Tutorial

Assemble to Reference

November 1, 2019

Sample to Insight

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

Tutorial

Assemble to Reference

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 using the *CLC Main Workbench*. For high-throughput sequencing data, use the *CLC Genomics Workbench* specific tutorials on De Novo assemblies and read mapping.

The data used in this tutorial are the sequence reads found in the Example Data folder that can be imported to the Navigation Area from the **Help** menu.

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.

1. In the Toolbox, go to: **Molecular Biology Tools** | **Sanger Sequencing Analysis ( )** | **Trim Sequences ( )**
2. Select the 9 sequences found in Example Data | Sequencing data | Sequencing reads and click **Next** (see figure 1).

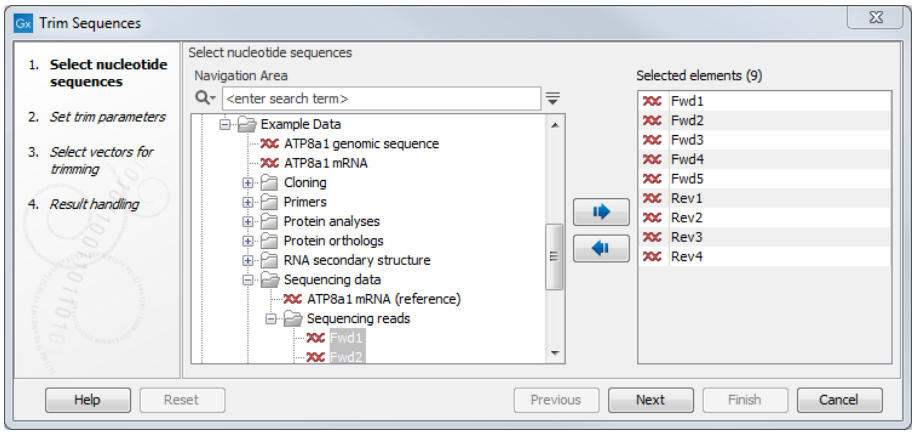

Figure 1: *Specifying how sequences should be trimmed. A stringent trimming of 0.02 is used in this example.*

3. In the next dialog, you will be able to specify how this trimming should be performed (see figure 2). For this data, we wish to use a more stringent trimming, so we set the limit of the quality score trim to 0.02.

   There is no vector contamination in these data, so we only trim for poor quality.
4. Choose to **Save** the results and click **Finish**.

When the trimming is performed, the parts of the sequences that are trimmed are actually annotated, not removed (see figure 3). By choosing **Save**, the Trim annotations will be saved directly to the input sequences, without opening them for you to view first.

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

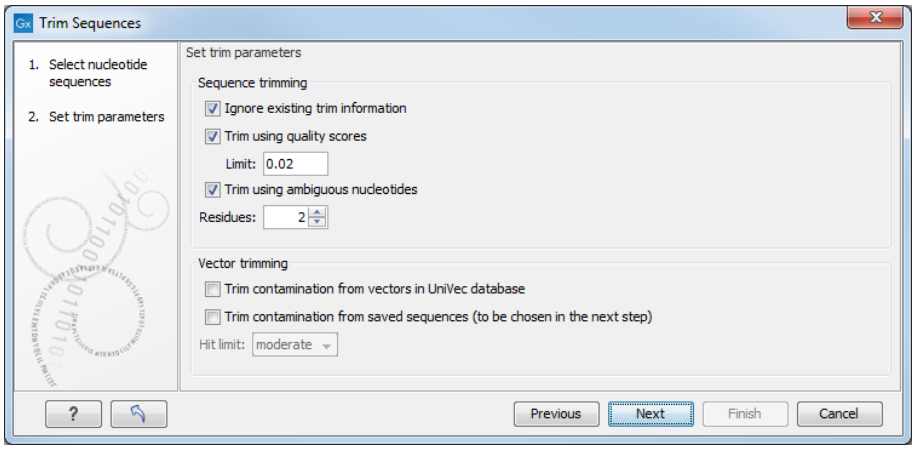

Figure 2: *Specifying how sequences should be trimmed. A stringent trimming of 0.02 is used in this example.*

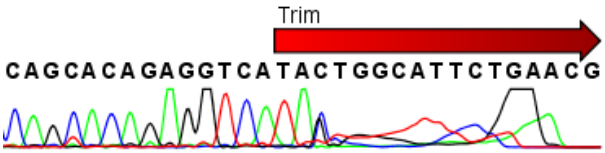

Figure 3: *Trimming creates annotations on the regions that will be ignored in the assembly process.*

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.

