Microsynth

mRNA Sequencing for Differential Gene Expression Analysis

Compare gene expression profiles obtained under different conditions

Study genetic profiles from transcript to pathway level

Introduction

RNA sequencing, the analysis of gene expression profiles by next generation sequencing (NGS), has become a powerful tool to investigate the transcriptome of an organism. The comparison of two or more conditions (e.g. mutant vs. wild type) allows the identification of differentially expressed genes that are up- or downregulated under a specific condition. Typical applications include the comparison of gene expression profiles between normal and cancer tissue, cells in high and low nutrient environments, stressed and unstressed cells or cells from distinct developmental stages. Major prerequisites for any RNA sequencing study are the availability of an annotated reference genome or transcriptome and sufficient numbers of biological replicates.

Microsynth’s Competences and Services

Differential gene expression analysis by RNA sequencing is one of Microsynth’s core competences. Based on years of experience, we provide a one-stop service from experimental design to bioinformatics analysis (see ***Figure 1***). You can either outsource the whole process or only single steps to us. To ensure a high reliability and accuracy of our product, the whole RNA sequencing pipeline including the bioinformatics analysis was evaluated and validated based on RNA spike-ins designed by the External RNA Control Consortium (ERCC). For further information on our validation process and possibilities to validate your own study, please contact us.

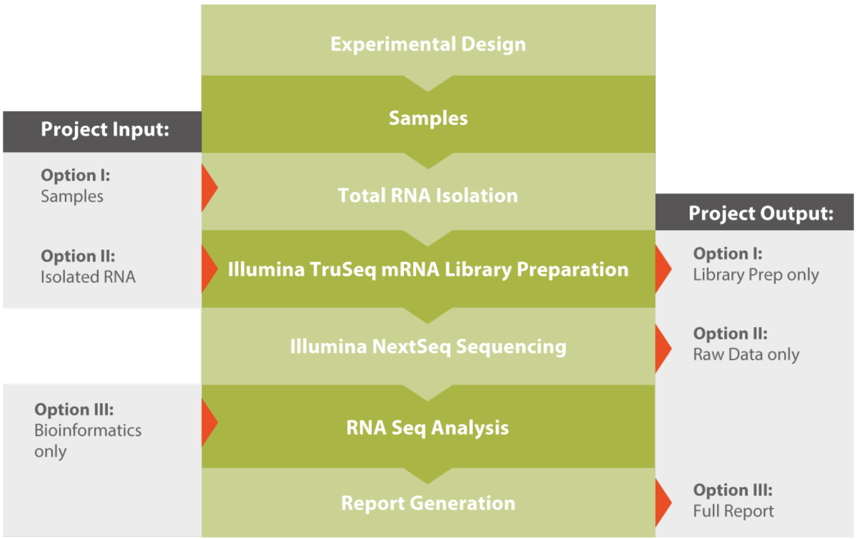


Figure 1. *Microsynth's workflow for RNA sequencing projects. The workflow can be entered and exited at various steps dependent on the customer's requirements.*

Experimental Design

The gain and impact of a study highly depends on its experimental design. The use of controls as well as appropriate sampling and RNA isolation methods are only a few examples of points to consider. What is important for any RNA sequencing project is the number of biological replicates. To obtain statistically valid results for a differential gene expression analysis, we usually advise to include at least three biological replicates per condition. Make use of our experience – our NGS specialists are happy to assist you from the start.

RNA Isolation

You can either perform the extraction yourself or outsource this critical step to us. Microsynth has extensive experience in RNA isolation from various demanding tissues and matrices.

Library Preparation and Sequencing

Following a quality check of the RNA, we will either perform poly(A) enrichment or ribosomal RNA (rRNA) depletion, depending on the organism or the target RNA to be studied. This step is essential since total RNA includes a large fraction of rRNA and other non-mRNA species and sequencing should be restricted to mRNA to avoid losses in sequencing depth. A stranded Illumina cDNA library is created by reverse transcription including the ligation of sequencing adaptors with barcodes. Finally, the libraries are pooled and sequenced on the Illumina NextSeq platform with a typical single-end read length of 75 bp. The sequencing depth per sample highly depends on the studied organisms and the desired sensitivity. Benchmarks for complex eukaryotic genomes (e.g. human, rat, mouse) are 100-150 million reads for high sensitivity and 20-30 million reads for typical sensitivity. Prokaryotic genomes require approximately 5-fold less reads.

