Advanced RNA-Seq

USER MANUAL

User manual for
Advanced RNA-Seq 1.5

Windows, Mac OS X and Linux

November 2, 2016

This software is for research purposes only.

QIAGEN Aarhus
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark

QIAGEN

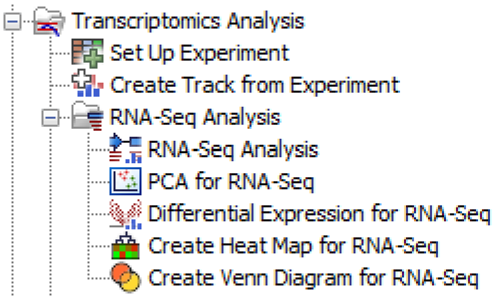
Contents

Chapter 1

Introduction to the Advanced RNA-Seq plugin

The Advanced RNA-Seq plugin provides new tools for statistical analysis and exploratory visualization of RNA-Seq data.

The plugin adds four new tools to the RNA-Seq Analysis folder (see figure 1.1).



Transcriptomics Analysis
- Set Up Experiment
- Create Track from Experiment
- RNA-Seq Analysis
  - RNA-Seq Analysis
  - PCA for RNA-Seq
  - Differential Expression for RNA-Seq
  - Create Heat Map for RNA-Seq
  - Create Venn Diagram for RNA-Seq

Figure 1.1: *The Toolbox with the advanced RNA-Seq tools*

The **PCA for RNA-Seq tool** provides 2D Principal Component Analysis with support for metadata visualization.

The **Differential Expression for RNA-Seq tool** is a new multi-factorial statistical analysis tool based on a negative binomial model. It uses a generalized linear model approach influenced by the multi-factorial EdgeR method [Robinson et al., 2010]. This tool produces a new kind of result, the statistical comparison track, which can be visualized at specific genomic locations together with other track types. The statistical comparison track also offers a volcano plot view.

The **Create Heat Map tool** shows a two dimensional heat map of expression values. Each column corresponds to one sample, while each row corresponds to a feature (a gene or a transcript). The samples and features are both clustered hierarchically.

The **Create Venn Diagram tool** shows the differentially expressed genes shared between statistical comparison tracks. The genes or transcripts considered to be differentially expressed can be controlled by setting appropriate p-value and fold change thresholds.

The **Create Expression Browser tool** allows expression values for several samples to be compared, together with annotations and statistical comparison tracks.

All of the new tools are designed for, and will only work on, RNA-Seq data (Expression tracks or the new statistical comparison tracks). They also make use of the new, unified metadata framework.

1.1 Input data and normalization

The Advanced RNA-Seq plugin requires RNA-Seq input data in the form of Expression Tracks. There is no support for microarray data or the Experiment objects used by the classic Transcriptomics Analysis tools.

Expression Tracks can be created using the RNA-Seq Analysis tool with a choice of possible expression values:

- **RPKM** This is the expression value measured in RPKM [Mortazavi et al., 2008]: $\text{RPKM} = \frac{\text{total exon reads}}{\text{mapped reads(millions)} \times \text{exon length (KB)}}$.
- **TPM** Transcripts Per Million This is computed as $\frac{\text{RPKM} \cdot 10^6}{\sum \text{RPKM}}$, where the sum is over the RPKM values of all genes/transcripts (see `http://bioinformatics.oxfordjournals.org/content/26/4/493.long`
- **Total counts** This is all the reads that are mapped to this gene - both reads that map uniquely to the gene or its transcripts and reads that matched to more positions in the reference which were assigned to this gene.
- **Unique counts** This is the number of reads that match uniquely to the gene or its transcripts.

The choice of expression value only affects how Expression Tracks are visualized in the track view. The results from the Advanced RNA-Seq plugin are not affected by this choice, as the most appropriate expression value is automatically selected for the analysis being performed. For detection of differential expression this is the 'Total counts' value, and for the other tools this is a normalized and transformed version of the 'Total counts' as described below.

1.1.1 Normalization

Since the sequencing depth might differ between samples, a per-sample library size normalization must be performed before samples can be compared. In contrast to the classic Transcriptomics Analysis tools, this normalization is automatically applied by the tools.

All of the tools in the Advanced RNA-Seq plugin use the TMM (trimmed mean of M values) normalization method [Robinson and Oshlack, 2010] to calculate effective libraries sizes, which are then used as part of the per-sample normalization. TMM normalization is the normalization used in EdgeR [Robinson et al., 2010].

TMM normalization adjusts library sizes based on the assumption that most genes are not differentially expressed. Therefore, it is important not to make subsets of the count data before doing statistical analysis or visualization, as this can lead to differences being normalized away.

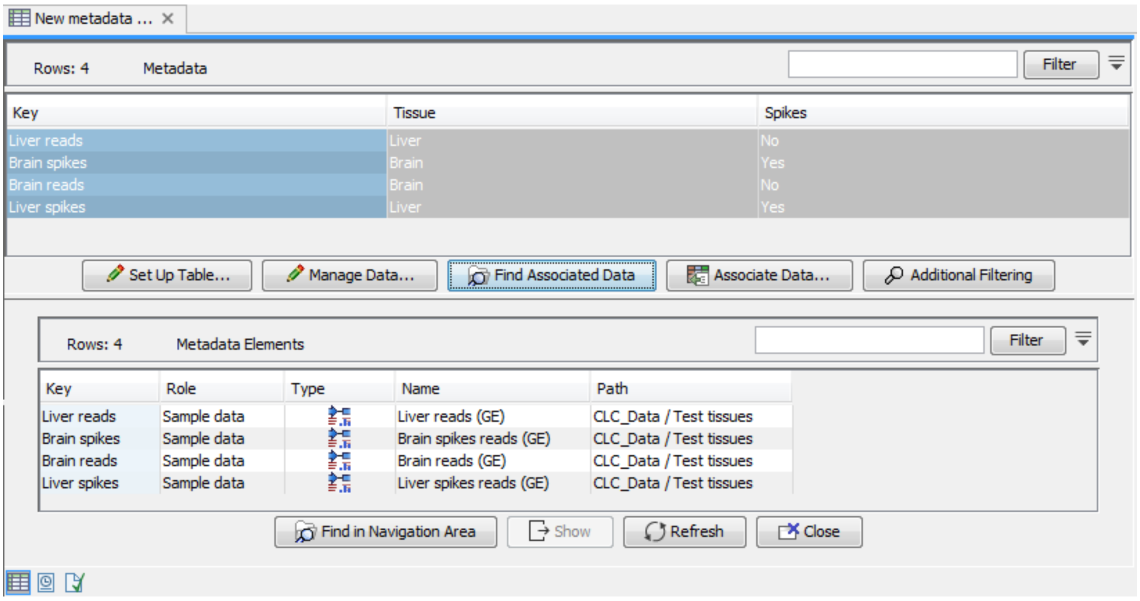
For the expression visualization tools (Create Heat Map and PCA for RNA-Seq) additional filtering and normalization are performed:

- 'log CPM' (Counts per Million) values are calculated for each gene. The CPM calculation uses the effective library sizes as calculated by the TMM normalization.
- After this first normalization, a second one is performed across samples for each gene: the counts for each gene are mean centered, and scaled to unit variance.
- Genes or transcripts with zero expression across all samples or invalid values (NaN or +/- Infinity) are removed.

1.1.2 Metadata

The new statistical analysis and visualization tools make extensive use of the metadata system. For example, metadata are required when defining the experimental design in the Differential Expression for RNA-Seq tool, and can be used to add extra layers of insight in the Create Heat Map and PCA for RNA-Seq tools.

To get the most out of these tools we recommend that all input expression tracks have associated metadata, as shown in figure 1.2. For information about how to use and setup metadata, please see http://clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Metadata.html



Key	Tissue	Spikes
Liver reads	Liver	No
Brain spikes	Brain	Yes
Brain reads	Brain	No
Liver spikes	Liver	Yes

Key	Role	Type	Name	Path
Liver reads	Sample data		Liver reads (GE)	CLC_Data / Test tissues
Brain spikes	Sample data		Brain spikes reads (GE)	CLC_Data / Test tissues
Brain reads	Sample data		Brain reads (GE)	CLC_Data / Test tissues
Liver spikes	Sample data		Liver spikes reads (GE)	CLC_Data / Test tissues

Figure 1.2: *An example of expression tracks with associated metadata.*

1.2 PCA for RNA-Seq

Principal Component Analysis makes it possible to project a high-dimensional dataset (where the number of dimensions equals the number of genes or transcripts) onto two or three dimensions. This helps in identifying outlying samples for quality control, and gives a feeling for the principal causes of variation in a dataset. The analysis proceeds by transforming a large set of variables (in this case, the counts for each individual gene or transcript) to a smaller set of orthogonal principal components. The first principal component specifies the direction with the largest variability in the data, the second component is the direction with the second largest variation, and so on.

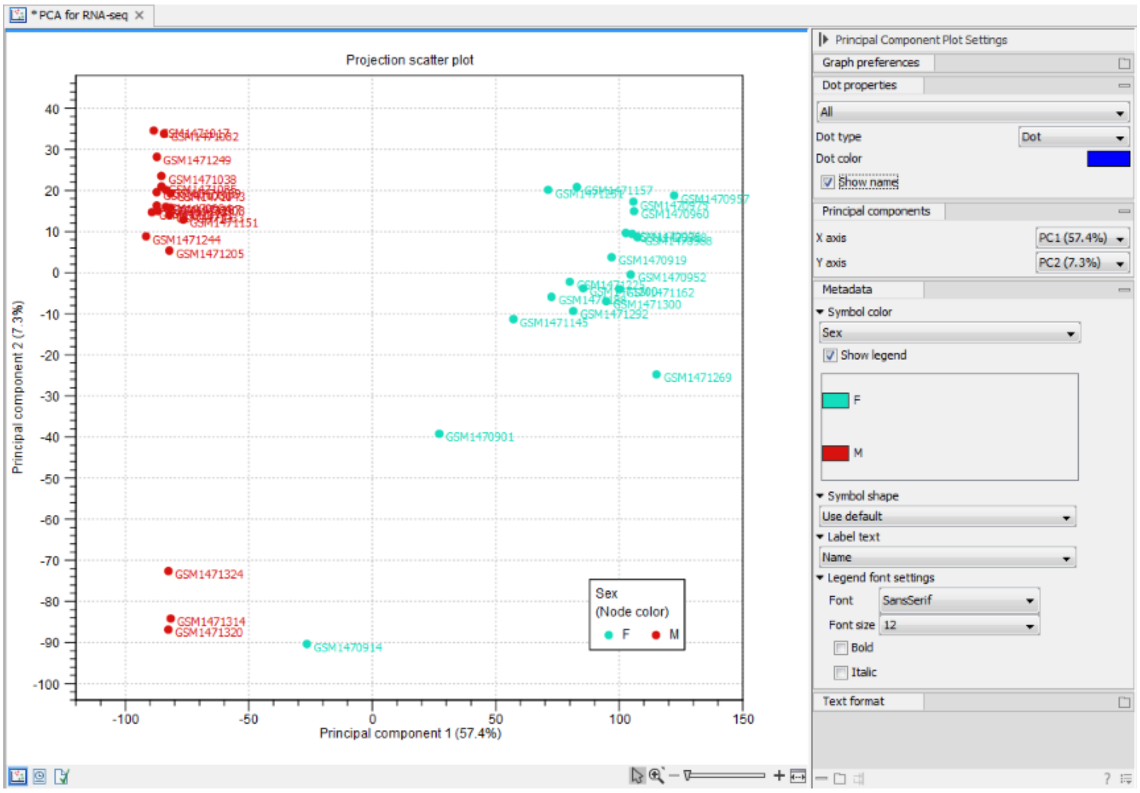
To start the analysis:

Toolbox | **Transcriptomics Analysis** ( )| **RNA-Seq Analysis** | **PCA for RNA-Seq** ( )

Select a number of expression tracks ( ) and click **Next**.

1.2.1 Principal component analysis plot (2D)

The default view is a two-dimensional principal component plot as shown in figure 1.3.



