Application Note

Illumina RNA Sequencing For Differential Gene Expression Analysis

Introduction

Next-generation sequencing (NGS) is a powerful technique to perform genome-wide transcriptional analysis of any biological organism (often also called RNA-Seq). By comparing two or more conditions RNA-Seq permits to find differentially expressed genes – genes that are up- or down-regulated under specific conditions. Typical examples include the comparison of transcription profiles from normal tissues versus cancer tissues, cells in high versus low nutrient environments, unstressed versus stressed cells or from distinct developmental stages of an organism. A prerequisite for any RNA-Seq study is the availability of an annotated reference genome or a reference transcriptome.

Why RNA-Seq?

The advantage of RNA-Seq over conventional microarray studies is that (i) no prior knowledge about gene models is necessary and (ii) an increased dynamic range is observed with overall higher sensitivity, reliability and reproducibility levels. In addition, many RNA-Seq protocols allow to analyze both the sense as well as the natural antisense transcripts (NATs) of genes. NATs are widespread in eukaryotic and procaryotic genomes and are now acknowledged as important modulators of gene expression.

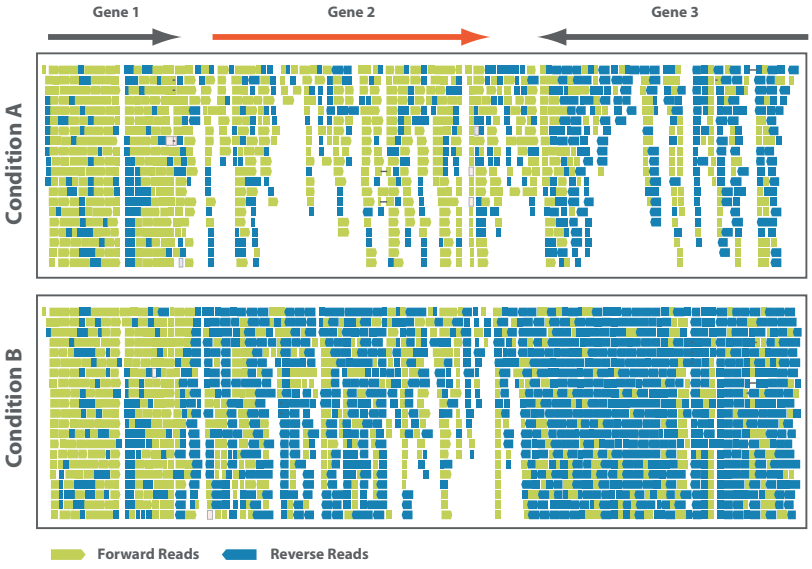


Figure 1. Example for sense/antisense expression levels (based on mapped forward and reverse reads) from fungal cultures treated under two different conditions - especially evident for gene 2. Where as condition A favors the generation of sense transcripts, condition B produces primarily natural antisense transcripts.

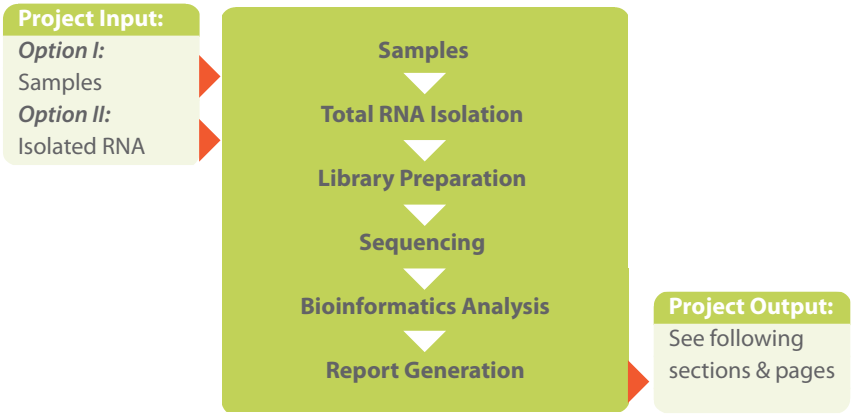
Microsynth Competences and Services

Experimental Design: As an expert in the area of RNA-Seq, Microsynth is able to provide a full service (from experimental design consulting up to bioinformatics analysis). Important for any RNA-Seq project is the number of biological replicates. To finally obtain statistical significance for your differential gene expression analysis, we usually advise to include at least 3 biological replicates per condition. **RNA Isolation:** Either you leave it up to Microsynth or you use a commercial kit to isolate total RNA. **Library Preparation and Sequencing:** Following a quality check of your samples, Microsynth will perform an mRNA enrichment or an rRNA depletion depending on the studied organism. This step is essential because the fraction of rRNA is high and sequencing should be restricted to mRNA (or miRNA). An Illumina

