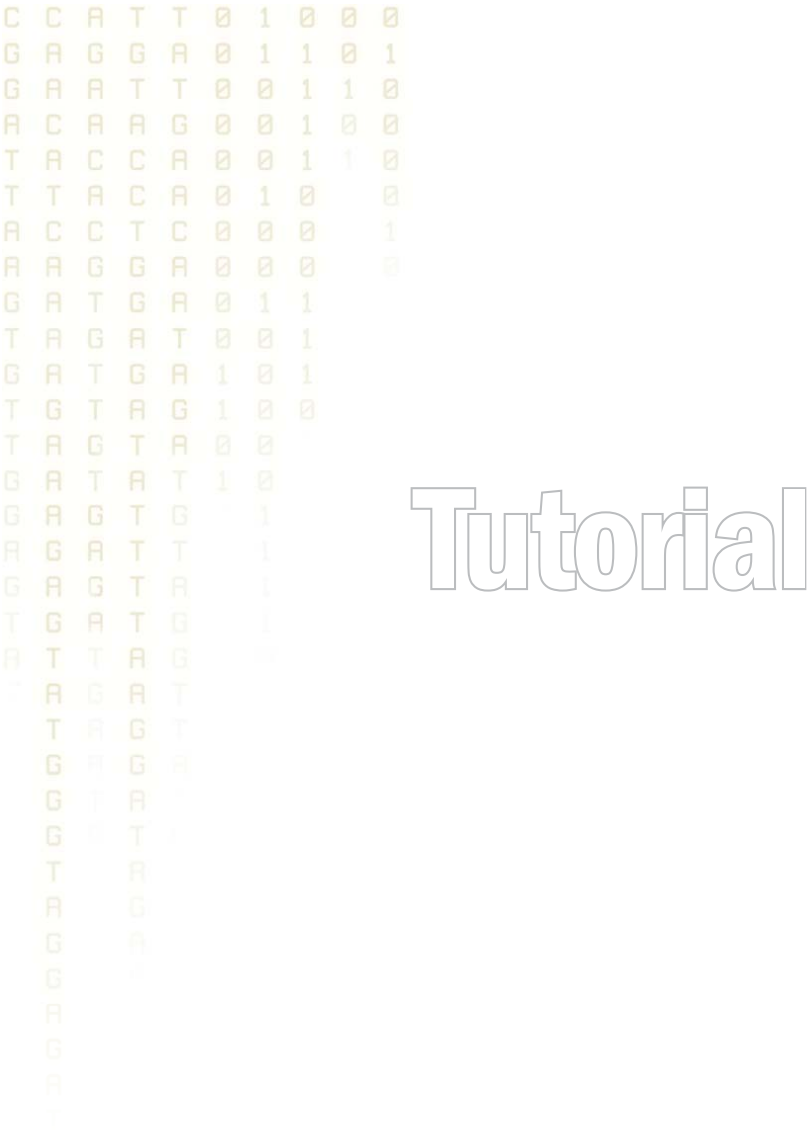


Tutorial: Assembly

June 21, 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 · support@clcbio.com

Tutorial: Assembly

In this tutorial, you will see how to assemble data from automated sequencers into a contig and how to find and inspect any conflicts that may exist between different reads.

This tutorial shows how to assemble sequencing data generated by conventional "Sanger" sequencing techniques. For high-throughput sequencing data, we refer to the *CLC Genomics Workbench* (see http://www.clcbio.com/genomics).

The data used in this tutorial are the sequence reads in the "Sequencing reads" folder in the "Sequencing data" folder of the **Example data** in the **Navigation Area**. If you do not have the example data, please go to the **Help** menu to import it.

Trimming the sequences

The first thing to do when analyzing sequencing data is to trim the sequences. Trimming serves a dual purpose: it both takes care of parts of the reads with poor quality, and it removes potential vector contamination. Trimming the sequencing data gives a better result in the further analysis.

Toolbox in the Menu Bar | Sequencing Data Analyses ( ) | Trim Sequences ( )

Select the 9 sequences and click **Next**.

In this dialog, you will be able to specify how this trimming should be performed.

For this data, we wish to use a more stringent trimming, so we set the limit of the quality score trim to 0,02 (see figure 1).

