Tutorial

ChIP Sequencing

November 21, 2017

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com ·
AdvancedGenomicsSupport@qiagen.com

Tutorial

ChIP Sequencing

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, such as the ones occuring in genomic regions that are less accessible, repeated regions or copy number aberrations.

This tutorial takes you through a complete ChIP sequencing workflow using *CLC Genomics Workbench* 10.You can also use Biomedical Genomics Workbench to go through most of the steps included in the tutorial. We will focus on how to run the analysis but we will not go through the technical details of how the Transcription Factor ChIP-Seq analysis is implemented. Click the **Help** button in the tool dialogs (see below) to read more about the tools used in the tutorial.

The workflow consists of five parts:

- Importing the raw sequencing data.
- Mapping the reads to a reference genome.
- Calling peaks.
- Visualizing the results.

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. The complete datasets can be found at the UCSC website.

Importing the raw sequencing data

1. First, download the data set from our website: `http://resources.qiagenbioinformatics.com/testdata/raw_data/ChIP-seq_NRSF_chr21_bx.zip`. Unzip the file somewhere on your computer (on the Desktop for example).
2. Start the *CLC Genomics Workbench* and import the sequencing data:

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

This will bring up the dialog shown in figure 1:

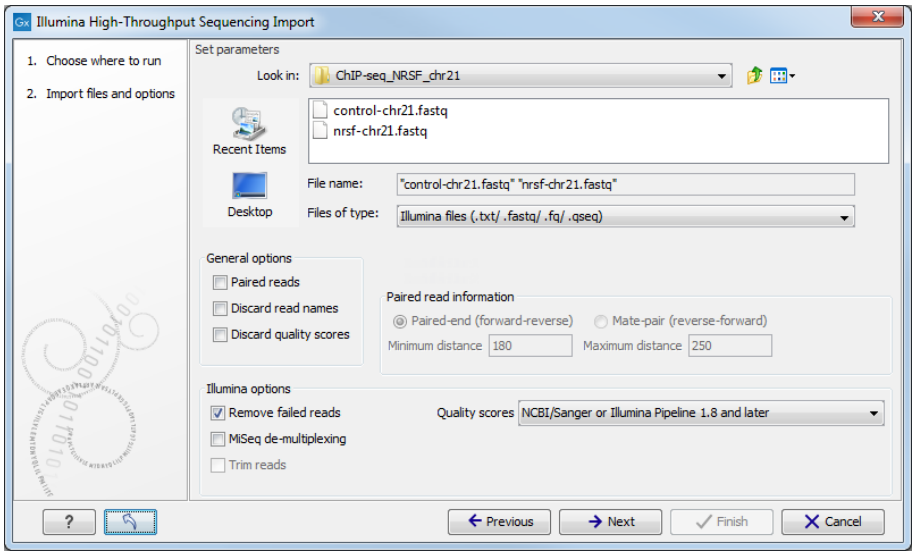

Figure 1: *Import raw reads. When analyzing your own data, you should select the sequencing technology appropriate for your data. This dataset consists of two fastq files obtained using an Illumina sequencer, so the Illumina importer should be chosen.*

3. Select the `nrsf-chr21.fastq` and `control-chr21.fastq` files. Make sure the **Paired reads** checkbox is not checked, as the options to discard read names and quality scores (These last options are not significant in this context because of the relatively small amount of reads). Click **Next**, **Save** the imported reads list in a folder created for this tutorial and click **Finish**.

4. Next, import the reference genome sequence that was also included in the zip file in clc format: drag and drop in the Navigation Area the files `NC_000021 (Genome).clc` and `NC_000021 (Gene).clc`, which are the genomic chromosome 21 reference sequence and the gene annotation track for chromosome 21, respectively. You can also use the Standard Import tool with the option **Automatic import**:

   Import ( ) | **Standard Import ( )**

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. Go to:

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

2. The first wizard window 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` (see figure 2). Click on the button labeled **Next**

