Tutorial

Expression Analysis using RNA-Seq

March 2, 2017

Tutorial

Expression Analysis using RNA-Seq

This tutorial will show how the RNA-Seq Analysis tools facilitate the expression analysis of RNA-Seq data.

Prerequisites   For this tutorial, you must be working with the CLC Genomics Workbench 10.0 or higher.

Download and import data   This tutorial will use RNA-Seq data for male and female *Drosophila melanogaster* from 3 different strains, collected at different time points during the year, and sequenced using 2 different RNA preparation protocols. Our goal is to find genes that are differentially expressed between the different strains, while controlling for the effects of the other factors.

To allow the completion of the tutorial in a reasonable amount of time, we have reduced the samples to only 10 individuals out of the 726 present in the original dataset [Lin et al., 2016], and have reduced each sample to 250,000 reads mapping to chromosome 2R.

The data set includes:

- The reads from 10 Drosophila samples.

- An Excel spreadsheet that contains the metadata associated with each individual.

    - **SRR_ID** is an SRA identifier unique for each individual.
    - **DGRP_Number** describes the strain of the fly.
    - **Sex** stated as M for males and F for females.
    - **Environment** stated as 2 and 3 for different calendar times for collecting the flies.
    - **RNA_Prep_Method** using QIAGEN RNeasy kit in all cases but following either the centrifuge or the vacuum based protocol.
    - **Lane** of the sequencer on which the sample was loaded.

- A workflow to rapidly and efficiently process statistical analyses on the RNA-Seq results.

- A GO annotation file for Drosophila.

Go through the following steps to download and import the data into the Workbench.

1. Download the sample data from our website: `http://resources.qiagenbioinformatics.com/testdata/RNA_Seq_Droso2.zip`.

2. Unzip the zip file on your local computer.

3. Start the *CLC Genomics Workbench*.

4. First, import the workflow and the gene association file we will need later in this tutorial: **File | Import** (📥) | **Standard Import** (📥) and choose the files called "RNA-Seq statistics workflow" and "gene_association.fb". Save the imported data in a new folder you can create in your Navigation Area (called for example RNA-Seq Drosophila).

5. Next, import the reads via the toolbar: **File** | **Import** (🗂) | **Illumina** (🖼). Choose the 10 *.fastq files, and **leave the option "Paired reads" unchecked**. You can however "Discard read names", "Discard quality scores" and "Remove failed reads" (figure 1). Click **Next**.
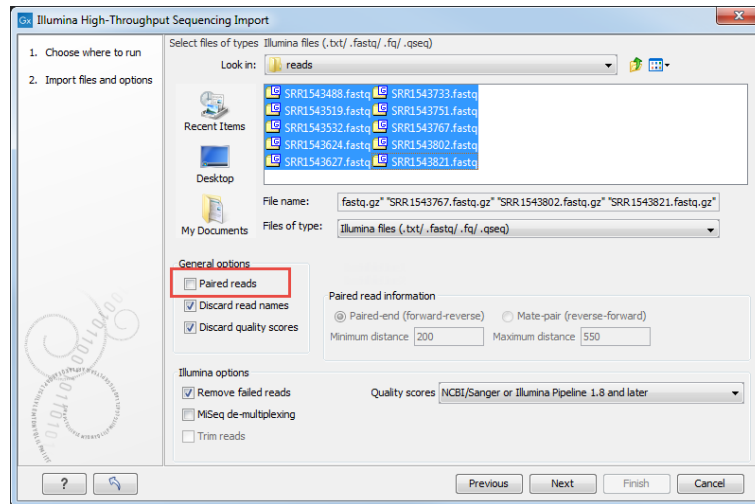

Figure 1: *In the Import wizard, select all 10 reads files.*

6. Now import the metadata table via the toolbar: **File** | **Import** (🗂) | **Import Metadata** (🖼).

7. A wizard opens. In the first dialog, select the spreadsheet saved on your local computer that contains the sample information "DrosophilaMetadata.xls". The contents of the Excel spreadsheet populates the table situated at the bottom of the dialog (figure 2). Click **Next**.
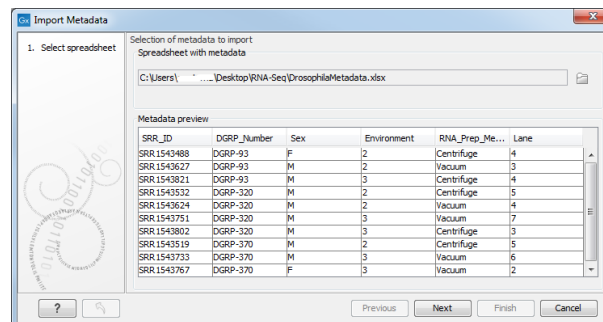

Figure 2: *In the Import Metadata wizard, select the spreadsheet with the sample information.*

8. Click on the Navigation button next to "Location of data", and find the imported reads in your Navigation Area. Right-click the name of the folder and select "Add folder contents"(figure 3). Click **OK**.

9. In the wizard, now you can see that the data association is successful by the fact that the metadata has filled each row of the table located at the bottom of the dialog. Click **Next**.

10. Select the "RNA-Seq Drosophila" folder to save the resulting metadata table.

11. Now open the metadata table you just saved (called "Samples" as can be seen in the upper part of figure 4). Select all rows in the table, and click **Find Associated Data**. This will show that all rows are associated correctly with the corresponding reads.
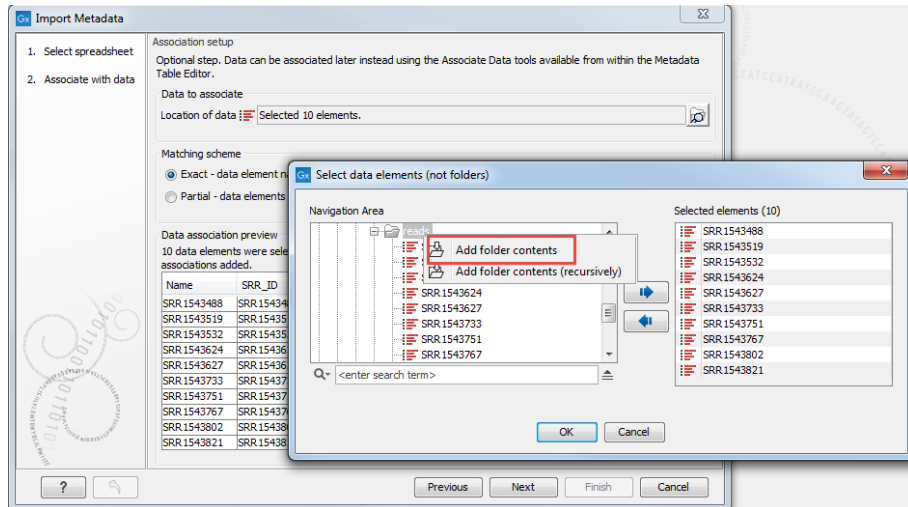
Tutorial

