CLC Academy

RNA-Seq - *CLC Genomics Workbench*

CLC bio
Finlandsgade 10-12 · 8200 Aarhus N · Denmark
Telephone: +45 70 22 55 09 · Fax: +45 70 22 55 19
www.clcbio.com · info@clcbio.com

CLC Academy

Contents

1 RNA-Seq analysis part I: Getting started **3**

1.1 Downloading and importing the data 3

1.2 Running the RNA-Seq analysis 3

1.3 Interpreting the brain spikes analysis result 6

2 RNA-Seq analysis part II: Non-specific matches and expression measures **10**

2.1 Running the same data set with and without non-specific matches 10

2.2 Comparing the data in a scatter plot 11

2.3 The RPKM expression measure 15

3 RNA-Seq analysis part III: Exon discovery **17**

3.1 Creating two samples for comparison 17

3.2 Identifying new and differentially expressed splice isoforms 18

4 RNA-Seq analysis part IV: Spikes and quality control **21**

4.1 Inspecting the spike reads 21

4.2 Checking within and between group variability 22

CLC Academy

Chapter 1

RNA-Seq analysis part I: Getting started

This tutorial is the first part of a series of tutorials about RNA-Seq. The aim of the tutorials is to take you from start to end of an RNA-Seq analysis including mapping of reads, interpreting results, checking quality and finally doing statistical analysis. Along the way, we will focus on illustrating the effect of the parameters and choices made during the analysis.

The data used is from a study reported in [**?**]. The data set consists of RNA-Seq data from three types of Mouse tissue: Brain, Liver and Skeletal muscle. Each of the tissues has been sampled twice, so there are 6 samples all in all.

1.1 Downloading and importing the data

At http://www.clcbio.com/ngsexampledata you find the following data:

Subset of the full data set This file can be imported using the standard import and includes a subset of the full data set including a region of chromosome 16 for use as a reference. When running the full data set, we extracted all the reads that matched the genes of this part of chromosome 16. Download and import this data set (using the normal import) for use in these tutorials.

Experiments with the full data set Later on, we will work on experiments generated from the full data set. Download and import this data set (using the normal import) for use in these tutorials.

Once downloaded and imported, you should have the following folders and data in the **Navigation Area** (see figure 1.1).

1.2 Running the RNA-Seq analysis

Now, you can start the actual analysis. First step is to transform the list of reads into what we call an RNA-Seq sample which is basically a list of genes with expression measures. To do this, go to:

Toolbox | **High-throughput Sequencing** ( ) | **RNA-Seq Analysis** ( )

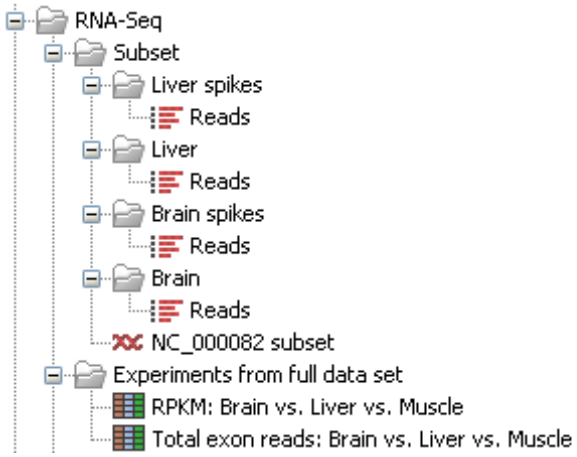


Figure 1.1: *The subset of the full data set has been imported together with the experiments generated from the full data set.*

This opens a dialog where you select the sequencing reads from the *Brain spike* sample, as shown in figure 1.2.

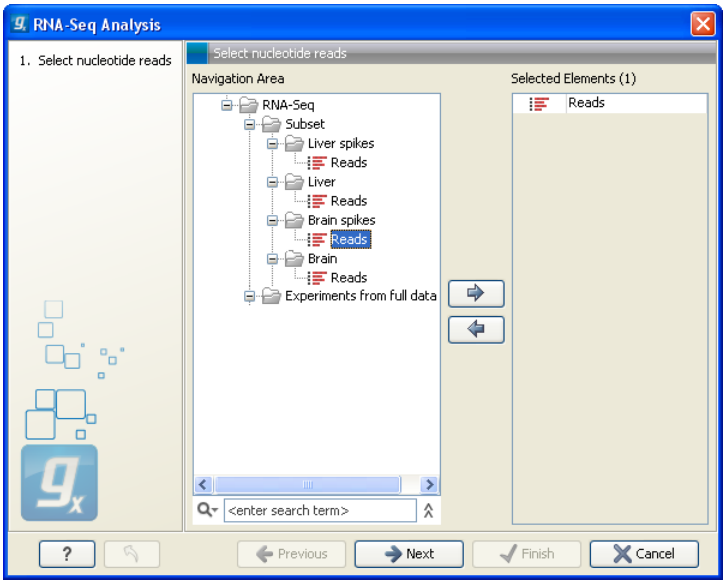


Figure 1.2: *Selecting the Brain spikes sample for RNA-Seq analysis.*

Click **Next** when the data is listed in the right-hand side of the dialog.

You are now presented with the dialog shown in figure 1.3.

Since we are using (part of) the ref-seq annotated mouse genome, choose **Use reference with annotations**. Click ( ) to select the reference sequence *NC_000082 subset*.

The last part of the dialog concerns parameters for the mapping. Leave these settings at their default - we will focus on these later on. (You can set the parameters to default by clicking the button ( ) at the bottom of the dialog, but note that this will also remove the selection of the reference sequence).

Clicking **Next** will show the dialog in figure 1.4.

The choice between **Prokaryote** and **Eukaryote** is basically a matter of telling the Workbench

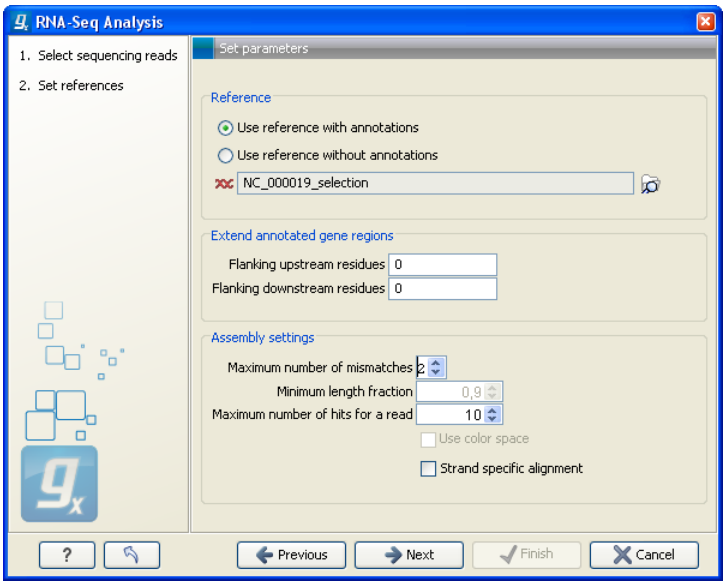


Figure 1.3: *Choosing the annotated reference sequence.*

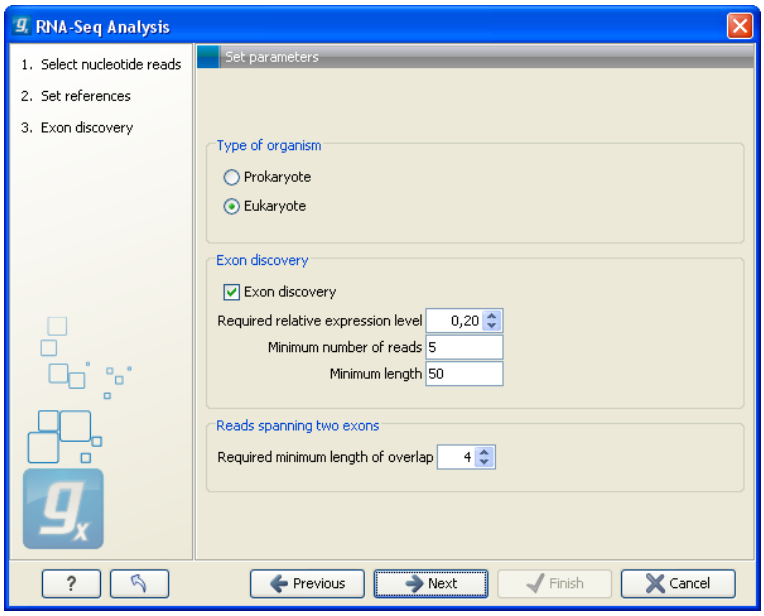


Figure 1.4: *Exon discovery.*

whether you have introns in your reference. In order to select **Eukaryote**, you need to have reference sequences with annotations of the type mRNA (this is the way the Workbench expects exons to be defined). The reference sequence provided with this tutorial includes mRNA annotations (they are the green annotations), so you select **Eukaryote** in this wizard.

Below you can specify settings for discovering novel exons. We will investigate this in detail later on.

Clicking **Next** will allow you to specify the output options as shown in figure 1.5.

Uncheck the **Create list of un-mapped reads**, **Create report** and **Make log** and click **Finish**.

The standard output is a table showing mapping statistics on each gene.

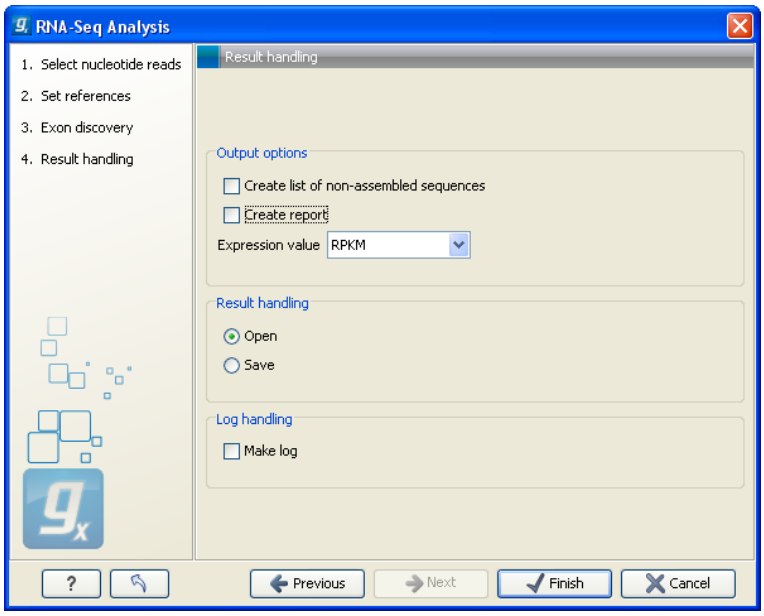


Figure 1.5: *Selecting the output of the RNA-Seq analysis.*

1.3 Interpreting the brain spikes analysis result

The result of the RNA-Seq analysis is shown in figure 1.6.

