Tutorial

ChIP Sequencing

February 1, 2024

QIAGEN Aarhus A/S · Kalkværksvej 5, 11. · DK - 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Sample to Insight

Tutorial

ChIP Sequencing

The purpose of this tutorial is to demonstrate how *CLC Genomics Workbench* can be used to analyze ChIP-Seq data.

We focus on the following:

- Import the raw sequencing data.
- Map the reads to a reference genome.
- Call peaks.
- Visualize the results.

Data used in this tutorial

This tutorial uses a ChIP-Seq dataset for the transcription factor NRSF (Neural Restrictive Silencer Factor) generated in the human lymphoblastoid cell line GM12878 [Rye et al., 2011]. NRSF represses neural genes in non-neuronal cells, and its ChIP-Seq peaks are therefore expected to be associated with genes involved in neural activity. A control dataset, in which the immunoprecipitation step was omitted, is included for bias correction.

To complete the tutorial in a reasonable amount of time, only a subset of the reads mapping to chromosome 21 are used here.

Prerequisites

For this tutorial, you must be working with *CLC Genomics Workbench* 24 or higher. Note that higher versions may produce slightly different results than those shown here.

General tips

- Throughout this tutorial, we provide links to relevant manual pages, which we recommend exploring for additional details.
- Tools can be found in the Toolbox, but it is often easier to launch them using Quick Launch (), found in the top toolbar (shortcut Ctrl+Shift+T or ⌘ +Shift+T on Mac). Quick Launch displays the full Toolbox path, making it easy to identify the location of the tool if needed.
- The in-built manual can be accessed by clicking the **Help** button on wizards or by selecting the **Help** option under the **Help** menu.
- Within wizards, the **Reset** button can be used to change settings to their default values.
- Columns in tables can be hidden by unchecking their name in the Side Panel.
- Columns in tables can be used to sort the rows, by successively clicking on the column name until the desired order (indicated by an arrow next to the column name) is achieved.
- Many data elements produced by *CLC Genomics Workbench* tools have multiple views, indicated as icons in the lower left corner of elements opened in the View Area. Clicking on one of the view icons while pressing the Ctrl (⌘ on Mac) key will open in split view such that both views are visible at the same time. Often, if viewing a table and a graphical representation in split view, selecting entries in the table will highlight them in the graphical representation. The order of the views can be changed using drag and drop, see Arrange views in View Area.

Importing the raw sequencing data

1. Download the tutorial data and unzip it.
2. Start *CLC Genomics Workbench*.
3. Launch the Illumina importer ( ) using Quick Launch ( ).
4. In the first wizard step, "Import files and options", click on the **Add files** button, locate the tutorial data, and select the `nrsf-chr21.fastq` and `control-chr21.fastq` files (figure 1).

   Make sure the **Paired reads** checkbox is not checked.

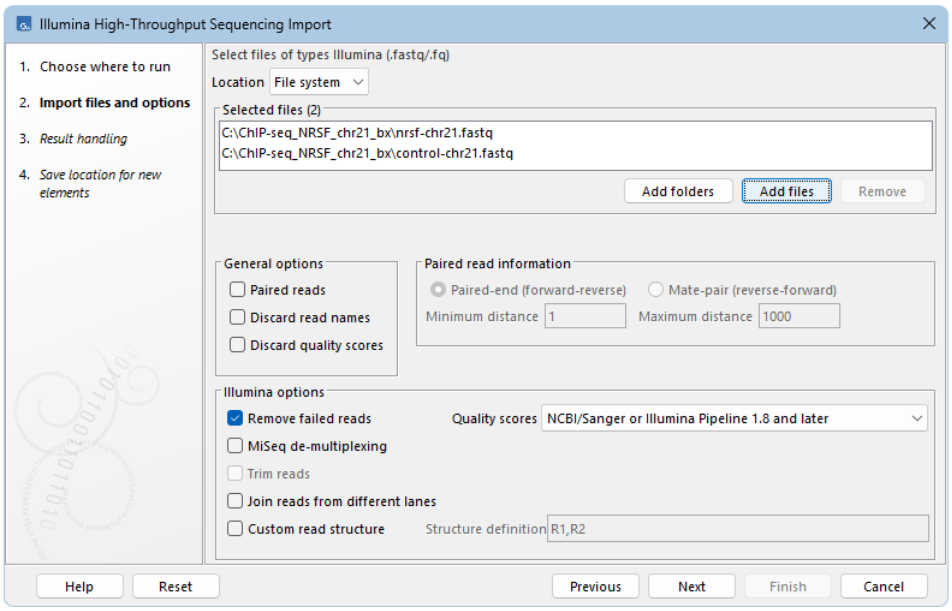

Figure 1: *Import raw reads. When analyzing your own data, you should select the sequencing technology appropriate for your data. This dataset consists of two fastq files obtained using an Illumina sequencer, so the Illumina importer should be chosen.*

