ChIP-seq practical: peak detection and peak annotation

Mali Salmon-Divon
Remco Loos
Myrto Kostadima

March 2012

Introduction

The goal of this hands-on session is to perform some basic tasks in the analysis of ChIP-seq data. We will align ChIP-seq data to the mouse genome using Bowtie, then we will find immuno-enriched areas using the peak caller MACS. 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 one data set in this practical, which can be found in the ChIP-seq directory on your desktop. This directory also contains an electronic version of this document, which can be useful to copy and paste commands.

The data files are the following:

1. oct4.fastq
   This file is based on Oct4 ChIP-seq data published by Chen et al. 2008. This is a fastq file of reads that will align to a single mouse chromosome. "oct4.fastq" contains reads from the ChIP sample.

2. "gfp.fastq" contains reads from the control GFP mock-IP.

   You will have to identify peaks, annotate the enriched regions and hopefully to find the Oct4 motif.

Basic analysis: From sequences to peaks

Alignment

There are a number of competing tools for short read alignment, each with its own set of strengths, weaknesses, and caveats. Here we will try Bowtie, a widely used ultrafast, memory-efficient short read aligner.

- Open a new terminal window and go to the Desktop/ChIP-seq directory.
- Type 'bowtie' to view the different parameters. Bowtie uses indexed genome for the alignment in order to keep its memory footprint small. The indexed genome is generated using the command: `bowtie-build [genome_fasta_file] [index_name]`
  We already generated the index for the mouse genome, and placed it under the bowtie-index subdirectory. The first argument for bowtie is the basename of the index for the genome to be searched, in our case is 'mm9'. The second argument is the name of the fastq file.
  `bowtie bowtie_index/mm9 [filename.fastq] >outfile`
- Before you run bowtie, you have to know which fastq format you have. This can be Solexa (if you have an old data file), Illumina 1.3+, or Sanger

fastq files, each of which encode the quality scores differently. For more details see *http://en.wikipedia.org/wiki/FASTQ_format*
The default Bowtie behaviour is to use Sanger scores. If you have a different encoding, you'll have to specify that. The fastq files we have are of Sanger format (Can you tell why? clue: look at the first few quality lines. You can use the command `head gfp.fastq` to do this.)

- In ChIP-seq analysis (unlike other applications such as RNA-seq) we usually would like to exclude all reads that map to more than one location in the genome. In order to do that we will have to use the `-m 1` flag. Here we will get the output in Bowtie's default format, which is fairly basis and easy to understand. In practice, it is often useful to use the SAM format (option -S), which is a standard format that most of NGS analysis tools support. Now you are ready to run Bowtie

  ```
  bowtie -m 1 bowtie-index/mm9 [filename.fastq] >[outfile_name]
  ```

  Do this for both the Oct4 and the control gfp file.

  Bowtie will take 2-3 minutes to align each file. This is fast compared to other aligners, such as Maq, which sacrifice some speed to obtain higher sensitivity.

  Look at the output file, can you tell to which chromosome the reads mapped?

Finding enriched areas using MACS

MACS (*http://liulab.dfci.harvard.edu/MACS/index.html*) stands for Model based analysis of ChIP-seq. It was designed for identifying transcription factor binding sites. MACS captures the influence of genome complexity to evaluate the significance of enriched ChIP regions, and improves the spatial resolution of binding sites through combining the information of both sequencing tag position and orientation. MACS can be easily used for ChIP-Seq data alone, or with a control sample to increase specificity.

- The input for MACS can be in ELAND, BED, SAM, BAM or BOWTIE formats (you just have to set the `--format` flag).
- Type "macs14" in order to see the various parameters. Those you will have 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`.

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

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

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

`--mfold` = MACS uses the 'best' peaks to build its model. This parameter indicates how good the peaks need to be to be used (expressed as n-fold enrichment with respect to the background). While it is running, Maq will tell you how many peaks it found satisfying this criterium. If it finds very few peaks, it is a good idea to lower this parameter to get a better model. Here, we will use the default value.

`--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:
```
macs14 -t [Oct4_aligned_file] -c [gfp_aligned_file] --format=BOWTIE
--name=Oct4 --gsize=138000000 --tsize=26 --diag --wig
```

- Look at the output saturation table (`Oct4_diag.xls`). To open Excel files, right-click on them, choose Open with and select OpenOffice. Do 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.

