LiqDB Manual

liqDB is a browsable and interactive database for small RNA abundance in biofluids.

The database can be browsed at a SRA study level and the users can analyse customized sample set and compare them to their own data.

Customized microRNA expression matrices can be downloaded, and the impact of confounding variables can be addressed. MicroRNAs can be analysed individually as a function of several variables like RNA extraction, phenotype, sex, library preparation or biofluid providing both, the microRNAs with highest and lowest fluctuation. MicroRNAs with low fluctuation can be valuable molecules for qPCR controls.

Scope

The description of all studies can be found under the following link:

http://bioinfo5.ugr.es/liqdb/studies

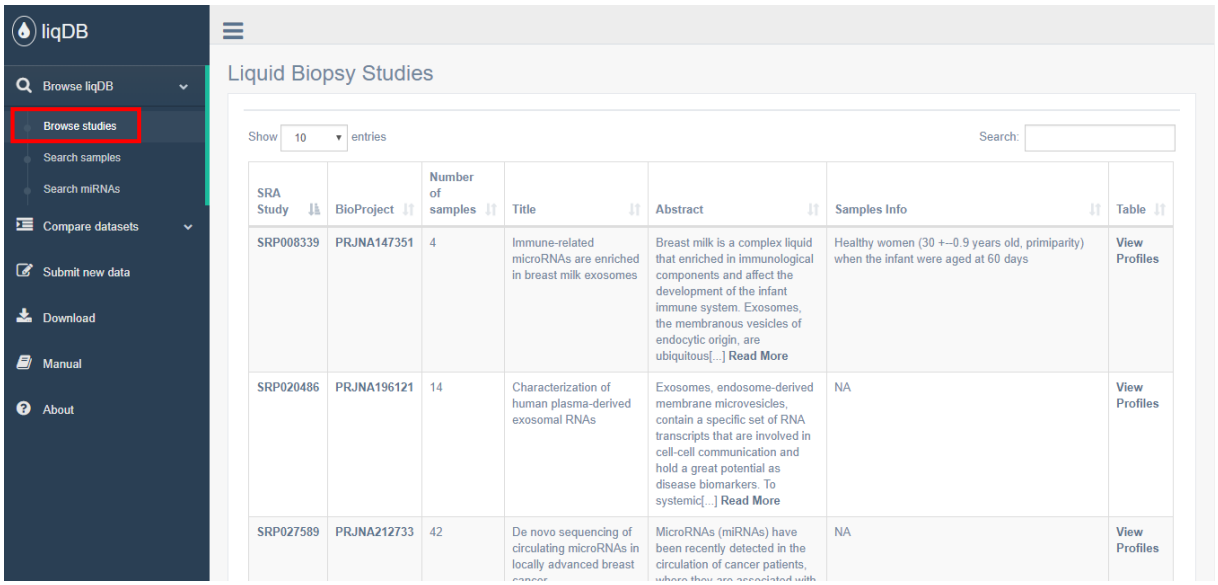

Figure 1. ***'Browse studies' section****.*

Modes of usage

Browse a study

A typical result page can be seen in this test data (*study SRP029994*).

The user can analyse miRNA expression profiles (**miRNA profiles tab**), analyse the relative frequencies of the different small RNAs (or fragments of other small ncRNAs like tRNAs, yRNAs or vRNAs), assess differential expression (**differential expression tab**) and check whether the samples contain material from bacteria or virus (**species distribution tab**).

miRNA profiles tab:

Three box plots are generated:

- The 20 microRNAs with highest expression values.
- The 20 microRNAs with highest CV (microRNAs most affected by any of the known or unknown variables). Note that not necessarily these microRNAs are associated statistically to the variable of interest (like health status like health/cancer).
- The 20 microRNAs with lowest coefficient of variation - CV (putative molecules for standardizing qPCR validation experiments).