5. In the next step, "Result handling", choose **Save**.
6. In the last step, "Save location for new elements", select a suitable location in the Navigation Area to save the imported data and click on **Finish**.

Next, import the reference genome files that were also included in the downloaded zip file by either:

- Dragging and dropping the files `NC_000021 (Genome).clc` and `NC_000021 (Gene).clc` into the Navigation Area.

  These files correspond to the genomic chromosome 21 reference sequence and the gene annotation track for chromosome 21, respectively.
- Using the Standard Import tool with the option **Automatic import** to import the two chromosome 21 reference files.

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:

1. Launch Map Reads to Reference ( ) using Quick Launch ( ).
2. In the first wizard step, "Select sequencing reads", select the `ChIP-Seq` folder where the imported sequence lists are stored (figure 2).

   Since we want to map two lists simultaneously, we must check the **Batch** option.

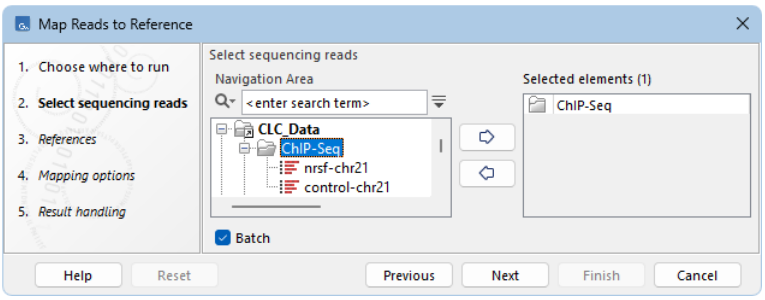

Figure 2: *Select the sequence lists containing the reads we want to analyze. Since we want to map two lists, we choose the batch mode.*

3. In the next step, "Batch overview", verify that only the two sequence lists `control-chr21` and `nrsf-chr21` are selected (figure 3).

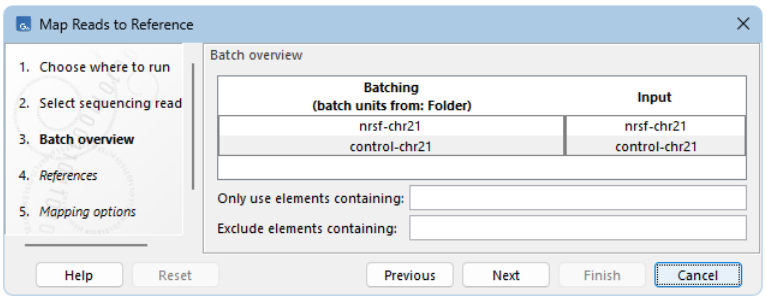

Figure 3: *Verify that the correct reads are used as input for the mapping.*

4. In the next step, "References", select the reference sequence `NC_000021 (Genome)` ( ) (figure 4).
5. In the next step, "Mapping options", keep the default settings under "Read alignment" and select **Ignore** under "Non-specific match handling" (figure 5).
6. In the next step, "Result handling", Select **Create reads track** to create track-based results (figure 6).

   Also check **Create report** to obtain a detailed report about the read mapping.

   Choose to **Save in specified location**.
7. In the last step, "Save location for new elements", select a suitable location in the Navigation Area to save the results and click on **Finish**.

You can follow the progress of the mapping both in the status bar at the bottom left corner and under the Processes tab in the Toolbox.

