

Differential Expression Analysis Results Tutorial

Summary: This tutorial provides you with guidelines on how to interpret the results of the [differential expression \(DE\) analysis module](#) of Oasis. The output consists of quality metrics and several different plots, such as principle component analysis (PCA), heatmaps and MA plots, which are useful for exploratory analysis of your data and results. It further includes an interactive web report that allows you to see which DE micro RNAs (miRNAs) target which genes, and gives you the opportunity to test for enrichment of various gene ontologies (GO) and pathways.

In addition, given that this module allows for the correction of expression values by including covariate information such as age and gender, this tutorial also includes an example that shows the improvement of results when carrying out DE analysis with covariates.

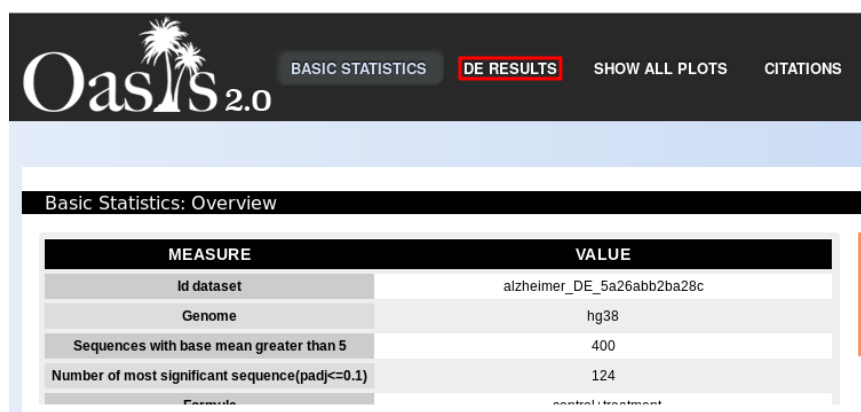
Most of the descriptions throughout this tutorial are based on the DE analysis output of the [Renal cancer demo dataset](#) (Osanto et al., 2012). For analysis with covariate information, the Alzheimer demo dataset (Leidinger et al., 2013) is used [with covariate correction](#) and [without covariate correction](#).

Overview

The first part of the output contains an overview table with quality metrics (Fig. 1). This table reports the number of sRNAs with at least 5 mean reads for all biological conditions, the number of DE sRNAs at the 10% level of significance and the formula used to run the DE analysis.

Since not all sRNAs are effectively expressed for particular biological conditions, Oasis filters out such sRNAs to improve on the DE analysis. Therefore, Oasis only keeps sRNAs with an average of 5 reads for either condition.

Including the number of DE sRNAs at the lenient threshold of 10% will give you an idea of the overall proportion of sRNAs whose expression changes between biological conditions. So, for example, in Fig. , out of 400 sequences with support of at least 5 reads, 124 are DE at a 10% significance level, resulting in ~31% DE sRNAs, which is reasonable.



| MEASURE | VALUE |
|---|----------------------------|
| Id dataset | alzheimer_DE_5a26abb2ba28c |
| Genome | hg38 |
| Sequences with base mean greater than 5 | 400 |
| Number of most significant sequence (padj<=0.1) | 124 |
| Formula | control:treatment |

Figure 1: Overview section in the output of the DE analysis module

Basic Statistics Plot

In addition to the Overview table, the output includes a principal component analysis (PCA) plot, heatmaps, MA plots and p-value distribution, which offer different visualisations of the sample similarities and the DE results.

Principal Components Analysis (PCA) plot: (Figure 2) this plot shows how well samples cluster together based on the similarity of their sRNA expression. Unlike the PCA in the [sRNA Output Tutorial](#), this PCA distinguishes between samples based on their tested condition to see whether they are spatially clustered. In general, outliers and samples clustering in the wrong group will be an indication of few DE sRNAs, so removing such samples should improve on the DE analysis and generate more DE sRNAs. In the example renal data, we can see the samples cluster reasonably well based on their condition, indicating an expectancy of quite a few DE sRNAs (Fig. 2). For more information on PCA plots, please refer to our [sRNA Output Tutorial](#).

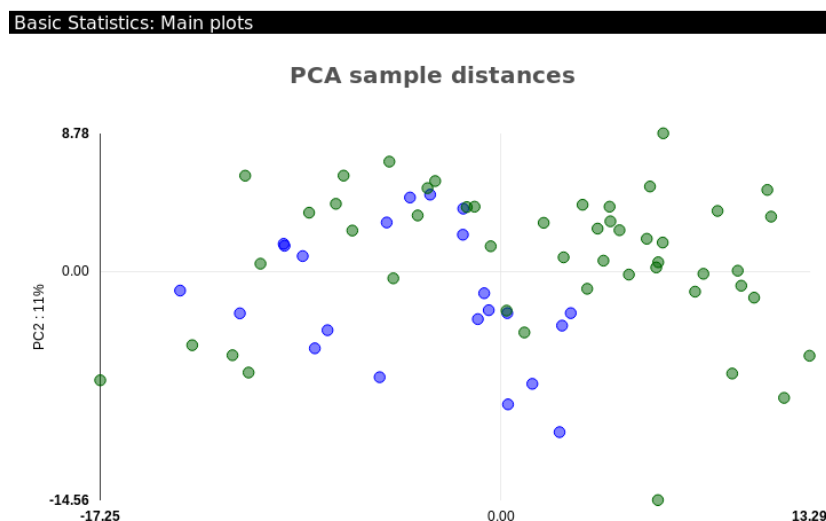


Figure 2: PCA plot of alzheimer data.

