CLC Academy

ChIP-Seq - *CLC Genomics Workbench*

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

CLC bio
a result of science

CLC Academy

Contents

1 ChIP sequencing: the basics **3**

1.1 Importing the data . . . 3

1.2 Mapping the reads to the reference . . . 4

1.3 Running the ChIP sequencing analysis . . . 6

2 ChIP sequencing: Understanding the details **9**

2.1 Data set . . . 9

2.2 Getting the right layout . . . 9

2.3 Looking for known genes . . . 10

2.4 Going into detail with the parameters . . . 10

2.5 Extracting the peak regions . . . 12

CLC Academy

Chapter 1

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 map the reads to a reference. Finally, you use the ChIP sequencing tool to detect significant peaks in the sample.

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.

1.1 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**

This will bring up the dialog shown in figure 1.1

Select the `ChIP-seq reads.fa` file and make sure the **Paired reads** checkbox is NOT checked.

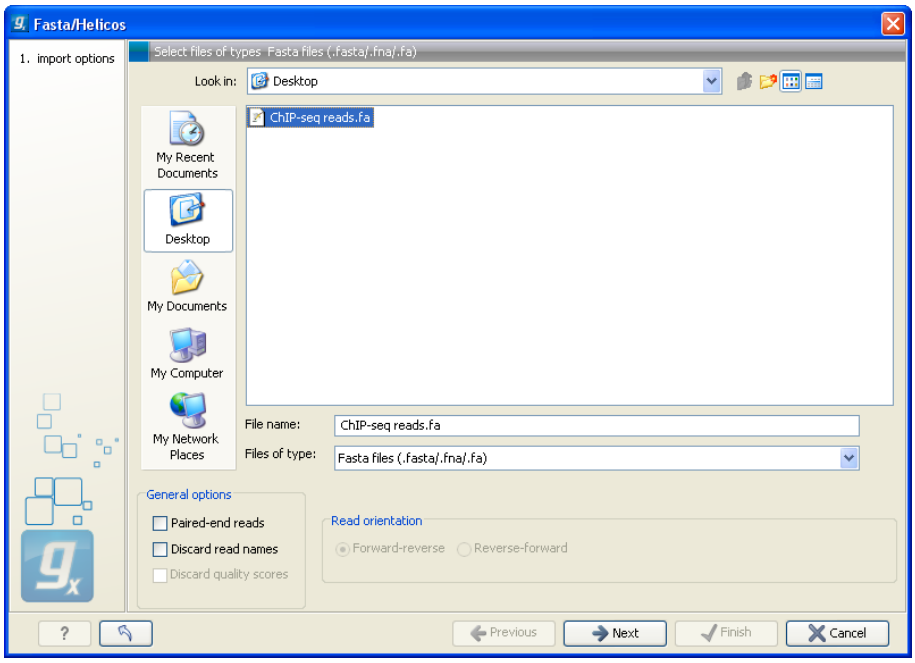


Figure 1.1: *When analyzing your own data, you would select the sequencing technology appropriate for your data. This data set consists of a fasta file, so you select Fasta.*

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).

1.2 Mapping the reads to the reference

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

Toolbox | **High-throughput Sequencing** ( ) | **Map Reads to Reference** ( )

This shows the dialog in figure 1.2).

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 1.3.

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 1.4.

For ChIP-seq, we recommend stringent mapping settings as shown in figure 1.4. Setting the limit

