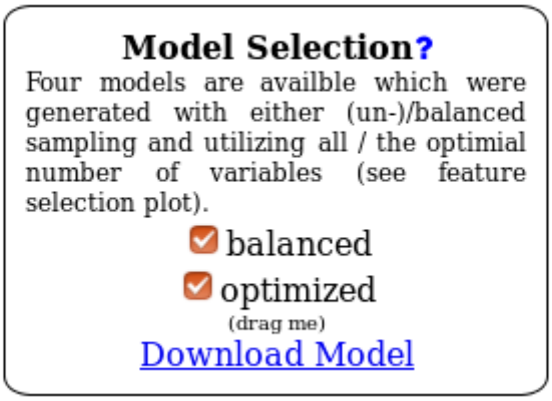
sRNA-based disease classification tutorial

Summary: The oasis classification module distinguishes between 4 cases which are defined by two dimensions: Balanced versus unbalanced model and optimized versus non-optimized model. The classification generates four Random Forest models with regard to these 4 cases. In the HTML view you can choose any combination of configurations and the whole web page will update accordingly.

As a recap, the Classification module performs a binary classification on the count files input as control or treatment by applying a Random Forest classifier. The output includes different types of plots: data exploration, classifier performance and feature importance plots, along with an interactive web report of known and predicted micro RNAs (miRNAs) that allows you to subsequently perform functional analyses. The analysis example throughout this tutorial is the classification output of the Psoriasis demo dataset (Joyce et al., 2011). This tutorial aims at helping the user interpret the results of the Classification module. Further information on how to submit data to the Oasis Classification module can be found on the classification tutorial page.

Overview

The most important element of the results is the **Model Selection** displayed in figure 1. There you can decide which model configuration you want to investigate more closely. Furthermore, you can use the `Download model` link to get the underlying R model which you then can use in your own R programs.



Model Selection?

Four models are availble which were generated with either (un-)/balanced sampling and utilizing all / the optimial number of variables (see feature selection plot).

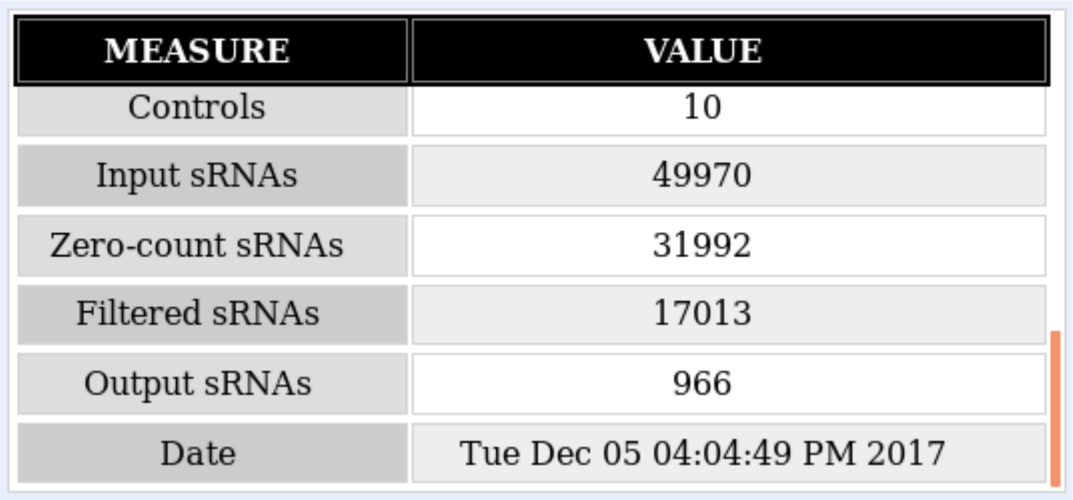
☑ balanced

☑ optimized

(drag me)

Download Model

Figure 1: Model selection dialog. Use the checkboxes to update the numbers and plots on the entire page to a specific model configuration. You can also download the R model for use in your own R programs.



