Tutorial: Small RNA analysis using Illumina data

January 10, 2011

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

Tutorial: Small RNA analysis using Illumina data

This tutorial shows how to go through the initial parts of analyzing a small RNA data set. It shows how to analyze one sample including how to trim off adapter sequences and extract the small RNAs, how to count the small RNAs, inspect and check the results and finally how to annotate the small RNAs to identify known miRNAs and other non-coding RNAs.

The tutorial is based on the study published in [Stark et al., 2010]. In this study, 12 samples from melanoma cell cultures were sequenced using an Illumina Genome Analyzer II. In this tutorial, we will analyze one of these samples, the MELB sample, which represents a primary melanocyte cell.

Downloading and importing the raw data

First, download the fastq file containing all the raw data from this sample: `http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?cmd=dload&run_list=SRR038853&format=fastq`.

Now, import the data set using the Illumina importer:

File | **Import High-throughput Sequencing** | **Illumina**

This will bring up the dialog shown in figure 1.

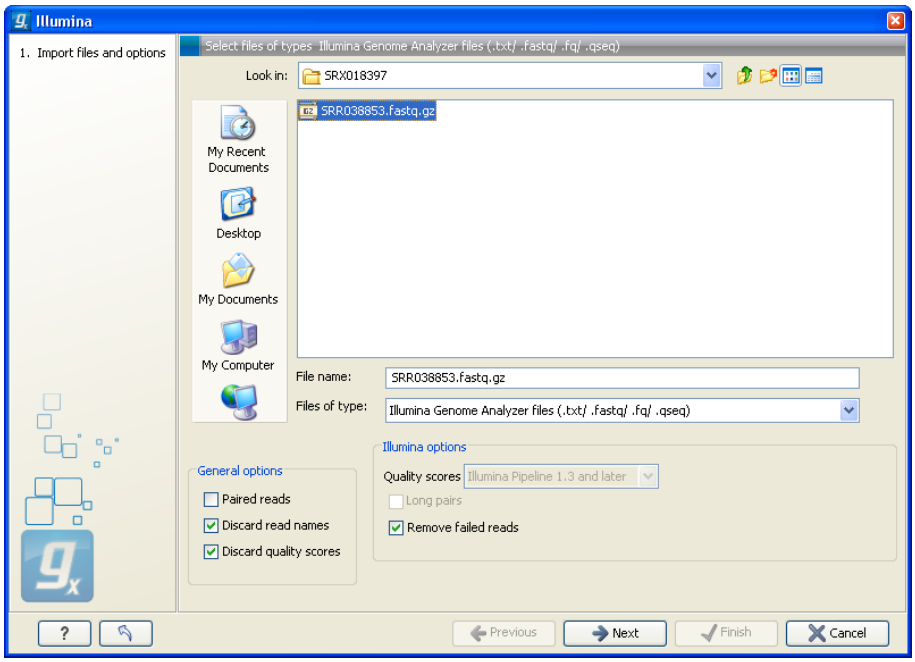


Figure 1: *Selecting the fastq file for import.*

Select the `SRR038853.fastq.gz`. Make sure the **Discard quality scores** and **Discard read names** checkboxes are checked. Information about quality scores and read names are not used in this analysis anyway, so it will just take up disk space when importing the data. Click **Next**, choose to **Save** and click **Finish**.

After a short while, the reads have been imported. Open the file you imported by double-clicking and place your mouse on the tab. After one second, you will see a small tool tip with information

about the number of reads in the file as shown in figure 2.

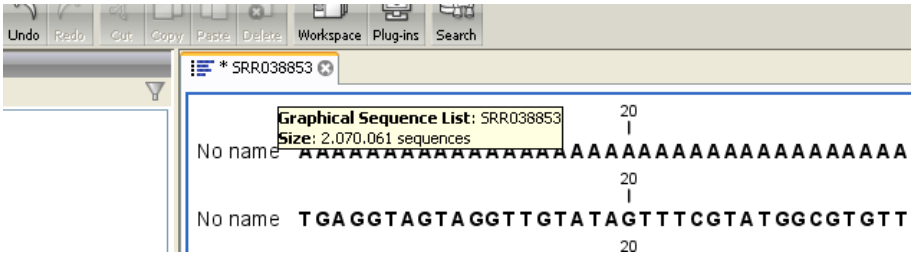


Figure 2: *This data set contains about two million reads.*

Close the view again.

Trimming adapters and counting the reads

The next step in the analysis is to trim off the partial adapter sequences and subsequently to count how many copies there are of each of the resulting small RNAs.

Toolbox | High-throughput Sequencing ( ) | Small RNA Analysis ( ) |Extract and Count ( )

This opens a dialog where you select the `SRR038853` sample as shown in figure 3.

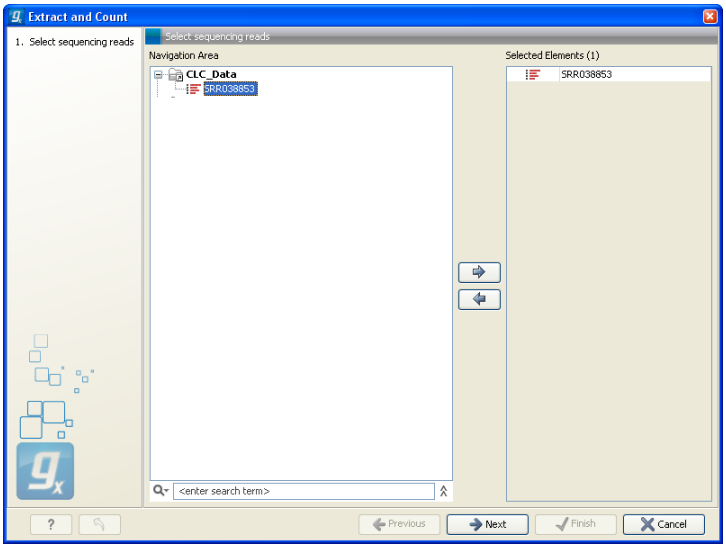


Figure 3: *Selecting the sample for extracting and counting the small RNAs.*

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 4.