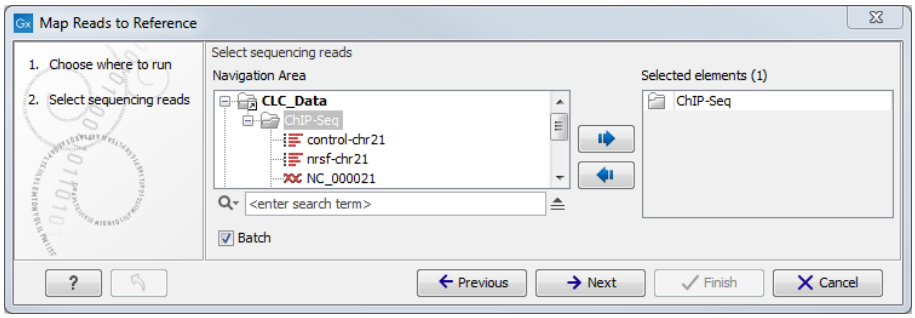

Figure 2: *Select sequence list containing the reads. Since we want to map two lists, we choose the batch mode.*

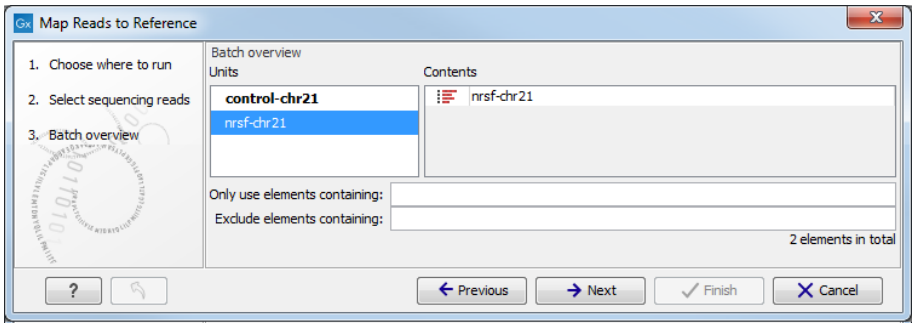

Figure 3: *Check that all reads are used as input for the mapping.*

3. In the next window, check that only the two lists `control-chr21` ( ) and `nrsf-chr21` ( ) are selected (figure 3). Click **Next**.

4. You can now select the reference sequence `NC_000021 (Genome)\icon{clcobjects/sequenc` as shown in figure 4. Click **Next**.

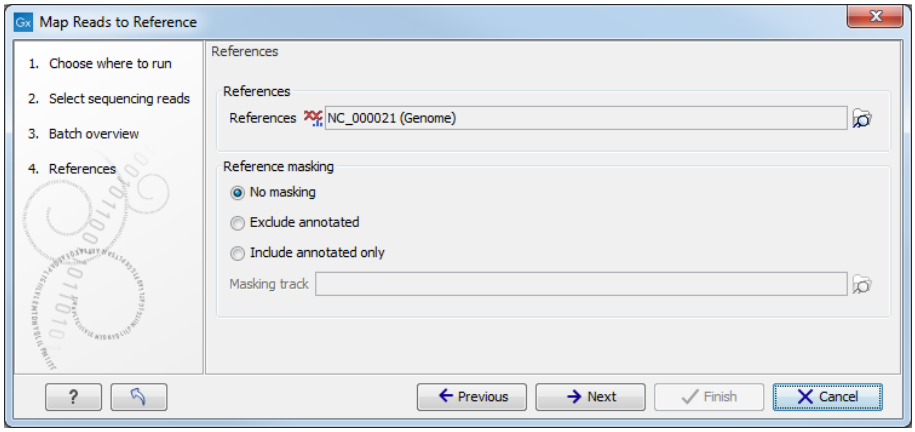

Figure 4: *Specifying the reference sequences and masking parameters.*

5. Leave mapping options as default (figure 5). You can use the Reset button if you are not sure whether you have previously changed the parameters for the tool. Then select to ignore the non-specific matches and click **Next**.