Principal Component Plot Settings

Graph preferences

Dot properties

All

Dot type: Dot

Dot color: 🟦

☑ Show name

Principal components

X axis: PC1 (57.4%)

Y axis: PC2 (7.3%)

Metadata

Symbol color

Sex

☑ Show legend

🟩 F

🟥 M

Symbol shape

Use default

Label text

Name

Legend font settings

Font: SansSerif

Font size: 12

☐ Bold

☐ Italic

Text format

Figure 1.3: *A principal component plot.*

The plot shows the projection of the samples onto the two-dimensional space spanned by the first and second principal components of the covariance matrix. The expression levels used as input are normalized log CPM values, see section 1.1.

The view settings can be adjusted using the **Side Panel**. Under **Graph preferences**, you can adjust the general properties of the plot.

- **Lock axes** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame** Shows a frame around the graph.
- **Tick type** Determines whether tick lines should be shown outside or inside the frame.
- **Tick lines at** Choosing Major ticks will show a grid behind the graph.
- **Horizontal axis range** Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.

- **Vertical axis range** Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **y = 0 axis** Draws a line where y = 0 with options for adjusting the line appearance.

Below the general preferences, you find the **Dot properties**:

- **Drop down menu** In this you select the expression tracks to which following choices apply.
- **Dot type** Allows you to choose between different dot types.
- **Dot color** Click the color box to select a color.
- **Show name** This will show a label with the name of the sample next to the dot.

Note that the Dot properties may be overridden when the Metadata options are used to control the visual appearance (see below).

The **Principal Components** group determines which two principal components are used in the 2D plot. By default, the first principal component is shown for the X axis and the second principal component is shown for the Y axis. The value after the principal component identifier (for example "PC1 (72.5 %)") displays the amount of variance explained by this particular principal component.

The **Metadata** group allows metadata associated with the Expression tracks to be visualized in a number of ways:

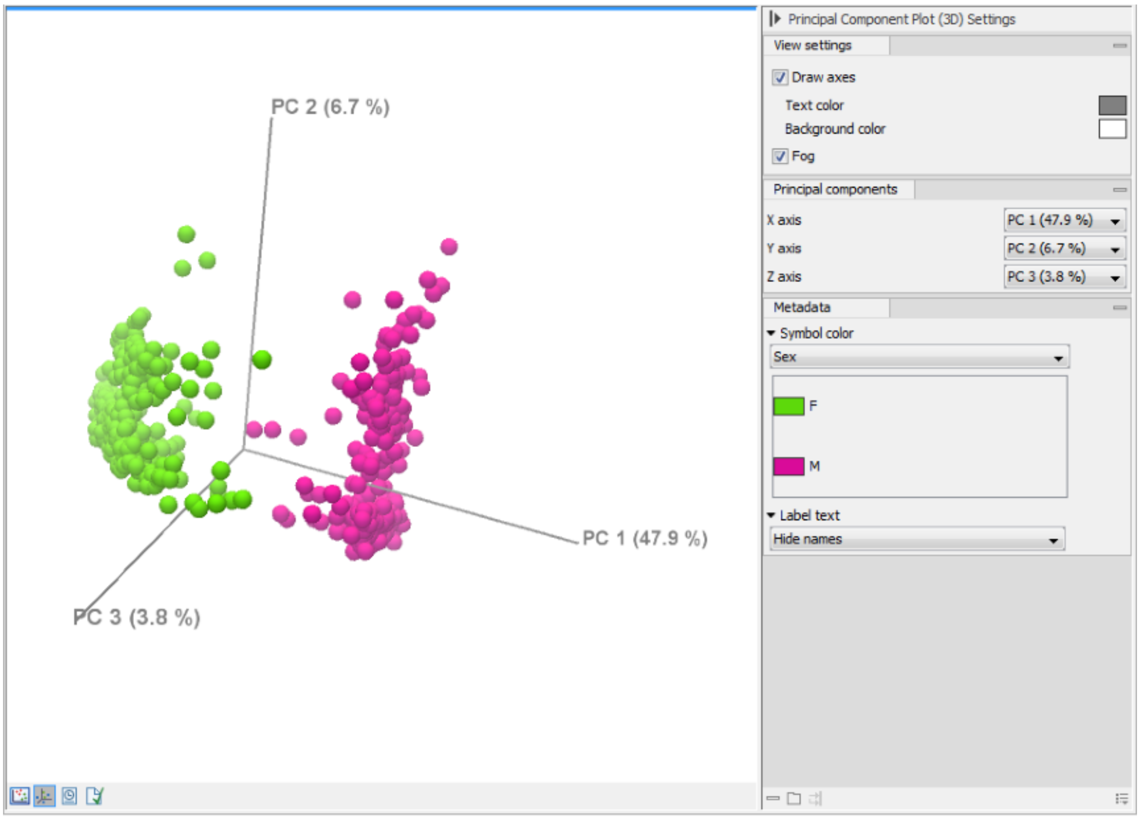
- **Symbol color** Colors are assigned based on a categorical factor in the metadata table.
- **Symbol shape** Shape is assigned based on a categorical factor in the metadata table.
- **Label text** Dots are labeled according to the values in a given metadata column.
- **Legend font settings** contains options to adjust the display of labels.

The graph and axes titles can be edited simply by clicking them with the mouse. These changes will be saved when you **Save** (⇤) the graph - whereas the changes in the **Side Panel** need to be saved explicitly (http://clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Saving_removing_applying_saved_settings.html).

1.2.2 Principal component analysis plot (3D)

The principal component plot may also be displayed in 3D. The 3D view is accessible through the view buttons at the bottom of the panel.

> *Notice that the 3D PCA rendering feature requires a graphics card capable of supporting OpenGL 2.0. Please make sure the latest driver for the graphics card is installed. Indirect rendering (such as x11 forwarding through ssh), remote desktop connection/VNC, and running in virtual machines is not supported.*



Principal Component Plot (3D) Settings

View settings

☑ Draw axes

Text color

Background color

☑ Fog

Principal components

X axis PC 1 (47.9 %)

Y axis PC 2 (6.7 %)

Z axis PC 3 (3.8 %)

Metadata

Symbol color

Sex

🟩 F

🟪 M

Label text

Hide names

Figure 1.4: *A principal component 3D plot.*

The 3D view may be rotated by dragging on the view with the left mouse button pressed. It is possible to pan the view by dragging with the right mouse button pressed. Zooming can be done either using the mouse scroll wheel, or by dragging with both left and right mouse button pressed. It is also possible to center and zoom to a sample simply by clicking on it. Clicking outside any sample (or clicking with the right mouse button) restores the zoom and centering.

The **Side Panel** offers a number of options to change the appearance of the 3D principal component plot:

The **View settings** group makes it possible to toggle the coordinate system on and off, and adjust the text and background color. It is also possible to enable **Fog**, which dims distant objects in order to improve the depth perception.

The **Principal Components** group determines which principal components are used in the 3D plot. The value after the principal component identifier (for example "PC 1 (72.5 %)") displays the amount of variance explained by this particular principal component.

The **Metadata** group allows metadata associated with the Expression tracks to be visualized using color or as text:

- **Symbol color** Colors are assigned based on a categorical factor in the metadata table.
- **Label text** Samples are labeled according to the values in a given metadata column. If 'Show names' is selected, the samples will be labeled according to their name (as shown in the **Navigation Area**).

To save the current view as an image, press the **Graphics** button in the Workbench toolbar. Next, select the location where you wish to save the image, select file format (PNG, JPEG, or TIFF), and provide a name, if you wish to use another name than the default name.

It is possible to save the current view settings (including camera settings) using the **Side Panel** view settings options, see `http://clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Saving_removing_applying_saved_settings.html`.

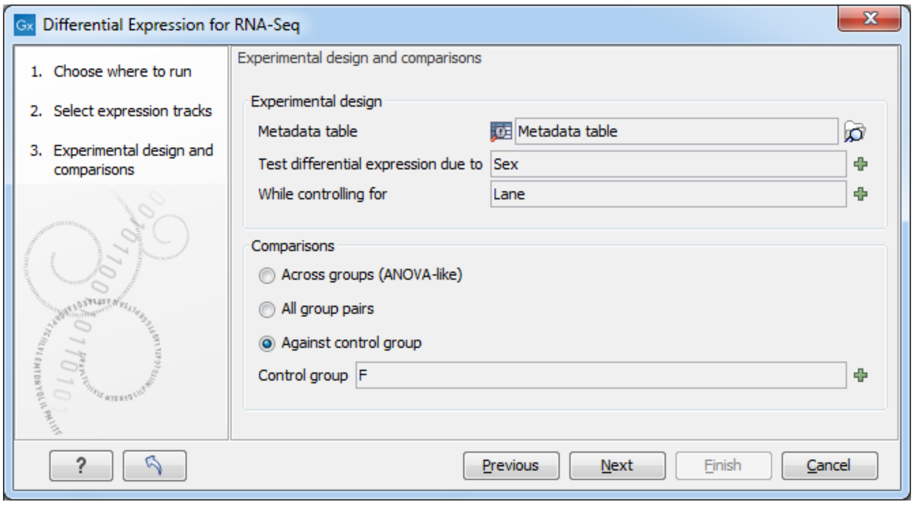
1.3 Differential Expression for RNA-Seq

The **Differential Expression for RNA-Seq tool** performs a statistical differential expression test for a set of Expression Tracks. The statistical analysis is described in more detail in section 1.3.1.

To run the Differential Expression for RNA-Seq analysis:

> **Toolbox | Transcriptomics Analysis ()| RNA-Seq Analysis | Differential Expression for RNA-Seq**

Select a number of Expression tracks () and click **Next**.



Differential Expression for RNA-Seq

1. Choose where to run
2. Select expression tracks
3. Experimental design and comparisons

Experimental design and comparisons

Experimental design

Metadata table: Metadata table

Test differential expression due to: Sex

While controlling for: Lane

Comparisons

○ Across groups (ANOVA-like)

○ All group pairs

● Against control group

Control group: F

Figure 1.5: *Setting up the experimental design and comparisons*

This will display the wizard shown in figure 1.5.

In the **Experimental design** panel, a Metadata table must be selected that describes the factors and groups for all the samples.

- **Metadata table** The metadata table describing the factors for the selected Expression tracks.
- **Test differential expression due to** Specify the one factor differential expression is tested for.
- **While controlling for** Specify confounding factors, i.e., factors that are not of primary interest, but may affect gene expression.

The **Comparisons** panel determines the number and type of statistical comparison tracks output by the tool (see section 1.3.2 for more details).

How many replicates do I need? The Differential Expression for RNA-Seq tool is capable of running without replicates, but this is not recommended and the results should be treated with caution. In general it is desirable to have as many biological replicates as possible – typically at least 3. Replication is important in that it allows the 'within group' variation to be accurately estimated for a gene. In the absence of replication, the Differential Expression for RNA-Seq tool assumes that genes with similar average expression levels have similar variability.

Technical or biological replicates? [Auer and Doerge, 2010] illustrates the importance of *biological replicates* with the example of an alien visiting Earth. The alien wishes to know if men are taller than women. It abducts one man and one woman, and measures their heights several times i.e. performs several *technical replicates*. However, in the absence of *biological replicates*, the alien would erroneously conclude that women are taller than men if this was the case in the two abducted individuals.

1.3.1 The statistical model

Each gene is modeled by a separate Generalized Linear Model (GLM). The use of the GLM formalism allows us to fit curves to expression values without assuming that the error on the values is normally distributed. Similarly to EdgeR and DESeq, we assume that the read counts follow a Negative Binomial distribution.

The Negative Binomial distribution can be understood as a 'Gamma-Poisson' mixture distribution i.e. the distribution resulting from a mixture of Poisson distributions, where the Poisson parameter λ is itself Gamma-distributed. In an RNA-Seq context, this Gamma distribution is controlled by the **dispersion** parameter, such that the Negative Binomial distribution reduces to a Poisson distribution when the dispersion is zero.