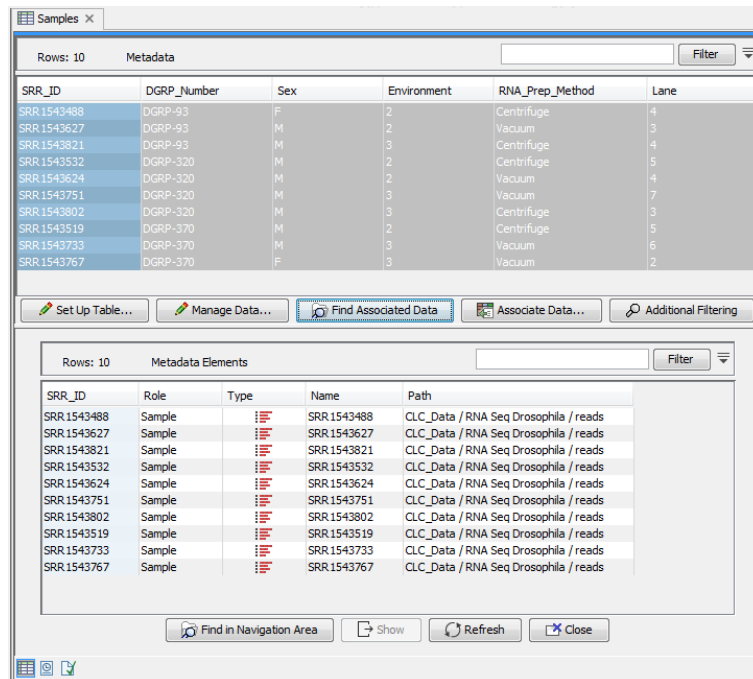Figure 3: *In the Import Metadata wizard, select the reads.*

Figure 4: *In the Metadata table, check that associations between reads and the metadata have happened successfully.*

Data management configuration

In order to go through this tutorial, you need the reference sequence for *Drosophila melanogaster*. In order to access this data, you can download the relevant data directly from the workbench.

1. Click on the button labeled **Download** ( ) in the top right corner of the Workbench, and choose the option "Download Reference Genome Data".

2. In the wizard, check the options "Animals (others)", and select "Drosophila melanogaster" from the drop down menu (figure 5). Click **Next**.
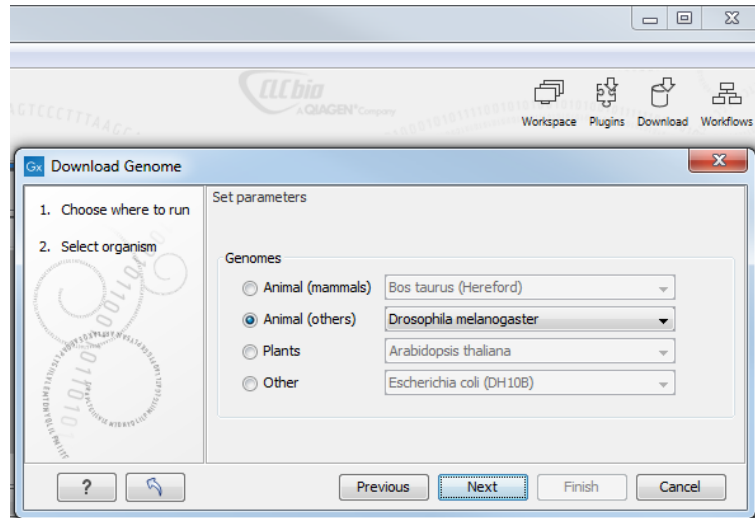
Figure 5: *Download reference data for* Drosophila melanogaster*.*

3. In the following wizard window, check "Download genome sequence" and click **Next**.

4. In the window called "Select annotations", add a check to the row "Genome Annotations". We do not need variant data for this tutorial (figure 6). Click **Next**.
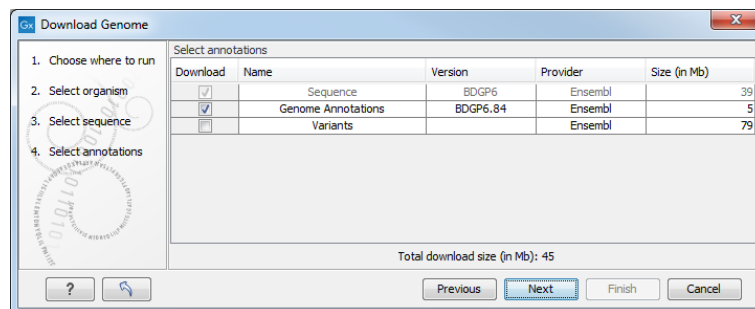

Figure 6: *Choose to download the sequence and the genome annotations for* Drosophila melanogaster*.*

5. **Save** the references in a subfolder you can call "references" in the "RNA-Seq Drosophila" folder.

Once the data and references have been downloaded and imported, you should see the folder and files in the Navigation Area as shown in figure 7.

RNA-Seq quantification step

To map the reads to the reference, we will use the RNA-Seq tool. Go to:

> **Toolbox | RNA-Seq Analysis (**📊**) | RNA-Seq Analysis (**📊**)**

1. Depending on your setup you may be asked where to run the tool. Choose the appropriate option for you and click **Next**.
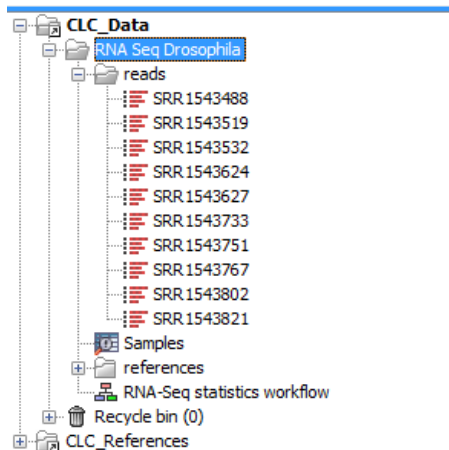
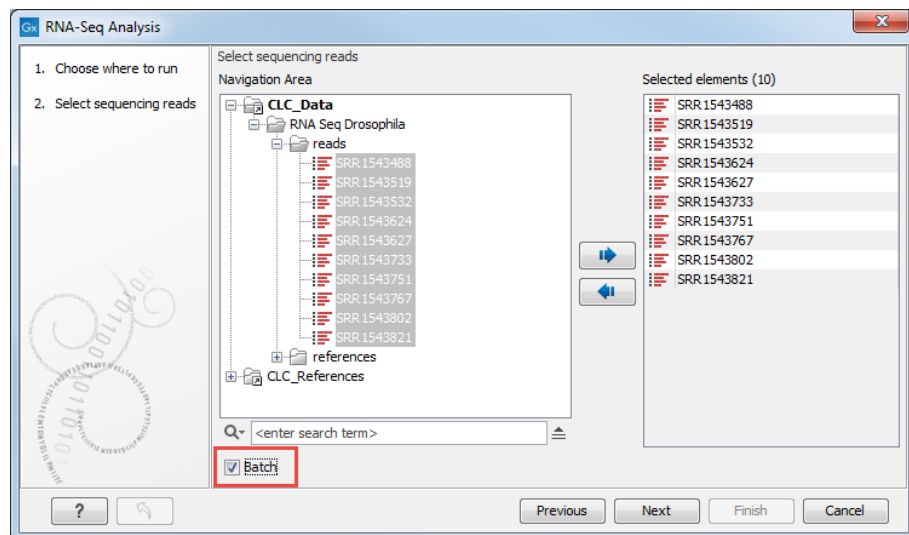Figure 7: *The data set has been imported.*

Figure 8: *Do not forget to check the "Batch" mode option to make sure that the 10 samples are processed independently of each other, and not as the reads of a single individual.*

2. In the next window, **check the "Batch" box** and select the 10 reads files (figure 8).

3. In the next wizard window review the batch units and click **Next**. Failing to see this window means you have not checked the "Batch" option during the previous step.