6. For CLC Genomics Workbench users, the dialog shown in figure 6 appears in the next window. Select **Create reads track** to create track-based results (it is the default option in Biomedical Genomics Workbench). Check **Create report** to obtain a detailed report about the read mapping and leave **Collect un-mapped reads** unchecked since we are not

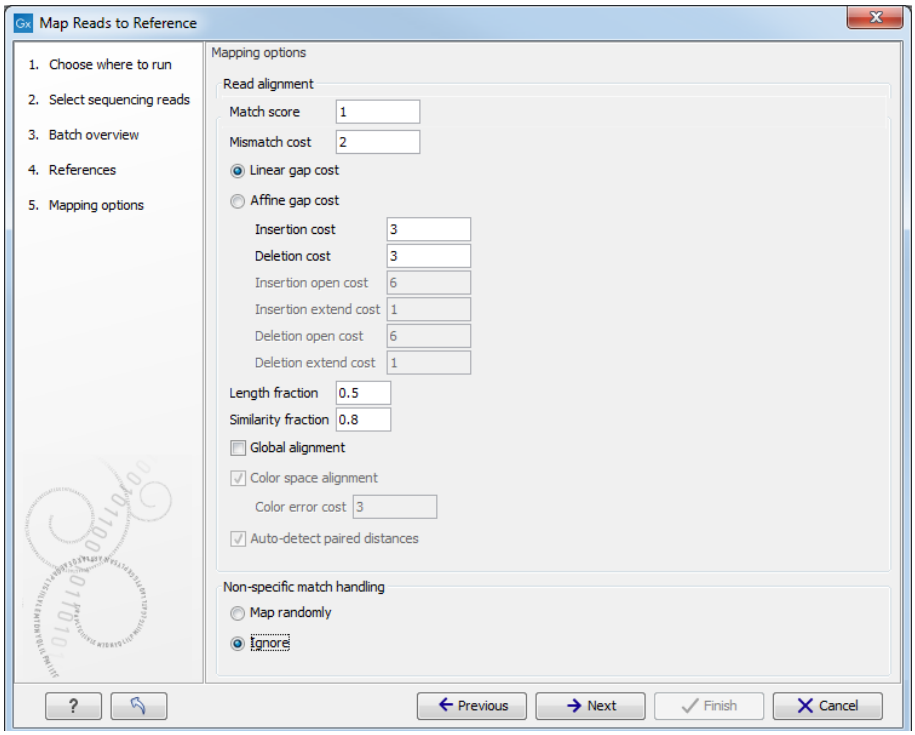

Figure 5: *A stringent read matching is desired for ChIP-Seq.*

interested in those reads.

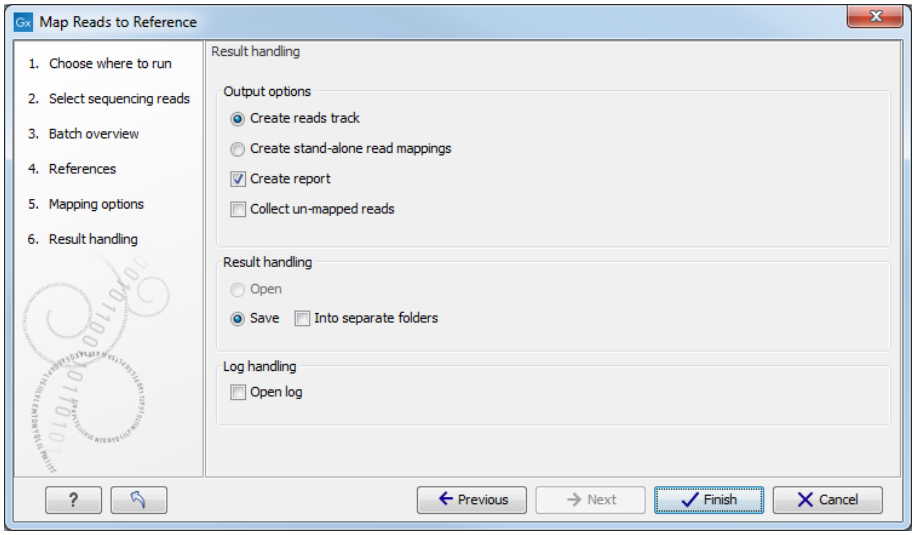

Figure 6: *Select Create reads track, Create report, and Save.*

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.