Make sure the checkbox is selected and click **Next**.

You will now see the dialog shown in figure 5.

In the list of adapter sequences, select the Illumina adapter. You can see in the preview panel below how many matches that are found for this adapter among the first 1000 reads in the input file. We will see more statistics on this for the full data set later on - this preview is just intended to support the user when defining the adapter trim setting.

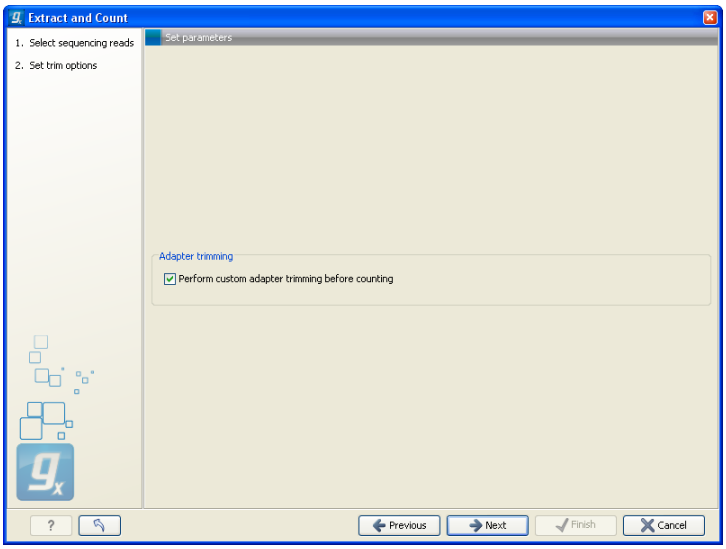


Figure 4: *Choosing to trim for adapter sequence.*

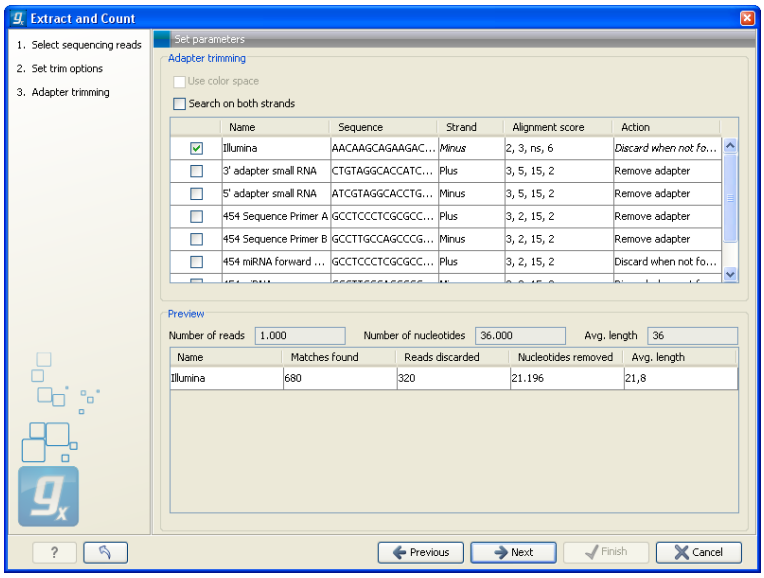


Figure 5: *Trimming the raw data for adapter sequence.*

Since the trim settings are already set right, click **Next**. Note that you could have changed both Strand, Alignment score and action in this panel by clicking/double-clicking the adapter.

You will now see the dialog shown in figure 6.

The most important choices in this dialog are that you can set a minimum and maximum length on the tags that you want to include when counting, and that you can decide how many copies there have to be in order to include the tag in the output. Leave these settings at the default and click **Next**.

This will allow you to specify the output options as shown in figure 7.

The default is to output a **Sample** which is the table of all the small RNAs and their counts, and to create a report showing summary statistics. Leave the settings at default and click **Finish**

