Tutorial

ChIP sequencing: the basics

December 15, 2009

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

ChIP sequencing: the basics

This tutorial takes you through a ChIP sequencing work flow using the *CLC Genomics Workbench*.

ChIP-Sequencing is used to analyze the interactions of proteins with genomic DNA. After a cross-binding step that tightly links proteins and DNA, ChIP-Seq uses chromatin immunoprecipitation (ChIP) to fish out the relevant pieces of genomic DNA. By subsequent massively parallel DNA sequencing and mapping to the reference genome it is possible to precisely identify binding sites of DNA-associated proteins. It can be used to precisely map global binding sites for any protein of interest but a practical limitation is the existence of good antibodies for the ChIP step. A natural next step bioinformatic analysis is to extract the binding regions and perform pattern discovery to learn about any conserved binding motif in the DNA. For further information, see the Wikipedia entry at http://en.wikipedia.org/wiki/Chip-Seq

For this tutorial, we use an artificial data set which was used for a ChIP-seq analysis competition announced at http://seqanswers.com/forums/showthread.php?t=1039, known as the "ChIP-seq challenge". We only use a subset of the original data set for this tutorial, since the purpose is to learn the basic principles.

The work flow consists of three parts: first, you import the data. Next, you perform a reference assembly of the reads. Finally, you use the ChIP sequencing tool to detect significant peaks in the contig.

In this tutorial we will not go through the details of the ChIP-seq analysis. The user manual already explains the details of the algorithm: Click the **Help** (?) button in the dialog (see below) to read this or go to http://www.clcbio.com/index.php?id=1330&manual=ChIP_sequencing.html.

Importing the data

First, download the data set from our web site: http://download.clcbio.com/testdata/raw_data/chip-seq.zip. Unzip the file somewhere on your computer (e.g. the Desktop).

Start the *CLC Genomics Workbench* and import the data:

File | **Import High-throughput Sequencing Data** | **Fasta/Helicos**

This will bring up the dialog shown in figure 1

Select the `ChIP-seq reads.fa` file and make sure the **Paired-end reads** checkbox is NOT checked. The option to discard read names is not significant in this context because of the relatively small amount of reads. Click **Next**, **Save** the imported reads list and click **Finish**.

After a short while, the 142,000 reads in the file have been imported. Next, import the reference genome sequence also included in the zip file:

File | **Import ()** | **Locate "NC_000021.gbk"** | **Select**

Note that this is a genbank file imported using the "normal" import tool. Next-generation sequencing data needs to be imported using the special tool in the toolbox because they often have a more complex structure (in this case you could actually have used the normal import, because it is a simple fasta file).

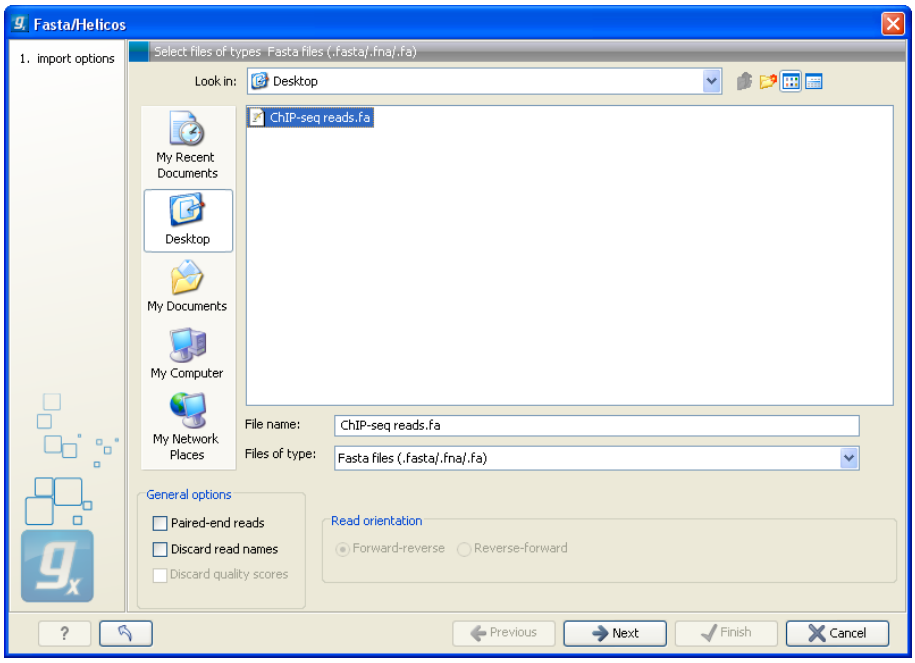


Figure 1: *When analyzing your own data, you would select the sequencing technology appropriate for your data. This data set consists of a fasta file, so you select Fasta/Helicos.*

Mapping the reads to the reference

First step in the analysis is to map the reads to the reference genome:

Toolbox | **High-throughput Sequencing** ( ) | **Reference Assembly** ( )

This shows the dialog in figure 2).

