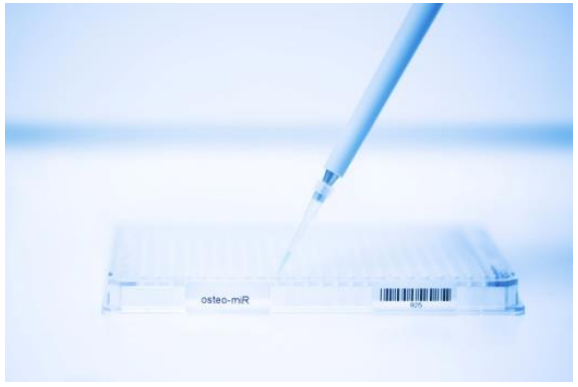
TAmiRNA
stability for life.

Exemplary report

Version 1



Date: 23.05.2017

Responsible Author: NN

Outline

1. Experiment overview
2. Data quality
3. Exploratory data analysis
4. Differential gene expression analysis
5. Project summary & next steps

Experiment overview *(example)*

- Aim: Exploring dosage-dependent effects of a novel anti-oxidative stress agent on intra- and extracellular microRNA levels in a human in-vitro model
- 36 samples were provided by customer as defined at project start
- Quality control was performed at TAmiRNA prior to sample selection
- 18 samples cell pellets and supernatants were sequenced, 36 in total
  → see table on the following slide

Outline

1. Experiment overview
2. Data quality
3. Exploratory data analysis
4. Differential gene expression analysis
5. Project summary & next steps

NGS data quality

- We dedicate at least 1 page to various quality aspects of the analytical workflow and data analysis for NGS experiments. In addition, raw data, figures, tables etc. are made available to the customer on a flash drive.

1. NGS library QC: purity and quantity were checked for each samples individually as well as the sequencing pool
2. Sequencing reads QC: quality score thresholds
3. Read size distribution, total read distribution and mapping statistics to reference data sets
4. Total number of microRNAs identified in each sample above threshold

