Tutorial

Small RNA Analysis using Illumina Data

October 5, 2016

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com ·
AdvancedGenomicsSupport@qiagen.com

Tutorial

Small RNA Analysis using Illumina Data

This tutorial shows how to go through the initial steps 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 by 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

1. Download the fastq file containing all the raw data using the **Download | Search for Reads in SRA** button.

    - In the SRA search window that opens in the View Area (figure 1), enter SRR038853 in the search field and click on **Start search**.

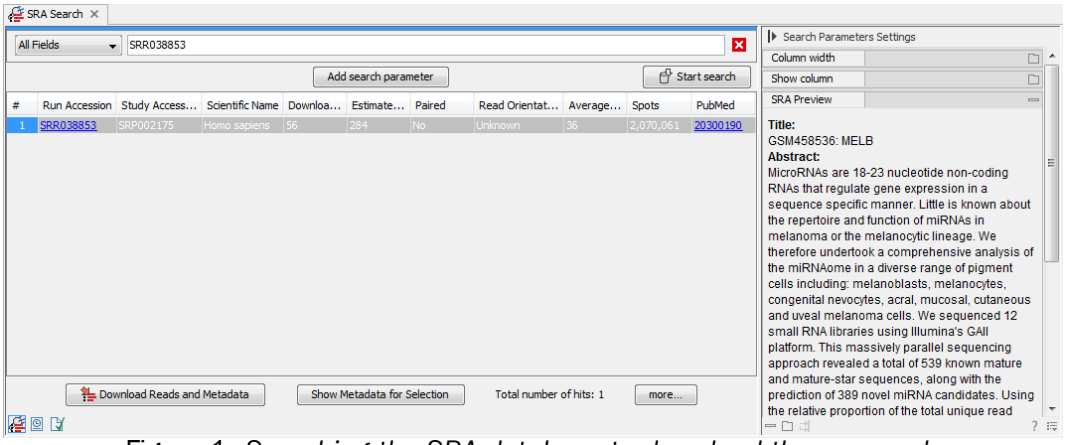


Figure 1: *Searching the SRA database to download the raw reads.*

    - Clicking on the Run Accession number will display the title and abstract of the publication associated with the selected run in the "SRA Preview" panel on the right hand side.
    - Click on the button "Download Reads and Metadata".
    - Make sure the **Discard read names** and **Discard quality scores** checkboxes are checked. Information about read names and quality scores are not used in this analysis so it would just take up disk space if imported with the data. Click **Next**.
    - Choose to **Save** the reads in a new folder you can call **Small RNA tutorial** 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.

2. Download the list of homo sapiens non-coding RNAs from ENSEMBL

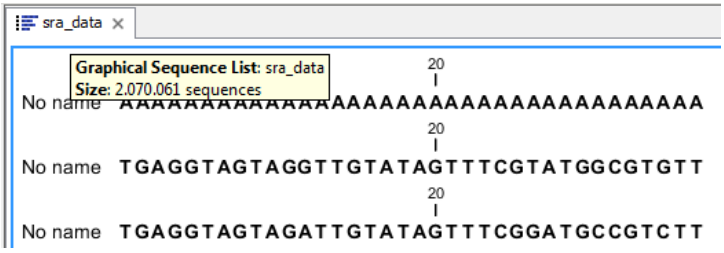


Figure 2: *This data set contains about two million reads.*

ftp://ftp.ensembl.org/pub/release-84/fasta/homo_sapiens/ncrna/Homo_sapiens.GRCh38.ncrna.fa.gz.

3. Import the list with the standard import option:

   File | **Import** | **Standard import** .

   Select the compressed **Homo_sapiens.GRCh37.75.ncrna file**. Click Next, choose to **Save** and click **Finish**.

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. For this you will have to create an adaptor list using the oligonucleotide sequences disclosed in the Illumina Customer Sequence Letter found at this address: http://support.illumina.com/downloads/illumina-customer-sequence-letter.html. [1]

1. To trim an adapter, you have to search for the reverse complement of the relevant adapter sequence on the minus strand. Here we want to trim the 3'RNA adapter of the section "Oligonucleotide sequences for the v1 and v1.5 small RNA kits". In this particular case, the reverse complement of the 3'RNA adapter is the same sequence as the RT Primer of the section "Oligonucleotide sequences for the v1 and v1.5 small RNA kits". A short cut to to creating the Trim Adapter List is then to simply copy the RT primer sequence from the letter.

   Otherwise, in the CLC Genomics Workbench, you could use the following instructions to create a reverse complement of an adapter sequence:

   - Copy the sequence of **3'RNA adapter** from the Illumina Customer Sequence Letter (from the first U nucleotide to the final U).
   - In the workbench, click on **File | New | Sequence**
   - Paste the adapter sequence, give it a name and choose the radio button for **RNA**. Click OK.
   - The adapter opens as a sequence in the View Area. Save it in the Navigation Area.
   - Launch ( ) the **Reverse Complement Sequence** tool.
   - Select the adapter sequence you just saved and click Next.
   - Choose to open the reverse complement of the adapter.you can now copy it.

[1] Oligonucleotide sequences © 2007-2013 Illumina, Inc. All rights reserved. No sponsorship or affiliation. Link provided for convenience. QIAGEN not responsible for content at link.

