

C C A T T 0 1 0 0 0
G A G G A 0 1 1 0 1
G A A T T 0 0 1 1 0
A C A A G 0 0 1 0 0
T A C C A 0 0 1 1 0
T T A C A 0 1 0 0 0
A C C T C 0 0 0 1 0
A A G G A 0 0 0 0 0
G A T G A 0 1 1 0 0
T A G A T 0 0 1 0 0
G A T G A 1 0 1 0 0
T G T A G 1 0 0 0 0
T A G T A 0 0 0 0 0
G A T A T 1 0 0 0 0
G A G T G 1 0 0 0 0
A G A T T 1 0 0 0 0
G A G T A 1 0 0 0 0
T G A T G 1 0 0 0 0
A T T A G 1 0 0 0 0
T A G A T 1 0 0 0 0
G A G A 1 0 0 0 0
G T A 1 0 0 0 0
G A T 1 0 0 0 0
T A G 1 0 0 0 0
A G A 1 0 0 0 0
G A G 1 0 0 0 0
A G 1 0 0 0 0
T 1 0 0 0 0

Tutorial

Tutorial: RNA-Seq analysis part I: Getting started

August 9, 2012



Tutorial: RNA-Seq analysis part I: Getting started

This tutorial is the first part of a series of tutorials about RNA-Seq. In this tutorial, we cover the basics steps of running an RNA-seq analysis with an annotated reference genome.

The data used is from a study reported in [Mortazavi et al., 2008]. 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.

Downloading and importing the data

At <http://www.clcbio.com/ngsexampledta> 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).

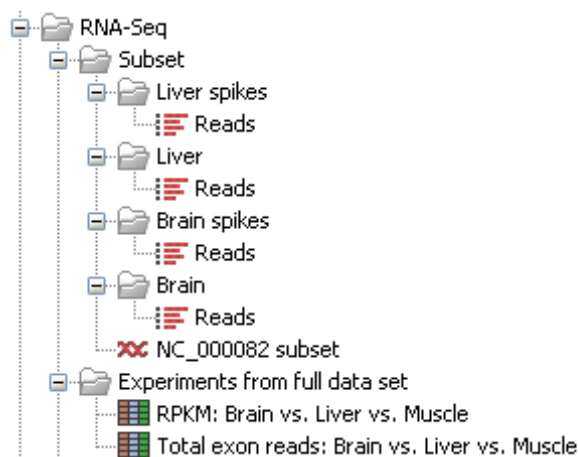


Figure 1: The subset of the full data set has been imported together with the experiments generated from the full data set.

Running the RNA-Seq analysis

The first step in the analysis is to use the list of reads and the annotated reference sequence to generate an **RNA-Seq sample**. This is a data type that basically contains a list of genes with expression values. To do this, go to:

Toolbox | Transcriptomics Analysis (🇺🇸) | RNA-Seq Analysis (🇺🇸)

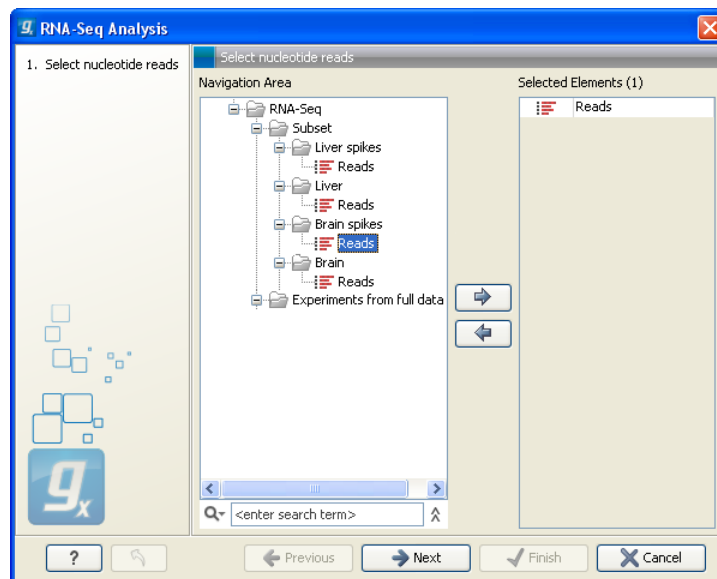


Figure 2: Selecting the *Brain spikes* sample for RNA-Seq analysis.