1. In the Toolbox, go to:
   Molecular Biology Tools | **Sanger Sequencing Analysis** ( ) | **Assemble Sequences to Reference** ( )
2. In the first dialog, select the nine sequencing reads again and click **Next**.
3. Then select the reference sequence. Click the **Browse and select** button ( ) and select the `ATP8a1 mRNA (reference)` from the Sequencing data folder (see figure 4). You can leave the other options in this window set to their defaults before clicking **Next**.
4. In the Set assemble parameters dialog, leave most options at their default values, but choose under the Trimming options section to "Use existing trim information" (that you have just added). This is shown in figure 5. Click **Next**.

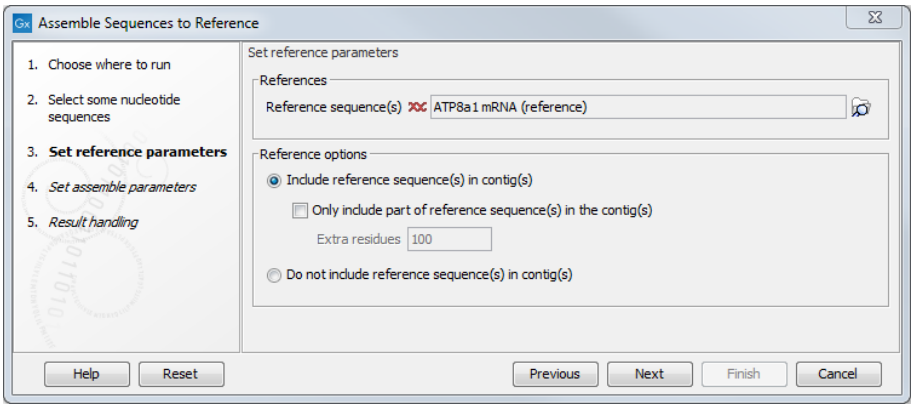

Figure 4: *The ATP8a1 mRNA (reference) sequence selected as reference sequence for the assembly.*

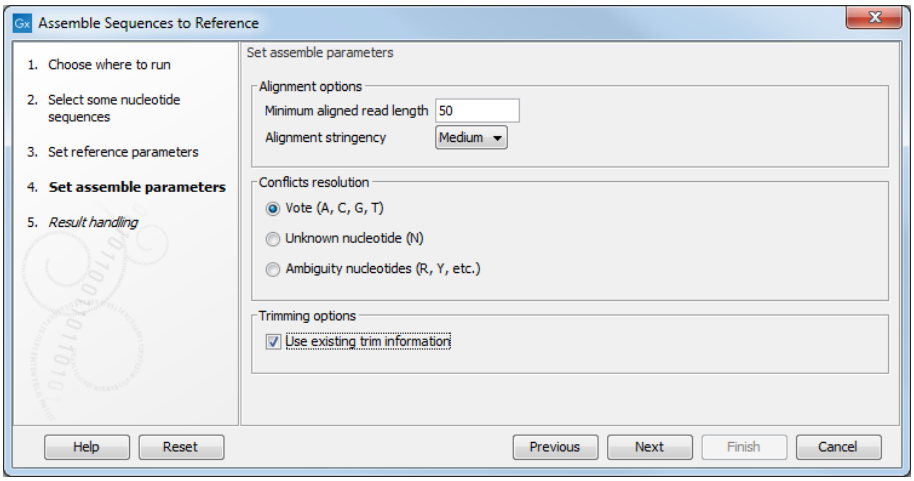

Figure 5: *Use the default settings and tick the box "Use existing trim information".*

5. Choose where you would like to save the results and click **Finish**. The assembly process will begin.

Getting an overview of the contig

The result of the assembly is a **Contig**, i.e., an alignment of the nine reads to the reference sequence. Open the file from the Navigation Area and click **Fit width** ( ) to see an overview of the contig.

Figure 6: *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. Set the compactness to Not compact in the Side Panel so you can see the Trace data of each read. 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; you can also use `','` and `'.'` keys to move back and forth between conflicts (see figure 7).

Figure 7: *Using the Find Conflict button highlights conflicts.*

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 "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" sequence by placing the cursor to the left of it and dragging the cursor across the T. Press **Delete**. If a warning pops up with the text **Edit Warning**, press **OK**. This will resolve the conflict.