2. For both workbenches, go to **File** | **New** | **Trim Adapter List (**  **)**

3. Click on the button **Add Row** () found at the bottom of the View Area in the New Adapter Trim List.

4. CLC Genomics Workbench users can paste the sequence of the reverse complement of the **3'RNA adapter** they just generated, and for Biomedical Workbench users just the RT Primer of the section "Oligonucleotide sequences for the v1 and v1.5 small RNA kits". Give it a name, adjust "Strand" to **Minus** and "Action" to **Discard when not found**. Leave the "Alignment scores costs" as default but in the "Match thresholds section", uncheck **Allow internal matches** and increase the **Minimum score at end** to 6 for "Allow end matches" (figure 3). Click on the button labeled **Finish** to create the adapter trim list.

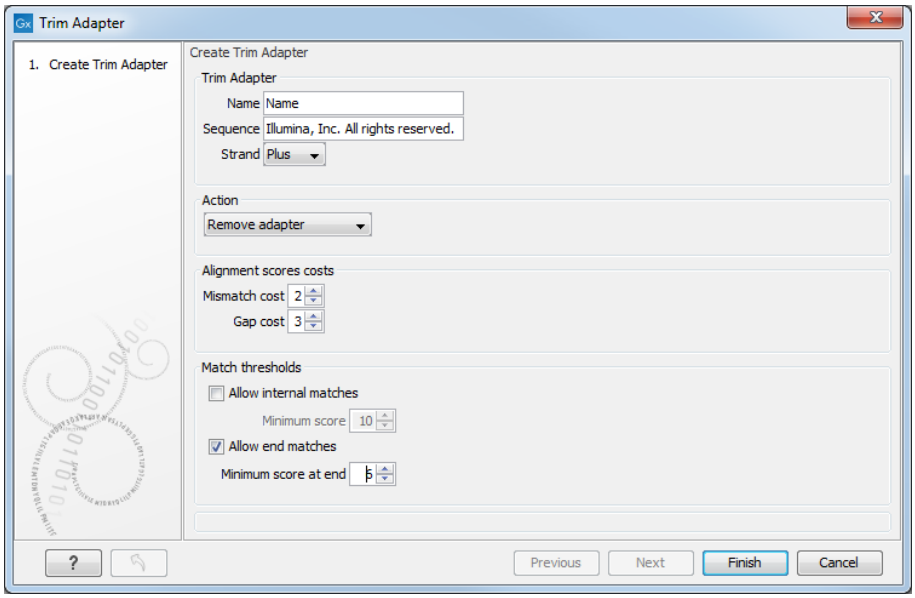


Figure 3: *Creating the adapter list.*

5. Save the generated adapter trim list as **New_Trim_Adapter_List** in the **Navigation Area**. You can do this by clicking on the tab and dragging and dropping the adapter trim list to the desired destination, or you can go to **File** in the menu bar and the choose **Save as**.

6. Now we are ready to trim the adapters from our reads. Launch () the **Extract and Count** () tool.

   This opens a dialog where you select the `SRR038853` sample as shown in figure 4. Click **Next**.

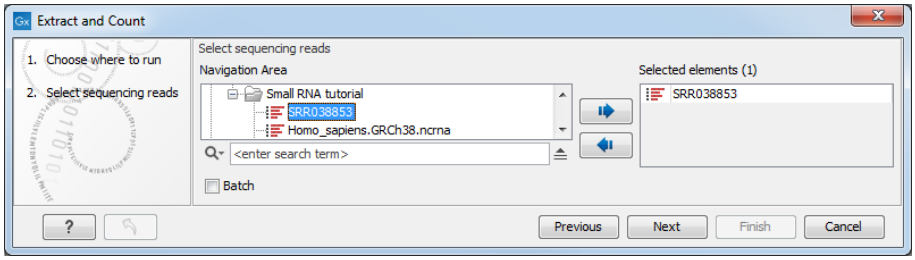


Figure 4: *Selecting the sample for extracting and counting the small RNAs.*

7. In the "Set trim options" dialog, make sure the checkbox for "Perform custom adapter trimming before counting" is checked (it is by default) and click Next.

8. In the next window (figure 5), select the adapter trim list.

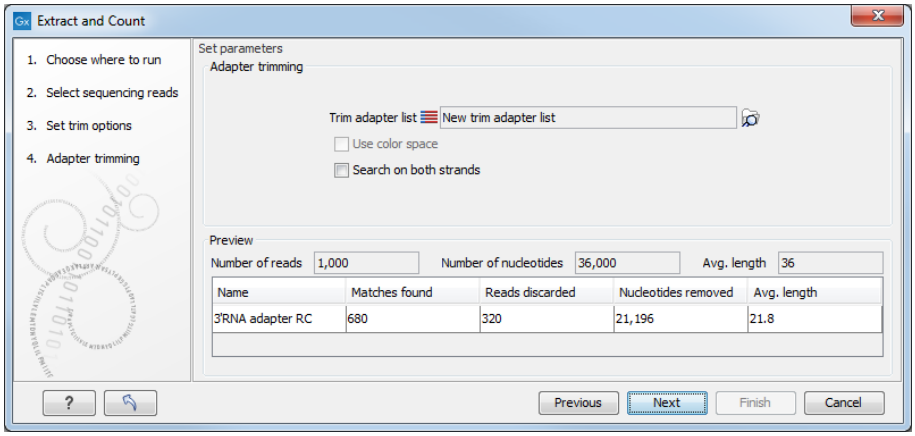


Figure 5: *Trimming the raw data for adapter sequence.*

In the preview panel below you can see the number of matches found among the first 1000 reads for this adapter. 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. Click **Next**.

