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

February 9, 2010

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

Tutorial: 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.

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 assembled. If you compare these numbers, you can see the first sample has 214903 unassembled reads whereas the second run without multihit reads has 227949 unassembled reads (figure 1).

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

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.

The scatter plot now shows the expression levels of the two samples. Since the RNA-Seq

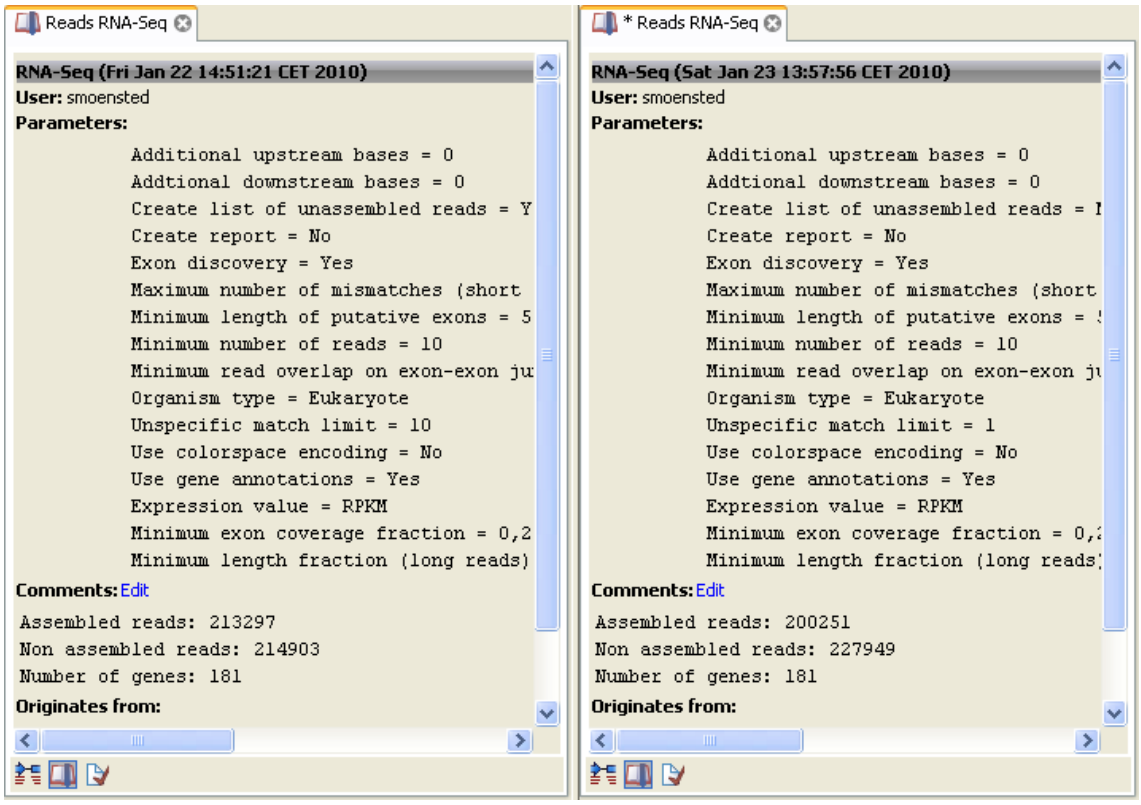


Figure 1: *Comparing the history entries for the two samples.*

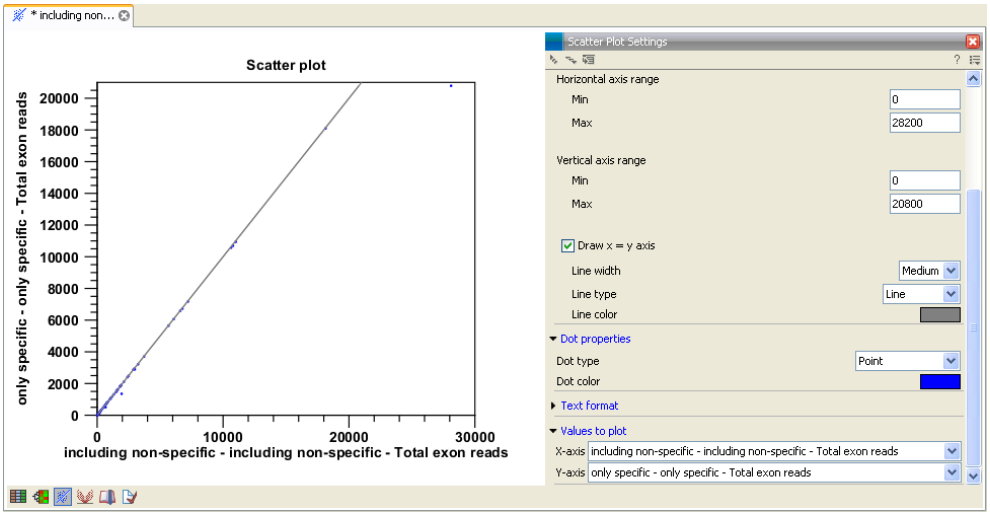


Figure 2: *A scatter plot showing the effect of including non-specific matches in the expression measure.*