1. Go to:

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

2. This opens a dialog where you select the `nrsf-chr21 (Reads)` ( ) and click **Next** (see figure 7).

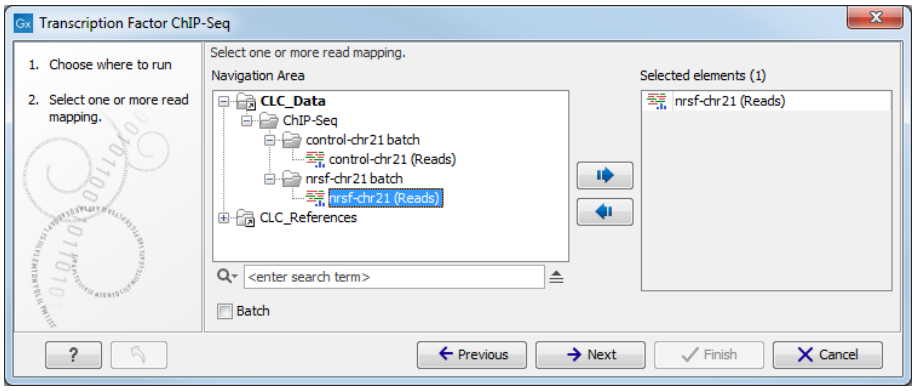

Figure 7: *Select the input reads.*

3. You can now choose `control-chr21 (Reads)` ( ) as control data (see figure 8). You can leave the **Maximum P-value for peak calling** to the default value of 0.1. 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.

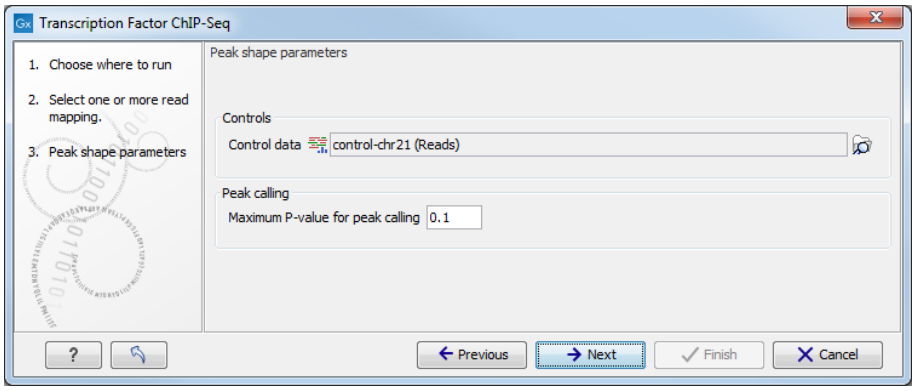

Figure 8: *Choose control data.*

4. You can now choose the output data to be generated (see figure 9). 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.

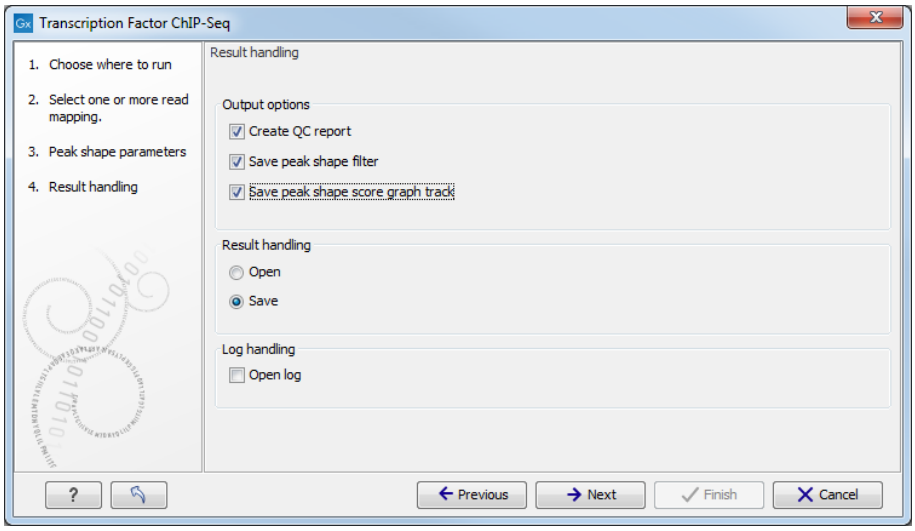

Figure 9: *Select the output data to be generated.*

- `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**. 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 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. However, in 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 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.