This opens a dialog where you select the sequencing reads from the *Brain spike* sample, as shown in figure 2.

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 3.

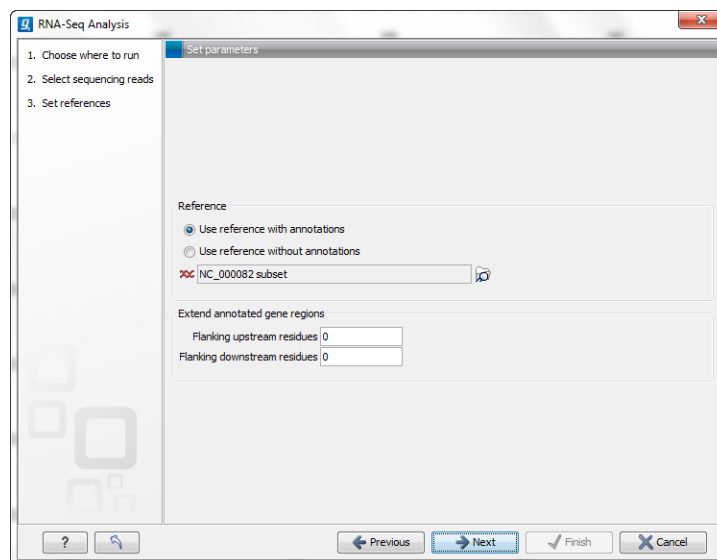


Figure 3: Choosing the annotated reference sequence.

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

Click **Next** where you can set parameters for the mapping. Leave these settings at their default. We focus on these in a later tutorial. If you have changed the parameters from the defaults during earlier work, please set the parameters to default by clicking the button (👉) at the bottom of the dialog. If you do this, you will have to define the reference sequence again.

Clicking **Next** will show the dialog in figure 4.

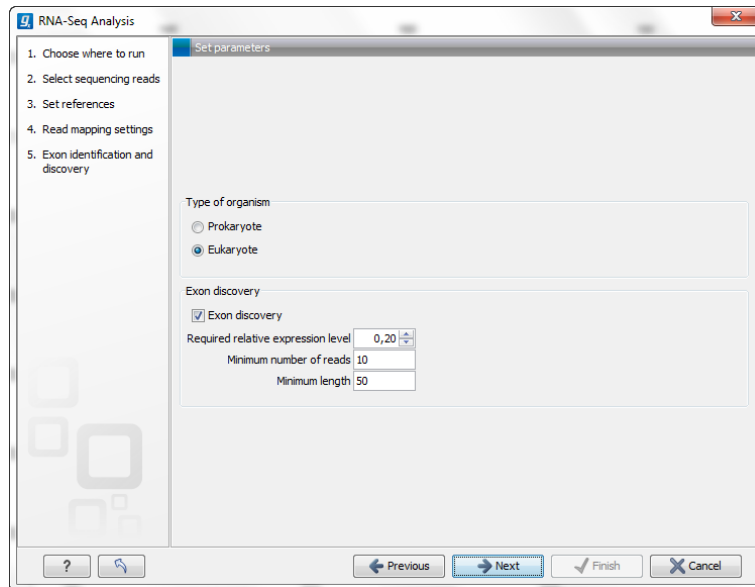


Figure 4: Exon discovery.

The choice between **Prokaryote** and **Eukaryote** is basically a matter of telling the Workbench 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. Please refer to the user manual for more in-depth details about this.

The reference sequence provided with this tutorial includes mRNA annotations. If you view the reference sequence in the viewing pane, these are the green annotations.

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 5.

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

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

Interpreting the brain spikes analysis result

The result of the RNA-Seq analysis is shown in figure 6.

The **Expression values** column is per default based on Genes:RPKM. For this tutorial, please change the measure to use **Genes:Total exon reads** instead by clicking at the bottom of the view. Expression measures are discussed in more detail in the second of this series of tutorials.