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 3, you can see the feature ID (gene name) and the x and y values of the dot.

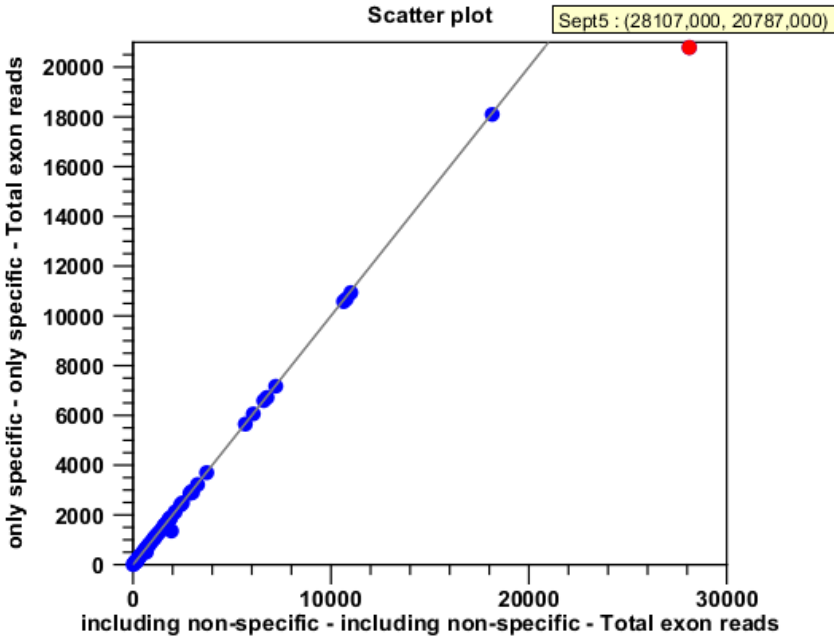


Figure 3: *Gene Sept5 is one of the genes showing a notable difference in expression.*

Open the *including non-specific* RNA-Seq sample and locate this gene by typing *Sept5* in the filter at the top. Double-click to open the contig. When you zoom out ( ) and scroll along to the end of the contig, 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 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.

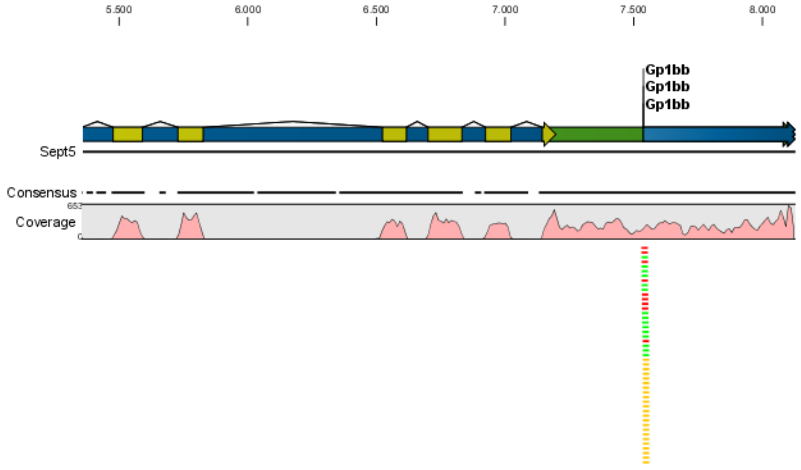


Figure 4: *Gp1bb is overlapping the end of the Sept5 gene.*

If you opened the *Gp1bb* contig 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 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 5

