Tutorial

ChIP Sequencing

November 27, 2015

Sample to Insight

CLC bio, a QIAGEN Company · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.clcbio.com · support-clcbio@qiagen.com

ChIP Sequencing

This tutorial takes you through a complete ChIP sequencing workflow using the *CLC Genomics Workbench*. This tutorial makes use of the peak-shape based Transcription Factor ChIP-Seq tool present in *CLC Genomics Workbench* 7.5 and higher.

ChIP-Sequencing is used to analyze the interactions of proteins with genomic DNA. After a cross-linking step that covalently links proteins and DNA, ChIP-seq uses chromatin immunoprecipitation (ChIP) to fish out the relevant pieces of genomic DNA. By subsequent massive parallel DNA sequencing and mapping to the reference genome it is possible to identify binding sites of DNA-associated proteins. It can be used to accurately map global binding sites for any protein of interest when specific antibodies are available. A natural next step bioinformatics analysis is to extract the binding regions and perform pattern discovery to learn about any conserved binding motif in the DNA. Usually a control experiment is performed where the immuno-precipitation step is left out. This control data is typically used to correct for sequencing biases, e.g. genomic regions that are more accessible, repeated regions or copy number aberrations. For further information, see the Wikipedia entry at http://en.wikipedia.org/wiki/Chip-Seq.

The workflow consists of five parts:

- Importing the raw sequencing data.
- Mapping the reads to a reference genome.
- Calling peaks.
- Visualizing the results.
- Extracting the DNA sequences of the peak regions.

In this tutorial we will focus on how to run the analysis and we will not go through the technical details of how the Transcription Factor ChIP-Seq analysis is implemented. 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://clcsupport.com/clcgenomicsworkbench/current/index.php?manual=ChIP_Seq_Analysis.html.

We will look at a subset of a ChIP-seq dataset for the transcription factor NRSF (Neural Restrictive Silencer Factor) on the human cell line Gm12878. Also known as REST (RE1-Silencing Transcription factor), NRSF is a transcription factor involved in the repression of neural genes in non-neuronal cells, such as the lymphoblastoid cell line Gm12878. We therefore expect NRSF ChIP-seq peaks to be associated with genes involved in neural activity. The data was collected by the Myers Lab at the HudsonAlpha Institute for Biotechnology. This dataset is well studied and has been often used to evaluate the performances of ChIP-seq algorithms [Rye et al., 2011]. In addition to the NRSF ChIP-seq dataset, we will also use a control experiment where the immuno-precipitation step is left out. In this tutorial, we will look at only a subset of the data, namely only the reads of the NRSF and control experiments mapping to human chromosome 21.

Importing the raw sequencing data

First, download the data set from our web site: http://download.clcbio.com/testdata/raw_data/ChIP-seq_NRSF_chr21.zip. Unzip the file somewhere on your computer (e.g. the Desktop).

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

File | **Import** ( ) | **Illumina...** ( )

This will bring up the dialog shown in figure 1:

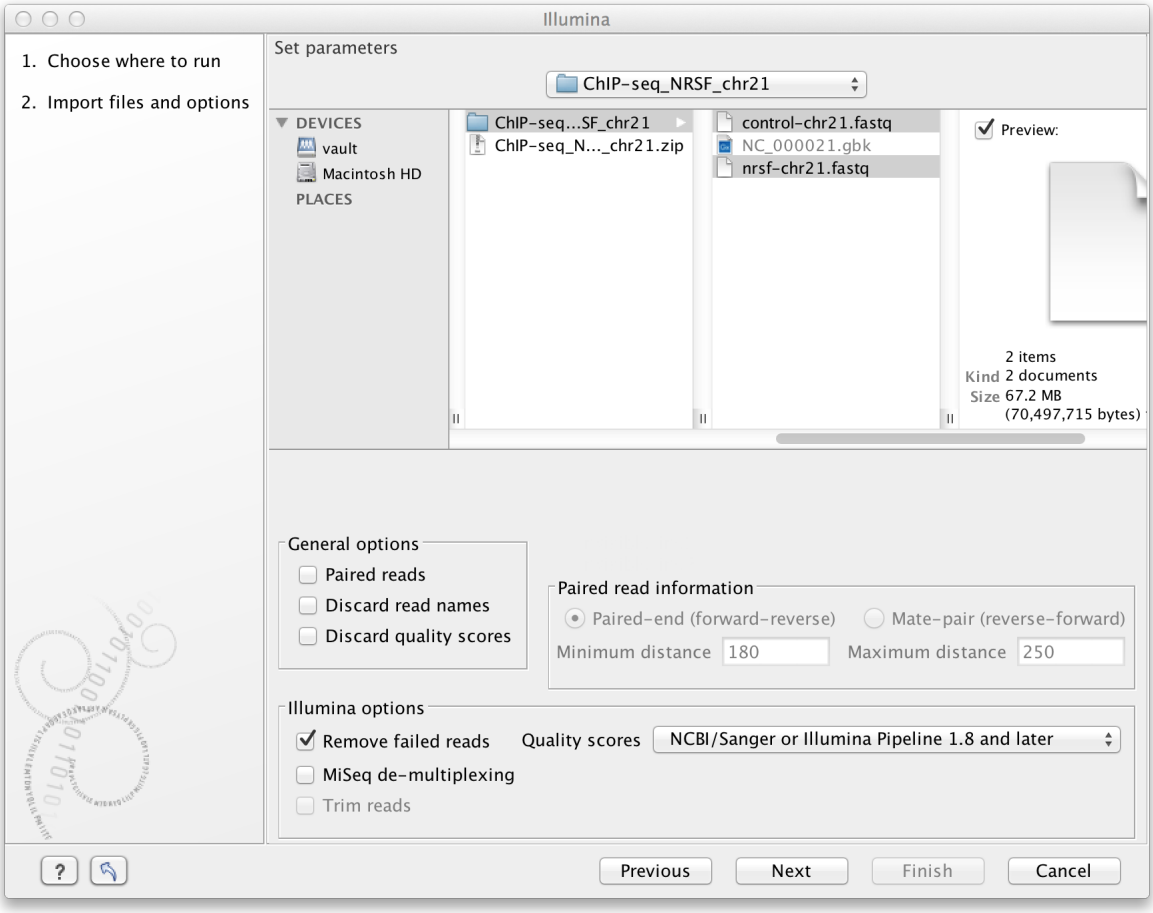


Figure 1: *Import raw reads. When analyzing your own data, you should select the sequencing technology appropriate for your data. This dataset consists of two fastq files obtained using an Illumina sequencer, so the Illumina importer should be chosen.*

