Tutorial miFRame

Data Input

miFRame gets NGS data along with a parameter set as input. The parameter input can be easily done via the web-interface and for NGS data different file types care supported as input. First, raw NGS files can be uploaded (Illumina fastq single end read files). These files are then analyzed using miRDeep2 with standard parameters. For each miRNA precursor, an annotated file is generated. This file contains the read counts per base and per sample for the respective miRNA. Additionally, users can carry out the miRDeep2 analysis locally and directly upload the miRDeep result files (option 2, miRDeep2 arf files), it is however also possible to use any other miRNA annotation tool and to generate the annotation files for upload with miFRame (option 3). While the third option allows for largest flexibility and transfer of small data sets in the range of few megabytes, substantial local pre-processing by the user is required. While the first option, NGS raw data, is most straightforward, large data sets in the range of several gigabytes have to be uploaded by users. Our tool offers also the potential to share data with collaborators. Each analysis gets a unique ID. With that ID and an optional password, collaboration partners can directly access the analysis results. A more detailed description of the different input formats is available on the miFRame main webpage (http://main.ccb.uni-saarland.de/miframe/).

As parameters the user can specify significance threshold, select whether just single positions or many bases across the precursor have to be significantly changed, a window size for detection of iso-microRNAs and the different quantitative and qualitative analyses to be carried out. Moreover, the user can specify the statistical test to be applied (parametric t-test for normally distributed data or Wilcoxon Mann-Whitney test) and the graphics output format (pdf vector graphics, jpeg or png). Finally, the user can decide whether the data should be used for the aggregation analysis.

Analysis

Expression analysis: Here, miFRame tests the hypothesis that all reads mapping to the precursor are differentially expressed between a case and control cohort. The mean read count per sample across the precursor is calculated. By applying the hypothesis test specified by the user, average count in case and control cohort is compared to each other

and a significance value per miRNA precursor is calculated.

Mature expression analysis: In this step, miFRame tests the hypothesis that miRNA precursors are differentially expressed between cases and controls. To this end, the average read count per sample on each mature 3’ and 5’ miRNA is calculated and by applying the hypothesis test specified by the user, a significance value is calculated.

3’ to 5’ expression analysis ratio: Here, miFRame searches for miRNAs that show either opposite or same differential expression between cases and controls on the 3’ compared to the 5’ mature forms of miRNAs. First, miRNAs with just a single mature form are omitted, first. Then, for each remaining miRNA the fold changes and significance values for the 3’ and 5’ forms are determined analogously to the mature expression analysis. From these two fold changes, the difference between the 3’ and the 5’ mature form is calculated. miRNA precursors with resulting expression quotients close to 1 show the same up- or down-regulation of both mature forms, while those miRNAs with fold changes >> 1 are more up-regulated in the 5’ mature form and miRNAs with fold changes << 1 are more up-regulated in the 3’ form.

Per-base expression analysis: In this analysis, miFRame checks how many base positions across the precursor of each miRNA are significantly differentially expressed in cases compared to controls. Most computational approaches just consider either maximal or average expression across the mature forms of miRNAs. In several cases, however, just few positions, especially at the edge of the mature miRNAs are significantly different between two cohorts. These sites are detected by the per-base expression analysis but are usually not discovered by considering the full precursor or mature miRNAs (see also window analysis).

Novel mature miRNA analysis: For a substantial amount of miRNAs in the miRBase just a single mature form is known. However, in many cases a significant number of reads maps exactly to positions matching the second mature form of known miRNAs. To this end, miFRame calculates for those miRNAs where just a single mature form is known whether reads are matching to the second mature form. Additionally, miFRame calculates whether this second mature form shows significantly different expression in cases as compared to controls and outputs how many reads are located on this potential novel mature miRNA.

Window analysis: As described in the per base analysis section, it is often the case that only a few single positions, predominantly at the edge of miRNAs are significant or show at least a higher significance value than the remaining part of the precursor or mature forms

of the miRNA. These sites are frequently not taken into account in quantitative analyses or are down-weighted by averaging across the mature miRNAs. miFRame allows the user to select a window size in the parameter selection step. Then, it calculates the percentage of bases covered in this window with respect to the 3p and 5p end of the 3’ and 5’ mature miRNA form. In consequence, for each precursor with two mature forms 4 percentages in cases and controls are calculated along with the difference between the two cohorts. Additionally, miFRame also outputs the respective motifs at the edge of the miRNAs. Thus, mature miRNAs with isoforms in specific traits can be easily detected.

Aggregation analysis: In case the user decided to carry out the aggregation analysis, just the relevant findings from the current study are stored anonymously in a local database. In return, the relevant findings from that user are compared to the database such that it becomes clear whether e.g. a novel mature miRNA is specific for a certain trait or is reported in many experiments, making a false positive finding by NGS artifacts more likely. This analysis may help to prioritize the replication and validation experiments by selecting findings that are less frequently discovered.