

ChIP-seq practical: peak detection and peak annotation

Mali Salmon-Divon and Christiana Spyrou

November 26th, 2009

Introduction

The goal of this hands-on session is to perform some basic tasks in the analysis of ChIP-seq data. We will find immuno-enriched areas using two peak calling methods, MACS and BayesPeak. We will visualize the data in a genome browser and perform annotation and motif analysis on the predicted binding regions.

Example Data

We will use two different data sets in this practical, which can be found at `/g/chrplas/course/ChIP-Seq`

Please connect to the EMBL cluster as follows

```
ssh -X studentNN@clnodeNNN
```

using student1 - student16, clnode143 - clnode147 and the password `htse2009`.

Then type

```
bash
```

```
source /g/huber/software/path.bash
```

```
gnome-terminal &
```

Now go to the folder you created yesterday, by typing

```
cd /g/chrplas/course/users
```

```
cd "your folder"
```

and link all the files we will be using with the command

```
ln -s ../../ChIP-Seq/* .
```

The corresponding files are the following:

1. `oct4.mapview` and `gfp.mapview`
These files are based on Oct4 ChIP-seq data published by Chen et al. 2008. These are Maq output files of reads aligned to the mouse chromosome 1. `oct4.bowtie` contains reads from ChIP sample and `gfp.bowtie` contains reads from the control GFP mock-IP. You will identify the Oct4 peaks using MACS. Afterwards you will annotate the enriched regions and look for presence of the Oct4 motif in these regions.
2. `H3K4me3-chr16.5M.bed` and `Input-chr16.5M.bed`
These are H3K4me3 data taken from Wilson et al. 2008. Today we will analyze a sample region from chromosome 16. The two `bed` files contain reads from H3K4me3 ChIP and Input DNA that were mapped to chromosome 16 of the mouse genome. You will use both MACS and BayesPeak to identify peaks and then check for agreement between them. Then you will view these enriched regions on a genome browser and finally split the broader peak regions into individual nucleosomes.

Oct4 analysis

Find enriched areas using MACS

MACS - <http://liulab.dfci.harvard.edu/MACS/index.html>

1. The input for MACS can be in ELAND, BED, SAM, BAM and BOWTIE formats (you just have to set the `--format` flag). Since the tool does not support Maq `mapview` format, you will have to convert the `mapview` files to `bed` format. The `bed` format is simple, it contains the chromosome number, start and end position of each read, and also information on which DNA strand it maps to.

More information about `bed` format can be found at <http://genome.ucsc.edu/FAQ/FAQformat#format1>

The strand sign should appear in the 6th column of the file, hence you will have to fill columns 4 and 5 with some text such as `“*”`. Also note that Maq chromosomal locations are 1-based (the first base of a chromosome is numbered 1), whereas `bed` files are 0-based, so when you convert the `mapview` format to `bed`, you have to remove 1 from the start locations. You can use the command `awk` to do the conversion, i.e.

```
awk '{print $2 "\t" ($3-1) "\t" ($3+25) "\t" "*" "\t" "*" "\t" $4}'
oct4.mapview > oct4.bed
```

```
awk '{print $2 "\t" ($3-1) "\t" ($3+25) "\t" "*" "\t" "*" "\t" $4}'
gfp.mapview > gfp.bed
```

Type `head oct4.bed` to see how this file looks like

2. Now you are ready to use MACS. Type MACS in order to see the various parameters. Those you will need to use include:

`-t` = to indicate the input ChIP file

`-c` = to indicate the name of the control file

`--format` = to change the file format. The default format is `bed`, thus if you choose to input Bowtie files, you have to set this flag accordingly.

`--name` = to set the name of the output files

`--gsize` = to define the mappable mouse genome size. With a read length of 26 bases, 70% of the genome length is a fair estimation. Since in this analysis we include only reads from chromosome 1, we will use as `gsize` 70% of the chromosomes length (197 Mb).

`--tsize` = to set the read length (look at the `fastq` or `mapview` files to check the length)

`--mfold` = to set the fold-ratio of reads in enriched and background areas, in order to select some initial peaks and build the model to calculate the read-shift size. This is not involved in the final peak detection.

The default value of 32 might be high, and you will have to decrease this number until a reasonable model is built. Usually it is not recommended to set `mfold` to be less than 10. In cases when even `mfold=10` cannot give a good model, it is recommended to use the `--nomodel` option, set the `shiftsize` to be half the fragment length and `--bw` to be the fragment length.

For this practical, you will set the `mfold` to be 5 (just for practice). Be aware that for this data set, the modelled `d` (distance between forward and reverse strand peaks) should be around 120 bases. Once the model is built, you can view it to check whether it looks reasonable. A PDF file of the model can be

generated using the command

```
R --vanilla < NAME_model.r
```

where NAME_model.r is one of the output files generated by MACS.

You can view the model that has been generated based on the full data set (reads from the whole genome and not only chromosome 1) at

<http://www.ebi.ac.uk/mali/Data/ChIP-seq>.

This is shown in the file `Oct4_model.pdf`.