Bioinformatics Analysis

After quality control, the sequencing reads are mapped to the reference genome using the software Salmon [1] or STAR [2], which both address the difficulty of mapping spliced reads. As input for statistical analysis the reads that uniquely map to a gene are counted. The identification of differentially expressed genes is performed by specialized statistical software such as DESeq2 [3]. Read counts are normalized, and the variance based on the replicates per condition is calculated. Finally statistical testing is applied to identify differentially expressed genes that are significantly up- or downregulated.

For organisms with available pathway information, a complementary service is provided identifying significantly up- or downregulated pathways (Gene ontology and KEGG classification).

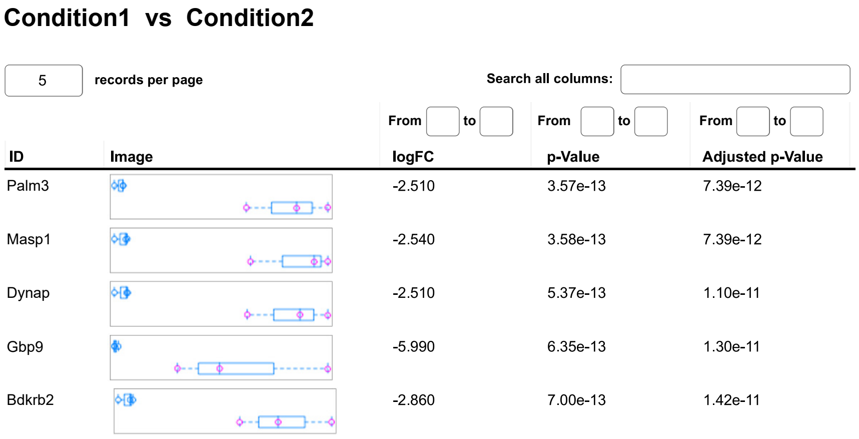
Example Results

The major outcomes of the differential gene expression analysis are the observed fold changes for each gene together with statistical measures and the normalized counts per sample (see ***Figure 2A***). Statistics for all genes are additionally summarized in an interactive table together with boxplots visualizing the observed fold changes (see ***Figure 2B***). The interactive table allows you to sort the data by any represented measure or to search for specific features of interest.

A

							Normalized Counts Condition01				Normalized Counts Condition02			
										Average				Average
ID	***baseMean***	***log2FC***	***lfcSE***	***stat***	***pvalue***	***padj***	***Rep01***	***Rep02***	***Rep03***	***Condition01***	***Rep01***	***Rep02***	***Rep03***	***Condition02***
7157_TP53	105974.5	4.556	0.169	27.01	1.1E-160	8.5E-158	218081.2	150818.2	242498.3	203799.2	8100.9	7822.9	8525.7	8149.8
367_AR	3063.3	4.063	0.109	37.18	1.3E-302	5.3E-299	6204.4	5565.9	5588.9	5786.4	335.1	367.1	318.3	340.1
80326_WNT10A	609.5	3.437	0.141	24.37	3.9E-131	1.3E-128	996.2	1190.9	1167.4	1118.2	92.3	109.2	100.8	100.8
5083_PAX9	707.0	3.362	0.193	17.43	4.6E-68	2.4E-66	927.7	1588.8	1369.3	1295.3	121.9	125.0	109.6	118.8
1535_CYBA	3275.4	3.274	0.192	17.05	3.5E-65	1.6E-63	5510.2	4391.7	8003.6	5968.5	611.0	584.2	551.5	582.2
2253_FGF8	38443.1	3.269	0.149	22.00	2.8E-107	4.6E-105	75902.7	52390.7	81335.6	69876.3	7641.3	6975.6	6412.7	7009.9
3730_KAL1	569.8	3.129	0.154	20.29	1.5E-91	1.7E-89	942.4	1202.6	932.9	1026.0	123.6	119.7	97.3	113.5
10913_EDAR	3440.5	3.096	0.345	8.98	2.6E-19	1.7E-18	6664.0	2986.4	9183.0	6277.8	555.3	647.3	606.7	603.1
4128_MAOA	68960.8	2.889	0.099	29.04	2.2E-185	2.9E-182	115356.3	121088.8	128663.9	121703.0	15206.0	15655.0	17794.8	16218.6

B



Condition1 vs Condition2

5 records per page

Search all columns:

		From to	From to	From to
ID	**Image**	**logFC**	**p-Value**	**Adjusted p-Value**
Palm3		-2.510	3.57e-13	7.39e-12
Masp1		-2.540	3.58e-13	7.39e-12
Dynap		-2.510	5.37e-13	1.10e-11
Gbp9		-5.990	6.35e-13	1.30e-11
Bdkrb2		-2.860	7.00e-13	1.42e-11

Figure 2. *Summary tables resulting from the differential gene expression analysis. 2A. Excerpt of a table summarizing the results of the analysis for two conditions with three replicates each. ID: gene ID; baseMean: average number of read counts; log2FC: log2 transformed fold change between conditions; lfcSE: standard error of log fold change; p-value: Wald test p-value; padj: p-value adjusted for multiple testing. 2B. The statistical results are also available as an interactive html-table allowing sorting and searching for specific features.*

For each comparison in the differential gene expression analysis, various graphs are provided for an easy and visual overview of the results (see ***Figure 3***). The graphs include MA plots visualizing the distribution of the differentially expressed genes, heatmaps for sample to sample distances and the top up- and downregulated genes as well as principal component analysis plots.

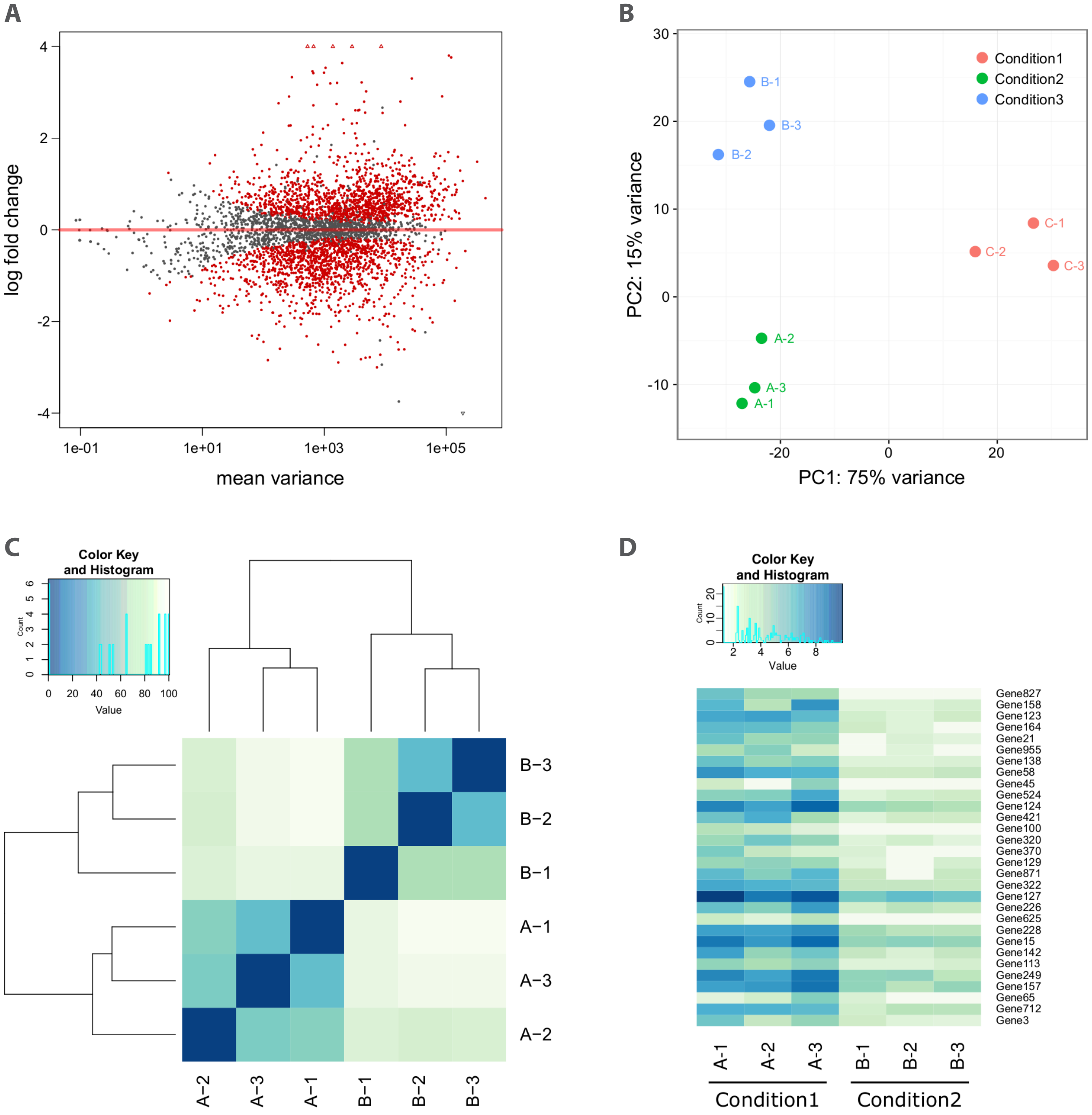


Figure 3. *Examples of provided overview plots. 3A. MA-plot visualizing the distribution of differentially expressed genes by plotting the mean expression against the log fold change. 3B. Principle Component Analysis plot to visualize sample clustering. 3C. Heatmap showing sample to sample distances for given conditions. 3D. Heatmap displaying the 30 most upregulated genes. A similar heatmap is provided for the most downregulated genes, both allowing an easy identification of putative candidate genes.*