Fitting a GLM to expression data

It is easiest to understand how the GLM model works through an example. Imagine an experiment looking at the effect of two drug treatments while controlling for the gender of a patient:

- **Test differential expression due to** Treatment with three groups: drugA, drugB, placebo
- **While controlling for** Gender with groups: Male, Female

In an abuse of mathematical notation, the underlying GLM for each gene looks like

$$\log y_i = (\mathrm{placebo\ and\ Male}) + \mathrm{drugA} + \mathrm{drugB} + \mathrm{Female} + \mathrm{constant_i} \tag{1.1}$$

where y_i is the expression level for the gene in sample i; the combined term $(\mathrm{placebo\ and\ Male})$ describes an arbitrarily chosen baseline expression level (of males being given a placebo); and the other terms drugA, drugB and Female are numbers describing the effect of each group with respect to this baseline. The $\mathrm{constant_i}$ accounts for differences in the library size between samples. For example, if a patient is male and given a placebo we predict the expression level to be

$$\log y_i = (\text{placebo and Male}) + \text{constant}_\text{i}.$$

If instead he had been given drug B, we would predict the expression level y_i to be augmented with the drugB coefficient, resulting in

$$\log y_i = (\text{placebo and Male}) + \text{drugB} + \text{constant}_\text{i}.$$

We assume that the expression levels y_i follow a Negative Binomial distribution. This distribution has a free parameter, the dispersion. The greater the dispersion, the greater the variation in expression levels for a gene.

The most likely values of the dispersion and coefficients, drugA, drugB and Female, are determined simultaneously by fitting the GLM to the data. To see why this simultaneous fitting is necessary, imagine an experiment where we observe counts {3,10,4} for Males and {30,20,8} for Females. The most natural fit is for the coefficient Female to have a two-fold change and for the dispersion to be small, but an alternative fit has no fold change and a larger dispersion. Under this second fit the variation in the counts is greater, and it is just by chance that all three Female values are larger than all three Male values.

Refining the estimate of dispersion

Much research has gone into refining the dispersion estimates of GLM fits. One important observation is that the GLM dispersion for a gene is often too low, because it is a *sample* dispersion rather than a *population* dispersion. We correct for this using the Cox-Reid adjusted likelihood, as in the multi-factorial EdgeR method [Robinson et al., 2010]. [1]

A second observation that can be used to improve the dispersion estimate, is that genes with the same average expression often have similar dispersions. To make use of this observation, we follow [Robinson et al., 2010] in estimating genewise dispersions from a linear combination of the likelihood for the gene of interest and neighboring genes with similar average expression levels. The weighting in this combination depends on the number of samples in an experiment, such that the neighbors have most weight when there are no replicates, and little effect when the number of replicates is high.

Statistical testing

The final GLM fit and dispersion estimate allows us to calculate the total likelihood of the model given the data, and the uncertainty on each fitted coefficient. The two statistical tests each make use of one of these values.

Wald test Tests whether a given coefficient is non-zero. This test is used in the **All group pairs** and **Against control group** comparisons. For example, to test whether there is a difference between patients treated with a placebo, and those treated with drugB, we would use the Wald test to determine if the drugB coefficient is non-zero.

[1] To understand the purpose of the correction, it may help to consider the analogous situation of calculation of the variance of normally distributed measurements. One approach would be to calculate $\frac{1}{n}\sum(x_i - \overline{x})^2$, but this is the *sample* variance and often too low. A commonly used correction for the *population* variance is: $\frac{1}{n-1}\sum(x_i - \overline{x})^2$.

Likelihood Ratio test Fits two GLMs, one with the given coefficients and one without. The more important the coefficients are, the greater the ratio of the likelihoods of the two models. This test is used in the Across groups (ANOVA-like) comparison. If we wanted to test whether either drug had an effect, we would compare the likelihoods of the GLM described in equation 1.1 with those in the reduced GLM $\log y_i = (\mathrm{Male}) + \mathrm{Female} + \mathrm{constant_i}$.

1.3.2 Output of the Differential Expression for RNA-Seq tool

The Differential Expression for RNA-Seq tool produces different numbers and types of statistical comparison tracks depending on the settings of the **Comparisons** panel. Depending on the choice either a Wald test or a Likelihood Ratio test is used. For example, assume that we test a factor called ‘Tissue’ with three groups: skin, liver, brain.

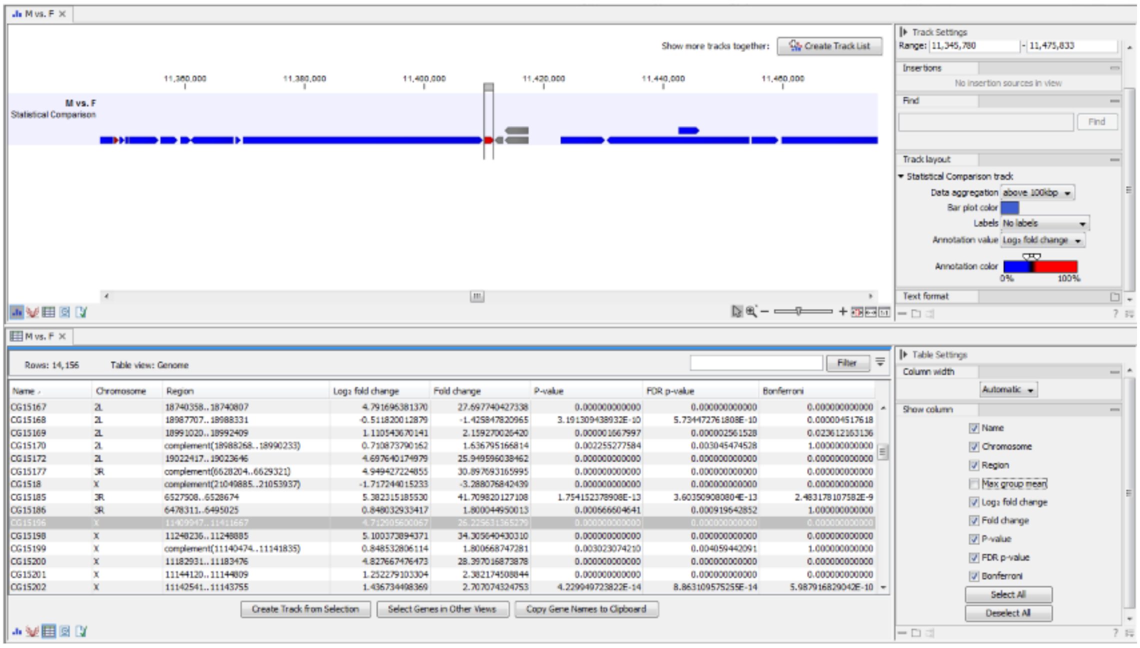
- **Across groups (ANOVA-like)** This mode tests for the effect of a factor across all groups.
  - *Outputs produced*: "Due to Tissue"
  - *Test used*: Likelihood ratio test
  - *Fold change* reports: The maximum pairwise fold change between any two of the three tissue types.
  - *Max of group means* reports: The maximum of the average group RPKM values among any of the tissue types for a gene.
- **All group pairs** tests for differences between all pairs of groups in a factor.
  - *Outputs produced*: "skin vs. liver", "skin vs. brain", "liver vs. brain"
  - *Test used*: Wald test
  - *Fold change* reports: The fold change in the defined order between the named pair of tissue types.
  - *Max of group means* reports: The maximum of the average group RPKM values between the two named tissue types.
- **Against control group** This mode tests for differences between all the groups in a factor and the named reference group. In this example the reference group is skin.
  - *Outputs produced*: "liver vs. skin", "brain vs. skin"
  - *Test used*: Wald test
  - *Fold change* reports: The fold change in the defined order between the named pair of tissue types.
  - *Max of group means* reports: The maximum of the average group RPKM values between the two named tissue types.

Note: Fold changes are calculated from the GLM, which corrects for differences in library size between the samples and the effects of confounding factors. It is therefore not possible to derive these fold changes from the original counts by simple algebraic calculations.

1.3.3 Statistical comparison tracks

The Differential Expression for RNA-Seq tool will output one or more statistical comparison tracks.

An example of a statistical comparison track is shown in figure 1.6. Statistical comparison tracks make it possible to show differential expression data alongside other kinds of tracks in a genomic context.



Track Settings
Range: 11,345,780 - 11,475,833
Insertions: No insertion sources in view
Track layout
Statistical Comparison track
Data aggregation: above 100kbp
Bar plot color
Labels: No labels
Annotation value: Log₂ fold change
Annotation color: 0% 100%

Rows: 14,156 Table view: Genome

Name	Chromosome	Region	Log₂ fold change	Fold change	P-value	FDR p-value	Bonferroni
CG15167	2L	18740358..18740807	4.791696381370	27.697740427338	0.000000000000	0.000000000000	0.000000000000
CG15168	2L	18987707..18988331	-0.511820012879	-1.425847820965	3.191309438932E-10	5.734472761808E-10	0.000004517618
CG15169	2L	18991020..18992409	1.110543670141	2.159270026420	0.000001667997	0.000002561528	0.023612163136
CG15170	2L	complement(18988268..18990233)	0.710873790162	1.636795166814	0.002255277584	0.003045474528	1.000000000000
CG15172	2L	19022417..19023646	4.697640174979	25.949596038462	0.000000000000	0.000000000000	0.000000000000
CG15177	3R	complement(6628204..6629321)	4.949427224855	30.897693165995	0.000000000000	0.000000000000	0.000000000000
CG1518	X	complement(21049885..21053937)	-1.717244015233	-3.288076842439	0.000000000000	0.000000000000	0.000000000000
CG15185	3R	6527508..6528674	5.382315185530	41.709820127108	1.754152378908E-13	3.603509080804E-13	2.483178107582E-9
CG15186	3R	6478311..6495025	0.848032933417	1.800044950013	0.000666604641	0.000919642852	1.000000000000
CG15196	X	11409947..11411667	4.712905600067	25.225631365279	0.000000000000	0.000000000000	0.000000000000
CG15198	X	11248236..11248885	5.100373894371	34.305640430310	0.000000000000	0.000000000000	0.000000000000
CG15199	X	complement(11140474..11141835)	0.848532806114	1.800668747281	0.003023074210	0.004059442091	1.000000000000
CG15200	X	11182931..11183476	4.827667476473	28.397016873878	0.000000000000	0.000000000000	0.000000000000
CG15201	X	11144120..11144809	1.252279103304	2.382174508844	0.000000000000	0.000000000000	0.000000000000
CG15202	X	11142541..11143755	1.436734498369	2.707074324753	4.229949723822E-14	8.863109575255E-14	5.987916829042E-10

Create Track from Selection | Select Genes in Other Views | Copy Gene Names to Clipboard

Table Settings
Column width: Automatic
Show column
☑ Name
☑ Chromosome
☑ Region
□ Max group mean
☑ Log₂ fold change
☑ Fold change
☑ P-value
☑ FDR p-value
☑ Bonferroni
Select All
Deselect All

Figure 1.6: *Statistical comparison track view.*

The track layout of the statistical comparison track can be customized as follows:

- **Data aggregation** Allows you to specify whether the information in the track should be shown in detail or whether you wish to aggregate data. By aggregating data you decrease the detail level, but increase the speed of the data display process, which is of particular interest when working with big data sets. The threshold (in bp) for when data should be aggregated can be specified with the drop-down box. The threshold describes the unit (or "bucket") size in base pairs, above which the data will start being aggregated. The bucket size depends on the track length and the zoom level. Hence, a data aggregation threshold with a low value will only show details when zoomed in, whereas a high value means that you can see details even when zoomed out. Please note that when using the high values, it will take longer time to display the data on the screen.
- **Bar plot color** Selects the color of aggregated data.
- **Labels** Determines where the gene name should be shown.
- **Annotation value** The value that is graphically shown in detail view:
  - **Max group means** For each group in the statistical comparison, the average RPKM is calculated. This value is the maximum of the average RPKM's.
  - **Log$_2$ fold change** The logarithmic fold change.