4. Set up the wizard step so that it looks like in figure 9. Keep "Genome annotated with genes and transcripts" checked. The Reference sequence, Gene track and mRNA track were previously downloaded using the Download manager, and can now be found in your "references folder".

   Click **Next**.

5. Use default parameters for the "Mapping options" dialog, except for the Strand specific option that should be set to **Reverse** (figure 10). Click **Next**.

6. In the "Expression level options" dialog, leave all parameters as default as well. You can
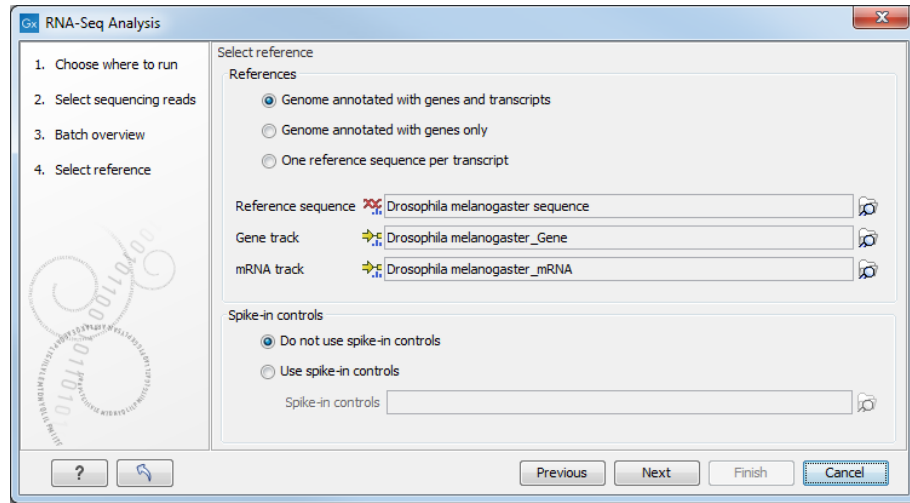
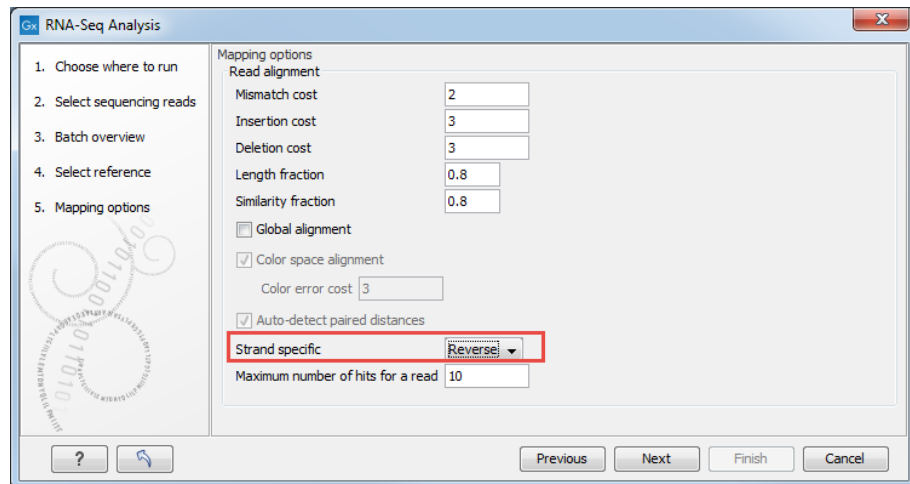Figure 9: *Genome annotated with genes and transcripts is checked.*

Figure 10: *The RNA-Seq was sequenced using a reverse protocol, so set the strand specificity to "Reverse" for the mapping.*

use the "Set to default" button (🖉) if you are not sure whether you have previously changed parameters for this tool. Click **Next**.

7. Save the results in a specified location (a new RNA-Seq results folder for example) and create subfolders per batch unit.

After starting the tool, you can use the progress bar to see how it is progressing (figure 11). The mapping for these 10 samples takes around 20 minutes on a standard laptop computer.

When the RNA-Seq Analysis tool has completed, you can click on the Refresh button of the Metadata Elements table (bottom part) of the "Samples" file. You can see that after analysis, we now have many more files associated to the initial samples (figure 12).
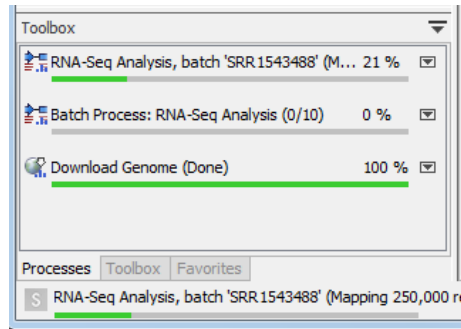
Figure 11: *The process called "Batch Process" indicates how many batches have been completed, while the one called "RNA-Seq Analysis" shows the analysis progress of a particular batch unit.*

PCA plot

We now use the PCA plot to check for outlying samples of poor quality that may need excluding from further analysis. We will also use the plot to see which factors listed in the metadata may confound the detection of genes that are differentially expressed due to fly strain.

1. Click the "Role" column title of the Metadata Elements table to sort the elements by title and select the 10 rows whose role is "Gene expression" (figure 12). Alternatively, you can enter "Gene expression" in the Filter box and select all remaining rows.
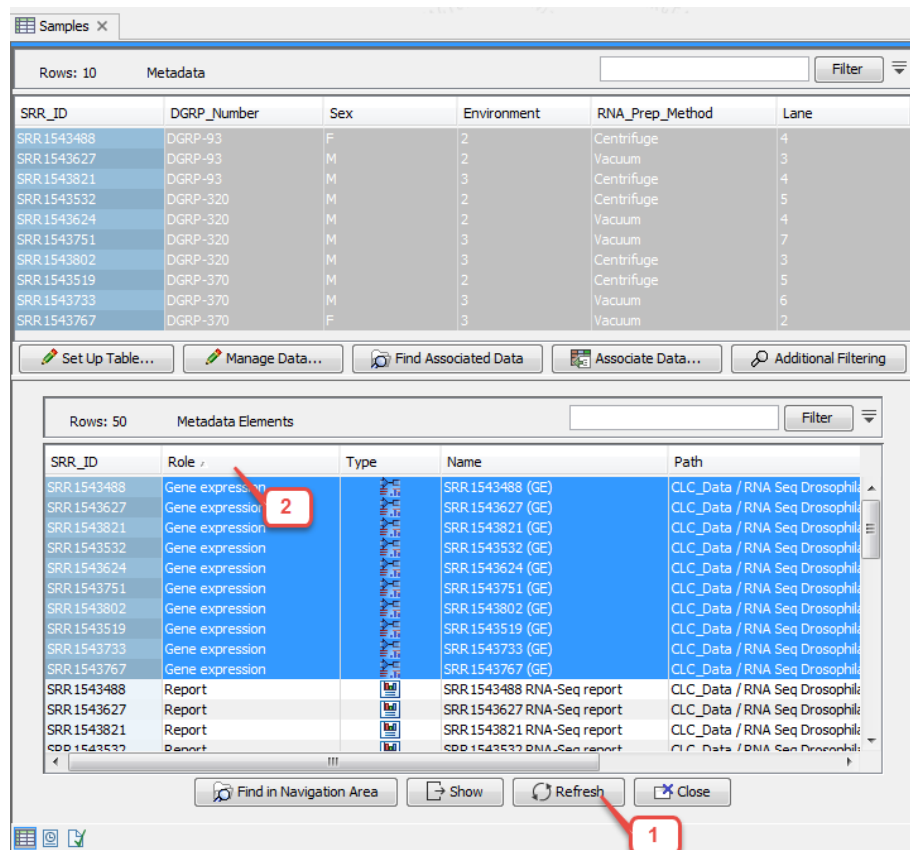