Heatmap: (Figure 3) this plot depicts the similarities between samples, using a predefined colour convention. The first heatmap in the DE output corresponds to the matrix of Euclidean distances computed between all samples, where the distances are a function of the regularised-log (rlog) transformed counts of each sample. The heatmap contains an auxiliary "Color Key Histogram" on the top left to indicate the distances between the samples (red indicates distance 0 between same samples) and dendrograms on the x and y axes show how samples cluster together. The renal data shows the treatment samples cluster well with each other, and the control samples occasionally cluster with the treatment samples, which should result in a reasonable amount of DE sRNAs (Fig. 3).

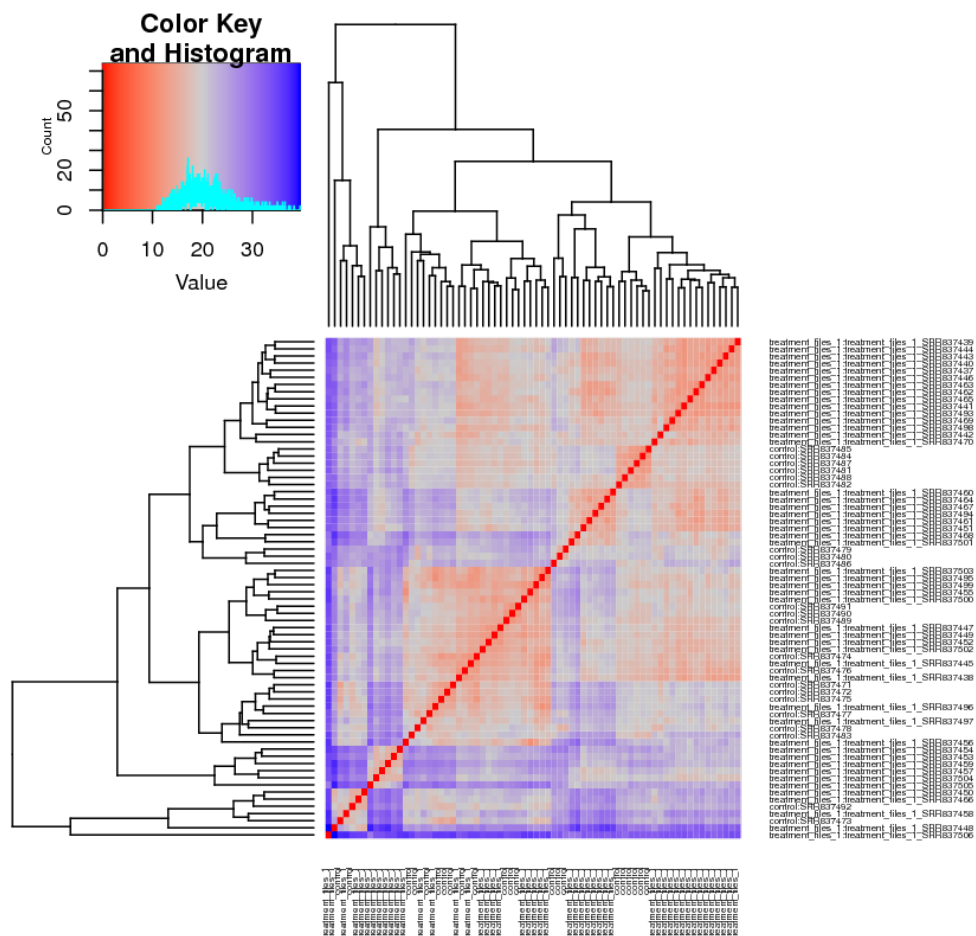


Figure 3: Heatmap of Alzheimer's disease data

MA plots: MA plots are a particular instantiation of the method proposed in (Altman & Bland, 1983), which consists of graphically representing the agreement or disagreement of applying two experimental conditions according to some variable of interest. The method consists of plotting the recorded difference in the value of the variable as a function of the average value of such variable when applying the experimental conditions. In the particular case of DE analysis, the variable of interest is the expression level transformed by \log_2 . As such, the dependent variable is the \log_2 fold change of the expression, and the independent variable is the mean expression. More information on the interpretation of this plot can be found in MA plots in the DESeq2 manual (Love, Huber, & Anders, 2014). In the output show expression changes, attributable to applying the treatment condition as a function of average expression. The assumption is that in a well-carried experiment, the expression of most sRNAs between control and treatment conditions should remain constant, while only several of them would be considerably different. For example, see figures 4 and 5.