Ensembl genome browser

It is often useful to look at your data in a genome browser, like Ensembl or the UCSC browser. This will allow you to get a 'feel' for the data, as well as detecting abnormalities and problems. Also, it may give you ideas for further analyses.

Launch an internet bowser and go to the Ensembl website at
http://www.ensembl.org/index.html

- Choose the genome of interest (that is, mouse) on the left side of the page.

- Click on the `Manage your data` link on the left, then choose `Attach remote file`. One option is to browse to the `wig` file MACS generated (usually under MACS_wiggle/treat) and to upload it to Ensembl (option `Upload data`). However, since wig files describing sequencing signal are very big, they would take a long time to upload, and the browsing process will be very slow. In addition, Ensembl has a size limit for uploaded files.

As a better alternative, wig files can be converted to an indexed binary format and put into a web accessible server (http, https, or ftp) instead on the Ensembl server. This makes all the browsing process much faster. Detailed instructions of how to generate bigWig from wig type files can be found at:
http://genome.ucsc.edu/goldenPath/help/bigWig.html.

We have generated bigWig files in advance for you to upload to the Ensembl browser. They are at the following URL: `http://www.ebi.ac.uk/~remco/ChIP-Seq_course/Oct4.bw`. To visualise the data:

- Paste the location above in the field `File URL`
- Choose data format `Bigwig`
- Choose some informative name and in the next window choose the colour of your preference.
- Click save and close the window to return to the genome browser.

Repeat the process for the control sample, located at `http://www.ebi.ac.uk/~remco/ChIP-Seq_course/gfp.bw`.

Go to a region on chromosome 1 (e.g. 1:34823162-35323161), and zoom in and out to view the signal and peak regions. Be aware that the data scale automatically, so bigger-looking peaks need not actually be bigger. Always look at the values on the left hand side axis.

What can you say about the profile of Oct4 peaks? Compare it with H3K4me3 histone modification `wig` file we have generated at `http://www.ebi.ac.uk/~remco/ChIP-Seq_course/H3K4me3.bw`.

Jump to 1:36066594-36079728 for a sample peak. Do you think H3K4me3 peaks regions contain one or more modification sites? What about Oct4?

Annotation: From peaks to biological interpretation

Peak Annotation

In order to biologically interpret the results of ChIP-seq experiments, it is usually recommended to look at the genes and other annotated elements that are located in proximity to the identified enriched regions. This can be easily done using PeakAnalyzer (available from *http://www.ebi.ac.uk/bertone/software*).

- Go to the PeakAnalyzer directory and launch the program by typing `java -jar PeakAnalyzer.jar` in the terminal.

- The first window allows you to choose between split application (which we will try next) and peak annotation.
  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 don't have 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 running, you will have the option of generating plots. (You can do this if R is installed on your computer. Otherwise, if you dont 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.

This list of closest downstream genes can be the basis of further analysis. For instance, you could look at the Gene Ontology terms associated with these genes to get an idea of the biological processes which may be affected. Web-based tools like DAVID (*http://david.abcc.ncifcrf.gov*) or GOstat (*http://gostat.wehi.edu.au*) take a list of genes and return the enriched GO categories.

Motif analysis

It is often interesting to find out whether we can associate identified the binding sites with a sequence pattern or motif. 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 wider than the actual binding sites. Sub-dividing the enriched areas by accurately partitioning enriched loci into a finer-resolution set of individual binding sites, and fetching sequences from the summit region where binding motifs are most likely to appear enhances the quality of the motif analysis. Sub-peak summit 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. 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.
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 set some parameters for splitting the peaks.

  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 5.
     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 internet bowser and go to the MEME website at
http://meme.sdsc.edu/meme/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), or just paste in the sequences.
- the number of motifs we expect to find (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.

You will receive the results by e-mail. This usually doesn't take more than a few minutes. When you receive the e-mail containing the MEME results:

- Open the e-mail and click on the link that leads 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 `Submit this motif to`. Click the `TOMTOM` button to compare to known motifs in motif databases, 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 there is time left, and you have brought your own data, this is the time to start analyzing it.

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

Remco *remco@ebi.ac.uk*
Myrto *kostadim@ebi.ac.uk*

Bibliography

Chen, X *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell.* Jun 13;133(6):1106-17 (2008).