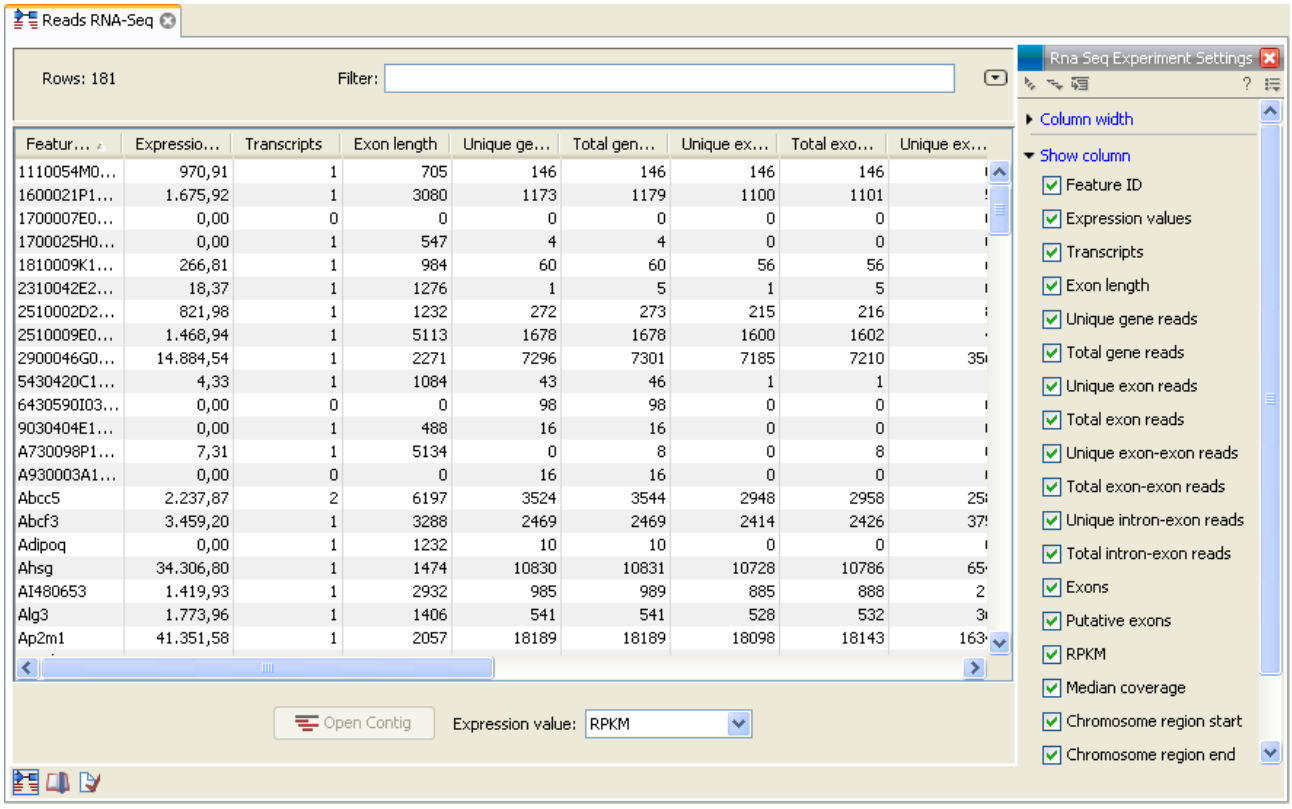


Figure 1.6: *The reads mapped to the Ashg gene.*

The **Expression values** column is now based on the RPKM. Change the measure to use **Total**

exon reads instead by clicking at the bottom of the view (we will go into more details with expression measures in part II). Now sort the table on the new expression value by clicking the column header twice. Find the *Ahsg* gene (4th from the top of the list) and double-click.

When the result is opened, you need to do a few customizations to make the view better suited for interpretation. In the **Side Panel**, under **Sequence layout**, set **Compactness** to Compact. Below, under **Text format**, set the font size to small or tiny. To save these customizations so that they take effect next time you open a mapping, click the **Save/Restore Settings** button ( ) at the top of the **Side Panel** and click **Save Settings**. Give your settings a name and make sure the check box to **Always apply these settings** is checked.

Double-click the tab of the view (or press Ctrl + M) to **Maximize the view** and click **Fit Width** ( ) in the tool bar to zoom out to see the full gene. You should now have a view similar to figure 1.7.

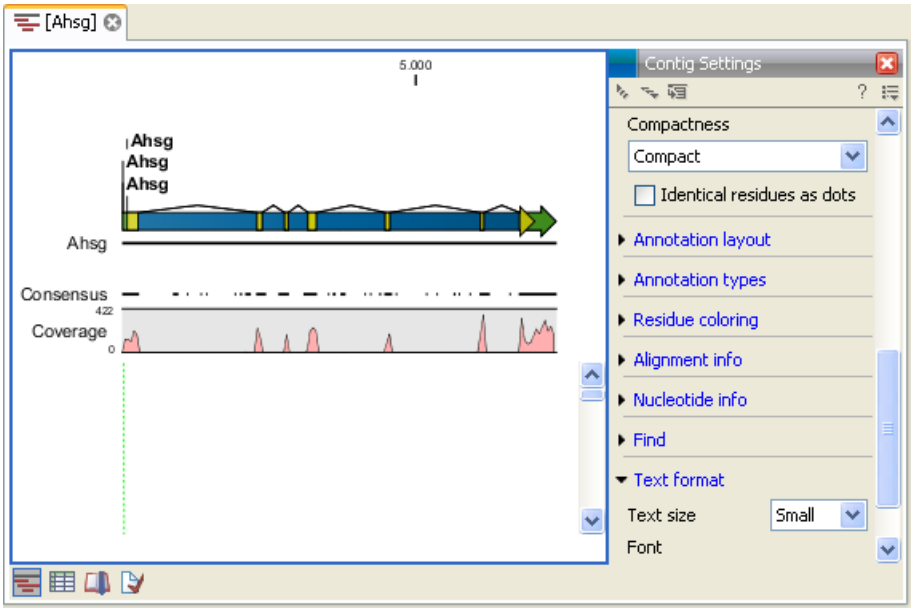


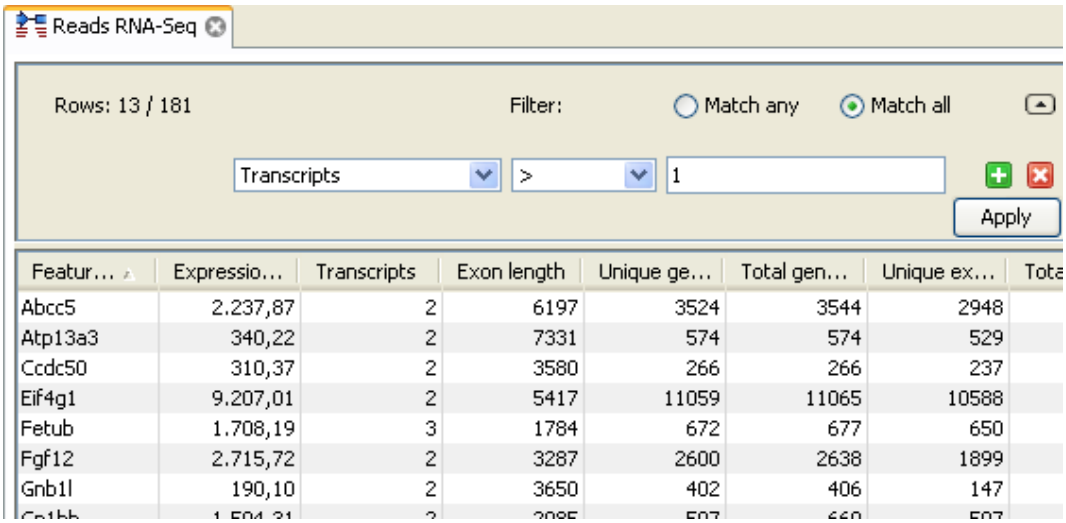
Figure 1.7: *The reads mapped to the Ashg gene.*

You can now see distinct peaks of coverage below the exons which are marked in green. Scroll slowly down on the scroll bar at the right hand side of the view. You will begin to see reads that have been mapped across exon-exon boundaries.

Click **Zoom in** ( ) and click-and-drag a rectangle around one of the exons. In this way you can zoom in to see more details of a particular exon. If you zoom all the way in, you will be able to see the nucleotide level and the alignment of the reads (to see the actual nucleotide level of the reads, you need to set the **Compactness** settings back to **Not compact**).

Close the view and go back to the RNA-seq sample. In the 'Transcripts' column you can see that the *Ahsg* gene only has one transcript annotated. Use the **Advanced filter** ( ) at the upper right hand part of the RNA-seq sample table view) to identify genes with more than one transcript annotated (set the filter to `Transcripts > 1` and press **Apply** as shown in figure 1.8).

The *Fetub* gene has three transcripts annotated. Open the mapping for this gene and press **Fit width** ( ) to zoom out completely and get an overview of the mapping to this gene. In the Side Panel under **Annotation Layout**, set the **Offset** to **Least offset**.



Featur...	Expressio...	Transcripts	Exon length	Unique ge...	Total gen...	Unique ex...	Tota
Abcc5	2.237,87	2	6197	3524	3544	2948	
Atp13a3	340,22	2	7331	574	574	529	
Ccdc50	310,37	2	3580	266	266	237	
Eif4g1	9.207,01	2	5417	11059	11065	10588	
Fetub	1.708,19	3	1784	672	677	650	
Fgf12	2.715,72	2	3287	2600	2638	1899	
Gnb1l	190,10	2	3650	402	406	147	

Figure 1.8: *Using the advanced filter to only show genes with more than one annotated transcript.*

One of the three transcripts annotated for *Fetub* uses a different first exon from the other two transcripts. There is no coverage in this exon at all, and thus no evidence for expression of the alternative first exon isoform. The other two transcripts have the same first exon but one skips the second exon of the other. If you scroll down you see both reads that span from exon 2 to exon 3 and reads that span from exon 2 to exon 4. Thus, there is evidence for both of these splice variants (see figure 1.9).