Select the `nrsf-chr21.fastq` and `control-chr21.fastq` files and make sure the **Paired reads** checkbox is not checked. The option to discard read names and quality scores are not significant in this context and can be safely set to false because of the relatively small amount of reads. Click **Next**, **Save** the imported reads list and click **Finish**.

After a short while, the raw reads from both files have been imported. Next, import the reference genome sequence that was also included in the zip file.

In this tutorial, we will use only the human chromosome 21 as reference. The reference is provided in genbank format in the file `NC_000021.gbk`. Since this file contains both sequence and annotations, we first import it using the Standard Import tool and later we extract separate tracks to store sequence and gene annotations.

To import the genbank file, drag and drop the `NC_000021.gbk` into the *CLC Genomics Workbench* or use the Standard Import tool:

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

Select the default option **Automatic import**. The *CLC Genomics Workbench* will correctly recognize that the file is in genbank format. Then, press **Next** and choose a folder where the result will be saved.

Next, we want to extract the sequence and gene information for chromosome 21 and store these as tracks:

Toolbox | Track Tools ( ) | Convert to Tracks ( )

Select `NC_000021` ( ) as input, press **Next** and in the next window (figure 2) select both the options **Create sequence track** and **Create annotation tracks**. Then press on the green plus icon ( ) and from the list of possible annotations, add **Gene** by selecting it and pressing the right arrow button ( ) as shown in figure 3.

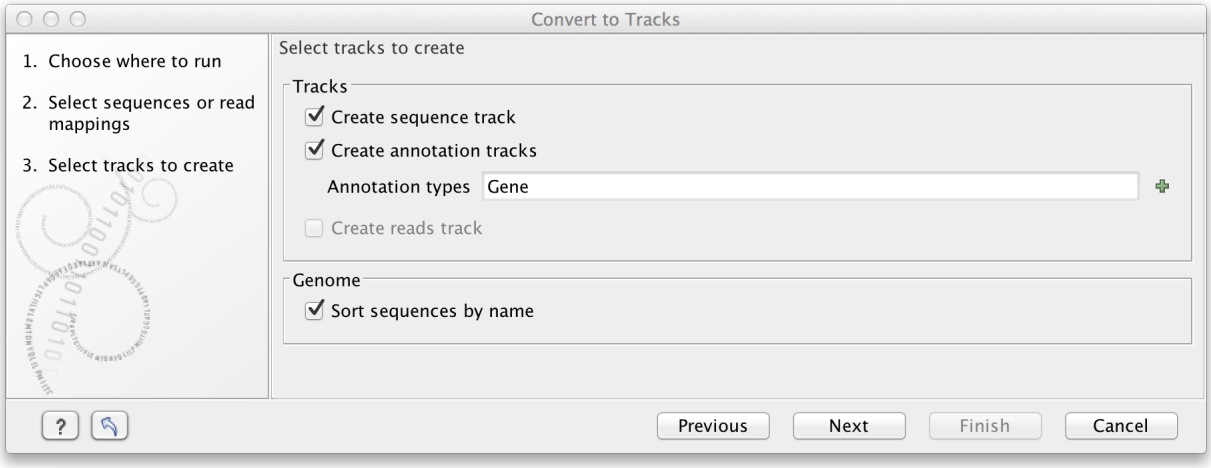


Figure 2: *Extract sequence and gene annotations to tracks.*

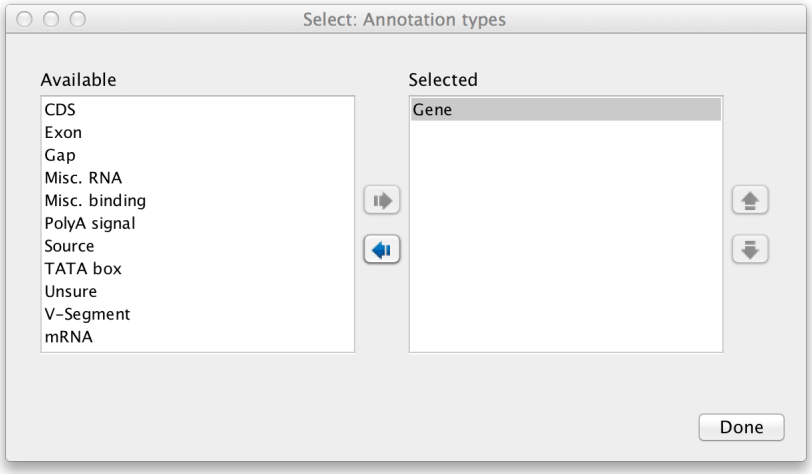


Figure 3: *Select the annotation type Gene.*

The option **Sort sequences by name** is irrelevant, as we are only looking at one chromosome here.

Click **Next**, **Save** the tracks and click **Finish**. The output will be the sequence track

NC_000021 (Genome) ( ) which stores the sequence of chromosome 21 and the annotation track NC_000021 (Gene) ( ), which stores gene annotations for chromosome 21. You should now have the files depicted in figure 4:

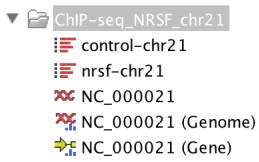


Figure 4: *The files created after the importing step is done.*

Mapping the reads to the reference genome

Once the data has been imported, the next step in the analysis is to map the reads to the reference genome:

Toolbox | NGS Core Tools ( ) | Map Reads to Reference ( )

The dialog shown in figure 5 allows you to choose the files containing the raw reads. Since we want to map two lists, we check the **Batch** option to enable the batch mode and select the folder where the sequence lists are stored (ChIP-seq_NRSF_chr21 in figure 5).

