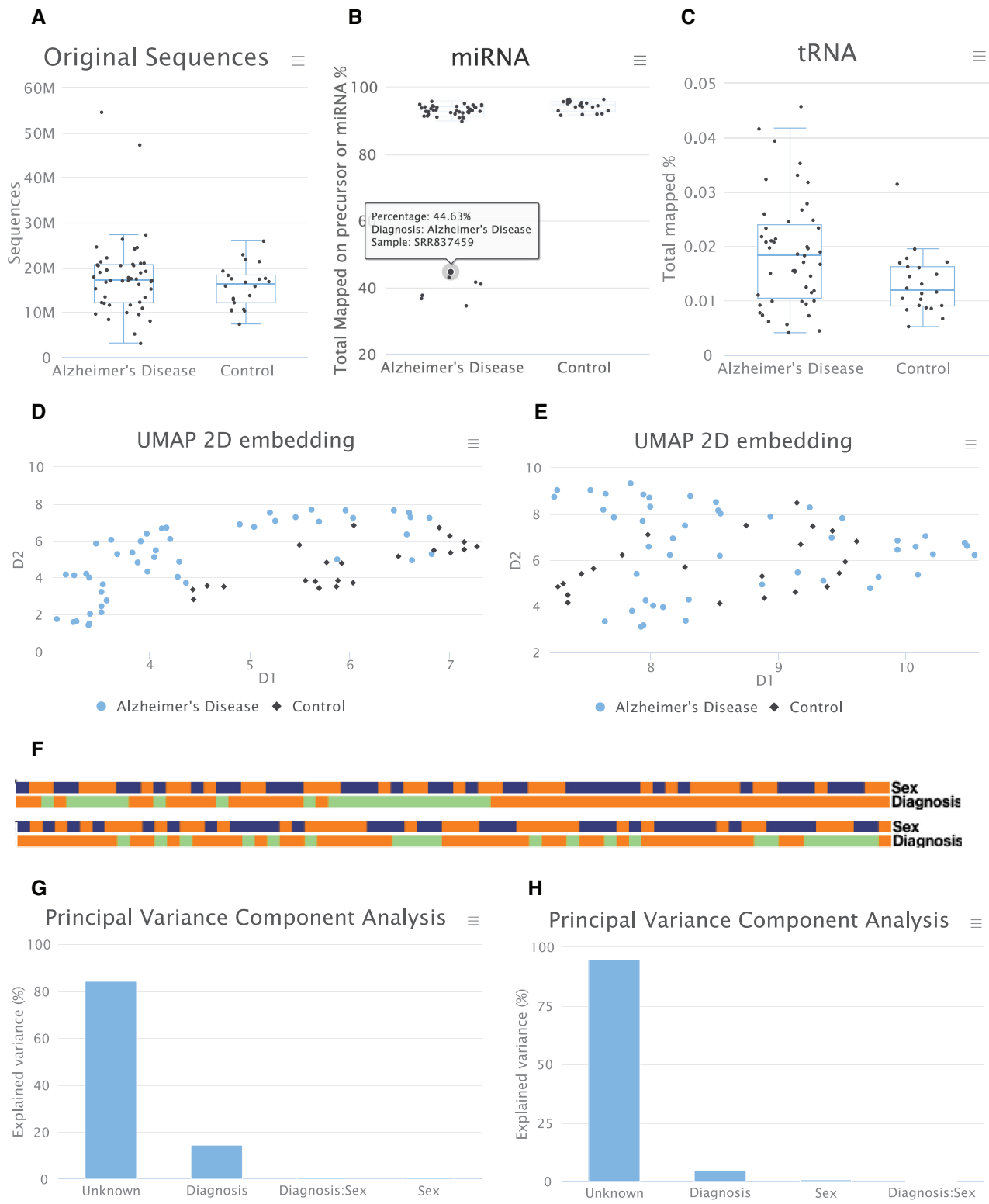
Figure 2. Selected result for the data pre-processing. The presented data are taken from the online demo data set on Alzheimer's Disease (AD). (**A**) Number of reads in the data set. Each dot represents a single sample. No difference between AD and controls exists. (**B**) Mapping to microRNAs. One point is highlighted. This feature can be used to identify outliers. (**C**) Mapping to tRNAs. In the overall distribution we observe here differences between the two classes. (**D**) Embedding of the samples using UMAP, colored by the disease phenotype using miRNAs. A clustering in the two groups can be recognized in this embedding. (**E**) The same embedding for tRNAs. In the case of this RNA class, no clear clustering is present. (**F**) Color-coded clustering for the sex and disease phenotype for microRNAs (top) and tRNAs (bottom). The two classes don't show clustering with respect to sex but for microRNAs a clustering of AD samples can be observed. (**G**) Results of the PVCA for miRNAs. Around 15% of the total variance can be explained by the disease phenotype. (**H**) The same results for the tRNAs. Here, a lower percentage of variance is explained by the disease phenotype.

ality to correct for batch effects. If experimental batches exist, a batch correction by the user is therefore recommended.

