

Next Generation Sequencing Workshop

ChIP-seq Hands-on Exercise

Remco Loos, EMBL-EBI (remco@ebi.ac.uk)

Myrto Kostadima, EMBL-EBI (kostadim@ebi.ac.uk)

General information

The following standard icons are used in the hands-on exercises to help you locating:



Important Information



General information / notes



Follow the following steps



Questions to be answered



Warning – PLEASE take care and read carefully



Optional Bonus exercise



Optional Bonus exercise for a champion

Resources used

Bowtie: <http://bowtie-bio.sourceforge.net/index.shtml>

Samtools: <http://samtools.sourceforge.net/>

BEDTools: <http://code.google.com/p/bedtools/>

UCSC tools: <http://hgdownload.cse.ucsc.edu/admin/exe/>

IGV genome browser: <http://www.broadinstitute.org/igv/>

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

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

MEME: <http://meme.sdsc.edu/meme/cgi-bin/meme.cgi>

TOMTOM: <http://meme.sdsc.edu/meme/cgi-bin/tomtom.cgi>

Additional resources:

Ensembl: <http://www.ensembl.org>

DAVID: <http://david.abcc.ncifcrf.gov>

GOstat: <http://meme.sdsc.edu/meme/cgi-bin/tomtom.cgi>

Original Data from: <http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-11431>

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

Introduction



The goal of this hands-on session is to perform some basic tasks in the analysis of ChIP-seq data. The first step includes an unspliced alignment for a small subset of raw reads. We will align raw sequencing data to the mouse genome using *Bowtie* and then we will manipulate the SAM output in order to visualize the alignment on the *IGV browser*. Then based on these aligned reads we will find immuno-enriched areas using the peak caller *MACS*. We will then perform functional annotation and motif analysis on the predicted binding regions.

Prepare the environment



We will use one data set in this practical, which can be found in the ChIP-seq directory on your desktop. Throughout this practical we will try to identify potential transcription factor binding sites of Oct4 in mouse embryonic stem cells.



Open the Terminal.

First, go to the right folder, where the data are stored.

```
cd ~/Desktop/ChIP-seq
```



The .fastq file that we will align is called Oct4.fastq. This file is based on Oct4 ChIP-seq data published by Chen et al. (2008). We will align these reads to the mouse chromosome.

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.



Bowtie has a number of parameters in order to perform the alignment. To view them all type

```
bowtie --help
```

Bowtie uses indexed genome for the alignment in order to keep its memory footprint small. Because of time constraints we will build the index only for one chromosome of the mouse genome. For this we need the chromosome sequence in fasta format. This is stored in a file named `mm10`, under the subdirectory `bowtie_index`.

The indexed chromosome is generated using the command:

```
bowtie-build bowtie_index/mm10.fa bowtie_index/mm10
```

This command will output 6 files that constitute the index. These files that have the prefix `mm10` are stored in the `bowtie_index` subdirectory. To view if they files have been successfully created type:

```
ls -l bowtie_index
```



Now that the genome is indexed we can move on to the actual alignment. The first two arguments make sure that the output is in SAM format using the `'-S'` parameter and that Bowtie reports only uniquely mapped reads using the `'-m 1'` option. The following argument for bowtie is the basename of the index for the genome to be searched; in our case is `mm10`. The last argument is the name of the fastq file.



Align the Oct4 reads using Bowtie:

```
bowtie -m 1 -S bowtie_index/mm10 Oct4.fastq \  
> Oct4.sam
```

The above command outputs the alignment in SAM format and stores them in the file `Oct4.sam`.



In general before you run *Bowtie*, you have to know which fastq format you have. The available fastq formats in bowtie are:

--phred33-quals input quals are Phred+33 (default)
--phred64-quals input quals are Phred+64 (same as --solexa1.3-quals)
--solexa-quals input quals are from GA Pipeline ver. < 1.3
--solexa1.3-quals input quals are from GA Pipeline ver. >= 1.3
--integer-quals qualities are given as space-separated integers (not ASCII)

The fastq files we are working on is of Sanger format (Phred+33), which is the default for *Bowtie*.

Bowtie will take 2-3 minutes to align the file. This is fast compared to other aligners that sacrifice some speed to obtain higher sensitivity.



Look at the SAM format by typing:

```
head -n 10 Oct4.sam
```



Can you distinguish between the header of the SAM format and the actual alignments?

What kind of information does the header provide you with?

To which chromosome are the reads mapped?

Manipulate SAM output



SAM files are rather big and when dealing with a high volume of NGS data, storage space can become an issue. We can convert SAM to BAM files (their binary equivalent files that are not human readable) that occupy much less space.



Convert SAM to BAM using *samtools* and store the output in the file `Oct4.bam`. You have to instruct *samtools* that the input is in SAM

format (-S), the output should be in BAM format (-b) and that you want the output to be stored in the file specified by the -o option:

```
samtools view -bSo Oct4.bam Oct4.sam
```

Visualize alignments in IGV



It is often instructive to look at your data in a genome browser. Here, we use IGV, a stand-alone browser, which has the advantage of being installed locally and providing fast access. Web-based genome browsers, like Ensembl or the UCSC browser, are slower, but provide more functionality. They do not only allow for more polished and flexible visualisation, but also provide easy access to a wealth of annotations and external data sources. This makes it straightforward to relate your data with information about repeat regions, known genes, epigenetic features or areas of cross-species conservation, to name just a few. As such, they are useful tools for exploratory analysis.

Visualisation will allow you to get a 'feel' for the data, as well as detecting abnormalities and problems. Also, exploring the data in such a way may give you ideas for further analyses.

