Application Note

Expression analysis of mRNA-Seq reads from *Arabidopsis* tissues

Application note on mRNA-Seq analysis using CLC Genomics Workbench

CLC bio © Copyright

CLC bio

Expression analysis of mRNA-Seq reads from *Arabidopsis* tissues

CLC bio's RNA-Seq tool allows the mapping of transcriptome reads to reference genomes and de novo assemblies. A set of Expression Analysis tools provides for the comparison, statistical analysis, and visualization of multiple RNA-Seq datasets.

Data

This application note is based on the publically available sets of reads from *Arabidopsis thaliana* mRNA libraries[1,2]. The annotated *A. thaliana* chromosomes used as references are available from GenBank[3].

[1] mRNA libraries submitted to GenBank SRA (Short Read Archive) by Whitehead Institute researchers
[2] The reads data from embryo tissue is available at http://www.ncbi.nlm.nih.gov/sra/SRR307073, and the reads data from endosperm is available at http://www.ncbi.nlm.nih.gov/sra/SRR307076
[3] Using the Search button in CLC Genomics Workbench -> Search for Sequences at NCBI -> type NC_00307*

Figure 1: Selection of file with sequencing reads in CLC Genomics Workbench

Analysis Workflow

RNA-Seq Analysis

While running the RNA-Seq tool, the reads (Figure 1) and the annotated *Arabidopsis* reference chromosomes are selected. In the tool wizard, you can specify the mapping parameters, e.g. the number of mismatches allowed and the maximum number of hits for a read. In this experiment, we run the assembly with the default mapping parameters allowing for a maximum of two mismatches and the maximum of ten hits for a read.

The reads set we are using here contains short IGA reads, 36 nt long. CLC Genomics Workbench detects the read length automatically and selects the short read aligner for this mapping.

We also select the default parameters for exon discovery - RPKM as the reporting expression value - and we also map the embryo dataset in order to obtain two mappings from different tissues for comparison.

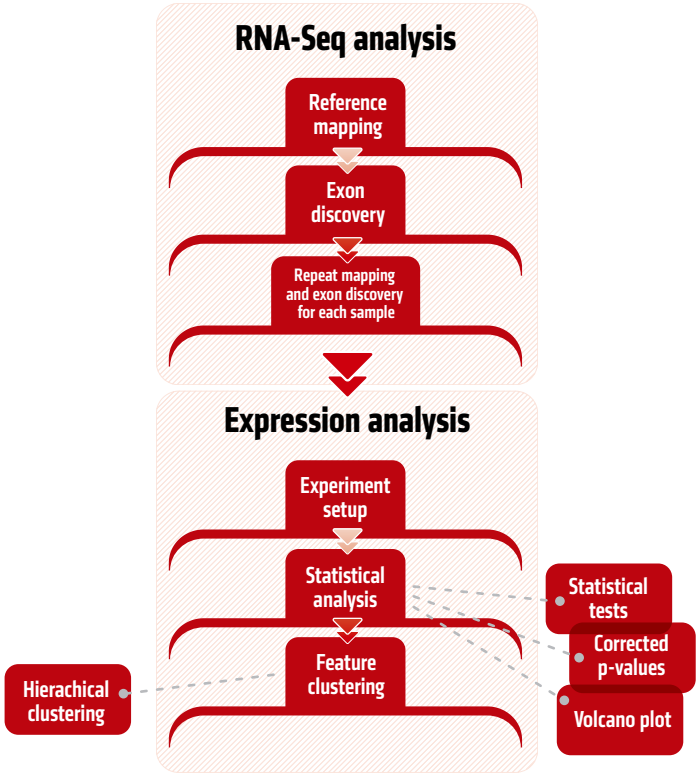

Figure 2: Workflow for RNA-Seq and Expression analysis using CLC Genomics Workbench