Now sort the table on the new expression value by clicking the column header twice. Find the *Ahsg* gene, which should be fairly near the top of the list. We wish to open the mapping of the sample reads against this gene. To do this, either click on the *Ahsg* gene and then click on the **Open mapping** button, or just double click the *Ahsg* gene line.

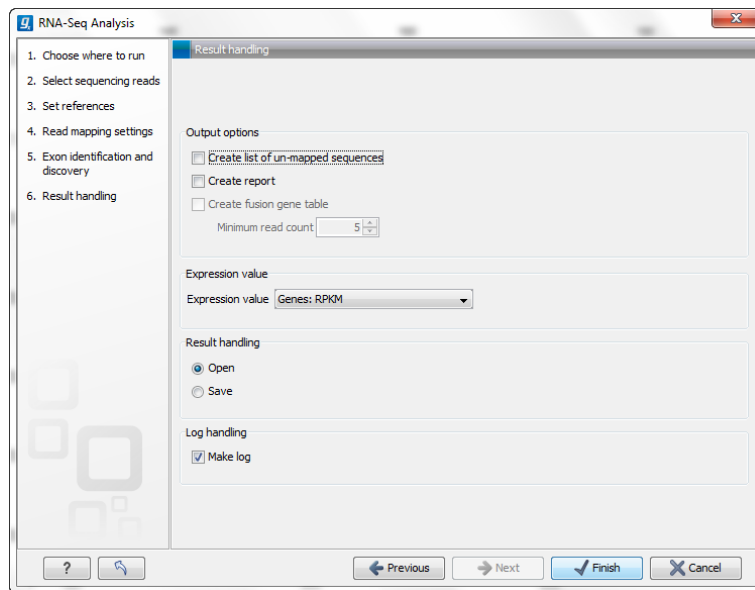


Figure 5: Selecting the output of the RNA-Seq analysis.

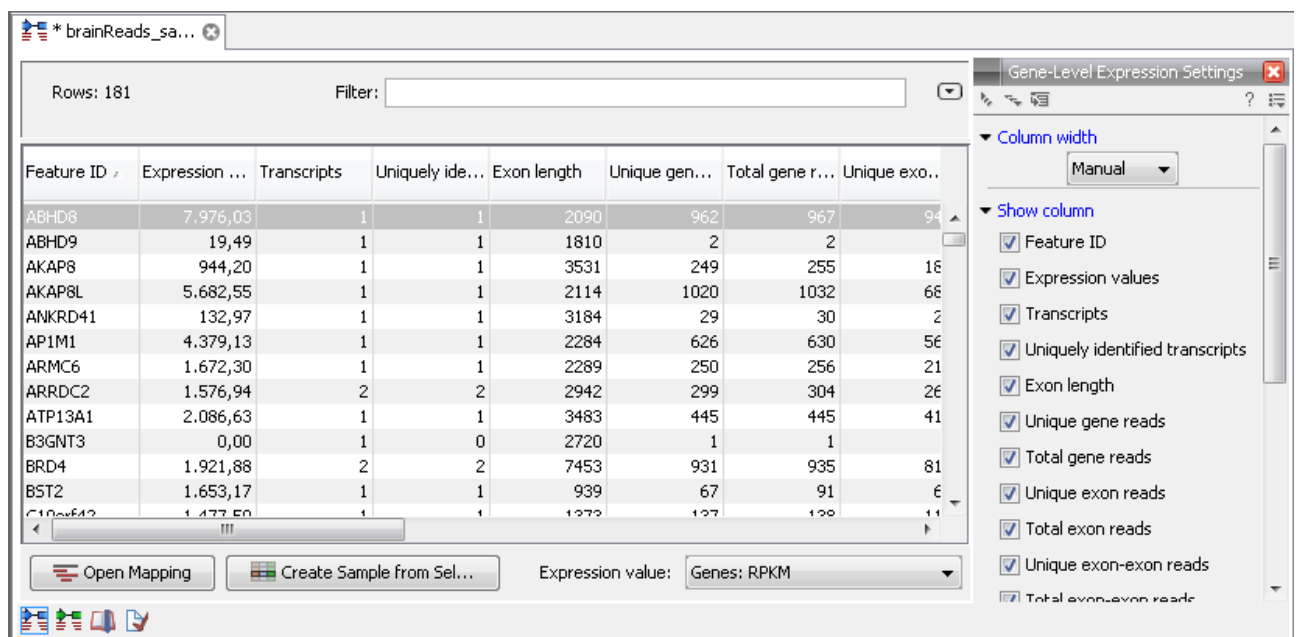



Figure 6: A table with expression values for all genes. Here, the name of the results generated was changed to BrainSpikes-RNAseq. Part of the name can be seen in the tab, so using meaningful names can help if you will be examining many output files. To change the name of an object, just click on it in the Navigation area, and then click again. You should then be able to change the name.

