Tutorial

Getting Started

November 21, 2017

Sample to Insight

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

Contents

Getting Started

This tutorial will take you through the most basic steps of working with the CLC Workbenches. All following chapters are independent of each other and can be done separately and in the order of your choice. For an interactive version of this tutorial, go to `http://www.clcbio.tv/video/854220/getting-started`

When you open *CLC Genomics Workbench* for the first time, the user interface looks like figure 1.

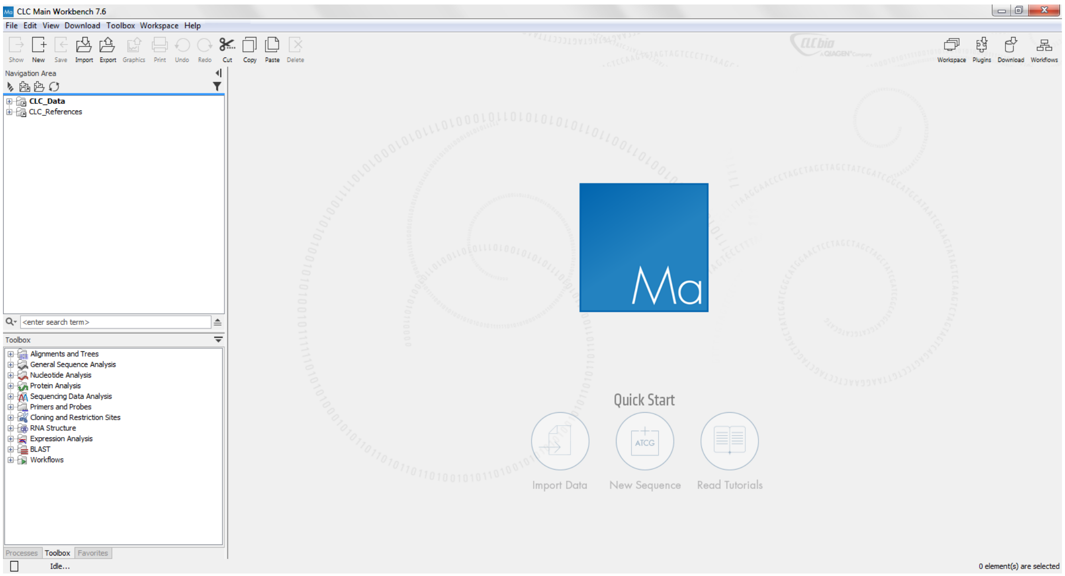


Figure 1: *The user interface as it looks when you start the program for the first time.*

At this stage, the important features are the **Navigation Area** and the **View Area**.

The **Navigation Area** to the left is where you keep all your data for use in the program. Most analyses of *CLC Genomics Workbench* require that the data is saved in the Navigation Area. There are several ways to get data into the Navigation Area, and this tutorial describes later how to import existing data.

The **View Area** is the main area to the right. This is where the data can be 'viewed'. In general, a **View** is a display of a piece of data, and the View Area can include several Views represented by tabs. The View Area can also be "split" to display two views simultaneously. It is possible to save a View by dragging and dropping the tab from the View Area to the Navigation Area.

Importing data in the workbench

This section shows how to create a folder and how to import data into the program.

Creating a a folder When *CLC Genomics Workbench* is started there is one element in the Navigation Area called **CLC_Data**: this element is a **Location**. A location points to a folder on your computer where your data for use with the workbench is stored.

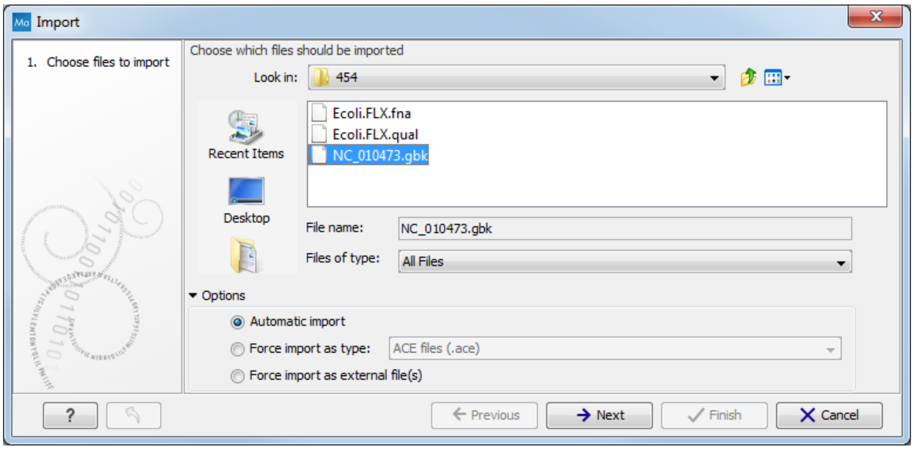
The data in the location can be organized into folders. Create a folder:

File | **New** | **Folder** ()
or **Ctrl + Shift + N (⌘ + Shift + N on Mac)**

Name the folder 'Example Data' and press Enter.

Import data As an example, first generation sequence data - as well as high throughput sequencing data - can be downloaded from `http://www.qiagenbioinformatics.com/support/example-data/` under "raw Data".

1. Download Roche/454 pyrosequencing genome data from E. coli commensal strain K-12. This file is chosen for demonstration purposes only - you may have another file on your desktop, which you can use to follow this tutorial. Unzip the file in the location of your choice (the desktop of your computer for example).
2. Click on the button Import in the top left corner of the workbench, or in the File menu. You can also use the quick command Ctrl + Shift + T (or ⌘ + Shift + T on Mac) to find the Import tool.
3. Choose Standard Import if you are given the option. In any case, the window shown in figure (figure 2) opens.