Read length

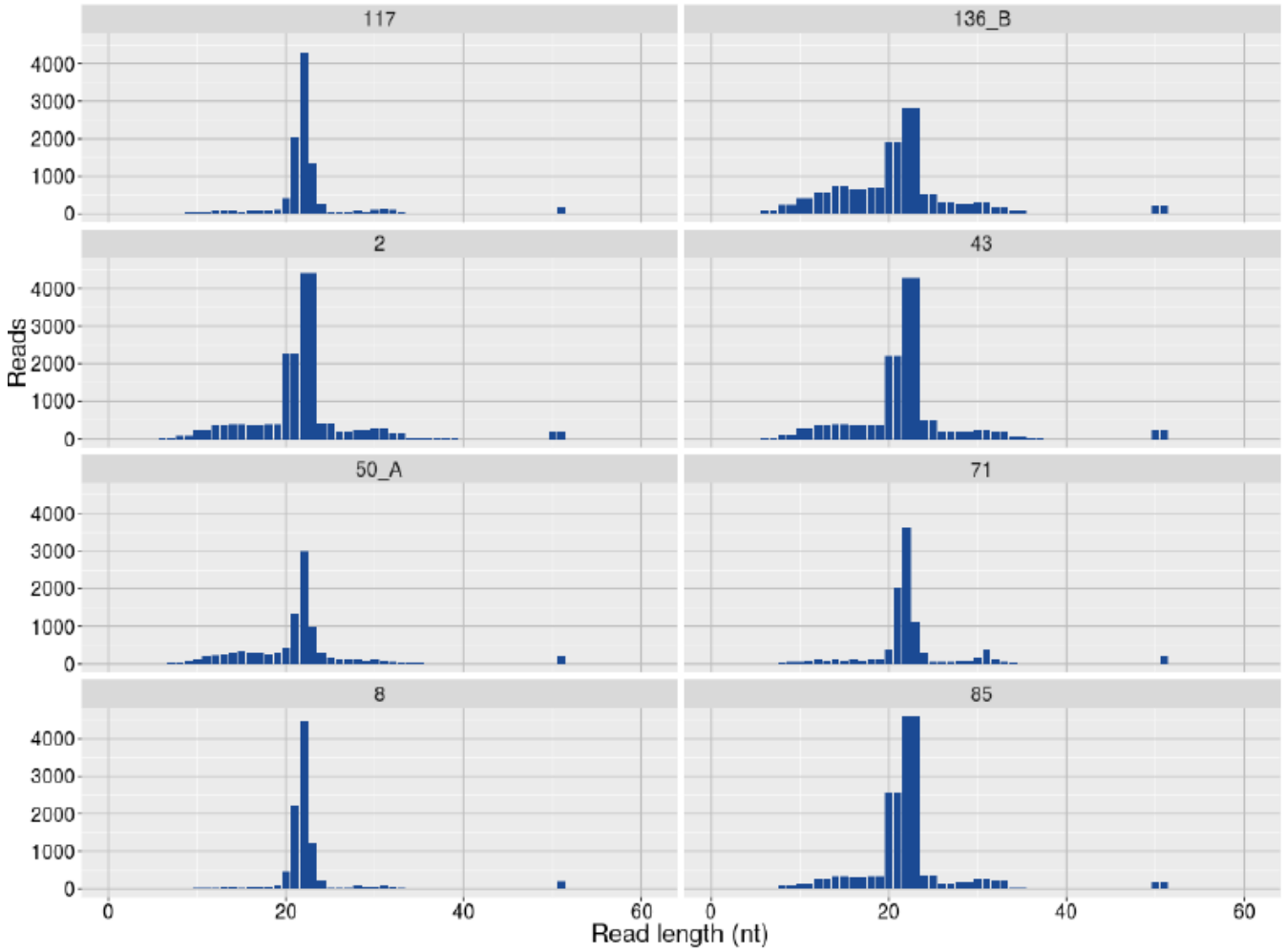


Figure 1: Read length and count information are shown as bar chart for the first 10 uterus samples.

We observe a peak at ~21/22 nucleotides, which corresponds to the presence of mature microRNAs in the data

Total number of reads

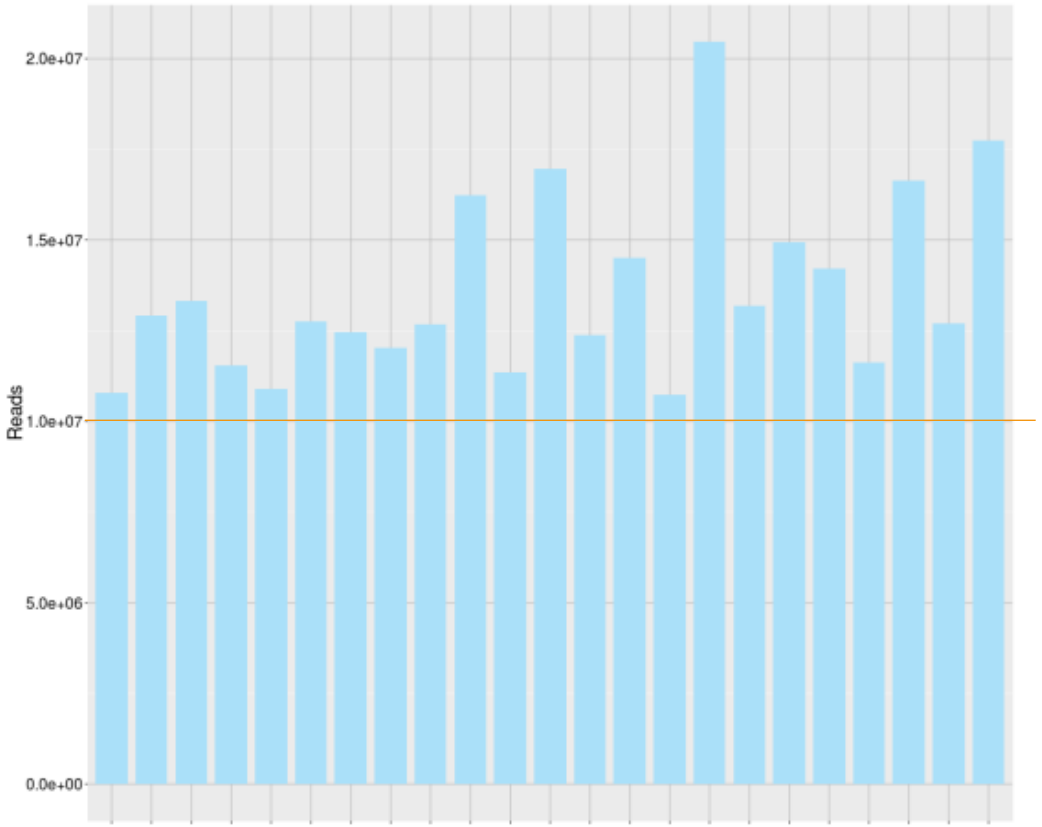


Figure 2: Total reads identified in the dataset

We aim for a total read count of at least 10 million reads per sample.