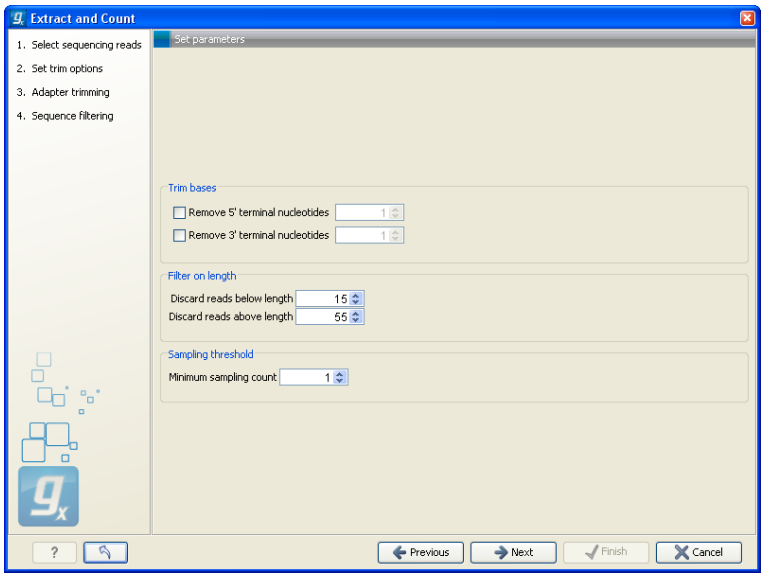


Figure 6: *Adjusting options for counting the small RNAs.*

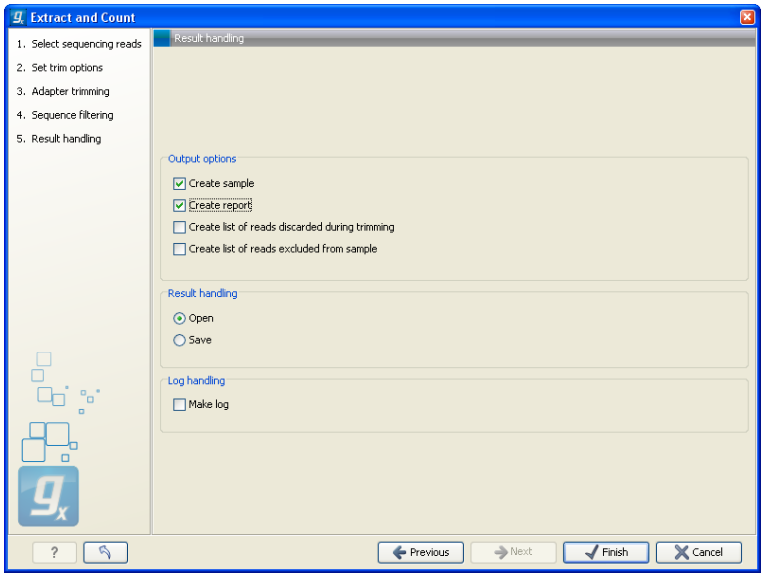


Figure 7: *Selecting the results to output.*

Interpreting the adapter trim report

Once the analysis is complete, two tabs will be opened. First, we take a look at the report.

The top part of the report is shown in figure 8.

The report is meant to be used as a quality check, mainly to see that the adapter trimming worked as expected. In this example, it shows that out of 2 million reads, 1.7 million reads passed the adapter trim. The trim settings meant that if no adapter sequence was found, the read would have been discarded. So this means that (part of) the adapter sequence was found in all these 1.7 million reads.

There is also a graph showing a distribution of the read lengths after trimming. In this example, there is a very nice distribution with a peak around 22 bp which is expected for miRNAs.

1 Trim summary

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed	Avg.length after trim
SRR038853	2,070,061	36.0	1,720,241	83.1%	21.9

2 Read length before / after trimming

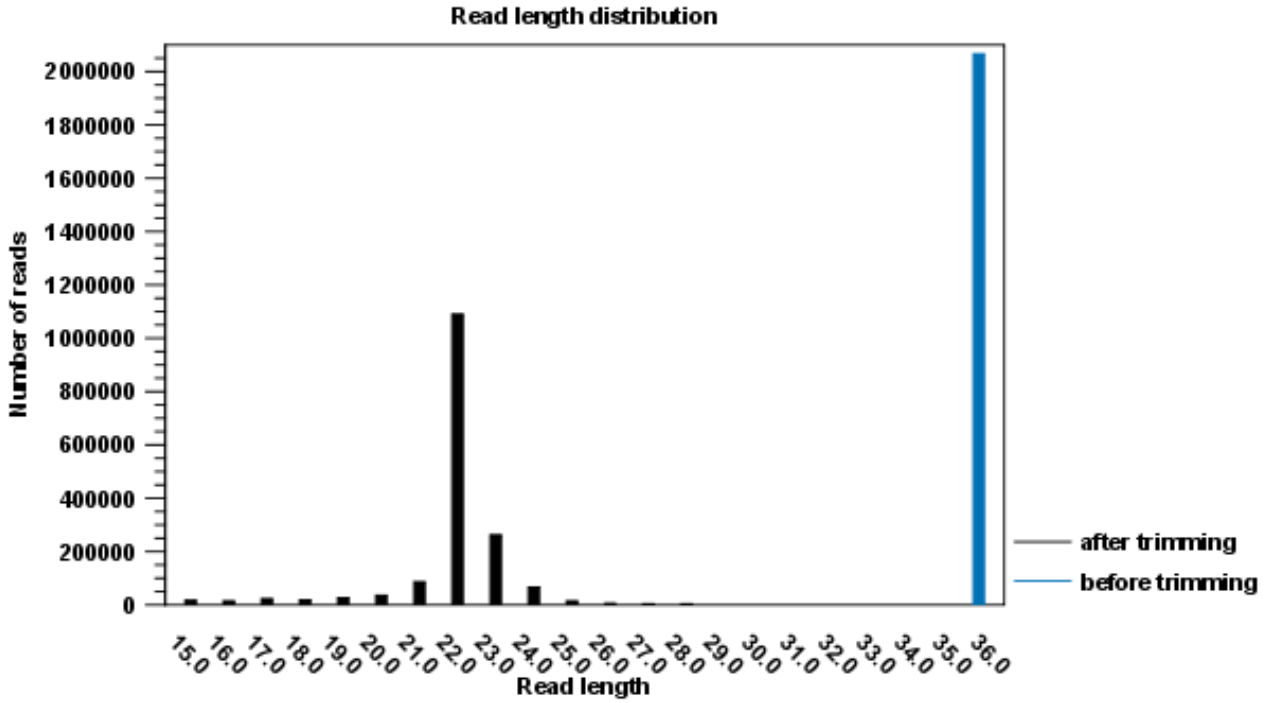


Figure 8: *The small RNA counting report.*

Investigating the small RNA sample

Save and close the report, and you should now see the small RNA sample as shown in figure 9.