Differential expression, downstream analysis and APIs to other tools

The previous analyses are based on sets of non-coding RNAs without focusing on single feature expression or differential expression, one of the key features of miRMaster 2. First, the expression of each miRNA is computed and shown in an aggregated and per-sample manner. The miRNAs can be sorted according to expression levels and pathway analyses using miEAA (all miRNAs) or miRTargetLink (single miRNA-gene interactions) are available. The results table can be downloaded as excel file or in csv format. If an annotation file was specified and a differential expression variable defined, volcano plots are generated. These show the negative decade logarithm of the P-value (adjusted Wilcoxon Mann–Whitney test) versus the $\log_2$ fold change (Figure 3A). From the interactive results table, miRNAs can be selected and boxplots detailing the expression of each sample are computed (Figure 3B/C). This table contains the raw and adjusted P-values of the Wilcoxon Mann–Whitney test as well as the results for a Shapiro Wilk normality test and t-test. If the data are normally distributed, t-test P-values can be used instead of the Wilcoxon Mann–Whitney test P-values. In addition, the table lists ANOVA and Kruskal–Wallis test P-values in case of more than two categories need to be compared per variable. As measure for the effect size, in addition to the fold change, the area under the receiver operator characteristics curve (AUC) is computed, as well as Cohen's d. For the de-regulated miRNAs, miRNA set enrichment analysis as well as over-representation analysis is facilitated through miEAA. The same metrics on differential expression are computed for each feature and all other RNA classes. In the case of other ncRNAs, however, no pathway enrichment analysis is available currently. This limitation might however be solved in the future by adding enrichment analyses for target genes or by integrating future data bases on pathways for other ncRNA classes that are developed.

A key question for researchers is to select the most valid de-regulated candidates and to exclude likely false positives. From our experience, several factors contribute to the success in validating de-regulated miRNAs (41). The miRNA has to show sufficient expression, relevant effect sizes between cases and control and at best a statistically significant difference. To allow users filtering for the best candidates, the results table offers the option to add several filter criteria that are connected by a logical 'and'. Authors can filter for those miRNAs present with at least 1 RPM, having an effect size of at least 0.7 and a P-value <0.05. By adapting the parameters, users can balance for rather specific results or rather sensitive results, depending on the underlying biological question.

Remarkably, miRMaster 2 supports the analysis of multiple comparisons at the same time. If the annotation column that is selected by the user has for example four groups, all pair-wise comparisons ($4 \times 3/2 = 6$) are carried out and presented to the user. Currently, however, only one annotation column can be used at a time for an analysis to avoid too many computations and complicated result representations.

The last analysis module comprises the prediction of new miRNAs. Again, the same information on expression and de-regulation as for the 10 RNA classes is calculated and presented to the user. Here, novel precursor miRNAs found in the data can be selected. An example is a precursor miRNA where the 5′ mature form is annotated while the 3′ that is expressed in the user's data is not yet annotated. For each miRNA candidate, the secondary structure and free energy is computed and presented (Figure 3D). Finally, the expression of reads on the 5′ and 3′ mature form (Figure 3E) is provided as bar-plot along the precursor. Here, users can verify the read stacks manually and in principle observe potential 3′ heterogeneity. Next, all the mapping reads are presented, containing potential isomiRs or other RNA fragments (Figure 3F). This module offers full-download capabilities as enumerated for the other analysis modules.

We clearly consider the respective potentially new miR-NAs as candidates only. It is frequently hard to distinguish between molecules of the different ncRNA classes (e.g. miRNAs that are actually tRNA fragments) or to exclude artifacts. The candidates call for an in-depth experimental validation before they should be considered as miRNAs (42). The validation rate is tightly coupled with the quality of the pre-selected precursors. In our previous study we reached a validation of almost 20%.

In the following sections we provide three use cases, first the analysis of a sncRNA-seq data set from *Mus musculus*, second, human sncRNA data from dementia patients and finally the analysis of single-cell small non-coding RNA data using unique molecular identifiers. All use cases are available on the miRMaster homepage.