MEASURE	VALUE
Controls	10
Input sRNAs	49970
Zero-count sRNAs	31992
Filtered sRNAs	17013
Output sRNAs	966
Date	Tue Dec 05 04:04:49 PM 2017

Figure 2: Overview of the classification job parameters.

Data exploration plots

In addition to the overview table the Classification module returns principle component analysis (PCA) plot (Fig. 3) and Outlier plot (Fig. 4) for the initial exploratory analysis of your data.

PCA plots are useful to understand whether control and treatment groups cluster separately, but are also good for understanding if other noticeable features in the data are present. The psoriasis data, for example, clusters into two biological conditions, where control samples appear in red and treatment samples appear in blue (Fig. 3). The distinction between the samples based on their group is poor, which can have many reasons, most notably technical problems or problems with the biological specimen.

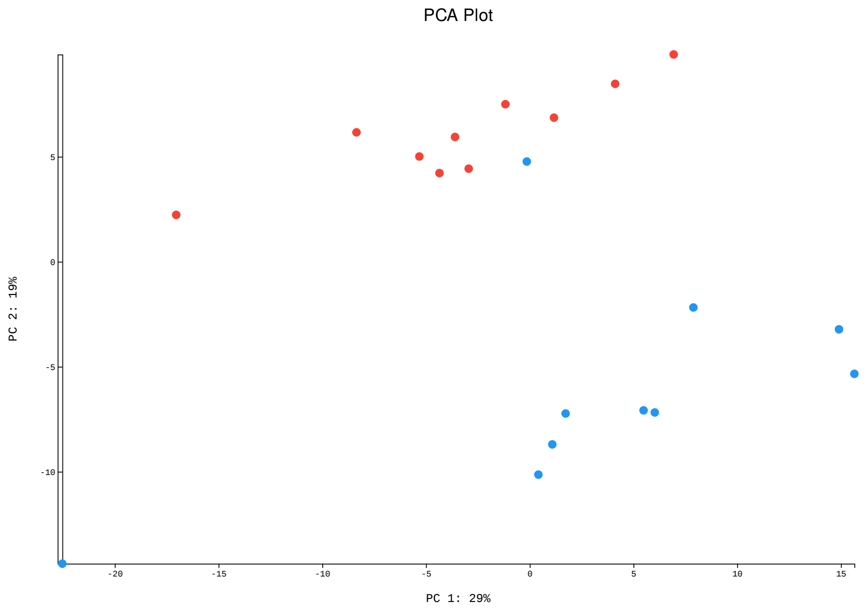


Figure 3: Principal components analysis (PCA) plot. The first two principal components are shown which represent the largest and second-largest variance in the dataset.

Technically, PCA plots show the two principal components of the samples used to train the classifier. The x- and y-axes indicate the level of variance "explained" by each principal component.

A complementary form of understanding how data is organized is given by the Outlier plot (Fig. 4). While the PCA shows a multitude of misclassified samples, here samples 14 and 20 clearly show suspiciously high scores. Technically, the Outlier plot shows the modified Z-scores on the proximities reported by the random forest. In random forests, proximities are a measure of how similar samples are to each other (Breiman, 2001), so they can be used as input for methods designed to detect outliers. The modified Z-score has been proposed as a measure for outlier detection and is a function of the sample's median absolute deviation (Iglewicz & Hoaglin, 1993).