Close the view and you are ready for part II: Non-specific matches and expression values.

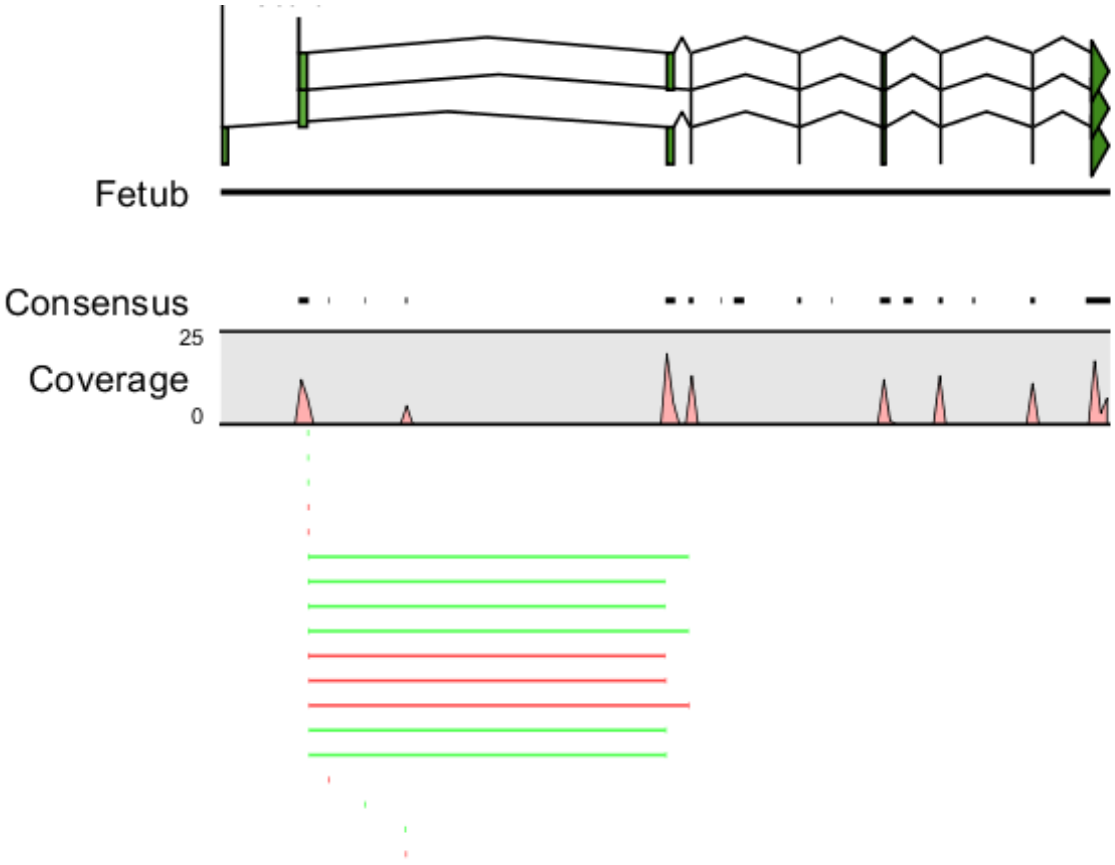


Figure 1.9: *Reads showing evidence for expression of two isoforms.*

Chapter 2

RNA-Seq analysis part II: Non-specific matches and expression measures

This tutorial is the second part of a series of tutorials about RNA-Seq analysis. We continue working with the data set introduced in the first tutorial.

In this tutorial we will first explain how non-specific matches are treated, and second we will explain the effect of using different expression measures.

2.1 Running the same data set with and without non-specific matches

Imagine a situation where you have nine reads that match equally well on two different genes. Since it is not possible to tell which transcript the reads actually came from, the Workbench has to decide where to place them. Based on other reads that are matched *uniquely* to the genes, the Workbench estimates the expression of each gene and use that as a weight to distribute the reads. In a situation where one of the genes have twice the number of unique matches, it will on average receive six of the nine reads whereas the other one would get three.

Now, we will show the effect of including these non-specific matches in the analysis. In the analysis of the first tutorial, the **Maximum number of hits for a read** was set to 10. This means that all reads that match in 10 or fewer places will be included in the mapping, with multi-hitting reads being distributed as described above. Now, run a new **RNA-Seq Analysis (** **)** on the *Brain spike* sample, but this time set the **Maximum number of hits for a read** to 1. You find this setting in step 2 - leave the rest of the settings as they are.

You will now end up with an RNA-Seq sample where all reads matching in more than one position are excluded. If you go to the **History** ( ) view of this new sample, you can see how many reads were not mapped. If you compare these numbers, you can see the first sample has 214631 unmapped reads whereas the second run without multihit reads has 229011 unmapped reads (figure 2.1).

Save ( ) the two samples with meaningful names, e.g. *including non-specific* and *only specific*.

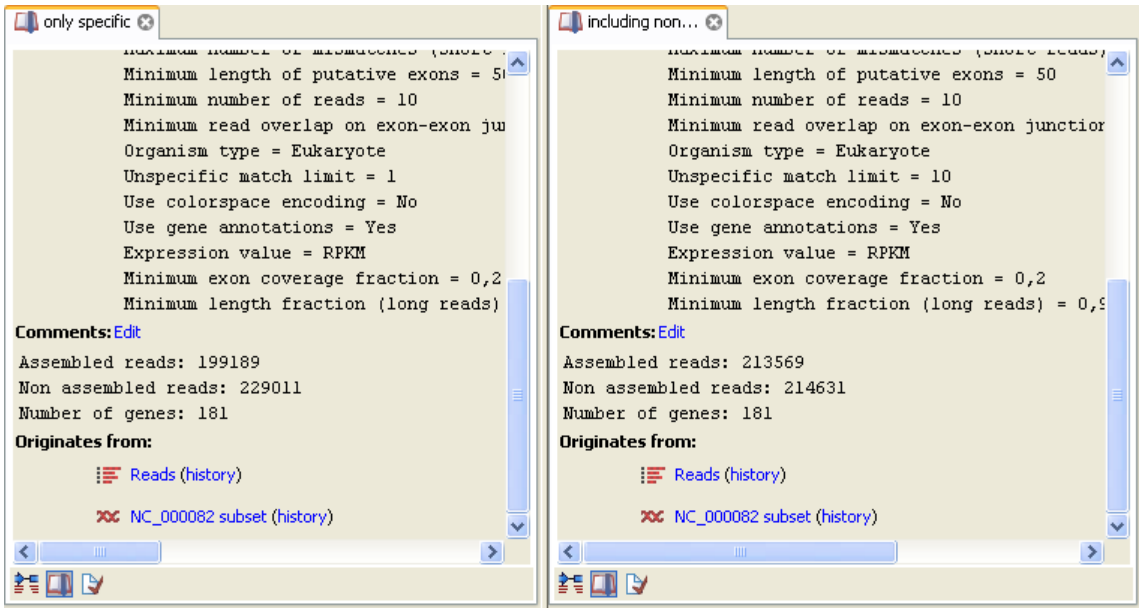


Figure 2.1: *Comparing the history entries for the two samples.*

2.2 Comparing the data in a scatter plot

Now, we want to see what this difference means in terms of the expression values. In order to compare the two samples, we set up an experiment:

Toolbox | **Expression Analysis** ( ) | **Set Up Experiment** ( )

Select the two RNA-Seq samples ( ) that you have just saved and click **Next**. Choose an un-paired, two-group experiment, set the **Value to use in experiment** to to **Total exon reads** and click **Next**. Name the groups *including non-specific* and *only specific* and click **Next**. Right-click each of the samples and assign it to the appropriate group. Click **Finish**.

You will now have an experiment based on the two samples. We will go into more details with the experiment later - for now we are interested in looking at the scatter plot. Click the **Scatter plot** ( ) icon at the bottom of the view.

At the bottom of the **Side Panel** you select the values to plot. Select *including non-specific Total exon reads* versus *only specific Total exon reads*, and you will see a view as shown in figure 2.2.

The scatter plot now shows the expression levels of the two samples. Since the RNA-Seq analysis was run on the same data set with the only difference being the treatment of non-specific matches, you can now see the direct effect of using and distributing the non-specific matches in this way.

You can now see that many of the genes have close to identical expression measures (they are located along the x=y line in the plot), but there are some that show higher expression in the sample including non-specific matches. To see the outliers more clearly, set the **Dot type** under **Dot properties** in the **Side Panel** to **Dot**.

The most outlying gene is *Sept5*. If you place your mouse on the dot as shown in figure 2.3, you can see the feature ID (gene name) and the x and y values of the dot.