Use Case (1): *Mus musculus* sncRNA atlas

To demonstrate the analysis scope of miRMaster for a non-human species, we analyzed sncRNA sequencing data from *M. musculus* (43). Here, for 11 organs and up to 14 replicates of both sexes, sncRNA data using Illumina sequencing have been analyzed. We started miRMaster with the 272 FASTQ files downloaded from SRA, used the tissue and mouse ID as annotation and configured the sex as differential expression variable. The complete results were available after 4 hours. The pre-processing page highlights a large span of sequencing depth, going from a few thousands up to 25 million reads, with a median of 1.9 million reads for the male samples and 5.1 million reads for the female samples. The mapping statistics show overall similar patterns between male and female samples for all considered RNA classes. Interestingly, the miRNA mapping statistics show the highest variability, going from less than 5% mapped reads for some samples, up to over 90% of mapped reads. This variability can mainly be attributed to the RNA composition of the different tissues, as shown by the batch analysis, where 73% of the observed variance can be explained by the tissue variable (Figure 4A). We find that for all other RNA classes, although varying in the proportion of explained variance, the tissue is the strongest factor, except for circRNAs, where most of the observed variance is explained by the sex (Figure 4B). These results are reflected by the
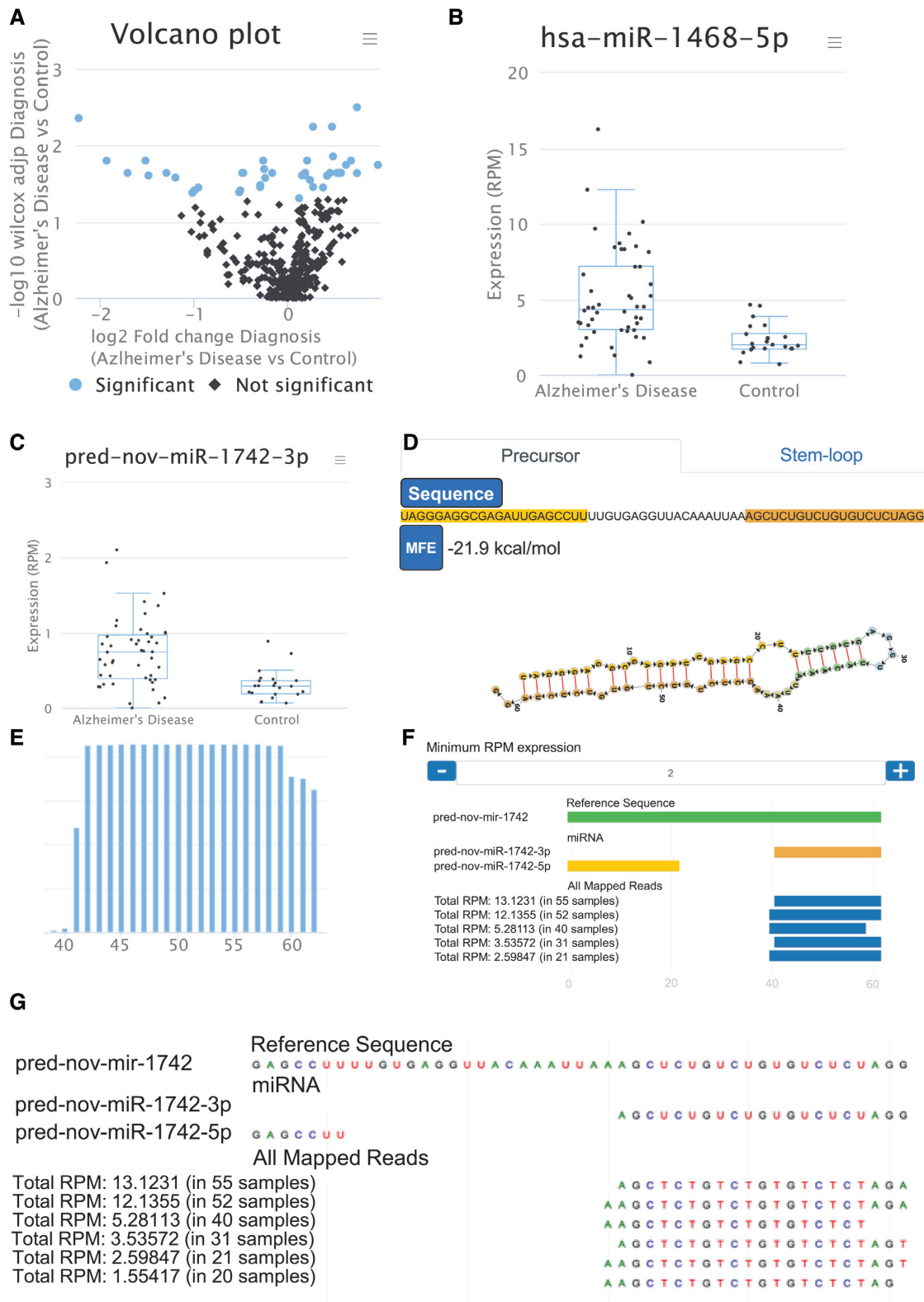