`--wig` = to generate signal wig files for viewing in a genome browser. Since this process is time consuming, it is recommended to run MACS first with this flag off, and once you decide on the values of the parameters (such as `mfold`), run MACS again with this flag on.

`--diag` = to generate a saturation table, which gives an indication whether the sequenced reads give a reliable representation of the possible peaks.

Now run macs using the following command:

```
macs -t oct4.bed -c gfp.bed --name=oct4 --gsize=137900000 --tsize=26  
--mfold=5 --diag --wig
```

3. By looking at the output saturation table (`oct4_diag.xls`) would you think that more sequencing is necessary?

Open the Excel peak file and view the peak details. Note that the number of tags (column 6) refers to the number of reads in the whole peak region and not the summit height.

Peak Annotation

In order to biologically interpret the results of ChIP-seq experiments, we need to consider the genes and other annotated elements that are located in the proximity of the identified enriched regions. This can be easily done using PeakAnalyzer.

Launch the program by typing `PeakAnalyzer`.

To install it on your local computer, you can download it from

<http://www.ebi.ac.uk/bertone/software>

and launch it by double clicking on the `PeakAnalyzer.jar` file, or by typing `java -jar PeakAnalyzer.jar` at the terminal.

The first window allows you to choose between split application and peak annotation. First choose the peak annotation option and click Next.

We would like to find the closest downstream genes to each peak and the genes that overlap with the peak region. For that purpose you should choose the NDG option and click next.

Fill in the location of the peak file `oct4_peaks.bed`, and choose the mouse GTF as the annotation file. You do not need to define a symbol file since gene symbols are included in the GTF file. Choose the output directory and run the program.

When the program has finished, you will have the option of generating plots.

(You can do this if R is installed on your computer. Otherwise, if you do not want to install R, you can generate similar plots with Excel using the output files that were generated by PeakAnalyzer.)

A pdf file with the plots will be generated in the output folder.

Motif analysis

We will use MEME for motif analysis. The input for MEME should be a file in **fasta** format containing the sequences of interest. In our case, these are the sequences of the identified peaks that probably contain Oct4 binding sites. Since many peak-finding tools merge overlapping areas of enrichment, the resulting peaks tend to be much larger than the actual binding sites. Sub-dividing the enriched areas by accurately partitioning enriched loci into a finer-resolution set of individual binding sites, enhances the quality of the motif analysis. We will view broad peak regions in a genome browser later on. In addition, binding motifs are most likely to appear at or near sub-peak summit regions, and these sequences can be retrieved directly from the Ensembl database using PeakAnalyzer.

1. Running PeakAnalyzer

If you have closed the PeakAnalyzer running window, open it again by typing **PeakAnalyzer**. If it is still open, just go back to the first window. Choose the “split peaks” utility and click Next.

The input consists of files generated by most peak-finding tools: a file containing the chromosome, start and end locations of the enriched regions, and a **.wig** signal file describing the size and shape of each peak. Before you upload the **oct4_peaks.bed** file, open it with

```
nedit oct4_peaks.bed
```

and delete the header line, without leaving the line empty.

Then, fill in the location of both files **oct4_peaks.bed** and the **wig** file generated by MACS, which is under “oct4_MACS.wiggle/treat” folder, check the option to fetch sequences and click Next.

In the next window you have to choose some parameters for splitting the peaks. We will explain what these parameters are soon. For now, just change the organism name from the default human to mouse and run the program.

Since the program has to read large **wig** files, it will take a few minutes to run. Once the run is finished, two output files will be produced. The first describes the location of the sub-peaks, and the second is a **fasta** file containing 300 sequences of length 61 bases, taken from the summit regions of the highest sub-peaks.

2. Running MEME

Open the “opera” browser and go to the MEME website at <http://meme.sdsc.edu/meme4.3.0/cgi-bin/meme.cgi>, and fill in the necessary details, such as:

- your e.mail address
- the sub-peaks **fasta** file **oct4_peaks.bestSubPeaks.fa** (will need uploading)
- the number of sequences we expect to see (“0 or 1” per sequence)

- the width of the desired motif (between 6 to 20)
- the maximum number of motifs to find (3 by default). For Oct4 one classical motif is known.

While we are waiting for the MEME results, let's start the analysis of the histone modification data.

H3K4me3 ChIP-seq analysis

In this part of the practical, we will find regions that are enriched with a trimethylation modification on Lysine 4 of histone H3 (H3K4me3). The `bed` files contain reads that were mapped to the mouse chromosome 16.

We will use two peak callers, BayesPeak and MACS, that use two different approaches for finding peaks.

You are already familiar with MACS from the previous part of this practical, so please run it on this dataset, using

```
macs -t H3K4me3-chr16.5M.bed -c Input-chr16.5M.bed --nomodel --shiftsize=75
--bw 150 --name=H3K4me3 --gsize=5000000 --tsize=36 --wig
```

BayesPeak

BayesPeak is a peak-calling tool that is currently being converted into an R package and will soon be available to download from <http://www.compbio.group.cam.ac.uk/Resources/BayesPeak>.