The results for the complementary pathway analysis include graphical representations of the up- and downregulated pathways for both gene ontology and KEGG terms (see ***Figure 4***). The pathway analysis module is available for most common model organisms such as human, mouse, rat or *E. coli*. Information is provided for all genes involved in a pathway count and log fold change, allowing a more detailed analysis of the identified pathways.

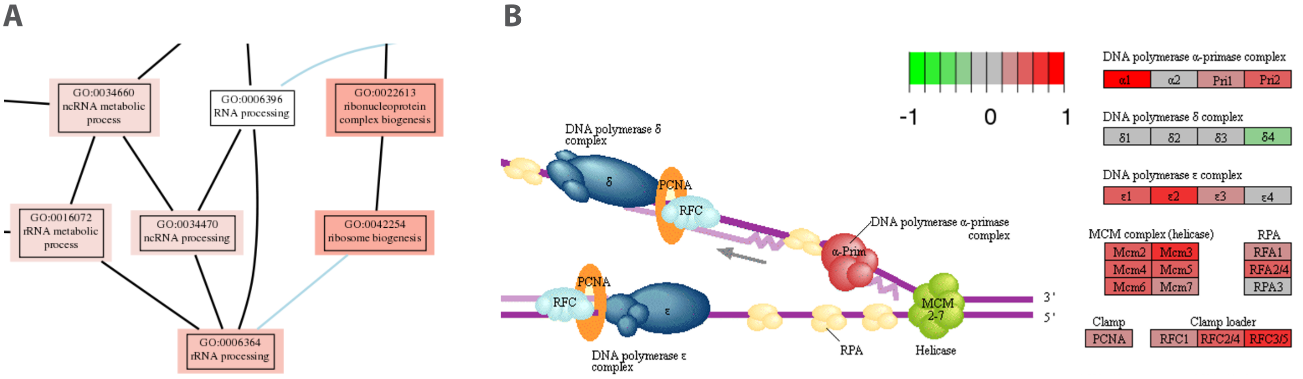

A
GO:0034660 ncRNA metabolic process
GO:0006396 RNA processing
GO:0022613 ribonucleoprotein complex biogenesis
GO:0016072 rRNA metabolic process
GO:0034470 ncRNA processing
GO:0042254 ribosome biogenesis
GO:0006364 rRNA processing
B
-1
0
1
DNA polymerase δ complex
δ
PCNA
RFC
DNA polymerase α-primase complex
α-Prim
MCM 2-7
3'
5'
PCNA
RFC
ε
DNA polymerase ε complex
RPA
Helicase
DNA polymerase α-primase complex
α1
α2
Pri1
Pri2
DNA polymerase δ complex
δ1
δ2
δ3
δ4
DNA polymerase ε complex
ε1
ε2
ε3
ε4
MCM complex (helicase)
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
RPA
RFA1
RFA2/4
RPA3
Clamp
PCNA
Clamp loader
RFC1
RFC2/4
RFC3/5

Figure 4. *Graphical output of the pathway analysis module. 4A. Excerpt from a network graph for upregulated pathways where colored nodes represent significantly upregulated gene ontology terms. 4B. Detail of a KEGG pathway analysis result. Both are available for up- and downregulated pathways.*

Related Services

For the analysis of the small RNA fraction (e.g. microRNA) of an organism, please refer to our small RNA sequencing service. If you are interested in the transcriptional analysis of whole microbial communities, our shotgun metatranscriptomics service offers an attractive solution.

References

[1] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods.

[2] Dobin et al. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

[3] Love et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15. https://doi.org/10.1186/s13059-014-0550-8