1. Choose files to import

Choose which files should be imported

Look in: 454

Ecoli.FLX.fna
Ecoli.FLX.qual
NC_010473.gbk

File name: NC_010473.gbk
Files of type: All Files

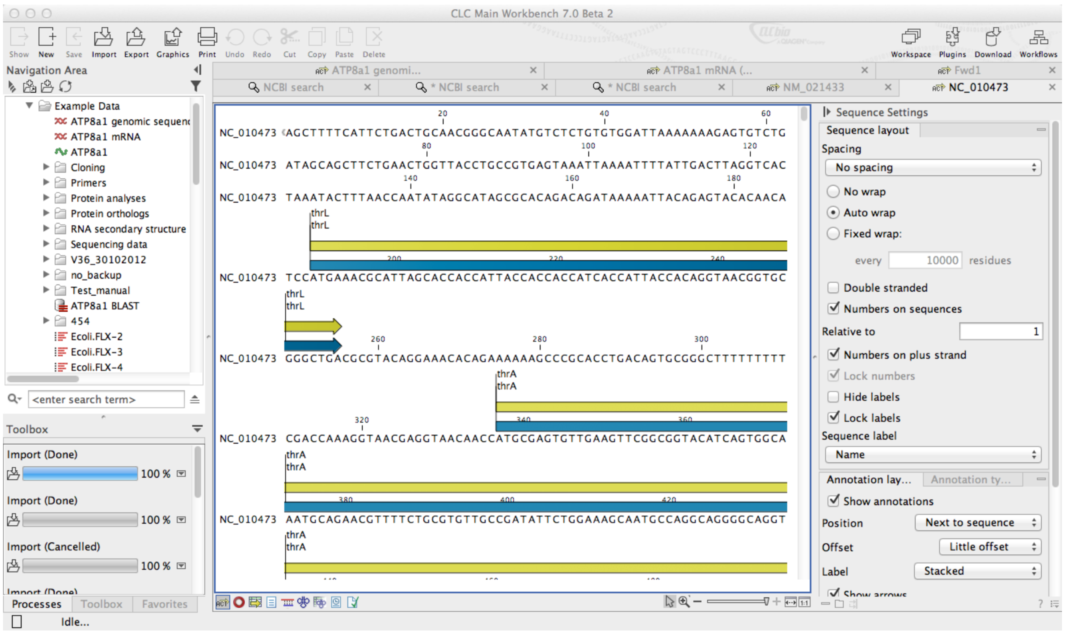
Options
● Automatic import
○ Force import as type: ACE files (.ace)
○ Force import as external file(s)

Figure 2: *Import a sequence in your Navigation Area.*

4. Choose the NC_010473 sequence you just saved (or any sequence of your choice), and click on **Next**.
5. Choose a location in the Navigation Area where you would like to save the data you are importing (in this case "Example Data"). Note that if you do not specify a location, the

data will be directly imported into the folder that was selected in the Navigation Area before you clicked **Import** - it is usually the folder containing the last file you opened and that is highlighted in the Navigation Area.

6. Click on Finish to proceed to the import.
7. The sequence is now saved in the Navigation Area. Double click on NC_010473 sequence: the result looks like figure 3.



Sequence layout
Spacing: No spacing
○ No wrap
● Auto wrap
○ Fixed wrap: every 10000 residues
☐ Double stranded
☑ Numbers on sequences
Relative to: 1
☑ Numbers on plus strand
☑ Lock numbers
☐ Hide labels
☑ Lock labels
Sequence label: Name
☑ Show annotations
Position: Next to sequence
Offset: Little offset
Label: Stacked

Figure 3: *The NC_010473.gbk (GenBank format) file is imported and opened.*

Import high throughput data The Genomics and Biomedical Genomics workbenches have special import options designed specifically for current NGS technologies. We will show here how to import Illumina reads. If you are working with CLC Main Workbench or CLC Sequence Viewer, you can skip this section and resume the tutorial at the next.

Note: When importing Illumina paired data, you are required to specify if the reads are Mate pair or Pared end. You can also specify reads distances: these are usually defined during the library preparation of your sequencing experiment. When in doubt about distances, you can enter default values: for paired-end the distances distances are between 1 and 1000 bp while mate-pair reads typically have longer distances between 1000-5000 bp (and sometimes up to 10000). Note that the tools usually used subsequently to process Illumina reads (such as "Map Reads to Reference" or "RNA-Seq Analysis") have an "Auto-detect paired distances" option that is enabled by default. As long as this option is used, mis-specifying the distances during import should bear no consequences.

1. Download Illumina genomic data from Pseudomonas aeruginosa. This file is chosen for demonstration purposes only - you may have another file on your desktop, which you can use to follow this tutorial. Unzip the file in the location of your choice (the desktop of your computer for example).
2. Click on the **Import** button and choose the option for **Illumina** (figure 4).

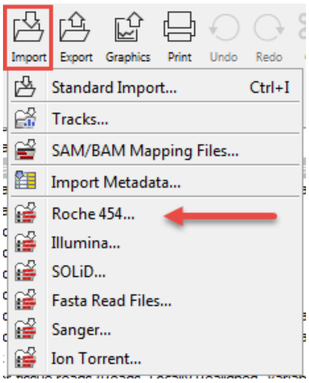
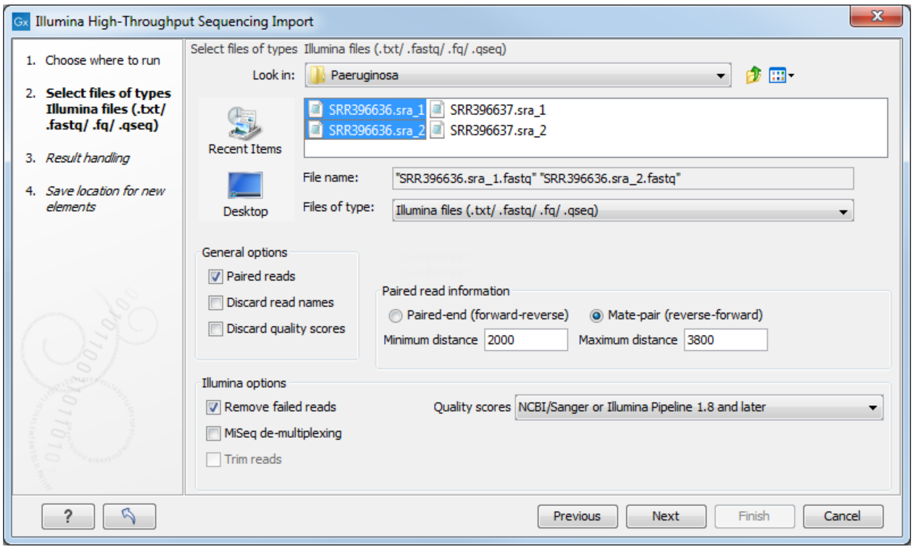


Figure 4: *Importing NGS data options.*

3. The first window of the Import wizard requires that you select two (or a multiple of two) files to import. The "Files of type" is set to **"Illumina files"** (figure 5) and look for the files you saved on your computer earlier. Check "Paired reads" in the General options box.
4. `SRR396636.sra_1.fastq` and `SRR396636.sra_2.fastq` are Mate-pair reads. A distance range 2000 to 3800 is appropriate for import. Click **Next**.