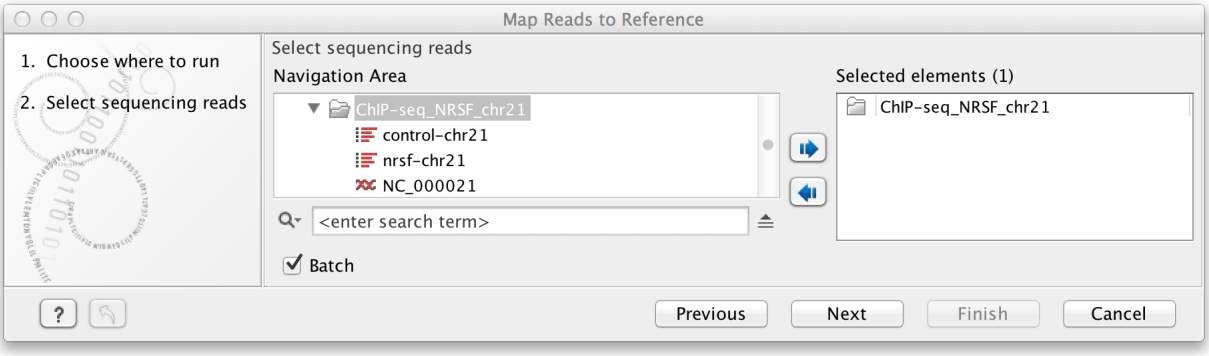


Figure 5: *Select sequence list containing the reads. Since we want to map two lists, we choose the batch mode.*

We then press **Next** and check that only the two lists control-chr21 ( ) and nrsf-chr21 ( ) are selected (figure 6).

Clicking **Next** will allow you to select a reference sequence as shown in figure 7.

At the top you select NC_000021 (Genome) ( ) 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 tutorial we are using only chromosome 21. Note that both the sequence track NC_000021 (Genome) ( ) that we just selected and the sequence NC_000021 ( ) could be used as reference.

Click **Next** and set mapping options as shown in figure 8.

For ChIP-seq, we recommend stringent mapping settings. Setting the length fraction to 0.5 specifies the minimum length fraction of a read that must match the reference sequence, and setting the similarity fraction to 0.8 specifies the minimum fraction of similarity between the read and the reference sequence. The mismatch, insertion, and deletion costs are here set at 2, 3

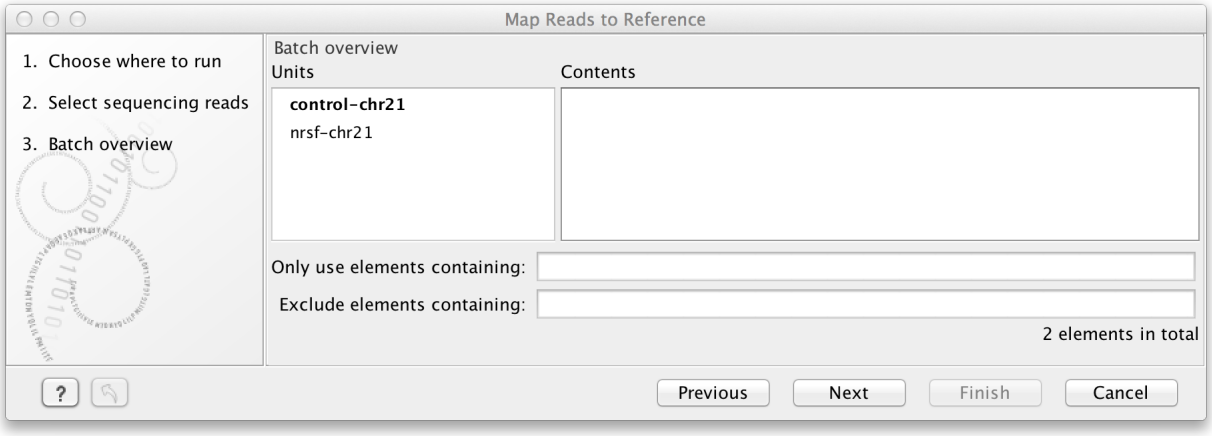


Figure 6: *Check that all reads are used as input for the mapping.*

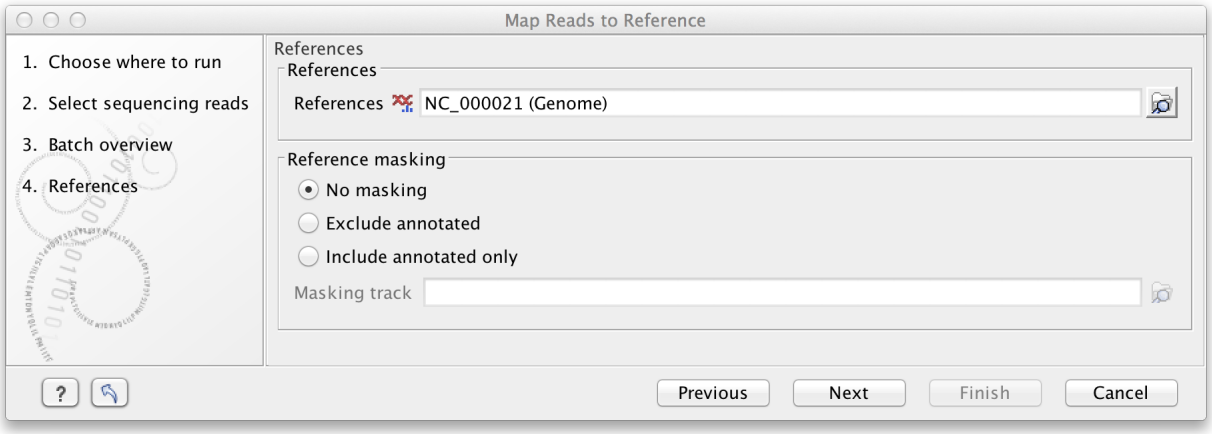


Figure 7: *Specifying the reference sequences and masking.*

and 3. Next, select to ignore the non-specific matches. The settings are not important for the result of this tutorial, but when you work with your own data, this may be important. For more information about the other settings, please click the **Help** (?) button.