Figure 12: *Finding all results having "Gene Expression" as Role using the metadata table.*

2. Go to **Toolbox** | **RNA-Seq Analysis** (📑) | **PCA for RNA-Seq** (🔳).

3. In the first wizard window, all Gene Expression files (having the (GE) suffix) should be pre-selected.

4. Choose to Save the PCA plot.

Open the PCA plot. The samples are plotted according to their DGRP_Number. Choose in turn each of the different values from the side panel option: **Metadata | Symbol color**. It becomes evident that the principal variation between the samples (PC1) is due to their sex, while the second greatest source of variation (PC2) emphasizes the importance of using the same RNA-Seq preparation protocol throughout an experiment (figure 13). The variation due to the RNA-Seq preparation protocol is an example of a batch effect. Batch effects can arise for many reasons, e.g. between different lanes of a sequencer, or samples prepared on different days. It is good practice to log as much metadata as possible about an experiment, so that all potential batch effects can be detected and corrected for in the statistics.
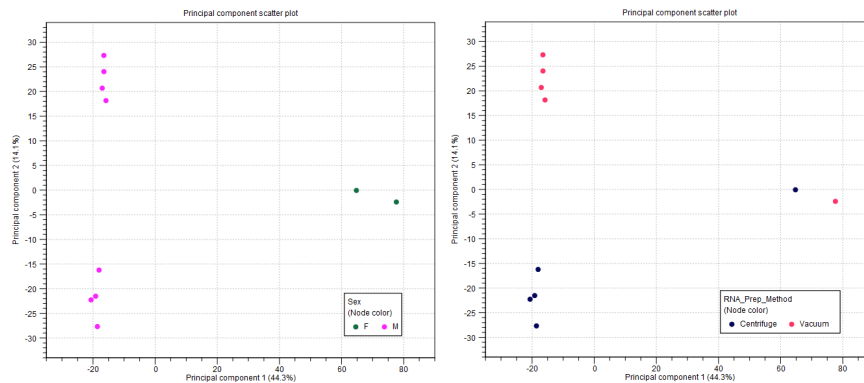

Figure 13: *Sex of the flies, and RNA preparation protocols are the 2 main factors explaining the differences between individuals.*

You can also investigate more remote components of the PCA by choosing to display the plot with another principal component as axis using the drop down menu in the side panel. The differences in the fly strains appear as PC3 and PC4. You can also see the PCA in 3D using the icon in the lower left corner (figure 14).

Going back to the initial representation of the PCA (figure 13), it is evident that fly sex is so dominant that we would benefit from removing this large confounding effect for further analyses. In general, it is often a matter of judgement whether to keep samples and try to correct for the factor in the statistics, or to remove samples so that no correction is necessary. In this particular case, we can proceed by removing the female individuals without sacrificing much data. We will thus perform the rest of the analysis on the 8 male samples. As for the other large confounding effect, the RNA preparation protocol, we will show how to correct for it in the next section.

Using a workflow to perform additional statistical analyses

To pursue our investigation of the data, we will use the workflow we imported earlier. Open the workflow by double clicking on its name in the Navigation Area (figure 15).
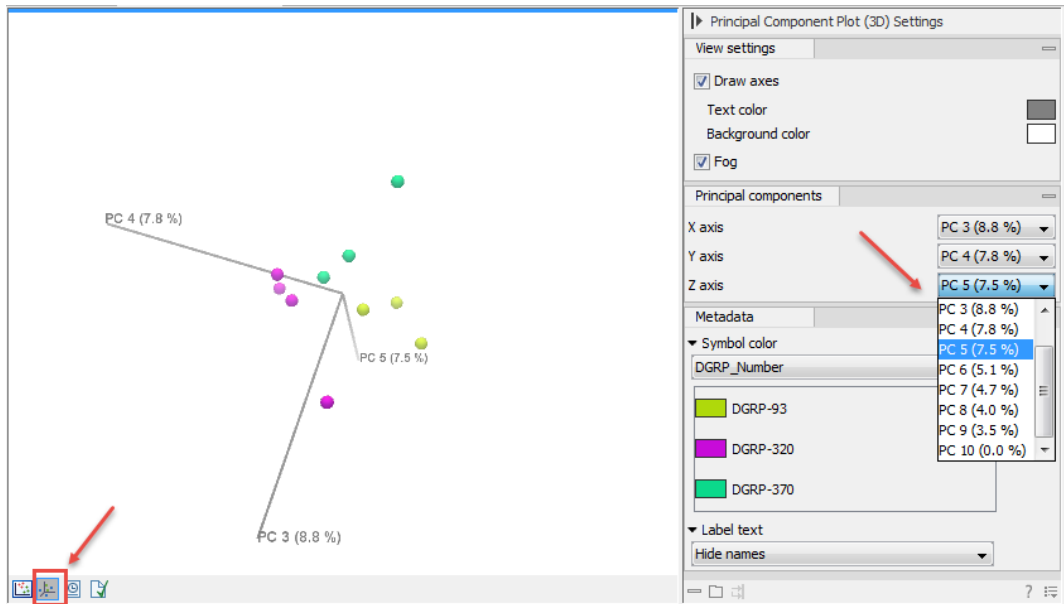
Figure 14: *Choose another Principal Component as axis x or y (and z) for your PCA plot, and view the PCA plot in 3D.*
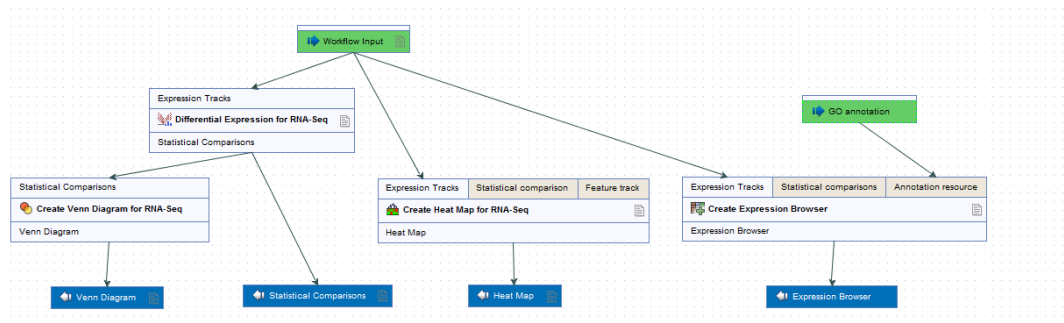

Figure 15: *View of the RNA-Seq statistics workflow.*

We can run the workflow from the Navigation Area, but in order to be able to use the functionalities of the metadata table to help select the relevant input files, it is recommended to install the workflow in your workbench.

1. In the lower right corner of the View Area, click on the button "Installation".

2. A pop up window opens containing the workflow information. Click on **Next**.

3. In the following window, choose to "Install the workflow on your local computer" before clicking **Finish**.

The workflow is now part of your toolbox in the "Workflow" folder. Before running it, we will remove the two female individuals from the dataset. This can be done easily from the metadata table "Samples".