You should now have a split view, with one half of your viewing area showing your results table, and the other half showing you the mapping.

A few customizations of the mapping view will help make the view better suited for interpretation. In the **Side Panel** beside the mapping, under **Text format**, set the text 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 7.

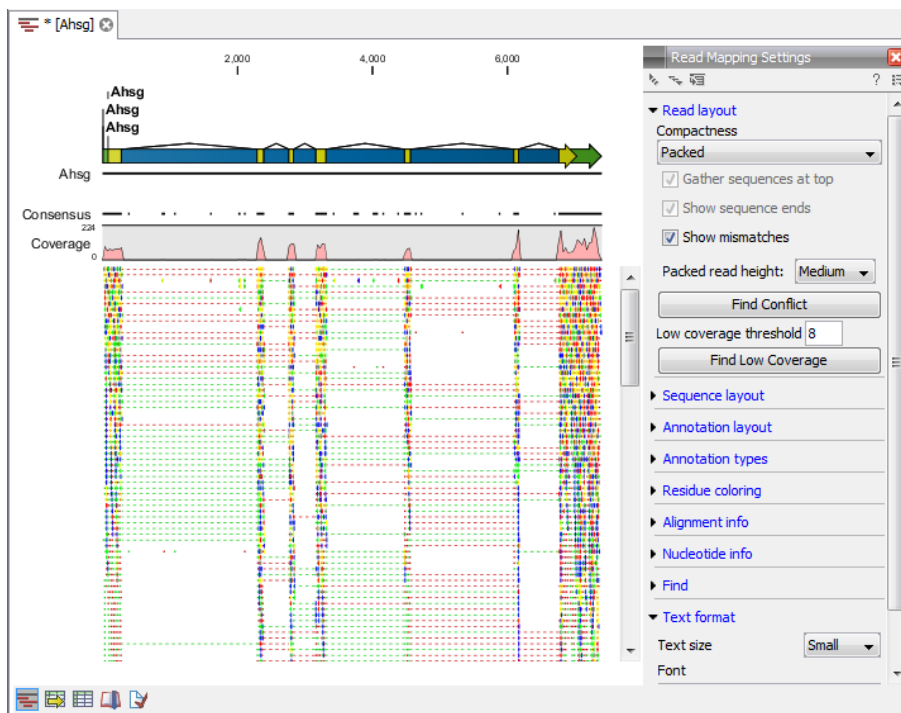





Figure 7: The reads mapped to the *Ahsg* gene.

You can now see distinct peaks of coverage below the exons which are marked in green. You should see that some reads map within exon regions, while others, that have thin dotted lines connecting each end of the read, have been mapped across exon-exon boundaries.

Click the **Zoom in** () button in the top right toolbar and click-and-drag a rectangle around one of the exons. In this way you can zoom in to see more details of that particular exon. If you zoom all the way in, you will be able to see the nucleotide level and the alignment of the reads.

Close the view by clicking in the small red X in near the tab name for the view, and go back to the RNA-seq sample. In the 'Transcripts annotated' column (which may appear to only have Transcripts... as the column heading in the default view) you can see that the *Ahsg* gene only has one transcript annotated.

Click on the **Advanced filter** () button at the upper right hand part of the RNA-seq sample table view), and use filtering to identify genes with more than one transcript annotated. To do this, set the filter to `Transcripts > 1` and press **Apply** as shown in figure 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.