After clicking **Next**, the dialog shown in figure 9 now appears.

Select **Create reads track** to create track-based results. Check **Create report** to obtain a detailed report about the read mapping and leave **Collect un-mapped reads** unchecked since we are not interested in those reads. Select the output options 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 reference sequence (Human chromosome 21, with a size of 47 Mbp), it may take a few minutes to map the data.

Calling peaks

The results of the read mapping are now used as input to the Transcription Factor ChIP-Seq tool to detect significant peaks:

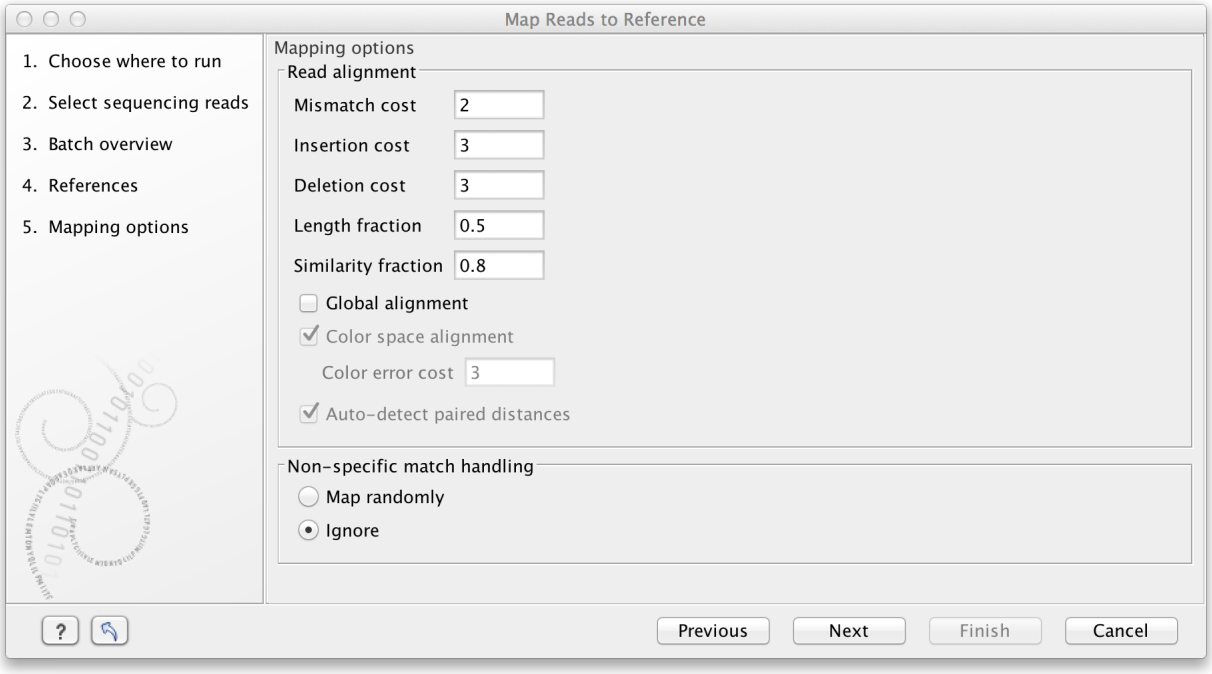


Figure 8: *A stringent read matching is desired for ChIP-seq.*

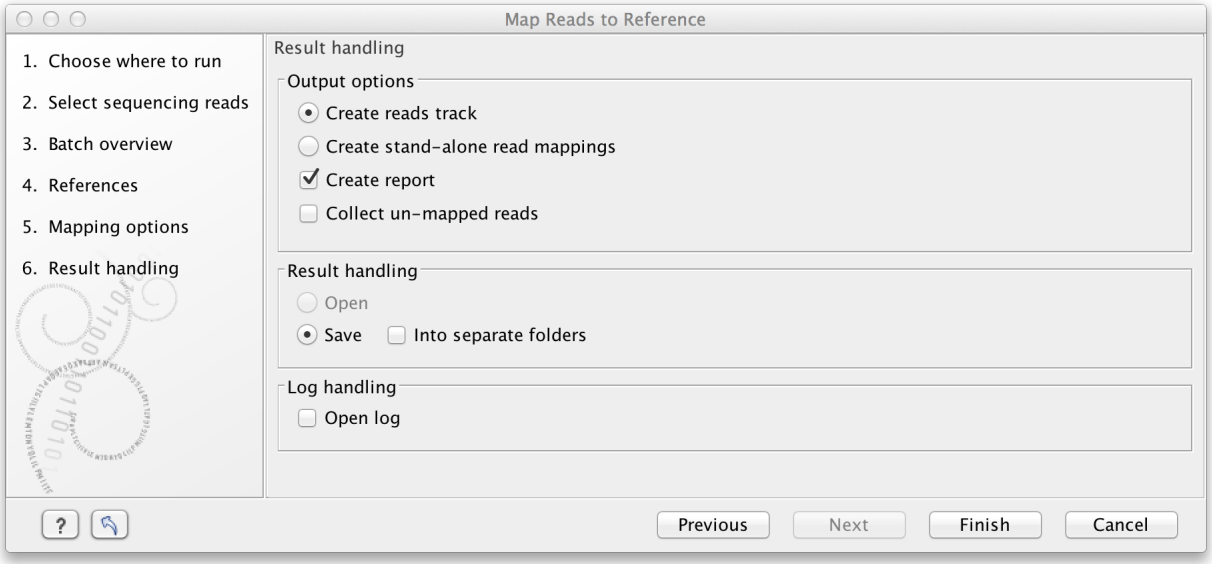


Figure 9: *Select Create reads track, Create report, and Save.*

Toolbox | **Epigenomics Analysis (**) | **Transcription Factor ChIP-Seq (**)

This opens a dialog where you select the `nrsf-chr21 (Reads)` ( ) and click **Next**.