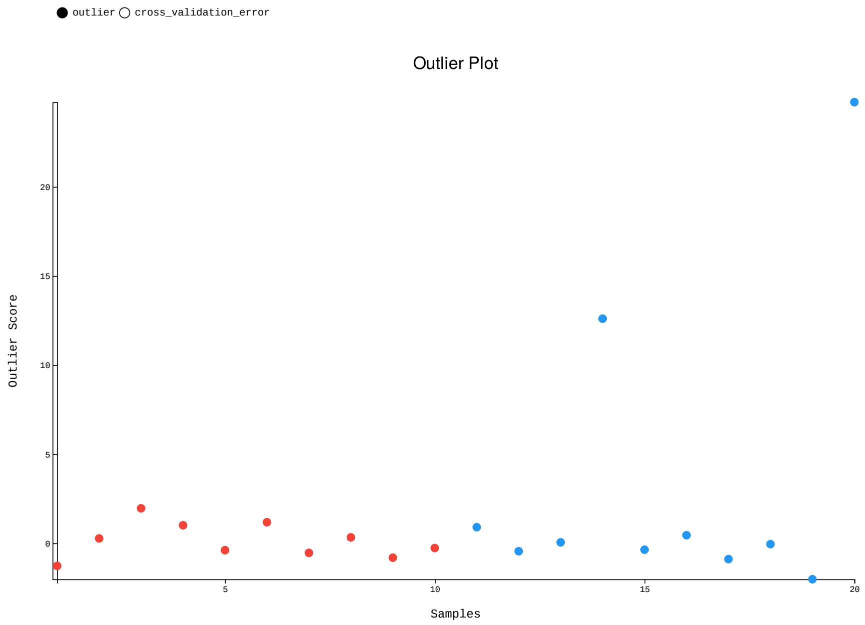


Figure 4:

Please note that for ease of visualization, the outlier plot in Oasis shows your samples with colors associated to the class they belong to: red for controls and blue for treatments. However, do not confuse this plot with a depiction of how your samples separate.

Some notes about outlier detection:

1. **How about to detect outlier** as suggested by (Breiman, 2001), modified Z- score values greater than 10 could be potential outliers. In the example of Fig. 4, sample 20 is a strong outlier candidate whereas sample 14 is a borderline case.

Performance plots

While the PCA and Outlier plots should be used to assess if your samples cluster into their biological conditions, ROC (receiver operating characteristic)

(Fig. 5) and precision-recall plots (Fig. 6) inform you about the actual classification performance.

A ROC curve shows the true-positive rate as a function of the false-positive rate. In a nutshell, the TPR (true-positive rate) is the rate of predicted true positive (treatment) samples over all positive (treatment) samples. The FPR (false-positive rate) is the rate of predicted false positives (control) samples over all negative (control) samples. An ideal classifier would have an ROC curve covering the whole "ROC space" and reach the upper-left corner of the space, corresponding to the point (0,1). A classifier that reaches such point would correctly predict all positive observations without incorrectly labeling any negative observation. In simple terms, the closer the red line to the upper left corner of the plot, the better is your classifier performance. The closer the red line to a diagonal line from the bottom left to the upper right corner of the plot, the worse is your classification performance (a diagonal line would signify random classification, i.e. the samples are randomly assigned as treatment or control, and not based on some sRNA measurements).