IGV is a stand-alone genome browser. Please check their website (<http://www.broadinstitute.org/igv/>) for all the formats that IGV can display. For our visualization purposes we will use the BAM and bigWig formats.



When uploading a BAM file into the genome browser, the browser will look for the index of the BAM file in the same folder where the BAM files is. The index file should have the same name as the BAM file and the suffix .bai. Finally, to create the index of a BAM file you need to make sure that the file is sorted according to chromosomal coordinates.



Sort alignments according to chromosome position and store the result in the file with the prefix `Oct4.sorted`:

```
samtools sort Oct4.bam Oct4.sorted
```

Index the sorted file.

```
samtools index Oct4.sorted.bam
```

The indexing will create a file called `Oct4.sorted.bam.bai`. Note that you don't have to specify the name of the index file when running *samtools*.



Another way to visualize the alignments is to convert the BAM file into a bigWig file. The bigWig format is for display of dense, continuous data and the data will be displayed as a graph. The resulting bigWig files are in an indexed binary format.



The BAM to bigWig conversion takes place in two steps. Firstly, we convert the BAM file into a bedgraph, called `Oct4.bedgraph`, using the tool *genomeCoverageBed* from *BEDTools*:

```
genomeCoverageBed -bg -ibam Oct4.sorted.bam \  
-g bowtie_index/mouse.mm10.genome > Oct4.bedgraph
```

Then we convert the bedgraph into a binary graph, called `Oct4.bw`, using the tool *bedGraphToBigWig* from the *UCSC tools*:

```
bedGraphToBigWig Oct4.bedgraph \  
bowtie_index/mouse.mm10.genome Oct4.bw
```



Both of the commands above take as input a file called `mouse.mm10.genome` that is stored under the subdirectory `bowtie_index`. These genome files are tab-delimited and describe the size of the chromosomes for the organism of interest. When using the

UCSC Genome Browser, Ensembl, or Galaxy, you typically indicate which species/genome build you are working. The way you do this for *BEDTools* is to create a “genome” file, which simply lists the names of the chromosomes (or scaffolds, etc.) and their size (in basepairs).

BEDTools includes pre-defined genome files for human and mouse in the **/genomes** directory included in the *BEDTools* distribution.



Now we will load the data into the *IGV* browser for visualization. In order to launch *IGV* type the following on your terminal:

igv.sh &

On the top left of your screen choose from the drop down menu `Mus musculus (mm10)`. Then in order to load the desired files go to:

File > Load from File

On the pop up window navigate to `Desktop > ChIP-seq` folder and select the file `Oct4.sorted.bam`.

Repeat these steps in order to load `Oct4.bw` as well.

Select `chr1` from the drop down menu on the top left. Right click on the name of `Oct4.bw` and choose `Maximum` under the `Windowing Function`. Right click again and select `Autoscale`.

In order to see the aligned reads of the BAM file, you need to zoom in to a specific region.



Look for gene `Lemd1` in the search box.

Can you see an `Oct4` binding site in the `Lemd1` gene?

Using the ‘+’ button on the top right zoom in more to see the details of the alignment.



What is the main difference between the visualization of BAM and bigWig files?

What do you think the different colors mean?

Alignment of control .fastq file



In the ChIP-seq folder you will find another .fastq file called `gfp.fastq`. Follow the steps described above for this dataset in order to align the control reads to the mouse genome as well.

Finding enriched areas using MACS



MACS 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.



Consult the MACS help file to see the options and parameters.

`macs --help`



The input for MACS can be in ELAND, BED, SAM, BAM or BOWTIE formats (you just have to set the `--format` flag). Options that 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)
- --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, 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 aligned bam file] \  
      -c [gfp aligned bam file] \  
      --format=BAM --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 OpenOffice Calc. On the pop up window select 'Separated by tab' only.



Do 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.



MACS generates its peak files in a file format called bed file. This is a simple text format containing genomic locations, specified by chromosome, begin and end positions, and some more optional information.

See <http://genome.ucsc.edu/FAQ/FAQformat.html#format1> for details.

Bed files can also be uploaded to IGV or other genome browsers.



Optional Bonus Section

Try uploading the peak file generated by MACS to IGV. Find the first peak in the file (use the **head** command to view the beginning of the bed file), and see if the peak looks convincing to you.

Annotation: From peaks to biological interpretation



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.



Go to the PeakAnalyzer directory and launch the program by typing

```
java -jar PeakAnalyzer.jar &
```

The first window allows you to choose between the 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, by pressing **Generate plots** (You can only do this if R is installed on your computer, as is the case here. Otherwise, if you don't 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 (contained in the file 'Oct4_peaks.ndg.bed') 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 that 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.



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_noheader.bed' (which we have generated for you under Desktop/Chlp-seq) and the wig file generated by MACS, which is under 'Oct4_MACS_wiggle/treat', check the option to **Fetch subpeak sequences** and click **Next**.

In the next window you have to set some parameters for splitting the peaks.

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.

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.



Open a web browser, go to the MEME website at <http://meme.sdsc.edu/meme/>, and choose the 'MEME' tool. 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.



Start Search. You will receive the results by e-mail. This usually doesn't take more than a few minutes.



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. We will use TOMTOM for this. Scroll down until you see the option **Submit this motif to**. Click the TOMTOM button to compare to known motifs in motif databases, and on the new page choose to compare your motif to those in the JASPAR and UniPROBE database.



Which motif was found to be the most similar to your motif?



CONGRATULATIONS! You've made it to the end of the practical.

Hope you enjoyed it!

Don't hesitate to ask any questions and feel free to contact us any time (email addresses on the front page).