9. You will now see the dialog shown in figure 6. The most important settings here are the minimum and maximum lengths for the tags you wish to include when counting and how many copies there have to be for a particular tag to be included in the output.

   Leave these options set to the default values and click the button labeled **Next**.

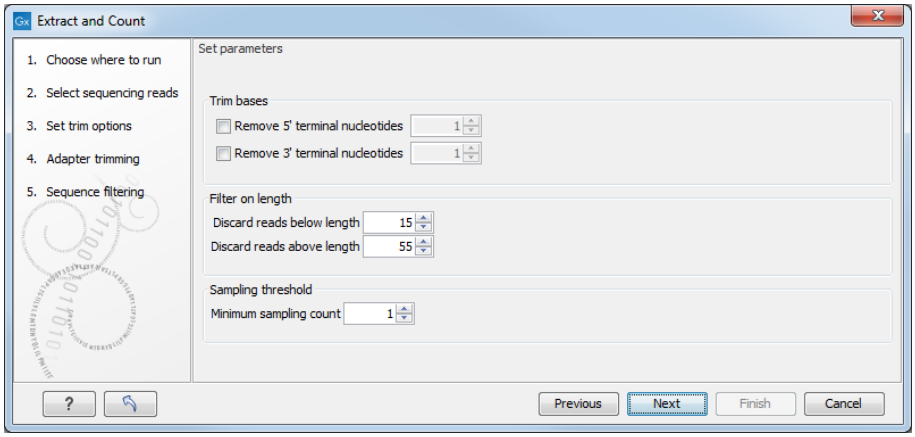


Figure 6: *Adjusting options for counting the small RNAs.*

10. You are now asked 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 default settings, choose to **Open** the results rather than Save them, and click on the button labeled **Finish**.

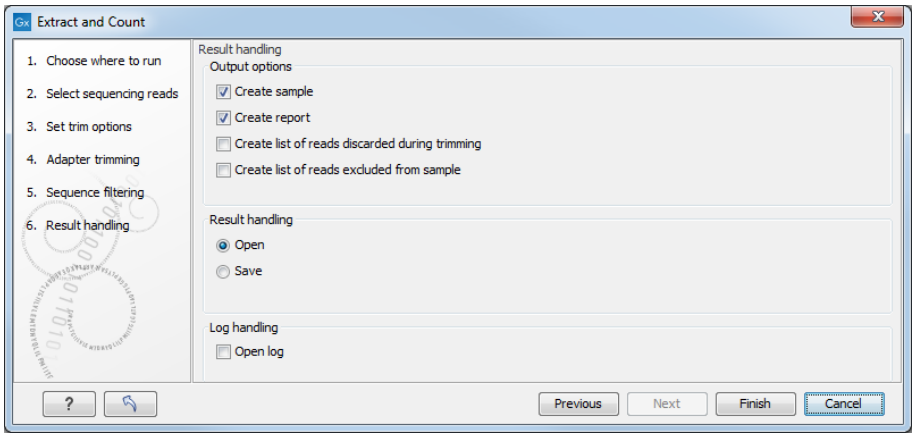
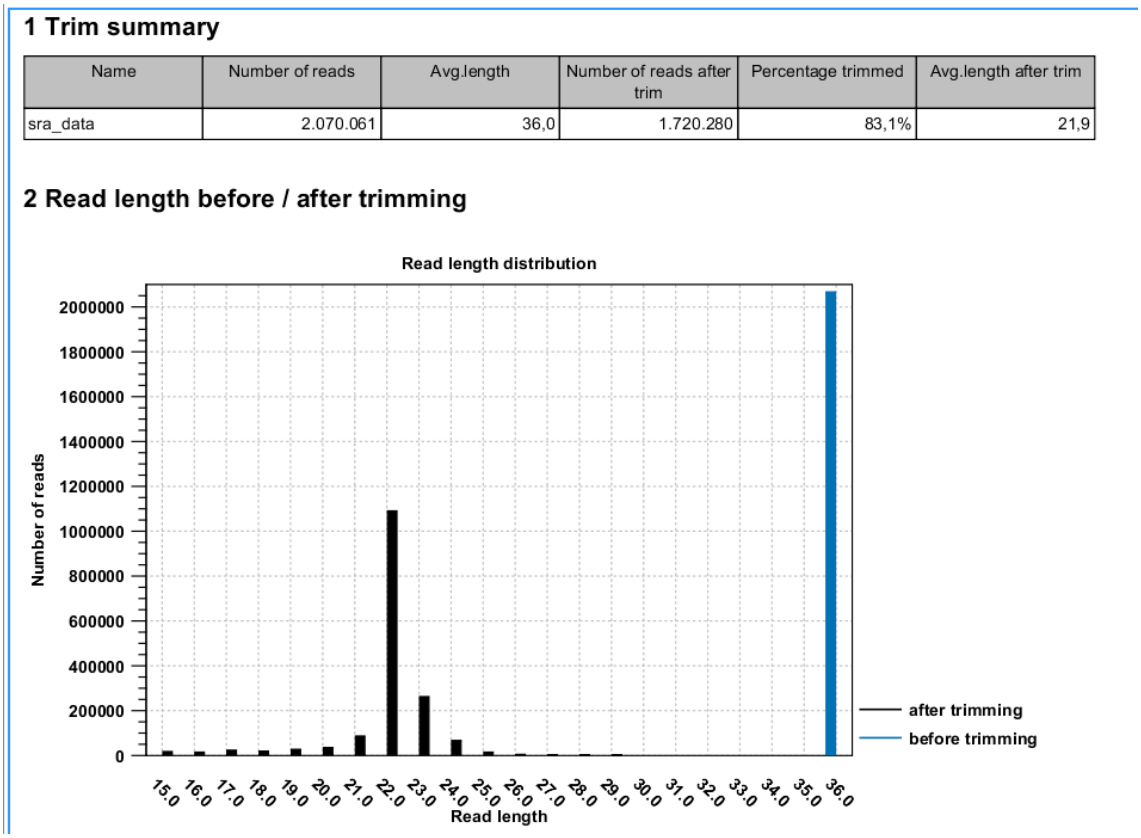