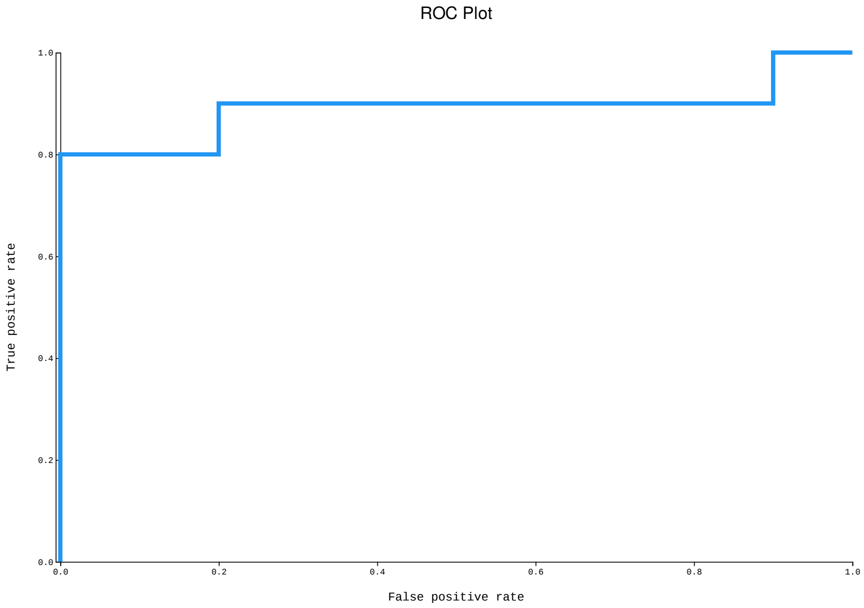


Figure 5: Receiver operating characteristic (ROC) plot. The area under the curve (AUC) is an important measure of classifier importance, because it encapsulates in a single number the ability of the classifier to separate the data. While the AUC measures classifier importance, the OOB error is an estimate of the expected classification error you will get if you classify a completely new sample.

The ROC space is normalized and covers an area of 100%, so one very simplified way to represent classifier performance is to use the area under the ROC, or AUC, reported in the Overview section and on top of the ROC plot. As a rule of thumb, AUC values above 0.9 signify good classification performance.

Possibly a more intuitive way to graphically 5). assess your classification performance is the precision-recall plot (Fig. Recall is equivalent to the TPR (see above), indicating how many of all treatment samples classify correctly. The precision is the rate of true positive (treatment) samples over the sum of all classified samples (true-positive and false-positive). Given a good classification, a precision-recall plot will pair high precision with high recall.

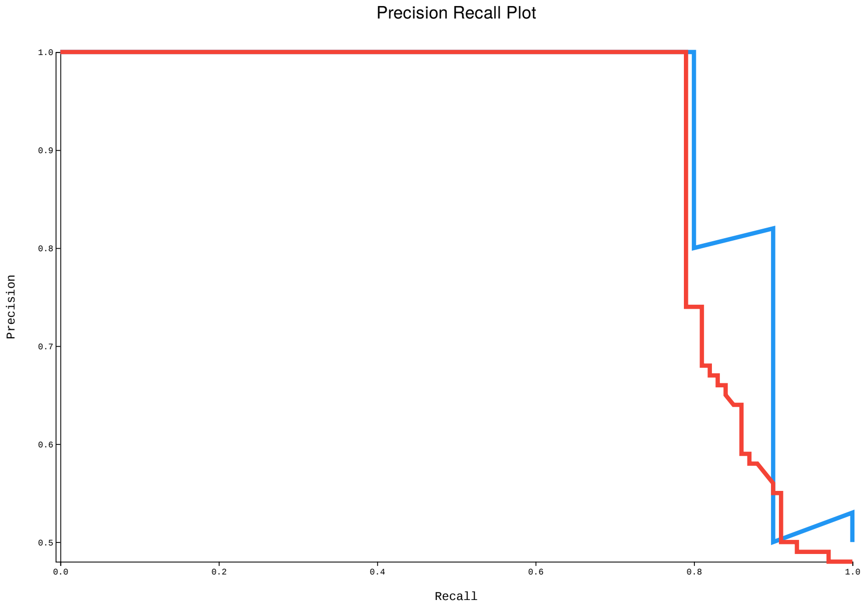


Figure 6:

More information on ROC plots can be found at Fawcett, 2003.

Feature importance plot

Biomarker identification is about selecting the most important set of sRNAs that might be useful to make a classification. Oasis uses the gini index as a measure of variable importance with higher indexes indicating higher importance of the sRNA. Oasis' variable importance plot reports the gini indices for the 10 most important sRNAs found by the random forest classifier, in order of decreasing importance (Fig. 7).