Including regions that have been trimmed off

Clicking the **Find Conflict** button again until you find the conflict shown in figure 8. The conflict comes from the fact that Rev1 has an extra A in a region that does not look trustworthy. You can first slide the boundary of the Rev1 Trim region to include the A that creates the conflict.

Now, only one sequence contributes to the consensus sequence. 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. To include part of the trimmed region of Fwd2 in the contig, move the vertical slider to include the 7 additional nucleotides AGAAAGG (see figure 9).

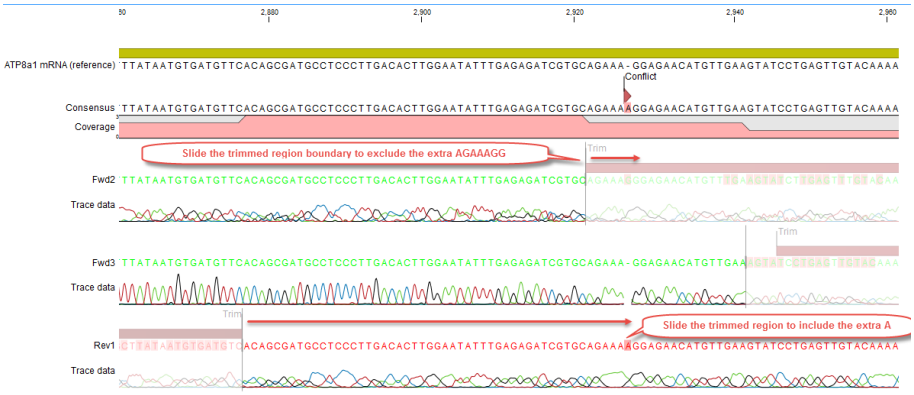

Figure 8: *Dragging the edge of trimmed regions.*

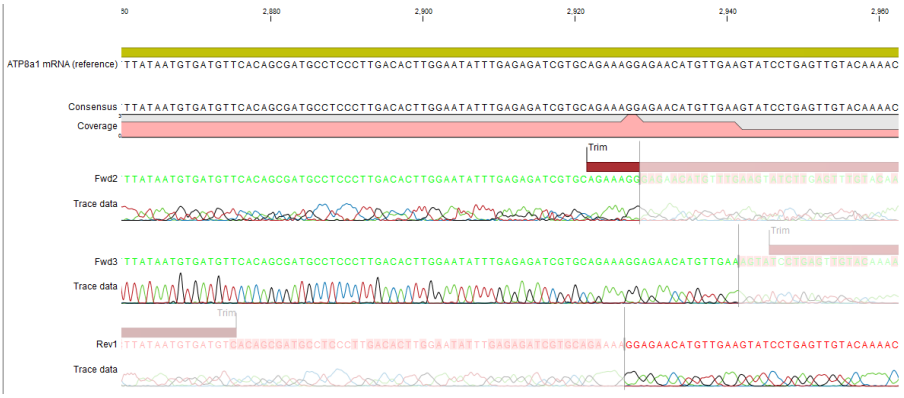

Figure 9: *Dragging the edge of the trimmed region.*

What we have demonstrated here is the relevance of keeping the trim regions on the sequences to resolve conflict and increase coverage. 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 until you find the conflict shown in figure 10, which is at position 2564.

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 using the Tool Bar zoom function ( ). Click on the selected base several times. Now you have zoomed in on the trace (see figure 10).

This gives more space between the residues, but if we would like to inspect the peaks even more, simply drag the peaks up and down with your mouse (see figure 11).

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 using the **Nucleotide info** in the Side Panel. In the option **Translation**, check **Show** and select **ORF/CDS** from the drop down menu (figure 12).