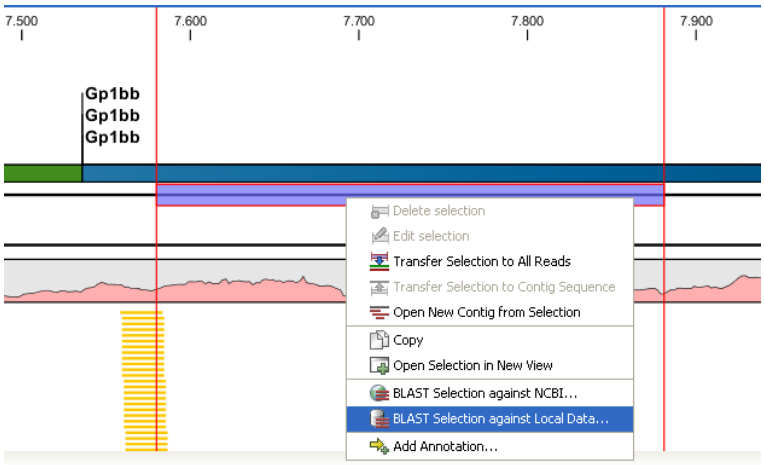


Figure 5: *BLAST against the reference.*

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 Main 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 contig 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 6).

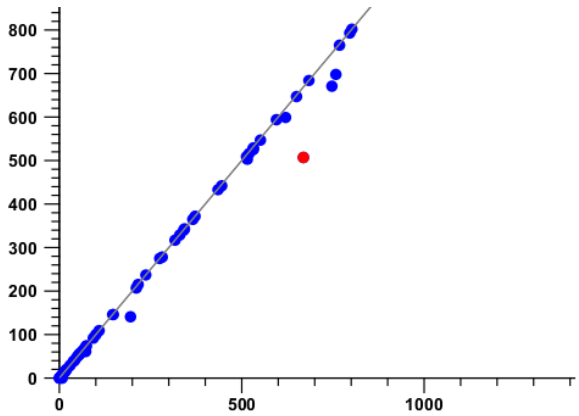


Figure 6: *Zooming in on the Gp1bb gene in the scatter plot.*

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* contig from the *including non-specific* sample, and you can see that there are fewer yellow reads than in the *Sept5* contig. As explained above, the non-specific reads are distributed according to the number of unique reads, and when you compare the two contigs, 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 7).

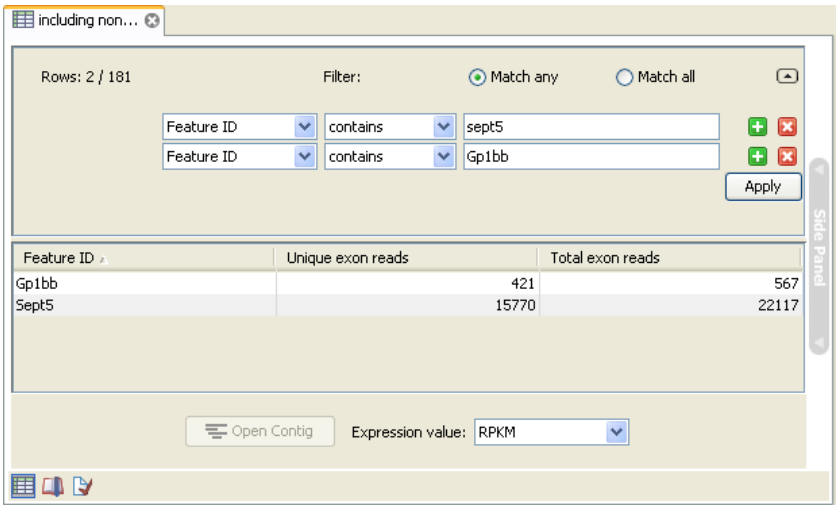


Figure 7: *Comparing the number of unique reads between Gp1bb and Sept5.*

From the scatter plot in figure 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 [Mortazavi et al., 2008] for an in-depth discussion of this topic.

The RPKM expression measure

Normalizing for sample size

The observations made from figure 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 213,297 and 200,251 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).

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. 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 8.

