ABGP 2017

ChIP-seq homework.

Robert Trumbly

Provide answers for all the questions in bold face.

This homework assignment will follow the tutorial protocol you received along with the lecture:

“Chip-seq Analysis with Galaxy: from reads to peaks (and motifs)”.

For most of the steps, you can simply follow the tutorial, which was demonstrated in the chip-seq lecture. Any changes will be noted. Some of the steps on Galaxy may take a long time to run, depending on how busy the server is. Be prepared to start a task, and come back later when the task should be completed. If you have problems, please contact me.

The procedure will be as in the lecture and tutorial except that a different dataset from the same experiments will be used.

Step 2-Obtaining the raw data.

Since the data are not kept at ArrayExpress, go directly to the ENA database (European Nucleotide Archive): <http://www.ebi.ac.uk/ena/>.

In the query box, enter the study accession number, SRP000217, and click Search. This study contains all the chip-seq data from the Chen 2008 paper.

On the first page of the study, there is a list of chip-seq datasets. To get to the one we want, click on next at the bottom of the page twice to get to the third page. Near the middle of the page, see SRR002011, in the Run accession column, and click on it. In the description you will see “Illumina sequencing of Mouse ES\_nanog genomic fragment library”, indicating it was a chip-seq experiment using antibodies against the nanog TF. Follow the tutorial’s method for uploading the file, but substitute SRR002011 for SRR002012 wherever it appears.

Step 3. Continue to follow the tutorial, but substituting SRR002011 for SRR002012 wherever it appears Upload the reads in the Galaxy server. Where the tutorial says, Get Data > Upload File, on Galaxy it now says “Get Data >Upload File from your computer”. In the popup window, choose Paste/Fetch data, and enter the URL in the window.

The following is an alternate way that is simpler, and has worked previously, but now has problems: “From the SRR002011 page at ENA, near the bottom of the page you will see Fastq files (galaxy). If you click on file1 below that, it will send the file to Galaxy, and open an instance of Galaxy on your browser. It would be a good idea to log in, to save your work. If you upload the file this way, you will have to specify the attributes of the file after uploading rather than before uploading as was done in the tutorial. In the box with the name of the uploaded file, there are three icons: an eye, a pencil, and an X. Click on the pencil icon to edit attributes. For database build, choose Mouse July 2007 mm9, then Save. Under the datatype tab at the top of the edit attributes page, choose fastqsanger, then save.” The problem with this method at present is that the file remains in the compressed .gz format, and cannot be read by Bowtie. For that reason, use the method described in the tutorial.

**How big is the file?**

Step 4 - Some statistics on the raw data. Note: the NGS tools are no longer under the NGS Toolbox Beta menu as described in the tutorial, they are now listed in the main menu.

Under the NGS: QC and manipulation menu, click on Compute quality statistics, then execute. There may be a warning “This is a new dataset and not all of its data are available yet”, but this step should work anyway.

**Analyze your results by clicking on the eye icon in the green box: how many lines are there in the file?**

**How many reads are there in the file (check the column count)?**

Step 5 - Mapping the reads with Bowtie

Follow all the steps in this section as written. In part 11, the full name of the tool to use is Filter SAM on bitwise flag values. In this tool, chose “insert flag” instead of “add a new flag” as in the tutorial.

**How many lines are there in this final file from Filter SAM?**

**This represents the number of mapped reads. Calculate the percentage of mapped reads for this experiment, the number of mapped reads divided by the total reads.**

Step 6 - Peak calling with MACS

Follow all the steps in the tutorial, except: Tag size should be set to 26, not 26bp.

**How many peaks ("regions") were detected by MACS ?**

Step 7 - Retrieving the peak sequences **(OMIT)**

Step 8 - Visualize the peak regions in UCSC genome browser

Go to the UCSC browser from the MACS bed file results. The initial browser window should be mouse chromosome 1, with many peaks in the MACS track at the top. To see a specific gene, enter the gene name Pou5f1, click Go. **Are there any MACS peaks in this window?**

Expand the window by clicking on zoom out 3X.

**Are there any MACS peaks visible now?**

**How many?**

**What is the position of the peak(s) in relation to the Pou5f1 gene (upstream, downstream, etc.)?**

**Do any of the peak(s) correspond to the nanog peaks found in the Chen 2008 paper (see Fig 1).**

**What do you think could account for any differences in your and their results?**