There are 88,460 unique small RNAs in the sample. You can filter and sort the sample, and you can extract subsets using the buttons at the bottom of the view. As an example, we will try to open the trimmed reads of one of the small RNAs: Sort the table on Length (clicking the column header) and click the row at the top. Then click the **Extract Reads** button and click **Finish** in the dialog that is opened. You should now be able to see the original read sequence with a trim annotation as shown in figure 10.

Clicking the **Double stranded** checkbox in the Side Panel to the right, you can see the minus strand as well, and you can see that the adapter sequence, `CAAGCAGAAGACGGCATACGA` has a perfect match here.

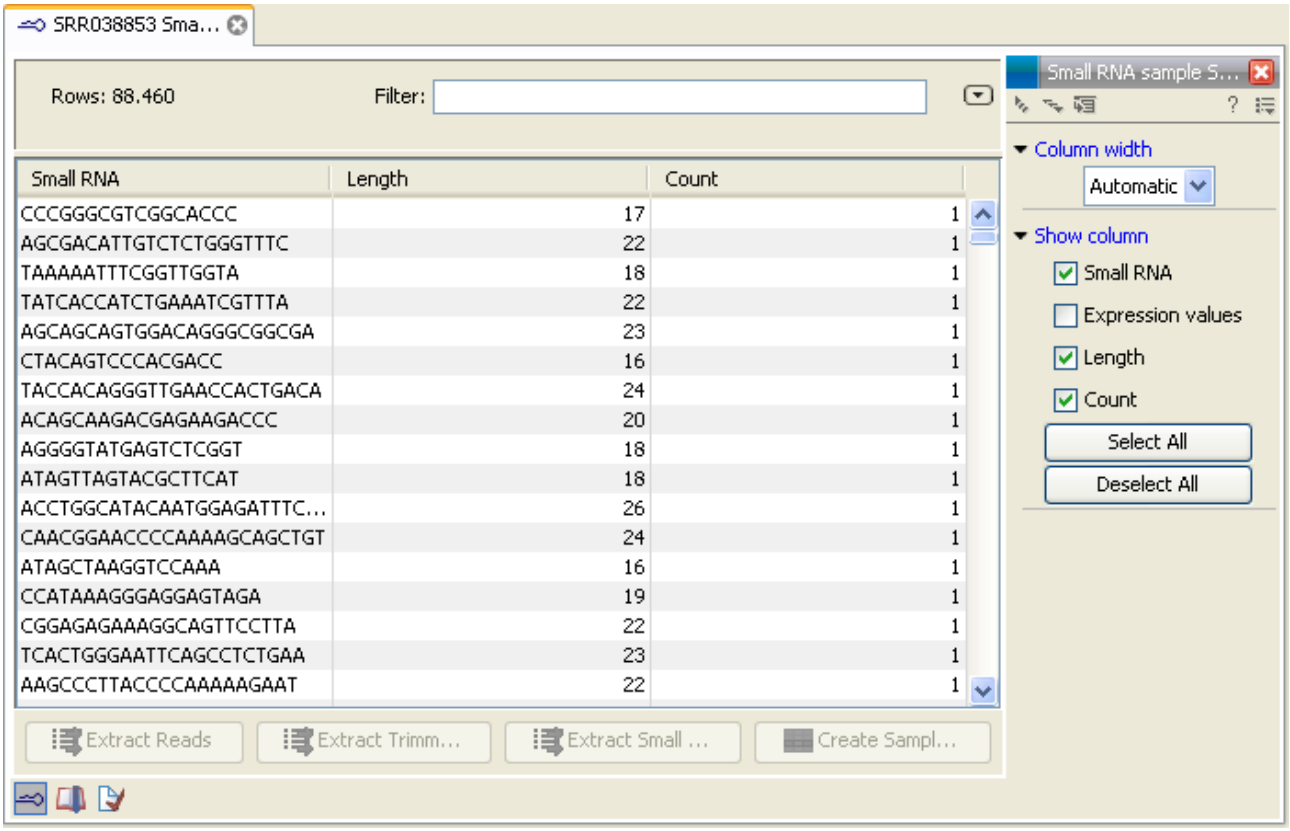


Figure 9: *The small RNA sample.*

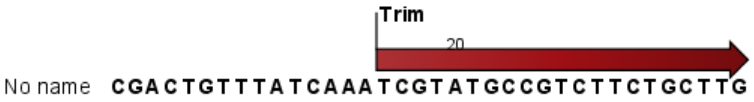


Figure 10: *A sequencing read displaying the trim annotation.*

Downloading miRBase and annotating the sample

The next step in the analysis is to annotate the small RNA sample to identify known small RNAs. We use two sources for the annotation: first, miRBase is used to identify known miRNAs and second a set of other known non-coding RNAs.