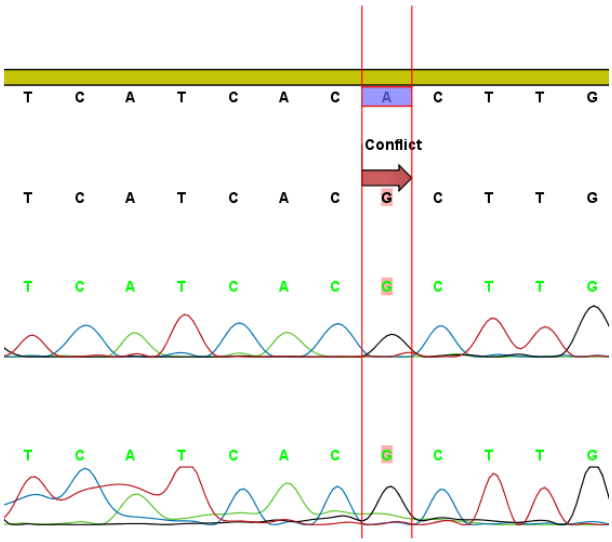

Figure 10: *Now you can see all the details of the traces.*

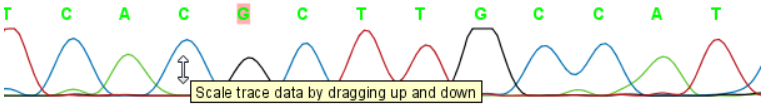

Figure 11: *Grab the traces to scale.*

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 orange. If it was a non-synonymous substitution, it would be colored in red.

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 (or ⌘ on Mac) and 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 for example 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 **History** ( ) icon at the bottom of the view.

In the history, you can see the details of each change (see figure 13).

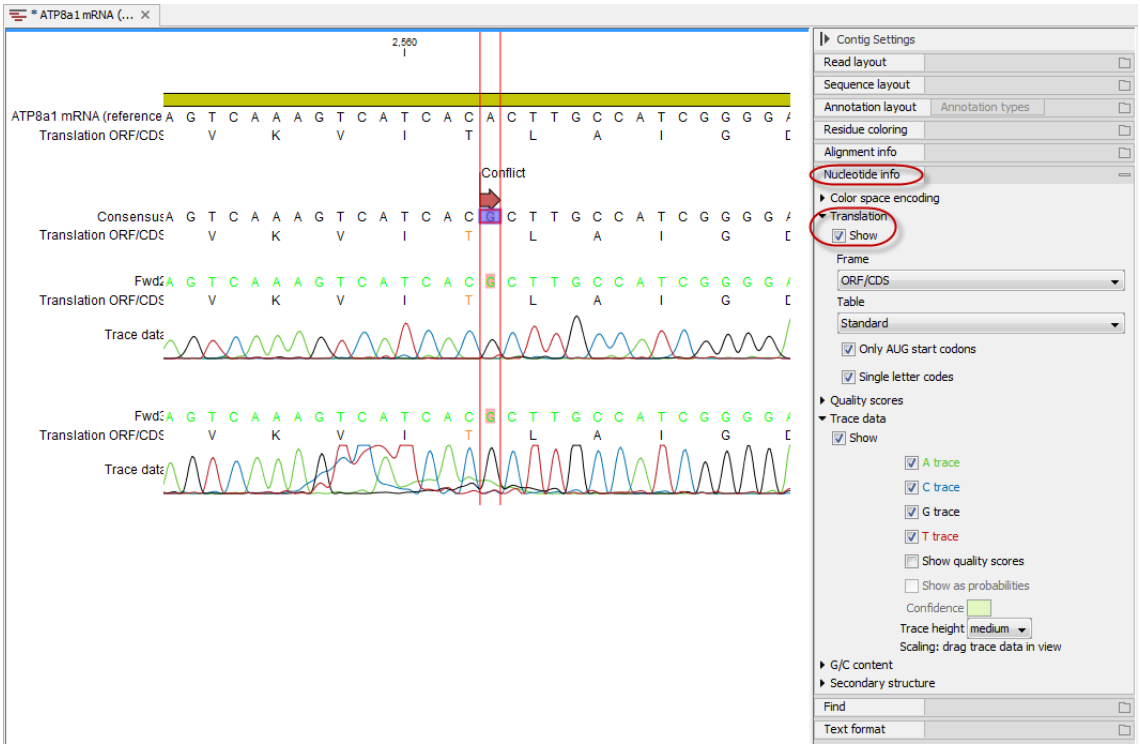

Figure 12: *Showing the translation along the contig.*

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: Right-click the name "Consensus" and choose to **Open Sequence**, which you can then save.

This will make it possible to use this sequence for further analyses in the 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 14).