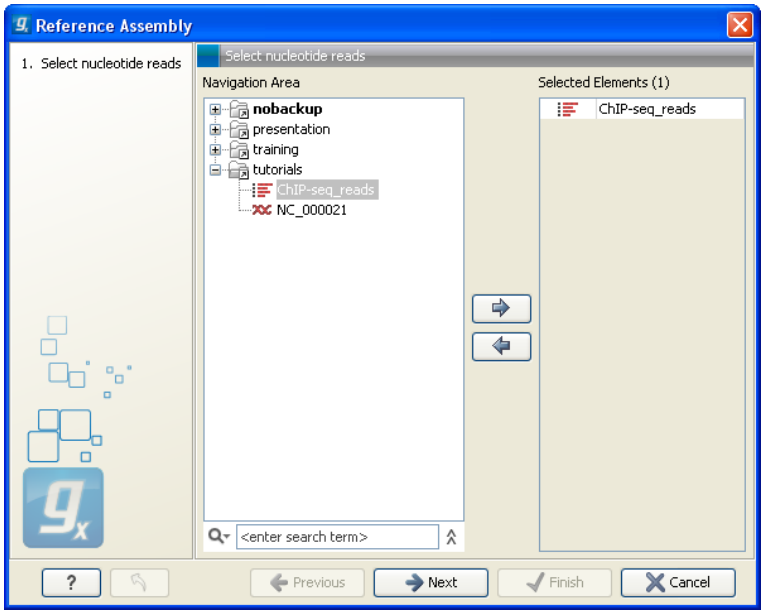


Figure 2: *Select sequence list containing the reads. The reference sequence will be selected in the next step.*

Select the `ChIP-seq reads` ( ) sequence list and add it to the panel to the right. Clicking

Next will allow you to select a reference sequence as shown in figure 3.

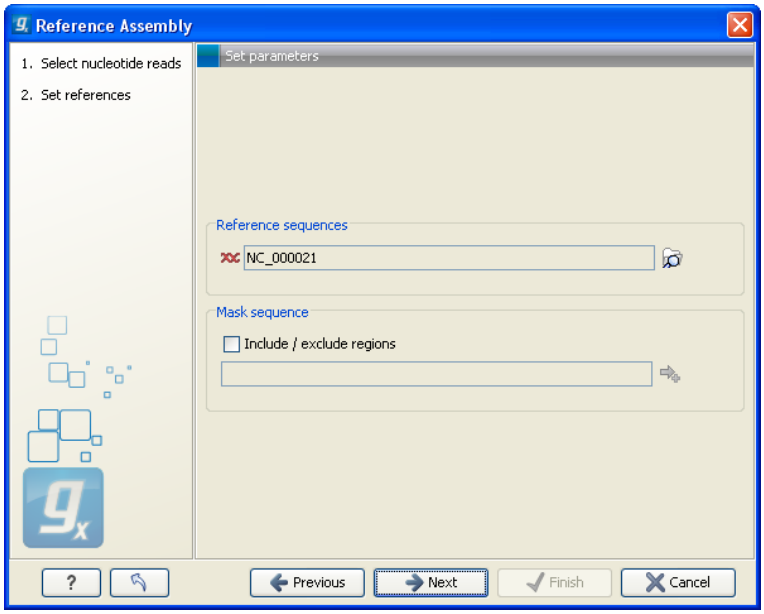


Figure 3: *Specifying the reference sequences and masking.*

At the top you select the `NC_000021` () by clicking the **Browse and select element** () button. You can select either single sequences or a list of sequences as reference sequences, but in this case just select this single chromosome.

Clicking **Next** to select the assembly settings as shown in figure 4.