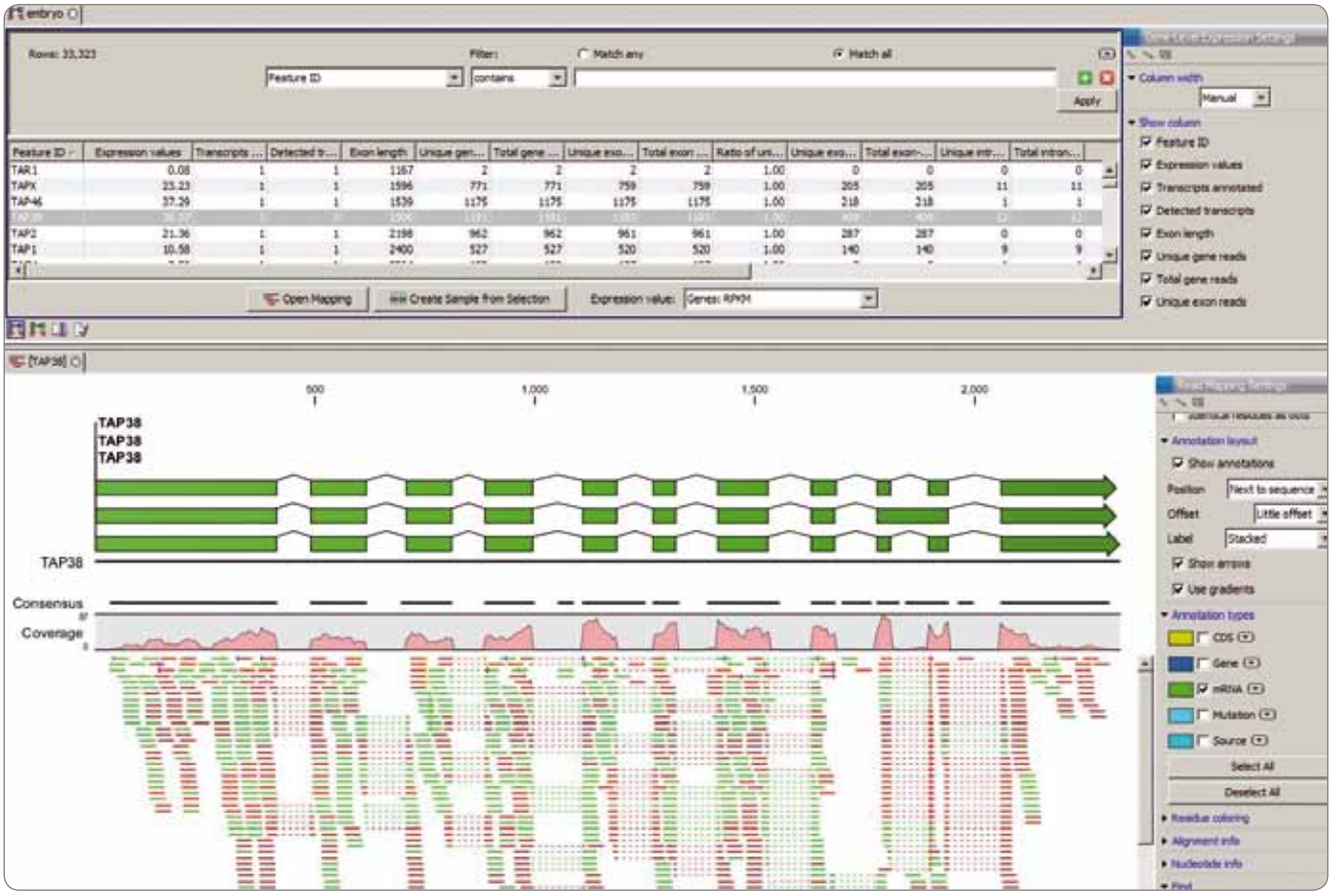

Figure 3: RNA-Seq table (top) and the mapping to the reference gene (TAP38) displayed in the read mapping view (bottom)

Exploring RNA-Seq tables and individual mappings

An RNA-Seq Expression table is produced upon completion of each mapping. The table can be searched, sorted, and exported, and every individual mapping can be opened and explored in the read mapping view (Figure 3).

Expression Analysis Tools

The first step of the expression analysis in the Workbench is to set up an experiment using your data. We select the embryo and the endosperm RNA-Seq samples for a two group comparison of unpaired samples and this tool produces a new table showing embryo vs. endosperm.

To identify differential expression, you have the choice of a number of standard statistical tests that are suitable for different data types and different types of experimental settings. There are two main categories of tests: tests that assume that the data has Gaussian distributions and compare means (Gaussian tests) and tests that compare proportions and assume that data consists of counts (tests on proportions).

For the statistical analysis, we test on proportions and run Kal's Z-test which is the test for comparing a single sample against another single sample. The tool adds new columns to the embryo vs. endosperm experiment table containing the results of the Z-test. By adding filters to the table view, we can search the table and select genes that satisfy specific criteria (Figure 4). The list of genes that appears after filtering can be saved as a new experiment table for a better overview. In this example, we manage to reduce the