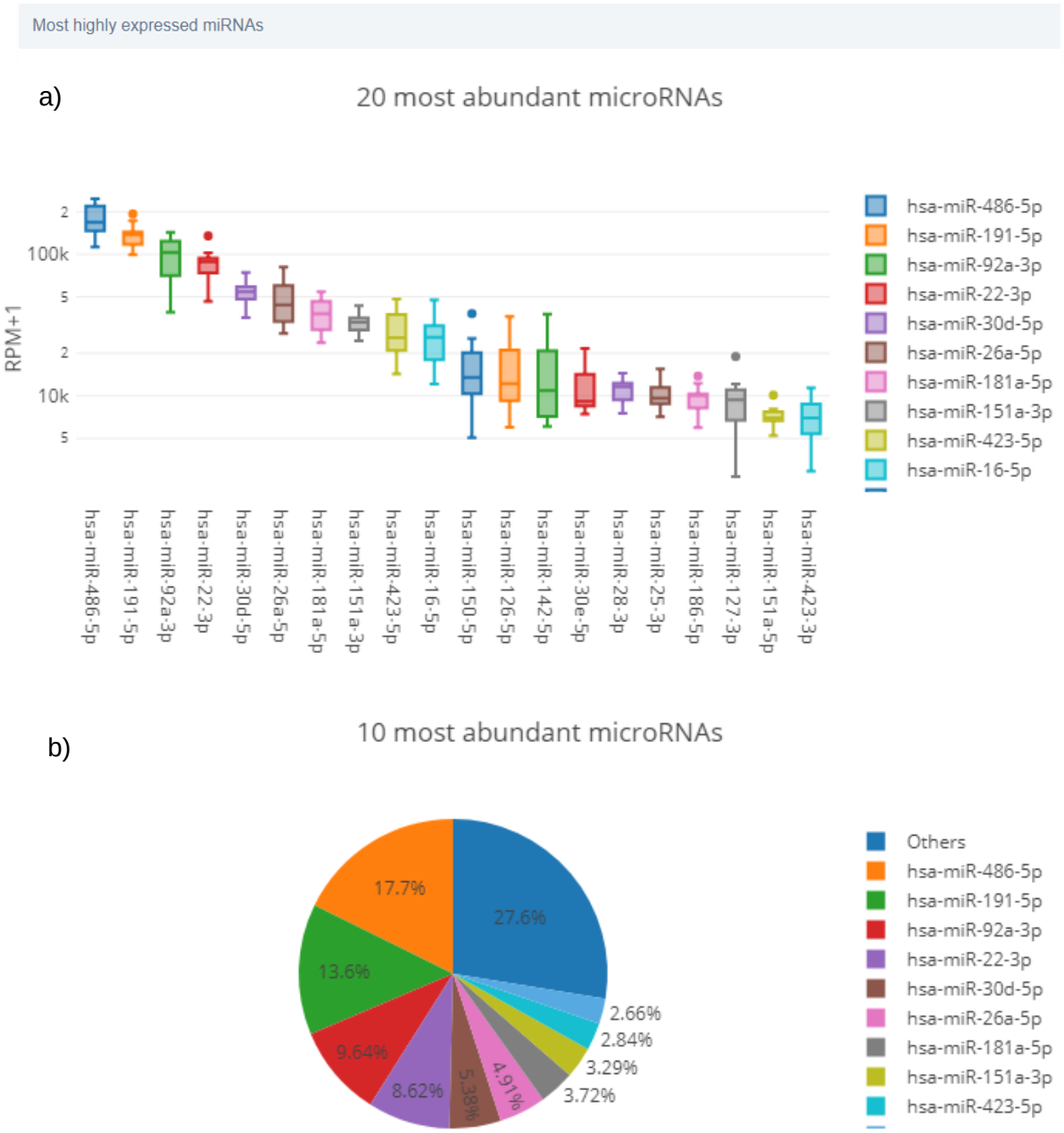

Figure 2. ***'miRNA profiles' tab – Most abundant miRNAs chart and pie chart****.* **a)** It shows the 20 most abundant miRNAs distribution in the study sorted by Reads Per Million+1. **b)** This pie chart allows for seeing how expressed a few miRNAs are, comparing to the total amount of them. In this example, the 10 most expressed miRNAs form the 72.4% of all miRNA abundance.