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	1.009	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	2.483	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	1.191	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	2.376	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.*

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:

1. Go to: **Toolbox** | **Epigenomics Analysis (**) | **Annotate with Nearby Gene Information (**)
2. Select first the track to annotate (`nrsf-chr21 (Reads) (Peaks)` ( )) (figure 12), and click **Next**.
3. In the dialog shown in figure 13 choose `NC_000021 (Gene)` ( ) as the reference gene track and click **Next**.
4. **Save** the result. The file `nrsf-chr21 (Reads) (Peaks, Annotated)` ( ) will be generated.

Visualizing the results

The best way to visualize the results is to create a Track List (called Genome Browser View in Biomedical Genomics Workbench):

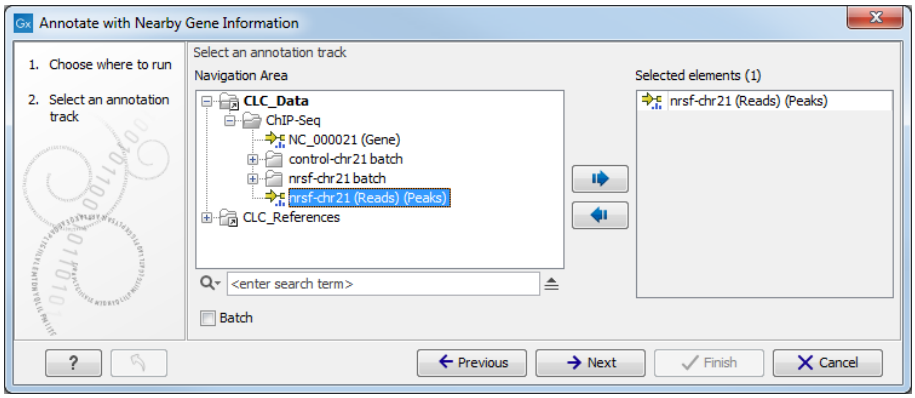

Figure 12: *Select the track to annotate.*

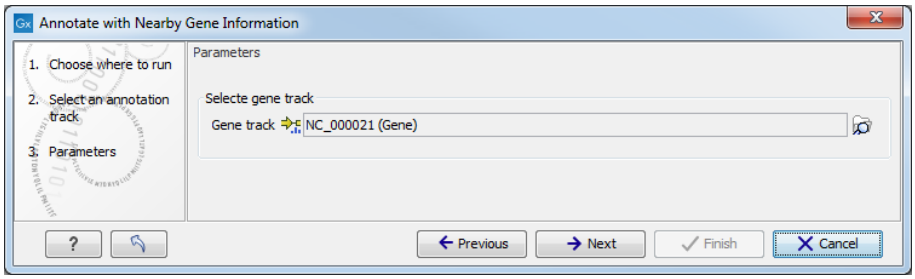

Figure 13: *Select the annotation track to be used as gene reference.*

New | **Track List** or **Genome Browser View** ( )

Select the tracks we created so far as shown in figure 14 and then press **Finish**.