Figure 3. Downstream analyses for AD and control samples. (**A**) Volcano plot displaying each microRNA as a dot. Colored dots are statistically significant (adjusted *P*-value < 0.05). (**B**) For one significant marker (hsa-miR-1468–5p) the data are presented as boxplots. Again, single samples can be highlighted by moving the mouse over. (**C**) Box-plot for a novel miRNA candidate. (**D**) Precursor structure of this novel miRNA candidate and the minimum free energy. Users can switch between the representation of the precursor and the stem-loop. (**E**) Distribution of reads on the mature -3p miRNA of the same candidate precursor. Towards the end of the mature miRNA, the –3p heterogeneity that is typical for miRNAs can be recognized. (**F**) Representation of isoforms. The green bar denotes the precursor, the yellow bar the –5p mature form, the orange bar the –3p form. Each blue bar shows an isoform. The number of reads supporting the isoform can be dynamically adjusted by the user (here, at least 2 RPM are required). (**G**) If users zoom in the representation, the single base resolution per isoform is displayed (in this example, 1 RPM coverage is sufficient).
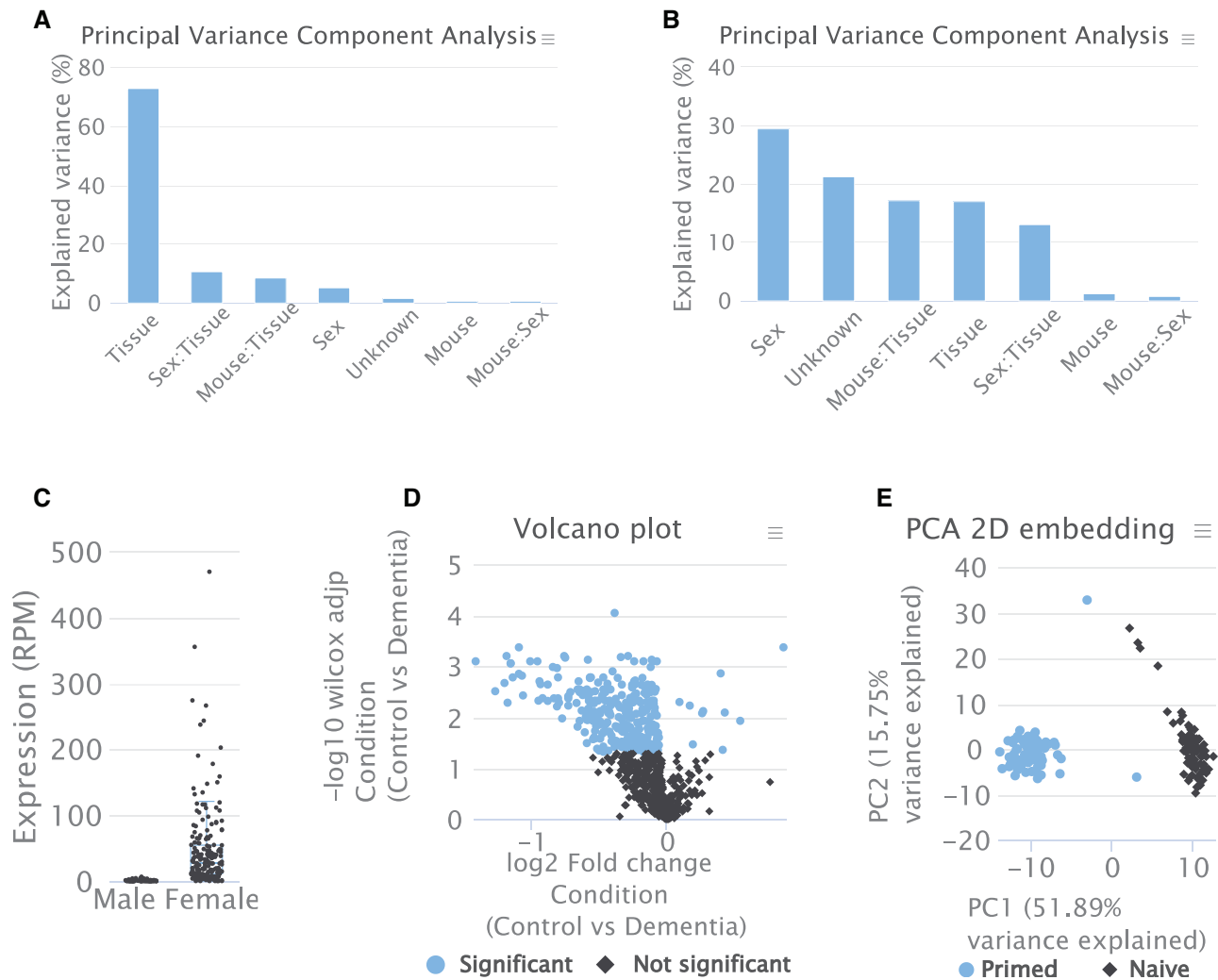
Figure 4. Downstream analysis results for the use cases. (**A**) Principal Variance Component Analysis based on the miRNA expression matrix, showing most of the variance explained by the mouse tissue. (**B**) Principal Variance Component Analysis based on the circRNA expression matrix, showing most of the variance explained by the mouse sex. (**C**) Reads per million (RPM) normalized expression of mmu_circ_0004351. (**D**) Volcano plot showing significantly deregulated miRNAs in dementia patients. (**E**) PCA embedding of the miRNA expression matrix showing a perfect separation between primed and naïve hESCs.

