Tutorial

ChIP sequencing: Understanding the details

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

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.

Data set

The data set used derives from a study reported in Nielsen et al., 2008. 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 assembled against the mouse refseq genome, and two regions of chromosome 7 have been taken out for use in this tutorial. The reference sequence of the contig used is 10 Mbp, and there are 23,600 reads of 32 bp each.

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

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 contig 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 14 peaks.

When the result is opened, you need to do a few customizations to make it better suited for interpretation. In the **Contig Settings 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 contig view.

Your screen should now look like figure 1.

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 two rows in the table representing one peak between *Plin* and *Pex11a* and one peak in an intron of *Pex11a*. Click one of the rows to have the contig 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 another peak between the two genes that is not found by the

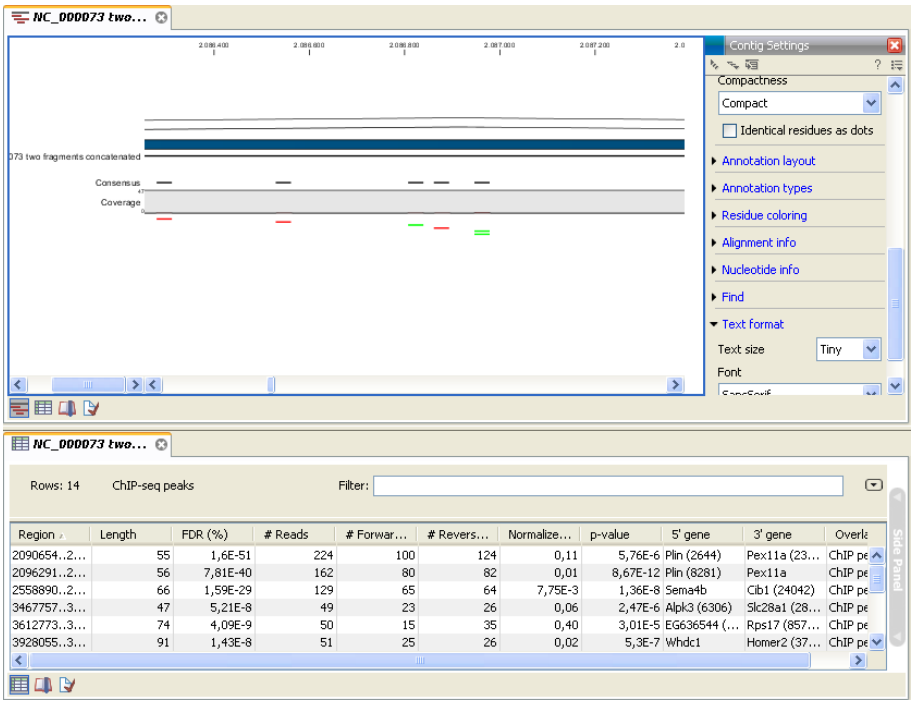


Figure 1: *The view has now been set up.*

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 again 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.

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 contig in the upper view, and click **Undo** ( ). This removes the peak annotations of the contig.

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 101 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 11 peaks.

Next, sort the table on p-value. If you click the peak with a p-value of 0.75, you will see that the distribution of the reads is different from the other peaks we have seen (see figure 2).

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.



Figure 2: *A peak with a random distribution of forward and reverse reads leading to a 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 0.0004 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.

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 contig and click **Open This Sequence**. This is because annotations can only be extracted from a sequence alone - not a contig.

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.



Tutorial

Bibliography

[Nielsen et al., 2008] Nielsen, R., Pedersen, T., Hagenbeek, D., Moulos, P., Siersbæk, R., Megens, E., Denissov, S., Børgesen, M., Francoijs, K., Mandrup, S., et al. (2008). Genome-wide profiling of PPARγ: RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic pathways and changes in RXR dimer composition during adipogenesis. *Genes & Development*, 22(21):2953.