One of the three transcripts that were annotated on this reference sequence for *Fetub* uses a different first exon from the other two transcripts. We can see in our mapping that there is no coverage in this exon, or, in other words, there is no evidence for expression of the alternative first exon isoform in this sample. There is evidence of expression of the other two transcripts

Rows: 13 / 181 Filter: Match any Match all

Transcripts > 1

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

though. These have the same first exon (exon 2), but one skips the third exon, while the other includes it. You can see this in the mapping, as you can see reads that span from exon 2 to exon 3, and also reads that span from exon 2 to exon 4. See figure 9).

Close this view. You can now continue with the second tutorial in this series: Non-specific matches and expression values.

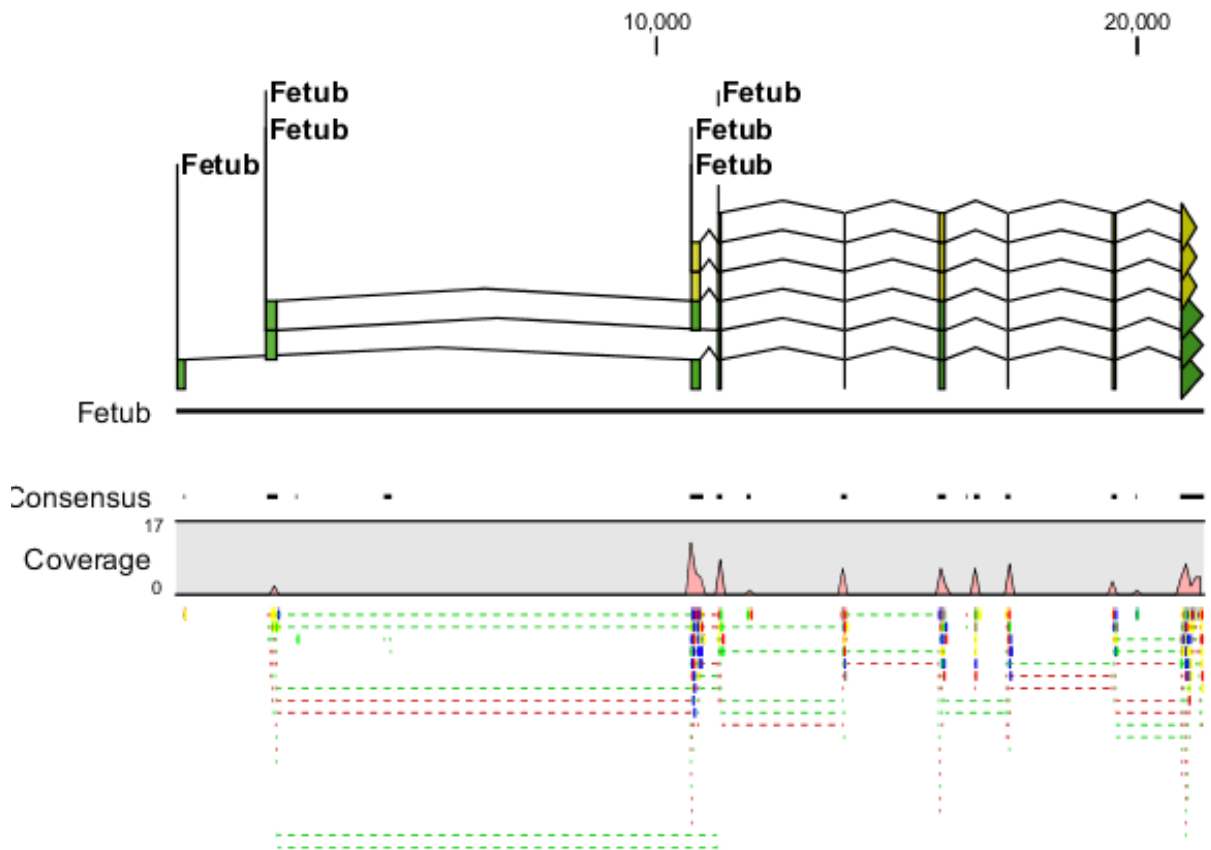


Figure 9: Reads showing evidence for expression of two isoforms. Exons shown here are the second, third, fourth and fifth of the *fetub* gene. The bottom reads shown here indicate reads that map across the borders of exons 2 and 4. Many other reads have been mapped across the border of exons 2 and 3.



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.