1. In the upper table of the "Samples" view, click on the arrow next to "Filter" to open up the advanced filter box.

2. Select **Sex** in the first field, **=** in the second, and type **M** in the last.

3. Click on the button "Filter".

4. Select the 8 remaining rows and click on the button "Find Associated Data".

5. Now, only the Metadata Elements associated with male individuals remain in the lower table as well.

6. Select all elements whose role is "Gene expression", and right click on the selection.

7. In the menu that just opened in your View Area, go to **Toolbox | Workflows | RNA-Seq statistics workflow** (figure 16)
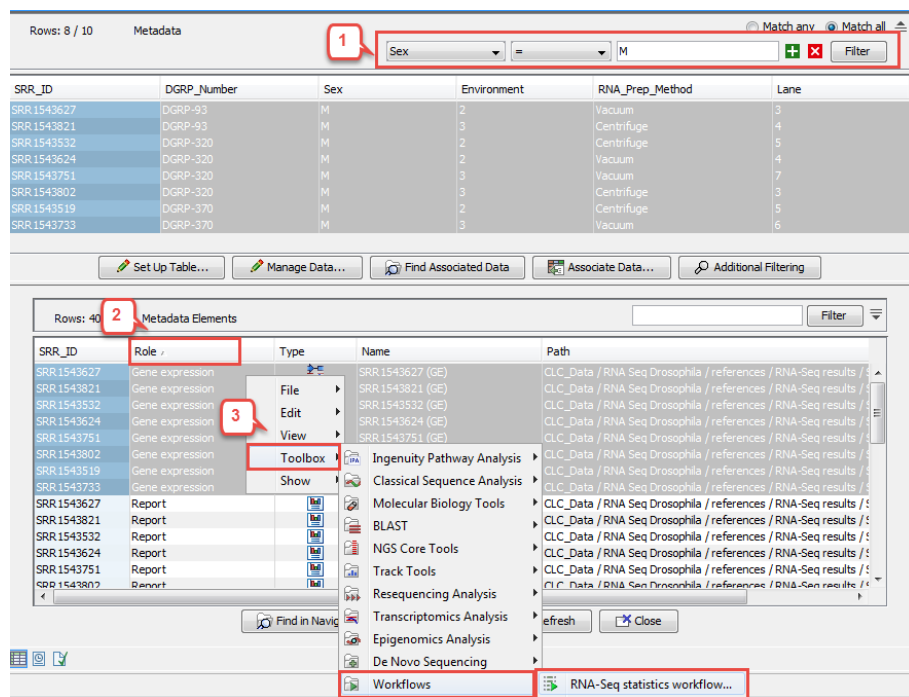

Figure 16: *Start the workflow from the metadata table with only a subset of your previous results.*

8. A wizard window opens for the workflow. If you are asked, choose where to execute the workflow and click **Next**.

9. Click **Next** again to accept the selected tracks. Note that if you start the workflow from the Toolbox, the tracks selected in the metadata table will also be automatically used as input, but not if you start the workflow from the Navigation Area.

10. In the "Select GO annotation" dialog, select the file "gene_association.fb" you imported earlier. Click **Next**.

11. In the "Differential Expression for RNA-Seq" window, set the parameters as shown in figure 17. We want to find genes that are differentially expressed between the different strains of flies (DGRP_Number) while controlling for the RNA preparation protocol (RNA_Prep_Method), that we now know introduces a bias in the data. We want to produce a comparison between each pair of fly strains, so we select the 'All group pairs' comparison. When finished, click **Next**.
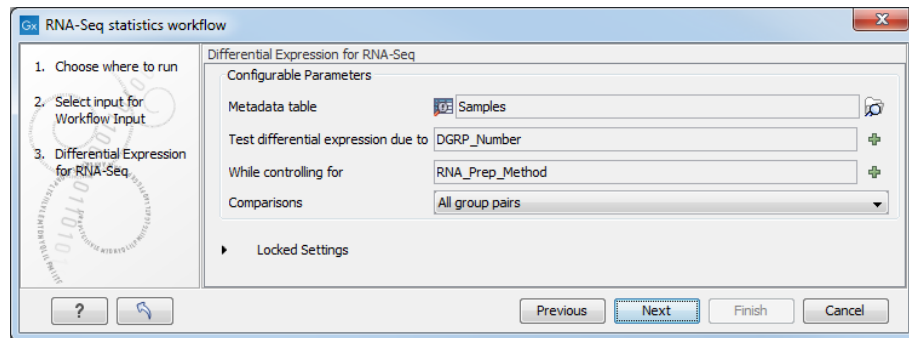
Figure 17: *Parameters for the RNA-Seq statistical analyses.*

12. In the "Result handling" window, choose a location to save the results (for example a new subfolder you can call "statistical analyses"), and click **Finish**.

The workflow will now execute.

Results visualization and interpretation

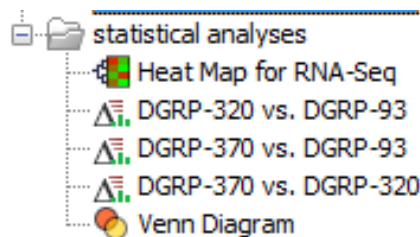When the workflow is completed, several new objects are created (figure 18)

Figure 18: *The results as seen in your Navigation Area.*

Venn diagram   Open the Venn diagram (figure 19).

The Venn diagram shows how many transcripts were detected to be differentially expressed in each comparison.

Select one of the diagram's intersections, where genes are differentially expressed in two comparisons. With the selection maintained, go to the table view of the Venn diagram. The same genes will still be selected in the table. It can be hard to find them, however you can create a new selection only containing the selected genes by clicking the "Create from selection" button. This will open a new view containing only the genes of the section that was selected in the Venn diagram.

Statistical comparison   Now we will show you how to track the genes of interest throughout your statistical results.

1. Make a new Track List containing all 3 statistics results, as well as the sequence, gene and mRNA reference tracks: go to **File | New | Track list** and select the statistical comparisons and the reference as indicated in figure 20. Click **Finish**.
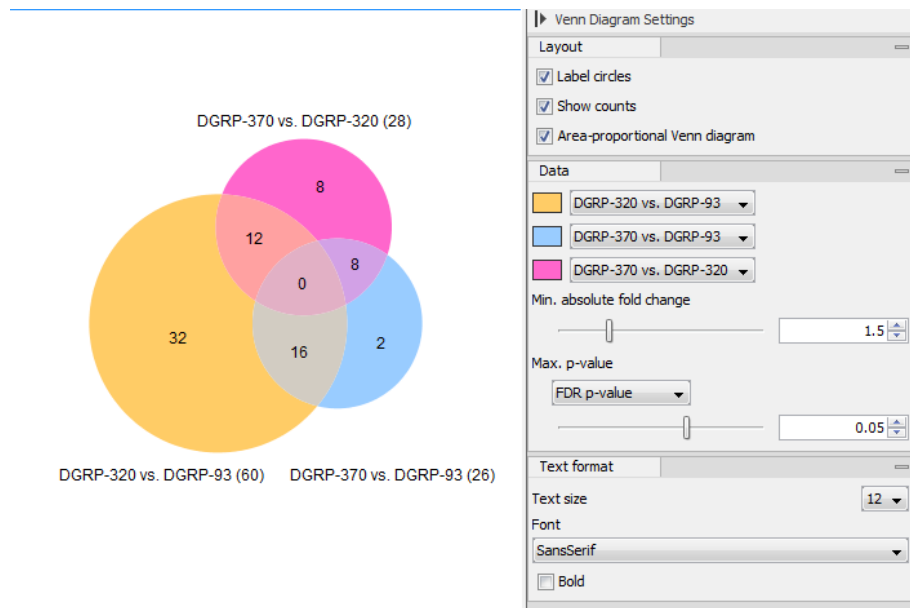
Figure 19: *An example of Venn Diagram. Note that the values you obtain in this diagram could differ slightly from the ones seen in this figure.*
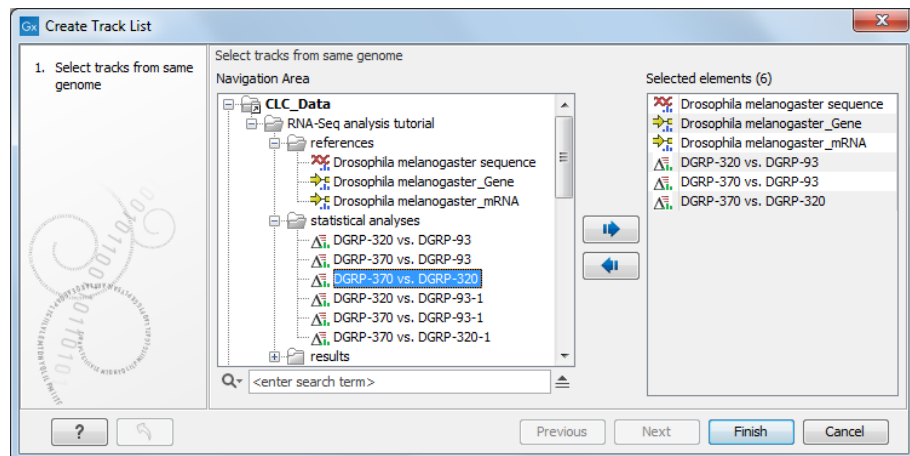