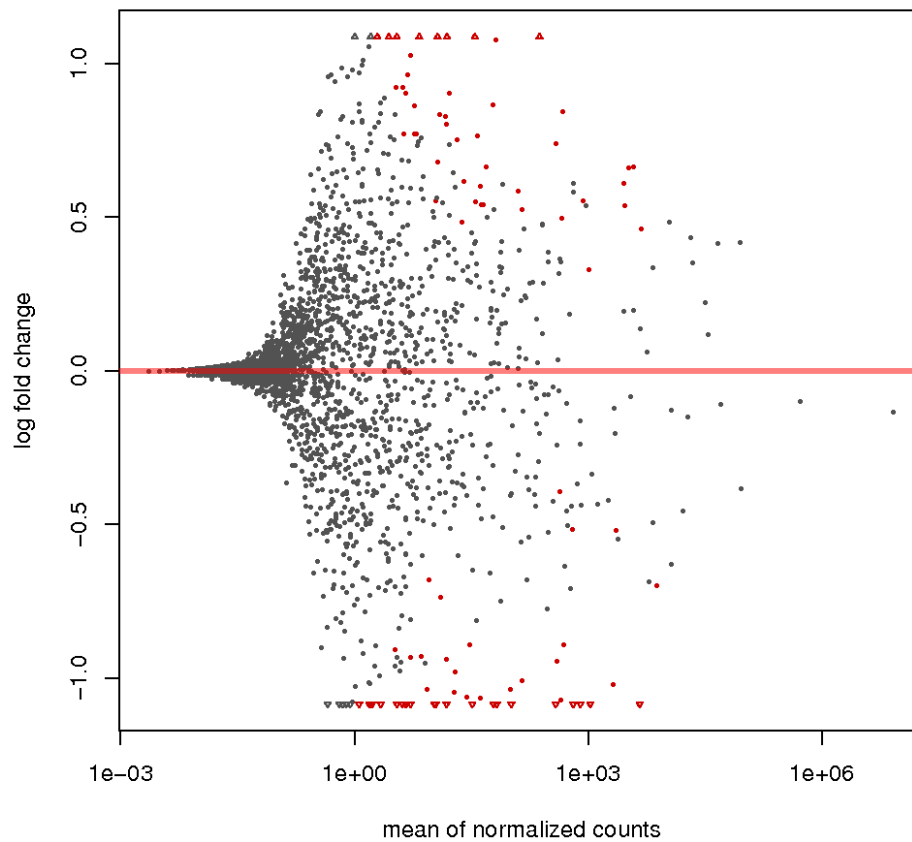


Figure 4: MA plot of log₂-fold changes of the renal data (without filtering). Red points correspond to sRNAs whose differential expression values have an adjusted p-value ≤ 0.1 . Triangles correspond to points falling out of the plot area. The default MA plot produced by DESeq2 and is obtained from raw count expression values of all sRNAs. For the alzheimer data, a considerable portion of points lie around the $y=0$ line, indicating that a considerable amount of sRNAs have 'constant' expression across experiments.

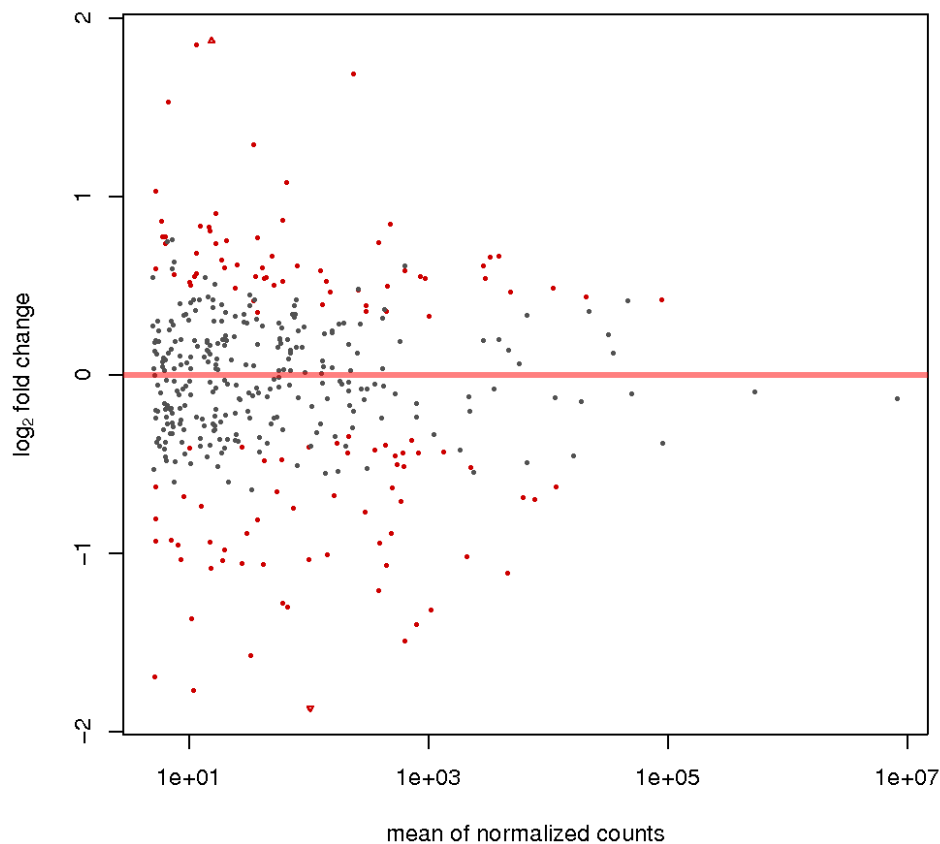


Figure 5: MA plot (after filtering). The standard MA plot output by DESeq2 (Fig. 4) might be too coarse and difficult for interpretation, so Oasis generates an additional MA plot after applying pre-processing. The pre-processing consists of (1) filtering out sRNAs with less than 5 reads on average per condition and (2) applying Benjamini-Hochberg correction to p-values of the remaining sRNAs. Applying this pre-processing to the renal data results in fewer points and a more scattered pattern.

PValue distribution: in order to see the connection between the significance of sRNAs among different conditions and their expressions, the distribution of p-values is plotted for sRNAs that did or did not pass the filtering criteria (5 reads on average per condition). This plot is useful to see how many sRNA have a statistically significant p-value, but are not supported by enough evidence (read coverage) to be considered as such. In the case of the renal data, while many sRNAs do not seem to pass the filter criteria, it is mostly in the higher p-values, while lower p-values are more associated with sRNAs that pass the filtering criteria. Therefore, the expected correlation between low p-values and high sRNA coverage is clear in the renal data. The p-value distribution is depicted in figure 6.

