ChIP-Seq analysis

Introduction

This tutorial describes the ChIP-Seq analysis workflow in Avadis NGS using a transcription faction regulation study. It assumes that Avadis NGS is installed and that the steps listed in the 'Getting Started' tutorial have been completed.

Dataset

The aim of this ChIP-Seq study is to find the regulatory role of NRSF/REST. The details of the study can be found at http://www.sciencemag.org/content/316/5830/1497.abstract. Briefly, a Control sample and a ChIP sample were sequenced using the Illumina GaIIx machine, to generate single-end reads with read length 25bp. The dataset zip file (`chipseq-small-dataset-illumina.zip`) contains the reads from chr1 of the "`_results.txt`" files from these samples. For the purpose of this tutorial, we will consider data that aligned to chr1 of the genome.

Goals

By comparing the coverage patterns in the samples, we want to find which genes are being regulated by the NRSF TF. This translates to finding the peaks in the ChIP sample against the Control sample, any significant motifs in these binding sites, and the genes regulated by the binding sites.

Sample Import

The samples were aligned against the hg18 assembly. The ChIP-Seq analysis experiment can be created (from the Project → New Experiment menu) with RefSeq annotations and other parameters as shown in the image below.

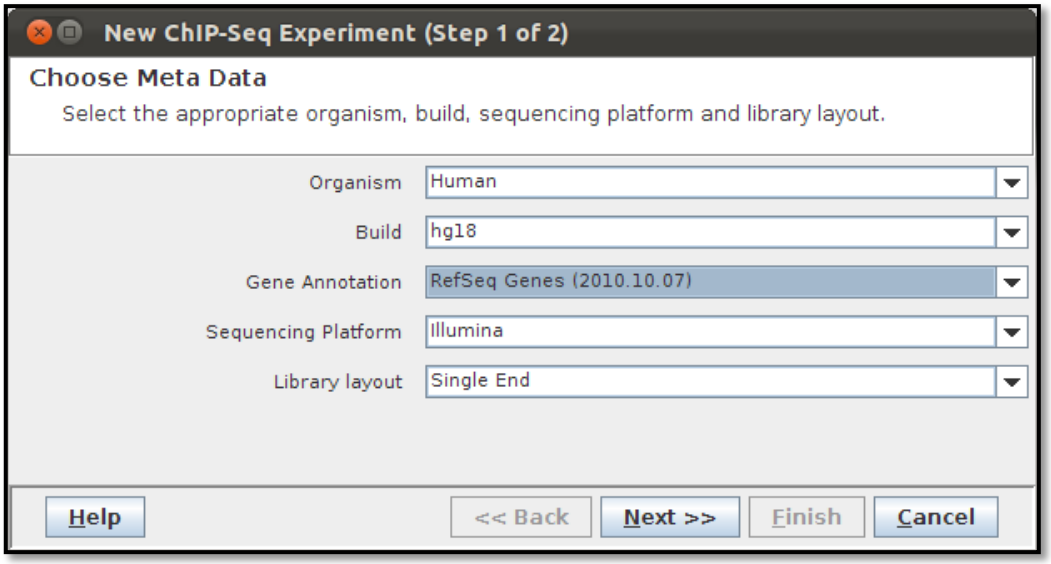


Figure 1: Experiment creation parameters

While loading the sample files in the next page of the wizard, the default sample names can be changed to chip1862 and mock1862, for brevity.

