

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

Tutorial

Tutorial: RNA-Seq analysis part I: Getting started

February 9, 2010



Tutorial: 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 [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/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).

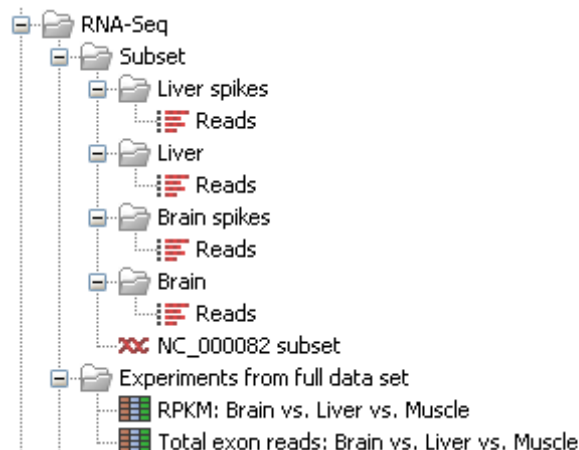


Figure 1: The subset of the full data set has been imported together with the experiments generated from the full data set.

Running the RNA-Seq analysis

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 ()

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

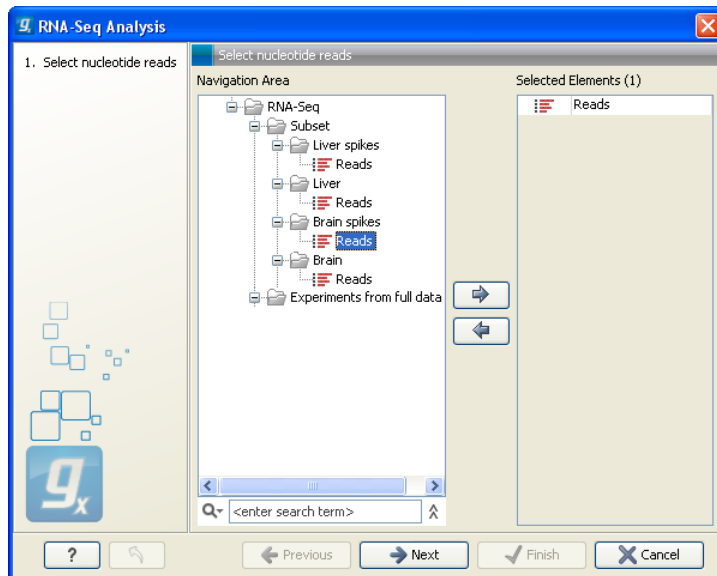


Figure 2: Selecting the *Brain spikes* sample for RNA-Seq analysis.

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 3.

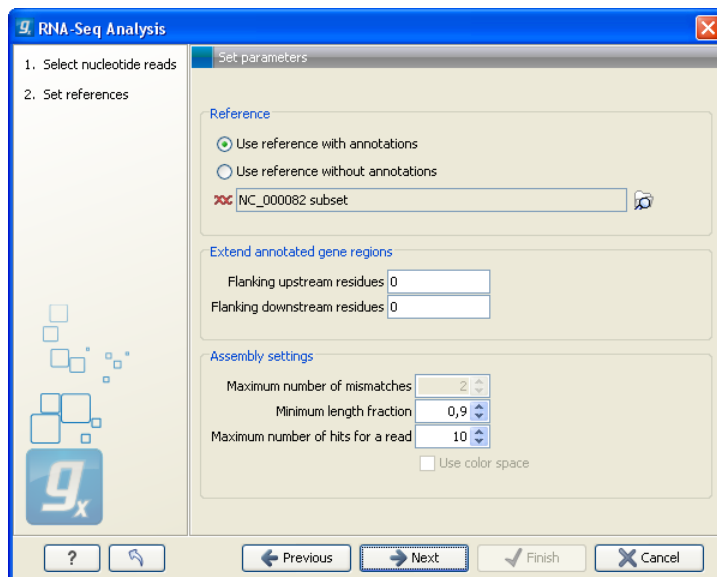


Figure 3: Choosing the annotated reference sequence.

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 assembly. 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 4.