You can now choose `control-chr21 (Reads)` ( ) as control data (see figure 10). You can leave the **Maximum P-value for peak calling** to the default value of 0.10. A smaller P-value can be specified to obtain a smaller number of high-quality peaks, while a higher P-value threshold can be set to obtain a higher number of peaks.

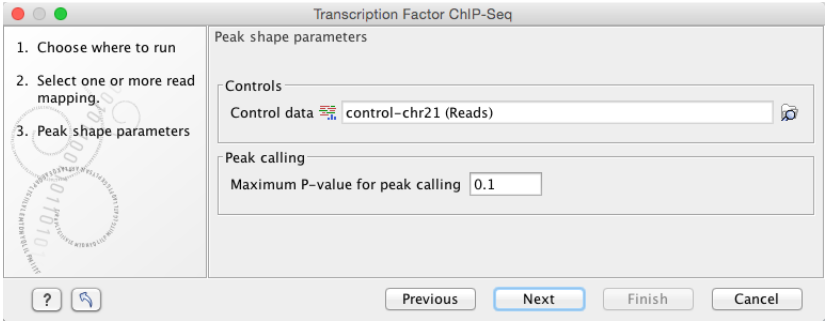


Figure 10: *Choose control data.*

After clicking **Next** you can choose the output data to be generated (see figure 11):

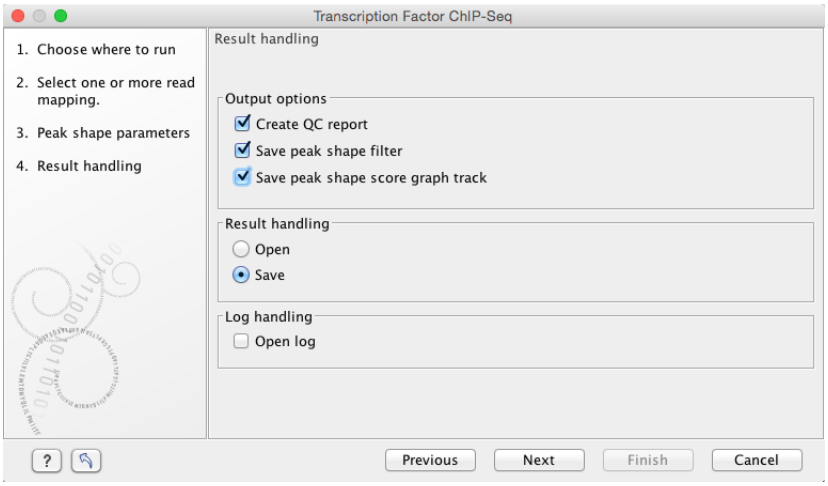


Figure 11: *Select the output data to be generated.*

In this tutorial, we select all the output which the Transcription Factor ChIP-Seq tool can generate. After a few minutes, the analysis will complete and the following results will appear:

- `nrsf-chr21 (Reads) (Peaks)` ( ) the list of all called peaks.
- `nrsf-chr21 (Reads) (QC Report)` ( ) The quality control reports. The QC report contains metrics about the quality of the ChIP-seq experiment.
- `nrsf-chr21 (Reads) (Peak shape filter)` ( ) The peak shape filter contains the peak shape that was learned during the ChIP-seq analysis.
- `nrsf-chr21 (Reads) (Peak shape score)` ( ) A graph track containing the peak shape score. The track shows the peak shape score for each genomic position.

Before continuing the analysis or looking at the results, we recommend to look at the quality control report. The most important sections of the report are the tables containing **Quality measures**. The report `nrsf-chr21 (Reads) (QC Report)` ( ) will contain one table for the NRSF dataset (figure 12) and one for the control dataset (figure 13).

For each of the 3 measures, the table provides the name, the value, notes to better understand the meaning of the measure and a status, which can assume the value **OK** if the value is reflective of sufficient quality and **Low** (or **Very Low**) if the value is lower than the quality threshold. For

Measure	Value	Status	Notes
Number of reads	486,047	Very low	For mammalian cells, this value should be at least 10 million reads. For smaller organisms (e.g. worm and fly), this value should be at least 2 million reads
Relative strand correlation	1.012	OK	The relative strand correlation describes the ratio between the fragment-length peak and the read- length peak in the cross-correlation plot. This value should be greater than 0.8 for transcription factor binding sites, but can be lower for ChIP-seq input or for histone marks
Normalized strand coefficient	2.488	OK	The normalized strand coefficient describes the ratio between the fragment-length peak and the background cross-correlation values. This value should be greater than 1.05 for ChIP-seq experiments

Figure 12: *Table of quality measures for the NRSF ChIP-seq dataset.*

more details on how the quality thresholds were determined, see Landt et al., 2012 and Marinov et al., 2014. In figure 12, the values for the relative strand correlation and the normalized strand coefficient are OK, while the number of reads is classified as **Very Low**. This should not be surprising or worrisome because the data used in this tutorial is a small subset of a ChIP-seq experiment. In fact, the full datasets consists of about 16 millions reads, which is significantly higher than the threshold value[1]. However, in normal circumstances, a small number of reads would be a strong indicator that the ChIP-seq experiment is of low quality.