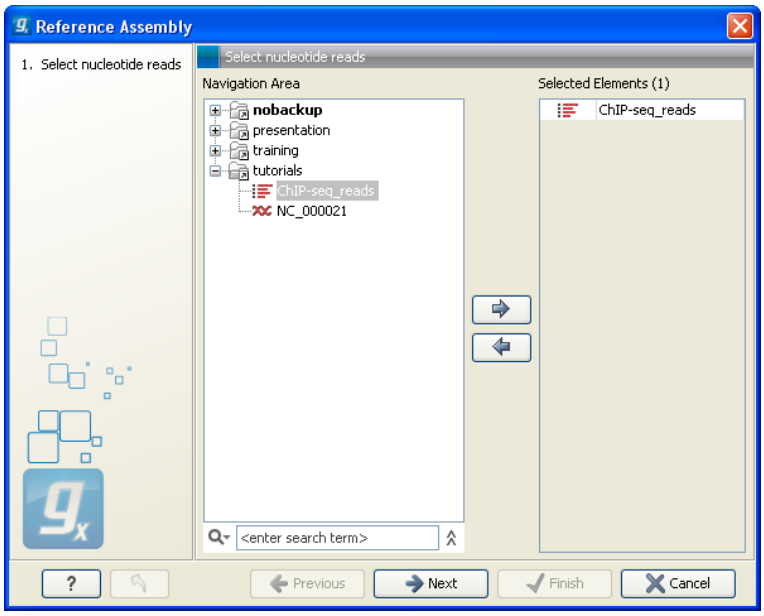


Figure 1.2: *Select sequence list containing the reads. The reference sequence will be selected in the next step.*

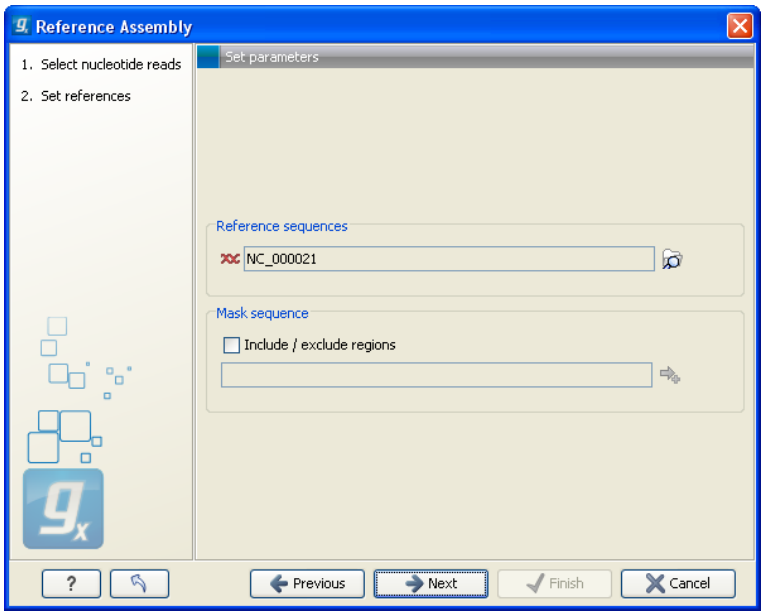


Figure 1.3: *Specifying the reference sequences and masking.*

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 1.5.

In this dialog, select to ignore the non-specific matches. Click **Next** and **Finish**.

You can follow the progress of the mapping 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