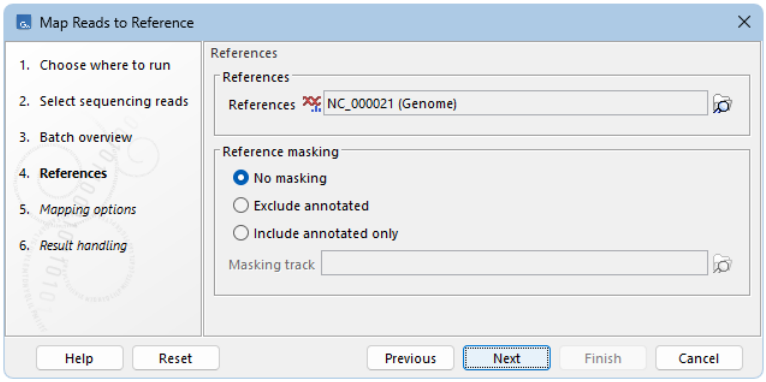

Figure 4: *Specifying the reference sequence and masking parameters.*

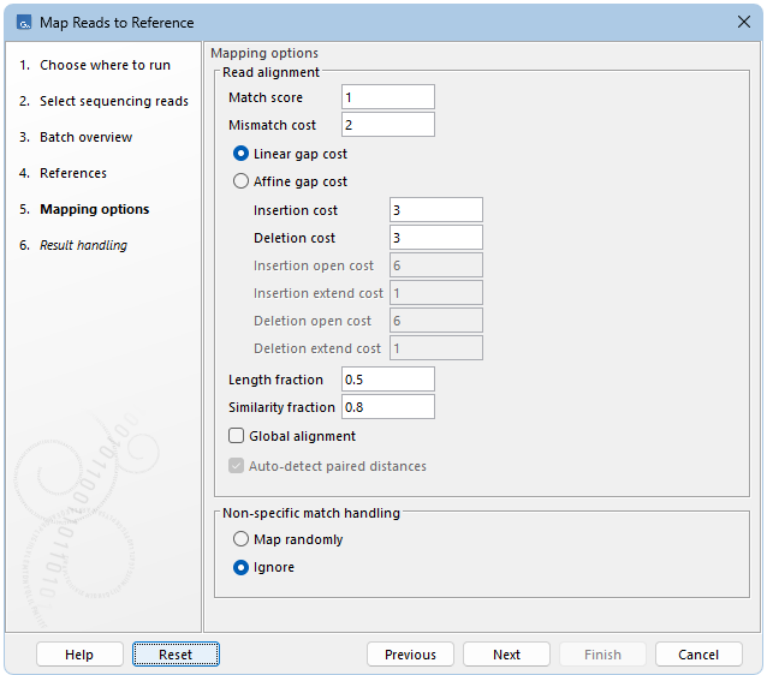

Figure 5: *Use default settings for the read mapping but choose to ignore non-specific matches.*

Calling peaks

We will now use the read mapping results as input to the Transcription Factor ChIP-Seq tool to detect significant peaks.

1. Launch Transcription Factor ChIP-Seq ( ) using Quick Launch ( ).
2. In the first step of the wizard that opens, "Select one or more read mappings", select the `nrsf-chr21 (Reads)` ( ) (figure 7).
3. In the next step, "Peak shape parameters", choose `control-chr21 (Reads)` ( ) as **Control data** (figure 8).

   Keep the default value of 0.1 for **Maximum P-value for peak calling**. 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.

Tutorial

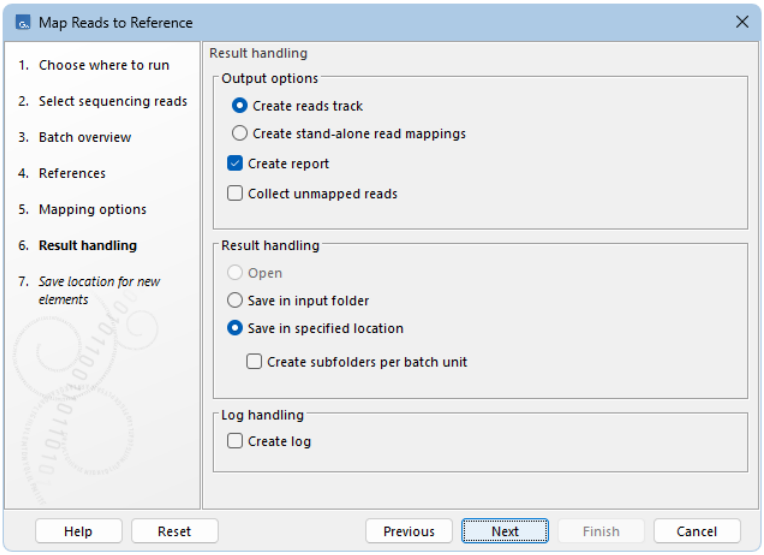

Figure 6: *Select Create reads track, Create report, and Save in specified location.*