  - **Fold change** The (signed) fold change. Genes/transcripts that are not observed in any sample have undefined fold changes and are reported as NaN (not a number).
  - **P-value** Standard p-value. Genes/transcripts that are not observed in any sample have undefined p-values and are reported as NaN (not a number).
  - **FDR p-value** The false discovery rate corrected p-value.
  - **Bonferroni** The Bonferroni corrected p-value.

- **Annotation color** Determines how the annotation value is mapped onto a color.

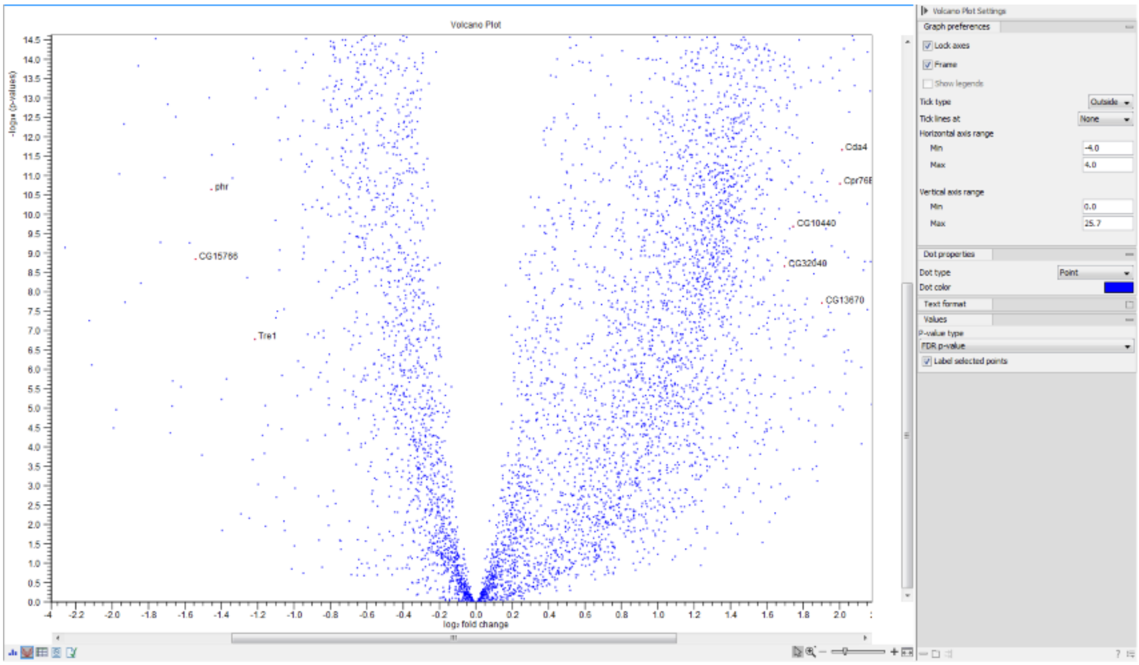
The expression track table view has three button.

- The "Create track from Selection" will create a Track using selected rows.
- The "Select Genes in Other Views" button finds and selects the currently selected genes and transcripts in all other open expression track table views.
- The "Copy Gene Names to Clipboard" button copies the currently selected gene names to the clipboard.

1.3.4 The volcano plot

Statistical comparison tracks also offer a volcano plot view.

An example of a volcano plot is shown in figure 1.7.



Volcano Plot Settings

Graph preferences

☑ Lock axes

☑ Frame

□ Show legends

Tick type: Outside

Tick lines at: None

Horizontal axis range
Min: -4.0
Max: 4.0

Vertical axis range
Min: 0.0
Max: 25.7

Dot properties

Dot type: Point

Dot color

Text format

Values

P-value type: FDR p-value

☑ Label selected points

Figure 1.7: *Volcano plot.*

The volcano plot shows the relationship between the p-values of a statistical test and the fold changes among the samples. The $\log_2$ fold changes are plotted on the x-axis, and the $-\log_{10}$ p-values are plotted on the y-axis. Features of interest are typically those in the upper left and right hand corners of the volcano plot, as these have large fold changes (lie far from $x = 0$) and are statistically significant (have large y-values). It is possible to change the type of p-value from the side panel (see below).

The view settings can be adjusted using the **Side Panel**. Under **Graph preferences**, you can adjust the general properties of the volcano plot

- **Lock axes** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame** Shows a frame around the graph.
- **Show legends** Shows the data legends.
- **Tick type** Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Horizontal axis range** Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range** Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.

Below the general preferences, you find the **Dot properties** and **Text format**, where you can adjust the coloring and appearance of the dots and text.

At the bottom are options for choosing which values to display:

- **P-value type** Selects which type of p-value to use.
- **Label selected points** Chooses whether selected points should be labeled.

Note that if you wish to use the same settings next time you open a volcano plot, you need to save the settings of the **Side Panel** (see http://clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Saving_removing_applying_saved_settings.html).

1.4 Create Heat Map for RNA-Seq

The **Create Heat Map tool** shows a two dimensional heat map of expression values. Each column corresponds to one sample, and each row corresponds to a feature (a gene or a transcript). The samples and features are both hierarchically clustered.

1.4.1 Clustering of features and samples

The hierarchical clustering clusters features by the similarity of their expression profiles over the set of samples. It clusters samples by the similarity of expression patterns over their features.

Each clustering has a tree structure that is generated by

1. Letting each feature or sample be a cluster.
2. Calculating pairwise distances between all clusters.
3. Joining the two closest clusters into one new cluster.
4. Iterating 2-3 until there is only one cluster left (which contains all the features or samples).

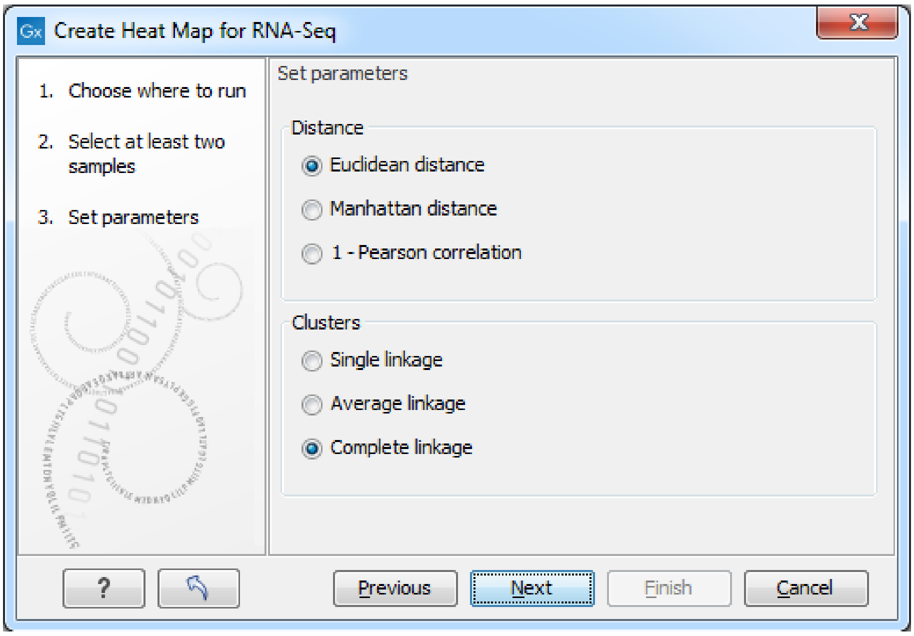
The tree is drawn so that the distances between clusters are reflected by the lengths of the branches in the tree.

To create a heat map:

> **Toolbox** | **Transcriptomics Analysis** ()| **RNA-Seq Analysis** | **Create Heat Map for RNA-Seq** ()

Select at least two expression tracks () and click **Next**.

This will display the wizard shown in figure 1.8. The hierarchical clustering algorithm requires that you specify a distance measure and a cluster linkage. The distance measure is used to specify how distances between two features or samples should be calculated. The cluster linkage specifies how the distance between two clusters, each consisting of a number of features or samples, should be calculated.



Set parameters

Distance
- ● Euclidean distance
- ○ Manhattan distance
- ○ 1 - Pearson correlation

Clusters
- ○ Single linkage
- ○ Average linkage
- ● Complete linkage

Figure 1.8: *Parameters for Create Heat Map.*

There are three kinds of **Distance measures**:

- **Euclidean distance** The ordinary distance between two points - the length of the segment connecting them. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Euclidean distance between u and v is

$$|u - v| = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}.$$

- **Manhattan distance** The Manhattan distance between two points is the distance measured along axes at right angles. If $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, then the Manhattan distance between u and v is

$$|u - v| = \sum_{i=1}^{n} |u_i - v_i|.$$

- **1 - Pearson correlation** The Pearson correlation coefficient between two elements $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ is defined as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{x_i - \overline{x}}{s_x})(\frac{y_i - \overline{y}}{s_y})$$

  where $\overline{x}/\overline{y}$ is the average of values in x/y and s_x/s_y is the sample standard deviation of these values. It takes a value $\in [-1, 1]$. Highly correlated elements have a high absolute value of the Pearson correlation, and elements whose values are un-informative about each other have Pearson correlation 0. Using $1 - |Pearsoncorrelation|$ as distance measure means that elements that are highly correlated will have a short distance between them, and elements that have low correlation will be more distant from each other.

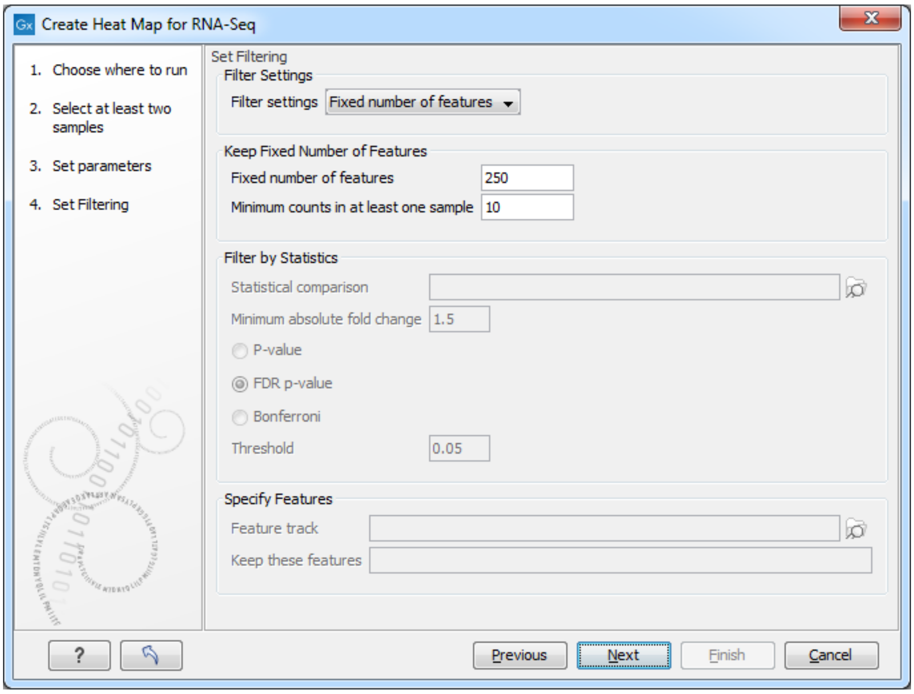
The possible cluster linkages are:

- **Single linkage** The distance between two clusters is computed as the distance between the two closest elements in the two clusters.
- **Average linkage** The distance between two clusters is computed as the average distance between objects from the first cluster and objects from the second cluster. The averaging is performed over all pairs (x, y), where x is an object from the first cluster and y is an object from the second cluster.
- **Complete linkage** The distance between two clusters is computed as the maximal object-to-object distance $d(x_i, y_j)$, where x_i comes from the first cluster, and y_j comes from the second cluster. In other words, the distance between two clusters is computed as the distance between the two farthest objects in the two clusters.

After having selected the distance measure, click **Next** to set up the feature filtering options as shown in figure 1.9.

Genomes usually contain too many features to allow for a meaningful visualization of all genes or transcripts. Clustering hundreds of thousands of features is also very time consuming. Therefore it is recommend to reduce the number of features before clustering and visualization.