PCA and UMAP embeddings, as well as the sample correlation and expression clustering. The circRNA showing the most significant change between male and female mice was mmu_circ_0004351, with a fold change of 17.5 and an FDR adjusted Wilcoxon Mann-Whitney *P*-value of 4.6 × 10^{-32} (Figure 4C). miRNAs for which only 5% of the observed variance could be attributed to the mouse sex, were also partially differentially expressed, and have been previously reported in literature for single tissues or other species, such as mmu-miR-27a-3p and miR-27b-3p (44,45) and miR-16-5p and miR-21-3p (46).

Use Case (2): Differential sncRNA expression in neurodegeneration

We previously analyzed the role of microRNAs in Alzheimer's disease (4,47,48). Moreover, the examples shown above are from Alzheimer's patients and controls (Figures 2 and 3). Recently, we published a data set com-

posed of dementia patients, including Alzheimer's disease and controls with a new technology, termed CoolMPS (49). To demonstrate the functionality of miRMaster for this sequencing assay, we analyzed the 216 case and control samples. The pre-processing page shows that all samples have been sequenced with more than 13 million reads, obtaining a median of 27.5 million for all samples. The genome mapping statistics show high mapping rates with a median of 95.5% and only few samples exhibiting a mapping rate below 90%. Since the RNAs of this data set were size selected to enrich for miRNAs, we observe the expected high mapping rates to miRNAs as well, with a median of 90.0%. The batch effect analysis suggests that most of the variation cannot be explained by any of the provided variables, followed by the age group and condition for most RNA classes. This is reflected by the sample embeddings as well as sample and expression clusterings, since no separation according to any of the provided annotations can be achieved. The differential

expression results of the RNA classes show a general trend of over-expression in dementia patients, since most RNAs are down-regulated in control patients. We find that 695 miRNAs were expressed with at least 3 reads in more than 50% of either control or dementia patients and that 270 miRNAs were significantly de-regulated with an adjusted Wilcoxon Mann-Whitney *P*-value below 0.05 (Figure 4D). Upon triggering an enrichment analysis with miEAA 2.0, we find 1,790 affected categories comprising known dementia and Alzheimer's disease related pathways such as the positive regulation of endoplasmic reticulum unfolded protein response (FDR adjusted *P*-value of 0.003, (50)) and Rab GTPase binding (FDR adjusted *P*-value 0.001, (51)).

Use Case (3): analysis of data with unique molecular identifiers

Due to the increasing popularity of single cell small non-coding RNA sequencing, we also demonstrate the capabilities of miRMaster on 168 primed and naïve hESCs sequenced with a UMI-based protocol presented by Faridani *et al*. (52). Based on the pre-processing of miRMaster we find that the sequencing depth distribution is similar between primed and naïve hESCs and ranges from 1 million reads up to 44 million with a median of 4.2 million. As expected for small input protocols, and especially in the context of single cell miRNA-seq, only about 50% of the reads were kept after adapter trimming and quality filtering, which can often be attributed to adapter dimerization. The genome mapping statistics show a large variability going from 4.4% up to 88.3%, where on median 0.9% of the reads mapped against microRNAs. The most represented RNA class are rRNAs (median 6.3%) followed by snoRNAs with a median of 3.8%. The batch effect analysis shows that most of the variance in miRNA and snoRNA counts is associated with the cell state (76.3% and 82.9%), in contrast to the other RNA classes, where most of the variance cannot be explained by the cell state. This separation is clearly displayed by the miRNA PCA embedding, where the first component almost perfectly splits the primed from the naïve cells (Figure 4E). The miRNA sample correlation matrix shows high similarity (>0.7) between the cells of each state, while the correlation between the groups is in the range between 0.2 and –0.2. The correlation clustering shows similar patterns for the snoRNAs, however, the correlation coefficients inside the same cell state are higher and in-between cell states vary in the range of 0.7 and 0.85. As highlighted in the publication by Faridani *et al*., the most up-regulated miRNA in naïve hESCs is hsa-miR-371a-5p (fold change of 797, FDR adjusted *P*-value of 1.11×10^{-27}), whereas the most down-regulated miRNA is hsa-miR-363–3p (fold change of 0.003, FDR adjusted *P*-value of 1.07×10^{-29}). It is evident that the current comparably shallow single cell small RNA data sets are not sufficient to detect all the non-coding RNA molecules present in a single cell. Especially the lower abundant non-coding RNA classes might not be sufficiently represented given the current experimental limitations. Nonetheless, miRMaster 2 identifies the molecules that are present in the sequencing data, representing a functional single cell non-coding RNA analysis pipeline.