Figure 7: *Selecting the results to output.*

Interpreting the adapter trim report

Once the analysis is complete, two tabs will be opened. First, we take a look at the report (figure 8).



1 Trim summary

Name	Number of reads	Avg.length	Number of reads after trim	Percentage trimmed	Avg.length after trim
sra_data	2.070.061	36,0	1.720.280	83,1%	21,9

2 Read length before / after trimming

Figure 8: *The small RNA counting report.*

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 more than 83% of the reads were trimmed. The trim settings set earlier implied that if no adapter sequence was found, the read would have been discarded. A graph showing a distribution of the read lengths before and after trimming. In this example, there

is a very nice distribution with a peak around 22 bp which is expected for miRNAs.

Save and close the report.

Investigating the small RNA sample

You should now see the small RNA sample. There are more than 88,000 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 open the trimmed reads of one of the small RNAs as shown in figure 9.

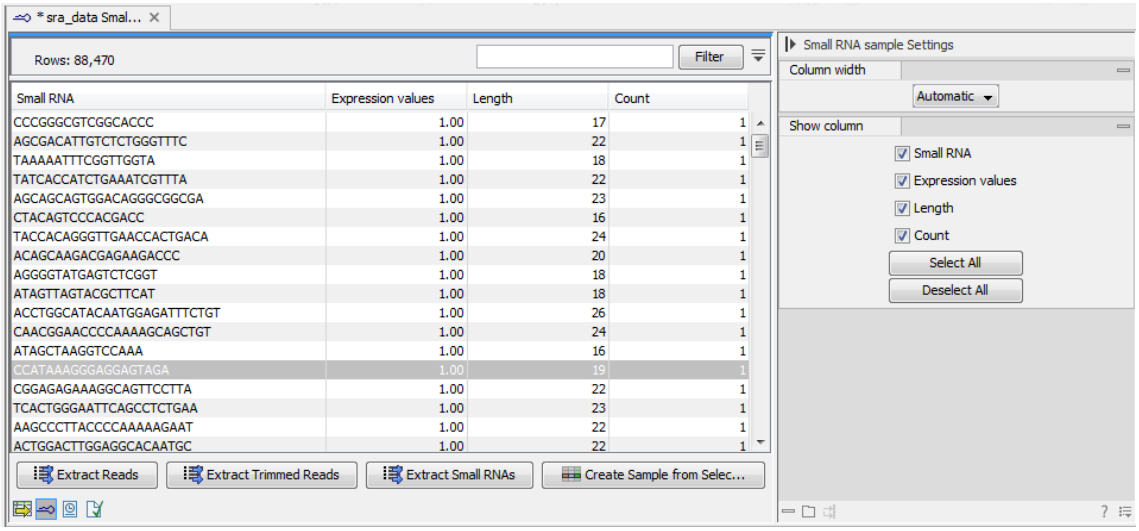


Figure 9: *The small RNA sample tabel with all possible columns selected for view.*

1. Sort the table on **Length** by clicking on that column header once. This should sort the table so that the entries with the shortest lengths are at the top. Now click the row at the top of the table so it is selected.

2. Click on the **Extract Reads** button. Keep the option **DNA** and choose to **Open** the result. Click on the button labeled **Finish**.

   You should now be able to see the original read sequence(s) and a trim annotation as shown in figure 10.

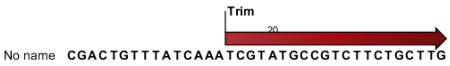


Figure 10: *A sequencing read with the trim annotation.*

3. Clicking the **Double stranded** checkbox in the Side Panel to the right under **Sequence Layout**, you can see the minus strand as well, and you can see that the adapter sequence has a perfect match in the reverse orientation here (figure 11).

4. Save and close the small RNA sample.

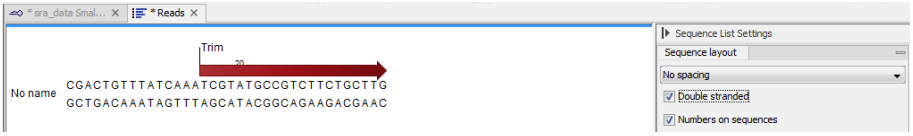


Figure 11: *A sequencing read shown as double stranded with the trim annotation.*

Downloading miRBase and annotating the sample

The next step in the analysis is to annotate the small RNA sample so known small RNAs can be identified. We use two sources for the annotation here:

- miRBase, which will be used to identify known miRNAs
- a set of other known non-coding RNAs, which were downloaded at the beginning of this tutorial.

