Tutorial: RNA-Seq Analysis Part III: Exon Discovery

March 15, 2013

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

Tutorial: RNA-Seq Analysis Part III: Exon Discovery

This tutorial is the third part of a series of tutorials about RNA-Seq analysis. In this tutorial we will focus on discovery of new putative exons.

We continue working with the data set introduced in the first tutorial and assume here that you have worked through the first two parts of this series, and thus already have the RNA-seq analyses outputs that are used here. Please note that you need to have chosen to run those analyses with the **Exon discovery** option turned on. If it was not turned on, please re-run the analyses of the previous tutorials again, with this option turned on.

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 1.

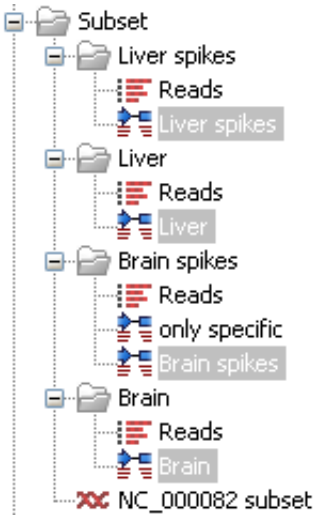
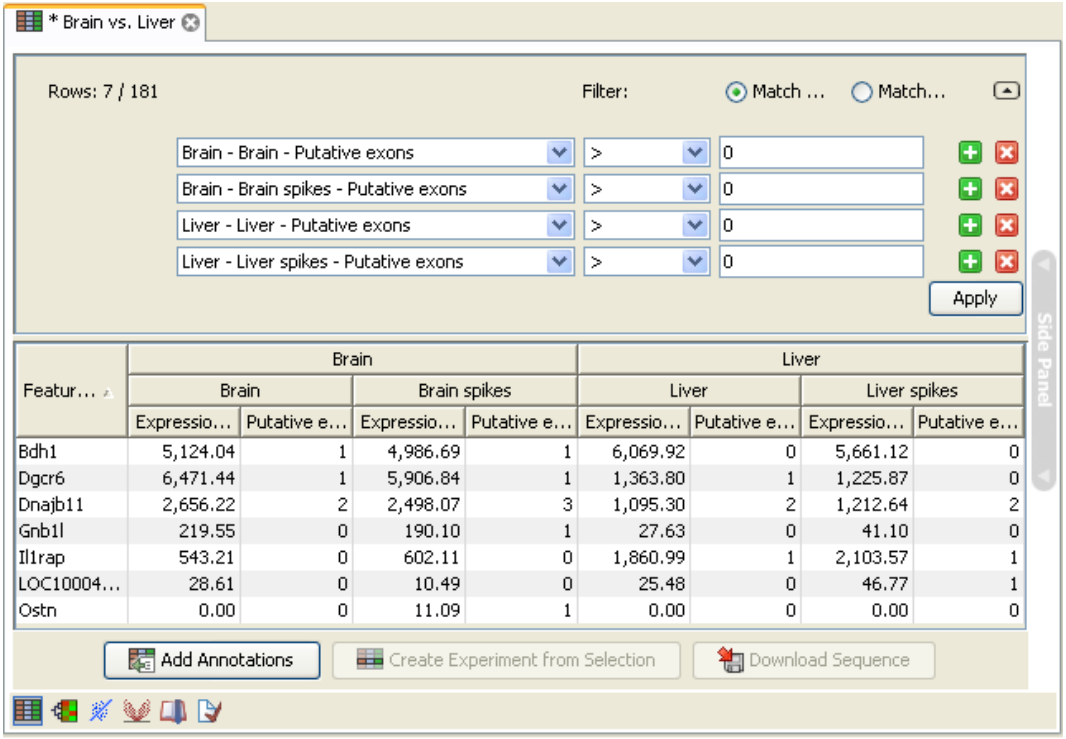


Figure 1: *Four samples have now been created.*

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 2.

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 (213,569 vs 117,429).



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 2: *Comparing putative exons between the groups.*

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).

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.

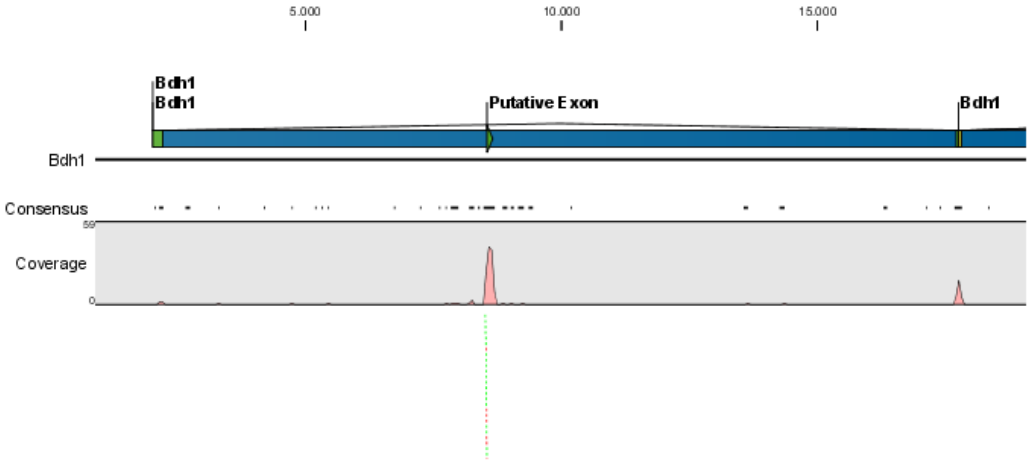


Figure 3: *A putative exon in the brain sample.*

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.

Open the liver spikes sample and look at the *Bdh1* mapping (see figure 4).

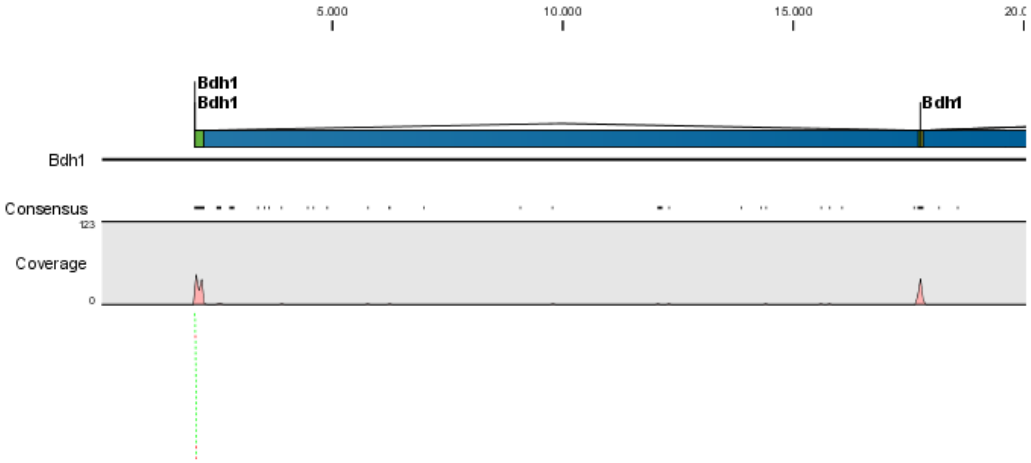


Figure 4: *In the liver sample, there is no putative exon.*

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 **Find** 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.