There are several different **Filter settings** to filter genes or transcripts:



Create Heat Map for RNA-Seq

1. Choose where to run
2. Select at least two samples
3. Set parameters
4. Set Filtering

Set Filtering

Filter Settings

Filter settings: Fixed number of features

Keep Fixed Number of Features

Fixed number of features: 250

Minimum counts in at least one sample: 10

Filter by Statistics

Statistical comparison:

Minimum absolute fold change: 1.5

○ P-value

● FDR p-value

○ Bonferroni

Threshold: 0.05

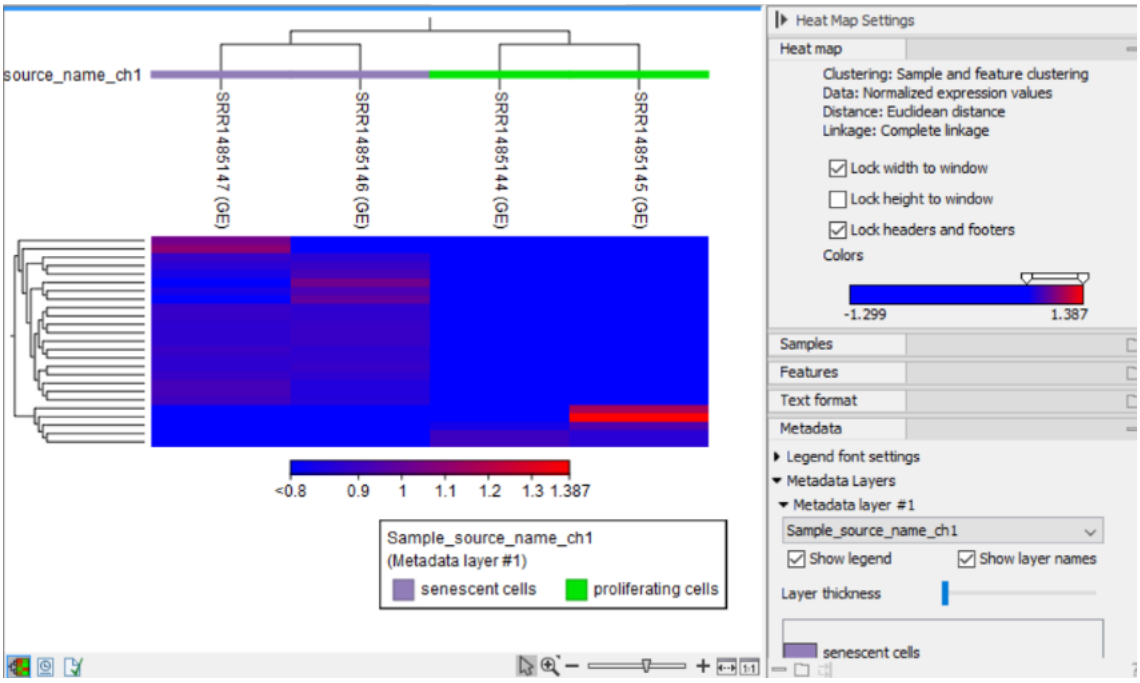
Specify Features

Feature track:

Keep these features:

Figure 1.9: *Feature filtering for Create Heat Map.*

- **No filtering** Keeps all features.
- **Fixed number of features**
  - **Fixed number of features** The given number of features with the highest coefficient of variation (the ratio of the standard deviation to the mean) are kept.
  - **Minimum counts in at least one sample** Only features with more than this number of counts in at least one sample will be taken into account. Notice that the counts are raw, un-normalized values.
- **Filter by statistics** Keeps features that are differentially expressed according to the specified cut-offs.
  - **Statistical comparison** A single statistical comparison track output by the Differential Expression for RNA-Seq tool.
  - **Minimum absolute fold change** Only features with a higher absolute fold change are kept.
  - **Threshold** Only features with a lower p-value are kept. It is possible to select which type of p-value to use.
- **Specify features** Keeps a set of features, as specified by either a feature track or by plain text.
  - **Feature track** Any genes or transcripts defined in the feature track will be kept.
  - **Keep these features** A plain text list of feature names. Any white-space characters, and ",", and ";" are accepted as separators.



Heat Map Settings

Heat map

Clustering: Sample and feature clustering
Data: Normalized expression values
Distance: Euclidean distance
Linkage: Complete linkage

☑ Lock width to window

□ Lock height to window

☑ Lock headers and footers

Colors

-1.299 1.387

Samples

Features

Text format

Metadata

Legend font settings

Metadata Layers

Metadata layer #1

Sample_source_name_ch1

☑ Show legend ☑ Show layer names

Layer thickness

senescent cells

Figure 1.10: *The 2D heat map.*

1.4.2 The heat map view

After the tool completes, a heat map like the one shown in (figure 1.10) is produced. In the heat map each row corresponds to a feature and each column to a sample. The color in the i'th row and j'th column reflects the expression level of feature i in sample j (the color scale can be set in the side panel). The expression values used are normalized log CPM values, see section 1.1.

There are a number of options to change the appearance of the heat map. At the top of the **Side Panel**, you find the **Heat map** group (see figure 1.10).

- **Lock width to window** When you zoom in the heat map, you will per default only zoom in on the vertical level. This is because the width of the heat map is locked to the window. If you uncheck this option, you will zoom both vertically and horizontally. Since you normally have more features than samples, it is useful to lock the width since you then have all the samples in view all the time.

- **Lock height to window** This is the corresponding option for the height. Note that if you check both options, you will not be able to zoom at all, since both the width and the height are fixed.

- **Lock headers and footers** This will ensure that you are always able to see the sample and feature names and the trees when you zoom in.

- **Colors** The expression levels are visualized using a gradient color scheme, where the right side color is used for high expression levels and the left side color is used for low expression levels. You can change the coloring by clicking the box, and you can change the relative coloring of the values by dragging the two knobs on the white slider above.

Below you find the **Samples** and **Features** groups. They contain options to show names, color legends, and trees above or below the heat map. The tree options also control the **Tree size**, including the option of showing the full tree, no matter how much space it will use.

The **Metadata** group makes it possible to visualize metadata associated with the Expression tracks:

- **Legend font settings** adjusts the label settings.
- **Metadata layers** Adds a color bar to the hierarchical sample tree, colored according to the value of a chosen metadata table column.

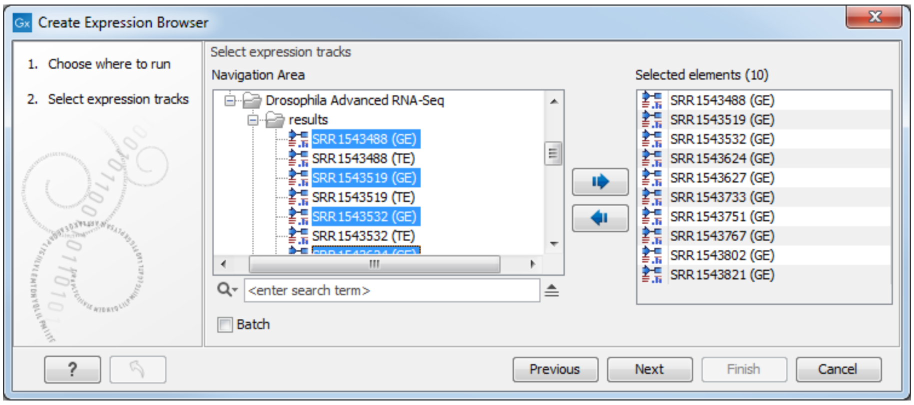
1.5 Create Expression Browser

The **Create Expression Browser** tool makes it possible to inspect gene and transcript expression level counts and statistics for many samples at the same time.

To run the tool:

> **Toolbox | Transcriptomics Analysis ( )| RNA-Seq Analysis | Create Expression Browser**

Select a number of expression tracks ( ), either Gene level (GE) or Transcript level Expression (TE) tracks but not a combination of both, and click **Next** (see figure 1.11).



Create Expression Browser

1. Choose where to run
2. Select expression tracks

Select expression tracks

Navigation Area

- Drosophila Advanced RNA-Seq
  - results
    - SRR1543488 (GE)
    - SRR1543488 (TE)
    - SRR1543519 (GE)
    - SRR1543519 (TE)
    - SRR1543532 (GE)
    - SRR1543532 (TE)

Selected elements (10)

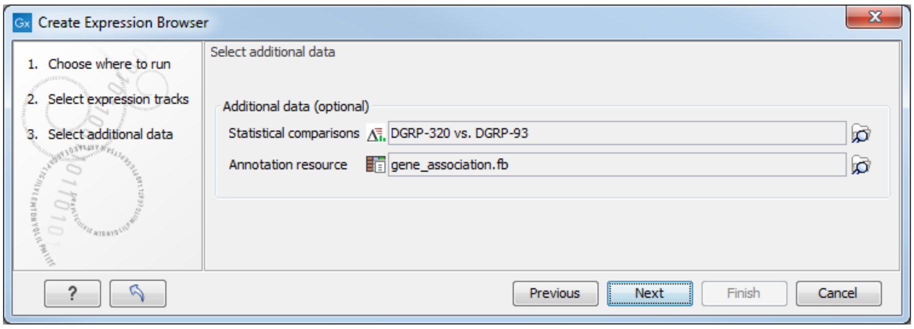
- SRR1543488 (GE)
- SRR1543519 (GE)
- SRR1543532 (GE)
- SRR1543624 (GE)
- SRR1543627 (GE)
- SRR1543733 (GE)
- SRR1543751 (GE)
- SRR1543767 (GE)
- SRR1543802 (GE)
- SRR1543821 (GE)

<enter search term>

□ Batch

Figure 1.11: *Select expression tracks, either GE or TE.*

In the second wizard dialog (see figure 1.12), you can decide to add one or more statistical comparisons or an annotation source.



Create Expression Browser

1. Choose where to run
2. Select expression tracks
3. Select additional data

Select additional data

Additional data (optional)

Statistical comparisons: DGRP-320 vs. DGRP-93

Annotation resource: gene_association.fb

Figure 1.12: *It is optional to provide statistical comparisons and annotation resources.*

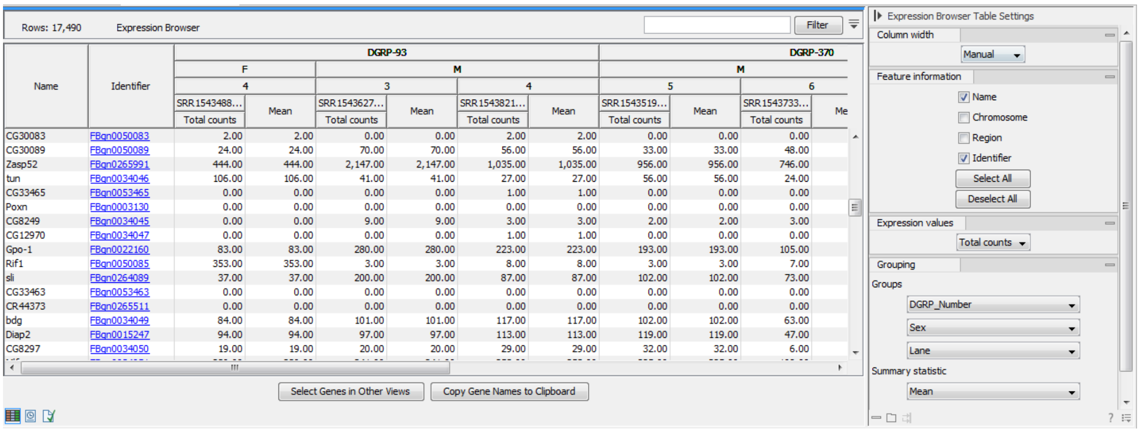
Statistical comparisons are generated by the tool 'Differential Expression for RNA-Seq'. You can only choose a comparison that was generated by the same kind of expression tracks as the ones you selected in the previous step. This means that when you want to create an expression browser for GE expression levels, you can only input in the second optional step a statistical comparison generated using GE tracks as well.