CONCLUSION AND FUTURE DIRECTION

One of the most important novel features is certainly the support for multiple species, with which we expect to largen our userbase. Furthermore, we added circRNAs as new RNA class. Moreover, the scope of the analysis modules has been widened and we offer new data analysis aspects, such as (i) sample correlation, (ii) expression clustering, (iii) embedding, (iv) batch effect assessment and (v) differential expression analyses. Finally, we extended the adapter trimming procedure to support more major sncRNA-seq library protocols and other custom protocols. In the same context, we implemented support for UMI based analysis, thereby making the tool ready for single cell sncRNA data. In addition to the new features, we performed a major update of the underlying data bases to their current standards, improved the user-experience as well as the representation of results in tables and as interactive plots.

While we incorporated a lot of feedback from researchers that used miRMaster in this update and therefore implemented new features that were missing also from our own experience, we still see potential to further develop the usability and scope of miRMaster. One aspect that we plan to improve is automated quality control. This includes a prediction of the uploaded sample type. By using annotated data, the solid tissue or body fluid from which the data were generated can be predicted with over 90% accuracy (43). A further step is to improve the functional support for viruses and bacteria. Originally, we implemented this step for detecting contamination or reporting the presence of exogenous species. While we tested this feature only using expression data from Myxobacteria, we now aim to design a dedicated analysis module for sRNAs from microorganisms. In our ambition to develop an AI-based quality control we plan to implement automated outlier detection and propose users to perform the computational analyses after excluding flagged outliers.

To further advance the development of miRMaster, we encourage the community to continue providing us constant feedback as well as to propose new features that are of broad interest.

DATA AVAILABILITY

miRMaster 2 is freely available at https://www.ccb.uni-saarland.de/mirmaster2.

REFERENCES

1. Bartel,D.P. (2018) Metazoan microRNAs. *Cell*, **173**, 20–51.
2. Stavast,C.J. and Erkeland,S.J. (2019) The non-canonical aspects of microRNAs: many roads to gene regulation. *Cells*, **8**, 1465.
3. Fehlmann,T., Backes,C., Kahraman,M., Haas,J., Ludwig,N., Posch,A.E., Wurstle,M.L., Hubenthal,M., Franke,A., Meder,B. *et al.* (2017) Web-based NGS data analysis using miRMaster: a large-scale meta-analysis of human miRNAs. *Nucleic Acids Res.*, **45**, 8731–8744.

4. Fehlmann,T., Meese,E. and Keller,A. (2017) Exploring ncRNAs in Alzheimer's disease by miRMaster. *Oncotarget*, **8**, 3771–3772.

5. Chen,L., Heikkinen,L., Wang,C., Yang,Y., Sun,H. and Wong,G. (2019) Trends in the development of miRNA bioinformatics tools. *Brief. Bioinform.*, **20**, 1836–1852.

6. Schmartz,G.P., Kern,F., Fehlmann,T., Wagner,V., Fromm,B. and Keller,A. (2020) Encyclopedia of tools for the analysis of miRNA isoforms. *Brief. Bioinform.*, doi:10.1093/bib/bbaa346.

7. Kesharwani,R.K., Chiesa,M., Bellazzi,R. and Colombo,G.I. (2019) CBS-miRSeq: a comprehensive tool for accurate and extensive analyses of microRNA-sequencing data. *Comput. Biol. Med.*, **110**, 234–243.

8. Kanke,M., Baran-Gale,J., Villanueva,J. and Sethupathy,P. (2016) miRquant 2.0: an expanded tool for accurate annotation and quantification of MicroRNAs and their isomiRs from small RNA-sequencing Data. *J Integr Bioinform*, **13**, 307.

9. Aparicio-Puerta,E., Lebron,R., Rueda,A., Gomez-Martin,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.

10. Vitsios,D.M. and Enright,A.J. (2015) Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics*, **31**, 3365–3367.

11. Shi,J., Dong,M., Li,L., Liu,L., Luz-Madrigal,A., Tsonis,P.A., Del Rio-Tsonis,K. and Liang,C. (2015) mirPRo-a novel standalone program for differential expression and variation analysis of miRNAs. *Sci. Rep.*, **5**, 14617.