1. You can download the latest version of miRBase directly via the Workbench by using the **Download miRBase ( )** tool.
   - Search for the tool using the Launch ( ) button.
   - Choose to **Save** the files in the folder created for this tutorial and click on **Finish**.
2. Launch the **Annotate and Merge Counts ( )** tool to start annotating.
3. This opens a dialog where you select the **sra-data Small RNA sample** as shown in figure 12. Click **Next**.

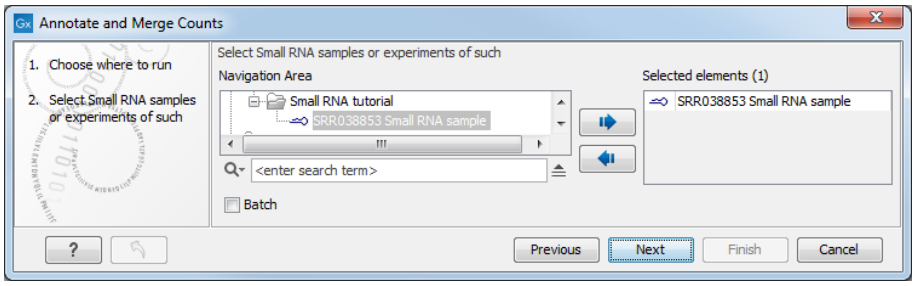


Figure 12: *Selecting the sample for annotating the small RNAs.*

4. You are now presented with the dialog shown in figure 13. At the top, check to use miRBase and select ( ) the **miRBase - Release 21** file that you have just downloaded. Below, check the **Use other resource** and select ( ) the **Homo_sapiens.GRCh38.ncrna** file that you imported in the beginning of this tutorial. Leave the other option set such that miRBase has the highest priority.

   The miRBase file contains a list of precursor sequences with specification of the mature 5' and in some cases the mature 3' regions. This information is used to categorize the annotated small RNAs. The "Other resource" 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 had other sources of non-coding small RNAs.

   Click **Next**.

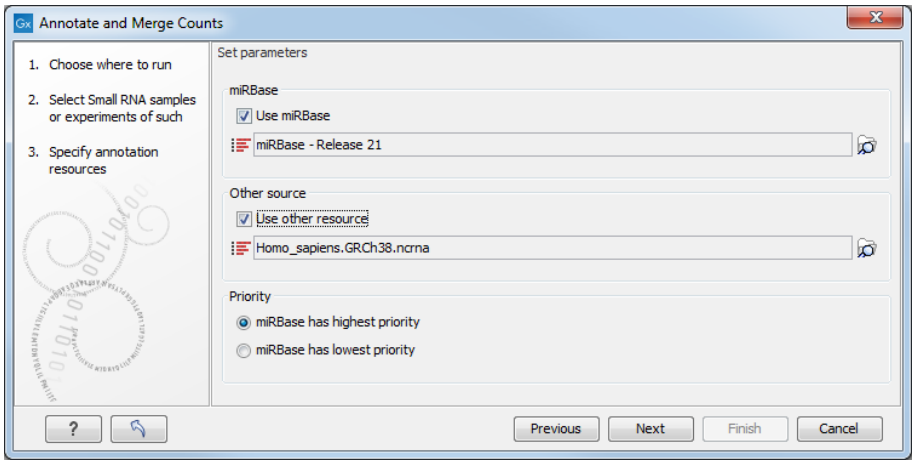


Figure 13: *Setting miRBase and the other non-coding RNAs as annotation sources.*

5. In this window (figure 14), select **Homo sapiens** and then select **Mus musculus**. Note that *the order in which these are listed is important* as it is the order in which the resources will be used for the purpose of annotation. The sample is human, so human annotations should have the first priority and thus should be the first source listed. Since there may be miRNAs that have not yet been identified in human but have an ortholog in mouse, including the mouse miRNAs may provide useful information. Click **Next**.