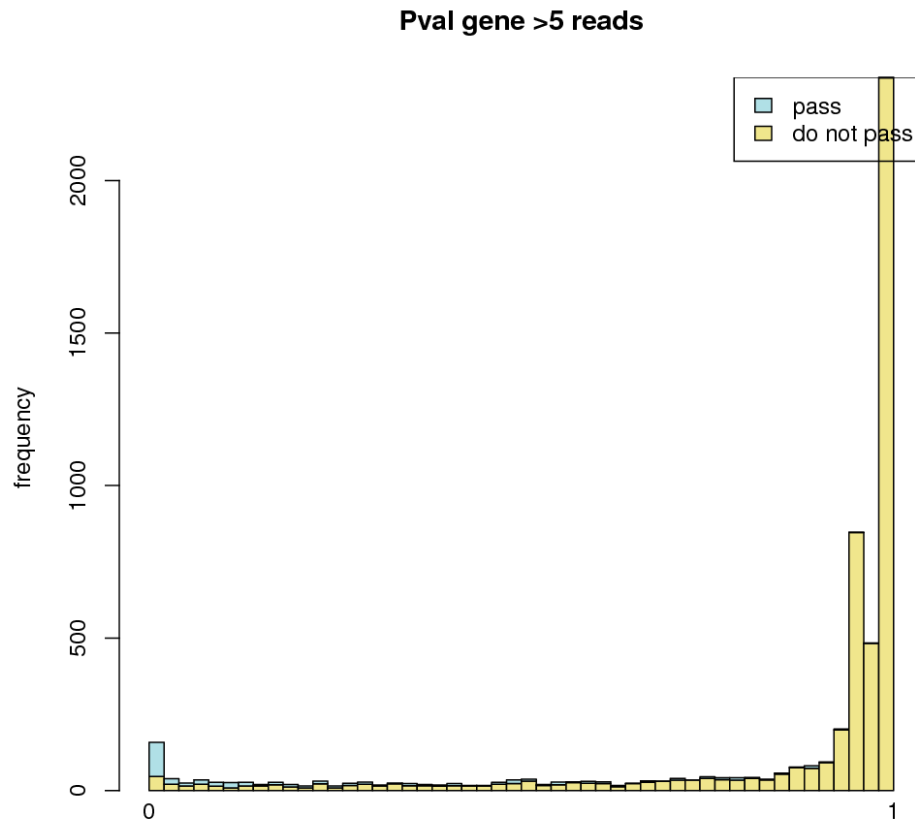


Figure 6: Histogram of p-values associated to the reported sRNAs. Yellow bins represent the distribution of p-values whose corresponding sRNAs passed Oasis' filtering stage and in blue, atop of the yellow bins, the distribution of p-values that did not pass the filtering stage.

Top 30 most expressed sequences: these 2 heatmaps show the top 30 most expressed sRNAs from raw counts and rlog transformed counts. For the renal cancer demo data, a single sRNA with massive raw counts (red square) overshadows all other sRNAs with somewhat lower expression (blue squares) (Fig. 7), while the rlog transformed counts for the sRNAs are spread across a reasonable scale (smooth flow from low expressions in blue to high expressions in red; Fig. 8). This suggests that the rlog transformation normalises the data in such a way that the change in expression is gradual, not drastic like in the raw counts.