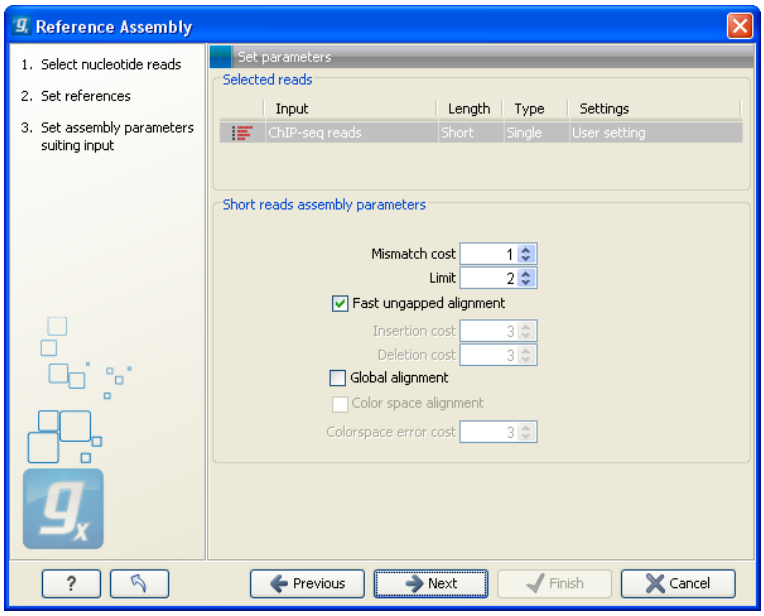


Figure 1.4: *A stringent read matching is desired for ChIP-seq.*

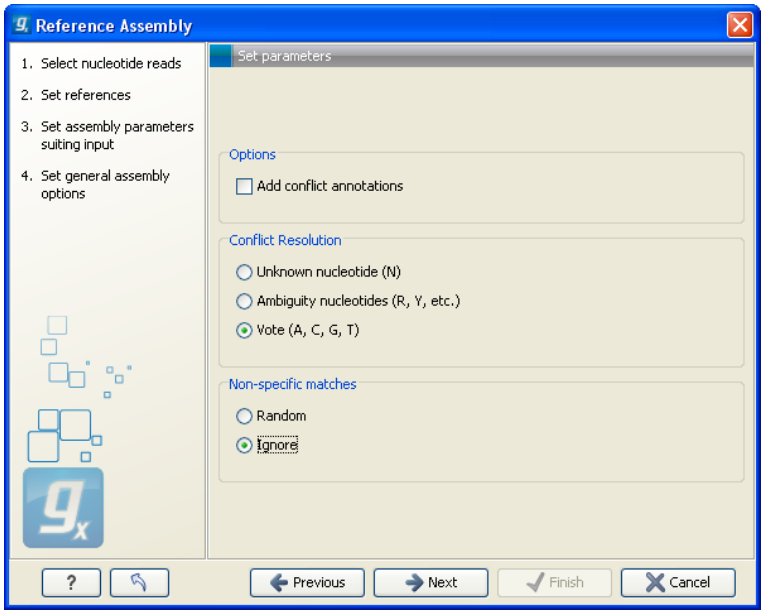


Figure 1.5: *Selecting to ignore non-specific matches.*

reference sequence (Human chromosome 21, with a size of 47 Mbp), it takes a while to map[1].

Save ( ) the result into a folder in the **Navigation Area**.

1.3 Running the ChIP sequencing analysis

The result of the read mapping 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-Seq Analysis** ( )

[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

This opens a dialog where you select the `ChIP-Seq reads mapping` ( ) 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 read the advanced tutorial.

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

Your screen should now look like figure 1.6.