Figure 20: *Creating a list of tracks with some references and the statistical comparisons.*

2. Open a table view of one of the statistical comparison by double-clicking on the name of the track.

3. Now re-open the Venn diagram, and select one of the "shared by two" sections.

4. Go to the table view of the Venn diagram and click "Select Genes in Other views". Because diagram and table are linked, the genes present in the central section of the Venn diagram are automatically selected in the table view, and by clicking on "Select genes in other view", they are now also selected in the table view of the statistical comparison track.

5. Going back to the statistical comparison we opened from the track list, switch to the volcano plot view (figure 21).
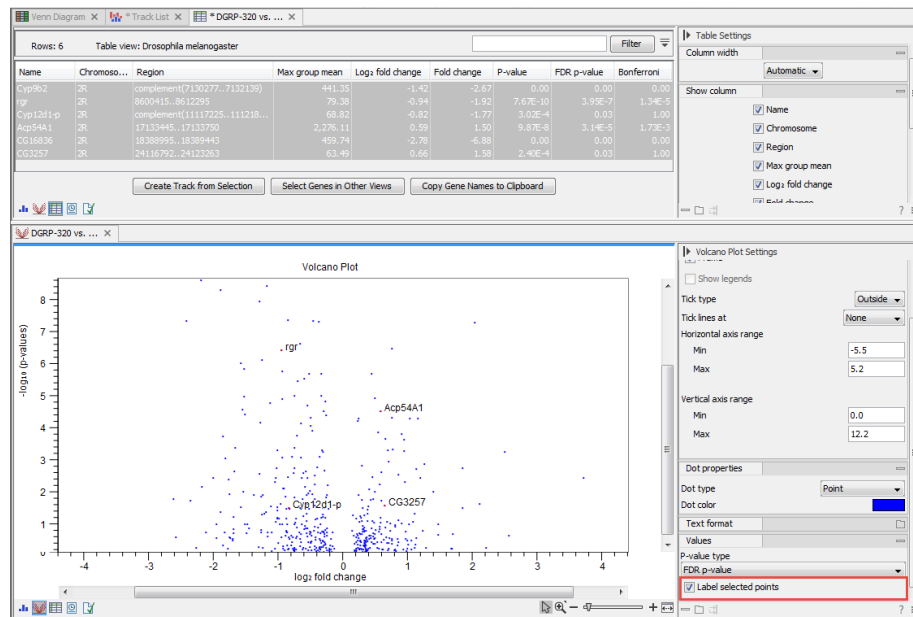
Figure 21: *The genes selected in the Venn diagram are now highlighted in red in the volcano plot. To enhance visibility, check the "Label selected points" in the side panel settings to the right of the volcano plot.*

We have demonstrated that a selection you make in the table will also be selected in the volcano plot, but note that it also works the other way round, so a gene selected in the volcano plot will also be selected in the table. The fact that table and plot are linked allows us to highlight the genes in the plots by filtering them in the table. Also, if you select only one gene in the volcano plot, the track list should jump to the gene (use the **Zoom to Selection** (⊞) button for optimal rendering of the track list).

Expression Browser   The Expression Browser makes it possible to inspect gene and transcript expression level counts and statistics for many samples at the same time. In addition, the GO biological process column will list for each gene or transcript one or several biological processes.

Again, it is possible to highlight in the Expression Browser genes that were selected as being differentially expressed between the different strains in the Venn diagram.

1. Open the Expression Browser (⊞) from the Navigation Area.

2. Now re-open the Venn diagram where one of the "shared by two" sections should still be selected.

3. In the table view of the Venn diagram, click **Select Genes in Other Views**.

4. The selected genes are now also selected in the Expression Browser, but they can be hard to find. Under the Expression Browser table, click on the button **Copy Gene Names to Clipboard**.

5. Then above the table, open the Filter box using the arrow. Choose **Name** in the first field drop-down menu, **is in list** in the second field, and copy the list of Gene Names you just saved in the third field.

6. Click on **Filter**.

The Expression Browser now only displays genes with the names of the previously selected genes (figure 22). This means that if two genes have the same name, both will be in that table, even if only one was in your original selection. Check that that number of rows displayed in the Expression Browser is identical to the sum of the genes selected in the Venn diagram (in the table below we selected all 3 sections of genes differentially expressed between strains).
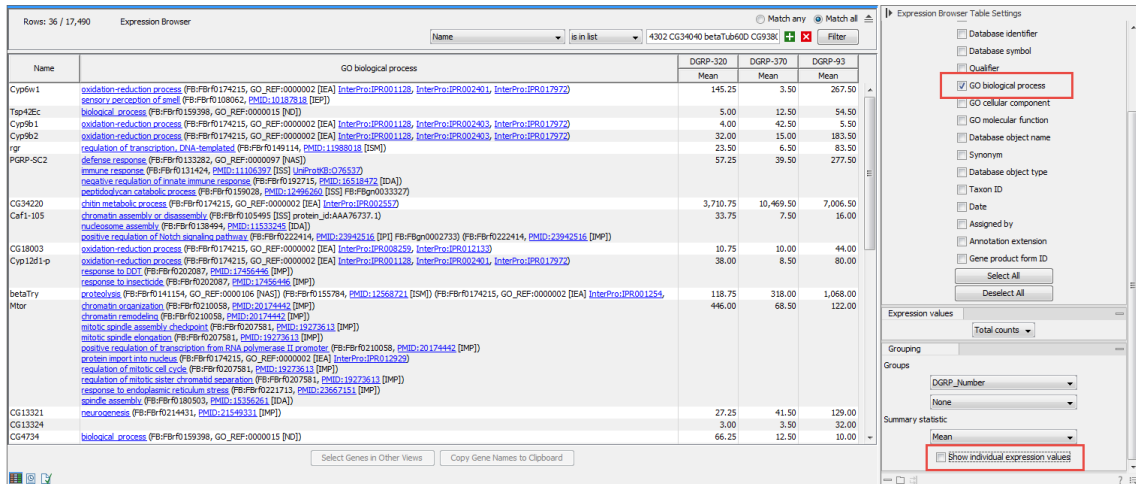

Figure 22: *Expression browser table displaying the genes found to be differentially expressed between the different strains. The table is set up to only show the name of the gene, the GO biological processes associated to this particular gene, and the mean of total counts for each strain.*

You can then 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 PMID, RefSeq, HGNC or UniProt accession number when available.