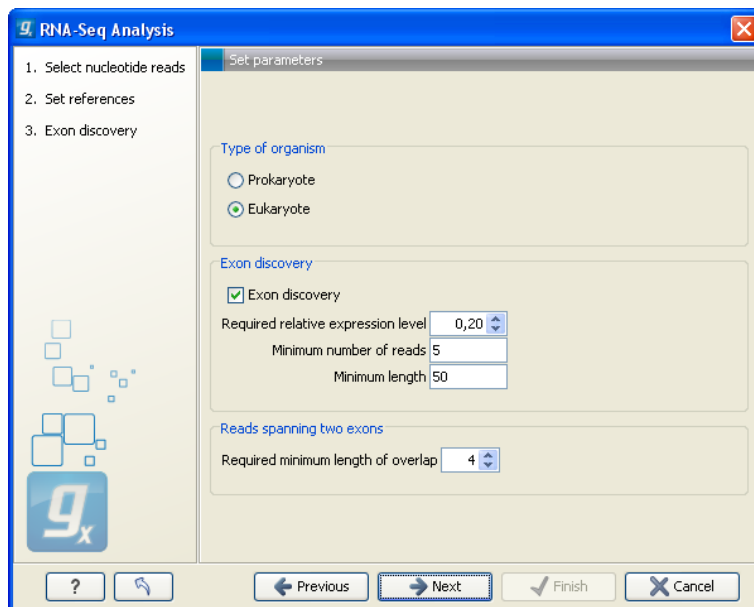


Figure 4: Exon discovery.

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). 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 5.

Uncheck the **Create list of unassembled sequences** and **Make log** and click **Finish**.

The standard output is a table showing mapping statistics on 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 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 **Contig Settings 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 contig, click the **Save/Restore**

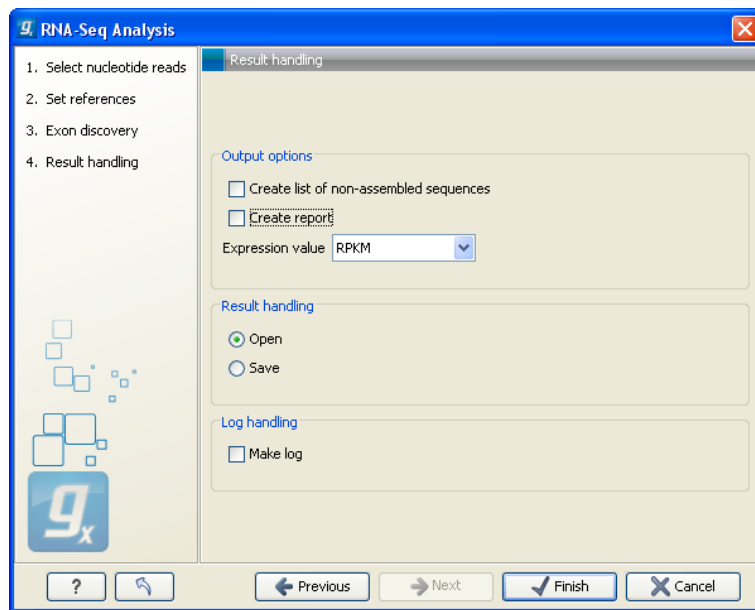


Figure 5: Selecting the output of the RNA-Seq analysis.