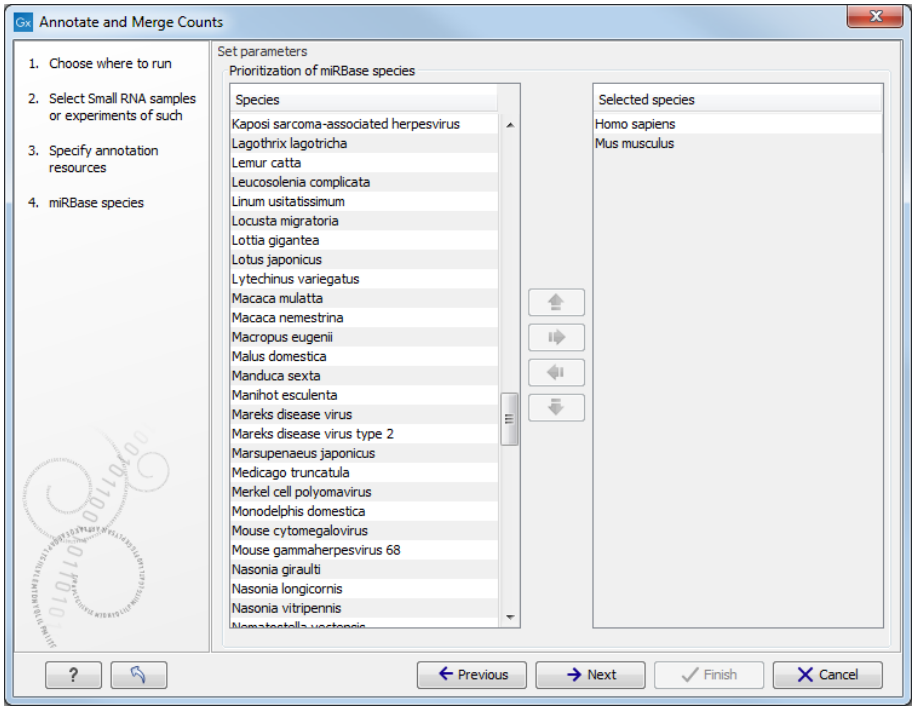


Figure 14: *Prioritizing species for annotation.*

6. In the "match parameters" dialog (figure 15), leave all settings at their default values and click **Next**.

7. In the dialog shown in figure 16, make sure all options except the "Create unannotated samples/experiments" are checked, choose to **Save** the results and click **Next**. You can now select the folder in which to save them. Click on the button labeled **Finish**.

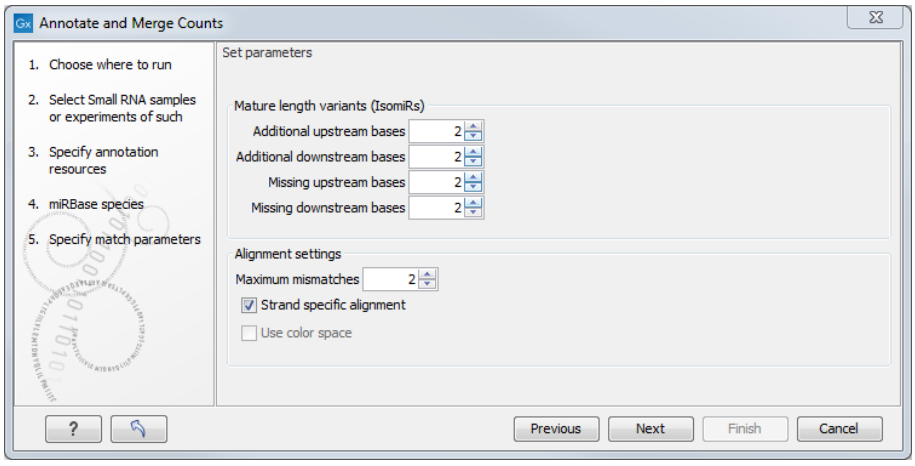


Figure 15: *Thresholds for annotating.*

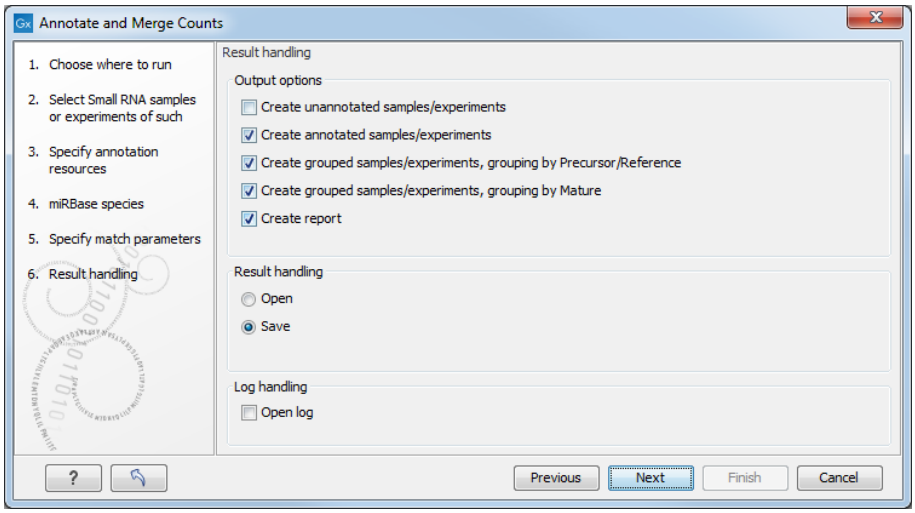


Figure 16: *Select the sample grouped on mature 5'.*

Analyzing the annotated samples

For a detailed description of the output, please refer to the user manual (press the **F1** key to display it).

In this example we focus on a few specific miRNAs to illustrate how the annotation and grouping of samples work and show some possibilities for interacting with the data.

Looking at the grouped sample Open the **SRR038853 small RNA Sample annotated** ( ). We want to look at *mir-29a*, so type this into the filter at the top of the table as shown in figure 17 and click on the button labeled **Filter**.

This will list all the tags that have been mapped to the mir-29a precursor sequence from miRBase. If you sort the table by **Count** (clicking the count column header twice to show the highest count on top of the table), you can see that most of these are exact matches of the mature 3' miRNA (Match type is set to Mature 3'). 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 `sra-data Small RNA sample grouped` ( ) and type in mir-29a in the filter. You now have two lines representing all the tags that have been annotated with mir-29a.)