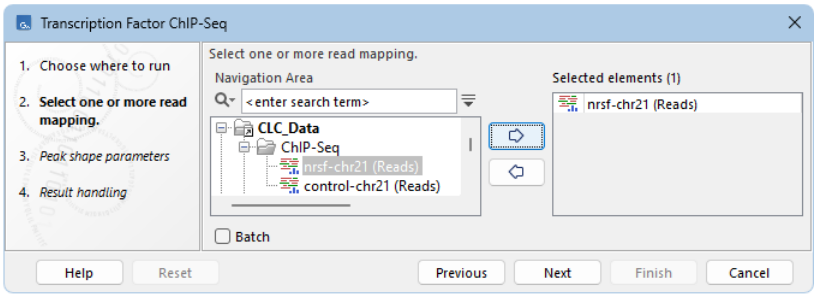

Figure 7: *Select the read mapping to use for peak detection.*

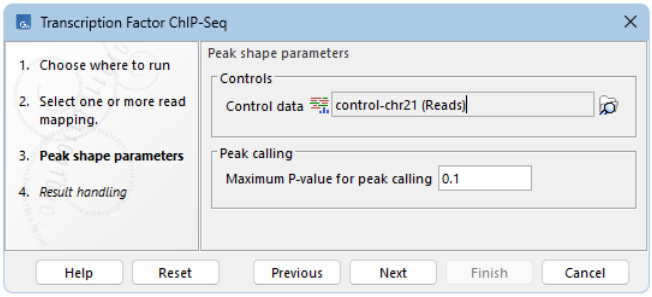

Figure 8: *Choose control data.*

4. In the next step, "result handling", check all three output options (figure 9) and choose **Save**.

5. In the last step, "Save location for new elements", select a suitable location in the Navigation Area to save the results and click on **Finish**.

When the analysis completes, the following results will appear in the selected location in the Navigation Area:

- `nrsf-chr21 (Reads) (Peaks)` (): The list of all called peaks.
- `nrsf-chr21 (Reads) (QC Report)` (): The quality control report. This report

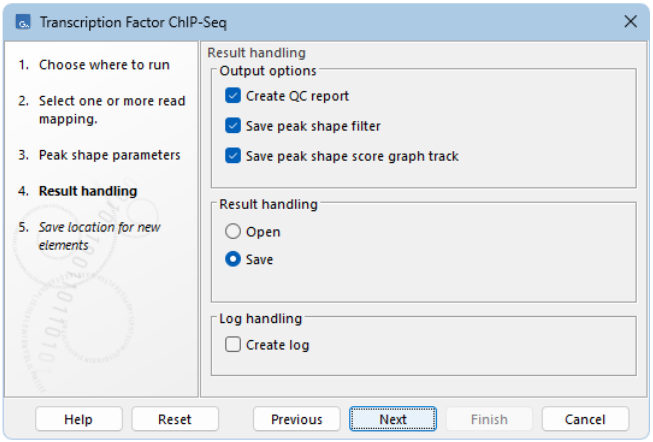

Figure 9: *Select which outputs to generate.*

contains metrics about the quality of the ChIP-Seq experiment.

- `nrsf-chr21 (Reads) (Peak shape filter)` (): The peak shape filter results 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 looking at the quality control report. The most important sections of the report are the tables containing **Quality measures**. For each of the 3 quality measures, the table provides the name, the value, notes to better understand the meaning of the measure, and a status. The status will be **OK** if the quality value is sufficient, or **Low** (or **Very Low**) if the value is lower than the quality threshold. For more details on how the quality thresholds were determined, see Landt et al., 2012 and Marinov et al., 2014.

In figure 10, 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 is not surprising or worrisome because the data used in this tutorial is a small subset of a ChIP-Seq experiment. In fact, the full dataset consists of about 16 million reads, which is significantly higher than the threshold value. However, under normal circumstances, a small number of reads would be a strong indicator that the ChIP-Seq experiment is of low quality.

The quality measures table for the control experiment (figure 11) can be interpreted in a similar way. 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.