Look in: Paeruginosa

File name: "SRR396636.sra_1.fastq" "SRR396636.sra_2.fastq"

Files of type: Illumina files (.txt/ .fastq/ .fq/ .qseq)

☑ Paired reads

☐ Discard read names

☐ Discard quality scores

○ Paired-end (forward-reverse) ● Mate-pair (reverse-forward)

Minimum distance 2000 Maximum distance 3800

☑ Remove failed reads

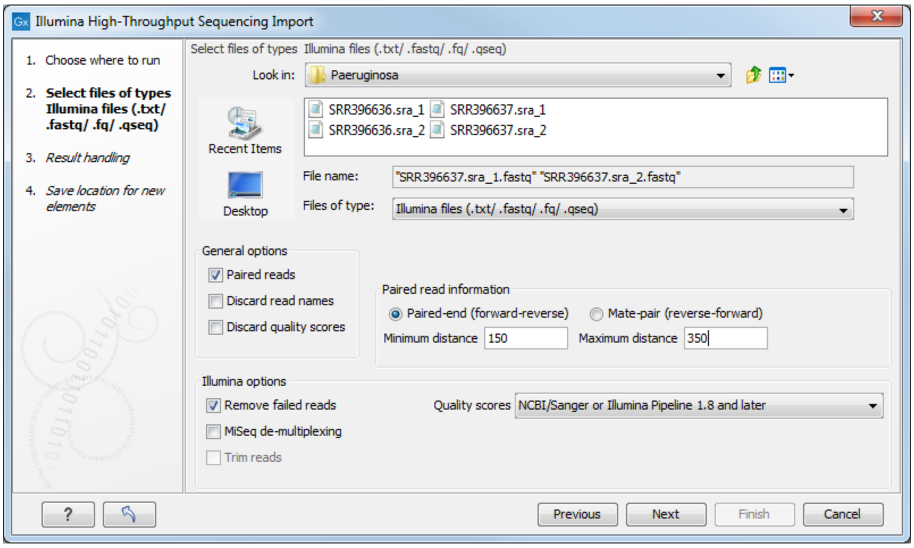
☐ MiSeq de-multiplexing

☐ Trim reads

Quality scores NCBI/Sanger or Illumina Pipeline 1.8 and later

Figure 5: *Select the 2 mate-pair files you downloaded and saved on your desktop.*

5. In the Result handling window, choose to Save the result of the Import and click **Next**.
6. You are prompted with a window that allows you to choose the location where you would like the reads to be saved. You can create a New folder for it. Click on **Finish**.
7. `SRR396637.sra_1.fastq` and `SRR396637.sra_2.fastq` are Paired-end reads. A distance range 150 to 350 is appropriate for import. Click **Next**.
8. In the Result handling window, choose to Save the result of the Import and click **Next**.
9. You are prompted with a window that allows you to choose the location where you would like the reads to be saved. You can create a New folder for it. Click on **Finish**.



Illumina High-Throughput Sequencing Import

1. Choose where to run
2. **Select files of types Illumina files (.txt/ .fastq/ .fq/ .qseq)**
3. *Result handling*
4. *Save location for new elements*

Select files of types Illumina files (.txt/ .fastq/ .fq/ .qseq)

Look in: Paeruginosa

SRR396636.sra_1 SRR396637.sra_1
SRR396636.sra_2 SRR396637.sra_2

File name: "SRR396637.sra_1.fastq" "SRR396637.sra_2.fastq"

Files of type: Illumina files (.txt/ .fastq/ .fq/ .qseq)

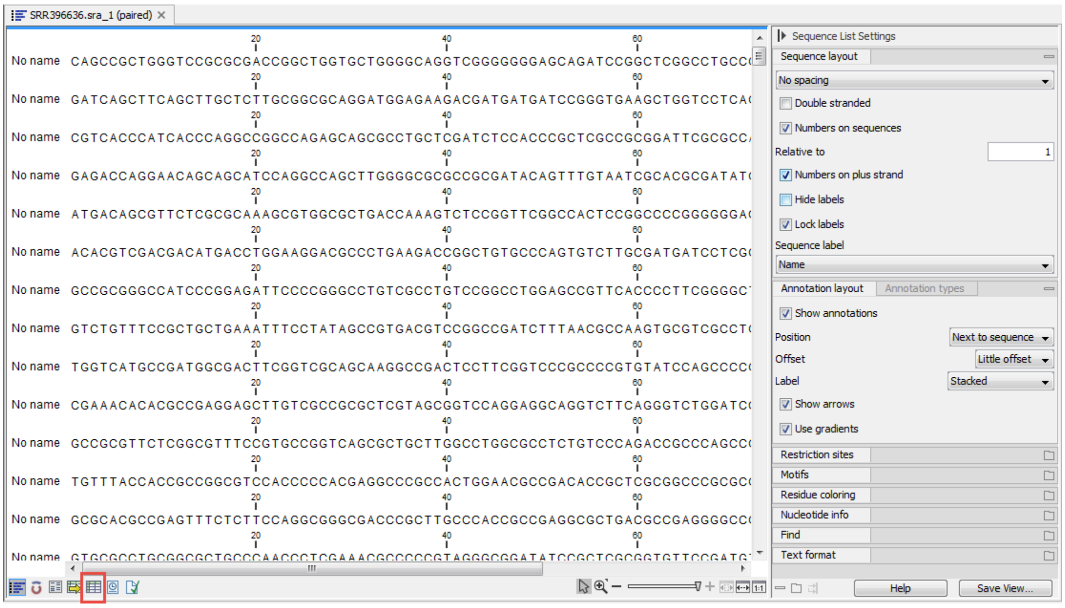
General options
☑ Paired reads
☐ Discard read names
☐ Discard quality scores