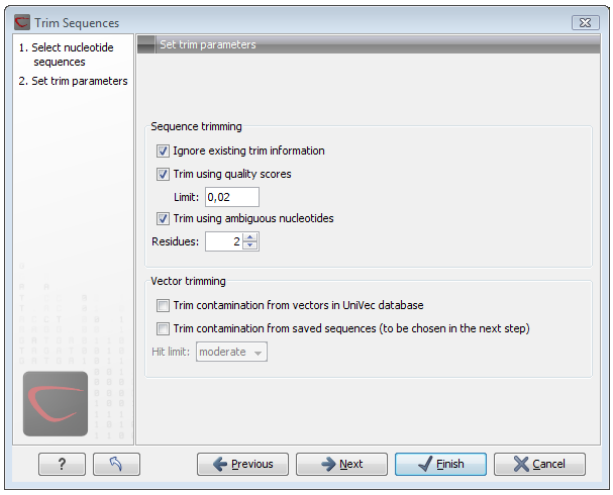


Figure 1: *Specifying how sequences should be trimmed. A stringent trimming of 0,02 is used in this example.*

There is no vector contamination in these data, se we only trim for poor quality.

If you place the mouse cursor on the parameters, you will see a brief explanation.

Click **Next** and choose to **Save** the results.

When the trimming is performed, the parts of the sequences that are trimmed are actually annotated, not removed (see figure 2). By choosing **Save**, the Trim annotations will be saved directly to the sequences, without opening them for you to view first.

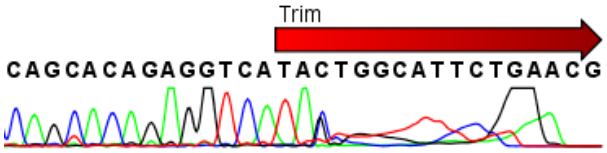


Figure 2: *Trimming creates annotations on the regions that will be ignored in the assembly process.*

These annotated parts of the sequences will be ignored in the subsequent assembly.

A natural question is: Why not simply delete the trimmed regions instead of annotating them? In some cases, deleting the regions would do no harm, but in other cases, these regions could potentially contain valuable information, and this information would be lost if the regions were deleted instead of annotated. We will see an example of this later in this tutorial.

Assembling the sequencing data

The next step is to assemble the sequences. This is the technical term for aligning the sequences where they overlap and reverse the reverse reads to make a contiguous sequence (also called a contig).

In this tutorial, we will use assembly to a reference sequence. This can be used when you have a reference sequence that you know is similar to your sequencing data.

> **Toolbox in the Menu Bar | Sequencing Data Analyses ( ) | Assemble Sequences to Reference ( )**

In the first dialog, select the nine sequencing reads and click **Next** to go to the second step of the assembly where you select the reference sequence.

Click the **Browse and select** button ( ) and select the "ATP8a1 mRNA (reference)" from the "Sequencing data" folder (see figure 3). You can leave the other options in this window set to their defaults.

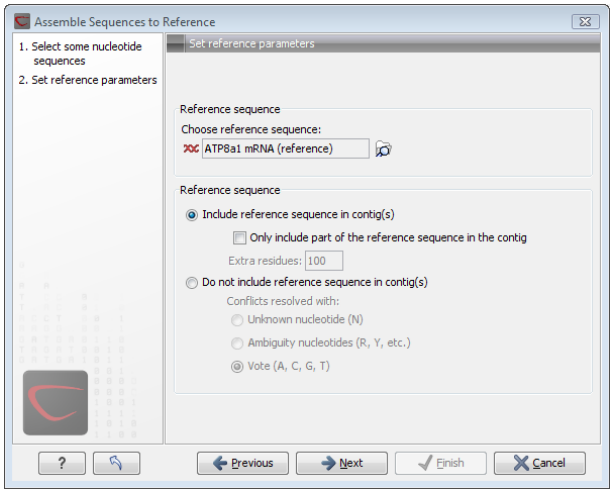


Figure 3: *The "ATP8a1 mRNA (reference)" sequence selected as reference sequence for the assembly.*

Click **Next** and choose to use the trim information (that you have just added).

Click **Next** and choose to Save your results. The next step will ask you for a location to save the

results to. You can just accept the default location, or you could use the left hand icon under the "Save in folder" heading to create a new folder to save your assembly into.

Click **Finish** and the assembly process will begins.