12. Lu,Y., Baras,A.S. and Halushka,M.K. (2018) miRge 2.0 for comprehensive analysis of microRNA sequencing data. *BMC Bioinformatics*, **19**, 275.

13. Wan,C., Gao,J., Zhang,H., Jiang,X., Zang,Q., Ban,R., Zhang,Y. and Shi,Q. (2017) CPSS 2.0: a computational platform update for the analysis of small RNA sequencing data. *Bioinformatics*, **33**, 3289–3291.

14. Kuksa,P.P., Amlie-Wolf,A., Katanic,Z., Valladares,O., Wang,L.S. and Leung,Y.Y. (2018) SPAR: small RNA-seq portal for analysis of sequencing experiments. *Nucleic Acids Res.*, **46**, W36–W42.

15. Lott,S.C., Schafer,R.A., Mann,M., Backofen,R., Hess,W.R., Voss,B. and Georg,J. (2018) GLASSgo - automated and reliable detection of sRNA homologs from a single input sequence. *Front Genet*, **9**, 124.

16. Seguin,J., Otten,P., Baerlocher,L., Farinelli,L. and Pooggin,M.M. (2016) MISIS-2: a bioinformatics tool for in-depth analysis of small RNAs and representation of consensus master genome in viral quasispecies. *J. Virol. Methods*, **233**, 37–40.

17. Ebert,M.S., Neilson,J.R. and Sharp,P.A. (2007) MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat. Methods*, **4**, 721–726.

18. Zhang,J., Liu,L., Xu,T., Xie,Y., Zhao,C., Li,J. and Le,T.D. (2019) miRspongeR: an R/Bioconductor package for the identification and analysis of miRNA sponge interaction networks and modules. *BMC Bioinformatics*, **20**, 235.

19. Desvignes,T., Batzel,P., Sydes,J., Eames,B.F. and Postlethwait,J.H. (2019) miRNA analysis with Prost! reveals evolutionary conservation of organ-enriched expression and post-transcriptional modifications in three-spined stickleback and zebrafish. *Sci. Rep.*, **9**, 3913.

20. Zhang,Y., Zang,Q., Zhang,H., Ban,R., Yang,Y., Iqbal,F., Li,A. and Shi,Q. (2016) DeAnnIso: a tool for online detection and annotation of isomiRs from small RNA sequencing data. *Nucleic Acids Res.*, **44**, W166–W175.

21. Chen,Y., Ruan,Z.R., Wang,Y., Huang,Q., Xue,M.Q., Zhou,X.L. and Wang,E.D. (2018) A threonyl-tRNA synthetase-like protein has tRNA aminoacylation and editing activities. *Nucleic Acids Res.*, **46**, 3643–3656.

22. Winek,K., Lobentanzer,S., Nadorp,B., Dubnov,S., Dames,C., Jagdmann,S., Moshitzky,G., Hotter,B., Meisel,C., Greenberg,D.S. *et al.* (2020) Transfer RNA fragments replace microRNA regulators of the cholinergic poststroke immune blockade. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 32606–32616.

23. Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.

24. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G.,

Bennett,R. *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.

25. Consortium,R.N. (2021) RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res.*, **49**, D212–D220.

26. Chan,P.P. and Lowe,T.M. (2016) GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic Acids Res.*, **44**, D184–D189.

27. Glazar,P., Papavasileiou,P. and Rajewsky,N. (2014) circBase: a database for circular RNAs. *RNA*, **20**, 1666–1670.

28. Zhao,Y., Li,H., Fang,S., Kang,Y., Wu,W., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. *et al.* (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.*, **44**, D203–D208.

29. Koster,J. and Rahmann,S. (2018) Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, **34**, 3600.

30. Fernandez,N.F., Gundersen,G.W., Rahman,A., Grimes,M.L., Rikova,K., Hornbeck,P. and Ma'ayan,A. (2017) Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Sci Data*, **4**, 170151.

31. 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.

32. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

33. Desvignes,T., Loher,P., Eilbeck,K., Ma,J., Urgese,G., Fromm,B., Sydes,J., Aparicio-Puerta,E., Barrera,V., Espin,R. *et al.* (2020) Unification of miRNA and isomiR research: the mirGFF3 format and the mirtop API. *Bioinformatics*, **36**, 698–703.

34. Friedlander,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.

35. Backes,C., Meder,B., Hart,M., Ludwig,N., Leidinger,P., Vogel,B., Galata,V., Roth,P., Menegatti,J., Grasser,F. *et al.* (2016) Prioritizing and selecting likely novel miRNAs from NGS data. *Nucleic Acids Res.*, **44**, e53.

36. Solomon,J., Kern,F., Fehlmann,T., Meese,E. and Keller,A. (2020) HumiR: web services, tools and databases for exploring human microRNA Data. *Biomolecules*, **10**, 1576.