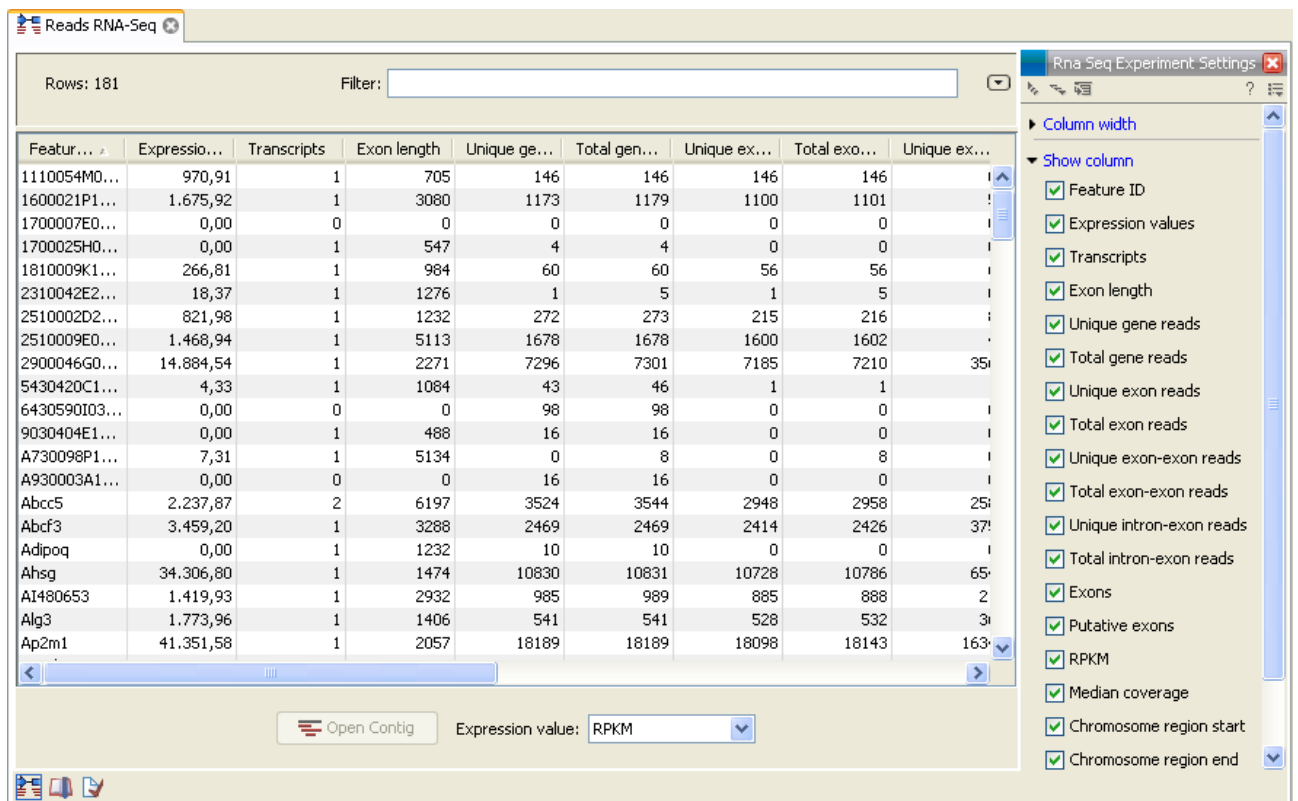


Figure 6: The reads assembled to the Ashg gene.

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 contig. You should now have a view similar to figure 7.

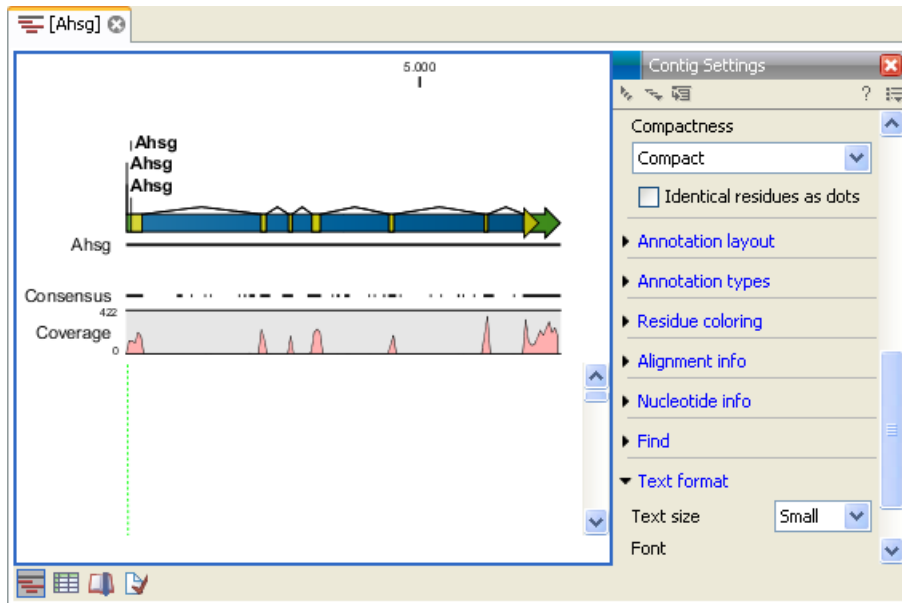


Figure 7: The reads assembled to the Ashg gene.