Getting an overview of the contig

The result of the assembly is a **Contig** which is an alignment of the nine reads to the reference sequence. Click **Fit width** ( ) to see an overview of the contig. To help you determine the coverage, display a coverage graph (see figure 4):

Alignment info in Side Panel | Coverage | Graph

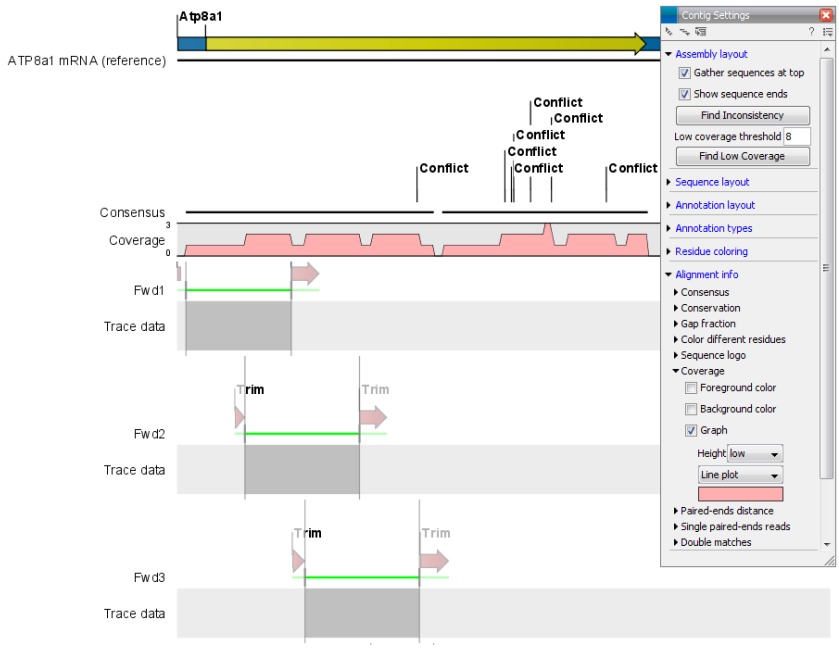


Figure 4: *An overview of the contig with the coverage graph.*

This overview can be an aid in determining whether coverage is satisfactory, and if not, which regions a new sequencing effort should focus on. Next, we go into the details of the contig.

Finding and editing conflicts

Click **Zoom to 100%** ( ) to zoom in on the residues at the beginning of the contig. Click the **Find Conflict** button at the top of the **Side Panel** or press the **Space** key to find the first position where there is disagreement between the reads (see figure 5).

In this example, the first read has a "T" (marked with a light-pink background color), whereas the second line has a gap. In order to determine which of the reads we should trust, we assess the quality of the read at this position.

A quick look at the regularity of the peaks of read "Rev2" compared to "Rev3" indicates that we should trust the "Rev2" read. In addition, you can see that we are close to the end of the end of "Rev3", and the quality of the chromatogram traces is often low near the ends.

Based on this, we decide not to trust "Rev3". To correct the read, select the "T" in the "Rev3"



Figure 5: *Using the Find Conflict button highlights conflicts.*

sequence by placing the cursor to the left of it and dragging the cursor across the T. Press **Delete** ( ).

This will resolve the conflict.

Including regions that have been trimmed off

Clicking the **Find Conflict** button again will find the next conflict.

This is the beginning of a stretch of gaps in the consensus sequence. This is because the reads have been trimmed at this position. However, if you look at the read at the bottom, *Fwd2*, you can see that a lot of the peaks actually seem to be fine, so we could just as well include this information in the contig.

If you scroll a little to the right, you can see where the trimmed region begins. To include this region in the contig, move the vertical slider at position 2073 to the left (see figure 6).

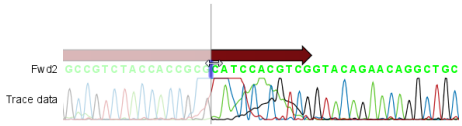


Figure 6: *Dragging the edge of the trimmed region.*

You will now see how the gaps in the consensus sequence are replaced by real sequence information.

Note that you can only move the sliders when you are zoomed in to see the sequence residues.

Inspecting the traces

Clicking the **Find Conflict** button again will find the next conflict.

Here both reads are different than the reference sequence. We now inspect the traces in more detail. In order to see the details, we zoom in on this position:

Zoom in in the Tool Bar ( ) | Click the selected base | Click again three times

Now you have zoomed in on the trace (see figure 7).

This gives more space between the residues, but if we would like to inspect the peaks even