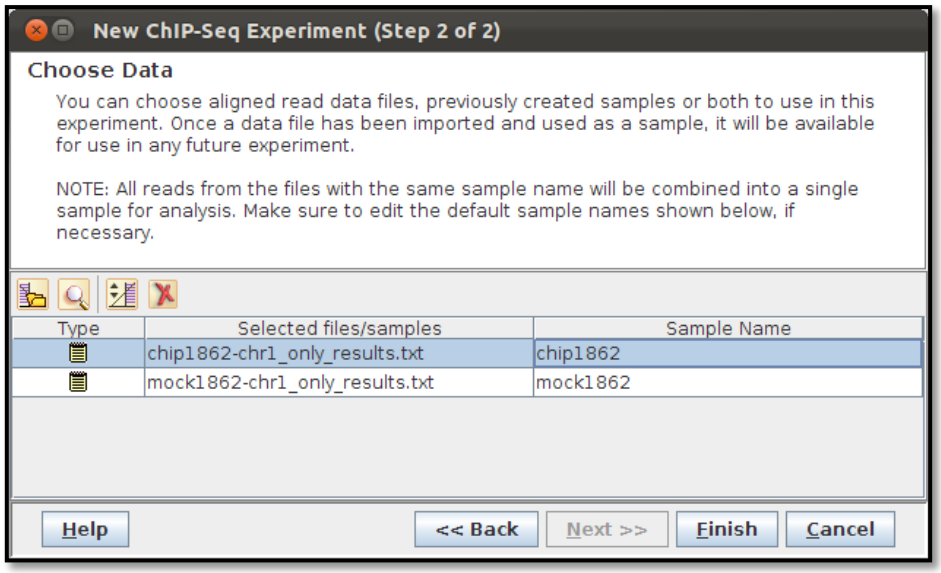


Figure 2: Sample loading

Once the experiment creation is finished, we can double-click on the 'All Aligned Reads' object in the experiment navigator on the LHS to check that there are 153,551 reads in chr1 of the chip1862, and 204,689 reads in mock1862. In addition, the genome browser view is also launched. By default, the genome browser shows one track each for the two samples, and one track with the RefSeq transcript annotations.

Figure 3: Genome browser view of the reads list

Quality Control and Filtering

In ChIP-Seq analysis, duplicate reads removal is the most commonly used filter. When this filter is invoked (from the Filters → Filter on Duplicates in the workflow navigator) with the default cut-offs,

duplicate reads that are the artefact of PCR will be removed, and we will be left with 143,263 reads in the chip1862 sample and 202,224 reads in the mock1862 sample.

Peak Calling

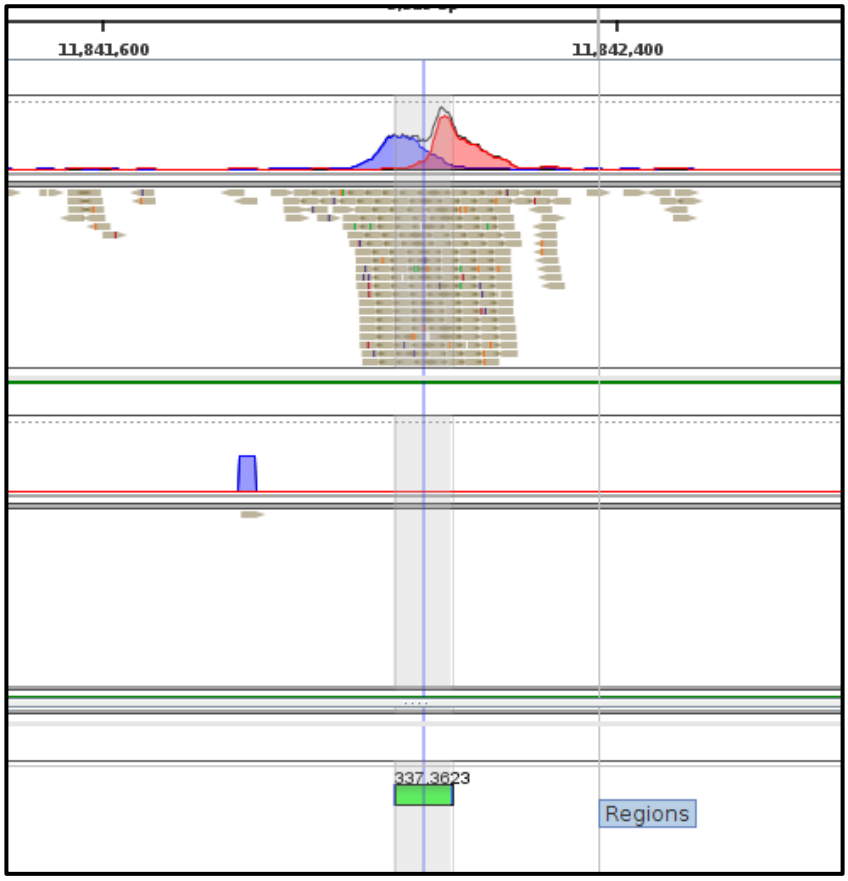
To find the transcription factor binding sites using peak detection, Avadis includes three algorithms:

1. Enriched Region Detection
2. Probabilistic Inference for ChIP-Seq (PICS)
3. Model based Analysis for ChIP-Seq (MACS)

For the current tutorial, we will run the PICS algorithm (from the Analysis→ Peak Detection workflow step) on the filtered read list with the default parameters. PICS is more commonly used for transcription factor binding, while MACS is used for histone modification. The Enriched Region Detection algorithm is a quick and dirty algorithm to identify enriched sites.