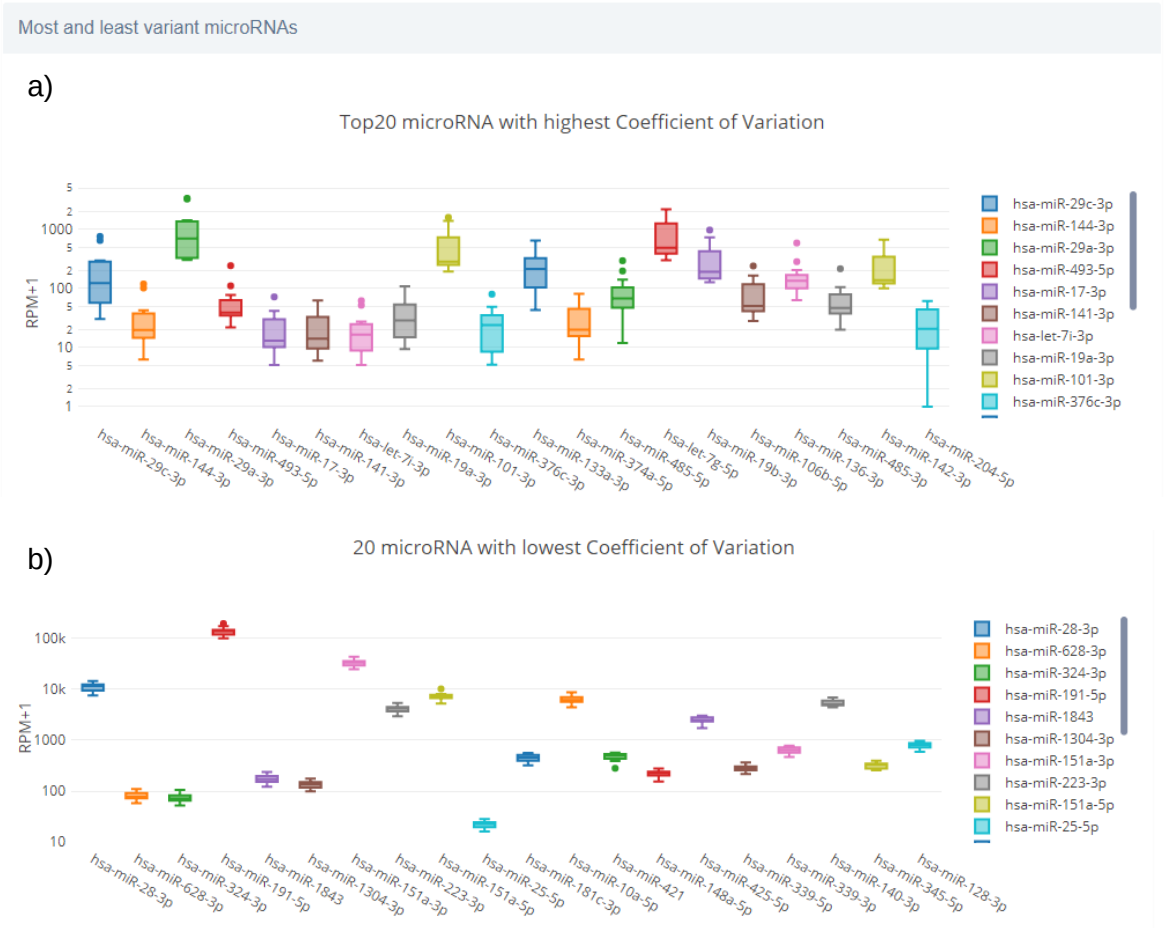

*Figure 3. '**miRNA profiles' tab – miRNAs with highest and lowest CV**.* It shows the 20 miRNAs with highest (**a**) and lowest (**b**) Coefficient of Variation, so, the user can visualize the dispersion of RPM values in a not depending on the magnitude measure. In this example, we could study some miRNAs (such as *hsa-miR-28-3p* or *hsa-miR-628-3p*) to use as reference microRNAs to normalize or standardize qPCR validation experiments.

Species distribution tab:

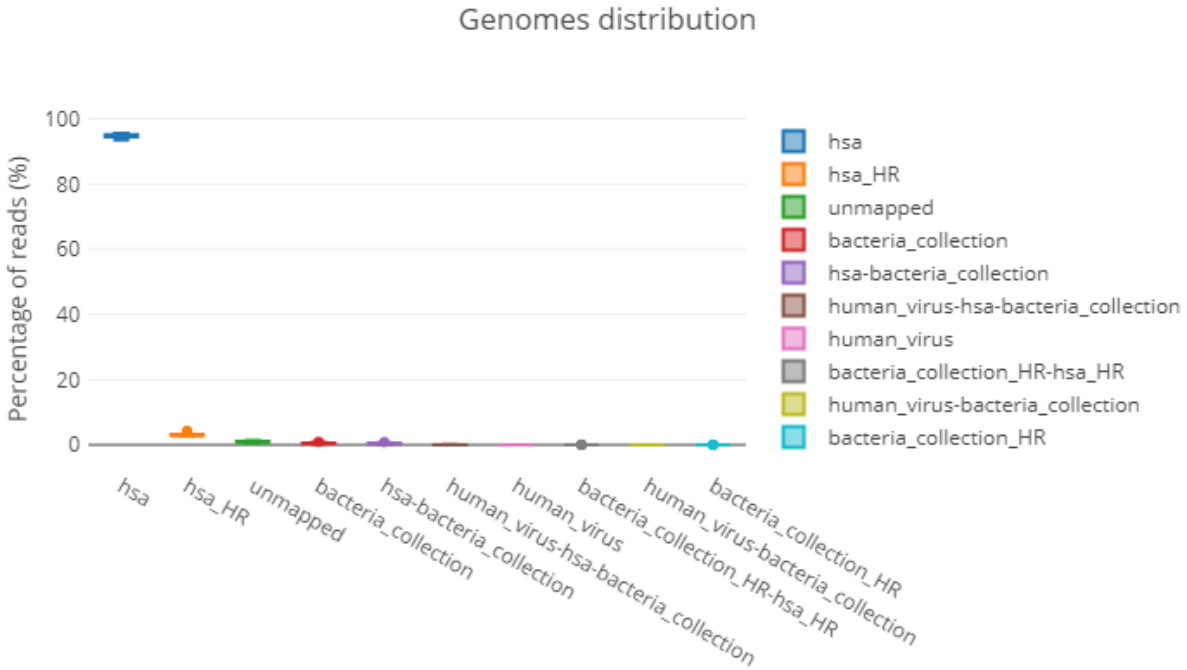

*Figure 4. **'Species distribution' tab**.*

Differential expression tab:

Figure 5. ***'Differential expression' tab with heatmap (one study)****.* This tab shows miRNA expression that has been found to be differentially expressed between some conditions (non-fatigued individuals or individuals with chronic fatigue syndrome, in this case). The user can also visualize a hierarchical clustered heatmap, which shows relationship between experiments and miRNAs expression.

Download tab:

In this section, there are three different files to download:

- Expression matrix for Read Per Millions of multiple mapping adjusted read counts (library normalized).

- Expression matrix for Raw reads (adjusted for multiple mapping).
- Complete data analysis of the study or query samples. It contains several files that are used to generate tables, such as:
    - Expression matrices (normal, sorted and with Coefficient of Variation) for Read Per Millions of multiple mapping adjusted read counts (library normalized)
    - Expression matrices (normal and annotated) for Raw reads (adjusted for multiple mapping).
    - Expression matrices (normal and sorted) for Genome Distribution: (HR = reads that map more than 20 times to the genome).
    - Expression matrices (normal and sorted) for RNA distribution. Percentages of RNA types found in the sample/study.
    - A log file of the analysis process.

Search for a microRNA

For each microRNA detected in the sample, the user can visualize its distribution by means of box plots for the following variables: biofluid type (*figure 7*), health state, gender, RNA extraction protocol, library preparation protocol, microvesicle extraction (yes or no) and study.