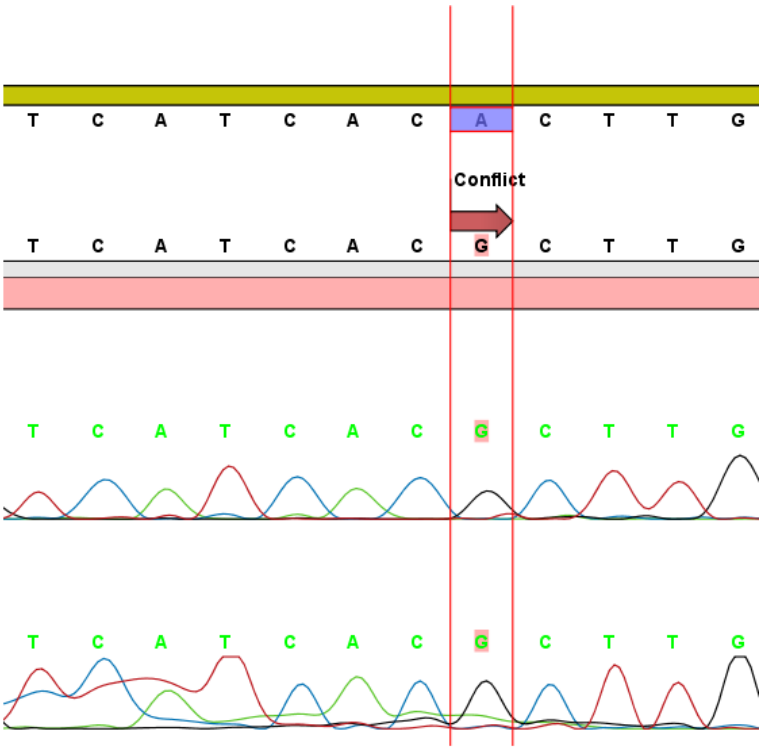


Figure 7: *Now you can see all the details of the traces.*

more, simply drag the peaks up and down with your mouse (see figure 8).

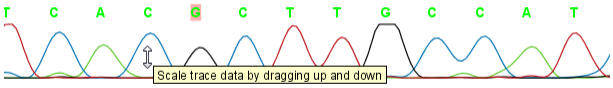


Figure 8: *Grab the traces to scale.*

Synonymous substitutions?

In this case we have sequenced the coding part of a gene. Often you want to know what a variation like this would mean on the protein level. To do this, show the translation along the contig:

> **Nucleotide info in the Side Panel | Translation | Show | Select ORF/CDS in the Frame box**

The result is shown in figure 9.

You can see that the variation is on the third base of the codon coding for threonine, so this is a synonymous substitution. That is why the T is colored yellow. If it was a non-synonymous substitution, it would be colored in red.

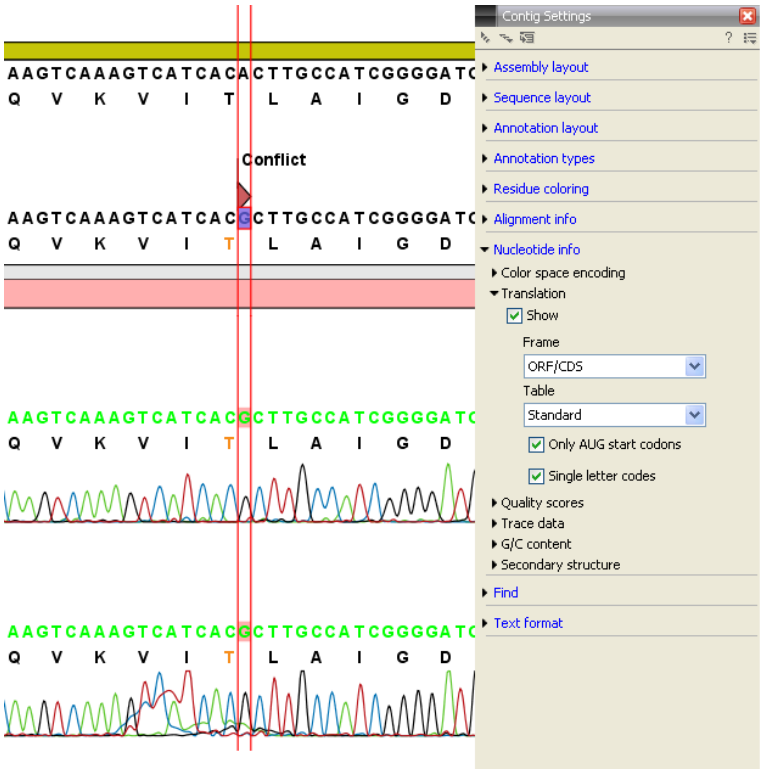


Figure 9: *Showing the translation along the contig.*

Getting an overview of the conflicts

Browsing the conflicts by clicking the **Find Conflict** button is useful in many cases, but you might also want to get an overview of all the conflicts in the entire contig. This is easily achieved by showing the contig in a table view:

> **Press and hold the Ctrl-button (⌘ on Mac) | Click Show Table ( ) at the bottom of the view**

This will open a table showing the conflicts. You can right-click the **Note** field and enter your own comment. In this dialog, enter a new text in the **Name** and click **OK**.