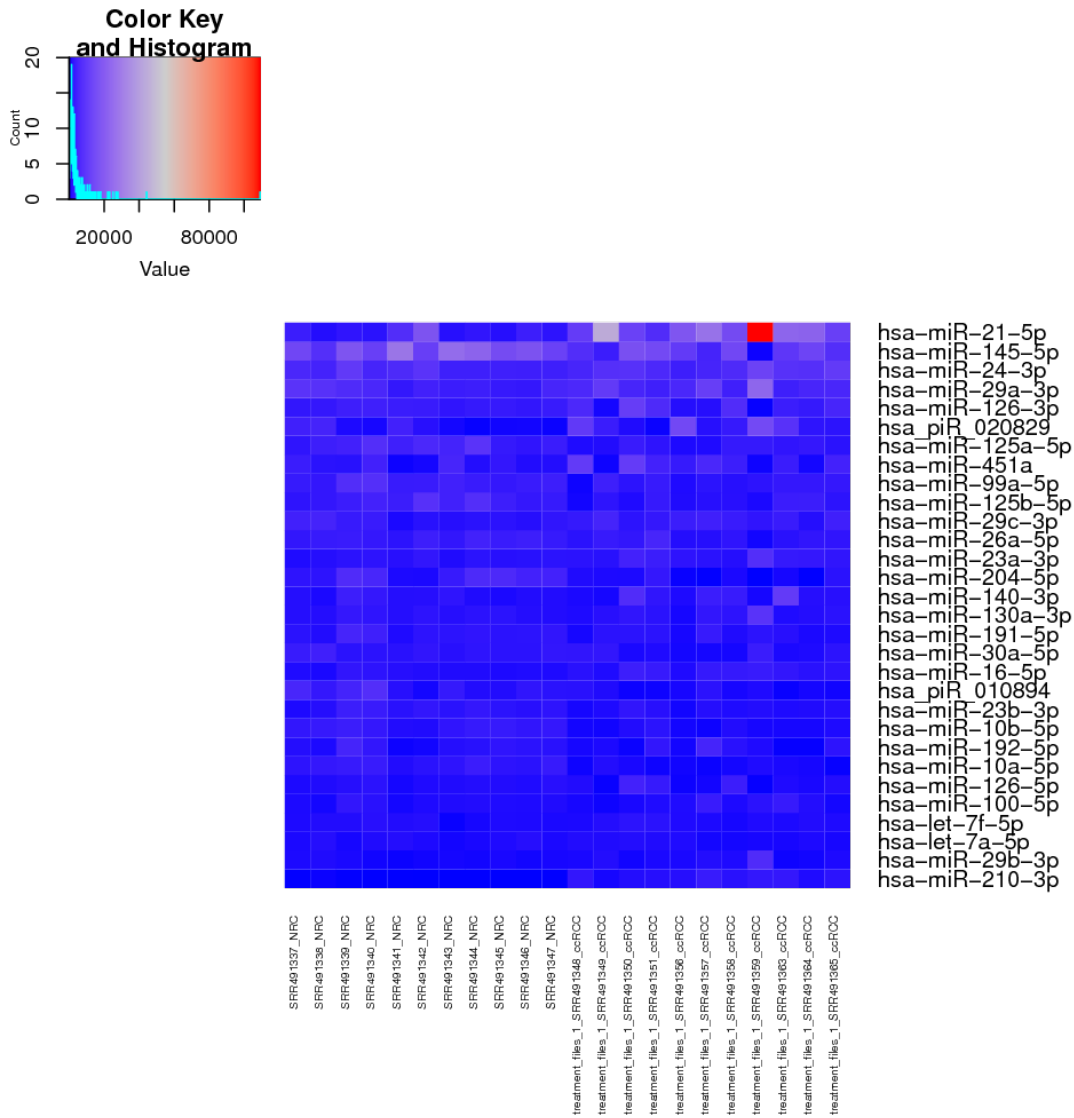


Figure 7: Top 30 most expressed sRNAs (raw counts)

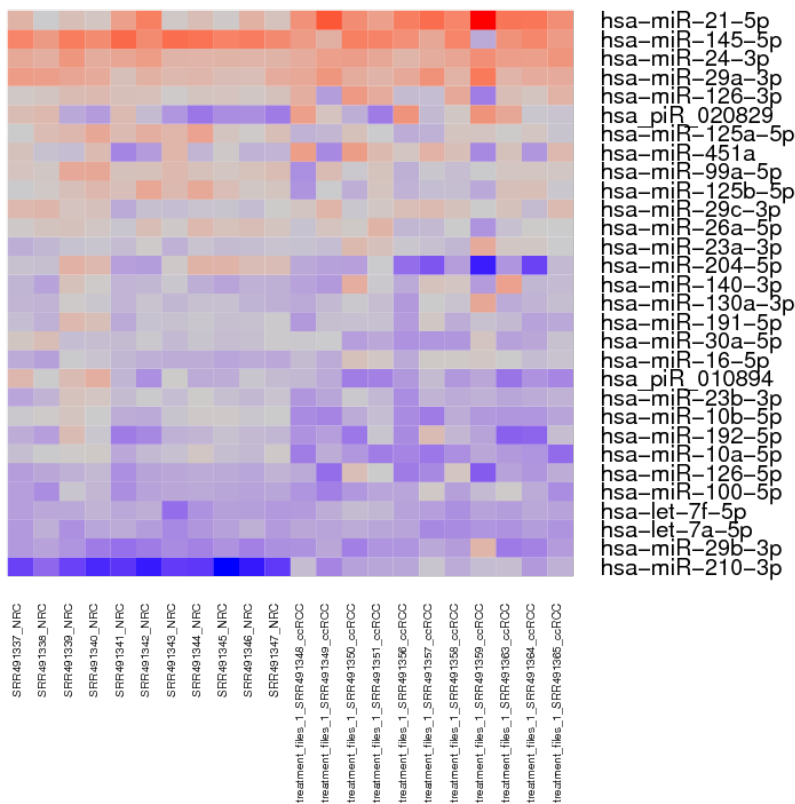
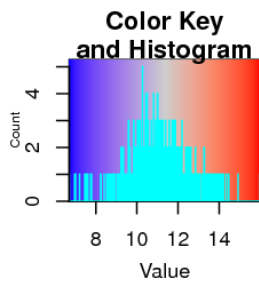


Figure 8: Top 30 most expressed sRNAs (rlog transformed counts)

All significant sequences: this heatmap contains the top 30 most significant sRNAs with adjusted p-values below 0.1, as well as the sample clustering based on the expression of those sRNAs alone. For the renal data, while the samples had an incomplete clustering to “control” and “treatment” groups when plotting a heatmap for all sRNAs (Fig. 3), NRC samples (representing control samples) still show mixed clustering with ccRCC samples when using a subset of top 30 most significant sRNAs (representing treatment samples) (Fig. 9).