> 10 million reads were generated in all 24 samples

Read mappings

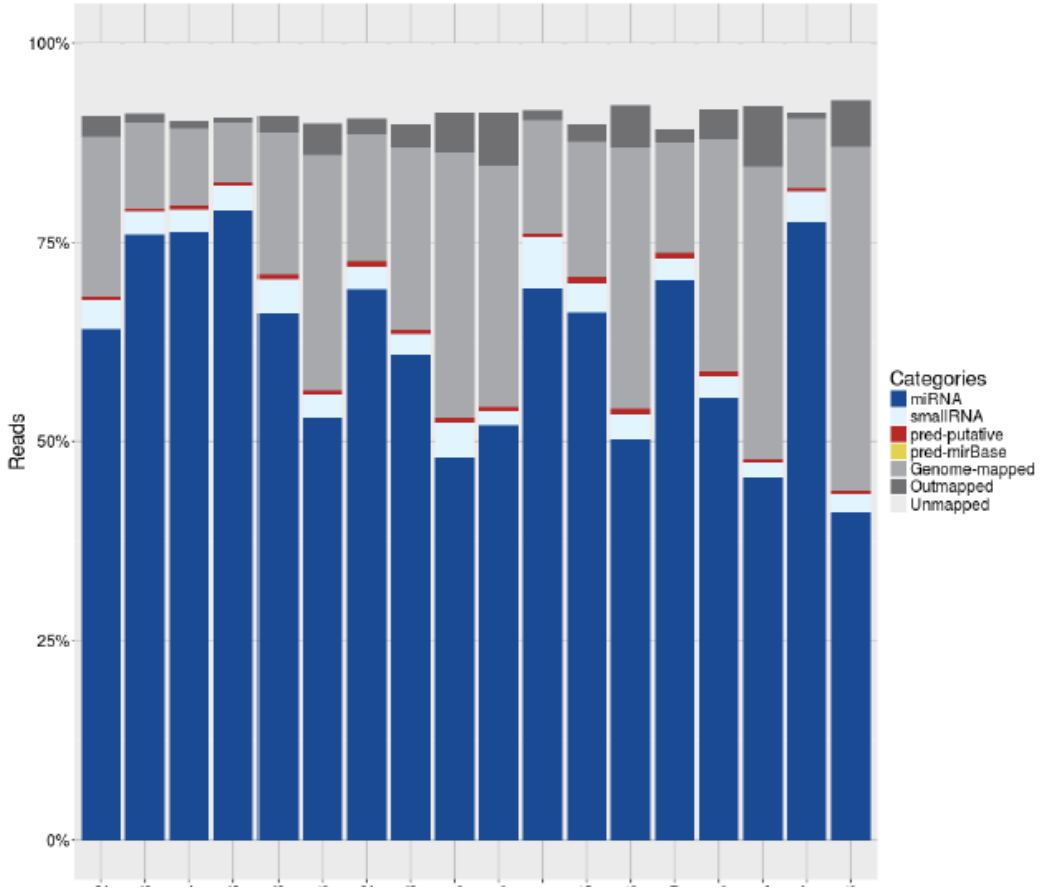


Figure 3: Read annotation

Sequencing reads are sequentially annotated as

- microRNAs, using miRBase as reference
- Putative novel microRNAs, using miRPara (pred-putative) algorithm
- Novel miRNAs with conserved orthologs in miRBase (pred-miRBase)
- Small RNAs, as annotated in Rfam
- Genome mapped reads
- Outmapped reads to repetitive elements in the genome
- Unmapped reads without alignment to any of the above mentioned references