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:

1. Launch Annotate with Nearby Information () using Quick Launch ().

1.1 Quality measures

Measure	Value	Status	Notes
Number of reads	486,301	Very low	For mammalian cells, this value should be at least 10 million reads. For organisms with smaller genomes (e.g. worm and fly), this value should be at least 2 million reads
Relative strand correlation	1.068	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	2.564	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 10: *Table of quality measures for the NRSF ChIP-Seq dataset.*

2.1 Quality measures

Measure	Value	Status	Notes
Number of reads	307,932	Very low	For mammalian cells, this value should be at least 10 million reads. For organisms with smaller genomes (e.g. worm and fly), this value should be at least 2 million reads
Relative strand correlation	1.196	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	2.377	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 11: *Table of quality measures for the control ChIP-Seq dataset.*

2. In the first step of the wizard that opens, "Select one annotation track", select the track to annotate (`nrsf-chr21 (Reads) (Peaks)` ()) (figure 12).

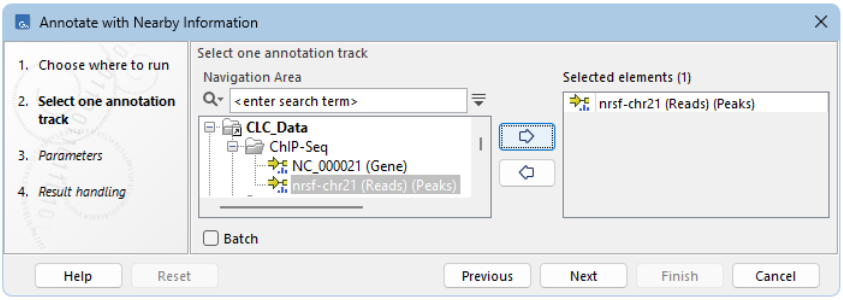

Figure 12: *Select the track to annotate.*

3. In the next step, "Parameters", choose `NC_000021 (Gene)` () as the reference **Annotation track** (figure 13), and click **Next**.

4. In the next step, "Result handling", choose to **Save** the results.

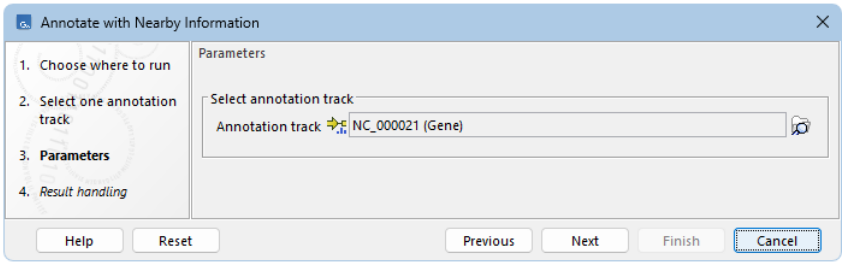

Figure 13: *Select the annotation track to be used as gene reference.*

5. In the last step, "Save location for new elements", select a suitable location in the Navigation Area to save the results and click on **Finish**.

   The file `nrsf-chr21 (Reads) (Peaks, Annotated)` ( ) will be generated.

Visualizing the results

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

1. Launch Track List ( ) using Quick Launch ( ).
2. In the wizard that opens, select the tracks we created so far (figure 14) and then click on **Finish**.

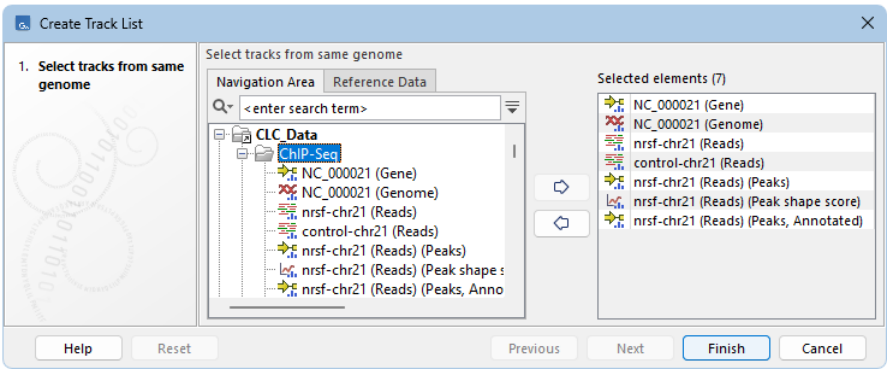

Figure 14: *Create a Track List to visualize the results.*

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)` (figure 15). Sort the table according to P-value so that we can look at the top peak. You will then be able to browse through the peaks by clicking in the table, jumping in the track list to the position of the peak selected in the table. Click on the 1:1 zoom in the bottom right corner of the track list to zoom in on the peak of interest, and zoom out as needed to see the closest gene. You can browse through all of the peaks found for this sample by selecting in the table.