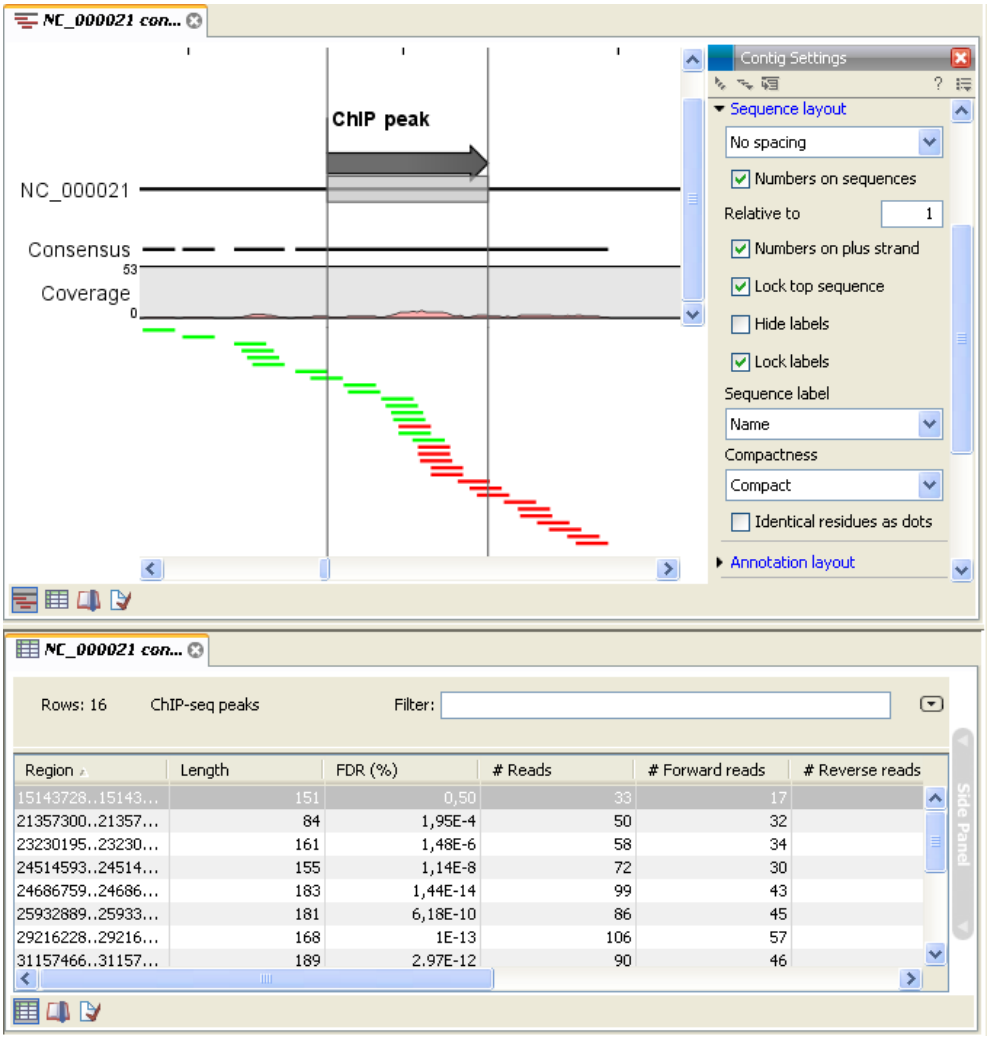


Figure 1.6: *A split view showing the peak table and the mapping.*

Note the nicely distributed green (forward) and red (reverse) reads for this peak. You can browse through all the 16 peaks found for this sample 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** 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.

CLC Academy

Chapter 2

ChIP sequencing: Understanding the details

This tutorial takes you through some of the details of the ChIP-seq analysis in the *CLC Genomics Workbench*. We recommend having a look at the basic ChIP-seq tutorial first (see http://www.clcbio.com/tutorials), since we will skip part of the work flow here.

2.1 Data set

The data set used derives from a study reported in **?**. In order to make the data set comprehendible and make the computing time and requirements low, the original full data set has been reduced.

First of all, only one of the 18 samples have been used. It is the sample of *PPAR*γ on day 6. This sample has been mapped against the mouse refseq genome, and two regions of chromosome 7 have been taken out for use in this tutorial. The reference sequence used is 10 Mbp, and there are 23,600 reads of 32 bp each.

The import and mapping was covered in the basic tutorial, so we proceed directly to the ChIP-seq analysis.

2.2 Getting the right layout

First, **Import** () the `chip-seq_pparg-subset.zip` file that you can download from http://download.clcbio.com/testdata/raw_data/chip-seq_pparg-subset.zip. Now, open the **ChIP-sequencing analysis** () dialog from the **Toolbox**, select the mapping result and click **Next**.

Make sure that all parameters are set to default. You can set the parameters to default by clicking the button () at the bottom of the dialog. Uncheck the **Shift reads based on fragment length** setting at this step.

Click **Next** and set parameters to default, click **Next** and set parameters to default, deselect the **Make log** checkbox, and click **Finish**. We will come back to learn more about the parameters later on. With these settings, you should be able to detect 12 peaks.

When the result is opened, you need to do a few customizations to make it better suited for

interpretation. In the **Side Panel**, under **Sequence layout**, set **Compactness** to Compact. Below, under **Text format**, set the font size to small or tiny.

Next, click **Zoom out** ( ) in the Toolbar and click 6-8 times in the view.

Your screen should now look like figure 2.1.