Paired read information
● Paired-end (forward-reverse) ○ Mate-pair (reverse-forward)
Minimum distance 150 Maximum distance 350

Illumina options
☑ Remove failed reads
☐ MiSeq de-multiplexing
☐ Trim reads
Quality scores NCBI/Sanger or Illumina Pipeline 1.8 and later

Figure 6: *Select the 2 paired-end files you downloaded and saved on your desktop.*

In this tutorial, I chose to save the high throughput data in the folder called **Paeruginosa**. Double click on one of the file name. The View looks like figure 7, but you can choose to see the data as a table by clicking on the Table icon (▦) at the bottom of the View Area.



SRR396636.sra_1 (paired)

```
No name  CAGCCGCTGGGTCCGCGCGACCGGCTGGTGCTGGGGCAGGTCGGGGGGGAGCAGATCCGGCTCGGCCTGCC
No name  GATCAGCTTCAGCTTGCTCTTGCGGCGCAGGATGGAGAAGACGATGATGATCCGGGTGAAGCTGGTCCTCA
No name  CGTCACCCATCACCCAGGCCGGCCAGAGCAGCGCCTGCTCGATCTCCACCCGCTCGCCGCGGATTCGCGCC
No name  GAGACCAGGAACAGCAGCATCCAGGCCAGCTTGGGGCGCGCCGCGATACAGTTTGTAATCGCACGCGATAT
No name  ATGACAGCGTTCTCGCGCAAAGCGTGGCGCTGACCAAAGTCTCCGGTTCGGCCACTCCGGCCCCGGGGGA
No name  ACACGTCGACGACATGACCTGGAAGGACGCCCTGAAGACCGGCTGTGCCCAGTGTCTTGCGATGATCCTCG
No name  GCCGCGGGCCATCCCGGAGATTCCCCGGGCCTGTCGCCTGTCCGGCCTGGAGCCGTTCACCCCTTCGGGGC
No name  GTCTGTTTCCGCTGCTGAAATTTCCTATAGCCGTGACGTCCGGCCGATCTTTAACGCCAAGTGCGTCGCCT
No name  TGGTCATGCCGATGGCGACTTCGGTCGCAGCAAGGCCGACTCCTTCGGTCCCGCCCCGTGTATCCAGCCCC
No name  CGAAACACACGCCGAGGAGCTTGTCGCCGCGCTCGTAGCGGTCCAGGAGGCAGGTCTTCAGGGTCTGGATC
No name  GCCGCGTTCTCGGCGTTTCCGTGCCGGTCAGCGCTGCTTGGCCTGGCGCCTCTGTCCCAGACCGCCCAGCC
No name  TGTTTACCACCGCCGGCGTCCACCCCCACGAGGCCCGCCACTGGAACGCCGACACCGCTCGCGGCCCGCGC
No name  GCGCACGCCGAGTTTCTCTTCCAGGCGGGCGACCCGCTTGCCCACCGCCGAGGCGCTGACGCCGAGGGGCC
No name  GTGCGCCTGCGGCGCTGCCCAACCCTCGAAACGCCCCCGTAGGGCGGATATCCGCTCGCGGTGTTCCGATG
```

Sequence List Settings

Sequence layout
No spacing
☐ Double stranded
☑ Numbers on sequences
Relative to 1
☑ Numbers on plus strand
☐ Hide labels
☑ Lock labels
Sequence label: Name

Annotation layout
☑ Show annotations
Position: Next to sequence
Offset: Little offset
Label: Stacked
☑ Show arrows
☑ Use gradients

Figure 7: *The Illumina data imported and opened.*

View a DNA sequence

This section will take you through some different ways to display a sequence in the program. The tutorial introduces zooming on a sequence, dragging tabs, and opening selection in new view. For an interactive version of this tutorial, go to `http://www.clcbio.tv/video/634026/opening-a-dna-sequence`.

We will be working with the sequence called *pcDNA3-atp8a1* located in the 'Cloning' folder in the Example data.

1. Download the Example Data folder by clicking on "Import Example Data" in the Help menu.
2. Double-click on the *pcDNA3-atp8a1* sequence in the Navigation Area to open it. The sequence is displayed with annotations above it (see figure 8).

```
pcDNA3-atp8a1 «GACGGATCGGGAGATCTCCCGATCCCCTATGGTCGACTCTCAGTACAATCTGCTC
pcDNA3-atp8a1  TGATGCCGCATAGTTAAGCCAGTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCT
pcDNA3-atp8a1  GAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACCGACAATT
pcDNA3-atp8a1  GCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGC
pcDNA3-atp8a1  CAGATATACGCGTTGACATTGATTATTGACTAGTTATTAATAGTAATCAATTACG
pcDNA3-atp8a1  GGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTTACGGTAA
pcDNA3-atp8a1  ATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGAC
```

Sequence Settings

Sequence layout

Spacing: No spacing

○ No wrap

● Auto wrap

○ Fixed wrap: every 10000 residues

☐ Double stranded

☑ Numbers on sequences

Relative to 1

☑ Numbers on plus strand

☑ Lock numbers

☐ Hide labels

☑ Lock labels

Sequence label: Name

Annotation layout | Annotation types

☑ Show annotations

Position: Next to sequence

Offset: Little offset

Figure 8: *Sequence pcDNA3-atp8a1 opened in a view.*

   As default, *CLC Genomics Workbench* displays a sequence with annotations (colored arrows on the sequence like the green CMV promoter region annotation) and zoomed in to see the residues.

3. In this tutorial we want to have an overview of the whole sequence. There are different ways to zoom out: for example, right click on the magnifier (🔍) in the Toolbar and select the Zoom out mode (🔍) (figure 9). Click on the sequence until you can see the whole sequence.