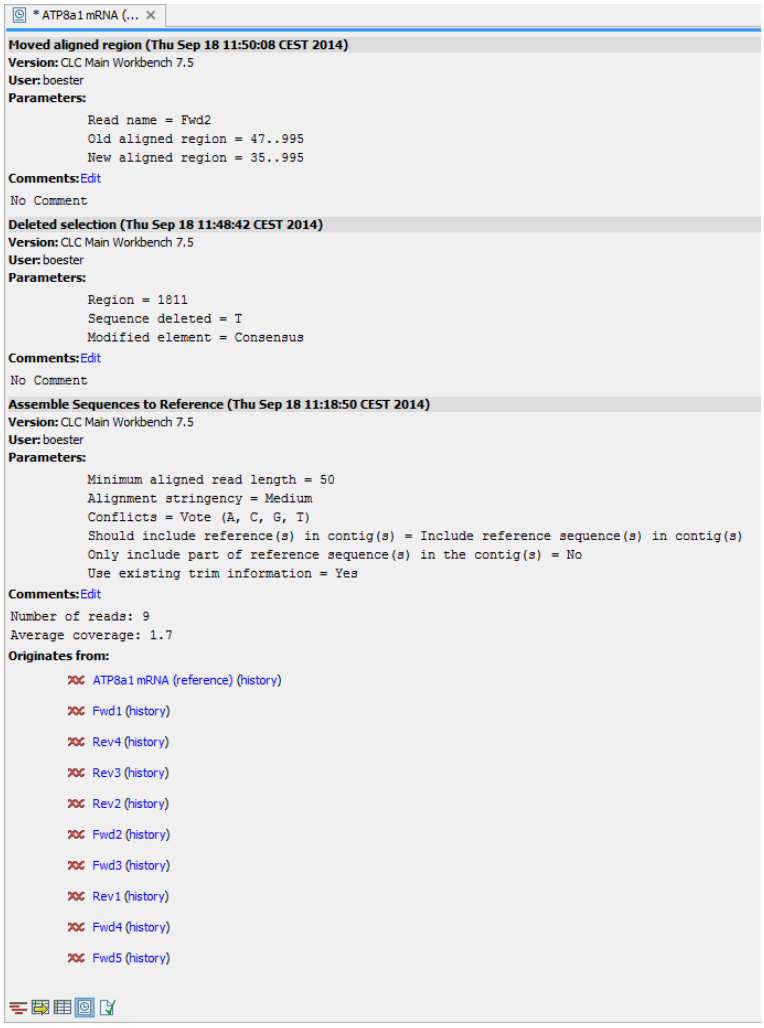

Figure 13: *The history of the contig showing that a "T" has been deleted and that the aligned region has been moved.*

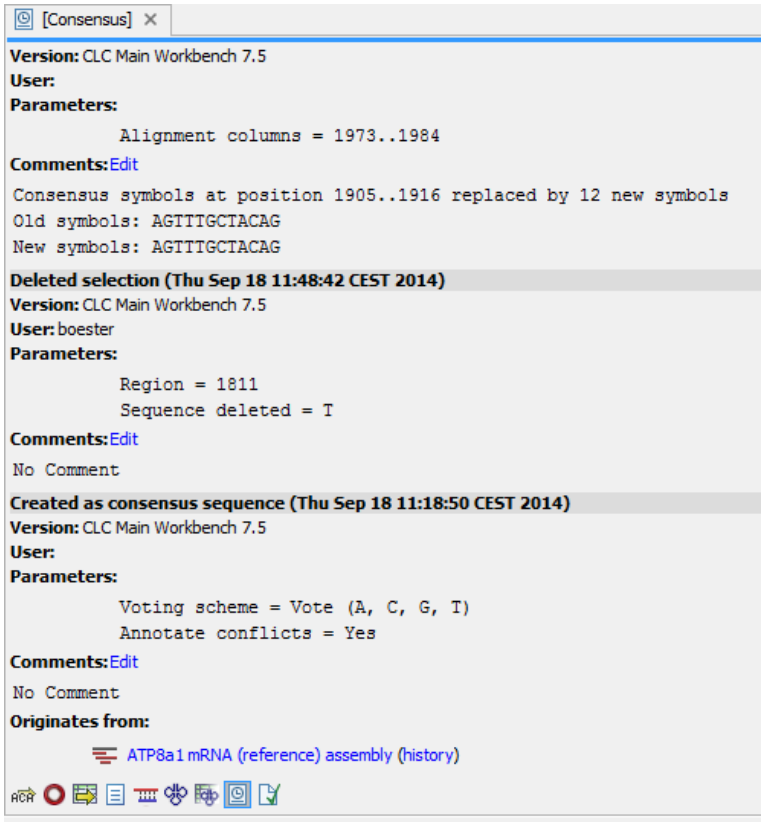

Figure 14: *The history of the consensus sequence, which has been extracted from the contig. Clicking the blue text "Reference contig" will find and highlight the name of the saved contig in the Navigation Area. Clicking the blue text "history" to the right will open the history view of the earlier contig. From there, you can choose other views, such as the Read mapping view, of the contig.*