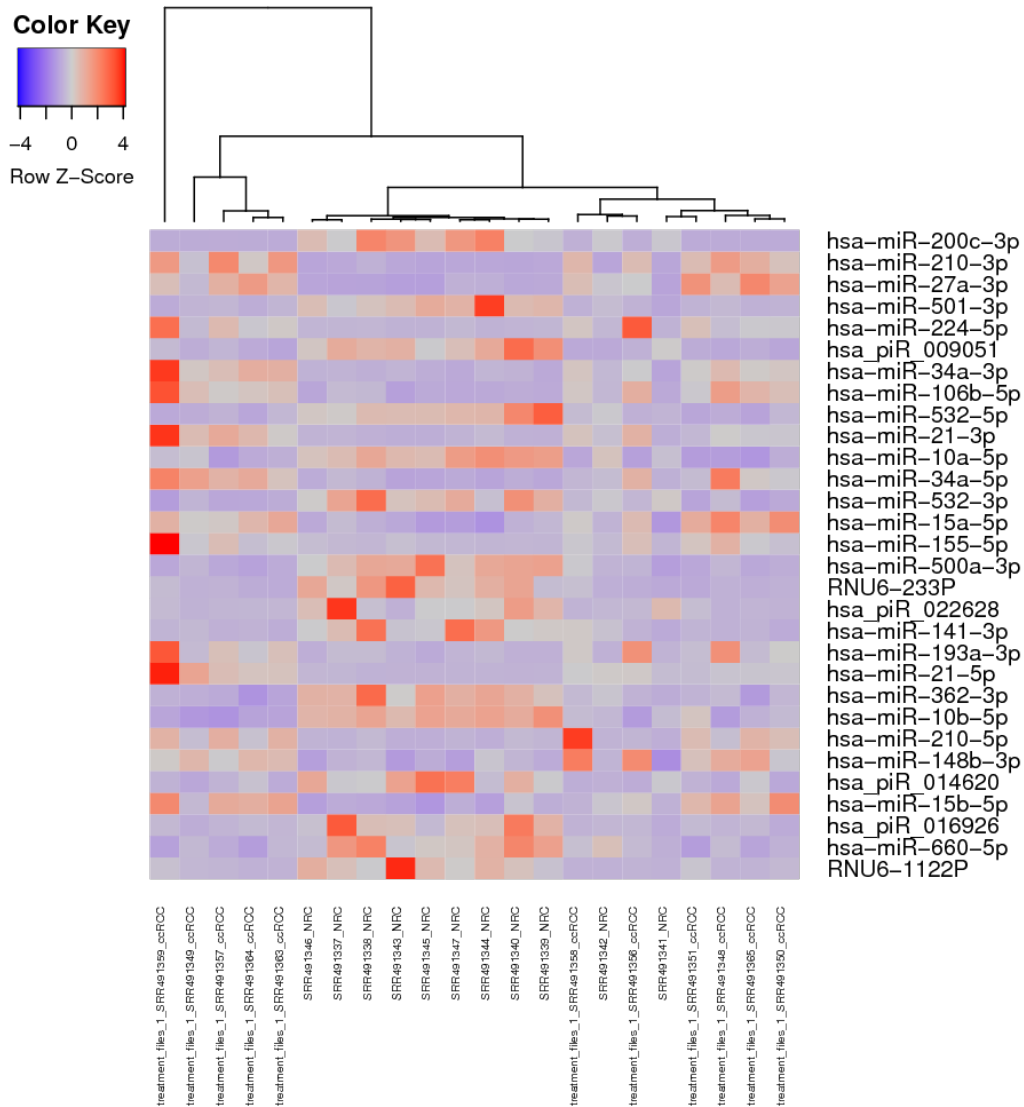


Figure 9: Heatmap with the top 30 most significant sRNAs (p-value <= 0.1)

Interactive web report

Apart from the statistical plots, the Oasis DE module returns an interactive web table that contains detailed information on each sRNA, and allows for the functional enrichment analysis of target sRNAs. To access this information, press the option "DE results" at the top of the HTML page (Fig. 10, red box). A new web page will open (Fig. 10), containing two functional parts: functional analysis options (Fig. 10, blue box) and a table of all DE analysis results for the sRNAs (Fig. 10, red box).

Oasis_{2.0}
INTERACTIVE RESULT TABLE

click here to open tutorial

1) Select by adj-pval :

2) Targets type :

3) Select the Enrichment Analysis : (1000 targets maximum)
 gProfiler (GO, KEGG, etc) Genemania (Interactome, GO) DAVID (KEGG, GO, etc) STRING (interactome) STITCH (interactome)