Open the *including non-specific* RNA-Seq sample and locate this gene by typing *Sept5* in the filter

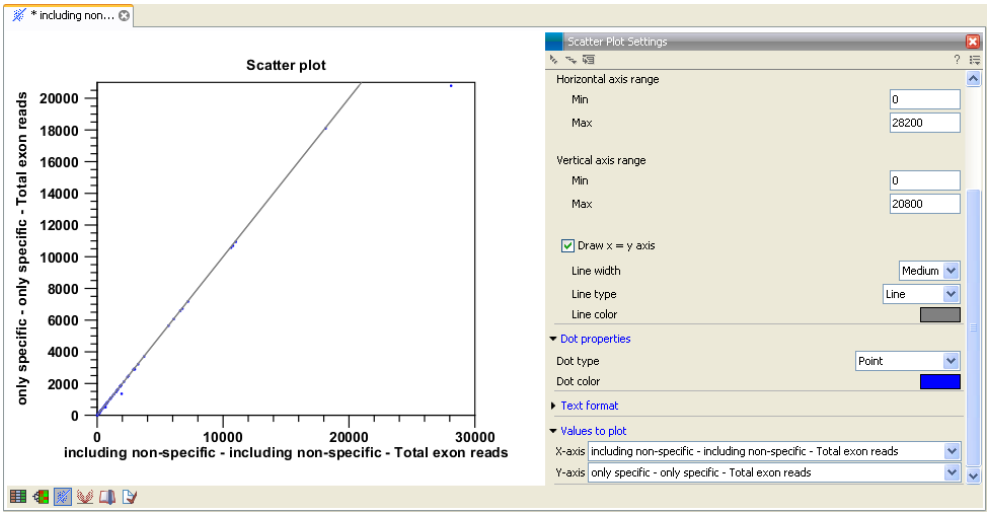


Figure 2.2: *A scatter plot showing the effect of including non-specific matches in the expression measure.*

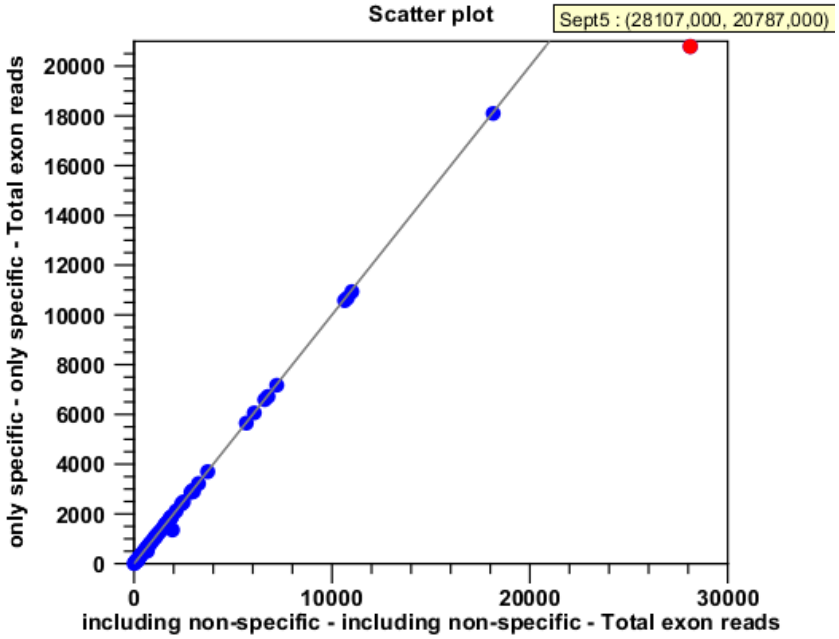


Figure 2.3: *Gene Sept5 is one of the genes showing a notable difference in expression.*

at the top. Double-click to open the mapping. When you zoom out ( ) and scroll along to the end of the gene, you will see a lot of reads that are yellow. Yellow is the color used for non-specific reads. In this case, all these yellow reads are the ones contributing to a higher expression measure when you compare in the scatter plot.

By looking at the gene annotations, you can also see the reason why there are so many non-specific matches. As shown in figure 2.4, there is an overlapping gene near the end. This means that all the reads that map to this part of the *Sept5* gene also map equally well to the beginning of the *Gp1bb* gene. These reads are then treated as non-specific matches.

If you opened the *Gp1bb* mapping you would also see the non-specific reads at the beginning where it overlaps with **Sept5**. Because we can see the overlap, we know why we have non-specific matches, but it could be that these reads would also match other places on the reference. It's

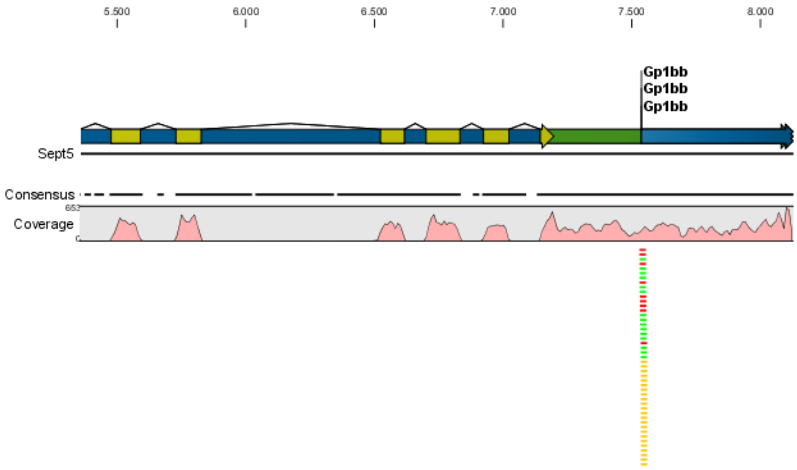


Figure 2.4: *Gp1bb is overlapping the end of the Sept5 gene.*

easy to check if the same region is present other places in the reference by conducting a BLAST search:

select the relevant part of the gene|right-click the selection|BLAST against Local Data ( )

This is illustrated in figure 2.5

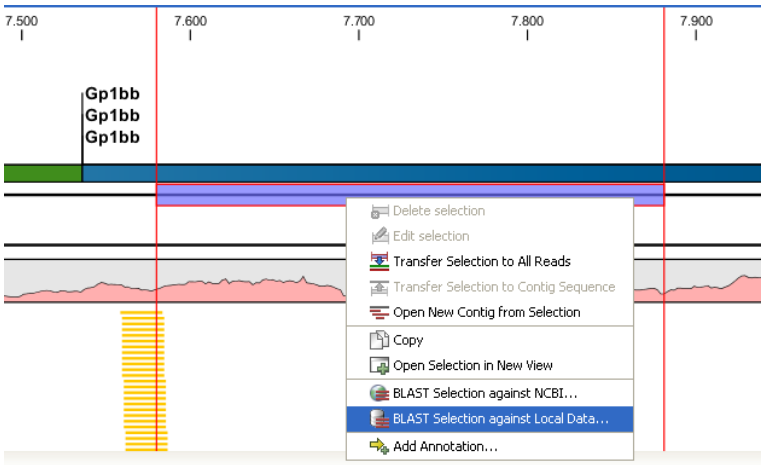


Figure 2.5: *BLAST against the reference.*

Close the BLAST dialog again since we are not going into details with BLAST search in this tutorial. Please read the BLAST tutorial to learn more about BLAST in *CLC Genomics Workbench*. The idea with BLASTing this selection would be to use the reference sequence as target and see how many hits you would find. In this case there is only one good hit, but if you have a region of non-specific matches that are not due to overlapping genes, you can use this approach to try to identify which other gene is "competing" for these reads.

Close ( ) the mapping view, go back to the experiment and switch to the table view ( ). Enter *Gp1bb* in the filter and click with your mouse on the *Gp1bb* gene. Switch back to the scatter plot ( ) and *Gp1bb* will now be high-lighted with a red color. Click **Zoom in** ( ) and click a couple of times on the gene to zoom in (see figure 2.6).

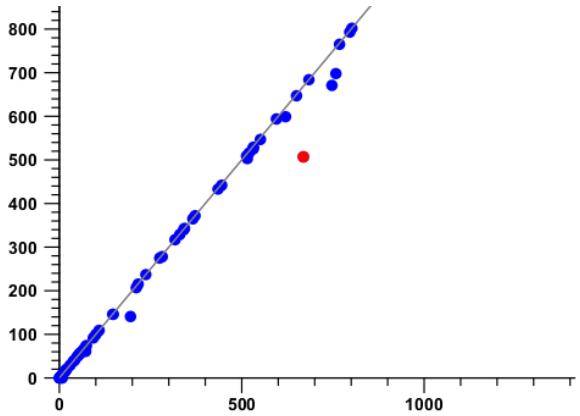


Figure 2.6: *Zooming in on the Gp1bb gene in the scatter plot.*

You can now see that this gene also exhibits differential expression between these two samples, but to a lesser degree than the *Sept5* gene. Open the *Gp1bb* mapping from the *including non-specific* sample, and you can see that there are fewer yellow reads than in the *Sept5* mapping. As explained above, the non-specific reads are distributed according to the number of unique reads, and when you compare the two results, it is evident that there are many more unique reads in the *Sept5* gene (you can easily see the difference in the RNA-Seq result table as shown in figure 2.7).

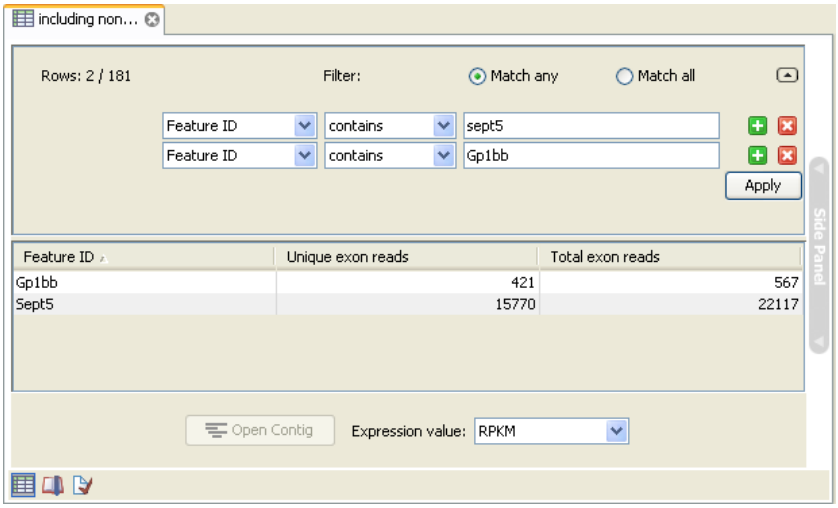


Figure 2.7: *Comparing the number of unique reads between Gp1bb and Sept5.*

From the scatter plot in figure 2.2, it is obvious that the decision on whether to include non-specific matches or not is very important. For some genes, the difference in expression is highly significant. This trend becomes even more evident when looking at the full data set where the proportion of non-specific matches is even higher (with the full reference transcriptome, there is a greater chance of finding sequences that are represented more times, e.g. arising through gene duplications).

It is hard to make general recommendations on how to treat non-specific matches. One of the pitfalls when including non-specific matches is that the number of unique matches can be too low to ensure a reliable distribution of the non-specific matches. One way of approaching this problem would be to run the same data set with different settings as shown in this tutorial. That will enable you to perform random checks of the genes whose expression is significantly altered,

and you will be able to identify this kind of pattern. On the other hand, if you completely disregard non-specific reads, you may underestimate the expression levels of genes in gene families.

We refer to [?] for an in-depth discussion of this topic.