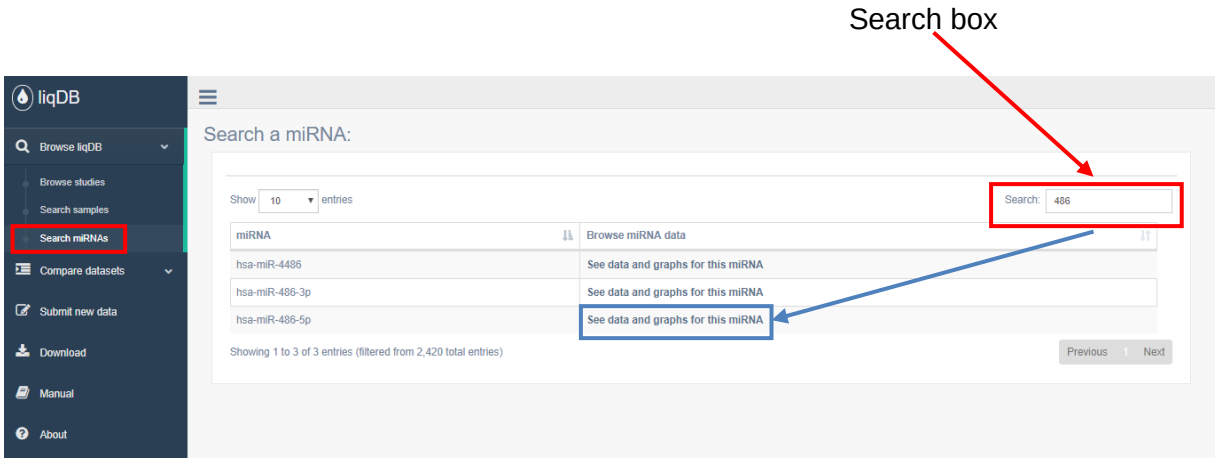

Figure 6. ***'Search miRNAs' section.***

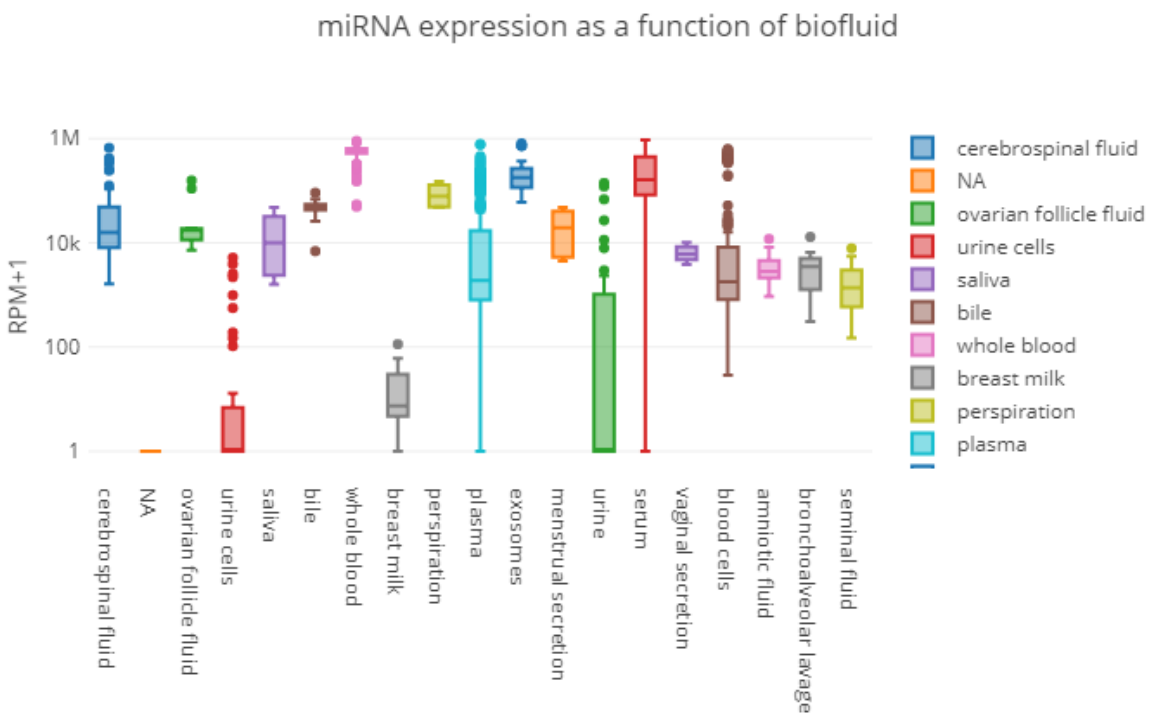

*Figure 7. **miRNA expression as a function of biofluid**.*

Explore a set of samples

The user can generate a set of samples analysing several features:

- The microRNA expression profiles by means of a heatmap.
- Percentage of reads that map to the human genome, to bacteria, virus or are unmapped.
- Relative frequency of the different RNA types like miRNA, vault-RNA, yRNA, tRNA fragments, other ncRNA, etc.