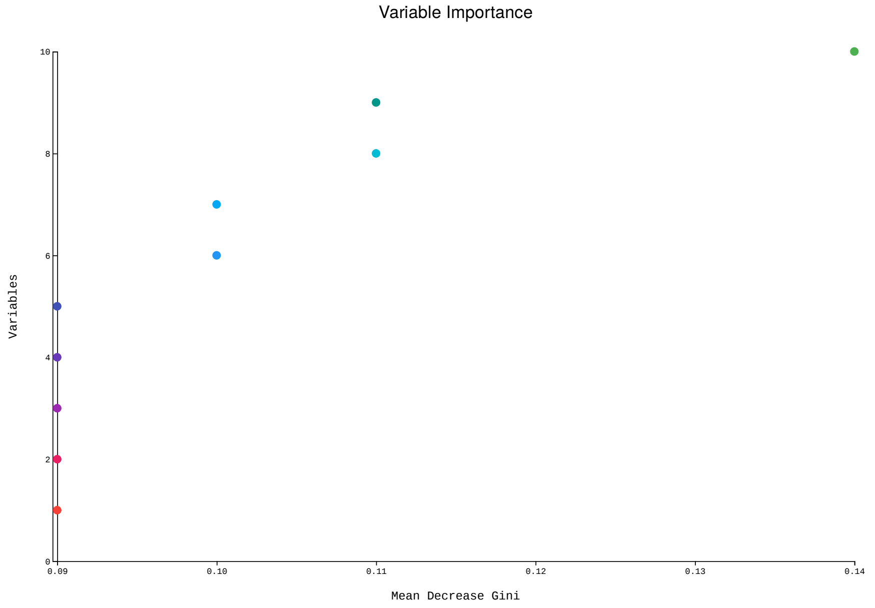


Figure 7:

The variable importance can be interpreted as follows:

1. **Gini indices:** as a "rule of thumb", a break in the variable importance plot will indicate the set of most important features. For example, in Fig. undefined the two most important breaks occur between hsa-miR-21-3p and hsa-miR-203a-5p, and between hsa-miR-574-3p and hsa-miR-21-5p.

A valuable plot for the identification of the ‘optimal’ biomarker set of sRNAs is the feature selection plot (Fig. 7). It reports the cross-validated prediction error of random forest models trained by increasingly adding sRNAs, with the order of additions being determined by the sRNA gini indices. This feature selection strategy has been applied in previous studies (Ashlock & Datta, 2012; Erho et al., 2013).

Notes on feature selection:

1. It has been suggested in (Svetnik, Liaw, & Tong, 2004) that the optimal set of features is given by the point in which the curve of feature selection plot is minimized. For the psoriasis example, it seems that the top 57 sRNAs obtain optimal classification results (10% OOB error, 1:57 features).
2. This plot is obtained by applying the strategy of (Svetnik et al., 2004) and consists of computing the cross-validated prediction error obtained from training different random forest models by iteratively adding features, with the order of additions being defined by the ranking of features obtained from the random forest trained with the full feature set. The number of folds used to partition the data is set to k=10.

Error rate plot

The last plot that Oasis' Classification module provides is the error rate plot (Fig. 8). It shows the Out-of-bag (OOB) error of the random forest model trained with the full set of features. This plot is useful to let you know whether the parameter NTREE of the forest has been correctly set. A sufficiently large NTREE will be reflected by OOB error rates for treatments, controls, and both classes that are parallel to the x-axis for higher NTREE values (green, black and red). Parallel error lines indicate constant classification performance for increasing number of trees.