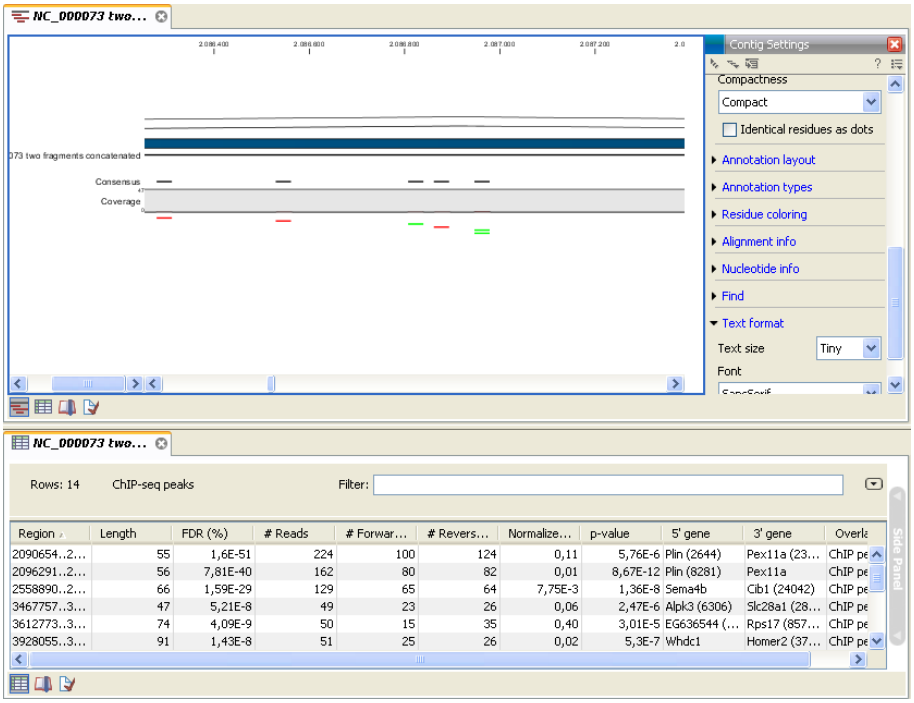


Figure 2.1: *The view has now been set up.*

2.3 Looking for known genes

The paper comments on the gene *perilipin (Plin)*, so we will now take a look at the binding sites surrounding that gene. In the peak table, simply type `Plin` into the **Filter** textbox, and you will now see one row in the table representing one peak located in an intron of *Pex11a* which sits next to *Plin*. Click the row to have the mapping view jump to see the peak.

You can now inspect the peak and if you zoom out you will see the adjacent *Plin* and *Pex11a* genes. Note that there is also two peaks between the two genes that is not found by the ChIP-sequencing tool. We will get back to that later.

Another gene commented in the paper is *Pnpla2* gene. Type this into the filter, and you will see two rows. The filter is very simple - it searches all the information in the table for the text you have input. Since the *Pnpla2* gene is represented in the columns for nearest gene 5' for one peak and as the nearest 3' gene in another peak, these two records are shown.

2.4 Going into detail with the parameters

Next, we will go through some of the settings for the ChIP-seq analysis. First, click the tab of the mapping in the upper view, and click **Undo** ( ). This removes the peak annotations of the mapping.

Run the analysis again, but this time make sure to uncheck all check-boxes on step 3 except the **Boundary refinement**. All other options remain the same.

The resulting table should report 79 peaks.

Now sort the table on the **Normalized difference** column. The one with the highest difference has 2 forward reads and 13 reverse reads. Usually, a peak like that would not be trusted because you would expect roughly the same number of forward and reverse reads. When you run the ChIP-sequencing tool, you can enter a maximum limit for this value. The default is 0.4. You can see in the table that applying this setting would have excluded 14 peaks.

Next, sort the table on p-value. If you click the peak with a p-value of 0.6, you will see that the distribution of the reads is different from the other peaks we have seen (see figure 2.2).