The Workbench lets you download the latest version of miRBase directly:

> **Toolbox** | **High-throughput Sequencing** ( ) | **Small RNA Analysis** ( ) |**Download miRBase** ( )

Choose **Save**, **Next** and **Finish**. Next, download and **Import** the set of other non-coding RNAs from `ftp://ftp.ensembl.org/pub/release-57/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh37.57.ncrna.fa.gz`. You should now have the two annotation files represented as sequence lists in the Workbench, and you are ready to start the annotation:

> **Toolbox** | **High-throughput Sequencing** ( ) | **Small RNA Analysis** ( ) |**Annotate and Merge Counts** ( )

This opens a dialog where you select the `SRR038853 Small RNA sample` as shown in figure 11.

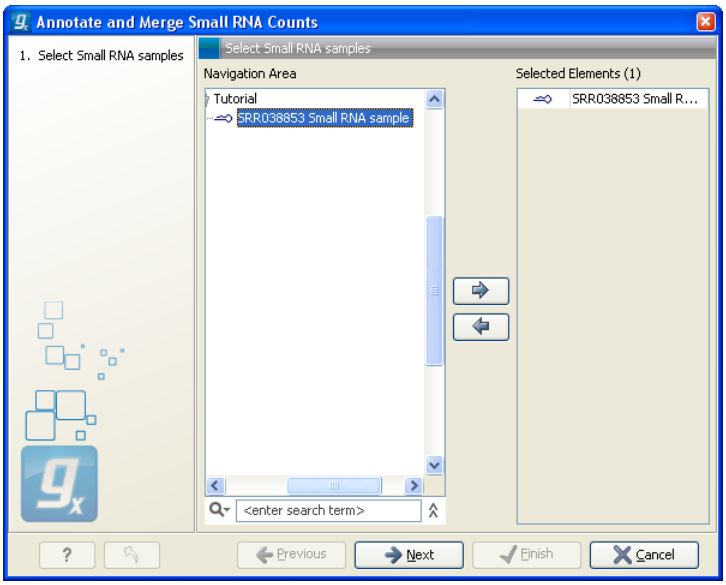


Figure 11: *Selecting the sample for annotating the small RNAs.*

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 12.

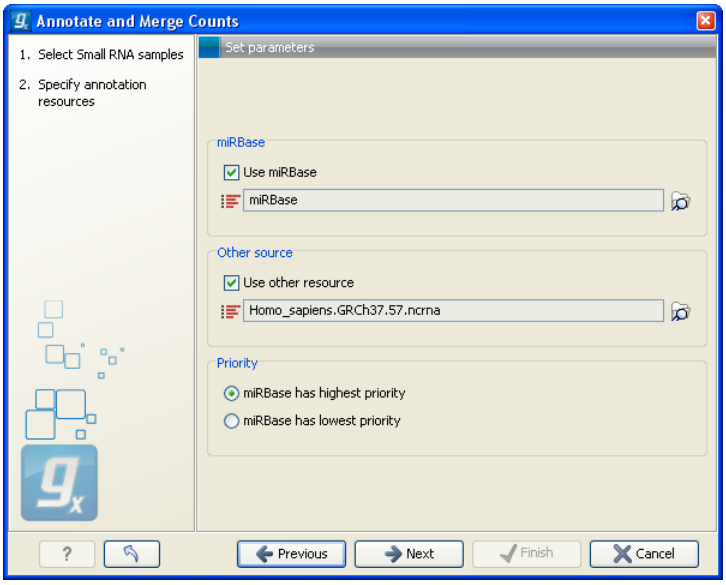


Figure 12: *Setting miRBase and the other non-coding RNAs as annotation source.*

At the top, select to use miRBase and select () the miRBase file that you downloaded previously. Below, check the **Use other resource** and select () the `Homo_sapiens.GRCh37.57.ncrna` file that you imported previously.

The miRBase file contains a list of precursor sequences with specification of the mature and in some cases the mature * regions. This information is used to categorize the annotated small RNAs. The **Other resources** does not include this kind of information and is used here in order to identify known small RNAs that are not miRNAs. Note that you could include several sequence lists here if you have other sources of non-coding small RNAs.

Click **Next** will show the dialog shown in figure 13.

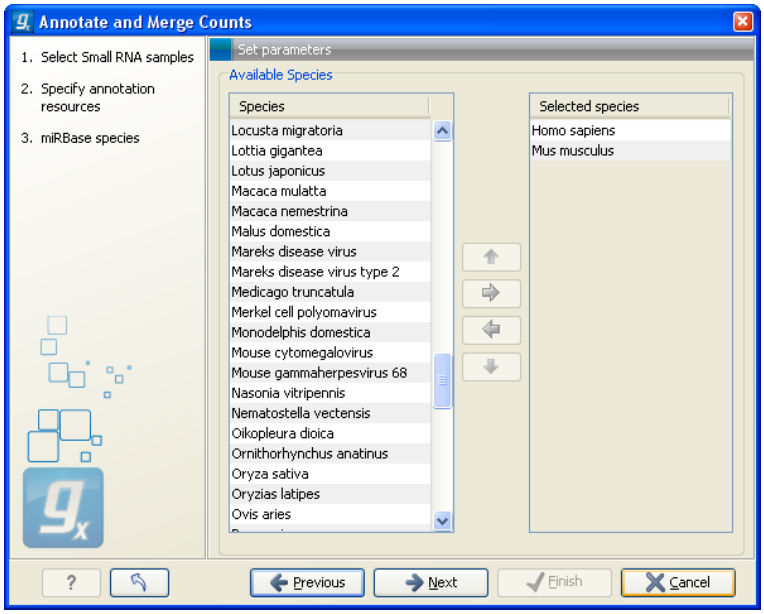


Figure 13: *Prioritizing species for annotation.*

Select first *Homo sapiens* and second *Mus musculus*. The sample is human, so that should be the first priority as annotation source, and mouse should be second in the list. Since there may be miRNAs that have not yet been identified in human but have an ortholog in mouse, it is interesting to include the mouse miRNAs as well.