Compare datasets

Compare two sets of samples contained in liqDB

The user can select two different sets analysing all features provided in the 'Explore a set of samples' section but additionally differential expression is assessed (*figure 8*). The following data is provided:

- A table with the comparisons of all groups (sometimes more than one group exist) providing group means and standard deviations, fold-change, p-value and Bonferroni corrected p-values. The statistically significance is assessed by means of the t-test.

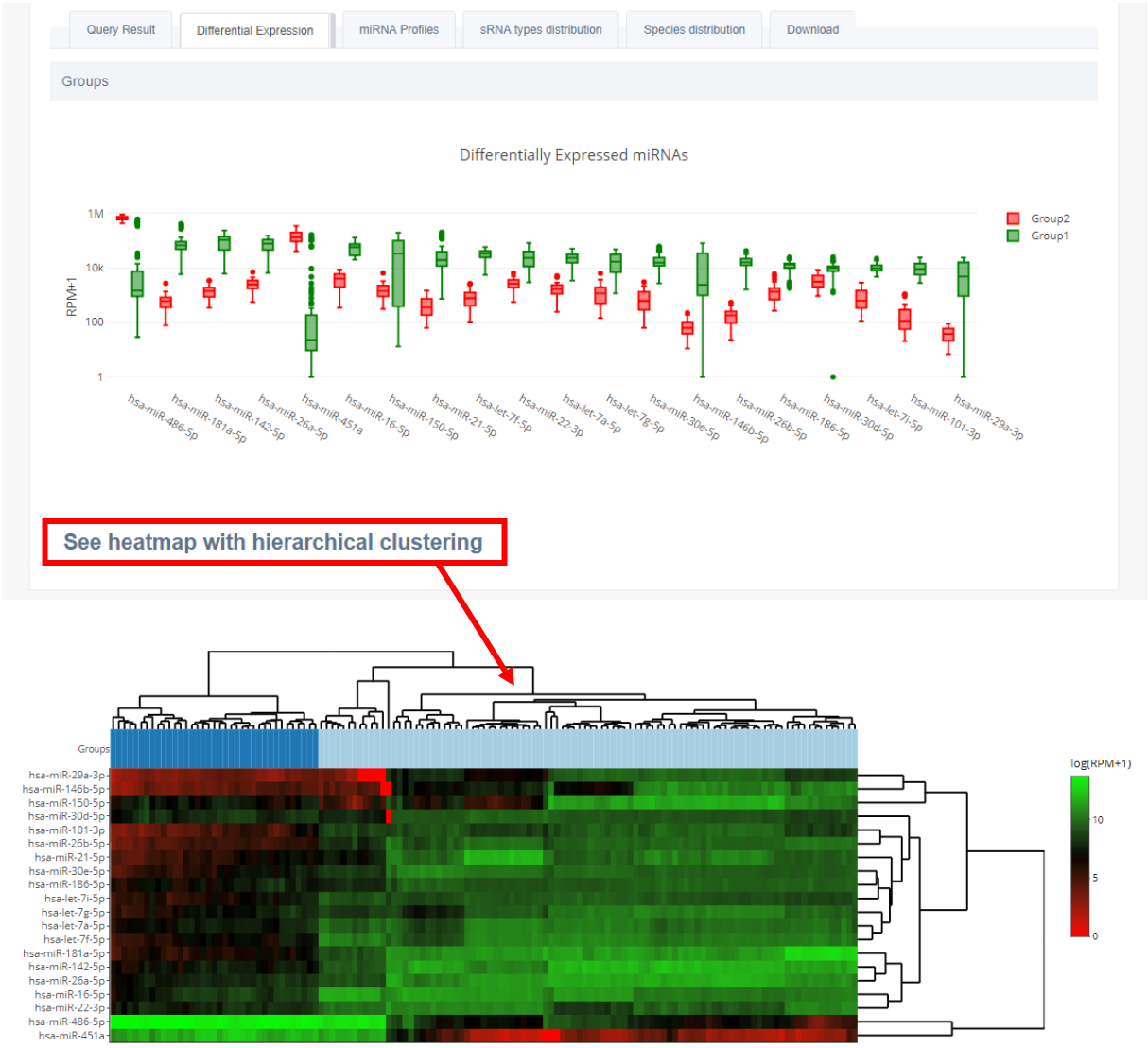

Figure 8. ***'Differential expression' tab with heatmap (comparing two sets of samples)****.* This section shows miRNA expression that has been found to be differentially expressed between some conditions. The user can also visualize a hierarchical clustered heatmap.