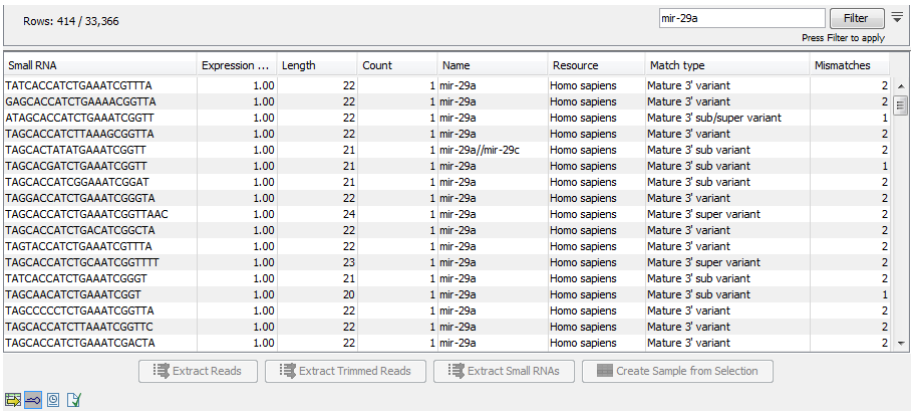


Figure 17: *Showing all tags annotated with mir29a.*

The upper row shows mir-29a from human and the lower row shows mir-29a from mouse. The total number of counts for the human mir-29a is 36,778. The number of reads in different categories are shown, e.g. 30,689 for the exact mature 3' corresponding to the number from the ungrouped sample in figure 17. 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 mapping of all the tags to the precursor sequence (see figure 18).

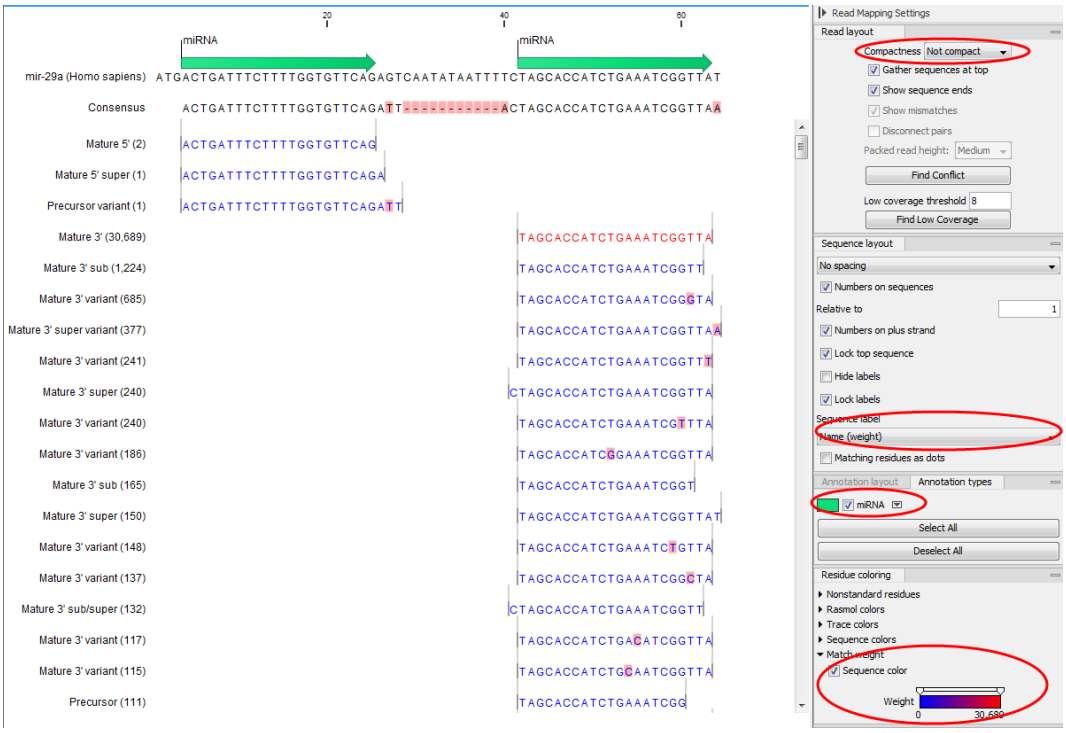


Figure 18: *Showing the alignment against the mir29a precursor sequence.*

The tags are colored to reflect the counts, which are also shown in numbers next to the name to the left. As the exact mature 3' is very dominant in terms of count, it is the only one standing out in a different color. Note that it may be necessary to adjust the user settings in the side panel to obtain the same view as shown in figure 18.

Tracking back from the sample grouped on mature

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 5' 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 to the numbers given in the sample grouped on mature: 314,813 + 300,337 = 615,161 (see the **Mature 5'** column in figure 19).