The strongest peak is close to the gene SYNJ1 (synaptojanin 1). This gene encodes a phospho-inositide 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.

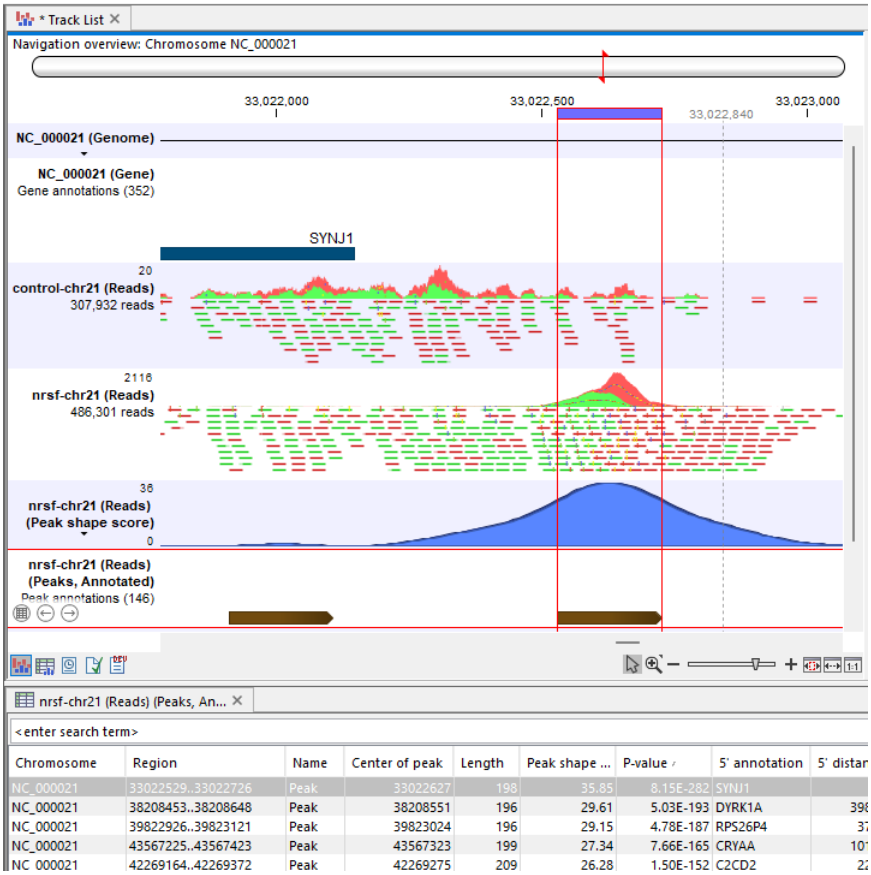

Figure 15: *A very strong peak near the gene SYNJ1.*

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. Motif discovery can then be performed using external applications such as TRANSFAC®.

To extract the sequences related to peak regions:

1. Launch Extract Annotated Regions ( ) using Quick Launch ( ).
2. In the first step of the wizard that opens, "Select annotated sequences OR a track of annotations, variants or statistical comparisons", select the peak file as input `nrsf-chr21 (Reads) (Peaks, Annotated)` ( ) (figure 16).
3. In the next step, "Set parameters", select the `NC_000021 (Genome)` ( ) track as **Reference sequence track** (figure 17).

   Click on the green plus icon ( ) at the right side of **Annotation types** and select **Peak**.

   Under "Naming of result sequences", check the **Include annotation region** and **Include annotation chromosome** options to give informative names to the resulting sequences.
4. In the next step, "Result handling", choose to **Save** the sequences.
5. In the last step, "Save location for new elements", select a suitable location in the Navigation Area to save the results and click on **Finish**.