Measure	Value	Status	Notes
Number of reads	307,787	Very low	For mammalian cells, this value should be at least 10 million reads. For smaller organisms (e.g. worm and fly), this value should be at least 2 million reads
Relative strand correlation	1.192	OK	The relative strand correlation describes the ratio between the fragment-length peak and the read- length peak in the cross-correlation plot. This value should be greater than 0.8 for transcription factor binding sites, but can be lower for ChIP-seq input or for histone marks
Normalized strand coefficient	2.375	OK	The normalized strand coefficient describes the ratio between the fragment-length peak and the background cross-correlation values. This value should be greater than 1.05 for ChIP-seq experiments

Figure 13: *Quality measures for the control ChIP-seq dataset.*

The quality measures table for the control experiment (figure 13) can be interpreted in a similar fashion. We note that, since this is a control experiment, the value of the relative strand correlation is not important and the status would be OK also for low values. As for NRSF, the fact that the number of reads is very low is due to the fact that only a small subset of the data was used.

The quality report contains additional information that could be used for troubleshooting. For example, if the relative strand correlation or the normalized strand coefficient were classified as low, the cross-correlation plots should be examined in more details. More information regarding the cross-correlation plots and the Transcription Factor ChIP-Seq tool can be found in the user manual. Click the **Help** (?) button or go to `http://clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Running_ChIP_Seq_Analysis_tool.html`.

[1] In this tutorial, we used only the subsets of the data mapping to chromosome 21. The complete datasets can be found at the UCSC website. The complete NRSF dataset is available at `http://hgdownload-test.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeHudsonalphaChipSeq/release1/wgEncodeHudsonalphaChipSeqRawDataRep1K562Nrsf.fastq.gz`. The complete control dataset is available at `http://hgdownload-test.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeHudsonalphaChipSeq/release1/wgEncodeHudsonalphaChipSeqRawDataRep1K562Control.fastq.gz`. You can download the human reference genome for hg18 from the *CLC Genomics Workbench* using the command **Download** | **Download Genome Data...** ( ) | **Animal (mammals)** | **Homo sapiens (hg18)**

After having verified that the quality of the ChIP-seq datasets is acceptable, the next step is to annotate them with information about their nearest upstream and downstream genes. This can be done using the Annotate with Nearby Gene Information tool:

Toolbox | **Epigenomics Analysis** ( ) | | **Annotate with Nearby Gene Information** ( )

Select first the track to annotate (`nrsf-chr21 (Reads) (Peaks)` ( )), and after clicking **Next**, the dialog shown in figure 14 will appear:

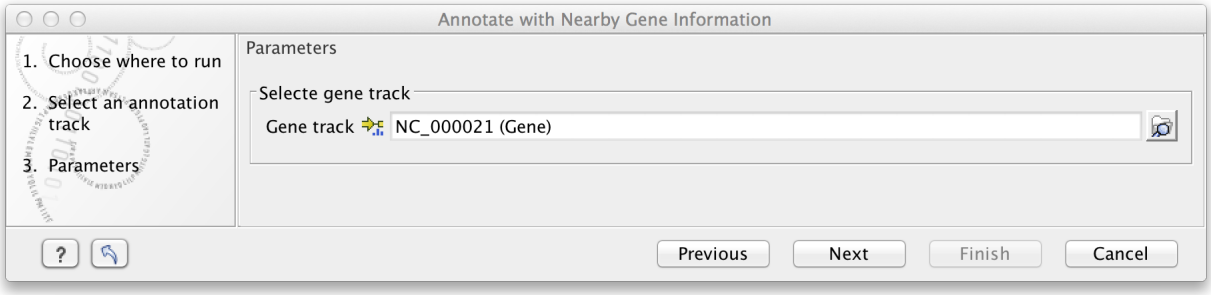


Figure 14: *Select the annotation track to be used as gene reference.*

Choose `NC_000021 (Gene)` ( ) as the reference gene track, then click **Next** and **Save** the result. The file `nrsf-chr21 (Reads) (Peaks) (Annotated)` ( ) will be generated.

You should now have the files depicted in figure 15:

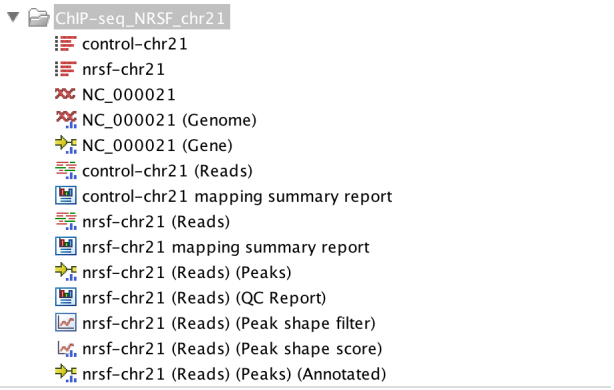


Figure 15: *All files created after the Transcription Factor ChIP-Seq analysis is done.*

Visualizing the results

The best way to visualize the results is to create a Track List:

Toolbox | **Track Tools** ( ) | **Create Track List** ( )

Select the tracks we created so far as shown in figure 16 and then press **Finish**.

Once the Track list is created, the easiest way to explore peaks is to make a split view of the table and the peak annotation track by double-clicking on the label `nrsf-chr21 (Reads) (Peaks) (Annotated)`. You will then be able to browse through the peaks by clicking in the table, as the peak annotation track and the table are connected. As a result the view will jump to the position of the peak selected in the table. You can browse through all the 144 peaks found