Compare a set of samples from liqDB to user provided data

The user can compare own data to the profiles stored in the database. liqDB is linked to sRNAtoolbox and therefore the user can first analyse the raw data (fastq or other suitable formats) with sRNAbench. The samples can then be specified by means of the sRNAbench jobID. The output is exactly the same as explained in the last section (Compare two set of samples...).

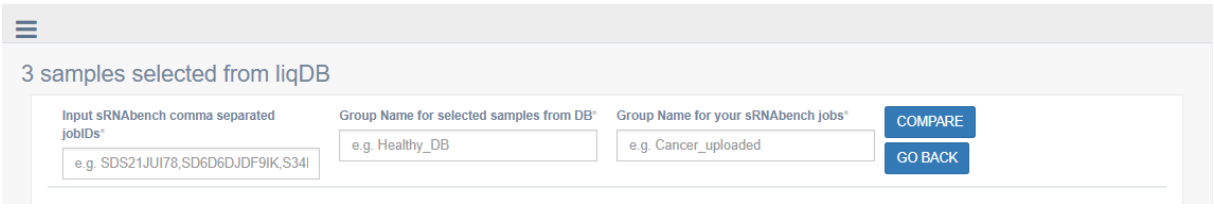

Figure 9. ***'Compare with your data' section - User provided data****.*

Basic statistics

In this section, the user can visualize some statistics tables of miRNA-mapping read count grouped by five different variables (SRA study, sex, library preparation method, biofluid and extraction method).

The statistical parameters provided are the number of samples, mean, standard deviation, minimum and maximum value, percentile 25, median and percentile 75.

Data generation

Below we describe briefly how the raw data was processed and how the expression profiles were generated.

Used expression values and relative abundances

All microRNAs plots and tables are based on two different expression values:

Multiple mapping adjusted raw reads counts:

These values are generated by dividing the read count (RC) by the number of times the read maps to a certain annotation. For reference sequences with several loci in the genome, the adjusted read counts are finally summed up giving only one value per reference sequence (i.e. mature microRNA sequence). The expression matrices of the adjusted raw reads can be downloaded.

RPM values using library normalization:

The adjusted read counts are divided by the total number of reads mapped to a certain annotation (microRNAs) times 1,000,000. All plots (box plots and heatmaps) and the differential expression analysis are based on this measure.

Differential expression

Expression matrices are generated using the RPM values and the differential expression is assessed by means of the t-test. If more than one group exists (like when comparing 3 different biofluids), all possible combinations are calculated.

Data processing protocol

The data was downloaded from SRA and processed with sRNAbench. The adapters were either obtained from the sample descriptions or determined using the ***guessAdapter*** option implemented in sRNAbench.

For expression profiling, the following protocol implemented into sRNAbench was used:

- Detect and remove the adapter (and barcodes or random adapters).
- Collapse the fastq into unique reads (UR) assigning a read count (the number of times a molecule was sequenced, RC) to each unique read.
- Map the reads to the human genome (GRCh38, patch 10), a collection of bacterias and human virus sequences using bowtie1 (seed alignment with seed length 19nt

and one mismatch). This setting was selected in order to detect isomiRs as well.
- Use reference sequence explained below to annotate the genome mapped reads. The annotations are performed in the exact order described below. After each library, the assigned reads are removed in order to avoid cross-library assignments.

Reference libraries

For profiling, the following libraries were used:

- ***miRNAs***: miRBase v22 & MirGeneDB v2.0 (only sequences that are not in miRBase).
- tRNAs: GtRNAdb, a genomic tRNA database (http://gtrnadb.ucsc.edu/).
- vault-RNA, yRNA and guide RNAs extracted from RefSeq (https://www.ncbi.nlm.nih.gov/refseq/).
- Non-coding RNAs from Ensembl (Release 91).
- Non-coding RNAs from RNAcentral release 9 (http://rnacentral.org/).
- RNA sequences of coding genes from Ensembl (Release 91).
- Bacteria genome sequences collection from Bacteria Ensembl (Release 39).
- Human-hosted virus sequences collection from EnsemblGenomes.