Click **Next** will show the dialog in figure 14.

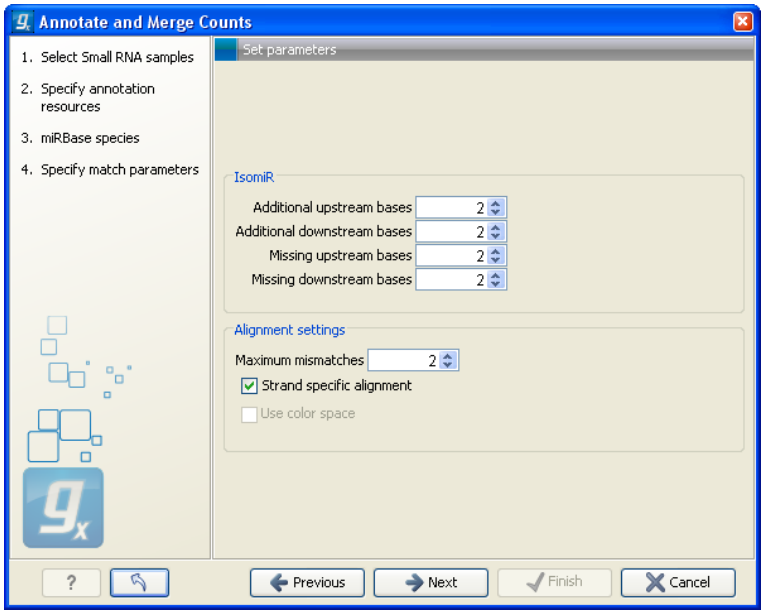


Figure 14: *Thresholds for annotating.*

Leave these settings at their default and click **Next** to display the dialog shown in figure 15.

Make sure all options except the unannotated sample are checked and click **Finish**.

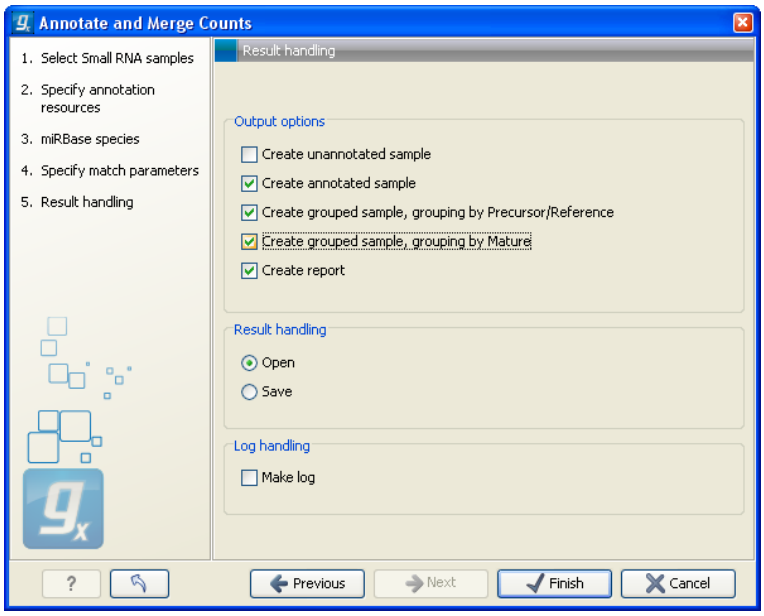


Figure 15: *Select the sample grouped on mature.*

Analyzing the annotated samples

For a detailed description of the output, please have a look in the user manual (press **F1** to display). In this example we focus on a few specific miRNAs that will illustrate how the annotation and grouping of samples work and show some of the possibilities you have for interacting with the data.

Looking at the grouped sample

Save ( ) and close all the views except for the un-grouped annotated sample ( ). We want to look at *mir-29a*, so type this into the filter at the top of the table as shown in figure 16.

This will list all the tags (414 out of 31,841) that have been mapped to the mir-29a precursor sequence from miRBase. If you sort the table by **Count** (clicking the count column header), you can see that most of these are exact matches of the mature miRNA. The rest are variants and length variants.

For expression analysis, it can make sense to look at all the variants of the same miRNA as one entity rather than 414 as it is the case here. Open the `SRR038853 Small RNA sample grouped` ( ) and type in mir-29a in the filter. You now have one line representing all the tags that have been annotated with mir-29a with a total count of 36,600. The number of reads in different categories are shown, e.g. 30,689 for the exact mature corresponding to the number from the ungrouped sample in figure 16. You also see a few tags annotated with the mouse ortholog, but this could be noise due to sequencing errors.

Double-click the human mir-29a row to open the alignment of all the tags to the precursor sequence (see figure 17).

The tags are colored to reflect the counts which are also shown in numbers next to the name to the left. Since the exact mature is very dominant in terms of count, it is the only one standing out in a different color.