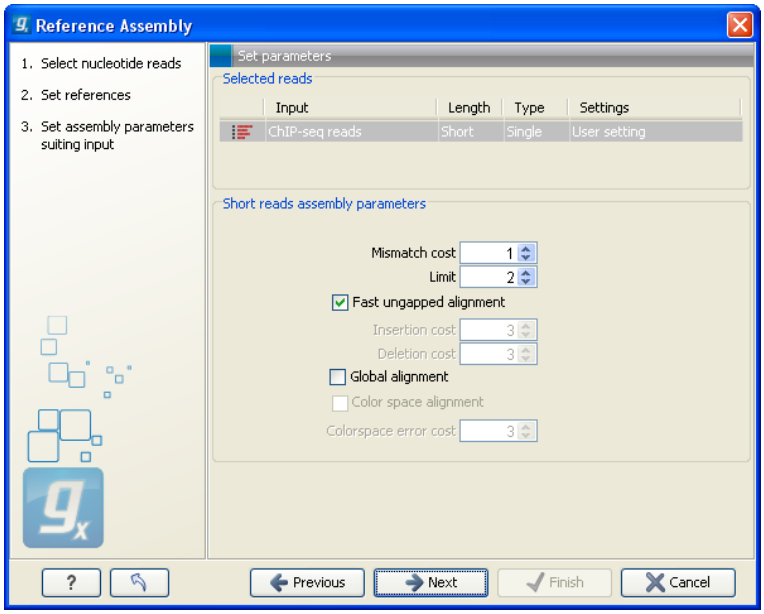


Figure 4: *A stringent read matching is desired for ChIP-seq.*

For ChIP-seq, we recommend stringent assembly settings as shown in figure 4. Setting the limit to 2 and mismatch cost to 1, you allow one mismatch per read or 2 un-aligned nucleotides at the ends. Since this data set is artificial, the settings are not important for the result of this tutorial, but when you work with your own data, this is important. For more information about the other

settings, please click the **Help** (?) button.

Click **Next** and you will see the dialog shown in figure 5.

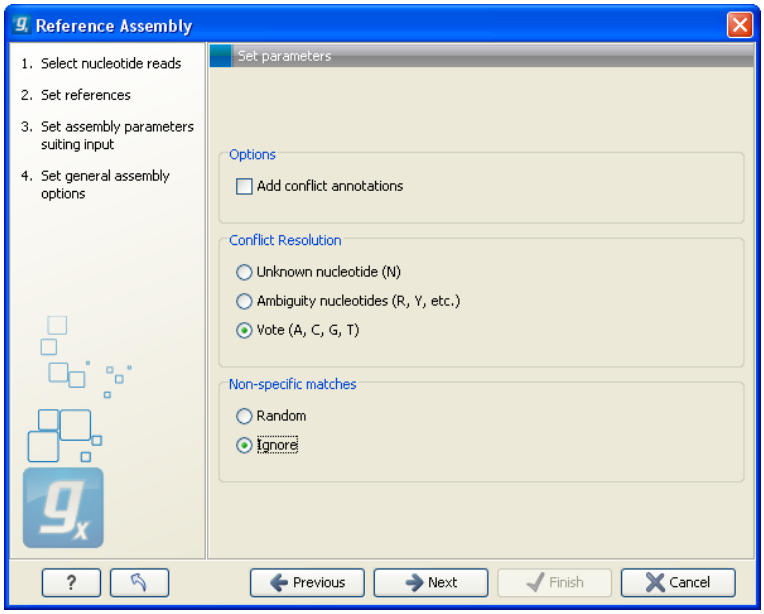


Figure 5: *Selecting to ignore non-specific matches.*

In this dialog, select to ignore the non-specific matches. Click **Next**, deselect the **Create overview table including all contigs** and click **Finish**.

You can follow the progress of the assembly both in the status bar at the bottom left corner and under the **Processes** tab. There is also a log showing the progress. Because of the quite big reference sequence (Human chromosome 21, with a size of 47 Mbp), it takes a while to assemble[1].

The result of the assembly is a contig that you now **Save** ( ) into a folder in the **Navigation Area**.

Running the ChIP sequencing analysis

The contig produced from the reference assembly is now used as input to the ChIP-seq function which surveys the pattern in coverage and read orientation to detect significant peaks:

Toolbox | **High-throughput Sequencing ( )** | **ChIP Sequencing ( )**

This opens a dialog where you select the `NC_000021 contig` ( ) and click **Next**.

Make sure that all parameters are set to default and click **Next**. You can set the parameters to default by clicking the button ( ) at the bottom of the dialog. Repeat this procedure until you can click **Finish**.

Remember that you can get details of the ChIP-seq analysis from the user manual: Click the **Help** (?) button to read this or go to `http://www.clcbio.com/index.php?id=1330&manual=ChIP_sequencing.html`.