Annotation resources can be obtained from different sources:

- In the majority of cases, annotations are obtained from Gene Ontology consortium at http://geneontology.org/page/download-annotations. Download the *.gz file of your choice from the website on to your computer, and import it in the workbench using the Standard Import tool.
- You may have your own annotation data in a spreadsheet. Learn how to import such files here. http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Generic_expression_annotation_data_file_formats.html.
- Annotations can be generated by the "Blast2GO PRO" plugin. See http://www.clcbio.com/clc-plugin/blast2go-pro/ for more information.
- Some annotations for human, mouse and rat are also bundled together with Reference Data Sets provided in Biomedical Genomics Workbench.

1.5.1 The expression browser

An Expression Browser is shown in figure 1.13.



Rows: 17,490 Expression Browser

		DGRP-93						DGRP-370				
		F		M				M				
Name	Identifier	4		3		4		5		6		
		SRR1543488...	Mean	SRR1543627...	Mean	SRR1543821...	Mean	SRR1543519...	Mean	SRR1543733...		Me
		Total counts		Total counts		Total counts		Total counts		Total counts		
CG30083	FBgn0050083	2.00	2.00	0.00	0.00	2.00	2.00	0.00	0.00	0.00	
CG30089	FBgn0050089	24.00	24.00	70.00	70.00	56.00	56.00	33.00	33.00	48.00	
Zasp52	FBgn0265991	444.00	444.00	2,147.00	2,147.00	1,035.00	1,035.00	956.00	956.00	746.00	
tun	FBgn0034046	106.00	106.00	41.00	41.00	27.00	27.00	56.00	56.00	24.00	
CG33465	FBgn0053465	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	
Poxn	FBgn0003130	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
CG8249	FBgn0034045	0.00	0.00	9.00	9.00	3.00	3.00	2.00	2.00	3.00	
CG12970	FBgn0034047	0.00	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	
Gpo-1	FBgn0022160	83.00	83.00	280.00	280.00	223.00	223.00	193.00	193.00	105.00	
Rif1	FBgn0050085	353.00	353.00	3.00	3.00	8.00	8.00	3.00	3.00	7.00	
sli	FBgn0264089	37.00	37.00	200.00	200.00	87.00	87.00	102.00	102.00	73.00	
CG33463	FBgn0053463	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
CR44373	FBgn0265511	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
bdg	FBgn0034049	84.00	84.00	101.00	101.00	117.00	117.00	102.00	102.00	63.00	
Diap2	FBgn0015247	94.00	94.00	97.00	97.00	113.00	113.00	119.00	119.00	47.00	
CG8297	FBgn0034050	19.00	19.00	20.00	20.00	29.00	29.00	32.00	32.00	6.00	

Expression Browser Table Settings

Column width: Manual

Feature information
☑ Name
☐ Chromosome
☐ Region
☑ Identifier

Expression values: Total counts

Grouping

Groups: DGRP_Number, Sex, Lane

Summary statistic: Mean

Figure 1.13: *Expression browser table when no statistical comparison or annotations resources were provided.*

Each row represents a gene or a transcript, defined by its name, the chromosome and the region where it is located, as well as an identifier linking to the relevant online database.

The expression values for each sample - or aggregation of samples - can be given by total counts, RPKM, TPM or CPM (Counts Per Million). These measurements differ from each other in two key ways:

1. RPKM and TPM measure the number of *transcripts* whereas total counts and CPM measure the number of *reads*. The distinction is important because in an RNA-Seq experiment, more reads are typically sequenced from longer transcripts than from shorter ones.

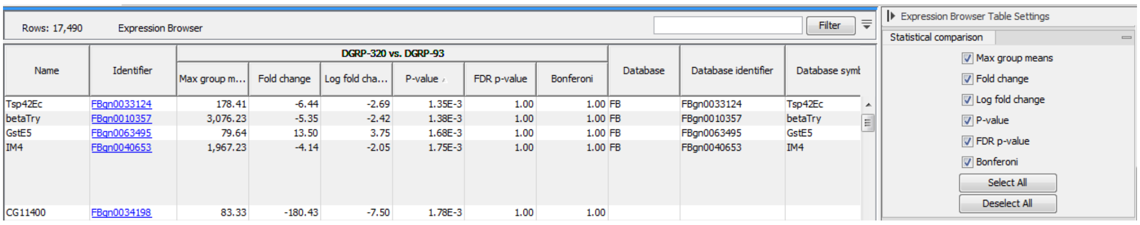
2. RPKM, TPM and CPM are normalized for sequencing-depth so their values are *comparable* between samples. Total counts is not normalized, so values are *not comparable* between samples.

> **How do I get the normalized counts used to calculate fold changes?** The CPM expression values are most comparable to the results of the Differential Expression for RNA-Seq tool. However, normalized counts are not used to calculate fold changes; instead the Differential Expression for RNA-Seq tool works by fitting a statistical model (which accounts for differences in sequencing-depth) to raw counts. It is therefore not possible to derive these fold changes from the CPM values by simple algebraic calculations.

It is possible to display the values for individual samples, or for groups of samples as defined by the metadata. Using the drop down menus in the "Grouping" section of the right-hand side setting panel, you can choose to group samples according to up to three metadata layers as shown in figure 1.13.

When individual samples are aggregated, an additional "summary statistic" column can be displayed to give either the mean, the minimum, or the maximum expression value for each group of samples. The table in figure 1.13 shows the mean of the expression values for the first group layer that was selected.

If one or more statistical comparisons are provided, extra columns can be displayed in the table using the "Statistical comparison" section of the Settings panel (figure 1.14). The columns correspond to the different statistical values generated by the "Differential Expression for RNA-seq" as detailed in section 1.3.3.



Rows: 17,490 Expression Browser

Name	Identifier	DGRP-320 vs. DGRP-93						Database	Database identifier	Database symt
		Max group m...	Fold change	Log fold cha...	P-value	FDR p-value	Bonferoni			
Tsp42Ec	FBgn0033124	178.41	-6.44	-2.69	1.35E-3	1.00	1.00	FB	FBgn0033124	Tsp42Ec
betaTry	FBgn0010357	3,076.23	-5.35	-2.42	1.38E-3	1.00	1.00	FB	FBgn0010357	betaTry
GstE5	FBgn0063495	79.64	13.50	3.75	1.68E-3	1.00	1.00	FB	FBgn0063495	GstE5
IM4	FBgn0040653	1,967.23	-4.14	-2.05	1.75E-3	1.00	1.00	FB	FBgn0040653	IM4
CG11400	FBgn0034198	83.33	-180.43	-7.50	1.78E-3	1.00	1.00			

Expression Browser Table Settings

Statistical comparison

☑ Max group means
☑ Fold change
☑ Log fold change
☑ P-value
☑ FDR p-value
☑ Bonferoni

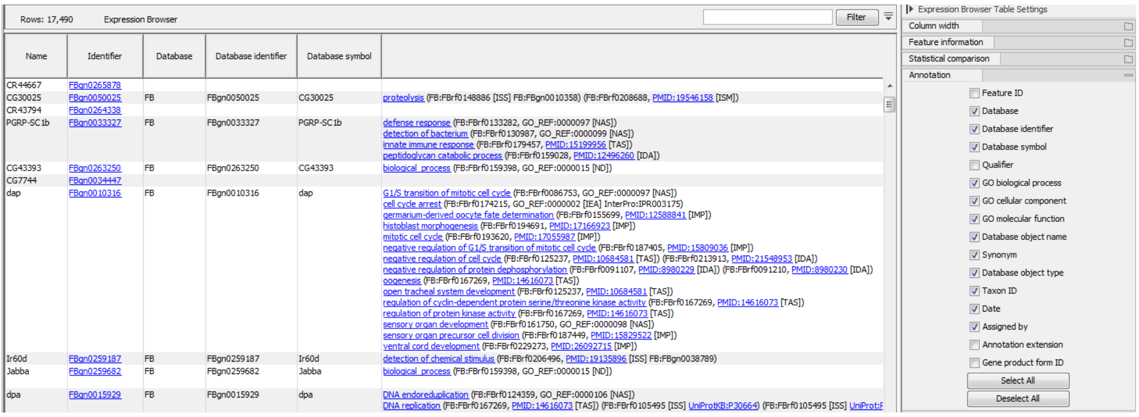
Figure 1.14: *Expression browser table when a statistical comparison is present.*

If an annotation database is provided, extra columns can be displayed in the table using the "Annotation" section of the Settings panel (figure 1.15). Which columns are available depends on the annotation file used. When using a GO annotation file, the GO biological process column will list for each gene or transcript one or several biological processes. Click on the process name to open the corresponding page on the Consortium for Gene Ontology webpage. It is also possible to access additional online information by clicking on the the PMID, RefSeq, HGNC or UniProt accession number when available.

Select the genes of interest and use the button present at the bottom of the table to highlight the genes in other views (volcano plot for instance) or to copy the genes of interest to a clipboard.

1.6 Create Venn Diagram for RNA-Seq

The **Create Venn Diagram tool** makes it possible to compare the overlap of differentially expressed genes or transcripts in two or more statistical comparison tracks. The genes



Rows: 17,490 Expression Browser

Filter

Name	Identifier	Database	Database identifier	Database symbol	
CR44667	FBgn0265878				
CG30025	FBgn0050025	FB	FBgn0050025	CG30025	proteolysis (FB:FBrf0148886 [ISS] FB:FBgn0010358) (FB:FBrf0208688, PMID:19546158 [ISM])
CR43794	FBgn0264338				
PGRP-SC1b	FBgn0033327	FB	FBgn0033327	PGRP-SC1b	defense response (FB:FBrf0133282, GO_REF:0000097 [NAS]) detection of bacterium (FB:FBrf0130987, GO_REF:0000099 [NAS]) innate immune response (FB:FBrf0179457, PMID:15199956 [TAS]) peptidoglycan catabolic process (FB:FBrf0159028, PMID:12496260 [IDA])
CG43393	FBgn0263250	FB	FBgn0263250	CG43393	biological_process (FB:FBrf0159398, GO_REF:0000015 [ND])
CG7744	FBgn0034447				
dap	FBgn0010316	FB	FBgn0010316	dap	G1/S transition of mitotic cell cycle (FB:FBrf0086753, GO_REF:0000097 [NAS]) cell cycle arrest (FB:FBrf0174215, GO_REF:0000002 [IEA] InterPro:IPR003175) germarium-derived oocyte fate determination (FB:FBrf0155699, PMID:12588841 [IMP]) histoblast morphogenesis (FB:FBrf0194691, PMID:17166923 [IMP]) mitotic cell cycle (FB:FBrf0193620, PMID:17055987 [IMP]) negative regulation of G1/S transition of mitotic cell cycle (FB:FBrf0187405, PMID:15809036 [IMP]) negative regulation of cell cycle (FB:FBrf0125237, PMID:10684581 [TAS]) (FB:FBrf0213913, PMID:21548953 [IDA]) negative regulation of protein dephosphorylation (FB:FBrf0091107, PMID:8980229 [IDA]) (FB:FBrf0091210, PMID:8980230 [IDA]) oogenesis (FB:FBrf0167269, PMID:14616073 [TAS]) open tracheal system development (FB:FBrf0125237, PMID:10684581 [TAS]) regulation of cyclin-dependent protein serine/threonine kinase activity (FB:FBrf0167269, PMID:14616073 [TAS]) regulation of protein kinase activity (FB:FBrf0167269, PMID:14616073 [TAS]) sensory organ development (FB:FBrf0161750, GO_REF:0000098 [NAS]) sensory organ precursor cell division (FB:FBrf0187449, PMID:15829522 [IMP]) ventral cord development (FB:FBrf0229273, PMID:26092715 [IMP])
Ir60d	FBgn0259187	FB	FBgn0259187	Ir60d	detection of chemical stimulus (FB:FBrf0206496, PMID:19135896 [ISS] FB:FBgn0038789)
Jabba	FBgn0259682	FB	FBgn0259682	Jabba	biological_process (FB:FBrf0159398, GO_REF:0000015 [ND])
dpa	FBgn0015929	FB	FBgn0015929	dpa	DNA endoreduplication (FB:FBrf0124359, GO_REF:0000106 [NAS]) DNA replication (FB:FBrf0167269, PMID:14616073 [TAS]) (FB:FBrf0105495 [ISS] UniProtKB:P30664) (FB:FBrf0105495 [ISS] UniProt:F