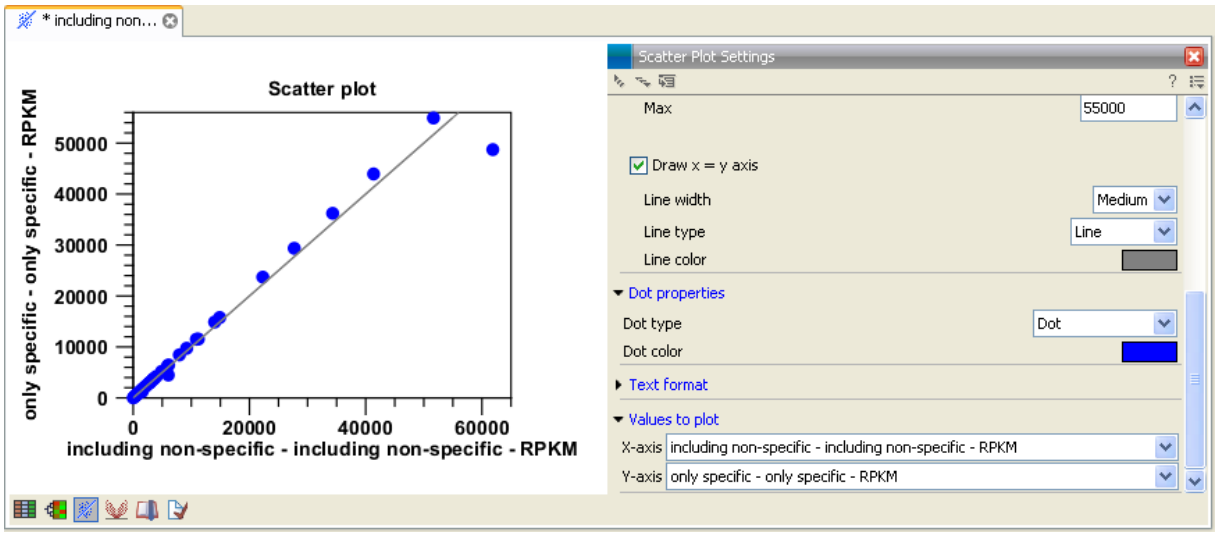


Figure 8: *The effect of including non-specific reads compared using RPKM.*

Where figure 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 $RPKM = \frac{\textit{total exon reads}}{\textit{mapped reads(millions)} \times \textit{exon length (KB)}}$.

Let's investigate two of the genes in the scatter plot. First, identify the two genes closest to the center of the scatter plot — on above the $x = y$ axis, one below. One of them is the *Sept5* gene that we have previously investigated. This still shows higher expression 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 6590, 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).

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 9).

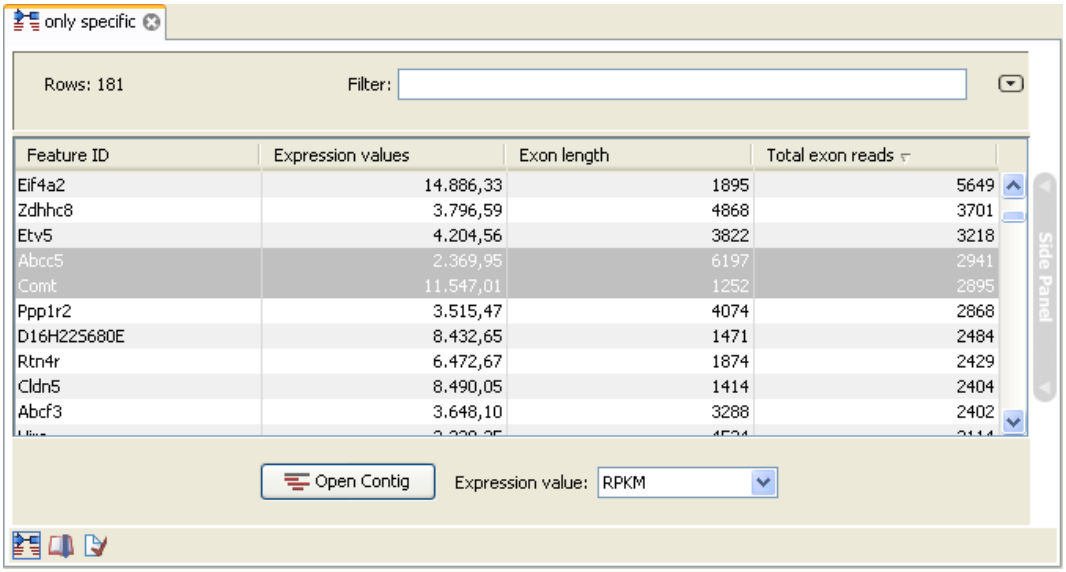


Figure 9: *Nearly the same number of total exon reads for two genes leads to widely different RPKM values because of the difference in transcript lengths.*

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.



Bibliography

[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.