37. Backes,C., Khaleeq,Q.T., Meese,E. and Keller,A. (2016) miEAA: microRNA enrichment analysis and annotation. *Nucleic Acids Res.*, **44**, W110–W116.

38. Kern,F., Fehlmann,T., Solomon,J., Schwed,L., Grammes,N., Backes,C., Van Keuren-Jensen,K., Craig,D.W., Meese,E. and Keller,A. (2020) miEAA 2.0: integrating multi-species microRNA enrichment analysis and workflow management systems. *Nucleic Acids Res.*, **48**, W521–W528.

39. Kern,F., Amand,J., Senatorov,I., Isakova,A., Backes,C., Meese,E., Keller,A. and Fehlmann,T. (2020) miRSwitch: detecting microRNA arm shift and switch events. *Nucleic Acids Res.*, **48**, W268–W274.

40. Hamberg,M., Backes,C., Fehlmann,T., Hart,M., Meder,B., Meese,E. and Keller,A. (2016) MiRTargetLink–miRNAs, genes and interaction networks. *Int. J. Mol. Sci.*, **17**, 564.

41. Backes,C., Sedaghat-Hamedani,F., Frese,K., Hart,M., Ludwig,N., Meder,B., Meese,E. and Keller,A. (2016) Bias in high-throughput analysis of miRNAs and implications for biomarker studies. *Anal. Chem.*, **88**, 2088–2095.

42. Alles,J., Fehlmann,T., Fischer,U., Backes,C., Galata,V., Minet,M., Hart,M., Abu-Halima,M., Grasser,F.A., Lenhof,H.P. *et al.* (2019) An estimate of the total number of true human miRNAs. *Nucleic Acids Res.*, **47**, 3353–3364.

43. Isakova,A., Fehlmann,T., Keller,A. and Quake,S.R. (2020) A mouse tissue atlas of small noncoding RNA. *Proc. Natl. Acad. Sci. U.S.A.*, **117**, 25634–25645.

44. Guo,D., Ye,Y., Qi,J., Tan,X., Zhang,Y., Ma,Y. and Li,Y. (2017) Age and sex differences in microRNAs expression during the process of thymus aging. *Acta Biochim. Biophys. Sin. (Shanghai)*, **49**, 409–419.

45. Hermenean,A., Trotta,M.C., Gharbia,S., Hermenean,A.G., Peteu,V.E., Balta,C., Cotoraci,C., Gesualdo,C., Rossi,S., Gherghiceanu,M. *et al.* (2020) Changes in retinal structure and ultrastructure in the aged mice correlate with differences in the expression of selected retinal miRNAs. *Front Pharmacol*, **11**, 593514.

46. Hasakova,K., Bezakova,J., Vician,M., Reis,R., Zeman,M. and Herichova,I. (2017) Gender-dependent expression of leading and passenger strand of miR-21 and miR-16 in human colorectal cancer and adjacent colonic tissues. *Physiol. Res.*, **66**, S575–S582.
47. Keller,A., Backes,C., Haas,J., Leidinger,P., Maetzler,W., Deuschle,C., Berg,D., Ruschil,C., Galata,V., Ruprecht,K. *et al.* (2016) Validating Alzheimer's disease micro RNAs using next-generation sequencing. *Alzheimers Dement*, **12**, 565–576.
48. Ludwig,N., Fehlmann,T., Kern,F., Gogol,M., Maetzler,W., Deutscher,S., Gurlit,S., Schulte,C., von Thaler,A.K., Deuschle,C. *et al.* (2019) Machine learning to detect Alzheimer's disease from circulating non-coding RNAs. *Genomics Proteomics Bioinformatics*, **17**, 430–440.
49. Li,Y., Fehlmann,T., Borcherding,A., Drmanac,S., Liu,S., Groeger,L., Xu,C., Callow,M., Villarosa,C., Jorjorian,A. *et al.* (2021) CoolMPS: evaluation of antibody labeling based massively parallel non-coding RNA sequencing. *Nucleic Acids Res.*, **49**, e10.
50. Scheper,W. and Hoozemans,J.J. (2015) The unfolded protein response in neurodegenerative diseases: a neuropathological perspective. *Acta Neuropathol.*, **130**, 315–331.
51. Zhang,X., Huang,T.Y., Yancey,J., Luo,H. and Zhang,Y.W. (2019) Role of Rab GTPases in Alzheimer's sisease. *ACS Chem. Neurosci.*, **10**, 828–838.
52. Faridani,O.R., Abdullayev,I., Hagemann-Jensen,M., Schell,J.P., Lanner,F. and Sandberg,R. (2016) Single-cell sequencing of the small-RNA transcriptome. *Nat. Biotechnol.*, **34**, 1264–1266.