   You can also directly use the Zoom to a selection icon (⬚) (figure 10). You can switch back to a view where you can see the residues by clicking on the Zoom to base level icon (1:1).

4. This sequence is circular, which is indicated by << and >> at the beginning and the end of the sequence. We can see a circular view of the sequence by clicking on Show as Circular (○) at the bottom of the view. You can see both linear and circular view at once by holding the Ctrl button on the keyboard (⌘ on Mac) while clicking on the Show as Circular (○) button (figure 11).

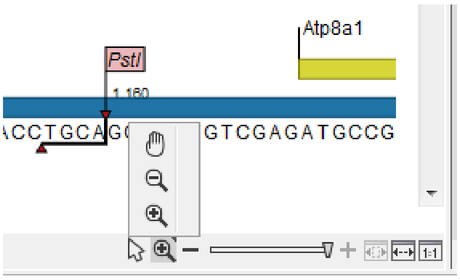


Figure 9: *Buttons displaying the different ways you can zoom on your sequence.*

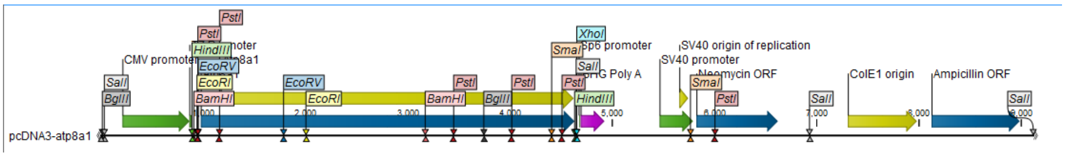


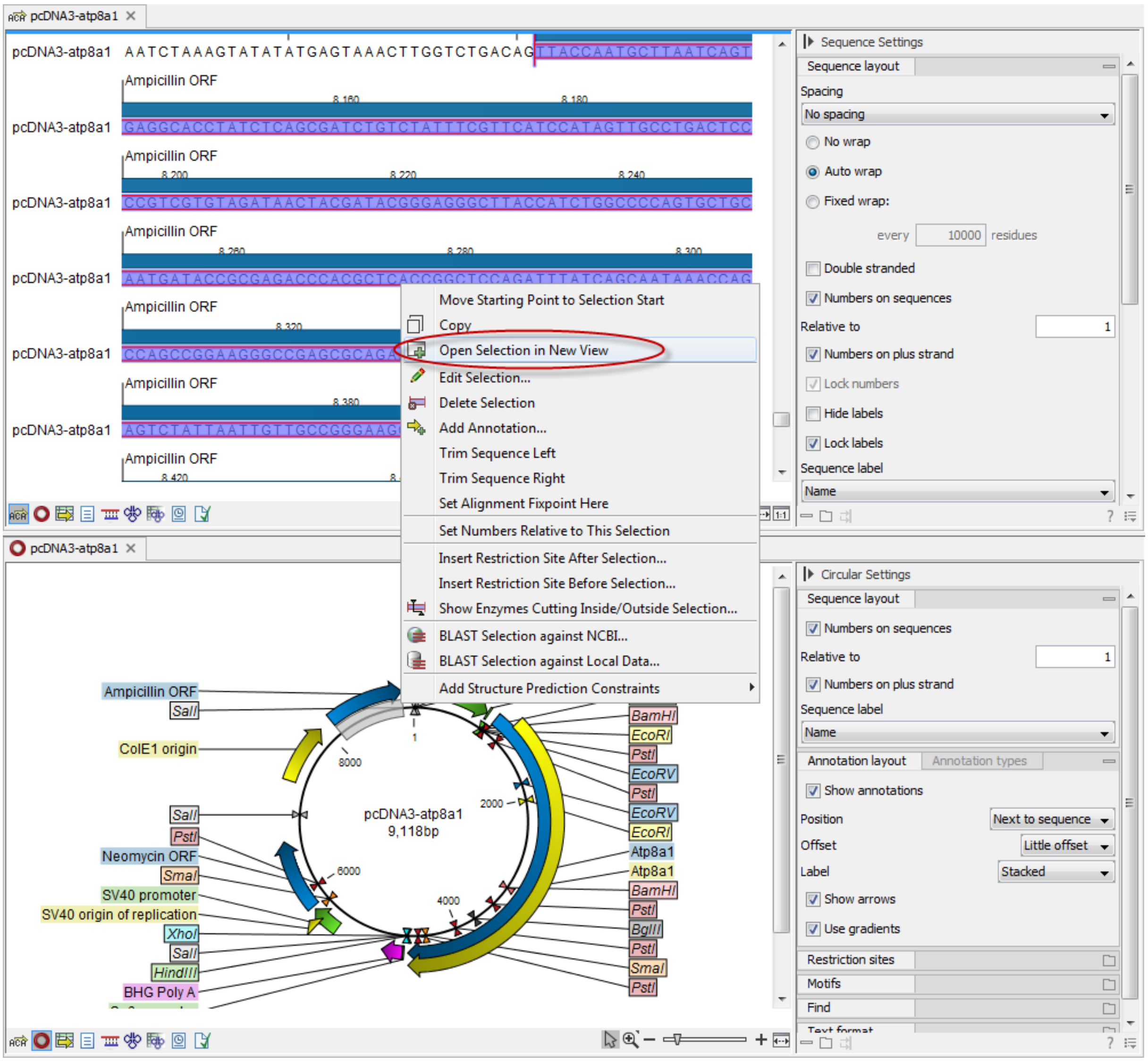
Figure 10: *Sequence pcDNA3-atp8a1 opened in a linear view and zoomed out to fit the screen.*

5. You may want to change the text size in the top panel to see more of the sequence. Scroll down in the **Sequence Settings** panel to **Text format** and change text size to 7. To learn more about what you can do in the side panel, see the "side Panel Settings" section of this tutorial.

6. Make a selection on the circular sequence (remember to switch to the **Selection** ( ) tool in the tool bar). Note that this selection is also reflected in the linear view above, and that the selection coordinates appear at the bottom right corner of the screen (in figure 11 the **Ampicillin ORF** was selected).