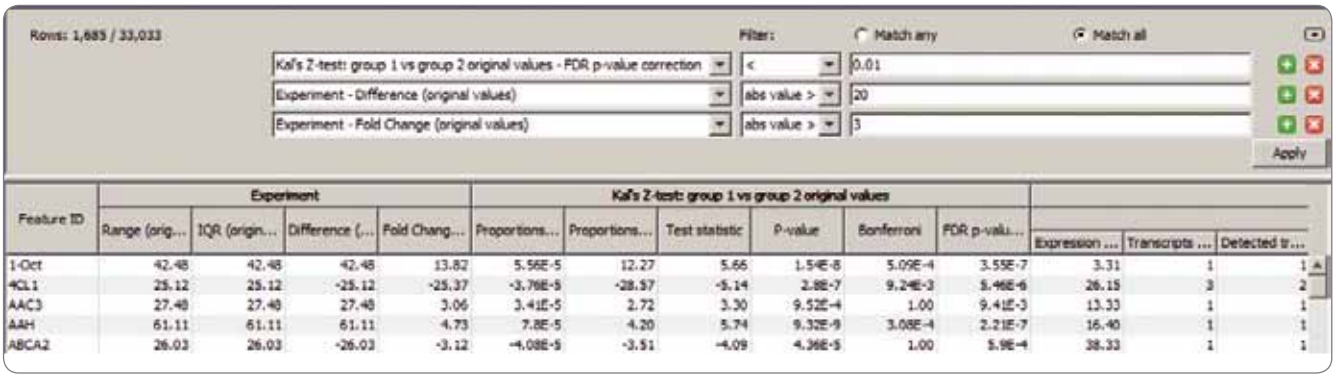

Figure 4: Expression Experiment Table: Filtering the data

number of genes from approximately 33,000 to 1,685 genes by adding three filters as shown in Figure 4.

The Workbench offers you the options of transformation and normalization of the expression values. This leads to a compressed dynamic range of the expression values and allows for easier visualization.

We transform the expression values for these 1,685 genes using the Logarithm transformation (Log10) as transformation method. The tool adds the new columns Transformed values under each of the samples in the experiment table. Afterwards, we cluster the features (genes) based on the transformed expression values. For clustering, we select Euclidean distance for Distances measure, and Average linkage for Clusters linkage criteria.

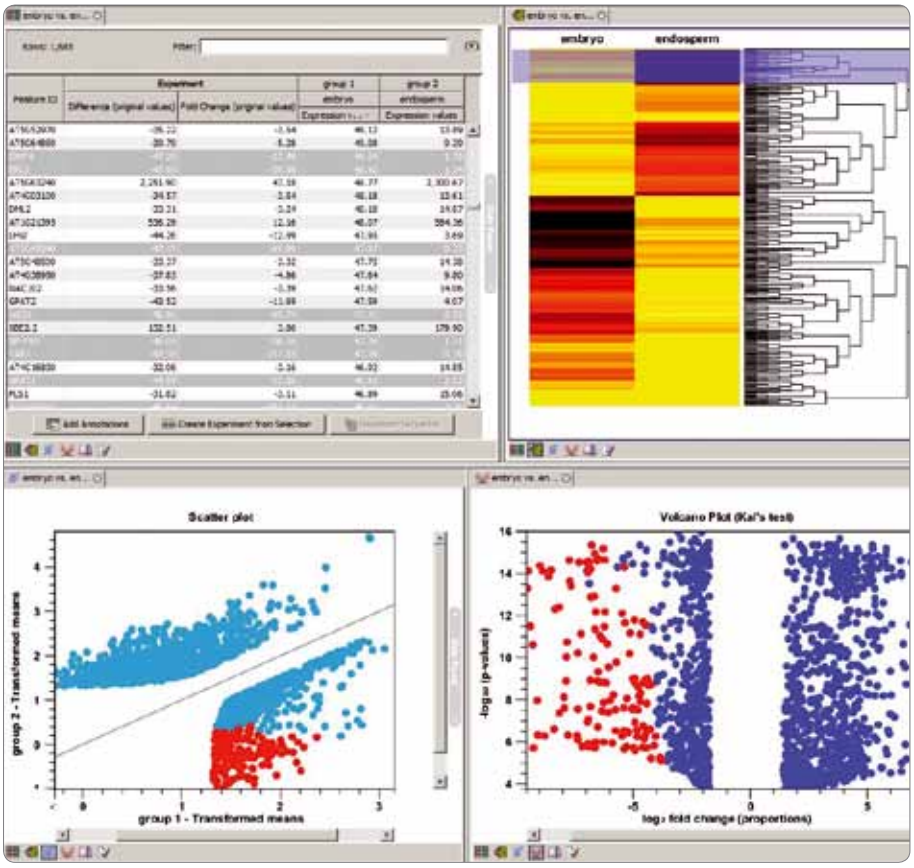

Figure 5: Visualization of selected genes in multiple linked views

Results

The resulting experiment table can be opened in multiple views using the icons in the lower left corner of the window. Each view is interactive and linked to other views of the same data. As an example you can see the on-the-fly changes illustrated on the right. Selecting the cluster of genes which are down-regulated in endosperm when compared to embryo (Figure 5) in the table view, will instantly select the same genes in other open views showing the same data (in this illustration in the heatmap, in the scatter plot and in the volcano plot). From the table view, you can easily extract this group of genes to a new experiment.

CLC bio · EMEA
Finlandsgade 10-12
Katrinebjerg · DK-8200 Aarhus N
Denmark
Phone: +45 7022 5509

CLC bio · Americas
10 Rogers St # 101
Cambridge · MA 02142
USA
Phone: +1 (617) 945 0178

CLC bio · AsiaPac
69 · Lane 77 · Xin Ai Road · 7th fl.
Neihu District · Taipei · Taiwan 114
Taiwan
Phone: +886 2 2790 0799

SCAN FOR MORE