Enrichment analysis

Display Targets

Clear selection

Open Total Results in excel Open Filtered Results in excel

| Select | mature | structure | position | strand | sequence | baseMean | baseMean control | log2FC | adjusted p-value | Validated Targets (miRTarBase and miRecords) | Predicted targets (miRanda) |
|--------------------------|-----------------|-----------|--------------------------|--------|---------------------|----------|------------------|--------|------------------|--|-----------------------------|
| <input type="checkbox"/> | hsa-miR-200c-3p | miRNA | chr12:6963742-6963764 | + | UAAUACUGCCGGUAAUGA | 115.53 | 225.33 | -5.19 | 2.28e-32 | 52 | 224 |
| <input type="checkbox"/> | hsa-miR-210-3p | miRNA | chr11:568112-568133 | - | CUGUGCUGUGACAGCGG | 1134.13 | 129.94 | 3.88 | 1.12e-30 | 54 | 1 |
| <input type="checkbox"/> | hsa-miR-27a-3p | miRNA | chr19:13836447-13836467 | - | UUCACAGUGGCUAAGUUC | 612.24 | 274.55 | 1.76 | 6.34e-12 | 104 | 1 |
| <input type="checkbox"/> | hsa-miR-501-3p | miRNA | chrX:50009772-50009793 | + | AAUGCACCCGGCAAGGAL | 34.76 | 57.93 | -2.11 | 1.12e-11 | 6 | 0 |
| <input type="checkbox"/> | hsa-miR-224-5p | miRNA | chrX:151958631-151958651 | - | CAAGUCACUAGUGGUUC | 74.74 | 7.51 | 3.65 | 2.12e-11 | 17 | 1 |
| <input type="checkbox"/> | hsa-miR-34a-3p | miRNA | chr1:9151693-9151714 | - | CAAUCAGCAAGUUAUCUGC | 51.86 | 15.34 | 2.37 | 1.30e-10 | 0 | 0 |
| <input type="checkbox"/> | hsa-miR-106b-5p | miRNA | chr7:100094043-100094063 | - | UAAAGUCUGACAGUGCAC | 402.02 | 180.74 | 1.73 | 1.73e-10 | 215 | 1 |
| <input type="checkbox"/> | hsa-miR-532-5p | miRNA | chrX:50003167-50003188 | + | CAUGCCUUGAGUGAGGAC | 226.07 | 353.46 | -1.84 | 2.31e-10 | 0 | 0 |
| <input type="checkbox"/> | hsa-miR-21-3p | miRNA | chr17:59841311-59841331 | + | CAACACCAGUCGUAUGGCU | 52.91 | 9.09 | 3.17 | 5.64e-10 | 2 | 1 |
| <input type="checkbox"/> | hsa-miR-10a-5p | miRNA | chr17:48579904-48579926 | - | UACCCUGUAGAUCCGAAUL | 1724.19 | 2680.67 | -1.77 | 7.75e-10 | 297 | 62 |
| <input type="checkbox"/> | hsa-miR-34a-5p | miRNA | chr1:9151735-9151756 | - | UGGCAGUGUCUAGCUGG | 819.74 | 227.83 | 2.51 | 9.70e-10 | 498 | 1 |
| <input type="checkbox"/> | hsa-miR-532-3p | miRNA | chrX:50003204-50003225 | + | CCUCCACACCCAAGGCU | 87.05 | 139.47 | -2.03 | 1.71e-09 | 20 | 0 |
| <input type="checkbox"/> | hsa-miR-15a-5p | miRNA | chr13:50049167-50049188 | - | UAGCAGCACAUAAUGGUUL | 120.20 | 60.05 | 1.40 | 5.92e-09 | 126 | 1 |
| <input type="checkbox"/> | hsa-miR-155-5p | miRNA | chr21:25573983-25574005 | + | UUAUUGCUAUUGUGUAUC | 49.28 | 6.66 | 3.37 | 8.26e-09 | 728 | 1 |
| <input type="checkbox"/> | hsa-miR-500a-3p | miRNA | chrX:50008482-50008503 | + | AUGCACUUGGGCAAGGAUL | 10.94 | 17.90 | -2.24 | 8.36e-09 | 14 | 0 |
| <input type="checkbox"/> | hsa-miR-141-3p | miRNA | chr12:6964155-6964176 | + | UAACACUGUCUGUAAAGA | 15.01 | 25.32 | -2.57 | 2.38e-07 | 37 | 225 |
| <input type="checkbox"/> | hsa-miR-193a-3p | miRNA | chr17:31560050-31560071 | + | AACUGCCUACAAGUCCC | 98.99 | 31.07 | 2.31 | 2.59e-07 | 3 | 1 |
| <input type="checkbox"/> | hsa-miR-21-5p | miRNA | chr17:59841273-59841294 | + | UAGCUUAUCAGACUGAUGL | 15830.19 | 4323.93 | 2.51 | 3.90e-07 | 492 | 1 |
| <input type="checkbox"/> | hsa-miR-362-3p | miRNA | chrX:50009005-50009026 | + | AACACACCUAUUCAAGGAU | 53.56 | 75.20 | -1.29 | 4.23e-07 | 3 | 0 |
| <input type="checkbox"/> | hsa-miR-10b-5p | miRNA | chr2:176150329-176150351 | + | UACCCUGUAGAACCGAAUL | 1885.79 | 2774.76 | -1.45 | 5.77e-07 | 124 | 113 |
| <input type="checkbox"/> | hsa-miR-210-5p | miRNA | chr11:568150-568171 | - | AGCCCCUGCCCACCGCAC | 20.91 | 2.99 | 2.87 | 6.05e-07 | 0 | 0 |

Figure 10: Interactive web report

Within the table, columns “mature”, “structure”, “position” and “strand” indicate the name of the reported sRNA, the sRNA species it belongs to (for elaboration on sRNA species, see the sRNA Output Tutorial) and its position and strand, respectively. The columns “baseMean” and “baseMean control” represent the mean expression values for each condition. In the case of multiple treatment groups, just the “baseMean control” column is included. “log2FC” and “adjusted p-value” follow suit with the log2 fold-change in expression between conditions and the adjusted p-value, respectively. Finally, “Validated Targets” and “Predicted Targets” indicate how many validated and predicted gene targets are linked to the sRNA. Clicking on one of the target entries will show a list of the gene targets as gene names.

Enrichment analysis

The enrichment analysis of Oasis allows the user to select specific miRNAs (either manually or by filtering based on adjusted p-value threshold), and use the associated genes to find which gene ontology (GO) categories, KEGG pathways or protein-protein interactions are enriched for them.

1. Select the miRNAs for functional enrichment manually or by using a particular adjusted p-value threshold (10%, 5%, 1% or 0.1%).

1) Select by adj-pval :

2. Select the type of miRNA targets you would miRecords like to Only analyse. validated The “Targets targets Predicted type” selector targets box gives you three options:(miRTarBase Both. (Hsu et al., 2014) and (Xiao et al., 2009)); (miRanda (Betel, Wilson, Gabow, Marks, & Sander, 2008)), or While there might be less validated targets for a given miRNA, we recommend using only validated targets, as target predictions have a notoriously high false-positive rate.

2) Targets type :

3. Select the enrichment analysis tool(s) you may want to use. At the moment, you are free to choose from the following web-services: gProfiler (Reimand, Arak, & Vilo, 2011), Genemania (Warde-Farley et al., 2010), DAVID (Huang, Lempicki, & Sherman, 2009), STRING (Snel, Lehmann, Bork, & Huynen, 2000), and STITCH (Kuhn et al., 2014). A brief description of each of these tools can be found at the end of this document. Check the boxes to select the enrichment analysis you would like to perform.

3) Select the Enrichment Analysis : (1000 targets maximum)

gProfiler (GO, KEGG, etc) Genemania (Interactome, GO) DAVID (KEGG, GO, etc) STRING (Interactome) STITCH (Interactome)

Finally, submit your enrichment analysis by clicking the Enrichment analysis button (Fig. 10, green box). You may submit jobs to multiple enrichment tools at once, but consider that a new browser window will open for each tool you use, and it will take longer to run. If the links do not open, try disabling your popup blocker.