   You can zoom on the selected portion of the sequence by clicking on the Zoom to a selection button ( ), or you can open a third view of just the selected part of the sequence by right-clicking anywhere in the highlighted sequence text in the top panel and choosing **Open Selection in New View** as shown in figure 11.

7. You can rearrange the View Area by dragging tabs of the different Views between the top and bottom panels. You can also save the selection by dragging the tab to the "Example Data" folder in the Navigation Area.



Sequence Settings

Sequence layout

Spacing: No spacing

○ No wrap

● Auto wrap

○ Fixed wrap:

every 10000 residues

☐ Double stranded

☑ Numbers on sequences

Relative to 1

☑ Numbers on plus strand

☑ Lock numbers

☐ Hide labels

☑ Lock labels

Sequence label: Name

Circular Settings

Sequence layout

☑ Numbers on sequences

Relative to 1

☑ Numbers on plus strand

Sequence label: Name

Annotation layout | Annotation types

☑ Show annotations

Position: Next to sequence

Offset: Little offset

Label: Stacked

☑ Show arrows

☑ Use gradients

Restriction sites

Motifs

Find

Text format

Figure 11: *Horizontally split View Area with linear and circular views of the sequence. Both views are linked and selecting a region in one view will select the same region in the other. A new view panel for just the selected sequence can be created by right-clicking the highlighted sequence and choosing Open Selection in New View.*

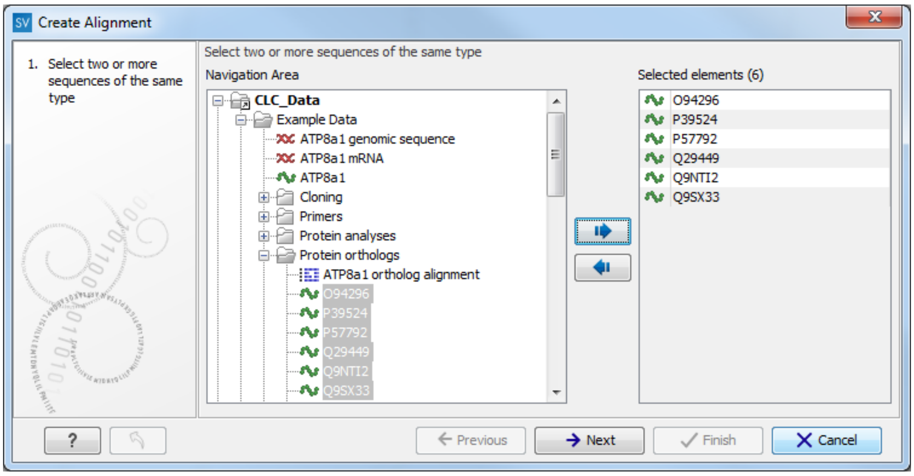
Running a Tool

In this section we will show you how to run a tool. As an example we will use the **Create Alignment** tool.

You can open a tool from the Toolbox by double clicking on its name or you can use the Launch button (🚀) and type in the name of the tool (in this case the **Create Alignment** tool).

A pop up window opens. Through a succession of windows you will enter the data you want to analyse, the parameters of the analysis you want to perform and how you want to handle the results of the analysis. You can navigate between windows by clicking the buttons Next and Previous at the bottom of the window. If you are not sure what to do, you can also click on the Help button and read the section of the manual relevant to the tool you are using. The **Reset** button will reset all parameters from the pop up window to their default values.

1. In the first window select the *Protein orthologs* folder in the **Example data** folder in the Navigation Area and choose the six sequences as shown in figure 12. The arrows can be used to add the sequences as well as remove the sequences from the **Selected Elements** list. Click on **Next**.



1. Select two or more sequences of the same type

Select two or more sequences of the same type

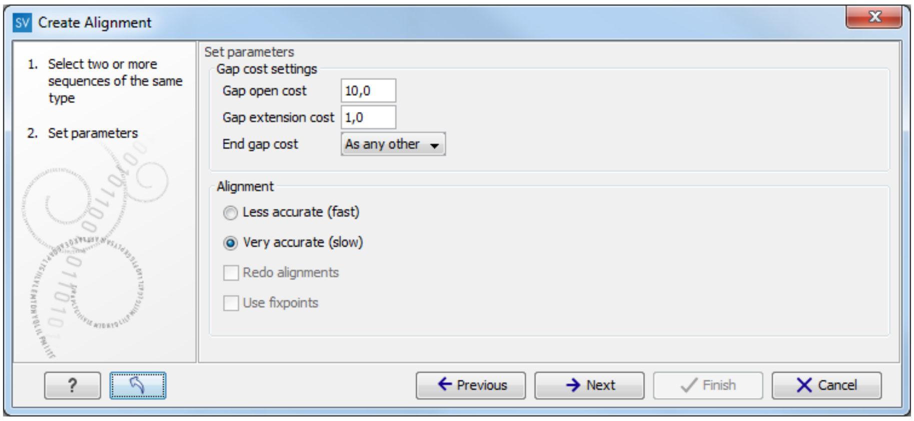
Figure 12: *Select the six sequences from the protein ortholog.*

2. The parameters should be set to default as shown in figure 13. To set all parameters as default click the **Reset** button. An explanation of the parameters can be found by clicking the Help button. Click on **Next**.

3. Specify where you would like to save the results. For example you can create a new folder outside the **Example data** folder.

4. Once you click on **Finish**, the tool starts. It is usually fast, but for heavier processes you can check on the progress of the tool in the "Processes" area, accessible by clicking on the "Processes" tab below the Toolbox.

5. When done, the status of the workbench will be set to "Idle..." in the lower left corner of the workbench. If you had chosen to "Open" the results, the alignment would open automatically in the View Area and you would have to save it manually to the Navigation Area. In this case we chose to Save results, so to check your alignment, double click on the name of the result file in the Navigation Area. If you do not know the name of the file



Create Alignment

1. Select two or more sequences of the same type
2. Set parameters

Set parameters

Gap cost settings

Gap open cost 10,0

Gap extension cost 1,0

End gap cost As any other

Alignment

○ Less accurate (fast)

● Very accurate (slow)

☐ Redo alignments

☐ Use fixpoints

Figure 13: *The parameters should be set to default.*

the tool just generated, or do not remember where you saved it, you can use the arrow next to the process that you want to view the results for (in this case the "Create Alignment" process as seen in figure 14) and choose the option "Show results" (the results will open in the View Area) or "Find results" (results will be highlighted in the Navigation Area).