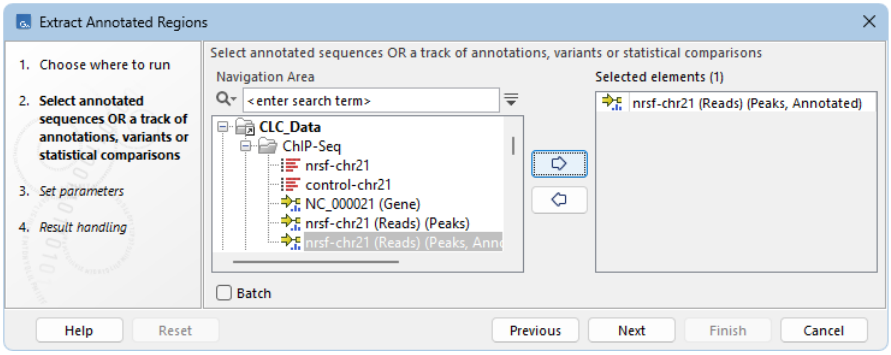

Figure 16: *Select the annotation track.*

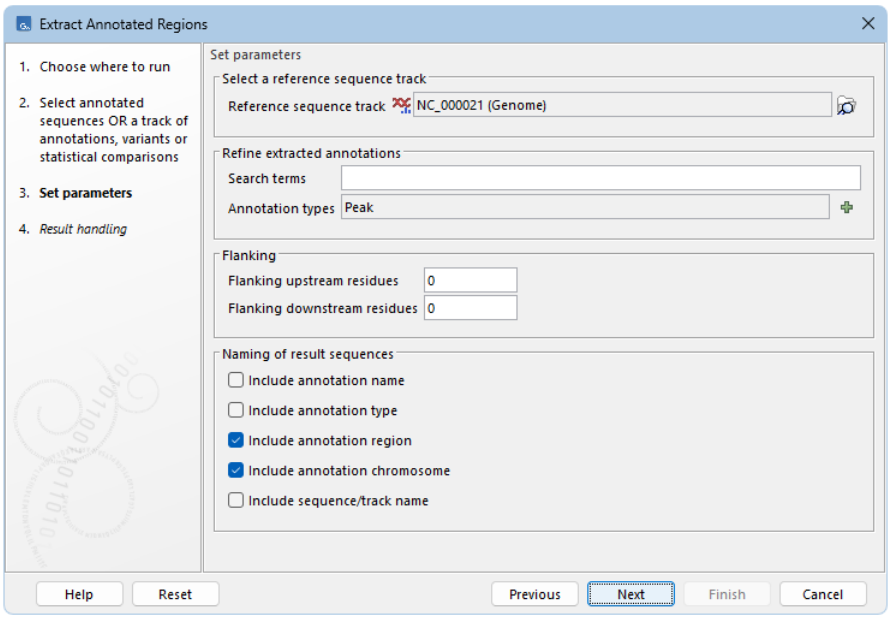

Figure 17: *Options for the Extract Annotated Regions tool.*

After a few seconds, the sequences will be saved to a file named `Extracted annotations` ( ).

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, then choose the **fasta** format (figure 18), select the file `Extracted annotations` ( ) and finally select the output file name.

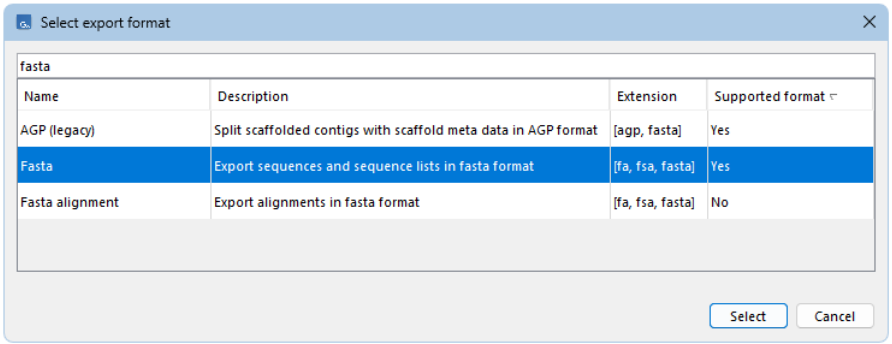

Figure 18: *Options for the Export tool.*

Bibliography

[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.

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