When you edit a comment, this is reflected in the conflict annotation on the consensus sequence. This means that when you use this sequence later on, you will easily be able to see the comments you have entered. The comment could be e.g. your interpretation of the conflict.

Documenting your changes

Whenever you make a change like deleting a "T", it will be noted in the contig's history. To open the history, click the fHistory ( ) icon at the bottom of the view.

In the history, you can see the details of each change (see figure 10).

Using the result for further analyses

When you have finished editing the contig, it can be saved, and you can also extract and save the consensus sequence:

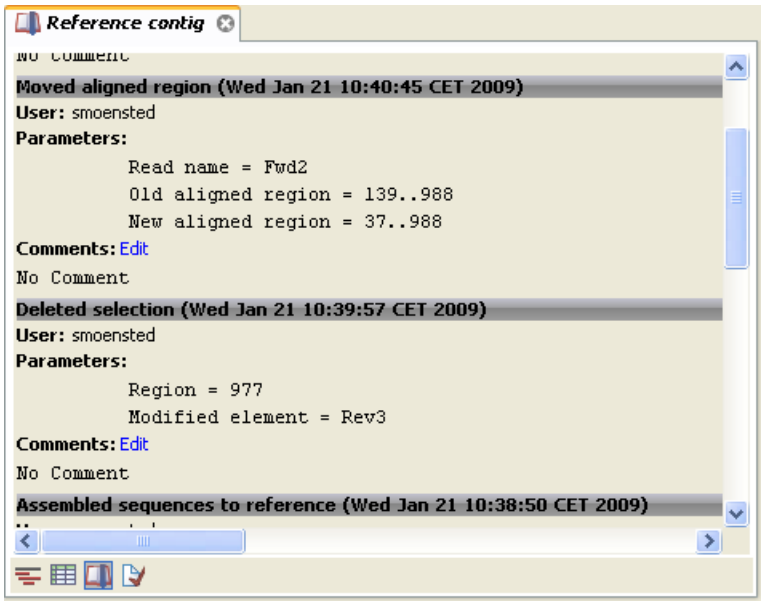


Figure 10: *The history of the contig showing that a "T" has been deleted and that the aligned region has been moved.*

Right-click the name "Consensus" | Open Copy of Sequence | Save ( )

This will make it possible to use this sequence for further analyses in the *CLC DNA Workbench*. All the conflict annotations are preserved, and in the sequence's history, you will find a reference to the original contig. As long as you also save the original contig, you will always be able to go back to it by choosing the Reference contig in the consensus sequence's history (see figure 11).

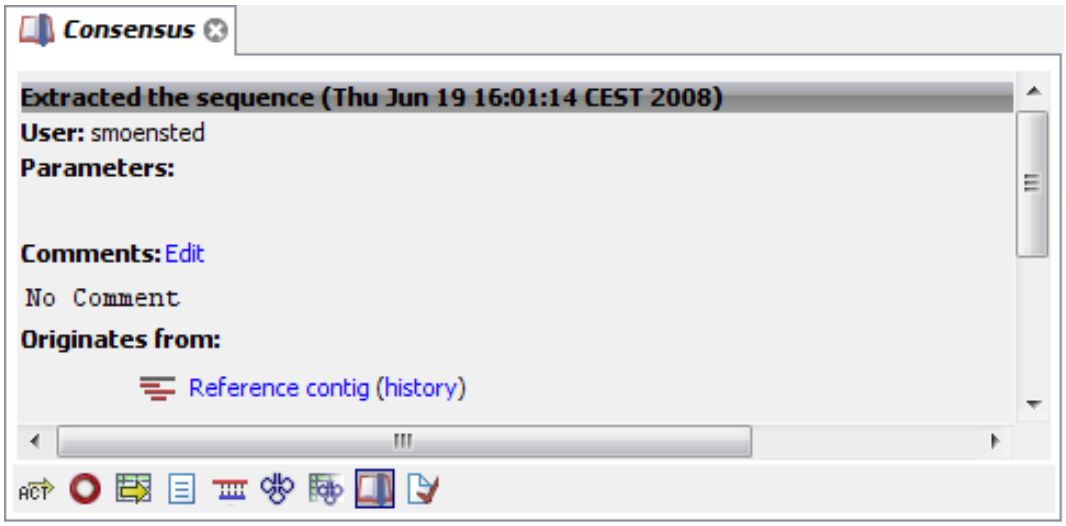


Figure 11: *The history of the consensus sequence, which has been extracted from the contig. Clicking the blue text "Reference contig" will find and highlight the name of the saved contig in the Navigation Area. Clicking the blue text "history" to the right will open the history view of the earlier contig. From there, you can choose other views, such as the Read mapping view, of the contig.*