cDNA library is generated by reverse-transcription including specific sequencing adaptors with barcodes. Finally, the libraries are pooled and sequenced on the Illumina machine. The envisaged number of reads per library depends on the organism under study and the desired sensitivity. Whereas the benchmark for complex eukaryotic genomes (e.g. human, rat, mouse) requires 100-150 M reads (high sensitivity) or 20-30 M reads (low sensitivity) per sample, a 10-fold less amount of reads is required for bacteria. **Bioinformatics Analysis:** Reads derived from the sequencing are mapped against the reference genome of the organism under study using the Bowtie2, TopHat or STAR software. TopHat primarily addresses the difficulty of mapping spliced reads in eukaryotic genomes (i.e. reads spanning two exons). Finally the reads per gene are counted and used as input for statistical analysis. Specific statistical software packages are used to seek for differentially expressed genes. These packages first normalize the data, then calculate the variance based on the replicates for each condition and finally compute statistical tests to find differentially expressed genes. **Provided Output Files:** You will receive a report with following content:

- raw counts of the mapping
- differential analysis results
- heatmap with top 30 genes
- sample clustering

Besides, raw sequence data, BAM mapping files and a brochure describing some statistical details, will be provided.



Examples for Most Important Output Files Provided by Microsynth

A

							Normalized Counts Condition01				Normalized Counts Condition02			
ID	baseMean	log2FC	lfcSE	stat	pvalue	padj	Rep01	Rep02	Rep03	Average Condition01	Rep01	Rep02	Rep03	Average 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

Condition01 vs Condition02

records per page Search all columns:

From to From to From to

ID	Image (Link to Image)	LogFC	p-Value	Adjusted p-Value
7157_TP53		4.556	1.07E-160	8.45E-158
367_AR		4.063	1.33E-302	5.25E-299
80326_WNT10A		3.437	3.87E-131	1.28E-128
5083_PAX9		3.362	4.61E-68	2.37E-66
1535_CYBA		3.274	3.46E-65	1.56E-63
2253_FGF8		3.269	2.77E-107	4.57E-105
3730_KAL1		3.129	1.49E-91	1.74E-89

Figure 2. Summary tables resulting from the differential gene expression analysis. **Figure 2A.** Extract of a table summarizing the main results of the analysis for two conditions including three replicates each. Similar tables are provided for each pairwise comparison in the experiment. In addition, a summary table for all pair-wise analyses is included. Whereas in the first column the gene ID (gene name) is indicated the column ***baseMean*** lists the average read counts for conditions 01 and 02. The column ***log2FC*** lists the logarithmic fold change between Condition01 and Condition02. Columns **pvalue** and ***padj*** list the p value as well as the adjusted p values for multiple testing. Finally, the normalized read counts for the replicates in Condition01 and Condition02 are listed. **Figure 2B.** The statistic part of the table is also available as .html version allowing to sort and search for specific features and links to box plots.