2.3 The RPKM expression measure

2.3.1 Normalizing for sample size

The observations made from figure 2.2 lead to another important consideration when dealing with RNA-Seq analysis: you have to decide which expression measure you want to use. When you have several samples (as in this study with six different samples), these will have different numbers and qualities of reads. You will often see that there is quite a big difference between the samples in the number of reads that can be matched. This means that it can be hard to compare the expression of the same gene in different samples simply by looking at the number of reads matched (i.e. total exon reads). When comparing the groups *including non-specific* versus *only specific total*, you can see this effect too, since they have 213,569 and 199,189 mapped reads, respectively. This means that you have an asymmetry in the scatter plot when using total exon reads as the expression measure (see we could see in figure 2.2).

There is another expression measure, RPKM (Reads Per Kilobase of exon model per Million mapped reads), which seeks to normalize for the difference in number of mapped reads between samples. We will now investigate RPKM in greater detail. Go back to the scatter plot in figure 2.2. Change the values to be plotted from total exon reads to RPKM for the two samples. You should now see a scatter plot as shown in figure 2.8.

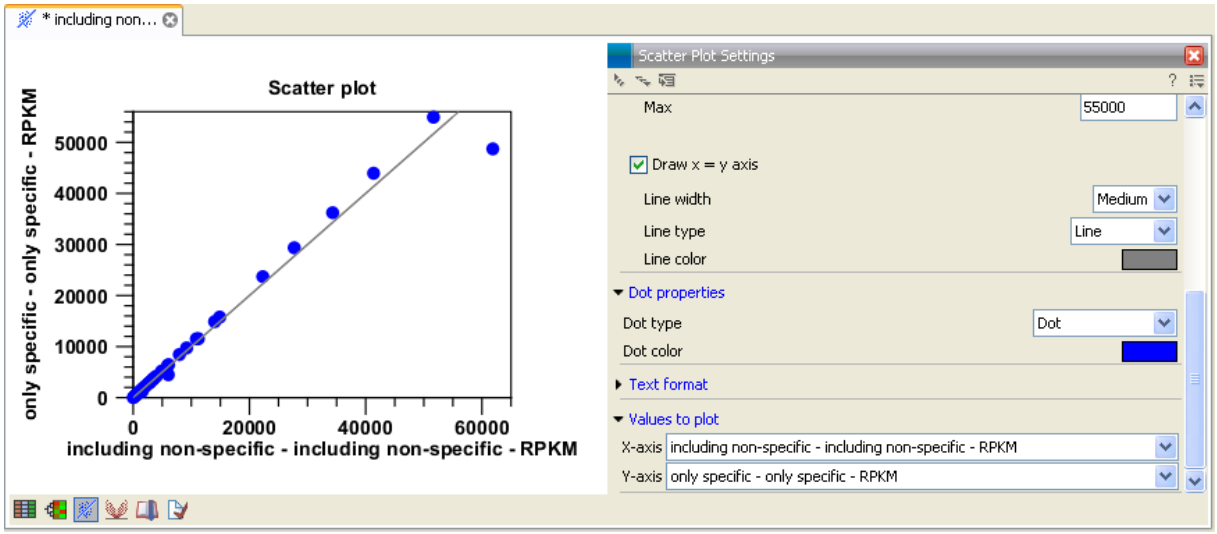


Figure 2.8: *The effect of including non-specific reads compared using RPKM.*

Where figure 2.2 showed either dots falling on the $x = y$ axis or below, you now see dots falling primarily slightly above $x = y$ axis or below. This is because the RPKM takes into account that the total number of mapped reads is higher in the *including non-specific* sample than in the *only specific* sample. RPKM is defined as $RPKM = \frac{total\ exon\ reads}{mapped\ reads(millions) \times exon\ length\ (KB)}$.

Let's investigate two of the genes in the scatter plot. First, identify the two genes at the top of the scatter plot — one above the $x = y$ axis, one below. One of them is the *Sept5* gene that we have previously investigated. This still shows higher expression in the *including non-specific sample*

because of the high number of non-specific matches. The other gene is *Sst*. Switch back to the experiment table ( ) and compare the total exon reads for both samples (you can deselect sample columns under **Sample level** in the **Side Panel**, that will ease the overview). The value is 6600 and 6514, respectively, so the number of total exon reads is almost identical. What is then the reason that the PRKM value is higher for the *only specific* sample? This is because this sample has a lower number of mapped reads, and the RPKM will thus be higher (see definition of RPKM above).

2.3.2 Normalizing for transcript length

In a sample, if two transcripts are present in the same number of copies, and the sequencing is unbiased, you would expect the same number of reads from each transcript. But if one transcript is short and the other is long, you would expect the long transcript to yield more reads. So if you wish to compare the expression of transcripts within the same sample, you need to take the transcript length into account.

If you look at the definition of RPKM above, you can see that besides number of mapped reads, the *exon length* is also considered. The idea behind this is to make it possible to compare expression levels of different transcripts.

Open the *only specific* sample and sort the table on total exon reads, you can see that the genes *Abcc5* and *Comt* (number 15 and 16 from the top) have almost the same number of reads (2,941 and 2,895). However, their expression value measured in RPKM is 2,369.95 and 11,547.01, respectively (see figure 2.9).

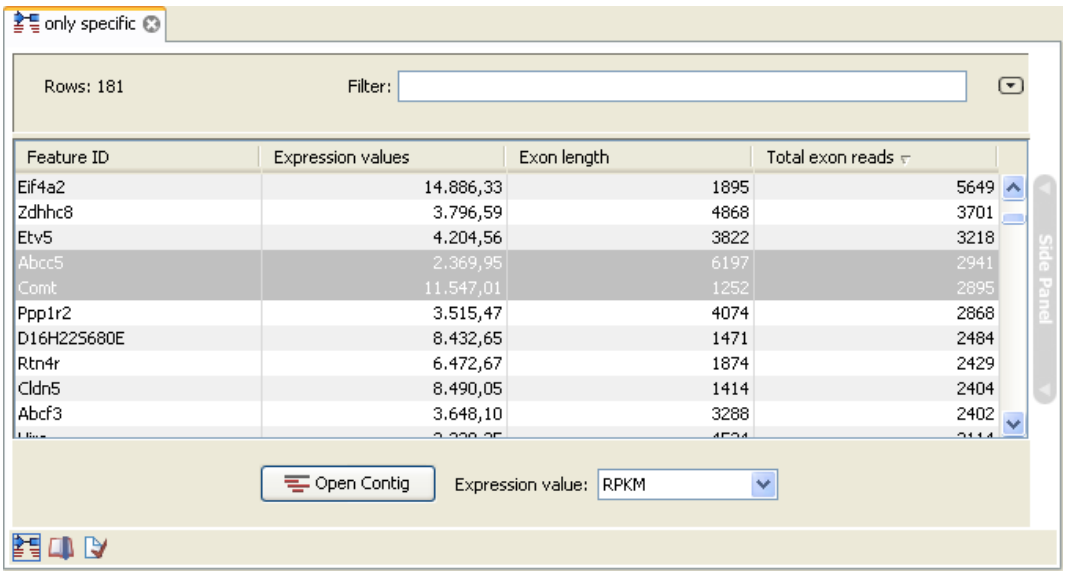


Figure 2.9: *Nearly the same number of total exon reads for two genes leads to widely different RPKM values because of the difference in transcript lengths.*

This is simply due to the difference in transcript length which you can also see in the table under **Exon length** (which sums the lengths of all the annotated exons).

Close all open views, save the experiment, and you are ready for part III.

Chapter 3

RNA-Seq analysis part III: Exon discovery

This tutorial is the third part of a series of tutorials about RNA-Seq analysis. We continue working with the data set introduced in the first tutorial.

In this tutorial we will focus on discovery of new putative exons.

3.1 Creating two samples for comparison

First, we will use two samples to find new putative exons expressed in one sample but not the other. In the previous tutorial, you have already created a sample called *including non-specific*. Rename this to *Brain spikes*. Then run three new RNA-Seq analyses with identical parameters but based on the *Brain*, *Liver spike* and *Liver reads*. Remember to change the **Maximum number of hits for a read** value back to 10.

Save the results in the appropriate folders, and you should now have a folder structure like the one shown in figure 3.1.

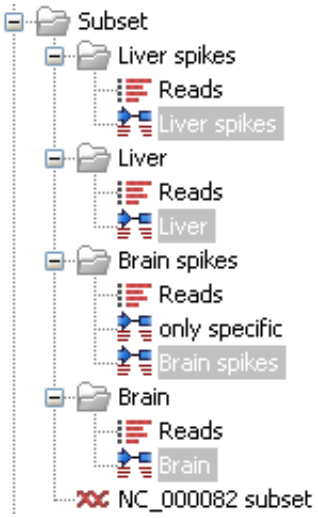
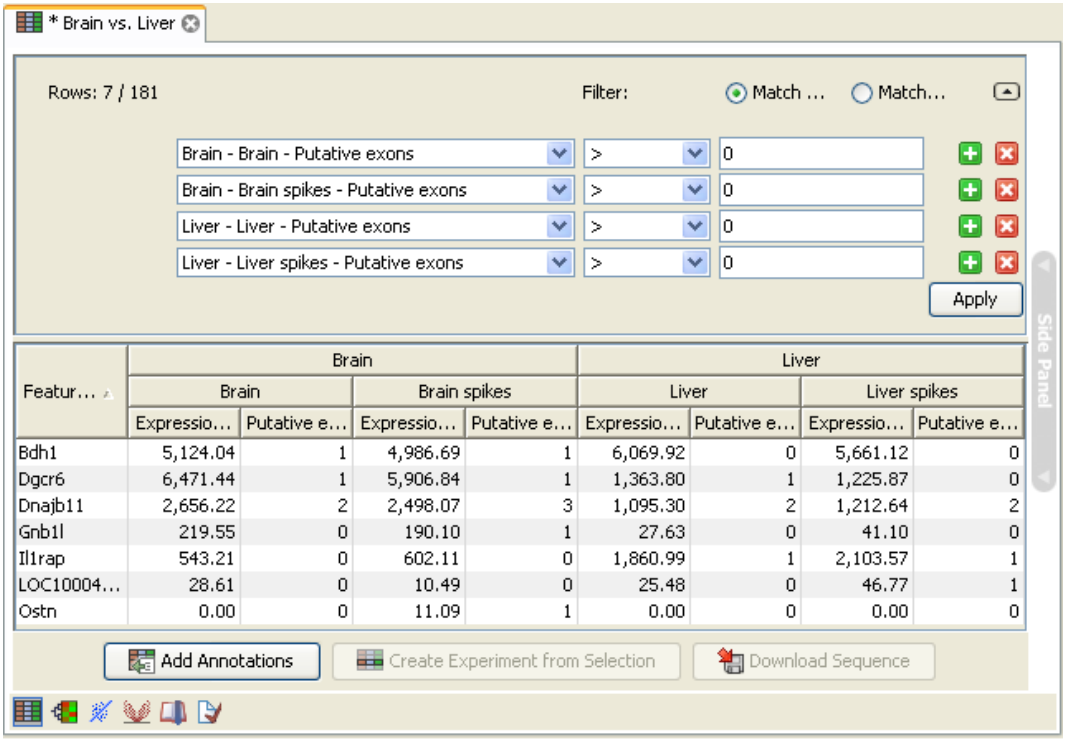


Figure 3.1: *Four samples have now been created.*

Now, set up an experiment with two groups, *Brain* and *Liver* and assign the four samples to the appropriate groups. Next, adjust the settings in the **Side Panel** and the **Advanced filter** ( ) to

look like figure 3.2.