Expression Browser Table Settings

Column width

Feature information

Statistical comparison

Annotation

☐ Feature ID
☑ Database
☑ Database identifier
☑ Database symbol
☐ Qualifier
☑ GO biological process
☑ GO cellular component
☑ GO molecular function
☑ Database object name
☑ Synonym
☑ Database object type
☑ Taxon ID
☑ Date
☑ Assigned by
☐ Annotation extension
☐ Gene product form ID

Select All

Deselect All

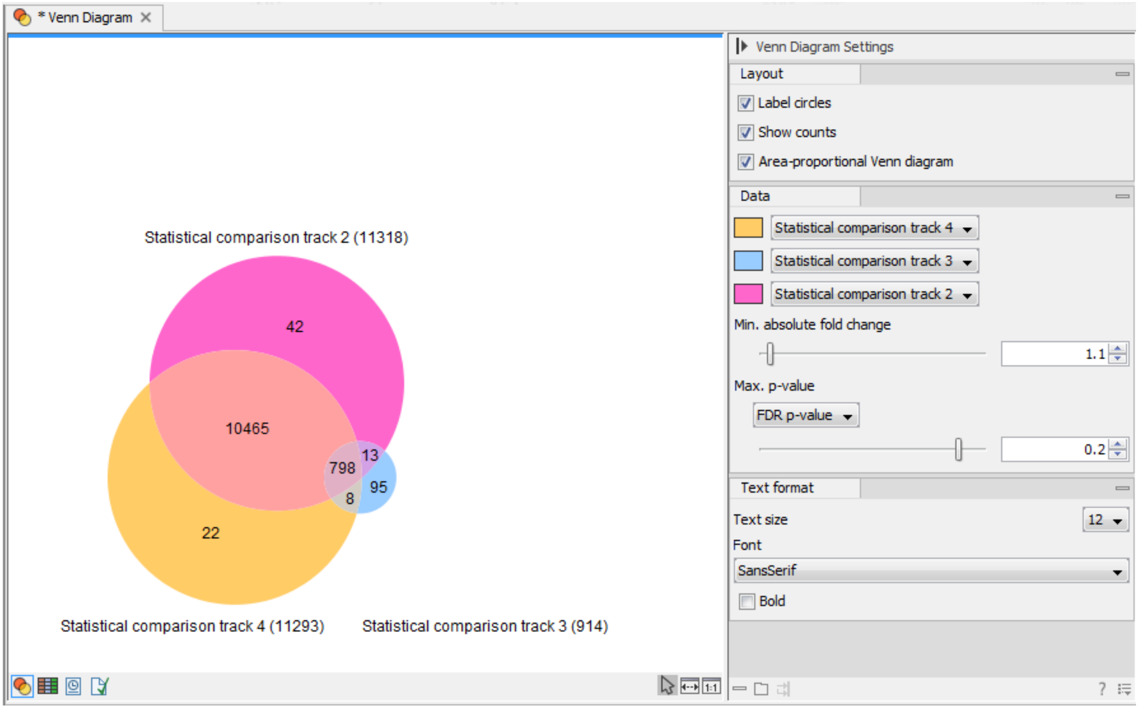
Figure 1.15: *Expression browser table when a GO annotation file is present.*

considered to be differentially expressed can be controlled by setting appropriate p-value and fold change thresholds.

To create the Venn diagram:

Toolbox | Transcriptomics Analysis ( )| RNA-Seq Analysis | Create Venn Diagram

Select a number of statistical comparison tracks ( ) and click **Next** (see figure 1.16).


* Venn Diagram
Statistical comparison track 2 (11318)
42
10465
13
798
95
8
22
Statistical comparison track 4 (11293)
Statistical comparison track 3 (914)

Venn Diagram Settings

Layout

☑ Label circles
☑ Show counts
☑ Area-proportional Venn diagram

Data

🟧 Statistical comparison track 4
🟦 Statistical comparison track 3
🟪 Statistical comparison track 2

Min. absolute fold change: 1.1

Max. p-value: FDR p-value, 0.2

Text format

Text size: 12

Font: SansSerif

☐ Bold

Figure 1.16: *The resulting Venn diagram.*

In the **Side Panel** to the right, it is possible to adjust the Venn Diagram settings. Under **Layout**, you can adjust the general properties of the plot.

- **Label circles** Toggles the names of the statistical comparison tracks.
- **Show counts** Toggles the display of gene or transcript counts.

- **Area-proportional Venn Diagram** When drawn as a Standard Venn Diagram, circles are drawn with fixed positions and identical size. When drawn in the default Venn Diagram mode, sizes and positions of the circles are adjusted in proportion to the number of overlapping features.

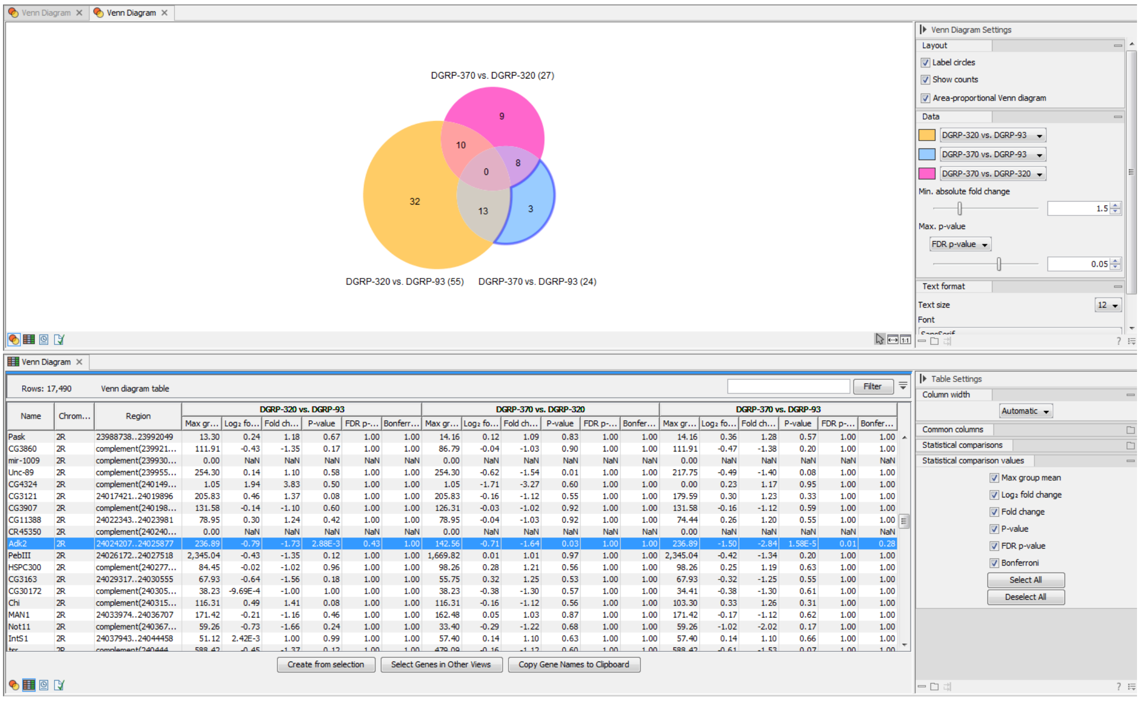
The **Data** side panel group makes it possible to choose the differentially expressed genes or features of interest. The set of statistical comparisons to be compared can be selected using the drop down combo boxes at the top of the group. It is possible to customize the color of a given statistical comparison using the color picker next to the drop down combo box.

- **Minimum absolute fold change** Only genes or transcripts with an absolute fold change higher than the specified threshold are taken into account.
- **Maximum P-value** Only genes or transcripts with a p-value less then the specified threshold will be taken into account. It is possible to select which p-value measure to use.

Finally, the **Text format** group makes it possible to adjust the settings for the count and statistical comparison labels.

1.6.1 Venn diagram table view

It is possible to inspect the p-values and fold changes for each gene or transcript individually in the Venn diagram table view (see figure 1.17).


Venn Diagram
Venn Diagram
DGRP-370 vs. DGRP-320 (27)
9
10
8
0
32
13
3
DGRP-320 vs. DGRP-93 (55)
DGRP-370 vs. DGRP-93 (24)
Venn Diagram
Filter
Create from selection
Select Genes in Other Views
Copy Gene Names to Clipboard
Select All
Deselect All

Venn Diagram Settings

Layout

☑ Label circles

☑ Show counts

☑ Area-proportional Venn diagram

Data

🟧 DGRP-320 vs. DGRP-93

🟦 DGRP-370 vs. DGRP-93

🟪 DGRP-370 vs. DGRP-320

Min. absolute fold change: 1.5

Max. p-value: FDR p-value 0.05

Text format

Text size: 12

Font

Rows: 17,490 Venn diagram table

Name	Chrom...	Region	DGRP-320 vs. DGRP-93						DGRP-370 vs. DGRP-320						DGRP-370 vs. DGRP-93					
			Max gr...	Log₂ fo...	Fold ch...	P-value	FDR p-...	Bonferr...	Max gr...	Log₂ fo...	Fold ch...	P-value	FDR p-...	Bonfer...	Max gr...	Log₂ fo...	Fold ch...	P-value	FDR p-...	Bonfer...
Pask	2R	23988738..23992049	13.30	0.24	1.18	0.67	1.00	1.00	14.16	0.12	1.09	0.83	1.00	1.00	14.16	0.36	1.28	0.57	1.00	1.00
CG3860	2R	complement(239921...	111.91	-0.43	-1.35	0.17	1.00	1.00	86.79	-0.04	-1.03	0.90	1.00	1.00	111.91	-0.47	-1.38	0.20	1.00	1.00
mir-1009	2R	complement(239930...	0.00	NaN	NaN	NaN	NaN	NaN	0.00	NaN	NaN	NaN	NaN	NaN	0.00	NaN	NaN	NaN	NaN	NaN
Unc-89	2R	complement(239955...	254.30	0.14	1.10	0.58	1.00	1.00	254.30	-0.62	-1.54	0.01	1.00	1.00	217.75	-0.49	-1.40	0.08	1.00	1.00
CG4324	2R	complement(240149...	1.05	1.94	3.83	0.50	1.00	1.00	1.05	-1.71	-3.27	0.60	1.00	1.00	0.00	0.23	1.17	0.95	1.00	1.00
CG3121	2R	24017421..24019896	205.83	0.46	1.37	0.08	1.00	1.00	205.83	-0.16	-1.12	0.55	1.00	1.00	179.59	0.30	1.23	0.33	1.00	1.00
CG3907	2R	complement(240198...	131.58	-0.14	-1.10	0.60	1.00	1.00	126.31	-0.03	-1.02	0.92	1.00	1.00	131.58	-0.16	-1.12	0.59	1.00	1.00
CG11388	2R	24022343..24023981	78.95	0.30	1.24	0.42	1.00	1.00	78.95	-0.04	-1.03	0.92	1.00	1.00	74.44	0.26	1.20	0.55	1.00	1.00
CR45350	2R	complement(240240...	0.00	NaN	NaN	NaN	NaN	NaN	0.00	NaN	NaN	NaN	NaN	NaN	0.00	NaN	NaN	NaN	NaN	NaN
Adk2	2R	24024207..24025877	236.89	-0.79	-1.73	2.88E-3	0.43	1.00	142.56	-0.71	-1.64	0.03	1.00	1.00	236.89	-1.50	-2.84	1.58E-5	0.01	0.28
PebIII	2R	24026172..24027518	2,345.04	-0.43	-1.35	0.12	1.00	1.00	1,669.82	0.01	1.01	0.97	1.00	1.00	2,345.04	-0.42	-1.34	0.20	1.00	1.00
HSPC300	2R	complement(240277...	84.45	-0.02	-1.02	0.96	1.00	1.00	98.26	0.28	1.21	0.56	1.00	1.00	98.26	0.25	1.19	0.63	1.00	1.00
CG3163	2R	24029317..24030555	67.93	-0.64	-1.56	0.18	1.00	1.00	55.75	0.32	1.25	0.53	1.00	1.00	67.93	-0.32	-1.25	0.55	1.00	1.00
CG30172	2R	complement(240305...	38.23	-9.69E-4	-1.00	1.00	1.00	1.00	38.23	-0.38	-1.30	0.57	1.00	1.00	34.41	-0.38	-1.30	0.61	1.00	1.00
Chi	2R	complement(240315...	116.31	0.49	1.41	0.08	1.00	1.00	116.31	-0.16	-1.12	0.56	1.00	1.00	103.30	0.33	1.26	0.31	1.00	1.00
MAN1	2R	24033974..24036707	171.42	-0.21	-1.16	0.46	1.00	1.00	162.48	0.05	1.03	0.87	1.00	1.00	171.42	-0.17	-1.12	0.62	1.00	1.00
Not11	2R	complement(240367...	59.26	-0.73	-1.66	0.24	1.00	1.00	33.40	-0.29	-1.22	0.68	1.00	1.00	59.26	-1.02	-2.02	0.17	1.00	1.00
IntS1	2R	24037943..24044458	51.12	2.42E-3	1.00	0.99	1.00	1.00	57.40	0.14	1.10	0.63	1.00	1.00	57.40	0.14	1.10	0.66	1.00	1.00
tsr	2R	complement(240444...	588.42	-0.45	-1.37	0.12	1.00	1.00	479.09	-0.16	-1.12	0.60	1.00	1.00	588.42	-0.61	-1.53	0.07	1.00	1.00