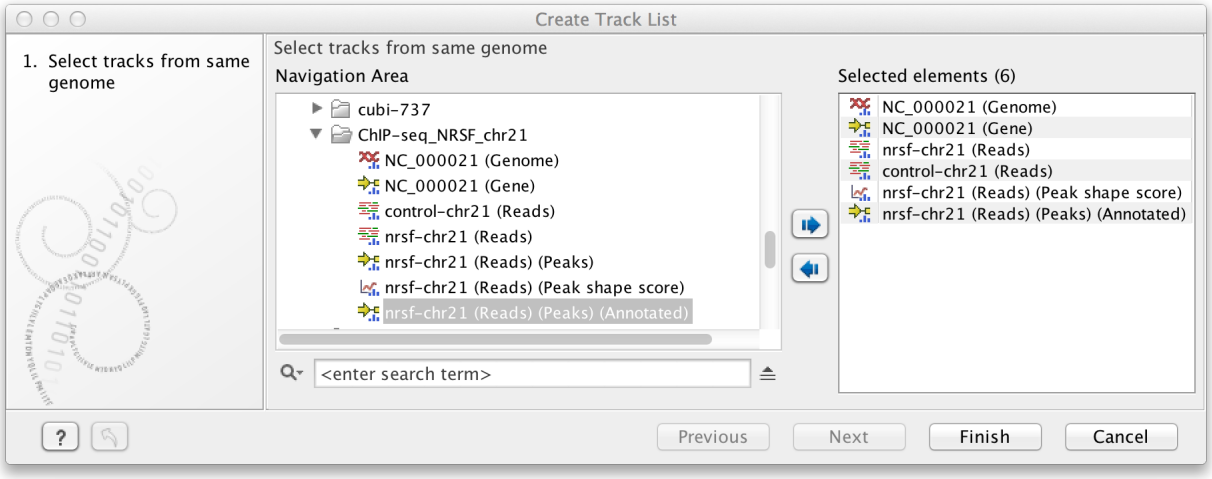


Figure 16: *Create a Track List to visualize the results.*

for this sample by selecting in the table. Next, we sort the table according to P-value so that we can look at the top peak (figure 17).

The strongest peak is close to the gene SYNJ1 (synaptojanin 1). This gene encodes a phosphoinositide phosphatase that regulates levels of membrane phosphatidylinositol-4,5-bisphosphate. The expression of this enzyme affects synaptic transmission and thus it is not a surprise that this gene is inhibited by NRSF, whose function is to repress neural genes in non-neuronal cells. Note the nicely distributed green (forward) and red (reverse) reads for this peak, this is a typical shape for transcription factors.

Extracting the DNA sequences of the peak regions.

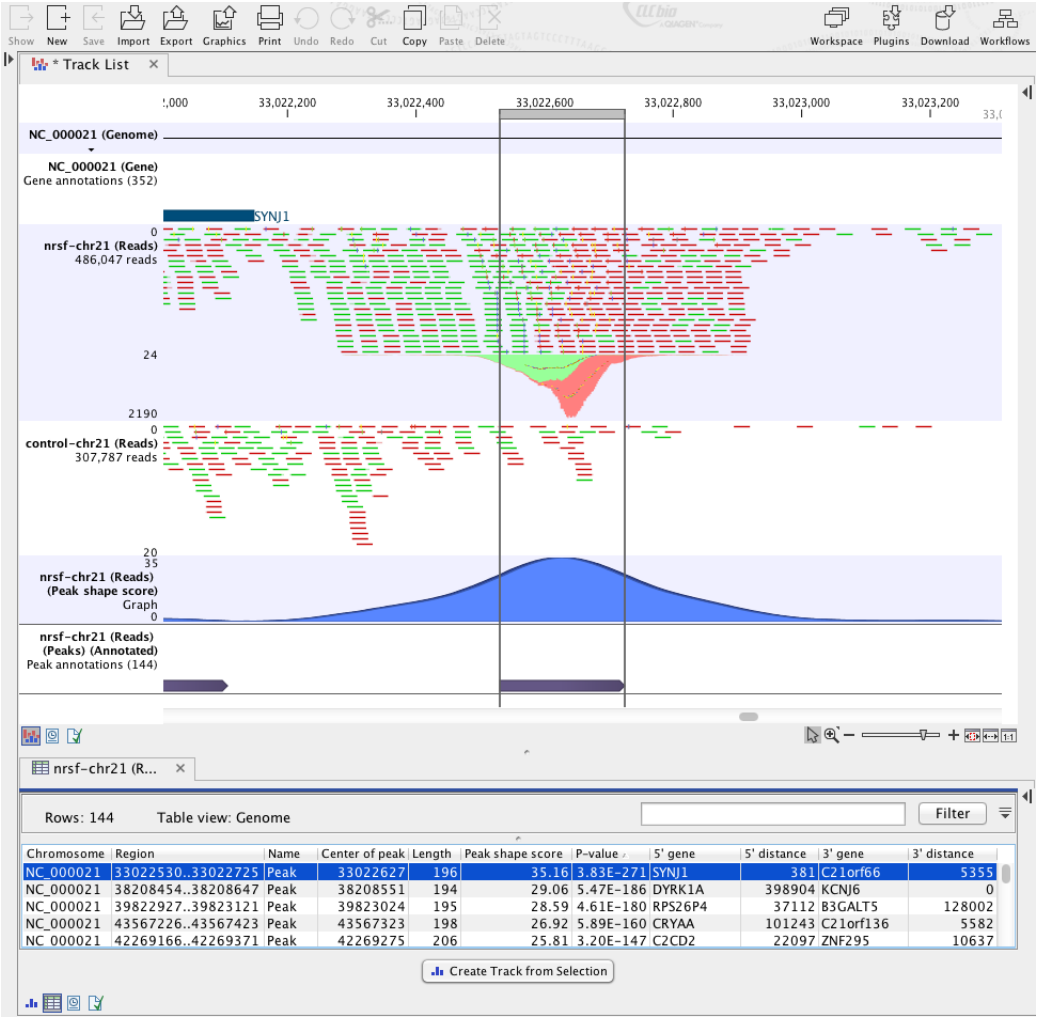
A common step in the analysis of ChIP-seq data is to extract the DNA sequences associated with peaks in the ChIP-seq data. These sequences are typically enriched with respect to some DNA motif, especially when the protein under examination is a transcription factor such as NRSF. This tutorial only covers the step of extracting the sequences. The motif discovery can be then performed using external applications such as MEME (Bailey et al., 2006, `http://meme.nbcr.net/meme/intro.html`) or TRANSFAC® (Matys et al., 2006, `http://www.biobase-international.com/product/transcription-factor-binding-sites`).

