Microsynth

Transcriptome analysis - a daunting task?

Transcriptome analysis is a daunting task due to the amount and fragmentation of the data. We will show you how to cope with this complexity in this short bulletin.

Introduction

To study a transcriptome can be a daunting task and especially so for eukaryotic organisms. The most recent human reference assembly GRCh38 has for instance 244550 unique exons with a length of up to 91671 bases, a mean of 330 bases and a total length of 80 million bases scattered across the 3 billion bases of the human genome. There are 52000 transcripts from 26475 genes, 22302 associated GO terms and 318 KEGG pathways. Studying the effect of a stressor on the transcriptomic level, finding your gene or pathway of interest and be confident in your findings needs preparation.

Example Study

In the following we will discuss an example and the decisions behind it, to guide you through an RNA sequencing project. We will demonstrate how to cope with the complexity described in the introduction and how to invest your finite time and resources to get reliable results.

Consider the following experimental setup. A stressor and its effect on the human transcriptome need to be studied *in vitro*. There are cells undergoing the treatment and cells kept as control. To be able to cope with variations and not be misled by outliers, biological replicates are recommended for the treated cells as well as for the control cells.

In the following we will describe a standard procedure at Microsynth given the experimental setup above.

1) After total RNA isolation, poly(A) enrichment is undergone in the construction of the sequencing libraries to remove non-coding RNA from the total RNA.

2) The main goal in this example is to count molecules sequenced from the libraries to quantify gene expression levels. Therefore, there is no need to sequence long stretches of the molecules but just enough as to be sure that a present molecule fits uniquely to a certain gene. A good compromise is to sequence single-end reads of 75 bases in length.

3) The levels of expression between genes vary in orders of magnitude. House-keeping genes will make up the bulk of the sequenced reads. To counter this uneven distribution and still be able to detect molecules fitting to genes of interest, which might have small or moderate expression levels, 30 million reads per sample are a good start. Roughly a fourth of the genes will have no detectable molecules and the range for the other genes will be from one counted molecule up to tens of thousands or even hundreds of thousands of molecules. So if we consider a treatment batch with three replicates and a control batch also with three replicates, 180 million single-end reads each with 75 bases in length will be sequenced.

4) To help you analyze all this data, Microsynth will carry out bioinformatics analysis. In this analysis the reads are mapped to the transcriptome in a splice-aware fashion, the mapped reads are counted for each gene and then normalized across your experiment. These normalized values are then statistically analyzed comparing the replicates from the two different conditions (treatment and control) to each other. In this way statistical significance of the fold changes and up- or down-regulation of any gene is calculated (see ***Figure 1***). Optionally a pathway analysis may be added, indicating which pathways are up- or down-regulated as well (see ***Figure 2***).

Microsynth AG, Switzerland
Schützenstrasse 15 · P.O. Box · CH-9436 Balgach · Phone +41 71 722 83 33 · Fax +41 71 722 87 58 · info@microsynth.ch · www.microsynth.ch

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

Condition1 vs Condition2

5 records per page Search all columns:

ID	Image	From to logFC	From to p-Value	From to 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 1. *Summary tables resulting from the differential gene expression analysis. 1A. 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. 1B. The statistical results are also available as an interactive html-table allowing sorting and searching for specific features.*

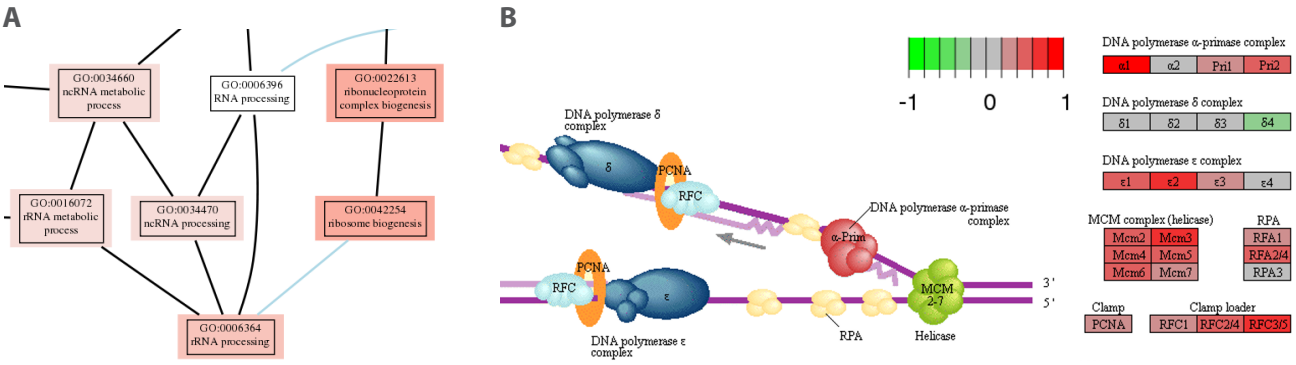

Figure 2. *Graphical output of the pathway analysis module. 2A. Excerpt from a network graph for upregulated pathways where colored nodes represent significantly upregulated gene ontology terms. 2B. Detail of a KEGG pathway analysis result.*

Conclusion

We hope to have shown that the daunting task of transcriptome analysis can be feasible, instructive and even fun if done scientifically correct. Please contact us with questions or requests. Our next generation team will gladly help you.

Need more information?
Send us your project requests via our contact form

Microsynth AG, Switzerland
Schützenstrasse 15 · P.O. Box · CH-9436 Balgach · Phone +41 71 722 83 33 · Fax +41 71 722 87 58 · info@microsynth.ch · www.microsynth.ch