Featur...	Brain				Liver			
	Brain		Brain spikes		Liver		Liver spikes	
	Expressio...	Putative e...	Expressio...	Putative e...	Expressio...	Putative e...	Expressio...	Putative e...
Bdh1	5,124.04	1	4,986.69	1	6,069.92	0	5,661.12	0
Dgcr6	6,471.44	1	5,906.84	1	1,363.80	1	1,225.87	0
Dnajb11	2,656.22	2	2,498.07	3	1,095.30	2	1,212.64	2
Gnb1l	219.55	0	190.10	1	27.63	0	41.10	0
Il1rap	543.21	0	602.11	0	1,860.99	1	2,103.57	1
LOC10004...	28.61	0	10.49	0	25.48	0	46.77	1
Ostn	0.00	0	11.09	1	0.00	0	0.00	0

Figure 3.2: *Comparing putative exons between the groups.*

To specify the **Advanced filter**, click the button ( ) at the top right corner of the view and click the **Add new filter criterion** ( ) button to add more criteria to the filter. Notice that we want all genes that have putative exons in at least one sample so we choose the **Match any** option for the filter.

This view enables you to identify the patterns of putative exons across all the samples. You can see that there is not complete consistency between the replicates in each tissue - the brain spikes sample has in general more putative exons than the other brain sample. This may be due to a difference in the number of mapped reads for the two samples (213,297 vs 117,302). When there are less assembled reads, there will typically be less evidence for a putative exon, since you specify a minimum number of reads for a putative exon when you run the RNA-Seq analysis (default is 10).

3.2 Identifying new and differentially expressed splice isoforms

If you take the gene at the top of the experiment, *Bdh1*, you can see that there is a putative exon in both of the brain samples but none in the liver samples. Now open the *Brain spikes* RNA-Seq result ( ) and open the *Bdh1* mapping. Zoom out and you will be able to see the putative exon as shown in figure 3.3.

If you zoom more in, you can see a very clear signal with many reads mapping in this region. The coverage level is almost identical between this and the other exons in the gene, except the first one which has very low coverage. This could indicate an alternative start site of this transcript and it seems like this new isoform is prevalent in the brain tissue.

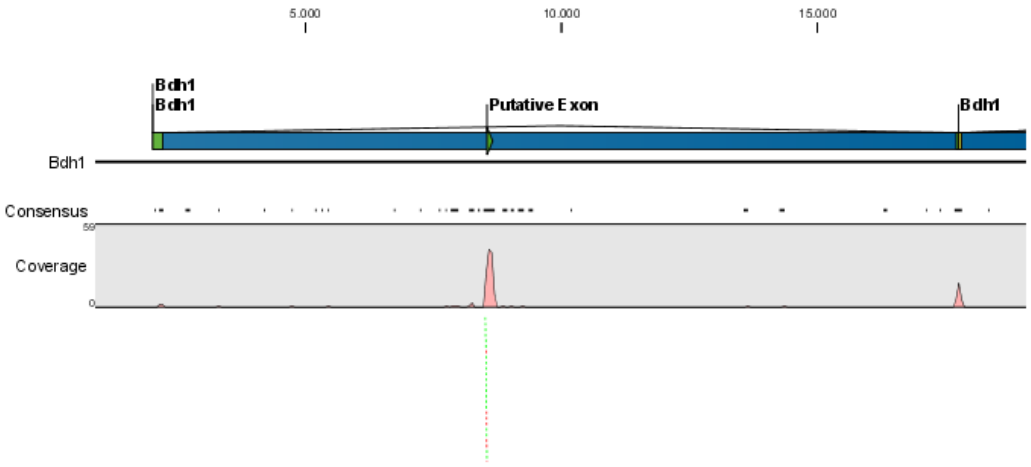


Figure 3.3: *A putative exon in the brain sample.*

Open the liver spikes sample and look at the *Bdh1* mapping (see figure 3.4).

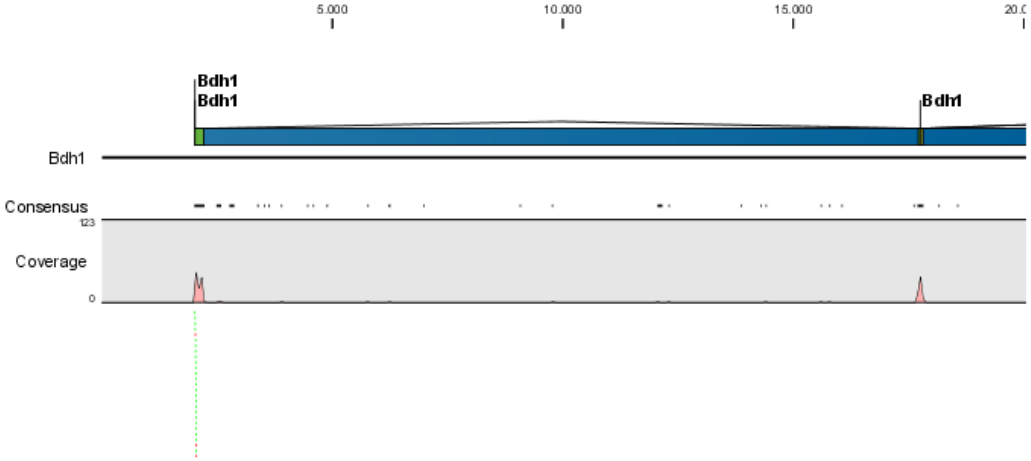


Figure 3.4: *In the liver sample, there is no putative exon.*

Notice that the peak seen in the putative exon region of the brain sample is completely absent in the liver sample (use the position on the reference sequence to guide you - the peak should have been around position 8500. You can also use the **Search** field in the **Side Panel** to find this position). In both samples there are reads spanning the junction between the first and second exons. When you look at the first exon, you can see that the liver sample has good coverage here which indicates that the splice isoform in liver is actually the one annotated, whereas, in addition to the annotated isoform, a new splice isoform may be expressed in the brain sample.

Note that the reads within putative exons do not count in the total exon reads number (and as a consequence not in RPKM either). Also, reads are only mapped to exon-exon junctions of annotated mRNA transcripts. If you want to include these reads in the analysis, you need to annotate the original reference sequence with this new transcript by adding a new mRNA annotation which is identical to the existing one but with the new exon included. We cannot be sure that this is right, but judging from the coverage levels of the other exons, it would be a plausible explanation. Then run the RNA-Seq analysis again using the modified reference sequence.

Close all open views, and you are ready for part IV.

CLC Academy

Chapter 4

RNA-Seq analysis part IV: Spikes and quality control

This tutorial is the fourth part of a series of tutorials about RNA-Seq analysis. We continue working with the data set introduced in the first tutorial, although in this tutorial we are no longer considering only a subset of the data.

In this tutorial we will focus on quality control. First, we will examine spike-ins in the data set, and second you will learn about general quality control tools.

4.1 Inspecting the spike reads

The data set includes six samples that come from three groups corresponding to the three different types of tissue. Within each group of samples, one sample has six spike-in genes, where controlled amounts of mRNA from *Arabidopsis* and phage lambda templates have been introduced in the sample prior to sequencing. We will now inspect the six spike-in genes and check if they are expressed in the spiked samples only as we would expect.

For this and the rest of the tutorial, we will work on the full data set. The data set is so large that on a powerful eight-core workstation computer with 32 GB RAM, it takes more than an hour to run one sample. Therefore, we have performed the analysis beforehand and included the final result with the files you downloaded in the first tutorial. The full RNA-Seq results with mappings of all the mouse genes take up too much space, so they are not included. Instead, we have set up experiments with all six samples.

Open the experiment in the *RPKM* folder and type "spike" in the filter to only include the six spike genes (when we ran the full analysis, we made an artificial sequence called "Spike chromosome" containing the six spike genes). Next, click the **IQR** column header to sort the genes according to the interquartile range of expression. You should now have a view similar to figure 4.1.

The important thing to check here is that we see expression of the spike genes in the spike samples but not in the rest. The six spike transcripts were added in different concentrations to the spike samples, and this is why we see the varying levels of expression.