Table Settings

Column width: Automatic

Common columns

Statistical comparisons

Statistical comparison values

☑ Max group mean

☑ Log₂ fold change

☑ Fold change

☑ P-value

☑ FDR p-value

☑ Bonferroni

Figure 1.17: *The Venn diagram table view.*

Clicking a circle segment in the Venn Diagram plot will select the genes or transcript in the table view. It is also possible to create a subset list of genes or transcripts using the **Create from selection**.

In the **Side Panel** to the right it is possible to adjust the Table settings. It is possible to adjust the column layout, and select which columns should be included in the table.

Chapter 2

Installation of the Advanced RNA-Seq plugin

Plugins are installed using the plugin manager. In order to install plugins on Windows, the Workbench must be run in administrator mode: Right-click the program shortcut and choose "Run as Administrator". Then follow the procedure described below.

Help in the Menu Bar | **Plugins... ( )**

or **Plugins ( ) in the Toolbar**

The plugin manager has two tabs at the top:

- **Manage Plugins.** This is an overview of plugins that are installed.
- **Download Plugins.** This is an overview of available plugins on CLC bio's server.

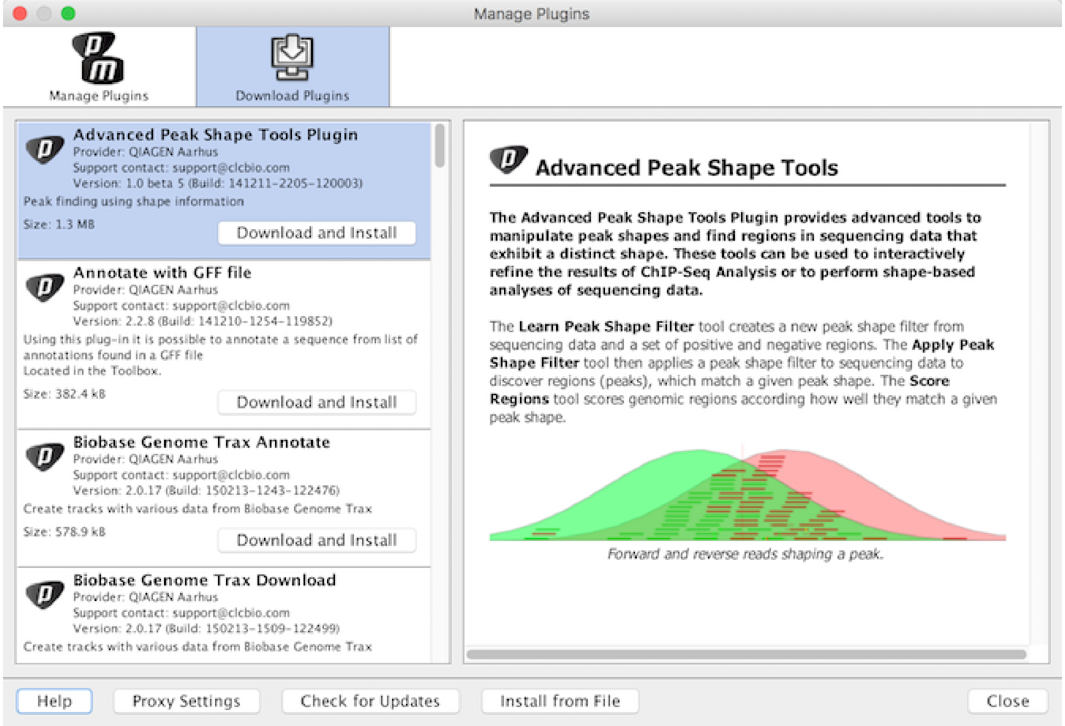
To install a plugin, click the **Download Plugins** tab. This will display an overview of the plugins that are available for download and installation (see figure 2.1).

Clicking a plugin will display additional information at the right side of the dialog. This will also display a button: **Download and Install**.

Click the Advanced RNA-Seq plugin and press **Download and Install**. A dialog displaying progress is now shown, and the plugin is downloaded and installed.

If the Advanced RNA-Seq plugin is not shown on the server, and you have it on your computer (for example if you have downloaded it from our website), you can install it by clicking the **Install from File** button at the bottom of the dialog. This will open a dialog where you can browse for the plugin. The plugin file should be a file of the type ".cpa".

When you close the dialog, you will be asked whether you wish to restart the CLC Workbench. The plugin will not be ready for use until you have restarted.



Advanced Peak Shape Tools Plugin
Provider: QIAGEN Aarhus
Support contact: support@clcbio.com
Version: 1.0 beta 5 (Build: 141211-2205-120003)
Peak finding using shape information
Size: 1.3 MB

Annotate with GFF file
Provider: QIAGEN Aarhus
Support contact: support@clcbio.com
Version: 2.2.8 (Build: 141210-1254-119852)
Using this plug-in it is possible to annotate a sequence from list of annotations found in a GFF file
Located in the Toolbox.
Size: 382.4 kB

Biobase Genome Trax Annotate
Provider: QIAGEN Aarhus
Support contact: support@clcbio.com
Version: 2.0.17 (Build: 150213-1243-122476)
Create tracks with various data from Biobase Genome Trax
Size: 578.9 kB

Biobase Genome Trax Download
Provider: QIAGEN Aarhus
Support contact: support@clcbio.com
Version: 2.0.17 (Build: 150213-1509-122499)
Create tracks with various data from Biobase Genome Trax

Advanced Peak Shape Tools

The Advanced Peak Shape Tools Plugin provides advanced tools to manipulate peak shapes and find regions in sequencing data that exhibit a distinct shape. These tools can be used to interactively refine the results of ChIP-Seq Analysis or to perform shape-based analyses of sequencing data.

The **Learn Peak Shape Filter** tool creates a new peak shape filter from sequencing data and a set of positive and negative regions. The **Apply Peak Shape Filter** tool then applies a peak shape filter to sequencing data to discover regions (peaks), which match a given peak shape. The **Score Regions** tool scores genomic regions according how well they match a given peak shape.

Figure 2.1: *The plugins that are available for download.*

Chapter 3

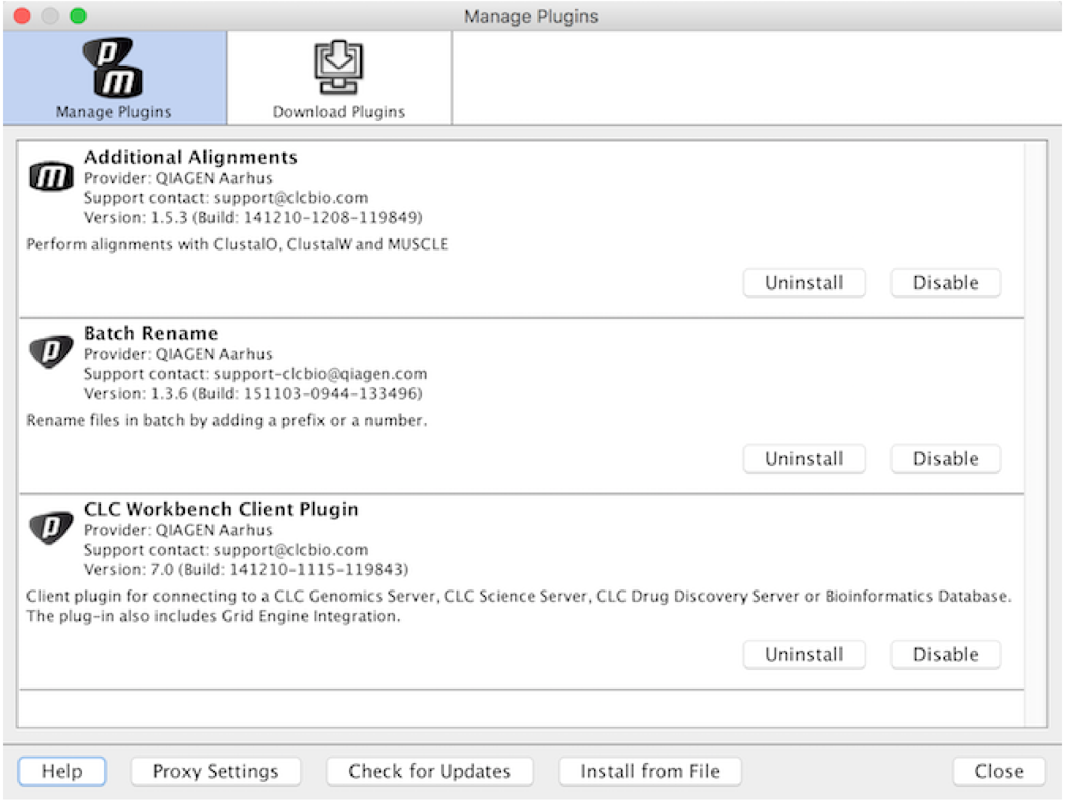
Uninstall

Plugins are uninstalled using the plugin manager:

Help in the Menu Bar | Plugins... ( )

or **Plugins ( ) in the Toolbar**

This will open the dialog shown in figure 3.1.



Additional Alignments
Provider: QIAGEN Aarhus
Support contact: support@clcbio.com
Version: 1.5.3 (Build: 141210-1208-119849)
Perform alignments with ClustalO, ClustalW and MUSCLE

Batch Rename
Provider: QIAGEN Aarhus
Support contact: support-clcbio@qiagen.com
Version: 1.3.6 (Build: 151103-0944-133496)
Rename files in batch by adding a prefix or a number.

CLC Workbench Client Plugin
Provider: QIAGEN Aarhus
Support contact: support@clcbio.com
Version: 7.0 (Build: 141210-1115-119843)
Client plugin for connecting to a CLC Genomics Server, CLC Science Server, CLC Drug Discovery Server or Bioinformatics Database.
The plug-in also includes Grid Engine Integration.

Figure 3.1: *The plugin manager with plugins installed.*

The installed plugins are shown in this dialog. To uninstall:

Click the Advanced RNA-Seq plugin | Uninstall

If you do not wish to completely uninstall the plugin but you don't want it to be used next time you start the Workbench, click the **Disable** button.

When you close the dialog, you will be asked whether you wish to restart the workbench. The plugin will not be uninstalled until the workbench is restarted.

Bibliography

[Auer and Doerge, 2010] Auer, P. L. and Doerge, R. (2010). Statistical design and analysis of rna sequencing data. *Genetics*, 185(2):405–416.

[Mortazavi et al., 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628.

[Robinson et al., 2010] Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

[Robinson and Oshlack, 2010] Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, 11(3):R25.

Index