Explore

We can explore and inspect the peaks predicted by PICS by dragging and dropping the output of PICS from the navigator on the LHS into the Genome Browser. We can use the next arrow in the PICS track to navigate from peak to peak or we can also right click on the output list in the experiment and use the 'Navigate in Genome Browser' functionality.



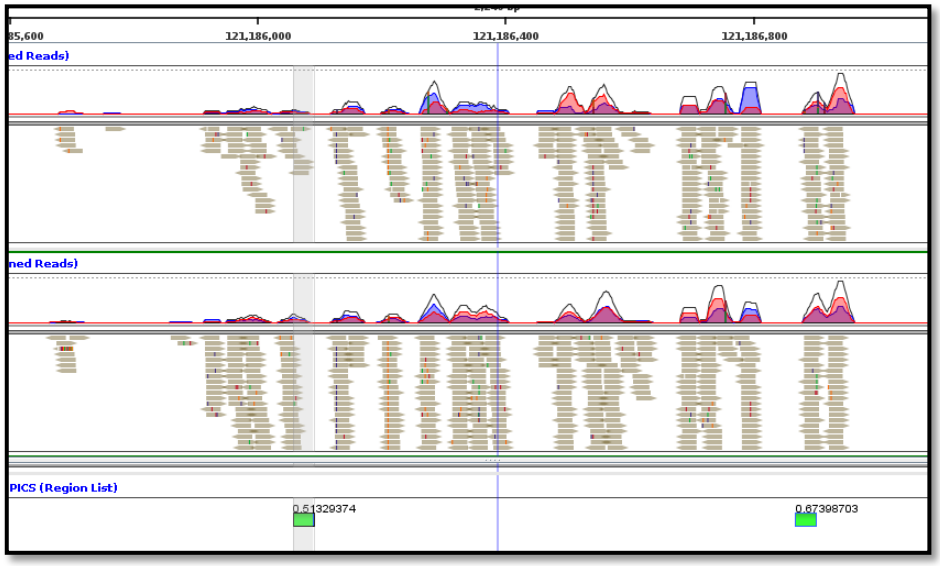


Figure 5: Low Quality Peaks

Downstream analysis

We can find genes in the vicinity of these binding sites (using the Results Interpretation → Translate regions to genes workflow step). For the current dataset, we would find 81 genes within 1000bp of the binding sites.

We can also find out the GO terms enriched by these 81 genes (using the Results Interpretation → GO analysis workflow step). Ion channel activity and synapse related terms would appear among the significant GO terms in this dataset.

Motif Detection using GADEM

To see why the transcription factor bound to these regions, we can analyse the DNA sequence at these locations to see if there is a dominant motif(s). We can run the GADEM Motif detection algorithm (from the Analysis → Motif Detection-GADEM workflow step) to identify the top few motifs.