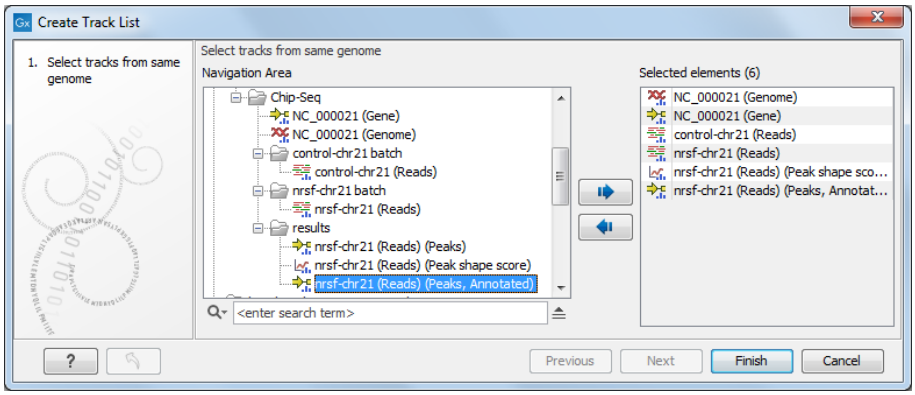

Figure 14: *Create a Track List (called Genome Browser View in Biomedical Genomics Workbench) to visualize the results.*

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)`. 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 highlighted in figure 15 to zoom in on the peak of interest, and zoom out as needed to see the closest gene. You can browse through all the 144 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 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

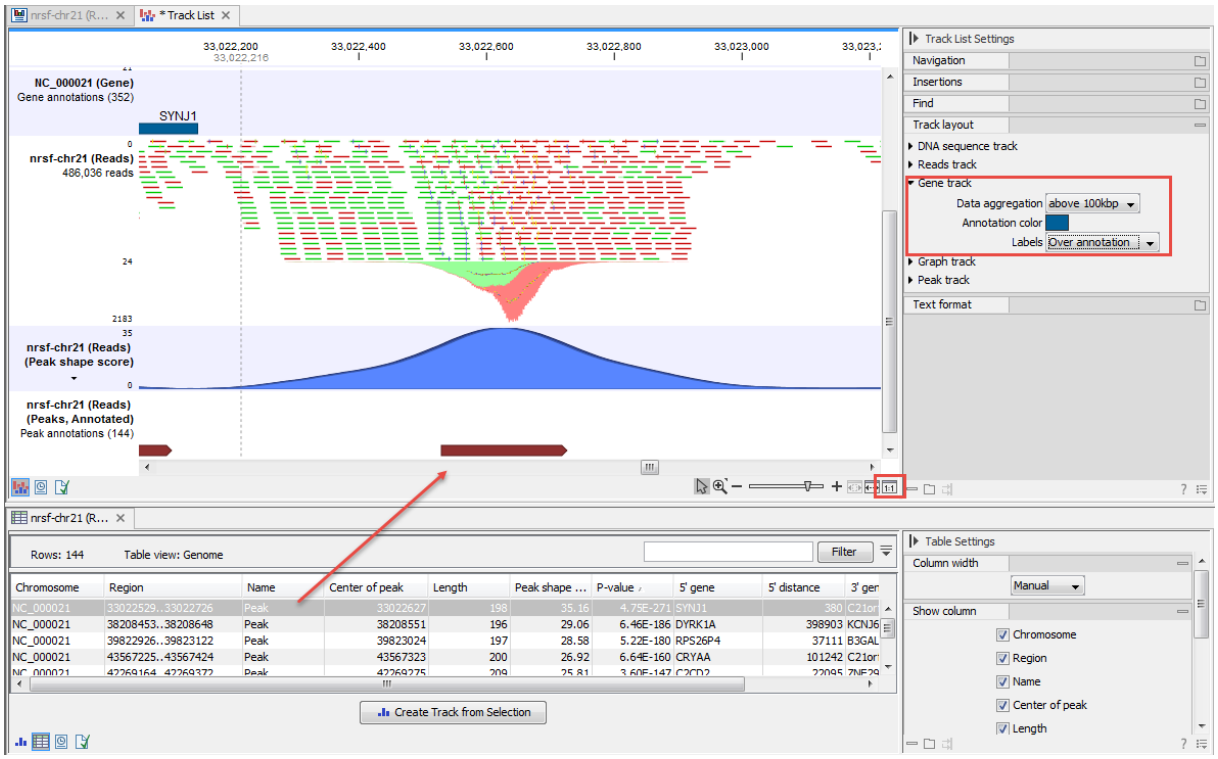

Figure 15: *A very strong peak near the gene SYNJ1.*

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.

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.