microRNA read counts

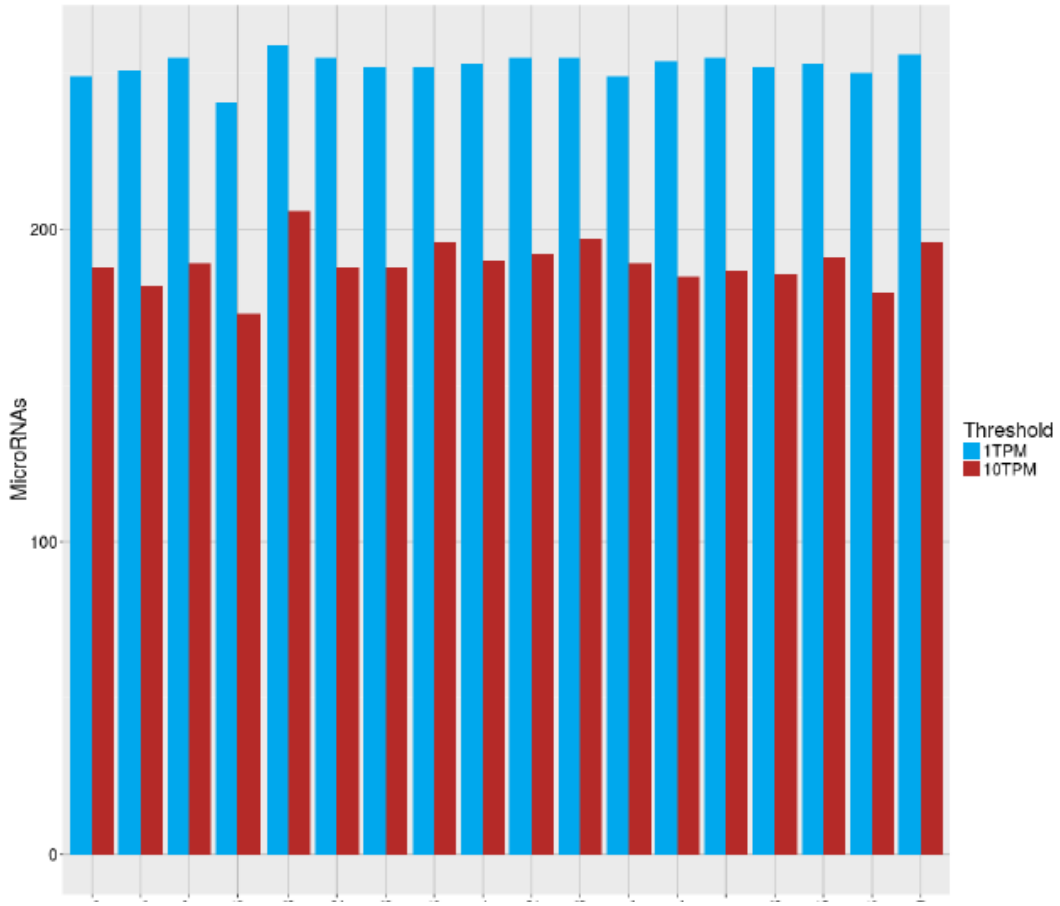


Figure 4:
Bar chart depicting the number of microRNAs identified per sample.

Expression levels are measured as Tags Per Million ("TPM").

TPM is a unit used to measure expression levels in NGS experiments and **can be used for any non-fragmenting library protocol.** The number of reads that map to a particular RNA species is divided by the total number of mapped reads in the sample and subsequently multiplied by one million. This is **a simple normalization procedure that corrects for the sequencing depth** and provides a **very transparent measure of quantity for each RNA species.**

Red: TPM > 10 used as cut-off
Blue: TPM > 1 used as cut-off

qPCR data quality

- We dedicate at least 1 page to various quality aspects of the analytical workflow and data analysis for qPCR experiments. In addition, raw data, figures, tables etc. are made available to the customer on a flash drive.

1. Total RNA recovery: spike-in controls are used to monitor constant recovery of total RNA from biofluids, small extracellular vesicles, cells and tissues
2. RNA purity: spike-in controls are used to monitor the presence of enzyme inhibitors
3. PCR efficiency and specificity: spike-in control, melting curve analysis and calculation of PCR efficiencies are performed

Outline

1. Experiment overview
2. Data quality
3. Exploratory data analysis
4. Differential gene expression analysis
5. Project summary & next steps