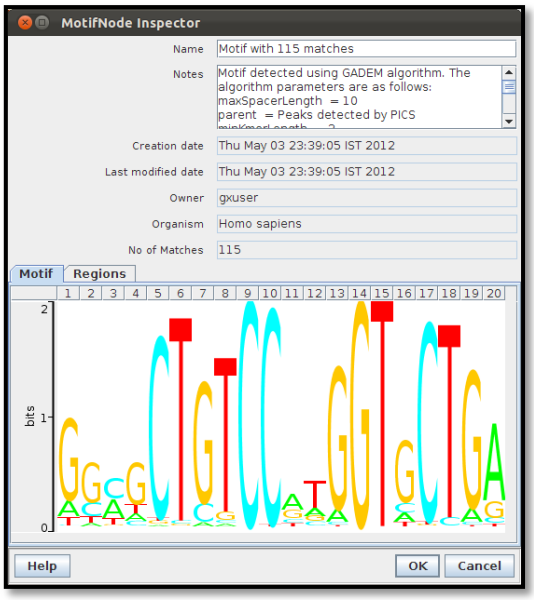
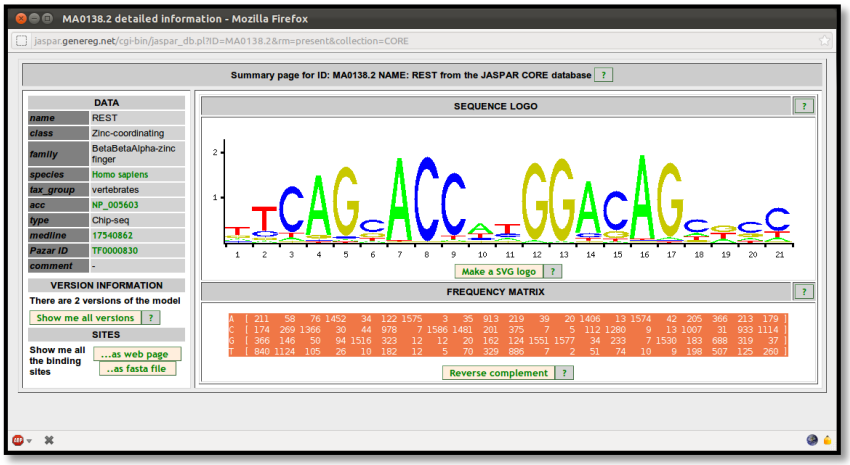
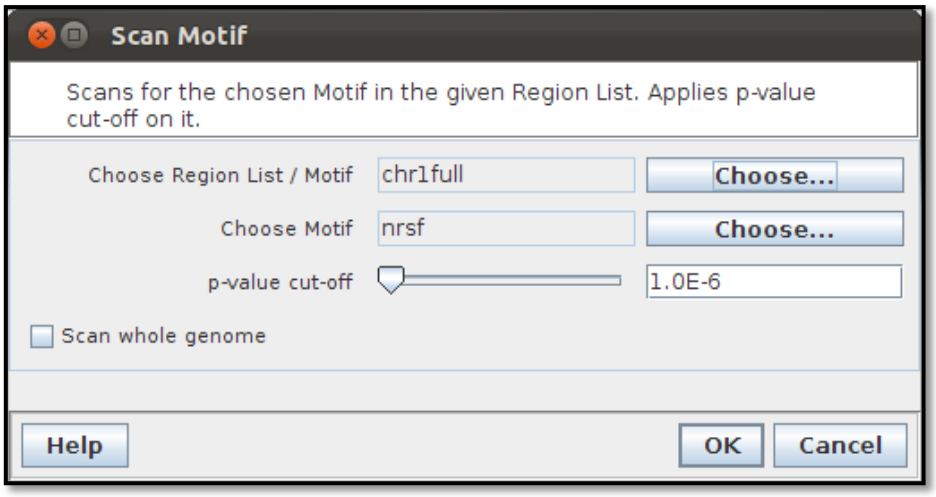


Figure 6: Motif node inspector

We can also search the JASPAR database (http://jaspar.binf.ku.dk/cgi-bin/jaspar_db.pl?rm=browse&db=core&tax_group=vertebrates) for the NRSF/REST transcription factor motif to see if it matches the one detected by GADEM.



The motif shown in the JASPAR database has been downloaded as a text file (called `nrsf.JASPAR` in `chipseq-small-dataset-illumina.zip`). The motif can be imported into Avadis (using the Utilities → Import Motif workflow step). After that, we can scan chr1 to find all occurrences of the motif.



Note that we are limiting the scan to chr1, by specifying a region list chr1full (which is nothing more than a single region encompassing the whole chromosome). This is also provided as a .bed file in the zipped folder (`chr1full.bed`). This can be imported using the 'Import Region List' functionality under 'Utilities'.

The motif occurrences would number about 1605. All of these could be considered potential binding sites. PICS/MACS give you the actual binding sites under the specific experimental conditions, but the scan reveals what the TF could do under different circumstances.

The samples in the tutorial have been modified to contain only the reads from chr1. The whole genome ChIP-seq samples are also included in the `chipseq-full-dataset-illumina.zip` file, should you want to try the analysis on the whole dataset.

This is a very brief overview of the ChIP-Seq experiment in Avadis NGS. For more details or clarifications, please revert back (sales@avadisngs.com or support@avadisngs.com) and we will address your queries.