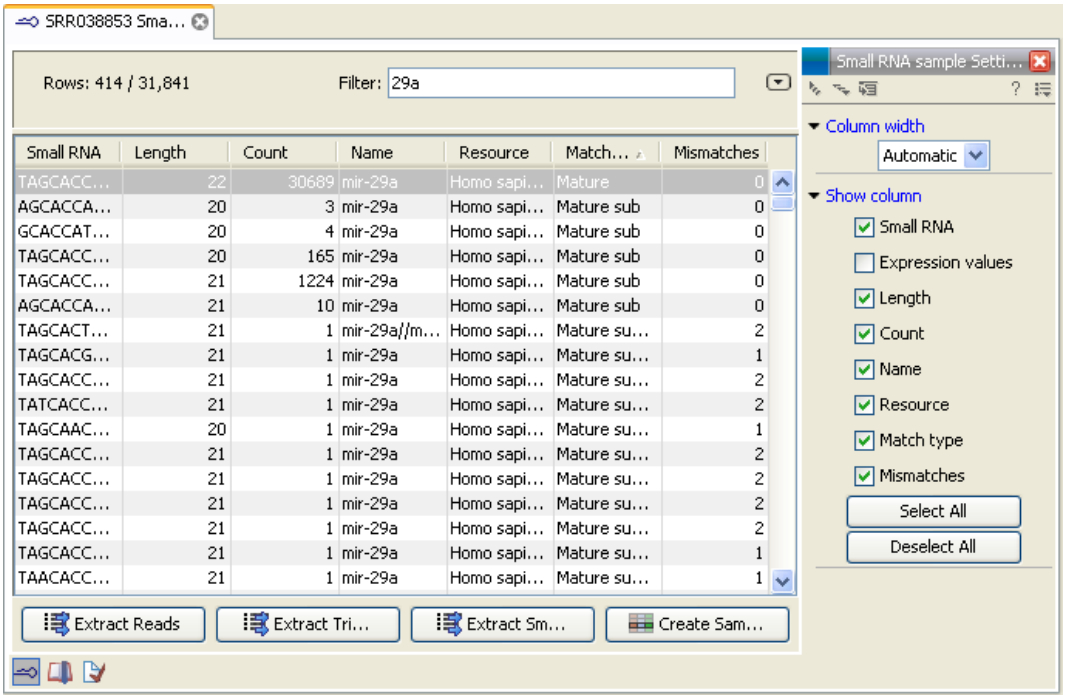


Figure 16: *Showing all tags annotated with mir29a.*

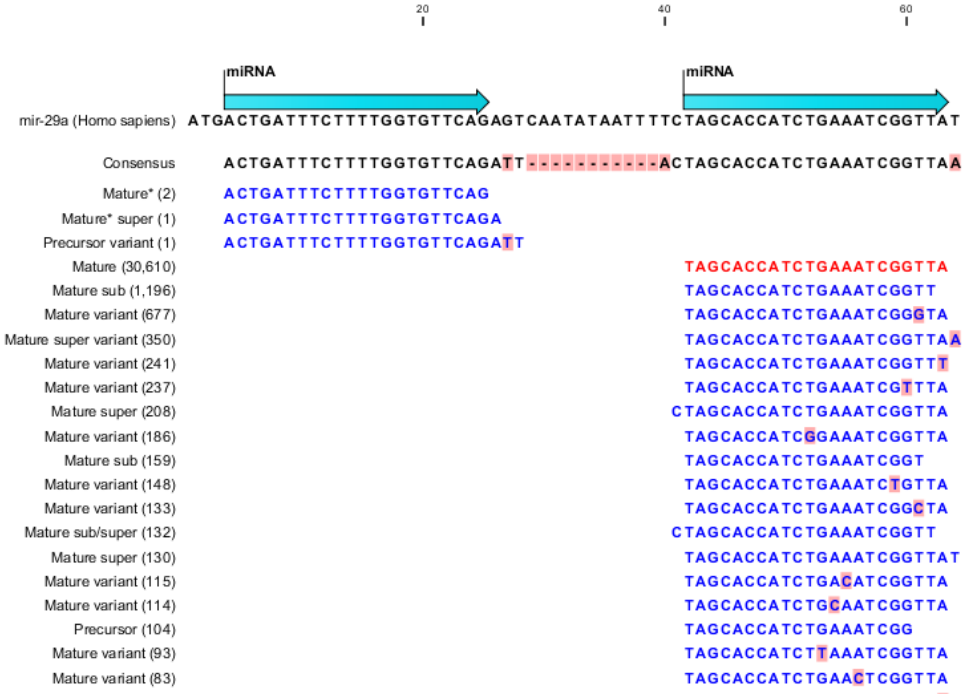


Figure 17: *Showing the alignment against the mir29a precursor sequence.*

Taking advantage of the RNA folding opportunities

One of the advantages of *CLC Genomics Workbench* is the integration between various tools. We are now going to explore the RNA secondary structure prediction tool for this miRNA. First, right-click the mir-29a label in the mapping view and select **Open This Sequence**. This will open this sequence in a new view but it is still part of the mapping and the grouped sample (this is

denoted by the square brackets around its name). We will now predict the secondary structure of this sequence:

Toolbox | **RNA Structure (**) | **Predict Secondary Structure (**)

Click **Next** and **Next** using the default settings, uncheck the option to add annotations and click **Finish**. Then switch to the **Secondary Structure 2D View** ( ) to see the predicted structure (see figure 18).