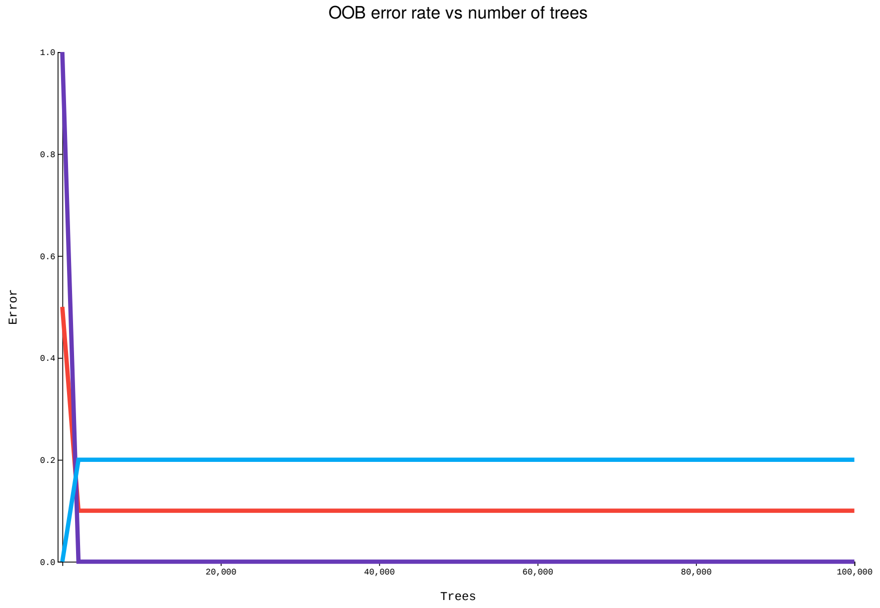


Figure 8: Out-of-Bag error plot. If the error lines converge, it means that the number of trees in the forest was correctly set.

Feature selection table

The classification works with the expression level of sRNA molecules in your dataset. Those small RNAs that are the most different between healthy samples and disease samples will show up as the ones with the highest feature importance.

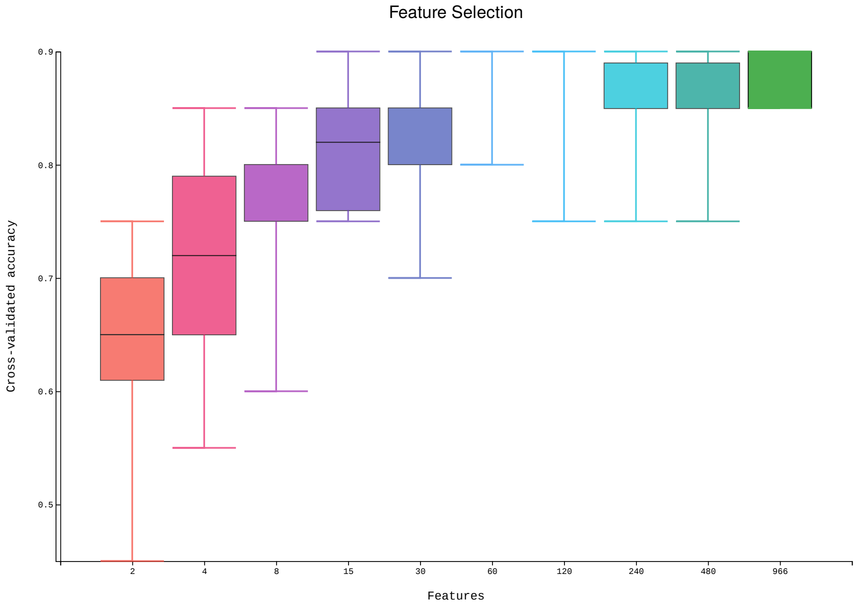
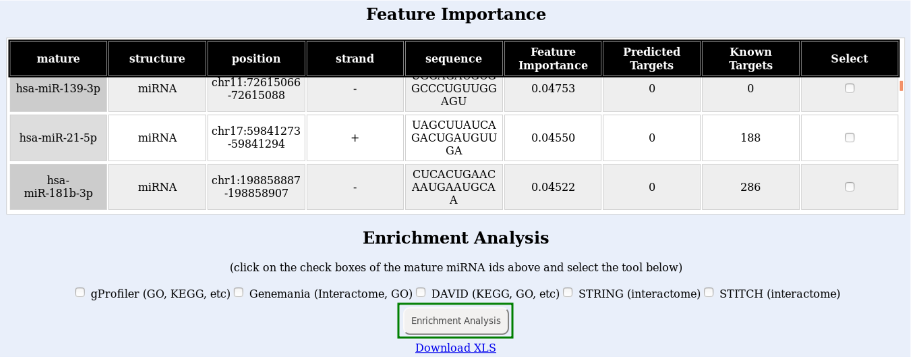


Figure 9: Feature selection plot showing how accurate each feature is for predicting the disease status.