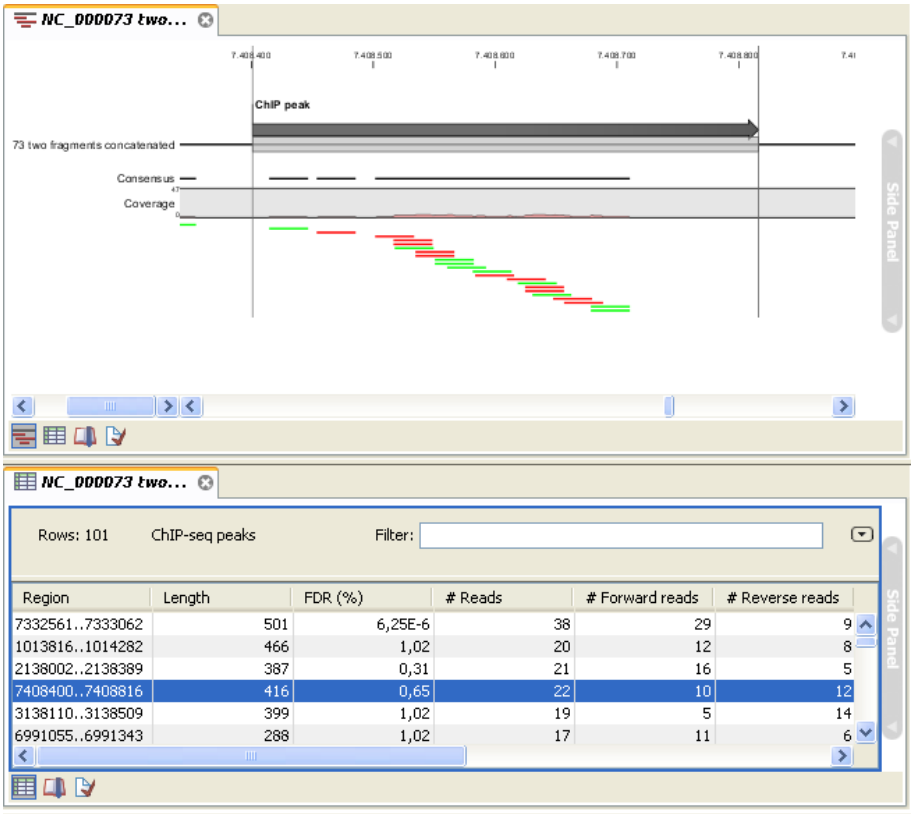


Figure 2.2: *A peak with a random distribution of forward and reverse reads leading to a high p-value.*

When the fragments from the ChIP are sequenced, they are sequenced from the ends which means that you would expect forward reads upstream of the binding site and reverse reads downstream. This peak has a very random distribution of the reads and that is the reason behind the high p-value.

When we did the first round of analysis, we looked at the *Plin* gene and saw a peak in coverage that was not detected. Now that we have relaxed the parameters, we can go back and inspect the peaks again. Type in `Plin` in the filter text box. You can see that the peak just next to the start site of *Plin* is also detected. It has a p-value of 0.0008 and is thus above the default threshold of 0.0001 used in the first round of analysis.

We recommend running the ChIP-sequencing analysis several times using different parameters

to get an idea of the best setting. If you start by running an analysis with relaxed parameters (e.g. not using the Normalized difference or the distribution P-value) you will be able to do a visual inspection of the peaks that would otherwise have been excluded and by interpretation narrow down the right parameter settings.

2.5 Extracting the peak regions

Once you have decided on the right settings for the analysis, make sure that all existing peak annotations are removed (right-click and delete annotations of type binding site) before you run the final analysis. In this way, the annotations on the sequences are right.

If you want to extract the sequence of the binding sites to do motif discovery, you can do this based on the binding site annotation. First, you need to download a plug-in to the Workbench called **Extract annotations**. You need to be connected to the internet to do this:

Help Menu Bar | **Plug-ins and Resources... (****)**

Click the **Download Plug-ins** tab. This will display an overview of the plug-ins that are available for download. Find the **Extract annotations** plug-in and click **Download and Install**. A dialog displaying progress is now shown, and the plug-in is downloaded and installed.

When you close the dialog, you will be asked whether you wish to restart the *CLC Genomics Workbench*. The plug-in will not be ready for use before you have restarted.

When you have re-started you can now extract all the parts of your genome that are covered by a "Binding site" annotation. First, right-click the name of the reference sequence in the mapping view and click **Open This Sequence**. This is because annotations can only be extracted from a sequence alone - not a mapping view.

Toolbox | **General Sequence Analyses (****)** | **Extract Annotations (****)**

Click through using the default settings. You will now see a list of all the binding sites which you can e.g. **Export** () in fasta format for use in a motif discovery tool.