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 contig. You will begin to see reads that have been assembled 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 8).

Featur...	Expressio...	Transcripts	Exon length	Unique ge...	Total gen...	Unique ex...	Tote
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	
Gnb1l	1.504,21	2	3085	507	507	507	

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

The *Fetub* gene has three transcripts annotated. Open the contig for this gene and press **Fit width** () to zoom out completely and get an overview of the assembly to this gene.

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 9).

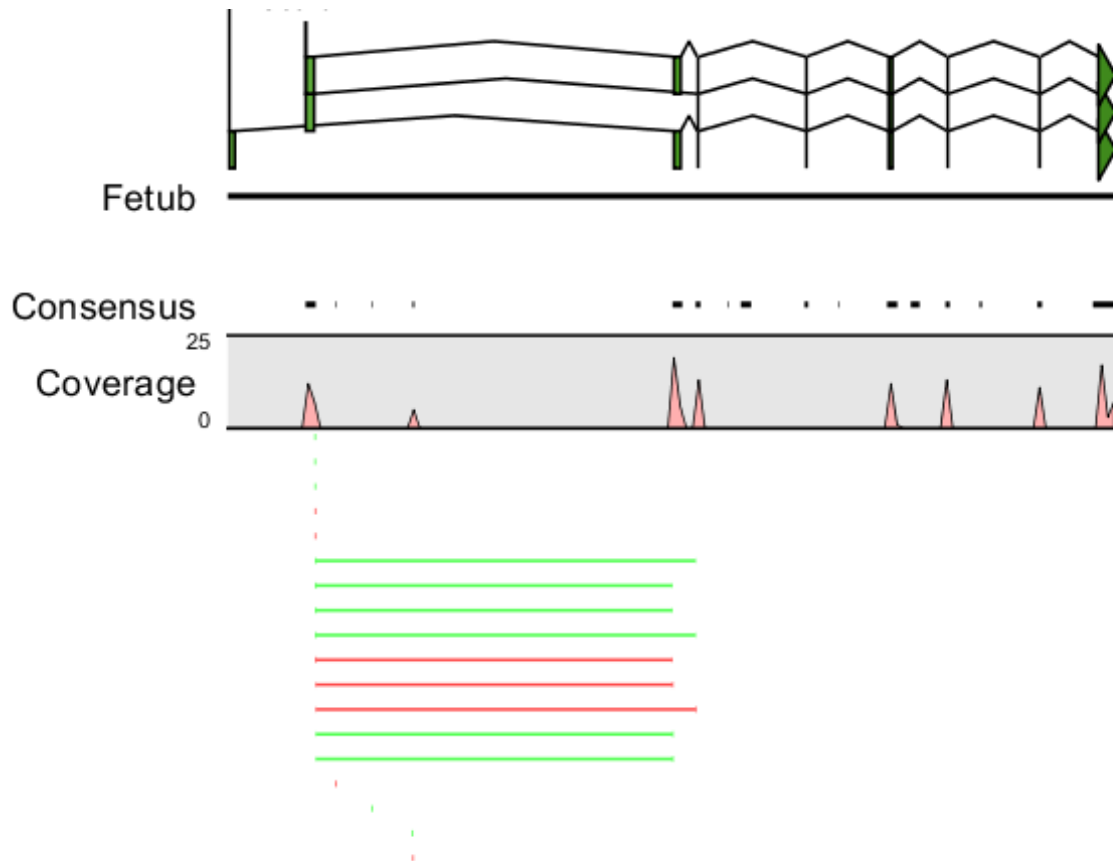


Figure 9: Reads showing evidence for expression of two isoforms.

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



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.