Heat map   Finally, open the Heat Map for RNA-Seq from the Navigation Area. The hierarchical clustering algorithm is pre-configured in the workflow to select the 25 "most interesting" genes in the heat map based on the coefficient of variation (relative standard deviation). The samples are also clustered on the horizontal axis, in unsupervised clustering.

In the right hand side panel, choose each metadata element in turn (see figure 23). You will notice that the 25 most interesting genes are differentially expressed due to the RNA Preparation method. This is what we should expect, because we know from the PCA plot that variation due to RNA preparation is greater than between the different strains.

One of the options of the Heat Map for RNA-Seq tool is to "Filter by statistics": using any of our statistical results with this option would lead to a heat map where samples clustered by fly strain.

1. Go back to the "samples" metadata table, and make sure that the "Gene expression" files for the 8 male individuals are selected.

2. From the toolbox, or directly from a right click on the selected elements in the metadata table, choose to run the "Create Heat Map for RNA-Seq" tool (figure 24). This opens the tool's wizard.
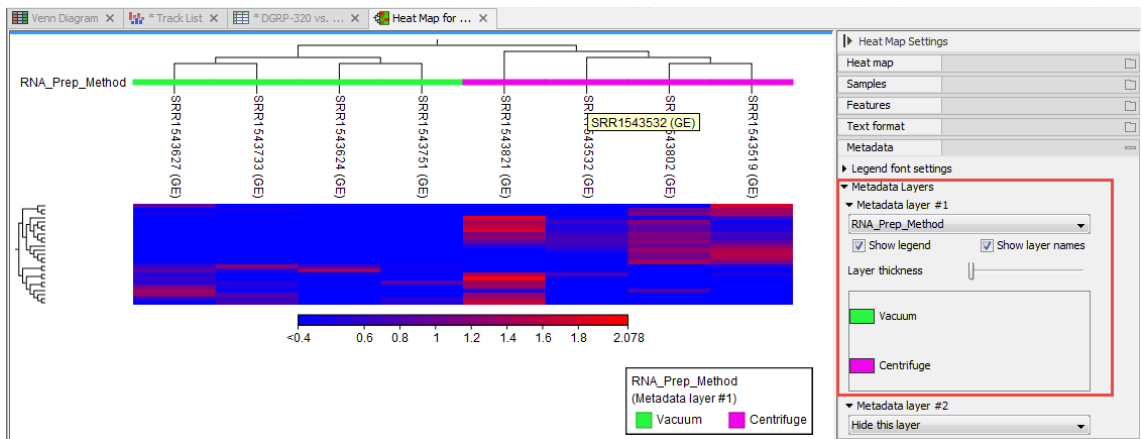
Tutorial

Figure 23: *Choose among the different metadata layers.*
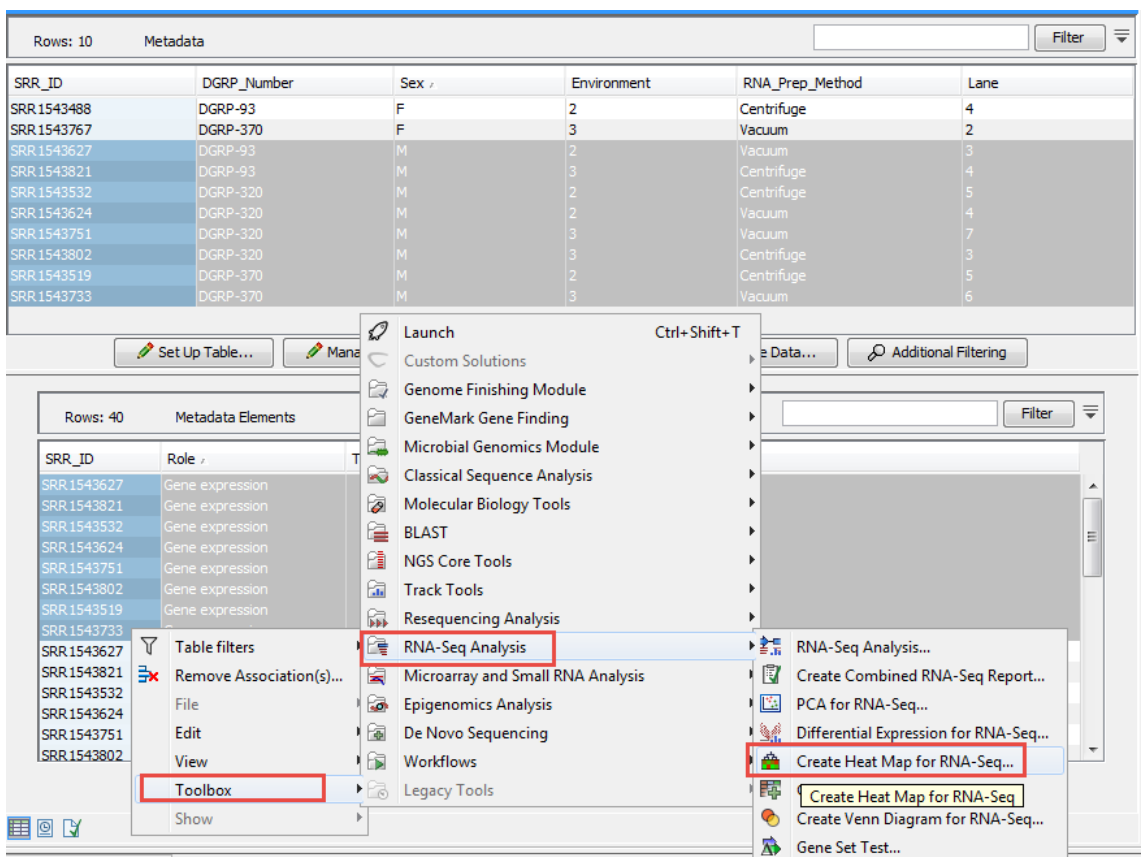

Figure 24: *Run the "Create Heat Map for RNA-Seq" tool on the 8 samples we chose to keep in the statistical analyses.*