Feature Importance

mature	structure	position	strand	sequence	Feature Importance	Predicted Targets	Known Targets	Select
hsa-miR-139-3p	miRNA	chr11:72615066 -72615088	-	GCCCUGUUGG AGU	0.04753	0	0	□
hsa-miR-21-5p	miRNA	chr17:59841273 -59841294	+	UAGCUUAUCA GACUGAUGUU GA	0.04550	0	188	□
hsa-miR-181b-3p	miRNA	chr1:198858887 -198858907	-	CUCACUGAAC AAUGAAUGCA A	0.04522	0	286	□

Enrichment Analysis

(click on the check boxes of the mature miRNA ids above and select the tool below)

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

Figure 10: Feature importance table with a list of all small RNA molecules from the datasets. You can click into the first column to open the external database record for microRNAs. Furthermore, you can click into the cells for the `predicted targets` or `known targets` to get the list of targets for a microRNA. If you want to perform enrichment analysis with any of the microRNA targets, you can use the checkboxes in the last column and then select any number of enrichment services at the bottom of the table.

Finally, submit your enrichment analysis by clicking the `Enrichment analysis` button (Fig. undefined, 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 case you want to analyse the results manually, you can download Excel tables of all analysed sRNAs (by clicking on the link `Download XLS`)

Enrichment services integrated into the Oasis results

gProfiler:
: Returns enriched GO categories, KEGG and REACTOME pathways, TRANSFAC regulatory motifs, miRBase miRNAs, CORUM protein complexes, Human Phenotype Ontologies and BioGRID protein- protein Genemania: interactions.

Genemania:
: Returns a protein-protein network, showing the protein products of the selected gene targets and how they associate with each STITCH other, as well STRING: as significantly enriched GO categories.

STITCH and STRING:
: Returns a single image of a protein-protein network. On top of having protein-protein interactions, STITCH includes small molecules, DAVID: drugs and ATPs associated with the target proteins as well.

DAVID
: Runs an enrichment test for all target genes, using various functional annotations (GO categories, KEGG pathways, BIOCARTA pathways, protein domains and so on).

References

- Ashlock, W., & Datta, S. (2012). Distinguishing endogenous retroviral LTRs from SINE elements using features extracted from evolved side effect machines. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(6), 1676–1689*. http://doi.org/10.1109/TCBB.2012.116
- 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
- Breiman, L. (2001). Random forests. *Machine Learning, 45(1), 5–32.* http://doi.org/10.1023/A:1010933404324
- Erho, N., Crisan, A., Vergara, I. A., Mitra, A. P., Ghadessi, M., Buerki, C., … Jenkins, R. B. (2013). Discovery and Validation of a Prostate Cancer Genomic Classifier that Predicts Early Metastasis Following Radical Prostatectomy. *PLoS ONE, 8(6).* http://doi.org/10.1371/journal.pone.0066855
- Fawcett, T. (2003). ROC Graphs : Notes and Practical Considerations for Data Mining Researchers. *HP Invent, 27.* http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf
- 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

- Iglewicz, B., & Hoaglin, D. C. (1993). How to Detect and Handle Outliers. *ASQC Quality Press (Vol 16). Retrieved from* https://books.google.de/books?id=siInAQAAIAAJ&q=isbn:087389247X+9780873892476&dq=isbn:087389247X+9780873892476&hl=de&sa=X&ved=0ahUKEwixosPei9XKAhVBPg8KHQDjDB8Q6AEIIjAA
- Joyce, C. E., Zhou, X., Xia, J., Ryan, C., Thrash, B., Menter, A., ... Bowcock, A. M. (2011). Deep sequencing of small RNAs from human skin reveals major alterations in the psoriasis miRNAome. *Human Molecular Genetics, 20(20), 4025...40.* https://dx.doi.org/10.1093%2Fhmg%2Fddr331