Exploratory expression analysis

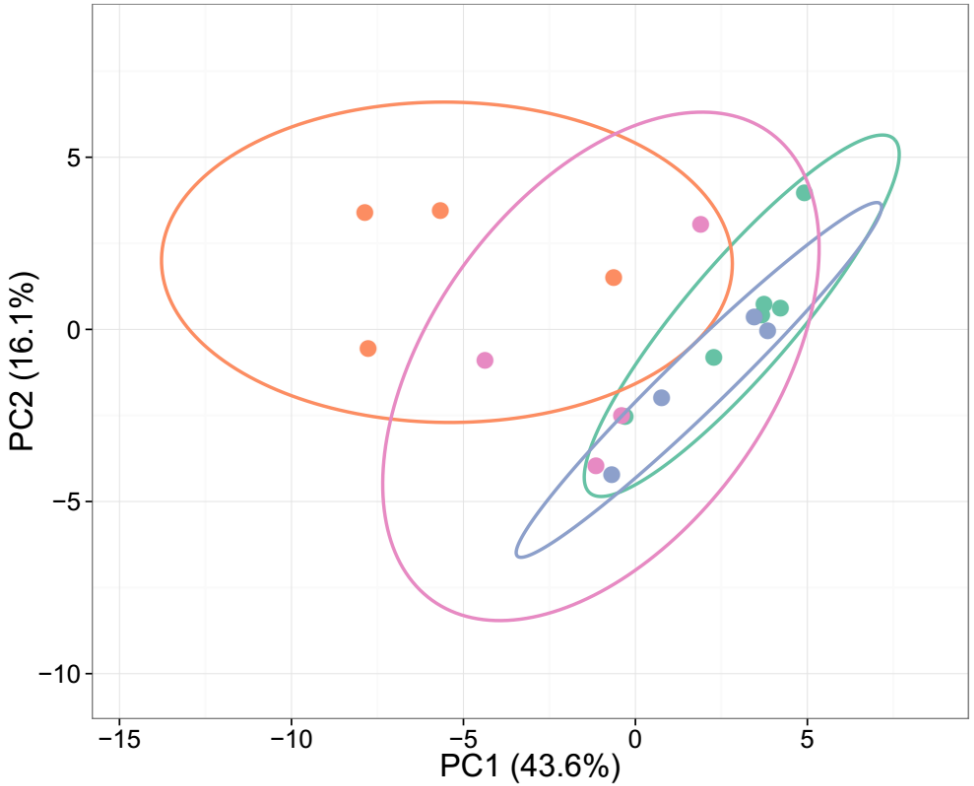
Aim: identify whether the experimental groups are the major determinant of microRNA expression in tissue, or whether there might be other confounding factors that impact microRNA expression.

- Exploratory analysis was performed using two datasets:
  - all microRNAs in the dataset with a TPM > 10
  - microRNAs in the dataset with a TPM > 10 and CV (%) > 50%

- **Principal component analysis (PCA):** PCA reduces the dimensionality of the dataset to few principal components, which explain a certain percentage of the variance in a dataset. Visualization was done as Scatter Plot with 95% confidence bands.

- **Hierarchical clustering:** Clustering of microRNAs and samples based on Pearson correlation coefficients. Visualization was done as heatmap with dendrograms.

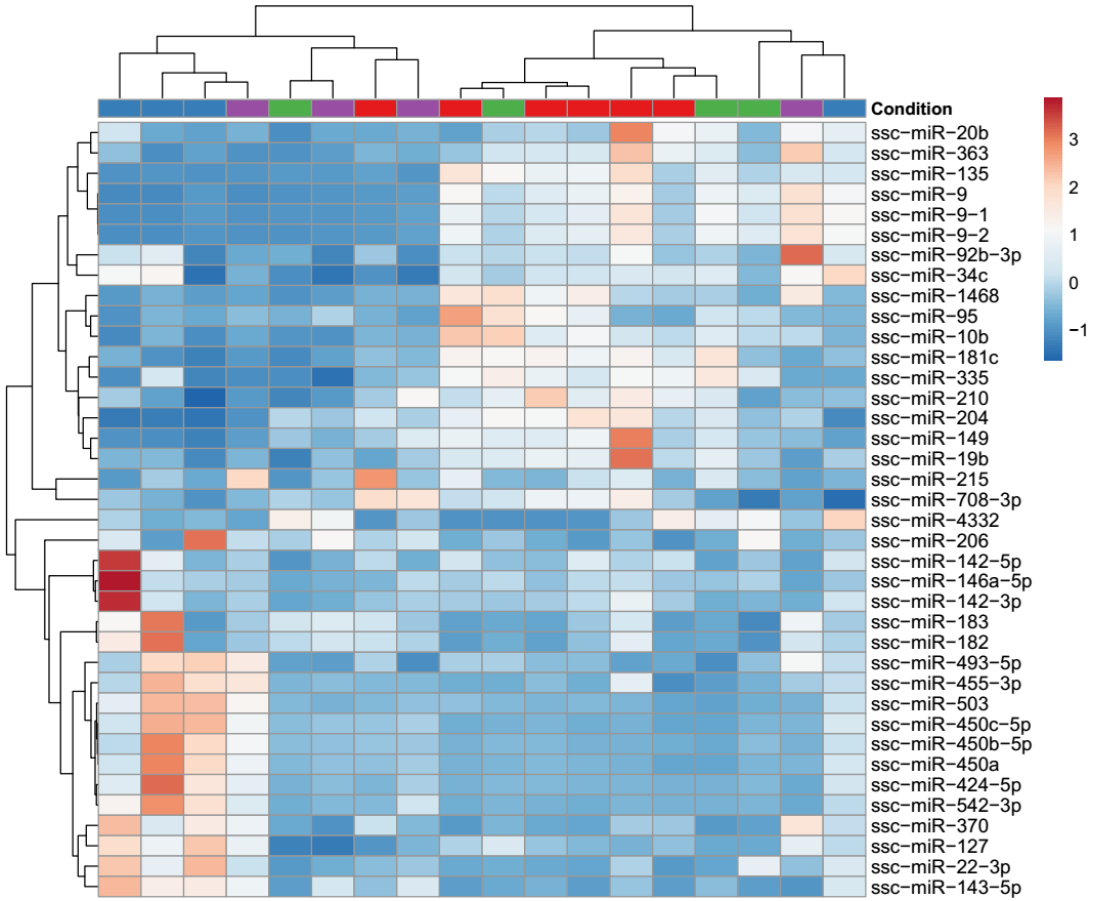
Principal component analysis *(example)*