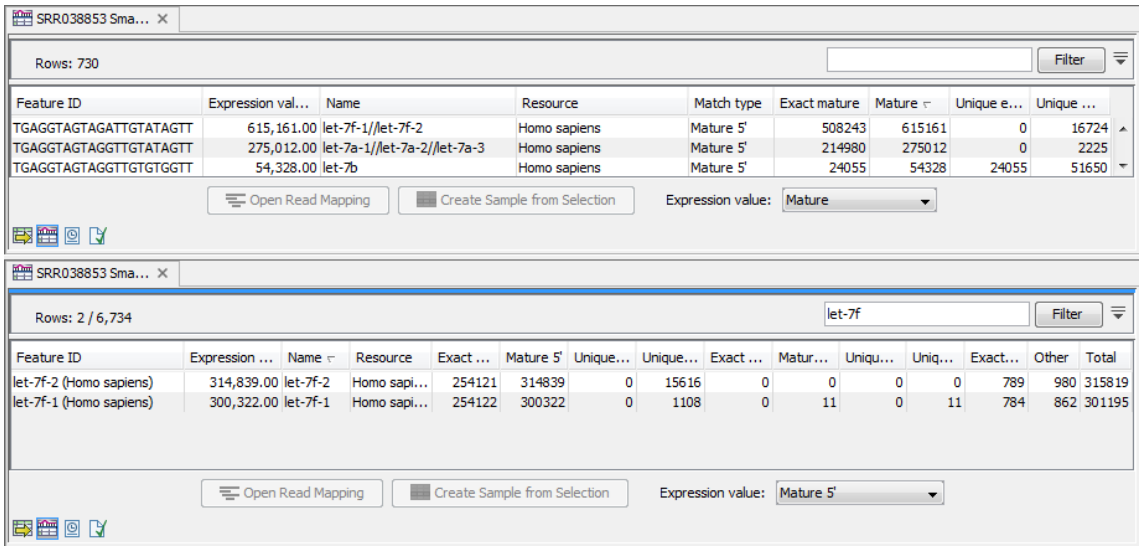


Figure 19: *The sample grouped on mature joins the counts of precursors sharing the same mature sequence.*

Taking advantage of the RNA folding opportunities in the CLC Genomics Workbench

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 using this miRNA.

1. Right-click on let-7f-2 (Homo sapiens) in the filtered **SRR038853 Small RNA sample grouped** and choose to "Open read mapping" (see figure 20).
2. In the read mapping, right-click the let-7f-2 (Homo sapiens) label at the top of the mapping and select **Open Sequence**.
3. This will open this sequence in a new view. However, it is still part of the mapping and the grouped sample (this is denoted by the square brackets around its name in the tab of this new view).
4. We can now predict the secondary structure of this sequence by launching the tool **Predict Secondary Structure** ( ) (choose the tool that belongs to the RNA Structure folder).
5. The sequence let-7f-2 is preselected. Click **Next** and **Next** using the default settings. Uncheck the output option to add annotations and click on the **Finish** button.

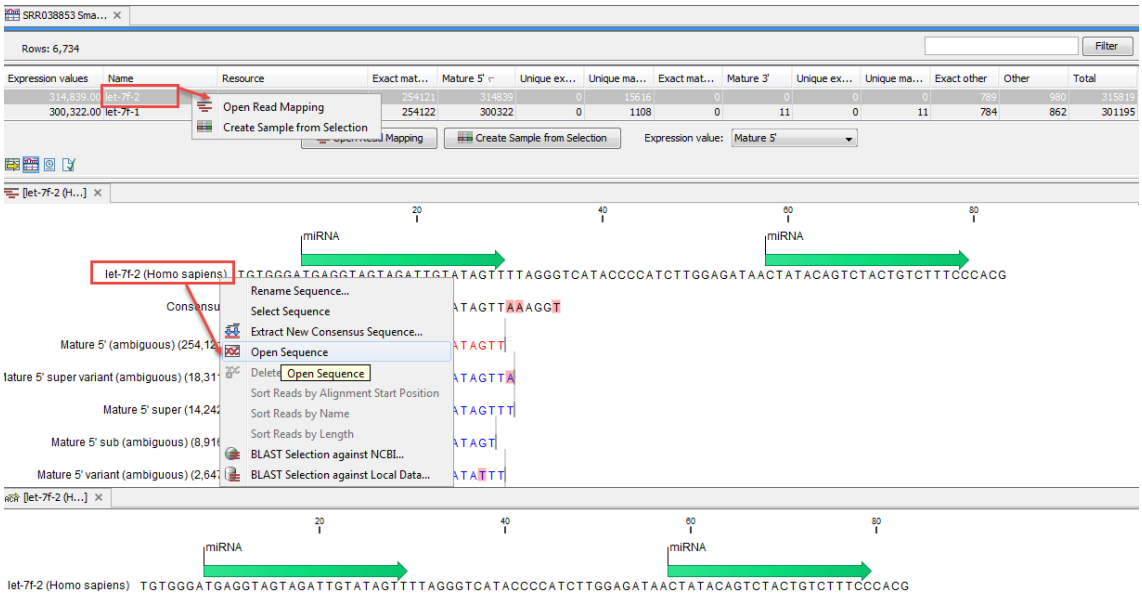


Figure 20: *Showing the folding of the mir-29a precursor.*

6. Switch to the **Secondary Structure 2D View** ( ) to see the predicted structure.

7. If your views are not already split, drag the tabs of the views to create a set-up as shown in figure 21 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 21: *A split view showing the secondary structure of the RNA together with the length variants.*

8. Close the views. You are prompted to save the changes, which in this case is the addition of the secondary structure to the precursor 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.