Select all six rows, and switch to the scatter plot ( ) to visualize this trend. Choose to compare expression value of *Brain spikes* and *Brain*, and you will be able to see some red points in the

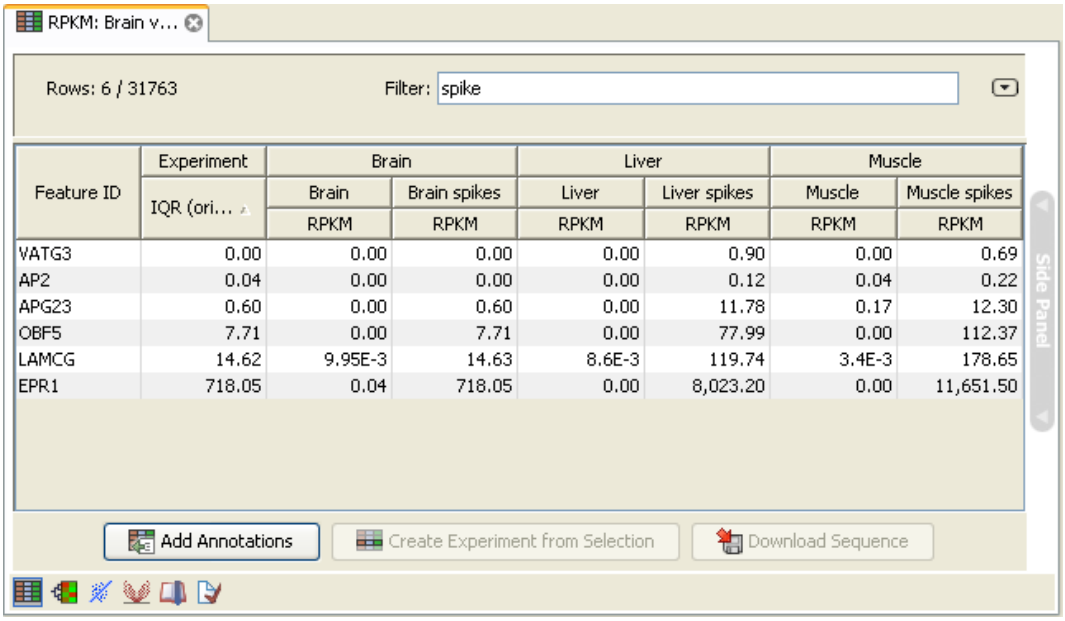


Figure 4.1: *Comparing the expression of the spikes.*

scatter plot. Change the **Dot type** to **Dot** in the **Side Panel** to see the red dots more clearly. The scatter plot is shown in figure 4.2.

At first glance, you only see one clear outlier at the bottom, whereas the other red dot (or rather, 'dots') seem to have more similar expression in the two samples (close to zero in both). But if you **Zoom in** ( ) on these dots, you can see that 3 of them are outliers, although to a smaller extent (see figure 4.3).

4.2 Checking within and between group variability

We have now confirmed that the spike controls look fine (you can check the other two groups in the scatter plot if you like). The next step in our quality control efforts is to check whether the overall variability of the samples reflect their grouping. In other words we want the samples from the same group to be relatively homogenous and distinguishable from the samples of the other groups.

4.2.1 Box plot

To examine and compare the overall distribution of the expression values in the samples you may use a **Box plot** ( ). Box plots can be used to get an overall impression of the locations of the distributions, and to some extend the spread of the distributions.

First, create a box plot based on the RPKM-based experiment:

Toolbox | **Expression Analysis** ( ) | **Quality Control** | **Create Box Plot** ( )

Select the experiment that is based on RPKM expression values and click **Next** and **Finish**.

The box plot is shown in figure 4.4.

The median and mean lines (you can display the mean in the **Side Panel**) show the median and mean expression values in the samples and the boxes extend from the p'th to the 100-p'th

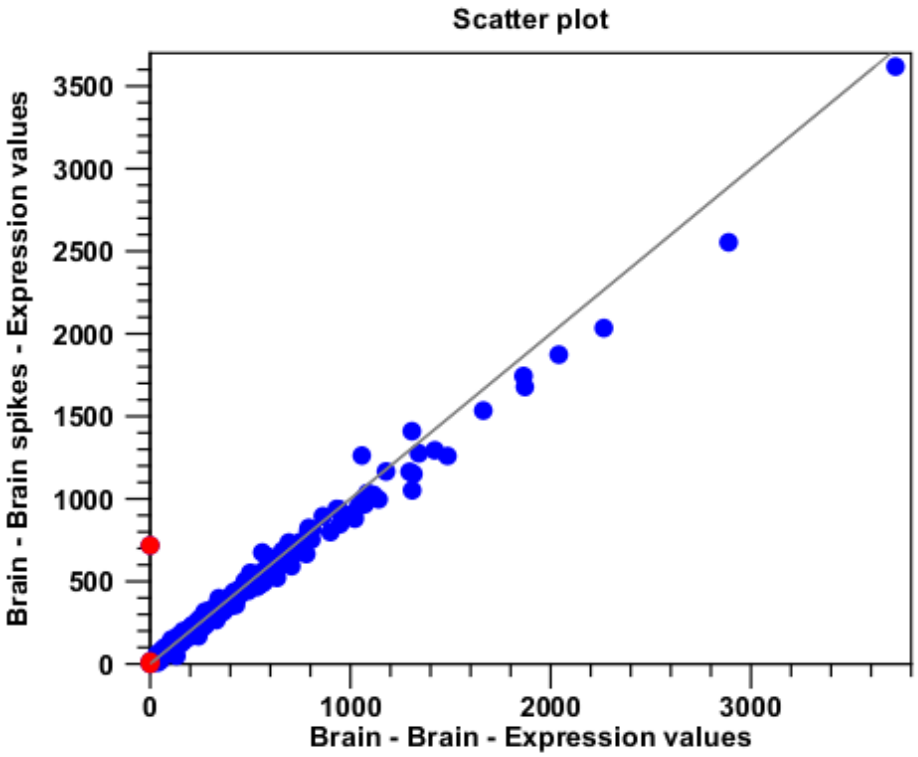


Figure 4.2: *Comparing the expression of the spikes in a scatter plot.*

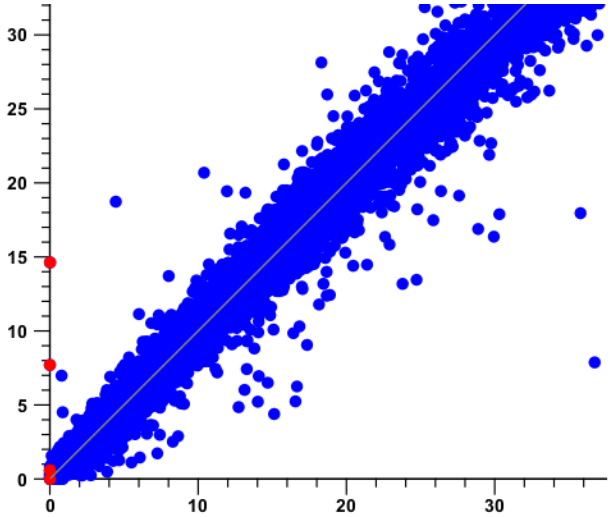


Figure 4.3: *Comparing the expression of the spikes in a scatter plot, zoomed in.*

percentile of the sets of expression values in the samples. Thus, the box plot view is rather sensitive to the choice of the percentile value, and you may get a better impression of how the distributions compare by trying different percentiles.

The distributions of the expression values are dominated by a lot of really small values and much fewer but much larger values. To diminish the effect on the box plots of the few very high values, you can square root transform the values and create box plots for the transformed values. First,

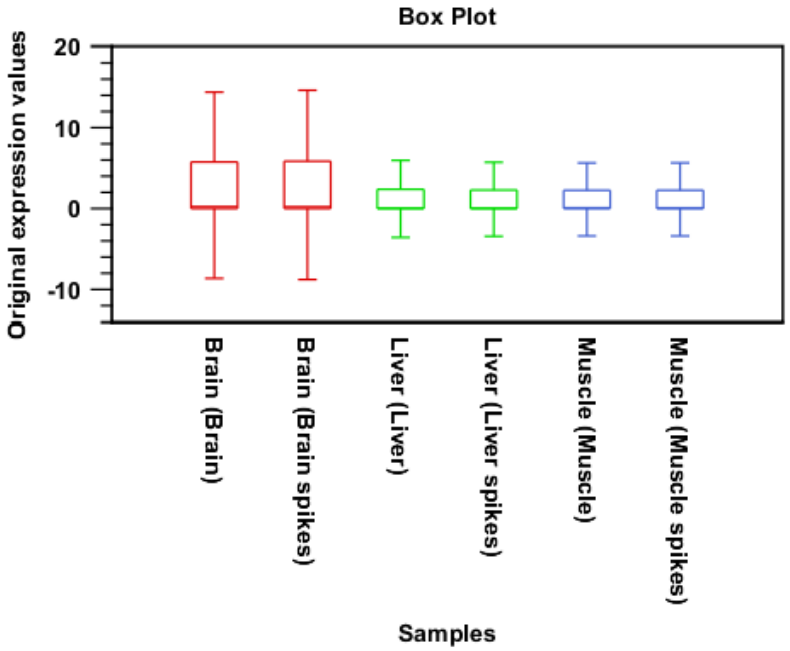


Figure 4.4: *A box plot of the 6 samples in the experiment, colored by group.*

transform the values in the experiment:

Toolbox | **Expression Analysis** ( ) | **Transformation and Normalization** | **Transform** ( )

Select the experiment, click **Next** and select **Square root transformation**. Click **Finish**. Now, create a new box plot, but this time make sure to select **Transformed expression values** in the second step. Figure 4.5 shows the box plot before and after square root transformation.

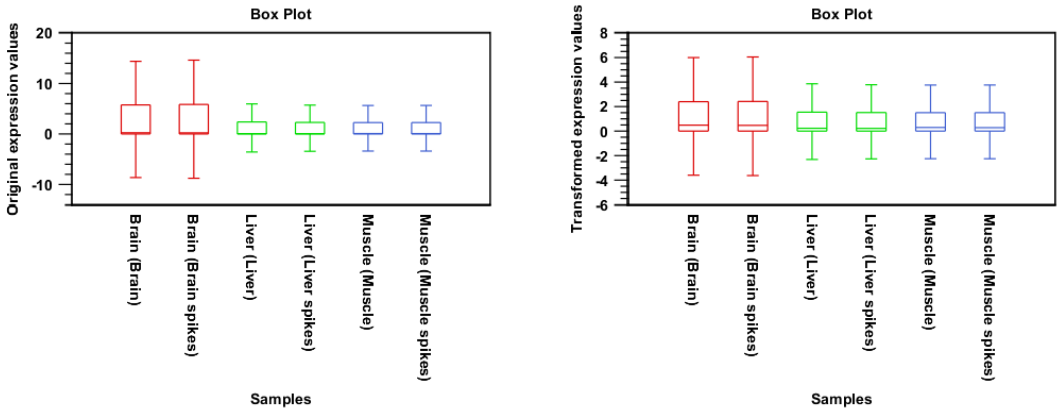


Figure 4.5: *To the left: original box plot based on RPKM values. To the right: box plot based on square root transformed expression values.*

After square root transformation, the distributions appear more a bit more similar.