Interpretation: PCA shows that experimental groups are a major determinant of microRNA expression in the experimental groups.

PC1 is relatively powerful, as it explains 43.6% of the variance in the dataset.

Hierarchical clustering *(example)*



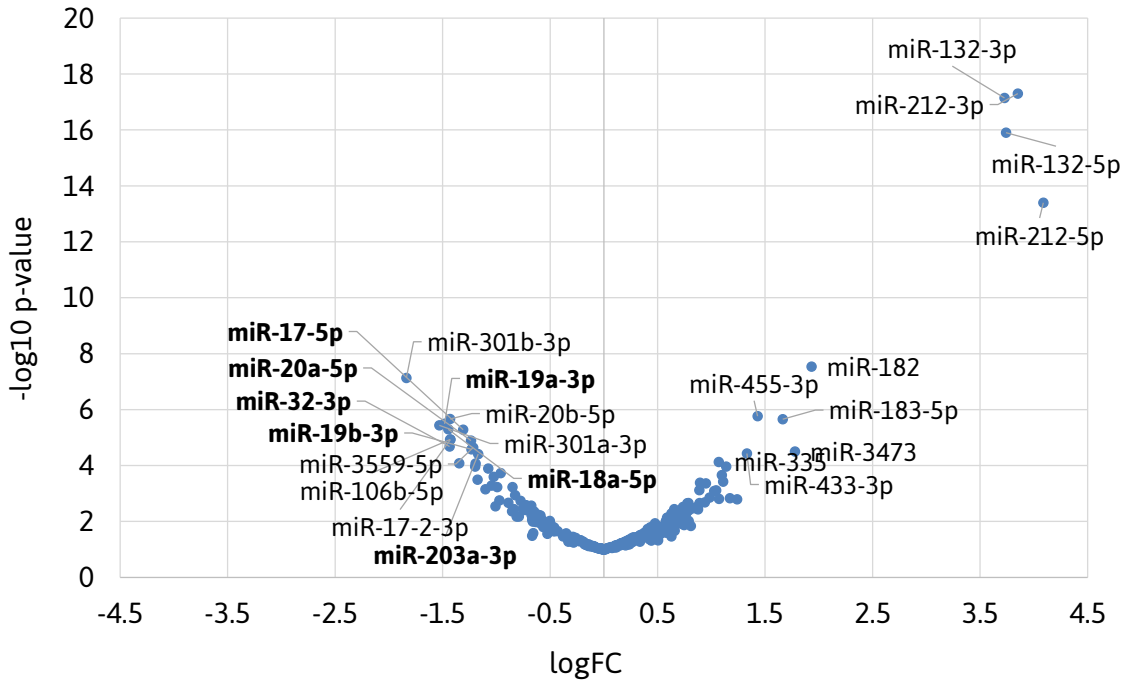
Interpretation:

Each row represents the expression of a specific microRNA in the analyzed samples. Red color indicates high expression level, blue color low expression level.

Dendrograms at the top indicate the similarity between samples.
Dendrograms on the right indicate the corrrelation between microRNAs in the dataset.

Color labels are added to highlight the assocation of a sample to a specific experimental group.

Volcano plot (*example*)



Interpretation:

For a given comparison (e.g. case vs control) a **volcano plot** depicts the **effect size** (log-transformed fold change, x-axis), and **significance level** (p-value, y-axis) for each microRNA.

This allows us to estimate the balance between up- and downregulation of microRNAs as well as the maximum effect sizes that have been identified.

The volcano plot is a useful graph to present promising biomarker candidates identified in a screening experiment.