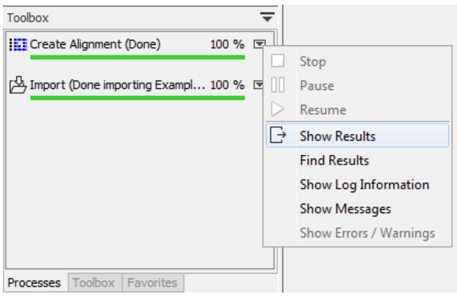


Figure 14: *The results can be viewed by clicking on the resently finished process and clicking Show Results.*

6. Once the alignment is open, it should look like displayed in figure 15.

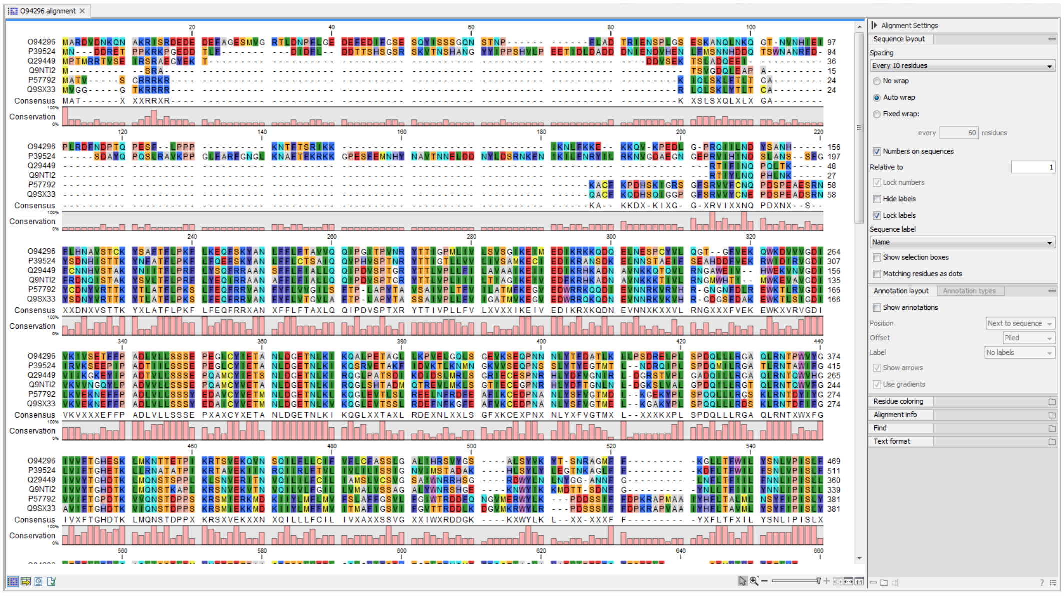
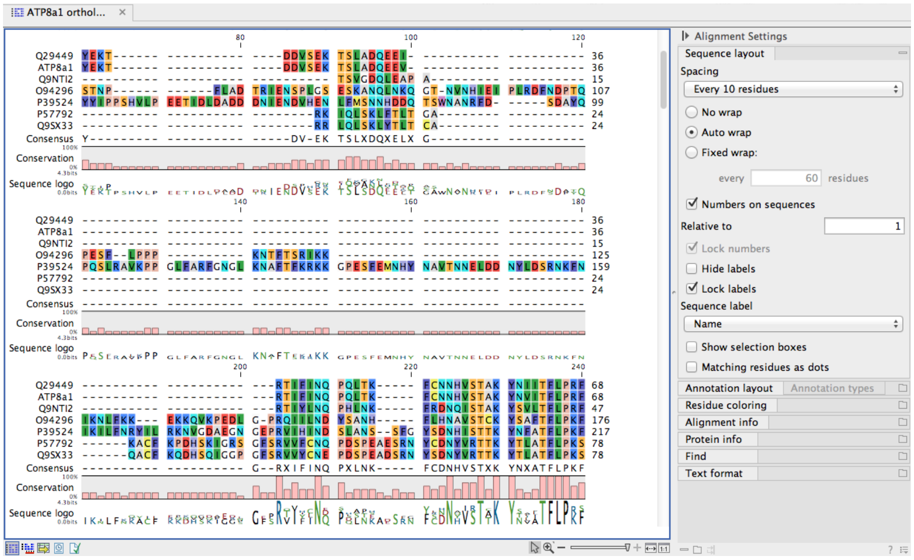


Figure 15: *The resulting alignment.*

Side Panel Settings

In this section we will show you how to use the **Side Panel** to change the way your sequences, alignments and other data are shown. You will also see how to save the changes that you made in the **Side Panel**. For an interactive version of this tutorial, go to http://www.clcbio.tv/video/632099/adjust-and-save-view-settings.

Open the protein alignment (▦) located under *Protein orthologs* in the **Example data** folder. The initial view of the alignment has colored the residues according to the Rasmol color scheme, and the alignment is automatically wrapped to fit the width of the view (figure 16).

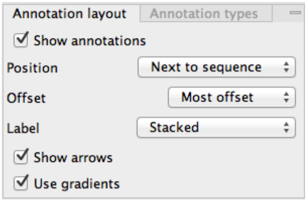


Sequence layout
Spacing: Every 10 residues
○ No wrap
● Auto wrap
○ Fixed wrap:
every 60 residues
☑ Numbers on sequences
Relative to 1
☑ Lock numbers
☐ Hide labels
☑ Lock labels
Sequence label: Name
☐ Show selection boxes
☐ Matching residues as dots

Figure 16: *The protein alignment as it looks when you open it with background color according to the Rasmol color scheme and automatically wrapped.*

Now, we are going to modify how this alignment is displayed. For this, we use the settings in the **Side Panel** to the right. All the settings are organized into groups, which can be expanded / collapsed by clicking the name of the group.