Notes:

1. There is a limit of 1000 miRNA targets allowed in an automatic submission, so if you need to use more than 1000 targets for your enrichment analysis, we recommend you copy the list of targets and directly use it on the software of your preference.
2. In order to reset miRNA selections, click on the button "Clear Selection" (Fig. 10, yellow box).
3. The table can be sorted by any of its columns by clicking on the column header.
4. A click on the sRNA ID will redirect you to a detailed annotation on mirbase (miRNA), UCSC genome browser tracks (novel miRNA) or genecards (Rappaport et al., 2014) (other sRNAs).
5. In case you want to analyse the results manually, you can download Excel tables of all analysed sRNAs (by clicking on the link [Open Total Results in excel](#)) or of sRNAs that passed the filtering stage (by clicking the link [Open programs Filtered Results in excel](#)).

Enrichment programs provided

When having a set of DE genes associated with particular miRNAs, the most direct way to derive some biological context from them is to run an enrichment analysis. This analysis finds particular biological terms or pathways which contain have a strong representation of DE genes within them. There are various tools to do such test, and here are several which are available through Oasis:

1. *gProfiler*: returns enriched GO categories, KEGG and REACTOME pathways, TRANSFAC regulatory motifs, miRBase miRNAs, CORUM protein complexes, Human Phenotype Ontologies and BioGRID protein-protein interactions.
2. *Genemania*: returns a protein-protein network, showing the protein products of the selected gene targets and how they associate with each other, as well as significantly enriched GO categories.
3. *STITCH* and *STRING*: returns a single image of a protein-protein network. On top of having protein-protein interactions, STITCH includes small molecules, drugs and ATPs-associated with the target proteins as well.
4. *DAVID*: Runs an enrichment test for all target genes, using various functional annotations (GO categories, KEGG pathways, BIOCARTA pathways, protein domains and so on).

Analysis with covariate information

For covariate-based DE analysis, we have a demo Alzheimer's dataset (Leidinger et al., 2013). First of all, it is important to mention that this dataset contains 3 outlier samples (SRR837453, SRR837503 and SRR837506) (see [sRNA detection results for the Alzheimer's dataset](#)). As such, those samples are removed from the DE analysis, resulting in 22 control and 45 Alzheimer's samples. Second of all, using the covariate table as mentioned in the Oasis DE tutorial, the analysis can be run with the covariate information or without it. This means the analysis is run once with the samples only, and once with the covariate table and the formula $-Gender+Age+DiseasePheno$ (for more details, see the [Oasis DE tutorial](#)). When comparing both results, they seem fairly similar, where the analysis with covariates has 179 sRNAs after coverage filtering (minimum 5 reads on average) and the other analysis has 178 sRNAs after coverage filtering. In addition, the analysis without covariates has 70 DE sRNAs, whereas the other analysis has 67 DE sRNAs. The reason 3 DE sRNAs without the covariate information are not DE with the covariate information is because those sRNAs are DE due to the variations of gender or age between the different samples, but the distinction between the disease phenotype (AD or C) is actually non-significant when the gender and age variations are corrected for. Therefore, the covariate information allows removing false positive DE sRNAs, which are only DE due to factors not relating to the actual tested conditions.

References

- Altman, D. G., & Bland, J. M. (1983). Measurement in Medicine : the Analysis of Method Comparison Studies. *The Statistician*, 32(July 1981), 307-317. <http://doi.org/10.2307/2987937>
- Betel, D., Wilson, M., Gabow, A., Marks, D. S., & Sander, C. (2008). The microRNA.org resource: Targets and expression. *Nucleic Acids Research*, 36(SUPPL. 1). <http://doi.org/10.1093/nar/gkm995>
- Hsu, S. Da, Tseng, Y. T., Shrestha, S., Lin, Y. L., Khaleel, A., Chou, C. H., ... Huang, H. Da. (2014). MiRTarBase update 2014: An information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research*, 42(D1). <http://doi.org/10.1093/nar/gkt1266>
- Huang, D. W., Lempicki, R. a, & Sherman, B. T. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), 44-57. <http://doi.org/10.1038/nprot.2008.211>
- Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T. H., Von Mering, C., Jensen, L. J., & Bork, P. (2014). STITCH 4: Integration of protein-chemical interactions with user data. *Nucleic Acids Research*, 42(D1). <http://doi.org/10.1093/nar/gkt1207>
- Leidinger, P., Backes, C., Deutscher, S., Schmitt, K., Mueller, S. C., Frese, K., ... Keller, A. (2013). A blood based 12-miRNA signature of Alzheimer disease patients. *Genome Biology*, 14(7), R78. <http://doi.org/10.1186/gb-2013-14-7-r78>
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. <https://doi.org/10.1186/s13059-014-0550-8>
- Osanto, S., Qin, Y., Buermans, H. P., Berkers, J., Lerut, E., Goeman, J. J., & van Poppel, H. (2012). Genome-wide microRNA expression analysis of clear cell renal cell carcinoma by next generation deep sequencing. *PLoS ONE*, 7(6). <http://doi.org/10.1371/journal.pone.0038298>
- Rappaport, N., Twik, M., Nativ, N., Stelzer, G., Bahir, I., Stein, T. I., ... Lancet, D. (2014). MalaCards: A Comprehensive automatically-mined Database of human diseases. *Current Protocols in Bioinformatics*, 2014, 1.24.1-1.24.19. <http://doi.org/10.1002/0471250953.bi0124s47>
- Reimand, J., Arak, T., & Vilo, J. (2011). G:Profiler - A web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, 39(SUPPL. 2). <http://doi.org/10.1093/nar/gkr378>
- Snel, B., Lehmann, G., Bork, P., & Huynen, M. a. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Research*, 28(18), 3442-3444. <http://doi.org/10.1093/nar/28.18.3442>
- Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., ... Morris, Q. (2010). The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Research*, 38(SUPPL. 2). <http://doi.org/10.1093/nar/gkq537>
- Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., & Li, T. (2009). miRecords: An integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37(SUPPL. 1). <http://doi.org/10.1093/nar/gkn851>