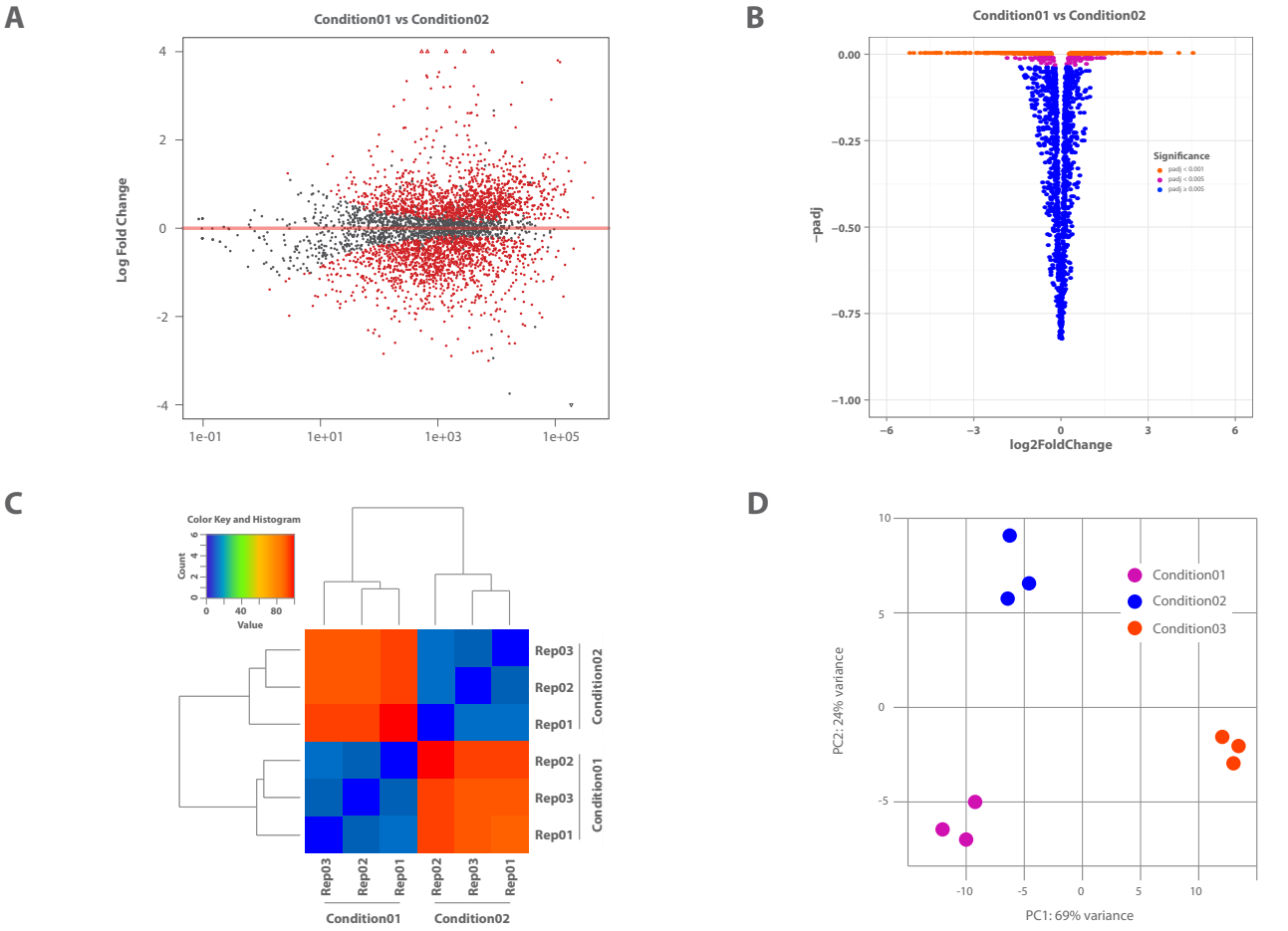


Figure 3. Examples of two plots provided. **Figure 3A.** MA-Plot visulazing the distribution of differentially expressed genes by plotting their normalized mean expression against the Log Fold Change. **Figure 3B.** Vulcano plot visualizing differentially expressed genes by plotting the Log Fold Change against the adjusted p-value. **Figure 3C.** Heatmap showing the sample-to-sample distances for a given condition. This analysis is helpful in detecting possible outliers from a sample pool. **Figure 3D** Principle Component Analysis supplements the heatmap to visualize sample clustering.

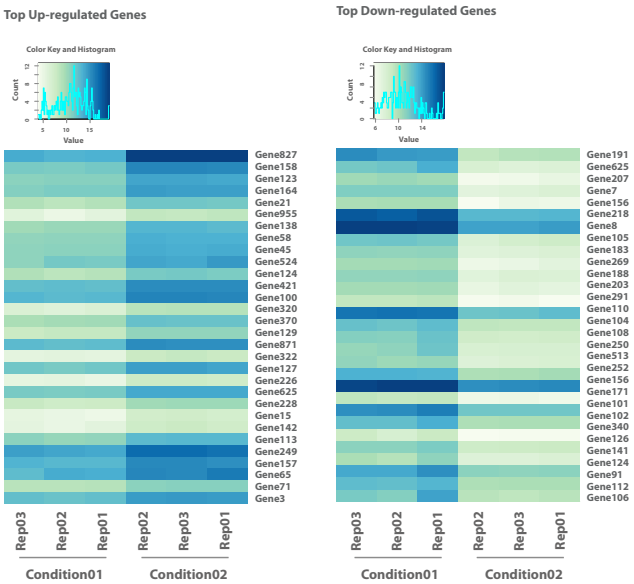


Figure 4. Heatmap of up- and down-regulated genes. For each studied condition, customers will obtain a heatmap where the 30 top up- and downregulated genes are displayed helping in identifying putative candidate genes.