Now, create similar box plots for the experiment based on total exon reads. You can see the result in figure 4.6 that again shows the box plot before and after square root transformation,

this time based on total exon reads rather than RPKM.

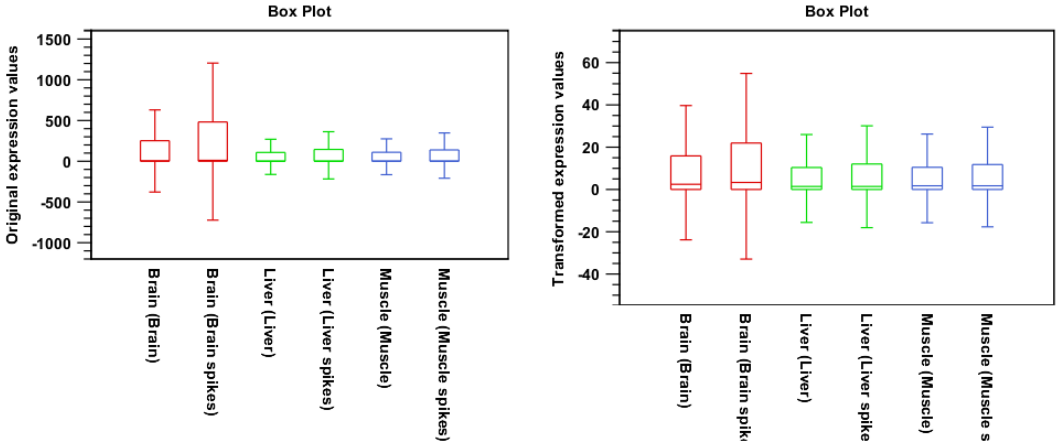


Figure 4.6: *To the left: original box plot based on total exon reads. To the right: box plot based on square root transformed expression values.*

Overall, the box plots indicate that the *locations* of the distributions of the expression values in the samples are similar both for RPKM and for Total Exon Reads, but there is considerable difference in the *spread* of the values. The distributions of RPKM values for samples for the same tissue are highly similar. This is expected, as the samples are technical replicates and the sample size is factored out in the RPKM (see part III of the tutorials). The high variability of the 'Total exon reads' counts is obviously related to the numbers of (mapped) reads in the samples.

4.2.2 Principal component analysis (PCA)

We continue to work with the experiment based on RPKM expression values. Next, we perform a **Principal Component Analysis (PCA)**:

> **Toolbox** | **Expression Analysis** ( ) | **Quality Control** | **Principal Component Analysis** ( )

Select the experiment, click **Next** and **Finish**. This will create a PCA plot as shown in figure 4.7.

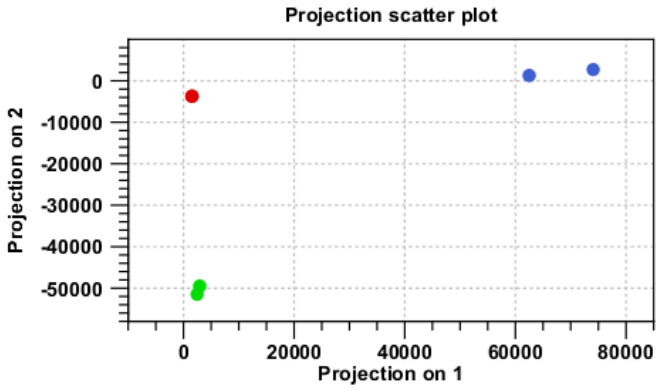


Figure 4.7: *A principal component analysis colored by group.*

The plot shows the projection of the samples onto the two-dimensional space spanned by the first and second principal component. (These are the orthogonal directions in which the data exhibits the largest and second-largest variability).

The dots are colored according to the groups, and they group very nicely in the plot (the two red red dots are on top of each other).

You can display the names of the samples in the plot using the settings under **Dot properties** in the **Side Panel** to the right of the view (see the result in figure 4.8).

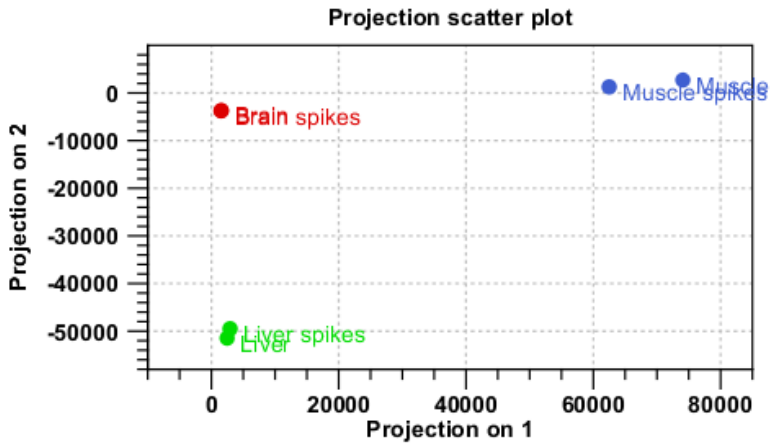


Figure 4.8: *Displaying the sample names.*

This PCA was done on RPKM-based expression values. As you saw in the previous tutorial, the RPKM normalizes for the sample size. In order to show the importance of doing this for this kind of analysis, perform a new PCA on the experiment located in the *total exon reads* folder.

The result is shown in figure 4.9.

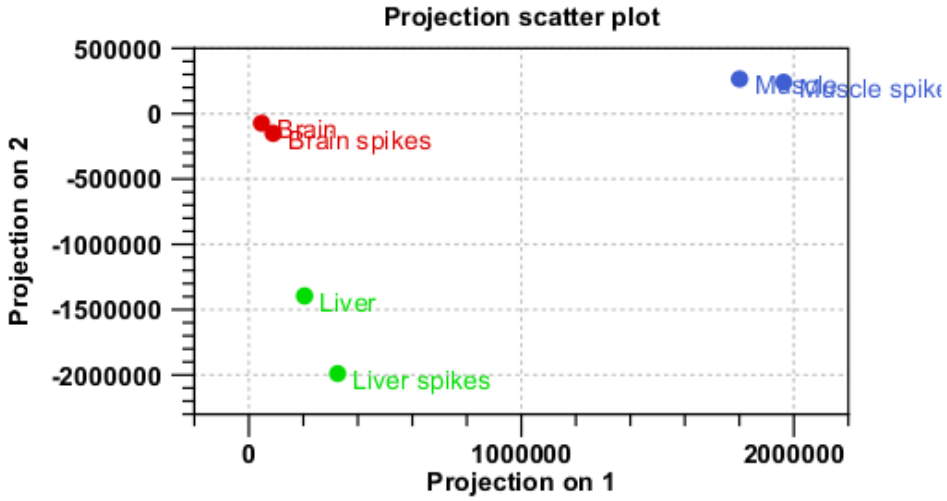


Figure 4.9: *PCA plot using total exon reads.*

Although the replicates still group together, the grouping is not near as clear as in figure 4.8 which uses the RPKM-based expression values.

4.2.3 Hierarchical clustering

In order to complement the principal component analysis, we will also do a hierarchical clustering of the samples to see if the samples cluster in the groups we expect:

> **Toolbox** | **Expression Analysis** ( ) | **Quality Control** | **Hierarchical Clustering of Samples** ( )

Select the experiment from the *RPKM* folder and click **Next**. Leave the parameters at their default and click **Finish**.

This will display a heat map showing the clustering of samples at the bottom (see figure 4.10).

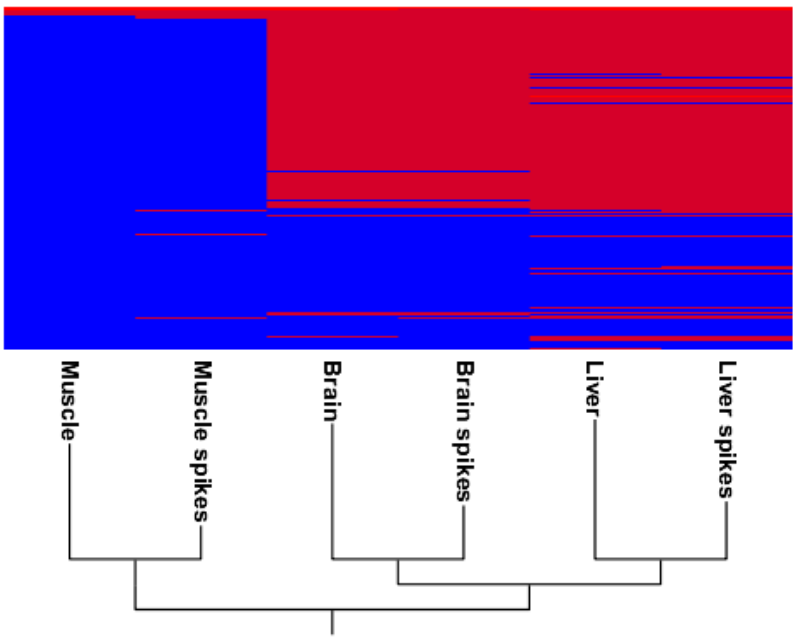


Figure 4.10: *Sample clustering. Adjusting the settings in the Side Panel to put the sample names at the bottom.*

The replicates cluster nicely together as expected, and it looks like the pattern of expression is more similar between brain and liver than between muscle and the other two groups.

Note that the heat map is not a new element to be stored in the **Navigation Area** - it is just another way of looking at the experiment (note the buttons to switch between different views in figure 4.11.

As a conclusion to this fourth tutorial on RNA-Seq, we can confirm that the spike signals are clear and unambiguous, and we can conclude that both the PCA and hierarchical clustering shows that the replicates are fairly homogenous. Just as in the previous tutorial, we can see the effect of using RPKM over total exon reads as expression measure. When doing this kind of quality control, it may be advisory to use the RPKM expression value as this is implicitly standardized



Figure 4.11: *Different views on an experiment.*

between samples (see tutorial part II). Alternatively, you can choose to use the total exon reads counts and then normalize them before performing quality control (you find the **Normalize** ( ) tool here: Toolbox -> Expression Analysis -> Transformation and Normalization).

One of the reasons why the samples cluster so nicely is that there is no biological variation in the samples. The two samples in each group are technical replicates, so what we are really measuring is the quality of the method. In experiments with biological replicates, you would expect to see much more intra-group variability.

CLC Academy