1. The first group is **Sequence Layout** which is expanded by default. Select **No wrap** in the **Sequence Layout**. This means that each sequence in the alignment is kept on the same line. To see more of the alignment, you now have to scroll horizontally.
2. Next, expand the **Annotation Layout** group and select **Show Annotations**. Set the **Offset** to "More offset" and set the **Label** to "Stacked".
3. Click on the **Annotation Types** tab. Here you will see a list of the types annotation that are carried by the sequences in the alignment (see figure 17).



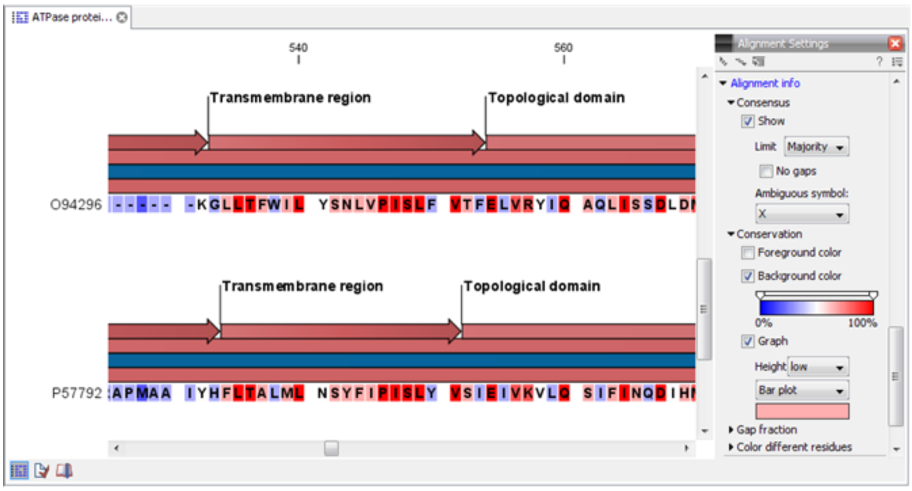
☑ Show annotations
Position: Next to sequence
Offset: Most offset
Label: Stacked
☑ Show arrows
☑ Use gradients

Figure 17: *The Annotation Layout and the Annotation Types tabs in the Side Panel.*

Check the "Region" annotation type, and you will see the regions as red annotations on the sequences.

4. Next, we will change the way the residues are colored. Click the **Alignment Info** group and under **Conservation**, check "Background color". This will use a gradient as background color for the residues. You can adjust the coloring by dragging the small arrows above the color box.

Now the alignment should look similar to figure 18.



Consensus

☑ Show

Limit Majority

☐ No gaps

Ambiguous symbol: X

Conservation

☐ Foreground color

☑ Background color

0% 100%

☑ Graph

Height low

Bar plot

Figure 18: *The alignment when all the above settings have been changed.*

Saving the settings in the Side Panel At this point, if you just close the view, the changes made to the **Side Panel** will not be saved. This means that you would have to perform the changes again next time you open the alignment. To save the changes to the **Side Panel**, click the **Save/Restore Settings** button (☰) at the bottom of the **Side Panel** and click **Save Alignment View Settings** (see figure 19). Choose the option **For Alignment Views in General**.

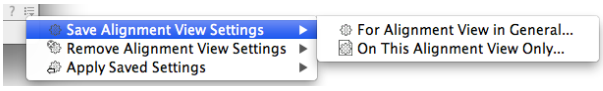


Figure 19: *Saving the settings of the Side Panel either generally or this particular alignment only.*

This will open the dialog shown in figure 20.

In this way you can save the current state of the settings in the **Side Panel** so that you can apply them to alignments later on. If you check **Always apply these settings**, these settings will be applied every time you open a view of the alignment.

Type "My settings" in the dialog and click **Save**.

Remove alignment view settings When you click the **Save/Restore Settings** button (☰) and select **Remove Alignments View Settings**, you can choose whether this should be applied generally or on this alignment view only (see figure 21).

This will open the dialog shown in figure 22 and allow you to remove specific settings.

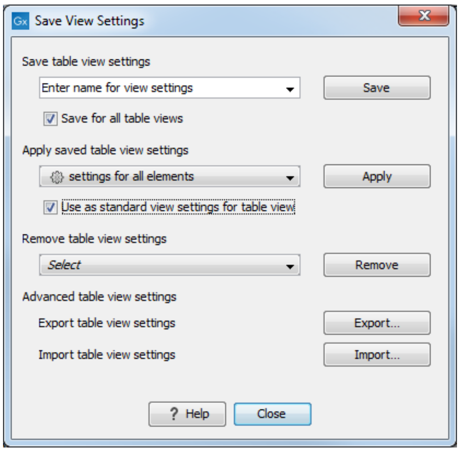


Figure 20: *Dialog for saving the settings of the Side Panel.*

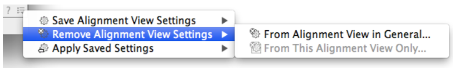


Figure 21: *Menu for removing saved settings.*

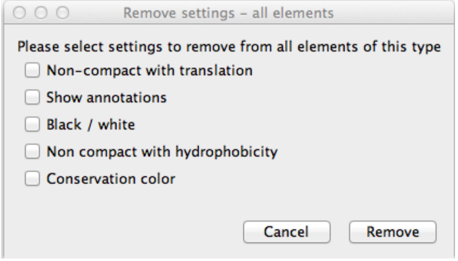


Figure 22: *Menu for removing saved settings.*

Applying saved settings When you click the **Save/Restore Settings** button (☰) again and select **Apply Saved Settings**, you will see "My settings" in the menu together with some pre-defined settings that the *CLC Genomics Workbench* has created for you (see figure 23).

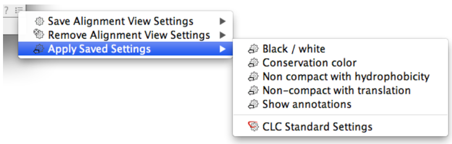


Figure 23: *Menu for applying saved settings.*

Whenever you open an alignment, you will be able to apply these settings. Each kind of view has its own list of settings that can be applied.

At the bottom of the list you will see the "CLC Standard Settings" which are the default settings for the view.