Outline

1. Experiment overview
2. Data quality
3. Exploratory data analysis
4. Differential gene expression analysis
5. Project summary & next steps

Differential gene expression analysis

Aim: identify microRNAs that are significantly regulated between experimental conditions

Differential gene expression analysis is performed using the EdgeR package under R/Bioconductor.

Explanations:

- **The log2-transformed fold change (log2FC)** describes the (relative) fold-difference in microRNA levels between controls and treatment groups. Positive log2FC describes up-regulation, negative log2FC describes down-regulation.
- **The p-value** describes the significance level for the fold change.
- **The adjusted p-value** describes the significance level after adjusting for multiple testing using Benjamini-Hochbergs method for false-discovery rate (FDR).

Differential expression table *(example)*

names	logFC	logCPM	PValue	FDR
ssc-miR-424-5p	-3.3441352	5.84290997	3.1258E-22	7.0017E-20
ssc-miR-542-3p	-3.6562921	9.49612798	2.6436E-20	2.9609E-18
ssc-miR-450c-3p	-3.1808596	2.30640141	2.5244E-15	1.8849E-13
ssc-miR-424-3p	-2.9004307	3.7162399	1.59E-14	8.2638E-13
ssc-miR-450a	-2.8715333	8.42023823	1.8446E-14	8.2638E-13
ssc-miR-503	-2.5859899	5.67352305	2.5084E-14	9.3647E-13
ssc-miR-450c-5p	-3.1105376	9.65424262	5.2532E-14	1.681E-12
ssc-miR-450b-5p	-2.9053701	9.44494482	1.6519E-12	4.6254E-11
ssc-miR-204	2.03649868	9.7265404	2.2033E-12	5.4839E-11
ssc-miR-22-3p	-1.6929472	8.0032043	6.7473E-08	1.374E-06
ssc-miR-143-5p	-1.6285796	10.4666689	6.4521E-08	1.374E-06
ssc-miR-369	-1.501783	3.80244533	3.5644E-07	6.6536E-06
ssc-miR-140-3p	-1.4455388	13.1998334	4.875E-07	8.4E-06
ssc-miR-432-5p	-1.5270197	3.94184501	3.3749E-06	5.0399E-05
ssc-miR-758	-1.6679067	4.04826321	3.3041E-06	5.0399E-05
ssc-miR-127	-1.4753565	7.38508512	5.3461E-06	7.4845E-05
ssc-miR-149	1.49986915	7.97122433	6.1055E-06	8.0449E-05
ssc-miR-493-3p	-1.6276142	3.45618654	6.6406E-06	8.2638E-05
ssc-miR-125a	1.2838786	12.0422003	8.6125E-06	0.00010154
ssc-miR-136	-1.4499058	3.7205668	1.0005E-05	0.00011206

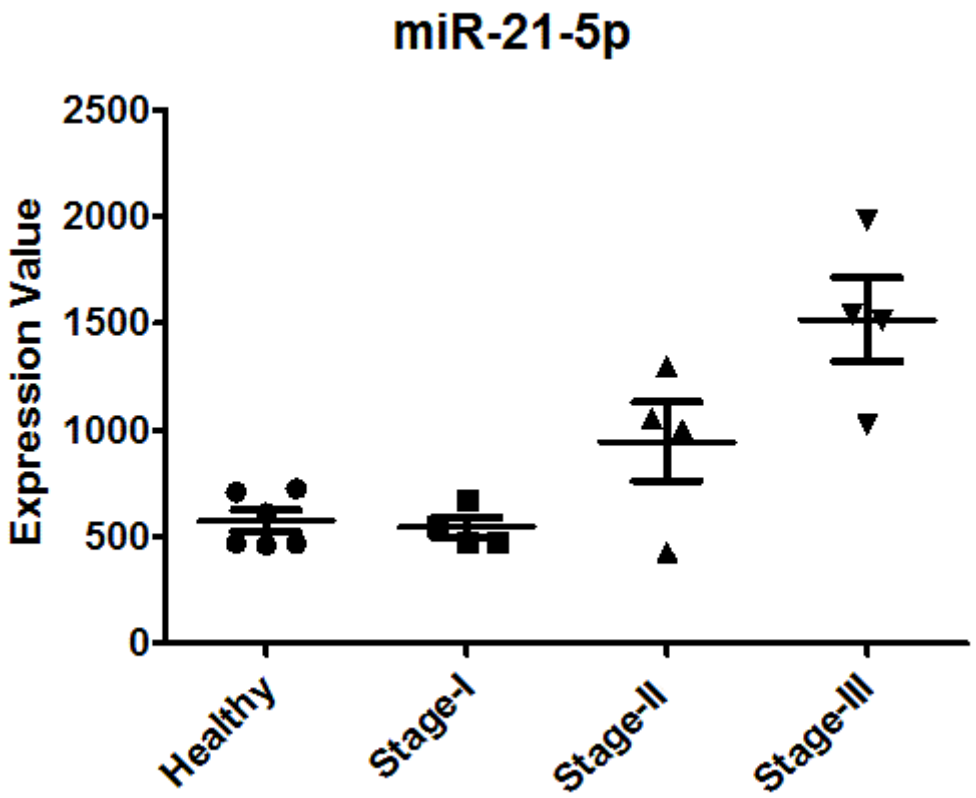
Table depicting the format in which the differential expression analysis results have been reported.