[1] If you think it takes too long on your computer to finish, there is a smaller subset available at `http://download.clcbio.com/testdata/raw_data/chip-seq-subset.zip`

As a result of the analysis, annotations are added to the contig's reference sequence where significant peaks are detected, and a table is displayed below the contig. Click the first row in the table in order to jump to the corresponding position on the contig. At this point you may want to adjust a few view settings to get a better overview: In the **Contig Settings Side Panel**, under **Sequence layout**, set **Compactness** to Compact. Next, click **Zoom out** ( ) in the Toolbar and click 6-8 times in the contig view.

Your screen should now look like figure 6.

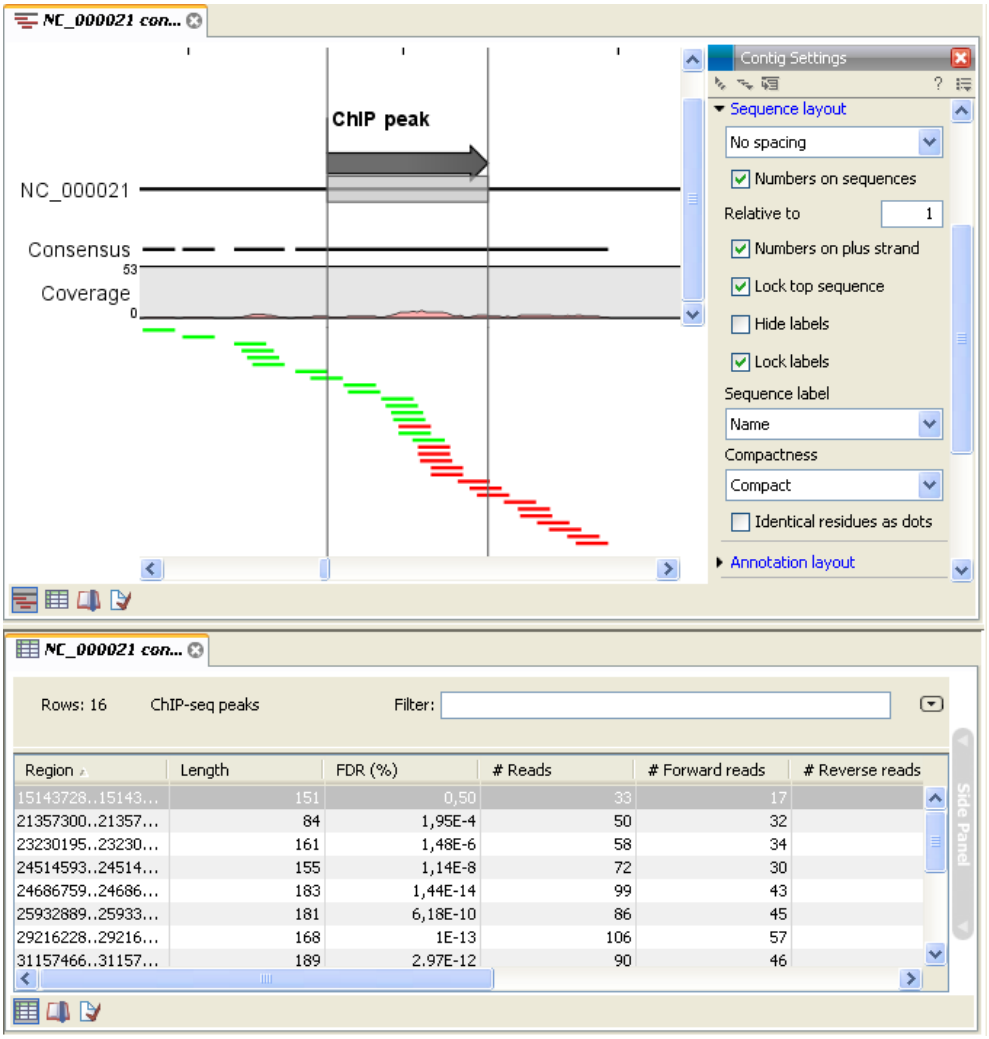


Figure 6: *A split view showing the peak table and the contig.*

Note the nicely distributed green (forward) and red (reverse) reads for this peak. You can browse through all the 16 peaks found for this contig by selecting in the table.

Since this is an artificial data set, we actually know what the results should be. If you go to the Side Panel in the contig and click **Annotation types**, you will find an annotation type called "Misc. binding". If you click that, the spiked-in peaks will be revealed on the reference sequence. You will see that all the peaks found by the *CLC Genomics Workbench* are covered by spike annotations (you can also see that in the table).

There are a few spiked in peaks that are not found since these only contain a very small number of reads - you can browse through them by clicking the small arrow ( ) next to the checkbox.