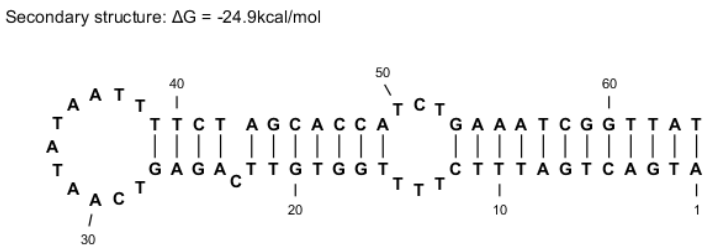


Figure 18: *Showing the folding of the mir-29a precursor.*

If your views are not already split, drag the tabs of the views to create a set-up as shown in figure 19 and select using the mouse either in the secondary structure or the reference in the mapping view and you will be able to follow the selections across the views.

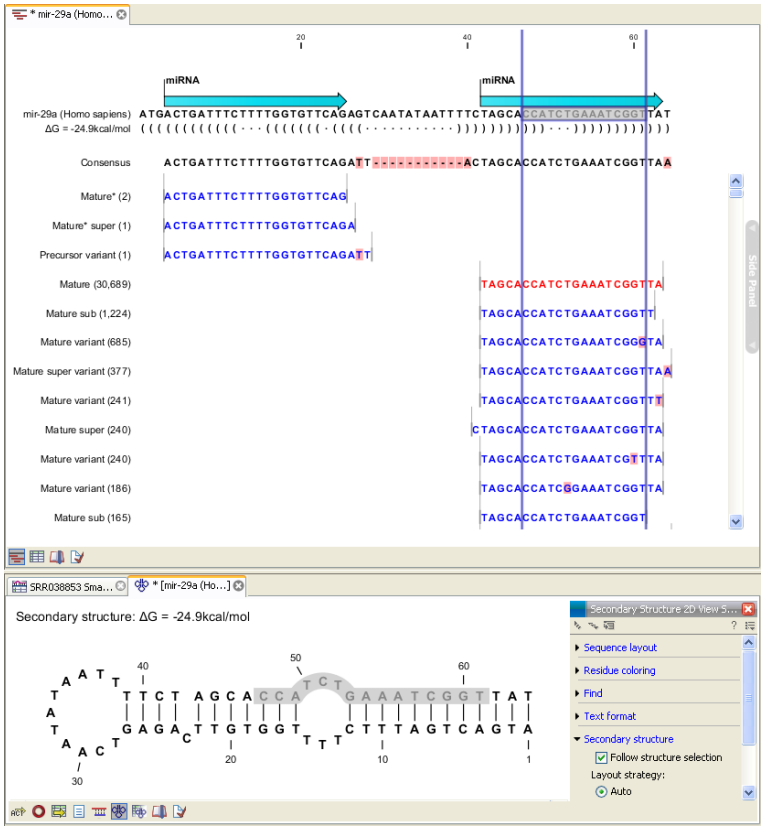


Figure 19: *A split view showing the secondary structure of the RNA together with the length variants.*

Close the views. You are prompted to save the changes which in this case is the adding of the secondary structure to the precursor sequence.

Tracking back from the mature sample

Now, open the `SRR038853 Small RNA sample grouped on mature` () and look at the first row with `let-7f-1//let-7f-2` in the **Name** column. The let-7f miRNA is annotated in miRBase with two different precursor sequences. This means that when the tags are annotated, they are assigned either to let-7f-1 or let-7f-2. The sample grouped on mature merges the tags from precursors sharing the same mature sequence (the sequence itself is shown in the **Feature id** column).

Open the `SRR038853 Small RNA sample grouped` and enter let-7f in the filter. The two precursor variants are now displayed, and you can see that the numbers sum up compared to the mature sample: 254,122 + 254,121 = 508,243 (see the **Mature** column in figure 20).

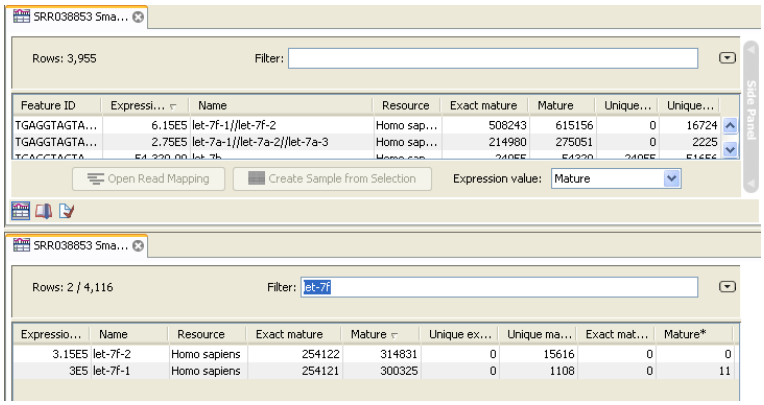


Figure 20: *The mature sample joins the counts of precursors sharing the same mature sequence.*



Bibliography

[Stark et al., 2010] Stark, M. S., Tyagi, S., Nancarrow, D. J., Boyle, G. M., Cook, A. L., Whiteman, D. C., Parsons, P. G., Schmidt, C., Sturm, R. A., and Hayward, N. K. (2010). Characterization of the melanoma mirnaome by deep sequencing. *PLoS One*, 5(3):e9685.