Down-regulation (logFC < 0) means down-regulation in control compared to treatment.

LogCPM is the log2-transformed read average read count for all samples in the analysis.

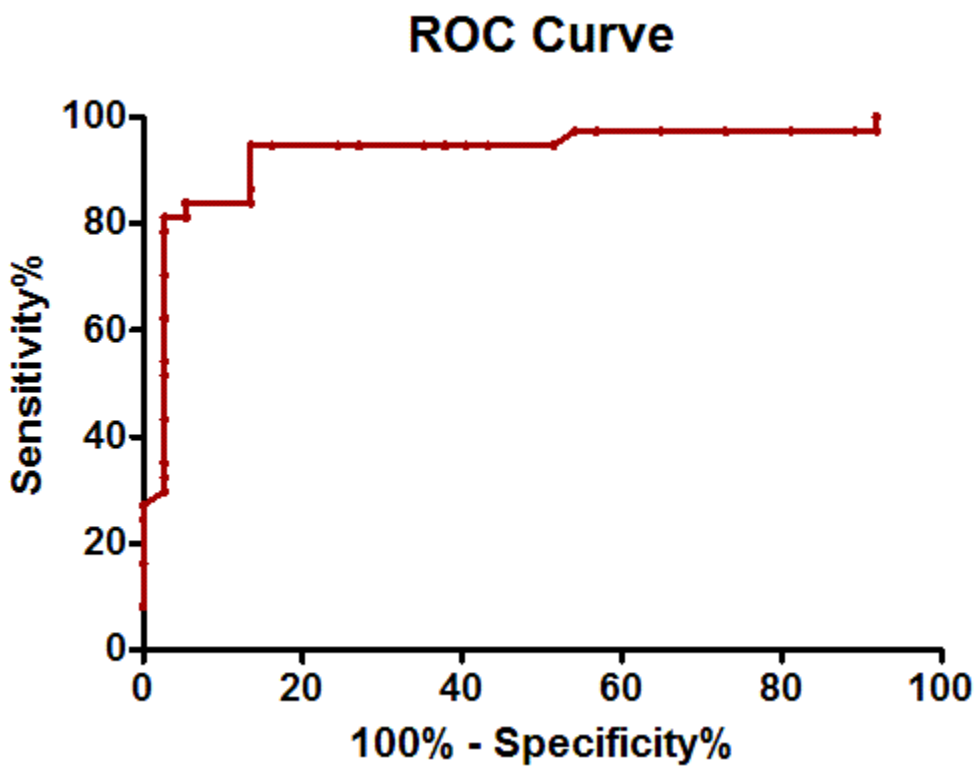
P-value and FDR describe the significance of the observed regulation.

Scatter plots (*example*)



Scatter plots or Boxplots will be generated to show the effect size of microRNA regulation as well as the intra-group variation in the experimental groups.

ROC analysis (*example*)



ROC analysis show the relation between sensitivity and specificity of a univariate or multivariate biomarker.

The area-under-the-curve (AUC) allows us to estimate the classification performance.

Depending on the type of disease/comparison,

- biomarkers with AUC values between 0.5-0.6 are considered to have low classification performance.
- biomarkers with AUC values between 0.6-0.8 are considered to have good classification performance.
- biomarkers with AUC values above 0.8 are considered to have very good classification performance.

Outline

1. Experiment overview
2. Data quality
3. Exploratory data analysis
4. Differential gene expression analysis
5. Project summary & next steps

Project summary & next steps

- Summary of our findings in terms of quality, exploratory analysis and differential expression analysis.

- Discussion of possible next steps to continue the project for example by validating an independent cohort.

Contact information

TAmiRNA GmbH
Muthgasse 18
1190 Vienna, Austria

www.tamirna.com
matthias.hackl@tamirna.com
+43 660 420 58 56