3. In the first window of the wizard, the elements are automatically pre-selected. Click **Next**.

4. In the next window, leave the parameters as they are selected by default (see figure 25) and click **Next**.

5. In the "Set filtering" window, choose the option "Filter by statistics". Then click on the Navigation icon next to the "Statistical comparison" field and select one of the comparison
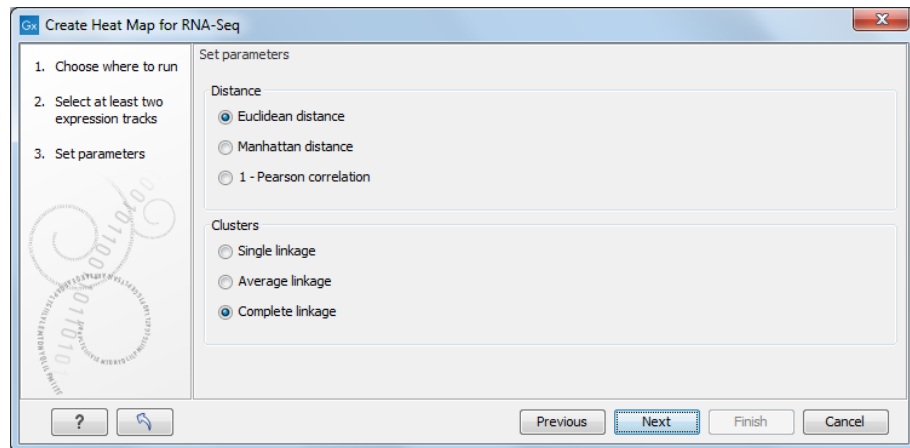
Figure 25: *Choose "Eucledian distances" and "Complete linkage" to run the "Create Heat Map for RNA-Seq" tool.*

generated by the workflow (figure 26). Click OK. Leave all other values as default and click **Next**.
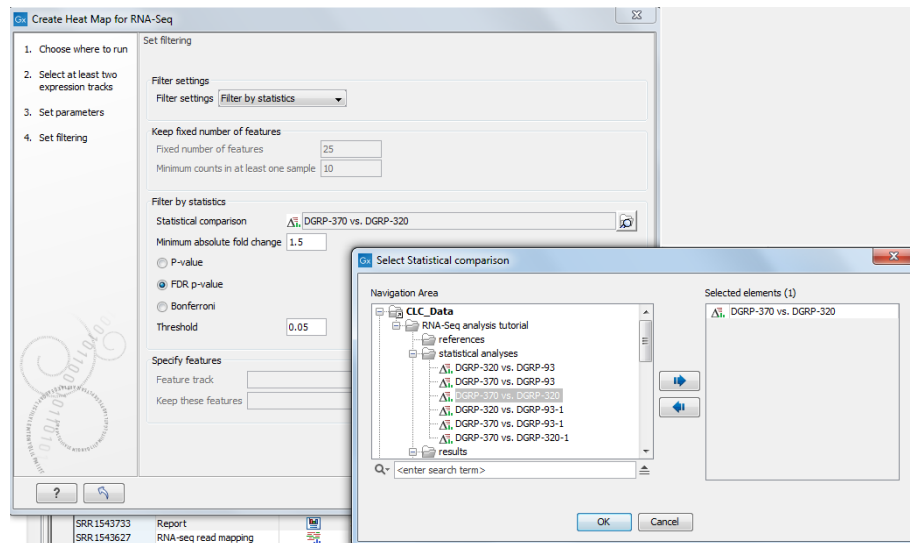

Figure 26: *Add a statistical comparison to filter by statistics.*

6. Choose to **Open** the result and click on **Finish** to start the tool.

The newest heat map opens as in figure 27, and it is now apparent that the samples are clustered by strains. There are also more differentially expressed genes included in the heat map, and these are different from the ones output by the tool when we did not correct for statistics.

Using the RNA-Seq analysis tools, we have been able not only to to detect batch effects, but also to control for these during subsequent analyses, allowing us to find strain-specific differences that were much smaller than differences due to gender or sample preparation.
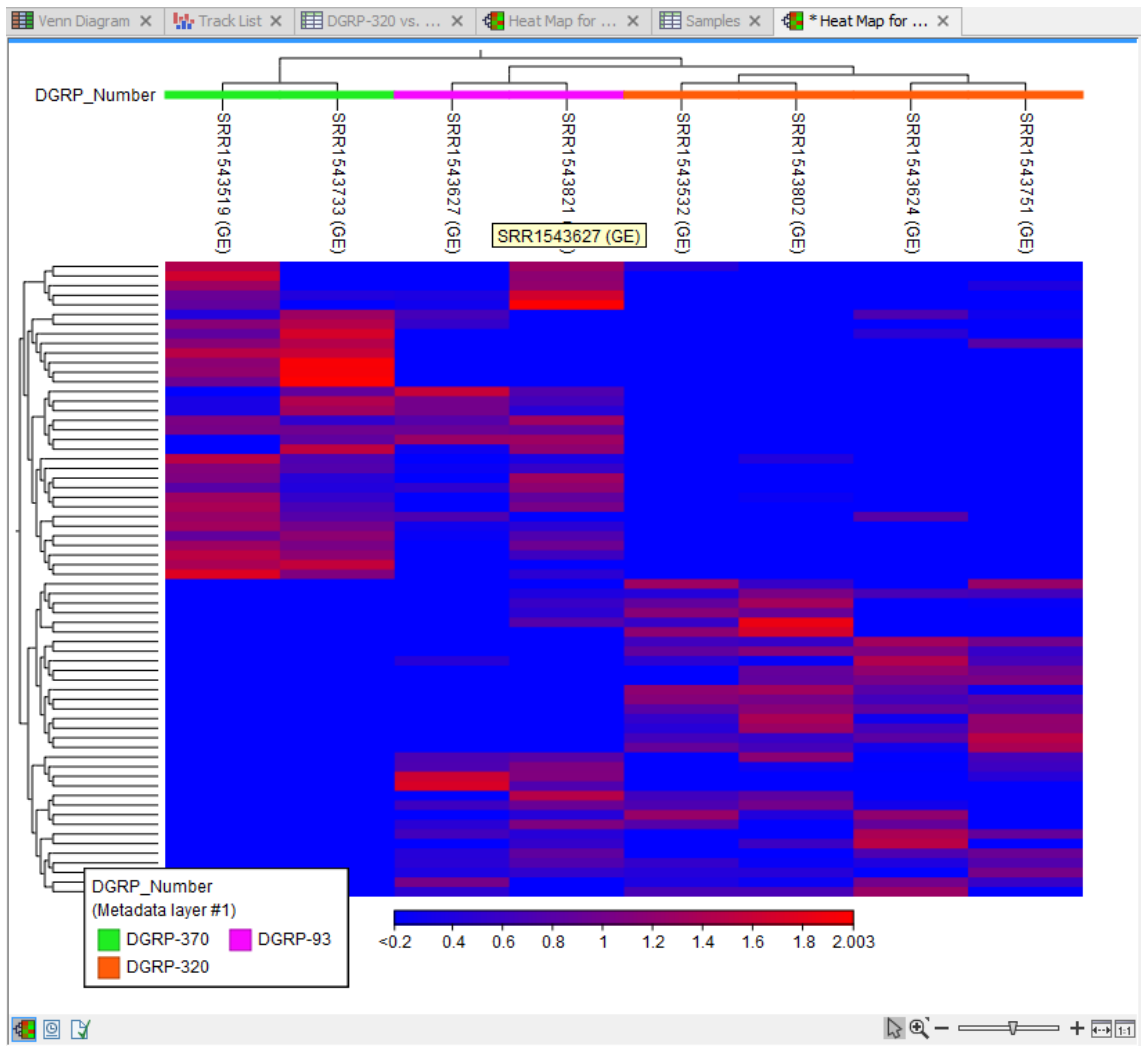
Tutorial

Figure 27: *The heat map after correcting for the sample preparation method bias.*

Bibliography

[Lin et al., 2016] Lin, Y., Golovnina, K., Chen, Z.-X., Lee, H. N., Negron, Y. L. S., Sultana, H., Oliver, B., and Harbison, S. T. (2016). Comparison of normalization and differential expression analyses using rna-seq data from 726 individual drosophila melanogaster. *BMC genomics*, 17(1):1.

_____