You can then use the Extract Annotations tool to extract the sequences related to the peak regions:

> **Toolbox** | **Classical Sequence Analysis** () | **General Sequence Analysis** () | **Extract Annotations** ()

After selecting the peak file `nrsf-chr21 (Reads) (Peaks) (Annotated)` () as input, the dialog shown in figure 18 will appear:

Select the sequence track `NC_000021 (Genome)` () and choose the annotation type Peak by clicking the green plus icon (). Choose **Include annotation region** and **Include annotation chromosome** to give informative names to the resulting sequences, then Click **Next**, choose to **Save** the sequences and click **Finish**. After few seconds, the sequences will be exported to a file named `Extracted Annotations` ().



Chromosome	Region	Name	Center of peak	Length	Peak shape score	P-value	5' gene	5' distance	3' gene	3' distance
NC_000021	33022530..33022725	Peak	33022627	196	35.16	3.83E-271	SYNJ1	381	C21orf66	5355
NC_000021	38208454..38208647	Peak	38208551	194	29.06	5.47E-186	DYRK1A	398904	KCNJ6	0
NC_000021	39822927..39823121	Peak	39823024	195	28.59	4.61E-180	RPS26P4	37112	B3GALT5	128002
NC_000021	43567226..43567423	Peak	43567323	198	26.92	5.89E-160	CRYAA	101243	C21orf136	5582
NC_000021	42269166..42269371	Peak	42269275	206	25.81	3.20E-147	C2CD2	22097	ZNF295	10637

Figure 17: *A very strong peak near the gene SYNJ1.*

Most external sequence-based analysis tools require an input in fasta file format. To export the sequence as fasta, you can run the Export tool:

File | **Export** ( )

Then, choose the **fasta** format, select the file `Extracted Annotations` ( ) and finally select the output file name.

You have now performed a complete ChIP-seq Analysis on a small dataset using the *CLC Genomics Workbench*, starting from raw sequencing data and ending with the inspection of the called peaks and the extraction of the sequences within peak regions. The ChIP-seq workflow described here does not require any tweaking of parameters and can be readily applied to larger ChIP-seq datasets.

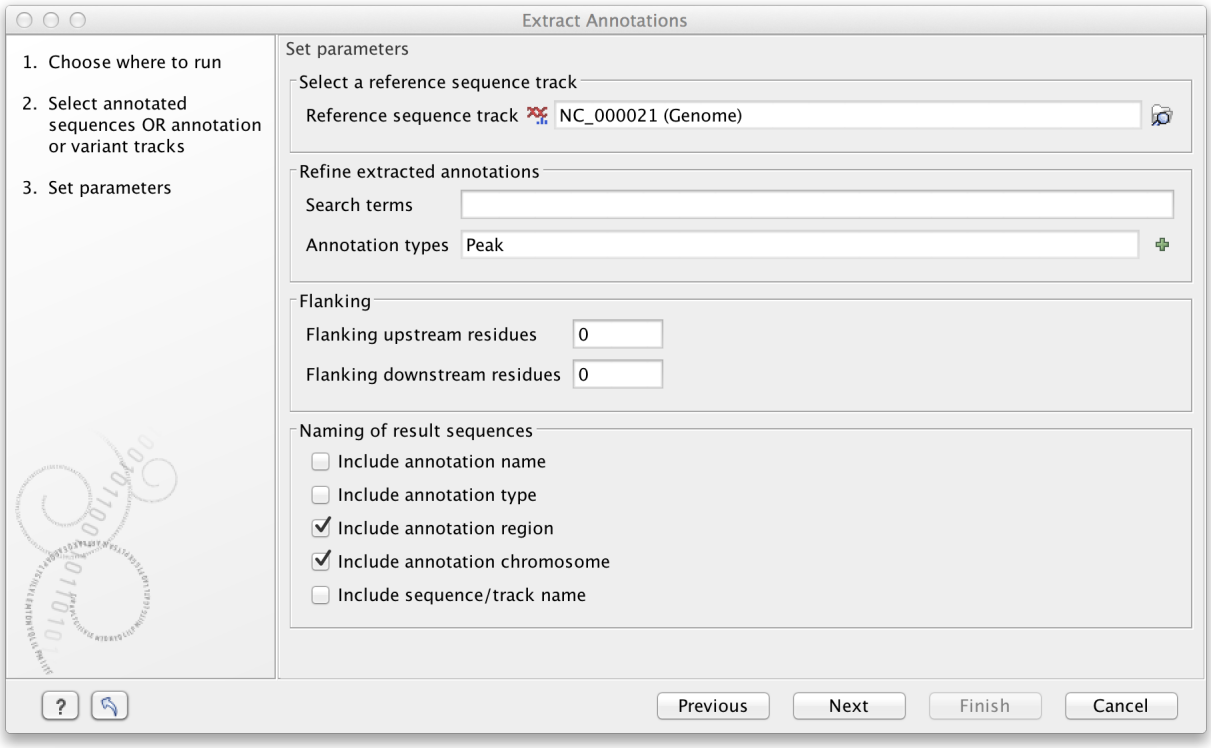


Figure 18: *Options for the Extract Annotations tool.*

Bibliography

[Bailey et al., 2006] Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(suppl 2):W369–W373.

[Landt et al., 2012] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res*, 22(9):1813–31.

[Marinov et al., 2014] Marinov, G. K., Kundaje, A., Park, P. J., and Wold, B. J. (2014). Large-scale quality analysis of published ChIP-seq data. *G3 (Bethesda)*, 4(2):209–23.

[Matys et al., 2006] Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A. E., and Wingender, E. (2006). TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–10.

[Rye et al., 2011] Rye, M. B., Sætrom, P., and Drabløs, F. (2011). A manually curated ChIP-seq benchmark demonstrates room for improvement in current peak-finder programs. *Nucleic Acids Res*, 39(4):e25.