We have already installed a preliminary version of it for you to use.

Open a session in R from the terminal by typing `R` and load the BayesPeak package by typing

```
> library(bayespeak)
```

Next you can use the help function `?` to find out about each command. For example,

```
> ?bayespeak
```

gives you some information on the function(s), and so does

```
> ?combine.peaks
```

You can close the help window by typing `q`.

Run the program by typing

```
> enriched.windows = bayespeak("H3K4me3-chr16.5M.bed", "Input-chr16.5M.bed",
chr = "chr16", start = 90E6, end = 95E6)
```

The program takes a few minutes to run. When it finishes, you will get a new prompt `>`.

This will give you a large set of genomic windows that the program identifies as enriched, as well as the posterior probabilities with which they were detected. These can be combined to produce the final regions of enrichment using the command

```
> final.peaks = combine.peaks(enriched.windows)
```

Now the `final.peaks` object contains the regions of enrichment. You can save these peaks to a file by using the command

```
> write.table(final.peaks, file = "bpeaks.txt", quote = FALSE, row.names = FALSE, sep = "\t").
```

Now lets close the R session by typing

```
> q()
```

(you dont have to save the workspace).

How many significant enrichment areas did MACS find? How many were found using BayesPeak? You can find this out using the unix command

```
wc H3K4me3_peaks.bed
```

Lets now find the number of enrichment locations found by both programs. This can be done using PeakAnalyzer. Run PeakAnalyzer and choose the “peak annotation” utility. Next, choose the “ODS” option to find overlap between two peak files. Upload the two peak files and run the program.

The output of “ODS” consists of three tab-delimited files. One of them is the “overlap” file. Each line in this file describes a region that is common in the two peak-files (a peak region from one data set can overlap with many regions in the other data set). The other two files (one for each input peak file) describe the peaks that are unique to each identified set.

How many regions identified by BayesPeak (the file that was uploaded first to the “ODS” utility) overlap with those that MACS calls? You can find that using the command

```
cut -f 1,2,3 "overlap_file_name" | sort | uniq |wc
```

This command sorts the first three columns of the “overlap” file and then counts how many unique lines there are. Since the “overlap file” contains one header line, you need to subtract 1 from the “number of overlap peaks” you got.

As you have probably noticed, many peaks found by MACS were not found by BayesPeak. Why is that? Are these false positives? Low modification level?

Lets view some of these peaks in the UCSC genome browser. Launch “opera” and go to the UCSC website at

<http://genome.ucsc.edu/index.html>

- Go to Genomes (left link on the upper links panel), and choose the genome of interest and the assembly (mouse, latest build).
- Click on the Add custom tracks link and browse to the wig file MACS generated (usually under MACS_wiggle/treat)
- Go to the genome browser, and change the histone track view to be full (under Custom tracks)
- Jump to the position on chr16: 90,217,994 - 90,224,852.

What can you say about the profile of H3K4me3 peaks? You can upload the Oct4 wig file from the previous session to compare with a transcription factor peak profile. Do you think this peak region contains one modified site or more? Keep in mind that a nucleosome core particle consists of approximately 147 bp of DNA.

We can split H3K4me3 peaks into sub-peaks using PeakAnalyzer. Launch PeakAnalyzer and choose the “split peak” utility. Upload the `H3K4me3_peaks.bed` peak file (dont forget to remove the track header line) and the `wig` file exists under the `H3K4me3_MACS/treat` folder , and set the output directory. This time we do not need the sequences so keep this option off. Now we have to set the following split parameters:

1. Separation float - This value determines when a peak will be separated into sub-peaks. This is the ratio between a valley and its neighbouring summit (the lower summit of the two). For example, if you set this height to be 0.5, two sub-peaks will be separated only if the height of the lower summit is twice the height of the valley. Keep the default value.
2. Minimum height - only sub-peaks with at least this number of tags in their summit region will be separated. Set this to be 10. After setting these parameters you can run the program.

Lets check the performance of the split procedure. Upload the “subpeak” file `H3K4me3_peaks.subpeaks.bed` that was generated into UCSC genome browser (but first remove the header line). Click on “manage custom track”, then “add custom track” and dont forget to change the view of the new track to be “full”. Go to the same region we looked at before. How many sub-peaks are found?

This is the time to check if you have received the e.mail containing the MEME results. Open the e.mail and click on the link that lead you to the html results page.

Scroll down until you see the first motif logo. We would like to know if this motif is similar to any other known motif. In order to do that we will use the “TOMTOM” program. Scroll down until you see the option “Compare to known motifs in motif databases using Tomtom”, click the “TOMTOM PSPM 1” button, and in the new page choose to compare your motif to those in the “TRANSFAC” database.

Which motif was found to be similar to your motif?

If you have brought your own data with you, this is the time to start analyzing it.

We hope you have enjoyed this tutorial. If you have questions now or in the future, you are welcome to contact us at

Mali mali@ebi.ac.uk

Christiana C.Spyrou@statslab.cam.ac.uk